Exploration of How Hate is Framed on Social Media

Rakshitha Rao Ailneni, Sanda M. Harabagiu
Human Language Technology Research Institute, The University of Texas at Dallas
Richardson, TX, USA
{rxa220074, sanda}@utdallas.edu

Abstract

Understanding how hate is framed in multimodal social media content is crucial for developing interpretable and robust hate detection systems. We present the MM-HateFrames Dataset, a large-scale resource encoding 2,298 Hate Frames (HFs) and their corresponding rationales discovered from two benchmark datasets—Hateful Memes and MMHS150K—comprising over 11K+ social media multimodal posts. This novel dataset allowed us to explore several generative and non-generative methods to automatically discover the way hate is framed when relying on MM-HateFrames. Experimental evaluations show that few-shot LMM prompting generates the most coherent and sound frame articulations. Therefore, the MM-HateFrames Dataset provides a valuable foundation for future research in hate speech understanding, frame articulation, and explainable multimodal NLP, enabling models to interpret not only *whether* content is hateful but also *how* hate is conceptually framed.

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

Keywords: Framing, Hate speech, Multimedia, Large Language Models

1. Introduction

Even though social media platforms such as Instagram, Facebook, Reddit, and X have empowered individuals by amplifying their voices and facilitating freedom of expression, they have simultaneously become fertile ground for hate speech and other forms of online harassment. Hate speech is defined as any form of communication that conveys, promotes, or has the capacity to incite hatred against an individual or group based on a shared characteristic or group membership (Her, 2012). The identification of hate speech is essential not only for safeguarding the groups targeted by such social media communications but also for enabling the development of counter-narratives aimed at mitigating its harmful effects; cf. (Alyahya and Aldayel, 2024; Albanyan et al., 2023).

However, the identification of hate speech remains a complex challenge for three primary reasons. First, there is inconsistency in the terminology and definitions of hate speech across disciplines, which undermines efforts to build reliable computational models, as it remains unclear which criteria they should be trained to detect. Second, hate speech relies heavily on cultural influences, leading to varying individual interpretations. Third, although hate speech may target a wide range of groups based on demographics characterized by religion, gender, family status, or race and may vary significantly in its mode of delivery, prior research has disproportionately concentrated on narrow subsets of targets (e.g., women and immigrants) (Basile et al., 2019) or specific forms of expressing hate (e.g., implicit vs. explicit hate) (ElSherief et al., 2021). To address this lack of definitional clarity, (Korre et al., 2025) used Semantic Componential Analysis (SCA) to identify *definitional components* of hate speech. SCA is a linguistic technique, based on the principle of semantic compositionality, used to break down the meaning of words or phrases into their constituent parts or features. SCA enabled derivation of a hierarchy of Hate Defining Elements (HDEs) for hate speech informed by 493 definitions spanning five diverse cultural domains. (Korre et al., 2025) revealed that the main three dimensions of hate speech are (1) the *Target* of hate, (2) the *Intention/Purpose* of hate, and (3) the *Act/Means* of

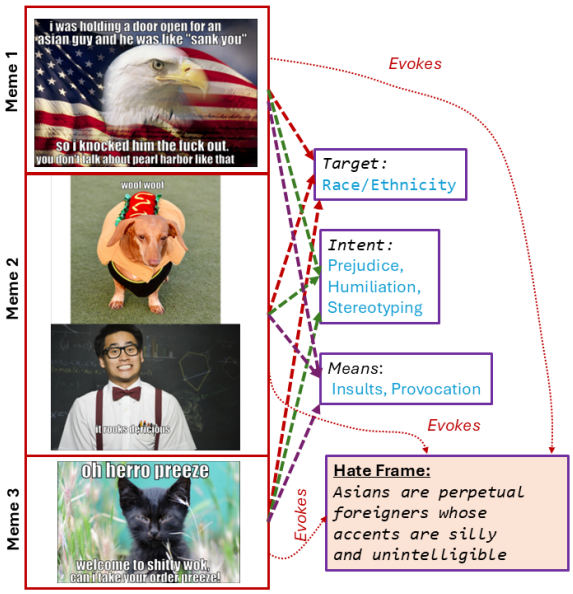

Figure 1: Hateful memes evoking the same Hate Frame (HF). The HF reveals the hate *Target*, *Intention* and *Means*.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8400–8414
11–16 May 2026.

delivering hate speech. Figure 1 illustrates three different memes and the HDEs corresponding to them along the Target, Intent, and Means dimensions.

We find that adopting the hierarchy of HDE for hate speech is necessary but not sufficient when processing hate speech. Just identifying the HDE from a MultiMedia Posting (MMP) does not inform **the way hate is framed**, especially when the communications contain not only text but also images, as shown in Figure 1. The HF evoked by the three memes illustrated in the figure points to *who* is the Target of hate, namely Asians, and articulates *why* the memes are hateful, because the usage of prejudice, humiliation, and stereotyping is evident due to the statement of the HF from the figure, and moreover, the HF exemplifies *how* hate is framed through insults and provocation. The discovery of HFs allows us to analyze which kind of hate is prevalent in a social media platform and to pinpoint all communications that evoke such HFs. Moreover, HFs can inform the design of counter-narratives tailored to address specific forms of hate (Guest et al., 2021; Chung et al., 2019).

Hate Framings (HF) are special forms of framings. Framing is a concept central to communication sciences. The definition of framings provided by Entman (1993) notes that *"to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."* Based on this definition, the Hate Frames (HFs) address the various problems raised when hate speech is performed, encoded by the Hate Defining Elements (HDE). HFs also provide a causal interpretation of the problems they address through their articulation.

In this paper we introduce **MM-HateFrames**, a corpus of multimodal media postings (MMPs) annotated with (a) the Hate Framings (HF) evoked by each multimodal posting as well as (b) Hate Defining Elements (HDEs) characterizing the HF. An example of such annotation is provided in Figure 1. Although we are the first to introduce and annotate HFs, we were inspired by recent work on frame discovery and articulation reported in (Weinzierl and Harabagiu, 2024) and (Ailneni and Harabagiu, 2025). (Weinzierl and Harabagiu, 2024) introduced a method for automatically discovering and articulating frames of communication evoked in text-based social media postings addressing vaccine hesitancy. (Ailneni and Harabagiu, 2025) proposed a method for discovering and articulating framings of misogyny from misogynistic memes without requiring prior knowledge of misogyny-related problems. Both these methods highlighted the fact that multiple postings evoke the same frame and that, in addition, relations such as *paraphrasing*, *specializing,* and *contradiction* span the discovered frames. Following these observations, MM-HateFrames also links the HFs through such relations. MM-HateFrames uses memes originating from two major hate datasets: (1) the Hateful Memes Dataset (Kiela et al., 2021) and (2) the MMHS150K dataset (Gomez et al., 2019)

MMP = Multimodal Media Posting	**LMM** = Large Multimodal Model
HF = Hate Frame	**LLM** = Large Language Model
HDE = Hate Defining Element	**CoT** = Chain of Thought
3 *different* **CoT** (Chain of Thought) prompting methods:	
1] CoT^{HDE} = CoT prompting for discovering HDEs	
2] CoT^{HF} = CoT prompting for discovering HFs.	
3] CoT ^{Rel} =CoT prompting for discovering relations between HFs.	

Figure 2: Acronyms used in the paper.

This paper makes the following contributions:
◁**1**▷ We introduce the first dataset of hate frames generated from multimodal hate speech, encompassing the full spectrum of targets, intents, and actions, rather than being limited to a narrow subset of hate phenomena. We release the MM-HateFrames annotations and software on GitHub[1].
◁**2**▷ For each Hate Frame (HF) available in MM-HateFrames, we provide expert-verified rationales that explain why the HF is being articulated, thereby enabling deeper interpretability.
◁**3**▷ We present an annotation procedure and guidelines for discovering and verifying Hate Frames (HFs) that are not confined to a single form of hate and that can be applied to any multimodal hate speech dataset.
◁**4**▷ We describe experiments with using the annotations available in MM-HateFrames for learning to discover and articulate Hate Frames (HFs).

2. Background and Related Work

2.1. Frame Discovery and Annotation

Early research aiming at frame discovery from social media has generally relied on unsupervised approaches (Russell Neuman et al., 2014; Meraz and Papacharissi, 2013) which revealed interesting framing patterns highlighted by lexical terms. However, these methods neither discovered the frame problems nor did they articulate the causes of those problems; therefore, they do not have the framing definition from Entman (1993). The Media Frames Corpus (MFC) (Card et al., 2015) was the first effort that annotated fifteen dimensions of policy frames, addressing such problems as Constitutionality and Jurisprudence or Security and Defense. MFC enabled supervised approaches to frame problem

[1] https://github.com/rak55/MM-HateFrames-Dataset

discovery, using sequential and pretrained models (Naderi and Hirst, 2017; Khanehzar et al., 2019).

Mendelsohn et al. (2021) considered the analysis of frames in social media by identifying immigration policy problems with multi-label classification methods, relying on RoBERTa (Liu et al., 2019). However, (Weinzierl and Harabagiu, 2024) were the first to leverage the reasoning capabilities of Large Language Models (LLMs) for automatically discovering and articulating frames from text-based social media. Ailneni and Harabagiu (2025) introduced Dis-MP&F, a method for discovering and articulating misogyny frames from memes, which also generates a data-driven Taxonomy of Misogyny (ToM) and is the first to articulate 1089 Frames of Misogyny (FoMs) without prior knowledge of all Misogyny Problems (MPs). The annotation of HFs in MM-HateFrames is inspired by the work reported in (Weinzierl and Harabagiu, 2024) and (Ailneni and Harabagiu, 2025), but it is also informed by the taxonomy provided in (Korre et al., 2025).

2.2. Identifying Hate on Social Media

Hate speech detection on social media initially relied on text-only datasets and classifiers, focusing on overtly abusive language (Fortuna and Nunes, 2018; Davidson et al., 2017; Waseem and Hovy, 2016). However, with the rise of multimodal content, research has shifted toward capturing the interplay between text and imagery. The Hateful Memes dataset (Kiela et al., 2021) marked a key milestone in hate detection research, serving as a benchmark for evaluating multimodal transformer architectures such as UNITER (Chen et al., 2020), VisualBERT (Li et al., 2019), and CLIP (Radford et al., 2021) in modeling cross-modal reasoning. Similarly, MMHS150K (Gomez et al., 2019) explored hate in Twitter image–text pairs, where multimodal fusion models were compared against text-only baselines, revealing challenges in aligning the two modalities. The MAMI dataset (Fersini et al., 2022) extended this effort to misogyny detection, inspiring fusion architectures that combine visual and textual encoders. Moving beyond explicit hate, HateXplain (Mathew et al., 2022) emphasized interpretability by providing target-group and rationale annotations, enabling attention-based models to explain decisions, while ToxiGen (Hartvigsen et al., 2022) addressed implicit hate through synthetic adversarial examples generated by LLMs. (Bui et al., 2025) introduces a multimodal, multilingual, and multicultural dataset for hate speech detection across languages and cultural contexts. However, none of the prior methods considered the analysis of how hate is framed, and none enables the understanding not only of *who* is hated and *why* but also of *how* they are hated. In contrast, our work focuses on modeling how hate is conceptually framed by providing annotations of Hate Frames, Hate Defining Elements, rationales, and inter-frame relations, enabling interpretability beyond detection.

3. Annotating the MM-HateFrames

The MM-HateFrames dataset consists of a corpus of MMPs, which we selected from existing hate speech datasets. For each MMP, the HF it evokes was annotated, as well as the corresponding HDEs, which were available from the taxonomy of HDEs released in (Korre et al., 2025). First, we annotate the HDEs discovered for each posting from our corpus. Then we articulate and verify the HFs evoked by postings. Finally, we also annotate in MM-HateFrames the relations that span the annotated HFs, enabling us to consolidate the most relevant HFs while also allowing the inspection of the various ways hate is framed on social media as well as relations between HFs, e.g., contradiction or subsumption.

3.1. The Corpus

The selection of the multimedia corpus of social media postings results from surveying prominent hate speech datasets previously released in the literature and identifying those that meet three key criteria: (1) multimodal composition (containing both image and text modalities), (2) English language content, and (3) diversity across targeted demographic groups. Based on these criteria, we selected two benchmark datasets that satisfy the above requirements: the *Hateful Memes* Dataset released by Facebook AI (Kiela et al., 2021) and *MMHS150K* introduced by (Gomez et al., 2019).

The Hateful Memes Dataset: is comprising 10k memes generated artificially by placing meme text over a new underlying stock image. Each meme was rated by three different annotators on a scale of 1-3, with a 1 indicating definitely hateful, a 2 indicating not sure, and a 3 indicating definitely not hateful. Memes with disagreement were annotated by an expert. The memes encompass hate directed at diverse demographics, including religion, nationality, sex, socio-economic status, etc.

Preprocessing the Hateful Memes Dataset: The dataset is partitioned into 8,500 memes for training, 500 for validation, and 1,000 for testing. Among the 9,000 memes comprising the training and validation splits, 3,300 are labeled as hateful. These hateful instances form the basis of our frame articulation analysis. Since the dataset already provides transcribed meme text, additional optical character recognition (OCR) extraction or further preprocessing was not required.

Although the examples are artificially generated,

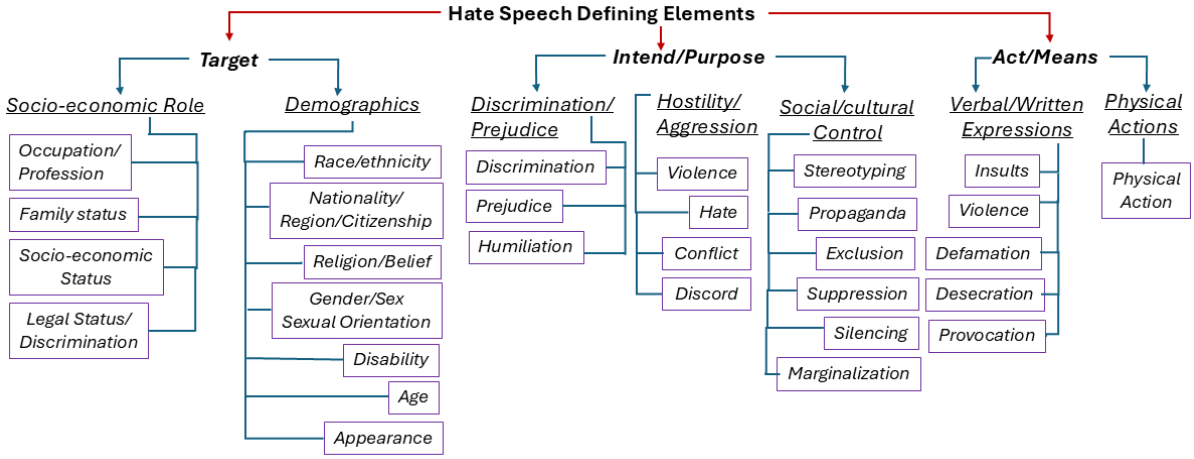

Figure 3: The three levels of the Hierarchy of Hate Defining Elements, cf. (Korre et al., 2025).

the dataset was explicitly designed to mimic realistic meme-style communication and to capture subtle multimodal interactions that occur in online hate speech. In this work, we use the dataset primarily as a controlled benchmark for studying multimodal hate framing. To complement this synthetic setting, we also include MMHS150K, which contains naturally occurring social media posts, allowing our analysis to cover both curated benchmark data and real-world multimodal hate expressions.

The MMHS150K dataset: The MMHS150K dataset comprises 150,000 multimodal tweets (text + image) collected via the Twitter API between September 2018 and February 2019. Tweets containing any of the 51 Hatebase terms (ElSherief et al., 2018) were retained and crowdsourced into six categories: *no attack, racist, sexist, homophobic, religion-based attack,* or *attack on other communities*.

Preprocessing the MMHS150K dataset: Each entry in the dataset contains the tweet text, its associated image, and an array of three integer labels (ranging from 0–5), one provided by each of three Amazon Mechanical Turk annotators. The label mapping is as follows: 0—Not Hate, 1—Racist, 2—Sexist, 3—Homophobic, 4—Religion-based, and 5—Other Hate. To isolate hateful instances, we retained posts where the majority of annotators agreed on a hate label, yielding 8,579 hateful posts from a total of 149,823 entries. Posts were excluded from the hateful subset if at least one annotator labeled them as not hate. The textual content of the post was further preprocessed to remove URLs and extraneous characters to ensure linguistic consistency.

3.2. A Taxonomy of Hate Defining Elements

Korre et al. (2025) processed 493 hate speech definitions drawn from five distinct domains—hate speech laws, Wikipedia, dictionaries, NLP research papers, and online—to generate a taxonomy of HDEs through Semantic Componential Analysis (SCA). Keywords were extracted and annotated as components of an HDE taxonomy organized around three main elements: (1) the hate target, (2) the hate intent/purpose, and (3) the hate act/means. Each of these dimensions is specialized along two additional levels of the hierarchy, as shown in Figure 3. But more importantly, the HDEs encoded in this hierarchy enable us to consider the problems addressed by the HFs that are discovered in Phase B in the following way: (1) the Target and all its subsumed HDEs reflect *who* is hated; (2) the Intent/Purpose and all its subsumed HDEs reflect *why* hate is framed in a certain way; while (3) the Act/Means and all its subsumed HDEs reflect *how* hate is framed. Therefore, the problems addressed by the HFs annotated in MM-HateFrames are addressing who is hated, how they are hated, and why they are hated. The HFs articulate these problems addressed through the HDEs.

3.3. Hate Frames Articulation

HFs are discovered and articulated from both datasets of our corpus in four phases:

◁**1**▷ In Phase A, all the MultiModal Posts (MMPs) are annotated with HDEs available from the hierarchy illustrated in Figure 3 using a zero-shot CoT prompting (Wei et al., 2022), namely CoT^{HDE}. All HDEs discovered in this way are verified and corrected when necessary by human experts.

◁**2**▷ In Phase B, HFs are discovered and articulated, given the HDE already annotated in each

Frames of communication select particular aspects of an issue and make them salient in communicating a message. Social science stipulates that discourse almost inescapably involves framing – a strategy of highlighting certain issues to promote a certain interpretation or attitude. It has been argued that "to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote problem definition, causal interpretation, moral evaluation, and/or treatment recommendation."

Task Overview:

You will be tasked with identifying and articulating hate framings on the social media postings.
For each input post, you will be provided with the Hate Defining Elements (HDEs) that the post addresses, categorized across three dimensions: 1) the Target of hate, (2) the Intention/Purpose of hate and (3) the Act/Means of delivering hate speech.

Procedure:

You should first discuss your reasoning and then provide your final decision.
Each post may or may not contain one or more Hate Frames (HFs). Your analysis proceeds as follows:
(a) Reason about whether the posting contains a frame (or more frames) or just states something factual or an experience. This reasoning should encompass the three HDEs. If the posting contains a frame, the next step is (b) Articulate that frame succinctly. You will perform these steps until the answer to (a) is false, either because there are no frames in the posting, or because you have already articulated all the frames.

Figure 4: System prompt used in Phase B.

MMP, using a second CoT prompting, namely CoT^{HF}, which relies on retrieval-based in-context learning.

◁**3**▷ In Phase C, possible relations between the HFs revealed in Phase B are discovered using few-shot CoT prompting of an LLM, namely CoT^{rel}. Non-relevant HFs are also filtered out in this phase.

◁**4**▷ In Phase D, the relevant HFs emerging from Phase C are evaluated by five annotators for clarity and soundness, metrics defined in (Weinzierl and Harabagiu, 2024) and the HFs deemed to have low scores using these metrics are edited by a human expert annotator.

3.3.1. Phase A: Annotating Hate Defining Elements

We leverage the zero-shot reasoning capabilities of LLMs (Kojima et al., 2022) to predict the target, intent, and means of each MMP, using CoT prompting (Wei et al., 2022). In this prompting setup, CoT^{HDE} instructs a Large Multimodal Model (LMM) to consider the HDE hierarchy to select the applicable targets, intents, and actions of a given MMP while also providing a rationale for each selected HDE. The LMM receives both the meme image and its associated text as input. In this way, for each MMP, this process yields a set of HDEs across the three dimensions (target, intent, means) along with their corresponding rationales.

Subsequently, three annotators review these predicted HDEs and edit them when necessary. An annotated HDE is retained only if a majority of annotators agree on its applicability to the MMP after considering the rationale generated alongside it. Inter-annotator reliability, measured using Krippendorff's α (Krippendorff, 2011), was 0.82, indicating strong agreement among annotators.

3.3.2. Phase B: Retrieval-Based In-Context Learning of Hate Frame Articulation

The in-context few-shot learning ability of LLMs has enabled them to generalize to unseen cases without additional fine-tuning, thereby unlocking a range of new technological possibilities. However, as (Liu et al., 2022) showed, an LLM's in-context performance can vary markedly with the choice of examples. This observation has motivated research in prompt retrieval, where, for a given test instance, training examples are selected for inclusion in the prompt according to a similarity metric. Prior work has either relied on off-the-shelf unsupervised similarity metrics or trained prompt retrievers to select examples based on surface-level similarity (Wang et al., 2024; Das et al., 2021). Building on these approaches, we retrieve demonstrations for each meme by leveraging CLIP embeddings (Radford et al., 2021) to capture deeper semantic similarity with the input.

To enable this retrieval, we first construct a demonstration bank comprising 250 MMPs annotated with HFs and their rationales by our group of experts. The bank is designed to ensure broad coverage, with at least five demonstrations spanning every HDE illustrated in Figure 3. Each candidate demonstration is embedded into a joint vision–language space by computing the element-wise average of normalized image and text embeddings derived from CLIP (Radford et al., 2021). The resulting vector representations are stored in a FAISS index (Douze et al., 2025) to facilitate efficient nearest-neighbor search.

During inference, for each input MMP, we restrict the candidate pool to demonstrations sharing at least one Hate Dimension Element (HDE) across all three hate dimensions shown in Figure 3. We compute the unified CLIP embedding of the input MMP and retrieve the top-k nearest demonstrations from the FAISS index. The **GPT-5-mini** model is then prompted using CoT reasoning

Figure 5: Examples of relations between HFs.

(CoT^{HF}) (Wei et al., 2022) to articulate the underlying HFs and their rationales (see Figure 4). In this way, 8,000 unique HFs were generated for the MM-HateFrames dataset and 12,214 HFs for MMHS150K.

3.3.3. Phase C: Discovering Relations Between Hate Frames

After articulating HFs in Phase B, we aimed to identify relations among them. Following (Weinzierl and Harabagiu, 2024), two HFs are labeled as *paraphrases* if they share identical HDEs and underlying causes, and as *contradictions* if their causes conflict. When two HFs share the same HDEs but one conveys a more specific cause, it is marked as a *specialization*. Examples of these relation types are shown in Figure 5. For each HF_A, we retrieve its top-k most similar frames S^A_{HF} using Sentence-BERT (Reimers and Gurevych, 2019). Each candidate pair (HF_A, HF_B), where $HF_B \in S^A_{HF}$, is then evaluated by **GPT-5-mini** with few-shot CoT reasoning (CoT^{Rel}) to determine the relation type and provide a supporting rationale. The prompts follow (Ailneni and Harabagiu, 2025) and include two examples each of paraphrase, contradiction, specialization, and unrelated pairs.

We apply this procedure to both datasets, merging paraphrased HFs to form a consolidated set. The relevance of each HF in the consolidated set is then computed as the total number of MMPs that evoke it, as in (Weinzierl and Harabagiu, 2024). Only HFs evoked by at least two MMPs are retained in the final set. Table 1 summarizes the number of articulated HFs from Phase B, the discovered relations (paraphrase, contradiction, specialization) in this phase, and the final set of relevant HFs for both datasets. To ensure the reliability of relation discovery, each predicted relation is manually verified by three annotators, who review and revise the LLM-generated type and rationale when necessary. A relation is finalized only if a majority of annotators agree on its validity. Inter-annotator reliability, measured using Krippendorff's α (Krippendorff, 2011), was 0.85 for verifying the validity of a relation.

Category	Hateful Memes	MMHS150K
Articulated HFs	8000	12214
Paraphrases	16572	22374
Specializations	53	67
Contradictions	43	32
Final HFs	1085	1213

Table 1: Number of HFs discovered in Phase A; number and type of relations between HFs discovered in Phase B; final number of HFs selected in Phase B using our method.

Figure 6: Examples of Memes with articulated HFs and rationales evaluated for clarity and soundness.

3.3.4. Phase D: Hate Frame Refinement

Once the final set of HFs was obtained for our corpus, we evaluated them using the criteria introduced in (Weinzierl and Harabagiu, 2024): (a) the *soundness* of the rationale generated by the LMM when prompted with CoT^{HF}, and (b) the *clarity* of the HF articulation. Five graduate students with near-native English proficiency and a computing background conducted these evaluations. The guidelines given to the annotators include the definition of an HF and an HDE, the hierarchy of HDEs (Korre et al., 2025) and few examples demonstrating the annotation procedure for some example HFs and their corresponding rationales.

For clarity, annotators were instructed to verify whether the HF was truly a frame rather than a factual statement and whether it represented a correct and contextually grounded interpretation of the input meme. For soundness, they assessed whether the rationale was relevant to the input post by referencing the target, intent, and means of hate.

For example, in Figure 6, the articulated HF abstracts the meme's message into a broader framing rather than repeating surface-level content. Moreover, the rationale explicitly identifies the target (male officers) and correctly characterizes the in-

Target Category	Hateful Memes	MMHS150K
Race / Ethnicity	491	540
Religion / Belief	383	124
Gender/Sex/Sexual Orientation	346	382
Nationality/Region/Citizenship	256	209
Appearance	153	189
Age	105	109
Disability	90	340
Legal Status/Discrimination	61	73
Occupation/Profession	61	145
Socio-economic Status	50	219
Family Status	32	16

Table 2: Distribution of HFs across targets for both datasets from MMHateFrames.

tent as humiliation through derogatory stereotyping, thereby grounding the frame in the meme's hateful purpose.

In contrast, the right-side meme fails at both criteria. The HF is merely a factual reformulation (e.g., using "framed as"), which does not capture a causal narrative. Additionally, the rationale misidentifies the intent as mere stereotyping, whereas the meme conveys exclusionary and hateful messaging. Because both the articulation and the identification of intent are inaccurate, the example fails both clarity and soundness.

Each HF was evaluated for clarity and soundness using binary (*yes*/*no*) labels. To address potential bias inherent in semi-automatic annotation settings, model-generated outputs were treated strictly as initial proposals rather than final labels. Annotators were instructed to independently assess each HF and its rationale against the multimodal content. Frames or rationales flagged by at least two annotators as failing either criterion were subsequently reviewed and revised by an expert annotator, resulting in revisions to 189 HFs from the Hateful Memes dataset and 205 from MMHS150K. This multi-level verification procedure ensured that the final annotations reflect human judgment and critical evaluation rather than passive acceptance of model suggestions.

The inter-annotator agreement measured as Krippendorff's alpha α (Krippendorff, 2011) is 0.75 for clarity and 0.73 for soundness, which are considered to be strong agreement.

4. Analysis of the MM-HateFrames dataset

MM-HateFrames annotates 24,122 HDEs from the hierarchy illustrated in Figure 3, out of which 4,449 correspond to hate targets, 13,918 correspond to hate intent/purpose, and 5,755 correspond to hate acts/means. But more importantly, each of these HDEs was addressed by 2,298 HFs that we have also annotated. The HFs can be inspected along several criteria: (1) the hate targets they address, (2) the hate intent/purposes, and (3) the hate act/means. For example, Table 2 illustrates the number of HFs for hate target categories on both datasets from our corpus. Our analysis reveals that on both datasets of our corpus, when Race/Ethnicity was the target of hate, the largest number of HFs were articulated, indicating that there are an extremely large number of ways to frame hate against this Target. On the other extreme, the smallest way of framing hate was obtained when the target was the family status. The distribution of HFs along the Intent/Purpose dimension or along the Means/Act dimension of hate reveals which specific form of hate purpose or hate means is most prolific in framing. This information is available on GitHub[2], revealing a subtle understanding of the many ways hate is framed on social media.

To further allow inspection of HFs, we have implemented an interface that considers, for each of the HFs encoded in MM-HateFrames, a method of browsing the relations that connect it to other HFs. This will allow researchers to find which HFs contradict each other and which HFs specialize a given HF.

5. Using MM-HateFrames

Methodology: To systematically evaluate the impact of our annotated HFs in the MM-HateFrames dataset, we study the task of automatically articulating Hate Frames (HFs) from multimodal memes. Given a meme (image and OCR text) along with its associated target, intent, and means, each method generates a concise HF that abstracts the underlying causal framing of hate expressed in the meme. We benchmark against seven representative methods spanning both non-generative and generative paradigms. The non-generative methods include **HAC-SBERT**, **HAC-CLIP**, and **HAC-LLaMA**,

[2] https://github.com/rak55/MM-HateFrames-Dataset

System	Discovered HFs	Paraphrases	Contradictions	Specializations	Final HFs
HAC-SBERT	-	-	-	-	551
HAC-CLIP	-	-	-	-	609
HAC-LLaMA	-	-	-	-	607
LLaVA (Zero-shot)	6890	9365	2	19	719
LLaVA (Few-shot)	6756	9198	8	25	728
InternVL (Zero-shot)	6076	8689	7	17	701
InternVL (Few-shot)	5993	8598	10	23	693

Table 3: Number of HFs discovered in Phase A; number and type of relations between HFs discovered in Phase B; final number of HFs selected in Phase B.

System	*Z*	*A*	*R*	R_K	F_1	P_A
HAC-SBERT	-	0.39	0.30	0.33	0.34	0.19
HAC-CLIP	-	0.41	0.32	0.34	0.36	0.25
HAC-LLaMA	-	0.44	0.40	0.39	0.42	0.29
LLaVA (Zero-shot)	0.53	0.67	0.62	0.59	0.64	0.33
LLaVA (Few-shot)	0.58	0.70	0.64	0.61	0.67	0.39
InternVL (Zero-shot)	0.62	0.69	0.62	0.60	0.65	0.35
InternVL (Few-shot)	**0.66**	**0.71**	**0.67**	**0.64**	**0.69**	**0.44**

Table 4: Evaluation results of the final sets of HFs.

while the generative multimodal methods comprise **LLaVA** and **InternVL**, each evaluated under zero-shot and few-shot settings. All experiments were conducted on the HatefulMemes dataset (Kiela et al., 2021), which consists of 3300 hateful memes. For clustering-based approaches, each meme M_i is encoded either as: (a) a text-only representation using Sentence-BERT (Reimers and Gurevych, 2019) embeddings of the OCR text (**HAC-SBERT**), or (b) a multimodal representation using the element-wise average of normalized CLIP (Radford et al., 2021) text and image embeddings (**HAC-CLIP**):

$$p_i = \tfrac{1}{2}\left(CLIP_{\text{text}}(t_i) + CLIP_{\text{img}}(I_i)\right)$$

We then apply Hierarchical Agglomerative Clustering (HAC) with Ward's linkage (Jr., 1963) and a variance-gain threshold of 1.1. For each cluster C_j, the OCR text of the meme nearest to the centroid is considered as a HF. As an extension to this method, the OCR texts of the top-k cluster members (k=5) are concatenated and summarized using a zero-shot LLaMA 3.1 (7B) (Touvron et al., 2023) prompt (**HAC-LLaMA**), producing a short, descriptive HF. This non-generative baseline measures thematic grouping ability without explicit reasoning; therefore, it does not produce any associated rationales or discover relations between HFs.

For generative multimodal evaluation, we employ **LLaVA-OneVision-1.5 (8B)** (Liu et al., 2023), a LMM capable of joint image–text reasoning. Using the same system prompt as MM-HateFrames, it integrates the hierarchy of HDE and the definition of a HF to guide articulation. Each input comprises the meme image, its OCR text, and the corresponding target, intent, and means. We consider both zero-shot (**LLaVA Zero-shot**) and few-shot (**LLaVA Few-shot**) settings, the latter employing five diverse multimodal exemplars that include the meme, OCR text, verified HF, and rationale. We further evaluate **InternVL-3.5 (8B)** (Chen et al., 2024), adopting identical zero-shot and few-shot prompting strategies (**InternVL Zero-shot**, **InternVL Few-shot**).

All generated HFs (after paraphrase elimination for generative baselines) were independently evaluated by three trained linguists. Following (Weinzierl and Harabagiu, 2024), we compute the following metrics: All generated HFs (after paraphrase elimination for generative baselines) were independently evaluated by three trained linguists. Following (Weinzierl and Harabagiu, 2024), we compute the following metrics:

Quality of Reasoning (Z). Measures the proportion of sound rationales:

$$Z = \frac{N_S}{N_T} \tag{1}$$

Here, N_S denote the number of HFs judged sound and N_T is the total number of generated HFs. A HF is considered sound if it satisfies the *soundness* criteria defined in Section 3.3.4.

Quality of Articulation (A). Measures the proportion of clear, well-formed HFs:

$$A = \frac{N_C}{N_T} \tag{2}$$

Here, N_C denote the number of HFs judged clear. A HF is considered clear if it satisfies the *clarity* criteria defined in Section 3.3.4.

Recall of Clear and Known HFs.

$$R = \frac{N_C}{N_C + N_F - N_K} \tag{3}$$

$$R_K = \frac{N_K}{N_F} \tag{4}$$

Here, R captures recall of clear HFs, while R_K measures recall of known HFs from the MM-HateFrames reference dataset. $N_F = 1085$ is the number of reference HFs; and N_K is the number of generated HFs deemed known by the annotators.

Combined F_1 Score. Captures the clarity–recall trade-off:

$$F_1 = \frac{2AR}{A + R} \tag{5}$$

Novel HF Precision (P_A). Measures clarity of newly discovered HFs:

$$P_A = \frac{N_C - N_K}{N_T - N_K} \tag{6}$$

Evaluation Results: Table 3 lists the number of HFs discovered in Phase A when using each system, the number of paraphrases and contradictions discovered in Phase B, and the final number of HFs selected in Phase B. Table 4 lists the results of all the evaluation metrics mentioned earlier across all baselines for discovering HFs. Among the non-generative clustering baselines, HAC-LLaMA achieves slightly higher clarity (A = 0.44) and recall (R = 0.40) than HAC-SBERT and HAC-CLIP, indicating that summarizing clustered meme texts with a generative model modestly improves interpretability. However, these clustering approaches still perform poorly overall, reflecting their inability to capture the causal and moral reasoning required for true frame articulation.

In contrast, multimodal LLMs—LLaVA and InternVL—produce markedly clearer and more coherent HFs. Both models benefit from few-shot demonstrations, with InternVL (Few-shot) obtaining the best overall results, suggesting that fine-grained visual grounding and even a few examples for in-context learning can substantially enhance the articulation of HFs.

These findings confirm that explicit multimodal reasoning is crucial for generating sound and novel frames beyond surface-level textual similarity.

6. Conclusion

We introduced the MM-HateFrames Dataset, a large-scale resource capturing how hate is conceptually framed in multimodal social media. Three CoT promtps are used to produce the annotations, which are also verified and edited by human experts. First, HDEs are annotated, which allowed us to combine Retrieval-based In-Context Learning with expert-guided refinement, to discover, articulate and curate HFs paired with their rationales.

Through extensive linguistic evaluation and benchmarking against both clustering-based and large multimodal baselines, we demonstrated that current LMMs, particularly under few-shot settings, are capable of producing coherent and contextually grounded frame articulations but still fall short in reasoning about the underlying causal and moral dimensions of hate.

In future work, we aim to enrich the dataset by linking HFs to underlying human values, enabling the study of how hateful narratives violate or distort shared moral principles. This value-grounded extension will facilitate the development of counter-hate narratives that challenge these framings constructively.

Limitations

First, while the MM-HateFrames dataset provides a valuable foundation for understanding how hate is conceptually framed in multimodal content, we cannot guarantee that it captures the full spectrum of possible HFs. Our analysis is limited to two benchmark datasets—Hateful Memes and MMHS150K—which, although diverse, represent only a subset of online hate expressions. Nevertheless, this serves as a strong starting point for systematically studying multimodal hate framing.

Second, our current framework considers only three types of relations between HFs—paraphrase, specialization, and contradiction. Future work should explore other potential relations to more comprehensively capture the complexity of hate discourse. Third, we restrict our analysis to image–text pairs. Other media modalities common on social platforms—such as audio, video, or animated GIFs—remain unexplored. Expanding to such modalities, as well as to platforms like Reddit or YouTube where posts and threads are longer and context-rich, would enhance the generalizability of our framework. Fourth, although we adopt the hierarchy of Hate Defining Elements (HDEs) proposed by (Korre et al., 2025) as the foundation for our annotation, we do not assume that it is complete or final. During annotation, annotators were instructed to flag cases in which the predefined categories did not fully capture the hate expression in the multimodal post. In practice, we did not observe systematic gaps at the level of the three primary hate dimensions (target, intent, and means). However, some borderline cases required careful interpretation within existing subcategories, indicating that future work could refine or extend the hierarchy

based on empirical findings. We therefore use the HDE hierarchy as a structured framework to ensure consistency and comparability, while acknowledging that it may evolve as new forms of multimodal hate expression are studied. Finally, we do not analyze the temporal dynamics of Hate Frames or how they evolve over time. Temporal modeling could provide valuable insights into the progression and diffusion of hateful narratives across social and cultural contexts.

Ethical Statement

Our work adheres fully to the ACL Code of Ethics[3]. We employ two publicly available datasets— the Hateful Memes and MMHS150K—which contain multimodal social media content expressing various forms of hate directed toward different demographic groups. Given the sensitivity of such material, we implemented strict ethical safeguards to ensure responsible data handling and analysis. The study received approval from the Institutional Review Board (IRB) at University of Texas at Dallas for research involving social media data. To ensure reliability and fairness, we followed a rigorous annotation and validation process, including inter-annotator agreement checks to maintain consistency and quality. While we acknowledge the potential harm associated with engaging with hateful or offensive content, our research aims to advance the understanding of online hate and to support the development of tools that help identify, interpret, and counteract harmful narratives. Ultimately, this work contributes to the ethical advancement of multimodal NLP and the broader goal of fostering safer, more inclusive online spaces.

Bibliographical References

2012. *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge University Press.

Rakshitha Rao Ailneni and Sanda M. Harabagiu. 2025. Automatically discovering how misogyny is framed on social media. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12189–12208, Albuquerque, New Mexico. Association for Computational Linguistics.

Abdullah Albanyan, Ahmed Hassan, and Eduardo Blanco. 2023. Not all counterhate tweets elicit the same replies: A fine-grained analysis. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 71–88, Toronto, Canada. Association for Computational Linguistics.

Ghadi Alyahya and Abeer Aldayel. 2024. Hatred stems from ignorance! distillation of the persuasion modes in countering conversational hate speech. In *International Conference on Web and Social Media*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 104–120, Berlin, Heidelberg. Springer-Verlag.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks.

[3] https://www.aclweb.org/portal/content/acl-code-ethics

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring hate speech detection in multimodal publications.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection.

Joe H. Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Shima Khanehzar, Andrew Turpin, and Gosia Mikolajczak. 2019. Modeling political framing across policy issues and contexts. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66, Sydney, Australia. Australasian Language Technology Association.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198, Albuquerque, New Mexico. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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.

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

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

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*, abs/1907.11692.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. Hatexplain: A benchmark dataset for explainable hate speech detection.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Sharon Meraz and Zizi Papacharissi. 2013. Networked gatekeeping and networked framing on #egypt. *The International Journal of Press/Politics*, 18(2):138–166.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

W. Russell Neuman, Lauren Guggenheim, S. Mo Jang, and Soo Young Bae. 2014. The dynamics of public attention: Agenda-setting theory meets big data. *Journal of Communication*, 64(2):193–214.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

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.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Discovering and articulating frames of communication from social media using chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1617–1631, St. Julian's, Malta. Association for Computational Linguistics.

7. Appendices

A. Annotation Protocol and Annotator Details

A.1. Annotator Background and Roles

The annotation process involved multiple annotators participating across different phases of the pipeline. All annotators were graduate-level students affiliated with our institution and had academic backgrounds in computer science, natural language processing, or computational linguistics.

Annotators demonstrated near-native English proficiency and received training on the definitions of Hate Frames (HFs), Hate Defining Elements (HDEs), and the overall annotation objectives before beginning the task. Three annotators participated in Phase A (HDE verification) and Phase C (relation verification), while five annotators participated in Phase D, which focused on evaluating clarity and soundness of the articulated frames. In addition to these annotators, a senior expert annotator supervised the annotation process and performed adjudication and revisions whenever disagreements or low-quality outputs were identified.

A.2. Annotation Training and Guidelines

Before annotation began, annotators were provided with written guidelines describing the definitions of HFs and HDEs, the hierarchy of HDEs used in this work, and several example multimodal posts illustrating correct and incorrect annotations. Training emphasized that model-generated outputs should be treated as suggestions rather than ground truth. Annotators were instructed to independently evaluate each prediction based on the multimodal content and to revise or reject outputs whenever necessary. The goal of the training process was to ensure consistent interpretation of framing, causal reasoning, and hate dimensions across annotators.

A.3. Annotation Interface and Workflow

Annotation was conducted through an internal annotation interface designed to support multimodal inspection and editing. For each multimodal post, annotators were presented with the meme image, the associated OCR or textual content, the model-generated predictions, and the accompanying rationale. The interface allowed annotators to directly edit labels, modify rationales, remove incorrect predictions, or add missing annotations. Annotators were required to actively review each instance and could not simply accept predictions without inspection. This workflow was designed to minimize anchoring bias and to ensure that final annotations reflected human judgment rather than automatic model outputs.

A.4. Phase A: Human Verification of Hate Defining Elements

In Phase A, an LMM generated initial predictions for the target, intent/purpose, and act/means dimensions of hate, together with rationales. Annotators reviewed these predictions by verifying whether each HDE was supported by the multimodal content and whether the accompanying rationale correctly justified the selection. Annotators were instructed to remove unsupported HDEs, correct inaccurate assignments, and add missing elements when required. An HDE annotation was retained only when at least two of the three annotators agreed on its validity. Inter-annotator agreement measured using Krippendorff's α was 0.82, indicating strong agreement.

A.5. Phase B: Human Verification of Hate Frame Articulation

Phase B produced candidate Hate Frames and rationales using retrieval-based in-context prompting. Annotators reviewed the generated frames to ensure that each articulation represented a generalized framing rather than a literal restatement of the meme content. They were instructed to verify that the articulated frame captured a causal or interpretive perspective and that the rationale was grounded in the identified target, intent, and means dimensions. Annotators were explicitly instructed not to accept outputs solely because they were model-generated and to revise articulations whenever necessary.

A.6. Phase C: Verification of Relations Between Hate Frames

In Phase C, candidate relations between pairs of Hate Frames were predicted automatically and then verified by human annotators. Annotators evaluated whether each relation corresponded to paraphrase, contradiction, specialization, or no relation by considering both the frame articulation and the provided rationale. Definitions of each relation type, together with illustrative examples of paraphrases, specializations, and contradictions between frames, were provided in the annotation guidelines to ensure consistent interpretation. A relation was finalized only when a majority agreement among annotators was achieved. Inter-annotator agreement for this phase, measured using Krippendorff's α, reached 0.85, indicating strong consistency.

A.7. Phase D: Evaluation of Clarity and Soundness

In Phase D, five annotators evaluated each Hate Frame using two binary criteria: clarity and soundness. Clarity measured whether the articulation represented a genuine frame rather than a factual reformulation, while soundness assessed whether the rationale correctly connected the meme content to the identified hate dimensions. Frames flagged by at least two annotators as failing either criterion were reviewed and revised by an expert annotator. Inter-annotator agreement was 0.75 for clarity and 0.73 for soundness, reflecting strong agreement levels for subjective evaluation tasks.

A.8. Disagreement Resolution and Expert Adjudication

Disagreements were resolved through majority voting whenever possible. Cases were identified as low quality based on clarity and soundness in Phase D were escalated to a senior expert annotator for revision. This procedure ensured consistency across annotations while maintaining human oversight over model-generated suggestions throughout the pipeline.

A.9. Annotation Disagreement Analysis

To promote transparency on annotation uncertainty, we retain and release annotator-level disagreement information across annotation phases. Disagreements occurred primarily in two stages: Phase A (HDE verification) and Phase D (clarity and soundness evaluation of Hate Frames).

Figure 7 illustrates a disagreement case from Phase A during HDE verification. All three annotators agreed that the primary hate target falls under Race/Ethnicity (Demographics), and the majority also agreed on Stereotyping as an intent and Provocation under verbal/written expressions as the means of delivery. However, disagreement emerged when evaluating additional intent subcategories within the Social/Cultural Control dimension.

Specifically, one annotator marked Marginalization as applicable, while the other two rejected it. Conversely, Exclusion was accepted by one annotator but rejected by the remaining two. These disagreements reflect interpretive differences at the subcategory level rather than disagreement about the overall hateful framing. While annotators consistently identified the meme as targeting race/ethnicity through stereotypical provocation, they differed in whether the framing implied social marginalization or explicit exclusion. Following the majority-agreement protocol, only the categories endorsed by at least two annotators were retained in the final annotation.

Figure 8 illustrates a disagreement case from Phase D during evaluation of the clarity criterion. The initially articulated HF stated that *Muslim women are unfit for public life and should be ridiculed for appearing in public roles*. While two annotators accepted this articulation as clear, three annotators rejected it on the grounds that it did not accurately reflect the interpretation intended in the meme.

The meme mocks a woman wearing a religious head covering by comparing it to a diaper and describing her as intellectually deficient. However, it does not explicitly frame Muslim women as unfit for public life more broadly; rather, it ridicules women who wear religious head coverings by portraying them as mentally inferior. The original frame therefore overgeneralized the meme's message, attributing a broader exclusionary narrative about public participation that was not directly articulated in the content.

Because clarity requires that an HF provide a correct and contextually grounded abstraction of the hateful frame of the meme, three annotators judged the initial articulation to be an inaccurate interpretation rather than a faithful frame. The case was subsequently reviewed by an expert annotator, who revised the frame to *Women who wear religious head coverings are mentally deficient*, align-

Targets: Race/Ethnicity (Demographics)
Annotator 1: **Accept**
Annotator 2: **Accept**
Annotator 3: **Accept**
Final targets: Race/Ethnicity (Demographics)
Intents: Stereotyping (Social/Cultural Control)
Annotator 1: **Accept**
Annotator 2: **Accept**
Annotator 3: **Accept**
Intents: Marginalization (Social/Cultural Control)
Annotator 1: **Accept**
Annotator 2: **Reject**
Annotator 3: **Reject**
Intents: Exclusion (Social/Cultural Control)
Annotator 1: **Reject**
Annotator 2: **Accept**
Annotator 3: **Reject**
Final Intents: Race/Ethnicity (Demographics)
Actions: Provocation (Verbal/Written Expressions)
Annotator 1: **Accept**
Annotator 2: **Accept**
Annotator 3: **Accept**
Final Actions: Provocation (Verbal/Written Expressions)

Figure 7: Example of annotator disagreement in Phase A (HDE verification)

Hate Frame: Muslim women are unfit for public life and should be ridiculed for appearing in public roles.
Criteria: Clarity
Annotator 1: Accept **Annotator 2:** Accept
Annotator 3: Reject
Annotator 4: Reject **Annotator 4:** Reject
Expert Revision: Women who wear religious head coverings are mentally deficient.

Figure 8: Phase D disagreement example for the clarity criterion

ing the articulation more closely with the meme's expressed ridicule.