Discerning Culturally-Aware Hate Framings for Hate Speech Identification in Memes

Anonymous ACL submission

Abstract

Warning: *this paper contains content that may be offensive or upsetting.*
The ability of Large Visual Language Models (LVLMs) to identify hate speech in memes across several cultures requires (1) access to cultural knowledge that is not readily available to them; and (2) the ability to reason with this knowledge. This paper shows how culturally-aware *Hate Frames* can be acquired to inform hate speech identification across various cultures. The paper also shows that counterfactual reasoning is ideal for identifying hate speech and generating explanations. Promising results show not only improved hate speech identification in each culture, but also reduction of the performance gap of hate speech identification between different cultures. [1]

1 Introduction

Social media platforms have empowered individuals by amplifying their voices and facilitating freedom of expression. However, these platforms have also 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 individuals or groups (Herz and Molnár, 2012). Therefore, the identification of hate speech is essential for safeguarding those targeted by such social media communications.

However, the identification of hate speech in social media remains a complex challenge because: (1) hate speech identification methods need to operate not only on texts (Waseem and Hovy, 2016; Davidson et al., 2017), but also on multimodal content (Gomez et al., 2019; Hossain et al., 2022; Bhandari et al., 2023; Kiela et al., 2020); and (2) hate speech relies heavily on cultural influences, leading to varying interpretations. Recently, (Bui et al., 2025) has investigated how well current Large Vision-Language Models (LVLMs) are capable to address these challenges for identifying hate speech across languages and cultures, introducing the $Multi^3Hate$, a parallel dataset of memes sampled across 5 cultures and languages. Figure 1 illustrates an example from this dataset. $Multi^3Hate$ enabled experiments prompting LVLMS to identify hate speech, (Bui et al., 2025) concluding that culture-motivated limitations exist, especially in some non-English, non-Western cultures. This

Figure 1: Example of Hate Frame discerned for a Meme interpreted in the German culture.

[1] Code:https://anonymous.4open.science/r/Cultural-Hate-Frames-2EB0/

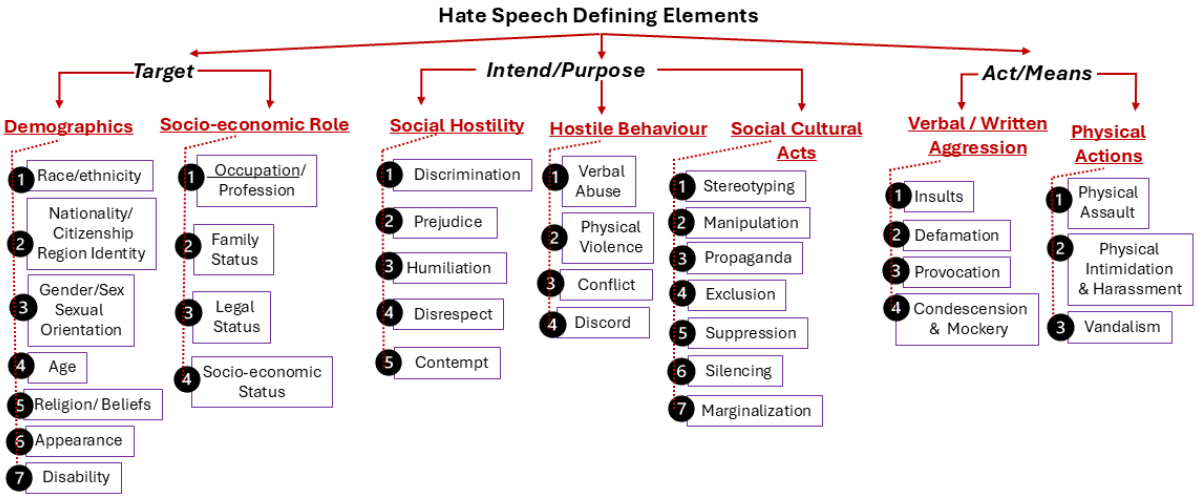


Figure 2: The Hierarchy of Hate Defining Elements, cf. (Korre et al., 2025).

prompted us to consider the following research question:

Q1: What kind of cultural knowledge should be available to LVLMs to enable them to accurately identify hate speech across various cultures?

To answer this question, we consider **Hate Frames**, which encode (a) cultural knowledge, which includes (i) knowledge about historical events; (ii) knowledge of the social norms and cultural values; and (iii) knowledge concerning minority groups in a culture; as well as (b) knowledge about how hate is defined. For example, Figure 1 illustrates a meme from the $Multi^3Hate$ dataset along with the Hate Frame discerned for its interpretation in German culture, which is the only culture in which the meme is interpreted as hateful.

While encoding cultural knowledge in Hate Frames is intuitive, encoding knowledge about how hate is defined is less so, but in this paper we show that it is equally important. For example, the Hate Frame illustrated in Figure 1 is informed by the defining elements of hate available from the Taxonomy of Hate Defining Elements illustrated in Figure 2. The Taxonomy has three dimensions: (1) the *hate target* of the meme; (2) the *intend or purpose* of the meme; and (3) the *means* through which hate is expressed. In Figure 1 the Hate Frame is populated for each dimension of the taxonomy with Hate Defining Elements (HDEs), e.g., Race/Ethnicity; as well as values inferred from the meme for these HDEs, e.g., North Africans. But, most importantly, Figure 1 shows how provocation, an HDE for means of hate, is inferred based on cultural knowledge, which is also explicit in the Hate Frame. Moreover, the Hate Frame indicates that the meme from Figure 1 is inferred as hateful in German culture, providing an explanation for the inference informed by the knowledge encoded in the Hate Frame. This brings us to the second research question that we considered:

Q2: What form of inference should LVLMs use to (a) identify hate speech and (b) explain it?

In this paper, we show that counterfactual reasoning, which provides arguments in favor of identifying hate speech and contrasts them with arguments that dismiss hate speech identification, is an ideal form of reasoning available to LVLMs.

Through comprehensive experiments, we show in this paper that culturally-aware Hate Frames informing counterfactual reasoning, not only provide valid answers to our two research questions, but also generate improved performance for hate speech identification on the $Multi^3Hate$ dataset. The results are superior in each culture and moreover, reduce the performance gap in hate speech identification between different cultures.

2 Related Work

Hate Speech identification is not a new research problem (Waseem and Hovy, 2016; Davidson et al., 2017; Fortuna and Nunes, 2018; Gomez et al., 2019; Kiela et al., 2020), but only recently has cultural knowledge been considered to have an important impact on the quality of hate speech detection. For example, this became evident when (Lee et al., 2024) considered several English-speaking cultures, aiming to detect hate speech in posts from YouTube and Reddit. More recently, (Bui et al.,

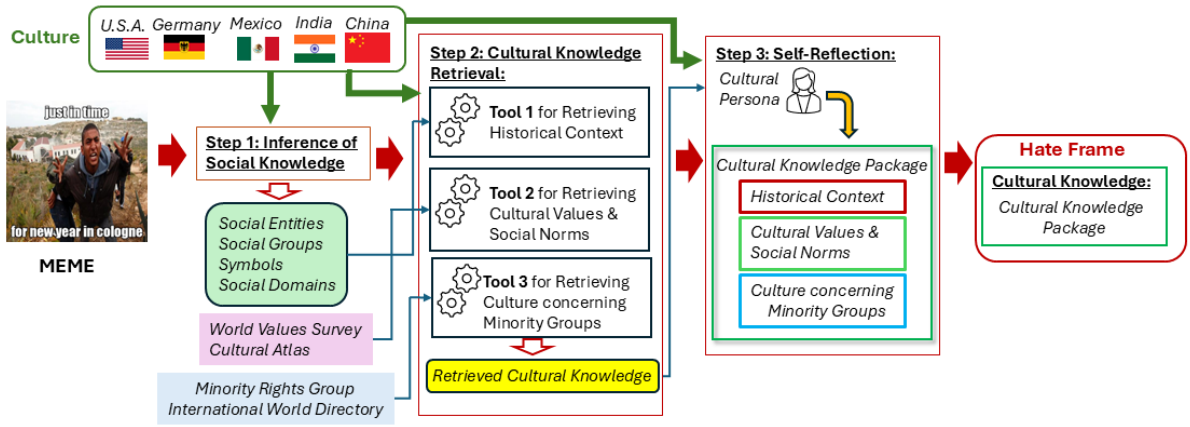


Figure 3: The Culture Knowledge Acquisition Agent (CKA-Agent).

2025) extended the study of the impact of cultural knowledge on hate speech identification in memes by considering several non-English speaking cultures, concluding that there are significant performance gaps when identifying hate speech in various cultures.

In this paper, we use the same dataset introduced in (Bui et al., 2025), discerning Hate Frames, in which several forms of knowledge are acquired by LVLMs, enabling them to perform counterfactual reasoning. Our method, informed by culturally-aware Hate Frames enables the inference of cultural nuances accounting for the identification of hateful memes across various cultures.

3 A Taxonomy of Hate Defining Elements

To define hate, Korre et al. (2025) processed 493 hate speech definitions drawn from five distinct domains: (1) hate speech laws, (2) Wikipedia, (3) dictionaries, (4) NLP research papers, and (5) online texts —to generate a taxonomy of **Hate Defining Elements (HDEs)**, illustrated in Figure 2. The HDE taxonomy was generated through Semantic Componential Analysis (SCA). SCA is a linguistic method that decomposes the meanings of words or phrases into their constituent features, identifying potential cross-domain and cross-cultural differences. The HDE taxonomy is organized around three main elements: (1) the hate target, (2) the hate intent/purpose, and (3) the hate act/means. Each of these three dimensions is specialized along two additional levels of the hierarchy, producing 34 distinct HDEs.

4 The Method

The System for identifying Cultural Hate and inferring Explanations informed by Cultural Knowledge Retrieval and Counterfactual Reasoning **(CHER)** generates for each meme M_i and each culture C_j a separate *Hate Frame*$_i^j$ which indicates if M_i is hateful in culture C_j, along with an explanation for the hate speech identification. *Hate Frame*$_i^j$ provides also evidence of (a) the cultural knowledge that supports the hate speech identification; and (b) the HDEs inferred from M_i in C_j. We note that even when hate speech is not identified, *Hate Frame*$_i^j$ encodes an explanation for this inference.

The **CHER** system discerns a *Hate Frame*$_i^j$ for each meme M_i in each culture C_j from the $Multi^3Hate$ dataset in three phases:
Phase 1: cultural knowledge acquisition;
Phase 2: identification of the Hate Defining Elements (HDEs), taking into account the acquired cultural knowledge. The identified HDEs and their values discerned from the meme are encoded in *Hate Frame*$_i^j$, along with the cultural knowledge;
Phase 3: performs *counterfactual reasoning* of the LVLM to infer if meme M_i is interpreted as hateful in culture C_j, generating a culturally-aware explanation of the inference. The inference and explanation are also encoded in *Hate Frame*$_i^j$.

4.1 Cultural Knowledge Acquisition

For each culture C_j, the **CHER** system acquires a **Cultural Knowledge Package (CKP)** through the operation of Culture Knowledge Acquisition *(CKA)-Agent*, illustrated in Figure 3. The CKA-Agent is implemented as a ReAct-style agent (Yao et al., 2023), which combines, in an interleaved

manner, the capabilities of VLMs to reason (e.g. through Chain-of-Thought (CoT) prompting) and act (e.g. retrieving information from Wikipedia). This is because we believe that the acquisition of diverse cultural knowledge that is useful for identifying hate speech should benefit from access to knowledge-acquisition tools.

The CKA-Agent receives a main prompt containing: (1) instructions for inferring social knowledge from the meme; (2) descriptions of each tool available to it along with their expected inputs and outputs; (3) the description of the persona for each culture C_j; and (4) instructions for performing self-reflection over the retrieved cultural knowledge, in order to assemble the CKP. The prompt is detailed in Appendix A. As shown in Figure 3, the CKA-Agent performs three different steps, as instructed in the prompt it receives:

Step 1: infers *social knowledge* by instructing the VLM to analyze the meme M_i visual content as well as its caption to infer (1) *social entities* (including people, locations, dates), (2) visible *social groups* (e.g., groups identifiable through age, skin tone, gender presentation, race, etc), (3) symbols, and (4) the *social domain* reflected in M_i. A social domain refers to a specific context such as family, school, or workplace that influences how individuals interact, behave, and form relationships. Appendix B shows examples of social knowledge inferred for various memes across several cultures.

Step 2: relies on several tools that the CKA-Agent can use *in parallel* to retrieve cultural knowledge:

Tool 1 for Retrieving Historical Context uses the social knowledge inferred in Step 1. It prompts the LLM to generate culture-specific Wikipedia queries, which are then sent to the Wikipedia API. The API returns five of the most relevant article summaries for each query. Appendix B provides an example comparing the Historical Context retrieved for the same meme across two cultures, together with the generated Wikipedia queries and the retrieved articles that pass the self-reflection criteria. Appendix A additionally provides the complete prompt used by Tool 1 for query generation.

Tool 2 for Retrieving Cultural Values and Social Norms performs semantic retrieval over (1) the World Values Survey (WVS) (Haerpfer et al., 2020) for cross-cultural attitudes and values, and (2) the Cultural Atlas (Mosaica, 2024) for socially documented practices and behavioral expectations. The queries used in the semantic retrieval are formed from the social domain inferred at Step 1, The semantic retrieval is made possible by encoding (i) each element of the social domain and (ii) the questions from WVS (or the section names from the Cultural Atlas) using the Sentence-BERT model (Reimers and Gurevych, 2019) and considering their cosine similarity. Only knowledge whose similarity to the inferred social domain exceeds a threshold τ is retrieved. Appendix B provides an example comparing the Cultural Values and Social Norms retrieved for the same meme across two cultures.

Tool 3 for Retrieving Knowledge concerning Minority Groups in a culture C_j relies on a publicly accessible resource that maintains country-specific profiles documenting the demographic composition, legal status, and historical circumstances of minority and indigenous communities worldwide [2]. For each of culture C_j, the tool uses three fields from the MRG country profile: (1) the main languages spoken; (2) the main religions practiced; and (3) the enumerated minority and indigenous population groups. Since minority group classifications, dominant religions, and main languages are stable demographic facts that do not vary by meme content, this information is compiled once per culture C_j, and does not need to be further evaluated. Appendix B provides examples of the knowledge concerning minority groups for each culture C_j.

Step 3: implements the *self-reflection* of the CKA-Agent, evaluating the cultural knowledge retrieved in Step 2. The self reflection uses a Cultural Persona CP_j, considered for each culture C_j. Each CP_j is constructed by providing the LVLM with cultural knowledge available from the Cultural Atlas *core concepts* and *Do's and Don'ts* sections for each culture C_j, as in (Rao et al., 2025; Kamruzzaman et al., 2026). The self-reflection step evaluates the quality of the retrieved cultural knowledge using two criteria: (1) *cultural consciousness*, determining whether the knowledge is likely to be recognized within the mainstream public consciousness of the culture C_j, and (2) *meme relevance*, determining whether the knowledge directly explains the meme's visual or textual content rather than being only tangentially related. Any retrieved cultural knowledge that does not meet both criteria is filtered out, allowing the CKA-Agent to assemble the CKP and encoded in the Hate Frame.

[2] Minority Rights Group International (MRG) World Directory https://minorityrights.org/

4.2 Identification of Hate Defining Elements

The Cultural Knowledge Package (CKP) informs the identification for each meme M_i in each culture C_j of the Hate Defining Elements (HDEs) pertaining to (1) the Hate Target; (2) Hate Intend/Purpose; and (3) Hate Acts/Means, encoded in the Taxonomy of Hate Defining Elements.

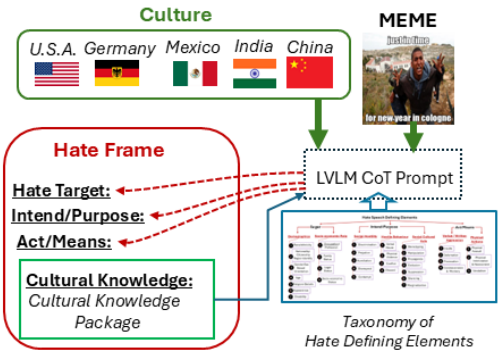


Figure 4: The Inference of HDEs.

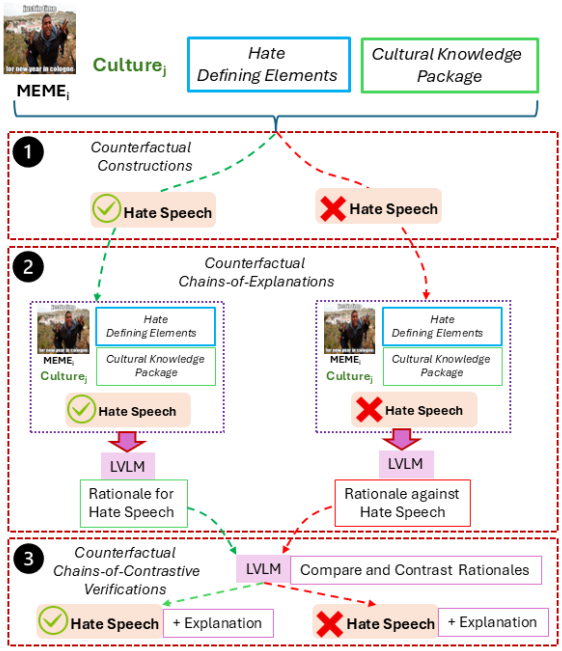


Figure 5: Counterfactual Prompting for Hate Speech Identification.

The identification of these HDEs results from Chain-of-Thought (CoT) prompting (Wei et al., 2022) of the LVLM, as shown in Figure 4. The prompt, detailed in Appendix A, includes the meme M_i, the complete HDE taxonomy described in Section 3, and the CKP assembled by the CKA-Agent for culture C_j. The LVLM is instructed to propose only those HDEs supported by evidence in the CKP and to generate a rationale explaining how each identified HDE is grounded in the CKP.

Since the identification of HDEs is performed independently for each culture C_j, the same meme M_i may lead to the identification of different HDEs in different cultures, as the CKP assembled for different cultures may vary. Additionally, multiple HDEs may be identified for the same meme M_i when the meme simultaneously refers to multiple Hate Targets, Intents, or Acts. Nevertheless, hate speech shall not be identified for any meme M_i in a culture C_j unless a Target HDE, an Intend/Purpose HDE and an Act/Means HDE have been inferred and used to populate the corresponding $Hate\ Frame_i^j$.

4.3 Counterfactual Reasoning for Hate Speech Identification

After the HDEs are identified for a meme M_i interpreted in a culture C_j, the decision of whether the meme contains hate speech is made through counterfactual reasoning, as shown in Figure 5. Rather than directly populating the $Hate\ Frame_i^j$, each identified HDE is independently evaluated through competing hateful and non-hateful interpretations grounded in the CKP.

□ Step 1 involves constructing a tree of counterfactuals for each identified HDE: one branch assumes that the HDE supports a hateful interpretation of M_i, while the other assumes that the same HDE does not support hate speech. Both branches are always constructed regardless of the content of M_i, since the validity of each interpretation must be evaluated in later stages. This allows **CHER** to explicitly contrast competing culturally grounded interpretations of the same meme.

□ Step 2 prompts the LVLM to explain why, given a particular HDE and the relevant cultural knowledge from the CKP, the meme should be interpreted as hateful or non-hateful. The prompting approach seeks to generate post hoc rationalizations for both hypotheses associated with each HDE. In Appendix A, we detail this Chain-of-Explanation (CoE) prompting strategy, inspired by (Weinzierl and Harabagiu, 2024). Through CoE prompting, the LVLM generates counterfactual explanations e_+^{HS} and e_-^{HS} for each HDE, where e_+^{HS} argues why the meme should be interpreted as hateful through the proposed HDE, while e_-^{HS} argues why the same HDE should not be interpreted as hateful.

□ Step 3 takes inspiration from Chain-of-Verification (CoVe) prompting proposed by (Dhu-

Model	Method	US	DE	MX	IN	CN
GPT-5	Vanilla	78.3±0.5	75.2±0.7	73.1±0.8	68.4±0.9	71.2±0.7
	CHER	**82.3±0.4**	**81.5±0.6**	**80.4±0.7**	**78.9±0.8**	**79.5±0.7**
Gemini 2.5 Pro	Vanilla	73.4±0.6	72.1±0.8	71.3±0.8	67.9±0.9	68.8±0.8
	CHER	**77.8±0.3**	**80.2±0.6**	**78.9±0.7**	**77.6±0.8**	**75.4±0.5**
Qwen2.5-VL 72B	Vanilla	73.2±0.7	65.4±0.9	67.8±0.9	61.3±1.0	61.8±1.0
	CHER	**77.4±0.6**	**71.8±0.3**	**72.9±0.8**	**70.6±0.9**	**68.4±0.9**
LLaVA OV 73B	Vanilla	70.3±0.8	63.2±1.0	65.1±1.0	59.4±1.1	60.2±1.0
	CHER	**74.1±0.7**	**69.3±0.9**	**70.4±0.5**	**66.7±0.5**	**68.1±0.2**
InternVL3 78B	Vanilla	67.8±0.9	62.4±1.0	63.5±1.0	58.2±1.1	57.6±1.1
	CHER	**71.2±0.8**	**68.1±0.3**	**68.4±0.9**	**66.8±0.8**	**64.3±0.2**

Table 1: Macro-F1 scores for five LVLMs under vanilla prompting and CHER across five cultures. Values are mean ± standard deviation across three independent LVLM runs.

liawala et al., 2024). CoVe prompting enables LVLMs to deliberate over generated responses through introspection. We employ the counterfactual explanations e_{+}^{HS} and e_{-}^{HS} produced for each HDE in Step 2 within a Counterfactual Chain-of-Contrastive Verification (CCoCV) prompting, as in (Weinzierl and Harabagiu, 2024). The LVLM compares both the counterfactual rationales and compares them with the CKP to contrast each hypothesis for hate speech inferred or not inferred from M_i.

The same CCoCV prompt additionally generates the final *Hate* $Frame_i^j$, including: (1) the final hate decision, assigned as hateful if hate speech is inferred for at least one identified HDE; (2) the subset of HDEs for which hate speech is inferred; (3) the corresponding relevant cultural knowledge selected from the CKP; (4) the explanation generated for the decision; and (5) the alternative non-hateful interpretation when hate speech is not inferred. The complete CCoCV prompt is detailed in Appendix A.

5 The Dataset

We use $Multi^3$*Hate* (Bui et al., 2025), a multimodal, multilingual, and multicultural parallel hate speech benchmark annotated by native annotators across five cultures: the United States (US), Germany (DE), Mexico (MX), India (IN), and China (CN). The dataset contains 1,500 memes spanning 15 topics across five sociopolitical categories: religion, political issues, nationality, ethnicity, and LGBTQ+ discourse. Each culture contains 300 annotated memes, with 154 hateful memes for the US, 179 for Germany, 167 for Mexico, 180 for India, and 190 for China. Each meme was annotated by at least five native annotators per culture using majority voting. Inter-annotator agreement, measured using Krippendorff's α, ranges from 0.3895 to 0.4686 across cultures, comparable to prior hate speech annotation studies (Ross et al., 2016; Lee et al., 2024).

6 Experimental Results

Implementation Details We use GPT-5[3](Singh et al., 2026) as the backbone LVLM for **CHER**. In accordance with the GPT-5 model family, we use a fixed temperature of 1.0. For each retrieved Wikipedia article, the first five lines of the article summary are retained as contextual knowledge. For Tool 2, the semantic similarity threshold τ is set to 0.6 to maximize cultural knowledge coverage.

Quantitative Results **CHER** identifies as hateful 153 memes for US culture; 206 memes for DE culture; 186 memes for MX culture; 188 memes for IN culture and 200 memes for CN culture. Across all cultures, Phase 2 of **CHER** generates 2,509 candidate HDEs, of which 1,902 are encoded in the 1500 Hate Frames discerned by **CHER**. Detailed error analysis for **CHER** is provided in Appendix F.

Table 1 reports the Macro-F1 performance of **CHER** when identifying hate speech across five cultures. In addition to GPT-5, we evaluate **CHER** when using several additional LVLMs, including Gemini 2.5 Pro[4], Qwen2.5-VL 72B (Bai et al., 2025), LLaVA OneVision 73B (Li et al., 2024), and InternVL3 78B (Zhu et al., 2025). We compare the performance of **CHER** against a *Vanilla* baseline that follows the prompt used in (Bui et al., 2025). Unlike **CHER**, which operates solely on English-language memes, the *Vanilla* baseline receives both the meme and its caption in the original language corresponding to the target culture. The

[3]API Version: gpt-5-2025-08-07
[4]API Version: gemini-2.5-pro

detailed prompt is provided in Appendix C. To account for the stochasticity of LLM generation, we report the mean and standard deviation of Macro-F1 over three independent runs with a sampling temperature of 1.0.

Table 1 shows that **CHER** consistently outperforms the vanilla baseline across all five models and all five cultures, with particularly pronounced gains for non-US cultures. For GPT-5, the largest absolute improvement occurs for IN (+10.5 F1 points), followed by DE (+6.3) and CN (+8.3). These results indicate that **CHER** not only improves the overall hate speech identification, but also substantially narrows the cross-cultural performance gap. Among the five LVLMs evaluated, GPT-5 achieves the strongest overall performance.

In addition, we have evaluated the quality of hate speech identification against several baselines:

◇ *(Baseline 1) Zero-Shot Prompting (ZS-Prompt)*: GPT-5 is prompted with meme M_i in ulture C_j using the same prompt used in (Bui et al., 2025). This baseline only identifies if the meme is hateful or not. It does not generate any Hate Frame.

◇ *(Baseline 2) Zero-Shot Chain-of-thought Prompting (ZS-CoT)*: GPT-5 is prompted using Chain-of-Thought(CoT), in a zero-shot manner with the meme M_i in a culture C_j, by extending the prompt used in (Bui et al., 2025) with step-by-step reasoning instructions that ask the model to identify relevant cultural knowledge, HDEs and their values, and generate a hate explanation before decising if M_i is hateful in C_j.

◇ *(Baseline 3) Zero-Shot Chain-of-thought Prompting with Cultural Persona (ZS-CoT + CP)*: Similar to ZS-CoT, but the prompt is augmented with a *cultural persona* constructed from the *Core Concepts* and *Do's and Don'ts* provided by Cultural Atlas for culture C_j.

◇ *(Baseline 4) Few-Shot Chain-of-thought Prompting (FS-CoT)*: Similar to ZS-CoT, but augments the prompt with three in-context demonstration memes M_i^x hateful in culture C_j, provinding also hate explanations, HDEs and their values, as well as cultural knowledge.

◇ *(Baseline 5) Few-Shot Chain-of-thought Prompting with Cultural Persona (FS-CoT + CP)* Similar to FS-CoT, but additionally augments the prompt with a *cultural persona*, like in ZS-CoT + CP.

◇ *(Baseline 6) Anthropological Prompting (AP-Prompt)*: For each culture C_j, we employ anthropological prompting (AlKhamissi et al., 2024) to prompt GPT-5 with the meme M_i. The anthropological framework encourages the model to reason through emic perspectives, cultural relativism, and socioeconomic context. We extend the original prompt with additional instructions to identify HDEs and determine whether the meme is hateful. To capture diverse cultural viewpoints, we construct demographic personas by sampling from a predefined set of demographic profiles; the full list is provided in Appendix D. The final hate speech decision is obtained through majority voting across three persona-based runs.

The prompts for all baselines are provided in Appendix C. Table 2 shows that **CHER** consistently outperforms all baselines, demonstrating the effectiveness of identifying hate speech through counterfactual reasoning informed by Hate Frames.

Method	US	DE	MX	IN	CN
CHER	**82.3**	**81.5**	**80.4**	**78.9**	**79.5**
ZS-Prompt	78.3	75.2	73.1	68.4	71.2
ZS-CoT	79.2	76.4	74.5	69.8	72.8
ZS-CoT + CP	79.8	77.2	75.8	72.3	74.2
FS-CoT	80.4	78.2	77.0	73.7	75.1
FS-CoT + CP	80.6	78.8	77.3	74.2	76.1
AP-Prompt	80.5	77.1	76.1	72.8	74.7

Table 2: Macro-F1 comparison between CHER and multiple prompting-based baselines for hate speech identification across five cultures.

Qualitative Results To conduct human evaluation of *Hate* $Frame_i^j$ for a meme M_i interpreted in a culture C_j, we randomly sampled 100 memes identified as hateful by **CHER** in each culture C_j, obtaining a set S_E of 500 *Hate* $Frame_i^j$. For each For each *Hate* $Frame_i^j$ from S_E, three native annotators who have resided in the country of C_j for a minimum of ten years have independently evaluated the *Hate* $Frame_i^j$ produced by **CHER** and each baseline across five dimensions:

Quality of the discovered HDEs: Annotators evaluate whether the HDEs for Hate Target, Hate Intend and Hate Acts and their values are correct independently using (Yes/No) judgments. The results are listed in Table 3, indicating that the hate targets were better identified than the other HDEs. However, all HDEs are discovered with accuracy over 75%. Appendix E details the results of HDE identification for each culture.

Quality of Acquired Cultural Knowledge Annotators assess the cultural knowledge along two dimen-

	HDE Quality (Accuracy %)			Cultural Knowledge (Mean 1–3)		Hate Explanation	
Method	**Target (%)**	**Intent (%)**	**Act (%)**	**Factual Accuracy**	**Meme Relevance**	**Plausibility (1–3)**	**Coherence (%)**
CHER	**81.0**	**75.2**	**77.1**	**2.78**	**2.72**	**2.74**	**81.3**
ZS-CoT	66.8	56.4	61.4	1.93	1.86	2.04	65.9
ZS-CoT+CP	69.6	59.3	64.7	2.07	2.00	2.14	69.0
FS-CoT	71.7	61.1	66.8	2.16	2.09	2.23	71.7
FS-CoT+CP	73.6	64.4	68.4	2.35	2.20	2.28	74.4
AP-Prompt	72.5	63.1	67.4	2.22	2.15	2.30	73.1

Table 3: Human evaluation results macro-averaged across five cultures. Accuracy (%) is reported for Target, Intent, Act, and coherence, while mean Likert scores (1–3) are reported for factual accuracy, meme relevance, and plausibility.

sions: (1) factual accuracy, evaluated on a 3-point Likert scale (*Accurate*, *Partially Accurate*, *Inaccurate*), measuring whether the retrieved knowledge correctly reflects real historical, social, or political information from culture C_j; and (2) relevance to the meme, evaluated using (*Directly Relevant*, *Tangentially Relevant*, *Not Relevant*), measuring whether the retrieved knowledge meaningfully supports the hateful interpretation of the meme. The results are listed in Table 3, indicating that **CHER** acquired cultural knowledge that has high factual accuracy and is deemed to be relevant to memes.

Quality of the Hate Explanation Annotators evaluate whether the generated hate explanation *plausibly* and *coherently* explains why the meme is hateful or non-hateful within C_j. Plausibility is evaluated using a 3-point Likert scale (*Plausible*, *Somewhat Plausible*, *Implausible*), while coherence is evaluated using binary judgments (*Coherent / Incoherent*). The results, listed in Table 3, indicate that **CHER** generates the most plausible explanations, which are also highly coherent.

For all qualitative evaluations, the Inter-Annotator Agreement (IAA) was measured using Krippendorff's α (Krippendorff, 2011), with results ranging from 0.61 to 0.79 across all evaluation dimensions and cultures, indicating substantial agreement throughout. Detailed annotation guidelines and per-culture IAA results are provided in Appendix G.

Ablation study Table 4 presents the ablation results for **CHER** across five cultures. Among all components, removing the cultural Knowledge Package (CKP) causes the largest drop of F1-score across all cultures, particularly for India and China, highlighting the importance of culture-specific knowledge. Within CKP, removing Tool 1 (Historical Context) results in the most significant degradation among all tools, while removing Tool 3 has the smallest impact. It is to be noted that when ignoring the HDEs, significant reduction of the F1-scores can also be observed, indicating the role of encoding HDEs in the Hate Frames. The counterfactual reasoning and self-reflection modules also provide substantial impact in hate speech identification by **CHER** across all five cultures. Details of the error analysis are provided in Appendix F.

Method	**US**	**DE**	**MX**	**IN**	**CN**
CHER	**82.3**	**81.5**	**80.4**	**78.9**	**79.5**
w/o Tool 1	81.1	78.7	78.9	74.6	77.3
w/o Tool 2 (Cultural Atlas)	81.1	80.2	78.8	77.6	78.5
w/o Tool 2 (WVS)	80.9	80.3	79.0	77.9	78.3
w/o Tool 3	80.0	80.1	79.7	77.8	79.1
w/o self-reflection	81.5	79.3	78.0	76.3	78.1
w/o HDE identification	81.6	79.4	78.1	75.5	77.4
w/o counterfactual reasoning	81.3	78.9	77.5	75.1	77.1
w/o CKP	79.3	75.9	74.2	69.9	71.7

Table 4: Macro-F1 ablation study evaluating the contribution of retrieval tools, self-reflection, HDE identification, and counterfactual reasoning in **CHER**.

7 Conclusion

This paper presents a method for acquiring culturally-aware Hate Frames that encode knowledge that is not readily available to Large Visual Language Models (LVLMs), enabling them to identify successfully hate speech in memes across several cultures. Hate speech identification results from counterfactual reasoning of LVLMs, which generate plausible and coherent explanations. Experiments performed on the $Multi^3Hate$ dataset indicate improved hate speech identification across all cultures, when compared with the results reported in (Bui et al., 2025). The results also indicate a significant reduction of the gap in performance across cultures.

8 Limitations

While CHER demonstrates strong performance in culturally grounded hate speech interpretation, several limitations remain.

Dataset Scale and Coverage. Our experiments rely on $Multi^3$*Hate*, which is currently the only dataset providing parallel multilingual and multicultural hate speech annotations across multiple cultures. Although the dataset enables controlled cross-cultural evaluation, it remains relatively limited in scale and cultural coverage. The benchmark contains only five target cultures and 45 meme templates, which may not fully capture the diversity, evolution, and regional variability of online hateful discourse. Expanding culturally aligned hate speech datasets across additional countries, languages, and sociopolitical contexts remains an important direction for future work.

Static Cultural Knowledge. CHER relies on externally retrieved cultural knowledge sources such as Wikipedia, Cultural Atlas, World Values Survey (WVS), and minority-group knowledge repositories. These resources are largely static and may fail to capture rapidly evolving online slang, emerging political events, internet subcultures, or newly formed stereotypes. As cultural discourse changes over time, retrieved knowledge may become incomplete, outdated, or insufficiently representative of real-world online communities.

Evaluation Scale. The human evaluation of hate frame quality is conducted on 100 memes per culture with three annotators, which, while consistent with comparable studies, constitutes a relatively small evaluation pool.

Computational Cost. CHER involves multiple sequential reasoning stages, including social knowledge inference, multi-source cultural knowledge retrieval, self-reflection, HDE identification, counterfactual explanation generation, and contrastive verification. This multi-stage pipeline introduces substantially higher computational cost, which may limit practical deployment in real-time content moderation settings, where latency constraints are critical.

Broader Multimedia Contexts. CHER is designed for the image-text meme format as annotated in $Multi^3$*Hate*. In practice, culturally grounded hate speech also appears in short-form videos, livestreams, audio clips, and multimodal conversational interactions that contain temporal, auditory, and contextual information beyond static meme content. Extending CHER to richer multimedia settings remains an important avenue for future work.

9 Ethical Statement

We respected the privacy and confidentiality of individuals whose social media content appeared in the $Multi^3$*Hate* dataset (Bui et al., 2025). All examples used in this work were handled in accordance with the original dataset usage policies, and no personally identifiable information was collected, stored, or redistributed as part of this study. This work received approval from the Institutional Review Board (IRB) at ANONYMIZED.

To support transparency and reproducibility, we document all experimental settings, prompts, configurations, and annotation procedures in the paper, supplemental material, and linked GitHub repository. Human evaluations were conducted to assess the quality of the generated Hate Frames, and we report detailed inter-annotator agreement results across evaluation criteria and cultures.

Given the harmful and offensive nature of the dataset, annotators were informed beforehand about the content and provided explicit consent prior to participation. Annotators were encouraged to take breaks whenever necessary during the annotation process.

While exposure to hateful content may pose potential psychological risks, the goal of this work is to advance the understanding of culturally grounded online hate and support the development of systems capable of identifying, interpreting, and mitigating harmful narratives across cultures. We hope this work contributes toward the responsible advancement of multimodal NLP and the creation of safer and more inclusive online spaces.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wen-

bin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Aashish Bhandari, Siddhant Bikram Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1994–2003.

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.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. 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, Lluís Gómez, and Dimosthenis Karatzas. 2019. Exploring hate speech detection in multimodal publications. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen, editors. 2020. *World Values Survey: Round Seven – Country-Pooled Datafile*. JD Systems Institute and WVSA Secretariat, Madrid, Spain and Vienna, Austria.

Michael Eric Herz and Peter Molnár. 2012. The content and context of hate speech: Rethinking regulation and responses. Cambridge University Press.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Mahammed Kamruzzaman, Hieu Minh Nguyen, Nazmul Hassan, and Gene Louis Kim. 2026. 'a woman is more culturally knowledgeable than a man?': The effect of personas on cultural norm interpretation in LLMs. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 220–237, Rabat, Morocco. Association for Computational Linguistics.

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

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.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Mosaica. 2024. The cultural atlas.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large 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 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *ArXiv*, abs/1701.08118.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, and et al. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

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 H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880, Bangkok, Thailand. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations (ICLR-2025)*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Prompts used in CHER

Figure 6 shows the main prompt for the CKA-Agent. Figure 7 presents the prompt used by Tool 1 for generating culture-specific Wikipedia queries to retrieve historical context. Figure 8 shows the prompt used for Hate Defining Element (HDE) identification. Figure 9 presents the Chain-of-Explanation prompt used for generating counterfactual hateful and non-hateful explanations, while Figure 10 shows the Counterfactual Chain-of-Contrastive Verification prompt used to judge the validity of the generated counterfactual explanations for each HDE.

You are a Cultural Knowledge Acquisition Agent for the {country} culture.
You are given:
1. The meme
2. The meme caption
STEP 1 — INFERENCE OF SOCIAL KNOWLEDGE
Analyze the meme's visual content and caption to infer the following categories of social knowledge.
1. SOCIAL ENTITIES including people, locations, dates, or named references identifiable in the image or caption.
2. VISIBLE SOCIAL GROUPS: Groups identifiable through observable characteristics such as age, skin tone, gender presentation, race, or physical features.
3. SYMBOLS: Objects, signs, icons, or visual elements that carry culturally encoded meaning beyond their literal appearance.
4. SOCIAL DOMAIN: The specific social context the meme touches such as family, school, workplace, religion, or governance that influences how individuals interact, behave, and form relationships.
STEP 2 — KNOWLEDGE ACQUISITION
Using the social knowledge inferred in Step 1, call the available tools to gather cultural knowledge.
AVAILABLE TOOLS:
Tool 1: Historical Context Retriever
Description: Generates queries, searches Wikipedia and returns the summaries of the top-*k* most relevant articles.
Input: Social entities, Visible social groups, symbols, social domain, *k*
Output: Article names and plain text summaries of the top-*k* most relevant Wikipedia articles.
Tool 2: Cultural Values and Social Norms Retriever
Description: Retrieves culturally relevant social norms and value-oriented knowledge for a given social domain and target culture by performing semantic retrieval over the World Values Survey (WVS) and Cultural Atlas. From WVS, the tool returns semantically matched survey questions together with their statistical summaries. From Cultural Atlas, it returns the matched section names, subsection names, and corresponding paragraph content. If no WVS questions or Cultural Atlas sections exceed the similarity threshold τ, the tool falls back to querying Wikipedia using the social domain and {country}.
Input: Social domain, τ
Output: Semantically matched WVS survey questions together with their statistical summaries, and matched Cultural Atlas section names, subsection names, and corresponding paragraph content. If no entries from WVS or Cultural Atlas exceed the similarity threshold, the tool instead returns the top-*k* Wikipedia article summaries retrieved using the social domain and {country}.
Tool 3: Minority Group Knowledge Retriever
Description: Returns the pre-compiled minority group information for the target culture from the Minority Rights Group World Directory.
Input: {country}
Output: Cultural profile containing three fields:
- Main languages: main languages spoken in {country}.
- Main religions: main religions practiced in {country}.
- Minority groups: enumerated minority and indigenous population groups.
STEP 3 — SELF-REFLECTION
You are a Cultural Knowledge Verifier.
Your cultural background:
{cultural_atlas_core_concepts}
Key cultural sensitivities in {culture}:
Do's: {cultural_atlas_dos} Don'ts: {cultural_atlas_donts}
For each entry returned by Historical Context Retriever (Tool 1), evaluate:
(1)Cultural consciousness: Is this information known in {country}'s mainstream public discourse?
(2) Meme relevance: Does this information directly connect to this specific meme's visual or textual content, rather than being only topically adjacent?
For the Cultural Values and Social Norms Retriever (Tool 2), evaluate only Meme relevance.
Discard any entry that fails either criterion. Do not apply self-reflection to Minority Group Knowledge Retriever entries — these are pre-validated resources.
STEP 4 — CULTURAL KNOWLEDGE PACKAGE
Return the Cultural Knowledge Package (CKP) as a structured output consisting of three fields, constructed only from entries that satisfy both self-reflection criteria:
1.Historical Context: Return **article name, short summary, relevance** (an explanation of its specific connection to this meme's content), and **mainstream reach** (High, Medium, Low).
2.Cultural Values and Social Norms:
- Relevant **survey questions** along with corresponding **statistical summaries** from World Values Survey.
- Relevant **section** and **subsection name(s)**, and **corresponding content** in the subsection from the Cultural Atlas.
- For Wikipedia, return Wikipedia article name and short summary of the article.

3. Minority Group Knowledge: Main languages, Main religions, Minority groups
If no entries retrieved from a tool satisfy the Cultural Consciousness and Meme Relevance criteria during self-reflection, the corresponding knowledge field for that tool should be returned as NULL.

Figure 6: Main prompt for CKA-Agent.

You are a historical context retrieval assistant for {country} culture.
You have been given the following information identified from a meme:
Social entities (people, locations, dates): {social_entities}
Visible social groups (e.g., groups identifiable through age, skin tone, gender presentation, race, etc.): {visible_social_groups}
Symbols: {symbols}
Social domain (additional contextual information describing the setting or interactional context of the meme, such as workplace, school, family, religion, politics, or public life): {social_domain}
Your task is to construct one Wikipedia search query per extracted social entity, visible social group, or symbol. When constructing each query:
- Use the country name in some form to make the query culture-specific.
- Use the social domain as additional contextual information whenever necessary to make the query culturally and socially specific.
Return generated queries in a list.

Figure 7: Prompt used by Tool 1 to generate culture-specific queries for retrieving historical context.

You are a cultural hate interpretation analyst. Your task is to identify culturally grounded Hate Defining Elements (HDEs) for a meme in {country} culture.
An HDE consists of three components:
1. **Target** — the social group, identity, or attribute toward which hate is directed.
2. **Intent** — the social function or harmful intention underlying the hateful interpretation.
3. **Act** — the observable form through which the hate is expressed in the meme.
You are given:
1. The meme
2. The meme caption
3. The complete HDE taxonomy
4. A **Cultural Knowledge Package (CKP)** retrieved for the {country} culture
Your goal is to propose candidate HDEs that are explicitly supported by evidence in the CKP. Use the following HDE taxonomy to represent hateful interpretations as Target–Intent–Act triples.

INSTRUCTIONS
---
1. ONLY propose HDEs supported by evidence in the CKP.
2. Multiple HDEs may be proposed if the meme simultaneously involves multiple targets, intents, or acts.
3. If the meme does not support any hateful interpretation grounded in CKP, return NULL.
4. Focus specifically on how the meme could be interpreted within the {country} culture.
5. For each HDE, identify:
- Target: which group is targeted and its specific value.
- Intent: what is the hateful purpose and its specific value.
- Act: how hate is expressed and its specific value.
6. Generate a **rationale** explaining how each identified HDE is grounded in the CKP.

OUTPUT FORMAT

Return a JSON list.
Each entry must contain *target*, *intent*, *act*, and a *rationale*.

Figure 8: Prompt used for the identification of Hate Defining Elements.

You are a cultural hate interpretation analyst.
Your task is to evaluate multiple proposed Hate Defining Elements (HDEs) for a meme in {country} culture.
An HDE consists of three components:
1. **Target** — the social group, identity, or attribute toward which hate is directed.
2. **Intent** — the social function or harmful intention underlying the hateful interpretation.
3. **Act** — the observable form through which the hate is expressed in the meme.
You are given:
1. The meme
2. The meme caption
3. A list of HDE(s) proposed for the meme in {country} culture.
4. A rationale for each HDE explaining how it is grounded in the Cultural Knowledge Package (CKP).
5. A **Cultural Knowledge Package (CKP)** retrieved for the {country} culture.
Generate:
(1) a *hateful* rationale explaining what culturally grounded reasoning would make the meme hateful under the proposed HDE, and
(2) a *non-hateful* rationale explanation describing what reasoning would support interpreting the meme as non-hateful within the same cultural context,
for EACH proposed HDE.
Both explanations must be grounded in the cultural knowledge package.

INSTRUCTIONS
----
1. For each HDE, the *hateful* explanation must explain:
- why the meme could reasonably be interpreted as hateful, specifically through the proposed Target, Intent, and Act.
- use the associated rationale to identify what knowledge from the CKP motivates the proposed HDE.
2. For generating the *non-hateful* explanation for each HDE you must examine:
- whether the rationale provided alongside each HDE sufficiently justifies the proposed HDE.
- whether the rationale is genuinely grounded in the CKP.
- if the rationale is weak or insufficiently grounded, use this as the basis for the *non-hateful* explanation.
- even when the HDE is grounded in the CKP, consider whether an alternative non-hateful interpretation of the meme is still possible in the {country} culture
3. Do NOT invent historical events, or cultural claims not present in the CKP.
4. Keep reasoning specific to the {country} culture.

OUTPUT FORMAT

Return a JSON list.
Each entry must contain: the HDE, a *hateful* explanation, and a *non-hateful* explanation.

Figure 9: Chain-of-Explanation prompt used for generating counterfactual explanations.

You are a cultural hate interpretation analyst.
You are given:
1. A meme and its caption
2. A list of proposed Hate Defining Element(s) proposed for the meme in {country} culture.
3. Two **counterfactual explanations** for each HDE:
   - a hateful explanation arguing why the meme should be interpreted as hateful through the proposed HDE
   - a non-hateful explanation arguing why the meme should not be interpreted as hateful through the same HDE
4. A **Cultural Knowledge Package (CKP)** retrieved for the {country} culture.

An HDE consists of three components:
1. **Target** — the social group, identity, or attribute toward which hate is directed.
2. **Intent** — the social function or harmful intention underlying the hateful interpretation.
3. **Act** — the observable form through which the hate is expressed in the meme.

Your task is to evaluate the competing hateful and non-hateful explanations for each proposed HDE and determine whether the HDE represents a valid hateful interpretation of the meme within the {country} culture.

You must then synthesize the final *Hate Frame* for the meme.
A Hate Frame consists of the following components:
1. **Hate label:** *hateful* or *not hateful*
2. **Identified HDE(s):** HDEs accepted as valid hateful interpretations after comparing the hateful and non-hateful explanations. Return NULL if no HDE is accepted.
3. **Cultural Knowledge:** Relevant cultural knowledge from the CKP that supports the accepted hateful interpretation(s). Return NULL if no HDE is accepted.
4. **Hate Explanation:** A final explanation justifying why the meme is interpreted as hateful or non-hateful within the {country} culture based on the evaluated counterfactual explanations.
5. **Alternative Interpretation:** The non-hateful interpretation of the meme if the meme is judged as not hateful. Return NULL if the meme is hateful.

INSTRUCTIONS
---
1. Evaluate each proposed HDE independently by comparing the hateful and non-hateful explanations.
2. Accept an HDE only when:
   - the hateful explanation is strongly grounded in the CKP,
   - the hateful interpretation coherently explains the meme through the proposed Target, Intent, and Act,
   - and the non-hateful explanation fails to better explain the meme within the {country} culture.
3. Reject an HDE when:
   - the hateful explanation is weak, indirect, speculative, or insufficiently grounded in the CKP,
   - or the non-hateful explanation provides a more plausible interpretation of the meme within the target culture.
4. Multiple HDEs may be accepted if the meme simultaneously supports multiple culturally grounded targets, intents, or acts within the {country} culture.
5. Use only cultural knowledge present in the CKP to construct Hate Frame.
6. The final Hate Explanation must explain why the hateful or non-hateful interpretation is ultimately more convincing within the {country} culture.
7. If no HDE is accepted:
   - assign the meme a not hateful label,
   - set Accepted HDE(s) and Cultural Knowledge to NULL,
   - and provide the alternative non-hateful interpretation of the meme.

OUTPUT FORMAT

Return a JSON object containing: hate label, accepted HDEs, cultural knowledge, hate explanation, and alternative interpretation.

Figure 10: Counterfactual Chain-of-Contrastive Verification prompt used to judge the validity of counterfactual explanations generated for each HDE.

B Qualitative Examples of Cultural Knowledge Retrieval and Hate Frame Generation

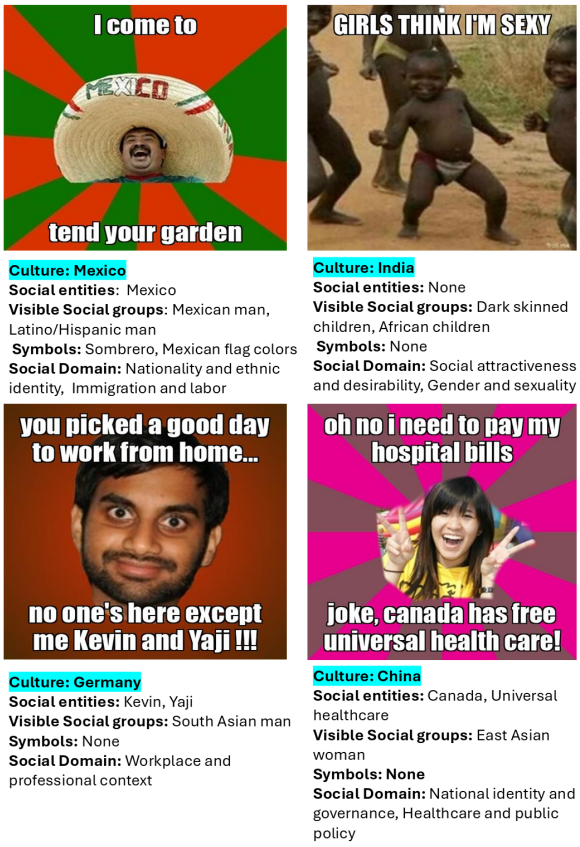


Figure 11: Examples of social knowledge inferred for various memes across several cultures.

Figure 11 presents examples of social knowledge inferred by CHER across different memes and cultures. For each meme, the system extracts four categories of culturally relevant information: *social entities*, *visible social groups*, *symbols*, and the *social domain*. For instance, the Mexico example identifies culturally salient symbols such as the sombrero and Mexican flag colors alongside immigration and labor-related social domains, while the Germany example captures workplace-related social context and identifies named entities such as "Kevin" and "Yaji." Similarly, the India example focuses on visible racialized social groups and themes of attractiveness and desirability, whereas the China example identifies governance and healthcare-related social domains connected to national identity and public policy discourse. Overall, the figure demonstrates CHER's ability to extract culturally grounded contextual signals that later guide CKP retrieval and Hate Frame generation. Figure 12 presents example outputs from Tool 1 (Historical Context Retriever) for the same meme interpreted in Germany and India. The figure illustrates how culture-specific Wikipedia queries generate different retrieved knowledge depending on the target culture, and how the self-reflection step filters retrieved entries using the *Cultural Consciousness* and *Meme Relevance* criteria.

For Germany, the query *Kevin Germany* retrieves the article *Kevinismus*, a culturally recognized German stereotype associating the name "Kevin" with lower socioeconomic status and poor educational background. This entry is marked as both culturally conscious and meme relevant, allowing it to remain in the final Cultural Knowledge Package (CKP). In contrast, the query *Yaji Germany* retrieves an unrelated article about *Peter Gilmour*, which fails both filtering criteria and is discarded.

Similarly, for India, the query *Indian workplace hierarchy* retrieves the article *Caste*, which reflects socially recognized hierarchical structures relevant to workplace and class interpretations. Although culturally conscious, it is judged insufficiently relevant to the meme and therefore excluded from the final CKP. Queries such as *Kevin India* and *Yaji India* retrieve unrelated entities and are rejected by the self-reflection module.

Overall, the figure demonstrates how CHER performs culture-specific historical retrieval while using self-reflection to remove culturally irrelevant or weakly grounded knowledge before Hate Frame generation.

Figure 13 presents example outputs from Tool 2, which retrieves culture-specific social norms and value-oriented knowledge using the inferred social domain *Workplace and professional context*. The figure compares the retrieved World Values Survey (WVS) statistics and Cultural Atlas business culture knowledge for Germany and India.

The top section shows semantically matched WVS questions along with their statistical summaries. Although both cultures strongly associate work with social importance and duty, the responses reveal notable cultural differences. For example, agreement with the statement *Work should always come first* is substantially higher in India (83%) than in Germany (28%), reflecting differing workplace value systems.

The bottom section presents retrieved Cultural Atlas subsections related to business culture and workplace interaction norms. For Germany, the retrieved knowledge emphasizes professionalism, adherence to rules, expertise-based hierarchy, and direct com-

munication styles. In contrast, the India retrieval highlights hierarchical workplace structures based on age and position, indirect communication patterns, relationship-oriented negotiation practices, and stronger sensitivity to interpersonal hierarchy. Overall, the figure illustrates how Tool 2 retrieves culturally grounded workplace norms and social values that differ substantially across cultures despite sharing the same inferred social domain. These retrieved signals later help contextualize culture-specific hateful interpretations during Hate Frame generation.

Table 5 summarizes demographic and cultural background knowledge about minority groups collected from [5] for the five cultures used in this work. For each culture, the table presents major spoken languages, dominant religious traditions, and socially relevant minority groups that may become targets of culturally grounded hateful discourse. The listed minority groups include ethnic, racial, religious, migrant, and indigenous communities that are historically or socially significant within each cultural context.

Figure 14 presents an end-to-end example of CHER's counterfactual reasoning process for the same meme interpreted in two different cultures: Germany and India. The figure illustrates how CHER generates culturally grounded Hate Defining Elements (HDEs), produces competing hateful and non-hateful explanations, and synthesizes the final Hate Frame through contrastive verification.

For Germany, CHER identifies a culturally grounded HDE targeting *socioeconomic status* through the documented stereotype of *Kevinismus*, which associates the name "Kevin" with lower socioeconomic status, poor educational attainment, and reduced professionalism in German discourse. The generated hateful explanation argues that the meme indirectly portrays "Kevin" and "Yaji" as lower-status workers who remain in the office while others work remotely, reinforcing workplace hierarchy stereotypes. The non-hateful explanation instead frames the meme as ordinary workplace humor related to remote work attendance. After comparing both interpretations, CHER accepts the hateful interpretation and generates a final Hate Frame labeled as hateful.

In contrast, for India, CHER fails to identify any culturally grounded HDE supported by the retrieved Cultural Knowledge Package (CKP). Since no relevant historical associations, social norms, or minority-related sensitivities are relevant for the meme, the system does not identify any HDEs and produces a final Hate Frame labeled as not hateful, interpreting the meme as general workplace humor.

[5] Minority Rights Group International (MRG) World Directory https://minorityrights.org/

Country: Germany

Query 1: South Asian workers in Germany
Article Name: Gastarbeiter
Article Summary: In German-speaking countries, the term Gastarbeiter (pronounced [ˈɡastˌʔaʁbaɪtɐ] ; both singular and plural) refers to foreign or migrant workers, particularly those who had moved to West Germany between 1955 and 1973, seeking work as part of a formal guest worker program (German: Gastarbeiterprogramm). Other countries had similar programs: in the Netherlands and Belgium it was called the gastarbeider program; in Sweden, Denmark and Norway it was called arbetskraftsinvandring (workforce-immigration); and in East Germany such workers were called Vertragsarbeiter. == Terminology ==In Nazi Germany, the term used to refer to foreign workers was Fremdarbeiter. However, the term obscured the compulsory nature of the work, and after World War II, the term has since been used to distinguish worker policies of the Nazi era to the Gastarbeiter programs of the Federal Republic of Germany.
Cultural Consciousness: Yes
Meme Relevance: No
Query 2: Kevin Germany
Article Name: Kevinismus
Article Summary: Kevinismus ("Kevinism") is a German term for the practice of giving children trendy, exotic-sounding names as opposed to traditional German ones. It is often considered to be an indicator of low social class. The prototypical example is Kevin, which like most such names came to Germany from Anglo-American culture. Specifically, Kevin McCallister, the protagonist of the 1990 comedy film Home Alone (titled Kevin – Allein zu Haus in the German release) is credited with making Kevin the most popular boys' name chosen in Germany in 1991.
Cultural Consciousness: Yes
Meme Relevance: Yes
Query 3: Yaji Germany
Article Name: Peter Gilmour
Article Summary: Peter Gilmour (born 25 January 1960) is an Australian sailing skipper. Gilmour is an America's Cup veteran and was named Western Australian Sportsman of the Year in 1987. He is also the only 4 time World Match racing Tour Champion. He currently resides in Perth, Western Australia.
Cultural Consciousness: No
Meme Relevance: No

Country: India

Query 1: Indian workplace hierarchy
Article Name: Caste
Article Summary: In India, a caste is a (usually endogamous) social group where membership is decided by birth. Broadly, Indian castes are divided into the Forward Castes, Other Backward Classes, Scheduled Castes, and Scheduled Tribes. Indian Christians and Indian Muslims are also function as castes (a full list of castes can be found at the end of this article). With castes separating individuals into different social groups, it follows that each group will have conflicting interests; oftentimes putting those with lower social standing in less favorable positions.
Cultural Consciousness: Yes
Meme Relevance: No
Query 2: Kevin India
Article Name: Chloe East
Article Summary: Chloe East (born February 16, 2001) is an American actress. She started her career as a child actress, appearing in two 2013 episodes of True Blood. She starred in the first season of Ice (2016–2017) and in the drama series Generation (2021). She has also featured in Steven Spielberg's semi-autobiographical The Fabelmans (2022), and in the horror film Heretic (2024). **Cultural Consciousness: No**
Meme Relevance: No
Query 3: Yaji India
Article Name: Bhumihar
Article Summary: Bhūmikā (Sanskrit: भूमिका) is derived from the word, Bhūmi, meaning earth, soil, ground or character. In the Upanishads, the Bhumikas are seven stages in the development of wisdom.== Hinduism ==The Upanishads speak about the seven bhūmikās or jñānabhūmis (fields of knowledge) or the seven stages of development of wisdom, and about the four kinds of jivanmuktas. In Chapter IV of the Varahā Upanishad, with regard to the characteristics of jivanmukti, Ribhu informs Nigadha of these seven:Subheccha or good desiresVichārāna or inquirytanumānasī or pertaining to the thinned mindSattvāpatti or the attainment of sattvaAsamśakti or non-attachmentPadārthabhāvanā or analysis of objectsTuriya or the final stage.
Cultural Consciousness: No
Meme Relevance: No

Figure 12: Most relevant article summaries retrieved by Tool 1 (Historical Context Retriever) for Germany and India, together with the corresponding Wikipedia search queries.

Social Domain: Workplace and professional context

Country: Germany	Country: India
World Values Survey (Question + Statistical Summary)	
1. **How important is work in your life? 82% say it is important.** 2. **Work is a Duty Towards Society. 74% agree.** 3. **Work Should Always Come First. 28% agree.**	1. **How important is work in your life? 93% say it is important.** 2. **Work is a Duty Towards Society. 84% agree.** 3. **Work Should Always Come First. 83% agree.**
Cultural Atlas (Section: Business Culture -> Subsection: Considerations)	
Subsection Content:	
▪ Workplaces in Germany are hierarchical based on experience and position. Although they respect those in authority, they dislike control or leadership that is solely based on status as opposed to expertise. ▪ Germans display a great amount of respect for those with education and experience, and will seek to know how your position relates to theirs. ▪ Expect a German to closely adhere to any regulations or rules that relate to the task. They may be hesitant to do work with those who show a tendency to cut corners. ▪ When doing business, Germans may develop new binding protocols during negotiations to avoid uncertainty and inconsistency, as well as to assure your reliability. ▪ When negotiating, written communication is used to record discussions and uphold agreements. ▪ Displays of passionate emotion, exaggerations or promises that sound too good to be true are likely to make Germans hesitant or suspicious of doing business with you. ▪ The 2023 Corruption Perception Index ranked Germany 9th out of 180 countries, receiving a score of 78/100.[1] This metric suggests that the country's public sector is quite clean from corruption. • Germans commonly see themselves as critical thinkers, constantly asking questions and seeking new perspectives to guide their decision-making. They can be flexible and will speak out to suggest a new idea if they believe there's a better way. • You may find that a German company has a well-founded confidence in their own way of doing things. This can come across as them being overly sure of themselves. Avoid interpreting this as meaning that they are closed-off to new ideas.	▪ Workplaces in India are hierarchical, based on age and position. ▪ Most Indians do not feel comfortable speaking in a direct, frank manner with those they do not have a strong relationship with. Therefore, expect indirect, circuitous communication to occur to reach an agreement or deal. ▪ They dislike rejecting business but will do so if it means future gain. ▪ In India, the polite way to say no is to say, "I'll see what I can do", or something to that effect, no matter how impossible the task may be. After they have been queried several times concerning their success, an answer of "I'm still checking" or something similar means "no". Such an indirect response also means "I am still your friend/ally, I tried". Therefore, be as transparent as possible with the questions in order to receive a clear yes or no answer (but, be cautious in doing this, or you will harm the relationship). ▪ It is common for senior members of a meeting to take phone calls. However, it is considered rude for juniors to follow this practice. ▪ If you would like to contact your Indian counterpart through a phone call, it is usually best to call 10am onwards but no later than 9pm in the evening. ▪ Everything is subject to change in India, so Indians can regard contracts more as agreements of willingness to do business. The actual specifications of limitations and regulations of a deal are negotiable depending on the business relationship. Try to be flexible – as they are – and be aware that a request for a more comprehensive document could jeopardise trust. ▪ On the Corruption Perception Index (2017), India ranks 81st out of 180 countries, receiving a score of 40 (on a scale from 0 to 100). This perception suggests that the country's public sector is somewhat corrupt.

Figure 13: Example of Cultural Values as well as Social Norms retrieved for the social domain *Workplace and professional context*.

	Main Languages	Main Religions	Minority Groups
US	English, Spanish	Christianity, Judaism, Islam	Hispanic and Latino communities, Mexicans, Puerto Ricans, Cubans, African Americans, Asian Americans, Asian Indians, Chinese Americans, Filipino Americans, Japanese Americans, Korean Americans, Vietnamese Americans, Native Americans, American Indians, Alaska Natives, Native Hawaiians, Pacific Islanders, Arab Americans, Middle Eastern Americans, Iranian Americans, Armenian Americans, Assyrian / Chaldean / Syriac communities, Turkish Americans, Jewish communities, Muslims, Amish, Quakers, Dukhobor and Bruderhoffer Christians, Christian Scientists, Hare Krishnas, Mormons
Germany	German, Danish, Frisian, Sorbian, Low German, Arabic, Kurdish, Polish, Russian, Turkish	Lutheran Christianity, Roman Catholicism, Islam, Judaism	Danes, Frisians, Sorbs, Roma / Sinti, Turkish migrant communities, Syrian refugee communities, Iraqi refugee communities, Afghan refugee communities, Jewish communities
Mexico	Spanish, Zapotec, Yucatec Maya, Nahuatl, Mixtec	Roman Catholicism, Indigenous religions	Indigenous Peoples (including Nahua, Maya, Zapotec, Mixtec, Tzotzil, Tzeltal, Otomi, Totonaca, Chol, Mazatec, Rarámuri / Tarahumara, Purépecha, Tepehuanes / Ódami, Wixárika / Huichol, Mixe), Afro-Mexicans, Afro-descendants, Black Mexicans, Indigenous migrants, Salvadoran migrants, Honduran migrants, Afro-Hondurans, Guatemalan migrants, Zapatista Indigenous communities
India	Hindi, Bangla, Telugu, Marathi, Tamil, Urdu, Gujarati, English	Hinduism, Islam, Christianity, Sikhism, Buddhism, Jainism, Zoroastrianism, Judaism	Muslims, Christians, Sikhs, Dalits, Adivasis, Anglo-Indians, Kashmiris, Jews, Andaman Islanders
China	Mandarin Chinese, Cantonese (Yue), Shanghainese (Wu), Fuzhou (Minbei), Hokkien-Taiwanese (Minnan), Qiang, Gan, Hakka, Uyghur, Tibetan	Buddhism, Roman Catholicism, Islam, Tibetan Buddhism (Lamaism), Protestantism, Shamanism, Animism, Taoism, Dongba religion	Zhuang, Manchu, Hui, Miao, Uyghurs, Yi (Lolo), Tujia, Mongols, Tibetans, Mosuo, Chuanqing, Gaoshan Indigenous peoples, Sherpas, Mang, Khmu, Minzú communities, Inner Mongolians, Xinjiang minorities, Tibetan minorities

Table 5: Knowledge concerning Minority Groups for all five cultures

Country: Germany

HDEs Identified:
Target(s): Socioeconomic Status **Target Value(s):** Lower socioeconomic background
Intent(s): Stereotyping
Intent Value(s): Workers from lower socioeconomic backgrounds are portrayed as socially undesirable and having lower status within professional environments.
Act(s): Insults
Act Value(s): Implying Kevin and Yaji are lower-status workers.
Rationale: The rationale is supported by CKP historical knowledge related to "Kevinismus," a documented German class-based stereotype that associates the name "Kevin" with lower socioeconomic status, poor educational attainment, and reduced professionalism.

Counterfactual Reasoning and Hate Speech Identification:
Socioeconomic Status → Stereotyping → Insults
Hate Explanation **(e_+^{HS}):**
In German discourse, "Kevin" is a documented class-associated name through the phenomenon of *Kevinismus*, where the name is linked to lower socioeconomic status, poor educational outcomes, and socially undesirable behavior. The meme presents a workplace situation in which Kevin and the depicted individual remain in the office while others work remotely, which may reinforce stereotypes associating lower-status individuals with reduced workplace privilege and inferior professional standing.
Non-Hateful **Explanation (e_-^{HS}):**
Although *Kevinismus* is a recognized stereotype in German culture, the meme functions as general workplace humor concerning remote work and office attendance. "Kevin" and "Yaji" are presented as individual colleagues remaining in the office, and the humor arises from the relatable inconvenience of being left behind rather than from hostility toward any socioeconomic group.

Hate Frame:
Hate label: Hateful
Accepted HDE(s):
Target(s): Socio-economic Status **Target Value(s)**: Lower socioeconomic background
Intent(s): Stereotyping
Intent Value(s): Workers from lower socioeconomic backgrounds are portrayed as socially undesirable and having lower status within professional environments.
Act(s): Insults
Act Value: Implying Kevin and Yaji are lower-status workers.
Cultural Knowledge: In Germany, "Kevin" functions as a culturally loaded name through the documented Kevinismus stereotype, which associates the name with lower educational achievement and undesirable social status.
Hate Explanation: The meme activates the German "Kevinismus" stereotype by implying that Kevin and Yaji remain in the office while others work remotely, indirectly portraying them as lower-status workers with reduced workplace privilege.

Country: India

HDEs Identified: None

Hate Frame:
Hate label: Not Hateful
Accepted HDE(s): None
Cultural Knowledge: None
Alternate Interpretation: Workplace humor
Hate Explanation: The retrieved cultural knowledge for India contains no relevant stereotype associations, historical context, or minority-group sensitivities connected to the meme. The meme is therefore interpreted as ordinary workplace humor related to remote work and office attendance.

Figure 14: Example of HDE identification, counterfactual explanations and Hate Frame.

Definition of Hate Speech:
Any kind of communication that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.
Answer if this meme is hate or not in {country} with a single alphabet letter among given answer choices a and b.
You are given:
1. Meme
2. Caption inside the meme image
Choose:
a: Hate
b: Non-hate

Figure 15: Zero-shot prompt used in $Multi^3$*Hate* (Bui et al., 2025).

C Prompts used for Baseline Methods,

Figure 15 shows the standard zero-shot prompting baseline from $Multi^3$*Hate* (Bui et al., 2025). They introduce six simple prompt variations for hate speech interpretation, and we use the first template as the representative zero-shot baseline in our experiments. Figure 16 presents the Zero-shot Chain-of-Thought (ZS-CoT) prompting strategy, which encourages step-by-step reasoning during hate speech interpretation. Figure 17 shows the Zero-shot Chain-of-Thought with Cultural Persona (ZS-CoT+CP) prompt, where the model is additionally guided to reason from the perspective of the target culture.
For each culture C_j, we employ anthropological prompting (AlKhamissi et al., 2024) to prompt GPT-5 with meme M_i. The prompt encourages the model to reason through anthropological concepts such as emic and etic perspectives, cultural relativism, socioeconomic background, lived experience, and regional context when interpreting the meme. We further extend the original anthropological prompting framework with additional instructions for cultural knowledge generation, Hate Defining Element (HDE) identification, hate explanation generation, and final hate speech classification. Figure 18 presents the complete anthropological prompt used in our experiments.

Definition of Hate Speech:
Any kind of communication that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.
Use the following HDE taxonomy to represent hateful intent as Target–Intent–Act triples.
{HDE Taxonomy}
1. Target — the social group, identity, or attribute toward which hate is directed.
2. Intent — the social function or harmful intention underlying the hateful interpretation.
3. Act — the observable form through which the hate is expressed in the meme.
Given an input meme, determine whether it constitutes Hate Speech or Non-Hate Speech in {country}.
Meme Caption: {caption}
Reason through the following steps:
Step 1 — Cultural Knowledge:
What cultural knowledge about {country} is relevant to interpreting this meme? Consider historical context, social norms, stereotypes, or minority group dynamics specific to {country} that a native would recognize.
Step 2 — HDE Identification:
If this meme is hateful in {country}, identify:
- Target: which group is targeted and what specific value (e.g., “Immigrants").
- Intent: what is the hateful purpose and its specific value.
- Act: how hate is expressed and its specific value.
If not hateful, set all fields to None.
Step 3 — Hate Explanation:
In one sentence, explain why this meme is or is not hateful in {country}, grounded in the cultural knowledge from Step 1.
Step 4 — Final Label:
Provide your final answer as *Hate* or *Non-Hate*.
Provide your response in the following format:
Cultural Knowledge: [relevant cultural context about {country} that informs this judgment]
Target Category: [taxonomy label or None]
Target Value: [specific group or attribute being targeted, or None]
Intent Category: [taxonomy label or None]
Intent Value: [specific hateful purpose in this cultural context, or None]
Act Category: [taxonomy label or None]
Act Value: [how the hate is expressed in this meme, or None]
Hate Explanation: [one sentence explaining why this meme is or is not hateful in {country}]
Label: Hate or Non-Hate

Figure 16: Zero-shot Chain-of-thought Prompt.

You are a culturally aware hate speech analyst tasked with interpreting memes within the social, historical, and cultural context of the {country} culture.
Your cultural background:
{cultural_atlas_core_concepts}
Key cultural sensitivities in {country}:
Do's: {cultural_atlas_dos}
Don'ts: {cultural_atlas_donts}
Definition of Hate Speech:
Any kind of communication that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.
Use the following HDE taxonomy to represent hateful intent as Target–Intent–Act triples.
{HDE Taxonomy}
1. Target — the social group, identity, or attribute toward which hate is directed.
2. Intent — the social function or harmful intention underlying the hateful interpretation.
3. Act — the observable form through which the hate is expressed in the meme.
Given an input meme, determine whether it constitutes Hate Speech or Non-Hate Speech in {country}.
Meme Caption: {caption}
Reason through the following steps:
Step 1 — Cultural Knowledge:
What cultural knowledge about {country} is relevant to interpreting this meme? Consider historical context, social norms, stereotypes, or minority group dynamics specific to {country} that a native would recognize.
Step 2 — HDE Identification:
If this meme is hateful in {country}, identify:
- Target: which group is targeted and what specific value (e.g., "Immigrants").
- Intent: what is the hateful purpose and its specific value.
- Act: how hate is expressed and its specific value.
If not hateful, set all fields to None.
Step 3 — Hate Explanation:
In one sentence, explain why this meme is or is not hateful in {country}, grounded in the cultural knowledge from Step 1.
Step 4 — Final Label:
Provide your final answer as *Hate* or *Non-Hate*.
Provide your response in the following format:
Cultural Knowledge: [relevant cultural context about {country} that informs this judgment]
Target Category: [taxonomy label or None]
Target Value: [specific group or attribute being targeted, or None]
Intent Category: [taxonomy label or None]
Intent Value: [specific hateful purpose in this cultural context, or None]
Act Category: [taxonomy label or None]
Act Value: [how the hate is expressed in this meme, or None]
Hate Explanation: [one sentence explaining why this meme is or is not hateful in {country}]
Label: Hate or Non-Hate

Figure 17: Prompt for Zero-shot Chain-of-thought with Cultural Persona.

The following is a framework adapted from the toolkit of anthropological methods:
1. Emic and Etic Perspectives: there are in-group ways of answering or thinking about a question and out-group ways.
2. Cultural Context: cultural context is pivotal in understanding different questions. This includes where people come from, what language they speak, where they live, and their kinship networks.
3. Individual Values and Personal Experience: experience and personal values play a big role in subjective understandings of day to day life.
4. Socioeconomic Background: income, family wealth, class, and socioeconomic background factor into answers.
5. Cultural Relativism: no culture is objectively better than another; understanding cultural relativism is crucial in understanding different personas.
6. Space and Time: age and place are important factors.
7. Nuance: each person will understand and answer questions based on the nuanced phrasing of the question.
Imagine you are a {marital_status} {sex} from {region}, {country}. You are {age} years of age and completed {education} education level. You consider yourself part of the {social_class}. Answer the following question from this perspective. Others will read what you choose; your goal is to convince them it was chosen from the perspective of the persona described above.
Definition of Hate Speech:
Any kind of communication that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.
Use the following HDE taxonomy to represent hateful intent as Target–Intent–Act triples.
{HDE Taxonomy}
1. Target — the social group, identity, or attribute toward which hate is directed.
2. Intent — the social function or harmful intention underlying the hateful interpretation.
3. Act — the observable form through which the hate is expressed in the meme.
You will analyze the following meme for {country}:
Meme image: {meme_image}
Caption: {caption}
Reason through the following steps:
Step 1 — Cultural Knowledge:
Drawing on the anthropological framework above and your personal background and lived experience in {country}, describe in one coherent paragraph what cultural knowledge is relevant to interpreting this meme — including historical context, social norms, stereotypes, or minority group dynamics that a native of {country} would recognize.
Step 2 — HDE Identification:
If this meme is hateful in {country}, identify:
- Target: which group is targeted and its specific value.
- Intent: what is the hateful purpose and its specific value.
- Act: how hate is expressed and its specific value.
If not hateful, set all fields to None.
Step 3 — Hate Explanation:
In one sentence, explain why this meme is or is not hateful in {country}, grounded in the cultural knowledge from Step 1.
Step 4 — Final Label:
Provide your final answer as Hate or Non-Hate.
Provide your response in the following format:
Cultural Knowledge: [one paragraph grounded in the anthropological framework and your persona]
Target Category: [taxonomy label or None]
Target Value: [specific group or attribute or None]
Intent Category: [taxonomy label or None]
Intent Value: [specific hateful purpose or None]
Act Category: [taxonomy label or None]
Act Value: [how hate is expressed or None]
Hate Explanation: [one sentence]
Label: Hate or Non-Hate

Figure 18: Anthropological Prompt.

D Anthropological Prompting

To capture diverse intra-cultural viewpoints for anthropological prompting, we construct demographic personas by sampling combinations of region, sex, age, social class, marital status, and education level from predefined demographic profiles shown in Table 6. The demographic profiles include culturally representative regions for each country, male and female identities, age groups between 20–39, upper-middle, lower-middle, and working-class socioeconomic backgrounds, multiple marital statuses, and varying education levels. For each meme-culture pair (M_i, C_j), the model is prompted from the perspective of three different personas, and the final hate speech prediction is obtained through majority voting across the persona-specific outputs.

E Additional Quantitative Results

Table 7 reports F1 scores for HDE identification across all baselines and CHER, evaluated on 100 hateful memes per culture using human-annotated ground truth. For each meme, three native annotators from the corresponding culture independently annotate the Target, Intent, and Act components, and the final HDE labels are obtained through majority voting. Annotators were provided with the complete HDE taxonomy, detailed descriptions of the HDE components, and examples explaining different Target, Intent, and Act categories. Inter-annotator agreement, measured using Krippendorff's α, averages 0.71 for Target, 0.69 for In-

Attribute	US	DE	MX	IN	CN
Region	Ohio / Texas / California	North Rhine-Westphalia / Bavaria / Berlin	Mexico City / Jalisco / Nuevo León	Maharashtra / Uttar Pradesh / Tamil Nadu	Guangdong / Beijing / Shanghai
Sex	Male / Female				
Age	20 – 39				
Social Class	Upper middle, Lower middle, Working				
Marital Status	Single, Married, Divorced				
Education	Lower, Middle, Higher				

Table 6: Demographic profiles for Anthropological Prompting.

	Target					Intent					Act				
Method	US	DE	MX	IN	CN	US	DE	MX	IN	CN	US	DE	MX	IN	CN
CHER	**84.2**	**82.1**	**80.8**	**78.4**	**79.6**	**78.1**	**76.4**	**74.9**	**73.9**	**72.8**	**80.4**	**76.8**	**78.2**	**74.3**	**75.9**
ZS-CoT	72.4	68.3	65.8	63.4	64.2	62.8	57.4	55.2	51.8	54.6	67.3	63.1	61.8	57.9	56.8
ZS-CoT+CP	74.1	70.4	70.8	65.4	67.3	64.2	60.8	59.1	54.7	57.9	69.8	66.4	64.7	60.1	62.4
FS-CoT	76.3	72.8	72.4	67.2	69.8	66.1	62.4	61.3	57.4	58.4	71.9	67.4	68.3	62.1	64.3
FS-CoT+CP	77.8	74.6	74.1	69.8	71.6	68.1	65.4	63.1	61.4	63.8	73.4	69.7	68.4	65.4	64.9
AP-Prompt	76.9	74.2	72.8	68.4	70.1	67.1	64.3	61.8	59.6	62.8	72.4	68.7	67.4	63.1	65.2

Table 7: F1 scores for HDE identification by CHER and prompting-based baselines, measured against human-annotated ground truth across 100 hateful memes per culture.

tent, and 0.71 for Act across all cultures, indicating substantial agreement. Across all methods, target identification achieves the highest scores, followed by Act identification, while Intent identification remains the most challenging task.

CHER consistently achieves the best performance across all cultures and HDE dimensions, demonstrating strong cross-cultural generalization. Among the baseline methods, FS-CoT+CP performs the strongest overall, suggesting that combining few-shot reasoning with culturally aware prompting substantially improves HDE identification. In contrast, zero-shot methods, particularly ZS-CoT, exhibit noticeably lower performance across all settings. The performance gap is more pronounced for India and China. Overall, the results demonstrate the effectiveness of CHER in capturing culturally grounded hateful discourse signals more accurately than existing prompting-based approaches.

F Error Analysis

Table 8 presents the confusion matrices and class-wise performance of CHER across all five cultures on $Multi^3Hate$. Across all cultures, CHER achieves substantially higher F_1 scores for the Hate class compared to the Non-Hate class.

Figure 19 presents examples where CHER's predictions diverge from the $Multi^3Hate$ ground-truth annotations. We observe that many of these disagreements arise from annotation ambiguity rather than clear reasoning failures.

Figures 19a and 19b illustrate cases where CHER predicts hateful interpretations for Germany despite non-hateful ground-truth labels. In Figure 19a, the meme invokes the globally recognizable stereotype of Indian call-center workers and associates it with technological incompetence. This stereotype is widely circulated within broader Western online discourse, including Germany, making the hateful interpretation culturally plausible. Similarly, Figure 19b relies on implicit racial caricature by comparing a dancing African child to Jay-Z. Although some annotators may interpret the meme as absurd celebrity humor, it can also plausibly activate anti-Black racial stereotypes historically present in Western internet culture.

Figure 19c highlights ambiguity surrounding na-

United States (US)

	Predicted		
Actual	Hate	Non-Hate	Total
Hate	**127**	27	154
Non-Hate	26	**120**	146
Total	153	147	300

Germany (DE)

	Predicted		
Actual	Hate	Non-Hate	Total
Hate	**167**	12	179
Non-Hate	39	**82**	121
Total	206	94	300

Mexico (MX)

	Predicted		
Actual	Hate	Non-Hate	Total
Hate	**148**	19	167
Non-Hate	38	**95**	133
Total	186	114	300

India (IN)

	Predicted		
Actual	Hate	Non-Hate	Total
Hate	**154**	26	180
Non-Hate	34	**86**	120
Total	188	112	300

China (CN)

	Predicted		
Actual	Hate	Non-Hate	Total
Hate	**167**	23	190
Non-Hate	33	**77**	110
Total	200	100	300

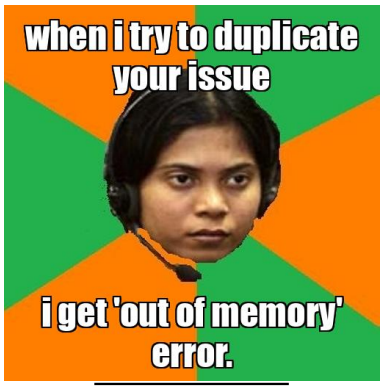
Metric	**US**	**DE**	**MX**	**IN**	**CN**
F_1^H	84.9	86.8	83.9	83.7	85.6
F_1^N	79.7	76.3	76.9	74.1	73.3
Macro-F1	82.3	81.5	80.4	78.9	79.5

Table 8: Confusion matrices for CHER across five cultures on $Multi^3Hate$. Rows = actual labels, Columns = predicted labels. F_1^H = F1 for Hate class, F_1^N = F1 for Non-Hate class.

tionalism and linguistic superiority. The meme may be interpreted either as humor or as mockery directed toward non-English-speaking groups through implicit cultural superiority claims. In contrast, Figure 19d illustrates a case where CHER assigns non-hateful labels for China and Mexico despite hateful ground-truth annotations. The meme's hateful interpretation is highly implicit, and it is not immediately clear why the meme should be considered hateful within the Chinese and Mexican cultural settings. This example therefore reflects ambiguity in how annotators perceive and interpret culturally grounded harmful intent across cultures. Figure 19e similarly reflects annotation ambiguity. The meme appears to mock overweight individuals through exaggerated humor involving fast food.

Figure 19e similarly reflects annotation ambiguity. The meme appears to mock overweight individuals through exaggerated humor involving fast food and aggressive behavior. While the meme is annotated as non-hateful in the Chinese setting, it is not immediately clear why the hateful interpretation should be rejected, since the meme plausibly targets body weight and reinforces negative stereotypes about overweight individuals. This example highlights how annotators may differ in whether they interpret appearance-based ridicule as hateful or merely as internet humor.

Importantly, the $Multi^3Hate$ benchmark (Bui et al., 2025) itself discusses annotation ambiguity and translation-related disagreements in cross-cultural hate speech interpretation. Several disagreements observed in Figure 19 may therefore stem not only from model prediction errors, but also from differences in annotator sensitivity and interpretations of humor, offensiveness, and harmful intent.



Culture	GT	Pred
US	1	1
DE	0	**1***
MX	0	**1***
CN	1	1
IN	0	0

(a)

Culture	GT	Pred
US	1	1
DE	0	**1***
MX	1	1
CN	1	1
IN	1	1

(b)

Culture	GT	Pred
US	1	1
DE	0	**1***
MX	1	1
CN	1	1
IN	0	0

(c)

Culture	GT	Pred
US	0	0
DE	0	0
MX	1	**0***
CN	1	**0***
IN	0	0

(d)

Culture	GT	Pred
US	1	1
DE	1	1
MX	1	1
CN	0	**1***
IN	1	1

(e)

Figure 19: Error analysis: memes where CHER diverges from Multi^{3}Hate ground truth (GT). * = prediction differs from GT.

G Human Evaluation Details

G.1 Annotator Profiles

We recruited three native annotators per culture, totalling fifteen annotators across the five target cultures. Each annotator was required to have resided continuously in their target country for a minimum of ten years and to demonstrate native proficiency in the culture's primary language. Annotators were recruited through academic and professional networks. The annotator pool spans a range of professional backgrounds including social sciences, computer science, linguistics, and communications, with all annotators holding at least a Bachelor's degree. The annotator pool across all cultures included both male and female participants.

G.2 Annotator Instructions

Annotators received a detailed annotation guide, presented in Figure 20, prior to starting the task. The guide covered the definition of hate speech used in this study, the HDE taxonomy, a description of the Hate Frame structure (Target, Intent, Act, Cultural Knowledge, Hate Explanation), and the evaluation criteria and scales for each dimension. Annotators were additionally provided with detailed descriptions of the Target, Intent, and Act components, along with examples explaining dif-

Dimension	Scale	US	DE	MX	IN	CN	Avg
Target Correctness	Binary	0.73	0.71	0.72	0.69	0.70	0.71
Intent Correctness	Binary	0.71	0.71	0.71	0.62	0.69	0.69
Act Correctness	Binary	0.72	0.73	0.70	0.66	0.74	0.71
Factual Accuracy	3-pt Likert	0.79	0.77	0.78	0.71	0.79	0.77
Meme Relevance	3-pt Likert	0.72	0.75	0.73	0.65	0.69	0.71
Plausibility	3-pt Likert	0.70	0.64	0.69	0.61	0.65	0.66
Coherence	Binary	0.77	0.77	0.75	0.70	0.74	0.75
HDE Quality avg		**0.72**	**0.72**	**0.71**	**0.65**	**0.71**	**0.70**
Cultural Knowledge avg		**0.75**	**0.76**	**0.75**	**0.68**	**0.74**	**0.74**
Hate Explanation avg		**0.73**	**0.70**	**0.72**	**0.65**	**0.69**	**0.70**

Table 9: Inter-annotator agreement measured using Krippendorff's α across cultures and evaluation dimensions, based on annotations from three native annotators per culture.

ferent Target, Intent, and Act categories and their corresponding values. Annotators were informed beforehand that the task involved exposure to potentially harmful hate speech content. All annotators completed a mandatory calibration round on ten held-out examples before proceeding to the main annotation task. Disagreements during calibration were discussed collectively to align understanding of the annotation criteria. For each evaluated pair (M_i, C_j), the annotators were shown the meme image, the meme caption, and the Hate Frame produced by the system under evaluation. They were asked to independently evaluate the frame using the criteria described in Section 6.

G.3 Inter Annotation Agreement Details

We measure inter-annotator agreement using Krippendorff's α (Krippendorff, 2011), computed with nominal weighting for binary dimensions (Target, Intent, Act, Coherence) and ordinal weighting for Likert dimensions (Factual Accuracy, Meme Relevance, Plausibility). Table 9 reports Krippendorff's α for each evaluation dimension across all cultures. Overall, the results demonstrate substantial inter-annotator agreement and consistent annotation reliability for evaluating culturally grounded hate frames. Agreement is highest for Factual Accuracy and Coherence, indicating strong consensus among annotators regarding the correctness of the retrieved cultural knowledge and the consistency of the generated explanations. Among the HDE dimensions, Target and Act achieve higher agreement than Intent, suggesting that identifying the underlying hateful intention of a meme is comparatively more subjective and culturally nuanced.

Definition of Hate Speech:
Any kind of communication that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.
Evaluate culturally grounded *Hate Frames* generated for memes. Each Hate Frame represents the system's interpretation of whether a meme is hateful within a specific cultural context.
You are given:
1. The meme
2. The meme caption
3. The generated Hate Frame.

The Hate Frame contains the predicted hate label, identified Hate Defining Elements (HDEs), retrieved cultural knowledge, a hate explanation, and an alternative interpretation if the meme is predicted as non-hateful.
Hate Defining Elements (HDEs)
A Hate Defining Element (HDE) consists of three components:
1. Target: The social group, identity, or attribute toward which hate is directed. Target values are short phrases identifying the group or characteristic being negatively portrayed in the meme. Examples of target values include *immigrants*, *Asians*, *lower socioeconomic background, etc.*
2. Intent: The underlying hateful purpose or social function of the meme. Intent values are short descriptive phrases associated with an intent category from the HDE taxonomy. Examples include *workers from lower socioeconomic backgrounds are portrayed as socially undesirable*, *immigrants are framed as outsiders*.
3. Act: The observable way through which hate is expressed in the meme. Act values are short descriptive phrases associated with an act category from the HDE taxonomy. Examples include *implying Kevin and Yaji are lower-status workers* and *mocking immigrants as culturally incompatible*.
{HDE Taxonomy}
Evaluation Criteria:
•**Criterion 1: Target Value Correctness** Evaluate whether the generated Target value correctly identifies the social group, identity, or attribute being negatively portrayed in the meme within your culture.
Use binary labels: Yes / No.
•**Criterion 2: Intent Value Correctness** Evaluate whether the generated Intent value correctly captures the underlying hateful purpose or social function of the meme within your culture. The Intent value should plausibly describe what harmful social meaning the meme conveys about the target group.
Use binary labels: Yes / No.
•**Criterion 3: Act Value Correctness** Evaluate whether the generated Act value correctly captures how the hateful meaning is expressed in the meme. The Act value should describe the observable insulting, stereotyping, exclusionary, or harmful expression mechanism used in the meme.
Use binary labels: Yes / No.
•**Criterion 4: Cultural Knowledge Factual Accuracy** Evaluate whether the retrieved cultural knowledge accurately reflects real historical, social, political, or cultural information from your culture.
Use a 3-point Likert scale:
Accurate (3), Partially Accurate (2), Inaccurate (1).
•**Criterion 5: Cultural Knowledge Meme Relevance** Evaluate whether the retrieved cultural knowledge meaningfully supports the hateful interpretation of the meme within your culture.
Use a 3-point Likert scale:
Directly Relevant (3), Tangentially Relevant (2), Not Relevant (1).
•**Criterion 6: Hate Explanation Plausibility** Evaluate whether the generated hate explanation plausibly explains why the meme is interpreted as hateful within your culture using the provided HDE values and cultural knowledge. Consider whether the explanation is culturally convincing or whether it overreaches beyond what the meme and cultural knowledge reasonably support.
Use a 3-point Likert scale:
Plausible (3), Somewhat Plausible (2), Implausible (1).
•**Criterion 7: Hate Explanation Coherence (HEC)** Evaluate whether the generated hate explanation is logically consistent, understandable, and internally coherent.
Use binary labels: Coherent / Incoherent.

Figure 20: Annotation Guidelines for Human evaluation of Hate Frames.