scientific reports

OPEN

Admissions in the age of AI: detecting AI-generated application materials in higher education

Yijun Zhao✉, Alexander Borelli, Fernando Martinez, Haoran Xue & Gary M. Weiss✉

Recent advances in Artificial Intelligence (AI), such as the development of large language models like ChatGPT, have blurred the boundaries between human and AI-generated text. This has led to a pressing need for tools that can determine whether text has been created or revised using AI. A general and universally effective detection model would be extremely useful, but appears to be beyond the reach of current technology and detection methods. The research described in this study adopts a domain and task specific approach and shows that specialized detection models can attain high accuracy. The study focuses on the higher education graduate admissions process, with the specific goal of identifying AI-generated and AI-revised Letters of Recommendation (LORs) and Statements of Intent (SOIs). Detecting such application materials is essential to ensure that applicants are evaluated on their true merits and abilities, and to foster an equitable and trustworthy admissions process. Our research is based on 3755 LORs and 1973 SOIs extracted from the application records of Fordham University's Master's programs in Computer Science and Data Science. To facilitate the construction and evaluation of detection models, we generated AI counterparts for each LOR and SOI using the GPT-3.5 Turbo API. The prompts for AI-generation text were derived from the admission data of the respective applicants, and the AI-revised LORs and SOIs were generated directly from the human-authored versions. We also utilize an open-access GPT-wiki-intro dataset to further validate our hypothesis regarding the feasibility of constructing domain-specific AI content detectors. Our experiments yield promising results in developing classifiers tailored to a specific domain when provided with sufficient training samples. Additionally, we present a comparative analysis of the word frequency and statistical characteristics of the text, which provides convincing evidence that ChatGPT employs distinctive vocabulary and paragraph structure compared to human-authored text. The code for this study is available on GitHub, and the models can be executed on user-provided data via an interactive web interface.

The rapid and remarkable progress in generative AI technologies, such as ChatGPT[1], has revolutionized automated content creation, giving rise to a new and complex challenge: the accurate differentiation between human-authored and AI-created content. As AI models become increasingly sophisticated at replicating human-authored text, images, audio, and video, the distinction between what is authentically human and what is artificially generated has become progressively more difficult to discern. These AI models are versatile and are being used to generate problematic content, such as deepfakes[2], fake news[3], and even fake art[4].

Distinguishing between human-authored and AI-crafted text is particularly important in the higher education domain, especially within college applications. As aspiring students compete for desired spots in universities, their application materials, such as essays, portfolios, and creative projects, serve as tangible demonstrations of their abilities and potential. However, with the advent of generative AI, admissions committees now find themselves challenged with not only evaluating the quality of applicants' work, but also with recognizing the impact of AI tools on content creation. To ensure fairness and integrity in the admissions process, institutions must balance between acknowledging the creative augmentation AI can offer and upholding the principles of meritocracy. Thus, developing robust AI detection mechanisms tailored to distinguish between human-authored and AI-generated content has become an essential part of modern higher education. Furthermore, because written communication skills are often relevant to the admission decision (especially for international students with a different first language), it is also important to be able to distinguish between human-authored text and AI-revised text, where the AI program modifies the text to improve the writing style, grammar, and/or vocabulary.

Computer and Information Sciences Department, Fordham University, New York, NY 10023, USA. ✉email: yzhao11@fordham.edu; gaweiss@fordham.edu

With the realism of AI-generated content reaching an unprecedented level across various domains, the task of detecting such content presents a formidable challenge. Despite extensive research and tool development[5–8], achieving a *universally effective* detection model remains elusive. This study explores the more limited objective of constructing a domain-specific AI content detection model. The focus is on two distinct types of admission documents within the education domain: Letters of Recommendation (LORs) submitted on behalf of an applicant and Statements of Intent (SOIs) submitted by applicants to explain their rationale for applying to the program. LORs are a standard requirement for most academic programs, while SOIs are more prevalent in graduate programs. We cast our detection problem as two classification tasks: distinguishing between human-authored and AI-generated text, and distinguishing between human-authored and AI-revised text. The rationale for this division is provided in the "Methods" section.

The data utilized by this study include a proprietary dataset comprising LORs and SOIs extracted from graduate applications at Fordham University, along with the open-access GPT-wiki-intro dataset[9]. The detection models used in this study employ two traditional machine learning models (Naïve Bayes[10] and Logistic Regression[11]) and two transformer-based LLMs (BERT[12] and DistilBERT[13]). Our experiments reveal that, while all models demonstrate excellent performance when trained with domain-specific data, the LLMs exhibit an advantage over the traditional models, especially for detecting AI-revised documents.

This study makes the following contributions. First, at the application level, it helps to preserve academic integrity and promote fair evaluations in response to the increasing use of writing assistance tools based on LLMs. This effort is particularly relevant for SOIs, which serve not only to understand applicants' motivations but also as a venue to assess their communication abilities. Second, this study contributes to the evidence that training effective detectors tailored to specific tasks using domain-specific data produces excellent results, while accurate universal AI-content detectors are currently unattainable. This study additionally shows that we can accurately identify AI-revised content. Finally, this study sheds light on some characteristics (e.g., vocabulary, paragraph structure, etc.) that help to distinguish between AI-generated and human-authored content.

Related work

Distinguishing between AI-generated and human-authored text is similar to detecting plagiarism. Both are motivated by the need to detect student cheating. In fact, many plagiarism detection companies, such as Turnitin[6], have expanded their focus to include AI-generated text. However, unlike plagiarism, the ethical considerations around AI tools for text generation are less clear-cut[14]. The guidelines for acceptable use of AI in academic writing are still evolving[15,16]. While some schools have banned AI tools outright and view them as a significant threat[17], others advocate for a more nuanced approach, acknowledging the potential role of AI in education[18]. Here we focus on the tools and strategies used to detect AI-generated content, as they are most relevant to our study.

Detection methods and approaches

Strategies to address the emergence of AI-generated text are of great interest. Three high-level approaches have been identified[16]: task design, institutional policy, and automatic identification of AI content. Task design involves creating assignments to thwart the effectiveness of AI text generation tools[19] while institutional policy relies on guidelines[14] and education[20] to prevent the unethical use of such tools. For automatic detection, methods can be divided into feature-based, neural language model-based, and domain-specific approaches[21].

Feature-based methods detect differences between AI-crafted and human-authored text through statistical analysis of features like word frequency, sentence and paragraph length, and the frequency of different linguistic features such as parts of speech[22,23]. Detection methods that rely on neural networks, especially those based on Transformer-based neural language models, are quite common. They can be divided into zero-shot methods that use the pre-trained models without modification[24] and those that fine-tune the pre-trained language models. When GPT-2 is constrained to pick among the top 40 words, OpenAI's zero-shot approach detects GPT-2 generated text with 1.5 billion parameters with an accuracy between 83% and 85%[25]. GROVER, a generative text model that generates fake news articles, detects AI generated fake news using the one-shot approach with 92% accuracy, but does poorly when evaluated on text from domains not present in the training set[3]. Zero-shot approaches have been outperformed by simple TF-IDF methods when detecting output generated by a model trained on another domain[25].

The neural language model-based approach is based on fine-tuning large bidirectional language models[25]. RoBERTa[26], which is based on BERT[12], uses fine-tuning to distinguish between AI-crafted and human-authored text by training on samples from each. Research shows that fine-tuning using even a few hundred samples can dramatically improve performance[27].

There is also a substantial amount of research that focuses on domain-specific detection. For example, A detector based on RoBERTa[26] trained to identify physics papers was fine-tuned to identify biomedicine papers using a few hundred examples[27]. Work in social networks has shown that detection performance is highly related to the data set used to train the model[28]. Fake Yelp reviews can be detected by a customized GPT-2 model fine-tuned on the Yelp reviews[29]. In all of these efforts, a general model was not used to detect the AI generated text. The findings indicate detectors built, or adapted, for the domain will provide better results.

Our study introduces models specifically designed to detect AI-generated content in academic environments. By concentrating on domain-specific application materials, such as LORs and SOIs, we tailor our detection models to better capture the unique characteristics of these texts. Instead of pursuing a general-purpose model – which may struggle with accuracy across diverse content types – we advocate for a specialized approach that proves more effective. Our findings demonstrate that targeting specific document types significantly enhances detection accuracy, offering a practical solution to the challenge of AI text detection in higher education admissions processes.

Practical tools and challenges

Simple classification techniques, such as logistic regression classifiers trained on TF-IDF features, have also been employed for AI detection. For instance, OpenAI achieved 88% and 74% accuracy on detecting GPT-2 text with 124 million and 1.5 billion parameters, respectively. Constraining GPT-2 to select from the top 40 words during text generation further increased detection accuracy to 97% and 95%[25]. While these methods show promise, commercially available tools still struggle with reliability. A non-academic study by TechCrunch contains its conclusion in its title "Most sites claiming to catch AI-written text fail spectacularly"[30]. OpenAI, one of the tools included in that study, has stated "Our classifier is not fully reliable" and that it detects only 26% of AI generated text as likely AI-written while incorrectly labelling human-generated text as AI-written 9% of the time[31]. General purpose AI text detection systems are not yet sufficiently accurate to be truly useful.

Additionally, it is worth noting that there are common methods in use for distinguishing between human and artificial users (i.e., bots) when the originator of the text can be directly queried. These include CAPTCHAs[32], which often involve asking a user to recognize distorted letters and digits or identify a set of images that contain particular objects. However, such methods do not apply in offline settings, such as determining whether a submitted essay is AI generated or revised, and hence are not relevant to our study.

Our study makes a contribution by developing a practical tool to fill the gap of detecting AI-text in academic materials. By tailoring our detection models to high-stakes documents like LORs and SOIs, we aim to provide a reliable and effective approach that may better support integrity in academic evaluations.

Data and preprocessing

This section outlines the data and preprocessing steps employed in this study and describes the proprietary admissions dataset and open-access Wikipedia dataset utilized for training and validating the classification models. This study was approved by Fordham's Institutional Review Board, which waived the requirement for informed consent based on the details of the study. All procedures were carried out following relevant guidelines and regulations. To ensure compliance with the Family Educational Rights and Privacy Act (FERPA), the student and recommender identities were anonymized in our proprietary dataset by automatically redacting their names and removing affiliations.

Admissions dataset

The classification tasks described in this study require human-authored, AI-revised, and AI-generated LORs and SOIs. The human-authored data are sourced from an admissions dataset provided by Fordham University, as described in the "Human-authored instances" section. The AI-generated and AI-revised versions were generated using the GPT-3.5 Turbo API based on the corresponding application materials and human-authored documents, respectively.

Human-authored instances

The human-authored LORs and SOIs were extracted from the application packages of Fordham University's Master's programs in Computer Science and Data Science, both of which are administered by the Computer and Information Sciences Department. Approximately 96% of the applicants in our dataset are under age 30, with two-thirds of them being under age 24. Of these applicants, 36% are female. Additionally, 84% did not list English as their first language, which is partially due to the high number of international applicants. Each application contains one SOI and up to three LORs. Our data consist of 3755 LORs and 1973 SOIs. All applications were submitted prior to the release of ChatGPT and the widespread availability of LLMs.

The human-authored samples sometimes include anomalous text, such as nonsensical tokens and words, and may also contain extra spaces and lines. These may have been generated as part of the process of packaging the application materials. Since these may aid the models in distinguishing between human-authored and AI-crafted content, a thorough data-cleaning process was undertaken to remove these potential clues, resulting in a data set better equipped to evaluate the true capabilities of the AI detection method. Specifically, a program was developed to eliminate or substitute the tokens indicative of the text's origin and to standardize the document layout and formatting.

AI-generated instances

This section describes the process for creating the AI-generated LORs and SOIs. The first step involves automatically generating prompts like those presented in Table 1, by passing specific details (e.g., age, gender, undergraduate major, GPA, and work experience) from the student application packages to two Python scripts—one for generating SOI prompts and the other for generating LOR prompts. To reflect the diversity and variety

Write a statement of intent for a master's in Data Science at Fordham University. My undergrad is in Mathematics, my GPA is 3.45, and I know python, java, matlab, software, calculus, linear algebra
Assume you are applying for a graduate program in Data Science at Fordham University. Write a statement of intent telling a story that explains your reasons for pursuing this program, and how your undergraduate major in Computer Science, and knowledge (java, c++, database, software) have prepared you for success in this mater's program
Please write a recommendation letter for 931000356 who want to pursue his master degree in MSDS at Fordham University. Please describe 931000356 has consistently demonstrated his hard work, creativity, and dedication in his role, and our relationship is Work, The statement should have around 343 words
Please write a recommendation letter for 722000185 who desire a master degree in MSCS at Fordham University. Please describe his passion for machine learning, and performance and our relationship is Academic, The statement should have around 400 words

Table 1. Sample SOI and LOR prompts for generating AI instances.

typically found in these documents, the SOI script includes five basic prompt templates and the LOR script includes four basic prompt templates; the template matching the document type is selected randomly each time a new prompt is generated. The information from the application materials is used to fill in the templates. To further the amount of diversity in the prompts, most of the template components are included in the prompts with a probability selected randomly from a prespecified range (e.g., 0.2 to 0.8). The prompts for the generated LORs and SOIs also specify a desired length in words, which is set to the length of the corresponding human-authored document. This is done to ensure that document length does not serve as a clue as to the origin of the document. The code for these scripts is available along with the rest of the code for our experiments[33].

The generated prompts are then passed to the GPT-3.5 Turbo API with the temperature parameter set to 0.7 (same default value for ChatGPT), to generate corresponding versions of the human-authored SOIs and LORs. Names, titles, and locations are omitted from the AI-generated text to minimize the amount of sensitive information provided to our research assistants. Given the 1-to-1 correspondence between human- and AI-generated documents, the dataset is balanced for our classification task. Sample AI-generated documents corresponding to the prompts listed in Table 1 are provided in the yy file.

AI-revised instances

The GPT-3.5 Turbo API was utilized to revise, or polish, the human-authored SOIs and LORs. This was accomplished via the simple prompt "revise the following text," which was then followed by the full text of the SOI or LOR. The temperature parameter setting was set to 0.7 to be consistent with the default value used by ChatGPT. We observed that consecutive requests to revise a document tends to create two notably different versions; thus we generate two revisions for each human-authored document, which allows us to capture more of the diversity introduced by the AI. We address the resulting class imbalance by oversampling the human-authored instances when forming the dataset used for training and testing the models. Names, titles, and locations were redacted from all instances just as they were for the original human-authored LORs and SOIs. Sample AI-revised documents are not provided because they retain a substantial amount of details from human-authored versions, which would raise privacy concerns.

Open-access Wikipedia data

The GPT-wiki-intro dataset[9] is a specialized collection of text data curated to help with the detection of AI-generated content, especially content generated by OpenAI's GPT models. It focuses on the introductory sections of Wikipedia articles, which typically provide concise factual and informative overviews of the topic. The dataset contains actual Wikipedia introduction data and a matching GPT generated entry. As discussed earlier, the main purpose for this data is to assess the ability of the LOR and SOI trained detection model to identify AI content from a different domain. However, in some of our experiments we also incorporate the Wikipedia data into the training of the detection model to verify that this yields a substantial improvement when trying to detect AI content in the same domain.

Methods

This section outlines our research approach and the machine learning models/techniques employed in this study. Two traditional machine learning models, Naïve Bayes and Logistic Regression, are used as our baseline models. Their performance is compared against two state-of-the-art transformer-based models, BERT and DistilBERT, which account for the temporal dependencies in the text data. The code for these models are publicly available[33] and all four models can be executed on text input via the online tools described in the "Interactive detection tools" section.

Approach and experimental design

As discussed in the introduction, the AI-content detection problem is decomposed into two classification tasks: distinguishing between human-authored and AI-generated text and between human-authored and AI-revised text. The rationale for this separation is twofold. Firstly, certain universities may permit AI-revised application materials but prohibit AI-generated ones. This distinction arises from the consideration that AI-revised documents can be viewed as originating from humans, with AI tools helping to correct grammar mistakes or refine writing styles. Hence, it becomes necessary to differentiate between these two types of AI-crafted content. Secondly, while we could approach the problem using a multi-class classifier for the three document types, the approach is more challenging due to similarities between AI-generated and AI-revised text, leading to ambiguity in class boundaries. Furthermore, evaluating the performance of multi-class classifiers is less straightforward, as metrics like precision, recall, and F1-score are most effectively designed for binary classification models.

For each classification task, we conducted two experiments to evaluate the effectiveness of domain-specific models and their cross-domain generalizability. In the initial experiment (results detailed in Table 2), we exclusively trained the models using educational data (i.e., LORs and SOIs). The training and test datasets were created through a random 4:1 split. Model performance was assessed using five metrics: overall accuracy, recall, specificity, precision, and F-1 score. In addition to evaluating the models on the combined test data (i.e., SOI+LOR), we analyzed their performance on individual document types (LOR and SOI) and 12,000 cross-domain, balanced examples from the GPT-wiki-intro dataset.

In the second experiment, we replicated the same procedure but augmented the training data with a disjoint set of 48,000 balanced instances from the GPT-wiki-intro dataset. Models trained with this mixed-domain dataset (i.e., LORs+SOIs+Wiki data) showed substantial improvements on the Wiki dataset with minimal impact on the educational data. This outcome reinforces our hypothesis that developing an AI-content detector within a specific domain is feasible, even with limited data resources. The findings from the second experiment are presented in Table 3.

Test set	Category	Model	Accuracy (%)	F-1 (%)	Precision (%)	Recall (%)	Specificity (%)
SOI	AI-generated	LR	99.81	99.71	99.43	100.00	99.71
		NB	99.23	98.84	100.00	97.70	100.00
		BERT	100.00	100.00	100.00	100.00	100.00
		DistilBERT	100.00	100.00	100.00	100.00	100.00
	AI-revised	LR	96.06	96.04	97.60	94.53	97.63
		NB	85.36	86.95	79.06	96.58	73.92
		BERT	99.86	99.86	100.00	99.73	100.00
		DistilBERT	99.72	99.73	99.73	99.73	99.72
LOR	AI-generated	LR	99.93	99.93	100.00	99.87	100.00
		NB	99.87	99.87	99.74	100.00	99.74
		BERT	99.93	99.93	99.87	100.00	99.87
		DistilBERT	99.87	99.87	99.74	100.00	99.74
	AI-revised	LR	95.77	95.63	97.95	93.41	98.09
		NB	85.45	86.97	78.05	98.18	73.01
		BERT	99.67	99.66	99.86	99.46	99.87
		DistilBERT	99.07	99.08	98.75	99.40	98.73
SOI + LOR	AI-generated	LR	99.90	99.89	99.89	99.89	99.91
		NB	99.71	99.68	99.78	99.57	99.82
		BERT	99.90	99.89	99.79	100.00	99.82
		DistilBERT	99.95	99.95	99.89	100.00	99.91
	AI-revised	LR	95.87	95.76	97.84	93.78	97.94
		NB	85.42	86.96	78.38	97.65	73.30
		BERT	99.73	99.73	99.91	99.55	99.91
		DistilBERT	99.28	99.29	99.07	99.51	99.05
Wiki	AI-generated	LR	49.98	÷0	÷0	0.00	100.00
		NB	49.98	÷0	÷0	0.00	100.00
		BERT	51.42	6.99	82.33	3.65	99.22
		DistilBERT	53.05	11.61	99.46	6.16	99.97
	AI-revised	LR	49.34	0.36	68.75	0.18	99.92
		NB	54.15	33.66	63.20	22.94	86.26
		BERT	53.40	30.01	57.78	20.27	85.60
		DistilBERT	54.01	25.33	63.67	15.81	91.22
SOI + LOR + Wiki	AI-generated	LR	57.24	23.61	99.89	13.39	99.99
		NB	57.21	23.54	99.78	13.35	99.97
		BERT	58.47	28.25	95.99	16.56	99.32
		DistilBERT	59.86	31.55	99.69	18.74	99.94
	AI-revised	LR	61.93	40.03	97.62	25.17	99.37
		NB	62.61	53.65	71.63	42.89	82.70
		BERT	65.94	54.86	79.50	41.88	89.45
		DistilBERT	66.26	53.25	85.07	38.76	93.31

Table 2. Model performance trained exclusively on application (SOI and LOR) data.

Machine learning algorithms

This section briefly introduces the machine learning models utilized in this study. We selected BERT and DistilBERT for their broad adoption and proven reliability in NLP tasks. Although these models are older, they are sufficient for our task and provide both computational efficiency and accessibility. The same rationale applies to our choice of machine learning models. We used Naive Bayes (NB) and Logistic Regression (LR), both of which demonstrated their effectiveness in detecting AI-generated content.

Naïve Bayes

Naïve Bayes (NB)[10] is a probabilistic classification algorithm built upon Bayes' theorem and relies on the "naïve"assumption that the features $\{x_1, x_2, \dots, x_n\}$ are conditionally independent, given the class label *y*. Mathematically,

$$P(x_1, x_2, \dots, x_n|y) = \prod_{i=1}^{n} P(x_i|y)$$

While the above assumption may not hold in all real-world scenarios, NB often serves as a strong baseline for text classification tasks. NB uses Bayes' theorem to calculate the probability of each class given the observed features, and then predicts the class for unlabelled data with the highest probability $\hat{y}$, i.e.,

Test set	Category	Model	Accuracy (%)	F-1 (%)	Precision (%)	Recall (%)	Specificity (%)
SOI	AI-generated	LR	99.23	98.86	98.30	99.43	99.14
		NB	99.81	99.71	100.00	99.43	100.00
		BERT	100.00	100.00	100.00	100.00	100.00
		DistilBERT	99.62	99.43	98.86	100.00	99.42
	AI-revised	LR	94.61	94.48	96.25	92.78	96.43
		NB	83.98	85.75	76.87	96.94	71.15
		BERT	99.86	99.87	100.00	99.73	100.00
		DistilBERT	99.79	99.80	99.87	99.73	99.86
LOR	AI-generated	LR	98.95	98.95	98.30	99.60	98.30
		NB	99.21	99.21	98.44	100.00	98.43
		BERT	99.93	99.93	99.87	100.00	99.87
		DistilBERT	99.28	99.28	98.56	100.00	98.56
	AI-revised	LR	93.38	93.33	94.76	91.95	94.83
		NB	83.92	85.96	76.82	97.56	70.05
		BERT	99.27	99.27	98.62	99.93	98.60
		DistilBERT	99.80	99.80	99.80	99.80	99.80
SOI + LOR	AI-generated	LR	99.02	98.93	98.30	99.57	98.56
		NB	99.36	99.30	98.72	99.89	98.92
		BERT	99.95	99.95	99.89	100.00	99.91
		DistilBERT	99.36	99.31	98.62	100.00	98.83
	AI-revised	LR	93.78	93.70	95.24	92.21	95.35
		NB	83.94	85.89	76.84	97.36	70.41
		BERT	99.46	99.47	99.07	99.87	99.05
		DistilBERT	99.80	99.80	99.82	99.78	99.82
Wiki	AI-generated	LR	90.75	90.66	91.61	89.72	91.78
		NB	74.23	68.31	88.69	55.55	92.91
		BERT	99.32	99.32	99.48	99.15	99.48
		DistilBERT	97.70	97.73	96.34	99.17	96.23
	AI-revised	LR	89.80	89.80	91.09	88.55	91.09
		NB	76.46	73.50	85.62	64.39	88.88
		BERT	97.52	97.61	95.50	99.80	95.16
		DistilBERT	98.91	98.93	98.63	99.23	98.58
SOI + LOR + Wiki	AI-generated	LR	91.95	91.78	92.54	91.04	92.84
		NB	77.88	73.29	90.70	61.49	93.85
		BERT	99.44	99.44	99.58	99.29	99.59
		DistilBERT	97.96	97.96	96.66	99.29	96.65
	AI-revised	LR	90.88	90.85	92.20	89.53	92.25
		NB	78.48	77.49	82.26	73.25	83.84
		BERT	98.04	98.10	96.44	99.82	96.22
		DistilBERT	99.15	99.16	98.95	99.38	98.92

Table 3. Model performance trained on application and Wiki data.

$$\hat{y} = *arg\, max_i(P(y) \cdot P(x_i|y))$$

In the data preprocessing phase for the NB model, we applied Term Frequency-Inverse Document Frequency (TF-IDF)[34] to vectorize the text input. TF-IDF transforms raw text data into numerical features by considering two key factors: the frequency of a term within a document (Term Frequency) and its significance across the entire dataset (Inverse Document Frequency). This method allows the model to prioritize highly discriminative terms for classification tasks by capturing the relative importance of words while reducing the influence of common terms.

Logistic regression

Logistic Regression (LR[11]) is a widely used algorithm in machine learning and statistics. It uses the Sigmoid function, $\sigma(z) = \frac{1}{1+e^{-z}}$, to model the relationship between the input features and the probability of belonging to the positive class (class 1). The input z to the Sigmoid function is modeled as a linear combination of the independent variables, $\{x_1, x_2, \ldots, x_n\}$, i.e.,

$$z = w_0 + w_1 * x_1 + w_2 * x_2 + \cdots + w_n * x_n$$

where $\{w_0, w_1, w_2, \ldots, w_n\}$ are model parameters.

The LR model generates predictions for new data by computing the conditional probability associated with the positive class based on the observed input features (i.e., $P(y = 1|(x_1, x_2, \ldots, x_n))$. If this probability is greater or equal to a predetermined threshold (typically 0.5), the model classifies it as class 1; otherwise, it predicts class 0. Logistic Regression is valued for its simplicity and interpretability, but its assumption of a linear relationship between features and the log of the target variable is not valid in all cases. We applied the same TF-IDF technique that was used for Naïve Bayes to prepare the training data for the LR model.

BERT

Bidirectional Encoder Representations from Transformer (BERT)[12] is among the most notable pre-trained language models in the NLP domain. This model's innovation lies in its ability to capture the bidirectional context of words in a sentence, enabling it to comprehend the intricacies of language, including nuances, word meanings, and context. As a result, BERT surpasses unidirectional models such as RNN or LSTM in a wide range of NLP tasks, including sentiment analysis, question answering, language translation, and text summarization.

BERT's architecture is built upon the Transformer model[35], which introduced the concept of self-attention mechanisms. These mechanisms enable BERT to assign varying levels of importance to different words in a sentence, facilitating the extraction of essential information and context. BERT's pre-training involves two key tasks: masked language modeling and next sentence prediction. In the former, BERT learns to predict missing words in a sentence, forcing it to understand the relationships between words in context. In the latter, BERT learns to determine whether a pair of sentences logically follows one another, enhancing its grasp of document-level context.

One of BERT's unique features is its ability to be fine-tuned for specific NLP tasks with relatively small amounts of task-specific data. This adaptability has made BERT the go-to choice for researchers and developers in various applications[36–39]. In this study, we fine-tuned the pre-trained BERT-base-uncased model with a 55% dropout rate for the final layer. This dropout level was selected empirically.

DistilBERT

DistilBERT[13] is a variant of the BERT model[35], designed to be more compact and computationally efficient while retaining comparable performance. DistilBERT is built on the same transformer architecture as BERT, which uses a stack of transformer encoder layers to process and encode input text data. The output is subsequently used for various downstream NLP tasks, such as text classification, sentiment analysis, and named entity recognition.

The main innovation in DistilBERT is the use of knowledge distillation, which involves training a smaller "distilled" model to mimic the behavior of a larger, pre-trained model. DistilBERT is trained to mimic BERT's outputs and achieves its compactness and efficiency by reducing the number of parameters compared to BERT. It typically uses 40% fewer parameters, which makes it faster to train and classify examples, but yet it retains much of BERT's performance. This makes it a preferred option for situations with constrained computational resources, such as deploying NLP models in environments with limited resources. Like the BERT model, we trained our DistilBERT model with a 55% dropout rate applied to the final hidden layer.

Results

This section presents the performance of AI-content detection models for our two classification tasks. We begin with the results of models trained on educational data, followed by outcomes for models trained on cross-domain data (SOI, LOR, and Wiki). In all experiments, the AI samples constitute the positive class.

Models trained on SOIs and LORs

Table 2 presents the performance of models trained with educational data. The results show that all four models demonstrate nearly perfect accuracy when evaluated on AI-generated SOI and/or LOR test data. The precision, recall, and F-1 scores further affirm their efficacy in distinguishing between AI-generated and human-authored academic documents.

For the AI-revised documents, both BERT and DistilBERT models exhibit comparable performance, achieving over 99% accuracy across both academic document types. However, the performance of the non-temporal models (LR and NB) declines when compared to their performance with the AI-generated cases. Specifically, LR achieves 96.06% overall accuracy in classifying SOIs and 95.77% in classifying LORs, whereas its performance on the corresponding AI-generated data is 99.81% and 99.93%, respectively. The NB model shows a similar trend with a more pronounced performance drop (85.36% vs. 99.23% and 85.45% vs. 99.87%) for the AI-revised category. Similar conclusions can be drawn for the F-1 and Precision metrics. Notably, all models maintain high recall across both document types. Therefore, the differences in performance are primarily attributed to misclassifications of human-authored documents.

All of the models, however, exhibit poor performance when assessed on the Wiki data, with accuracy values that are close to random guessing. The LR and NB models predict all AI-generated instances as human-authored (i.e., recall of 0), while BERT and DistilBERT achieve only slightly higher recall (3.65% and 6.16%, respectively). Model performance on the AI-revised Wiki data are similarly very poor so that the detection models lack practical value. Evaluating these models on all three forms of data (SOI+LOR+Wiki) yields better results due to their strong performance on SOIs and LORs.

The large performance gap indicates that the models are well-tuned for the academic documents that appeared in the training set, but are not suitable as general-purpose AI-content detectors. To further validate our hypothesis that effective classifiers can be trained with domain-specific data, we proceeded to train our models using a combined data set that included the SOIs, LORs, and the Wiki data.

Models trained on mixed domain data

The results for the models trained using the LOR, SOI, and Wiki data are presented in Table 3. When evaluated on the SOI and/or LOR data, the models built using the extended training set exhibit a marginal reduction in accuracy across both document types in comparison to the results detailed in Table 2, which were exclusively trained on SOI and LOR data. Specifically, a calculation of the average model accuracy over the AI-generated rows for SOI, LOR, and SOI+LOR blocks (i.e., twelve rows in Table 2) shows an average of 99.84% compared to 99.48% in the corresponding rows of Table 3. A similar calculation indicates the average accuracy across these blocks for the AI-revised category as 95.11% versus 94.30%.

This decline is expected due to the complexity and diversity introduced by the Wiki samples into the training data. What is noteworthy is the substantial improvement in the model performance for the Wiki dataset. The average accuracy of the four models over the Wiki block increased from 51.11% to 90.50% for the AI-generated data and from 52.73% to 90.67% for the AI-revised data. Furthermore, the transformer-based models (BERT and DistilBERT) exhibited a significant advantage over non-temporal models (LR and NB). This experiment suggests that AI-content detectors lack generalizability across domains. However, with sufficient training samples from each domain, effective domain-specific classifiers can be developed.

Analysis and comparison of linguistic characteristics

In this section, a variety of linguistic characteristics are analyzed and compared between text generated by AI, revised by AI, and authored by humans.

Total vocabulary and paragraph structure comparison

In this section we explore differences in the vocabulary size and the structure of the paragraphs produced by GPT-3.5 Turbo and humans. Table 4 provides a summary of these statistics. The first observation is that the AI-generated and AI-revised documents utilize a much smaller total vocabulary than the human-authored documents. Specifically, the AI-generated LORs (SOIs) use a vocabulary of 4909 (5593) words, while the human-authored LORs (SOIs) use 36,105 (36,641) words. The AI-revised documents show a similar but less extreme pattern, as the AI-revised LORs (SOIs) use a vocabulary of 18, 477 (18, 702) words. Thus, the total vocabularies for the human-authored LORs and SOIs are roughly six to seven times the size of the vocabularies for the corresponding AI-generated documents and twice the size of vocabularies used in the AI-revised documents. The smaller difference in vocabulary size for the AI-revised documents suggests that those documents retain many of the words from the original (human-authored) versions. This seems plausible as one would expect that a document created by AI from a short prompt would exhibit more of the AI's language characteristics than a revision to a complete, human-authored, document.

The second observation is the extreme discrepancy in exclusive word usage between AI-crafted and human-authored content. Table 4 shows that the AI-generated LORs (SOIs) have only 464 (391) words absent from their human-authored counterparts, while the human-authored LORs (SOIs) have 31,660 (30,439) words that do not appear in any of their AI-generated counterparts. This shows that the AI-generated documents not only use a much smaller total vocabulary, but the number of distinctive words is even smaller than would otherwise be expected, indicating the repeated use of favored words. The "exclusive words" results for the AI-revised documents parallel those for the AI-generated documents, but, as with the total vocabulary results, the pattern is much less extreme. These results again suggest that the modifications made by the AI text revision process is less extreme than for the AI text generation process.

Table 4 also presents a few statistics related to paragraph structure and shows that the AI-generated text has more, but shorter, paragraphs than the corresponding human-authored counterparts, possibly resulting in a more fragmented overall structure. This trend also holds for the AI-revised documents but is much less pronounced than the AI-generated documents. A careful examination of these statistics also reveals that for both types of AI-crafted documents, these differences are much more extreme for the LORs than the SOIs. The *p*-values in Table 4 demonstrate that all of the structural differences are statistically significant with a very high degree of confidence except those for the AI-revised SOIs.

From these statistics we conclude that the AI text generation process yields more, but shorter, paragraphs for both the LORs and SOIs. However, this trend is notably reduced for the AI-revised documents, indicating that

		AI-generated			AI-revised		
		AI	Human	*p*-val	AI	Human	*p*-val
LOR	Total vocabulary	4909	36,105		18,477	36,105	
	Exclusive words	464	31,660		1559	19,187	
	Avg (sentences/paragraph)	2.78	4.92	$< 10^{-5}$	4.12	4.92	$< 10^{-5}$
	Avg (# paragraphs)	4.87	2.56	$< 10^{-5}$	3.98	2.56	$< 10^{-5}$
SOI	Total vocabulary	5593	35,641		18,702	35,641	
	Exclusive words	391	30,439		1565	18,504	
	Avg (sentences/paragraph)	3.94	4.36	$< 10^{-5}$	4.32	4.36	0.420
	Avg (# paragraphs)	6.01	5.44	$< 10^{-5}$	5.75	5.44	0.002

Table 4. Vocabulary and paragraph statistics.

the revised text retains more of the characteristics of the human-authored documents. The reason for the more pronounced differences in the LORs is unclear. One potential explanation could be the observed differences between the LOR and SOI documents. Specifically, LORs tend to be more formal and structured, often with an opening paragraph explaining the relationship to the applicant and a closing paragraph summarizing the rationale and strength of the recommendation. In contrast, SOIs tend to be highly personal and involve the applicant discussing their life's journey or family history. These differences may make it easier to segment the LORs into shorter paragraphs.

Word frequency comparison

In order to gain a deeper and more detailed understanding of the differences in vocabulary between AI-crafted and human-authored text, we next focus on the words with the largest frequency differences between the two types of documents. Table 5 displays the top 15 words preferred by AI yet used infrequently by humans, and conversely, those preferred by humans but less so by AI. The "AI" and "Human" columns denote the total occurrences of the word in the respective AI and human documents. The word frequency statistics are calculated separately for LORs and SOIs and for the AI-generated and AI-revised documents. The "Ratio" column is calculated as "AI" / "Human" for the "AI-Preferred Words" and "Human" / "AI" for the "Human-Preferred Words." To avoid an undefined or infinite ratio, a "0" in the denominator is replaced by a "1."

Table 5 reveals a notable difference in repetition patterns between AI-preferred and human-preferred words. Twenty AI-preferred words are used at least 500 times, whereas only one human-authored word ("get") reaches such frequency. GPT-3.5 seems to heavily favor words like "additionally," "wholeheartedly," "unwavering," and "emphasis."

There are also key stylistic differences. The AI-crafted documents feature a formal and sophisticated vocabulary while the human-authored documents feature a simple and colloquial vocabulary. The human-preferred words include many simple words like "got," "get," "lot,", and "don't." If one scans down the two word columns it is clear that the AI-preferred words are on average longer. The human-preferred words are clearly more colloquial as they include 10 total, and 5 distinct, contractions (e.g., "I'm", "don't"), while there is not a single contraction among the AI-preferred words (note that formal writing generally avoids contractions). The AI-preferred words also include many highly descriptive adjectives, such as "unwavering," "insatiable,", "unparalleled," and "transformative," while such adjectives are almost totally lacking for the human-preferred words. Notably, the human-preferred list includes one abbreviation ("CS"), one misspelled word ("programing"), and many possessive words (e.g., people's), all of which are totally absent in the AI-preferred word list. These observations seem to hold equally for the LORs and SOIs, with no apparent differences in favored words.

Interactive detection tools

The detection models described and evaluated in this article are packaged as publicly accessible, easy to use tools. There are two variants: one tool for distinguishing between human-authored and AI-generated text[40] and a second for distinguishing between human-authored and AI-revised text[41]. Since the models employed by these tools were trained using LOR and SOI data, they will perform best in detecting these types of text.

As illustrated in Fig. 1, the interface allows users to input sample text and choose from the four models developed in this study. After executing a selected model on the provided text, the interface will display the prediction along with its corresponding confidence level. The "Model Explanation" button offers access to the Local Interpretable Model-Agnostic Explanations (LIME)[42] and Shapley Additive Explanations (SHAP[43]) values, which provide a deep understanding of the factors influencing the decisions made by the models.

Conclusion

This study evaluated the use of traditional classification models and transformer-based LLMs for distinguishing between AI-generated and human-authored content and between AI-revised and human-authored content. These results highlight the promise of specialized classification models tailored to specific application domains. The results illustrate the limitations of general AI content detectors and the potential of classification models tailored for specific application domains. Our findings further indicate that detecting AI-revised content is more challenging than detecting AI-generated content. Paragraph and word statistics suggest some concrete reasons for this observation, as AI-revised content tends to share more characteristics with human-generated content, such as a broader total vocabulary. Another observation is that advanced transformer-based temporal models generally outperform their non-temporal counterparts. However, these differences are marginal when evaluating the models withing the same training domain; thus simple models can be effective in these situations.

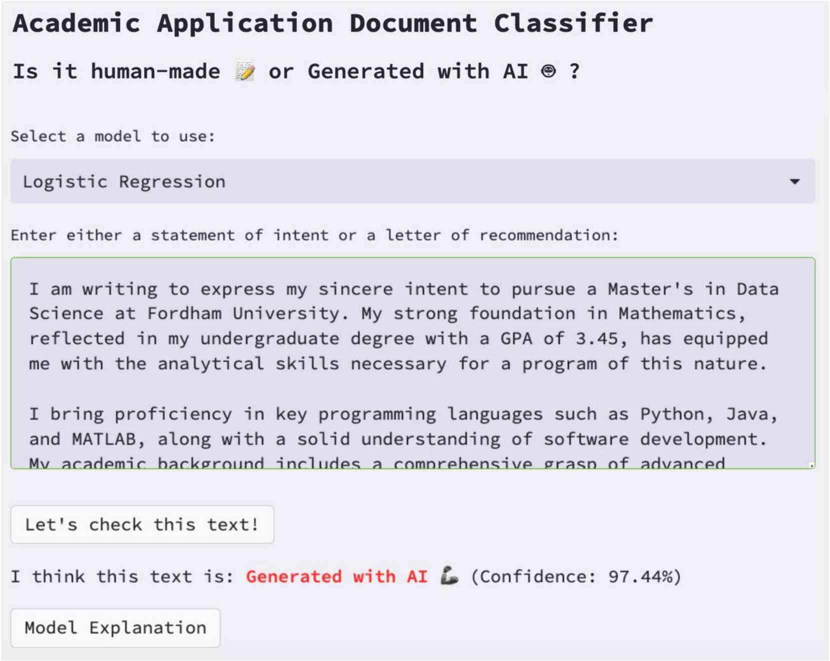
The word and paragraph analysis revealed distinct patterns in the language and structure employed by GPT 3.5. In particular, although AI-crafted documents exhibit natural and intelligent language, they utilize a notably narrower, more supplicated, repetitive, and formal vocabulary compared to human-written content - a style contrasts with the more diverse and colloquial vocabulary often favored by humans. Additionally, the paragraph analysis suggests that AI-generated content frequently features more but shorter paragraphs compared to human writing. A closer examination of the AI-generated documents further uncovers a significant overuse of adverbs by AI, contrasting with humans' preference for a simpler subject-predicate structure.

The findings of this research have important practical implications for higher education institutions. In the competitive world of academia, where institutions strive to identify the most deserving candidates, AI detection tools like the ones described in this study can be used to safeguard the fairness and integrity of the admissions process. This will ensure that the applicants are evaluated on their true merits and abilities, thereby fostering a more equitable and trustworthy admissions process. In practice, admissions officers or program directors will need to view the materials flagged as AI-generated or AI-revised with caution, potentially applying penalties if

Category	AI-preferred words				Human-preferred words			
	Word	AI	Human	Ratio	Word	AI	Human	Ratio
AI-Generated LORs	Fostering	461	2	154	Got	0	498	498
	Witnessing	1206	8	134	Get	1	455	455
	Fosters	114	0	114	Lot	0	374	374
	Unwavering	1878	16	110	Although	1	372	372
	Showcasing	549	5	92	Homework	1	319	319
	Advancements	748	9	75	Really	1	308	308
	Palpable	59	0	59	Gave	0	295	295
	Nontechnical	525	11	44	Though	1	280	280
	Showcases	304	6	43	I'm	0	271	271
	Prowess	565	13	40	Associate	0	248	248
	Hackathons	155	3	39	Reference	0	239	239
	Representations	114	2	38	Man	0	229	229
	Insatiable	378	9	38	Quite	2	453	227
	Unparalleled	414	11	35	Times	1	222	222
	Unyielding	68	1	34	Started	1	220	220
AI-revised LORs	Showcasing	561	5	94	Months	1	276	276
	Young	78	0	78	I'm	0	271	271
	Surpasses	46	0	46	University's	0	196	196
	Inquiries	501	11	42	September	0	144	144
	Showcased	651	19	33	Don't	0	143	143
	Self	1128	35	31	Weeks	1	129	129
	Fostering	80	2	27	Company's	0	95	95
	Wholeheartedly	2244	86	26	He's	0	94	94
	Willingly	288	11	24	Didn't	0	93	93
	Noting	205	8	23	June	0	88	88
	Unwavering	384	16	23	Cannot	0	86	86
	Surpassing	102	4	20	January	0	73	73
	Additionally	2589	129	20	Bachelors	1	68	68
	Provoking	138	6	20	Learnt	0	64	64
	Recipient	173	8	19	What's	0	56	56
AI-generated SOIs	Young	146	0	146	get	1	693	693
	Responsibly	41	0	41	etc	0	386	386
	Vibrant	948	22	43	Later	1	299	299
	Collaborations	209	5	42	CS	0	277	277
	Aligns	1266	39	32	Semester	1	236	236
	Transformative	257	9	29	Fall	0	216	216
	Partnerships	149	6	25	Three	2	409	205
	Collaborate	1139	46	25	Months	0	202	202
	Emphasis	1489	67	22	Graduated	1	199	199
	Fostering	172	8	22	Going	0	194	194
	Fosters	104	5	21	Lot	2	380	190
	Collaboratively	40	2	20	Five	0	181	181
	Meaningfully	119	6	20	Called	0	180	180
	Ethical	593	31	19	Paper	1	173	173
	Evolving	807	46	18	Interesting	1	169	169
Continued								

Category	AI-preferred words				Human-preferred words			
	Word	AI	Human	Ratio	Word	AI	Human	Ratio
AI-Revised SOIs	Young	162	0	162	I'm	0	355	355
	Aligning	56	1	28	University's	0	229	229
	Surpassing	25	0	25	Bachelor's	0	226	226
	Minded	72	2	24	Months	0	202	202
	Ran	23	0	23	Company's	0	177	177
	Solidifying	40	1	20	Learnt	0	161	161
	Showcasing	39	1	20	Today	0	157	157
	Revised	112	5	19	People's	0	134	134
	Fueling	52	2	17	Programing	0	128	128
	Rounded	130	7	16	Today's	0	126	126
	Fueled	308	18	16	Ago	1	97	97
	Self	648	40	16	Didn't	0	97	97
	Aligns	622	39	16	Cannot	0	91	91
	Noteworthy	15	0	15	What's	0	75	75
	Unwavering	75	4	15	Don't	0	75	75

Table 5. AI vs. human word frequency comparison for 15 most common words.



Academic Application Document Classifier

Is it human-made 📝 or Generated with AI 🤖 ?

Select a model to use:

Logistic Regression

Enter either a statement of intent or a letter of recommendation:

```
I am writing to express my sincere intent to pursue a Master's in Data
Science at Fordham University. My strong foundation in Mathematics,
reflected in my undergraduate degree with a GPA of 3.45, has equipped
me with the analytical skills necessary for a program of this nature.

I bring proficiency in key programming languages such as Python, Java,
and MATLAB, along with a solid understanding of software development.
My academic background includes a comprehensive grasp of advanced
```

I think this text is: Generated with AI 🦾 (Confidence: 97.44%)

Fig. 1. Interactive web interface for detecting AI-generated LORs or SOIs.

tool usage is prohibited. The utility of such tools hinges on their high accuracy, and our findings indicate that this is possible.

This study has several limitations. Firstly, using domain-specific models requires training data for each domain, which may not always be readily available. Secondly, prompt design is critical in shaping AI outputs. Our approach simulates typical real-world scenarios where users interact with AI tools using minimal prompt optimization, focusing on common user behavior rather than pushing the limits of prompt engineering. However, as studies like ours continue to identify distinctive traits of AI-generated content (e.g., sophisticated vocabulary, longer sentences, etc.), users may refine their prompts to thwart detection models. Like any evolving technology, these models require continuous fine-tuning to adapt to changing practices. Lastly, the experiments are limited to AI content generated by GPT-3.5. It is thus unclear how a detection model trained using GPT-3.5 generated data will perform on text generated by other LLMs or other versions of GPT. We plan to address these limitations in our future work. Similar to challenges faced by other detection problems (e.g., virus detection that must account for different operating systems), the detection of AI content demands ongoing maintenance as new

LLM's are introduced and new versions are released. Nevertheless, the findings of this study offer compelling evidence that domain-specific AI content detectors can achieve high accuracy and offer practical value.

Data availability

The data used in the current study is not publicly available due to its proprietary nature and compliance requirements with the Family Educational Rights and Privacy Act (FERPA). Researchers interested in accessing the data, please contact Dr. Yijun Zhao or Dr. Gary Weiss at Fordham University to discuss the possibility of gaining access through protocols such as a data usage agreement.

Received: 28 January 2024; Accepted: 25 October 2024
Published online: 02 November 2024

References

1. OpenAI. Chatgpt. https://chat.openai.com (2021). Accessed on September 3, 2023.
2. Westerlund, M. The emergence of deepfake technology: A review. *Technol. Innov. Manag. Rev.* **9** (2019).
3. Zellers, R. et al. Defending against neural fake news. *Adv. Neural Inf, Process. Syst.* **32** (2019).
4. Ragot, M., Martin, N. & Cojean, S. AI-generated vs. human artworks. a perception bias towards artificial intelligence? In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–10 (2020).
5. Pegoraro, A., Kumari, K., Fereidooni, H. & Sadeghi, A.-R. To chatgpt, or not to chatgpt: That is the question! arXiv preprint arXiv:2304.01487 (2023).
6. Turnitin. Turnitin's AI writing detection available now. https://www.turnitin.com/solutions/ai-writing (2023).
7. Guo, B. et al. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597 (2023).
8. Bleumink, A. G. & Shikhule, A. Keeping AI honest in education: Identifying gpt-generated text. *Edukado AI Res.*, 1–5 (2023).
9. Aaditya Bhat. Gpt-wiki-intro (revision 0e458f5) (2023).
10. Berrar, D. Bayes' theorem and naive bayes classifier. *Encycl. Bioinform. Comput. Biol. ABC Bioinform.* **403**, 412 (2018).
11. Kleinbaum, D. G., Klein, M., Kleinbaum, D. G. & Klein, M. Intro. to logistic regression. *Logistic Regression: A Self-learning Text* 1–39 (2010).
12. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
13. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:1910.01108 (2020).
14. Perkins, M. Academic integrity considerations of AI large language models in the post-pandemic era: Chatgpt and beyond. *J. Univ. Teach. Learn. Pract.* **20**, 07 (2023).
15. Dwivedi, Y. K. et al. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* **71**, 102642 (2023).
16. Lo, C. K. What is the impact of chatgpt on education? a rapid review of the literature. *Educ. Sci.* **13**, 410 (2023).
17. Johnson, A. Chatgpt in schools: Here's where it's banned-and how it could potentially help students (2023).
18. Grassini, S. Shaping the future of education: exploring the potential and consequences of AI and chatgpt in educational settings. *Educ. Sci.* **13**, 692 (2023).
19. Zhai, X. Chatgpt user experience: Implications for education. *Available at SSRN 4312418* (2022).
20. Halaweh, M. Chatgpt in education: Strategies for responsible implementation. *Contemp. Educ. Technol.* **15**(2), ep421 (2023).
21. Crothers, E., Japkowicz, N. & Viktor, H. L. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access* (2023).
22. Fröhling, L. & Zubiaga, A. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Comput. Sci.* **7**, e443 (2021).
23. Nguyen-Son, H.-Q., Tieu, N.-D. T., Nguyen, H. H., Yamagishi, J. & Zen, I. E. Identifying computer-generated text using statistical analysis. In *2017 Asia-Pacific Signal and Info. Processing Association Annual Summit and Conference*, 1504–1511 (IEEE, 2017).
24. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D. & Finn, C. Detectgpt: Zero-shot machine-generated text detection using probability curvature. arXiv preprint arXiv:2301.11305 (2023).
25. Solaiman, I. et al. Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203 (2019).
26. Liu, Y. et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019).
27. Rodriguez, J., Hay, T., Gros, D., Shamsi, Z. & Srinivasan, R. Cross-domain detection of gpt-2-generated technical text. In *Proc. 2022 Conf. North American Chapter of the Assoc. for Comp. Linguistics: Human Language Technologies*, 1213–1233 (2022).
28. Tourille, J., Sow, B. & Popescu, A. Automatic detection of bot-generated tweets. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, 44–51 (2022).
29. Stiff, H. & Johansson, F. Detecting computer-generated disinformation. *Int. J. Data Sci. Anal.* **13**, 363–383 (2022).
30. Wiggers, K. Most sites claiming to catch AI-written text fail spectacularly. *TechCrunch*. https://techcrunch.com/2023/02/16/most-sites-claiming-to-catch-ai-written-text-fail-spectacularly (2023).
31. Kirchner, J. H., Ahmad, L., Aaronson, S. & Leike, J. New AI classifier for indicating AI-written text. *OpenAI* (2023).
32. Von Ahn, L., Blum, M., Hopper, N. J. & Langford, J. Captcha: Using hard AI problems for security. In *Advances in Cryptology-EUROCRYPT 2003: International Conference on the Theory and Applications of Cryptographic Techniques, Warsaw, Poland, May 4–8, 2003 Proceedings 22*, 294–311 (Springer, 2003).
33. Lopez, F. M. AI admissions detector code. https://github.com/Fordham-EDM-Lab/AI-Admissions-Detector (2024).
34. Leskovec, J., Rajaraman, A. & Ullman, J. D. *Importance of Words in Documents (page 9)* (Cambridge University Press, 2022).
35. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding (2019). arXiv:1810.04805.
36. Rahman, M. W. U. et al. A bert-based deep learning approach for reputation analysis in social media. In *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, 1–8 (IEEE, 2022).
37. Jahan, M. S., Beddiar, D. R., Oussalah, M., Arhab, N. & Bounab, Y. Hate and offensive language detection using bert for English subtask a. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13–17, 2021* (RWTH Aachen University, 2021).
38. Nayak, A., Timmapathini, H., Ponnalagu, K. & Venkoparao, V. G. Domain adaptation challenges of bert in tokenization and sub-word representations of out-of-vocabulary words. In *Proc. 1st Workshop on Insights from Negative Results in NLP*, 1–5 (2020).
39. Geetha, M. & Renuka, D. K. Improving the performance of aspect based sentiment analysis using fine-tuned bert base uncased model. *Int. J. Intell. Netw.* **2**, 64–69 (2021).
40. Martinez, F. AI-generation admissions detector tool. https://huggingface.co/spaces/ferdmartin/GradApplicationDocsApp (2024).
41. Borelli, A. AI-revision admissions detector tool. https://huggingface.co/spaces/alexborelli/GradApplicationDocsApp (2024).

42. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should i trust you?" explaining the predictions of any classifier. In *Proc. of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144 (2016).
43. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30** (2017).

Author contributions

Conceptualization, Y.Z. and G.W.; methodology, Y.Z. and G.W.; software, A.B., F.M., H.X.; validation, Y.Z., A.B., F.M., H.X., and G.W.; formal analysis, Y.Z., A.B., F.M., H.X., and G.W.; investigation, Y.Z., A.B., F.M., H.X., and G.W.; resources, Y.Z., and G.W; data curation, A.B., F.M., H.X.; writing-original draft preparation, Y.Z. and G.W.; writing-review and editing, Y.Z. and G.W.; supervision, Y.Z. and G.W.; project administration, Y.Z. and G.W.; all authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-77847-z.

Correspondence and requests for materials should be addressed to Y.Z. or G.M.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.