Can LLMs distinct AI Essays from Student Essays

By: Jake Shapiro

Introduction

Title

Can GPTZero's AI Vocabulary Distinguish Between LLM-Generated and Student-Written Essays?

Authors

Veronica Juliana Schmalz and Anaïs Tack

Publication Year

2025

Significance

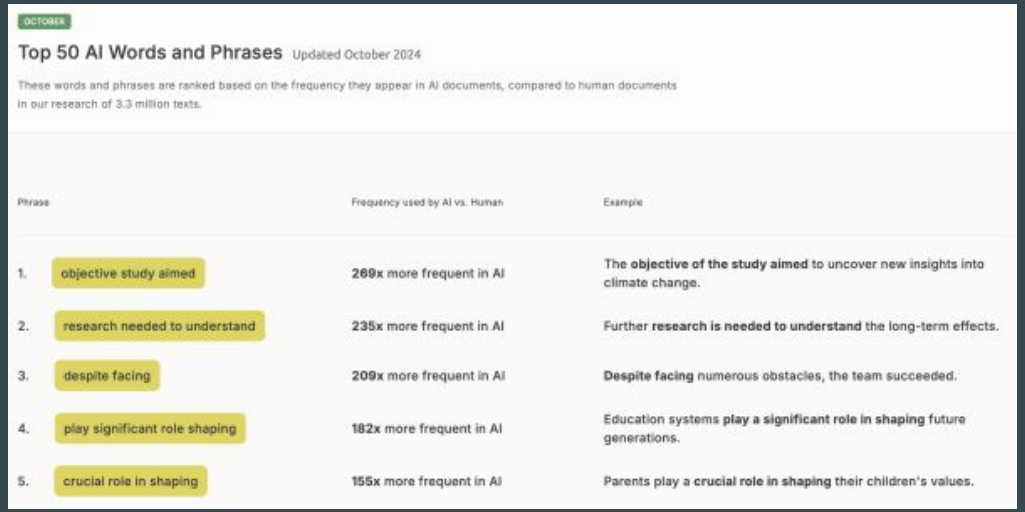
- Use of LLMs for writing essays
- Can teachers distinct LLM essays and student written essays?
- Are AI detectors accurate?

Background

Key Concepts

- Examined common phrases used by LLMs
  - Analyzed AI vocabulary lists published by GPTZero

- Tested on variety of LLMs



OCTOBER

Top 50 AI Words and Phrases Updated October 2024

These words and phrases are ranked based on the frequency they appear in AI documents, compared to human documents in our research of 3.3 million texts.

Phrase	Frequency used by AI vs. Human	Example
1. objective study aimed	**269x** more frequent in AI	The **objective of the study aimed** to uncover new insights into climate change.
2. research needed to understand	**235x** more frequent in AI	Further **research is needed to understand** the long-term effects.
3. despite facing	**209x** more frequent in AI	**Despite facing** numerous obstacles, the team succeeded.
4. play significant role shaping	**182x** more frequent in AI	Education systems **play a significant role in shaping** future generations.
5. crucial role in shaping	**155x** more frequent in AI	Parents play a **crucial role in shaping** their children's values.

Prior Research

- Study claimed GPTZero's effectiveness hasn't been "empirically validated"

- Journal of Korean Medical Science
  - Researcher: Farrokh Habibzadeh
  - Done in 2023

- Computer Information System Department
  - Researchers: Karen Paullet, Jamie Pinchot, Evan Kinney, Tyler Stewart
  - Done in 2024

Research Questions

Research Questions

- Primary Question
  - Can GPTZero's AI vocabulary distinguish between essays written by students and those generated by LLMs?

- Supporting Questions
  - How well do classifiers built solely on GPTZero's AI Vocabulary terms perform vs. classifiers trained on the full vocabulary of a dataset containing student and LLM-generated essays?
  - Do the AI Vocabulary lists generalize across different LLMs?
  - Which specific AI-related words and phrases contribute most to distinguishing LLM-generated texts from student written essays?

Methodology

Tools and Technologies Used

- GPTZero's AI Vocabulary lists (October 2024 - March 2025)

- Ghostbuster Dataset
  - 1,000 student essays
  - 1,000 ChatGPT essays
  - 1,000 Claude essays

- Python & scikit-learn
  - CountVectorizer
  - Naive Bayes classifiers

Step-by-Step

- Step 1
  - Collecting GPTZero's AI Vocabulary
- Step 2
  - Preparing the dataset
- Step 3
  - Feature Extraction (Bag-of-Words Approach)
    - Bernoulli vector
    - Multinomial vector
- Step 4
  - Training and Testing Models
- Step 5
  - Evaluation

Features	LLM	Vocabulary	Accuracy	Precision	Recall	F1	MCC
Bernoulli	All	GPTZero List: All	0.532	0.884	0.343	0.494	0.272
		GPTZero List: Oct	0.503	0.877	0.296	0.443	0.240
		GPTZero List: Nov/Dec	0.416	0.996	0.129	0.228	0.199
		GPTZero List: Jan/Feb/Mar	0.363	0.969	0.046	0.089	0.117
		Ghostbuster BoW	0.871	0.846	0.986	0.911	0.711
	Claude	GPTZero List: All	0.522	0.657	0.09	0.158	0.085
		GPTZero List: Oct	0.502	0.501	0.968	0.660	0.011
		GPTZero List: Nov/Dec	0.508	0.786	0.022	0.043	0.068
		GPTZero List: Jan/Feb/Mar	0.503	1.0	0.007	0.014	0.059
		Ghostbuster BoW	0.889	0.825	0.987	0.899	0.793
	GPT	GPTZero List: All	**0.755**	0.882	0.588	0.705	0.541
		GPTZero List: Oct	0.703	0.853	0.49	0.622	0.448
		GPTZero List: Nov/Dec	0.616	0.964	0.242	0.386	0.351
		GPTZero List: Jan/Feb/Mar	0.544	0.968	0.092	0.167	0.209
		Ghostbuster BoW	0.929	0.892	0.977	0.933	0.862
Multinomial	All	GPTZero List: All	0.517	0.891	0.314	0.464	0.263
		GPTZero List: Oct	0.452	0.910	0.197	0.324	0.212
		GPTZero List: Nov/Dec	0.410	0.968	0.119	0.213	0.191
		GPTZero List: Jan/Feb/Mar	0.363	0.969	0.046	0.089	0.117
		Ghostbuster BoW	0.901	0.955	0.893	0.923	0.787
	Claude	GPTZero List: All	0.518	0.673	0.072	0.130	0.082
		GPTZero List: Oct	0.504	0.538	0.064	0.114	0.019
		GPTZero List: Nov/Dec	0.508	0.786	0.022	0.043	0.068
		GPTZero List: Jan/Feb/Mar	0.503	1.0	0.007	0.014	0.059
		Ghostbuster BoW	**0.964**	0.976	0.951	0.964	0.928
	GPT	GPTZero List: All	**0.729**	0.884	0.527	0.660	0.501
		GPTZero List: Oct	0.654	0.895	0.350	0.503	0.390
		GPTZero List: Nov/Dec	0.604	0.964	0.216	0.353	0.330
		GPTZero List: Jan/Feb/Mar	0.539	0.964	0.081	0.149	0.194
		Ghostbuster BoW	0.912	0.942	0.877	0.909	0.825

Findings

Key Findings

1. GPTZero's AI Vocabulary Lists had limited coverage

2. Classifier performance was generally weak

3. Full-Vocabulary models performed much better

4. Individual "AI" words didn't help much

Main Takeaway

GPTZero's AI Vocabulary can't reliably distinguish student and AI-generated essays

Commentary

Critique & Evaluation

- Narrow Focus on Vocabulary
- No Human or Real-World Validation
- Reliance on a Limited Dataset

Quiz

1. What dataset did the researchers use for their experiments?

A. OpenAI's training dataset

B. A collection of Reddit essays

C. The Ghostbuster dataset

D. GPTZero’s propriety essay corpus

2. What was one major limitation noted by the researchers?

A. The study used too many deep learning models

B. The AI Vocabulary lists were not interpretable

C. The dataset included too many different LLMs

D. The vocabulary-based method didn't generalize well to other AI models

3. The researchers used advanced deep learning models to analyze essay text

True

False

4. What programming tool did the researchers use to implement their models?

A. TensorFlow

B. Scikit-learn

C. PyTorch

D. Keras

Questions?