# Detecting and Reasoning about Bias in Multimodal Content

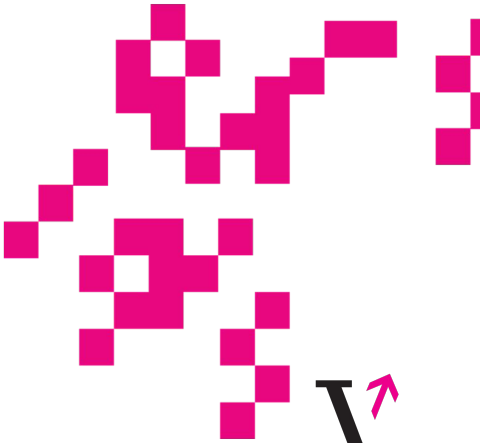
VECTOR INSTITUTE | INSTITUT VECTEUR

**Authors**: Shaina Raza, et al.

**Presenter**: Carolyn Chong

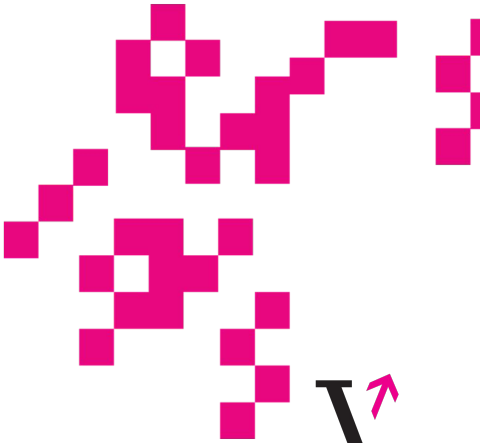
# The Problem: Framing “Bias”

- **Current Gap:** Text-only models miss the “Visual Framing” provided by images.
- **Definition:** Framing bias is the directional use of text, image, or their interaction to skew interpretation.
- **The Challenge:** We need models that don't just “label” bias but can reason through it like a human critic.



# The ViLBias Dataset

- **Scale:** 40,945 high-quality text-image pairs.
- **Premium Sources:** Driven by major global outlets including the Financial Times (4,895 pairs), USA TODAY (3,320), and CNN (3,067).
- **VQA Design:** Designed for both Classification (Yes/No) and Reasoning (Why?).



# Methodology: Human-in-the-Loop

### The Pipeline

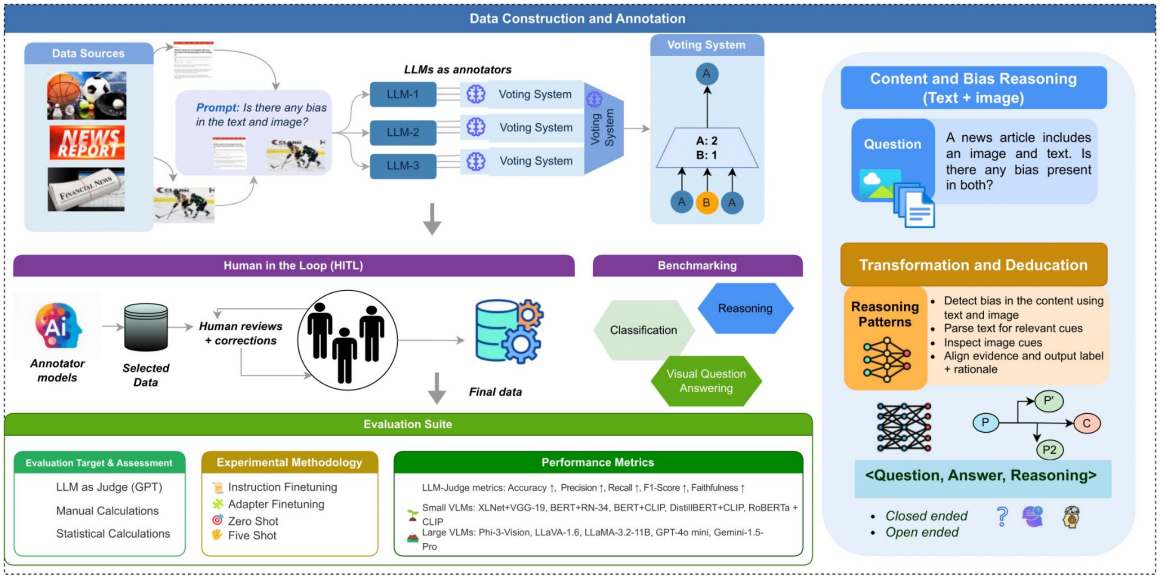
Multi-LLM voting ensures initial labeling consistency.

### Expert Oversight

Human reviewers correct AI “hallucinations” and refine reasoning rationales.

### Efficiency First

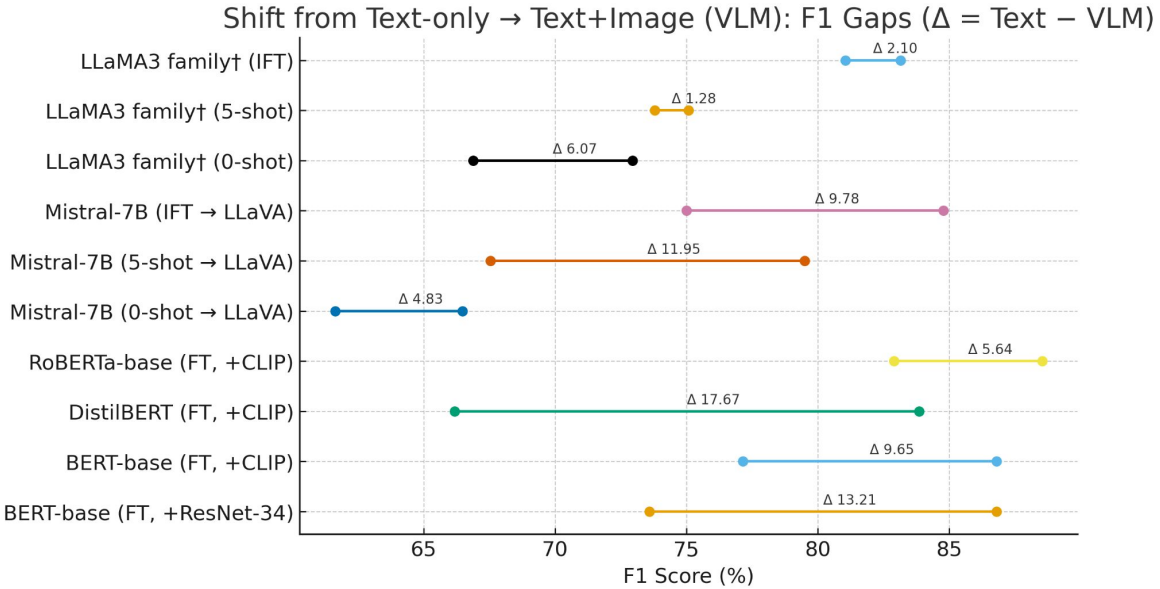
Evaluated using LoRA and QLoRA adapters, achieving state-of-the-art results with <5% trainable parameters.



# Results

## *Key Finding 1: Multimodal Gains*

**Adding images improves F1 by 3–5% over text-only; up to +17% for smaller models**




VECTOR INSTITUTE | INSTITUT VECTEUR

# Key Finding 2.1: Efficiency

**Parameter-efficient methods (LoRA/QLoRA/Adapters) recover 97–99% of full fine-tuning with <5% trainable parameters**

| Method | Train.% | GPU Mem. | Phi-3 | | LLaVA-1.6 | | LLaMA-3.2-11B | |
|---|---|---|---|---|---|---|---|---|
| | | | F1 | R. Acc. | F1 | R. Acc. | F1 | R. Acc. |
| Full FT | 100 | 24.0 | **82.0** | **67.9** | **78.8** | **65.5** | **80.4** | **66.1** |
| LoRA | 1.2 | 10.1 | 81.2 | 66.8 | 77.9 | 64.7 | 79.7 | 65.3 |
| QLoRA | 1.2 | 7.4 | 80.9 | 66.5 | 77.6 | 64.2 | 79.4 | 65.1 |
| Adapters | 5.0 | 12.8 | 81.4 | 67.1 | 78.1 | 65.0 | 79.9 | 65.7 |
| Prompt-Tuning | 0.1 | 6.9 | 79.8 | 64.9 | 76.2 | 62.8 | 77.8 | 63.5 |

# Key Finding 2.2: Efficiency

**RoBERTa + CLIP** show best bias detection performance after finetuning on vilbias benchmark.

Consistent performance improvement on both open and close source VLMs as we move from **0-shot → 5-shot → Instruction Finetuning**.

| Model | Config | Prec. (%) | Recall (%) | F1 Score (%) | Acc. (%) |
|---|---|---|---|---|---|
| *Small VLMs* | | | | | |
| XLNet + VGG-19 | FT | 72.0 | 68.2 | 70.1 | 77.1 |
| BERT + RN-34 | FT | 75.8 | 71.5 | 73.6 | 79.4 |
| BERT + CLIP | FT | 81.3 | 73.4 | 77.2 | 84.2 |
| DistilBERT + CLIP | FT | 68.5 | 64.0 | 66.2 | 74.9 |
| RoBERTa + CLIP | FT | **84.5** | **81.4** | **82.9** | **83.6** |
| *Large VLMs* | | | | | |
| Phi-3-Vision | 0-shot | 70.4 | 66.0 | 68.1 | 69.8 |
| | 5-shot | 73.2 | 71.0 | 72.1 | 70.5 |
| | IFT | **76.8** | **78.1** | **77.4** | **74.0** |
| LLaVA-1.6 | 0-shot | 62.5 | 60.8 | 61.6 | 62.7 |
| | 5-shot | 68.1 | 67.0 | 67.5 | 65.2 |
| | IFT | 75.4 | 74.6 | 75.0 | 76.1 |
| LLaMA-3.2-11B | 0-shot | 65.0 | 68.9 | 66.9 | 68.4 |
| | 5-shot | 73.4 | 74.2 | 73.8 | 72.1 |
| GPT-4o mini† | 0-shot | 71.8 | 74.6 | 73.2 | 72.9 |
| | 5-shot | **77.9** | **79.8** | **78.8** | **77.2** |
| Gemini-1.5 Pro† | 0-shot | 70.9 | 73.1 | 72.0 | 71.5 |
| | 5-shot | 76.8 | 78.5 | 77.6 | 76.9 |

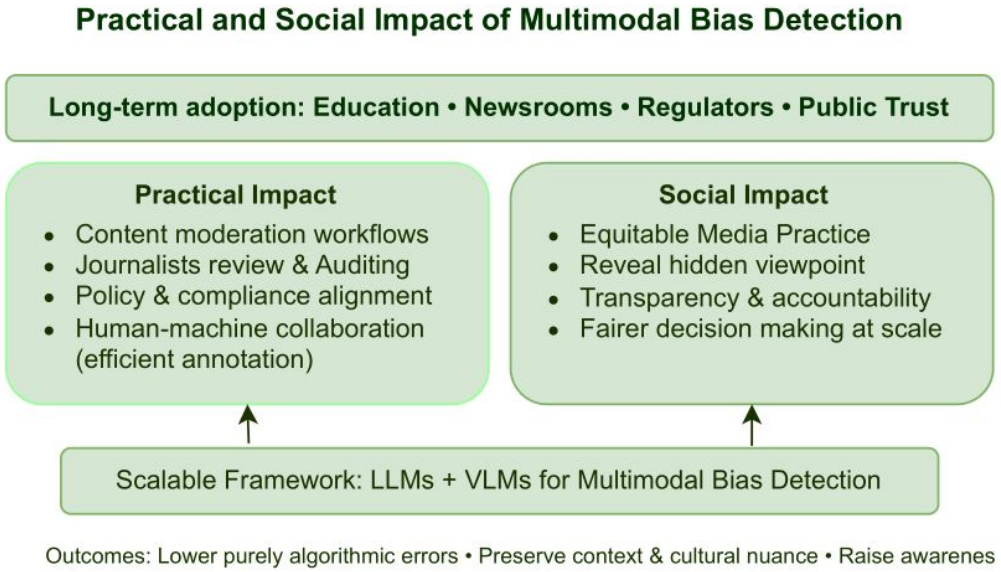
# Limitations

**Scope:** English/Western-centric (FT, CNN, BBC); limited cultural diversity

**Labels:** Binary classification simplifies multidimensional bias (partisan, sensationalist, representational)

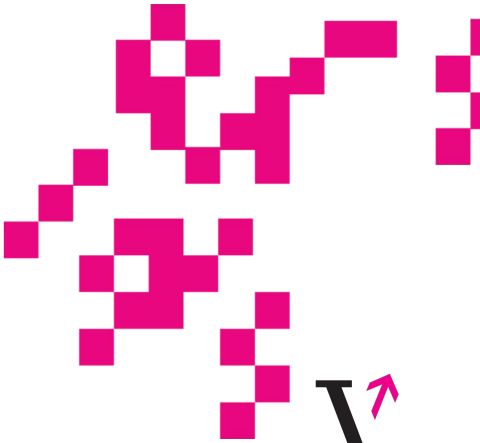
**Risks:** LLM-as-judge circularity, computational barriers for small labs, potential misuse by authoritarian regimes

# Impact



# Key Findings: The Multimodal Advantage

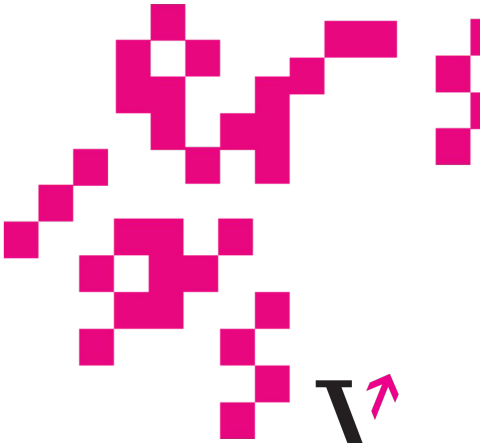
- **Multimodal Gains:** Adding images consistently improves the F1 score by 3–5%; smaller models see gains up to 17%.
- **Efficiency:** Parameter-efficient methods recover 97–99% of full fine-tuning performance.
- **The Reasoning Gap:** While models are good at labeling, a performance gap remains in their ability to generate “Faithful” rationales.



VECTOR INSTITUTE | INSTITUT VECTEUR

# Conclusion & Path Forward

- **Impact:** A scalable tool for content moderators to detect subtle media framing.
- **Current Limitations:** Focus is currently English/Western-centric; labels are binary.
- **Next Steps:** Expanding to multi-dimensional bias (e.g., partisan vs. sensationalist) and diverse cultural contexts.



VECTOR INSTITUTE | INSTITUT VECTEUR

# Thank you

Code: https://anonymous.4open.science/r/VILBias-367F

VECTOR INSTITUTE | INSTITUT VECTEUR

shaina.raza@vectorinstitute.ai

# What is AIXPERT

AIXPERT is an international research initiative funded by the European Union's Horizon Europe programme and the Swiss State Secretariat for Education, Research and Innovation (SERI). Our mission is to make AI smarter, safer, and more trustworthy across critical sectors such as healthcare, human resources, manufacturing, robotics, and the creative industries.

- Build an adaptable, explainable AI-agentic platform
- Define and assess AI trustworthiness
- Advance explainable multimodal foundation models
- Demonstrate real-world impact through pilot use cases (healthcare, recruitment, educational robotics, manufacturing and creative arts)

**Acknowledgements:**
Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. The AIXPERT Project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant No. 101214389, and from the Swiss State Secretariat for Education, Research and Innovation (SERI).

Funded by
the European Union