Exploring Vision-Language Model Explanations for Classifying Depressive Symptoms

Anonymous ACL submission

Abstract

Over the past years, memes have evolved from being exclusively a medium of humorous exchanges to one that allows users to express a range of emotions freely and easily. With the ever-growing utilization of memes for expressing depressive symptoms, we conduct a study on identifying depressive symptoms exhibited by memes shared by users of online social media platforms. We introduce RESTOR**Ex**as a vital resource for detecting depressive symptoms in memes on social media through the use of explanations generated by state-of-the-art VLMs and a set of expert-annotated gold-label explanations for these memes. We provide an extensive comparative survey of four VLMs annotated by domain experts in Computer Science and Psychology. Further, we validate the generated content through a therapeutic utility evaluation led by an expert in psychology.

1 Introduction

In the digital age, have emerged as a pervasive form of content on online social media platforms, underscoring their escalating significance. In 2020 alone, Instagram, a popular social media app, reported at least one million posts shared every day mentioning the word "meme" (Brown, 2022). Notably, memes may be an attractive form of communication for individuals experiencing depressive symptoms, providing them with a nuanced outlet for self-expression that may be less direct and confrontational. Akram et al. (2020) observe that sharing humor-intentional depressive memes may be beneficial for some individuals since a humorous take on a negative experience may give the perception of peer support among viewers who have experienced something similar. Considering the growing impact of mental health problems on economies worldwide (The Lancet Global Health, 2020), the proactive identification of depressive symptoms in users of online social media platforms serves as an important research question.

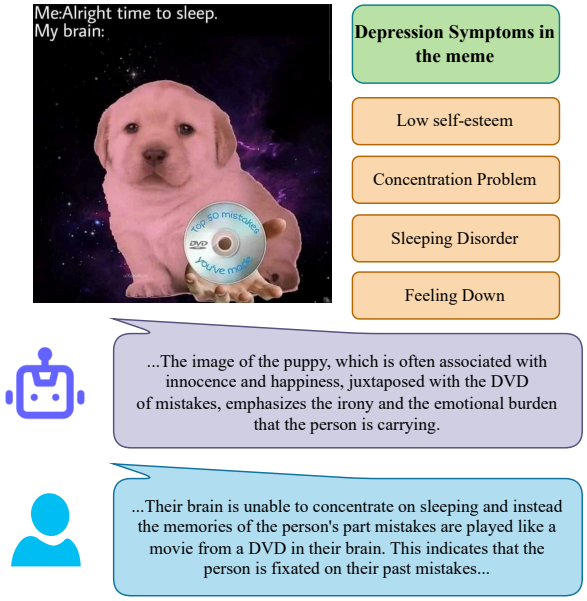


Figure 1: An example of a depressive meme from RESTOR**Ex** with the VLM generated explanation and expert-written explanation

Recognising this importance of memes in online social media today, work on meme analysis has gained traction starting with the Facebook Hateful Memes Challenge (Kiela et al., 2020). However, research in the are has been mostly focused on aspects such as harmfulness (Pramanick et al., 2021a) and cyber-bullying (Jha et al., 2024). In this work, we work on a relatively understudied area in meme analysis and apply it to the mental health domain. We work to identify seven fine-grained depressive symptoms - Feeling Down, Lack of Interest, Eating Disorder, Sleeping Disorder, Concentration Problem, Low Self Esteem, and Self Harm - based on the clinically established 9-scale Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001).

The task involves multi-label classification for the seven identified symptoms in the multimodal memes as proposed by Yadav et al. (2023). We derive a new dataset, RESTOR**Ex**, consisting of a

variety of explanations generated by state of the art Vision Language Models and supplemented with expert-annotated explanations for the memes.

The primary focus of our work is to examine the application of Vision Language Models (VLMs) in understanding complex multimodal data like memes in the context of mental health. We focus on utilising VLMs as an intermediate step by generating explanations that succinctly explain the content and meaning of the meme image. This analysis of meme understanding is especially important as memes involve a mix of sarcasm, irony and other types of figurative speech.

For example, in Figure 1, the meme has the caption *"Me: Alright time to sleep."*, *"My brain:"*, with a picture of a puppy holding a DVD that says, *"Top 50 mistakes you've made"*. The human annotator emphasises on the meme author's fixation on their past mistakes, correctly identifying the *'Low Self-Esteem'* symptom. Further, the VLM-generated explanation, recognising the image of the sad expression of the puppy (often used in a positive context) juxtaposed with negative emotions relating to the DVD, understands the irony of the situation to correctly assign the *'Feeling Down'* label to this meme. Overall, through a multi-dimensional understanding of the amalgamation of various figurative speeches in the meme, the ideal explanation should capture all four depressive symptoms expressed by the meme author – *Low Self-Esteem*, *Concentration Problem*, *Sleeping Disorder*, and *Feeling Down*.

In this work, we investigate the utility of state-of-the-art multimodal models by pushing their reasoning skills in understanding the complex contexts within meme content. We focus on answering questions about the strengths and weaknesses of each model in our study to provide accurate judgment about the current reasoning capabilities of these models through rigorous and comprehensive analysis. We supplement our analysis with expert-led Cognitive Analytic Therapy (CAT) (Parry et al., 2021) evaluation to provide a richer understanding of generated content beyond the typical linguistic view on these topics.

Our contributions are summarized as follows-

- **Novel Dataset:** Introducing the RESTOREx dataset, with multiple VLM-generated explanations and expert annotated gold label explanations.
- **In-depth Human Evaluation:** Conducted a human evaluation with domain experts to address key research questions regarding VLM-generated explanations across various aspects (fluency, relevance, figurative understanding, and persuasiveness) in the context of depressive meme understanding.
- **Therapeutic Utility Evaluation:** Present a Cognitive Analytic Therapy Competencies based analysis of VLM generated content conducted by a domain expert in Psychology.

2 Related Work

This section discusses relevant studies on meme analysis, the lack of discussion of the mental health domain in current work, provides background on existing studies on depression identification using social media data and discusses the scope of current work on the application of LLMs and VLMs.

2.1 Meme Analysis

In recent years, efforts towards the analysis of memes and subsequent datasets in the area have been focussed mainly on aspects such as hatefulness (Kiela et al., 2021), harmfulness (Pramanick et al., 2021b; Sharma et al., 2023), propaganda (Dimitrov et al., 2021), cyberbullying (Maity et al., 2022), trolling (Suryawanshi et al., 2020), and emotions (Sharma et al., 2020; Ramamoorthy et al., 2022; Mishra et al., 2023). The work by Yadav et al. (2023) extended meme research towards the mental health domain with the release of the RESTORE dataset for detecting fine-grain depressive symptoms in memes.

Recently, works such as those by Agarwal et al. (2024) and Zhong and Baghel (2024) have explored the application of explanations generated by VLMs and LLMs for explaining memes in various contexts. However, rigorous analysis of the generated content remains a challenge that we attempt to resolve with this work.

2.2 Depressive Symptom Analysis

Previous work on depressive symptom analysis has focussed mostly on textual data such as Tweets and Reddit posts, with limited work on multimodal content such as memes (Yadav et al., 2023). Works such as Yadav et al. (2021, 2018); Benton et al. (2017) have focussed on analyzing depressive symptoms by collecting data from online social media platforms. Works such as Yates et al. (2017) have served as important benchmarks for analyzing textual depressive data. This work aims to extend depression analysis to the multimodal meme analysis setting.

Dataset Split		Symptom								Total
		LOI	FD	ED	SD	LSE	CP	SH	LOE	
Train	RESTORE	471	2085	1939	1562	855	595	1516	122	8814
	RESTOREx	441	1781	1714	997	703	348	1357	0	7096
Test	RESTORE	97	294	98	99	128	100	106	95	662
	RESTOREx	66	219	85	78	114	73	82	0	520
Validation	RESTORE	45	195	49	45	85	42	61	54	361
	RESTOREx	37	130	54	36	79	28	54	0	310

Table 1: Data distribution in RESTOREx, compared to RESTORE

3 Dataset

We introduce RESTOREx, a dataset for the multi-label classification of seven fine-grained PHQ-9 depressive symptom categories – *Feeling Down, Lack of Interest, Eating Disorder, Sleeping Disorder, Concentration Problem, Low Self Esteem, and Self Harm*. We provide expert annotated human explanations and conduct extensive expert analysis of all generated content. Our dataset is derived from the original RESTORE dataset proposed by Yadav et al. (2023) by introducing significant additions and modifications.

3.1 Dataset Cleaning and Re-annotation

Upon careful examination of the existing dataset, we discovered inconsistencies in the original annotations of the RESTORE dataset and undertook the task of re-annotating the dataset in three ways:

1. Re-annotated the test and validation sets of the dataset with additional annotators
2. Removed the *'Lack of Energy'* label due to the lack of accurate training labels
3. Removed non-meme images such as quotes from the dataset.

The re-annotation of the test and validation dataset was undertaken by two annotators[1] who were trained to identify fine-grained depressive symptom labels based upon the definitions laid out originally by Yadav et al. (2023). The inter-annotator agreement was captured as the Krippendorff's Alpha coefficient (krippendorff, 2004), a metric widely used for measuring reliability in annotations for multi-label tasks. The coefficient is obtained as 0.833 using MASI distance (Passonneau, 2006) as the distance function, representing strong-agreement between the annotators.

As a result of automatic curation, the original dataset had erroneous training samples for the label *Lack of Energy*. The training set consisted exclusively of automatically identified training samples for this label which were inconsistent with the definition laid out by the authors. To this end, we drop this label from dataset and our analysis as we are unable to train models to accurately detect the category due to a lack of valid training examples. Further issues were observed in the dataset with the inclusion of text-only images or quotes which do not fit the definition of 'memes'. We manually removed all such examples leading to a decrease in about 20% of the dataset size. The final dataset distribution is shown in Table 1.

3.2 Introduction of Ground-truth Explanations

As a part of our work on extending the dataset, we introduce ground-truth human-written explanations for the test dataset. This ground-truth explanation serves as a resource for future research to compare model-generated explanations with a human-annotated dataset. The annotations are undertaken by psychology and computer science experts to present a representative mix of background information for this interdisciplinary task.

3.3 Annotation Guidelines

We provide the annotation guidelines based on the PHQ-9 depression questionnaire. The group of 3 independent domain-expert annotators used an online annotation tool for the annotation process. The annotators are assigned two annotation tasks:

1. For a given meme, the annotators are required to mark as per requirements one or more of the following seven depressive symptoms conveyed in the meme – Feeling Down, Lack of Interest, Eating Disorder, Sleeping Disorder, Concentration Problem, Low Self Esteem, and Self Harm. In case the annotators do not find any relevant symptom associated with the meme, they are instructed to leave the field as blank.
2. For each meme, the annotators are required to provide the human annotated gold label explanations in the provided text field.

3.4 Incorporating VLM Explanations

We propose to extend the original RESTORE dataset with VLM generated explanations as an additional text modality. These detail-rich explanations serve as a textual description of the cross-modal information in a meme image. As the pre-trained

[1]The annotators included a research assistant and a research volunteer who is a native English speaker. Both the annotators possess reading, writing, and speaking fluency in English.

Vision-Language Models used in our experiments are based upon recent pre-trained Large Language Models with over 7B parameters, their understanding of memes encompasses various aspects of commonsense reasoning, sarcasm and a nuanced thought process. These detailed explanations are deemed useful in the classification problem for both machine learning models as well as human annotators, with their exceptional understanding of the reasons behind why a meme author may be exhibiting symptoms of depression.

Model Selection For a representative variety in the generated content, we put careful consideration into the choice of our models. We take particular effort to utilise a mix of open-source and closed-source models. We also ensure a variety in the base foundational LLMs that these VLMs are built upon. Based on these criterias, we generate explanations from four popular VLMs:

1. LLaVA 1.5 (Liu et al., 2023), trained on top of Llama2-7B (Touvron et al., 2023)
2. LLaVA-NeXT (Liu et al., 2024), trained on top of Mistral-7B (Jiang et al., 2023)
3. MiniCPM-V (Hu et al., 2024), trained on top of SigLip-400M (Zhai et al., 2023) and Qwen2-7B (Yang et al., 2024)
4. GPT-4o (OpenAI et al., 2024), a closed-source model

3.5 Analysis of Generated Explanations

A well-documented issue in the content generated by LLMs is their tendency to hallucinate (Huang et al., 2023; Rawte et al., 2023). Hence, in the sensitive domain of mental health disorders, it is pivotal to ensure the legitimacy of explanations as determined by domain experts. To achieve this, we conduct a comprehensive comparative survey of the four generated explanations comprising of 24 questions answered by a mix of experts in the mental health domain and the computer science domain. The questions focus on fluency, relevance, accuracy and informativeness of the explanations in order to judge their quality. The domain experts further annotate a set of memes with gold-label human explanations. The annotators are not presented with model names corresponding to the explanations in an effort to extract unbiased opinions about the explanations. In addition, they are provided detailed guidelines on how to answer these questions as described in the Appendix A. Detailed analysis of our findings from this survey is described in section 5.5. The inclusion of domain experts in the evaluation of the generated explanations allows us to be confident in their effectiveness for our task.

The entire reannotated dataset, including VLM generated explanations and gold-label human explanations would be made available for researchers.

4 Experiments

In this section, we present details on our experiments utilising the diverse set of VLM generated and human annotated experiments for the mutli-label classification task of depressive symptom identification. We perform these experiments to provide quantifiable comparisons between the explanations provided by the chosen models. We also provide a comparison to the results of experiments on directly prompting the model for the classification task in order to understand the extent of information loss while using the model generated explanation as a proxy for the meme image. All experiments leverage the strong in-context learning capabilities of the GPT-4o model. We use the commercial OpenAI API for access to the model.

4.1 Zero-shot VLM Experiments

We prompt the GPT-4o model for the zero-shot classification task using a combination of system and user input prompts. A relevant system prompt is chosen manually and instructs the model to provide a list of one or more identified depressive symptoms based on the explanation/image data. The user prompt provides the meme specific data in the form of the image or explanation and asks the model to present it's output in the form of a json string containing a list of predicted symptoms.

4.2 One-shot VLM Experiments

For the one-shot prompting setting, the details in the zero-shot prompting experiment are followed with the addition of a randomly chosen input sample with it's correct corresponding labels provided as an input example to the model. This experiment is designed to align the model's output to the dataset characteristics which it can observe from the provided sample.

5 Results and Analysis

In this section, we discuss i) results and observations the zero-shot and few-shot experiments described in Section 4, ii) the domain expert analysis

	Experiment	FD	LOI	SH	ED	LSE	CP	SD	Macro	Micro	Weighted
Zero-shot	Image Only	0.66	0.39	0.66	0.79	0.52	0.53	0.83	0.63	0.61	0.63
	LLaVA 1.5 Explanations	0.55	0.37	0.34	0.48	0.38	0.47	0.67	0.47	0.48	0.48
	LLaVA-NeXT Explanations	0.62	0.43	0.60	0.71	0.46	0.53	0.79	0.59	0.57	0.59
	MiniCPM-V Explanations	0.62	0.40	0.53	0.73	0.51	0.53	0.78	0.59	0.58	0.59
	GPT-4o Explanations	0.63	0.39	0.58	0.73	0.46	0.54	0.86	0.60	0.57	0.60
One-shot	Image Only	**0.68**	**0.55**	**0.71**	**0.81**	**0.54**	0.57	0.85	**0.67**	**0.66**	**0.67**
	LLaVA 1.5 Explanations	0.47	0.31	0.43	0.44	0.35	0.50	0.62	0.44	0.44	0.45
	LLaVA-NeXT Explanations	0.64	0.49	0.64	0.69	0.47	**0.58**	0.80	0.62	0.60	0.62
	MiniCPM-V Explanations	0.64	0.43	0.55	0.73	0.48	0.55	0.77	0.59	0.59	0.60
	GPT-4o Explanations	0.60	0.41	0.61	0.73	0.46	0.54	**0.88**	0.61	0.58	0.60

Table 2: Class-wise results of the zero-shot and one-shot experiments with GPT-4o. All values are F-1 scores.

of the generated explanations, and iii) our observations from the comparative human study of the memes.

5.1 Experimental Results

As we can note in Table 2, our best results are obtained using the meme image as the input to the model. This may be expected as the breakdown of the image into VLM generated explanations leads to some reduction in the information available to the model. The success of the GPT-4o model in this in-context learning setup indicates it's prowess in understanding the figurative meaning of the meme. When we use the VLM generated explanation as proxy for the meme image, we notice a reduction in model performance due to the reduction in information captured. However, as we can observe, the decrement in performance is marginal for most depression categories for the LLaVA-NeXT, MiniCPM-V and GPT-4o generated explanations. This indicates that these explanations are able to retain much of the information required for ascertaining the corresponding depressive symptoms associated with a meme. This speaks to these models' capability of capturing the intended figurative meanings present in the meme image and presenting them in a textual format. These results also indicate how significant effort in modelling is required to effectively utilise the generated explanations.

5.2 Domain Expert Perspective

In order to evaluate the utility of the LLMs, it is crucial to understand the rationale for using humor and creating memes in the case of Depression.

1. **Humor as a Defense Mechanism:** Humor can be used to create emotional distance and to help people process feelings. This enables them to cope with stress, Depression, and anxiety, especially during important developmental transitions (Sarink and García-Montes, 2023; Erickson and Feldstein, 2006). While positive humor (affiliative, self-enhancing) is linked to stable psychological adjustment, negative humor (especially the self-defeating humor seen in memes) is associated with higher depressive symptoms (Erickson and Feldstein, 2006). Therefore, though it helps build social relationships and navigate depressive symptoms, it harms self-esteem in the process (Erickson and Feldstein, 2006). Specific to memes, humor provides an avenue for emotional catharsis by channeling dark or uncomfortable topics toward humor (Sarink and García-Montes, 2023).

2. **The Psychology Underlying Memes:** Memes allow people to express complex emotions surrounding Depression and anxiety in a relatable manner through their hypernarrativity (Wagener, 2021). They provide a platform for communal discourse and validation for individuals in similar situations by supporting a larger cultural framework to help process individual experiences (Wagener, 2021).

5.3 Depressive Meme Analysis using Cognitive Analytic Therapy

The Cognitive Analytic Therapy (CAT) Competencies (Parry et al., 2021) have been adapted to evaluate the LLMs' performance when interpreting memes on mental health themes. It is highly applicable here due to its structured approach to categorising and assessing both larger psychological patterns and more subtle emotional nuances and risk assessments. This grading can help us understand how different LLMs and their interpretation of memes are applicable in therapeutic contexts.

	Criteria	Description
(1)	Knowledge and Understanding of Mental Health Problems	Grasps mental health issues like depression and anxiety
(2)	Knowledge of Depression	Able to recognize, analyze, and describe depressive symptoms
(3)	Cultural Competence	Understands cultural or subcultural references (e.g., anime culture, Avenger's Infinity War) and the implications of their use
(4)	Ability to Engage the Subject/Client (Meme Creator)	Engages with the emotional content of the meme creator (the client in this case)
(5)	Ability to Foster Therapeutic Alliance	Understands the perspective of the subject of the meme
(6)	Ability to Work with Emotional Content	Navigates themes surrounding depression effectively
(7)	Ability to Assess and Manage Risk of Self-Harm	Recognizes and proactively addresses risk factors relating to self harm or suicide
(8)	Knowledge of Ethical Guidelines	Respects ethical considerations in the discussion of sensitive topics

Table 3: Criteria used for CAT evaluation

Understandably, due to the differing application of the CAT, selected criteria were not included, and the adapted criteria used here were derived primarily from the "General Therapeutic Competencies" section. This means that this grading is not an entirely accurate representation of the CAT and that specific grading criteria for LLMs need further development using this as a potential foundation. The criteria used for evaluation are described in Table 3. Each model was scored out of a total of eight points based on these criteria.

LLaVA 1.5 Score: 1/8 Failed to engage with mental health themes and lacked emotional or psychological insights. While it recognised some elements, it was unable to interpret deeper.

LLaVA-NeXT Score: 4/8 Showed moderate cultural competence and understanding of mental health but struggled to form connections. Lacked therapeutic alignment and risk assessment.

MiniCPM-V Score: 5/8 Demonstrated a fair understanding of mental health, engaging with depressive themes and isolation. Was emotionally sensitive and culturally aware but wasn't as accurate with risk assessment.

GPT-4o Score: 7/8 Demonstrated empathy and emotional depth by consistently recognising and explaining mental health themes like depression and isolation as well as their implications for the subject. While it recognised themes surrounding suicide, it could have benefited from a more explicit and urgent risk assessment.

Based on the CAT evaluations, the GPT-4o model outperformed the others in both its consistency and depth of understanding. MiniCPM-V and LLaVA-NeXT showed some ability to capture the required depth, while LLaVA 1.5 was highly inconsistent.

5.4 Depressive Meme Analysis using Visual Cues

Several themes of visual cues were either underexplored or missed entirely by the models. These revolved around subtle features like nuanced facial expressions, eye contact, and cultural experiences.

- **Facial Expressions and Emotional Masks:** The models generally could not recognize masked, more profound emotional struggles. For instance, in the "*Hide the Pain Harold*" meme shown in Figure 2(a), the forced smile was not interpreted as concealing emotional pain. Similarly, in another meme, the subject's change in eye contact, which could signify guilt, was missed. Recognizing such emotional masking is important, as it can often be used as a coping mechanism for depression.

Model	Fluency	Relevance	Figurative Meaning	Persuasiveness
LLaVA 1.5	4.33	2.27	35.3%	0.28
LLaVA-NeXT	4.41	3.80	84.3%	0.59
MiniCPM-V	4.14	3.76	84.3%	0.69
GPT-4o	4.84	4.53	86.3%	0.94

Table 4: Results of the human analysis study comparing generated explanations

- **Cultural References:** The models also missed specific cultural references and their emotional impact within the meme context. For example, references like Peter Parker's death in *Avenger's Infinity War* in Figure 2(b), and King of the Hill in Figure 2(c) were either misunderstood or referred to incorrectly, such as mistaking it for *The Simpsons*. The models also overlooked subtleties in social interactions, like the awkwardness of opening a chat while someone else is typing. These cultural details are essential for capturing the full emotional and social context of memes.

5.5 Human Analysis of Explanations

This section describes the our observations from our comparative survey where we focus on five research questions in order to comprehensively judge the quality of VLM generated explanations. A numeric breakdown of our results from this study are shown in Table 4. An example meme, complete with the four model explanations and human-annotated explanations, is provided in Figure 3.

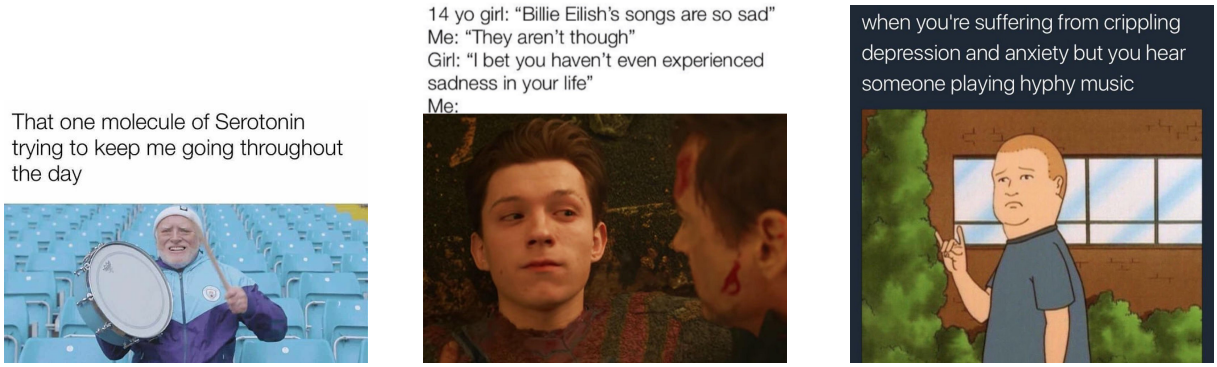
RQ1: Are VLMs fluent? Yes. As is evident from their average fluency scores, all four models are able to formulate their explanation in fluent English sentences. However, they do rarely introduce grammatical mistakes which do not meaningfully impact it's comprehensiveness to human readers. Of the four models, GPT-4o leads with it's fluency and clarity of generated content.

RQ2: Are generated explanations relevant? Mostly. The models range in relevance from 2.28 for the LLaVA 1.5 to 4.53 for the GPT-4o model. A score close to 2 for the LLaVA 1.5 model indicates that the explanations are mostly irrelevant to the meme content and do not discuss their relation to the depressive symptoms. The LLaVA-NeXT and MiniCPM-V models have a score of 3.80 and 3.75 indicating that these explanations are descriptive about the meme content and describe the depressive symptoms being discussed in the meme. However, they do not go deep into relating all aspects of the meme with depressive symptoms as achieved by the GPT-4o model with a relevance score of 4.54.

RQ3: Do VLMs capture figurative meaning? Yes. With the exception of LLaVA 1.5 the other generative models are able to capture the figurative speech that the meme authors is trying to convey with all four models capturing this figurative speech over 84% of times. For example, as shown in Figure 3, the LLaVA-NeXT correctly explains the irony in the presence of a 'puppy', often associated with innocence, and the presence of the 'DVD of mistakes' signifying the emotional burden of a person. The annotators point out that the models, especially the GPT-4o model, tend to over-explain and in turn bring out irrelevant metaphors in some cases. This phenomenon of over-explanation by the models merits further analysis in future work.

RQ4: Are VLMs persuasive? The persuasion score is calculated as a weighted average of persuasion scores for each example where the persuasion score is given in 0.25 increments with the least persuasive explanation receiving a score of 0.25 and the most persuasive one receiving a score of 1. As seen in Table 4 , we observe that GPT-4o scores highest in persuasiveness. This can be attributed to it's detailed explanations which list down individual symptoms and their relation to the meme. In some cases, it is observed that the human annotators find the model convincing enough to move away from their initial classification of the depressive symptoms in favour of supplemental symptoms pointed out by the language model.

RQ5: Understanding the appeal of some explanations. A key observation during the annotation process is that the human annotators are able to identify the technical superiority of the GPT-4o model over other models without model identifiers during the annotation process. We hypothesise that some features of model generations stand out to human annotators and improve their confidence in the models. We ask annotators to specifically mark standout features from the generated explanations in order to get a better understanding of human perspective on the generations. A repeated theme in these standout features is the discussion of references to popular culture. Another aspect that stands out to annotators is deep discussion about metaphors in the meme. For example, annotators



(a) Emotional Mask

(b) Cultural Reference to the Avengers Movie

(c) Cultural Reference

Figure 2: Examples on visual cues

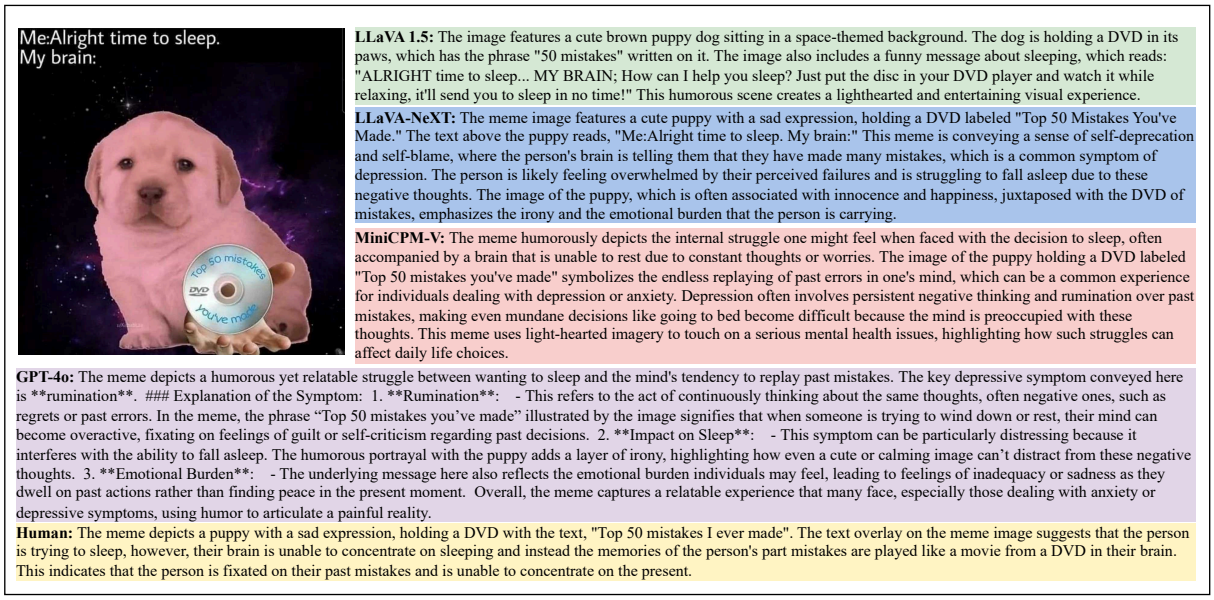


Figure 3: Example meme with LLaVA 1.5, LLaVA-NeXT, MiniCPM-V, GPT-4o and Human explanations

point to the discussion of rumination as a standout feature for the GPT-4o explanation (c.f. Figure 3).

6 Conclusion and Future Work

This work focuses on providing an understanding of the current state of VLM reasoning for complex data like memes in the mental health domain. We present a novel resource in the form of our dataset, RESTOR**Ex**, providing a mix of VLM generated and human annotated explanations for depressive memes. We provide a comprehensive study to validate the quality of explanations generated by VLMs in an effort to pave the way for their future application in automatically annotating depression related content. To enhance the soundness of our analysis, we work with experts in the domain of psychology and analyse our set of depressive memes from the lens of psychological theories prevalent in the field. Our application of the CAT evaluation provides a framework for future work in the inter-disciplinary area of meme and mental health research to provide analysis backed by literature in the field of Psychology.

We recommend future work on this topic to focus on techniques for effective utilisation of these model generated explanations in enhancing the performance of classification task. Techniques which enhance the understanding of figurative speech, a common feature of memes, such as the use of knowledge graphs and summarisation need further exploration. The rich information present in these explanations needs may better captured by such methods.

Limitations

In this work, we aim to further research in the detection of depressive symptoms in memes shared on social media platform. As we utilise memes shared by users online, we have no clinical evidence that the user suffers from the adjudged depressive symptom. Further, given the complex nature of the task, we cannot be assured of annotation quality as it may contain the biases and prejudices of the human annotators. We recommend that there must be further validations of the dataset system before it may be deployed publicly given the sensitive nature of the mental health domain. Further, as we utilise LLMs and VLMs for generating content, we must be aware of their tendency to hallucinate and the explanations provided by these models may not be considered as the gold standard without human validation. With our study, we do not intend to provide a diagnosis of mental health disorders for users and it is designed only to assist interested parties such as clinicians.

Ethical Considerations

As the base dataset that we extend upon is curated automatically, there are concerns of privacy and confidentiality of users online. We ensure that the original authors committed to ethical practices and maintained user privacy before utilisation of the dataset. Further, we assure that we store the data safely and do not attempt to de-anonymise users in any way possible. Further, the usage and training of large language models requires heavy computational resources which may have serious environmental impacts. However, we remained committed to limiting our utilisation of computational and energy resources and attempt to use these models in inference setups requiring significantly less resources. The resources we have used have been obtained for research purpose only and after due permission of the original authors. Any biases in the source dataset are attributed to the original authors and any newly introduced biases are unintentional. Researchers may access our dataset after due permission of the authors of the original datasets as we merely extend on the same.

References

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. MemeMQA: Multimodal question answering for memes via rationale-based inferencing. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5042–5078, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Umair Akram, Jennifer Drabble, Glhenda Cau, Frayer Hershaw, Ashileen Rajenthran, Mollie Lowe, Carissa Trommelen, and Jason G. Ellis. 2020. Exploratory study on the role of emotion regulation in perceived valence, humour, and beneficial use of depressive internet memes in depression. *Scientific Reports*, 10(1):899.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Helen Brown. 2022. The surprising power of internet memes — bbc.com. [Accessed 15-10-2024].

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Sarah J. Erickson and Sarah W. Feldstein. 2006. Adolescent humor and its relationship to coping, defense strategies, psychological distress, and well-being. *Child Psychiatry and Human Development*, 37(3):255–271.

Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, dahai li, Zhiyuan Liu, and Maosong Sun. 2024. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. 2021. The hateful memes challenge: Competition report. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR.

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.

Klaus krippendorff. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Quality and Quantity*, 38(6):787–800.

K. Kroenke, R. L. Spitzer, and J. B. Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1739–1749, New York, NY, USA. Association for Computing Machinery.

Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes. *Preprint*, arXiv:2303.09892.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff aznd Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach,

Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Glenys Parry, Dawn Bennett, Anthony D. Roth, and Stephen Kellett. 2021. Developing a competence framework for cognitive analytic therapy. *Psychology and Psychotherapy: Theory, Research and Practice*, 94(S1):151–170. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/papt.12306.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P. Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes (short paper). In *DE-FACTIFY@AAAI*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

FSM Sarink and JM García-Montes. 2023. Humor interventions in psychotherapy and their effect on levels of depression and anxiety in adult clients, a systematic review. *Frontiers in Psychiatry*, 13:1049476.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role labelling in memes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

The Lancet Global Health. 2020. Mental health matters. *The Lancet Global Health*, 8(11):e1352.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Andreas Wagener. 2021. The postdigital emergence of memes and gifs: Meaning, discourse, and hypernarrative creativity. *Postdigital Science and Education*, 3:831–850.

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, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya, and Amit Sheth. 2018. Multi-task learning framework for mining crowd intelligence towards clinical treatment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 271–277, New Orleans, Louisiana. Association for Computational Linguistics.

Shweta Yadav, Usha Lokala, Raminta Daniulaityte, Krishnaprasad Thirunarayan, Francois Lamy, and Amit Sheth. 2021. "when they say weed causes depression, but it's your fav antidepressant": Knowledge-aware attention framework for relationship extraction. *PLOS ONE*, 16(3):1–18.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, Paris, France. IEEE.

Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal Understanding of Memes with Fair Explanations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2007–2017, Seattle, WA, USA. IEEE.

A Annotation Process and Guidelines

The interface used by the annotators for this annotation task is available at Figure 4.

The full details provided to annotators for the comparative survey are described in this section.

General Guidelines As annotators for the explanation analysis task, you will encounter three types of questions. There would be explanation-specific questions (which require you to answer specific questions about each explanation), comparative questions (which require you to rank the explanation for the given criteria) and descriptive questions (which require you to answer them in about four lines or 30-40 words each).

Q1. [Explanation x] [Fluency] Is the explanation well formed and fluent in English, irrespective of its relevance to the meme context? Guideline: The options are on a 5-point Likert scale, with 5 representing an extremely fluent and well-formed explanation and 1 representing an inarticulate or badly-formed explanation. This question essentially pertains to judging the language skill of the model. You do not need to consider the meme image for this question and should ideally judge the explanation solely on its language.

Q2. [Explanation x] [Relevance] Is the explanation relevant to the context of the meme? Guideline: The options are on a 5-point Likert scale, with 5 representing a highly relevant explanation and 1 representing an irrelevant explanation. This question is intended to match how aligned the explanation text is to the meme image. The question also intends to answer if the explanation is aligned with the depressive symptoms in the meme or if it is a generic overview of the meme. You should carefully consider the relationship between the meme image and the explanation while answering this question. Do not consider if the language is grammatically correct.

Q3. [Explanation x] [Literal Meaning] Does the explanation provide a literal interpretation of the meme content? Guideline: This is a binary 'Yes' or 'No' question. The answer is whether the explanation provides "a" literal meaning or interpretation for the given meme image. This would encompass the model explaining literally the objects of the meme and their actions, such as "a man with his hands on his face." This contrasts with a figurative interpretation, such as "a depressed man holding his head in distress." Please note that answering 'Yes' to this does not indicate that there are no figurative interpretations also present in the meme. The presence of these two is mutually independent. If you answer yes to the question above, please also enter the span that correlates to this kind of literal interpretation by copying and pasting that span (or sentence of text) into the provided text box.

Q4. [Explanation x] [Figurative Meaning] Does the explanation provide a figurative interpretation of the meme content? Guideline: This is a binary 'Yes' or 'No' question. The answer is whether the explanation provides "a" figurative meaning or interpretation for the given meme image. This would encompass the model explaining figuratively the objects of the meme and their actions, such as "The phrase "surf on depression" is a metaphorical expression that suggests the person is experiencing a deep sense of sadness or emotional turmoil, which is often associated with depression." This is in contrast to a literal interpretation that depression is a physical object like a surfboard that the person is riding. Please note that answering 'Yes' to this does not indicate that no literal interpretations are also present in the meme. The presence of these two is mutually independent. If you answer yes to the question above, please also enter the span that correlates to this kind of figurative interpretation by copying and pasting that span (or sentence of text) into the provided text box.

Q5. [Explanation x] [Depressive Symptoms] List the depressive symptoms identifiable from the explanation. Guideline: This is a multiple-select question with eight options referring to the eight depressive symptoms. The question aims to identify depressive symptoms using "only" the model-generated explanation. As your response to this question, you are required to select the depressive symptoms without considering the meme image as a data source in your judgement.

Q6. [Persuasiveness] Rank the explanations based on their persuasiveness. Guideline: Here, persuasiveness refers to the property of convincingness of an explanation to make you align your categorisation of the depressive symptoms with those pointed out by the explanation. For instance, a detailed description of the figurative meaning of one aspect of the meme may persuade you to consider its thought process as correct. In this question,

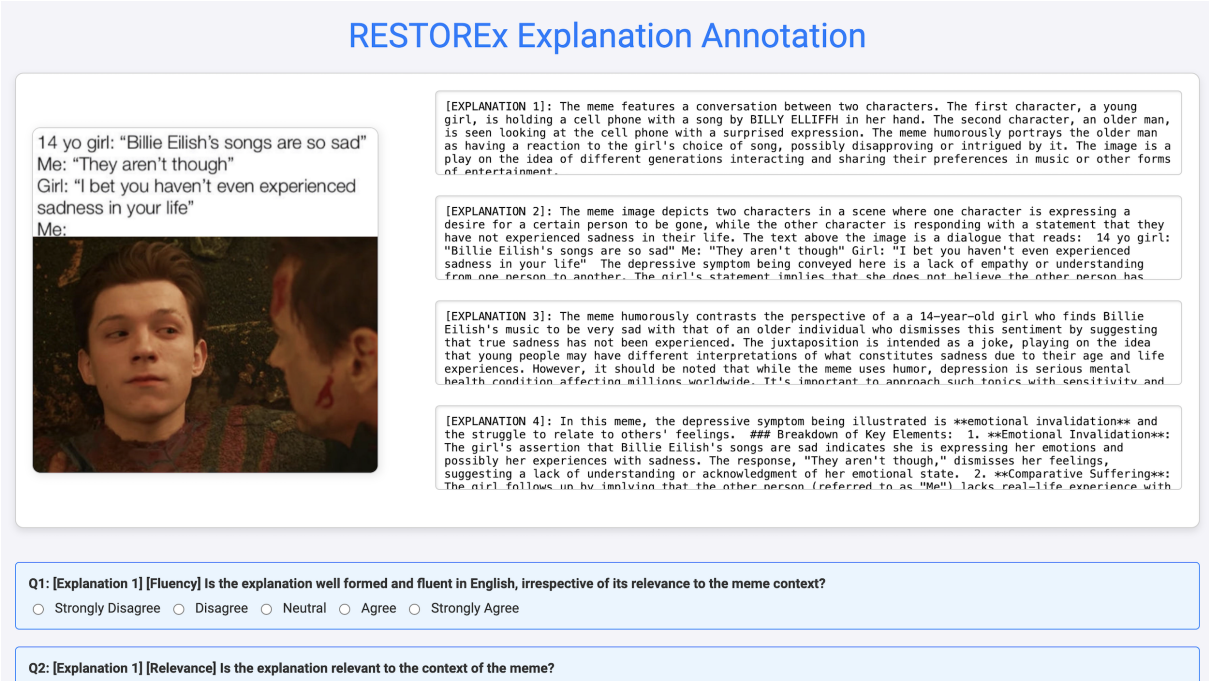


Figure 4: Annotation interface provided to the annotators

we are not as concerned about the correctness of the thought process as we are about the presentation of the model in convincing you that it is correct and shaping your opinion in a way. A good way to think about this question would be to ask the question yourself if the explanation introduced another viewpoint about the meme that you had not previously considered and how strong that feeling was.

Q7. [Correctness] Rank the correctness of each explanation in identifying depressive symptoms. Guideline: In this question, you are asked to rank the explanations based on your perceived correctness of their interpretations. This does not relate to confidence or convincingness of an explanation like the question about persuasiveness would. You should provide your ranking strictly based on which explanation provides the most evidence about the prevalence of the depressive symptoms that is correct and which allows you to mark categories from the available options. For example, direct references to the meme exhibiting sleeping disorder with the use of words such as “lack of sleep” would mean that the explanation exhibits high correctness when sleeping disorder is one of the correct labels according to you.

Q8. Provide your own explanation about the meme given the meme image. Guideline: In this question, you are asked to provide a short explanation of the type that the models have provided in your own words. Please try to capture your thought process in the explanation and do not rely on repeating the viewpoint mentioned by the model-generated explanation (unless that was part of your original thought process). The explanation you provide should be comparable in length to the other explanations or around 5 full sentences.

Q9. Describe standout features or sentences from one or more of the explanations. Guideline: In this prompt, you are asked to mention some parts of specific explanations that stood out from other explanations and helped distinguish them. The emphasis should be on sentences or elements that were not captured by other explanations. These should be placed in context with the other explanations, that is to say that a part of the explanation is standout with respect to the other explanations. If a really good analogy or metaphor is captured successfully by all the explanations then it is not a standout feature. These could be positive or negative points so you are required to mention if the feature was helpful or detrimental to your understanding of the meme. An example of this could be a metaphor identified by an explanation not captured by other explanations, making the meme easier to understand.