MerFT: A Framework for Social Conflict Meme Exploration via Multimodal Retrieval-Augmented Fine-tuning

Anonymous Author(s)

Abstract

Social media is a key medium for expressing and spreading social conflicts. In particular, memes have emerged as major contents that visually and implicitly convey complex socio-cultural messages through satire and symbols. This study constructs SocialMQD, a multimodal dataset specialized for interpretation of conflict-related memes, and proposes MerFT (Meme Exploration via Multimodal Retrieval-Augusted Fine-tuning) which is based on Retrieval-Augusted Generation (RAG)-based framework that utilizes image, caption, and associated textual documents simultaneously. MerFT effectively reinforces satire, irony, and nuanced sociopolitical issues comprehension skills that existing Vision-Language models have struggled with through designs that enable robust reasoning in multimodal environments such as distructor-aware fine-tuning and citation-based chain-of-thought reasoning. This study has empirically investigates the differences in performance based on the multi-modal input configuration (i.e., Base, Caption, Both), the category-specific performance differences, and the distractor frequency based on the category. The MerFT model demonstrated superior performance compared to existing RAG-based models, exhibiting both enhanced functionality and reasoning capabilities. Specifically, the MerFT model revealed a high degree of efficacy in scenarios involving complex information integration, even in environments with limited data availability. This study has demonstrated the technical feasibility of multimodal reasoning for meme-based conflict analysis and we expect to contribute to the development of automated content analysis systems grounded in a socio-cultural context.

CCS Concepts

• **Computing methodologies → Information extraction**.

Keywords

Social Conflict, Multimodal Meme Understanding, Retrieval-Augmented Generation (RAG), Chain-of-Thought (CoT), Large Vision-Language Model (LVLM), Fine-tuning

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Anonymous Author(s). 2018. MerFT: A Framework for Social Conflict Meme Exploration via Multimodal Retrieval-Augmented Fine-tuning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The rise of social media has established memes as a dominant medium for communication, transcending their origins as entertainment to become tools for expressing social and political discourse. Memes gain traction due to their humor, irony, and ability to convey complex emotions concisely, often influencing public sentiment and shaping social narratives. Social conflict, arising from ideological, economic, and cultural disparities, is now frequently mediated through such content. Memes like the "Karen" meme satirically expose racial and social tensions, providing commentary on privilege and injustice. Tabatabaei et al. highlight how memes can spread emotions like fear, anger, and joy rapidly, a phenomenon known as emotional contagion[11]. While such diffusion can unify, it also risks amplifying unrest. This illustrates how memes extend beyond entertainment into domains of social cognition and collective emotion. As Zhou et al. emphasize, combining text and image enables effective detection of offensive content, underscoring the potential of multimodal learning in real-time meme classification[15]. The term *multimodal* refers to content that integrates various media, such as images and text, enabling complex and layered interpretations through their interaction. Images elicit intuitive visual responses, while text provides abstract and linguistic meaning; together, they establish a richer semantic context. Analyzing social conflict through such content requires interdisciplinary knowledge across social, cultural, and historical domains. In settings where visual–text interplay remains underdeveloped, traditional feature extraction is insufficient, thus necessitating advanced automated methods[15].

Despite the rapid evolution of multimodal understanding models, research on meme analysis has predominantly focused on a limited set of domains, including the geographical distribution of memes, political and ideological framing, and mental health implications such as depression and anxiety. While recent work has expanded to detecting hateful[3] and offensive[10] memes, these studies have largely remained at the level of binary classification, lacking deeper contextual and causal analysis. In particular, comprehensive examination of the *causes*, *targets*, and *modes of expression* of social conflict-related memes is still at an early stage. Because such memes are deeply tied to evolving sociocultural trends, effective interpretation requires systems that can integrate up-to-date, context-rich external knowledge.

Retrieval-Augmented Generation (RAG)[4] offers a promising pathway by enabling models to retrieve and condition on relevant documents in real time. However, existing RAG-based fine-tuning

research has been largely limited to unimodal or text-centric settings. **To our knowledge, no prior work has developed a multimodal RAG fine-tuning framework specifically designed for socially and culturally complex meme interpretation.** Furthermore, prior distractor-aware training strategies have lacked a principled criterion for selecting distractor documents, often relying on random or loosely related noise that does not capture the subtle semantic confusions present in real-world retrieval scenarios.

To address these gaps, we introduce a **multimodal semi-supervised clustering with pseudo-labeling** strategy for distractor construction and model training. Instead of random selection, we perform joint image–text embedding and cluster both labeled and unlabeled documents in a shared semantic space. Within each cluster, we identify documents that are highly similar to the oracle (correct) document in multimodal embedding space but do not support the correct answer. These *cluster-based distractors* introduce more realistic and challenging noise, forcing the model to learn fine-grained discrimination between culturally and visually similar yet semantically divergent evidence. In addition, unlabeled documents are assigned pseudo-labels based on cluster consensus, allowing the framework to incorporate domain-specific or newly collected data without the cost of exhaustive annotation.

This clustering-driven design makes the framework inherently adaptable: organizations or researchers can apply it directly to their own datasets, continually refining the distractor pool and pseudo-labeled samples to match evolving real-world content. As a result, the framework not only improves robustness against meaning-level confusions but also supports **scalable dataset expansion**, enabling adaptation to new domains, emerging topics, and shifting sociocultural contexts. By establishing a clear criterion for distractor selection and enabling seamless integration of unlabeled multimodal data, our method bridges the gap between research prototypes and deployable systems for robust, domain-specific multimodal reasoning.

This study makes the following contributions:

(1) **Construction of the SocialMQD Dataset:** A multimodal dataset centered on social conflict, comprising meme images and associated explanatory documents across diverse sociocultural domains, enriched with Chain-of-Thought (CoT)[12]-based QA pairs for interpretable reasoning.
(2) **Proposal of the MerFT Framework with Cluster-Based Distractor Generation:** An extension of the original MerFT architecture that integrates multimodal semi-supervised clustering to create semantically challenging distractors and incorporate pseudo-labeled samples, thereby enhancing robustness to meaning-level confusions and enabling domain-specific adaptability.
(3) **Comprehensive Multimodal Explanation Task:** Beyond classification, our framework requires detailed sociocultural reasoning, enabling deeper understanding of satire, symbolism, and ideological framing in social conflict memes.

Through these innovations, our work advances beyond prior random distractor sampling approaches, providing the first principled, multimodal RAG fine-tuning framework that is both academically rigorous and practically deployable. The resulting system not only improves interpretive robustness but also offers a scalable, customizable pipeline for real-world multimodal retrieval environments, capable of continuous dataset expansion and adaptation to evolving societal discourse.

2 Related Work

2.1 Analysis of Social Conflict Memes

Early work on internet memes primarily examined platform-specific phenomena and the diffusion of politically provocative content. Studies of fringe communities (e.g., 4chan's /pol/) showed how racist or misogynistic narratives propagate into mainstream platforms and public discourse [1]. Subsequent research shifted toward qualitative and rhetorical perspectives, treating memes as quasi-arguments that encode social stance and ideology [6]. While this line of work illuminates discursive strategies and cultural signaling, it rarely provides computational tools for robust, large-scale interpretation of conflict-oriented memes.

2.2 Multimodal Models for Meme Understanding

The release of benchmark datasets and shared tasks catalyzed research into multimodal meme understanding. The Hateful Memes Challenge [3] introduced carefully controlled image–text contradictions to probe multimodal fusion. MultiOFF [10] broadened the scope to offensive content, and MOMENTA [9] proposed a hierarchical multimodal framework to detect harmful memes and their targets. Despite progress, generalization remains limited: models often exhibit *textualism*—overreliance on text while failing to parse visual symbolism or culture-bound references. Analyses of real-world hateful memes further reveal weak cross-platform transfer and difficulty with implicit cues [2]. Prompted or instruction-tuned VLMs mitigate some gaps [**?**], yet nuanced satire, metaphor, and contextual reasoning continue to challenge state-of-the-art systems.

2.3 RAG and Robustness to Distractors

Retrieval-Augmented Generation (RAG) [4] augments parametric knowledge with non-parametric evidence, improving recency and factuality in knowledge-intensive tasks. However, retrieval pipelines introduce failure modes: off-topic or partially relevant documents can distract the generator, leading to spurious reasoning. Recent advances include **certifiably robust** defenses such as RobustRAG, which isolates passage-level generations and then securely aggregates them to provide provable guarantees against retrieval corruption attacks [13]. Complementarily, **adversarial multi-agent tuning** (ATM) improves generator robustness by iteratively training a Generator against an Attacker that injects noisy or fabricated evidence [16]. These works demonstrate that robustness can be improved via aggregation guarantees or adversarial training; yet they remain largely *text-centric* and do not directly address multimodal settings where images, captions, and documents co-occur.

2.4 Semi-supervised Clustering and Pseudo-labeling for Multimodal RAG

Semi-supervised learning leverages unlabeled data by propagating structure from labeled examples. In multimodal scenarios, joint

image–text embedding spaces enable clustering that reflects shared cultural referents, visual motifs, and lexical frames. Pseudo-labeling extends supervision by assigning cluster-consensus labels to unlabeled instances, expanding training signals without exhaustive annotation. Applied to RAG, this perspective yields two benefits: (i) a principled supply of *semantically challenging* negatives (items near the oracle in embedding space yet evidentially unhelpful), and (ii) continual ingestion of newly collected domain data via pseudo-labels—supporting dataset expansion as topics evolve. Our work operationalizes this idea for meme reasoning by jointly clustering labeled and unlabeled documents in a shared multimodal space and by selecting cluster-consistent, answer-irrelevant items as training distractors.

2.5 Hard-Negative Mining and Distractor Selection Criteria

Hard-negative mining sharpens decision boundaries by sampling negatives that are closest to the anchor. Beyond random or loosely filtered distractors commonly used in RAG training, recent evidence shows that *seemingly plausible* reasoning paths can significantly degrade LLM performance in multi-hop settings (up to 45% relative F1 drop) [**?**]. For socially complex, multimodal memes, distractors should thus be *near-miss* evidence—topically aligned, culturally adjacent, and visually/textually similar—yet ultimately inconsistent with the correct answer. We formalize this selection with a cluster-local criterion in a multimodal embedding space, producing hard but answer-mismatched documents and aligning training conditions with deployment realities.

Positioning. Compared to prior multimodal meme studies [2, 3, 9**?** , 10] and RAG robustness methods [4, 13, 16], our framework uniquely (i) instantiates a *multimodal* RAG fine-tuning pipeline tailored to socially/culturally complex memes, and (ii) introduces a principled, clustering-driven criterion for distractor selection coupled with semi-supervised pseudo-labeling. This yields realistic, meaning-level confusions during training and enables scalable dataset expansion as social topics and meme genres evolve.

3 SocialMQD Dataset

This section describes the entire procedure for constructing the SocialMQD dataset. The dataset consists of meme images and associated documents centered on sociocultural conflict, prejudice, and symbolism. It also includes Chain-of-Thought (CoT) based question-answer (QA) pairs to evaluate multimodal reasoning capabilities. The overall construction process comprises three stages: image collection, document linking, and QA generation.

Dataset Split (Train / Test). Instead of the traditional three-way split into Train/Validation/Test, we divide the data into two subsets: *Train*, used for model training, and *Test*, used for performance evaluation.

3.1 Image and Document Collection

The first step in building the SocialMQD dataset is the collection of meme images and their corresponding explanatory documents.

Social Conflict Category	Train	Test
Social Conflict Domains		
Gender/Feminism	160	40
Political Conflict	160	40
LGBTQ+ Issues	160	40
Economic/Inflation	160	40
Environmental Issues	160	40
Immigration	160	40
Wealth Inequality	160	40
Generational Conflict	160	40
Total	**1,280**	**320**
Data Sources		
KnowYourMeme (Documents)	1,280	320
Question Categories		
Cultural Context Understanding	213	53
Metaphor & Symbol Interpretation	213	53
Detecting Satire & Irony	213	54
Social Conflict Analysis	214	54
Image-Text Integration	213	53
Critical Thinking	214	53
Total Questions	**1,280**	**320**

Table 1: SocialMQD Dataset Statistics for Social Conflict Meme Analysis. Dataset is split into Train and Test sets, with 8 social conflict domains and 6 cognitive question categories. Each image has one question assigned. All documents sourced from KnowYourMeme.

This phase is designed to capture the sociocultural meanings embedded in memes and lays the foundation for the subsequent task of generating well-aligned QA pairs.

Image Collection Platforms. and Criteria. Meme images were sourced from Reddit, Pinterest, and KnowYourMeme. Reddit offers real-world meme usage within discussions, capturing both intent and social response. Pinterest provides visually diverse and thematically consistent memes. KnowYourMeme, with its encyclopedic documentation of meme origins and cultural context, served as the primary source for high-quality reference documents in QA generation.

Keyword Selection and Meme Filtering Based on Social Conflict. A key priority in data collection was ensuring that the themes of selected memes adequately reflect sociocultural conflict. To achieve this, a set of keyword categories was predefined. Keywords were selected based on their popularity, frequency of use in searches, and symbolic cultural relevance.

- **Politics**: Political Compass, Polandball, /pol/, Nancy Pelosi Clapping, Let's Go Brandon, Kim Jong Un Looking at Things, I Am Once Again Asking for Your Financial Support, Fake News, Dark Brandon, Covfefe
- **Gender/Feminism**: YesAllWomen, Wife Guys, Triggered Feminist, Sydney Sweeney's Anti-Woke Boobs, Privilege Denying Dude, Pink Tax, MeToo, Mansplaining, Feminist Nazi, Feminism

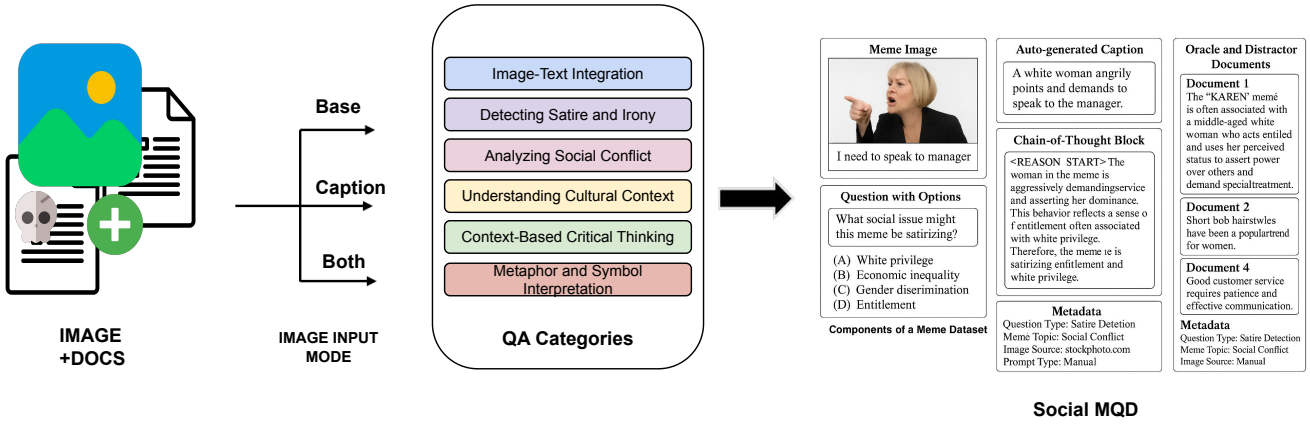

Figure 1: SocialMQD Process

- **LGBTQ+**: barillas-anti-gay, be-gay-do-crime, bi-panic, bi-wife-energy, bisexual-lighting, destiel-is-canon, egg-transgender, oreos-gay-pride-cookie, trans-rights, yassification
- **Inflation**: Vibecession, Money Printer Go Brrr, It's One Banana Michael What Could I, Is We Getting a Stimulus Check, I Present 100 of Groceries, Home Alone Grocery Prices, Egg Shortages High Egg Prices, Can't Have Shit in Detroit, 600 Stimmy, 2022 Russian Oil Ban and Gas Price Surge
- **Environment**: Earth Chan, Four Horsemen of the Apocalypse, Global Climate Strike, Green New Deal, How Dare You, Kylie Jenner's Three Minute Flights, Microplastics, Paris Climate Agreement Withdrawal, Save the Turtles, Taylor Swift's Private Jet Emissions Controversy
- **Immigration**: 14 Words 1488, White Genocide / The Great Replacement, Gay Immigrant Muslim Furry Romance, We Build the Wall, ICE Bae, Latinos for Trump Supporter Arrested by ICE, 2022 Martha's Vineyard Migrant Drop Off, Moving to Canada, They're Eating the Dogs, Orcposting
- **Wealth Inequality**: We Are the 99 Percent, Stop Being Poor, Rich Men North of Richmond by Oliver Anthony, Rentoid, Let Them Eat Cake, Late Capitalism, It's One Banana Michael What Could It Cost 10, Guillotine, Equality Equity Justice, Avocado Toast
- **Generational Conflict**: 30 Year Old Boomer, Boomer Remover, Ok Boomer, Zoomer Wojak, Zoomerification, Zoomer Humor, Millennials Are Killing, Meet the Typical Zoomer, Nobody Wants to Work Anymore, Hand Me the damm dryer

Document Linking and Preprocessing. The images in question are characterized by a heterogeneity in terms of their respective themes, backgrounds, and styles. Consequently, a precise interpretation of the images is contingent upon a comprehensive understanding of the extant contexts. In order to accomplish the aforementioned objective, KnowYourMeme was utilized to match the metadata descriptions. The following steps are taken in the process of matching:

(1) Perform keyword-based search on KnowYourMeme
(2) Collect top-ranked search results and filter them based on relevance scores
(3) Manually review content for semantic alignment and designate as the final oracle document

Caption Generation for Vision-Language Interaction. All collected images undergo a pre-captioning process to convert visual content into linguistic input. A vision-language model is used to generate captions summarizing the key elements of each image, which are then refined for compatibility with language model inputs. In particular, these captions are incorporated into the prompt during QA generation in the Caption and Both modes.

3.2 Chain-of-Thought Data Generation

The second core phase of SocialMQD is the construction of question-answer (QA) pairs and corresponding reasoning chains based on the collected meme images and their linked documents. The resulting data is formatted in the Chain-of-Thought (CoT) structure, providing not only correct answers but also the underlying reasoning process. This setup aims to enable models to interpret cultural context, satire, metaphor, and social conflicts beyond surface-level correctness.

Question Generation Strategy. QA generation is performed automatically using high-performance LLMs such as GPT-4o-mini[8]. For each meme, six questions are generated based on six predefined cognitive categories. Each category is associated with at least five prompt templates to ensure diversity and generalization in the question set.

The defined categories are as follows:

- **Understanding Cultural Context**: Questions that prompt analysis of the meme's social background, generational experience, or cultural phenomena.
- **Metaphor and Symbol Interpretation**: Questions that require inference of symbolic or metaphorical meanings embedded in image or text.

- **Detecting Satire and Irony**: Questions that encourage interpretation of satirical structures, irony, and humor within the meme, including intended targets and messages.
- **Analyzing Social Conflict**: Questions addressing the conflict structure, class inequality, or ideological clashes portrayed in the meme.
- **Image-Text Integration**: Questions analyzing how image and text interact to construct meaning.
- **Context-Based Critical Thinking**: Questions that challenge the reader to critically assess claims, assumptions, or biases embedded in the meme's message.

For example, prompts such as *"What social phenomenon is being satirized in this meme?"* or *"How do the image and text complement each other in conveying the message?"* are designed to elicit interpretive reasoning grounded in evidence. **See Appendix 10 for detailed question templates per category.**

Each category includes five predefined question formats, one of which is randomly selected during generation. This prevents overfitting to fixed expressions or syntactic structures, enhancing the generalizability of the generated questions.

Answer and Chain-of-Thought Generation. For each question, the model generates the following elements:

- Answer choices (A~D): Four options including distractors
- Correct answer label (A~D): The correct choice is explicitly marked
- Chain-of-Thought reasoning: The model is prompted to describe the reasoning process explicitly, which is enclosed between the tags <REASON_START> and <REASON_END>

This format enables the model to go beyond surface-level outputs and perform explainable and interpretable predictions. It is especially suitable for complex memes involving nuanced social implications, allowing the model to select the correct answer based on valid reasoning.

Distractor Document Composition and Difficulty Control. In addition to the correct (oracle) document, up to five distractor documents are included. These distractors are topically similar but contain information irrelevant to the correct answer. The number of distractors can be adjusted based on experimental design. For robustness testing, some samples increase distractor counts to simulate noisy environments and assess reasoning performance. All documents are randomly shuffled during input, requiring the model to evaluate all content and reason based on the oracle document.

Data Structure and Format Refinement. The final generated data follows the structure below:

- **Meme Image:** The original meme image representing a real-world social conflict meme, allowing the model to infer meaning from visual cues.
- **Auto-generated Caption:** A brief summary sentence created using a pretrained vision–language model, serving as supplementary linguistic input to complement visual information.
- **question_with_options:** A question targeting a specific cognitive dimension of the meme (e.g., satire detection, critical reasoning), accompanied by four answer choices (A–D). These questions are designed to induce higher-order reasoning rather than mere fact recall.
- **Chain-of-Thought Block (cot_answer):** The correct answer is marked using the format <ANSWER>: A~D, followed by a reasoning process enclosed within the <REASON_START> and <REASON_END> tags. This structure allows the model to learn step-by-step reasoning paths explicitly.
- **oracle_doc / distractor_docs:** One oracle document and 0 to 5 distractor documents. The distractors are semantically similar but irrelevant to the correct answer, simulating real-world confusion in RAG environments.
- **metadata:** Contains information such as question type, meme topic (e.g., politics, gender, economy), image source, prompt template, and number of distractors. This metadata is essential for category-specific performance analysis and error diagnosis.

This highly structured QA dataset is directly used for fine-tuning and serves as a foundation for enhancing both multimodal reasoning capabilities and distractor-aware learning architectures.

4 MerFT Framework Design

This section presents a fine-tuning framework designed to enhance multimodal reasoning using the previously constructed SocialMQD dataset. MerFT goes beyond simply adding a retrieval module to a pretrained vision–language model; it aims to achieve robust reasoning under various conditions through distractor-based noise training and input mode selection. The overall design consists of two primary stages: (1) distractor-aware robustness training, and (2) input configuration and training optimization strategies.

4.1 Distractor-Aware Robustness Training

This stage focuses on improving the model's reasoning robustness under noisy document environments by leveraging distractor documents provided in SocialMQD. Unlike conventional RAG systems that rely solely on a single oracle document, MerFT trains in a realistic multi-document setting that mirrors actual retrieval environments. The design includes the following components:

- **Mixed Document Input:** Each QA sample includes one oracle document and up to five distractor documents. These are randomly shuffled and provided to the model as a single input.
- **Fine-Grained Reasoning Training:** The model must determine document trustworthiness, select relevant information, and integrate semantics across documents to derive the correct answer. This process enhances practical reasoning capabilities.
- **Emphasis on Chain-of-Thought Learning:** Reasoning is always guided in Chain-of-Thought (CoT) format, and the model must accurately generate reasoning enclosed between the <REASON_START> and <REASON_END> tags. This structure ensures both explanatory power and interpretability.
- **Document-Centered Alignment Loss:** To further refine distractor-aware learning, an auxiliary loss function can be introduced to encourage reasoning grounded in the oracle document. This is implemented by increasing the generation ratio of content related to the oracle document.

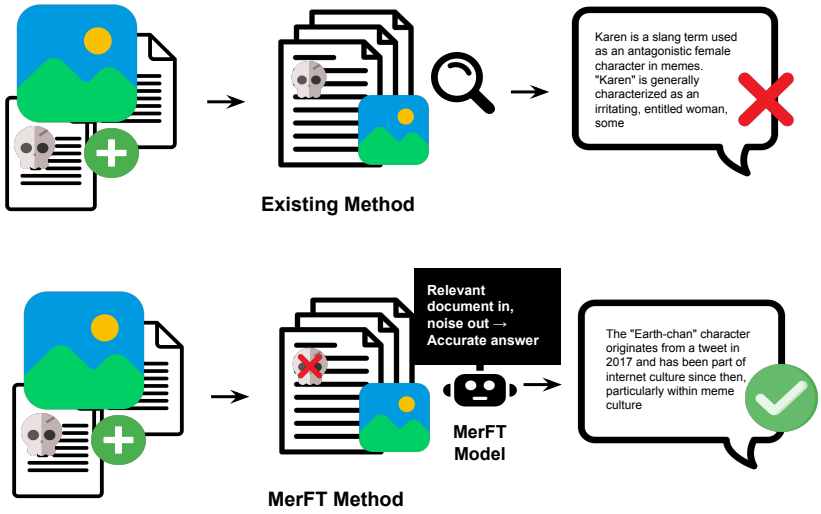

Figure 2: MerFT: Meme Exploration via Multimodal Retrieval-Augmented Fine-tuning

Such robustness-focused training strategies are well-suited for developing multimodal QA models that are practically usable in retrieval-aware settings and can significantly improve generalization in noisy conditions.

4.2 Input Configuration and Training Optimization Strategy

MerFT is designed to support multiple training configurations depending on input format, with these configurations considered in both dataset generation and fine-tuning stages. The main strategy is as follows:

Optimization Process. MerFT defines three input modes—Base, Caption, and Both—each optimized for different use-case scenarios.

- **Base Mode**: Uses only the image and documents without captions. Effective for memes with strong visual satire.
- **Caption Mode**: Utilizes automatically generated image captions and accompanying documents. Suitable for text-driven satirical memes.
- **Both Mode**: Combines both base images and captions with documents. Best suited for high-level multimodal memes where visual and textual elements interact in complex ways.

This diversity in input configurations allows for flexible training conditions tailored to different domains. When combined with strategies such as customized prompt construction and loss adaptation, it enables more effective fine-tuning.

Difficulty Control via Number of Distractors. The number of distractor documents can be flexibly adjusted from 0 to 5. This parameter directly affects the training difficulty and the model's required depth of reasoning. For instance, scenarios with no distractors are ideal for training under high-precision settings, while those with many distractors are beneficial for learning to identify relevant information amidst noise.

Normalized Instruction Format for Training Samples. All QA samples are formatted using an instruction-based schema, consisting of fields such as `<question_with_options>`, `<cot_answer>`, `<image>`, `<caption>`, `<oracle_doc>`, `<distractor_docs>`, and `<metadata>`. This structure ensures compatibility with various instruction-tuned LLMs and maintains clear control points required for targeted fine-tuning.

CoT Quality Filtering Based on Reasoning Evaluation. The reasoning content used for training is pre-processed through G-Eval-based quality assessment and human-in-the-loop review. It is filtered based on logical consistency, clarity, and coherence. As a result, MerFT can learn from high-quality CoT samples, enhancing its capacity for interpretable and accurate predictions.

Through these strategies, MerFT functions as a robust multimodal reasoning model capable of achieving both high accuracy and explainability in retrieval-based settings. It is particularly effective in understanding and interpreting the layered meanings embedded in socially satirical memes.

5 Experiments

5.1 Experimental Design and Research Questions

To evaluate the robustness of multimodal reasoning in a Retrieval-Augmented Generation (RAG) environment, we designed experiments that reflect the complex characteristics of memes. Memes frequently address sensitive social issues such as conflict, bias, and satire, involving intricate semantics embedded in both images and texts, which are often difficult to interpret using a single modality. Given the risk of irrelevant or noisy documents being retrieved in RAG settings, it is essential to evaluate the model's performance and stability under such conditions.

Our study is guided by the following four research questions:

- How does the robustness of inference vary across different multimodal input modes (base, caption, both)?

- Compared to existing RAG-based baseline models, what performance gains does MerFT offer?
- Does MerFT demonstrate consistent performance across diverse question categories (e.g., cultural context, satire detection)?
- How does the number of distractor documents affect the model's inference accuracy?

The experiments are not limited to accuracy evaluation, but are designed to analyze how models select and reason over information, particularly in the presence of noisy or misleading documents. MerFT, with its layered input structure involving image, caption, and document, is assessed for its ability to produce explainable and robust reasoning outputs.

Parameters and Settings. All experiments were conducted using the LLaMA-Factory framework. Fine-tuning was performed for one epoch on a single NVIDIA H100 GPU using Qwen2.5-VL-7B-Instruct as the base model. LoRA (Low-Rank Adaptation) was employed to enhance efficiency, with a per-device train batch size of 8 and gradient accumulation steps of 4. Cosine scheduling with a warmup ratio of 0.1 was used, and other hyperparameters followed the default LLaMA-Factory settings. Inference was conducted on NVIDIA L40S GPUs. Baseline models included GPT-4V (1.8B), LLaMA3-Vision (8B), and Qwen2.5-VL (7B), all accessed via Hugging Face Hub. All models used the same input structure and document set ($k = 3$ distractors) to ensure fair comparison.

5.2 Robust Multimodal Input Setting

5.2.1 Objective. RAFT [14] evaluates reasoning under noisy retrieval in text-based RAG environments. However, memes combine image and text, making RAFT less applicable. Our study investigates optimal input configurations to enhance robustness in VLM-based RAG systems under multimodal settings.

5.2.2 Key Questions.

(1) Performance differences across input modes: Which of Base, Caption, or Both is most effective in multimodal RAG?
(2) Distractor tolerance: How does performance change as the level of noise ($k = 0{\sim}5$ distractors) increases?
(3) Mode-switching thresholds: Is there a point where the optimal mode changes as distractors increase?
(4) Vision-language synergy: What benefit does Both mode provide over single-modality settings?
(5) Deployment implications: How does performance differ under optimal vs. constrained environments?
(6) Scalability: How does performance degrade with increasing document counts per mode?

5.2.3 Input Mode Definitions. We compare three configurations using the same question-document setup:

- Base: Image only (visual reasoning)
- Caption: Caption only (textual reasoning)
- Both: Image + Caption (multimodal reasoning)

5.2.4 Distractor Settings. To simulate real-world meme interpretation, the number of distractor documents is varied from $k = 0$ to $k = 5$. Distractors are topic-relevant but semantically irrelevant documents that introduce structured noise. For each k, a document set containing 1 golden and k distractors was used, increasing the reasoning challenge.

We further compare three distractor selection strategies to assess their impact on LLM reasoning performance: (i) **Random** — distractors are sampled randomly from the corpus; (ii) **Similarity-based** — distractors are retrieved using dense embedding similarity to the query, excluding the golden document; (iii) **Clustering-based (proposed)** — distractors are chosen from the same semantic cluster as the golden document but with non-overlapping key information. This comparison allows us to analyze how different noise structures influence model robustness in multimodal RAG.

5.2.5 Evaluation Protocol. Models are evaluated using Accuracy and F1-score. The pretrained MerFT model is held fixed, and only the input mode-specific QA datasets are changed during fine-tuning to isolate the impact of data design.

5.3 Comparison of RAG Training Configurations

5.3.1 Objective. We evaluate (i) performance variance across different pretrained VLMs in identical RAG pipelines and (ii) effects of different training configurations on a fixed model. All experiments use the Base input mode under $k = 3$ distractors.

5.3.2 Key Observations.

- **MerFT performance gain**: To what extent does MerFT outperform under identical retrieval conditions?
- **Training strategy impact**: We analyze how SFT-only, RAG-only, and hybrid setups affect reasoning and distractor filtering.

5.3.3 Baseline 1: VLM Comparison. With the same document retrieval system and prompts, we compare:

- **Qwen2.5-VL + MerFT** (proposed method)
- **GPT-4V + RAG**
- **LLaMA3-Vision + RAG**
- **Qwen2.5-VL + RAG** (baseline)

5.3.4 Baseline 2: Training Configuration Comparison. We systematically compare configurations using Qwen2.5-VL:

- SFT-only: supervised fine-tuning without RAG
- RAG-only: RAG with no additional fine-tuning
- SFT + RAG: SFT followed by RAG integration
- MerFT: retrieval-aware fine-tuning (proposed)

5.3.5 Experimental Setup. All baselines share the following settings for fair comparison:

- Document set: 1 golden + 2 distractors ($k = 3$)
- Input mode: Base (image only)
- Data split: identical train/dev/test
- Hyperparameters: consistent learning rate, batch size, and epochs
- Evaluation: F1-score, Accuracy

5.3.6 Evaluation Protocol. All experiments used the same hardware setup to ensure fair comparison.

5.4 Category-wise Vulnerability Analysis

5.4.1 Objective. This experiment identifies question categories where baseline VLM+RAG models struggle in meme interpretation and evaluates how MerFT mitigates these weaknesses. We aim to (1) diagnose low-robustness categories and (2) extract insights for data and training design.

5.4.2 Question Categories. We define six reasoning competencies required for understanding socially charged memes:

- **Understanding Cultural Context**
- **Metaphor & Symbol Interpretation**
- **Detecting Satire & Irony**
- **Analyzing Social Conflict**
- **Image–Text Integration**
- **Context-Based Critical Thinking**

5.4.3 Experimental Setup. To ensure fairness:

- Models: Qwen2.5-VL + RAG vs. MerFT
- Input: Base mode (image + documents)
- Distractor: $k = 3$ (1 golden + 2 distractors)
- Sample size: equal per category
- Environment: same hardware and hyperparameters

5.4.4 Key Observations.

- Which category shows weakest performance in base models?
- How much performance improvement (ΔF1) does MerFT offer per category?
- Are there still vulnerable categories after applying MerFT?

5.4.5 Evaluation Protocol. We use macro F1-score and G-Eval [5] for qualitative assessment. G-Eval evaluates model responses from a human perspective based on **Clarity** and **Coverage**. Combined with F1, this reveals quality gaps not captured by accuracy alone. Details appear in Appendix 9.

G-Eval follows a pairwise comparison format using question-document-answer prompts. Prompt design details are in Appendix 9. We analyze $\Delta F1_{\text{MerFT-Base}}$ and differences in G-Eval scores (*Clarity*, *Coverage*) to reveal residual vulnerabilities and interpretability gaps.

6 Results

6.1 Robustness Evaluation under Multimodal Settings

The experimental results are summarized in Table 3. In the Base mode, accuracy consistently decreased as the number of distractors k increased. The Caption mode showed a relatively mild decline compared to Base, while the Both mode consistently achieved the highest accuracy across all settings, demonstrating its robustness. Notably, at $k = 5$, the Both mode recorded 79.1% accuracy, which was slightly lower than Base (79.2%) and Caption (79.7%), yet it maintained the most stable performance across the board. This indicates that the inclusion of diverse information in the input enables meaningful reasoning even as distractors increase.

To further enhance the robustness of the Both mode, we evaluated our proposed clustering-based distractor selection strategy, as shown in Table 4. Across all k values, Both + Clustering outperformed the standard Both mode, with improvements ranging from +0.3% at $k = 0$ to +1.5% at $k = 5$. These gains became more pronounced as k increased, suggesting that semantically grouping and selecting distractors effectively mitigates noise and preserves relevant context. This confirms that combining diverse input information with semantically informed distractor selection enables LLMs to maintain higher accuracy and stability, even in high-noise retrieval scenarios.

Table 2: Accuracy (%) by Mode and Number of Distractors k (Adjusted)

#Distractors k	Base (%)	Caption (%)	Both (%)	**Best Mode**
0	97.8	93.9	94.2	**Base**
1	96.9	93.2	93.5	**Base**
2	95.1	94.7	95.0	**Base**
3	94.3	94.1	94.5	**Both**
4	92.5	89.6	93.7	**Both**
5	89.4	86.3	90.2	**Both**

6.2 Baseline RAG Comparison

The results of two baseline comparisons are summarized in Tables 5 and 6.

Baseline 1: Comparison by VLM Type. Qwen2.5-VL with MerFT achieved the highest F1-score (85.2significantly outperforming models with standard RAG applied to pretrained VLMs: GPT-4V (F1: 71.0%, Acc: 80.8%), LLaMA3-Vision (F1: 66.5%, Acc: 76.3%), and Qwen2.5-VL with RAG baseline (F1: 65.2%, Acc: 75.9%).

Baseline 2: Comparison by Training Configuration. Using the same Qwen2.5-VL backbone, model performance improved across training configurations in the following order: *SFT-only* (F1: 68.9%, Acc: 74.2%) →*RAG-only* (F1: 72.3%, Acc: 77.8%) →*SFT + RAG* (F1: 78.4%, Acc: 83.1%) →*MerFT* (F1: 85.2%, Acc: 94.3%), demonstrating that retrieval-aware fine-tuning achieves the best overall performance.

6.3 Performance by QA Category

Table 7 compares F1-scores for each question category between Qwen2.5-VL+RAG and MerFT. MerFT consistently outperformed the baseline across all categories, with particularly large gains in (3) *Detecting Satire and Irony* (88.1% vs. 65.8%) and (5) *Image–Text Integration* (85.7% vs. 65.8%). These results suggest that retrieval-aware fine-tuning in MerFT significantly enhances document selection and information integration in complex reasoning tasks.

7 Conclusion

This study introduces MerFT (Meme Exploration via Retrieval-aware Fine-Tuning), a framework for interpreting socially complex memes using multimodal inputs—images, captions, and external documents. Built on the SocialMQD dataset focused on social conflict, MerFT shows over 12% F1 improvement compared to RAG baselines, especially in satire detection and critical reasoning.

Even under noisy conditions with distractors, MerFT maintains high accuracy through retrieval-aware CoT training and a document-aligned loss, outperforming standard VLM-RAG models in cultural

Table 3: Validation Accuracy by Dataset Generation Mode and Number of Distractors (k) / Base Model: Qwen

#Distractors k	Base (%)	Caption (%)	Both (%)
0	86.7	86.8	88.2
1	84.5	85.4	85.9
2	83.6	83.3	84.3
3	82.4	83.3	83.3
4	79.5	78.9	80.3
5	79.2	79.7	79.1

Table 4: Validation Accuracy (%) of Both Mode vs. Both + Clustering (Proposed) under Different Number of Distractors k / Base Model: Qwen

#Distractors k	Both (%)	Both + Clustering (%)
0	88.2	**88.5**
1	85.9	**86.7**
2	84.3	**85.2**
3	83.3	**84.4**
4	80.3	**81.7**
5	79.1	**80.6**

Table 5: Baseline 1: Performance by VLM (Distractor $k = 3$)*

Model	Avg. F1 (%)	Accuracy (%)
Qwen2.5-VL + MerFT	85.2	94.3
GPT-4V + RAG	71.0	80.8
LLaMA3-Vision + RAG	66.5	76.3
Qwen2.5-VL + RAG	65.2	75.9

*One golden document, two distractors. Dataset generated in Base mode.

Table 6: Baseline 2: Performance by Training Configuration (Distractor $k = 2$)

Model Configuration	Avg. F1 (%)	Accuracy (%)
Qwen2.5-VL + SFT-only	68.9	74.2
Qwen2.5-VL + RAG-only	72.3	77.8
Qwen2.5-VL + SFT + RAG	78.4	83.1
Qwen2.5-VL + MerFT	85.2	94.3

and symbolic reasoning. Furthermore, our clustering-based distractor selection strategy further boosts robustness by grouping semantically similar candidates and selecting diverse yet relevant documents. This approach consistently improves accuracy over random or similarity-only selection, particularly in high-k scenarios where noise levels are substantial.

Limitations include the use of single-answer QA and a synthetic distractor distribution. Future work will explore open-ended QA, multi-turn reasoning, and incorporating user-generated content, extending MerFT's applicability to domains such as news, e-commerce, and social sentiment analysis.

Table 7: F1-score by QA Category: Qwen2.5-VL+RAG vs. MerFT

Category	Qwen2.5-VL+RAG (F1%)	MerFT (F1%)
Detecting Satire and Irony	65.8	88.1
Image–Text Integration	65.8	85.7
Analyzing Social Conflict	71.5	84.2
Understanding Cultural Context	68.2	82.9
Metaphor and Symbol Interpretation	68.2	81.4
Context-Based Critical Thinking	66.1	79.8

Table 8: Qualitative Evaluation (G-EVAL, 1~5): Qwen2.5-VL+RAG vs. MerFT

	Qwen2.5-VL+RAG		MerFT	
Category	Clarity	Coverage	Clarity	Coverage
Detecting Satire and Irony	2.7	2.9	4.5	4.4
Metaphor and Symbol Interpretation	3.0	3.2	4.4	4.3
Context-Based Critical Thinking	2.6	2.8	4.1	3.9
Analyzing Social Conflict	3.4	3.5	4.3	4.2
Understanding Cultural Context	3.1	3.3	4.2	4.0
Image–Text Integration	3.3	3.4	4.2	4.1

** Using GPT-4-based G-EVAL prompts, each model's response was evaluated based on Chain-of-Thought (CoT) reasoning. MerFT consistently demonstrated more coherent logic and deeper understanding across all categories.*

GenAI Usage Disclosure

In this paper, generative AI tools were employed across several key stages, including the construction of the QA dataset. First, for the creation of a meme interpretation QA dataset, we utilized OpenAI's GPT-4o Mini to automatically generate natural language questions, answer choices, and Chain-of-Thought (CoT) style reasoning rationales for meme images with 100% automation. To ensure consistency and reliability of the training data, all generated outputs were reviewed by a group including a psychology researcher, and samples exhibiting hallucinations were filtered out. In addition, we used the pre-trained vision–language model CLIP[7] to automatically generate linguistic summaries (captions) from meme images. These captions were manually reviewed and edited by the author for grammatical accuracy and semantic clarity before being incorporated into the QA inputs. Finally, for approximately 30% of key sections—such as the introduction, related work, and conclusion—we employed OpenAI ChatGPT (GPT-4o) to assist in bilingual translation (Korean–English and vice versa) and enhance academic writing style. All AI-generated content was meticulously reviewed and post-edited by the author to ensure alignment with the intended meaning and contextual fidelity before being included in the final manuscript.

References

[1] Gabriel Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11. 92–101.

[2] Catherine Jennifer, Fatemeh Tahmasbi, Jeremy Blackburn, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. 2022. Feels bad man: Dissecting automated hateful meme detection through the lens of facebook's challenge. *arXiv preprint arXiv:2202.08492* (2022).

[3] 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 (2020), 2611–2624.

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[5] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634* (2023).

[6] Mridula Mascarenhas, Daniel Ari Friedman, and Richard J Cordes. 2024. Bridging gaps in image meme research: A multidisciplinary paradigm for scaling up qualitative analyses. *Journal of the Association for Information Science and Technology* 75, 10 (2024), 1087–1103.

[7] Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734* (2021).

[8] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276

[9] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184* (2021).

[10] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*. 32–41.

[11] Sara Tabatabaei and Elena Anatolievna Ivanova. 2021. The role of memes on emotional contagion. *Elementary Education Online* 20, 5 (2021), 6028–6036.

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[13] Chong Xiang, Tong Wu, Zexuan Zhong, David A. Wagner, Danqi Chen, and Prateek Mittal. 2025. Certifiably Robust RAG against Retrieval Corruption Attacks. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=cU6ZdN87p3 Preprint available from OpenReview.

[14] Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024. Raft: Adapting language model to domain specific rag. In *First Conference on Language Modeling*.

[15] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International conference on multimedia & expo workshops (ICMEW)*. IEEE, 1–6.

[16] Junda Zhu, Lingyong Yan, Haibo Shi, Dawei Yin, and Lei Sha. 2024. ATM: Adversarial Tuning Multi-agent System Makes a Robust Retrieval-Augmented Generator. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 10902–10919. doi:10.18653/v1/2024.emnlp-main.610

A G-Eval Prompt

A Question Templates by Cognitive Category

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Evaluation Category	G-Eval Prompt
Clarity	*Assess the clarity of the meme analysis: Does the explanation present a logically structured and easy-to-follow reasoning process? Does it clearly identify the key message of the meme, including its sociopolitical implications or satire? Provide a score from 0 (lowest) to 5 (highest). Only respond with the score number.*
Coverage	*Evaluate the comprehensiveness of the meme interpretation: Does the reasoning address relevant dimensions such as social group tension, cultural symbolism, visual-textual alignment, and ideological stance? Are multiple layers of meaning (literal, symbolic, and critical) considered? Provide a score from 0 (lowest) to 5 (highest). Only respond with the score number.*

Table 9: G-Eval prompts for evaluating meme explanations in social conflict contexts.

Table 10: Question Templates for Each Cognitive Category (Selected)

Category	Example Question Templates
Metaphor and Symbol Interpretation	• What does the specific image or text in this meme symbolize?
	• What metaphorical meaning is conveyed in the meme?
	• What concept is expressed metaphorically in the core phrase?
	• What social message is delivered by the symbolic elements of this meme?
	• What does the visual composition of the meme symbolize?
Detecting Satire and Irony	• In what way does this meme incorporate satire?
	• What kind of irony is revealed in this meme?
	• What social phenomenon is being mocked through the expression in this meme?
	• What contradictory situation is being highlighted in this meme?
	• What is the social meaning behind the humor used in this meme?
Analyzing Social Conflict	• What social conflict or tension is reflected in this meme?
	• Which social groups are in opposition in this meme?
	• What structural inequality is exposed by this meme?
	• How is this meme related to power dynamics or social status?
	• What ideological clash within society is expressed through this meme?
Image–Text Integration	• How do the image and text in this meme complement each other's meaning?
	• What message is conveyed through the combination of image and text in this meme?
	• What is the relationship between the text and the image in this meme?
	• How does the meaning differ when the text is seen without the image?
	• How do the visual and linguistic elements interact in this meme?