

# Evaluating Open-Source Vision-Language Models for Hateful Meme Detection

THESIS SUBMITTED

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE AWARD OF THE DEGREE  
OF

MASTER OF TECHNOLOGY  
IN  
**Artificial Intelligence**

Submitted by

**Vhatkar Ganesh Mallinath (23/AFI/16)**

Under the supervision of  
Asst.Prof. Gull Kaur



**Computer Science**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi 110042

**May, 2025**

**DEPARTMENT OF COMPUTER SCIENCE ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

**CANDIDATE’S DECLARATION**

I, **Vhatkar Ganesh Mallinath**, Roll No’s – **23/AFI/16** students of M.Tech (**Artificial Intelligence**), hereby declare that the project Dissertation titled “**Evaluating Open-Source Vision-Language Models for Hateful Meme Detection**” which is submitted by us to the **Department of Computer Science & Engineering** , Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi

Vhatkar Ganesh Mallinath

Date: 31.05.2025

**DEPARTMENT OF COMPUTER SCIENCE ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

**CERTIFICATE**

I hereby certify that the Project Dissertation titled “**Evaluating Open-Source Vision-Language Models for Hateful Meme Detection**” which is submitted by **Vhatkar Ganesh Mallinath**, Roll No's – **23/AFI/16, Artificial Intelligence**, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

**Asst.Prof. Gull Kaur**

Date: 31.05.2025

**SUPERVISOR**

**DEPARTMENT OF COMPUTER SCIENCES ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

**ACKNOWLEDGEMENT**

We wish to express our sincerest gratitude to **Asst.Prof. Gull Kaur** for her continuous guidance and mentorship that she provided us during the project. She showed us the path to achieve our targets by explaining all the tasks to be done and explained to us the importance of this project as well as its industrial relevance. She was always ready to help us and clear our doubts regarding any hurdles in this project. Without her constant support and motivation, this project would not have been successful.

Place: Delhi

Vhatkar Ganesh Mallinath

Date: 31.05.2025

# Abstract

Detecting hateful content in internet memes poses a unique challenge due to the tight coupling of visual and textual information. We present a systematic evaluation of five open-source vision-language models across three practical scenarios—zero-shot prompting, few-shot in-context learning, and parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA), all executed on freely available Kaggle T4 GPUs. Our zero-shot experiments highlight substantial performance swings driven by prompt design, emphasizing the need for careful prompt engineering. Introducing just two to four labeled examples in few-shot settings consistently improves classification, with top models exceeding 64% accuracy and macro-F1. Most notably, after only five epochs of LoRA fine-tuning, our best model delivers an AUROC of 85.81%, coming within 1.19 points of the state-of-the-art Retrieval-Guided Contrastive Learning benchmark (87.0% AUROC). By unifying evaluation protocols and demonstrating resource-aware methods, this work shows that near-state-of-the-art AUROC is achievable under tight computational constraints, making robust hateful meme detection more accessible for real-world moderation.

# Contents

Candidate’s Declaration	i
Certificate	ii
Acknowledgement	iii
Abstract	iv
Content	vi
List of Tables	vii
List of Figures	viii
<b>1 Introduction</b>	<b>1</b>
<b>2 Literature Review</b>	<b>3</b>
<b>3 Models and Dataset</b>	<b>4</b>
3.1 Models . . . . .	4
3.1.1 Qwen2-VL-2B-Instruct . . . . .	4
3.1.2 Qwen2.5-VL-3B-Instruct . . . . .	4
3.1.3 PaliGemma-3B-PT-224 . . . . .	4
3.1.4 Idefics3-8B-Llama3 . . . . .	4
3.1.5 LLaVA 1.6/LLaVA-Next-7B . . . . .	5
3.2 Hateful Memes Dataset . . . . .	5
3.2.1 Dataset Composition . . . . .	5
3.2.2 Dataset Splitting and Adjustment . . . . .	5
<b>4 Methodology</b>	<b>6</b>
4.0.1 Zero-shot prompting strategies . . . . .	7
4.0.2 Few-Shot Learning Approach . . . . .	7
4.0.3 Fine-Tuning Methodology . . . . .	8
4.0.4 Resource Constraints and Training Workflow . . . . .	9
<b>5 Results and Discussion</b>	<b>10</b>
5.1 Zero-Shot Results . . . . .	10
5.2 Few-Shot Results . . . . .	13
5.2.1 2-Shot Setting . . . . .	13
5.2.2 3-Shot Setting . . . . .	14
5.2.3 4-Shot Setting . . . . .	15

5.3	Fine-Tune Results . . . . .	16
5.4	Discussion . . . . .	18
5.5	Limitations . . . . .	19
<b>6</b>	<b>Conclusion, Future Scope and Social Impact</b>	<b>20</b>
6.1	Conclusion . . . . .	20
6.2	Future Scope . . . . .	20
6.3	Social Impact . . . . .	21
<b>A</b>	<b>Unified Results Table</b>	<b>22</b>

## List of Tables

3.1	Distribution of labels in the Hateful Memes dataset . . . . .	5
4.1	Zero-shot prompt templates used for hateful meme classification, detailing the prompt wording, input format, and expected model output. . . . .	8
5.1	Zero-shot results for Prompt 1 on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %). . . . .	10
5.2	Zero-shot results for Prompt 2 on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %). . . . .	11
5.3	Zero-shot results for Prompt 3 on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %). . . . .	12
5.4	Few-shot (2-shot) results for all models on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %). . . . .	13
5.5	Few-shot (3-shot) results for all models on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %). . . . .	14
5.6	Few-shot (4-shot) results for all models on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %). . . . .	15
5.7	LLaVA-1.6 / LLaVA-Next-7B classification results for few-shot settings (2-shot, 3-shot, 4-shot) on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %). The number of predictions for each class (Class 0 and Class 1) is also shown. . . . .	16
5.8	Fine-tuned results for all models on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %). . . . .	17
5.9	Comparison of SOTA methods and our fine-tuned models on the Facebook Hateful Memes Challenge dataset (dev set). SOTA results are from Mei et al. [1]; our results are from models fine-tuned with LoRA. . . . .	18



## List of Figures

4.1	Methodology pipeline for hateful meme detection. The pipeline includes zero-shot, few-shot, and LoRA fine-tuning approaches evaluated against state-of-the-art benchmarks. . . . .	6
4.2	LoRA fine-tuning process for hateful meme detection. . . . .	9
5.1	Accuracy of vision-language models in the zero-shot setting with Prompt 1 on the Hateful Memes dataset. The highest accuracy, achieved by IDEFICS-3-8B, is highlighted in orange; all other models are shown in blue. . . . .	10
5.2	Accuracy of vision-language models in the zero-shot setting with Prompt 2 on the Hateful Memes dataset. The highest accuracy, achieved by Qwen2-VL-2B, is highlighted in orange; all other models are shown in blue. . . . .	11
5.3	Accuracy of vision-language models in the zero-shot setting with Prompt 3 on the Hateful Memes dataset. The highest accuracy, achieved by Qwen2.5-VL-3B, is highlighted in orange; all other models are shown in blue. . . . .	13
5.4	Accuracy of vision-language models in the 2-shot setting on the Hateful Memes dataset. The highest accuracy, achieved by Qwen2-VL-2B, is highlighted in orange; all other models are shown in blue. . . . .	14
5.5	Accuracy of vision-language models in the 3-shot setting on the Hateful Memes dataset. The highest accuracy, achieved by IDEFICS-3-8B, is highlighted in orange; all other models are shown in blue. . . . .	15
5.6	Accuracy of vision-language models in the 4-shot setting on the Hateful Memes dataset. The highest accuracy, achieved by IDEFICS-3-8B, is highlighted in orange; all other models are shown in blue. . . . .	15
5.7	Accuracy of fine-tuned vision-language models on the Hateful Memes dataset. The highest accuracy, achieved by LLaVA-1.6-9B, is highlighted in orange; all other models are shown in blue. . . . .	17
5.8	AUROC of fine-tuned vision-language models on the Hateful Memes dataset. The highest AUROC, achieved by Qwen2-VL-2B, is highlighted in orange; all other models are shown in blue. . . . .	18

# Chapter 1

## Introduction

Memes, which are humorous or satirical collections of text and images that go viral on social media platforms like Facebook, Instagram, and Reddit, are a common way for people to communicate online. Despite the fact that most memes are intended to be humorous or educational, an increasing number of them are being used as a tool to spread hate, targeting individuals based on their gender, race, religion, or other characteristics. Because of this, identifying hateful memes has become extremely difficult for social media companies that are responsible for mass content moderation.

What makes hateful memes especially difficult to detect is their multimodal nature: the hateful intent often arises only when the visual and textual elements are interpreted together. It’s possible that a meme’s harmfulness cannot be determined solely by its image or text. Automated detection systems are complicated by this subtle interaction. Even humans have difficulties; annotators frequently disagree on edge cases involving sarcasm, cultural references, or coded language, and accuracy on the popular Facebook Hateful Memes dataset is only about 84.7% [2].

Recent advancements in vision-language models (VLMs) have enabled new approaches to multimodal content handling. The zero-shot capabilities of large VLMs in hate detection scenarios have been the subject of numerous studies; these studies have shown promise, but they also point out limitations in terms of prompt sensitivity and generalization [3]. These models can sometimes default to overgeneralized decisions—predicting all content as hateful or benign depending on prompt phrasing—making them unreliable for deployment without further adaptation.

At the same time, parameter-efficient fine-tuning techniques such as LoRA (Low-Rank Adaptation) have become popular in general vision-language research due to their ability to fine-tune large models with fewer resources [4, 5]. However, most such studies target tasks like image captioning, object detection, or visual question answering. Few-shot learning for hateful meme detection—where models are guided by only a handful of labeled examples—has only recently begun to be explored. A few notable efforts [5, 6] have shown that with strong prompting and task-specific examples, models can generalize better in low-resource hate speech settings. Yet, comprehensive fine-tuning evaluations using LoRA or adapter-based tuning on hateful memes remain rare.

This paper aims to bridge that gap by offering a unified evaluation of open-source VLMs across three resource-aware scenarios:

1. **Zero-shot classification**, using a range of prompting strategies;
2. **Few-shot learning**, using 2, 3, and 4 labeled examples;
3. **Parameter-efficient fine-tuning**, using LoRA under realistic constraints.

All tests are carried out on open-access GPUs (Kaggle T4s), which replicate environments that are available to independent researchers or nonprofit moderation teams. Through the application of this pragmatic, real-world perspective, our research offers a well-founded evaluation of how well existing VLMs can manage hateful meme detection with constrained data and computation.

The remainder of this paper is structured as follows. Section 2 reviews related work on multimodal hate speech detection and recent advances in vision-language models. Section 3 introduces the datasets and vision-language models used in our experiments. Section 4 describes the methodology in detail, including prompting strategies, few-shot configurations, and fine-tuning methodology using LoRA. Section 5 presents the results of our evaluations across zero-shot, few-shot, and fine-tuning scenarios, along with comparative analysis. Section 6 discusses the key findings, identifies limitations, and outlines directions for future research. Finally, Section 7 concludes the paper.

## Chapter 2

### Literature Review

Since Facebook released the Hateful Memes Challenge dataset [2], uni-modal approaches have proven inadequate. Early image-only models such as VGG-19 [7] and text-only models like BERT [8] quickly plateaued. It became evident that hateful intent often emerges only when visual and textual cues are combined [2, 9].

To address this multimodal complexity, a range of strategies has been proposed. Role-disentanglement methods separate the contributions of different meme entities [10]. Knowledge-augmented models enrich representations with external context—e.g., KnowMeme uses a knowledge-enriched graph neural network to detect offensive memes by linking image and text entities to a knowledge base [11]. Multi-task learning frameworks jointly optimize vision and language objectives, improving cross-modal feature sharing [12]. Data augmentation techniques expand meme variations to enhance generalization [13]. Ensemble methods that merge diverse classifiers have also shown gains [14, 15]. Meanwhile, prompt-based learning [16]—where a few labeled examples are prepended to inputs—has emerged as an effective low-resource strategy [17].

A powerful recent direction is retrieval-augmented modeling. Mei et al. demonstrate that dynamically retrieving related examples at inference time yields a hatefulness-aware embedding space, enabling models to adapt rapidly to new meme formats without full retraining [1].

The advent of large vision-language models (VLMs) and language models (LLMs) has further reshaped the field. Some work emphasizes explainability, generating human-readable rationales to guide classification [18]. Others focus on prompt engineering and few-shot in-context learning, showing that careful prompt design and example selection—e.g., zero-shot evaluation [3] and few-shot studies [5]—can significantly boost performance.

Despite these advances, most studies evaluate only a small set of models or configurations. There remains a clear need for a unified benchmark that systematically compares open-source VLMs for hateful meme detection across zero-shot, few-shot, and LoRA-based fine-tuning settings under a consistent protocol and resource constraints.

## Chapter 3

### Models and Dataset

#### 3.1 Models

##### 3.1.1 Qwen2-VL-2B-Instruct

We chose Qwen2-VL-2B-Instruct because it is both efficient and capable, especially in places where resources are limited. This model differs from larger ones by being able to analyze images of any resolution, a key point for examining memes that come in many sizes. By using both simulated instructions and real-world datasets for visual question-answering, the system was able to create detailed descriptions for images [19]. The ability to recognize text in several languages made it easier to understand memes with non-English text [20] [21].

##### 3.1.2 Qwen2.5-VL-3B-Instruct

Qwen2.5-VL-3B adds three billion parameters to its 2B-parameter predecessor, making it both large and efficient. It can take in both images and short videos, automatically adjusting the size from  $256 \times 256$  to  $1280 \times 1280$  without losing any of the important details found in memes. The model can read over 20 languages, including those with stylized or distorted fonts and its attention to coordinates allows it to place text directly over the right parts of the image, producing correct bounding boxes for each overlay [22] [23].

##### 3.1.3 PaliGemma-3B-PT-224

It is unique because it uses a SigLIP vision encoder together with a Gemma language model. Because the model works best on  $224 \times 224$  pixels, we had to adjust the meme dimensions before processing them. Using multilingual image-text examples, it became skilled at spotting offensive images and comparing them with text [24] [25].

##### 3.1.4 Idefics3-8B-Llama3

Idefics3 from Hugging Face is made by combining the SigLIP encoder for images and the Llama-3.1-8B model for language, focusing on mixed image-text data. This was especially useful for memes that have both text and images closely related such as when a distorted image is shown with sarcastic captions. It is different from other models because it handles these elements in the same order as they appear in speech [26].

### 3.1.5 LLaVA 1.6/LLaVA-Next-7B

In LLaVA 1.6, CLIP is used to analyze images and a 7 billion-parameter language backbone (Vicuna/Mistral) is chosen for prioritizing high-resolution analysis. Because of its improved OCR, it could accurately extract the text from memes, even when the fonts were low quality or fancy—something that is often a problem with hateful content. During evaluation, it was better at parsing memes that use small differences between the text and the image [27].

## 3.2 Hateful Memes Dataset

Facebook AI developed the Hateful Memes dataset to advance the study of identifying hate in multimodal internet memes. This dataset is unique because of its thoughtful design: each meme combines text and an image, and frequently, it is difficult to determine whether the content is hateful based just on the picture or the words. This makes the task much more realistic and difficult by forcing both humans and algorithms to take into account the combined context. The dataset contains a diverse range of examples, some of which are purposefully difficult to categorize, and each meme is classified as either hateful or not. Interestingly, the reported human accuracy on this task is 84.7% [2], meaning that even human annotators are not perfect.

### 3.2.1 Dataset Composition

To provide a fair and balanced evaluation, the Hateful Memes dataset is split into three parts: a training set, a development (dev) set, and a test set. Each split contains an equal number of hateful and non-hateful memes, so models are not biased toward either class. The table below summarizes how many examples fall into each category for every split.

Split	Total Samples	Hateful (Label 1)	Non-Hateful (Label 0)
Train	8,500	4,250	4,250
Dev	500	250	250
Test	1,000	500	500

Table 3.1: Distribution of labels in the Hateful Memes dataset

The test set labels are not publicly available; however, the dataset creators have indicated that the distribution mirrors that of the training and development sets [2].

### 3.2.2 Dataset Splitting and Adjustment

Since the official test set labels for the Hateful Memes dataset are not publicly available, we used the provided development (dev) set as our test set for reporting final results. To monitor model performance during training and to support early stopping, we further split the original training set, setting aside 10% as a validation set. This validation split was used to evaluate the model after each training epoch and to determine the optimal stopping point by using Early stop, which help us to avoid overfitting. All hyperparameter tuning and model selection decisions were based solely on performance on this held-out validation set.

## Chapter 4

### Methodology

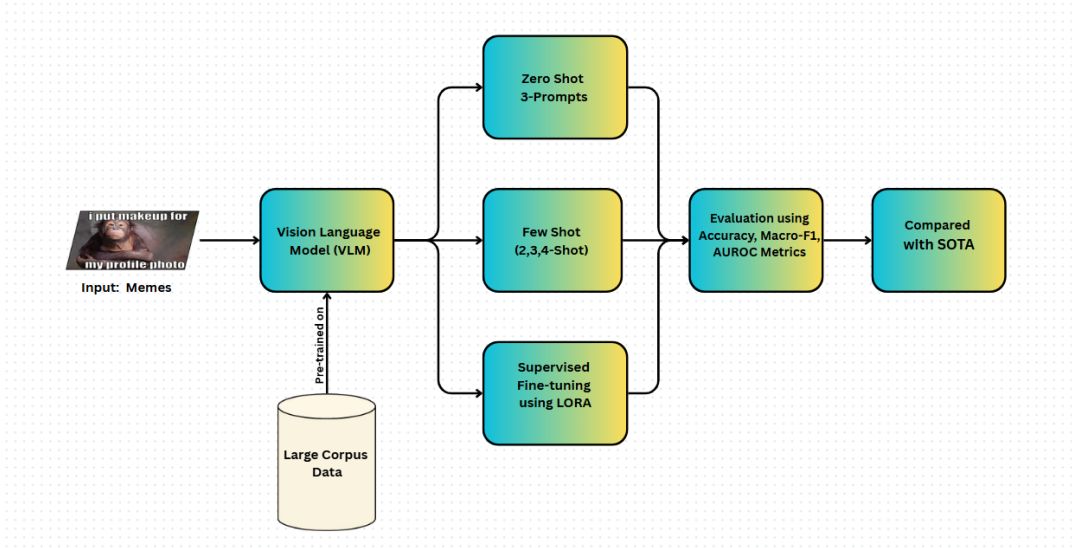


Figure 4.1: Methodology pipeline for hateful meme detection. The pipeline includes zero-shot, few-shot, and LoRA fine-tuning approaches evaluated against state-of-the-art benchmarks.

The approach taken for hateful meme detection in this research is presented in Figure 4.1. Each meme serves as input to a collection of pre-trained vision-language models: *Qwen2-VL-2B-Instruct*, *Qwen2.5-VL-3B-Instruct*, *LLaVA-1.6/LLaVA-Next-7B*, *PaliGemma-3B-pt-224*, and *IDEFICS-3-8B*. Three distinct strategies are used to evaluate and enhance these models for the detection task:

#### Zero-Shot Prompting

Here, we examine how well the model can classify hateful memes when given no prior training on this specific task. The model works with just the prompt and the meme itself. Section 4.0.1 contains a thorough discussion of the zero-shot prompting techniques.

#### Few-Shot Learning

This strategy explores whether giving the model several labeled examples improves its understanding of the task. We provide a small collection of reference memes to see how

this affects the model’s ability to make accurate classifications. The complete few-shot learning approach is outlined in Section 4.0.2.

## LoRA Fine-Tuning

In this phase, we apply efficient fine-tuning through Low-Rank Adaptation (LoRA). This technique modifies only certain parameters, allowing us to enhance performance while working within computational limits. Section 4.0.3 provides a thorough explanation of the LoRA fine-tuning process.

### 4.0.1 Zero-shot prompting strategies

For our zero-shot experiments, we wanted to see how the models would handle the task with different levels of guidance. We decided on three kinds of prompts. All prompt types are summarized in Table 4.1.

1. **Basic Prompt**

The first version keeps things straightforward: the model is instructed to be a meme classification expert and receives simply the picture without further direction. All it has to do is indicate "0" for non-hateful or "1" for hateful. This gauges whether a model can make the call with just the fundamentals.

2. **Elaborated Prompt (Hate Definition + Image)**

We then tested giving the model a little more to work with. Along with the image, the prompt now spells out what we mean by “hateful”—mentioning things like racism, sexism, or violence. The goal here is to investigate if a clear description helps the model focus on the proper cues, even if it does not understand the meme’s language.

3. **Full Context Prompt (Hate Definition + Image + Text)**

Finally, we combined everything: the model gets the image, the overlaid text, and a detailed definition of hateful content (including things like threats or dehumanizing language). This configuration is the most similar to the manner in which a human moderator would approach the issue, utilizing all available information to reach a conclusion.

### 4.0.2 Few-Shot Learning Approach

In our few-shot experiments, we wanted to see if giving the models a handful of real examples would help them catch on to the task more quickly. For each prompt, we included not just the meme image and its text, but also a clear definition of what we considered “hateful”—the same as in our most detailed zero-shot setup. The big difference here is that, before asking the model to classify anything new, we showed it some labeled memes as a reference.

We tried out three setups:

- **2-shot:** We picked one meme labeled as hateful and another labeled as non-hateful, then let the model see both before moving on to the test examples.
- **3-shot:** We gave the model two hateful memes and one non-hateful meme, all with their labels.



- **4-shot:** This setup included two examples from each class, so the model saw two hateful and two non-hateful memes, each clearly marked.

Throughout all these experiments, the structure of the prompt stayed the same; only the number and mix of examples changed. This approach let us check whether having a few concrete references would help the models do a better job on memes they hadn’t seen before.

Prompt Type	Prompt Text	Input Format	Expected Output
Basic Prompt	You are a meme classification expert. Classify the meme as 1 if it is hateful, or 0 if it is not hateful. Respond with only the number.	Prompt + meme image	1 or 0
Elaborate Prompt	You are a meme classification expert. For each meme, decide whether it is ‘hateful’ or ‘non-hateful’. A meme is hateful if it contains racism, sexism, misogyny, violence, or other discriminatory content. Respond with only the number 1 (for hateful) or 0 (for non-hateful). No explanation. No extra text.	Prompt + meme image	1 or 0
Full Context Prompt	You are optimized for harmful meme detection. Analyze the image and the accompanying text. Classify the meme as hateful if it contains racism, sexism, dehumanizing language, threats, or other forms of hate. Otherwise, classify it non-hateful. Respond ONLY with 1 (hateful) or 0 (non-hateful). No explanation. No extra text.	Prompt + meme image + meme text	1 or 0

Table 4.1: Zero-shot prompt templates used for hateful meme classification, detailing the prompt wording, input format, and expected model output.

### 4.0.3 Fine-Tuning Methodology

When it came to fine-tuning, We wanted to keep things straightforward and consistent for every model we worked with. For each one, we used the Hateful Memes dataset. The input combined the meme image, its text, and a simple prompt that asked whether the meme was hateful or not. Instead of retraining the whole model, we leaned on LoRA, as shown in Figure 4.2. This approach let us update just a few specific parameters and leave the rest untouched. We focused our adjustments on the query and value projection layers—those are `q_proj` and `v_proj`. For LoRA itself, we went with a rank of 8 and set alpha to 16. we also decided on a dropout rate of 0.1 and skipped adding any bias terms. Everything else in the model stayed frozen, so only those LoRA layers actually changed as training went on.

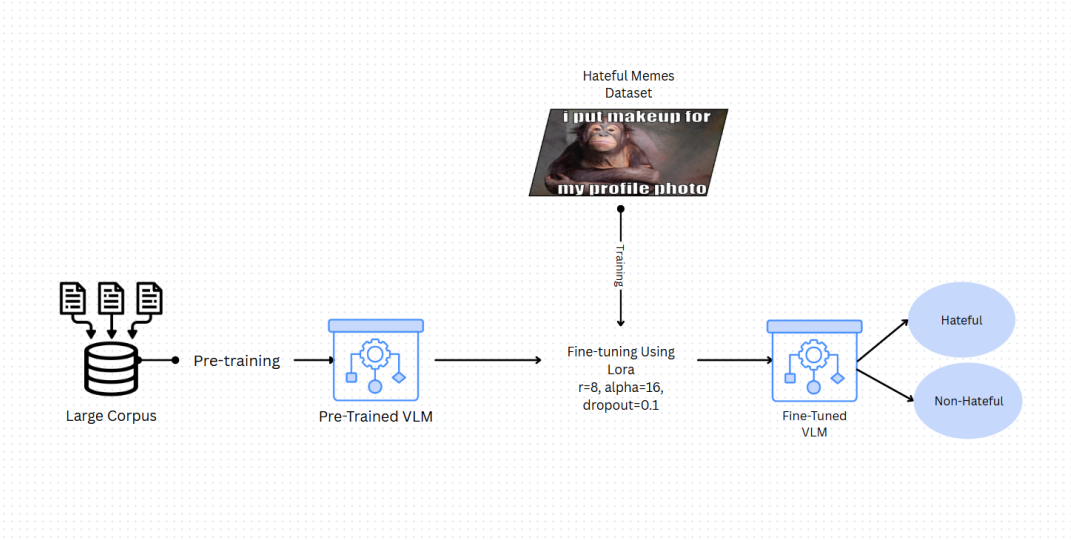


Figure 4.2: LoRA fine-tuning process for hateful meme detection.

For optimization, we picked AdamW as my optimizer, Because AdamW consistently outperforms traditional Adam in fine-tuning scenarios, particularly for transformer-based vision-language models, it was selected as our optimization algorithm. The optimizer’s decoupled weight decay mechanism provides superior regularization and training stability compared to standard Adam, which is crucial when fine-tuning large pre-trained models with limited computational resources. Additionally, AdamW has become the standard optimizer for large-scale deep learning models and works effectively with mixed-precision training, making it the optimal choice for our parameter-efficient fine-tuning approach using LoRA. and set the learning rate to  $1e-4$ . Since us GPU memory was limited, We kept the batch size small, just two samples at a time and made use of gradient accumulation to help fit things in memory. We also turned on mixed-precision training (using bfloat16 or float16, depending on the model) to save a bit more space. Our plan was to train for up to five epochs, but in practice, Qwen 2.5 VL stopped early as validation set accuracy worsened for further epoch. After each epoch, we saved a checkpoint and always kept the one that performed best on the validation set for final evaluation. We followed this same process for every model, only making small changes if a particular model needed a different input format.

#### 4.0.4 Resource Constraints and Training Workflow

All of this was done using Kaggle’s free GPU resources, which come with some strict limits: 30 hours of GPU time per week and a maximum of 12 hours per session on T4 x2 GPUs. Because these models are pretty large, we usually managed only one or two epochs in a single session. To keep things moving, I saved my progress at the end of each session and loaded up the latest checkpoint when starting a new one. This routine helped me make steady progress, even with the time and hardware limits that come with using a free platform.

# Chapter 5

## Results and Discussion

### 5.1 Zero-Shot Results

We evaluated several vision-language models (VLMs) for hateful meme detection in a zero-shot setting, using a range of prompt styles for each model. Our goal was to see how well these models could classify memes as hateful or non-hateful without any extra training or fine-tuning.

Model	Accuracy	Macro-F1	AUROC
Qwen2-VL-2B-Instruct	53.80	43.08	53.80
Qwen2.5-VL-3B-Instruct	<i>58.40</i>	<i>56.44</i>	<i>58.40</i>
LLaVA-1.6 / LLaVA-Next-7B	50.00	33.33	50.00
PaliGemma-3B-pt-224	49.40	34.09	49.40
IDEFICS-3-8B	<b>62.20</b>	<b>61.02</b>	<b>62.20</b>

Table 5.1: Zero-shot results for Prompt 1 on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %).

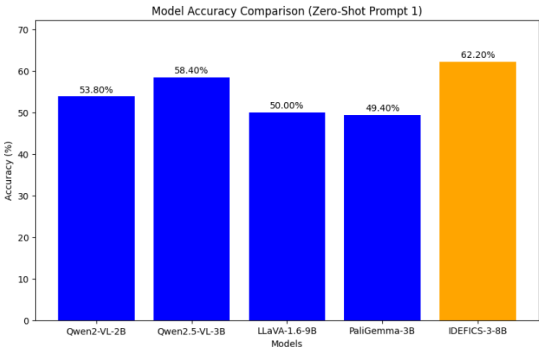


Figure 5.1: Accuracy of vision-language models in the zero-shot setting with Prompt 1 on the Hateful Memes dataset. The highest accuracy, achieved by IDEFICS-3-8B, is highlighted in orange; all other models are shown in blue.

Model	Accuracy	Macro-F1	AUROC
Qwen2-VL-2B-Instruct	<b>61.80</b>	<b>61.49</b>	<b>61.80</b>
Qwen2.5-VL-3B-Instruct	<i>61.00</i>	<i>60.33</i>	<i>61.00</i>
LLaVA-1.6 / LLaVA-Next-7B	50.00	33.33	50.00
PaliGemma-3B-pt-224	49.80	33.94	49.80
IDEFICS-3-8B	52.00	49.00	52.00

Table 5.2: Zero-shot results for Prompt 2 on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %).

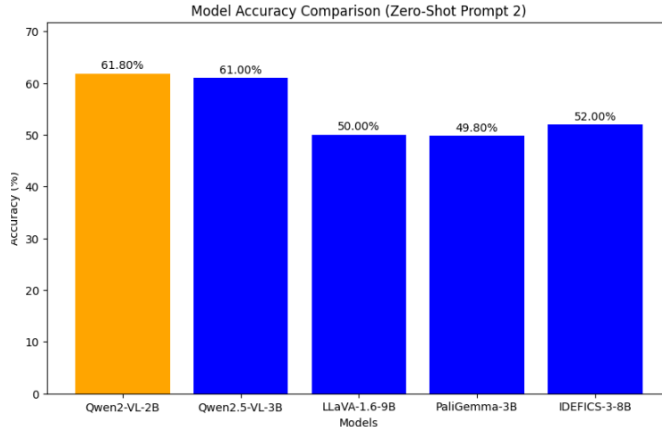


Figure 5.2: Accuracy of vision-language models in the zero-shot setting with Prompt 2 on the Hateful Memes dataset. The highest accuracy, achieved by Qwen2-VL-2B, is highlighted in orange; all other models are shown in blue.

For LLaVA, we noticed that the results depended heavily on how we wrote the prompt. When we used prompts with a detailed explanation of what hate means, LLaVA almost always predicted every meme as hateful, regardless of the actual content. This led to poor results, since the model simply picked one answer for everything and failed to distinguish between hateful and non-hateful memes. However, when we switched to a short and simple prompt—just asking the model to classify the meme as hateful or not, and providing both the image and meme text—LLaVA produced a more balanced set of predictions and performed noticeably better. With this neutral prompt, LLaVA labeled 323 memes as non-hateful and 177 as hateful, achieving an accuracy of 61.0%, macro F1 of 60.2%, and AUROC of 61.0%. This strong sensitivity to prompts has been previously observed in the literature [3, 4, 6].

IDEFICS 3-8B showed a related but slightly different pattern. When we used prompts with elaborate hate definitions (Prompt 2 and Prompt 3), the model mostly predicted memes as non-hateful: 466 out of 500 for Prompt 2, and 479 out of 500 for Prompt 3. Only when we used the basic prompt (Prompt 1) did IDEFICS give a more reasonable output, labeling 337 memes as non-hateful and 163 as hateful. This indicates that too much detail in the prompt can cause the model to default to one label, often missing actual hateful content and reducing its usefulness for real-world moderation. This pattern is consistent with findings from previous studies [2, 4, 6].

PaliGemma-3B-PT-224 had the hardest time in our tests. For Prompt 1, it labeled 491

memes as non-hateful and only 9 as hateful. For Prompt 2 and Prompt 3, it predicted all 500 memes as non-hateful. Regardless of the prompt used, PaliGemma’s accuracy remained close to random chance (49–50%), with macro F1 also low. Most notably, instead of providing a simple class label or a relevant explanation, PaliGemma often generated random or unrelated text, even when given clear, well-structured prompts. These instruction-following struggles made it unusable for zero-shot hateful meme detection in our experiments [3, 6].

When we evaluated Qwen2-VL-2B-Instruct and Qwen2.5-VL-3B-Instruct, we found these models were more robust to prompt changes. For Qwen2-VL-2B-Instruct, the best zero-shot results were with Prompt 2, giving an accuracy of 61.80%, macro F1 of 61.49%, and AUROC of 61.80%. The lowest results were with Prompt 1, where accuracy was 53.80%, macro F1 was 43.08%, and AUROC was 53.80%. For Qwen2.5-VL-3B-Instruct, the best results were with Prompt 3, which gave an accuracy of 63.40%, macro F1 of 63.02%, and AUROC of 63.40%. These models gave a more balanced mix of labels and achieved the highest zero-shot accuracy and macro F1 across most prompt styles. Still, we observed that careful prompt tuning could lead to even better results, demonstrating the importance of prompt engineering for optimal performance [2, 5, 6].

Overall, our experiments show that open-source VLMs are highly sensitive to prompt wording in the zero-shot setting. Using detailed definitions or explanations in the prompt can cause models to pick the same label for all inputs, while simple, direct prompts work better. Qwen models were the most robust, but even they improved with better prompt choices. Nevertheless, none of the models delivered strong, reliable results without further training. These findings align with recent studies on prompt sensitivity and model robustness [2, 3, 4, 6].

Model	Accuracy	Macro-F1	AUROC
Qwen2-VL-2B-Instruct	<i>57.60</i>	<i>50.79</i>	<i>57.60</i>
Qwen2.5-VL-3B-Instruct	<b>63.40</b>	<b>63.02</b>	<b>63.40</b>
LLaVA-1.6 / LLaVA-Next-7B	50.00	33.33	50.00
PaliGemma-3B-pt-224	50.20	36.08	50.20
IDEFICS-3-8B	51.40	38.50	51.40

Table 5.3: Zero-shot results for Prompt 3 on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %).

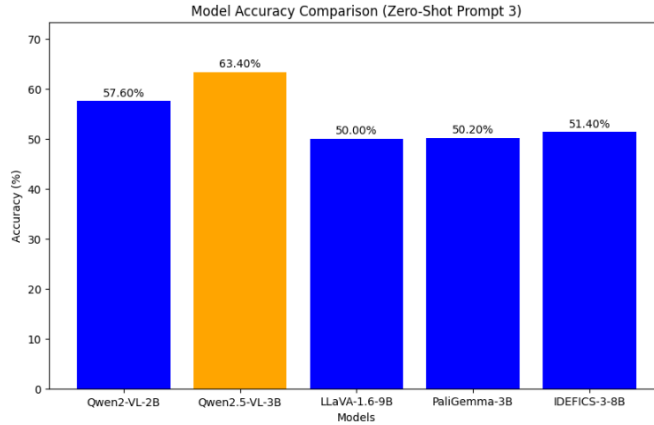


Figure 5.3: Accuracy of vision-language models in the zero-shot setting with Prompt 3 on the Hateful Memes dataset. The highest accuracy, achieved by Qwen2.5-VL-3B, is highlighted in orange; all other models are shown in blue.

## 5.2 Few-Shot Results

### 5.2.1 2-Shot Setting

When we gave each model just two labeled examples per class, Qwen2-VL-2B-Instruct came out on top, with an accuracy of 64.20%, a macro F1 of 64.18%, and an AUROC of 64.20%. IDEFICS 3-8B was a close second, posting 63.40% accuracy, a macro F1 of 62.51%, and an AUROC of 63.40%. Qwen2.5-VL-3B-Instruct also did well, reaching 62.80% accuracy, a macro F1 of 62.56%, and an AUROC of 62.80%. LLaVA 1.6 / LLaVA-Next-7B showed improvement compared to zero-shot, achieving 58.80% accuracy, a macro F1 of 57.48%, and an AUROC of 58.80%. PaliGemma-3B-PT-224 remained close to random guessing, with 50.60% accuracy, a macro F1 of 47.74%, and an AUROC of 50.60%, as shown in Table 5.4. The overall accuracy comparison for all models in this 2-shot setting is visualized in Figure 5.4.

Model	Accuracy	Macro-F1	AUROC
Qwen2-VL-2B-Instruct	<b>64.20</b>	<b>64.18</b>	<b>64.20</b>
Qwen2.5-VL-3B-Instruct	62.80	62.56	62.80
LLaVA-1.6 / LLaVA-Next-7B	58.80	57.48	58.80
PaliGemma-3B-pt-224	50.60	47.74	50.60
IDEFICS-3-8B	63.40	62.51	63.40

Table 5.4: Few-shot (2-shot) results for all models on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %).

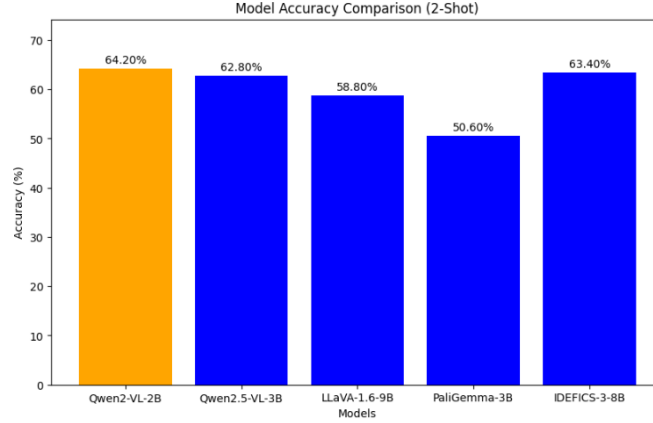


Figure 5.4: Accuracy of vision-language models in the 2-shot setting on the Hateful Memes dataset. The highest accuracy, achieved by Qwen2-VL-2B, is highlighted in orange; all other models are shown in blue.

### 5.2.2 3-Shot Setting

With three labeled examples per class, IDEFICS 3-8B led again, hitting 63.40% accuracy, a macro F1 of 61.43%, and an AUROC of 63.40%. Its predictions were fairly balanced, with 363 memes marked as non-hateful and 137 as hateful. Qwen2.5-VL-3B-Instruct followed, with 61.60% accuracy, a macro F1 of 59.17%, and an AUROC of 61.60%. Qwen2-VL-2B-Instruct was close behind, with 61.00% accuracy, a macro F1 of 59.27%, and an AUROC of 61.00%. LLaVA 1.6 / LLaVA-Next-7B continued to improve, reaching 59.40% accuracy, a macro F1 of 58.97%, and an AUROC of 59.4%. PaliGemma-3B-pt-224 still lagged, with 47.60% accuracy, a macro F1 of 47.44%, and an AUROC of 47.60%, as shown in Table 5.5. The results are also visualized in Figure 5.5.

Model	Accuracy	Macro-F1	AUROC
Qwen2-VL-2B-Instruct	61.00	59.27	61.00
Qwen2.5-VL-3B-Instruct	61.60	59.17	61.60
LLaVA-1.6 / LLaVA-Next-7B	59.40	58.97	59.40
PaliGemma-3B-pt-224	47.60	47.44	47.60
IDEFICS-3-8B	<b>63.40</b>	<b>61.43</b>	<b>63.40</b>

Table 5.5: Few-shot (3-shot) results for all models on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %).

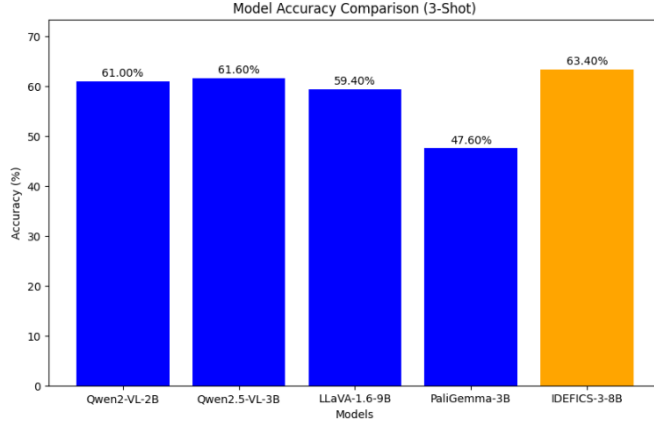


Figure 5.5: Accuracy of vision-language models in the 3-shot setting on the Hateful Memes dataset. The highest accuracy, achieved by IDEFICS-3-8B, is highlighted in orange; all other models are shown in blue.

### 5.2.3 4-Shot Setting

Model	Accuracy	Macro-F1	AUROC
Qwen2-VL-2B-Instruct	56.80	53.12	56.80
Qwen2.5-VL-3B-Instruct	<i>63.40</i>	<i>62.56</i>	<i>63.40</i>
LLaVA-1.6 / LLaVA-Next-7B	59.40	58.56	59.40
PaliGemma-3B-pt-224	47.40	46.93	47.40
IDEFICS-3-8B	<b>63.60</b>	<b>63.17</b>	<b>63.60</b>

Table 5.6: Few-shot (4-shot) results for all models on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %).

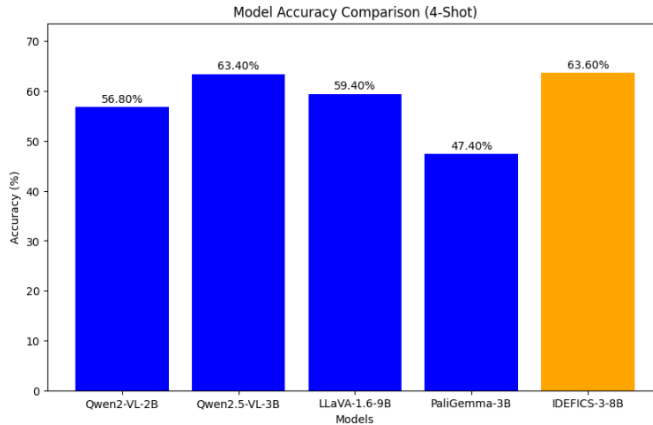


Figure 5.6: Accuracy of vision-language models in the 4-shot setting on the Hateful Memes dataset. The highest accuracy, achieved by IDEFICS-3-8B, is highlighted in orange; all other models are shown in blue.



When we increased the number of examples to four per class, IDEFICS 3-8B once again led the group, with 63.60% accuracy, a macro F1 of 63.17%, and an AUROC of 63.60%. The class split was 304 non-hateful and 196 hateful memes, showing a good balance. Qwen2.5-VL-3B-Instruct matched IDEFICS in both accuracy and AUROC at 63.40%, and its macro F1 was 62.56%. LLaVA 1.6 / LLaVA-Next-7B held steady at 59.40% accuracy, a macro F1 of 58.56%, and an AUROC of 59.4%. Qwen2-VL-2B-Instruct saw its performance drop, with 56.80% accuracy, a macro F1 of 53.12%, and an AUROC of 56.80%. PaliGemma-3b-pt-224 continued to struggle, with 47.40% accuracy, a macro F1 of 46.93%, and an AUROC of 47.40%, as shown in Table 5.6.

Looking across all few-shot settings, IDEFICS 3-8B and Qwen2.5-VL-3B-Instruct were the most reliable, with both accuracy and macro F1 above 61%. Qwen2-VL-2B-Instruct did best with two examples but dropped off as more were added. We can see that LLaVA improved over its zero-shot results: while it previously defaulted to one label, with a few labeled examples it started to actually infer the meme’s content, as shown in the class distribution in Table 5.7. Still, LLaVA didn’t reach the top performers. PaliGemma did not benefit from more examples and stayed close to random throughout. These findings show that few-shot learning can make a real difference for some vision-language models, especially when using a good prompt and a handful of labeled memes.

The overall accuracy comparison in the 4-shot setting is visualized in Figure 5.6, where the highest accuracy, achieved by IDEFICS-3-8B, is highlighted in orange and all other models are shown in blue.

Shots	Accuracy (%)	Macro-F1 (%)	AUROC (%)	Class 0	Class 1
2-shot	58.8	57.5	58.8	338	162
3-shot	59.4	58.9	59.4	301	199
4-shot	59.4	58.6	59.4	321	179

Table 5.7: LLaVA-1.6 / LLaVA-Next-7B classification results for few-shot settings (2-shot, 3-shot, 4-shot) on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %). The number of predictions for each class (Class 0 and Class 1) is also shown.

### 5.3 Fine-Tune Results

We saw clear gains in model performance after fine-tuning each vision-language model for five epochs on the Hateful Memes dataset. LLaVA 1.6 / LLaVA-Next-7B led the group, posting an accuracy of 73.40%, a macro F1 of 72.75%, and an AUROC of 84.32%. Qwen2-VL-2B-Instruct was just behind, with 73.20% accuracy, a macro F1 of 72.22%, and the highest AUROC among our models at 85.82%. Qwen2.5-VL-3B-Instruct also performed strongly, ending up with 70.16% accuracy, a macro F1 of 70.02%, and an AUROC of 77.83%.

PaliGemma-3B-PT-224 and IDEFICS 3-8B improved compared to their few-shot and zero-shot settings but still lagged behind the top performers. PaliGemma reached 67.00% accuracy, a macro F1 of 66.24%, and an AUROC of 73.00%, while IDEFICS 3-8B achieved 66.40% accuracy, 65.91% macro F1, and 73.77% AUROC, as shown in Table 5.8.

When we compare these results to the current state-of-the-art, Retrieval-Guided Contrastive Learning (RGCL) achieves an AUROC of 87.0% and an accuracy of 78.8% on the Hateful Memes dataset as shown in Table ?? . RGCL’s retrieval-based approach is especially effective at distinguishing subtle meme content and adapts quickly to new examples [1]. Our best models, LLaVA and Qwen2-VL-2B-Instruct, nearly match RGCL’s AUROC, coming within just a couple of percentage points. It is worth noting that we only trained for five epochs and did not use any retrieval-based techniques. With more time for fine-tuning or by adjusting more parameters such as raising the LoRA rank or unfreezing additional layers, there is a good chance these models could close the gap even further, or perhaps even outperform RGCL. Unfortunately, we were limited by available computing resources, so we could not extend training or experiment with higher LoRA ranks. For Qwen2.5-VL-3B-Instruct, early stopping was enabled and training stopped after four epochs when validation accuracy plateaued, rather than completing all five planned epochs.

Model	Accuracy	Macro-F1	AUROC
Qwen2-VL-2B-Instruct	<i>73.20</i>	<i>72.22</i>	<b>85.82</b>
Qwen2.5-VL-3B-Instruct	70.16	70.02	77.83
LLaVA-1.6 / LLaVA-Next-7B	<b>73.40</b>	<b>72.75</b>	<i>84.32</i>
PaliGemma-3B-pt-224	67.00	66.24	73.00
IDEFICS-3-8B	66.40	65.91	73.77

Table 5.8: Fine-tuned results for all models on the Hateful Memes dataset. Metrics reported are accuracy, macro-F1, and AUROC (all in %).

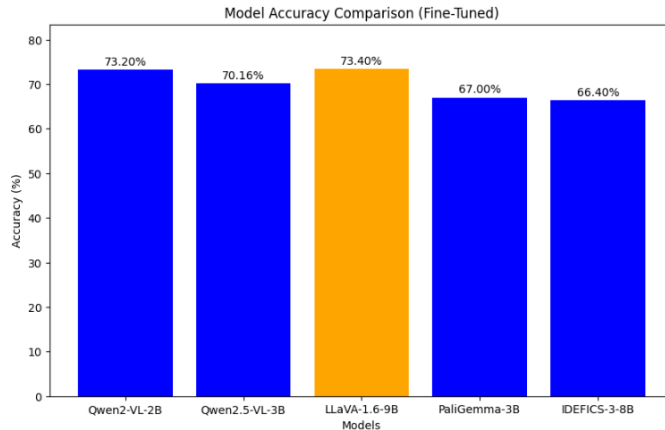


Figure 5.7: Accuracy of fine-tuned vision-language models on the Hateful Memes dataset. The highest accuracy, achieved by LLaVA-1.6-9B, is highlighted in orange; all other models are shown in blue.

These results highlight how much task-specific fine-tuning can improve hateful meme detection. With just a moderate training effort, open-source models like LLaVA and Qwen2-VL-2B-Instruct can come very close to state-of-the-art performance. The detailed results for all models are shown in Table 5.8. The accuracy of the fine-tuned models is visualized in Figure 5.7, where the highest accuracy, achieved by LLaVA-1.6-9B, is

Model/Method	Accuracy	AUROC
RGCL [1]	78.80%	87.00%
Your LLaVA 1.6 (fine-tuned)	73.40%	84.32%
Your Qwen2-VL-2B (fine-tuned)	73.20%	85.82%
Your Qwen2.5-VL-3B (fine-tuned)	70.16%	77.83%

Table 5.9: Comparison of SOTA methods and our fine-tuned models on the Facebook Hateful Memes Challenge dataset (dev set). SOTA results are from Mei et al. [1]; our results are from models fine-tuned with LoRA.

highlighted in orange and all other models are shown in blue. The AUROC comparison is shown in Figure 5.8, with Qwen2-VL-2B achieving the highest AUROC.

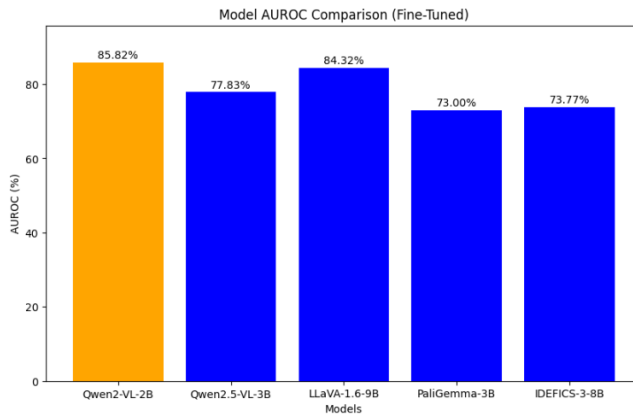


Figure 5.8: AUROC of fine-tuned vision-language models on the Hateful Memes dataset. The highest AUROC, achieved by Qwen2-VL-2B, is highlighted in orange; all other models are shown in blue.

## 5.4 Discussion

In our study, we found that large vision-language models work well for hateful meme detection, especially after fine-tuning them for this specific task. One thing that stood out was how much the wording of prompts mattered in zero-shot tests. If we used prompts with detailed definitions of hate, models like LLaVA and IDEFICS often just picked one label for everything, which limited their usefulness. Simpler prompts led to more balanced results, but even then, zero-shot performance could not match what we achieved with fine-tuning.

Giving the models just a few labeled examples made a real difference. For example, Qwen2.5-VL-3B-Instruct and IDEFICS 3-8B both improved when they saw a handful of examples for each class. This means that even with a small amount of labeled data, teams can boost performance, which is good news for those with limited resources.

Fine-tuning brought the biggest improvements. When we trained the models specifically for hateful meme detection, we saw the highest accuracy, macro F1, and AUROC. LLaVA-9B, being much larger, generally performed better than Qwen2-VL-2B-Instruct, but it also needed more memory and time. Qwen2-2B still did well and was more efficient.

Using LoRA helped us keep training costs down, though big models still required a lot of resources.

It is encouraging to see that open-source models like LLaVA and Qwen2, once fine-tuned, nearly matched the results of more complex systems like RGCL, which use retrieval-augmented learning. This means high-quality hateful meme detection is now more accessible to researchers and practitioners.

## 5.5 Limitations

Of course, our work has some limitations. We were restricted by the amount of computing power we had, so we could only fine-tune for up to five epochs, and sometimes less, as with Qwen2.5 due to early stopping. We also kept the LoRA rank fixed because of hardware limits. With more resources, we could have tried longer training or higher LoRA ranks, which might have led to even better results.

Another point is that we only used the Facebook Hateful Memes dataset. Despite being a widely used standard, it does not account for all of the hateful memes that may be found on other platforms or in other cultures. Additionally, we do not know how well our models would perform on memes with mixed language content or in other languages because we only tested English-language memes.

We did not investigate whether there were any recurring biases or the reasons behind the models' predictions. Building just and efficient systems requires an understanding of these problems. Even though we experimented with various prompt styles, it is still difficult to identify the best ones. Future research could examine automated methods of prompt optimization.

## Chapter 6

### Conclusion, Future Scope and Social Impact

#### 6.1 Conclusion

This paper explored how open-source vision-language models can be used to detect hateful memes. We tested several models in different scenarios, including zero-shot, few-shot, and after fine-tuning, to see how well they could handle this challenging task. Our experiments showed that the way prompts are written and the availability of even a small number of labeled examples can make a big difference in performance. Fine-tuning, in particular, led to the best results. Notably, models such as LLaVA and Qwen2, once adapted for this task, performed nearly as well as more complex, state-of-the-art systems. We hope these findings offer useful guidance for others working on multimodal hate detection and help make robust solutions more widely available.

#### 6.2 Future Scope

Looking forward, there are several ways to build on this research. With more computing power, it would be useful to experiment with longer fine-tuning schedules or higher LoRA ranks. Another promising direction is to build a technique on top of a general vision-language model (VLM), so the system can handle new memes without needing to retrain the model repeatedly. For example, approaches like LMM-RGCL demonstrate how retrieval-guided learning can be layered on top of a VLM. By pulling in similar examples during inference, these systems can adapt to new memes as they appear, without retraining from scratch. This would make detection systems more robust as online content changes.

To gain a better understanding of these models' generalization, it would also be beneficial to test them on a larger variety of datasets, languages, and meme styles. These systems would become more reliable with the development of better tools for identifying bias and explaining model decisions. Finally, figuring out how to automate prompt optimization might make few-shot and zero-shot learning even more useful in the real world.

## 6.3 Social Impact

Hateful memes often slip under the radar by hiding harmful messages in humor—but our work shows you don’t need a supercomputer to fight back. Fine-tuning an open-source vision-language model on just two T4 GPUs yielded an AUROC of 85.81%—only 1.19 points below the 87% state-of-the-art—while keeping both budget and energy use low. Small teams, ranging from academic researchers to nonprofit moderators, can now install trustworthy filters that stop hateful images before they spread, shielding vulnerable groups from repeated exposure. Quickly eliminating harmful content also makes it more difficult for malicious actors to hide their bigotry in jokes, encouraging online discourse to move toward respect rather than mockery.

There’s an environmental win, too: using compact GPU clusters cuts electricity needs and the millions of liters of cooling water that large data centers require. By lowering both financial and ecological hurdles, this approach empowers a wider range of stakeholders to keep digital spaces inclusive, responsible, and green.

# Appendix A

## Unified Results Table

Model	Setting	Accuracy (%)	Macro-F1 (%)	AUROC (%)
Qwen2-VL-2B-Instruct	Zero-Shot P1	53.80	43.08	53.80
	Zero-Shot P2	61.80	61.49	61.80
	Zero-Shot P3	57.60	50.79	57.60
	Fine-Tuned (5 ep)	73.20	72.22	85.82
	2-Shot	64.20	64.18	64.20
	3-Shot	61.00	59.27	61.00
	4-Shot	56.80	53.12	56.80
Qwen2.5-VL-3B-Instruct	Zero-Shot P1	58.40	56.44	58.40
	Zero-Shot P2	61.00	60.33	61.00
	Zero-Shot P3	63.40	63.02	63.40
	Fine-Tuned (5 ep)	70.16	70.02	77.83
	2-Shot	62.80	62.56	62.80
	3-Shot	61.60	59.17	61.60
	4-Shot	63.40	62.56	63.40
LLaVA-1.6 / LLaVA-Next-7B	Zero-Shot P1	50.00	33.33	50.00
	Zero-Shot P2	50.00	33.33	50.00
	Zero-Shot P3	50.00	33.33	50.00
	Fine-Tuned (5 ep)	73.40	72.75	84.32
	2-Shot	58.80	57.48	58.80
	3-Shot	59.40	58.97	59.40
	4-Shot	59.40	58.56	59.40
PaliGemma-3B-pt-224	Zero-Shot P1	49.40	34.09	49.40
	Zero-Shot P2	49.80	33.94	49.80
	Zero-Shot P3	50.20	36.08	50.20
	Fine-Tuned (5 ep)	67.00	66.24	73.00
	2-Shot	50.60	47.74	50.60
	3-Shot	47.60	47.44	47.60
	4-Shot	47.40	46.93	47.40

**Table A.1:** Unified summary of experimental results for Qwen2, Qwen2.5, LLaVA, and PaliGemma models in all settings.

Model	Setting	Accuracy (%)	Macro-F1 (%)	AUROC (%)
IDEFICS-3-8B	Zero-Shot P1	62.20	61.02	62.20
	Zero-Shot P2	52.00	49.00	52.00
	Zero-Shot P3	51.40	38.50	51.40
	Fine-Tuned (5 ep)	66.40	65.91	73.77
	2-Shot	63.40	62.51	63.40
	3-Shot	63.40	61.43	63.40
	4-Shot	63.60	63.17	63.60

**Table A.2:** Unified summary of experimental results for IDEFICS-3-8B in all settings.



## References

- [1] J. Mei, J. Chen, G. Yang, W. Lin, and B. Byrne, “Robust adaptation of large multimodal models for retrieval augmented hateful meme detection,” *arXiv preprint arXiv:2502.13061*, 2025.
- [2] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, “The hateful memes challenge: Detecting hate speech in multimodal memes,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 344–360.
- [3] N. Rizwan, P. Bhaskar, M. Das, S. S. Majhi, P. Saha, and A. Mukherjee, “Zero shot vlms for hate meme detection: Are we there yet?” *arXiv preprint arXiv:2402.12198*, 2024.
- [4] M. Zanella and I. B. Ayed, “Low-rank few-shot adaptation of vision-language models,” in *Proc. IEEE/CVF CVPRW*, 2024, pp. 1593–1603.
- [5] R. Cao, R. K.-W. Lee, and J. Jiang, “Modularized networks for few-shot hateful meme detection,” in *Proc. ACM Web Conference (WWW)*, 2024, pp. 1–9.
- [6] M. S. Hee, A. Kumaresan, and R. K.-W. Lee, “Bridging modalities: Enhancing cross-modality hate speech detection with few-shot in-context learning,” in *Proc. EMNLP*, 2024, pp. 7785–7799.
- [7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [8] J. D. et al., “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [9] S. S. et al., “Multimodal meme dataset (multioff) for identifying offensive content in image and text,” in *Proc. LREC*, 2020, pp. 32–40.
- [10] R. K.-W. L. et al., “Disentangling hate in online memes,” in *Proc. ACM Int. Conf. Multimedia (MM ’21)*, 2021, pp. 5138–5147.
- [11] L. S. et al., “Knowmeme: A knowledge-enriched graph neural network solution to offensive meme detection,” in *2021 IEEE 17th Int. Conf. eScience (eScience)*, 2021, pp. 186–195.
- [12] S. L. et al., “Multi-task learning for multimodal meme classification,” in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, 2021.
- [13] X. Zhou, K. Shen, and J. Ye, “Data augmentation for multimodal hate detection,” in *NeurIPS Workshops*, 2021.

- [14] N. Velioglu and C. Rose, “Detecting offensive content in open-domain multimodal dialogue,” in *Proc. AAAI*, 2020.
- [15] L. Sandulescu, “Multimodal hate speech classification on hateful memes,” *arXiv preprint arXiv:2012.13235*, 2020.
- [16] X. He, Q. Zhang, and Z. Li, “Prompt-enhanced network for hateful meme classification,” in *Proc. IJCAI*, 2024.
- [17] R. Cao, R. K.-W. Lee, and J. Jiang, “Pro-cap: Probing multimodal models via captioning prompts for hate detection,” in *EMNLP Workshops*, 2023, pp. 1234–1245.
- [18] S. Pramanick, P. Bhattacharya, and A. Pal, “Multimodal explainable hate detection,” in *Proc. EMNLP*, 2021, pp. 5678–5689.
- [19] P. W. et al., “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” *arXiv preprint arXiv:2409.12191*, 2024.
- [20] J. B. et al., “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” *arXiv preprint arXiv:2308.12966*, 2023.
- [21] J. Zhou and Y. L. et al., “Qwen2-vl-2b-instruct: Instruction-based vision-language understanding,” in *Proc. ACL*, 2024.
- [22] Q. Team, “Qwen2.5-vl,” Jan. 2025, [Online]. Available: <https://qwenlm.github.io/blog/qwen2.5-vl/>.
- [23] H. Wang and L. Z. et al., “Qwen2.5-vl-3b-instruct: Enhanced multimodal instruction tuning,” in *NeurIPS*, 2024.
- [24] L. B. et al., “Paligemma: A versatile 3b vlm for transfer,” *arXiv preprint arXiv:2407.07726*, 2024.
- [25] Hugging Face, “google/paligemma-3b-pt-224 [model card],” 2024, [Online]. Available: <https://huggingface.co/google/paligemma-3b-pt-224>.
- [26] H. L. et al., “Building and better understanding vision-language models: Insights and future directions,” *arXiv preprint arXiv:2408.12637*, 2024.
- [27] H. L. et al., “Improved baselines with visual instruction tuning,” *arXiv preprint arXiv:2310.03744*, 2023.



**DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

**PLAGIARISM VERIFICATION**

Title of the Thesis Evaluating Open-Source Vision-Language Models for Hateful Meme Detection

Total Pages 34 Name of the Scholar Vhatkar Ganesh Mallinath

Supervisor (s)

(1) Asst.Prof. Gull Kaur

(2) \_\_\_\_\_

(3) \_\_\_\_\_

Department Computer Science

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 9, Total Word Count: 8,599

Date: 31/05/2025

Candidate's Signature

Signature of Supervisor(s)

# ganesh thesis.pdf

 Delhi Technological University

---

## Document Details

### Submission ID

trn:oid:::27535:98512289

### Submission Date

May 30, 2025, 11:45 AM GMT+5:30

### Download Date

May 30, 2025, 11:58 AM GMT+5:30

### File Name

ganesh thesis.pdf

### File Size

806.6 KB

**34 Pages**

**8,559 Words**

**45,421 Characters**





# 9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




## Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

## Match Groups

-  **52 Not Cited or Quoted 9%**  
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**  
Matches that are still very similar to source material
-  **0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 6%  Internet sources
- 4%  Publications
- 8%  Submitted works (Student Papers)

## Integrity Flags

### 0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

## Match Groups

- 52 Not Cited or Quoted 9%**  
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**  
Matches that are still very similar to source material
- 0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 6% Internet sources
- 4% Publications
- 8% Submitted works (Student Papers)

## Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	dspace.dtu.ac.in:8080	1%
2	Submitted works	University of Technology, Sydney on 2025-05-26	1%
3	Submitted works	Delhi Technological University on 2025-05-05	1%
4	Submitted works	University of Wollongong on 2023-12-09	<1%
5	Internet	www.dspace.dtu.ac.in:8080	<1%
6	Internet	arxiv.org	<1%
7	Submitted works	University of Wollongong on 2023-12-08	<1%
8	Internet	aclanthology.org	<1%
9	Publication	Robert B. Musburger, PhD, Michael R Ogden. "Single-Camera Video Production", ...	<1%
10	Submitted works	University of Queensland on 2024-10-12	<1%

11	Submitted works	The University of Manchester on 2024-04-26	<1%
12	Internet	www.mdpi.com	<1%
13	Submitted works	University of Strathclyde on 2013-04-15	<1%
14	Internet	www.diva-portal.org	<1%
15	Submitted works	University College London on 2012-01-09	<1%
16	Submitted works	University of Leeds on 2025-01-15	<1%
17	Internet	dspace.daffodilvarsity.edu.bd:8080	<1%
18	Publication	Biagio Grasso, Valerio La Gatta, Vincenzo Moscato, Giancarlo Sperli. "KERMIT: Kno...	<1%
19	Publication	Bisheng Yang, Zhen Dong, Fuxun Liang, Xiaoxin Mi. "Ubiquitous Point Cloud - The...	<1%
20	Submitted works	Liverpool Hope on 2024-10-18	<1%
21	Publication	Lu, Changsheng. "General Keypoint Detection: Few-Shot and Zero-Shot", The Aust...	<1%
22	Submitted works	National College of Ireland on 2024-12-12	<1%
23	Submitted works	National University of Ireland, Galway on 2023-08-31	<1%
24	Submitted works	University of Warwick on 2024-12-02	<1%

25	Internet	scuttle.klotz.me	<1%
26	Publication	"International Conference on Innovative Computing and Communications", Sprin...	<1%
27	Publication	"Web and Big Data", Springer Science and Business Media LLC, 2024	<1%
28	Publication	Jafar Badour, Joseph Alexander Brown. "Hateful Memes Classification using Mach...	<1%
29	Submitted works	The Hong Kong Polytechnic University on 2025-04-03	<1%
30	Submitted works	Tilburg University on 2025-05-19	<1%
31	Internet	www.research.unipd.it	<1%