A THE USE OF LLM

We used a large language model (ChatGPT, OpenAI) solely for English copyediting, including grammar correction, wording and minor stylistic re-writes, and occasional LaTeX formatting help. The model was not used for idea generation, literature search, data collection/annotation, coding, analysis, or producing results. All scientific claims and contributions were written and verified by the authors, and no non-public data were shared with the model. The authors assume full responsibility for the content of the paper.

B DATA DETAILS

B.1 DATASET OVERVIEW

We study four datasets covering three evaluation tasks. **FHM** and **MultiOff** are annotated only for **Task 1**, whereas **HarM-c** and **HarM-p** additionally support **Task 2** and **Task 3**. Dataset sizes (train/test) are listed in Table 3.

Dataset	Train Samples	Test Samples
FHM	8,500	1,000
MultiOff	445	149
HarM-c	3,013	354
HarM-p	2,938	355

Table 3: Dataset statistics (train/test) used in our experiments.

Per-task label definitions are as follows:

Task 1 (harmfulness, binary): Decide whether a meme conveys harmful content considering image–text interaction (e.g., slurs, dehumanizing metaphors, threats or disparagement). Labels: 0=*not harmful*, 1=*harmful*.

Task 2 (harmfulness level, ordinal): Grade the severity of harmfulness when present; 1 reflects mild/implicit disparagement or sarcasm, while 2 indicates explicit or strongly implied hostility or incitement. Labels: 0=*not harmful*, 1=*somewhat harmful*, 2=*very harmful*.

Task 3 (influence target type): Identify the primary target of the harmful implication of the meme conditioned on harmfulness. If Task 1 = 0 or Task 2 = 0, set label of Task 3 to 0 (*none*); Labels: 0=*none*, 1=*individual*, 2=*organization*, 3=*community*, 4=*society*.

Dataset	**Task 1: harmful or Not**		**Task 2: Harmful level**			**Task 3: influence target type**				
	0	1	0	1	2	0	1	2	3	4
FHM	510	490	–	-	-	-	-	-	-	-
MultiOff	91	58	–	-	-	-	-	-	-	-
HarM-c	230	124	230	21	103	230	119	0	3	2
HarM-p	184	171	184	159	12	184	81	69	15	6

Table 4: Label distribution for each dataset across the three tasks.

B.2 DATASET-SPECIFIC DESCRIPTIONS AND PROMPT DESIGN

FHM provides a large-scale set of memes. It has 8,500 training examples and 1,000 test examples. and its goal is a binary classification of whether a meme is hateful.

Definition
A meme is considered hateful if it contains a direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. We define attack as violent or dehumanizing (comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. The mocking of hate crimes is also considered hate speech.

Instruction
Please analyze the provided image and respond ONLY with a JSON object in the following format:
{"is_hateful": <boolean>}
"is_hateful": Set to true if the image is harmful according to the definition, otherwise false.

Figure 6: Definition and instructions for FHM.

MultiOff contains 445 train and 149 test memes collected from Kaggle (2016 US election) and social media platforms, reflecting real-world political memes and emphasizing both image and text as context.

Definition
A meme is considered offensive if it contains content that is likely to offend, insult, or upset people based on factors such as race, ethnicity, religion, gender, sexual orientation, political views, or other personal characteristics. Offensive content may include offensive speech, discriminatory language, vulgarity, or content that mocks or belittles specific groups or individuals.

Instruction
Please analyze the provided image and respond ONLY with a JSON object in the following format:
{"is_offensive": <boolean>}
"is_offensive": Must be `true` if the image is offensive according to the definition, otherwise `false`.

Figure 7: Definition and instructions for MultiOff.

HarM includes two datasets: Harm-C (3,544 memes) and Harm-P (3,552 memes) are real-world meme datasets on COVID-19 and US politics, collected from web and social media, deduplicated and filtered, and annotated by experts for harm intensity and target type (individual, organization, community, society).

Definition
A multimodal unit consisting of an image and embedded text that has the potential to cause harm to an individual, an organization, a community, or society.

Instruction
Please analyze the provided image and respond ONLY with a JSON object in the following format:
{"harm_level": <string>, "target_type": <string>}
"harm_level": Must be one of ["not harmful", "somewhat harmful", "very harmful"].
"target_type": Must be one of ["individual", "organization", "community", "society"].
If "harm_level" is "not harmful", then "target_type" should be "none".

Figure 8: Definitions and instructions shared by HarM-c and HarM-p.

C Hyperparameter in Retrieval Stage

In the retrieval stage, we combine three modalities: image embeddings, elements, and generated textual descriptions. Let $\mathbf{v}$, $\mathbf{s}$, and $\mathbf{t}$ denote their respective representations. We form a fused vector

$$\mathbf{r} = \lambda_0 \mathbf{v} + \lambda_1 \mathbf{s} + \lambda_2 \mathbf{t}, \quad \text{with } \lambda_0 + \lambda_1 + \lambda_2 = 1,$$

and use $\mathbf{r}$ to compute retrieval similarity.

Table 5 lists five representative settings. **S1** uses vision only; **S2** and **S3** combine vision with symbolic or textual cues; **S4** assigns approximately equal weights to all modalities; **S5** relies purely on non-visual semantics (element + description).

Empirically, **S5** attains the best accuracy (71.75), while **S4** is a strong second (70.34). We hypothesize that harmfulness is primarily conveyed by semantic information including who is targeted, how,

and in what context, which are better captured by symbolic parses and generated descriptions than by appearance alone. Manual inspection of retrieved memes supports this view: memes that are visually similar can encode opposite meanings due to overlaid text or subtle symbolic cues, which can mislead vision-dominant retrieval (e.g., **S1**–**S3**). Overall, down-weighting the image channel (small λ_0) while allocating most weight to λ_1 and λ_2 tends to yield stronger retrieval for downstream reasoning.

Table 5: Parameter settings for associative retrieval with three modalities. λ_0, λ_1, and λ_2 denote the weights of image, symbolic elements, and generated descriptions, respectively.

Setting ID	λ_0 (Image)	λ_1 (Elements)	λ_2 (Description)	ACC
S1	1.0	0.0	0.0	69.77
S2	0.5	0.5	0.0	68.93
S3	0.5	0.0	0.5	67.80
S4	0.33	0.33	0.34	<u>70.34</u>
S5	0	0.5	0.5	**71.75**

D Scaling Language Models

We conducted a controlled study to examine how scaling LLM affects performance within our framework. Here, the VLM is kept fixed as the perceptual backbone, while the LLM responsible for reasoning is varied in size. To further test the generality of our approach, in addition to Qwen2.5-VL-7B used in the main experiments, we also include InternVL3.5-8B, another VLM that performs well in baseline evaluations. Table 6 reports the results on the HarM dataset across all three tasks.

We observe that introducing even a 7B LLM substantially improves performance over the VLM-only baseline, highlighting the effectiveness of symbolic reasoning. While model size has limited impact on the relatively easier binary harmful classification task, larger LLMs consistently yield gains on more fine-grained tasks. In particular, the 14B and 32B models provide notable improvements on harmfulness level prediction and target type identification, indicating that stronger reasoning capacity better supports nuanced inference.

Method		Harmful or Not			Harmfulness Level	Target type
VLM	LLM	ACC	BACC	MCC	M-F1	M-F1
Qwen2.5VL-7B	–	67.23	53.97	18.23	35.92	3.82
Qwen2.5VL-7B	Qwen2.5-7B	67.80	67.78	34.13	<u>47.68</u>	11.02
Qwen2.5VL-7B	Qwen2.5-14B	**70.34**	**71.78**	**41.59**	**51.90**	<u>15.94</u>
Qwen2.5VL-7B	Qwen2.5-32B	<u>68.08</u>	<u>70.79</u>	<u>39.73</u>	46.98	**16.01**
InternVL3.5-8B	–	64.41	69.82	38.99	42.86	12.25
InternVL3.5-8B	Qwen2.5-7B	**68.93**	69.21	36.83	47.63	12.67
InternVL3.5-8B	Qwen2.5-14B	67.80	**72.62**	**43.93**	**50.82**	<u>16.70</u>
InternVL3.5-8B	Qwen2.5-32B	<u>68.36</u>	<u>71.19</u>	<u>40.52</u>	<u>48.63</u>	**17.27**

Table 6: Performance of API-based models on the HarM dataset across three tasks.

E Symbolic Reasoning Module Design

Section 3.4 describes our symbolic reasoning pipeline. Due to space limitations, the main text illustrates only the Translator. Here we present the visualizations of the remaining two components—the *Planner* and *Solver*—referenced in Section 3.4.

Prompt Large-Language Model

"Please derive a step-by-step plan using the First-OrderLogic rule for determining the conclusion based on the context."

Input

"Symbolic Definition": ["Attack(x): x is an attack based on characteristics", "Violent(x): x is violent speech or actions", "Dehumanizing(x): x compares people to non-human things"]
"Translated Premise": { "dataset_content": ["∀x (Attack(x) ↔ (Violent(x) ∨ Dehumanizing(x) ∨ Inferiority(x) ∨ Exclusion(x) ∨ Segregation(x) ∨ MockingHateCrime(x)))…"],
"related_img1": ["∀x (TwoPanelLayout(x) ∧ Man(x) ∧ WearsSuit(x) ∧ Bald(x) ∧ TextTop('its their character', x)…"],
"related_img2": ["∀x (SinglePanelLayout(x) ∧ Man(x) ∧ WearsSuit(x) ∧ HoldsMicrophone(x) ∧ Smiling(x)…"]}
"Translated Statement": {"query_meme": {"Man(x) ∧ WearsSuit(x) ∧ Man(y) ∧ Bald(y)…"},
"question": "ContainsHateSpeech(x)"}

Output

"Step by Step Plan": ["Step 1: [Local] - Content Extraction - Parse symbols from input: two male portraits; TextTop('its their character')…",
"Step 2: [Local] - Symbol Mapping - From Symbolic Definition and Premise: Attack(x) ↔ (Violent(x) ∨ Dehumanizing(x) ∨ Inferiority(x) ∨ Exclusion(x) ∨ Segregation(x)…",
"Step 3: [Local] - Protected Attribute - Detect MentionsProtectedAttr(x, color) from 'color'. Using: Modus Ponens.",
"Step 4: [Related-Meme] - Retrieve related_img* by cues {'character', 'color', 'skin', 'race'}; compute CommonHarmfulFeature as intersection/majority over…",
"Step 5: [Synthesis] - If CommonHarmfulFeature = ∅ and no harmful predicates hold locally, infer ¬Attack(x); else Attack(x)…",
"Step 6: [Decision] - Define ContainsHateSpeech(x) ↔ (Attack(x) ∧ MentionsProtectedAttr(x, _)); output Harm Level and Target Type. Using: Modus Ponens + Consistency Check."]

Figure 9: Illustration of the Planner Stage

Prompt Vision-Language Model

"Your task is to carry out a logical solution based on the Step-By-Step Plan and input image."

Input

"img": "img/42953.png"
"Symbolic Definition": ["Attack(x): x is an attack based on characteristics", "Violent(x): x is violent speech or actions", "Dehumanizing(x): x compares people to non-human things"]
"Translated Premise": { "dataset_content": ["∀x (Attack(x) ↔ (Violent(x) ∨ Dehumanizing(x) ∨ Inferiority(x) ∨ Exclusion(x) ∨ Segregation(x) ∨ MockingHateCrime(x)))…"],
"related_img1": ["∀x (TwoPanelLayout(x) ∧ Man(x) ∧ WearsSuit(x) ∧ Bald(x) ∧ TextTop('its their character', x)…"],
"related_img2": ["∀x (SinglePanelLayout(x) ∧ Man(x) ∧ WearsSuit(x) ∧ HoldsMicrophone(x) ∧ Smiling(x)…"]}
"Translated Statement": {"query_meme": {"Man(x) ∧ WearsSuit(x) ∧ Man(y) ∧ Bald(y)…"},
"question": "ContainsHateSpeech(x)"}
"Step by Step Plan": ["Step 1: Identify and analyze the symbolic definitions provided in the Symbolic Definition section to understand the domain concepts…",
"Step 2: …"]

Output

"img": "img/42593.png"
"label": 0,
"solution": ["Step 1: [Local] - Content Extraction - TwoPanel; Man(x), Man(y); TextTop('its their character')…",
"Step 2: [Local] - Protected Attribute - MentionsProtectedAttr(m, color) detected (skin color)…",
"Step 3: [Local] - Harm Predicates - EqualityMessage(m) ⇒ ¬Violent(m) ∧ ¬Dehumanizing(m) ∧ ¬Inferiority(m) ∧ ¬Exclusion(m)…",
"Step 4: [Related-Meme] - Majority over related_img* shows equality/anti-discrimination…",
"Step 5: [Synthesis] - Attack(m) ↔ (Violent ∨ Dehumanizing ∨ Inferiority ∨ Exclusion ∨ Segregation ∨ MockingHateCrime)…",
"Step 6: [Decision] - ContainsHateSpeech(m) = Attack(m) ∧ MentionsProtectedAttr(m, color) ⇒ false…"]

Figure 10: Process Visualization of covid_memes_5560 form HarM

F Process Visualization

To provide a more intuitive understanding of how our framework operates, we visualize several representative examples that illustrate the intermediate steps in both the retrieval and reasoning stages. The visualizations show how relevant memories are retrieved across different modalities and how the retrieved information is subsequently integrated into the reasoning process. These examples are intended to complement the quantitative results by offering a step-by-step view of the model's internal workflow.

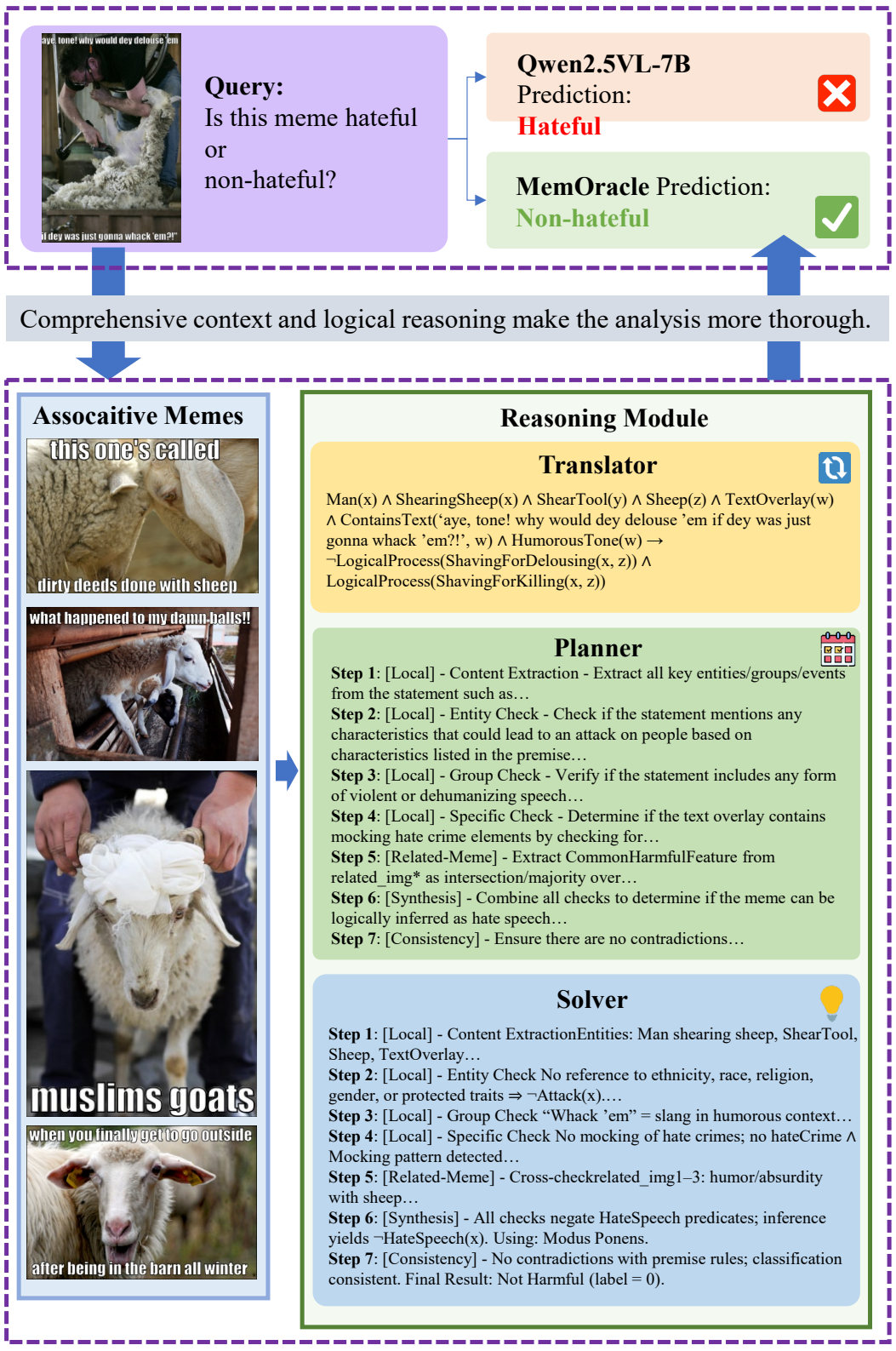

Figure 11: Process Visualization of 03164 form FHM

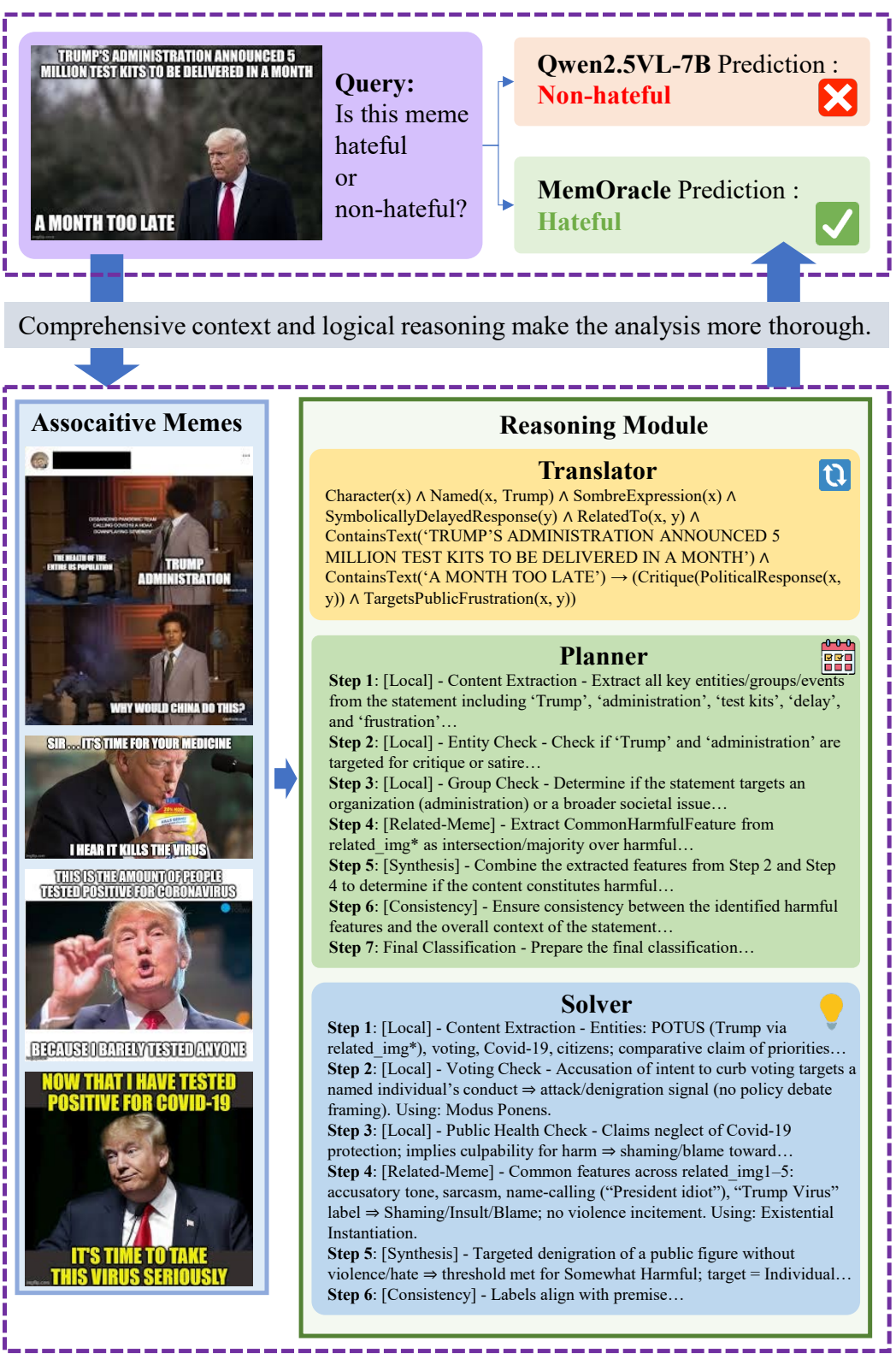

Figure 12: Process Visualization of covid_memes_5560 form HarM