BiasLens: An End-to-End Pipeline for Detecting and Mitigating Bias in Vietnamese AI-Generated Text

Nguyen Hai Long, Ho Gia Bao, Huynh Bao Khang, Nguyen Trieu Duong, Pham Minh Hoang · Academic Supervisor: **Tom Huynh** · Industry Supervisor: **Ginel Dorleon, Quang Dang**

ABSTRACT

Generative AI systems increasingly power chatbots, moderation tools, and writing assistants, yet in low-resource languages such as Vietnamese they can produce biased or culturally inappropriate content that existing developer tooling cannot detect. We present BiasLens, a modular pipeline that classifies AI-generated Vietnamese text across 14 bias categories using a zero-shot and LoRA-fine-tuned Qwen2.5 language model (building on an earlier fine-tuned PhoBERT baseline), automatically rewrites flagged text into safer alternatives via an LLM-based mitigation module, and exposes the full workflow through a FastAPI service and web interface for developers to inspect and act on.

1 INTRODUCTION

Generative AI is widely deployed for Vietnamese-language chat, moderation, and writing assistance, but may generate stereotyped or culturally demeaning content with no tooling in place to catch it. This gap motivates a purpose-built detection-and-mitigation pipeline for Vietnamese text.

2 OBJECTIVES

- Detect bias across multiple categories in Vietnamese AI-generated text
- Evaluate fairness and neutrality of LLM outputs
- Automatically mitigate flagged bias via AI-assisted rewriting
- Package the pipeline as a practical, developer-facing tool

3 KEY DELIVERABLES

- ✓ Detection & serving pipeline (FastAPI)
- ✓ 14-category multi-label bias classifier
- ✓ Labeled Vietnamese bias dataset
- ✓ LLM-based mitigation (bias-aware rewriting)
- ✓ Developer-facing web interface (BiasLens)

4 METHODOLOGY

Sentences are deduplicated and split into a task-grouped train/val/test set. A PhoBERT classifier serves as the initial single-label baseline across 13 bias categories; the current model re-frames detection as zero-shot classification refined with LoRA fine-tuning on Qwen2.5, producing multi-label scores across 14 categories that follow the structure of the VNFairness dataset currently being developed by our supervisors. Flagged text is passed to an LLM-based mitigation step that rewrites it into a neutral alternative.

5 RESULTS

Binary bias-presence detection reaches 49% macro F1 across all 14 categories, rising to 70% on the 7 categories with enough test data to trust the score. Tested on 292 held-out examples never seen in training - see selected screens at right.

6 SYSTEM ARCHITECTURE

Input
AI-generated Vietnamese text (chat, moderation, writing-assistant output)

↓

Detect
Zero-shot + LoRA-fine-tuned Qwen2.5 - 14 bias categories, multi-label

↓

Mitigate
LLM rewrites flagged text into a safer, non-biased version

↓

Evaluate
Dashboard reports the classifier's per-category precision, recall, and F1 on held-out test data

Served via FastAPI; the BiasLens Next.js UI drives the full flow end to end.

7 DISCUSSION

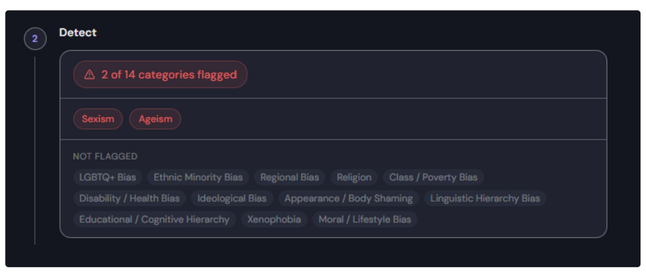
The LoRA-fine-tuned Qwen2.5 model reaches up to 90% F1 on well-represented categories such as LGBTQ+ bias and disability bias, showing that pairing zero-shot classification with lightweight fine-tuning can reliably catch clear, well-documented bias patterns. Categories with the fewest labeled examples score lower, a data-scarcity signal for future data collection rather than a modeling failure.

8 CONCLUSION

- A working end-to-end bias detection & mitigation pipeline for Vietnamese text
- LoRA fine-tuning on a pretrained multilingual LLM meaningfully outperforms from-scratch classifiers under limited labeled data
- Next: grow labeled examples for the most data-scarce categories

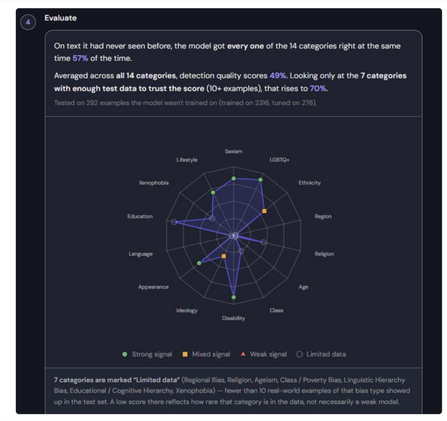
SCREENSHOTS

Input



Detect

Mitigate



Evaluate

Evaluate - per-category detail

ACKNOWLEDGMENTS
Scan to view this project on the RMIT Virtual Showcase website. With thanks to our academic supervisor, industry supervisor, and SSET for their guidance throughout this project.