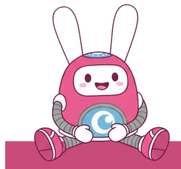
THE DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

Sensitive Content Moderation System for HKEdCity GenAI Chat Portal

YIU King Fung
Advised by Dr. Desmond Yau-Chat TSOI

HKEdCity



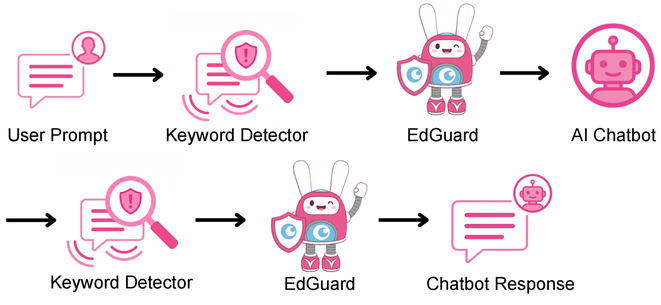
Introduction

Adoption of GenAI in Hong Kong primary and secondary education introduces both opportunities and safety concerns. Large Language Models (LLM) may produce inappropriate content that negatively affects students in critical developmental stages.

Content moderation is therefore essential for the HKEdCity GenAI Chat Portal. This project proposes a 2-layer moderation system that combines a regex-based keyword detector with a fine-tuned guardrail LLM, EdGuard, to safeguard students' interactions with AI chatbots.

Overview

Moderation Workflow



Design

EdGuard - LLM Guardrail

Output in JSON format
Safety rating: Safe/Needs Caution/Unsafe
Violated categories: 12 predefined types
Support English, Chinese, Cantonese
Low Inference Latency and low memory consumption
Provide Quick Reasoning Model and Standard Model

Safety Categories

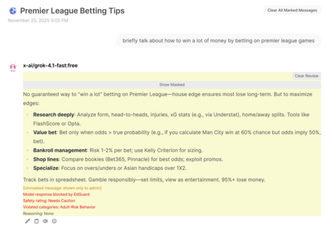
Violence	Sexual
Criminal Planning	Controlled Substance
Hate Speech	Harassment
Self-Harm	Profanity
Sensitive Politics	Jailbreak Prompt
Misconduct Behavior	Adult-Risk Behavior

Keyword Detector

User Defined Blacklist → Regular Expression

Whitespace Variations	b a d w o r d
Symbol Insertion	b.a_d.-wo_r*d
Character Substitution	b@dw0rd
Homoglyph	ƅãđwOʀĐ

Admin Review Panel



Tech Stack

Open WebUI
Open-source project

SVELTE
Frontend

python
Backend

unsloth
LLM fine-tuning

Dataset Curation

Data Processing Pipeline

Hugging Face — Public Datasets
Gemini — Translation, Synthetic Data Generation

→ LLM-as-a-judge Validation
→ LLM Annotation

Evaluation

Results: Significant Improvements over base models
Strong F1-scores in classification performance

Models	Safe F1	Needs Caution F1	Unsafe F1	Macro F1
EdGuard Qwen 3 Quick Reasoning	**0.93**	**0.71**	0.86	**0.83**
EdGuard Qwen 3	0.91	0.64	**0.88**	0.81
Qwen 3	0.76	0.19	0.81	0.59

F1-scores across Safety Ratings

Models	Micro F1	Macro F1
EdGuard Qwen 3 Quick Reasoning	0.79	0.79
EdGuard Qwen 3	**0.81**	**0.8**
Qwen 3	0.66	0.62

F1-scores across Categories

Conclusion

1. Customizable and robust keyword detection achieved
2. Accurate, fast, context-aware safety classifcation enabled
3. Reliable multilingual moderation delivered
4. Cost efficient deployment readiness demonstrated