

TEAM_CHEN@CC-MMD 2026: Culture-Conditioned Prompting for Zero-Shot Misogynistic Meme Detection — Asymmetric Shifts and a Feminist Probe

Yumeng Chen

School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
ec25993@qmul.ac.uk

Arkaitz Zubiaga 

School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
a.zubiaga@qmul.ac.uk

Abstract

CC-MMD 2026 Task B evaluates cross-cultural misogynistic meme detection. We use a fixed zero-shot VLM, Qwen2.5-VL-7B-Instruct, and vary the cultural framing of the prompt. On the English test split, positive prediction rates shift from 60.2% under a generic prompt to 61.9% under an Indian prompt, 55.0% under a Chinese prompt, and 83.8% under an Indian feminist prompt. Agreement is 78.1% between the Indian cultural and feminist settings, and 69% between the Indian prompt and a similar variant on a 100-sample subset. A Chinese sanity check shows that cross-split reuse is unreliable, with only 64% agreement. We provide a reproducible zero-shot audit framework and show that feminist prompting can serve as an effective probe of intra-cultural decision sensitivity.

CCS Concepts

• **Computing methodologies** → **Computer vision**; *Natural language processing*; • **Information systems** → **Multimedia information systems**; • **Social and professional topics** → *Hate speech*; Cultural characteristics.

Keywords

cross-cultural robustness; misogyny detection; meme detection; vision-language models; zero-shot prompting; cultural bias audit

ACM Reference Format:

Yumeng Chen and Arkaitz Zubiaga. 2026. TEAM_CHEN@CC-MMD 2026: Culture-Conditioned Prompting for Zero-Shot Misogynistic Meme Detection — Asymmetric Shifts and a Feminist Probe. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26)*, October 05–09, 2026, Napoli, Italy. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3776574.3832464>

1 Introduction

Misogynistic memes are difficult to detect because they often encode harmful meaning implicitly through the interaction of image, text, humour, irony, and culture-specific shorthand, rather than through explicit abusive language alone. Prior work has shown that harmful intent is frequently hidden behind subtle or latent

forms of abuse, making annotation and automatic detection difficult [5, 10, 20]. In cross-cultural settings, this challenge becomes even greater, because the same content may be interpreted differently across communities; recent work has documented substantial disagreement in both multimodal and text-only harmful-content annotation [3, 8].

CC-MMD 2026 Task B addresses this problem by evaluating cross-cultural misogynistic meme detection across English, Chinese, Tamil, and Malayalam, with robustness targets spanning Indian, Chinese, and Western (Ireland) cultural contexts [4, 13]. This task is further motivated by growing evidence that current vision-language models (VLMs) are not culturally neutral, but remain shaped by Western- or English-centric training and evaluation pipelines [2, 11, 18, 19].

To isolate the effect of cultural framing, we fix a zero-shot VLM, Qwen2.5-VL-7B-Instruct [14], keep the meme image and OCR transcription unchanged, and vary only the prompt framing. We compare a generic prompt, Indian and Chinese cultural prompts, and an Indian feminist prompt, and further test prompt sensitivity with a semantically similar Indian variant. Our contribution is threefold: a reproducible zero-shot audit framework for cross-cultural misogynistic meme detection; evidence that culture-conditioned prompting shifts decisions asymmetrically without improving aligned performance; and a demonstration that feminist reframing surfaces intra-cultural sensitivity beyond broad cultural framing. Figure 1 summarises the experimental pipeline.

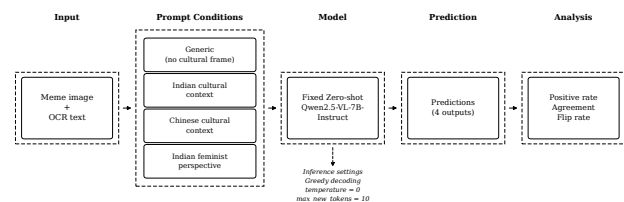


Figure 1: Overview of the zero-shot prompt audit pipeline.

2 Background and Related Work

2.1 Cross-cultural and Implicit Harmful Content Detection

Cross-cultural moderation is difficult because harmful meaning is culturally situated rather than universally stable, and VLMs remain shaped by Western- or English-centric training and evaluation



This work is licensed under a Creative Commons Attribution 4.0 International License. *ICMI '26, Napoli, Italy*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2318-6/26/10

<https://doi.org/10.1145/3776574.3832464>

pipelines [18]. Multicultural annotators can disagree substantially on the same content [3, 8], and implicit harm is often hidden behind misleading keywords, subtle formulations, or latent hate [5, 10, 20], compounded by limited multilingual resources and weak cross-lingual transfer [6]. VLMs also perform unevenly across cultures, with stronger results on Western than East Asian splits and US-centred interpretations of culturally variable gestures [2, 7, 19].

2.2 Misogyny Detection in Memes and Feminist Prompting as a Probe

Most misogyny-detection systems focus on supervised learning within a single language or cultural setting [17]; in multimodal meme detection, Fuse-MD targets Tamil and Malayalam [12] and CUET_Ignite targets Chinese memes with combined visual-textual encoders [15], but neither isolates prompt framing while holding model and input fixed. Our work instead relates to studies of how models reason about implicit misogyny: Muti et al. show that implied misogyny detection can depend more on internalized stereotypes than robust reasoning [9], Ovalle et al. show gender-exclusive harms can persist after alignment and go under-detected by mainstream evaluation [11], apparent cultural competence in VLMs may reflect superficial pattern matching rather than grounded understanding [21], and culturally sensitive reasoning remains unstable [1]. In this spirit, we use culture-conditioned and feminist prompts as lightweight audit tools: a feminist prompt is treated not as normative ground truth but as a probe of intra-cultural decision sensitivity, comparing a broad Indian cultural prompt against an explicitly feminist reframing to test how the model’s decisions change under a critical lens focused on patriarchy, sexism, and gender-based violence.

3 Method

3.1 Model Setup

We use Qwen2.5-VL-7B-Instruct [14] as a fixed zero-shot vision-language model, selected for its strong open-weight performance at a scale compatible with a single T4 GPU and its native image-text instruction-following, which suits zero-shot, prompt-only interventions without architectural changes. The model takes a meme image and its OCR transcription as input and outputs a binary label: 1 for misogynistic and 0 for non-misogynistic. No fine-tuning, parameter updates, or external supervision is applied. All inference uses greedy decoding with `max_new_tokens=10` and `temperature=0`. The only variable across conditions is the cultural framing in the prompt.

3.2 Prompt Design

We define four prompt conditions: **generic** (no explicit cultural frame), **Indian cultural**, **Chinese cultural**, and **Indian feminist**, which extends the Indian frame with an explicitly critical perspective on patriarchy, sexism, and gender-based violence. This set was designed to keep the study small but still informative: a generic baseline, two broad cultural prompts, and one more specific intra-cultural probe. All prompts ask the model to limit its output to only one digit (0 or 1).

Table 1: Prompt templates used in the experiments.

Prompt condition	Template (simplified)
Generic	Classify whether this meme is misogynistic (1) or not (0).
Indian cultural	Classify according to the cultural context of India.
Chinese cultural	Classify according to the cultural context of China.
Indian feminist	Classify from an Indian feminist perspective that critiques patriarchy, gender-based violence, and sexist stereotypes.

We did not include a Chinese feminist prompt mainly because each additional condition required a costly full inference run, which was difficult to accommodate within the available time and GPU budget. Instead, we prioritised one intra-cultural probe beyond broad cultural framing. We implemented this in the Indian setting because it allowed a clean within-context comparison between a broad cultural prompt and a more specific feminist reframing, and this comparison yielded a substantial additional behavioural shift in our results (positive rate 0.619 to 0.838; flip rate 0.219).

All prompt-conditioned analyses are performed exclusively on the English test split. The Tamil and Malayalam splits are used only for the official four-split evaluation and are not involved in the prompt comparison or flip-count analyses.

3.3 Prompt Sensitivity

To test whether model behaviour is sensitive to wording rather than only to the intended cultural frame, we construct a **semantic variant** of the Indian cultural prompt. This variant preserves the original meaning while changing surface wording. We run both prompts on a 100-sample subset of the English test split and compute agreement.

3.4 Evaluation Metrics and Sanity Check

In addition to the official Macro-F1, we report positive prediction rate, agreement, and flip rate. Macro-F1 measures alignment with target labels; positive prediction rate captures the model’s overall tendency to assign misogyny labels under different prompt conditions; agreement and flip rate quantify how often predictions remain stable or change across prompts for the same samples.

We also conduct a post-hoc sanity check on the Chinese split. The submitted version approximated the `indian_culture` column by reusing predictions across splits via image IDs. To quantify the resulting error, we randomly sample 50 Chinese memes, re-infer them with the correct Indian cultural prompt, and compare the results against the approximation.

4 Experimental Setup

We use the CC-MMD 2026 Task B dataset, which includes four language splits: English (1000 test samples), Chinese (340), Tamil (356), and Malayalam (200) [4]. The task evaluates cross-cultural robustness across Indian, Chinese, and Western (Ireland) contexts, and Macro-F1 is used as the official evaluation metric. The development split provides annotated labels and is used for target-aligned

internal evaluation, while the English test split is used for the main prompt-comparison experiments. In addition, we use a 100-sample subset of the English test split for prompt sensitivity testing and a randomly sampled 50-case subset of the Chinese split for the post-hoc sanity check.

All experiments are conducted in a zero-shot setting on a single T4 GPU with batch size 1, greedy decoding, `max_new_tokens=10`, and `temperature=0`. To maintain stable inference during long runs, GPU cache is periodically cleared.

5 Results

5.1 Official Performance and Development Diagnostics

We first report the official four-split test results, followed by internal development diagnostics. According to the official competition evaluation, our method achieves Macro-F1 scores of 0.7045 on English, 0.7627 on Chinese, 0.7369 on Tamil, and 0.6833 on Malayalam. As a purely zero-shot system, this approach is not intended to compete directly with the strongest supervised fine-tuning methods, but it provides a stable baseline for analysing prompt-conditioned behaviour across cultural settings.

Separately from the official evaluation, we conduct internal diagnostics on the English development split using organizer-provided target-label settings. When evaluated against Indian target labels, the generic proxy reaches a Macro-F1 of 0.6289, while the Indian-aligned prompt reaches 0.6292. When evaluated against Chinese target labels, the generic proxy reaches 0.7819, while the Chinese-aligned prompt reaches 0.7733. These development results suggest that culture-conditioned prompting does not clearly improve aligned development performance over the generic proxy. We therefore treat the development analysis as a behavioural diagnostic rather than as a replacement for the official four-split evaluation.

Table 2 summarizes the official scores and English-dev diagnostics. The latter show no clear gain from aligned cultural prompts over the generic proxy.

5.2 Prompt-Conditioned Decision Shifts on the English Test Split

We next examine how the model’s decision distribution changes when only the cultural framing of the prompt is varied. On the English test split, the positive prediction rate is 0.602 under the generic prompt, 0.619 under the Indian cultural prompt, 0.550 under

Table 2: Official results and English-dev diagnostics.

Setting	Split / target	Macro-F1
Official result	English test	0.7045
Official result	Chinese test	0.7627
Official result	Tamil test	0.7369
Official result	Malayalam test	0.6833
Generic proxy	English dev / Indian labels	0.6289
Indian prompt	English dev / Indian labels	0.6292
Generic proxy	English dev / Chinese labels	0.7819
Chinese prompt	English dev / Chinese labels	0.7733

Table 3: Positive prediction rates and flip rates under prompt-conditioned settings on the English test split.

Metric	Condition / transition	Value
Positive rate	Generic	0.602
Positive rate	Indian cultural	0.619
Positive rate	Chinese cultural	0.550
Positive rate	Indian feminist	0.838
Flip rate	Generic → India	0.049
Flip rate	Generic → China	0.086
Flip rate	India → Feminist	0.219

the Chinese cultural prompt, and 0.838 under the Indian feminist prompt.

Two robust patterns emerge. First, cultural prompting induces asymmetric decision shifts. Relative to the generic prompt, the Indian cultural prompt slightly increases the rate of misogyny predictions, whereas the Chinese cultural prompt decreases it. Consistent with this, Generic→India produces 49 flipped predictions, while Generic→China produces 86. This indicates that cultural framing does not operate as a uniform moderation signal; instead, different prompt frames move the model’s decision boundary in different directions. Second, the strongest effect comes from feminist reframing: the positive prediction rate rises from 0.619 to 0.838 relative to the Indian cultural condition (Table 3), suggesting that broad cultural labels do not fully capture variation in interpretive stance and that explicitly critical framing can substantially alter the treatment of borderline or implicit cases.

5.3 Feminist Reframing and Exploratory Prompt Sensitivity

We next compare the Indian cultural and Indian feminist prompts to assess whether intra-cultural reframing changes model behaviour. Agreement between the two settings is 0.781, a flip rate of 0.219 across 219 English test samples, confirming that feminist reframing produces a substantial behavioural shift beyond the increase in positive prediction rate alone. Table 3 also reports flip rates for other prompt transitions: India→Feminist induces the largest shift, exceeding both Generic→India and Generic→China, reinforcing that feminist reframing is a stronger behavioural intervention than broad cultural prompting alone.

We additionally conduct an exploratory prompt-sensitivity check using a wording variant of the Indian cultural prompt on a 100-sample subset, obtaining an agreement of 0.69. We interpret this cautiously as preliminary evidence that prompt formulation may affect model behaviour, rather than a definitive sensitivity estimate, and include it as an auxiliary diagnostic motivating future controlled analysis.

5.4 Chinese Sanity Check

Finally, we report a post-hoc sanity check on the Chinese split. This analysis tentatively reused Tamil and Malayalam predictions via image IDs to approximate an “Indian perspective” on Chinese memes. Because the ID spaces are disjoint across language partitions, this approximation is methodologically flawed and is not used in any

Table 4: Exploratory prompt sensitivity and Chinese sanity-check results.

Analysis	Metric	Value
Indian vs variant	Agreement	0.69
Chinese sanity check	Sample size	50
Chinese sanity check	Agreement	0.64
Chinese sanity check	Flip rate	0.36
Chinese sanity check	Approx. positive rate	0.30
Chinese sanity check	True positive rate	0.22

primary analysis. On 50 sampled cases, agreement between the approximated labels and true target-specific re-inference is only 0.64, with a flip rate of 0.36.

Table 4 summarizes the exploratory prompt-sensitivity result together with the Chinese sanity-check metrics.

We therefore do not interpret this disagreement as direct evidence of cultural difference. Instead, we use it only as a methodological caution showing that such cross-split reuse is unreliable. This issue does not affect the paper’s main results, all of which are computed on the original English test instances under prompt-controlled conditions.

6 Discussion

Culture-conditioned prompting changes model behaviour more clearly than it improves aligned development performance. Although prompt framing produces clear shifts in positive prediction rate and flip count, culturally aligned prompts do not outperform a generic proxy in development diagnostics. In our zero-shot setting, the primary effect of prompting is therefore behavioural rather than performance-driven.

These behavioural shifts are also asymmetric. The Indian cultural prompt slightly raises the positive prediction rate, the Chinese cultural prompt lowers it, and the feminist prompt produces the strongest increase, indicating that prompt framing does not act as a uniform sensitivity control but shifts the model’s decision boundary in different directions. Cultural prompting therefore changes both how often the model predicts misogyny and which examples cross the decision boundary.

Feminist reframing serves as the strongest probe in our experiments. Compared with the Indian cultural prompt, it produces both a large increase in positive prediction rate and a substantial number of flipped predictions. This does not imply that the feminist prompt is uniquely correct; rather, it suggests that broad cultural labels alone may not capture the full variation in interpretive stance. Although the high positive rate under the feminist prompt (83.8%) may partly reflect over-prediction rather than genuine sensitivity, the large flip count (219) demonstrates a non-trivial behavioural shift. Future work should combine targeted human annotation with prompt-based analysis to distinguish improved sensitivity from over-prediction under strongly framed prompts.

At the same time, prompt-based cultural evaluation should not be over-interpreted as evidence of robust cultural understanding. The wording-variant experiment reveals prompt sensitivity, while the Chinese sanity check shows that cross-split reuse is unreliable when

ID spaces are mismatched. We therefore view zero-shot culture-conditioned prompting as a lightweight audit tool rather than a complete solution for cross-cultural moderation. More broadly, future work should also consider whether simplified cultural prompting risks homogenising diverse viewpoints within a culture, a concern increasingly discussed in recent AI safety work on model homogenisation [16].

7 Conclusion

This paper presented a zero-shot audit of prompt-conditioned cross-cultural misogyny detection for CC-MMD 2026 Task B using Qwen2.5-VL-7B-Instruct, keeping the model and meme input fixed while varying only the cultural framing of the prompt. Our results show that culture-conditioned prompting reshapes model behaviour asymmetrically without a corresponding gain in aligned performance, that feminist reframing produces the strongest such shift, and that prompt-sensitivity and sanity checks reveal an instability that any prompt-based cultural evaluation must account for. We therefore argue that zero-shot culture-conditioned prompting is best understood as a lightweight audit framework rather than a complete solution to cross-cultural moderation.

8 Limitations

Our study has several limitations. First, we evaluate only a single zero-shot VLM, Qwen2.5-VL-7B-Instruct, so the observed behavioural patterns may not generalize to other model families; extending this audit to a second VLM (e.g., LLaVA-1.6-7B) under identical conditions is planned as follow-up work. Second, our development-set analysis is only an internal diagnostic and should not be treated as equivalent to the official four-split evaluation. Third, the wording-variant experiment is exploratory rather than fully controlled, so stronger claims about wording sensitivity require standardized parsing and reruns under identical inference and postprocessing settings. Fourth, the Chinese sanity check relies on a weak approximation across disjoint ID spaces and is included only as a methodological caution, not as direct evidence of cultural difference. Finally, our analysis focuses on binary label shifts and aggregate metrics; richer qualitative annotation and controlled human evaluation would be needed to distinguish improved sensitivity from over-prediction.

Ethics and Privacy Statement

This work studies misogynistic and culturally sensitive meme content for analytical purposes only. We do not present it as a deployable moderation system; rather, we audit how a fixed VLM responds to cultural framing under controlled conditions. The feminist prompt is used as an analytical probe rather than as normative ground truth.

All primary analyses use the original English meme instances and do not involve cross-language ID mapping. The only exception is the Chinese sanity check (Section 5.4), where Tamil and Malayalam predictions were tentatively reused via image IDs to approximate an “Indian perspective” on Chinese memes. Because these ID spaces are disjoint, the resulting low agreement (0.64) is treated only as evidence that such cross-split reuse is unreliable.

References

- [1] Zabir Al Nazi, G. M. Shahariar, Abrar Hossain, and Wei Peng. 2025. Are Vision Language Models Cross-Cultural Theory of Mind Reasoners? *arXiv preprint arXiv:2512.17394* (2025). arXiv:2512.17394 [cs.CL] <https://arxiv.org/abs/2512.17394>
- [2] Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. 2025. See It from My Perspective: How Language Affects Cultural Bias in Image Understanding. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=Xbl6t6zxZs>
- [3] Minh Duc Bui, Ming Shan Hee, Saliha Diallo, Bilel Trabelsi, Simon Razniewski, and Gerhard Weikum. 2025. Multi3Hate: Multimodal, Multilingual, and Multicultural Hate Speech Detection with Vision-Language Models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. <https://aclanthology.org/2025.naacl-long.490/>
- [4] CC-MMD 2026 Organizers. 2026. Cross-Cultural Misogynistic Meme Detection Challenge (CC-MMD) 2026. <https://www.codabench.org/competitions/14187/>. ICMI 2026 Grand Challenge task description and competition page.
- [5] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 conference on empirical methods in natural language processing*. 345–363.
- [6] Aiqi Jiang and Arkaitz Zubiaga. 2026. Cross-lingual Offensive Language Detection: A Systematic Review of Dataset, Approach and Challenge. *Comput. Surveys* 58, 10 (2026), 1–37.
- [7] Hyeokjun Lee, Jeonghyo Song, Geonhui Jang, and Youngjoon Yoo. 2025. Lost in Translation: A Position Paper on Probing Cultural Bias in Vision-Language Models via Hanbok VQA. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 4538–4545. https://openaccess.thecvf.com/content/ICCV2025W/WCCA/html/Lee_Lost_in_Translation_A_Position_Paper_on_Probing_Cultural_Bias_ICCVW_2025_paper.html
- [8] 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)*. 4205–4224.
- [9] Arianna Muti, Federico Ruggeri, Khalid Al-Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. 2024. Language is Scary when Over-Analyzed: Unpacking Implied Misogynistic Reasoning with Argumentation Theory-Driven Prompts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 21137–21154. doi:10.18653/v1/2024.emnlp-main.1174
- [10] Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 1997–2013.
- [11] Anaelia Ovalle, Krunoslav Lehman Pavasovic, Louis Martin, Luke Zettlemoyer, Eric Michael Smith, Adina Williams, and Levent Sagun. 2025. The Root Shapes the Fruit: On the Persistence of Gender-Exclusive Harms in Aligned Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. doi:10.1145/3715275.3732196
- [12] Rahul Ponnusamy, Saranya Rajiakodi, Bhuvaneswari Sivagnanam, Anshid Kizhakkeparambil, Dhruv Sharma, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2026. Fuse-MD: A culturally-aware multimodal model for detecting misogyny memes. *Natural Language Processing Journal* 14 (March 2026), 100197. doi:10.1016/j.nlp.2026.100197
- [13] Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Anshid A. Kizhakkeparambil, Ping Du, Saranya Rajiakodi, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2026. Overview of the CC-MMD 2026 Grand Challenge: Cross-Cultural Misogynistic Meme Detection in Multimodal Memes. In *Proceedings of the 28th International Conference on Multimodal Interaction (ICMI '26)*. Association for Computing Machinery, Napoli, Italy.
- [14] Qwen Team. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025). arXiv:2502.13923 [cs.CV] <https://arxiv.org/abs/2502.13923>
- [15] MD. Mahadi Rahman, Mohammad Minhaj Uddin, Mohammad Oman, and Mohammad Shamsul Arefin. 2025. CUET_Ignite@LT-EDI-2025: A Multimodal Transformer-Based Approach for Detecting Misogynistic Memes in Chinese Social Media. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*. Unior Press, Naples, Italy, 172–177. <https://aclanthology.org/2025.ltedi-1.28/>
- [16] Ian Rios-Sialer. 2026. Structure-Aware Diversity Pursuit as an AI Safety Strategy against Homogenization. *arXiv preprint arXiv:2601.06116* (2026). arXiv:2601.06116 <https://arxiv.org/abs/2601.06116>
- [17] Elena Shushkevich and John Cardiff. 2019. Automatic misogyny detection in social media: A survey. *Computación y Sistemas* 23, 4 (2019), 1159–1164.
- [18] Mo Wang, Kaixuan Ren, Pratik Jalan, Ahmed Ashraf, Tuong Vy Vu, Rahul Seetharaman, Shah Nawaz, and Usman Naseem. 2026. From Native Memes to Global Moderation: Cross-Cultural Evaluation of Vision-Language Models for Hateful Meme Detection. In *Proceedings of the ACM Web Conference 2026*. Association for Computing Machinery, 9788–9799. doi:10.1145/3774904.3793007
- [19] Akhila Yerukola, Saadia Gabriel, Nanyun Peng, and Maarten Sap. 2025. Mind the Gesture: Evaluating AI Sensitivity to Culturally Offensive Non-Verbal Gestures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 25041–25080. <https://aclanthology.org/2025.acl-long.1218/>
- [20] Wenjie Yin and Arkaitz Zubiaga. 2022. Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media* 30 (2022), 100210.
- [21] Haorui Yu, Yang Zhao, Yijia Chu, and Qifeng Yi. 2025. Seeing Symbols, Missing Cultures: Probing Vision-Language Models' Reasoning on Fire Imagery and Cultural Meaning. In *Proceedings of the 9th Widening NLP Workshop*. Association for Computational Linguistics, Suzhou, China, 1–8. doi:10.18653/v1/2025.wnlp-main.1

A Full Prompt Templates

This appendix provides the full prompt texts used in the main experiments.

A.1 Generic Prompt

You are an expert in detecting misogyny in memes. Given an English meme image and its OCR transcription, classify whether it is misogynistic (1) or not (0). Consider both the visual and textual content. Answer with ONLY a single digit: 1 for misogyny, 0 for not-misogyny. No extra text.
OCR transcription: ""[TRANSCRIPTION]""

A.2 Indian Cultural Prompt

You are an expert in detecting misogyny in memes. Given an English meme image and its OCR transcription, classify whether it is misogynistic (1) or not (0) according to the cultural context of India. Consider both visual and text. Be aware of cultural nuances, implicit language, humor, and symbolism. Answer with ONLY a single digit: 1 for misogyny, 0 for not-misogyny. No extra text.
OCR transcription: ""[TRANSCRIPTION]""

A.3 Chinese Cultural Prompt

You are an expert in detecting misogyny in memes. Given an English meme image and its OCR transcription, classify whether it is misogynistic (1) or not (0) according to the cultural context of China. Consider both visual and text. Be aware of cultural nuances, implicit language, humor, and symbolism. Answer with ONLY a single digit: 1 for misogyny, 0 for not-misogyny. No extra text.
OCR transcription: ""[TRANSCRIPTION]""

A.4 Indian Feminist Prompt

You are an expert in detecting misogyny in memes from an Indian feminist perspective that actively critiques patriarchy, gender-based violence, and sexist stereotypes. Given an English meme image and its

OCR transcription, classify whether it is misogynistic (1) or not (0) according to this feminist critical lens.

Consider both visual and text. Be aware of subtle forms of misogyny that may be normalised in mainstream culture. Answer with ONLY a single digit: 1 for misogyny, 0 for not-misogyny. No extra text.
OCR transcription: ""[TRANSCRIPTION]""