The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

MAMA-Memeia! Multi-Aspect Multi-Agent Collaboration for Depressive Symptoms Identification in Memes

Siddhant Agarwal[1], Adya Dhuler[2], Polly Ruhnke[1], Melvin Speisman[1], Md Shad Akhtar[3], Shweta Yadav[1]

[1] University of Illinois at Chicago, USA
[2] Creighton University, USA
[3] Indraprastha Institute of Information Technology Delhi, India
{sagarw38, shwetay}@uic.edu, shad.akhtar@iiitd.ac.in

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 in expressing depressive sentiments, 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 Large Language Model (LLM) generated and human-annotated explanations. We introduce *MAMA-Memeia*, a collaborative multi-agent multi-aspect discussion framework grounded in the clinical psychology method of Cognitive Analytic Therapy (CAT) Competencies. *MAMA-Memeia* improves upon the current state-of-the-art by 7.55% in macro-F1 and is established as the new benchmark compared to over 30 methods.

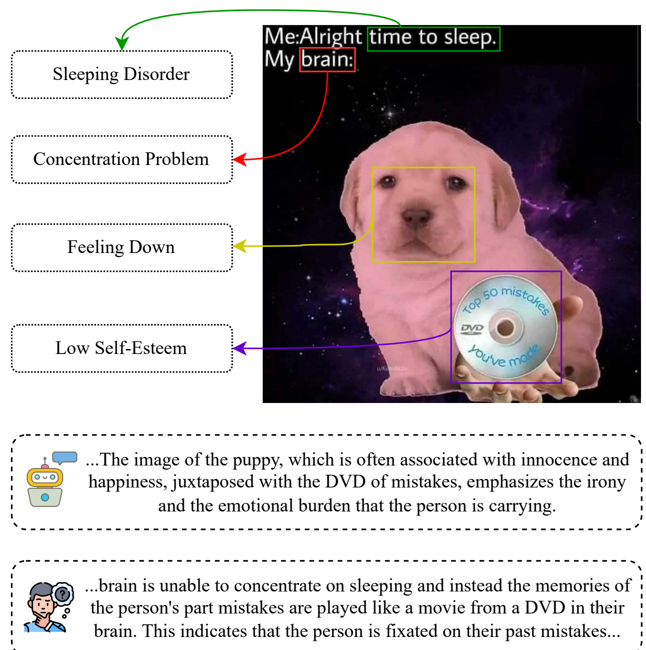


Figure 1: An example of a depressive meme from `RESTOR`**`Ex`** with labeled fine-grained depressive symptoms and *explanations* from the LLM agent and Human

1 Introduction

In the digital age, memes 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). Beyond entertainment, memes have grown to become an important outlet for expressive posting as well (Wang et al. 2019), as people have increasingly begun to use memes to communicate their emotional struggles, most particularly depression, where a simple Google search for "depression memes" shows more than $84,700,000$ results. Akram et al. (2020) observe that sharing humor-intended depressive memes can be beneficial for some individuals, as presenting a negative experience humorously may create a sense of peer support among viewers who have gone through similar situations. The importance of the visual modality is further cemented considering that the majority of social media posts today contain image or video content which increases engagement by around 100% (Cashyap 2024).

Recognizing the importance of memes in online social media today, meme analysis research has gained traction, starting with the Facebook Hateful Memes Challenge (Kiela et al. 2020). However, research in the area has been mostly focused on aspects such as harmfulness (Pramanick et al. 2021) and cyber-bullying (Jha et al. 2024b). 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 based on the clinically established 9-scale Patient Health Questionnaire (PHQ-9) (Kroenke, Spitzer, and Williams 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 *explanations* generated by state-of-the-art Multimodal Large Language Models and supplemented with human-annotated explanations for the memes. For this task, we examine the application of Multimodal Large Language Models in understanding complex multimodal data of memes in the context of mental health. This is a complex task for language models given the mix of sarcasm, irony and other types of figurative speech involved in memes. 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 emphasizes on the meme author's fixation on their past mistakes, correctly identifying the *'Low Self-Esteem'* symptom. Further, the LLM-generated explanation, recognizing 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 various figurative speeches in the meme, the method 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 introduce *MAMA-Memeia*, a Multi-Aspect Multi-Agent collaborative framework for identifying depressive symptoms from memes. We ground this proposed methodology in clinical psychology by adapting the Cognitive Analytic Therapy (CAT) Competencies (Parry et al. 2021) into our Multi-Aspect prompting setup which is used by medical professionals for analyzing thinking patterns. We supplement our analysis with extensive comparative results establishing the efficacy of *MAMA-Memeia* as the preeminent method for the task of depression symptom analysis.

Our contributions are summarized as follows-

- **Novel Dataset:** Introducing the RESTOR**Ex** dataset, with LLM-generated explanations and human annotated gold label explanations.
- **Novel Methodology:** Introducing the *MAMA-Memeia* framework, a novel state-of-the-art multi-agent discussion framework that builds upon the foundations of clinical psychology for multi-aspect prompting.
- **In-depth Human Evaluation:** Conducted a human evaluation with domain experts to address key research questions regarding LLM-generated content across five aspects.

2 Related Work

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. 2020), harmfulness (Sharma et al. 2023a), and emotions (Sharma et al. 2020; Mishra et al. 2023). Recently, Joshi, Ilievski, and Luceri (2024) attempted to understand the propagation of meme content across social media platforms.

While most work on modeling for meme analysis has focused on contrastive learning setups (Mei et al. 2024) and knowledge fusion techniques (Sharma et al. 2023b), the recent direction of the area has been towards the use of LLMs (Jha et al. 2024a). Works such as those by Agarwal et al. (2024) and Zhong and Baghel (2024) have explored the application of explanations generated by 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.

	LOI	FD	ED	SD	LSE	CP	SH	Total
Train	441	1781	1714	997	703	348	1357	7096
Test	66	219	85	78	114	73	82	520
Validation	37	130	54	36	79	28	54	310

Table 1: Class-wise distribution of fine-grained depression symptom labels in RESTOR**Ex**.

While research on memes expanded, their application in the mental health domain remained limited. Yadav et al. (2023) extended meme analysis research towards the mental health domain with the release of the RESTORE dataset for detecting fine-grain depressive symptoms in memes. Mazhar et al. (2025) expand this research by utilizing a knowledge-fusion and retrieval framework for depression symptom identification as well as introducing a dataset for detecting anxiety symptoms. We aim to expand on these works with a more elaborate method by utilizing LLM agents.

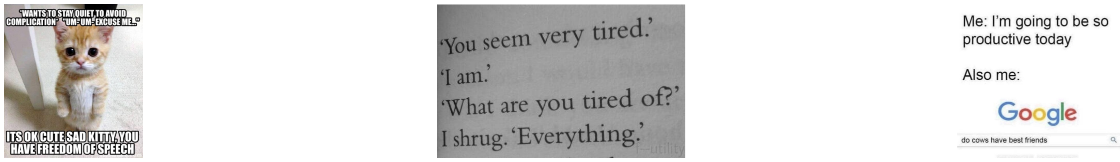
Previous work on depressive symptom analysis has focused mostly on textual data such as Tweets and Reddit posts, with limited work on multimodal content such as memes (Yadav et al. 2023). Works such as Yadav et al. (2021, 2018); Benton, Mitchell, and Hovy (2017) have focused on analyzing depressive symptoms by collecting data from online social media platforms, mostly on tweets of users for detection or tracking of depression symptoms in users. Works such as Yates, Cohan, and Goharian (2017); Yadav et al. (2020) have served as important benchmarks for analyzing textual depressive data. Typically text-based depression symptom analysis works utilize simpler BERT-based methods with specialized modules (Zhang et al. 2023). Recently, Wang, Inkpen, and Kirinde Gamaarachchige (2024) analyzed the utility of LLMs for estimation of the levels of depression by prompting LLMs with a questionnaire. Chen et al. (2024) have utilized LLMs for synthetic data creation in the form of interviews of depressed users highlighting the capabilities of LLMs to work in the mental health context. This work aims to extend depression analysis to the multimodal setting and applying advanced LLM-based methods, including multi-agent methods which have not yet been effectively utilized in depression related tasks.

3 The RESTOR**Ex** Dataset

We introduce RESTOR**Ex**, a dataset for the multi-label classification of seven fine-grained PHQ-9 depressive symptom categories – *Feeling Down (FD), Lack of Interest (LOI), Eating Disorder (ED), Sleeping Disorder (SD), Concentration Problem (CP), Low Self Esteem (LSE), and Self Harm (SH).* We provide expert annotated human explanations and conduct extensive expert analysis of all generated content. Table 1 provides label-wise distribution of RESTOR**Ex**.

3.1 Dataset Curation and Filtering

The RESTOR**Ex** dataset is derived from the RESTORE dataset, which collects meme images from the social media



(a) Misclassifications in Test Samples: Original- [FD, LOI], Updated- [FD, LSE]

(b) Non-meme Images: image that corresponds to a quote

(c) Inaccurate 'Lack of Energy' samples

Figure 2: Examples of meme samples from `RESTORE` that were re-annotated for the curation of `RESTOR`**`Ex`**

platforms, Twitter and Reddit. The dataset consists of human annotations for eight fine-grained depression symptom categories in the test and validation subsets of the dataset. The training subset is supplemented with automatically curated samples using keyword-based search (such as 'eating disorder memes', 'feeling down memes', etc.).

In the curation of the `RESTOR`**`Ex`** dataset, we identify and correct the following inconsistencies in the original `RESTORE` dataset: **(i)** Re-annotated the test and validation sets of the dataset with additional annotators to mitigate issues with misclassified labels as shown in Figure 2a. **(ii)** Removed non-meme images such as quotes from the dataset as shown in Figure 2b. leading to a reduction of about 20% in the dataset size. **(iii)** Removed the *'Lack of Energy'* label due to the lack of accurate training labels as shown in Figure 2c. This was done as all of the samples for this class in the training set (471 samples) were automatically curated with significant issues.

3.2 Explanations in RESTOREx

To enhance the utility of `RESTOR`**`Ex`** as a resource, we introduce ground-truth human-written explanations for the test subset (520 samples). This ground-truth explanation serves as a resource for future research to compare model-generated explanations with a human-annotated dataset. An *explanation* by the human-annotators, through their nuanced thought process, encapsulates important aspects such as figurative language (for example, sarcasm and metaphors), commonsense reasoning, and is grounded in their cultural awareness which allows explanations to be aware of the cultural phenomena of the time. These detail-rich explanations serve as a textual description of the cross-modal information in a meme image.

3.3 Annotation Guidelines

The curation of the `RESTOR`**`Ex`** dataset consists of two annotation tasks for dataset re-annotation and filtering, and for curation of human-written gold label explanations.

Task 1: For a given meme, the annotators are required to determine one or more of the seven depressive symptoms conveyed in the meme according to PHQ-9 definitions (Kroenke, Spitzer, and Williams 2001). Any non-meme image is filtered in this process. This annotation is performed by two annotators trained on the task of fine-grained depression symptom identification as per the guidelines of the original dataset curation (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.

Task 2: For each meme, the annotators are required to provide the human-annotated ground-truth explanations in the provided text field. The annotator is instructed to capture their thought process and incorporate details such as the use of figurative language, commonsense knowledge, cultural references, and visual cues in their provided explanations. This annotation process is performed by two domain-expert annotators.

3.4 LLM Generated Explanations

We propose to extend the dataset further with Multimodal LLM generated explanations as a proxy for human-written annotations for large datasets where human annotation is not feasible such as for the over 7000 samples in the train dataset. We generate explanations using three open-source and three closed-source LLMs and conduct a human analysis on these explanations to evaluate their utility as a proxy for human-written explanations.

Analysis on LLM Generated Explanations We focus on five research questions in order to comprehensively judge the quality of LLM generated explanations: **RQ1:** *Are LLMs fluent?*, **RQ2:** *Are generated explanations relevant?*, **RQ3:** *Do LLMs capture figurative meaning?*, **RQ4:** *Are LLMs persuasive?* and **RQ5**: *Understanding the appeal of some explanations*. We conduct this analysis on six selected model explanations and human-annotated explanations based on a detailed questionnaire on a subset of test samples. We choose three open source models – LLaVA 1.5, LLaVA-NeXT, MiniCPM-Vand three closed source models – GPT-4o, Claude 3.5 Sonnet and Gemini-2.0-flash. The results of this human analysis clearly establish the superiority of closed-source LLMs such as Gemini (Team 2024) over open-source models for the task of `RESTOR`**`Ex`** motivating our decision of utilizing such closed-source models as the backbone of our methodology.

4 Methodology

In this section we propose the *MAMA-Memeia* framework, a collaborative multi-agent discussion framework that utilizes clinical psychology principles to detect depressive symptoms. We first describe the domain expert's perspective on the task of depressive symptom classification and then

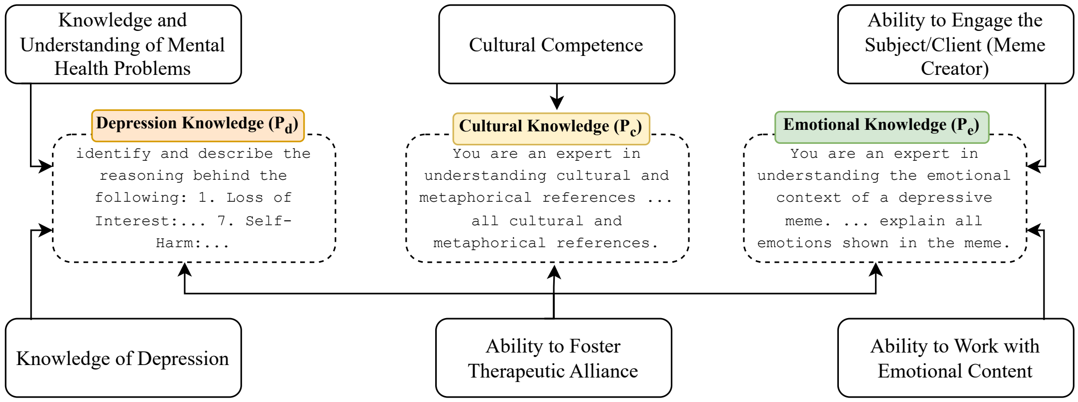


Figure 3: Cognitive Analytic Therapy (CAT) Competencies (Parry et al. 2021) adapted as guidelines form the basis for Multi-Aspect Prompting. Multiple criteria combine to form each Knowledge Aspect. We design three Aspect-specific prompts – P_d (Depression Knowledge), P_e (Emotional Knowledge), and P_c (Cultural Knowledge) – to generate the corresponding explanations – E_d, E_e, and E_c. We use these explanations individually and concatenated as $< E_d, E_e, E_c >$ for our experiments.

present our approach to integrating this perspective in a multi-agent setup through multi-aspect prompting.

4.1 Domain Expert Perspective

The psychology underlying meme with mental health is centered around emotional expressions and coping mechanisms (particularly, in the form of humor). Specifically, meme allows individuals to express complex emotions surrounding their mental health 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).

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 associated with stable psychological adjustment, negative humor (especially self-defeating humor seen in memes) is associated with higher depressive symptoms (Erickson and Feldstein 2006). Therefore, although it helps tops to 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).

Cognitive Analytic Therapy for Understanding Depressive Memes Cognitive Analytic Therapy (CAT) is a form of talking therapy that focuses primarily on identifying patterns of thinking, feeling, and behavior (NHS 2025). It is widely used in the clinical therapy setting for people living with depression, anxiety, or eating problems, who self-harm or have personal or relationship problems. In this study, we adapted CAT - General Therapeutic Competencies (Parry et al. 2021) to formulate evaluative guidelines for interpreting memes on mental health themes. It is highly applicable to the task due to its structured approach to categorizing and assessing both larger psychological patterns and more subtle emotional nuances and risk assessments. We leverage eight CAT criteria focusing on meme understanding – (i) *Knowledge and Understanding of Mental Health Problems*, (ii) *Knowledge of Depression*, (iii) *Cultural Competence*, (iv) *Ability to Engage the Subject/Client (Meme Creator)*, (v) *Ability to Foster Therapeutic Alliance*, (vi) *Ability to Work with Emotional Content*, (vii) *Ability to Assess and Manage Risk of Self-Harm*, (viii) *Knowledge of Ethical Guidelines*.

4.2 Multi-Aspect Prompting

We ground our approach in the CAT criteria described in section 4.1. These guidelines highlight three major aspects of knowledge required for understanding depressive memes:
(i) Depression Knowledge - This aspect deals with providing knowledge about the specific depressive symptoms to the model. We provide this knowledge with the use of definitions for the seven symptoms as laid out by (Yadav et al. 2023). These definitions ensure that the model avoids misinterpretation of the depressive symptoms.
(ii) Emotional Knowledge - A key consideration while determining depressive symptoms according to the CAT guidelines is the consideration of the emotional state of the user. To this effect, we include this knowledge aspect to ensure that the model considers the emotional states that may be associated with depressive themes in the meme.
(iii) Cultural Knowledge - This knowledge aspect aims to provide information about important cultural phenomena such as pop culture references or metaphorical interpretations that form a key part of meme understanding. With the inclusion of such knowledge it is expected that the model would be able to recognize the figurative language of the meme across the textual and visual modalities.

Based on these, we design three Aspect-specific prompts P_d (Depression Knowledge), P_e (Emotional Knowledge), and P_c (Cultural Knowledge) as described in Figure 3.

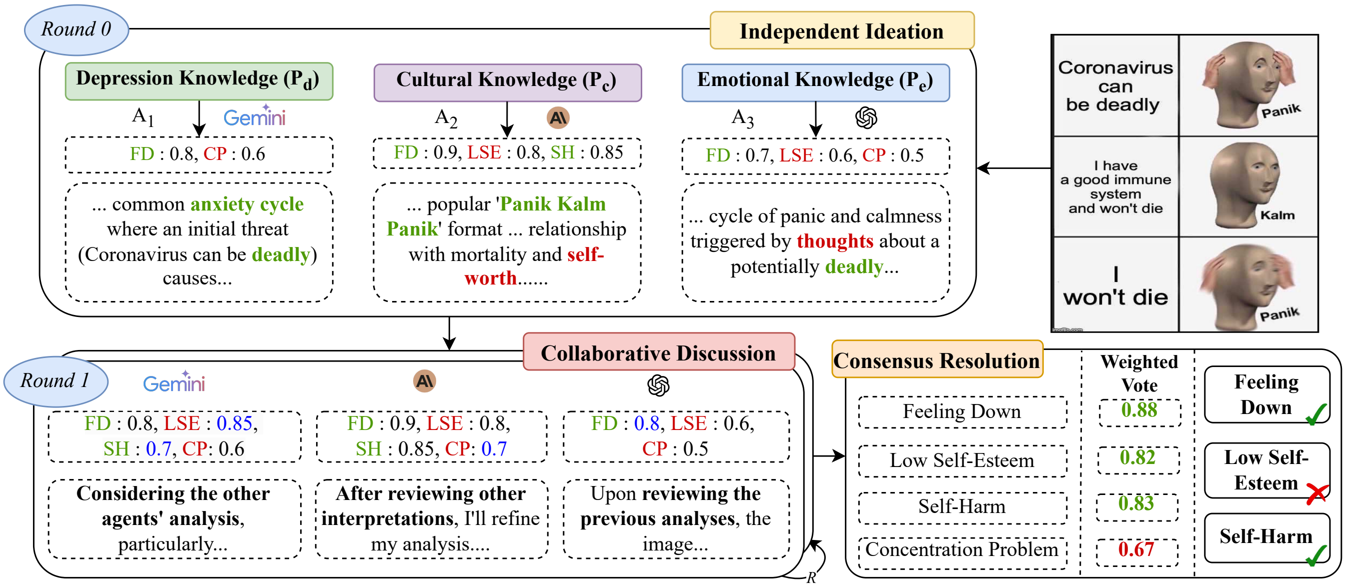


Figure 4: Overview of the *MAMA-Memeia* framework with Gemini-2.0-flash Gemini, Claude 3.5 Sonnet, GPT-4o, consisting of three phases- (1) *Independent Ideation*: generation of initial predictions and reasoning, (2) *Collaborative Discussion*: Agents collaborate and reconsider their predictions, (3) *Consensus Resolution*: Weighted vote to determine predicted labels.

4.3 *MAMA-Memeia*: A Multi-Agent Multi-Aspect Inference Framework

We introduce the *MAMA-Memeia* framework, which builds upon the Multi-Aspect Prompting setup by utilizing multiple agents that capture specific knowledge of depressive memes in the prediction of the final labels. Our approach is inspired from multi-agent works (Wu et al. 2024; Du et al. 2024; Chen, Saha, and Bansal 2024) which highlights the efficacy of multi-agent setups.

We utilize this knowledge in combination with our study of Multi-Aspect Prompting by deploying three recent state-of-the-art LLMs - GPT-4o (OpenAI 2024), Claude-3.5 Sonnet (Anthropic 2024) and Gemini-2.0-flash (Team 2024) - in a multi-agent setup. Our choice of these models is based on the human analysis provided in subsection 3.4. The *MAMA-Memeia* framework consists of three phases: *Independent Ideation, Collaborative Discussion and Consensus Resolution*. It is demonstrated in Figure 4 and described in detail below for n agents.

Independent Ideation The first step in the *MAMA-Memeia* framework is the generation of initial symptom predictions and reasoning for the given meme from each agent. Formally, given a list of agents $A = \{A_1, A_2, ..., A_n\}$ of size n, each agent $A_i \in A$ generates an initial prediction, $p_i^{(0)}$, which is a list of s_i symptoms predicted by it. The study (Xiong et al. 2024) has shown the effectiveness of LLM-derived confidence scores, therefore, we also generate confidence estimates, $c_{i,j}^{(0)} \in [0, 1] \ \forall \ j \in s_i$. Each agent, A_i, also generates an initial explanation, $e_i^{(0)}$ that explains the thought process behind it's predictions. To achieve this, A_i is designated a Aspect-Specific System Prompt, $P_{a,i} \in \{P_d, P_e, P_c\}$, for each agent to focus on one knowledge aspect. Each agent A_i is then prompted with the Aspect-specific Prompt, $P_{a,i}$, the meme image, M_{img}, and the user prompt, P_{usr} (detailed prompts in suppl.). For each agent, A_i, this can be formulated as:

$$p_i^{(0)}, c_i^{(0)}, e_i^{(0)} = A_i(P_{a,i}, M_{img}, P_{usr}) \tag{1}$$

Collaborative Discussion In the next phase of our framework, all agents undergo R rounds of collaborative discussion. This is done to ensure each agent is fairly exposed to the thought process of other agents before coming up with their final predictions. For each discussion round, r, every agent is provided a Discussion prompt, P_{dsc}, which has multiple characteristics. The discussion prompt first introduces the agents to the idea that they are in a collaborative environment by asking them to review the responses of other agents and then reconsider their response. We call this the initiator prompt, P_{ini}. Secondly, the discussion prompt for discussion round, r, consists of the responses of the other agents for round, $r-1$ (for $r = 1$ this refers to the Independent Ideation phase). Each agent is provided with both explanation, $e_j^{(r-1)}$, as well as predictions, $p_j^{(r-1)}$, generated by the other agents, $A \setminus A_i$. This is supplemented with the confidence estimates, $c_j^{(r-1)}$, for each prediction in $p_j^{(r-1)}$. The availability of all three components, explanation, predictions, and confidence estimates, allows the agent to condition on the reasoning and confidence associated with the predictions before it reconsiders its initial judgement. The round r discussion prompt, for agent A_i, is:

$$P_{dsc,i}^{(r)} = P_{ini} + \|_{j=1, j \neq i}^{n} \{e_j^{(r-1)}, p_j^{(r-1)}, c_j^{(r-1)}\}$$

This discussion prompt is then passed along with the entire conversation history of the agent, $h_i^{(r)}$, to obtain the predic-

Model	Macro-F1	Weighted-F1
Unimodal Text		
OCR:		
BERT (Devlin et al. 2019)	62.02	62.63
MentalBERT (Ji et al. 2022)	63.77	64.47
BART (Lewis et al. 2020)	44.71	49.46
MentalBART (Yang et al. 2023)	61.76	62.43
Explanation:		
BERT	63.39	63.96
MentalBERT	64.62	65.27
BART	55.81	57.11
MentalBART	64.96	64.30
Unimodal Image		
ViT (Dosovitskiy et al. 2021)	34.96	39.04
ResNet (He et al. 2015)	27.14	33.46
EfficientNet (Tan and Le 2019)	25.18	31.61
Multimodal		
Image + OCR		
CLIP (Radford et al. 2021)	45.83	48.09
VisualBERT (Li et al. 2019)	62.70	63.67
ViT + BERT	38.57	42.24
Image + Explanation		
CLIP	39.23	42.56
VisualBERT	62.37	62.25
ViT + BERT	37.78	40.58
Previous SOTA		
Yadav et al. (2023) †	65.18	64.67
MAMA-Memeia Gemini	**72.73**	**72.45**
$\Delta_{\textit{MAMA-Memeia}\ -\ \dagger}$	↑ 7.55%	↑ 7.78%

Table 2: Comparison of *MAMA-Memeia* with unimodal and multimodal baselines along with previous state-of-the-art.

tions, confidence estimates and explanation of the agent for round r as:

$$p_i^{(r)}, c_i^{(r)}, e_i^{(r)} = A_i(P_{dsc,i}^{(r)}, h_i^{(r)}) \tag{2}$$

Consensus Resolution At the end of $N = R + 1$ rounds, we determine the final predictions through our consensus resolution algorithm which performs a weighted vote between the agents based on the confidence estimates. For a given threshold, t, the final labels, L, are determined as follows:

$$L = \left[j \;\middle|\; \left(\frac{1}{n}\sum_{i=1}^{n} c_{i,j}^{(N)}\, \mathbb{I}\{j \in p_i^{(N)}\}\right) > t \right] \tag{3}$$

The consensus resolution algorithm is used in order to eliminate labels for which the agents have low confidence. It also ensures that there is a balance between the agents and extremely high confidence of one agent does not bias *MAMA-Memeia*.

5 Experiments: Benchmarking RESTOREx

Baselines To establish the effectiveness of the *MAMA-Memeia* framework, we first compare its performance to several relevant unimodal and multimodal baseline architectures as described in Table 2. These baselines use the following models: **(i) Unimodal Text**: BERT, MentalBERT, BART, MentalBART **(ii) Unimodal Image**: ViT, ResNet, EfficientNet **(iii) Multimodal**: CLIP, VisualBERT, BERT + ViT, SOTA (Yadav et al. 2023)

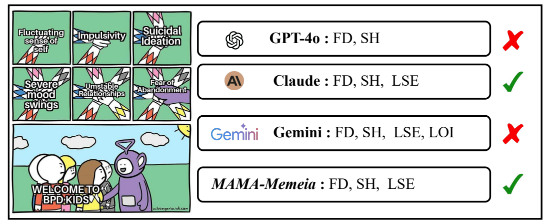


Figure 5: Meme example with predicted labels from various LLMs and *MAMA-Memeia*.

Baseline Results As shown in Table 2, *MAMA-Memeia* outperforms the previous State-of-the-Art results of (Yadav et al. 2023) by 7.55% in macro-F1 score and 7.78% in weighted-F1 score. The setup by Yadav et al. (2023) utilizes multimodal fusion based on conditional adaptive gating with pre-trained ResNet and BERT models which are fine-tuned for the task of depressive symptom classification. The superior performance of *MAMA-Memeia* highlights the effectiveness of Large Language Model based approaches over these traditional pre-trained models. This is further confirmed compared to the fine-tuned unimodal and multimodal baselines which sit well below the benchmark set by *MAMA-Memeia*. Further, Table 2 highlights the importance of the textual modality for the fine-grained depressive symptom analysis task as shown by the sharp improvement in performance from the unimodal image setup to the unimodal text setup. The utility of the LLM generated explanations is also established as a consistent increase in performance can be seen by substituting OCR with LLM generated explanations as the textual modality as seen with the BART model improving in macro-F1 from 44.71% with OCR to 55.81% with explanations.

Experimentation with Multimodal LLMs We experiment with a variety of different open-source and closed-source models to provide explanations for memes as shown in Table 3. These setups are compared in both Single-Agent and Multi-Agent setups, as well as with a variety of prompting setups including the Multi-Aspect prompting (cf. Section 4.2). Finally, we compare *MAMA-Memeia* and other LLM based approaches with results derived using gold-label human-written explanations. For a representative variety in the generated content, we put careful consideration into the choice of our models. We ensure a mix of open-source and closed-source models with various base foundational LLMs that these Multimodal LLMs are built upon. Based on these criterias, we generate explanations from six popular Multimodal LLMs - LLaVA 1.5 (Liu et al. 2023), LLaVA-NeXT (Liu et al. 2024), MiniCPM-V (Hu et al. 2024), GPT-4o (OpenAI 2024), Claude 3.5 Sonnet (Anthropic 2024), Gemini-2.0-flash (Team 2024).

Results and Ablation As seen in Table 3, the performance of closed-source models such as Claude 3.5 Sonnet is significantly better compared to the open-source models such as LLaVA 1.5. The significant boost for the GPT-4o and Gemini-2.0-flash models with the Multi-Aspect Prompting setup indicates towards the effectiveness of these

Method	FD	LOI	SH	ED	LSE	CP	SD	Macro-F1	Micro-F1	Weighted-F1
Human Explanations	63.00	47.00	77.00	72.00	54.00	57.00	81.00	65.00	62.00	64.00
Vanila LLM Explanations										
Open-Source LLMs										
LLaVA 1.5	55.00	37.00	34.00	48.00	38.00	47.00	67.00	47.00	48.00	48.00
LLaVA-NeXT	62.00	43.00	60.00	71.00	46.00	53.00	79.00	59.00	57.00	59.00
MiniCPM-V	62.00	40.00	53.00	73.00	51.00	53.00	78.00	59.00	58.00	59.00
Closed-Source LLMs										
GPT-4o	63.00	39.00	58.00	73.00	46.00	54.00	86.00	60.00	57.00	60.00
Cluade 3.5 Sonnet	70.78	48.25	77.71	81.01	54.35	57.92	88.76	68.39	66.19	68.74
Gemini-2.0-flash	63.09	32.02	75.86	75.47	49.22	60.47	80.22	62.34	58.19	62.55
Multi-Aspect Prompting										
Depression Knowledge (P_d)										
GPT-4o	65.13	**53.33**	79.52	78.21	45.02	58.46	86.22	66.56	62.67	65.66
Cluade 3.5 Sonnet	68.27	45.45	77.30	82.93	48.46	60.09	84.71	66.74	63.60	66.75
Gemini-2.0-flash	66.40	46.60	73.85	74.21	55.86	65.92	78.95	65.97	64.93	65.99
Cultural Knowledge (P_c)										
GPT-4o	69.55	49.43	69.50	79.22	49.87	60.60	86.06	66.31	64.67	66.60
Cluade 3.5 Sonnet	71.09	45.57	72.99	83.23	44.95	56.67	85.21	65.67	62.61	66.31
Gemini-2.0-flash	66.77	38.96	61.88	75.74	45.55	53.46	76.50	59.84	57.99	61.04
Emotional Knowledge (P_e)										
GPT-4o	65.52	48.61	76.39	76.39	49.87	68.57	83.87	67.03	64.36	66.31
Cluade 3.5 Sonnet	62.27	40.14	70.73	80.77	42.04	47.39	86.75	61.44	56.62	61.33
Gemini-2.0-flash	66.54	48.43	73.47	77.78	53.98	65.93	79.37	66.50	64.92	66.33
Combined ($P_d + P_c + P_e$)										
GPT-4o	64.94	51.37	79.52	79.77	44.73	59.69	86.75	66.68	62.62	65.74
Cluade 3.5 Sonnet	67.65	45.45	76.83	**83.64**	49.53	58.72	84.71	66.65	63.51	66.62
Gemini-2.0-flash	65.28	48.18	72.73	76.83	53.71	66.67	78.95	66.05	64.41	65.71
MAMA-Memeia	**72.46**	52.27	76.06	80.50	**59.87**	**77.37**	**90.57**	**72.73**	**71.15**	**72.45**
− Aspect-specific Prompting	70.95	49.70	**80.28**	82.12	58.27	73.91	87.90	71.88	70.31	71.51

Table 3: Comparison of *MAMA-Memeia* with Mulitmodal LLM based single-agent setups and Human-annotated explanations. *MAMA-Memeia* is the best performing method in macro, micro and weighted F1 scores. Vanilla LLM Explanations refer to the prompting setup without the inclusion of CAT knowledge based prompts. FD: Feeling Down, LOI: Lack of Interest, SH: Self-Harm, ED: Eating Disorder, LSE: Low Self-Esteem, CP: Concentration Problem, SD: Sleeping Disorder.

knowledge-based prompting setups. This is also noted with the ablation with removing aspect-specific prompting (referring to the Multi-Aspect setup). Compared to the Human Explanations, our framework achieves an improvement of more than 8% and Claude 3.5 Sonnet achieves an improvement of more than 4% in weighted-F1 score supporting our proposal for utilizing LLM generated explanations as an automated and low-resource alternative to human annotations.

Qualitative Analysis We perform a qualitative analysis of the outputs of the GPT-4o, Claude 3.5 Sonnet and Gemini-2.0-flash models along with the *MAMA-Memeia* framework to infer trends in predictions made by these models. An example of this analysis is provided in Figure 5. As seen in Figure 5, we observe the that the Gemini-2.0-flash model is prone to over-prediction while the GPT-4o model is prone to under-prediction. This behavior is partly explained when observing the length of the explanations generated by these models where the Gemini-2.0-flash model consistently outputs the lengthiest explanations among the set. However, given the black-box nature of these models, further analysis is required for detailed understanding of this phenomenon. This offers an explanation for the improved performance of the *MAMA-Memeia* framework where the averaging of the confidence of each of these models yields better results.

Analyzing the patterns in the multiple rounds of debate between the models, we observe that the models frequently correct and influence each other through their detailed explanations. For instance, as in Figure 4, only one of the models (Claude 3.5 Sonnet) correctly predicts the symptom of Self-Harm and corrects the other two models in their predictions over the rounds of debate. This self-correction tendency stands out as a major strength of *MAMA-Memeia*.

6 Conclusion

We introduced `RESTOR`**`Ex`** a dataset for fine-grained classification of depressive symptoms in memes augmented with ground-truth human explanations and LLM generated explanations for memes. We presented *MAMA-Memeia*, a Multi-Agent Multi-Aspect inferencing framework, grounded in literature in Psychology, for classifying depressive symptoms in multimodal meme data. We demonstrated that *MAMA-Memeia* consistently outperformed traditional methods and improved upon single-agent LLM approaches for the task.

While we utilize closed-source models for our *MAMA-Memeia* framework, we look forward to developing methods for effective utilization of open-source LLMs in future work given the pressing need for open and transparent research on LLMs in sensitive domains such as mental health.

Ethical Statement

This work builds upon the original `RESTORE` dataset (Yadav et al. 2023), which contains publicly available meme images collected from social media; the original authors ensured that all data were anonymized. Because our study concerns the sensitive domain of mental health content, we emphasize that the proposed system identifies depressive symptoms expressed within meme content only and is not intended for diagnosing or profiling any individual. Our multi-agent framework leverages large language models that may introduce additional biases, and we release all methods and data resources solely for research purposes. We caution against deploying such LLM-based systems in high-stakes mental health or content-moderation settings without substantial expert human oversight, particularly for critical symptoms such as self-harm. We further acknowledge that the closed-source models used in our framework provide limited transparency regarding their architectures and training data, which constrains interpretability. Finally, because our approach performs inference using existing pretrained models rather than training large models from scratch, its environmental impact is comparatively limited.

References

Agarwal, S.; Sharma, S.; Nakov, P.; and Chakraborty, T. 2024. MemeMQA: Multimodal Question Answering for Memes via Rationale-Based Inferencing. In *Findings of the Association for Computational Linguistics: ACL 2024*, 5042–5078. Bangkok, Thailand.

Akram, U.; Drabble, J.; Cau, G.; Hershaw, F.; Rajenthran, A.; Lowe, M.; Trommelen, C.; and Ellis, J. G. 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.

Anthropic. 2024. Introducing Claude 3.5 Sonnet.

Benton, A.; Mitchell, M.; and Hovy, D. 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*, 152–162. Valencia, Spain.

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

Cashyap, S. 2024. Research: Photo posts produce significantly more engagement than link posts on Facebook.

Chen, J.; Saha, S.; and Bansal, M. 2024. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7066–7085. Bangkok, Thailand.

Chen, Z.; Deng, J.; Zhou, J.; Wu, J.; Qian, T.; and Huang, M. 2024. Depression Detection in Clinical Interviews with LLM-Empowered Structural Element Graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8181–8194. Mexico City, Mexico.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Erickson, S. J.; and Feldstein, S. W. 2006. Adolescent Humor and its Relationship to Coping, Defense Strategies, Psychological Distress, and Well-Being. *Child Psychiatry and Human Development*, 37(3): 255–271.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, S.; Tu, Y.; Han, X.; Cui, G.; He, C.; Zhao, W.; Long, X.; Zheng, Z.; Fang, Y.; Huang, Y.; Zhang, X.; Thai, Z. L.; Wang, C.; Yao, Y.; Zhao, C.; Zhou, J.; Cai, J.; Zhai, Z.; Ding, N.; Jia, C.; Zeng, G.; dahai li; Liu, Z.; and Sun, M. 2024. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. In *First Conference on Language Modeling*.

Jha, P.; Jain, R.; Mandal, K.; Chadha, A.; Saha, S.; and Bhattacharyya, P. 2024a. MemeGuard: An LLM and VLM-based Framework for Advancing Content Moderation via Meme Intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8084–8104. Bangkok, Thailand.

Jha, P.; Maity, K.; Jain, R.; Verma, A.; Saha, S.; and Bhattacharyya, P. 2024b. 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)*, 930–943. St. Julian's, Malta.

Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; and Cambria, E. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 7184–7190. Marseille, France.

Joshi, S.; Ilievski, F.; and Luceri, L. 2024. Contextualizing Internet Memes Across Social Media Platforms. In *Companion Proceedings of the ACM Web Conference 2024*, 1831–1840. New York, NY, USA: ACM. ISBN 9798400701726.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, volume 33, 2611–2624. Curran Associates, Inc.

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

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

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.

Mazhar, A.; hasan shaik, Z.; Srivastava, A.; Ruhnke, P.; Vaddavalli, L.; Katragadda, S. K.; Yadav, S.; and Akhtar, M. S. 2025. Figurative-cum-Commonsense Knowledge Infusion for Multimodal Mental Health Meme Classification. arXiv:2501.15321.

Mei, J.; Chen, J.; Lin, W.; Byrne, B.; and Tomalin, M. 2024. Improving Hateful Meme Detection through Retrieval-Guided Contrastive Learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5333–5347. Bangkok, Thailand.

Mishra, S.; Suryavardan, S.; Patwa, P.; Chakraborty, M.; Rani, A.; Reganti, A. N.; Chadha, A.; Das, A.; Sheth, A. P.; Chinnakotla, M.; Ekbal, A.; and Kumar, S. 2023. Memotion 3: Dataset on sentiment and emotion analysis of codemixed Hindi-English Memes. In *DE-FACTIFY@AAAI*.

NHS. 2025. Cognitive analytic therapy (CAT). Accessed on 15.02.2025.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

Parry, G.; Bennett, D.; Roth, A. D.; and Kellett, S. 2021. Developing a competence framework for cognitive analytic therapy. *Psychology and Psychotherapy: Theory, Research and Practice*, 94(S1): 151–170.

Passonneau, R. 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.

Pramanick, S.; Dimitrov, D.; Mukherjee, R.; Sharma, S.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2783–2796. Online.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Sarink, F.; and García-Montes, J. 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.

Sharma, C.; Bhageria, D.; Scott, W.; PYKL, S.; Das, A.; Chakraborty, T.; Pulabaigari, V.; and Gambäck, B. 2020. SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 759–773. Barcelona (online).

Sharma, S.; Agarwal, S.; Suresh, T.; Nakov, P.; Akhtar, M. S.; and Chakraborty, T. 2023a. 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.

Sharma, S.; S, R.; Arora, U.; Akhtar, M. S.; and Chakraborty, T. 2023b. MEMEX: Detecting Explanatory Evidence for Memes via Knowledge-Enriched Contextualization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5272–5290. Toronto, Canada.

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114. PMLR. ISSN: 2640-3498.

Team, G. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Wagener, A. 2021. The Postdigital Emergence of Memes and GIFs: Meaning, Discourse, and Hypernarrative Creativity. *Postdigital Science and Education*, 3: 831–850.

Wang, Y.; Inkpen, D.; and Kirinde Gamaarachchige, P. 2024. Explainable Depression Detection Using Large Language Models on Social Media Data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, 108–126. St. Julians, Malta.

Wang, Y.; Li, Y.; Gui, X.; Kou, Y.; and Liu, F. 2019. Culturally-Embedded Visual Literacy: A Study of Impression Management via Emoticon, Emoji, Sticker, and Meme on Social Media in China. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E. E.; Jiang, L.; Zhang, X.; Zhang, S.; Awadallah, A.; White, R. W.; Burger, D.; and Wang, C. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *COLM 2024*.

Xiong, M.; Hu, Z.; Lu, X.; LI, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Yadav, S.; Caragea, C.; Zhao, C.; Kumari, N.; Solberg, M.; and Sharma, T. 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)*, 8890–8905. Toronto, Canada.

Yadav, S.; Chauhan, J.; Sain, J. P.; Thirunarayan, K.; Sheth, A.; and Schumm, J. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. *arXiv preprint arXiv:2011.06149*.

Yadav, S.; Ekbal, A.; Saha, S.; Bhattacharyya, P.; and Sheth, A. 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)*, 271–277. New Orleans, Louisiana.

Yadav, S.; Lokala, U.; Daniulaityte, R.; Thirunarayan, K.; Lamy, F.; and Sheth, A. 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.

Yang, K.; Zhang, T.; Kuang, Z.; Xie, Q.; and Ananiadou, S. 2023. MentalLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. arXiv:2309.13567.

Yates, A.; Cohan, A.; and Goharian, N. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2968–2978. Copenhagen, Denmark.

Zhang, T.; Yang, K.; Alhuzali, H.; Liu, B.; and Ananiadou, S. 2023. PHQ-aware depressive symptoms identification with similarity contrastive learning on social media. *Information Processing Management*, 60(5): 103417.

Zhong, Y.; and Baghel, B. K. 2024. Multimodal Understanding of Memes with Fair Explanations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2007–2017. Seattle, WA, USA: IEEE. ISBN 9798350365474.