Evaluating Rationale-Grounded Responses for Meme Images via Self-Scoring-Based Selection

Ryotaro Kanae, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology (NAIST), Japan
kanae.ryotaro.kr8@naist.ac.jp
{sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

Hate detection in memes involves an inherent trade-off between over-detection, which may suppress legitimate expression, and under-detection, which may allow harmful content to spread. Therefore, beyond improving detection accuracy, it is important to make the decision-making process transparent and assess the quality of its underlying reasoning. In this study, we extend REPS, a reasoning-enhancement framework, to the multimodal setting in order to improve both the explainability and detection sensitivity of vision–language models for context-dependent hateful meme detection. Our approach enables the model to iteratively select and learn from self-generated rationales. Experimental results show that explanation plausibility improves with an approximately 60% win rate while achieving a recall of 0.960. These findings demonstrate that reasoning-enhanced multimodal learning improves both explanation quality and detection sensitivity, while highlighting the trade-off introduced by over-reasoning.

1 Introduction

Memes on social media combine short text with visual cues, making their meaning highly context-dependent and enabling them to spread rapidly (Shifman, 2013). This property also allows memes to implicitly convey discriminatory content related to attributes such as race, ethnicity, gender, and religion, motivating the development of automated hate detection systems (Schmidt and Wiegand, 2017; Martinez Pandiani et al., 2025). Because hateful intent often emerges from interactions between textual and visual information (Kumar and Nandakumar, 2022), unimodal approaches are inherently insufficient (Gomez et al., 2020). However, even vision-language models (VLMs) may generate hallucinated or post-hoc rationales (Huang et al., 2025; Liu et al., 2025; Jain and Wallace, 2019), limiting the transparency and trustworthiness of their decision-making process.

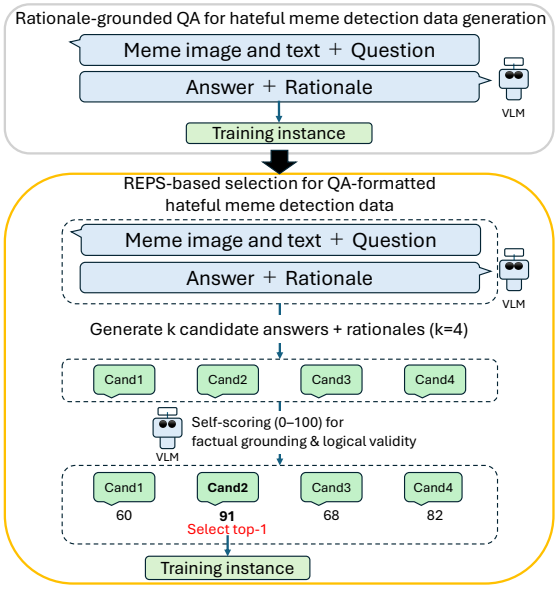


Figure 1: Overview of the proposed framework. QA-formatted supervision is constructed for hateful meme detection, and REPS is extended to the multimodal setting to select plausible QA rationales for training.

To make the decision process more transparent, we formulate hateful meme detection as a QA-formatted task, where a VLM predicts whether a meme is hateful while generating question–answer (QA) rationales that explain its decision. Figure 1 illustrates the overview of our approach. First, we transform the Hateful Memes Dataset (Kiela et al., 2020) into QA-formatted supervision for jointly learning hate detection and rationale generation. We then extend Rationale Enhancement through Pairwise Selection (REPS) (Kawabata and Sugawara, 2024) to the multimodal setting. For each meme, multiple QA rationale candidates are generated, and REPS selects the most plausible candidate as supervision for training the VLM. To isolate the effect of this rationale selection process, we compare our approach with a baseline trained on

Proceedings of the 19th International Natural Language Generation Conference: Short Papers, pages 124–130
October 17-21, 2026 ©2026 Association for Computational Linguistics

the same QA-formatted data without REPS-based selection. Finally, we evaluate whether rationale selection improves both hateful meme detection performance and the plausibility of the generated QA rationales.

Our experiments show that REPS-based selection substantially improves the plausibility of decision rationales compared with the baseline. We also observe high recall, indicating increased sensitivity to hateful content at the cost of more false positives, suggesting a trade-off between explanation quality and discriminative accuracy. These findings demonstrate that reasoning-enhanced multimodal learning can improve both explanation quality and detection sensitivity, while highlighting the trade-offs introduced by over-reasoning.

2 Background and Related Work

Hateful Memes Dataset and Multimodal Reasoning. The Hateful Memes Dataset (Kiela et al., 2020) is a benchmark designed to evaluate multimodal reasoning for hateful meme detection. A key characteristic of this dataset is the inclusion of benign confounders, where neither the image nor the text alone is hateful, but their combination conveys hateful intent. Consequently, solving this task requires VLMs to capture complex interactions between visual and textual information while incorporating relevant background knowledge and cultural context. These characteristics make the dataset a suitable benchmark for studying not only detection performance but also the quality of the reasoning underlying model predictions.

VLM Reasoning Processes and REPS. Improving the transparency of VLM reasoning remains an important research challenge. Chain-of-Thought (CoT) prompting (Wei et al., 2022; Shao et al., 2024) encourages models to generate intermediate reasoning steps together with their predictions and has been widely adopted to improve reasoning performance and interpretability. As an explicit reasoning approach, we adopt a QA-formatted rationale-generation framework for hateful meme detection, in which the model jointly predicts the label and its supporting rationale (Kawabata and Sugawara, 2024). To improve the quality of generated rationales, we build on Rationale Enhancement through Pairwise Selection (REPS) (Kawabata and Sugawara, 2024), which performs pairwise comparisons among multiple model-generated answer–rationale pairs and uses the model itself

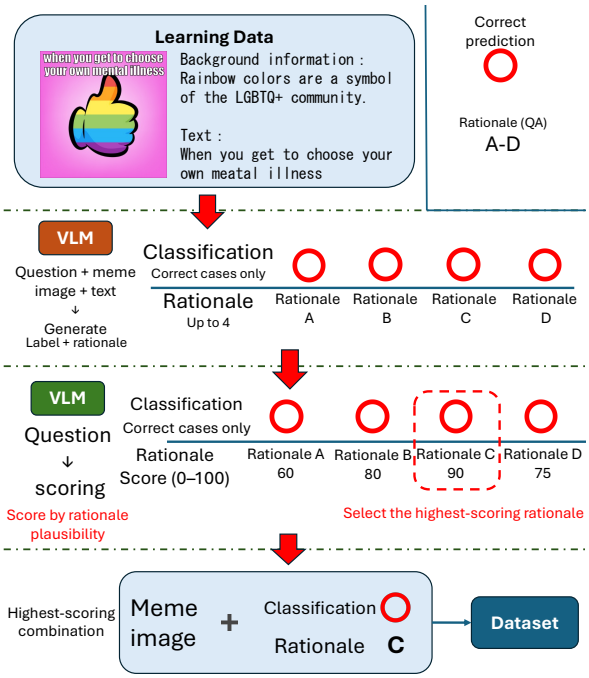


Figure 2: Examples of hateful meme images and the procedure for constructing the training dataset.

to select the most plausible rationale for training. By iteratively selecting higher-quality rationales, REPS improves the plausibility of explanations without requiring additional human annotations. We extend REPS to the multimodal setting and evaluate its impact on both hate detection performance and rationale quality.

3 Methods and Experimental Setup

We extend REPS to the visual domain and apply it to hateful meme detection. First, the QA generation model (M_{gen}) converts the Hateful Memes Dataset into a QA format compatible with the REPS framework. Next, the QA selection model (M_{sel}) scores the generated rationales, as illustrated in Figure 2, and selects the most plausible QA candidate for each training instance. The selected QA pairs are then used to fine-tune the student model (M_{stu}) to jointly predict the classification label and generate a rationale for its prediction. For evaluation, we measure not only classification accuracy but also Explanation Plausibility (Win Rate) using an evaluation model (M_{eval}).

3.1 Dataset Construction

Seed Dataset. First, we use the Hateful Memes Dataset (Kiela et al., 2020) as the seed dataset for our experiments. The training split (8,500 instances) is used for model training, while the valida-

Method	Number of training instances
Original data	8,500
Augmentation	29,886
Random	8,245
REPS	8,245

Table 1: Number of training examples per setting. For Random and REPS, one QA pair is used per image among the 8,245 images where generation succeeded.

tion split (500 instances) is reserved exclusively for final evaluation. Because the ground-truth labels of the official test split are not publicly available, we report results on the validation split.

Data Construction and Training Settings. To evaluate REPS, we first use M_{gen} to generate QA candidates for each training image following Figure 2. We sample up to four QA pairs per image by running generation up to 10 times, retaining only images for which at least one valid QA pair is generated. Table 1 shows the resulting dataset statistics. Based on the candidate pool, we construct training data using the following three settings:

1. **Baseline (Random):** One QA candidate is randomly selected for each image as supervision, corresponding to the distillation of teacher-generated outputs without selection.
2. **Baseline (Augmentation):** All generated QA candidates are used as training data, increasing the amount and diversity of supervision.
3. **REPS:** QA candidates are scored using a prompt-based evaluator according to answer appropriateness, internal consistency, and plausibility. The highest-scoring candidate is then selected to supervise each image.

Prompt templates for QA generation and evaluation are provided in Appendix A.

3.2 Model Configuration

As illustrated in Figure 3, we use four models with distinct roles for data generation, selection, training, and evaluation.

- **QA generation model (M_{gen}):** M_{gen} generates QA pairs consisting of a predicted label and its supporting rationale. We use `llama3-llava-next-8b-hf`[1] (Li et al., 2024).
- **QA selection model (M_{sel}):** M_{sel} selects the most plausible QA candidate for each training instance using REPS-style pairwise evaluation. We use `Qwen2.5-VL-7B-Instruct`[2] (Bai et al., 2025).
- **Student model (M_{stu}):** M_{stu} is fine-tuned on the constructed dataset for hateful meme detection. We primarily use `LLaVA-1.5-7B`[3] (Liu et al., 2024) as the backbone model. Since M_{stu} is the model that performs the actual inference, we additionally use `LLaVA-v1.6-Vicuna-13B`[4] (Liu et al., 2024) to examine whether our approach remains effective across different parameter scales.
- **Evaluation model (M_{eval}):** M_{eval} serves as a judge model for assessing the quality of generated explanations. We use `InternVL2-8B`[5] (OpenGVLab Team, 2024). Using a different VLM as the judge helps reduce preference bias toward the models used for data generation and selection.

The synthetic QA pairs generated by M_{gen} and selected by M_{sel} are used to construct the training data for M_{stu} and fine-tune it using LoRA (Hu et al., 2022) with batch size 32, 3 epochs with a learning rate of 2×10^{-4}. To ensure a fair comparison, all methods, i.e., REPS and Baseline (Augmentation and Random), are trained with the same hyperparameters, and we use half precision. Unless otherwise specified, all implementations are based on the Hugging Face Transformers library (Wolf et al., 2020) and use the default settings of the `Trainer` API. Training was conducted on a single NVIDIA A6000 GPU.

3.3 Evaluation Metrics

We evaluate model performance from two perspectives: quantitative classification accuracy and qualitative explanation capability.

Classification Metrics. We extract the predicted label (Hateful or Not Hateful) from the model-generated answer and report Accuracy, Precision, Recall, and F1 score.

Explanation Plausibility (Win Rate). In hate speech detection, the rationale for a decision is as important as the predicted label. To ensure reproducible evaluation, we adopt an automated evaluation framework that uses a VLM as a judge (LLM-

[1] https://hf.co/llava-hf/llama3-llava-next-8b

[2] https://hf.co/Qwen/Qwen2.5-VL-7B-Instruct

[3] https://hf.co/llava-hf/llava-1.5-7b-hf

[4] https://hf.co/llava-hf/llava-v1.6-vicuna-13b-hf

[5] https://hf.co/OpenGVLab/InternVL2-8B

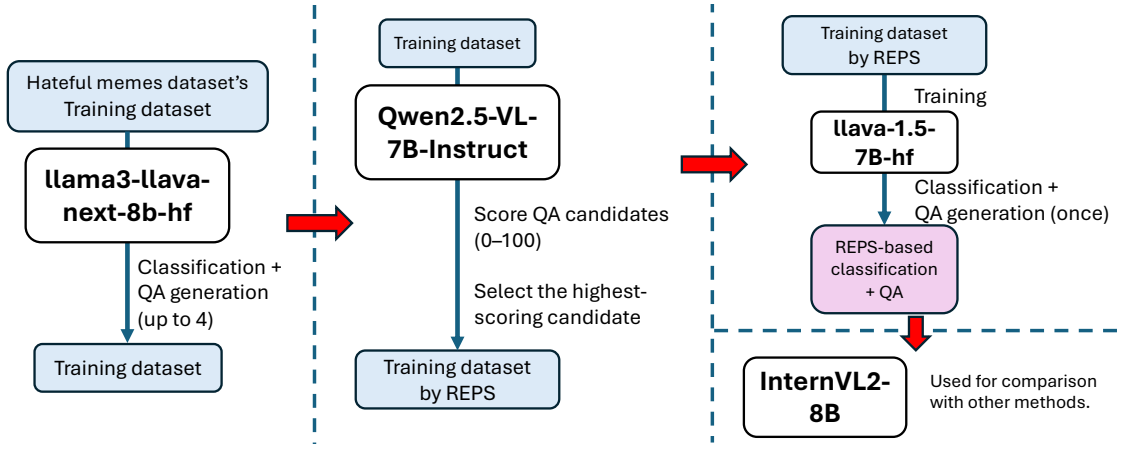


Figure 3: Models and Procedures for Each Stage

Method	Accuracy	Precision	Recall	F1 Score
Baseline (Rand.)	0.526	**0.514**	0.928	0.662
Baseline (Aug.)	0.518	0.511	0.872	0.644
REPS (Proposed)	**0.524**	0.513	**0.960**	**0.669**

Table 2: Classification performance of each method using the primary backbone model, `LLaVA-1.5-7B`. REPS achieves the highest Recall and F1 scores. "Rand." denotes Random, and "Aug." denotes Augmentation.

as-a-Judge) (Zheng et al., 2023) instead of human annotation. Specifically, on the validation set (500 instances), we perform pairwise comparisons of the QA outputs produced by the three models, i.e., Baseline (Random), Baseline (Augmentation), and REPS, using M_{eval}. We then compute the Win Rate, defined as the proportion of comparisons in which the QA generated by the model trained with REPS is preferred. By directly comparing the quality of two rationales, this evaluation measures how rationale quality changes before and after REPS-based training. To minimize preference bias, we use an independent VLM as M_{eval} that differs from the VLMs used in the other stages of our framework.

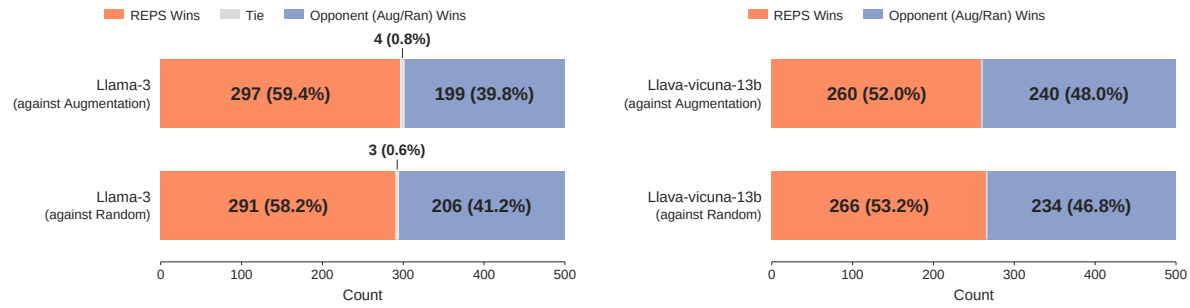
4 Results

Quantitative Evaluation: Classification Performance. Table 2 reports the classification performance of each method. First, we observe no substantial differences among methods in Accuracy and Precision. This is consistent with the findings of (Kawabata and Sugawara, 2024), who note that REPS is not primarily designed to yield large improvements in classification performance. In contrast, REPS improves Recall to 0.960, though this comes at the expense of increased false positives. Examining the confusion matrix on the validation set, among the 250 hateful instances, REPS misclassified 10 as not hateful, whereas Baseline (Augmentation) and Baseline (Random) misclassified 32 and 20 instances, respectively. These results suggest that REPS-based training substantially improves the model's sensitivity to hateful content.

Qualitative Evaluation: Explanation Plausibility. Next, we evaluate the quality of the generated decision rationales through pairwise comparisons using the evaluation model (M_{eval}). Figure 4 shows the results. REPS is preferred over Baseline (Random) and Baseline (Augmentation) in 58.2% and 59.4% of the comparisons, respectively. This improvement is consistent with the findings of Kawabata and Sugawara (2024), suggesting that the REPS selection process is also effective for improving explanation quality in the multimodal setting. Nevertheless, the results with the other backbone, `LLaVA-v1.6-Vicuna-13B`, show only marginal gains. Although the margins differ between the two backbone models, REPS consistently achieves win rates above chance, suggesting that its benefit is preserved across them.

5 Discussion

Divergence Between Explanation Quality and Classification Accuracy. REPS improves rationale plausibility but does not improve classification accuracy, consistent with prior findings that better-looking reasoning does not necessarily yield better predictions (Kawabata and Sugawara, 2024). Given the complex interplay between visual and textual information in VLMs, plausible explanations are not guaranteed to faithfully reflect the model's decision process or correspond to the ground-truth label (Jacovi and Goldberg, 2020; Lyu et al., 2024).



(a) `LLaVA-1.5-7B` (b) `LLaVA-v1.6-Vicuna-13B`

Figure 4: Pairwise evaluation results (Win Rate) by M_{eval}. The orange bars indicate the proportion of comparisons won by REPS, while the blue bars indicate the proportion won by the baseline.

Method	Setting	n	Word count		Hedge count	
			mean ± std	median [Q1,Q3]	mean ± std	median [Q1,Q3]
LLaVA-1.5-7B (main backbone)						
Baseline	augmentation	500	147.59 ± 24.99	$145\,[130, 160]$	2.848 ± 1.983	$3\,[1, 4]$
Baseline	random	500	138.15 ± 23.41	$138\,[121, 153.25]$	2.532 ± 1.742	$2\,[1, 4]$
REPS (Proposed)	reps	500	138.68 ± 23.33	$137.5\,[122, 154]$	2.066 ± 1.605	$2\,[1, 3]$
LLaVA-v1.6-Vicuna-13B (sub backbone)						
Baseline	augmentation	500	173.63 ± 30.43	$173\,[152, 200]$	1.898 ± 1.743	$1\,[1, 3]$
Baseline	random	500	135.82 ± 24.75	$135\,[118, 150]$	1.186 ± 1.356	$1\,[0, 2]$
REPS (Proposed)	reps	500	72.04 ± 14.38	$71\,[63, 81]$	0.392 ± 0.596	$0\,[0, 1]$

Table 3: Style statistics of generated rationales. We report mean ± std and median with quartiles $[Q1, Q3]$.



Figure 5: An example failure case.

Analysis of Higher Recall and Increased Over-Detection. To better understand the higher recall and increased false positives, we analyze representative errors and patterns of rationale. Figure 5 presents a representative failure case. Moreover, we analyze two surface properties of the generated rationales: (i) length (word count) and (ii) hedging usage (hedge count). Table 3 reports the mean±std and median [Q1, Q3] for each setting. Rationale lengths are broadly comparable across methods: REPS is similar to Baseline (Random) and slightly shorter than Baseline (Augmentation). In contrast, REPS uses fewer hedge words (mean 2.066 vs. 2.532–2.848; median 2 with a lower upper quartile), suggesting REPS encourages more assertive explanations without increasing rationale length.

6 Conclusion

In this study, we investigated the selection of REPS-based training data to improve VLM explainability for hateful meme detection. Experimental results show that REPS improves rationale plausibility (approximately 60% win rate over the baselines) while achieving high recall (0.960), indicating greater sensitivity to subtle image–text hate cues. However, this improvement comes at the cost of increased false positives, revealing a trade-off between explanation quality and classification performance. Future work includes incorporating supervision on benign cases and refining the selection process to reduce over-detection.

Limitations

First, we evaluated our approach only on the Hateful Memes Dataset and reported results on the validation split, because the official test labels are not publicly available. Therefore, the generalizability

of our findings to other meme datasets, platforms, languages, or deployment settings remains unclear.

Second, our pipeline depends on specific model choices for generation, selection, and evaluation (e.g., the selection model and the VLM-as-a-Judge). In particular, explanation plausibility is assessed through pairwise comparisons by a judge model, whose preferences may not fully align with human judgments. Moreover, plausibility does not guarantee faithfulness to the model's true decision process, so well-formed explanations may still be post hoc or incomplete. Therefore, further human evaluation would be ideal for practical validation.

Third, REPS increases detection sensitivity but also increases false positives, reflecting a trade-off that may lead to over-moderation in practice. We do not investigate threshold calibration or decision policies that explicitly control this trade-off.

Fourth, our training data is constructed from automatically generated QA candidates. A small portion of images (approximately 3%) fail during QA generation and are therefore excluded, potentially introducing bias into the constructed dataset.

Ethical Considerations

Dataset. We use the Hateful Memes Dataset in accordance with its license and terms of use and do not redistribute the dataset (including images or associated text) beyond what those terms permit. Because the benchmark contains potentially offensive content, we present qualitative examples and model outputs only when necessary and take care to minimize unnecessary exposure or amplification. Given the risk of false positives and potential misuse, the proposed system is not intended for fully automated enforcement and should be used only with human oversight.

AI-assistant. AI-assisted writing tools, including DeepL, Grammarly, and GPT-5, were used for language refinement, including wording improvements, translation, and grammatical correction. All research ideas, experimental design, implementation, analyses, and conclusions were developed and verified by the authors.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report. *Preprint*, arXiv:2502.13923.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1470–1478.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Akira Kawabata and Saku Sugawara. 2024. Rationale-aware answer verification by pairwise self-evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16178–16196, Miami, Florida, USA. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild. Blog post.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.

Zujing Liu, Junwen Pan, Qi She, Yuan Gao, and Guisong Xia. 2025. On the faithfulness of visual thinking: Measurement and enhancement. *Preprint*, arXiv:2510.23482.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Delfina S. Martinez Pandiani, Erik Tjong Kim Sang, and Davide Ceolin. 2025. 'toxic' memes: A survey of computational perspectives on the detection and explanation of meme toxicities. *Online Social Networks and Media*, 47:100317.

OpenGVLab Team. 2024. InternVL2: Better than the Best—Expanding Performance Boundaries of Open-Source Multimodal Models with the Progressive Scaling Strategy. Project blog post.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 8612–8642. Curran Associates, Inc.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Prompt Templates

We used three prompt templates: (i) a structured QA prompt that outputs JSON for the label and minimal evidence, (ii) an explanation prompt for generating a rationale, and (iii) a scoring prompt that assigns a score from 0 to 100 based on factual grounding and logical consistency. OCR-extracted meme text was included as context when available.

Structured QA Prompt (JSON)

You are a careful multimodal QA assistant. Output JSON only.
1) Q1: Does the meme express hate toward a protected group? - Answer with one of: "Yes", "No", or "Uncertain".
2) Q2: If yes or uncertain, which group appears targeted? - Short phrase, or "None".
3) Evidence: Quote minimal text or describe minimal visual elements (<=2 sentences).
Return JSON with keys: "q1": "...", "q2_group": "...", "evidence": "..." Do not add any text outside JSON.
[OCR/text in meme]: "<OCR_TEXT>"

Explanation Prompt

Provide a concise, fact-grounded rationale in 2-4 sentences about whether the meme is hateful and why.
[OCR/text in meme]: "<OCR_TEXT>"

Scoring(0–100) Prompt

You are a helpful assistant who evaluates explanations' factual accuracy and logical consistency.
Look at the image and the provided information. Evaluate the following explanation and assign a score from 0 to 100 based on how factually grounded it is in the image and how logically valid it is.
Scoring rubric: - 90-100: Fully grounded in the image/information, logically coherent, no unsupported claims. - 70-89: Mostly grounded; minor speculation or small logical gaps. - 40-69: Partially grounded; several unsupported claims or notable inconsistencies. - 10-39: Largely ungrounded; major hallucinations or faulty reasoning. - 0-9: Completely ungrounded or contradicts the image/information.
Output format: Score: <integer 0-100> Rationale: <1-2 sentences explaining the main evidence and the main issue, if any>