CS 613: Natural Language Processing
IIT Gandhinagar
Sem-I - 2025-26

ASSIGNMENT 1

Team Name: Machine Minds
Team members: B Keerthan Varma (23110068), K Dinesh Siddhartha (23110168), K Hemanth (23110170), M Lakshmi Manasa (23110193), Praveen Kumar (23110257), R Bhavana (23110274), V Venkat Akhilesh (23110348)

Introduction

Motivation:
Memes have become one of the most influential modes of online communication, combining text and visuals to spread rapidly across social media. While many memes are humorous, a growing portion are harmful, spreading misinformation, reinforcing stereotypes, or inciting hate. Detecting such harmful memes is critical because they can undermine public trust, damage reputations, and amplify social divides. Real world implications include safeguarding public health messaging, reducing online harassment, and ensuring safer digital spaces. Looking ahead, the ability to detect harmful memes zero shot (without labeled data) is essential for future AI moderation systems that must adapt to constantly evolving online content. If this project fails, harmful content will continue spreading unchecked, leaving platforms vulnerable to misinformation, hate speech, and cultural manipulation.

Relation with NLP:
This is fundamentally an NLP problem because memes often embed meaning through text, either in isolation or in combination with visual cues. Understanding harmful intent requires semantic analysis of language, contextual interpretation, and reasoning over multimodal content. The framework in this work leverages large multimodal models (LMMs) to extract, debate, and reason about meme text, showing how natural language processing is crucial for detecting implicit harmfulness. Thus, the project extends NLP beyond plain text, into multimodal reasoning for real world safety applications.

Problem Type:
The problem is best categorized as a classification task within NLP and multimodal AI. Specifically, it involves binary classification determining whether a meme is harmful or harmless. Unlike generative tasks, this focuses on prediction and reasoning, with the twist that it must operate under a zero shot setting, without annotated data. By framing harmful

meme detection as classification supported by insight retrieval and multi-agent debate, the approach positions itself as a robust NLP driven moderation tool.

Related Work:

A) State of the art:

The current best-performing zero-shot systems for harmful and hateful meme detection combine large multimodal models (LMMs) with reasoning prompts or retrieval. In closed-source models, GPT-4o and Gemini-1.5-Flash show the strongest zero-shot performance; among open-source LMMs, LLaVA-1.6-34B is the top baseline. The MIND paper proposes a multi-agent, retrieval, bidirectional insight, debate pipeline that substantially improves zero shot Macro-F1 over these baselines by leveraging similar sample retrieval and multi agent reasoning.

Classic / training-based baselines:
Older and widely cited approaches to harmful meme detection use multimodal two-stream architectures, late fusion ensembles, and task specific fine-tuning (examples: Late Fusion, MOMENTA and other multimodal/hateful-memes approaches). These training-based methods generally perform well when in domain labeled data is available but suffer from out-of-distribution or rapidly evolving memes that motivate zero-shot approaches like MIND.

Multi-agent and retrieval-augmented trends:
MIND builds on two active research trends: (1) retrieval-augmented reasoning (use similar examples to provide context) and (2) LLM / LMM multi-agent frameworks (debate, self-refine, specialized agents). The MIND contribution is to combine multimodal retrieval, bidirectional insight derivation, and an explicit multi-agent debate/judge stage for zero-shot harmful meme detection.

B) Baseline implementation availability:

Open source LMMs / toolkits: Implementations exist for many of the open models used as baselines (LLaVA variants, InstructBLIP, MiniGPT-v2, OpenFlamingo). These are available from their respective repositories and are commonly used as drop-in LMM backbones for research.

Closed source models: GPT-4o and Gemini-1.5-Flash are proprietary (accessible via APIs), not open-sourced.

MIND code: the paper links a GitHub project (https://github.com/destroy-lonely/MIND) in the header; however the paper also

states that due to privacy/ongoing research the code used for experiments was not included with the submission and will be shared upon acceptance. Check the project repo for runnable scripts and any released code.

C) Representative Results:

Dataset	HarM		FHM		MAMI	
Model	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
GPT-4o (Achiam et al., 2023)	67.51	60.29	68.80	68.25	81.00	81.00
Gemini-1.5-Flash (Team et al., 2024)	66.10	64.18	60.20	58.90	76.40	74.29
LLaVA-1.5-7B (Liu et al., 2024)	59.23	49.44	53.80	45.51	52.90	41.53
InstructBLIP-7B (Dai et al., 2023)	51.13	50.99	52.00	48.85	53.10	46.93
MiniGPT-v2-7B (Chen et al., 2023)	60.12	52.39	51.30	47.88	57.40	52.22
OpenFlamingo-9B (Awadalla et al., 2023)	63.42	54.36	50.50	49.52	54.70	49.88
LLaVA-1.5-13B (Liu et al., 2024)	62.28	50.45	55.20	53.01	60.10	55.52
InstructBLIP-13B (Dai et al., 2023)	64.92	49.61	55.40	51.89	60.00	57.97
LLaVA-1.6-34B (Liu et al., 2024)	67.51	61.59	**64.00**	**63.51**	**71.30**	**71.28**
MIND (LLaVA-1.5-13B)	**68.93**	**65.19**	60.80	60.71	68.90	68.84

Table 1: Zero-shot harmful meme detection results on three datasets. The accuracy and macro-averaged F1 scores (%) are reported as the metrics. All baseline models are equipped with Chain-of-Thought prompt. The best and second best results in open-source setting are in **bold** and underlined, respectively.

Model	HarM				FHM				MAMI			
	Accuracy		Macro-F_1		Accuracy		Macro-F_1		Accuracy		Macro-F_1	
	ori.	MIND	ori.	MIND	ori.	MIND	ori.	MIND	ori.	MIND	ori.	MIND
LLaVA-1.5-7B	59.23	62.71 (+3.48)	49.44	57.22 (+7.78)	53.80	54.00 (+0.20)	45.51	48.28 (+2.77)	52.90	53.90 (+1.00)	41.53	45.45 (+3.92)
LLaVA-1.5-13B	62.28	68.93 (+6.65)	50.45	65.19 (+14.74)	55.20	60.80 (+5.60)	53.01	60.71 (+7.70)	60.10	68.90 (+8.80)	55.52	68.84 (+13.32)
LLaVA-1.6-34B	67.51	69.49 (+1.98)	61.59	66.12 (+4.53)	64.00	66.40 (+2.40)	63.51	68.38 (+4.87)	71.30	73.60 (+2.30)	71.28	75.38 (+4.10)
Gemini-1.5-Flash	66.10	68.36 (+2.26)	64.18	66.92 (+2.74)	60.20	63.80 (+3.60)	58.90	62.50 (+3.60)	76.40	78.00 (+1.60)	74.29	77.89 (+3.60)

Table 2: Performance improvements of our proposed framework across different model scales and datasets for zero-shot harmful meme detection. Numbers in green indicate absolute improvements over original models.

Datasets:

Yes, the datasets are publicly available. The MIND paper evaluates on three benchmark datasets:

- HarM (Pramanick et al., 2021a): Memes related to COVID-19, labeled as very harmful, partially harmful, or harmless (merged into harmful vs harmless for binary classification).
- FHM (Kiela et al., 2020): The Facebook Hateful Memes Challenge dataset, released for multimodal hate speech detection.
- MAMI (Fersini et al., 2022): The Multimedia Automatic Misogyny Identification dataset, focusing on memes derogatory towards women.

Size / total number of instances:

From the paper:

Datasets	Test	
	#harmful	#harmless
HarM	124	230
FHM	250	250
MAMI	500	500

Table 4: Statistics of test sets.

Labels:
Yes, these datasets are already labeled (harmful vs. harmless). HarM originally had 3 labels (very harmful, partially harmful, harmless), but the harmful categories are merged for binary classification.

Bias in labels:
Yes, imbalance exists in the labeling across some datasets:

- HarM: Shows a significant imbalance, with only 124 harmful instances versus 230 harmless instances (a ratio of approximately 1:1.85).
- FHM and MAMI: The test sets for these two datasets are balanced (250/250 and 500/500, respectively).

Due to this imbalance in HarM, the paper emphasizes using the Macro-averaged F1 score as the dominant evaluation metric, as it provides a more competitive measure of performance than simple accuracy, which can be skewed by the majority class.

If unlabeled or if expansion needed:
If we were to extend beyond benchmark datasets, new memes could be crawled from Twitter, Reddit, or Facebook using their public APIs or open-source scrapers (subject to ethical approval). For labeling, a combination of crowdsourcing (human annotators) and weak supervision (e.g., prompting large language/multimodal models for label suggestions) can be used.

How many samples are enough?
The existing datasets (≈2,000 - 3,000 labeled memes total across splits) are sufficient for benchmark evaluation. For training robust systems or fine-tuning, typically 10,000 - 20,000 memes would be more reliable, but zero-shot setups like MIND specifically aim to operate without additional labels.

Time for data curation:
Using existing datasets: 2- 4 days (some datasets are not publicly available, so this will require emailing the paper authors).
Crawling & labeling new memes: 3 - 4 weeks depending on scope. Crawling via APIs/scripts may take ~1 week, and annotation (if using human workers) could take another 2 - 3 weeks depending on the labeling workforce and quality checks.

Experimental setting:

Goal: Compare the MIND multi-agent retrieval with insight and debate pipeline (zero-shot, no training) against standard zero-shot LMM prompting and other existing training-based baselines. Also run ablations (SSR off, RID off, unidirectional RID, no debate) to quantify each component's contribution.

Models / baselines:

- Open-source LMMs: LLaVA-1.5-13B and LLaVA-1.6-34B (backbones used in paper).
- Other open models available: InstructBLIP, MiniGPT-v2, OpenFlamingo (optional).
- Closed-source baselines (API): GPT-4o and Gemini-1.5-Flash (via API) if budget allows (paper includes them as comparisons).
- Our system (MIND): Implement SSR (CLIP ViT-L/14@336p embeddings with $\lambda v/\lambda t$ fusion), Forward + Backward RID, two debaters + judge arbitration. Keep K=3 as default (paper found K=3 optimal).

Decisions to test:

- Vary K (1 - 7) for retrieval.
- Compare embedding fusion weights (λv, λt) search.
- Compare prompt variants (CoT vs direct label).
- Ablation variants (w/o SSR, w/o RID, w/o IAI, forward-only, backward-only).

Train / Dev / Test split policy:

- Using the published splits wherever available (paper uses HarM / FHM / MAMI test sets; see Table 4 for test counts). For fair replication, adopting the same test splits.
- Reference / Unlabeled set (Sref): For zero-shot MIND, treating the original training data (images and text without labels) as the reference corpus Sref used for retrieval.
- Dev set:
    1. If the dataset includes an official validation split, use it for tuning hyperparameters (λv, λt, K, prompts).
    2. If not, create a dev set by randomly holding out 10 - 15% of the train/reference split (stratified by whatever label information you have, if any), or by using a small labeled subset (e.g., 100 - 300 examples) to tune retrieval and prompt hyperparameters.
- Test set: Use the official test set only for final evaluation (no peeking). Report metrics averaged across runs where randomness exists (e.g., API stochasticity set temperature=0 for determinism, as in the paper).
- For the Dataset created from web crawling, we will firstly divide it into categories and then make the train, test split in each category to maintain balance in the training and testing to eliminate unwanted biases.

Hyperparameter search (how to find best params):

- Small, discrete/hypers we will grid-search (cheap):
    1. $K \in \{1, 2, 3, 4, 5\}$ (paper found 3 is optimal).
    2. $\lambda v \in \{0.6, 0.7, 0.8, 0.9\}$, $\lambda t = 1 - \lambda v$ (paper used 0.8/0.2 as result of grid search).
    3. Prompt variants: CoT vs short vs note-informed debater/judge prompts.

Expensive / continuous search (better than brute force):

- Use Bayesian Optimization to tune continuous parameters like λv/λt and weighting schemes for retrieved examples. This reduces evaluations compared to grid search.
- For prompt templates, use a small combinatorial grid of template variants and then fine-tune top candidates with a few validation examples.
- When computing costs are high: apply multi fidelity search (Hyperband) quickly discard poor configs using a small dev subset, then expand promising configs to the full dev set.
- Evaluation during search: use dev macro-F1 as the objective.

Metrics:

- **Primary metric: Macro-averaged F1 score (Macro-F1).**
- **Why:** datasets are class-imbalanced (e.g., HarM test counts show imbalance). Macro-F1 treats each class equally and prevents inflated scores from majority-class bias. The MIND paper uses Macro-F1 as the dominant metric.
- Secondary metrics: Accuracy (for comparability), per-class precision/recall, confusion matrix (to understand types of mistakes), and calibration/error analysis (how many harmful memes were missed).
- Statistical significance: report 95% confidence intervals (bootstrap) across random seeds / prompt variants when relevant.
- Human evaluation (optional but valuable): sample cases where model asserts harmfulness and have human annotators verify; collect qualitative judgments from human moderators for interpretability of the derived “Thought” reasoning chains.

Experimental protocol & runs:

- Deterministic settings: set LMM temperature = 0 for deterministic outputs (paper uses temperature 0).
- Number of independent runs: for stochastic baselines (if any), run 3 seeds; for zero-shot deterministic systems, single run suffices with fixed temperature.
- Ablations: run the full model and each ablation on all datasets to produce the ablation table (like Table 3 in paper).

Resources & system effort (compute, time, people)

- Compute (recommended):
    1. For reproducibility with LLaVA-1.5-13B: at least 4 × 48GB GPUs (paper used four NVIDIA A40 48GiB GPUs).

2. For LLaVA-1.6-34B or closed-source APIs: use API access for GPT-4o/Gemini to avoid hosting costs.

- **Estimated processing time (empirical from paper):** HarM / FHM / MAMI processing times reported: using Gemini-1.5-Flash ~1.5h / 3h / 5h; using LLaVA-1.5-13B ~3h / 4.5h / 9h respectively. These are complete-run times (inference over dataset) reported in the paper which can be used to plan runtime budgets.
- **Inference cost overhead:** MIND requires multiple LMM calls per example (forward, backward, debaters, judge), leading to approximately ~8× inference calls vs simple zero-shot prompting (paper's estimate). This affects hourly GPU/API costs.

Data indexing & retrieval implementation:

- Embeddings: use CLIP ViT-L/14@336p (frozen) for Venc/Tenc and multimodal fusion $E = \lambda v \cdot V + \lambda t \cdot T$ (paper used $\lambda v=0.8$, $\lambda t=0.2$ after grid search).
- Index: FAISS or Milvus for fast nearest-neighbor retrieval of Sref embeddings.
- Storage: Document DB (SQLite / PostgreSQL) storing image paths, raw text, CLIP embeddings ($\mathbb{R}^d$).
- Retrieve K nearest per example (K tuned, default 3).

Final demo: scenario & system effort required:

Goal: interactive web demo where one can upload any meme image and see Model's final harmful/harmless judgment with reasons for the judgement.
Tech stack & deployment:

- Backend: FastAPI (Python), LMM calls, and reasoning. Containerize with Docker.
- LMM execution (for this assignment):
    1. Option A (research): host LLaVA models on GPU instances and call locally.
    2. Option B (lean/demo): use GPT-4o / Gemini APIs for LMM steps (cheaper dev effort, pay-as-you-go).
- Frontend: React (single-page) showing upload and results.

Alternative: lightweight command-line or Jupyter notebook with an interactive cell that shows the entire chain.

Effort Required:
Dataset Curation: 3-4 people work on Hindi and Telugu dataset curation for creating a dataset of around 8k instances for each language. 3-4 people work on Tamil and Kannada dataset curation for creating a dataset of around 8k instances for each language.

Baseline Testing and Finetuning: Zero shot testing of 10-15 existing baseline models with understanding and implementing open/closed source codes and further fine tuning the models. This will need each one in the group implementing 2-3 models.

Deliverables & success criteria:

Deliverables:

- Reproduction script, configs, and evaluation notebooks.
- Tables reproducing paper results + our ablations and tuned variants.
- Demo (web app or notebook) with sample queries and explanation traces.
- Short report summarizing findings.

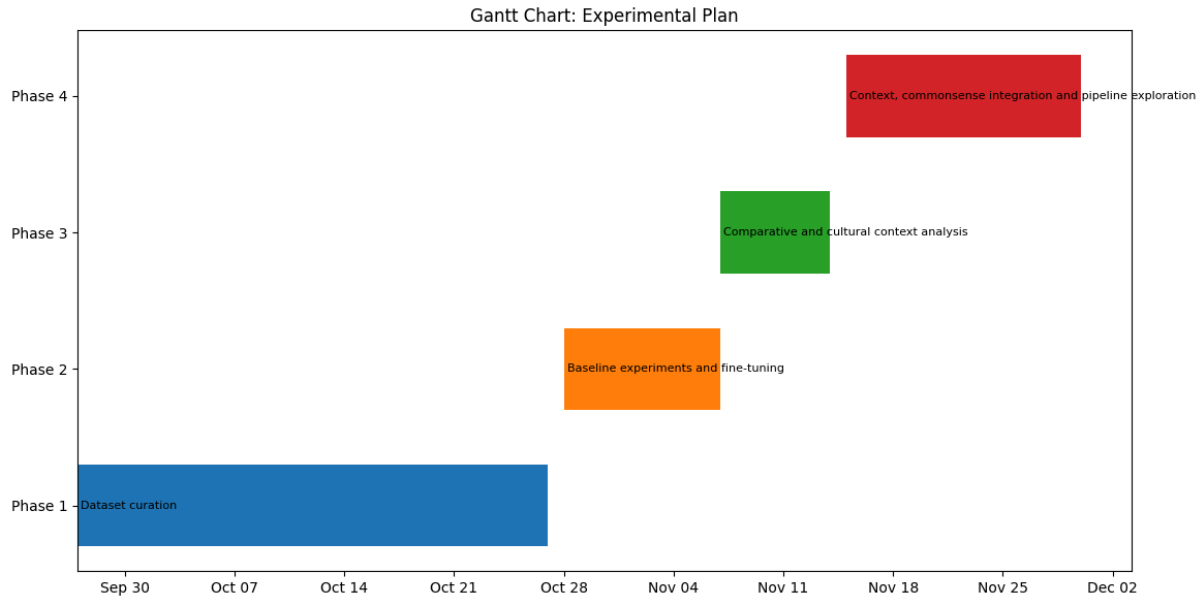
- Curated Dataset and Preprocessing Code
- Baseline Performance report and Model Checkpoints
- Interpretability and Bias Analysis
- Model Pipelines, Ablation Study and Performance results

Success criteria:

- Achieve Macro-F1 within ±3 - 5% of reported MIND results for at least one backbone (LLaVA-1.5-13B).
- Ablation table that demonstrates contributions of SSR, RID, and IAI.
- Working interactive demo showing retrieval, agent reasoning, and final decision.

Project Management:

Gantt Chart:



Computation Resources:

- GPUs: 4 × NVIDIA A40 48GB (or equivalent A100/H100) for hosting open LMMs like LLaVA.
- CPU Cores: 16 - 32 cores for preprocessing, FAISS indexing, and pipeline orchestration.
- RAM: 128 - 256 GB (for embeddings + dataset loading).
- Disk: ≥2 TB SSD (for model checkpoints, embeddings, and logs).

- Alternative (budget-friendly): If using GPT-4o/Gemini APIs instead of hosting models locally, only 1 CPU VM with 64 GB RAM is needed.

Success Criteria:

- **Quantitative:** Achieve Macro-F1 within ±3–5% of reported MIND results.
- **Notebook:** Deliver a working interactive demo (web app or notebook) that accepts memes, retrieves references, shows reasoning, and outputs final harmful/harmless predictions. (For the purpose of this assignment)

For Success of the Project:

- Curate a balanced dataset of at least 8000 memes in each of Kannada, Hindi, Telugu, and Tamil.
- Train baselines : IndicBERT + ResNet,CLIP, VisualBERT, ViLT, MOMENTA, KnowMeme, MIND and Zero shot evaluation. Record metrics (accuracy, precision, recall, F1 score and MMAE) before and after fine-tuning.
- Apply attention maps and LIME for interpretability. Analyze outputs for cultural appropriateness and biases with human evaluation.
- Integrate commonsense knowledge bases like ConceptNet. Evaluate improvements on the above mentioned metrics, Ablation study and perform interpretability analysis.

Biggest Risk & Mitigation:

- Risk: Retrieval quality or LMM limitations may cause poor judgments, leading to low Macro-F1.
- Mitigation: Early evaluation of retrieval quality, fallback strategies, and caching to reduce inference cost.
- The size of the Dataset and overfitting.

Experimentation:

Github link- https://github.com/DineshSiddhartha/MIND-code-base-

a) **Reproducing the Results in the Paper:**
   Results achieved in the Paper:

Dataset	HarM		FHM		MAMI	
Model	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
GPT-4o (Achiam et al., 2023)	67.51	60.29	68.80	68.25	81.00	81.00
Gemini-1.5-Flash (Team et al., 2024)	66.10	64.18	60.20	58.90	76.40	74.29
LLaVA-1.5-7B (Liu et al., 2024)	59.23	49.44	53.80	45.51	52.90	41.53
InstructBLIP-7B (Dai et al., 2023)	51.13	50.99	52.00	48.85	53.10	46.93
MiniGPT-v2-7B (Chen et al., 2023)	60.12	52.39	51.30	47.88	57.40	52.22
OpenFlamingo-9B (Awadalla et al., 2023)	63.42	54.36	50.50	49.52	54.70	49.88
LLaVA-1.5-13B (Liu et al., 2024)	62.28	50.45	55.20	53.01	60.10	55.52
InstructBLIP-13B (Dai et al., 2023)	64.92	49.61	55.40	51.89	60.00	57.97
LLaVA-1.6-34B (Liu et al., 2024)	67.51	61.59	**64.00**	**63.51**	**71.30**	**71.28**
MInD (LLaVA-1.5-13B)	**68.93**	**65.19**	60.80	60.71	68.90	68.84

Table 1: Zero-shot harmful meme detection results on three datasets. The accuracy and macro-averaged F1 scores (%) are reported as the metrics. All baseline models are equipped with Chain-of-Thought prompt. The best and second best results in open-source setting are in **bold** and underlined, respectively.

Model	HarM				FHM				MAMI			
	Accuracy		Macro-F_1		Accuracy		Macro-F_1		Accuracy		Macro-F_1	
	ori.	MInD	ori.	MInD	ori.	MInD	ori.	MInD	ori.	MInD	ori.	MInD
LLaVA-1.5-7B	59.23	62.71 (+3.48)	49.44	57.22 (+7.78)	53.80	54.00 (+0.20)	45.51	48.28 (+2.77)	52.90	53.90 (+1.00)	41.53	45.45 (+3.92)
LLaVA-1.5-13B	62.28	68.93 (+6.65)	50.45	65.19 (+14.74)	55.20	60.80 (+5.60)	53.01	60.71 (+7.70)	60.10	68.90 (+8.80)	55.52	68.84 (+13.32)
LLaVA-1.6-34B	67.51	69.49 (+1.98)	61.59	66.12 (+4.53)	64.00	66.40 (+2.40)	63.51	68.38 (+4.87)	71.30	73.60 (+2.30)	71.28	75.38 (+4.10)
Gemini-1.5-Flash	66.10	68.36 (+2.26)	64.18	66.92 (+2.74)	60.20	63.80 (+3.60)	58.90	62.50 (+3.60)	76.40	78.00 (+1.60)	74.29	77.89 (+3.60)

Table 2: Performance improvements of our proposed framework across different model scales and datasets for zero-shot harmful meme detection. Numbers in green indicate absolute improvements over original models.

Results Reproduced by us:

Model	HarM Dataset (entire test set)	FHM Dataset (20 samples of test set)	MAMI Dataset (30 samples of test set)
MInD (LLaVA-1.5-13B)	Accuracy- 68.64 F1 Score- 64.94	Accuracy- 50.00 F1 Score- 33.33	Accuracy- 73.33 F1 Score- 72.85
MInD (LLaVA-1.5-7B)	MLE	Accuracy- 50.00 F1 Score- 33.33	Accuracy- 73.33 F1 Score- 72.85

For the **FHM** and **MAMI** datasets, due to **resource limitations and the extensive inference time** required for each sample, evaluation was conducted on a **subset of 50 test instances** rather than the full set of 1000 samples. While this reduced the evaluation scope, the selected subset was chosen to maintain a representative distribution of labels and content diversity, ensuring that the observed performance trends remain indicative of overall model behavior.

b) Patterns in results and findings:

Across all three datasets HarM, FHM, and MAMI, a consistent performance hierarchy is observed among the models. The smaller LLMs, such as LLaVA-1.5-7B, InstructBLIP-7B, and MiniGPT-v2-7B, generally perform worse than their larger counterparts and proprietary models like GPT-4o and Gemini-1.5-Flash. This indicates a strong scaling trend, where increasing model size correlates with better zero-shot multimodal understanding and reasoning capabilities. Larger models benefit from richer visual-textual embeddings and more extensive instruction-tuning data, which allow them to generalize more effectively to complex meme content and nuanced forms of harm.

Despite the overall gap between smaller and larger models, there are noticeable differences even within the smaller LLM group. Models designed with stronger visual-language grounding such as LLaVA and InstructBLIP outperform lighter baselines like MiniGPT-v2. This shows that architectural choices and alignment strategies play a significant role in performance, sometimes more so than sheer parameter count. For example, InstructBLIP-7B consistently outperforms MiniGPT-v2-7B, highlighting that effective cross-modal alignment can enhance harmful content understanding, even for smaller models.

When these smaller models are integrated into the MIND framework, their performance improves substantially across all datasets. The observed gains, typically ranging from +3% to +6% in both accuracy and macro-F1, demonstrate that MIND enhances multimodal reasoning through improved alignment and inference mechanisms. This result suggests that model architecture and training strategy can partly offset limitations imposed by smaller parameter sizes. In essence, MIND helps smaller models better capture fine-grained relationships between textual and visual cues, which are essential for detecting subtle harmful content in memes.

The consistent improvements across HarM, FHM, and MAMI indicate that this enhancement is not dataset-specific but reflects a broader robustness. Thus, the findings emphasize a key insight: architectural intelligence and multimodal reasoning augmentation can complement model scaling, enabling smaller, more efficient LLMs to achieve competitive performance in complex multimodal understanding tasks.