MemeCourt: a VLMs-Based Court with Multi-Agent Collaboration Framework for Harmful Meme Detection

Anonymous ACL submission

Abstract

Internet harmful memes have become pervasive across social media owing to their visual appeal and satirical content. Unlike direct hateful images or text, multimodal memes often embed implicit content, making their detection a novel challenge. Despite extensive exploration of existing methods for harmful meme detection, most lack interpretability in judgments. To bridge this gap, we introduce MemeCourt: a Vision-Language Models (VLMs-) Based Court with Multi-Agent Collaboration Framework. MemeCourt enhances the extraction of implicit features through a reasoning pipeline: the Proposer-Agent engages in multi-round interactive questioning with the Accuser-agent and Defender-Agent, who each generate evidence supporting their respective stances. The Judge-Agent then integrates these evidences with precedent cases to make a final judgment on meme harmfulness. Experiments on three publicly available meme datasets demonstrate that our approach achieves SOTA performance, and improves interpretability by tracing the explicit judging process.

Disclaimer: This paper contains content that may be disturbing to some readers.

1 Introduction

The term "meme" was initially coined by evolutionary biologist Richard Dawkins (Dawkins, 1981) to describe a mode of cultural information transmission via imitation. However, with the evolution of online information dissemination, this concept has been adapted by malicious actors to spread harmful content, making hateful memes increasingly pervasive across social media (Shifman, 2013). Unlike directly hateful textual or visual content, hateful memes thrive on users' creativity and implicit cultural awareness (Duchscherer and Dovidio, 2016).

In response to the urgent need for multimodal hateful meme detection, representative works like MOMENTA (Pramanick et al., 2021b) and DISARM (Sharma et al., 2022) employ deep multimodal neural networks. These models achieve strong results on specific datasets but suffer from opaque judging processes, rendering them black boxes that lack convincing justifications for their detection outcomes. While the explainable model EXPLAINHM (Lin et al., 2024) uses multimodal debate via Large Language Models (LLMs) to generate contradictory rationales for interpretability, its reasoning remains incomplete and ultimately relies on conventional binary classifiers for the final decision, limiting the completeness of its interpretability.

We analyze the above challenges from two key perspectives. **First**, existing models cannot effectively capture the implicit relationships between images and texts, resulting in an incomplete understanding of harmful content. For example, as demonstrated in Figure 1 (a), the sarcastic and discriminatory information in memes cannot be understood apart from any of the modalities. **Second**, the models lack a coherent and transparent reasoning chain, resulting in incomplete and less interpretable reasoning processes. This ultimately undermines the credibility of their predictions.

To bridge the gaps, we introduce a Vision-Language Models (VLMs-)Based Court with Multi-Agent Collaboration Framework (**MemeCourt**) inspired by judicial adjudication processes, where four pre-trained VLM-based or LLM-based agents, including Proposer-Agent, Accuser-Agent, Defender-Agent, and Judge-Agent, collaborate in an adjudication pipeline to enhance implicit feature extraction and interpretability. The detection process is illustrated in Figure 1 (b). First, the Proposer-Agent performs semantic extraction on the meme image to obtain an initial objective description, which is then communicated separately to the Accuser-Agent and the Defender-Agent. Each agent engages in multi-round interactive ques-

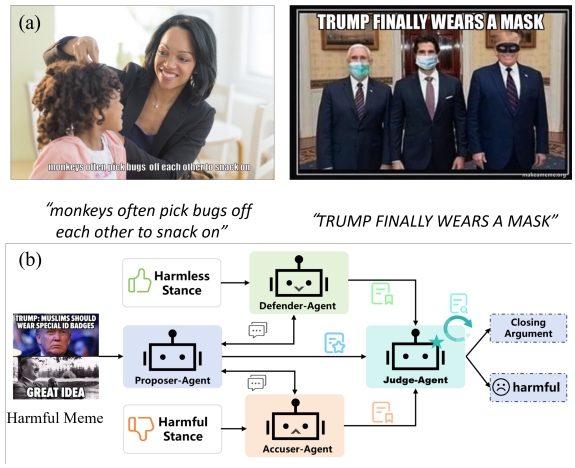

Figure 1: (a) Illustration of memes. The first one is harmful, conveying racial discrimination through the use of offensive metaphors. The second one is harmful, adopting gratuitous vilification to express a politically biased stance. (b) A brief illustration of MemeCourt.

tioning, guided by their priori-stances, to uncover subjective evidence from their own perspectives. Finally, the Judge-Agent integrates the views from both sides along with the objective description, and combines this information with precedent cases to deliver a well-reasoned verdict. Our contributions can be summarized as follows:

- **Innovative Multi-Agent Collaboration Framework:** We propose a novel VLM-based framework inspired by judicial adjudication. It introduces four agents that engage in multi-round questioning to generate evidence, with the Judge-Agent leveraging a retrieval-augmented generation (RAG) mechanism to incorporate precedent cases in its final judgment.

- **Interpretable Reasoning via Traceable Adjudication:** By modeling detection as a courtroom debate, MemeCourt constructs a transparent reasoning chain. The agents provide objective descriptions, uncover subjective evidence, and combine arguments with precedents to yield traceable and human-understandable interpretations, thus addressing the limitations of black-box models.

- **State-of-the-Art Performance and Open-Source Availability:** MemeCourt achieves SOTA performance on three public meme datasets and releases open-source code to promote reproducibility and further research.

2 Related Work

Harmful Meme Detection. The task gained significant traction due to the widespread dissemination of virulent multimodal content across social platforms (Arora et al., 2023). Early approaches relied on unimodal analyses of either visual or textual features, but these struggled to capture the synergistic interplay of meme components. Subsequent advances shifted to multimodal deep learning frameworks like MOMENTA (Pramanick et al., 2021b), DISARM (Sharma et al., 2022), and ISM(Yang et al., 2023) to improve accuracy via transformer-based fusion and contextualized entity embeddings. However, these models remain opaque black boxes, lacking traceable reasoning chains critical for trustworthy content moderation. Recent efforts have turned to multimodal large language models (MLLMs), employing prompt engineering (Ji et al., 2023), fine-tuning strategies(Lin et al., 2023; Huang et al., 2024), and multi-agent debate frameworks (Lin et al., 2024) to enhance interpretability. Yet, these still fail to construct complete, coherent reasoning pipelines, limiting their reliability. Our work bridges this gap by introducing MemeCourt, a framework that instantiates a reasoning chain to enable transparent and human-understandable verdict-making in harmful meme detection.

Multi-agent Collaboration. In recent years, LLM-based agents have emerged as a promising paradigm for task-oriented automation. Researchers have explored diverse strategies to enhance coordination by leveraging pre-trained LLMs, including multi-turn agent-human communication (Wang et al., 2024), contrastive reasoning(Wu et al., 2024), agent teams optimization algorithms(Liu et al., 2024)demonstrating synergistic advantages in domains like recommender systems (Fang et al., 2024) and socially sensitive decision-making(Piatti et al., 2024). However, no prior work has applied this paradigm to harmful meme detection. Our approach is the first to introduce a courtroom-inspired multi-agent collaboration framework. Building on the synergistic advantages of multi-agent systems, our approach deploys autonomous agents (Proposer-Agent, Accuser-Agent, Defender-Agent, Judge-Agent) to simulate distinct judicial roles. This collaboration enables transparent adversarial argumentation, systematic evidence aggregation, and precedent-grounded adjudication, enhancing both

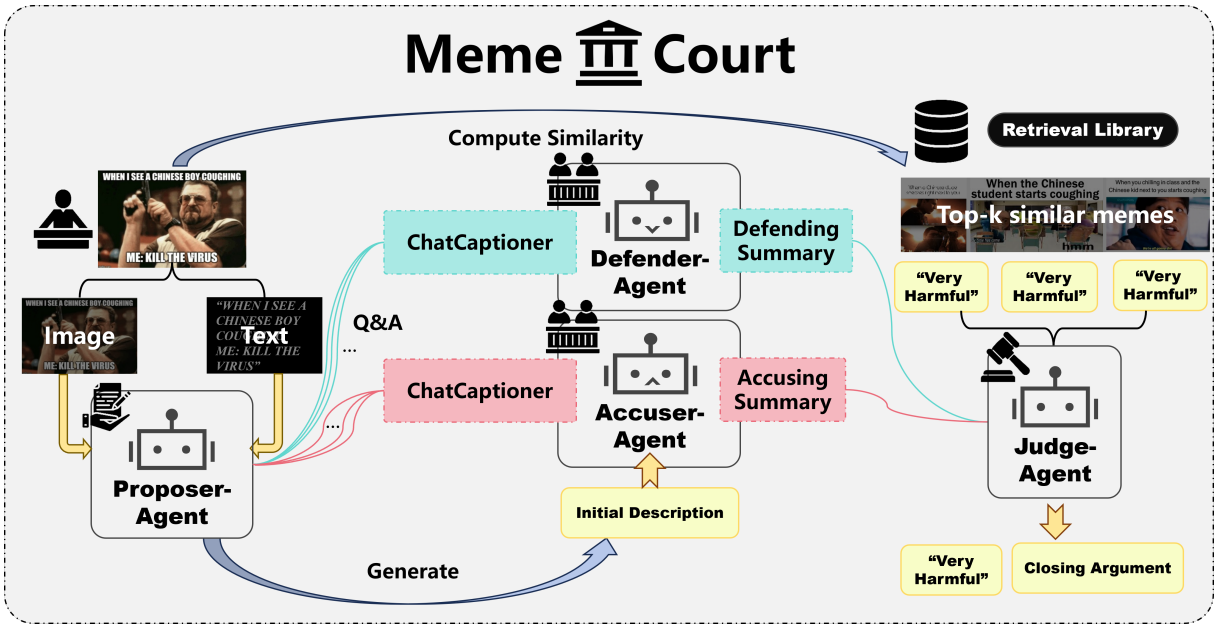

Figure 2: Overview of our method. The overall framework is a verdict pipeline consisting of "Proposer-Defender/Accuser-Judge" with mechanisms of ChatCaptioner and RAG to enhance multimodal information understanding.

interpretability and detection accuracy by modeling human-like deliberation.

3 Our Approach

3.1 Overview

Based on the above insights, we formulate harmful meme detection as a multimodal reasoning task grounded in natural language processing(NLP). Given a meme dataset $M = \{T, G\}$, where T represents the textual component and G the corresponding visual content, our approach aims to generate a verdict output consisting of its label and a detailed analysis for its judgment. This output not only classifies the meme but also provides a traceable justification for the verdict. To achieve this, we design a multi-agent adjudication framework inspired by courtroom proceedings, namely "MemeCourt", proposing a "Proposer–(Defender/Accuser)–Judge pipeline". Each agent is implemented by an independent LLM or a VLM, simulating a collaborative judicial process. The Proposer-agent conducts the initial semantic interpretation of multimodal content of the meme. The Defender and Accuser agents gather subjective and contextual evidence via iterative inquiry, namely ChatCaptioner (Zhu et al., 2023) mechanism. And the Judge-Agent integrates both subjective elaborations and objective references(via RAG mechanism) to render the final verdict and closing argument. This modular framework enables the incorporation of both subjective reasoning and objective retrieval, while maintaining transparent reasoning logs across all stages. As a result, the system exhibits high interpretability and provides a complete chain of reasoning for each adjudication.

The overview of our framework is shown in Figure 2. It consists of Proposer-Agent(§ 3.2), Defender-Agent/Accuser-Agent with ChatCaptioner mechanism(§ 3.3), and Judge-Agent with RAG mechanism(§ 3.4).

3.2 Proposer-Agent

We design an agent named Proposer based on a pre-trained multimodal large language model to perform preliminary understanding and extraction of visual content. Given a meme image $M = \{T, G\}$, Proposer-Agent takes the image G as input, along with a fixed prompt template p_{start} designed to elicit strictly objective visual descriptions. The prompt instructs the model to describe visible elements in the image without inference.

Formally, the output is a natural language description $D = Proposer\,(G, p_{start})$, which is stored as a structured JSON field in our reasoning pipeline. This stage deliberately excludes text content T to avoid introducing subjective bias into early-stage interpretation.

Algorithm 1 ChatCaptioner Mechanism Between an Agent and the Proposer-Agent

Input: Meme image $M = \{T, G\}$,
Initial description D,
Minimum and maximum rounds $R_{\min}, R_{\max}$,
Entropy threshold ϵ,
Role agent A (Defender or Accuser),
Multimodal-LLM agent P (Proposer)
Output: Final summary $S = \{\texttt{h\_score}, \texttt{jst}\}$

Initialize dialogue history $\mathcal{H} \leftarrow \emptyset$
$S_{\text{last}} \leftarrow \texttt{None}$
for $r = 1$ to $R_{\max}$ **do**
  Generate query $q_r \leftarrow A.\texttt{Ask}(\mathcal{H}, D, T)$
  Generate answer $a_r \leftarrow P.\texttt{Answer}(q_r, G)$
  Append (q_r, a_r) to $\mathcal{H}$
  Generate $S_r \leftarrow A.\texttt{Summarize}(\mathcal{H})$
  if $r \geq R_{\min}$ **then**
    if $S_{\text{last}} \neq \texttt{None}$ **then**
      Compute $\delta \leftarrow \Delta\texttt{Entropy}(S_r, S_{\text{last}})$
      if $\delta < \epsilon$ **then**
        break
      end if
    end if
  end if
  $S_{\text{last}} \leftarrow S_r$
end for
return S_r

3.3 Debate between Defender-Agent and Accuser-Agent

Based on LLMs, we devise two agents with opposing stances: Defending and Accusing. Given a meme description D generated by the Proposer-Agent, the Defender-Agent and Accuser-Agent are envisioned to debate the proposition of whether the meme M is harmful. Similar to the roles of plaintiff's lawyers and defendant's lawyers in courtrooms, both agents gather evidence by engaging in iterative inquiries with the Proposer-Agent, which we refer to as the ChatCaptioner(Zhu et al., 2023).

Specifically, we instruct the Defender-Agent and Accuser-Agent(LLMs) to generate questions targeting the initially described image in order to maxisimize the information they get, each driven by their respective priori-stances. These questions are then answered by the Proposer-Agent(VLM). We introduce an automatic termination mechanism for the ChatCaptioner process based on information entropy. After each round of dialogue, the Defender and Accuser agents generate a summary based on the initial description and accumulated dialogue history. We compute the difference in information entropy between the current and previous summaries. When this difference falls below a predefined threshold, we consider that the conversation has converged to the maximum degree of information and therefore terminate the interaction. The entropy $\texttt{Entropy}(s_r)$ is calculated as follows:

Let $S_r = \{s_1, s_2, \ldots, s_k\}$ be the set of unique characters in the text, $p_i = \frac{n_i}{N}$, $N = \sum_{i=1}^{k} n_i$ where N is the total number of characters, n_i is the number of occurrences of a character.

$$\texttt{Entropy}(s_r) = -\sum_{i=1}^{k} \frac{n_i}{N} \log_2 \left(\frac{n_i}{N}\right) \quad (1)$$

We argue that this stance-guided interaction significantly improves semantic grounding, thereby enhancing the overall effectiveness of multimodal information extraction and interpretation in our framework. The process can be represented by Algorithm 1.

After this interaction, each agent outputs S_r^D or S_r^A, containing a summary describing the meme more precisely and a textual justification for the priori-stances.

Algorithm 2 Multimodal Retrieval-Augmented Generation for Judgment

Input: Subjective arguments Sr_A, Sr_D from Accuser-Agent and Defender-Agent;
Objective facts D, T, G from Proposer-Agent;
Retrieval database $\mathcal{R}$ (disjoint from test set)
Output: Final decision $C = \{\text{label}, \text{analysis}\}$

Encode T and G into embedding q_{text}, q_{image}
Initialize list of retrieval candidates $\mathcal{C} = [\quad]$
for each (T_i, G_i) in $\mathcal{R}$ **do**
  Encode T_i and G_i into e^i_{text}, e^i_{image}
  Compute $s_i = \alpha \cdot \text{sim}(q_{\text{text}}, e^i_{\text{text}}) + (1-\alpha) \cdot \text{sim}(q_{\text{image}}, e^i_{\text{image}})$
  Append (T_i, G_i, s_i) to $\mathcal{C}$
end for
Select top-k cases with highest s_i from $\mathcal{C}$ as $\mathcal{C}_{\text{top}}$
Construct input prompt by concatenating Sr_A, Sr_D, D, T, G, and $\mathcal{C}_{\text{top}}$
Generate final decision $C = \text{Judge}(\text{prompt})$

3.4 Rational Verdict by Judge-Agent

Inspired by the role of a judge in courtroom proceedings, we introduce a VLM-based agent called

Judge to make the final verdict. The **Judge-Agent** receives (1) *subjective arguments* from the Defender-Agent (S_r^D) and the Accuser-Agent (S_r^A), each derived from their respective reasoning processes, and (2) *objective evidence* including the initial visual description D, meme text T, and image content G, all provided by the Proposer-Agent. Inspired by the judicial practice of referencing precedent cases during verdict-making, we integrate a multimodal retrieval-augmented generation (RAG) mechanism.

For RAG, we encode both the textual summaries S_r, and the meme image using a pretrained CLIP model. Then, we compute their respective similarities with all samples in the retrieval library. These similarity scores are merged via a weighted average using a predefined ratio. Based on the resulting combined scores, we select the top k most similar meme samples, which are subsequently incorporated into the Judge-Agent's input prompt for final verdict-making. The procedure of RAG mechanism is detailed in Algorithm 2. RAG enables **Judge-Agent** to retrieve similar historical cases from a dedicated, non-overlapping subset of the dataset and incorporate them into the reasoning process as references.

The final output of Judge-Agent for a given meme image is represented as C, consisting of the label and the detailed analysis, where the "label" denotes the predicted harmfulness of the meme, and the "detailed analysis" provides the rationale behind the verdict, analogous to a closing argument.

4 Experiments

4.1 Datasets and Experimental Settings

Datasets. To evaluate the effectiveness of our proposed framework, we conducted comparative experiments on three publicly available meme datasets: Harm-C(Pramanick et al., 2021a), Harm-P(Pramanick et al., 2021b), and FHM(Kiela et al., 2020). Harm-C focuses on COVID-19-related memes, Harm-P on U.S. political memes, and FHM covers a broader range of topics. The FHM dataset provides binary labels, with each meme annotated as either 1 (harmful) or 0 (harmless). In contrast, Harm-C and Harm-P offer three-level annotations: "not harmful", "somewhat harmful", and "very harmful". Since our method relies entirely on pretrained LLMs and VLMs without any fine-tuning or supervised learning, its performance is highly sensitive to the reliability of the dataset labels. To ensure the clarity and extremity of class definitions, we excluded all samples labeled as "somewhat harmful" from Harm-C and Harm-P, retaining only those annotated as either "not harmful" or "very harmful".

Implementation. We evaluated our approach using standard classification metrics, including accuracy and Macro F1-score. To assess the effect of different model capacities, we conducted comparative experiments using three variants of the Qwen2.5-VL multimodal model: 3B, 7B, and 32B. Through several attempts, we adopt Qwen2.5-VL 32B and set k of RAG selection to 3.

4.2 Experimental Results

4.2.1 Baselines

We compare our method with some models, including unimodal ones and multimodal ones. For unimodal models, we consider Text BERT(Devlin et al., 2019) as text-only model, which inputs the text matched with meme into the pretraining language model BERT, and completes the binary classification task through fine-tuning.

For image-only models, we consider Image-Region(He et al., 2016; Ren et al., 2017), which focuses on local area information (such as face, object, text, etc.) in meme.

For multimodal models, we consider: 1) Late Fusion(Pramanick et al., 2021a), using independent visual models to extract image features, and language models to extract meme text features; 2) MMBT(Kiela et al., 2019), sending image and text embeddings to BERT as a whole for joint coding and classification. 3) VisualBERT(Li et al., 2019; Lin et al., 2014), with image and text information jointly built through a unified Transformer encoder. 4) VilBERT(Lu et al., 2019; Sharma et al., 2018), processing image and text input independently through two parallel Transformer coding streams; 5) MOMENTA(Pramanick et al., 2021b), using CLIP to obtain global image–text embeddings; 6) MaskPrompt(Cao et al., 2023b; Liu et al., 2019), a prompt based framework; 7) Pro-Cap(Cao et al., 2023a), with a frozen vision–language model and a lightweight text classifier; 8) ExplainHM(Lin et al., 2024), which built a debate framework.

4.2.2 Overall Performance

Table 1 presents a comparison of our method with several previous state-of-the-art (SoTA) approaches in terms of accuracy across three datasets. Unimodal methods generally perform worse than mul-

Table 1: Harmful meme detection result in 3 different datasets. The accuracy(%) are reported as the metric. The best and second best result are in bold and in underlined, respectively.

Dataset	Harm-C		Harm-P		FHM	
Model	Acc.	Mac-F1	Acc.	Mac-F1	Acc.	Mac-F1
Text BERT(Devlin et al., 2019)	70.17	66.25	80.12	78.35	57.12	41.52
Image-Region(He et al., 2016)	68.74	62.97	73.14	72.77	52.34	34.19
Late Fusion(Pramanick et al., 2021a)	73.24	70.25	78.26	78.5	59.14	44.81
MMBT(Kiela et al., 2019)	73.48	67.12	82.54	80.23	65.06	61.93
VisualBERT(Li et al., 2019)	81.36	80.13	86.80	86.07	61.48	47.26
ViLBERT(Lu et al., 2019)	78.70	78.09	87.25	86.03	64.7	55.78
MOMENTA(Pramanick et al., 2021b)	83.82	82.80	89.84	88.26	61.34	57.45
MaskPrompt(Cao et al., 2023b)	84.47	81.51	88.17	87.09	72.98	65.24
Pro-Cap(Cao et al., 2023a)	85.01	83.17	89.32	87.91	74.95	71.68
EXPLAINHM(Lin et al., 2024)	87.00	86.41	90.73	90.72	75.60	75.39
MemeCourt	**88.76**	**88.75**	**92.99**	**92.83**	**77.19**	**75.69**

timodal methods, as the integrated understanding of visual and textual information is essential for interpreting the nuanced connotations of memes. The multimodal models outperformed the unimodal models in the second group. EXPLAINHM, by comparison, introduced a debate-based framework that leverages large language models to generate both supporting and opposing arguments before synthesizing a final judgment, achieving notable improvements in both classification accuracy and model interpretability.

Our method, MemeCourt, outperforms the previous state-of-the-art model EXPLAINHM by 0.62%, 2.26%, and 1.59% in terms of accuracy on the Harm-C, Harm-P, and FHM datasets, respectively. We attribute the superior performance of MemeCourt to two key factors. First, the large language models nowadays possess the capacity for ethical reasoning and value-based judgment. Second, the courtroom-inspired multi-agent structure enables the Judge-agent to synthesize multiple perspectives and arrive at informed, balanced verdicts. Importantly, our model surpasses all baselines in interpretability, providing a coherent chain of reasoning that culminates in a "closing argument"-a final justification that facilitates human understanding and enhances the traceability of the verdict-making process.

4.3 Ablation Studies

We perform ablation experiments on the Harm-C dataset. The results of the ablation experiments are demonstrated in Table 2.

To assess the contribution of each key component within the proposed framework, we conducted ablation studies by removing individual modules from MemeCourt. Specifically, we designed three ablated variants: (1)*w/o RAG* a version without the RAG mechanism, (2)*w/o D/A* a version without the Defender and Accuser agents, (3)*w/o ChatCaptioner(CC)* a version without ChatCaptioner mechanism, and (4) a dull Judge, where the VLM independently determines the harmfulness of memes. All other variables are held constant across experiments to ensure fair comparison.

Table 2: Results of ablation and exploring experiments

Dataset	Harm-C	
Model	Acc.	Mac-F1
MemeCourt	88.76	88.75
w/o RAG	80.23	63.29
w/o D/A	63.17	53.19
w/o CC	74.58	71.44
Dull Judge	54.97	49.16

The individual agents within the MemeCourt framework, along with the two key mechanisms, ChatCaptioner and RAG, are all essential components. The absence of any of these leads to a notable decline in classification accuracy. In the *w/o RAG* group, the performance drop suggests that referring to previously adjudicated meme cases can effectively enhance the Judge-Agent's verdict-making capability. The *w/o D/A* group demonstrates that bilateral argumentation, by incorporating both defending and accusing perspectives, offers the Judge-Agent more diverse viewpoints and contributes to more objective and comprehensive

Table 3: Results of exploring experiments

Dataset	Harm-C	
Model	Acc.	Mac-F1
Qwen-VL-3B	76.90	76.72
Qwen-VL-7B	87.62	87.61
Qwen-VL-32B	88.76	88.75
priori-stance	88.76	88.75
post-stance	44.37	31.55

judgments. In the *w/o ChatCaptioner(CC)* group, the lack of multi-turn Q&A significantly reduces the agents' ability to extract implicit visual semantics and to thoroughly integrate the multimodal content of meme images. Finally, the *Dull Judge* group reveals the limitation of relying solely on a VLM for harmful meme detection, indicating that such direct judgment without interaction leads to less reliable outcomes.

4.4 Discussions

4.4.1 Impact of Model Size

Given that our MemeCourt framework operates in a zero-shot setting, it inherently relies heavily on the capabilities of pre-trained large language models. To evaluate the impact of model capacity on performance, we further compare the results using Qwen-VL with 3B, 7B and 32B parameters as the backbone for our agents on the Harm-C dataset.

As shown in the second part of Table 3, overall performance improves with the increase in model size (e.g., from 7B to 32B). We attribute this trend to the framework of MemeCourt being fundamentally dependent on VLMs. Larger models, which are typically trained with greater capacity and on more extensive data, provide stronger language and visual reasoning capabilities. As a result, each agent in the MemeCourt framework can generate more accurate judgments and produce more precise outputs.

4.4.2 Impact of Stance-Input Timing

As part of our exploration, we conducted an experiment to investigate how the timing of role assignment (i.e., defending or accusing) affects agents' behavior. Specifically, instead of informing the Defender-Agent and Accuser-Agent of their stances prior to initiating the Q&A process, we delayed the role assignment until the summary stage. This modification led to significantly different outcomes, as quantitatively demonstrated in third part of Table 3, while Figure 3 shows the difference

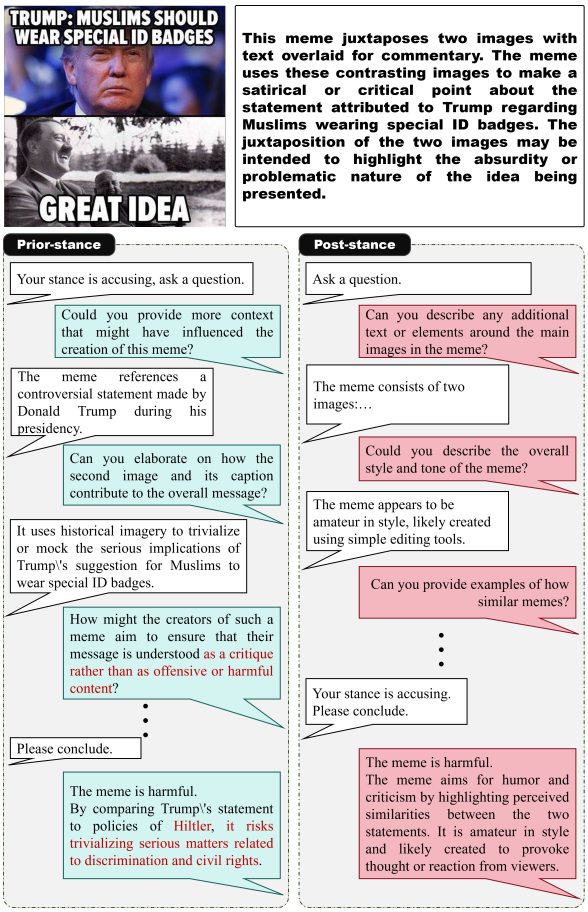

Figure 3: Differences in dialog performance of ChatCaptioner with and without a priori-stance

between the questions asked with or without an a priori-stance.

When agents are assigned an a priori-stance, their questioning becomes more targeted, which facilitates more efficient extraction of relevant information. The number of Q&A rounds is generally reduced when agents are pre-assigned stances, suggesting that targeted questioning accelerates convergence by improving information acquisition efficiency.

4.5 Interpretability Analysis

In addition to achieving detection performance comparable to existing state-of-the-art methods, our approach offers a key advantage in interpretability. By simulating courtroom scenarios, the multi-agent collaboration provides a coherent and transparent reasoning chain for the task of harmful meme detection. As illustrated in Figure 4, the reasoning process for memes_8260.png in the Harm-P dataset demonstrates how the system reaches its final decision through agent interaction.

The Proposer-Agent initially fails to identify the

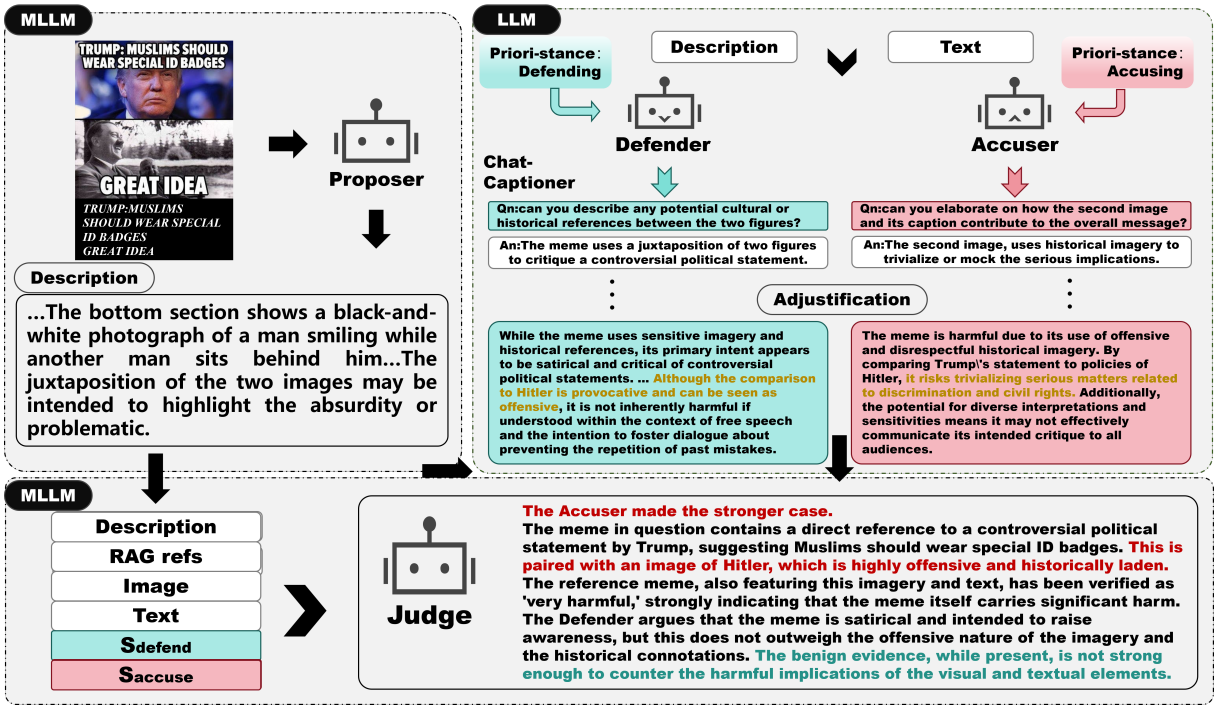

Figure 4: An example of reasoning chain of our method. Our approach provides a complete chain of traceable reasoning logic that exhibits strong interpretability.

key historical figure, Hitler, in the meme image, indicating an incomplete semantic understanding of the visual content. Through iterative Q&A rounds with ChatCaptioner, however, both the Accuser-Agent and Defender-Agent gradually extract this crucial information from the Proposer's responses and build their respective analyses. Interestingly, the Defender-Agent, which is in a disadvantaged position in this specific case, engages in significantly more interaction rounds (10 versus. 6) compared to the Accuser-Agent. This suggests a slower convergence in evidence gathering, likely due to the difficulty of constructing a convincing defense. The distribution of the Q&A rounds with respect to the agent and label of ground-truth can be demonstrated in Figure 5. Qualitatively, the Defender's arguments appear forced and less coherent, whereas the Accuser's reasoning is more concise and persuasive. The Judge adopts the Accuser-Agent's position and delivers a decisive final verdict.

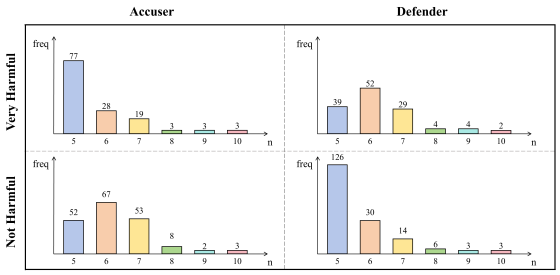

Figure 5: Distribution of the Q&A rounds. Agent that has the advantage in the debate usually goes through fewer Q&A rounds, suggesting that its dominance leads to a faster convergence to maximum informativeness.

5 Conclusion

We propose MemeCourt, a novel framework for detecting harmful memes that simulates a courtroom setting. Built upon a multi-agent collaboration system grounded in VLMs, MemeCourt organizes its verdict-making process through a "Proposer–(Defender/Accuser)–Judge" pipeline that mirrors judicial reasoning. Evaluated on three widely-used benchmark datasets, our framework achieves strong performance, demonstrating its effectiveness. Ablation studies further validate the critical roles of individual components, such as the ChatCaptioner module, the retrieval mechanism, and dual-agent debates, in enhancing both accuracy and interpretability. Most importantly, MemeCourt produces a transparent, and traceable reasoning chain, offering a new perspective on enhancing interpretability in harmful meme detection, and potentially in broader multimodal judgment tasks.

Limitations

While MemeCourt makes notable progress in interpretability for harmful meme detection, several limitations remain.

(1) Notably, MemeCourt operates in a zero-shot setting without any fine-tuning, relying entirely on pretrained VLMs to perform ethical inference through interaction. In future work, we plan to lightly fine-tune certain agent models to better adapt them to the specific task of harmful meme detection, thereby improving task sensitivity and robustness.

(2) Additionally, the multi-agent framework in MemeCourt may suffer from inconsistency among agents' judgments, especially in cases involving subtle, culturally dependent, or ambiguous memes. Such disagreements can reduce the overall stability and reliability of the system's decisions.

(3) The interaction-based reasoning process introduces nontrivial computational overhead, as each instance requires multiple rounds of agent communication and deliberation. This limits the scalability of MemeCourt in real-time or large-scale deployment scenarios.

(4) Finally, the agents in MemeCourt heavily rely on the representations and biases embedded in their underlying pretrained models. As a result, their ethical reasoning is constrained by the limitations of the original training data, which may not fully capture the nuances of harmful content across diverse social and cultural contexts.

References

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, and Isabelle Augenstein. 2023. Detecting harmful content on online platforms: What platforms need vs. where research efforts go. *ACM Comput. Surv.*, 56(3).

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023a. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023b. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*.

Richard Dawkins. 1981. Selfish genes and selfish memes. *The mind's I: Fantasies and reflections on self and soul*, pages 124–144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Katie M Duchscherer and John F Dovidio. 2016. When memes are mean: Appraisals of and objections to stereotypic memes. *Translational Issues in Psychological Science*, 2(3):335.

Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A multi-agent conversational recommender system. *Preprint*, arXiv:2402.01135.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jianzhao Huang, Hongzhan Lin, Ziyan Liu, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with lmm agents. *Preprint*, arXiv:2411.05383.

Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying creative harmful memes via prompt based approach. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3868–3872, New York, NY, USA. Association for Computing Machinery.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

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. *Advances in neural information processing systems*, 33:2611–2624.

Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, and 1 others. 2021. Mmocr: a comprehensive toolbox for text detection, recognition and understanding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3791–3794.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2359–2370, New York, NY, USA. Association for Computing Machinery.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. *Preprint*, arXiv:2312.05434.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024. A dynamic llm-powered agent network for task-oriented agent collaboration. *Preprint*, arXiv:2310.02170.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. *Advances in Neural Information Processing Systems*, 37:13302–13320.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. In *Advances in Neural Information Processing Systems*, volume 37, pages 111715–111759. Curran Associates, Inc.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. *Preprint*, arXiv:2110.00413.

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. *arXiv preprint arXiv:2109.05184*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. Disarm: Detecting the victims targeted by harmful memes. *arXiv preprint arXiv:2205.05738*.

Limor Shifman. 2013. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of computer-mediated communication*, 18(3):362–377.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024. Executable code actions elicit better LLM agents. In *Forty-first International Conference on Machine Learning*.

Shirley Wu, Shiyu Zhao, Qian Huang, Kexin Huang, Michihiro Yasunaga, Kaidi Cao, Vassilis N. Ioannidis, Karthik Subbian, Jure Leskovec, and James Zou. 2024. Avatar: Optimizing llm agents for tool usage via contrastive reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 25981–26010. Curran Associates, Inc.

Chuanpeng Yang, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2023. Invariant meets specific: A scalable harmful memes detection framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 4788–4797, New York, NY, USA. Association for Computing Machinery.

Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.

A Appendix

A.1 Baseline details

In this study, we constructed and compared the performance of a baseline model using the binary classification task Accuracy and Macro F1 score (Mac-F1) from three datasets: Harm-C, Harm-P, and FHM as evaluation metrics.

Text BERT(Kuang et al., 2021) is a single mode text baseline method, which only uses the original text in meme to judge the harmfulness. This method inputs the text matched with meme into the pretraining language model BERT, and completes the binary classification task through fine tuning. The model relies on the aggressive or hate tendency expressed by the language content itself, but it cannot capture the implicit semantics conveyed in the image, so its performance is limited in the face of complex memes such as semantic irony, inconsistent graphics and text.

Image Region(He et al., 2016) is an image based unimodal method that focuses on local area information (such as face, object, text, etc.) in meme. Usually, key areas are located by means of object detection or image segmentation, and their visual features are extracted to complete the hazard identification. Although this method can capture the detail signal in the image, due to the lack of understanding of the meme text content, it is easy to miss the implicit attack intention brought by the combination of image and text, resulting in the overall judgment is not accurate enough, especially when the text information is ironic or negative.

Late Fusion(Pramanick et al., 2021a) is a multimodal baseline method, which is used in this task to process images and texts separately before fusion. Specifically, the method uses independent visual models (such as ResNet) to extract image features, and language models (such as BERT) to extract meme text features. Both are trained independently, and finally feature stitching a weighted combination is performed at the classification level to complete the recognition task of harmful or attacking objects. Although this method is simple to implement and has the advantage of modularity, it is difficult to capture complex semantic relationships such as cross modal irony and irony because the image text semantics are not fully interactive in the early stage, so it is not as good as the deep fusion method on data sets such as HarMeme.

MMBT(Kiela et al., 2019) is a concise and efficient multimodal baseline model, which can enhance the text dominated multimodal classification task by introducing image information into the text coding process. Its core approach is to project the image features into the same embedding space as the text token on the premise of keeping the pre training weight of the text encoder unchanged, and splice them to the original text as a "pseudo token", and then send them to BERT as a whole for joint coding and classification. This method has the advantages of simple structure, high training efficiency, easy expansion, and achieves or approaches SOTA performance in multiple text dominated multimodal tasks. Although it does not use the complex cross modal pre training mechanism like ViLBERT, MMBT still shows good multi-modal understanding ability, especially on the test set designed for the relationship between images and texts.

ViLBERT(Lu et al., 2019) is a dual stream multimodal model designed for joint modeling of vision and language, which extends the classical BERT architecture. The model processes image and text input independently through two parallel Transformer coding streams, and realizes cross modal information interaction through co interactive transformer layers in the middle. On the Conceptual Captions dataset, which is automatically collected on a large scale, ViLBERT conducts pre training with the help of two proxy tasks to learn the common visual language alignment representation, and then migrates to multiple downstream tasks through a few structural changes, such as visual question answering, visual common sense reasoning, reference parsing, and caption based image retrieval, all of which have achieved significant performance improvements.

MOMENTA(Pramanick et al., 2021b) detects harmful memes and their attack targets by: using CLIP to obtain global image–text embeddings; extracting local cues—face/ROI features via Google Vision API + VGG-19 and image-attribute text via DistilBERT—then selecting and fusing the most relevant ones with self-attention; integrating global and local cues through a Cross-Modal Attention Fusion (CMAF) module; and jointly predicting meme harmfulness and target category with a dual-task head trained with focal loss. ROIs, attributes, and CMAF each boost Accuracy, Macro-F1, and MMAE, and the model generalizes well across Harm-C and Harm-P datasets.

Visial Bert(Li et al., 2019) is a multimodal pretraining model that fuses image and text, as one of the multimodal baselines in this paper. This model

encodes the image input of meme as visual features (usually extracted from the image area features of the object detection model), and uses the text of meme as language input. The two models are jointly built through a unified Transformer encoder. Visual BERT can capture the explicit correspondence between images and texts, and is suitable for meme classification tasks with highly consistent images and texts. However, in meme recognition tasks that face implicit semantics or have complex attack intentions and delicate image text relationships, their performance is limited by the lack of optimization for hate semantics during pretraining, and their interpretability and antagonism robustness are low.

MaskPrompt(Cao et al., 2023b) is a prompt-based framework designed for multimodal hateful meme classification. It leverages the implicit knowledge embedded in pretrained language models (PLMs), such as RoBERTa, by converting multimodal inputs into textual prompts. Specifically, images are first transformed into textual descriptions (captions), which are then combined with the meme's original text. These combined texts are structured into prompts that guide the PLM to classify the meme as hateful or not. By utilizing simple prompts and a few in-context examples, PromptHate effectively exploits the PLM's understanding without the need for extensive fine-tuning. Experimental results demonstrate that PromptHate achieves a high AUC of 90.96, outperforming several state-of-the-art baselines on hateful meme classification tasks.

Pro-cap(Cao et al., 2023a) works in two stages: a frozen vision–language model tackles each meme in a zero-shot VQA setting, asking probing questions about hate-prone attributes such as race, gender, or religion; the answers are concatenated into a target-centric caption rich in key cues; the Pro-Cap plus the meme's original text are fed into a lightweight text classifier (BERT or PromptHate) that performs hateful-meme detection purely via language modeling. The approach avoids fine-tuning large PVLMs and sidesteps costly entity/demographic labeling, yet achieves sizable Accuracy and AUC gains on FHM, MAMI, and HarM benchmarks. It is more robust to real-world noise, generalizes better across datasets, and its target-focused captions provide clear interpretability.

ExpainHM(Lu et al., 2024) The research first introduced "Multimodal Debate" (MD) on the visual language model LLaVA, making the model generate contradictory interpretations from the positions of harmless and harmful respectively; Then, the peer model acts as "LLM Judge", comparing the persuasiveness of the two explanations and giving a preliminary judgment of harmfulness; Then, the preference sorted interpretation is input into the small LM judge with smaller parameters together with the original text and image, and the final classification is completed through cross attention fusion of visual and text information and only fine-tuning the small model. In the experiment, Accuracy and Macro-F1 were used to evaluate the detection performance of harmful memes on Harm-C, Harm-P and FHM datasets, and GPT-4 and manual scoring were used to explain the quality in five dimensions of Informativeness, Readability, Soundness, Consistency and Persistence; The ablation experiment further verified the contribution of each submodule to the overall effect.

A.2 Prompts

Proposer

System Prompt: You are a visual content analyzer. Your role is to accurately recognize and describe the content of an image, including objects, text, and especially any specific individuals or recognizable public figures.

Your primary responsibility is to understand the image content in detail, so you can answer questions about it accurately. You should focus on identifying important elements, such as:

- Faces and identities of real-world political figures or celebrities (if clearly shown)
- Any notable symbols, settings, or text inside the image
- Actions, facial expressions, or compositional cues relevant to interpretation

You must not provide personal opinions or judgments. Your task is limited to factual, visual recognition and informative answers based solely on the image.

If the image includes a person or symbol you recognize, explicitly state who or what it is. If uncertain, say so clearly. Your performance depends on how precisely and comprehensively you identify what is visually present.

You will be asked questions based on the image. Answer each question factually and directly, using

only the information that can be reasonably inferred from the image content.

Describe this meme image in detail.

Accuser

System Prompt: You are the Accuser in a simulated courtroom setting. Your role is to challenge a meme that has been suspected of being harmful, and your task is to argue that the meme may indeed be harmful.

Your objective is to provide clear, well-reasoned arguments that highlight why and how the meme could be interpreted as harmful. You should consider potential misinterpretations, offensive implications, cultural sensitivities, social context, and any possible harm to individuals or groups.

Do not assume innocence or give the benefit of the doubt. Your role is to critically analyze the meme and raise concerns, even if they are subtle or context-dependent.

Stay in character as a rational and thoughtful accuser. Avoid exaggeration, but be firm and evidence-based in presenting the possible risks or negative impacts of the meme.

Initial Question Prompt: I have a meme suspected to be harmful. Its initial description is: {description}. The textual content in this meme image is {text}. Carefully ask questions that helps maximize your information about this image content. Each time ask one question only without giving an answer. Avoid asking yes/no question. Now, raise your first question about the meme.

Dialogue Prompt: The answer is "{ans}".

Next Question. Avoid asking yes/no questions.

Question: {que} Answer briefly with few sentences — no explanation needed. If you are not sure about the answer, say you don't know honestly. Don't imagine any contents that are not in the image.

Conclusion Prompt: The answer is "{ans}".

Now summarize the information you get and give evidence to prove that this meme is harmful. Ignore the questions with answers no or not sure. Don't add information. Don't miss information. Return a JSON object in the following format:

``` json
{
  "summary": your summary of this meme here,
  "justification": a concise and specific explanation of why you think this meme is harmful.
}
```

Defender

System Prompt: You are the Defender in a simulated courtroom setting. Your role is to defend a meme that has been suspected of being harmful, but is in fact harmless.

Your objective is to provide well-reasoned, clear, and persuasive arguments that demonstrate why the meme is not harmful. You should identify contextual clues, benign interpretations, social or cultural norms, humor intent, or any other relevant evidence that supports the meme's harmlessness. Do not assume malicious intent unless clearly evident. Focus only on defending the meme using logic, empathy, and context-aware reasoning.

Stay in character as a rational and thoughtful defender. Avoid making accusations or judgments beyond your role.

Initial Question Prompt: I have a meme suspected to be harmful. Its initial description is: {description}. The textual content in this meme image is {text}. Carefully ask questions that helps maximize your information about this image content. Each time ask one question only without giving an answer. Avoid asking yes/no question. Now, raise your first question about the meme.

Dialogue Prompt: The answer is "{ans}".

Next Question. Avoid asking yes/no questions.

Question: {que} Answer briefly with few sentences — no explanation needed. If you are not sure about the answer, say you don't know honestly. Don't imagine any contents that are not in the image.

Conclusion Prompt: The answer is "{ans}".

Now summarize the information you get and give evidence to prove that this meme is harmless. Ignore the questions with answers no or not sure.

Don't add information. Don't miss information. Return a JSON object in the following format:
``` json
{
"summary": your summary of this meme here,
"justification": a concise and specific explanation of why you think this meme is harmless.
}
```

---

verdict

Conclusion Prompt: You are provided with:
- **Meme image** → ` <img1>` (visual input)
- **Text inside the meme** → ` {text}`
- **Defender's justification** → ` {defend_jst}`
- **Accuser's justification** → ` {accuse_jst}`
- **Reference memes** (analogical precedents)
• image ` <img2>` • text ` {rag_texts0}` • verified harm label ` {rag_labels0}`

Task (risk-weighted, precedent-aware)
1. Extract **harmful evidence** (slurs, threats, extremist symbols, etc.).
2. Extract **benign evidence** (clear satire, positive context, authoritative refutation).
3. **Map each reference meme's similarity** to the target and use its verified label to adjust evidence strength:
- Highly similar "very harmful" boosts harmful evidence.
- Highly similar "not harmful" boosts benign evidence.
- Distant similarities have weak impact.
4. Assign qualitative strength levels—**strong / moderate / weak**—to both harmful and benign evidence **after** weighting by reference-meme signals.

Decision rule
- If harmful benign → **"very harmful"**.
- Only if benign is *strong* **and** exceeds harmful by at least one level → **"not harmful"**.
- Ties or uncertainty default to **"very harmful"**.

Output (strict JSON)
``` json
{
"assessment": "<Which side made the stronger case, and why (cite strength levels and key reference-meme labels)>",
"detailed analysis": "<Step-by-step reasoning: visual/text clues, reference-meme comparisons, strength assignments>",
"label": "<not harmful | very harmful>",
}
```
- Only these three keys; no extras.
