RajneetiDrishti: A Two-Stage Vision-Language Ensemble Framework for Political Meme Classification

PoliMemeDecode Datathon – CUET CSE FEST 2025 · Final Round - **NeuronX**

S.M. Shahriar[1]

Md Mobashir Hasan[2]

Sabit Ahamed Preanto[3]

Eiamin Hassan Shanto[4]

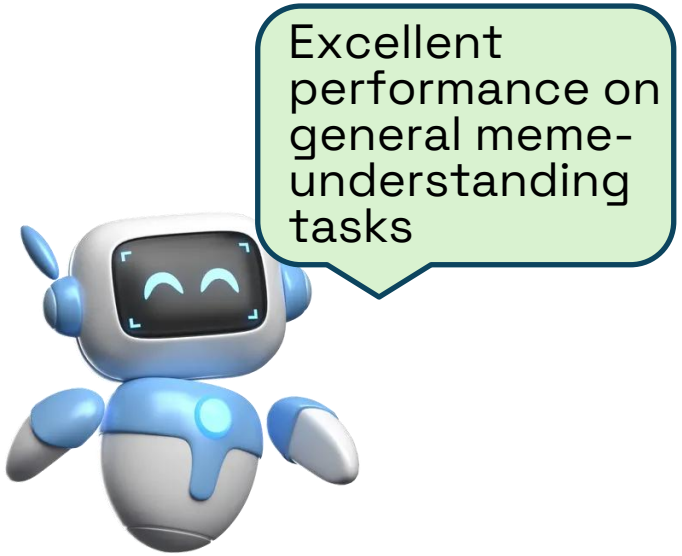
Chittagong University of Engineering and Technology[1]

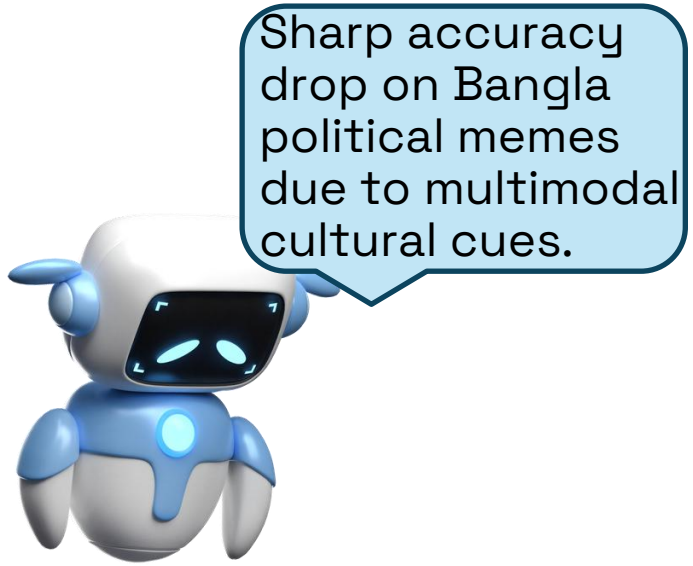
Daffodil International University[2,3,4]

1 | Introduction

#Problem Statement

”Can AI models truly understand political nuance in memes, or are they just matching patterns?”





120M+ Bangla social media users, yet political meme understanding remains challenging.

#Proposed Solution

Our Solution

RajneetiDrishti—the first two-stage vision–language ensemble framework for political meme classification

#Background & Motivation

What Makes RajneetiDrishti Unique?

Designed for Bangladeshi Multimodal Political Contexts

Specifically addresses the hardest elements of Bangladeshi memes—Banglish text, metaphors, humor cues, sarcasm, and cultural references—where generic VLMs fail.

Knowledge-Enhanced Two-Stage Reasoning

Pairs Qwen2.5-VL-7B for rich metadata classification with Phi-3-Vision-128k for political keyword-guided validation, correcting Stage-1 mistakes.

Resource-Efficient Framework

Operates entirely on test data—no training or fine-tuning—yet achieves 93.71% macro F1 using free-tier GPUs.

2 | Data Processing

#Embedded Features

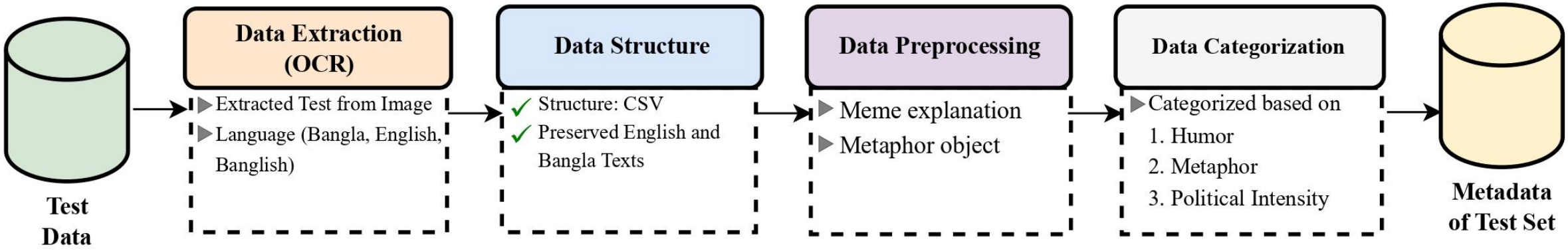


Figure 1: Metadata Creation

- **Data Extraction (OCR):** Extracted text from memes and detected language (Bangla, English, Banglish).
- **Data Structure:** Organized samples into CSV format while preserving all extracted texts. (Total 7 columns)
- **Data Preprocessing:** Added meme explanations and identified metaphor objects.
- **Data Categorization:** Labeled each meme based on Humor, Metaphor, and Political Intensity.

#Dataset Statistics

Dataset Statistics	Value
Training Set	
Total Samples	2,860
Non-Political	2,007 (70.2%)
Political	853 (29.8%)
Test Set	
Total Samples	330
Text Characteristics (OCR)	
Mean Characters	125.30
Max Characters	770
Min Characters	8
Mean Word Count	15.30
Max Word Count	128
Min Word Count	1

Table 1: Dataset overview and OCR text characteristics

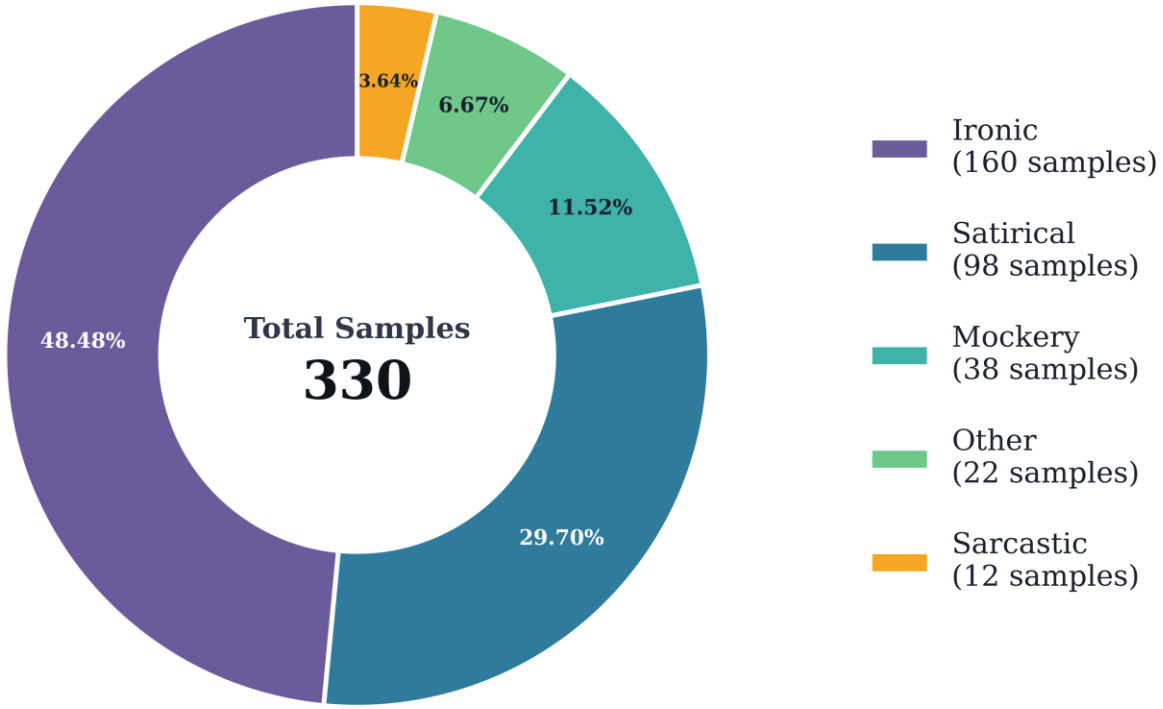


Figure 2: Humor category distribution across 330 test samples

#Dataset Statistics

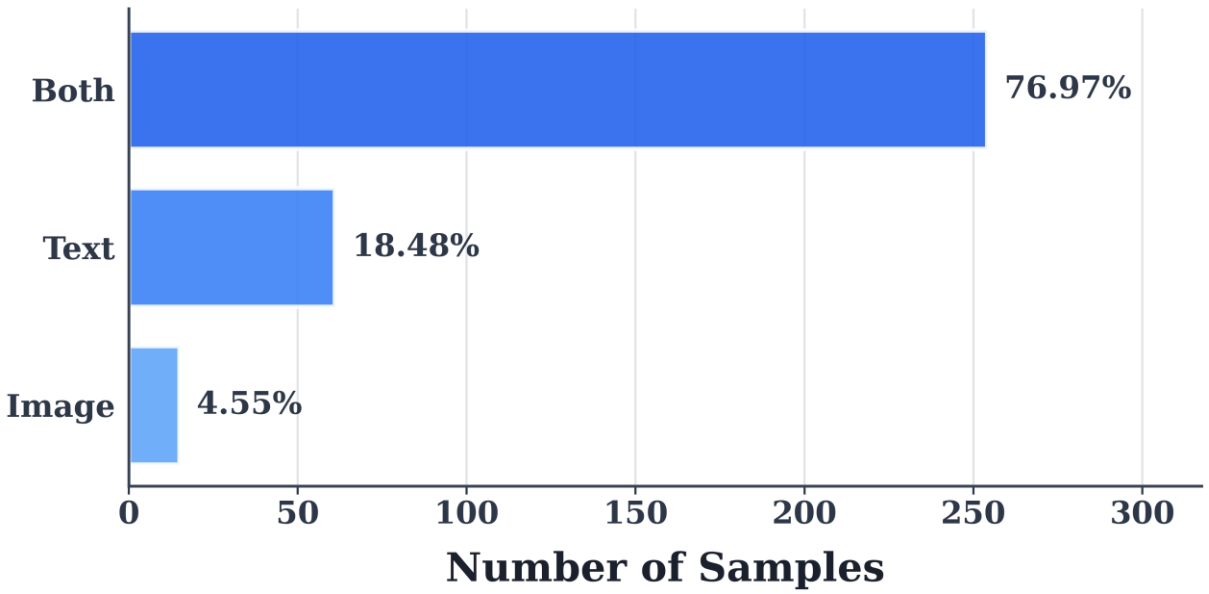


Figure 3: Distribution of image, text, and combined cues

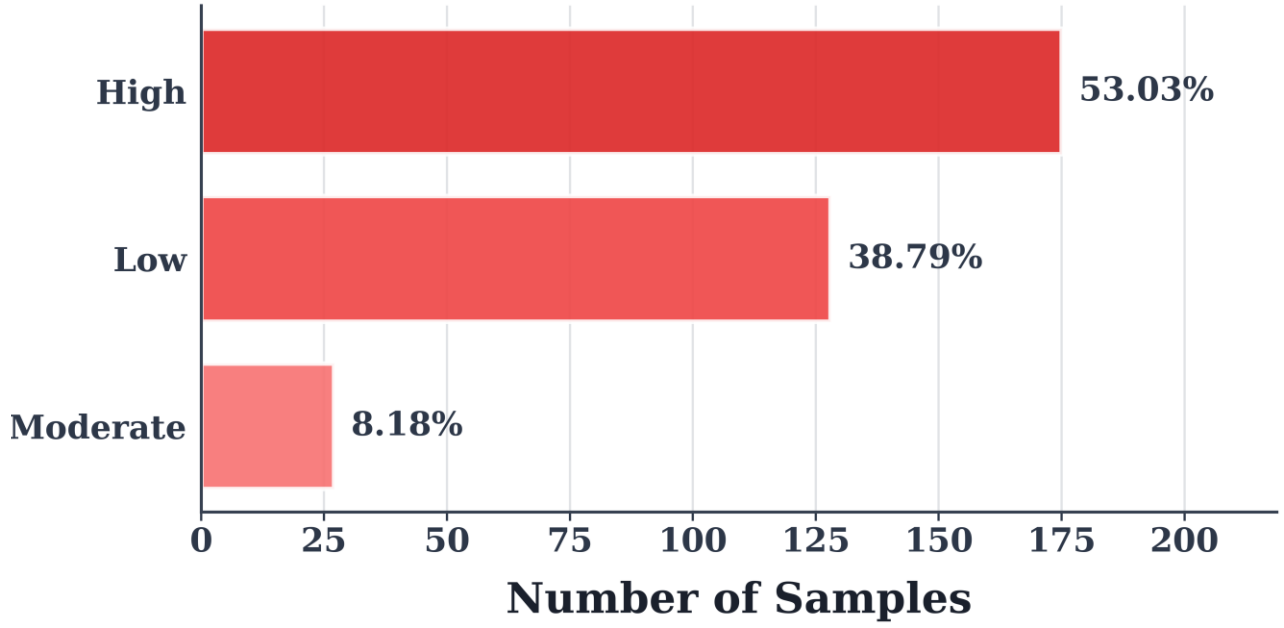


Figure 4: Distribution of political-intensity levels

3 | Methodology

#System Architecture

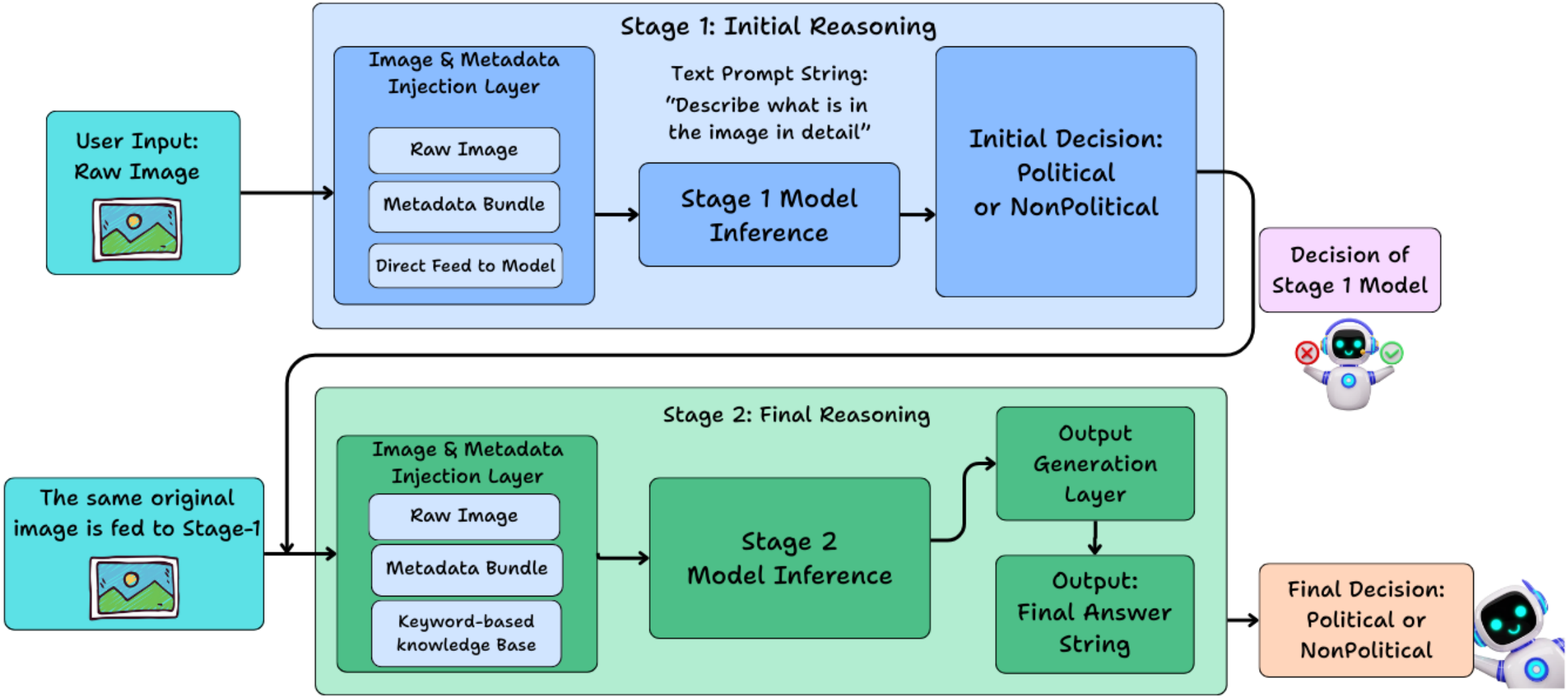
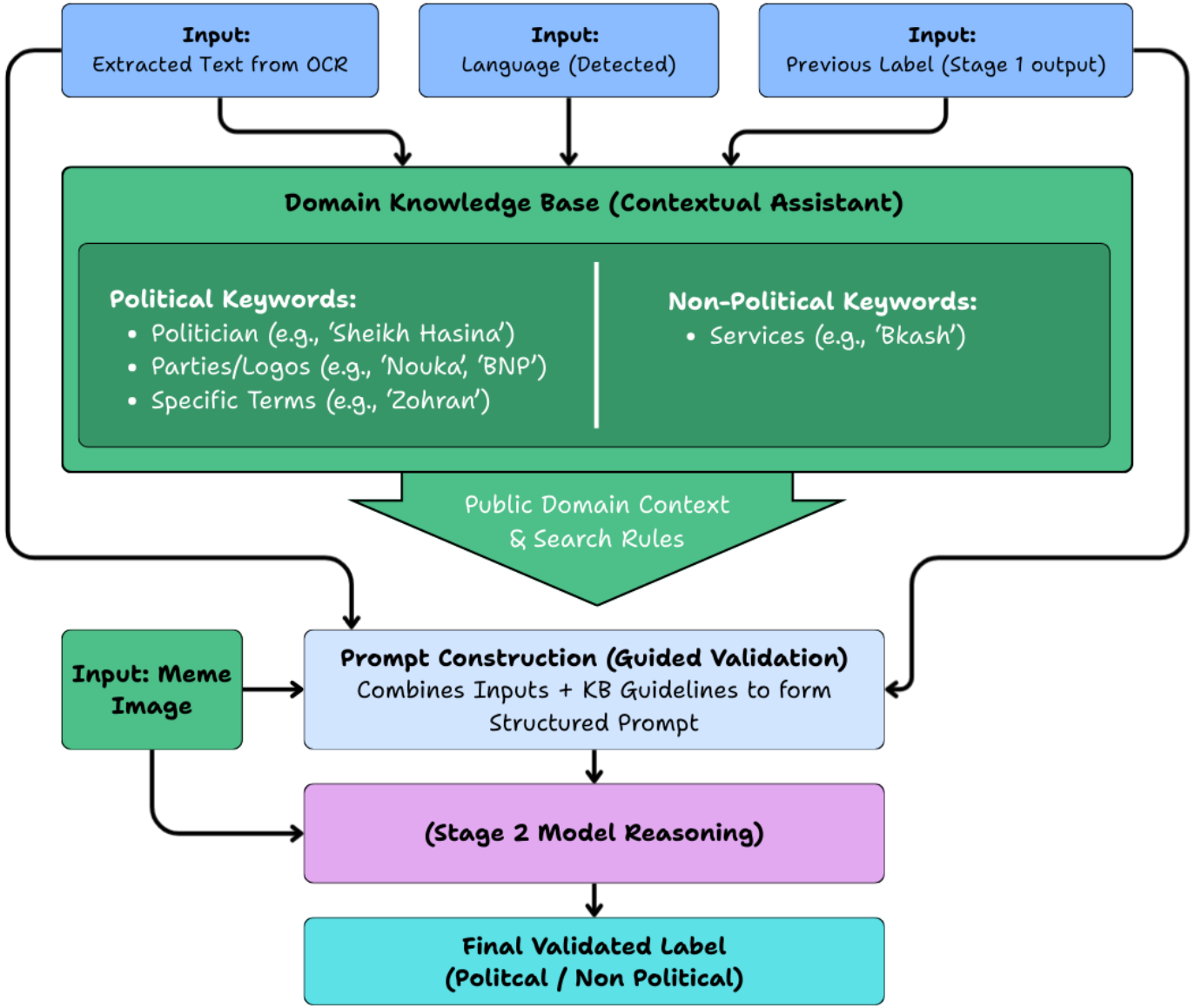


Figure 5: RajneetiDrishti two-stage vision–language ensemble pipeline

#System Architecture

- ✓ Stage-1 architecture performs metadata-aware initial classification.
- ✓ Stage-2 architecture validates or corrects outputs using a political knowledge-base.
- ✓ Pipeline reduces false negatives by combining semantic cues + domain knowledge.

#Knowledge-Base Injection



- ✓ OCR text, language, and Stage-1 output feed into a political/non-political keyword knowledge base.
- ✓ These inputs are merged into a structured prompt for guided validation.
- ✓ Phi-3 Vision uses this enriched prompt to refine or correct the final political classification.

Figure 6: Creation of Knowledge-base from common keywords

#Knowledge-Base Injection Benefits

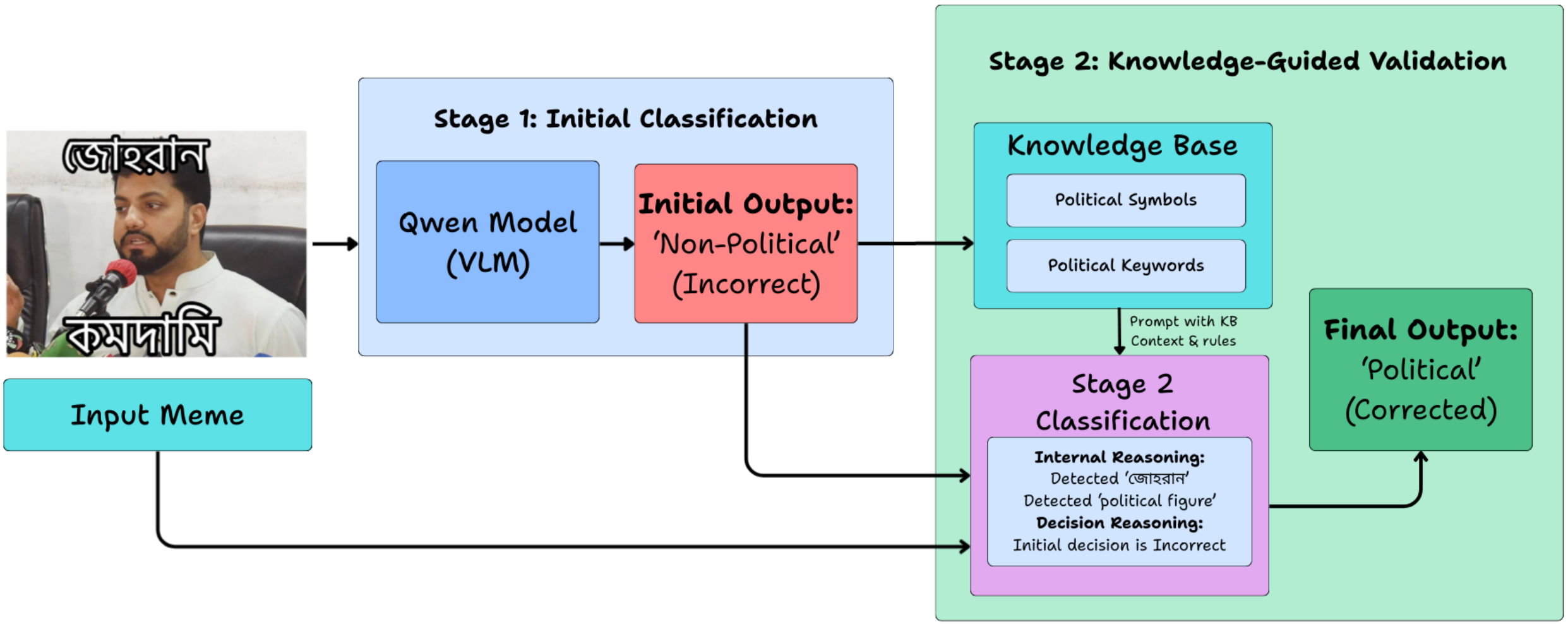


Figure 7: Knowledge-base guided validation

#Reasoning Overview

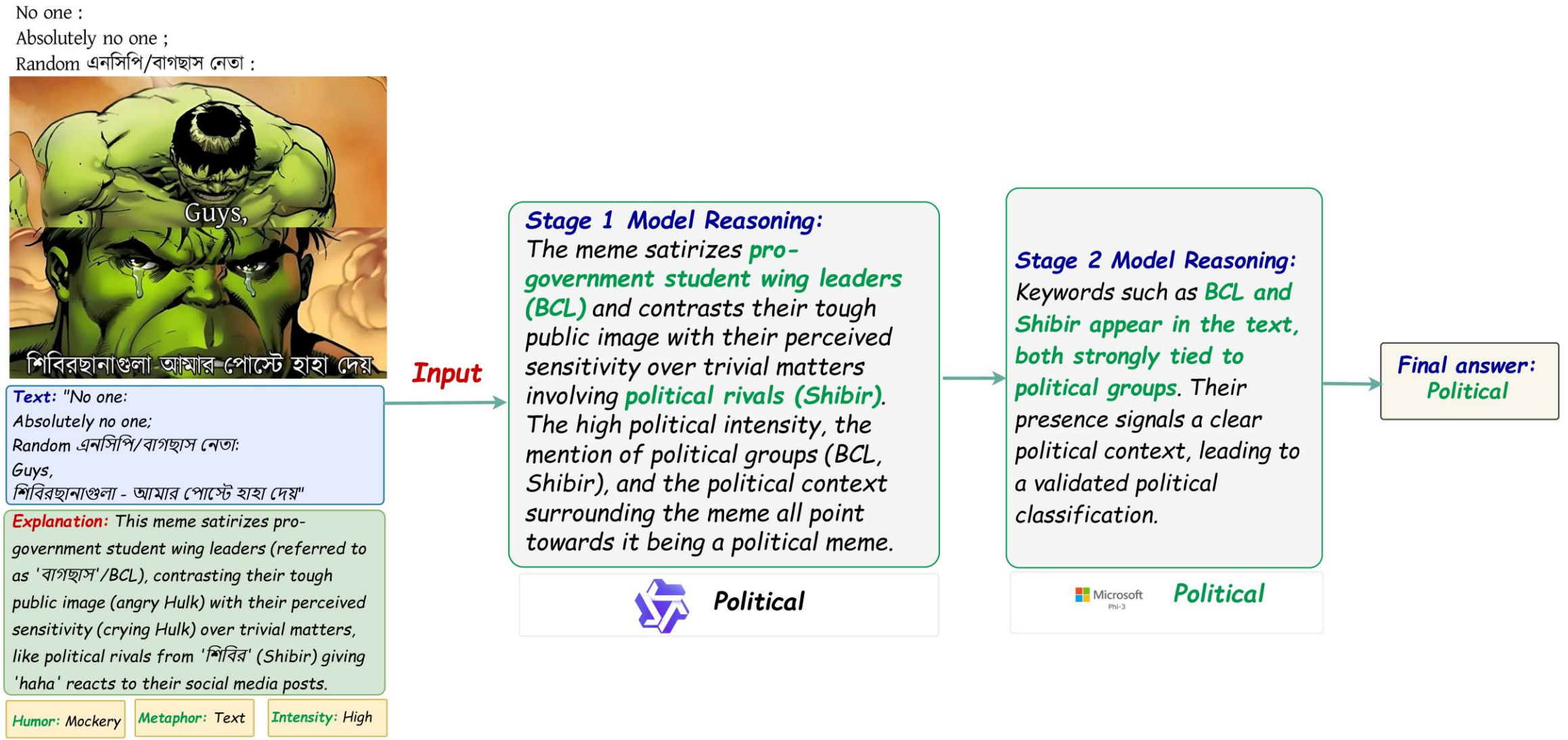


Figure 8: Example of Stage-1 and Stage-2 reasoning on a political meme

#Models Evaluated & LoRA Fine-Tuning

Single models evaluated: LLaVA-v1.5-7B, PaliGemma-3B-mix, Phi-4-multimodal, Qwen3-VL-8B, Phi-3-Vision-128k, Qwen2.5-VL-7B (base + LoRA).

PaliGemma

Phi-4

Qwen3-VL

Phi-3

Table 2: LoRA fine-tuning configuration applied to Qwen2.5-VL-7B

Rank	16
Alpha	32
Dropout	0.05
Learning Rate (LR)	2e-4
Batch Size	8
Epochs	3

4 Performance Evaluation

#Results Comparison

Table 3: Macro F1-Scores on the 330-sample PoliMemeDecode Test Set (public)

Individual Models		Ensemble Configurations (Stage 1: Qwen2.5-VL-7B)	
Qwen2.5-VL-7B	91.61%	+Phi-3-Vision-128k	93.71%
Fine-tuned Qwen2.5-VL-7B	91.20%	+Qwen3-VL-8B	92.63%
Phi-3-Vision-128k-instruct	89.66%	+Phi-4-multimodal	88.87%
Qwen3-VL-8B	87.67%	+PaliGemma-3b-mix	87.88%
Phi-4-multimodal-instruct	85.68%	+LLaVA-v1.5-7b	80.13%
PaliGemma-3b-mix-448	77.69%		
LLaVA-v1.5-7b	52.61%		

#Results Analysis

Overall Performance Results:

- ✓ Qwen2.5-VL-7B emerges as the strongest single model (91.61%).
- ✓ Best ensemble: Qwen2.5-VL-7B → Phi-3 Vision achieves 93.71%.
- ✓ Demonstrates value of sequential reasoning and model complementarity.

#Improvement

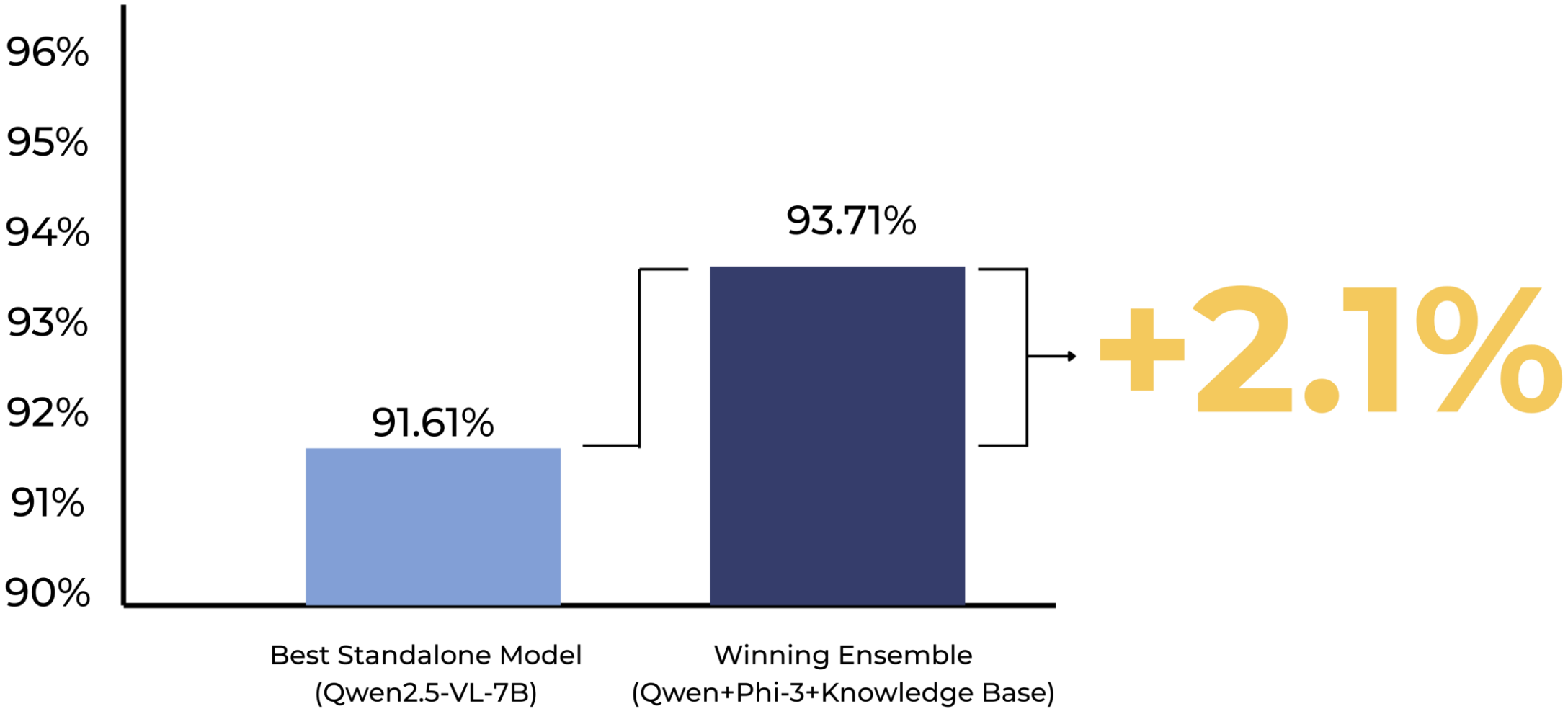


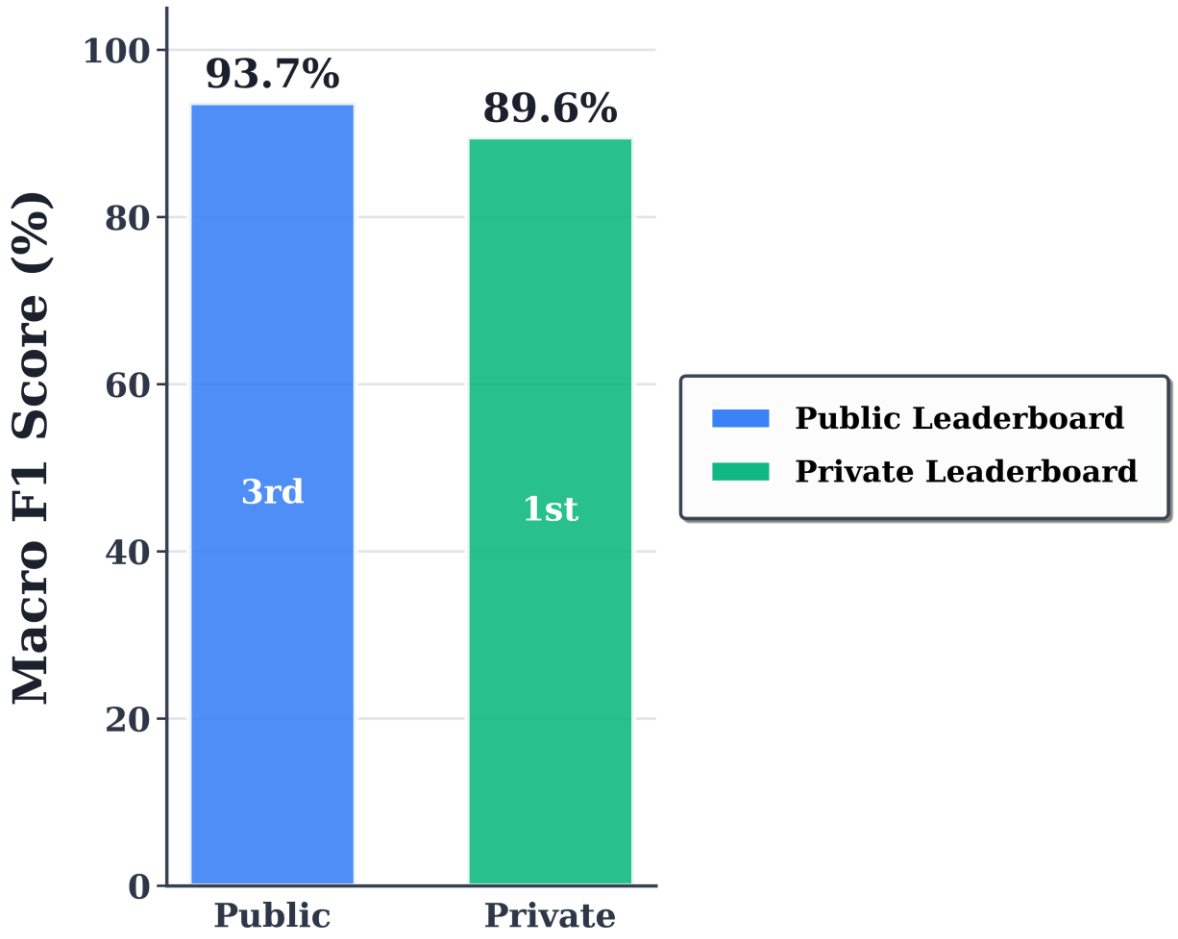
Figure 9: Standalone vs. Ensemble Performance

#Inference Time

Table 4: Computational efficiency shows 4s inference per sample on free-tier GPUs with no training required; the two-stage ensemble adds only 0.5s.

Configuration	Time/Sample	GPU Tier
Qwen2.5-VL-7B	3.5s	Free
Two-Stage Ensemble	4s	Free
Resource Requirements		
Test Samples Used	330	–
Training Required	No	–

#Score & Rank



✓ This illustrates the rank jump and score improvement in private evaluation.

Figure 10: Public vs private leaderboard performance

5 Error Analysis

#Challenging Scenarios & Model Confusion

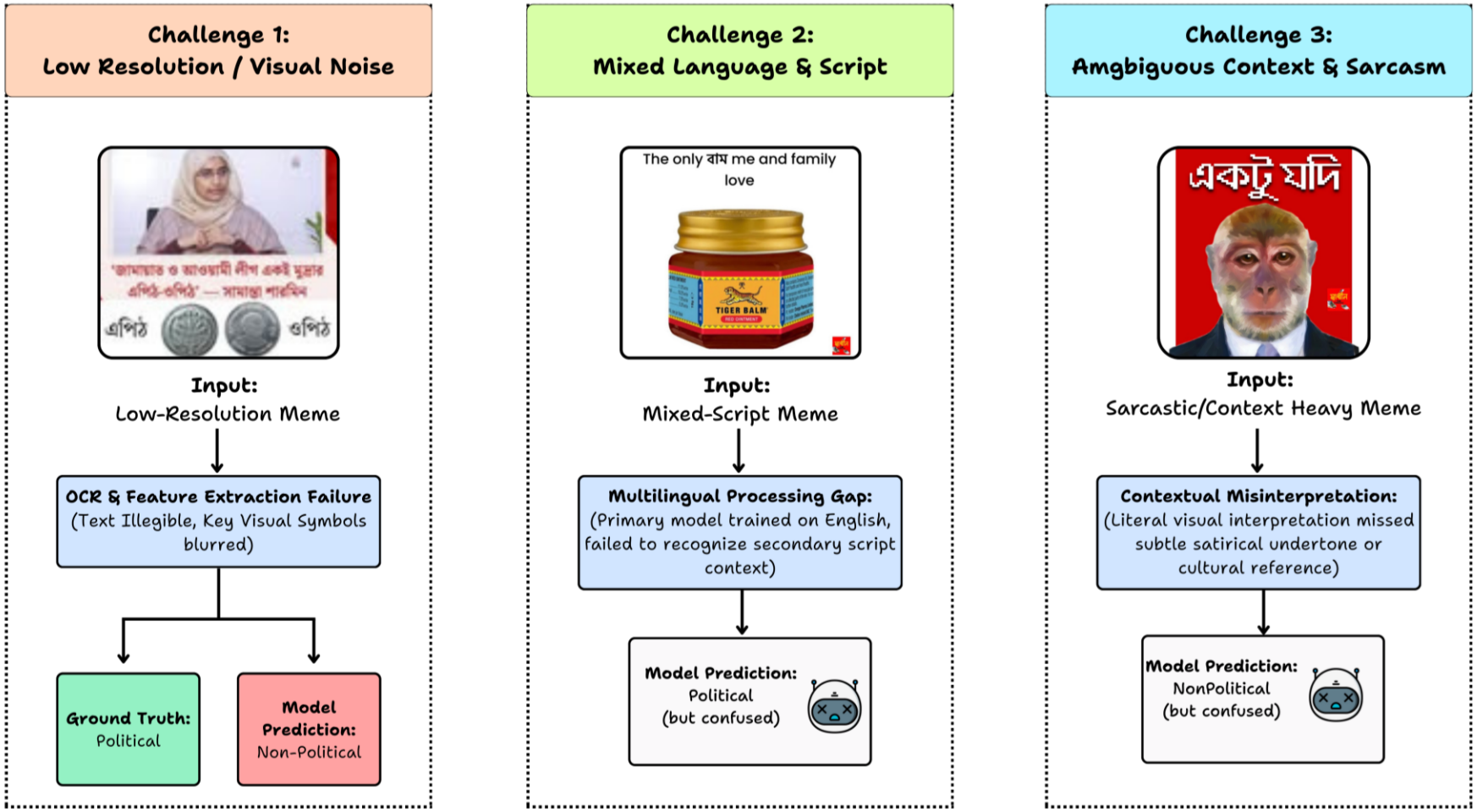


Figure 11: Examples of challenging meme cases

✓ Visualizes low-resolution, mixed-language, and sarcasm-driven meme cause misclassification.

#Challenging Scenarios & Model Confusion

Error Patterns Across Memes:

- Low-resolution images break OCR extraction and object detection.
- Mixed-language memes confuse models lacking Banglish representations.
- Sarcasm- and metaphor-heavy memes require cultural grounding.

#Findings

Ensemble Advantage

Sequential pairing of Qwen and Phi-3 consistently improves accuracy by correcting Stage-1 misclassifications through knowledge-guided reasoning.

CoT/VQA Prompting Advantage

Structured CoT-style prompting and image-aware VQA queries help models extract deeper contextual meaning, improving reasoning on subtle political cues that simple direct prompts fail to capture.

Knowledge Boost

Injecting political keyword context reduces false negatives and improves classification confidence in borderline cases.

Fine-Tuning Not Required

Pretrained VLMs outperform LoRA-adapted variants, reinforcing the efficiency of test-only, zero-training strategies.

#Limitations

Binary Classification Scope

Currently distinguishes only Political vs. Non-Political—no multi-label or party-specific categories.

Sarcasm & Deep Context Challenges

Highly layered satire or culturally implicit humor can still confuse the ensemble.

Sensitive to OCR Quality

Low-resolution or noisy memes still propagate OCR errors into Stage-1 and Stage-2 reasoning.

Static Knowledge Base

Keyword lists must be updated manually; emerging political entities or slang may be missed.

#THANK YOU

#QnA