Psychological Test and Assessment Modeling, Volume 65, 2023 (1), 125-144

Detection of AI-generated Essays in Writing Assessments

Duanli Yan, Michael Fauss, Jiangang Hao, Wenju Cui

Educational Testing Service, 660 Rosedale Road, Princeton, NJ, 08540

Abstract

The recent advance in AI technology has led to tremendous progress in automated text generation. Powerful language models, such as the GPT-3 and ChatGPT from OpenAI and BARD from Google, can generate high-quality essays when provided with a simple prompt. This paper shows how AI-generated essays are similar or different from human-written essays based on a set of typical prompts for a sample from a large-scale assessment. We also introduce two classifiers that can detect AI-generated essays with a high accuracy of over 95%. The goal of this study is for researchers to think and develop methodologies to address these issues and to ensure the quality of writing assessments.

Keywords: AI technology, GPT-3/ChatGPT, Writing assessment, Test security

Author Note

Correspondence concerning this article should be addressed to Duanli Yan, Educational Testing Service, Princeton, NJ 08541, dyan@ets.org

1 Introduction

With the advances in technology, the environments for learning and assessment have improved tremendously both in the classroom and at large-scale assessments, including Google Classroom, online assessments, and many apps for learning and assessment. These advances have enabled learners and test takers to improve their knowledge and skills in different subjects but provided opportunities for gaming the assessments (Holmes & Porayska-Pomsta, 2022). More recently, AI text generation demonstrated its capacity in dramatically improving writing quality, especially, for writers who need help on words, sentences, grammar, and other language mechanic aspects. With large language models like BERT, GPT-3 (Brown et al., 2020), and the more recent ChatGPT or BARD, students can learn and practice their writing skills using apps to generate text, start an introduction, raise an argument, draft a statement, and discuss a topic or write a whole essay. On the other side of the coin, AI-generated texts or essays may also be used for gaming assessments, especially remotely administered tests, leading to test security concerns.

Although student essays are often rated by experienced content experts, relying on human raters to identify AI-generated texts may not be feasible. Experiments suggest that humans rarely perform better than random guessing (60% - 65%) when asked to identify texts generated by modern AIs[1] (Clark et al., 2021; Ippolito et al, 2020). Moreover, most human raters are unable to explain their decisions beyond vague statements such as the text "rambles in a way that makes sense" or is "too natural to be AI-[generated]" (Clark et al., 2021; Ippolito et al., 2020). Evidence shows that human raters can be trained to identify better AI-generated text, however, accuracy generally remains below the levels that are acceptable in educational high-stakes assessment (Dugan et al, 2020). Nevertheless, it is likely just a matter of time until large language models can mimic diverse writing styles. Moreover, a detection based on shortcomings of an essay always has the problem that it only works in one direction. While a badly organized essay with many spelling mistakes is almost certainly human-written, a well-written, well-organized essay is not almost certainly AI-generated. Therefore, there is a strong need for systems that can detect AI-generated text automatically and reliably, especially in high-stakes assessment settings.

In recent years, the problem of automatically detecting AI-generated text has received considerable attention in literature. The proposed approaches can roughly be grouped into three families. The first family is detectors that use neural networks that are either fine-tuned or trained from scratch on large sets of human-written and AI-generated texts. The same pre-trained transformer models that are used to generate text often make for good detectors as well (Adelani et al., 2020; Fagni et al., 2020; Solaiman et al., 2019; Uchendu et al, 2020; Zellers et al., 2019). The second family is detectors

[1] Note that this development is relatively recent. ETS scientists have been researching to identify machine-created essays for quite a while. Since five years ago, machine-generated essays could be easily recognized, even by untrained readers (Cahill, Chodorow, & Flor, 2018).

based on explicit features extracted from the text. These features can range from simple words or n-gram frequencies to complex statistical and stylistic features designed by domain experts (Fröhling & Zubiaga, 2021; Gallé et al., 2021; Karumuri, 2022). Finally, the third family of detectors forms a middle ground because it uses large language models to calculate informative features. For example, the perplexity or likelihood that a specific large language model generates a given sequence of tokens can be used as a feature to detect AI-generated texts (Gehrmann et al., 2019; Hao, 2023; Solaiman et al., 2019; Tian, 2023).

All three families have their own strengths and weaknesses. While fine-tuned large transformer models generally yielded the best detection performance (Jawahar et al., 2020), they provide little to no human-interpretable evidence, making the detection process highly opaque. Traditionally, feature-based detectors are on the other end of the spectrum. They are transparent in that it is typically clear about which features led to a particular decision, but their performance is not as good as that of the fine-tuned large language model in many scenarios (Fröhling & Zubiaga, 2021; Yan et al., 2022). This also holds for the third family of detectors. For this family, the main bottleneck is that evaluating properties such as the perplexity or likelihood of a text with respect to a specific language model requires access to the weights of the underlying neural network. However, many state-of-the-art networks can only be accessed through limited APIs, if at all. In practice, open-source language models, such as GPT-2, are used as proxies for proprietary networks. That is, a text generated by ChatGPT is also likely to be generated by GPT-2. While this assumption is often justified today, it might become problematic in the future as language models evolve rapidly – often behind closed doors (Gershgorn, 2020).

Many applications and websites have been released recently with claims of detecting AI-generated texts (Slashdot, 2023). However, using these in the controlled environment of educational assessments could be problematic for various reasons. First, they are "all-purpose" detectors trained on vast data sets containing many types of text – from recipes to computer code – irrelevant to high-stakes writing assessments. Moreover, many of these tools seem to be released without clear performance metrics. This makes it difficult to justify and trust the detection result, especially in high-stakes test situations, where the detection outcome can have significant consequences for the test takers.

On the other hand, for those tools that come with transparent performance metrics, such as OpenAI's detector (OpenAI, 2023), it is still unknown how these metrics change if the tool is applied to the much narrower scope of essays for writing assessments. Finally, using a third-party detector raises privacy and information security issues that would have to be addressed and monitored. These difficulties are not unsurmountable, but they illustrate the advantages of a detector that is custom-made and custom-trained for the task of detecting AI-generated essays in high-stakes large-scale assessments.

In this paper, we explored the detection of AI-generated essays responding to large-scale writing assessment prompts by developing automated detectors and explored in-

depth the features of AI-generated essays for the studied prompts. To this end, we provided writing prompts to a state-of-the-art language model, elicited responses from the model, and compared its responses to those of humans. As a first study on this topic in educational assessment, this study intended to provide systematic investigation of the detection issue. Previous results are either anecdotal (@teddynpc, 2022) or based on more restricted tasks such as SAT-style analogies or sentence insertion problems (Wu & Bai, 2021). In this paper, we focus on the first two families of detectors. For the first approach, we fine-tuned a RoBERTa model, and for the second approach, we trained a variety of classifiers based on features generated by ETS's in-house scoring engine, e-rater®[2] (Attali & Burstein, 2006).

It is worth noting that reporting findings from test security research is very sensitive in the sense that test takers may potentially use the findings to game the tests. As such, we avoided specifically defining the writing features used below to characterize text. Instead, we simply refer to them as Feature 1, Feature 2, and Feature 3. The details of all e-rater® features can be found in the publicly available document (Attali & Burstein, 2006).

2 Method

This study consists of two parts. First, we explored a small-sample-size, in-depth study to examine various features from AI-generated essays and compare those with human-written essays based on the same prompt. Second, we conducted a large-sample-size study in which we trained and evaluated two detectors, one based on e-rater® features while the other using a large variant of the well-known RoBERTa language model.

Training and evaluating a detector with clear performance metrics requires a large set of representative and labeled data. To produce a sample of AI-generated essays, we used OpenAI's GPT-3 large language model.[3] The interested reader is referred to Brown et al. (2020) and the references in the paper for details on GPT-3's size, architecture, training data set and other technical aspects (OpenAI, 2023).

To explore the characteristics of AI-generated texts, we randomly selected four writing prompts from a large pool of writing assessment prompts and extracted a set of 1,000 essays per prompt from a large-scale writing assessment. The same prompts were used to produce several versions of GPT-3-generated essays by varying GPT-

[2] e-rater® is an automated scoring engine developed at ETS using the most recent NLP development and is based on millions of human raters. It went through extensive evaluations based on several hundreds of thousands of essays at each upgrade almost annually to ensure its scoring quality (Attali & Burstein, 2006).

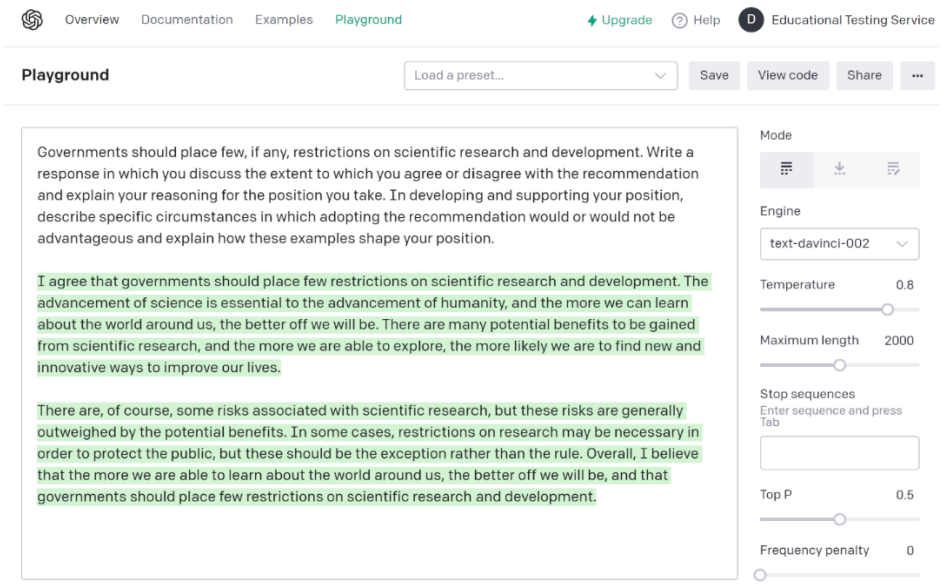
[3] During the preparation of this manuscript, ChatGPT was released by OpenAI, which replaced the traditional interface used by GPT-3 with a chat box, allowing the users to have more natural conversations with AI.

3's parameters including Temperature and Length. Figure 1 shows an example of a GPT-3-generated essay given a published GRE prompt. The exact parameter settings are elaborated below.

To obtain a baseline for the study, we mimicked the case when a test taker simply inputs a prompt to GPT and copies its answer. At this point, it is not clear how test takers have used and will use large language models to game assessments. Prompt tuning can be done and is just one option of many to obtain a desired text. The problem of having to modify a given text so that it passes as "handwritten" is not specific to AI-based cheating. Anyone copying from a template, such as Wikipedia or various cans, faces this "challenge."

Figure 1

A Published GRE Prompt with GPT-3 Generated Essay



2.1 Feature Exploration – Small Sample

Before training a detector on a large sample of essays, it is helpful to inspect a small sample of representative essays more in-depth to identify differences in individual features. The insights gained in this analysis help us identify informative features and generally allow for a more fine-grained assessment of which characteristics of GPT-

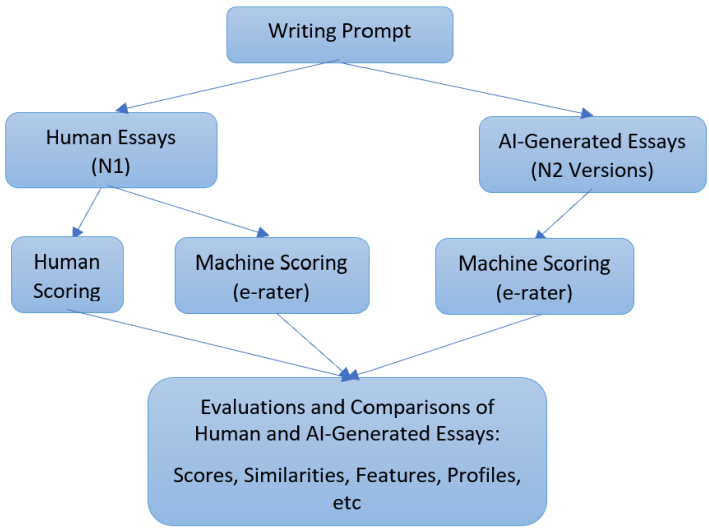
3-generated texts are indistinguishable from human writing and which, upon closer, machine-assisted inspection, differ.

For both the real test-takers' writing samples and AI-generated essays, we used ETS's automated scoring engine (e-rater®) to score them and produce writing features that characterize each essay. The analysis procedures (Attali & Burstein, 2006)(Attali & Burstein, 2006) are as follows (as seen in Figure 2):

- First, select writing prompts from a large-scale writing assessment.
- Next, extract some real test takers' writing samples for these prompts.
- Third, use GPT-3 to generate several versions of the essays based on the same prompts.
- Fourth, use e-rater® to score both human-written essays and AI-generated essays.
- Finally, evaluate and compare the features of both human-written and AI-generated essays.

Figure 2

Analysis Flowchart



To evaluate the AI-generated essays and compare the AI-generated essays with real test takers' written essays, we examined all features extracted from e-rater® on the feature values and distributions. We compared feature values and distributions for essays with different lengths from short, medium, and long for both human and AI-generated essays from a large-scale writing assessment.

We selected human essays with different lengths based on a specific prompt. The definition of short, medium, and long essays is based on their lengths. We used GPT-3 to generate AI essays to mimic human essays with corresponding lengths based on the same prompt. We compared GPT-3 generated essays with human essays at various lengths, examined the whole range of features from the AI-generated essays and human essays, and compared the characteristics of these essays in detecting the AI-generated essays.

2.2 Feature Exploration - Large Sample

To investigate the differences at a large-sample level, we used OpenAI's Python API to generate 1,000 essays for each of four different prompts, resulting in a training data set of 4,000 human-written and 4,000 AI-generated essays. There is no consensus among the machine learning community whether the training set for a binary classifier should be balanced (50% positives and 50% negatives) or match the (estimated) proportions of positive and negative samples in the real data. However, since we have yet to learn how extensive the use of large language models in writing assessments is, we decided to use a balanced data set for both training and evaluation.

We used the text-davinci-002 model and instructed GPT-3 to write at least 500 words. It is worthy of note that GPT-3 does not always comply with the given instructions. The resulting distribution of the true essay lengths is not under the control of the researchers. In order to generate a diverse sample of essays, we chose a relatively high sampling temperature of 1.2. Both the frequency and the presence penalty were set to zero. Moreover, to simulate copy-typing the AI-generated text by test-takers,[4] we added artificial typos to the essays using the typo library (Kumar, 2022) to add common spelling mistakes. Finally, based on the spelling-mistake statistics from Flor et al. (2015), we set the frequency of typos to 2%, which is approximately the average frequency of misspelled words in a GRE essay written by an English native speaker scoring a score of 3, or a non-native speaker scoring a score of 4.

Another important aspect worth exploring is the generated essay *similarity*. It is not clear that an advanced language model like GPT-3 can generate thousands of responses to the same prompt without repeating itself. If this were the case, it would be possible to detect AI-generated essays or at least GPT-3 generated essays by comparing a given sample to a catalog of representative AI-generated essays. To gain insight into how similar the AI-generated essays are, we randomly sampled a subset of 2,000 essays and calculated the pairwise 3-gram cosine similarities for all possible pairs. We then looked at the share of pairs that exceed a given similarity threshold of 0.05. For comparison, we repeated the same procedure for a random subset of 2,000 human-

[4] For most computer-based writing assessments, it is not possible to directly paste into the input form, the test taker must manually copy the text.

written essays and a mixed sample in which 1,000 essays were AI-generated and 1,000 were human-written.

Next, we evaluated the features extracted for the large sample of human-written and AI-generated essays. Note that instead of exact values, we provide means and standard deviations over the respective groups of 4,000 essays.

2.3 Detector based on a Fine-Tuned RoBERTa Model

Large, pre-trained language models have been used successfully to generate text and classify text. In particular, the RoBERTa model (Liu et al., 2019), a robustly trained variant of Google's BERT model (Devlin et al., 2018), has repeatedly shown excellent detection results (Minaee et al., 2022). Therefore, we chose RoBERTa to develop a classifier that detects AI-generated essays.

2.4 Detector based on e-rater® Features and Support Vector Machine

All human-written and AI-generated essays were scored by ETS's e-rater® engine. During this process, e-rater extracted nearly 200 features that we used for our study. For scoring, e-rater® uses a hierarchical approach, in which it combines groups of low-level features into fewer macro-features (Attali & Burstein, 2006). We used both low-level and aggregated features in the detector design to maximize the available information. Using the Python machine-learning package scikit-learn (Pedregosa et al., 2011), we trained several feature-based binary classifiers, i.e., combinations of features from e-rater®, including and compared their performance via five-fold cross-validation using a support vector classifier (SVC).

3 Results

3.1 Feature Exploration – Small Sample

For the study on the small sample, Table 1 presents an example comparing essay scores and features for both human-written and AI-generated essays. For the randomly selected writing prompt, we listed four human-written essays in different lengths (short, medium, long), their human rater scores (1-5), e-rater® scores (0-5), and their feature values. We also listed four versions of AI-generated essays in different lengths, e-rater® scores (3-4), and their feature values.

Table 1

Comparison of Human and AI-generated Essays

Essays	Length	Human	e-rater	Feature 1	Feature 2	Feature 3
H1	11	1	0	2.398	0.000	-0.603
H2	306	3	3	3.778	1.946	-0.198
H3	444	4	5	3.899	2.197	-0.116
H4	605	5	5	4.459	1.946	-0.122
GPT3_1	109	NA	3	3.998	0.693	0.000
GPT3_2	119	NA	3	4.086	0.693	0.000
GPT3_3	176	NA	4	3.379	1.792	0.000
GPT3_4	214	NA	4	3.979	1.386	0.000

The human scores range from 0 to 5 based on operational scoring rubrics. For example, a score equals 1 if an essay has 1) serious disorganization or underdevelopment 2) little or no detail, or irrelevant specifics, or questionable responsiveness to the task, 3) serious and frequent errors in sentence structure or usage; a score equals 5 if an

essay has 1) effectively addresses the topic and task, 2) is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details, 3) displays unity, progression and coherence, 4) displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors. e-rater® scoring ranges from 0 to 5 based on its scoring model, trained and evaluated with large data sets, containing a set of essay features that predict the human scoring for the same prompt.

For the human-written essays, H1–H4, the short essay contained 11 words, the medium essays contained 306 and 444 words, and the long essay contained 605 words. The human raters assigned scores of 1, 3, 4, and 5, while the e-rater® assigned scores of 0, 3, 5, and 5, respectively. For the AI-generated essays, GPT3_1 – GPT3_4, the short essay contained 109 words, the medium essays contained 119 and 176, and the long essay contained 214 words. Given a prompt, GPT-3 generated essays that were sufficiently long to adequately answer the question in the writing prompt (between 100 and 200 words). The e-rater® assigned scores of 3, 3, 4, and 4 on the GPT3-generated essays.

In reality, humans write essays in any length, while GPT-3 doesn't generate essays too short (e.g., 11 words). In general, it generates an essay with adequate length to answer a prompt unless you give more information (or a revised prompt) for a longer essay. Our results show that, given a prompt, the essays that GPT-3 generated were not greatly different in lengths with its parameter settings. GPT-3 did generate longer essays with modified prompt with additional information included in the prompt. In Table 1, GPT-3 generated essays with lengths from 109 to 214. It can't generate an essay with only 11 words, and it couldn't generate an essay with more than 600 words for the same unmodified prompt.

We examined the whole range of essay features including but not limited to grammar. Given the space limit, we only illustrated a few in the paper. On examining the essay features, for both human and AI-generated essays, all the essay features extracted were within the normal ranges of their numerical values. Features 1, 2, and 3 were part of the standard features from the automated scoring engine. Again, for test security, we did not list the specific feature names. These features generally correspond to writing style, mechanics, usage, and grammar. We refer readers to Attali & Burstein (2006) for more details.

The AI-generated essays demonstrated the same characteristics consistently across all essays with different lengths. However, there were some salient feature differences. For example, all the AI-generated essays had zeros on Feature 3, as compared to the negative values for human-written essays. Feature 3 is about grammatical errors. It was found that this discrepancy was consistent across all the AI-generated essays regardless of essay lengths. This indicates that AI essays (GPT3_1 - GPT3_4) did not commit any common human errors, i.e., zero scores on Feature 3, for which even good writers still make (e.g., high scored human essay H4) among other characteristics. Thus, AI essays would receive higher scores either scored by human raters or by e-rater® due to few or no common human errors. This result is consistent with other

research findings that AI text generators make less grammatical errors (Crothers, et al, 2023).

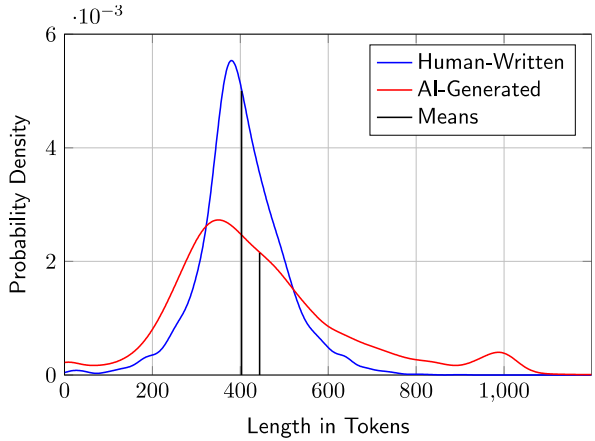
In this study, we didn't recruit human raters to score the AI-generated essays. Given the e-rater® engine is built based on millions of human raters' scores and extensively evaluated on several hundreds of thousands of human essays at each year over the last decades, it is expected that human scoring of AI-generated essays has a large variation compared to the e-rater® scoring of the same essays. Thus, human scoring of AI-generated essays could be very unreliable and is not appropriate nor cost effective.

3.2 Feature Exploration - Large Sample

For the study on the large sample, the distributions of empirical lengths of both sets of essays are plotted in Figure 3. By inspection, the AI-generated sample yielded a slightly larger mean and variance. However, both distributions showed approximately the same essay length range, which ensures that the AI-generated essays cannot simply be identified by the length.

Figure 3

Essay Length Distributions

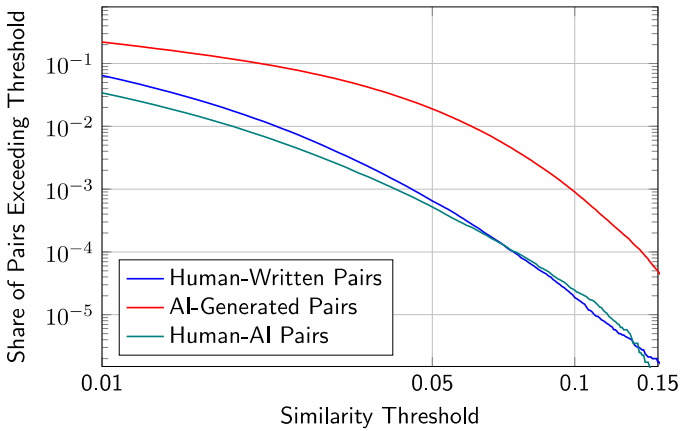


To gain insight into the similarity of the AI-generated essays, the share of pairwise 3-gram cosine similarities that exceed a given similarity threshold is plotted in red in Figure 4 as a function of the threshold. The shares of pairs exceeding a given threshold are also plotted in blue and teal, respectively, in Figure 4. In general, the AI-generated essays were more similar to each other than the human-written ones, but not by a large

margin. For example, only one in a thousand AI-generated pairs exceeded a cosine similarity threshold of 0.1. Moreover, the similarity profile of the mixed sample in teal, was almost identical to that of the human-only sample, particularly at threshold values larger than 0.05. However, although the 3-gram cosine similarity is widely used in practice, it might not be the best measure in this context. Ongoing research at ETS has shown that classifiers based on an ensemble of different similarity measures perform significantly better and can provide additional evidence in practice.

Figure 4

Essay Length Distributions vs. Essay Similarity Profile



Next, we evaluated the features in Table 1 for the large sample of human-written and AI-generated essays. Given the large sample size, rather than presenting the exact values for each essay feature, we provide the means and standard deviations of the feature values based on the 4,000 AI and human essays respectively. In particular, adding artificial typos to AI-generated essays reduced the number of zero values in Feature 3 as observed in the small-sample case, though there are still differences among other features.

Table 2

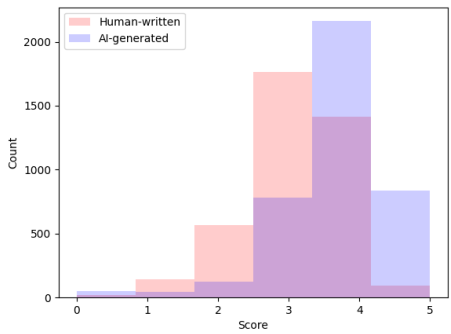
Comparison of Human and AI-generated Essays – Large Sample

Essays	LENGTH	Feature 1	Feature 2	Feature 3
HUMAN				
mean	340.4748	3.8294	1.9597	-0.1796
STD	84.7216	0.3456	0.3450	0.0704
AI				
MEAN	363.5910	3.7997	2.0014	-0.1715
STD	147.6294	0.3846	0.4669	0.0714

Moreover, the scores assigned to the essays by e-rater® contained some helpful information. For example, as seen in Figure 5, AI-generated essays were significantly more likely to get the best and second-best scores while also being more likely to get the worst score.[5] On the other hand, human-written essays were more likely to get average scores. These differences demonstrate that AI tools can help test takers to obtain spuriously high scores on essay exams. In our experiment, almost 75% of the AI-generated essays scored 4 or higher, with over 20% getting the top score of 5. In comparison, the top score was assigned to only 2.5% of the human essays.

Figure 5

Score Distributions



[5] Given the relatively high sampling temperature we used when generating the essays, some of these are cases where GPT-3 started generating nonsense or even random symbols after some time. This phenomenon is known as *text degeneration* and has been studied in, for example, (Holtzman et al., 2020).

3.3 Detector based on a fine-Tuned RoBERTa Model

We fine-tuned the large variant of the pre-trained RoBERTa model using a data set of 8,000 essays, 4,000 of which are AI-generated with spelling mistakes added. 60% (4,800 essays) of the data set was used for training, and 20% (1,600 essays) were used for validation and evaluation, respectively. We used the ADAMW variant of stochastic gradient descent (Loshchilov, 2017) with a batch size of 32 and a learning rate of $4x10^{-5}$ to fine-tune the network. The evaluation loss was minimal after four epochs. On our test set, the fine-tuned RoBERTa-based detector achieved an accuracy (percentage of correctly identified essays) of 99.75%, mislabeling two essays in each class.

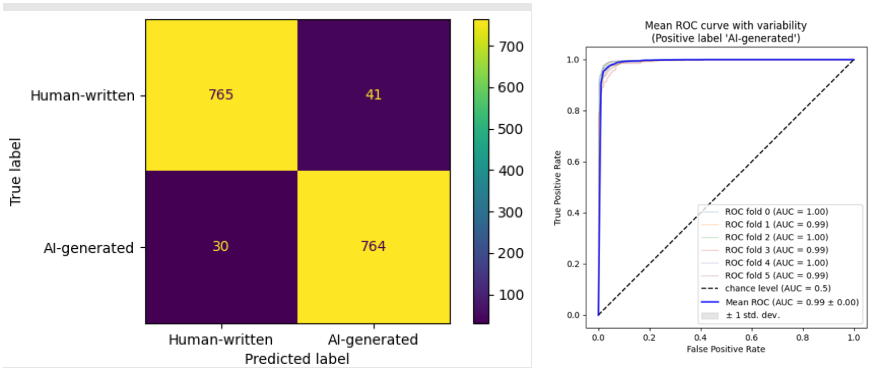
Although this high detection accuracy is promising, a drawback of the RoBERTa-based detector is that it is a black-box. Upon the development of the detector, there is little information about how the features and characteristics of the two classes of essays determine the detection results. However, given the high-stakes nature of many writing assessments, transparency is often critical for validity evidence collection. Therefore, it is worthwhile to further explore the development of a detector that uses explicit, well-defined features that provide insight into what features of a flagged essay characterize it as AI-generated.

3.4 Detector based on e-rater® Features and Support Vector Machine

A support vector classifier (SVC) using radial basis functions consistently performed well in terms of detection accuracy. The support vector classifier attained an average accuracy of approximately 96%. Figure 6 shows the detector's performance in terms of the confusion matrix (left) and ROC (right).

Figure 6

Performance of the Detector Based on e-rater® Features

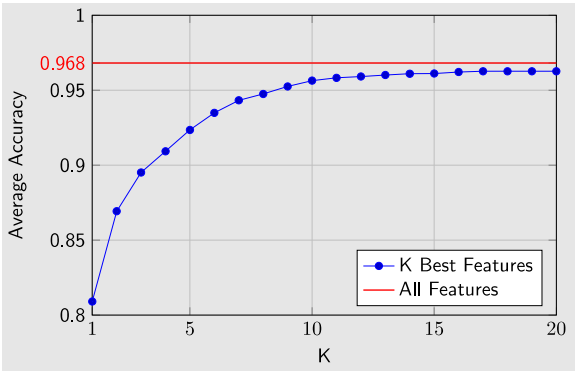


Note: *Left: Confusion matrix. Right: ROC curve*

Our further investigation also revealed that a relatively small subset of the e-rater® features already contained a large amount of information about whether an essay was AI-generated. This is illustrated in Figure 7 plotting the average (over five cross-validation folds) accuracy of the SVC detector as a function of the number of features. The best features were chosen greedily via sequential forward selection. The top three features were already enough to attain 90% accuracy. Moreover, the detector's performance started to plateau after approximately 15 features. These most discriminative features overlapped with the features we discussed before. For test security, the features are not listed explicitly to avoid potential gaming of our tests.

Figure 7

Classification Accuracy as the Number of Features Changes



4 Discussion

This paper presented an empirical study to investigate the systematic difference between AI (GPT-3) generated and human-written essays based on samples from a large-scale writing assessment. As the first research on detecting AI-generated essays in educational assessment, we also developed two detectors for AI-generated essays and established benchmark reference classification accuracy based on our samples. By applying ETS e-rater® to both human-written and AI-generated essays, AI-generated essays showed fewer grammar errors and other errors. Most AI-generated essays were scored higher by e-rater®, which justifies the motivation of using AI in gaming tests. Both the e-rater® feature-based and the pretrained-RoBERTa-based detectors can detect AI-generated essays with high accuracy. In summary, the main findings of this study are summarized as follows.

1. State-of-the-art large language models can generate essays in response to writing prompts that are, in many aspects, indistinguishable from human-written essays for untrained readers and general-purpose automated scoring systems such as e-rater®.
2. AI-generated essays showed statistical anomalies compared to their human-written counterparts. In particular, there are no spelling and grammar mistakes in AI-generated essays.
3. Fine-tuned large language model can detect AI-generated essays with great accuracy, exceeding 99% in our experiments.
4. Traditional classifiers based on e-rater® features with SVM do not perform as well as the pre-trained model but still reached accuracies around 95% in our experiments.
5. Some features extracted by e-rater® are quite different between human-written and AI-generated essays.

Though the detectors explored in this study are promising, the current study has several limitations. First, the AI-generated essays used in this study were based on limited variations of the prompts used to interact with the large language models. For example, we did not generate grammar mistakes, and assuming the error frequency to be a global parameter is an oversimplification. However, since our aim was not to model human writing but human copy-typing, we considered the chosen approach reasonable approximation. It is known that different ways of prompting the large language models could lead to different quality of the generated texts. For example, instead of only providing the prompt, GPT-3 could be given several examples of well-written essays. Alternatively, the model could be fine-tuned based on feedback on the generated essays. Hence, the samples used in this study (though a reasonably large sample) could underestimate the variation of AI-generated essays. In a future study, we will explore modified prompts with added prompt information, introduce more variations to the prompt texts, and increase the sample size to approximate wider range of real-world AI-generated essays.

Second, the AI-generated essays used in this study were mostly intact except for some typos being added. In real-world applications, people are more likely to make more revisions and edits on top of AI-generated texts, which will reduce the differences revealed in this paper. Therefore, it is possible that the difference will eventually diminish if enough edits are applied. Considering other information, such as the key-stroke writing process, could help resolve this issue (Hao, 2023).

Third, the length distribution of the essays generated from GPT-3 used in this paper did not completely match with that from human-written essays. This is because the GPT-3 does not strictly follow our requests and prompts for the number of words in its generated essays. An improved comparison could include some resampling to make the length distribution of the samples closer. Meanwhile, the comparison in this paper is based on four writing prompts and the generalizability of the findings to a broader range of prompts is unclear.

Finally, in this study, we did not comprehensively compare the essays generated by different large language models but mainly focused on GPT-3. It is plausible to expect that essays generated by different large language models could bear different features.

The findings from this study show that AI could generate human-like essays. Stake-holders such as schools and testing organizations need to get prepared on detecting potential AI-facilitated essays submitted by students and test-takers. When scoring and evaluating essays submitted in high-stake assessment, questions center around potential misuse of AI in assessment should be highlighted and addressed: is the essay human-written essay or AI-generated essay or a combination of human and AI-generated essay? With the debut of ChatGPT, the assessment field faces new challenges due to potential misuse of generative AI technology in both low-stakes and high-stakes assessment settings. New approaches and solutions are needed to allow assessment to take advantage of generative AI but not allow the gaming of assessment.

References

@teddynpc. (2022, 12 2). *I made ChatGPT take a full SAT test. Here's how it did.* Retrieved from Twitter: https://twitter.com/davidtsong/status/1598767389390573569?s=20&t=UqurMfQ5_X3wFGvDr4c-EQ

Adelani, D. I., Mai, H., Fang, F., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2020). Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection. *Proceedings of the 34th International Conference on Advanced Information Networking and Applications*, (pp. 1341–1354).

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

Cahill, A., Chodorow, M., & Flor, M. (2018). Developing an e-rater advisory to detect babel-generated essays. *Journal of Writing Analytics, 2*, 203-224.

Choi, I., Hao, J., Deane, P., & Zhang, M. (2021). Benchmark keystroke biometrics accuracy from high-stakes writing tasks. *ETS Research Report Series*, 2021(1), 1–13.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's 'human' is not gold: evaluating human evaluation of generated text. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, (pp. 7282–7296).

Crothers, E., Japkowicz, N., & Viktor, H. (2023). *Machine generated text: a comprehensive survey of threat models and detection methods*. https://arxiv.org/pdf/2210.07321.pdf.

Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks*. ETS Research Report Series, 2014(1), 1–23.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Dugan, L., Ippolito, D., Kirubarajan, A., & Callison-Burch, C. (2020). RoFT: A tool for evaluating human detection of machine-generated text. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (pp. 189–196).

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2020). *TweepFake: about detecting deepfake tweets.* Retrieved from CoRR: abs/2008.00036

Flor, M., Futagi, Y., Lopez, M., & Mulholland, M. (2015). Patterns of misspellings in L2 and L1 English: a view from the ETS spelling corpus. *Bergen Language and Linguistics Studies, 6*. https://doi.org/10.15845/bells.v6i0.811

Fröhling, L., & Zubiaga, A. (2021). Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7:e443.

Gallé, M., Rozen, J., Kruszewski, G., & Elsahar, H. (2021). *Unsupervised and distributional detection of machine-generated text.* Retrieved from arXiv: https://arxiv.org/abs/2111.02878

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). *GLTR: Statistical detection and visualization of generated text.* Retrieved from arXiv: https://arxiv.org/abs/1906.04043.

Gershgorn, D. (2020, 8 20). *GPT-3 is an amazing research tool. but Openai isn't sharing the code.* Retrieved from OneZero: https://onezero.medium.com/gpt-3-is-an-amazing-research-tool-openai-isnt-sharing-the-code-d048ba39bbfd.

Hao, J., (2023). *Detecting chatgpt-generated essays for high-stakes applications: What you should keep in mind*, LinkedIn, Retrieved February 10, 2023, from https://www.linkedin.com/pulse/detecting-chatgpt-generated-essays-high-stakes-applications-hao.

Hao, J. & Fauss, M., (2022, November). *Test security in remote testing age: perspectives from process data analytics and AI*, Presentation at the 20th annual Maryland Assessment Research Center (MARC) conference. College Park, Maryland.

Holmes, W., & Porayska-Pomsta, K. (Eds.). (2022). *The ethics of artificial intelligence in education: Practices, challenges, and debates* (1st ed.). Routledge. https://doi.org/10.4324/9780429329067

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 1808–1822).

Jawahar, G., Abdul-Mageed, M., & Lakshmanan, L. V. (2020). *Automatic detection of machine generated text: A critical survey.* Retrieved from arXiv: https://arxiv.org/abs/2011.01314.

Karumuri, R. T. (2022). *Interpretable features for distinguishing machine generated news articles.* A thesis presented in partial fulfillment of the requirement for the degree of Master of Science, Arizona State University.

Kumar, R. (2022). *A python package to simulate typographical errors in English language.* Retrieved from https://github.com/ranvijaykumar/typo

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692.

Loshchilov, I., & Hutter, F. (2017). *Decoupled weight decay regularization*. arXiv preprint arXiv:1711.05101.

Minaee, S. Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J. (2021). Deep learning--based text classification: a comprehensive review. *ACM Comput. Surv*. 54, 3. https://doi.org/10.1145/3439726

OpenAI. (2023). *New AI classifier for indicating AI-written text.* Retrieved from https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text/

Pedregosa *et al. (2011).* Scikit-learn: Machine Learning in Python, *JMLR* 12, pp. 2825-2830.

Slashdot (2023). *Best AI content detection tools of 2023*. https://slashdot.org/software/ai-content-detection/

Solaiman, I., Brundage, M., Clark, J., Askel, A., Herbert-Voss, A., Wu, J., . . . Wang, J. (2019). *Release srategies and the social impacts of language models.* Retrieved from CoRR: abs/1908.09203.

Tian, E. (2023). *GPTZero*. Retrieved from https://gptzero.me/.

Uchendu, A., Le, T., Shu, K., & Lee, D. (2020). Authorship attribution for neural text generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Wu, F., & Bai, X. (2021). *InsertGNN: Can graph neural networks outperform humans in TOEFL sentence insertion problem*? arXiv preprint arXiv:2103.15066.

Yan, D., Fauss, M., Cui, W., & Hao, J., (2022, October). *Detection of AI-generated essays*, Presentation at the Conference on Test Security, Princeton, NJ.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in Neural Information Processing Systems*, pp. 9054–9065.

Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features*. ETS Research Report Series, 2015 (2), 1–12.