AraHarMeme: A Multimodal Benchmark for Harmful Meme Detection in Arabic

Anonymous ACL submission

Abstract

In this work, we introduce AraHarMeme, a benchmark dataset for harmful content detection in Arabic memes. The dataset consists of 5,313 Arabic social media images collected from diverse platforms and manually annotated into three classes: harmful memes, non-harmful memes, and non-meme content. To the best of our knowledge, AraHarMeme is the first Arabic multimodal benchmark that explicitly addresses harmful meme detection. We conduct experiments using text-only, image-only, supervised multimodal models, as well as large vision-language models evaluated under zero-shot settings. The results show that Arabic-focused text models consistently outperform multilingual baselines, while image-only models are insufficient for reliable harmful content detection. Supervised multimodal models yield improvements when strong Arabic text encoders are combined with appropriate visual backbones, whereas zero-shot large vision-language models remain considerably less effective than supervised approaches on this dataset. The dataset and evaluation protocol will be publicly released to support future research in Arabic multimodal safety and content moderation.

Content Warning. This paper and the introduced dataset contain examples of harmful and offensive content (e.g., hate speech, harassment, or abusive imagery) included solely for the purpose of scientific research. Reader discretion is advised.

1 Introduction

Harmful content on social media remains a persistent challenge for automated moderation systems, particularly as abusive or hateful messages are increasingly expressed in indirect and context-dependent ways (Kiela et al., 2020). While much prior work has focused on textual abuse and hate speech (Talat and Hovy, 2016), a growing portion of harmful online content is conveyed through images combined with short textual overlays.

Internet memes are a common example of such content. Memes often rely on the interaction between image and text, as well as shared cultural knowledge, to convey meaning. As a result, harmful intent may not be apparent from either modality alone. This observation has motivated multimodal benchmarks for English, most notably the Hateful Memes dataset (Kiela et al., 2020), which demonstrated that unimodal approaches are insufficient for reliable detection. More recent work has further examined multimodal and large vision-language models for meme-based safety tasks, revealing mixed performance despite strong general multimodal capabilities (Alayrac et al., 2022; Liu et al., 2023).

In contrast, harmful meme detection in Arabic remains largely unexplored. Arabic social media content is characterized by extensive dialectal variation, frequent code-switching, and non-standard orthography, all of which complicate automatic analysis (Antoun et al., 2020; Abdul-Mageed et al., 2021). Existing Arabic datasets for abusive or hateful content focus almost exclusively on text-only settings and do not account for the multimodal structure of memes. Although recent Arabic multimodal resources have begun to emerge (Alwajih et al., 2025), none explicitly target harmful meme detection as a content moderation task.

To address this gap, we introduce **AraHarMeme**, a multimodal benchmark for harmful meme detection in Arabic. The dataset consists of 5,313 images collected from social media and annotated into three categories: harmful memes, non-harmful memes, and non-meme content. The inclusion of a non-meme class reflects realistic moderation scenarios in which visually similar but semantically unrelated images are common.

We accompany AraHarMeme with a standardized evaluation protocol and baseline experi-

ments spanning unimodal, multimodal, and large language-vision models. Together, these components provide a foundation for studying harmful meme detection in Arabic and for analyzing the limitations of current modeling approaches in a dialect-rich, multimodal setting. The main contributions are summarized as follows:

- We introduce **AraHarMeme**, a multimodal Arabic benchmark explicitly designed for harmful meme detection.
- We provide manually curated annotations covering harmful, non-harmful, and non-meme content.
- We establish strong baseline results, highlighting key challenges in Arabic harmful meme detection under multimodal settings.
- We assess the quality and stability of the benchmark through cross-dataset evaluation on an existing Arabic meme dataset (AraMeme) (Alam et al., 2024), analyzing the consistency of model performance across related tasks.
- We will publicly release the dataset and evaluation protocol to support future research on Arabic multimodal content moderation.

2 Related Work

Early research on harmful content detection focused primarily on text-based settings, addressing hate speech, abusive language, and offensive content on social media platforms (Talat and Hovy, 2016; Davidson et al., 2017). While effective for explicit abuse, text-only approaches struggle when harmful intent is implicit or context-dependent.

Memes present a particularly challenging case. As multimodal artifacts combining images with short, informal text, memes often convey meaning through interactions between modalities rather than through either modality alone. This has motivated growing interest in multimodal approaches to harmful content detection.

2.1 Multimodal Harmful Meme Detection

The Hateful Memes benchmark demonstrated that unimodal models are insufficient for reliable meme-based harm detection, establishing multimodal reasoning as a core requirement for this task (Kiela et al., 2020). Subsequent work explored multimodal fusion architectures that combine visual encoders with pretrained language models, generally outperforming text-only and image-only baselines (Goel and Poswal, 2024; Arya et al., 2024).

More recent studies have examined additional aspects of harmful meme detection, including explainability and robustness. Lin et al. (Lin et al., 2024) propose explainable multimodal models that expose conflicting visual and textual evidence, while evaluations of large vision–language models such as Flamingo and LLaVA show mixed results for harmful meme classification despite strong performance on general multimodal tasks (Alayrac et al., 2022; Liu et al., 2023). Recent analyses further highlight persistent safety limitations of large multimodal models in real-world meme settings (Lee et al., 2025). To address robustness concerns, retrieval-augmented and sentiment-aware approaches have been proposed, reporting improvements on hateful or harmful meme benchmarks (Mei et al., 2025; Duan et al., 2025). Most of this work, however, focuses on English data.

2.2 Arabic and Low-Resource Multimodal Benchmarks

Arabic NLP research has largely concentrated on text-only tasks such as sentiment analysis and hate speech detection, reflecting challenges related to dialectal variation and non-standard orthography (Pramanick et al., 2021). Multimodal resources for Arabic have only recently begun to emerge. AraMeme introduces an Arabic meme dataset focused on propagandistic content but is not designed for harmfulness detection as a moderation task (Alam et al., 2024).

More recent efforts expand Arabic multimodal coverage beyond memes, including culturally grounded instruction datasets and vision–language benchmarks that highlight performance gaps across dialects (Alwajih et al., 2025). Community-driven shared tasks have also released multimodal Arabic datasets for classification and propaganda-related challenges (Hasanain et al., 2024). Despite this progress, there is still no benchmark explicitly designed for harmful meme detection in Arabic that supports systematic evaluation across text-only, image-only, supervised multimodal, and zero-shot vision–language models. This gap motivates the introduction of AraHarMeme.

Figure 1: Example of an Arabic harmful meme requiring multimodal reasoning. The harmful intent arises from the interaction between the overlaid Arabic text and the image, while neither modality alone is sufficient to determine harmfulness.

3 AraHarMeme Benchmark

AraHarMeme is designed to reflect key linguistic, cultural, and orthographic characteristics of Arabic social-media discourse while capturing the multimodal nature of harmful content in memes. Figure 1 illustrates a representative instance in which harmful intent arises from the interaction between overlaid Arabic text and the image, while neither modality alone is sufficient to determine harmfulness. In this example, the text, glossed as "*Blacks are quiet*", can be interpreted as a descriptive statement in isolation, and the image alone depicts a person holding a machete-like object without explicit harmful framing. When interpreted jointly, however, the combination gives rise to a harmful, derogatory implication.

Each data instance therefore consists of a meme image, OCR-extracted text, and a manually assigned harm label, supporting both binary and multi-class evaluation settings. Figure 2 summarizes the AraHarMeme dataset construction and annotation pipeline.

In addition to multimodal complexity, the dataset exhibits substantial linguistic diversity: as shown in Figure 3 (Appendix C), AraHarMeme covers a wide range of Arabic dialects, including Egyptian, Gulf, Levantine, and Maghrebi Arabic, as well as Arabizi and code-switched Arabic–English content. Orthographic variation is also pervasive in the OCR-extracted text (Figure 4, Appendix C), with frequent non-standard spelling, character elongation, emojis, and visually stylized fonts characteristic of Arabic memes.

3.1 Dataset Collection

The AraHarMeme dataset was collected from multiple publicly accessible online platforms, including Twitter/X, Facebook public pages, Telegram public channels and Reddit communities (e.g., r/ArabMemes). These platforms were selected due to their high prevalence of informal, dialectal, and visually grounded content.

Data collection relied on keyword-based crawling combined with image scraping. Seed keywords were manually curated to cover a broad range of harmful and non-harmful meme content commonly observed in Arabic social media. These included (i) harm-related terms (e.g., insults, harassment, and derogatory expressions), (ii) dialectal and colloquial variants of such terms across major Arabic dialects, and (iii) meme-related cues frequently used in Arabic memes. Representative examples include insult terms such as "كلب" (dog), "حمار" (donkey), and dialectal variants (e.g., "كلبك" ،"حماررر"), sarcastic or provocative expressions such as e.g., "اسكت" (shut up) and "عار" (shame), as well as colloquial spellings and elongations. To reduce bias toward Modern Standard Arabic, keywords were expanded to include Egyptian, Gulf, Levantine, and Maghrebi forms, Arabizi spellings (e.g., kalb (dog) , 7mar (donkey)), and common code-switched Arabic-English expressions. The complete list of keywords is provided in Appendix A.

3.2 Dataset Filtering

Following collection, a structural filtering stage was applied to remove low-quality or unsuitable content while preserving multimodal diversity. Near-duplicate images were identified and removed using perceptual hashing, retaining a single representative instance per duplicate cluster. Images were discarded if they were severely corrupted, extremely low-resolution, or visually irrelevant to meme discourse. Importantly, no filtering decisions at this stage were based on harmfulness, which was determined exclusively during manual annotation.

3.3 OCR Text Extraction

Text embedded in images was extracted using *Tesseract OCR (v5.3)*[1] with Arabic language models. OCR was applied uniformly to all images regardless of font style, text density, or visual lay-

[1] https://github.com/tesseract-ocr/tesseract

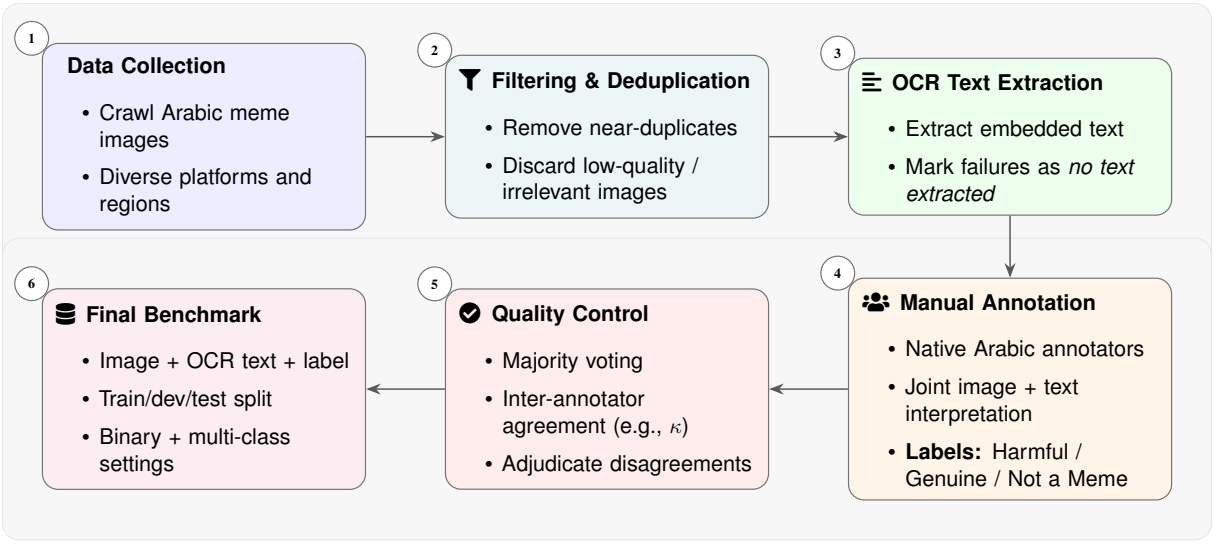

Figure 2: Overview of the AraHarMeme dataset construction and annotation pipeline.

out. We do not perform any automatic OCR post-processing or correction at this stage. In particular, we do not normalize spelling, remove noise, or manually edit OCR outputs during extraction. Instead, any necessary corrections are performed only during the manual annotation phase to ensure that the final text faithfully reflects what is visibly written on the meme.

An OCR result is considered *unsuccessful* if it:

1. contained only isolated characters or fragments that did not form interpretable words,
2. consisted of corrupted or nonsensical strings caused by stylized fonts, heavy noise, or low resolution, or
3. clearly did not correspond to any visible textual content in the image.

Instances meeting these criteria were labeled as *"no text extracted"*. This designation was assigned only after manual verification during annotation.

3.4 Manual Annotation

Instances were annotated manually by a team of three native Arabic speakers with academic training in Arabic linguistics and demonstrated expertise in regional Arabic dialects. Prior to annotation, annotators were provided with detailed written guidelines describing label definitions, examples of implicit harm, sarcasm, and culturally grounded references, as well as instructions for jointly interpreting visual and textual content.

Annotators were explicitly instructed to consider both modalities together when assigning labels. All harmfulness judgments were made exclusively at this stage and did not influence earlier filtering or OCR-related decisions. OCR-extracted text was verified and corrected when necessary during annotation, particularly in cases involving dialectal spelling, stylized fonts, or OCR noise. Corrections were strictly limited to faithfully reflecting the visible text in the image and did not involve reinterpretation, paraphrasing, or semantic modification.

The annotation schema consists of three classes:

- **Harmful**: Memes that convey offensive, abusive, harassing, hateful, or otherwise harmful content. Harmfulness may be expressed explicitly (e.g., direct insults or abusive language) or implicitly through image–text interaction, sarcasm, or contextual references.
- **Genuine**: Memes that are non-harmful and contain neutral, benign, or light humorous content without offensive or derogatory intent.
- **Not a Meme**: Images that lack meme-specific characteristics, including standalone photographs, decorative images, screenshots of unrelated content, or images without communicative or semantic intent.

3.5 Quality Control and Agreement

The dataset instances independently annotated by all three annotators. Final labels were determined

using majority voting. In cases where all three annotators disagreed, the instance was reviewed jointly and resolved through discussion in accordance with the annotation guidelines.

Inter-annotator agreement was measured using pairwise Cohens κ, yielding agreement scores of 0.72, 0.64, and 0.85 across annotator pairs, with an average $\kappa = 0.74$, indicating substantial agreement. This high level of agreement reflects consistent interpretation despite the presence of dialectal variation, implicit harm, and multimodal ambiguity.

3.6 Dataset statistics

The final AraHarMeme benchmark consists of **5,323 images**, split into 4,189 training instances and 1,134 test instances. The dataset includes 1,992 harmful memes, 2,039 genuine (non-harmful) memes, and 1,282 non-meme images, as summarized in Table 2.

Appendix C provides additional dataset statistics. Figure 3 shows the distribution of Arabic dialects, with most instances containing dialectal Arabic rather than Modern Standard Arabic. Figure 4 presents the character-level length distribution of OCR-extracted text, which varies substantially across instances due to informal writing styles and OCR-related noise, and Figure 5 reports the class distribution across the three labels.

3.7 Comparison with Prior Arabic Meme Datasets

Although multimodal content such as memes has been widely studied in high-resource languages like English, work on Arabic meme datasets is still emerging. One of the first resources to explicitly target Arabic meme understanding is ArMeme (Alam et al., 2024), a multimodal corpus of approximately 6,000 memes annotated for propagandistic content. ArMeme serves as an initial benchmark for multimodal research in Arabic and has been utilized in multimodal classification and shared tasks on propaganda detection (Haouhat et al., 2024; Shah et al., 2024). An extension of ArMeme, known as MemeXplain, augments the base dataset with bilingual explanations for the propaganda labels, enabling explainable multimodal classification research (Kmainasi et al., 2025).

Despite their value, these resources have limitations relative to AraHarMeme. ArMeme and MemeXplain are primarily geared toward detecting propagandistic content in memes, whereas AraHarMeme adopts a broader conception of harmful content, encompassing general offensive, abusive, and harmful themes beyond propaganda.

Other Arabic multimodal corpora, such as general visual-text pairs for sentiment analysis or cultural representation, exist but do not specifically address memes or harmful multimodal content relevant to content moderation, making AraHarMeme unique in its targeted task focus. Table 1 comparesAraHarMeme with existing Arabic multimodal meme datasets.

4 Experiments

4.1 Experimental Setup

All experiments are conducted on the AraHarMeme benchmark described in Section 3. Each instance consists of an image, OCR-extracted Arabic text (when available), and a manually assigned label from three classes: Harmful, Genuine, and Not a Meme. A fixed train-test split is used throughout all experiments. Test instances are strictly excluded from all training and fine-tuning procedures. The same test split is used for supervised learning and zero-shot evaluation to ensure consistent and comparable results across experimental settings. All supervised experiments are implemented using the HuggingFace Transformers library [2] and executed on a single GPU.

We report Accuracy, Precision, Recall, and F1-score for all supervised models. For zero-shot large language models, we additionally report macro-averaged Precision, Recall, and F1-score to better capture per-class behavior under imbalanced predictions.

4.2 Text-only Models

Text-only experiments are implemented using pretrained transformer models for sequence classification. We evaluate both **Arabic-specific** models, including ARBERT (Abdul-Mageed et al., 2021), MARBERT and MARBERTv2 (Abdul-Mageed et al., 2021), CamelBERT (Inoue et al., 2021), and Asafaya (Safaya et al., 2020), as well as **multilingual baselines** such as XLM-R (Conneau et al., 2020), mBERT (Devlin et al., 2019), and DistilMBERT (Sanh et al., 2019). All models are initialized from their publicly released pretrained checkpoints.

[2] https://huggingface.co

Table 1: Comparison of AraHarMeme with existing Arabic multimodal meme datasets (provided for reference).

Dataset	Size	Language Coverage	Modalities	Annotation Types	Primary Task(s)
ArMeme (Alam et al., 2024)	∼6K	Arabic (MSA + dialects)	Image+Text	Propaganda labels	Propaganda detection
MemeXplain (Kmainasi et al., 2025)	∼6K	Arabic (MSA + dialects)	Image+Text	Propaganda labels + explanations	Explainable propaganda classification
AraHarMeme (ours)	∼5K	Dialectal Arabic, Arabizi, code-switching	Image+Text	Harmful / Non-harmful / Not a Meme	Broad harmful meme detection

Statistic	Value
Total memes	5,313
Training set	4,189
Test set	1,134
Genuine (non-harmful)	2,039
Harmful	1,992
Not a meme	1,282

Table 2: Statistics of the AraHarMeme dataset.

Fine-tuning is performed using the HuggingFace `Trainer` API (Wolf et al., 2020). Models are trained for four epochs with a batch size of 16, using the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2×10^{-5}. A linear learning rate scheduler with warm-up over the first 10% of training steps is applied. We use standard cross-entropy loss for all classification experiments. No additional task-specific pretraining or data augmentation is applied. Evaluation is performed on the held-out test set using standard classification metrics.

4.3 Image-only Models

Image-only experiments are conducted using standard convolutional and transformer-based vision architectures, including ResNet-34 and ResNet-50 (He et al., 2016), ViT-Base (Dosovitskiy, 2020), DeiT-Base (Touvron et al., 2021), and ConvNeXt-Tiny (Liu et al., 2022). All models are initialized with ImageNet-pretrained weights.

Image models are trained for four epochs using the same optimizer, learning rate, and batch size as the text-based models to maintain a controlled comparison. Standard image preprocessing and resizing are applied. No textual or multimodal information is used in this setting.

4.4 Supervised Multimodal Models

Supervised multimodal experiments combine a pretrained text encoder with a visual backbone through a late fusion strategy. Textual and visual representations are extracted independently and concatenated before the final classification layer.

All multimodal models are trained using the same hyperparameters as unimodal models to ensure a controlled comparison: four epochs, batch size 16, the AdamW optimizer (Loshchilov and Hutter, 2017), and a learning rate of 2×10^{-5}. No modality-specific loss weighting, curriculum learning, or auxiliary supervision is applied. This design isolates the contribution of multimodal fusion without introducing additional training complexity.

4.5 Zero-shot Evaluation of Large Language–Vision Models

We evaluate a set of **Large Language–Vision Models (LLVMs)** in a zero-shot setting. All evaluated models are natively multimodal, i.e., capable of processing both visual and textual inputs. To better understand their behavior under different input conditions, we evaluate each model using three configurations: *text-only*, *image-only*, and *text–image*.

The evaluated models include **Gemma-3 (4B, 12B)**(Team et al., 2025), **LLaVA-7B**[3], **LLaVA-Phi3-3.8B**[4], **Moondream-1.8B**[5], **Qwen2.5-VL-3B**[6], and **Qwen3-VL-30B**[7]. These models are selected to cover a range of architectures and parameter scales representative of recent large language-vision systems.

All evaluations are conducted in a zero-shot manner. No task-specific fine-tuning, parameter updates, or in-context examples are used. Instead, models are prompted using a fixed instruction prompt that defines the task, the label space, and explicit decision rules. The prompt instructs

[3] https://llava-vl.github.io/
[4] https://ollama.com/library/llava-phi3:3.8b
[5] https://moondream.ai
[6] https://github.com/QwenLM/Qwen2.5-VL
[7] https://github.com/QwenLM/Qwen3-VL

the model to classify each input into one of three categories: *Harmful Arabic Meme*, *Not Harmful Arabic Meme*, or *Not a Meme / Not Arabic*. Models are explicitly instructed not to guess missing text and not to rely on external knowledge beyond the visible input. The full prompt used for all zero-shot experiments is provided in Appendix B.

In the *text-only* configuration, models receive only the OCR-extracted text when available. In the *image-only* configuration, only the image is provided. In the *text–image* configuration, both modalities are provided jointly. Model outputs are mapped deterministically to the target label set based on the predicted class number. Inference is performed using deterministic decoding to ensure reproducibility across runs.

5 Results and Discussion

This section analyzes the performance of supervised text-only, image-only, and multimodal models, as well as zero-shot large language–vision models, on the AraHarMeme benchmark. Results for supervised models are reported in Table 3, while zero-shot results are summarized in Table 4.

Text-only models. Among text-only approaches, Arabic-specific transformer models consistently outperform multilingual baselines. As shown in Table 3, ARBERT achieves the highest performance (Acc = 0.820, F1 = 0.819), followed closely by MARBERT (Acc = 0.818, F1 = 0.817). Other Arabic-focused models, including Asafaya and CamelBERT, show comparable performance, whereas multilingual models such as XLM-R, mBERT, and DistilMBERT perform noticeably worse. This gap indicates that Arabic-specific pretraining remains essential for modeling dialectal variation, informal usage, and non-standard orthography characteristic of Arabic social-media content.

Image-only models. Image-only models perform substantially worse than text-based classifiers. Although modern vision architectures such as ConvNeXt-Tiny and DeiT-Base achieve moderate performance (F1 around 0.72--0.73), they consistently lag behind text-only models. Both convolutional and transformer-based vision models exhibit limited recall, suggesting that visual information alone is often insufficient for identifying harmful intent, particularly in cases where harm is implicit or linguistically mediated.

Supervised multimodal models. Supervised multimodal models show more varied behavior. When strong Arabic text encoders are combined with appropriate visual backbones, multimodal fusion yields substantial gains over unimodal baselines. In particular, the ARBERT + ConvNeXt-Tiny model achieves the best overall performance (Acc = 0.951, F1 = 0.944), outperforming both text-only and image-only models. These improvements are reflected in balanced precision and recall, indicating that visual features can complement textual representations when linguistic modeling is sufficiently robust. In contrast, multimodal models built on MARBERT or multilingual encoders yield smaller or inconsistent gains, suggesting that multimodal fusion does not compensate for weaker linguistic representations.

Zero-shot large language–vision models. Table 4 reports zero-shot performance for large language–vision models evaluated under text-only, image-only, and multimodal input configurations. Across all models, zero-shot performance remains substantially lower than that of supervised text-based and multimodal approaches. Providing both image and text generally improves performance relative to unimodal inputs; however, the gains are limited. Qwen3-VL-30B achieves the strongest zero-shot results (Acc = 0.615, F1 = 0.579 in the multimodal setting), but still falls well short of fine-tuned Arabic-specific text models and supervised multimodal architectures. These results indicate that current large language–vision models are not competitive for Arabic harmful meme detection without task-specific adaptation.

To further assess the robustness of these observations, we evaluate the same subset of large language–vision models on the ArMeme dataset. The results, presented in Table 5 (Appendix D), show qualitatively similar behavior: the relative ranking of models is largely preserved, multimodal inputs generally help, and overall zero-shot performance remains limited. This cross-dataset consistency suggests that the performance patterns observed on AraHarMeme are not dataset-specific, but instead reflect broader limitations of current multimodal LLMs for Arabic meme understanding.

6 Conclusion

This paper introduced **AraHarMeme**, a multimodal benchmark for harmful meme detection in

Table 3: Performance comparison of text-only, image-only, and supervised multimodal models on the AraHarMeme benchmark. Macro-averaged metrics are reported for all models. Best results within each block are highlighted in bold.

Model	Acc	P	R	F1
Text-only Models				
ARBERT	**0.820**	**0.835**	**0.820**	**0.819**
MARBERT	0.818	0.835	0.818	0.817
Asafaya	0.808	0.822	0.808	0.807
CamelBERT	0.808	0.826	0.808	0.805
MARBERTv2	0.792	0.805	0.792	0.789
XLM-R	0.776	0.789	0.776	0.772
mBERT	0.742	0.753	0.742	0.736
DistilMBERT	0.748	0.758	0.748	0.744
Image-only Models				
ConvNeXt-Tiny	**0.729**	**0.726**	**0.729**	**0.727**
DeiT-Base	0.722	0.723	0.722	0.721
ResNet50	0.703	0.709	0.703	0.705
ViT-Base	0.687	0.693	0.687	0.689
ResNet34	0.686	0.679	0.686	0.676
Supervised Multimodal Models				
ARBERT + ConvNeXt-T	**0.951**	**0.947**	**0.941**	**0.944**
ARBERT + DeiT-B	0.895	0.887	0.879	0.883
MARBERT + DeiT-B	0.705	0.708	0.681	0.690
MARBERT + ConvNeXt-T	0.700	0.672	0.665	0.666
mBERT + DeiT-B	0.709	0.744	0.649	0.658
DistilMBERT + ConvNeXt-T	0.660	0.660	0.602	0.608
XLM-R + ConvNeXt-T	0.648	0.590	0.595	0.592
MARBERT + ViT-B	0.656	0.613	0.616	0.613
ARBERT + ResNet50	0.615	0.604	0.589	0.593

Table 4: Zero-shot performance of large language and multimodal LLMs on AraHarMeme. Macro-averaged metrics are reported. Best results within each block are in bold.

Model	Input	Acc	F1	P	R
Gemma-3 12B					
	Text	0.484	0.353	0.310	0.411
	Image	0.491	0.342	0.495	0.409
	Multi	**0.506**	**0.424**	**0.573**	**0.448**
Gemma-3 4B					
	Text	0.365	0.362	0.376	0.403
	Image	0.456	0.317	**0.538**	0.386
	Multi	**0.462**	**0.435**	0.454	**0.438**
LLaVA-7B					
	Text	**0.394**	**0.329**	**0.381**	**0.344**
	Image	0.261	0.242	0.320	0.329
	Multi	0.387	0.314	0.346	0.340
LLaVA-Phi3-3.8B					
	Text	0.354	0.269	0.330	0.330
	Image	**0.387**	0.248	0.219	0.342
	Multi	0.346	**0.270**	**0.394**	**0.345**
Moondream-1.8B					
	Text	**0.404**	0.194	**0.468**	**0.335**
	Image	0.307	**0.247**	0.210	0.328
	Multi	0.328	0.233	0.189	0.309
Qwen2.5-VL-3B					
	Text	0.317	0.260	0.209	0.378
	Image	0.305	**0.288**	**0.472**	**0.400**
	Multi	**0.378**	0.284	0.238	0.353
Qwen3-VL-30B					
	Text	0.595	0.548	0.589	0.547
	Image	0.553	0.502	0.615	0.507
	Multi	**0.615**	**0.579**	**0.633**	**0.569**

Arabic, designed to reflect the linguistic diversity, cultural context, and implicit multimodal nature of Arabic social-media content. Extensive experiments show that fine-tuned Arabic-specific text models substantially outperform multilingual baselines, that image-only models are insufficient for reliable harmfulness detection, and that supervised multimodal models are effective only when paired with strong Arabic-aware linguistic representations. In contrast, zero-shot large language–vision models remain markedly less competitive, even when provided with both image and text.

Future work will explore larger-scale Arabic multimodal pretraining, more fine-grained harm taxonomies, and improved cross-modal fusion strategies that better capture implicit and culturally grounded harmful content. We release AraHarMeme and the accompanying evaluation protocol to support further research on Arabic multimodal safety and content moderation.

Ethical Considerations

AraHarMeme consists exclusively of publicly available meme images and does not contain any user-identifying information. No private accounts, restricted groups, or personal metadata were accessed or collected, and the dataset therefore poses minimal privacy risk.

As with any dataset involving harmful or offensive content, the annotations reflect human judgment and may encode cultural or societal biases. We provide clear annotation guidelines and encourage users of the dataset to remain aware of these limitations when developing or deploying models. Models trained on AraHarMeme may support research on Arabic content moderation and contribute to tools for analysts, journalists, and platform moderators. However, we caution against using such models as fully automated decision-

makers without human oversight.

Limitations

AraHarMeme was collected from publicly accessible social media platforms and reflects the variability of real-world Arabic meme content. As a result, the benchmark is sensitive to OCR quality. Although OCR outputs were manually reviewed during annotation, stylized fonts, low-resolution images, emojis, and complex layouts may introduce noise in fully automated settings, potentially affecting model performance.

In addition, harmfulness annotation is inherently subjective. While annotations were performed by native Arabic speakers with familiarity across dialects, interpretations of offensive or harmful content can vary across regions and communities, which may limit generalizability.

Use of AI Assistance

In the preparation of this work, we used AI-assisted tools to support code development (e.g., GitHub Copilot) and for limited language-related assistance, such as spelling checks and minor stylistic refinements (e.g., Grammarly and ChatGPT). All scientific content, analysis, and conclusions were produced and verified by the authors.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Firoj Alam, Abul Hasnat, Fatema Ahmad, Md Arid Hasan, and Maram Hasanain. 2024. Armeme: Propagandistic content in arabic memes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21071--21090.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716--23736.

Fakhraddin Alwajih, Samar M. Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, Baraah Qawasmeh, Houdaifa Atou, Serry Sibaee, Hamzah A. Alsayadi, Walid Al-Dhabyani, Maged S. Al-shaibani, Aya El aatar, Nour Qandos, Rahaf Alhamouri, and 18 others. 2025. Pearl: A multimodal culturally-aware Arabic instruction dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23048--23079, Suzhou, China. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9--15, Marseille, France. European Language Resource Association.

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEe Access*, 12:22359--22375.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440--8451, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171--4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yuxiao Duan, Xiang Zhao, and Hao Guo. 2025. Sentiment-aware cross-modal semantic interaction model for harmful meme detection. *Decision Support Systems*, page 114509.

Aman Goel and Abhishek Poswal. 2024. Mmhs: Multimodal model for hate speech intensity prediction. In *International Conference on Speech and Computer*, pages 95--108. Springer.

Abdelhamid Haouhat, Hadda Cherroun, Slimane Bellaouar, and Attia Nehar. 2024. Modos at araieval shared task: Multimodal propagandistic memes classification using weighted sam, clip and arabiangpt. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 483--488, Bangkok, Thailand. Association for Computational Linguistics.

Maram Hasanain, Md. Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md. Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024. Araieval shared task: Propagandistic techniques detection in unimodal and multimodal arabic content. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, pages 456--466. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770--778.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92--104, Kyiv, Ukraine (Virtual). 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.

Mohamed Bayan Kmainasi, Abul Hasnat, Md Arid Hasan, Ali Ezzat Shahroor, and Firoj Alam. 2025. MemeIntel: Explainable detection of propagandistic and hateful memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30263--30279, Suzhou, China. Association for Computational Linguistics.

DongGeon Lee, Joonwon Jang, Jihae Jeong, and Hwanjo Yu. 2025. Are visionlanguage models safe in the wild? a meme-based benchmark study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 30533--30576. Association for Computational Linguistics.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359--2370.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892--34916.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976--11986.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. 2025. Robust adaptation of large multimodal models for retrieval augmented hateful meme detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23806--23828, Suzhou, China. Association for Computational Linguistics.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439--4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054--2059, Barcelona (online). International Committee for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Uzair Shah, Md. Rafiul Biswas, Marco Agus, Mowafa Househ, and Wajdi Zaghouani. 2024. Mememind at araieval shared task: Generative augmentation and feature fusion for multimodal propaganda detection in arabic memes. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 467--472, Bangkok, Thailand. Association for Computational Linguistics.

Zeerak Talat 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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé

Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347--10357. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38--45.

A Keyword-Based Crawling

Seed keywords were manually curated and grouped into categories to guide keyword-based crawling of Arabic memes. The list below provides representative examples used during data collection.

- **Harm-related insults (MSA and colloquial):**
  كلب، حمار، وسخ، قذر، غبي، حقير، نجس، متخلف، حثالة، زبالة، ساقط
- **Threatening or silencing expressions:**
  اسكت، اخرس، اطبق فمك، هدد، اقتلك، اضربك، دمرك، سحقك، نهايتك، انتبه لنفسك
- **Derogatory and humiliating phrases:**
  عار، عيب، فضيحة، مهزلة، مسخرة، ضحكة، عيب عليك، وصمة، خزي
- **Sarcastic and provocative cues:**
  برافو عليك، شاطر، عبقري زمانك، يا سلام، ذكي جداً، مثال يحتذى به
- **Dialectal variants and regional slang:**
  - **Egyptian:**
    حيوان، أهبل، عبيط، معفن، ولا مؤاخذة
  - **Levantine:**
    حمارك، غبيان، وسخين، بلا مخ
  - **Gulf:**
    تافه، فاشل، هطف، ما تسوى
  - **Maghrebi:**
    حماررر، زبلة، بهلول، مسخ
- **Orthographic variations and elongation:**
  حمااار، كلبيب، وسخخخ، غبيبي، قذذذر
- **Arabizi spellings (Latin-script Arabic):**
  kalb, 7mar, weskh, ghabi, wes5, 3ayb, 5ara, enta 7mar
- **Code-switched Arabic–English expressions:**
  stupid عرب dirty ناس crazy ناس idiot عربي ناس stupid كلام trash
- **Meme-related discourse markers:**
  لما، لما تكون، لما تشوف، هذا لما، توقع، تخيل، القصة باختصار (POV)

The keyword list was iteratively expanded to include dialectal spellings, colloquial usage, and non-standard orthography commonly observed in Arabic social-media memes. These keywords were used solely to guide crawling and candidate collection; final inclusion and labeling decisions were performed through manual annotation.

B Zero-shot Classification Prompt

This appendix presents the exact instruction prompt used for zero-shot evaluation in this work. The prompt was originally designed for multimodal input, where models receive both an image and OCR-extracted text. The same prompt was applied verbatim in text-only and image-only settings by providing only the available modality and omitting the missing one. No task-specific fine-tuning, in-context examples, or external knowledge were used.

Classification Instructions

You will receive an image and text. Classify the input into one of the following categories.

1 -> **Harmful Arabic Meme:** The image is a meme. The visible text is written exclusively in Arabic or Arabic dialects (e.g., Moroccan, Egyptian, Algerian, Tunisian). The content contains insults, profanity, hate, discrimination, mockery, or degrading humor.

2 -> **Not Harmful Arabic Meme:** The image is a meme. The visible text is written exclusively in Arabic or Arabic dialects. The content is neutral, harmless, or contains light humor.

3 -> **Not a Meme / Not Arabic:** The image is not a meme, contains no readable text, or the visible text is predominantly non-Arabic and unrelated to Arabic meme discourse.

Rules: If any non-Arabic text is present, output 3. Do not guess missing text. Do not use external knowledge.

Extracted text: {caption}

Output format: <class_number> <explanation>

C Dataset Statistics and Distributions

This appendix provides additional statistics and distributional visualizations for the AraHarMeme benchmark, complementing the dataset description in Section 3.

D Additional Experiments on ArMeme

We further evaluate the curated benchmark by testing a subset of large language models on the existing ArMeme dataset. Table 5 reports the zero-shot performance of this subset of LLMs.

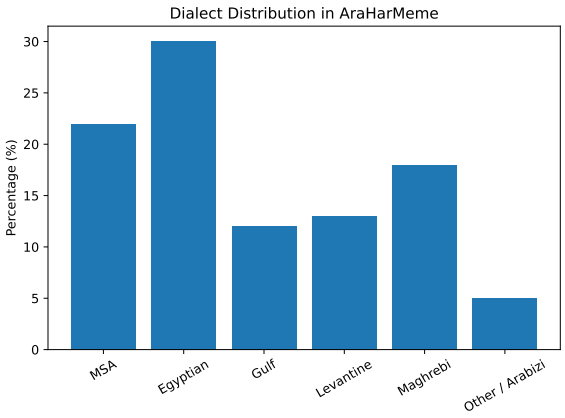

Figure 3: Dialect distribution in the AraHarMeme benchmark.

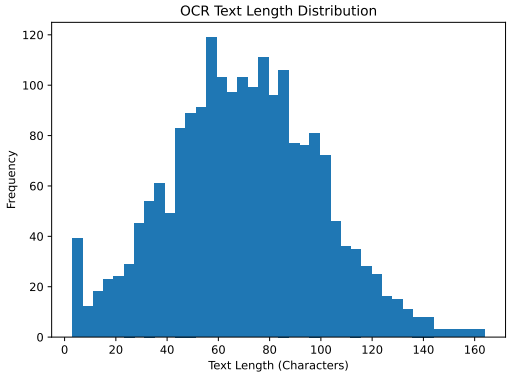

Figure 4: Character-level text length distribution of OCR-extracted meme text.

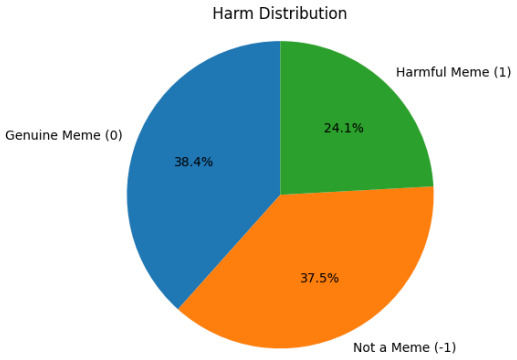

Figure 5: Label distribution across the three annotation categories.

Table 5: Zero-shot performance of selected LLMs on ArMeme. Macro-averaged metrics are reported. Best results within each model block are highlighted in bold.

Model Input	Acc	F1	P	R
Gemma-3 12B				
Text	0.123	0.136	**0.489**	0.271
Image	0.237	0.138	0.179	0.216
Multi	**0.278**	**0.229**	0.432	**0.343**
Gemma-3 4B				
Text	0.258	0.115	0.231	0.250
Image	**0.655**	0.200	0.193	0.250
Multi	0.499	**0.238**	**0.236**	**0.258**
LLaVA 7B				
Text	0.397	0.206	0.225	0.247
Image	**0.655**	0.198	0.164	0.250
Multi	0.580	**0.239**	**0.228**	**0.254**
LLaVA-Phi3 3.8B				
Text	0.438	0.222	0.228	0.254
Image	**0.635**	0.228	**0.244**	0.257
Multi	0.470	**0.232**	0.232	**0.260**
Moondream 1.8B				
Text	**0.658**	0.198	0.164	0.250
Image	0.621	**0.222**	**0.219**	0.249
Multi	0.489	0.226	0.220	**0.240**
Qwen2.5-VL 3B				
Text	0.303	0.156	0.230	0.244
Image	**0.604**	**0.240**	**0.233**	**0.257**
Multi	0.354	0.186	0.223	0.250
Qwen3-VL 30B				
Text	0.623	0.264	0.252	0.278
Image	0.627	0.243	0.242	0.261
Multi	**0.646**	**0.268**	**0.267**	**0.281**