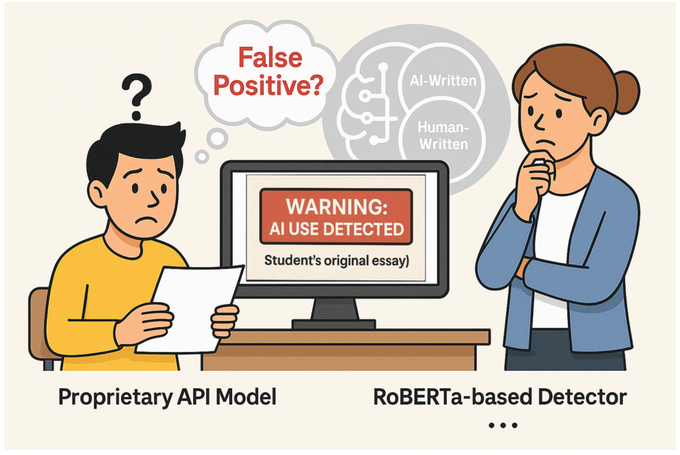
Detecting **AI Influence in Student Writing:** Toward Reliable and Interpretable Classifiers



Ibrahim R Hallac, Abdelaziz Qassi, Hasan Ogul

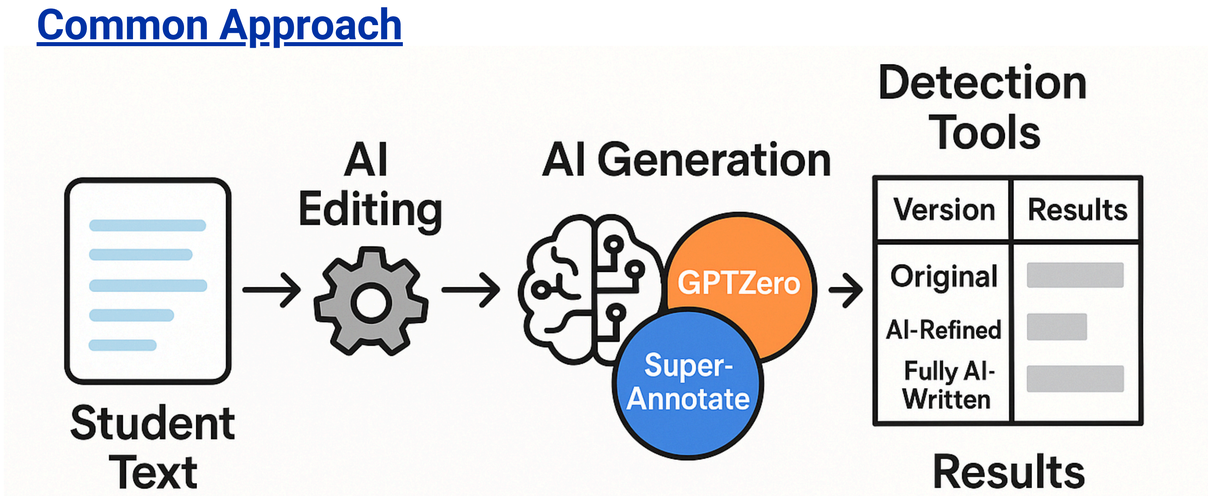
Department of Computer Science and Communication
Østfold University College, Norway

PROJECT OVERVEIW

Large Language Models (**LLMs**) are increasingly used in student writing, raising concerns about authorship, academic integrity, and fairness. This study investigates how to detect varying levels of AI involvement — from fully human-written to lightly AI-edited and fully AI-generated texts.

🎯 Goal: Support the development of fair, transparent, and educationally appropriate AI detection tools.

Common Approach



Study Summary

We conducted a series of manual experiments to observe how detection systems respond to different levels of AI intervention in student writing.

- Selected multiple authentic student-written paragraphs
- For each paragraph, created three versions:
  - Original (unaltered student text)
  - AI-refined (light grammar/style edits)
  - Fully AI-written (same meaning, different wording)
- Evaluated all versions using:
  - GPTZero (Proprietary detection API)
  - SuperAnnotate (Open-source RoBERTa-based model)

We observed consistent patterns across cases and include representative examples to illustrate key trends.

Detection Tools Compared

Tool	Type	Key Features
GPTZero	Proprietary	- Widely used in education - Outputs labels: AI / Mixed / Human
SuperAnnotate (RoBERTa-large)	Open-source	- Based on **RoBERTa-large** - Outputs a **probability score** (sigmoid) - **Runs locally** - Public on Hugging Face: https://huggingface.co/SuperAnnotate/roberta-large-llm-content-detector

Trade-offs between accessibility, interpretability, and practical deployment.

Representative Detection Examples

Some illustrative examples showing detection outcomes across versions.

Example A – Topic: Working Alone for Success

- **Original**: "...you must try to learn or do something new every single day of your life."
- **AI-Refined**: "Gaining deeper understanding requires trying something new every day."
- **Fully AI-Written**: "We must try to learn something new every day."

Version	SuperAnnotate Score	GPTZero Judgment
Original	0.0213	1% AI, 99% Human
AI-Refined	0.9431	6% AI, 94% Human
Fully AI-Written	0.9834	8% AI, 92% Human

Example B – Topic: Online Classes at Home

- **Original**: "...some students have to work and they are too busy to take online classes."
- **AI-Refined**: "Some of them have to work and are too busy."
- **Fully AI-Written**: "School should remain a place for focused learning and home for rest."

Version	SuperAnnotate Score	GPTZero Judgment
Original	0.9652	2% AI, 98% Human
AI-Refined	0.9765	42% AI, 58% Human
Fully AI-Written	0.5940	100% AI

Observation:

- SuperAnnotate often gives false positives, assigning high AI probabilities even to authentic student writing — regardless of fluency level.
- GPTZero is more conservative overall, but still shows increased suspicion for lightly edited texts, especially when grammar or clarity improves.
- Fully AI-generated texts are usually flagged correctly by both tools, indicating that clear-cut cases are detectable.

Dataset Contribution

We are building a carefully curated detection dataset that includes:

- Authentic student-written texts
- AI-edited and AI-generated variants (content-preserving)
- Detection outputs from both commercial and open-source tools

This goes beyond one-off LLM responses or synthetic benchmarks — it reflects real educational edge cases, shaped through careful versioning and evaluation

Why it matters:

- Can serve as a benchmark to test future LLMs and detection tools
- Enables reproducible research in AI influence detection
- Facilitates fairer models grounded in classroom realities

Impact & Outlook

This project contributes:

- A reusable, real-world benchmark dataset for detection studies
- A hands-on, comparative evaluation of detection strategies
- Interpretability, generalization, and educational alignment — essential for trustworthy AI in learning environments
- Insights into model behavior, fairness concerns, and deployment

ACKNOWLEDGMENTS

With support from the AI4AfL project (2022–2025), funded by the Research Council of Norway (Grant No. 326607).