Bridging Modalities: Enhancing Cross-Modality Hate Speech Detection with Few-Shot In-Context Learning

Ming Shan Hee[1,*] Aditi Kumaresan[1,*] Roy Ka-Wei Lee[1]
[1]Singapore University of Technology and Design
{mingshan_hee@mymail., aditi_kumaresan@, roy_lee@}sutd.edu.sg
[*]These authors contributed equally to this work.

Abstract

The widespread presence of hate speech on the internet, including formats such as text-based tweets and vision-language memes, poses a significant challenge to digital platform safety. Recent research has developed detection models tailored to specific modalities; however, there is a notable gap in transferring detection capabilities across different formats. This study conducts extensive experiments using few-shot in-context learning with large language models to explore the transferability of hate speech detection between modalities. Our findings demonstrate that text-based hate speech examples can significantly enhance the classification accuracy of vision-language hate speech. Moreover, text-based demonstrations outperform vision-language demonstrations in few-shot learning settings. These results highlight the effectiveness of cross-modality knowledge transfer and offer valuable insights for improving hate speech detection systems[1].

1 Introduction

Motivation. Hate speech in the online space appears in various forms, including text-based tweets and vision-language memes. Recent hate speech studies have developed models targeting specific modalities [Cao et al., 2023, Awal et al., 2021]. However, these approaches are often optimized to within-distribution data and fail to address zero-shot out-of-distribution scenarios.

The emergence of vision-language hate speech, which comprises text and visual elements, presents two significant challenges. First, there is a scarcity of datasets, as this area has only recently gained lots of attention. Second, collecting and using such data is complicated by copyright issues and increasingly stringent regulations on social platforms. Consequently, the limited availability of vision-language data hampers performance in out-of-distribution cases. In contrast, the abundance and diversity of text-based data offer a potential source for cross-modality knowledge transfer [Hee et al., 2024].

Research Objectives. This paper investigates whether text-based hate speech detection capabilities can be transferred to multimodal formats. By leveraging the richness of text-based data, we aim to enhance the detection of vision-language hate speech, addressing current research limitations and improving performance in low-resource settings.

Contributions. This study makes the following key contributions: (i) We conduct extensive experiments evaluating the transferability of text-based hate speech detection to vision-language formats using few-shot in-context learning with large language models. (ii) We demonstrate that text-based hate speech examples significantly improve the classification accuracy of vision-language hate speech. (iii) We show that text-based demonstrations in few-shot learning contexts outperform vision-language hate speech demonstrations, highlighting the potential for cross-modality knowledge transfer. These contributions address critical gaps in existing research and provide a foundation for developing robust hate speech detection systems.

[1]GitHub: `https://github.com/Social-AI-Studio/Bridging-Modalities`

Table 1: Statistical distributions of datasets, where "H" represents Hate and "Non-H" represents non-hate

Dataset	Support		Test	
	# H	# Non-H	# H	# Non-H
Latent Hatred	8189	13,921	–	–
FHM-FG	3,007	5,493	–	–
MAMI	–	–	246	254
	–	–	500	500

2 Research Questions

As all forms of hate speech share one definition, this study investigates the usefulness of using hate speech from one form, such as text-based hate speech, to classify hate speech in another form, such as vision-language hate speech. Working towards this goal, we formulate two research questions to guide our investigation.

RQ1: Does the text hate speech support set help with vision-language hate speech? Visual-language hate speech presents a distinct challenge compared to text-based hate speech, as malicious messages can hide within visual elements or interactions between modalities. It remains uncertain whether text-based hate speech can be useful for classifying visual-language hate speech. We investigate this uncertainty by performing few-shot in-context learning on large language models. This method allows the model to learn from text-based hate speech demonstration examples before classifying visual-language hate speech instances.

RQ2: How does the text hate speech support set fare against the vision-language hate speech support set? Intuitively, using vision-language hate speech demonstrations should result in superior performance. However, the effectiveness of text-based hate speech demonstrations compared to vision-language hate speech demonstrations remains an open question. To investigate this gap, we conducted another round of few-shot in-context learning on large language models with a vision-language hate speech support set.

3 Experiments

3.1 Experiment Settings

Models. We use the Mistral-7B[2] [Jiang et al., 2023] and Qwen2-7B[3] [Bai et al., 2023] models, both of which demonstrate strong performance across various benchmarks, in our primary experiments. Notably, their models on LMSYS's Chatbot Arena Leaderboard achieve high ELO scores [Chiang et al., 2024]. To facilitate reproducibility and minimize randomness, we use the greedy decoding strategy for text generation.

We conducted additional experiments to support the findings in our paper further with two additional models: LLaVA-7B[4] and Llama3-8B[5]. The results of these experiments are presented in Appendix I.

Test Datasets. The Facebook Hateful Memes (FHM) dataset [Mathias et al., 2021] contains synthetic memes categorized into five types of hate incitement: gender, racial, religious, nationality, and disability-based. The Multimedia Automatic Misogyny Identification (MAMI) [Fersini et al., 2022] dataset comprises real-world misogynistic memes classified into shaming, stereotype, objectification, and violence categories. Both datasets contain text overlay information, eliminating the need for an OCR model to

[2] `mistralai/Mistral-7B-Instruct-v0.3`
[3] `Qwen/Qwen2-7B-Instruct`
[4] `llava-hf/llava-v1.6-mistral-7b-hf`
[5] `meta-llama/Llama-3.1-8B-Instruct`

extract text.

For evaluation, we use the FHM's dev_seen split, which includes 246 hateful memes and 254 non-hateful ones, and the MAMI's test split, consisting of 500 hateful and 500 non-hateful memes.

Text Support Set. We use the Latent Hatred [ElSherief et al., 2021] dataset, which includes both explicit and implicit forms of hate speech, such as coded and indirect derogatory attacks. This dataset comprises 13,921 non-hateful speeches, 1,089 explicit hate speeches, and 7,100 implicit hate speeches.

Vision-Language Support Set. We use the FHM train split for evaluation, containing 3,007 hateful memes and 5,493 non-hateful memes.

3.2 Data Preprocessing

Image Captioning. To perform hateful meme classification with the large language models, we perform image captioning on the meme using the OFA [Wang et al., 2022] model pre-trained on the MSCOCO [Lin et al., 2014] dataset.

Rationale Generation. We prompt Mistral-7B to generate informative rationales that explain the underlying meaning of the content, providing additional context for the few-shot in-context learning. Specifically, the model generates rationales by using the content and ground truth labels (i.e., prompt + content $\rightarrow$ ground truth label $\rightarrow$ explanation). For the Latent Hatred dataset, we use post information and labels, while for the FHM dataset, we use meme text, captions, and labels. To mitigate noise from varying rationale formulations, we instruct the model to consider both textual and visual elements, focusing on target groups, imagery, and the impact of tweet/meme bias perpetuation. More details can be found in Appendix F.

3.3 RQ1: Does text hate speech help with vision-language hate speech?

To evaluate the effectiveness of the few-shot in-context learning approach and the Latent Hatred support set, we employed three sampling strategies: Random sampling, TF-IDF sampling, and BM-25 sampling. The TF-IDF and BM-25 strategies leverage the text and caption information of the test record to identify similar examples from the support set, focusing on either the text or the generated rationale. Table 2 shows the comparison of zero-shot and few-shot in-context learning experiment results with Latent Hatred support set.

The experimental results demonstrate that employing a few-shot in-context learning approach with text-based hate speech demonstrations is highly effective in classifying vision-language hate speech. Firstly, while the random sampling strategy could retrieve more irrelevant demonstrations compared to other strategies, the few-shot in-context learning with random sampling surpasses the zero-shot inference performance on both models across two datasets in terms of F1 score. Secondly, the TF-IDF and BM-25 sampling strategies exceed the zero-shot inference performance on both models within the MAMI dataset. Conversely, within the FHM dataset, we observed several instances where some sampling strategies in the few-shot in-context learning scenario performed worse than zero-shot inference. However, these sampling strategies consistently outperformed zero-shot inference when run with 16-shots in-context learning. Lastly, the best few-shot in-context learning performance within each dataset and each model shows significant improvement over zero-shot model performance. For example, the Mistral-7B model achieves an F1 score improvement of 0.64 and 1.23 on the FHM and MAMI datasets respectively.

3.4 RQ2: How does text hate speech support set fare against vision-language hate speech support set?

Table 3 shows the comparison of zero-shot and few-shot in-context learning experiment results with the FHM support set. The experimental results indicate that using the FHM support set can enhance model performance in some scenarios. However, it is noteworthy that in many instances, few-shot in-context learning performs worse than zero-shot model performance when compared against the Latent Hatred support set. Most significantly, the model encounters the most failures on the FHM test set despite using the FHM train set as a support set. We also observed that the best model performance with the Latent Hatred support set surpasses the best model performance with the FHM support set across all instances. We speculate that this discrepancy may stem from the oversimplification of visual information into image captions and the broader topic coverage provided by the Latent Hatred dataset. Nevertheless, this suggests that text-based data can serve as a valuable resource for improving performance on multimodal tasks, particularly in low-resource settings.

4 Few-Shot Demonstration Analysis

While including relevant few-shot in-context learning examples can improve model performance, the degree to which these examples benefit the model remains uncertain. To gain deeper insights, we examine the examples that got correctly classified and misclassified using the demonstration exemplars from the Latent Hatred support dataset.

The detailed analysis and case study examples, along with their few-shot in-context demonstrations, can be found in Appendices G and H.

Latent Hatred's Support Set We found that using relevant examples as demonstrations significantly improves classification, as the additional context aids the model in evaluating similar content more effectively. This approach enhances the model's ability to generalize across diverse hate speech contexts and formats, thereby helping to reduce false negatives in edge cases. However, we also observed that models sometimes misinterpret neutral content as hateful. This misinterpretation may arise from exposure to demonstration examples that contain dismissive or derogatory language on sensitive topics. Consequently, these examples can lead to an overgeneralization of what qualifies as hateful, causing content that was correctly classified in a zero-shot setting to be misclassified. This issue is similar to the problem of oversensitivity to specific terms found in fine-tuned multimodal hate speech detection models [Cuo et al., 2022, Hee et al., 2022, Rizzi et al., 2023].

5 Related Works

Numerous approaches have been proposed to tackle the online hate speech problem [Cao et al., 2023, Lee et al., 2021, Hee et al., 2023, Lin et al., 2024]. While these approaches demonstrate impressive performance, they often require large amounts of data for fine-tuning, and the rapid evolution of hate speech can quickly render these models outdated. Furthermore, a recent study indicated that these models are vulnerable to adversarial attacks [Aggarwal et al., 2023].

These challenges led to exploring few-shot hate speech detection approaches, where models learn using limited data [Meta, 2021, Awal et al., 2023]. Mod-HATE trains specialized modules on related tasks and integrates the weighted module with large language models to enhance detection capabilities [Cao et al., 2024]. Our approach contributes to this field by addressing the challenge of limited data availability, using the abundance and diversity of text-based hate speech as an alternative source for cross-modality knowledge transfer.

6 Conclusion

We investigated the possibility of cross-modality knowledge transfer using few-shot in-context learning with large language models. Our extensive experiments show that text-based hate speech demonstrations significantly improve the classification accuracy of vision-language hate speech, and using text-based demonstrations in few-shot in-context learning outperforms using vision-language demonstrations. For future works, we aim to extend our analysis to more datasets and explore other cross-modality knowledge transfer approaches such as cross-modality fine-tuning.

Acknowledgement

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award ID: MOE-T2EP20222-0010). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

Limitations

There are several limitations in this research study:
Model Coverage and Model Size. In this study, we evaluated and compared two large models containing 7B parameters. In the future, we aim to extend our analysis to other large models when more computational resources are available.
Large Language Model. In this study, we evaluated few-shot in-context learning in large language models. The experiments are designed in this manner, so to ensure that there can be a fair comparison between the different support sets. We recognize that using a vision-language support set for few-shot in-context learning with a large vision-language model could achieve better performance. However, evaluation using large vision-language models would then be unfair to text support set for few-shot in-context learning.

Ethical Considerations

Impact of False Positives. Developing a reliable and generalizable hate speech detection system is crucial, as false positives can significantly impact free speech and diminish user trust. Firstly, overly aggressive detection systems may mistakenly flag content that does not qualify as hate speech, thereby suppressing free speech and hindering meaningful discussions. Secondly, when users frequently encounter false positives, their confidence in the platform's moderation system may diminish. The reduced trust can result in decreased user engagement and a perception of bias within the platform.

References

[Aggarwal et al., 2023] Piush Aggarwal, Pranit Chawla, Mithun Das, Punyajoy Saha, Binny Mathew, Torsten Zesch, and Animesh Mukherjee. 2023. Hateproof: Are hateful meme detection systems really robust? In *Proceedings of the ACM Web Conference 2023*, pages 3734–3743.

[Awal et al., 2021] Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2021. Angrybert: Joint learning target and emotion for hate speech detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 701–713. Springer.

[Awal et al., 2023] Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. 2023. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Transactions on Computational Social Systems*, 11(1):1086–1095.

[Bai et al., 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

[Cao et al., 2023] Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252.

[Cao et al., 2024] Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4575–4584.

[Chiang et al., 2024] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

[Cuo et al., 2022] Keyan Cuo, Wentai Zhao, Mu Jaden, Vishant Vishwamitra, Ziming Zhao, and Hongxin Hu. 2022. Understanding the generalizability of hateful memes detection models against covid-19-related hateful memes. In *International Conference on Machine Learning and Applications*.

[ElSherief et al., 2021] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

[Fersini et al., 2022] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

[Hee et al., 2022] Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, pages 3651–3655.

[Hee et al., 2023] Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5995–6003.

[Hee et al., 2024] Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent advances in hate speech moderation: Multimodality and the role of large models. *arXiv preprint arXiv:2401.16727*.

[Jiang et al., 2023] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

[Lee et al., 2021] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147.

[Lin et al., 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

[Lin et al., 2024] 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 on Web Conference 2024*, pages 2359–2370.

[Mathias et al., 2021] Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the woah 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206.

[Meta, 2021] Meta. 2021. Harmful content can evolve quickly. our new ai system adapts to tackle it. Accessed on Oct 2, 2024.

[Rizzi et al., 2023] Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(5):103474.

[Wang et al., 2022] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052.

A Potential Risks

This project seeks to counteract the dissemination of harmful memes, aiming to protect individuals from prejudice and discrimination based on race, religion and gender. However, we acknowledge the risk of malicious users reverse-engineering memes to evade detection by CMTL-RAG AI systems, which is strongly discouraged and condemned.

B Licenses and Usage Scientific Artifacts

B.1 Models

All of the LLMs used in this paper contain licenses permissive for academic and/or research use.

- Mistral-7B Apache-2.0 License
- Qwen2-7B Apache-2.0 License
- LLaVA-7B Apache 2.0 License
- LLaMA-8B Llama 3.1 Community License

B.2 Datasets

All of the datasets used in this paper contain licenses permissive for academic and/or research use.

- Latent Hatred Dataset. MIT License
- Hateful Memes Dataset. MIT License
- Multimedia Automatic Misogyny Identification. Creative Commons License (CC BY-NC-SA 4.0)

B.3 Anonymity and Offensive Content

The datasets used in this research contain offensive content, which is crucial for addressing the research questions. Importantly, there are no unique identifiers for the individuals who authored the hateful content in these datasets.

C Computational Experiments

NVIDIA A40 GPUs were utilized for the work done in this paper.

C.1 Experimental Setup

We thoroughly discussed the experimental setup in the main body of the paper. This included descriptions of the models used (Mistral-7B and Qwen2), the number of shots (0-shot, 4-shots, 8-shots, 16-shots), and the different strategies employed for matching (Random, TF-IDF, BM-25) across two datasets (FHM and MAMI). Best-found hyperparameter values were highlighted in the results tables, such as the highest accuracy and F1 scores achieved for each experimental condition.

C.2 Use of Existing Packages

C.2.1 Large Language Models

- transformers 4.41.1

C.2.2 Matching and Retrieval Scoring

- rank-bm25 0.2.2 for BM-25 similarity matching
- scikit-learn 1.5.0 for TF-IDF similarity matching

D Few-Shot In-Context Learning

In this approach, we retrieve relevant labelled examples from a 'support dataset' using similarity metrics such as TF-IDF or BM-25 for a given meme from the inference dataset. These examples are then provided as demonstrations in a few-shot prompt to enhance the model's understanding of the meme. Finally, we prompt the model to classify the meme, leveraging the augmented context for improved accuracy.

E Similarity Metrics

E.1 TF-IDF

TFIDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus). By creating TF-IDF vectors for a 'support dataset', we can use cosine similarity to find the most similar records to a given inference record.

E.2 BM25

BM25 is an advanced version of the TF-IDF weighting scheme used in search engines. It incorporates term frequency saturation and document length normalization to improve retrieval performance. We generate vectors for each record in the 'support dataset' and use cosine similarity to identify records most similar to an inference record.

F Rationale Generation Details

We use the Mistral-7B model, a state-of-the-art language model known for its capabilities in language understanding and generation.

We implement a ten-shot prompting method to generate explanations for the hateful tweets in the Latent Hatred dataset. Specifically, we select five examples of hateful posts and five non-hateful posts for the ten-shot prompt demonstrations. Each demonstration in the prompt follows the following template, given the post text, the post label (hateful or not hateful) and the post rationale:

```
User: Determine whether the following post is hateful. Text: {text}:
Assistant: {label}
User: Briefly provide an explanation, in no more than three points, for the post
Your explanation should address the targeted group, any derogatory imagery or la
and the impact it has on perpetuating bias, stereotypes, prejudice, discriminati
Assistant: Answer: {rationale}
```

Similarly, for FHM/MAMI, we select five examples of hateful memes and five non-hateful memes for the ten-shot prompt demonstrations.

Each demonstration in the prompt follows the following template, given the meme text, the meme label (hateful or not hateful) and rationale:

```
User: Determine whether the following meme is hateful. Text: {text} Caption: {ca
Assistant: {label}
User: Briefly provide an explanation, in no more than three points, for the meme
Your explanation should address the targeted group, any derogatory imagery or la
and the impact it has on perpetuating bias, stereotypes, prejudice, discriminati
Assistant: Answer: {rationale}
```

The demonstration explanation follows a list format, where each list item addresses the targeted group, any derogatory imagery or language used, and the impact it has on perpetuating bias. Finally, to reiterate the classification of the post's hatefulness, the explanation concludes with the sentence, "In summary, this post/meme is {label}". During inference, the demonstrations provide a structured guide for the model to generate rationales based on the given post/meme's text and label.

G In-Context Demonstration Analysis

This section evaluates and compares the effectiveness of Latent Hatred in enhancing the models' multimodal hateful memes classification ability. Additionally, we identify situations where the context provided by these demonstrations can sometimes hinder the model's performance. The in-context demonstrations for each case study example can be found in Appendix H.

G.1 Case Study: Latent Hatred Cross-Modality Effectiveness

Latent Hatred - Correct Classifications. We examined and studied two cases where the Mistral-7B model failed to correctly classify the FHM meme in the 0-shot classification setting but succeeded when latent hatred demonstrations were introduced at 4, 8, and 16-shot levels.

Example 1 - Analysis. Demonstration 1 is crucial because it demonstrates how derogatory language targeting a religious group (comparing the Qur'an to marijuana) directly relates to the meme's use of sensitive imagery (a woman in a hijab) and the potential for offending religious or cultural groups.

Example 2 - Analysis. The meme implies that people with an afro, 'presumably from a poor country', have low IQs and that an 80 IQ is the best they can achieve due to poverty. Demonstration 2 highlights how generalizations about IQ can perpetuate stereotypes about intelligence and economic status.

Latent Hatred - Incorrect Classifications. We examined and studied two cases where the model's initial success in 0-shot classification declined when Latent Hatred demonstrations were included in 4, 8, and 16-shot settings.

Example 1 Analysis. Demonstration 1 features derogatory language that dehumanizes individuals by comparing them to baboons, which likely influenced the model to associate the meme's content with hatefulness. The model might have incorrectly applied the same reasoning to the meme with the baboon imagery, interpreting it as offensive or derogatory.

Analysis. The model misclassified the meme as hateful due to its reliance on examples like Example 3 and Example 4, which have similar wording, leading to a misinterpretation of the neutral photo. Example 3 involves dismissive language that trivializes hate crimes, potentially influencing the model to perceive similar tones or implications in the meme. Example 4 demonstrates how derogatory comparisons involving sensitive topics (like Jewish people and the Holocaust) can be classified as hateful.

H In-context Demonstrations

H.1 Latent Hatred Support Set - Correct Classification Case Study #1

```
### EXAMPLE 1
Content: A post containing ’my weed is like the qur’an.<repeat> i burn it until
Answer: Hateful
Rationale: Targeted Group: The post targets individuals who follow Islam by maki
comparison between their holy book, the Qur’an, and marijuana.
Derogatory Imagery/Language: The statement uses disrespectful and offensive lang
the Qur’an to a drug and suggesting that it should be burned.
Impact on Bias/Stereotypes: This post reinforces harmful stereotypes and prejudi
by belittling their religious beliefs and promoting disrespect towards their hol
incite fear and hostility towards Muslim individuals and communities.
In summary, this post is hateful.

### EXAMPLE 2
Content: A post containing ’" vile weed!"""’
Answer: Not Hateful
Rationale: Targeted Group: The post does not target a specific group based on at
race, religion, ethnicity, gender, sexual orientation, disability, or other dist
characteristics.
Derogatory Imagery/Language: There is no derogatory language or imagery used in
phrase "vile weed" could be interpreted as negative, but it does not specificall
Impact on Bias/Stereotypes: The post does not perpetuate harmful stereotypes or
any group. It does not contribute to bias, discrimination, or inciting harm.
In summary, this post is not hateful.

### EXAMPLE 3
Content: A post containing ’>>> americas future>>> white woman raped then stoned
in south africa’
Answer: Hateful
Rationale: Targeted Group: The post targets black individuals, particularly thos
by implying that they are responsible for a violent crime against a white woman.
Derogatory Imagery/Language: The use of the phrase "racist blacks" is derogatory
harmful stereotypes about black people, suggesting that they are inherently viol
discriminatory. The title "america’s future" also implies that the safety and we
people is at risk due to the actions of black individuals.
Impact on Bias/Stereotypes: This post reinforces harmful stereotypes and prejudi
individuals, potentially inciting fear and hostility towards them. It contribute
that portrays black people as dangerous and threatening, which can perpetuate ra
discrimination.
In summary, this post is hateful.

### EXAMPLE 4
```

```
Content: A post containing 'mandela's rainbow nation: white woman raped then sto
in south africa via'
Answer: Hateful
Rationale: Targeted Group: The post targets black individuals in South Africa, i
are violent and racist towards white women.
Derogatory Imagery/Language: The statement uses derogatory language by describin
as "racist" and suggesting that they are responsible for a violent act against a
Impact on Bias/Stereotypes: This post reinforces harmful stereotypes and prejudi
individuals, perpetuating the idea that they are violent and dangerous towards w
could incite fear and hostility towards black South Africans while promoting dis
divisive beliefs.
In summary, this post is hateful.
```

H.2 Latent Hatred Support Set - Correct Classification Case Study #2

[Content similar to above with four examples showing IQ-related posts with rationales]

H.3 Latent Hatred Support Set - Incorrect Classification Case Study #1

[Content similar to above with four examples including baboon comparison]

H.4 Latent Hatred Support Set - Incorrect Classification Case Study #2

[Content similar to above with four examples including dismissive language about hate crimes]

I Additional Experiments

I.1 LLaVA-7B

I.2 Llama3-8B

Table 2: Comparison of zero-shot and few-shot in-context learning with Latent Hatred support set across different demonstration sampling (Dem. Sampl.) strategies. Underlined represent the best results within a dataset for the given model and given few-shot setting, bold indicate the best results within a dataset for a given model across all few-shot settings and red denote few-shot in-context learning results below zero-shot performance.

Model	# Shots	Dem. Samp.	FHM			MAMI		
	Matching		Acc.	F1	# Invalids	Acc.	F1	# Invalids
Mistral-7B	0-shot	–	0.614	0.594	0	0.619	0.568	0
	4-shots	Random	0.618	0.613	0	0.655	0.636	0
		TF-IDF Text.	0.634	0.634	0	0.653	0.649	0
		TF-IDF Rationale	0.618	0.618	0	0.662	0.658	0
		BM-25 Text.	0.658	0.657	0	0.665	0.662	0
		BM-25 Rationale	0.598	0.596	0	0.676	0.671	0
	8-shots	Random	0.620	0.611	0	0.634	0.602	0
		TF-IDF Text.	0.642	0.641	0	0.665	0.658	0
		TF-IDF Rationale	0.626	0.625	0	0.657	0.649	0
		BM-25 Text.	0.660	0.658	0	0.685	0.680	0
		BM-25 Rationale	0.612	0.608	0	0.669	0.661	0
	16-shots	Random	0.618	0.610	0	0.642	0.611	0
		TF-IDF Text.	0.644	0.644	0	0.675	0.668	0
		TF-IDF Rationale	0.632	0.631	0	0.632	0.631	0
		BM-25 Text.	0.638	0.636	0	**0.705**	**0.701**	0
		BM-25 Rationale	0.614	0.611	0	0.665	0.659	0
Qwen2-7B	0-shot	–	0.624	0.609	0	0.614	0.574	0
	4-shots	Random	0.620	0.614	0	0.653	0.632	0
		TF-IDF Text.	0.632	0.631	0	0.650	0.641	0
		TF-IDF Rationale	0.634	0.633	0	0.663	0.653	0
		BM-25 Text.	0.644	0.642	0	0.672	0.664	0
		BM-25 Rationale	0.590	0.587	0	0.663	0.654	0
	8-shots	Random	0.632	0.628	0	0.645	0.622	0
		TF-IDF Text.	0.632	0.632	0	0.656	0.650	0
		TF-IDF Rationale	0.618	0.617	0	0.664	0.656	0
		BM-25 Text.	0.654	0.653	0	0.679	0.674	0
		BM-25 Rationale	0.604	0.603	0	0.654	0.646	0
	16-shots	Random	0.632	0.626	0	0.652	0.631	0
		TF-IDF Text.	0.628	0.628	0	0.656	0.651	0
		TF-IDF Rationale	0.632	0.631	0	0.665	0.659	0
		BM-25 Text.	0.624	0.624	0	0.678	0.674	0
		BM-25 Rationale	0.630	0.629	0	0.679	0.674	0

Table 3: Comparison of zero-shot and few-shot in-context learning experiment results with FHM support set across different demonstration sampling (Dem. Sampl.) strategies. Underlined represent the best results within a dataset for the given model and given few-shot setting, bold indicate the best results within a dataset for a given model across all few-shot settings and red denote few-shot in-context learning results below zero-shot performance.

Model	# Shots	Dem. Samp.	FHM			MAMI		
	Matching		Acc.	F1	# Invalids	Acc.	F1	# Invalids
Mistral-7B	0-shot	–	0.614	0.594	0	0.619	0.568	0
	4-shots	Random	0.622	0.617	0	0.656	0.642	0
		TF-IDF Text+Cap.	0.604	0.598	0	0.678	0.670	0
		TF-IDF Rationale	0.618	0.613	0	0.662	0.652	0
		BM-25 Text+Cap.	0.592	0.584	0	0.662	0.653	0
		BM-25 Rationale	0.620	0.617	0	0.667	0.659	0
	8-shots	Random	0.624	0.615	0	0.652	0.632	0
		TF-IDF Text+Cap.	0.618	0.611	0	0.675	0.664	0
		TF-IDF Rationale	0.628	0.622	0	0.681	0.670	0
		BM-25 Text+Cap.	0.606	0.599	0	0.672	0.661	0
		BM-25 Rationale	0.628	0.624	0	0.674	0.666	0
	16-shots	Random	0.620	0.614	0	0.668	0.651	0
		TF-IDF Text+Cap.	0.620	0.617	0	0.672	0.665	0
		TF-IDF Rationale	0.638	0.635	0	0.671	0.661	0
		BM-25 Text+Cap.	0.630	0.625	0	0.682	0.673	0
		BM-25 Rationale	0.634	0.633	0	**0.687**	**0.680**	0
Qwen2-7B	0-shot	–	0.624	0.609	0	0.614	0.574	0
	4-shots	Random	0.606	0.602	0	0.655	0.642	0
		TF-IDF Text+Cap.	0.620	0.620	0	0.659	0.657	0
		TF-IDF Rationale	0.636	0.636	0	0.650	0.646	0
		BM-25 Text+Cap.	0.616	0.616	0	0.676	0.674	0
		BM-25 Rationale	0.622	0.622	0	0.669	0.672	0
	8-shots	Random	0.592	0.581	0	0.642	0.624	0
		TF-IDF Text+Cap.	0.606	0.604	0	0.648	0.645	0
		TF-IDF Rationale	0.620	0.619	0	0.649	0.644	0
		BM-25 Text+Cap.	0.614	0.613	0	0.665	0.662	0
		BM-25 Rationale	0.624	0.623	0	0.669	0.664	0
	16-shots	Random	0.602	0.592	0	0.650	0.634	0
		TF-IDF Text+Cap.	0.610	0.610	0	0.649	0.648	0
		TF-IDF Rationale	0.604	0.604	0	0.656	0.653	0
		BM-25 Text+Cap.	0.610	0.610	0	0.654	0.653	0
		BM-25 Rationale	0.626	0.626	0	0.653	0.650	0

Table 4: Correct Classification - Case Study 1

Meme Text: life hack#23 how to get stoned with no weed Meme Caption: a young woman in a hijab kisses her mother on the cheek Label: Hateful

Table 5: Correct Classification - Case Study 2

Meme Text: when you have an 80 point iq but it's the highest in your village Meme Caption: a photograph of a woman with an afro Label: Hateful

Table 6: Incorrect Classification - Case Study 1

Meme Text: meanwhile in baltimore Meme Caption: a baboon mounting another baboon in the serengeti Label: Not Hateful

Table 7: Incorrect Classification - Case Study 2

Meme Text: you can smell disgust and hate from this photo Meme Caption: a black and white photo of two men standing next to each other Label: Not hateful

Table 8: Comparison of zero-shot and few-shot in-context learning experiment results with Latent Hatred support set across different demonstration sampling (Dem. Sampl.) strategies.

Model	# Shots	Dem. Samp.	FHM			MAMI		
	Matching		Acc.	F1	# Invalids	Acc.	F1	# Invalids
LLaVA-7B	0-shot	–	0.512	0.509	27	0.553	0.533	30
	4-shots	Random	0.592	0.576	0	0.606	0.559	0
		TF-IDF Text+Cap.	0.590	0.581	0	0.613	0.590	0
		TF-IDF Rationale	0.594	0.585	0	0.618	0.593	0
		BM-25 Text+Cap.	0.602	0.590	0	0.619	0.594	0
		BM-25 Rationale	0.588	0.575	0	0.635	0.608	0
	8-shots	Random	0.576	0.547	0	0.597	0.537	0
		TF-IDF Text+Cap.	0.592	0.582	0	0.603	0.627	0
		TF-IDF Rationale	0.594	0.581	0	0.634	0.611	0
		BM-25 Text+Cap.	0.612	0.599	0	0.636	0.611	0
		BM-25 Rationale	0.598	0.584	0	0.619	0.589	0
	16-shots	Random	0.576	0.547	0	0.583	0.514	0
		TF-IDF Text+Cap.	0.598	0.585	0	0.636	0.610	0
		TF-IDF Rationale	0.590	0.577	0	0.633	0.608	0
		BM-25 Text+Cap.	0.608	0.596	0	0.644	0.623	0
		BM-25 Rationale	0.596	0.578	0	0.622	0.594	0

Table 9: Comparison of zero-shot and few-shot in-context learning experiment results with FHM support set across different demonstration sampling (Dem. Sampl.) strategies.

Model	# Shots	Dem. Samp.	FHM			MAMI		
	Matching		Acc.	F1	# Invalids	Acc.	F1	# Invalids
LLaVA-7B	0-shot	–	0.512	0.509	27	0.553	0.533	30
	4-shots	Random	0.596	0.576	0	0.591	0.547	0
		TF-IDF Text+Cap.	0.578	0.554	0	0.611	0.581	0
		TF-IDF Rationale	0.594	0.571	0	0.621	0.600	0
		BM-25 Text+Cap.	0.576	0.551	0	0.626	0.599	0
		BM-25 Rationale	0.570	0.557	0	0.634	0.610	0
	8-shots	Random	0.594	0.575	0	0.600	0.556	0
		TF-IDF Text+Cap.	0.572	0.546	0	0.638	0.612	0
		TF-IDF Rationale	0.584	0.568	0	0.637	0.613	0
		BM-25 Text+Cap.	0.568	0.544	0	0.626	0.596	0
		BM-25 Rationale	0.584	0.573	0	0.635	0.616	0
	16-shots	Random	0.378	0.362	183	0.376	0.345	374
		TF-IDF Text+Cap.	0.426	0.393	113	0.509	0.480	208
		TF-IDF Rationale	0.416	0.389	150	0.486	0.447	232
		BM-25 Text+Cap.	0.420	0.395	146	0.453	0.422	279
		BM-25 Rationale	0.124	0.118	387	0.112	0.109	812

Table 10: Comparison of zero-shot and few-shot in-context learning experiment results with Latent Hatred support set across different demonstration sampling (Dem. Sampl.) strategies.

Model	# Shots	Dem. Samp.	FHM			MAMI		
	Matching		Acc.	F1	# Invalids	Acc.	F1	# Invalids
Llama3-8B	0-shot	–	0.614	0.586	7	0.634	0.596	5
	4-shots	Random	0.598	0.561	9	0.569	0.499	21
		TF-IDF Text	0.592	0.558	6	0.592	0.550	15
		TF-IDF Rationale	0.596	0.584	8	0.607	0.588	24
		BM-25 Text	0.592	0.550	15	0.628	0.606	14
		BM-25 Rationale	0.602	0.589	3	0.628	0.611	17
	8-shots	Random	0.612	0.579	3	0.592	0.531	1
		TF-IDF Text	0.600	0.571	2	0.601	0.559	2
		TF-IDF Rationale	0.608	0.599	17	0.629	0.609	20
		BM-25 Text	0.601	0.559	2	0.647	0.627	8
		BM-25 Rationale	0.576	0.564	17	0.626	0.613	24
	16-shots	Random	0.620	0.589	0	0.583	0.516	0
		TF-IDF Text	0.628	0.606	0	0.605	0.563	0
		TF-IDF Rationale	0.622	0.610	0	0.625	0.603	1
		BM-25 Text	0.605	0.563	0	0.663	0.647	0
		BM-25 Rationale	0.624	0.611	0	0.658	0.643	0

Table 11: Comparison of zero-shot and few-shot in-context learning experiment results with FHM support set across different demonstration sampling (Dem. Sampl.) strategies.

Model	# Shots	Dem. Samp.	FHM			MAMI		
	Matching		Acc.	F1	# Invalids	Acc.	F1	# Invalids
Llama3-8B	0-shot	–	0.614	0.586	7	0.634	0.596	5
	4-shots	Random	0.598	0.568	5	0.583	0.535	10
		TF-IDF Text	0.598	0.569	1	0.592	0.551	11
		TF-IDF Rationale	0.598	0.568	1	0.633	0.606	15
		BM-25 Text	0.606	0.574	2	0.602	0.564	14
		BM-25 Rationale	0.600	0.585	6	0.627	0.608	15
	8-shots	Random	0.576	0.550	40	0.525	0.487	110
		TF-IDF Text	0.564	0.535	29	0.546	0.518	137
		TF-IDF Rationale	0.566	0.545	26	0.547	0.526	131
		BM-25 Text	0.560	0.536	33	0.552	0.526	132
		BM-25 Rationale	0.592	0.578	28	0.599	0.581	82
	16-shots	Random	0.632	0.608	8	0.600	0.565	29
		TF-IDF Text	0.574	0.547	7	0.610	0.581	35
		TF-IDF Rationale	0.610	0.591	5	0.616	0.592	38
		BM-25 Text	0.590	0.563	4	0.614	0.590	40
		BM-25 Rationale	0.610	0.600	15	0.623	0.611	69