AuthenText

Machine-Generated Text Detection in Student Essays

Brendan Ho, Chris Wyble, Terence Pak

Introduction

Brendan Ho

Chris Wyble

Terence Pak

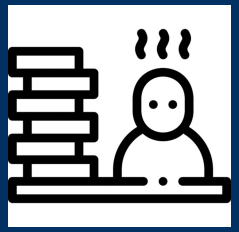
Primary Goal

This project aims to build an effective Machine-Generated Text (MGT) detection tool that determines if K-12th grade student essays are human-written or machine-generated text.

Problem Statement

MGT usage in student essays presents new and complex issues for educators

Millions of essays have been detected to have some MGT



Only 45% of MGT essays were correctly identified by educators

Available MGT detection tools are unreliable

Social Impact to Users

- **Users: K-12 educators**

- Uphold academic integrity
  - Educators can keep students accountable
  - Unethical usage of MGT will be deterred

- Educators can save time by not having to manually check every essay for MGT

- Improving our tool's accuracy will make students and educators less anxious of false results

Domain Expert Feedback

Gathered feedback from a K-12 educator to gather their personal experiences with MGT in student essays.

Key Takeaways:

- Key challenge of identifying MGT is that some students will use MGT partially in their essays (rephrasing or fill-in-the-blank)
- Highlighted text identifying what parts of an essay are MGT would be extremely helpful
- Fellow peers don't trust current MGT detection tools

Dataset

- DAIGT Proper Train Dataset (Kaggle)
  - ~160,000 essays
  - Collection of multiple essay datasets:
    - MGT from various generative text models (e.g. ChatGPT, Llama-70b, Falcon 180b)
    - K-12th grade human-written essays from various prompts
  - 72% human-written essays

ML Model Methodology

- Binocular score, B, is the ratio of perplexity and cross-perplexity

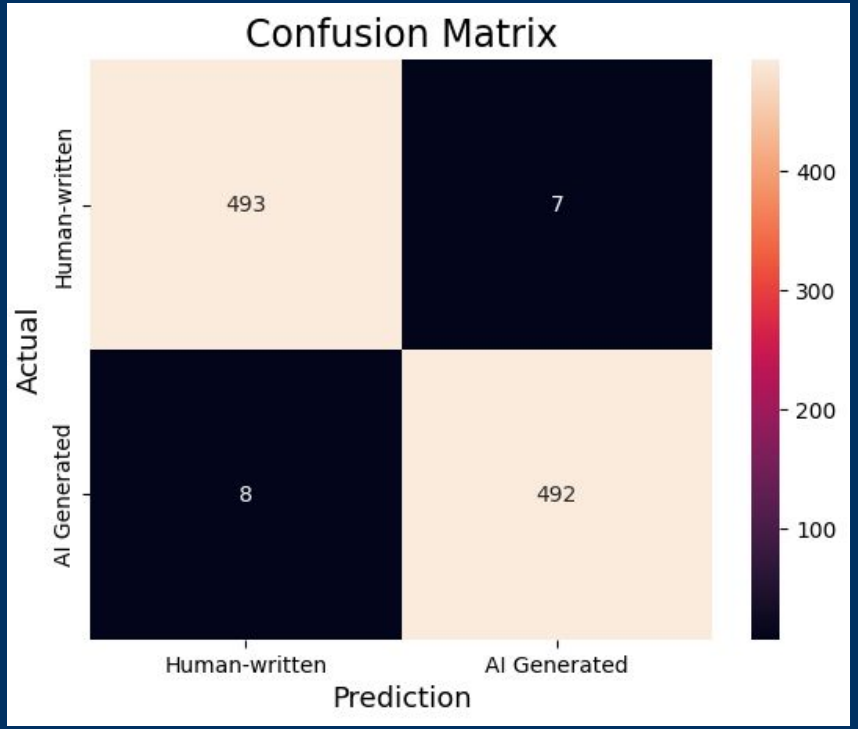
$$B_{\mathcal{M}_1,\mathcal{M}_2}(s) = \frac{\log \mathrm{PPL}_{\mathcal{M}_1}(s)}{\log \text{X-PPL}_{\mathcal{M}_1,\mathcal{M}_2}(s)}$$

  - Given 2 Language Models M_1 and M_2:
    - Perplexity: how surprising the next token in a given text is to M_1
    - Cross-perplexity: how surprising the next token prediction of M_2 is to M_1
- Intuition: Cross-perplexity helps normalizes how "surprising" a text is irrelevant of the prompt/context

Initial Model Testing

- Evaluated model on pure MGT vs human-written essays
  - 1000 samples
  - Balanced labels
  - Recall: 98%

- Building on domain expert feedback, we developed a way to test partial MGT

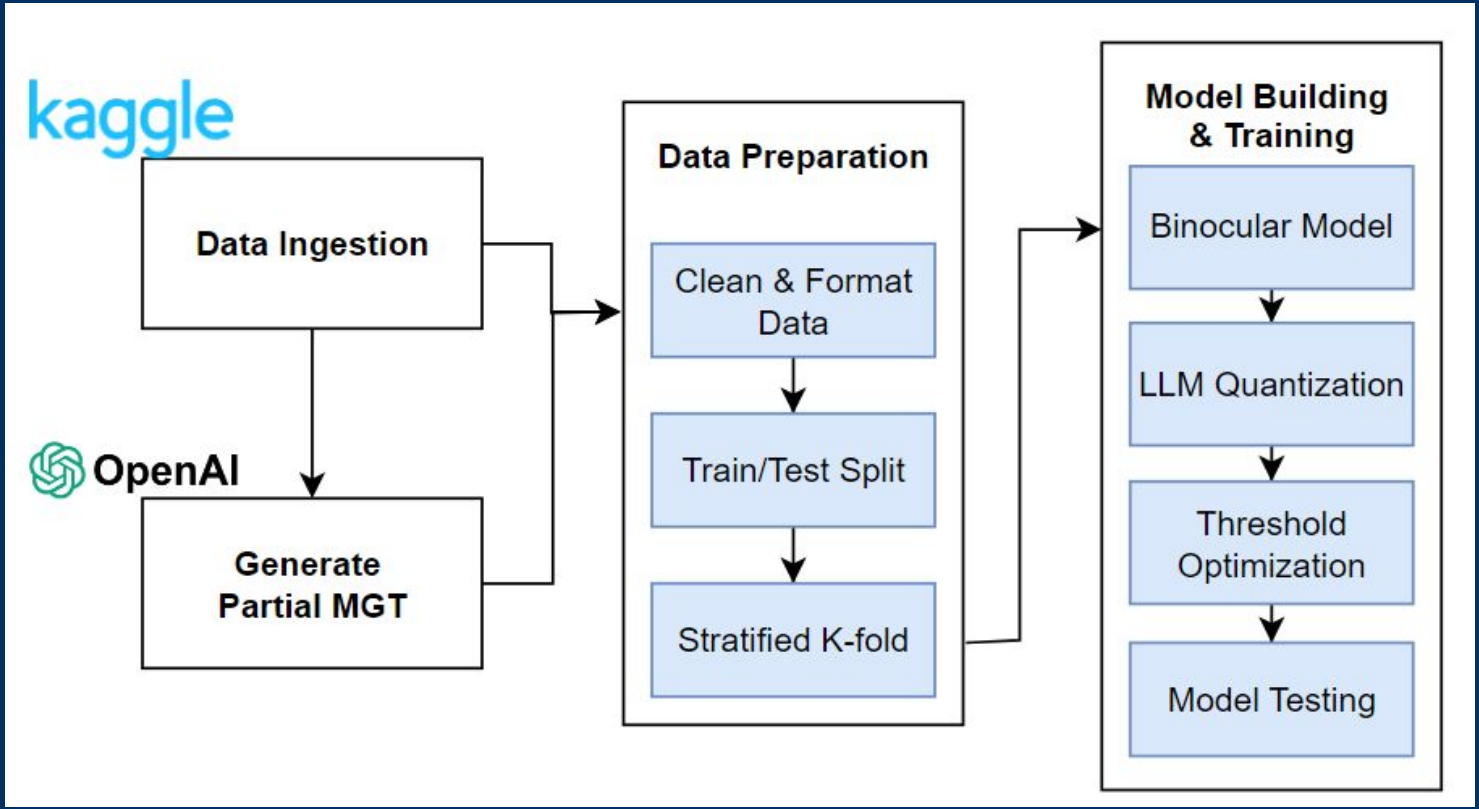


Dataset Generation

- Goal is to test the robustness of the Binoculars model on partial MGT dataset
- Each of 6 datasets contain 10,000 essays
  - 5,000 human-written essays
  - 5,000 essays from one of the partial-generation methods

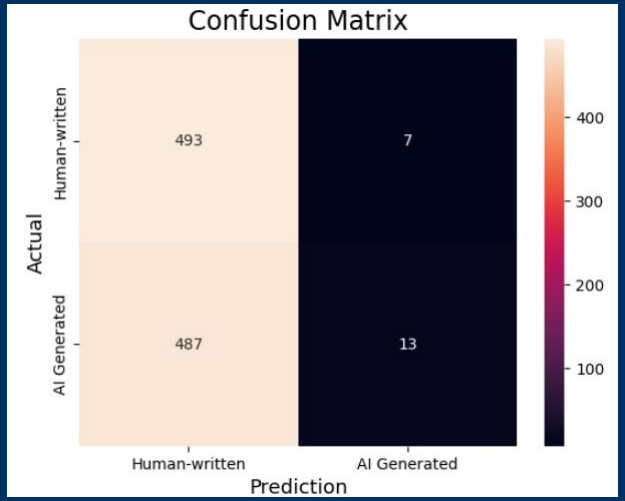
Method	Experiment		
Rephrasing	25% MGT	50% MGT	75% MGT
Fill-in-the-blank	25% MGT	50% MGT	75% MGT

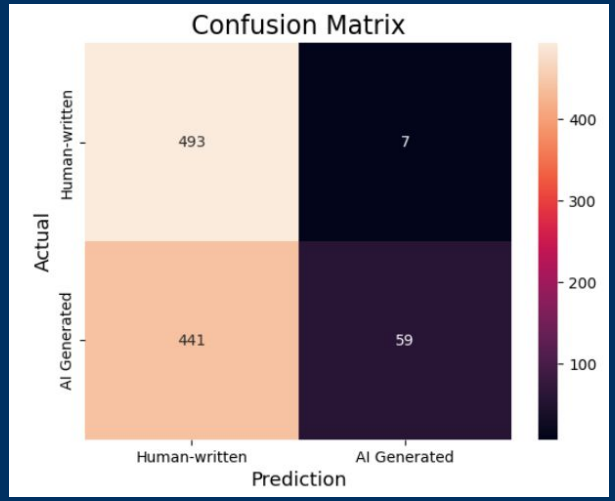
ML Pipeline

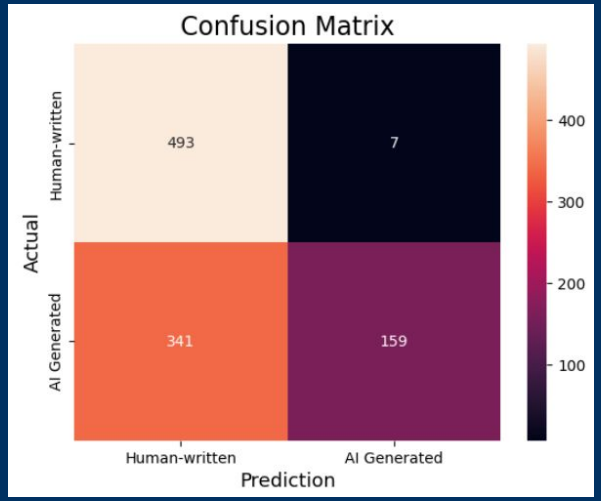

kaggle
Data Ingestion
OpenAI
Generate Partial MGT
Data Preparation
Clean & Format Data
Train/Test Split
Stratified K-fold
Model Building & Training
Binocular Model
LLM Quantization
Threshold Optimization
Model Testing

Partial Generation (Rephrase) Results

25% MGT	50% MGT	75% MGT
Recall: 51%	Recall: 55%	Recall: 65%

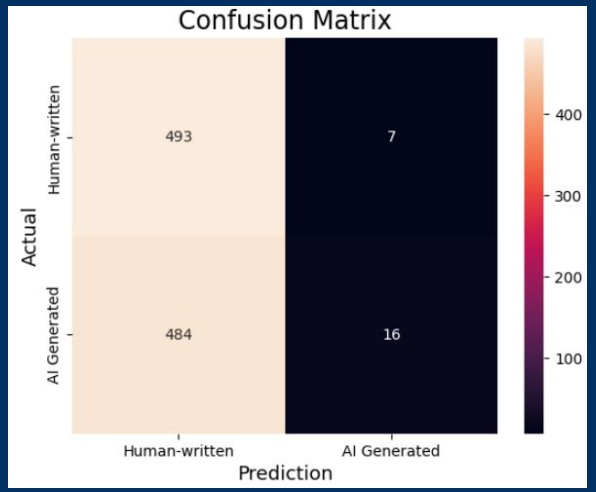


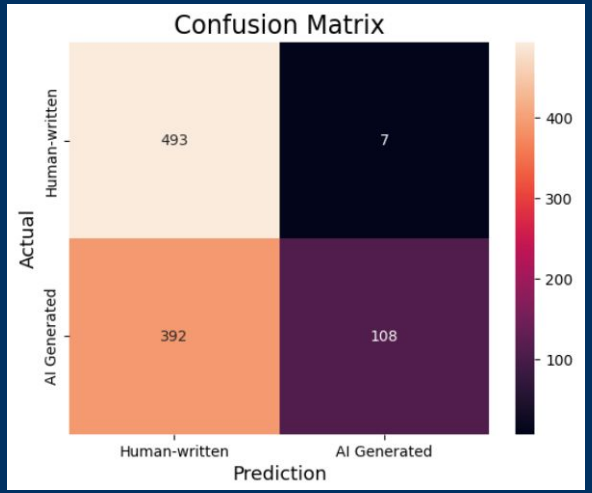


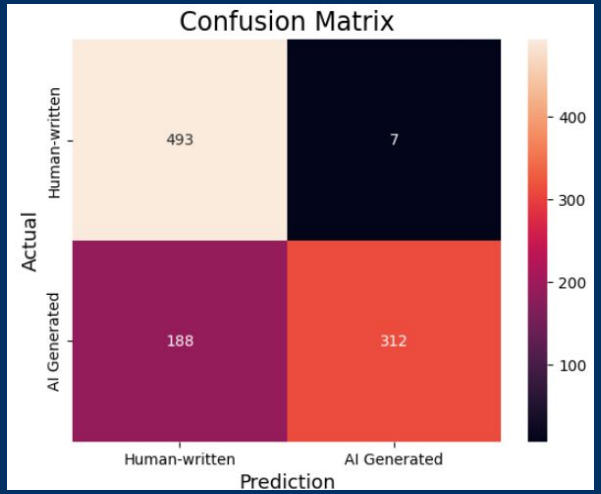


Partial Generation (Fill-in-the-Blank) Results

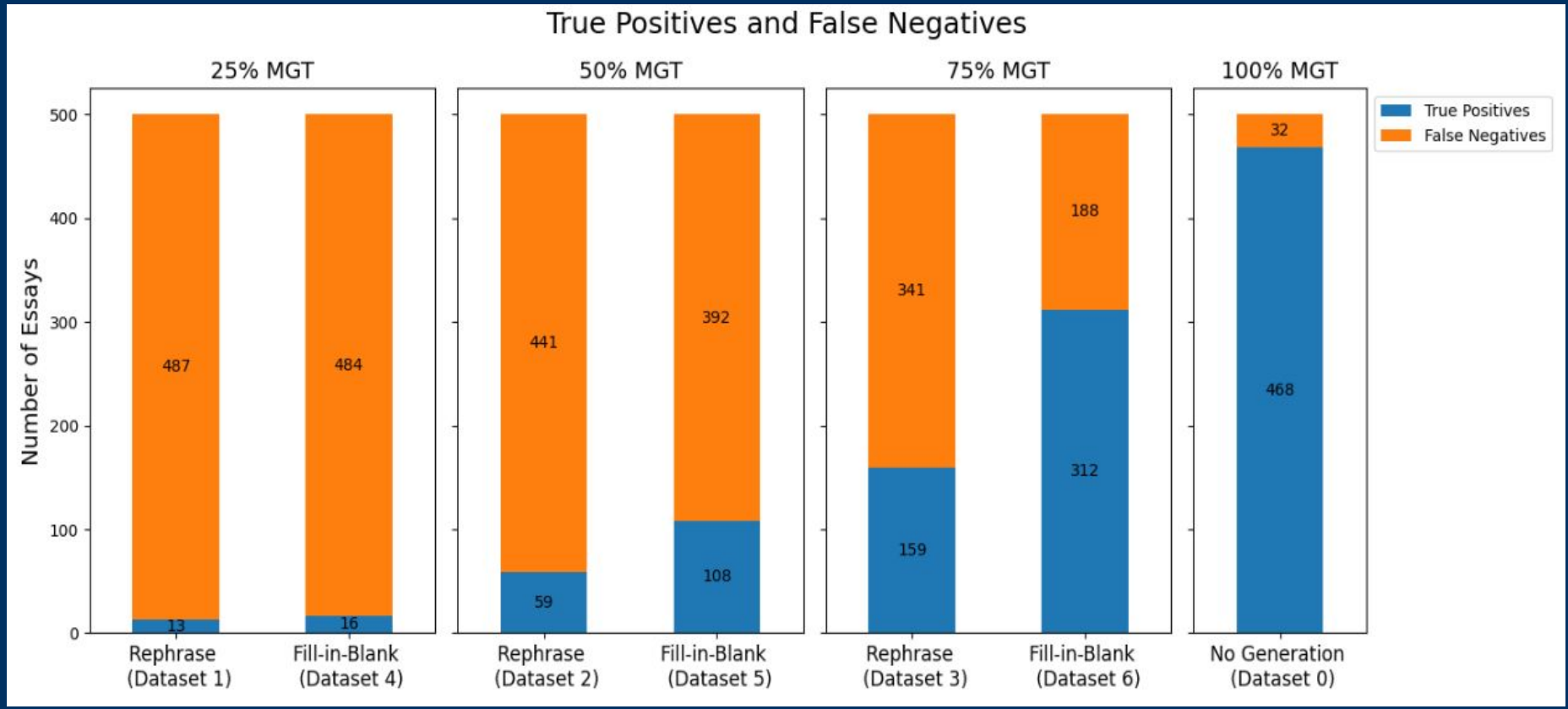
25% MGT	50% MGT	75% MGT
Recall: 51%	Recall: 60%	Recall: 80%



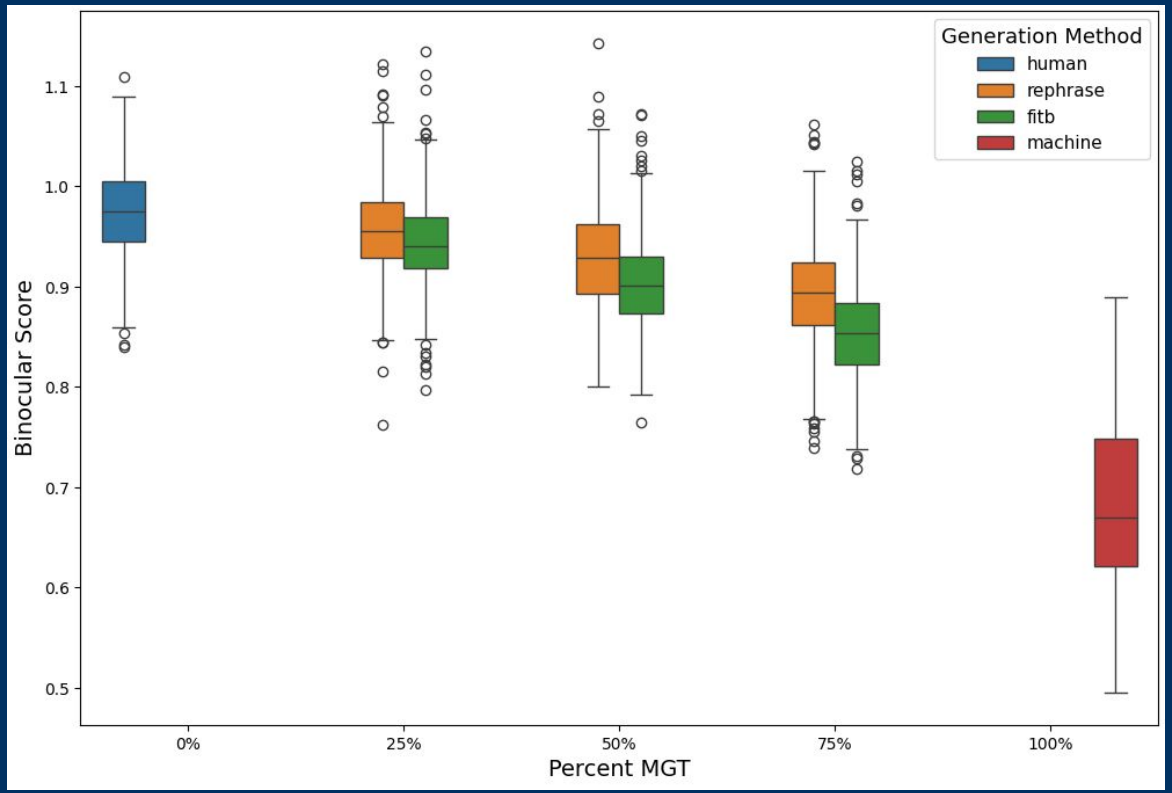




Partial Evaluation



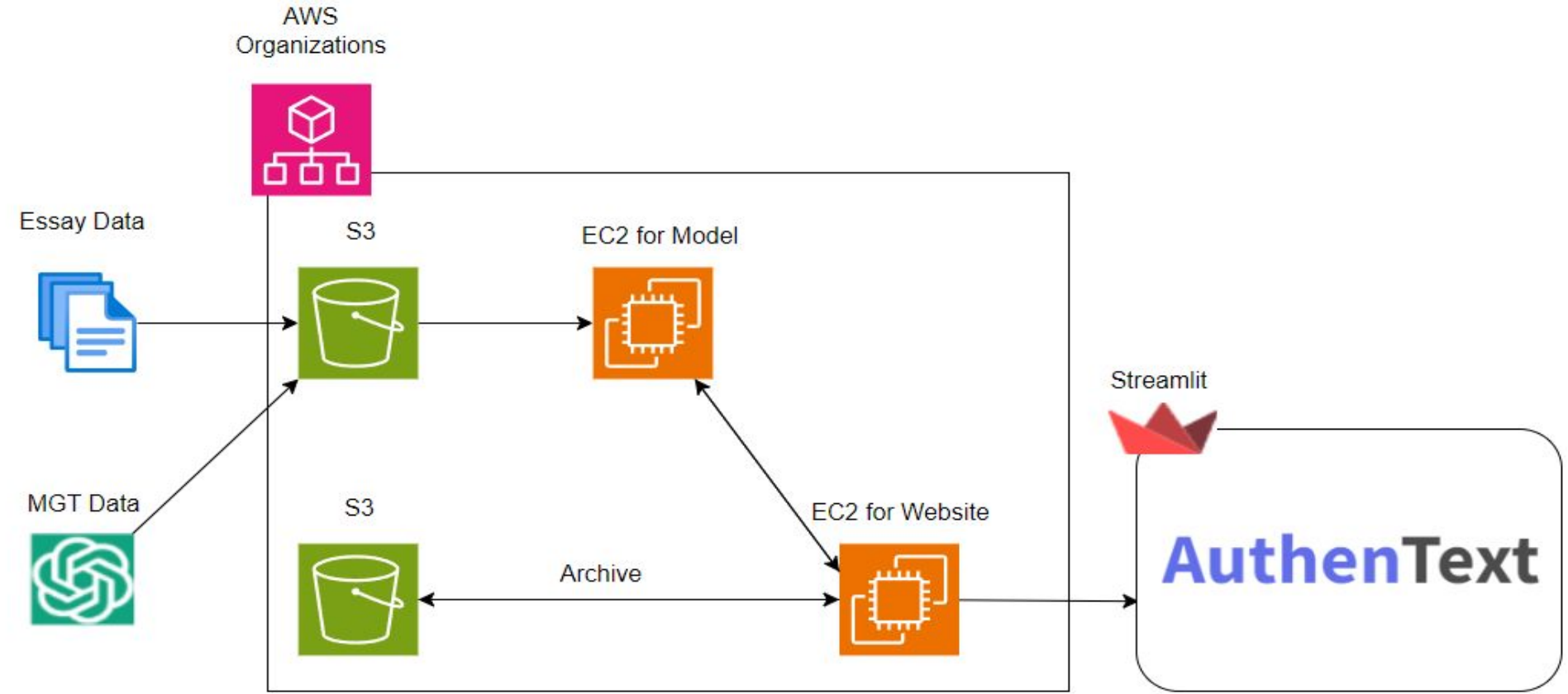
Partial Evaluation



Key Takeaways

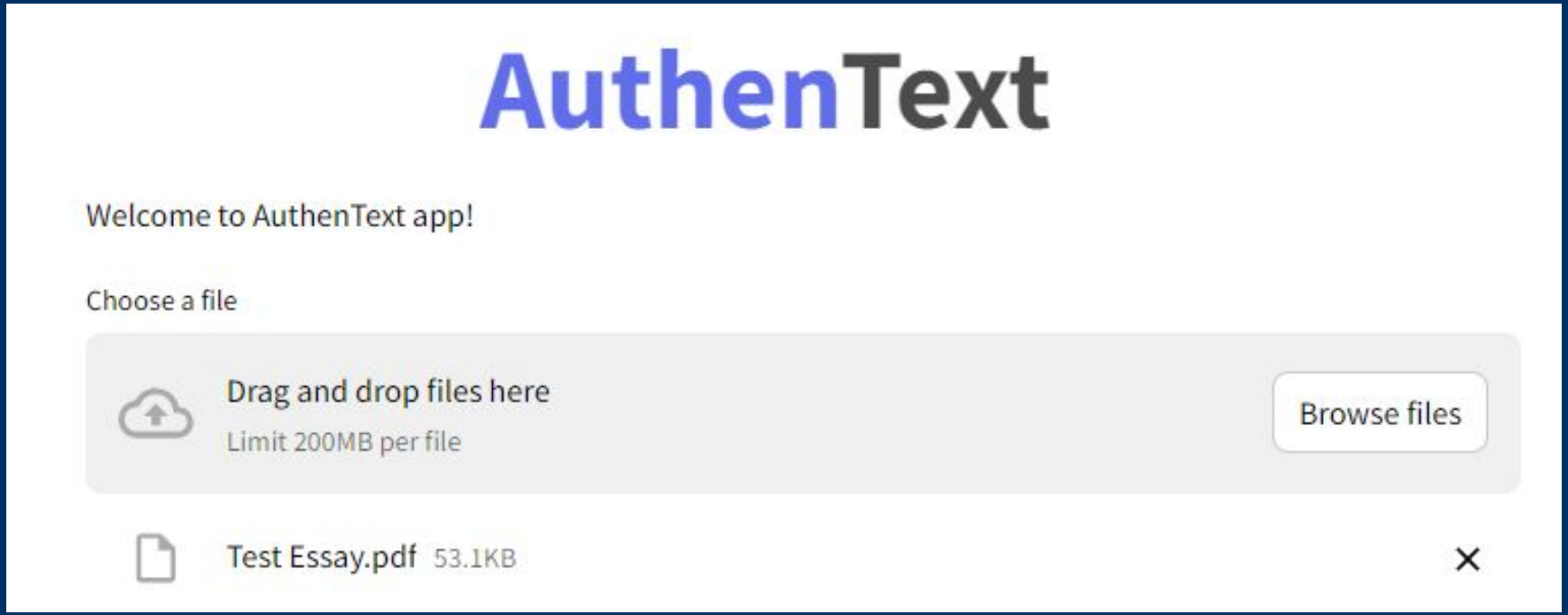
- Binoculars is not robust in classification of partial MGT essays but is robust in fully MGT essays
- Binoculars is able to detect “Fill-in-Blank” MGT more accurately than “Rephrase” MGT
- To Binoculars, there is more similarity between a completely human-written essay and an essay with a small amount of human text

Data Pipeline



MVP Demo

Website Link: **AuthenText**



Technical Challenges

- If provided a larger budget (non-AWS), we would have used Chat-GPT4 to generate MGT instead of GPT3.5
  - Applied data engineering techniques to cleanse and properly format essay data

- Hardware limitations - Requires 2 language models and thus requires a high GPU memory to run the model
  - Expanded memory to run (g5.12xlarge)

- Due to model size, we were unable to serve model on client side
  - Created an API to serve the model using a second EC2 instance

Beyond the Deadline

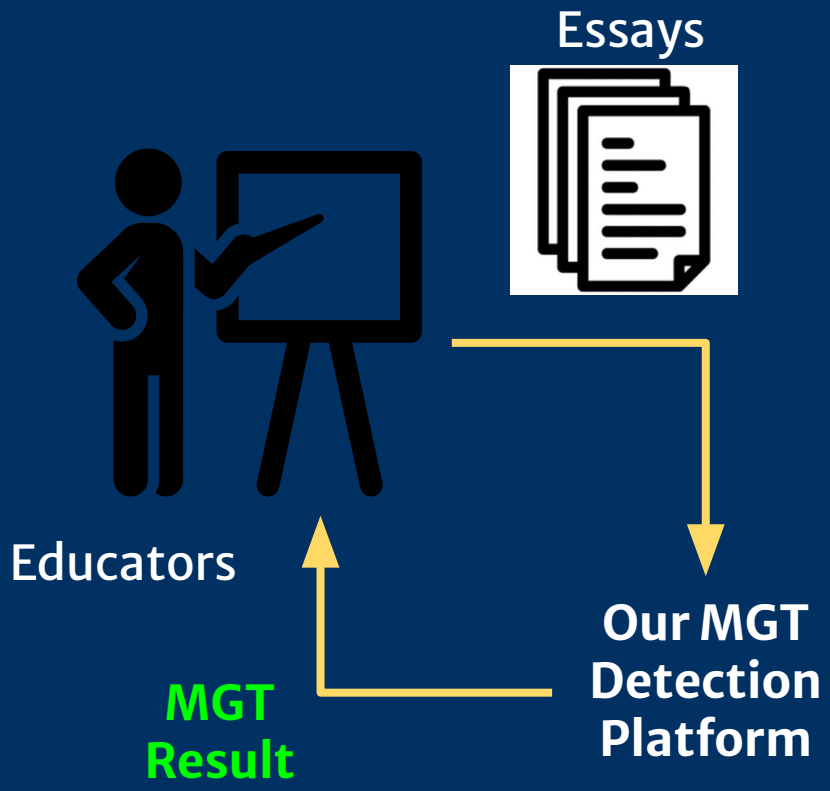
- Implement another set of datasets evaluating generation parameters and increase dataset size
- Experiment with more AI text detection models
- Establish multiple classifications rather than a binary value

Conclusion

Our mission is to help educators catch MGT in K-12th grade student essays with a Machine-Generated Text (MGT) detection tool.

Appendix

Product Overview



Key Function:
Educators will upload student essays to an interactive platform and receive an MGT detection result in the given text

Additional Feature:

- Highlighted phrases/text of highly suspect MGT

Limitation:

- It is not a plagiarism checker. It is an MGT detector
- It is not the decision maker. It is only a tool

Back to the User

- Educators can classify highly MGT essays.
- Educators should be cautious of partial MGT essays.
- For the sake of accuracy, some MGT can be permissible

Findings

There is an increase classification accuracy with a higher percentage of MGT

- Accuracy may not be the best metric (Is it good enough to tell a story?)
- Pure classification may not be ideal in a partial MGT dataset

Highlighting makes more sense?

- Go back into 2 of the same essay id and see how they appear on the highlight. Rephrasing should have darker coloring than fill-in-blank for given sentences
- Demo can include 5 essays

Include a binary determination

- Heavy caveat (Potential MGT vs Human-Written)
- Label is calculated by generate a binocular score for each token and sums it up. For given essay, each token should generate more perplexity than a machine.
- difficult to identify which token is MGT or not. It's all about signal. We don't know the threshold of a given token in order to classify as MGT or not. Finding max amount of signal can give us percentage of MGT signal.

References

1. Spotting LLMs With Binoculars
2. DAIGT Dataset
3. MGT Student Usage
4. MGT Identified by Educators
5. Available MGT detection tools are unreliable

Social Impact

- Uphold academic integrity
  - Educators can keep students accountable
  - Unethical usage of MGT will be deterred

- Educators can save time by not having to manually check every essay for MGT

- Improving our tool’s accuracy will make students and educators less anxious of false results

Target Users

Target Users: K-12 educators

It’s crucial to address cheating at an early age

- Address it before it becomes normalized
- Academic cheating can start as early as first grade
- 58% of high school students admitted to plagiarism

Technical Challenge

- LLMs tend to generate “unsurprising” text (low perplexity)
- However, without context from a prompt, it is difficult to assess how “surprising” a given text is
- Example below generated by GPT-4:

“Dr. Capy Cosmos, a capybara unlike any other, astounded the scientific community with his groundbreaking research in astrophysics. With his keen sense of observation and unparalleled ability to interpret cosmic data, he uncovered new insights into the mysteries of black holes and the origins of the universe. As he peered through telescopes with his large, round eyes, fellow researchers often remarked that it seemed as if the stars themselves whispered their secrets directly to him. Dr. Cosmos not only became a beacon of inspiration to aspiring scientists but also proved that intellect and innovation can be found in the most unexpected of creatures.” – GPT 4

Ethical Considerations

- Is there an effective method for a student to contest false positive results?
- Can this tool negatively impact relationship between student and educator?
- When can MGT be used ethically?

Proper Tool Usage

- Educators should always cross-check positive results
- Educators need to be cognizant of incorrect results, use social context, and must be the one to make the final decision
- Students should be informed about the existence of this tool and that their data will only be used for MGT evaluation