# Detecting AI-Generated Text

Targeting Academic Integrity Applications

Afnan Alabdulwahab • Carter Day • Brennan Thompson

https://github.com/AfnanAbdul/TuringLens

# TABLE OF CONTENTS


01 Motivation

02 Data

03 Model

04 Results

05 Future Work

# The Academic Integrity Challenge

**Current Challenges:**

- Growing difficulty in distinguishing between human-written and LLM-generated text
- Current detectors often struggle with bias: high accuracy for AI detection but poor performance on human text
- Real consequences: Students' legitimate work incorrectly flagged as AI-generated
    - OpenAI's Classifier and Turnitin's detector faced significant backlash
    - A Texas A&M professor accused an entire class of using ChatGPT


**Why This Matters:**

- **False positives harm students** - Incorrectly flagged work leads to academic penalties
- **Academic integrity systems require balance** - Both missing AI text and misclassifying human work have consequences

**Goal:**

To fine-tune and compare language models for detecting AI-generated text — with a specific focus on academic writing.

TechCrunch: "OpenAI scuttles AI-written text detector over 'low rate of accuracy'" (July 25, 2023)
Rolling Stone: "Texas A&M Professor Wrongly Accuses Class of Cheating With ChatGPT" (May 17, 2023)

# Datasets: Two-Stage Approach

Kaggle LLM - Detect AI Generated Text

**Purpose:** Primary training & baseline evaluation

**Size & Balance:** 29,000 essays (17,508 human / 11,637 AI)

**Content:**

- Student essays written by humans or generated by LLMs.

**Strength:**

- Lightweight — ideal for fast, initial fine-tuning
- Academic essay format matches our use case

RAID (Robust AI-generated text Detection)

**Purpose:** Evaluation & further fine-tuning.

**Scale:** Subset selected from 10+ million documents

**Content:**

- Domains include: News, Books, Abstracts, Reviews, Reddit, Recipes, Wikipedia, Poetry
- We filtered to focus on the “Abstract” domain

**Strength:**

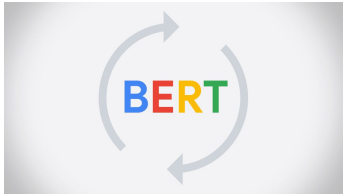
- 11 LLM source models and various writing styles
- Includes adversarial modifications to evade detection:
  - Paraphrasing, word substitutions, typo introduction
  - Designed to test robustness of detection models

# Model Exploration & Selection

We adopted a divide-and-conquer strategy to explore multiple model architectures in parallel.

**Models Explored:**

- GPT2: full fine-tuning
- DistilRoBERTa-base: full fine-tuning
- RoBERTa-LoRA: parameter-efficient fine-tuning


**Workflow**

- Each team member fine-tuned one of the base models on the Kaggle dataset.
- We then compared performance on a held-out test set.
- Models were compared based on Human detection accuracy, AI detection accuracy, and computational efficiency

**Model Selection**

- RoBERTa-LoRA achieved the best performance (99%) and efficiency (Required only 0.82% of trainable parameters).
- It was selected for further evaluation and domain-specific fine-tuning on the RAID dataset.

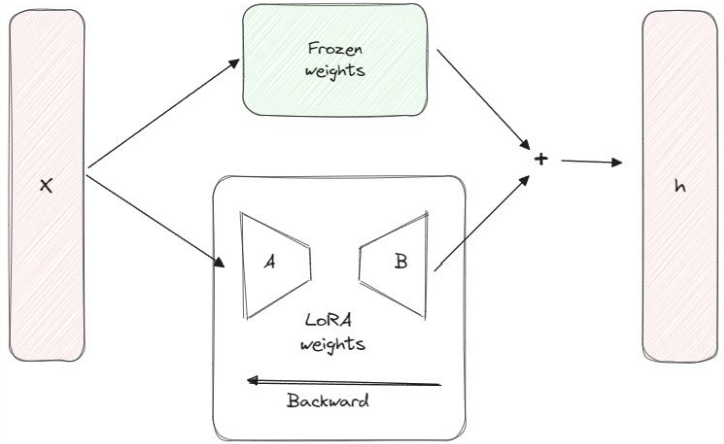
# RoBERTa with LoRA

**Base Model**: RoBERTa (Robustly optimized BERT approach) + Classification Head

- Transformer-based architecture with 125M parameters
- Strong baseline for text classification tasks
- Binary classification: Human (0) vs. AI-generated (1) text

**Parameter-Efficient Fine-Tuning**: Low-Rank Adaptation (LoRA)

- Trains only 0.82% of parameters (1.03M vs. 125M)
- Hyperparameters: r=8, alpha=32, dropout=0.1




LoRA

# Two-Stage Approach

**First Stage**: Fine-Tuning on Kaggle Dataset

**Training setup:** 5 epochs, batch size 16, learning rate $1\times10^{-4}$, AdamW optimizer, linear scheduler, no class weighting, evaluated on accuracy/F1.

**Evaluation on Test set:**

- Human Accuracy ≈ 99.5%
- AI Accuracy ≈ 99.8%

**Evaluation on External Dataset (RAID)**

- Sampled 3.5K abstracts from RAID
- (50% human, 50% across 11 AI models)
- Model over-predicted AI on human texts:
  - Human Accuracy: 16.8%
  - AI Accuracy: 97.3%

**Second Stage**: Further fine-tuning on RAID sample

**Training setup**: 10 epochs, batch size 16, learning rate $1\times10^{-4}$, linear scheduler, weighted loss (human:AI = 10:1), optimized for human accuracy.

**Addressing Classification Bias**

- Dataset: 1,766 human + 1,766 AI + 1,766 adversarial AI samples
- Applied 10:1 (human:AI) weighting ratio to penalize human misclassification
- Added Human-Focused Evaluation Metrics during training
  - Added specific tracking for human detection
  - Balanced accuracy instead of raw accuracy

# Results

## Model Accuracy on Kaggle Dataset

| Model | Human Accuracy | AI Accuracy | Overall Accuracy | Parameter Efficiency |
|---|---|---|---|---|
| GPT-2 | 95.5% | 99.8% | 97.2% | 100% params trained |
| DistilRoBERTa-base | 96.9% | 99.8% | 98.0% | 100% params trained |
| **RoBERTa-LoRA** | **99.5%** | **99.8%** | **99.6%** | **0.82% params trained** |

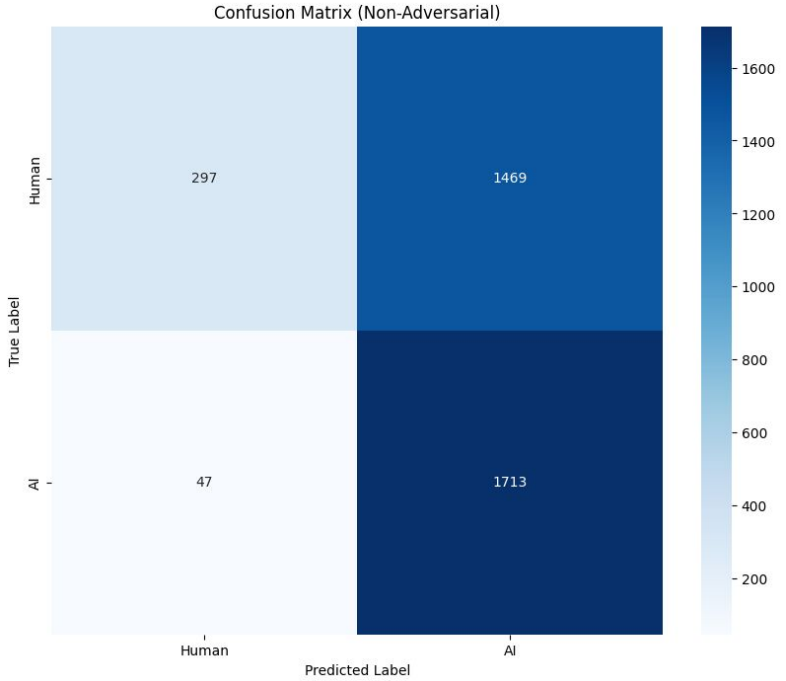
## RoBERTa-LoRA Evaluation (RAID Dataset):

Before/After Retraining RoBERTa LoRA on RAID dataset

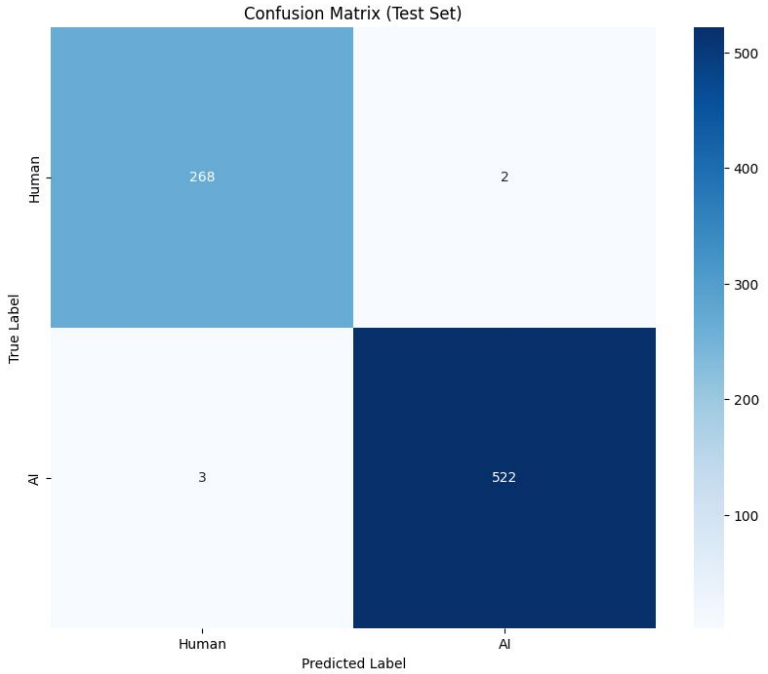
| Model | Human Accuracy | AI Accuracy | Overall Accuracy | Precision |
|---|---|---|---|---|
| **RoBERTa-LoRA (before)** | 16.8% | 97.3% | 57% | 53.8% |
| **RoBERTa-LoRA (after)** | 99.3% | 99.4% | 99.4% | 99.6% |

# Evaluation

Before Retraining RoBERTa-LoRA (RAID Evaluation)




After Retraining RoBERTa-LoRA (RAID Evaluation)



# Limitations & Success Factors

**Current Limitations:**

- Current models rely solely on binary labels (human vs. AI) without deeper linguistic or stylistic features.
- Model effectiveness may decrease over time as AI text generation improves, distinguishing features become more subtle.

**Success Factors:**

- Penalizing human misclassification (10:1 ratio) helped the model recover human accuracy without sacrificing AI detection.
- Training with human-focused metrics (human and balanced accuracy) guided better optimization than raw accuracy.

# Conclusion & Future Work

This project was designed specifically for the academic writing setting, focusing on detecting AI-generated content in essays and scholarly abstracts. Our approach reduced false positives of human-written academic text from 83.2% to 0.7% on the RAID dataset.

**Key Insights:**

- **False Positives Matter More Than Overall Accuracy**: In academic integrity contexts, misclassifying human work as AI-generated has serious ethical implications
- **Targeting the Right Metric Makes All the Difference**: Optimizing for human_accuracy rather than overall accuracy dramatically improved our model's usefulness in academic settings

**Opportunities for Future Work:**

- **Benchmark Against Existing Tools**
- **Check How Predictable the Writing Is**
  We can explore methods like GPTZero that look at how predictable or repetitive the text is — AI writing often feels more consistent or “too perfect” compared to humans.
- **Writing Style Measurements**
  - Add ways to measure how diverse the word choices are
  - Track sentence structure complexity to better distinguish human writing
- **Measure Vocabulary Depth**
  We can also look at how rich the vocabulary is — for example, how often rare words are used or how many different words show up — to spot subtle patterns between human and AI text.

# Thank You!