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

Veronica Juliana Schmalz **Anaïs Tack**

KU Leuven
[1] Faculty of Arts, Research Unit Linguistics
[2] imec research group itec
veronicajuliana.schmalz@kuleuven.be anais.tack@kuleuven.be

Abstract

Despite recent advances in AI detection methods, their practical application, especially in education, remains limited. Educators need functional tools pointing to *AI indicators within texts*, rather than merely estimating *whether* AI was used. GPTZero's new AI Vocabulary feature, which highlights parts of a text likely to be AI-generated based on frequent words and phrases from LLM-generated texts, offers a potential solution. However, its effectiveness has not yet been empirically validated.

In this study, we examine whether GPTZero's AI Vocabulary can effectively distinguish between LLM-generated and student-written essays. We analyze the AI Vocabulary lists published from October 2024 to March 2025 and evaluate them on a subset of the Ghostbuster dataset, which includes student and LLM essays. We train multiple Bag-of-Words classifiers using GPTZero's AI Vocabulary terms as features and examine their individual contributions to classification.

Our findings show that simply checking for the presence, not the frequency, of specific AI terms yields the best results, particularly with ChatGPT-generated essays. However, performance drops to near-random when applied to Claude-generated essays, indicating that GPTZero's AI Vocabulary may not generalize well to texts generated by LLMs other than ChatGPT. Additionally, all classifiers based on GPTZero's AI Vocabulary significantly underperform compared to Bag-of-Words classifiers trained directly on the full dataset vocabulary. These findings suggest that fixed vocabularies based solely on lexical features, despite their interpretability, have limited effectiveness across different LLMs and educational writing contexts.

1 Introduction

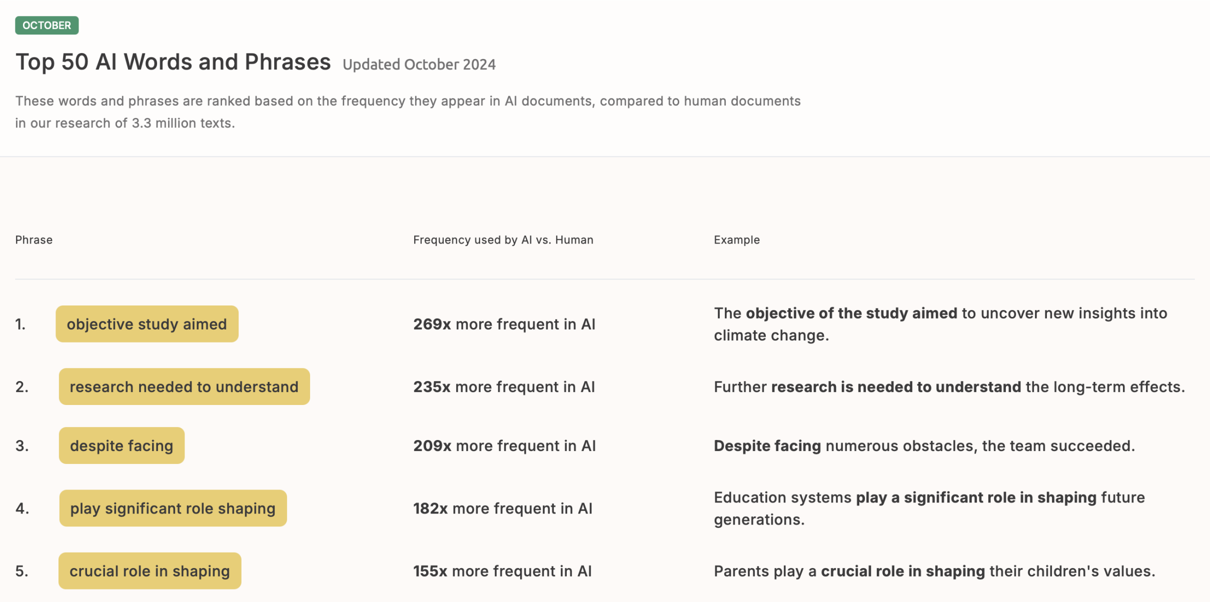
Recently, the introduction of user-friendly interfaces such as ChatGPT (OpenAI, 2023) has made a significant impact on education. An increasing number of students are using large language models (LLMs) to write essays (among other things), and this creates new challenges for educators to assess various skills and ensure academic integrity (Cotton et al., 2024). Even experienced teachers and those familiar with LLMs often struggle to tell apart student-written essays from those created by LLMs, as studies have shown (Fleckenstein et al., 2024; Waltzer et al., 2024; Perkins et al., 2024).

To address these challenges, numerous AI detection methods and tools have been developed (see Wu et al. 2025 for a review). However, as highlighted by Weber-Wulff et al. (2023), most detection tools in the market lack robustness with student texts and interpretability for non-expert users such as teachers. GPTZero (Tian and Cui, 2025), a popular AI detection tool, aims to offer a more transparent and interpretable solution. It analyzes texts for patterns, vocabulary and styles that are more common in AI-generated writing than in human writing, aiming to assist educators in verifying the authenticity of student work.

In October 2024, GPTZero introduced a new **AI Vocabulary**[1] feature (Figure 1), which highlights text parts that are likely to be AI-generated. This feature includes a list of the 50 words and phrases most commonly used by LLMs (Constantino, 2024), which can be interpreted as AI indicators, and is updated monthly. Each term is assigned a weight, indicating its frequency in AI-generated texts relative to human-written ones, and is accompanied by a contextual example. Since December 2024, the list has been expanded to include the top 100 words and phrases commonly used by AI. A key question, however, is whether this feature can be used to effectively distinguish LLM-generated essays from student-written ones. In this paper, we address this question by conducting

[1] https://gptzero.me/ai-vocabulary

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 937–952
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics



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.

Figure 1: Screenshot of GPTZero's AI Vocabulary Released on October 7, 2024.

the first systematic study that assesses GPTZero's AI Vocabulary feature in detecting LLM-generated content from educational contexts. Specifically, we integrate the AI Vocabulary lists (from October 2024 to March 2025) within supervised Bag-of-Words (BoW) classification models, namely two Naive Bayes classifiers trained on a subset of the Ghostbuster detector dataset (Verma et al., 2024), containing student and LLM-generated essays from ChatGPT (OpenAI, 2023) and Claude (Anthropic, 2023). We selected these models for their interpretability, as they allow us to inspect the contribution of each feature, namely the AI Vocabulary terms, to classification decisions. We position our work as a step toward evaluating the real-world utility of interpretable detection tools in educational contexts, where the use of AI is becoming increasingly widespread and, therefore, both reliable and efficient solutions are needed.

2 Background

Being able to differentiate between human-written and LLM-generated texts has recently become a much-discussed research topic, especially in academic and educational contexts. However, current AI detection methods present two main issues: (i) they often rely on non-transparent features, abstract and difficult for the average person to interpret and (ii) they have limited applicability for texts written by students, who are underrepresented in the training data.

Several current AI detection methods and systems prioritize model-based statistical metrics over basic linguistic features, such as perplexity (Vasilatos et al., 2023) and burstiness (Tian and Cui, 2025), log-probability (Solaiman et al., 2021) and high-dimensional neural representations (Guo et al., 2024). While highly performative, these methods do not offer interpretable justifications for their predictions, making it difficult for educators to reliably use and understand their outcomes (Ji et al., 2024). The underrepresentation of student texts in the detectors' training sets represents another significant challenge. Student writing can exhibit lower fluency, formulaic phrasing or genre-specific traits that differ from both typical human and LLM-generated outputs. This mismatch can lead to high false positive rates, as observed in recent evaluations (Weber-Wulff et al., 2023; Liang et al., 2023; Perkins et al., 2024), as well as numerous false negatives, particularly when texts undergo simple adversarial modifications to evade the detectors (Weber-Wulff et al., 2023; Perkins et al., 2024).

In response to these AI detectors' transparency issues, interpretable alternatives focusing on word frequency, n-gram patterns and stylometric indicators (Opara, 2024; Ciccarelli et al., 2024; Muñoz-Ortiz et al., 2024) have emerged to offer more transparent and pedagogically useful solutions. However, these methods are often less robust when applied to domain shifts or with LLM-generated texts modified to become less detectable. A hybrid detection tool, GPTZero (Tian and Cui, 2025), combines statistical features, such as perplexity and burstiness, with more interpretable metrics, including readability, text complexity and average sentence length. Although it does not fully disclose the rationale behind its classifications, GPTZero claims to be "the top AI detector for teachers" (Tian and Cui, 2025). As such, it has recently introduced AI Vocabulary lists that highlight in text terms disproportionately used by LLMs compared to human authors, as a way to enhance interpretability and better support educational use among teachers.

Beyond AI detection models, some studies have recently emerged that focus on quantifying and analyzing a significant increase in the use of certain words and phrases, especially in scientific writing, after the introduction of LLMs. Kobak et al. (2024) employed large-scale corpus analysis of medical abstracts to track excess word usage and revealed a sharp rise in usually less frequent terms such as "delve" and "intricacies". Juzek and Ward (2025) used model testing methods and human evaluators to explore why LLMs overrepresent certain terms, focusing on 21 "focal words". However, their results turned out to be inconclusive. Mingmeng and Roberto (2024) quantitatively compared academic texts before and after the spread of LLMs, documenting a general trend towards producing more complex and abstract texts. Liang et al. (2024) analyzed textual features and metadata from papers across different domains, linking higher rates of LLMs use with texts whose first authors published more preprints, shorter papers or in more popular research fields.

While these studies provide interesting insights into possible LLM-influenced term choices, they mostly remain focused on quantitative and comparative vocabulary studies that analyze the language of scientific publications and do not directly extend to student writing or educational contexts. Moreover, to the best of our knowledge, no existing works have systematically leveraged terms more frequently used by LLMs to build and evaluate AI detection models focused on the educational domain, where they would currently be highly needed. This highlights a critical research gap, in which to explore whether vocabulary-based AI detection methods could be applied to distinguish student-written texts from LLM-generated ones, supporting educators in a more interpretable and linguistically justified manner. To address this gap, in this paper we evaluate for the first time GPTZero's AI Vocabulary lists as a promising way to detect LLM-generated essays among student-written ones. To the best of our knowledge, these are the only publicly available, extensive AI-vocabulary lists derived from a significant number of documents that go beyond scientific publications and likely include student-written texts, given GPTZero's commitment to teachers and educational contexts [2]. Moreover, they also provide data concerning the different word and phrase frequencies found in human-authored and LLM-generated texts (see Figure 1), further increasing their relevance. Starting with this study we aim to work towards developing a more transparent AI detection methodology applicable in educational contexts, reliable and better aligned with educators' needs.

3 Method

3.1 GPTZero's AI Vocabulary Lists

We collected the AI Vocabulary lists published on the GPTZero website between October 2024 and March 2025.[3] Each month, we gathered a list of words and phrases together with their frequency estimates (see Appendix A), which had been estimated on 3.3 million texts (Tian and Cui, 2025). The October 2024 list featured the 50 most frequent AI-related terms, including single words and multi-word expressions. In November 2024, the same list ("Updated October 2024") remained online. In December 2024, a new list ("Updated November 2024"), now including 99 items, was published.[4] However, this updated list contained some errors, such as missing words, duplicate entries, and phrase variations. Subsequently, a corrected list with 100 items was published later on in December, 2024. This one still contained duplicates, so we removed exact double entries for the purpose of our experiments. In January 2025, a new list ("Updated January 2025") with 100 unique phrases was published. No new list was published in February 2025; instead, the January 2025 list remained online for that month. The March 2025 list was labeled as "updated", but it was identical to the January and February lists. As a result, there were only three distinct AI Vocabulary lists that we could use in our experiments: (a) the October 2024 list, (b) the November/December 2024 list, and (c) the January/February/March 2025 list. In addition, we constructed a *combined* list ("All")

[2] https://gptzero.me/educators

[3] The reader can retrieve these lists using https://web.archive.org/.

[4] https://web.archive.org/web/20241208223132/https://gptzero.me/ai-vocabulary

that merged all unique words and phrases from these three sources.

3.2 Data

To detect LLM-generated essays using GPTZero's AI Vocabulary, we used a subset of the data initially employed to train the Ghostbuster detector (Verma et al., 2024). This dataset originally contained 21,000 documents, including articles, creative writing pieces and student essays. For our experiments, we focus on a subset of 1,000 university student essays sourced from IvyPanda (IvyPanda, 2025), a platform where users can submit essays from high school and university levels concerning various topics and subjects, and 2,000 LLM-generated essays. To obtain the latter, Verma et al. (2024) used ChatGPT to generate the prompts corresponding to the unique 1,000 assignments based on which the student texts were written. These prompts were then used to generate 1,000 essays with ChatGPT and 1,000 essays with Claude. The desired essay length was also specified in them to match the human-written texts. The resulting median word count was 661 for student essays, 536 for ChatGPT-generated essays and 456 for Claude-generated essays. For the rest of the paper, we will refer to this subset of essays as the *Ghostbuster corpus*.

3.3 Models

We experimented with three classification models, each trained to predict whether an essay is written by a student or by (a) an LLM, (b) Claude, or (c) ChatGPT. To this end, we performed binary classification (using only the binary labels "AI" and "human") on different dataset partitions: 1,000 student essays with (a) all 2,000 LLM-generated essays, (b) 1,000 essays generated by Claude, or (c) 1,000 essays generated by ChatGPT. We used *scikit-learn* (Pedregosa et al., 2011) to implement and train these classification models.

For each of these three classifiers, we estimated separate models for each of the four AI Vocabulary lists (monthly or combined), computing different feature vectors for the words and phrases in the list. For each list, we counted the occurrences of its items in the corpus using a Bag-of-Words (BoW) approach. Each AI word or phrase was treated as a distinct feature. Since the vocabulary included multi-word units (i.e., AI phrases), we employed an n-gram vectorization strategy to capture these phrases, setting n to range from 1 token up to the maximum number of tokens found in the longest phrase in the list.[5]

We integrated the BoW features in a binary Naive Bayes classifier

$$P(c|w_1, ..., w_n) \propto P(c) \prod_{i=1}^{n} P(w_i|c) \qquad (1)$$

to predict the class c, namely whether an essay is generated by an AI (positive class) or written by a student (negative class). The models assumed independence between each word/phrase w and always used a uniform prior, assuming an equal chance (50%) that an essay belonged to one of the two[6] classes.

We experimented with two types of feature vectors: (a) a Multinomial feature vector indicating the counts of the words and phrases in the essay, and (b) a Bernoulli feature vector indicating the presence or absence of the words and phrases in the essay.

For comparison, we also trained binary Naive Bayes classifiers – using either Multinomial or Bernoulli feature vectors – based on an alternative Bag-of-Words approach. In this configuration, the vocabulary comprised all unigram word types found in the Ghostbuster dataset, which were used to construct the feature vectors. These models served as a baseline to assess the effect of using the curated vocabulary lists in contrast to the default vocabulary derived directly from the training data.

3.4 Experiments

We trained a total of 24 classification models (3 AI x 4 lists x 2 features) using GPTZero AI Vocabulary lists, along with 6 reference models (3 AI x 2 features) based on the vocabulary derived from the Ghostbuster training data. To ensure an exhaustive evaluation, all models were trained and tested using leave-one-out cross-validation.

3.5 Metrics

We evaluated the classifiers' performance using accuracy, (binary) precision, (binary) recall, (binary) F1-score, MCC (Matthews correlation coefficient) and AUROC (Area Under the Receiver Operating Characteristic curve) computed with *scikit-learn*

[5] This was implemented using `CountVectorizer`, with the `ngram_range` parameter set to `(1, max_phrase_length)`.

[6] It is important to reiterate that we did not perform any multiclass classification between the different LLMs in the dataset. We always compared LLM-generated to student-written, or Claude-generated to student-written, or GPT-generated to student-written (cf., *supra*).

(Pedregosa et al., 2011). Precision, recall and F1 score were computed for the positive class only (LLM-generated essays).

4 Results

4.1 GPTZero's AI Vocabulary terms' distribution in Ghostbuster

Table 1 lists the terms from GPTZero's AI Vocabulary found in the Ghostbusters dataset. Of the 245 distinct words and phrases published between October 2024 and March 2025, only 98 appeared in the entire dataset, 53 in the Claude subset and 91 in the ChatGPT subset. These low and different distributions suggest that many AI-specific vocabulary terms identified by GPTZero as salient AI indicators, such as "left an indelible mark" (ranked 8th in Table 7 but only found 9 times in our corpus), "a rich tapestry" (ranked 18th in Table 7 but only found 6 times in our corpus), "offers valuable insights" (ranked 9th in Table 7 but only found 7 times in our corpus), "despite facing " (ranked 3rd in Table 5 but only found 6 times in our corpus) and "study aims to explore" (ranked 6th in Table 5 but only found twice in our corpus) may not be frequently used in educational LLM-generated essays, in particular by models other than ChatGPT for most of the cases.

To assess the alignment between GPTZero's AI Vocabulary rankings and their usage in LLM-generated essays, we calculated Spearman rank correlations between each term's rank in the AI Vocabulary lists and its rank based on frequency of usage in LLM-generated texts (Claude and/or GPT). Our results (see Table 9) indicate generally weak or negative correlations between AI Vocabulary rankings and their occurrence across LLM-generated texts. There was, however, a significant positive correlation between the terms' ranks in the October lists and their usage in Claude-generated texts ($\rho = 0.501$, $p < .001$), as well as between the terms' ranks in the January-March lists and their usage in Claude-generated texts ($\rho = 0.476$, $p < .001$). In contrast, correlations with ChatGPT-generated texts remained low or negative, except for a modest positive correlation ($\rho = 0.211$, $p = .053$) with the November/December AI Vocabulary list [7]. Based on these findings, it is still unclear whether the actual ranks of the AI Vocabulary words and phrases in the list are informative and could consequently be used for AI text detection in education.

4.2 Classification performance with GPTZero's AI Vocabulary lists

In our experiments, we evaluated two types of Naive Bayes classifiers, one using a Bernoulli feature vector and one using a Multinomial feature vector, based on GPTZero's AI Vocabulary lists (from October 2024 to March 2025). We tested the classifiers in detecting AI-generated essays both individually, with each monthly AI Vocabulary list, and with a combined list containing 245 AI Vocabulary terms from all months. We evaluated both the full Ghostbuster essays corpus and subsets specific to Claude- and ChatGPT-generated texts.

Overall, classification results were close to random, with accuracy ranging from 0.363 to 0.755 for Bernoulli models and 0.363 to 0.729 for Multinomial models (see Table 2) using the different AI Vocabulary lists. However, we found more promising results when focusing specifically on ChatGPT-generated texts using the combined GPTZero's AI Vocabulary lists of all months. Here, the Bernoulli model achieved the highest accuracy (0.755), high precision (0.882), moderate recall (0.588) and an F1 score of 0.705, which indicates good performance in identifying LLM-generated texts, although it might have missed some positive cases. The high precision score signals that the model does not make numerous false predictions causing it to mislabel student texts as AI-generated (a significant risk in educational contexts as highlighted by Liang et al. 2023). However, the moderate recall also indicates that the model's sensitivity should increase in order to avoid some LLM-generated texts to go undetected (also particularly relevant in educational contexts as stressed by Fleckenstein et al. 2024; Weber-Wulff et al. 2023; Perkins et al. 2024). An MCC score of 0.541 supports our interpretation and an AUROC of 0.595 suggests that the model, despite being better than random, may struggle in more ambiguous cases. The Multinomial model, using the same feature set, yielded higher precision (0.884) and AUROC (0.705), indicating higher sensitivity to ChatGPT-generated content and a more balanced classification ability. This may be due to an overrepresentation of ChatGPT-generated texts in the datasets used by GPTZero to compile the AI Vocabulary lists. However, this model also reached

[7] These different correlation values suggest that while higher-ranked AI Vocabulary words tend to be relatively more frequent in Claude-generated essays compared to lower-ranked terms, their overall presence in such texts remains sparse.

	All	C	G		All	C	G
add an extra layer	1	0	1	meticulous attention to	4	0	4
add depth to	1	0	1	meticulously crafted	2	1	1
address issues like	1	1	0	navigate challenges	2	2	0
advancements	204	42	189	navigate the complex	6	0	6
aiding	19	3	18	offer valuable insights	8	0	8
aim to explore	1	1	0	offers numerous benefits	3	0	3
aims to enhance	3	1	2	offers valuable	12	1	12
aligns	78	25	66	offers valuable insights	7	0	7
an unwavering commitment	1	0	1	potentially leading	15	0	15
commitment to excellence	2	0	2	prioritize	247	39	227
consider factors like	1	1	0	prioritizing	71	17	62
continue to inspire	4	0	4	provide an insight	1	1	1
contribute to understanding	1	0	1	provide valuable insights	20	5	17
crucial role in shaping	34	1	33	provided valuable	6	3	3
crucial role in understanding	1	0	1	provided valuable insights	4	2	2
delve deeper	4	1	4	provides valuable	36	10	28
delve deeper into	4	1	4	provides valuable insights	24	5	20
despite facing	6	4	2	providing insights	2	1	2
emphasize the need	7	2	7	relentless pursuit	4	0	4
enduring legacy	3	0	3	remarked	3	3	2
ensure long term success	2	0	2	researchers aim	1	1	0
essential to recognize	19	0	19	researchers aimed	3	0	3
explores themes	3	0	3	rich tapestry	6	0	6
findings shed	1	0	1	sense of camaraderie	8	0	8
findings shed light	1	0	1	showcasing	52	8	49
fostering	249	23	236	significant advancements	8	1	8
fostering sense	23	0	23	sparking	5	1	4
gain comprehensive understanding	10	1	9	standout	7	1	7
gain deeper	32	5	27	stark reminder	9	1	8
gain deeper insights	1	0	1	struggles faced	15	0	15
gain deeper understanding	28	5	23	study aims to explore	2	2	0
gain valuable	23	6	17	study highlights the importance	1	1	0
gain valuable insights	17	1	16	study provides valuable	2	0	2
garnered significant	1	0	1	study sheds	1	0	1
highlight the need	3	1	2	study sheds light	1	0	1
highlight the potential	1	0	1	surpassing	9	6	7
highlight the significance	7	0	7	the complex interplay	8	7	1
highlighting the need	3	2	1	the multifaceted nature	12	1	11
hindering	47	7	47	the potential to revolutionize	10	0	10
holds significant	10	1	9	the relentless pursuit	1	0	1
impacting	62	31	45	the transformative power	9	2	7
indelible mark	11	0	11	tragically	6	4	4
indicating potential	1	0	1	underscore the importance	1	0	1
intricate relationship	3	0	3	understand the behavior	1	0	1
left an indelible mark	9	0	9	understand the complexity	2	0	2
left lasting	11	4	7	unwavering commitment	2	0	2
let delve	7	0	7	unwavering support	1	1	1
making it challenging	14	2	14	valuable insights	115	21	99
marked significant	4	0	4	vital role in shaping	9	0	9

Table 1: GPTZero's AI words/phrases with their counts in Ghostbusters (All), Claude (C), and GPT (G) subsets.

Features	LLM	Vocabulary	Accuracy	Precision	Recall	F1	MCC	AUROC
Bernoulli	All	GPTZero List: All	0.532	0.884	0.343	0.494	0.272	0.362
		GPTZero List: Oct	0.503	0.877	0.296	0.443	0.240	0.292
		GPTZero List: Nov/Dec	0.416	0.996	0.129	0.228	0.199	0.135
		GPTZero List: Jan/Feb/Mar	0.363	0.969	0.046	0.089	0.117	0.050
		Ghostbuster BoW	0.871	0.846	0.986	0.911	0.711	0.948
	Claude	GPTZero List: All	0.522	0.657	0.09	0.158	0.085	0.156
		GPTZero List: Oct	0.502	0.501	0.968	0.660	0.011	0.107
		GPTZero List: Nov/Dec	0.508	0.786	0.022	0.043	0.068	0.033
		GPTZero List: Jan/Feb/Mar	0.503	1.0	0.007	0.014	0.059	0.017
		Ghostbuster BoW	0.889	0.825	0.987	0.899	0.793	0.975
	GPT	GPTZero List: All	**0.755**	0.882	0.588	0.705	0.541	0.595
		GPTZero List: Oct	0.703	0.853	0.49	0.622	0.448	0.495
		GPTZero List: Nov/Dec	0.616	0.964	0.242	0.386	0.351	0.250
		GPTZero List: Jan/Feb/Mar	0.544	0.968	0.092	0.167	0.209	0.102
		Ghostbuster BoW	0.929	0.892	0.977	0.933	0.862	0.990
Multinomial	All	GPTZero List: All	0.517	0.891	0.314	0.464	0.263	0.604
		GPTZero List: Oct	0.452	0.910	0.197	0.324	0.212	0.550
		GPTZero List: Nov/Dec	0.410	0.968	0.119	0.213	0.191	0.549
		GPTZero List: Jan/Feb/Mar	0.363	0.969	0.046	0.089	0.117	0.518
		Ghostbuster BoW	0.901	0.955	0.893	0.923	0.787	0.957
	Claude	GPTZero List: All	0.518	0.673	0.072	0.130	0.082	0.517
		GPTZero List: Oct	0.504	0.538	0.064	0.114	0.019	0.506
		GPTZero List: Nov/Dec	0.508	0.786	0.022	0.043	0.068	0.509
		GPTZero List: Jan/Feb/Mar	0.503	1.0	0.007	0.014	0.059	0.503
		Ghostbuster BoW	**0.964**	0.976	0.951	0.964	0.928	0.991
	GPT	GPTZero List: All	**0.729**	0.884	0.527	0.660	0.501	0.705
		GPTZero List: Oct	0.654	0.895	0.350	0.503	0.390	0.597
		GPTZero List: Nov/Dec	0.604	0.964	0.216	0.353	0.330	0.589
		GPTZero List: Jan/Feb/Mar	0.539	0.964	0.081	0.149	0.194	0.530
		Ghostbuster BoW	0.912	0.942	0.877	0.909	0.825	0.953

Table 2: Performance of classifiers on leave-one-out cross-validation. The highest accuracy values are indicated in boldface.

lower accuracy (0.729) and F1 score (0.660), making it less reliable.

These results suggest that binary-feature BoW models like Bernoulli may be more effective at detecting ChatGPT-generated texts based solely on AI-related terms' presence, while frequency-based models like Multinomial may be better at identifying subtler vocabulary usage patterns. Finally, both Naive Bayes classifiers were significantly outperformed by a baseline Multinomial BoW classifier trained on the full vocabulary of the Ghostbuster dataset. This model achieved a maximum accuracy of 0.964 and an AUROC score of 0.991 (see Table 2) with Claude texts - differing from the previous highest results for ChatGPT-generated essays using the AI Vocabulary lists, possibly more effective given the absence or scarcity of Claude's generated data for the compilation of the AI Vocabulary lists [8]. This highlights the limitations of relying on fixed AI Vocabulary lists for AI detection, which might not reflect the language found in educational essays written by different LLMs and students.

4.3 AI Vocabulary terms contribution to classification

To better understand which specific AI Vocabulary terms influenced classification, we analyzed their log probabilities under our best-performing Naive Bayes models, namely the Bernoulli and Multinomial variants that achieved the highest classification results. These models were trained using the subset of ChatGPT-generated texts from the Ghostbuster corpus and the full combined AI Vocabulary list (with 245 terms from October 2024 to March

[8] See GPTZero's support article https://support.gptzero.me/hc/en-us/articles/15129377479959 for more

2025).

Phrase	Rank	Freq.	Count	OR	LP
fostering	245	9	236	11.88	-2.15
prioritize	238	11	227	8.54	-2.22
advancements	243	9	189	3.91	-2.65
valuable insights	19	230	99	5.53	-2.93
prioritizing	241	9	62	2.71	-3.43
aligns	234	17	66	2.19	-3.48
showcasing	232	21	49	4.31	-3.62
hindering	242	9	47	2.64	-3.74
crucial role in shaping	37	155	33	10.53	-3.83
impacting	237	12	45	2.07	-3.89
gain deeper	86	98	27	5.39	-3.97
provides valuable	117	86	28	2.28	-4.00
gain deeper understanding	50	131	23	2.82	-4.13
provides valuable insights	4	464	20	1.33	-4.27
essential to recognize	213	48	19	6.56	-4.27
fostering sense	45	138	23	2.11	-4.27
gain valuable	100	92	17	1.99	-4.43
gain valuable insights	175	59	16	1.64	-4.49
potentially leading	231	43	15	5.81	-4.55
aiding	244	9	18	3.28	-4.55
provide valuable insights	7	332	17	1.32	-4.62
struggles faced	224	46	15	5.93	-4.70
making it challenging	142	74	14	3.00	-4.70
indelible mark	11	275	11	2.33	-4.78
the potential to revolutionize	123	83	10	4.34	-4.86
offers valuable	128	81	12	1.23	-4.96

Table 3: Top 25 phrases contributing to LLM-generated text detection, ordered by log-probability from the best Bernoulli Naive Bayes classifier using all AI Vocabulary lists on ChatGPT-generated essays. The *Rank* and *Frequency* columns relate to the combined GPTZero's AI Vocabulary lists (245 phrases from October 2024 to March 2025), *Count* refers to the frequency in ChatGPT-generated texts, *OR* refers to the odds ratio in LLM-generated vs. human-authored texts and *LP* represents log probability of the phrase contribution to classification.

Phrase	Rank	Freq.	Count	OR	LP
fostering	245	9	236	7.49	-2.00
prioritize	238	11	227	5.49	-2.08
advancements	243	9	189	2.44	-2.33
valuable insights	19	230	99	4.76	-2.87
prioritizing	241	9	62	2.75	-3.42
aligns	234	17	66	2.05	-3.44
showcasing	232	21	49	4.14	-3.62
hindering	242	9	47	2.59	-3.71
crucial role in shaping	37	155	33	9.57	-3.90
impacting	237	12	45	2.13	-3.96
gain deeper	86	98	27	5.33	-4.09
provides valuable	117	86	28	2.36	-4.13
fostering sense	45	138	23	1.95	-4.25
gain deeper understanding	50	131	23	2.81	-4.25
essential to recognize	213	48	19	6.38	-4.43
provides valuable insights	4	464	20	1.31	-4.43
gain valuable	100	92	17	1.92	-4.54
aiding	244	9	18	3.15	-4.59
gain valuable insights	175	59	16	1.68	-4.59
provide valuable insights	7	332	17	1.22	-4.65
potentially leading	231	43	15	5.58	-4.65
struggles faced	224	46	15	5.41	-4.65
making it challenging	142	74	14	3.02	-4.86
offers valuable	128	81	12	1.28	-4.94
indelible mark	11	275	11	2.27	-4.94
the multifaceted nature	98	92	11	3.12	-4.94

Table 4: Top 25 phrases contributing to LLM-generated text detection, ordered by log-probability from the best Multinomial Naive Bayes classifier using all AI Vocabulary lists on ChatGPT-generated essays. The *Rank* and *Frequency* columns relate to the combined GPTZero's AI Vocabulary lists (245 phrases from October 2024 to March 2025), *Count* refers to the frequency in ChatGPT-generated texts, *OR* refers to the odds ratio in LLM-generated vs. human-authored texts and *LP* represents log probability of the phrase contribution to classification.

For both models, Bernoulli and Multinomial, the top 25 terms that contributed the most to classification were largely the same, although in slightly different order (see Table 3 and Table 4). Each table includes the terms' original *Rank* and *Frequency* in the combined AI Vocabulary list, their *Count* in ChatGPT-generated texts, the *OR* (indicating their relative likelihood in LLM vs. human text based on odds ratios) and the models' log probabilities, *LP* (reflecting the terms' contribution to the model decision; lower values imply weaker impact).

We noticed in the Bernoulli and Multinomial classifiers that several words and phrases found in numerous ChatGPT-generated texts, such as "fostering" (counted 236 times), "prioritize" (227), "advancements" (62), "aligns" (66) and "showcasing" (49), despite being found more frequently in Ghostbuster's texts than in GPTZero's ranking lists, were less effective in distinguishing AI-generated from student-authored essays given their low log probabilities. Similarly, when considering terms that were highly ranked and common in AI Vocabulary lists, such as "provides valuable" (4th), "provides valuable insights" (7th), "indelible mark"(11th) and "valuable insights" (19th), we observed also low log probabilities, apart from recurring phrases, meaning that they did not significantly contribute to classification.

We decided to maintain separate entries for morphological variants rather than indexing them together to investigate whether certain preferences exist in LLM-generated texts. In this way, we could check if verb tense, number, or grammatical person can also influence AI-generated text detection. By maintaining distinctions such as "provide" vs. "provides" (valuable insights) and "study shed" vs. "study sheds" we can evaluate whether specific variants display distributional biases in LLM-generated texts compared to student-authored essays. However, if these differences prove insignificant, as seems to be the case in our experiments, in future works we could consider lemmatization or stemming.

Our findings are in line with our previous observations, provided in Section 4.1, where term rankings and frequencies did not seem to notably support classification. They, nevertheless, confirm our classification results described in Section 4.2, highlighting the strengths of the BoW Bernoulli model over the Multinomial one, accounting for the terms' presence only, rather than for their frequency, to better distinguish LLM-generated texts from student-written ones.

5 Conclusion

In this study, we presented the first empirical evaluation of GPTZero's AI Vocabulary lists as a way to detect AI-generated texts in educational settings. Our findings show that these precompiled vocabulary lists, despite being transparent and easily interpretable for educators, have limited effectiveness in detecting LLM-generated texts among educational essays, especially beyond ChatGPT. Even for ChatGPT-generated texts, the classification performance of our Naive Bayes models based on AI Vocabulary lists was modest and only improved when using a combined list of 245 terms. We achieved better results with BoW models that used the full Ghostbuster dataset vocabulary, suggesting that broader language patterns may be more effective for AI detection with different LLMs.

Future research should focus on a deeper, more domain-specific analysis and comparison between student and LLM-generated texts in educational domains, including more diverse student samples and LLM-generated texts. Vocabulary-based AI detectors could benefit from the inclusion of additional functional and structural features, considering each term and phrase as linguistic constructions that reflect users' language more in detail.

Overall, although our results might not come close to state-of-the-art detectors, with this work we addressed a key research gap. To the best of our knowledge, no prior study has evaluated precompiled AI Vocabulary lists, publicly available and derived from a diverse set of texts beyond scientific articles, for AI detection in education. Our findings offer practical and detailed insights into the utility and accuracy of *transparent* linguistic features, such as AI Vocabulary lists, that can support educators in distinguishing LLM-generated and student-written texts. By doing so, this work contributes to the ongoing efforts to improve AI detection systems and lays a foundation for further investigation and refinement in educational contexts.

Limitations

Although our work provides useful evidence in the analysis of GPTZero's AI Vocabulary lists for AI detection, there are several limitations that need to be accounted for. First, we only tested two Bag-of-Words classifiers (Bernoulli and Multinomial) using a Naive Bayes approach. These are relatively simple models. More advanced machine learning and neural approaches could help to expand the testing framework and potentially improve detection accuracy. Second, the dataset used in this study, the Ghostbuster essay subcorpus, represents outputs from older versions of ChatGPT and Claude models. As new model versions are released, the vocabulary patterns of current AI systems may differ significantly. Moreover, due to the lack of metadata, we assume students to be native English speakers. Future studies should examine L2 students, who may rely more on LLMs and for whom current detectors might be less effective. Third, as LLMs continue to evolve, their outputs become closer to human language, making fixed vocabulary lists less effective over time. To remain useful, these AI Vocabulary lists would need to be updated more frequently and adapted across different writing domains, to reflect changes in language use.

Acknowledgments

The authors would like to thank Piet Desmet for for his insightful comments and feedback. The research presented in this paper was funded by the Research Foundation Flanders (FWO) through a doctoral fellowship awarded to Veronica Juliana Schmalz (1108723N).

References

Anthropic. 2023. Claude: An AI assistant.

Vittorio Ciccarelli, Cornelia Genz, Nele Mastracchio, Wiebke Petersen, Anna Stein, and Hanxin Xia. 2024. Team art-nat-HHU at SemEval-2024 Task 8: Stylistically Informed Fusion Model for MGT-Detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1690–1697, Mexico City, Mexico. Association for Computational Linguistics.

Tor Constantino. 2024. New List Ranks AI's 50 Most Overused Words - Updates Monthly. Accessed on March 22, 2025.

Deborah R. E. Cotton, Pauline A. Cotton, and James R. Shipway. 2024. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2):228–239.

Johanna Fleckenstein, Jennifer Meyer, Thorben Jansen, Stefan D. Keller, Olaf Köller, and Jens Möller. 2024. Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence*, 6:100209.

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. Detective: Detecting AI-Generated Text via Multi-Level Contrastive Learning. *Advances in Neural Information Processing Systems*, 37:88320–88347.

IvyPanda. 2025. Free Essay Examples and Writing Resources. `https://ivypanda.com/`. Accessed: March 10, 2025.

Jiazhou Ji, Ruizhe Li, Shujun Li, Jie Guo, Weidong Qiu, Zheng Huang, Chiyu Chen, Xiaoyu Jiang, and Xinru Lu. 2024. Detecting Machine-Generated Texts: Not Just "AI vs Humans" and Explainability is Complicated. *arXiv preprint*, 2406:18259.

Tom S. Juzek and Zina B. Ward. 2025. Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6397–6411, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dmitry Kobak, Rita González-Márquez, Emőke Ágnes Horvát, and Jan Lause. 2024. Delving into ChatGPT Usage in Academic Writing through Excess Vocabulary. *arXiv*, 2406(07016).

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the Increasing Use of LLMs in Scientific Papers. In *Proceedings of the COLM 2024 Conference*.

Geng Mingmeng and Trotta Roberto. 2024. Is ChatGPT Transforming Academics' Writing Style? *arXiv preprint*, 2404(08627).

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, 57(10):265.

Chidimma Opara. 2024. StyloAI: Distinguishing AI-generated Content with Stylometric Analysis. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pages 105–114. Springer.

OpenAI. 2023. Chatgpt (gpt-4).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mike Perkins, Jasper Roe, Darius Postma, James McGaughran, and Don Hickerson. 2024. Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics*, 22(1):89–113.

Basel Solaiman, Didier Guériot, Shaban Almouahed, Bassem Alsahwa, and Éloi Bossé. 2021. A New Hybrid Possibilistic-Probabilistic Decision-Making Scheme for Classification. *Entropy*, 23(1):67.

Edward Tian and Alexander Cui. 2025. GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods.

Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. 2023. HowkGPT: Investigating the Detection of ChatGPT-Generated University Student Homework through Context-Aware Perplexity Analysis. *arXiv preprint*, 2305:18226.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. "ghostbuster: Detecting text ghostwritten by large language models". In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717. Association for Computational Linguistics.

Talia Waltzer, Cecile Pilegard, and Gail D. Heyman. 2024. Can you spot the bot? Identifying AI-generated writing in college essays. *International Journal for Educational Integrity*, 20(1):11.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):1–39.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia S. Chao, and Derek F. Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, pages 1–66.

A GPTZero's AI Vocabulary Lists

This appendix contains lists of the top AI words and phrases from GPTZero, spanning from October 2024 to March 2025. Each monthly list includes frequently used AI-related terms, along with their frequency estimates. The November list was initially identical to the October list with 50 entries, but an update appeared in December with 99 entries. However, this updated version contained errors such as missing determiners and prepositions (e.g., *crucial role understanding* instead of *a crucial role in understanding*) and incongruencies, including duplicate entries (e.g., 4) *provide valuable insights - 464* and 84) *provide valuable insights - 86*). This list also contains numerous variations of the same phrase (e.g., *13) plays a crucial role in understanding - 247* and *14) play a crucial role in understanding- 242*) and longer phrases that are part of other shorter phrases, also appearing in the list (e.g., *24) plays a crucial role in shaping - 178* and *26) crucial role in shaping - 171*). The January, February and March lists were identical, so we report them in the same table. The frequency estimates indicate how many times more frequently a term appears in AI-written texts compared to human-written texts. For example, a term with a frequency estimate of 10 means it is ten times more common in AI texts than in human texts, based on a collection of 3.3 million documents (Tian and Cui, 2025).

	Phrase	**Freq.**
1	objective study aimed	269
2	research needed to understand	235
3	despite facing	209
4	play significant role shaping	182
5	crucial role in shaping	155
6	study aims to explore	144
7	notable works include	121
8	consider factors like	121
9	today's fast paced world	107
10	expressed excitement	93
11	highlights importance considering	89
12	emphasizing importance	74
13	making it challenging	74
14	aims to enhance	72
15	study sheds light	69
16	emphasizing need	68
17	today's digital age	68
18	explores themes	66
19	address issues like	65
20	highlighting the need	63
21	study introduce	60
22	notable figures	59
23	gain valuable insights	59
24	showing promising results	59
25	media plays a significant role	57
26	shared insights	56
27	ensure long term success	55
28	make a positive impact on the world	55
29	facing criticism	52
30	providing insights	49
31	emphasized importance	48
32	indicating potential	47
33	struggles faced	46
34	secured win	46
35	secure win	44
36	potentially leading	43
37	showcasing	21
38	remarked	18
39	aligns	17
40	surpassing	12
41	tragically	12
42	impacting	12
43	prioritize	11
44	sparking	11
45	standout	11
46	prioritizing	9
47	hindering	9
48	advancements	9
49	aiding	9
50	fostering	9

Table 5: GPTZero's Top AI Words and Phrases for October 2024

#	Phrase	Freq.
1	provided valuable insights	902
2	gain valuable insights	739
3	casting long shadows	561
4	provides valuable insights	464
5	gain comprehensive understanding	355
6	study provides valuable	340
7	provide valuable insights	332
8	left indelible mark	319
9	offers valuable insights	298
10	indelible mark	275
11	unwavering commitment	256
12	play crucial role shaping	250
13	plays crucial role understanding	247
14	played significant role shaping	239
15	left indelible	231
16	valuable insights	230
17	rich tapestry	227
18	offer valuable insights	207
19	opens new avenues	206
20	help feel sense	197
21	adds layer complexity	194
22	significant contributions field	188
23	plays crucial role shaping	178
24	research needed explore	171
25	crucial role shaping	171
26	intricate relationship	165
27	findings contribute	157
28	continue inspire	152
29	stark reminder	151
30	hung heavy	147
31	crucial role understanding	139
32	fostering sense	138
33	significant attention recent years	136
34	needed fully understand	133
35	pivotal role shaping	131
36	gain deeper understanding	131
37	study sheds light	130
38	continues inspire	129
39	implications various	129
40	highlights importance considering	124
41	let delve	123
42	holds significant	121
43	study sheds	120
44	garnered significant	120
45	advancing understanding	119
46	voice dripping sarcasm	119
47	conclusion study provides	117
48	findings shed light	116
49	commitment public service	116
51	provided valuable insights	113
52	mix fear	109
53	crucial role maintaining	106
54	serves reminder	106
55	voice dripping	106
56	gain deeper insights	104
57	insights potential	101
58	significant advancement	100
59	researchers aimed	100
60	significant advancements	98
61	gain deeper	98
62	began voice	98
63	findings shed	97
64	study provide valuable	96
65	plays crucial role regulating	96
66	left lasting	96
67	sense camaraderie	94
68	potential revolutionize	94
69	navigate challenges	94
70	voice surprisingly	92
71	gain valuable	92
72	understanding behavior	91
73	delve deeper	91
74	plays crucial role ensuring	91
75	relentless pursuit	90
76	significant role shaping	88
77	researchers aim	88
78	meticulously crafted	88
79	study shed light	87
80	dripping sarcasm	87
81	aims shed light	87
82	voice rising	87
83	provides valuable	86
84	play significant role shaping	85
85	renewed sense purpose	85
86	marked significant	85
87	enduring legacy	84
88	offers numerous benefits	84
89	commitment excellence	83
90	study shed	83
91	plays crucial role determining	83
92	significant attention recent	83
93	offers valuable	81
94	plays significant role shaping	79
95	play crucial role determining	78
96	despite chaos	78
97	paving way future	77
98	highlights significance	77
99	locals visitors alike	77

Table 6: GPTZero's Top AI Words and Phrases for November 2024 (*first version with repetitions and errors*)

#	Phrase	Freq.
1	provided valuable insights	902
2	gain valuable insights	739
3	casting long shadows	561
4	provides valuable insights	464
5	gain comprehensive understanding	355
6	study provides valuable	340
7	provide valuable insights	332
8	left an indelible mark	319
9	offers valuable insights	298
10	an indelible mark	275
11	an unwavering commitment	256
12	play a crucial role in shaping	250
13	plays a crucial role in understanding	247
14	play a crucial role in understanding	242
15	played a significant role in shaping	239
16	left an indelible	231
17	valuable insights	230
18	a rich tapestry	227
19	offer valuable insights	207
20	opens new avenues	206
21	help to feel a sense	197
22	adds a layer of complexity	194
23	significant contributions to the field	188
24	plays a crucial role in shaping	178
25	research needed to explore	171
26	crucial role in shaping	171
27	the intricate relationship	165
28	findings contribute to	157
29	continue to inspire	152
30	a stark reminder	151
31	hung heavy	147
32	crucial role in understanding	139
33	fostering sense	138
34	significant attention in recent years	136
35	needed to fully understand	133
36	pivotal role in shaping	131
37	gain a deeper understanding	131
38	study sheds light on	130
39	continues to inspire	129
40	implications of various	129
41	highlights the importance of considering	124
42	let us delve	123
43	holds a significant	121
44	study sheds light on	120
45	garnered significant	120
46	advancing the understanding	119
47	voice dripping with sarcasm	119
48	conclusion of the study provides	117
49	findings shed light on	116
50	commitment to public service	116
51	provided valuable	113
52	mix the fear	109
53	crucial role in maintaining	106
54	serves a reminder	106
55	voice is dripping	106
56	gain a deeper insights	104
57	insights into the potential	101
58	a significant advancement	100
59	the researchers aimed	100
60	significant advancements	98
61	gain a deeper	98
62	began to voice	98
63	findings shed light on	97
64	study provides valuable	96
65	plays a crucial role in regulating	96
66	left a lasting	96
67	sense of camaraderie	94
68	potential to revolutionize	94
69	navigate the challenges	94
70	the voice surprisingly	92
71	gain a valuable	92
72	understanding the behavior	91
73	delve deeper into	91
74	plays a crucial role in ensuring	91
75	relentless pursuit	90
76	significant role in shaping	88
77	researchers aim to	88
78	meticulously crafted	88
79	study shed light on	87
80	dripping with sarcasm	87
81	aims to shed light	87
82	voice is rising	87
83	provides valuable insights	86
84	play a significant role in shaping	85
85	renewed sense of purpose	85
86	marked a significant	85
87	an enduring legacy	84
88	offers numerous benefits	84
89	commitment to excellence	83
90	study shed light	83
91	plays a crucial role in determining	83
92	significant attention in recent	83
93	offers a valuable	81
94	plays a significant role in shaping	79
95	play a crucial role in determining	78
96	despite the chaos	78
97	paving the way for the future	77
98	highlights the significance	77
99	locals and visitors alike	77

Table 7: GPTZero's Top AI Words and Phrases for November 2024 (*corrected version published in December 2024*)

#	Phrase	Freq.	#	Phrase	Freq.
1	provide a valuable insight	468	51	understand the behavior	61
2	left an indelible mark	317	52	broad implications	61
3	play a significant role in shaping	207	53	a prominent figure	61
4	an unwavering commitment	202	54	study highlights the importance	60
5	open a new avenue	174	55	a significant turning point	60
6	a stark reminder	166	56	curiosity piques	59
7	play a crucial role in determining	151	57	today in the digital age	59
8	finding a contribution	139	58	implication to understand	59
9	crucial role in understanding	135	59	a beacon of hope	58
10	finding a shed light	121	60	pave the way for the future	58
11	gain a comprehensive understanding	120	61	finding an important implication	57
12	conclusion of the study provides	119	62	understand the complexity	57
13	a nuanced understanding	115	63	meticulous attention to	57
14	hold a significant	114	64	add a layer	57
15	gain significant attention	107	65	the legacy of life	56
16	continue to inspire	105	66	identify the area of improvement	56
17	provide a comprehensive overview	104	67	aim to explore	56
18	finding the highlight the importance	99	68	highlight the need	55
19	endure a legacy	99	69	provide the text	55
20	mark a significant	96	70	conclusion of the study demonstrates	55
21	gain a deeper understanding	95	71	a multifaceted approach	55
22	the multifaceted nature	92	72	provide a framework to understand	55
23	the complex interplay	89	73	present a unique challenge	55
24	study shed light on	89	74	highlight the significance	54
25	need to fully understand	88	75	add depth to	54
26	navigate the complex	87	76	a significant stride	53
27	a serf reminder	85	77	gain an insight	53
28	the potential to revolutionize	83	78	underscore the need	52
29	the relentless pursuit	79	79	the importance to consider	52
30	offer a valuable	77	80	offer a unique perspective	52
31	underscore the importance	76	81	contribute to understanding	52
32	a complex multifaceted	74	82	a significant implication	52
33	the transformative power	74	83	despite the challenge faced	52
34	today the fast pace of the world	74	84	enhances the understanding	51
35	a significant milestone	73	85	make an informed decision in regard to	50
36	delve deeper into	72	86	the target intervention	50
37	provide an insight	71	87	require a careful consideration	49
38	navigate the challenge	71	88	essential to recognize	48
39	highlight the potential	69	89	validate the finding	48
40	pose a significant challenge	69	90	vital role in shaping	47
41	a unique blend	68	91	sense of camaraderie	47
42	a crucial development	68	92	influence various factors	47
43	various fields include	67	93	make a challenge	46
44	commitment to excellence	65	94	unwavering support	46
45	sent shockwaves through	65	95	importance of the address	46
46	emphasize the need	65	96	a significant step forward	46
47	despite the face	65	97	add an extra layer	45
48	understanding the fundamental	64	98	address the root cause	44
49	leave a lasting	63	99	a profound implication	44
50	gain a valuable	62	100	contributes to understanding	44

Table 8: GPTZero's Top AI Words and Phrases from January 2025 to March 2025

B Spearman ranking correlation

In this appendix, we present the Spearman rank correlations between the term rankings in each AI Vocabulary list (*October, Nov/Dec, Jan/Feb/Mar and All*) and the rankings of the same terms based on their frequency in the Ghostbuster corpus, considering the ChatGPT (*GPT*) and Claude (*Claude*) subsets separately, as well as the entire dataset (*All*).

AI Vocabulary	**LLM**	ρ	p
All	All	-0.064	0.316
Oct	All	-0.149	0.299
Nov/Dec	All	0.065	0.529
Jan/Feb/Mar	All	0.124	0.219
All	Claude	-0.170	0.007
Oct	Claude	**0.501**	**0.000**
Nov/Dec	Claude	0.045	0.660
Jan/Feb/Mar	Claude	0.476	0.000
All	GPT	-0.095	0.135
Oct	GPT	-0.243	0.088
Nov/Dec	GPT	0.211	0.039
Jan/Feb/Mar	GPT	-0.010	0.914

Table 9: Spearman ranking correlation coefficients and p-values between GPTZero's AI Vocabulary terms and the odds ratios of those terms in LLM-generated terms from the Ghostbuster dataset (*All, Claude or GPT only*).