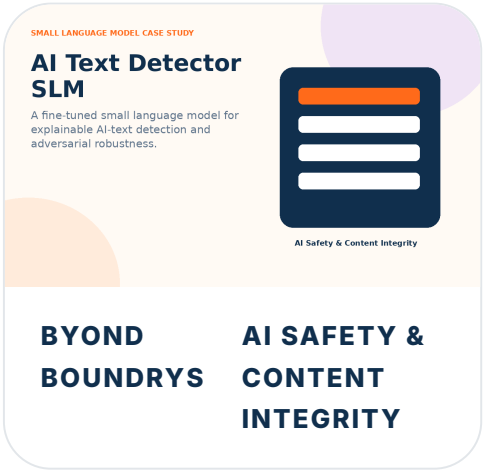
SMALL LANGUAGE MODEL CASE STUDY

AI Text Detector SLM

A fine-tuned small language model for explainable AI-text detection and adversarial robustness.

The rapid proliferation of AI-generated text across professional and institutional environments has introduced a new class of operational risk. Organizations can no longer assume that written submissions, published content, or support communications are human-authored and the consequences of misclassification vary significantly by context.

> **Project classification:** This prototype explores explainable and adversarially tested AI-text classification using a fine-tuned small language model.



INDUSTRY	CLIENT TYPE	PROJECT STAGE
AI Safety & Content Integrity	AI model product	Prototype
EVIDENCE	**DELIVERY SCOPE**	
Pilot-observed	Classification model, explainability and adversarial evaluation	

Explainable AI-text classification

The prototype produces explainable AI-text risk classifications rather than relying on a label alone.

Risk-based results rather than binary labels alone

Confidence outputs support risk-based review instead of automatic conclusions.

Improved resilience to rewritten AI text

Adversarial testing evaluates how the classifier behaves on paraphrased and rewritten AI text.

01 / BUSINESS CONTEXT

Who needed the solution

The rapid proliferation of AI-generated text across professional and institutional environments has introduced a new class of operational risk. Organizations can no longer assume that written submissions, published content, or support communications are human-authored and the consequences of misclassification vary significantly by context. Hiring teams encounter AI-generated job applications and assessments. Academic institutions handle AI-assisted submissions.

02 / CHALLENGE

What needed to change

Detection must remain accurate even after deliberate paraphrasing or humanization. This adversarial robustness requirement fundamentally reshapes training methodology, dataset design, and feature engineering.

Simple AI-or-human labels lacked useful confidence	Binary outputs without confidence most systems provide only AI/Human labels without risk scoring or uncertainty indicators.
Users could not understand why text was flagged	Lack of explainability stakeholders cannot understand or validate why text is flagged.
Rewriting and paraphrasing could bypass basic detectors	Vulnerability to rewrites paraphrasing or humanization tools can bypass standard detectors.
Structured or non-native writing created false positives	High false positives structured writing styles and non-native language patterns are often misclassified.
Large general models increased cost and latency	

Vulnerability to rewrites paraphrasing or humanization tools can bypass standard detectors.

03 / WORKFLOW TRANSFORMATION

Before and after

Previous workflow

- Simple AI-or-human labels lacked useful confidence
- Users could not understand why text was flagged
- Rewriting and paraphrasing could bypass basic detectors
- Structured or non-native writing created false positives
- Large general models increased cost and latency

Structured workflow

- Build a hybrid training set
- Fine-tune RoBERTa with LoRA
- Produce confidence and explanations
- Test adversarial robustness
- Review structured outputs

04 / SOLUTION

How we approached it

The solution is built on a fine-tuned RoBERTa classifier using LoRA (Low-Rank Adaptation), trained on a hybrid dataset and enhanced with explainability and robustness layers. Model Selection: RoBERTa with LoRA RoBERTa was selected for its strong sequence classification capabilities and sensitivity to stylistic patterns distinguishing AI from human text.

01

Build a hybrid training set

Combine public, human-written, AI-generated and rewritten examples across domains.

02

Fine-tune RoBERTa with LoRA

Adapt an efficient classifier without full-model retraining.

03

Produce confidence and explanations

Return risk scores and interpretable text-level signals instead of only a label.

04

Test adversarial robustness

Evaluate paraphrased and humanised text alongside false-positive behaviour.

05 / EXAMPLE WORKFLOW

From input to reviewable output

1 **Build a hybrid training set** → 2 **Fine-tune RoBERTa with LoRA** → 3 **Produce confidence and explanations** → 4 **Test adversarial robustness**

5 **Review confidence and explanation** → 6 **Human evidence-based decision**

06 / DELIVERY SCOPE

What the delivery covered

Model and dataset workflow

- Hybrid human and AI-text dataset construction
- Label and split preparation
- Adversarial rewrite test set
- Confidence and explanation outputs

Model engineering

- RoBERTa fine-tuning with LoRA
- Class-imbalance and calibration review
- Robustness and false-positive evaluation
- Structured inference API

Application engineering

- React review interface
- FastAPI service layer
- Pydantic validation
- Python evaluation pipeline

Scope boundary

The prototype produces probabilistic risk signals and must not be treated as proof of authorship or misconduct. No benchmark value is published without dataset and false-positive documentation.

---

07 / ARCHITECTURE AND CONTROLS

How the system is organised

RoBERTa

React

Python

FastAPI

Pydantic

PyTorch

Pandas

LoRA

Probabilistic result boundary

Detection output is a risk signal and must not be treated as conclusive proof of authorship.

False-positive evaluation

Human-written and adversarially rewritten text must be included when evaluating model behaviour.

Confidence and explanation

The interface exposes confidence and supporting indicators for reviewer interpretation.

Human review

Consequential decisions require independent evidence and qualified human review.

08 / BUSINESS VALUE

Value observed during validation

Pilot-observed

Explainable AI-text classification

The prototype produces explainable AI-text risk classifications rather than relying on a label alone.

2

Risk-based results rather than binary labels alone

Pilot-observed

Confidence outputs support risk-based review instead of automatic conclusions.

3

Improved resilience to rewritten AI text

Pilot-observed

Adversarial testing evaluates how the classifier behaves on paraphrased and rewritten AI text.

4

Reduced false-positive risk through broader evaluation

Pilot-observed

Broader evaluation helps identify false-positive risk on human-written content.

5

Efficient deployment with a fine-tuned smaller model

Pilot-observed

A fine-tuned smaller model demonstrates a path to lower-latency and lower-cost inference.

09 / TECHNOLOGY

Technology and component roles

RO **RoBERTa**
Fine-tuned text classification model

RE **React**
Reviewer or product user interface

PY **Python**
AI, data-processing and backend logic

FA	**FastAPI**	Backend APIs and workflow orchestration
PY	**Pydantic**	Structured schema and output validation
PY	**PyTorch**	Model training and inference
PA	**Pandas**	Tabular data processing and analysis
LO	**LoRA**	Parameter-efficient model fine-tuning
NU	**NumPy**	Numerical processing

10 / RELATED WORK

Explore related case studies

CONTENT TECHNOLOGY & AI WRITING

Humanizer SLM

Constraint-first text rewriting that improves naturalness while preserving names, numbers and critical facts.

CONTENT TECHNOLOGY & MARKETING

YourSocial

A structured content platform that moves users from thinking to outlining, writing and controlled refinement.

HEALTHCARE & CLINICAL RESEARCH

Clinical Data Harmonization

Hybrid RAG and fine-tuned small language models for explainable source-to-standard clinical data mapping.

Working on a similar challenge?

Discuss the workflow, data, integrations and evidence required for a practical first release.