Plagiarism Detection Using Keystroke Logs

Scott Crossley
Vanderbilt University
Nashville, Tennessee USA
scott.crossley@vanderbilt.edu

Yu Tian
Digital Harbor Foundation
Baltimore, Marylan USA
terry@levi.digitalharbor.org

Joon Suh Choi
Vanderbilt University
Nashville, Tennessee USA
joon.suh.choi@vanderbilt.edu

Langdon Holmes
Vanderbilt University
Nashville, Tennessee USA
langdon.holmes@vanderbilt.edu

Wesley Morris
Vanderbilt University
Nashville, Tennessee USA
wesley.g.morris@vanderbilt.edu

ABSTRACT

This study examines the potential to use keystroke logs to examine differences between authentic writing and transcribed essay writing. Transcribed writing produced within writing platforms where copy and paste functions are disabled indicates that students are likely copying texts from the internet or from generative artificial intelligence (AI) models. Transcribed texts should differ from authentic texts where writers follow a process that includes monitoring, evaluating, and revising texts. This study develops a transcription detection model by using keystroke logs within a machine learning model to predict whether an essay is authentic or transcribed. Results indicated that keystroke logs accurately predicted whether an essay was written or transcribed with 99% accuracy using a random forest model. Authentic writing included a greater number of pauses before sentences and words, had a greater number of insertions and longer insertions, deleted more words and characters, and had a greater number of revisions than transcribed writing. Transcribers, on the other hand, produced a greater number of writing bursts because they were simply copying language. Overall, the results indicated that authentic writing is a dynamic process where writers monitor their writing and evaluate whether the writing needs to be changed if problems are identified. Transcribed writing, on the other hand is much more linear. The results may have important implications for plagiarism detection.

Keywords

Plagiarism detection, Keystroke logging, AI detection, Student writing

1. INTRODUCTION

The advent of generative artificial intelligence (AI) has led to concerns in the educational community about the ability to assess student performance through writing. Fears that students may demonstrate performance in the absence of learning by using AI have been brought to the foreground with the release of generative AI tools like ChatGPT. Concerns that students would submit texts, ideas, and arguments that were not their own are not new. However, prior to generative AI, plagiarism was relatively easy to combat. For instance, many schools and universities subscribed to plagiarism detection tools like TurnItIn (www.turnitin.com). These plagiarism detection tools search the internet and large collections of writings submitted by other students for similarities between those texts and a student's writing, effectively detecting copied text [15]. The approach worked well because there were known databases of writing that the tools could search, and they could reliably return simple matches based on strings of words that were similar. Teachers, of course, still had to make judgments about whether the writing was plagiarized based on attributions, amount of writing that was flagged, and other factors.

Generative AI upended this approach to detecting plagiarism because tools like ChatGPT can produce writing samples that are indistinguishable from human writing [29]. These samples contain unique ideas and strings of words that do not occur in extant databases, making detection challenging if not impossible to do reliably [3]. The difficulty in detecting whether a writing sample is produced by generative AI raised fears that student plagiarism would increase exponentially. As a result, teachers and administrators have had to rethink how writing is taught, the use of AI writing assistance, and student expectations.

The advent of generative AI also required researchers and industrial applications to reconsider how to detect AI-produced writing with a flood of recent research focusing on using AI or stylometric approaches to detect whether a writing sample is produced by AI [9, 22]. While these attempts have shown some degree of success, there are two problems. First, the accuracy of the new models for AI detection are too low to be useful in practice because of the likelihood of producing a false positive (i.e., detecting a writing sample is plagiarized when it is not). Second, as generative AI models proliferate, mature, and learn to mimic a diversity of writing styles, detection accuracies have decreased [29].

Thus, teachers and administrators are left with few choices in terms of AI detection. The easiest solution is to embrace generative AI and support students in using the technology to enhance their ideas and their writing. After all, AI can assist students with planning, completing, and revising their ideas, which are all important components of the writing process [12]. A second solution is to refocus writing education on the writing process, not the written product [11]. Assessing writing as a process related to idea development,

S. Crossley, Y. Tian, J. S. Choi, L. Holmes, and W. Morris. Plagiarism detection using keystroke logs. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 476–483, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729864

outlining, writing, and revision emphasizes the behaviors and cognitive activities that lead to successful text generation, not the end product itself.

While the focus on writing products over writing process will remain the norm for the foreseeable future, the collection and analyses of writing process data has the added benefit of aiding plagiarism detection. Such an approach will be applicable in controlled writing settings like software systems that control applications and content and restrict users' ability to move information into the system from. Such virtual environments, in which students may be isolated from the wider internet and are not afforded the opportunity to copy and paste ideas from elsewhere, are often called walled gardens. While walled gardens are convenient and allow students to experiment and develop ideas [2], they do not preclude students from transcribing essays retrieved from the internet or produced by generative AI outside the system and passing those essays off as their own work within the system, thus circumventing attempts to minimize plagiarism.

It is within this context that the current study takes place. We recognize that AI detection models that focus on the product alone will not be successful at identifying or combatting plagiarism, especially as AI evolves. We also recognize that reverting to wet-ink assessment will not be easy, especially in a world where digital literacy skills are normalized and where students are expected to use digital resources to help them write successfully. Thus, we see an important space for more controlled digital writing environments like walled gardens that monitor student writing and control for copying and pasting, while allowing students to produce essays using word processors. Controlled digital writing spaces allow for tighter control of the writing process and help ensure that the information produced by the student reflects that student's writing ability. However, as mentioned earlier, these spaces suffer from an important limitation: they cannot detect when a user transcribes text from outside the system into the system.

Our solution to this problem is the use of keystroke logs to detect transcription. Thus, our approach is not based on the written product but on the writing process. Previous work in non-academic writing has indicated that keystroke logs represent authentic writing processes much differently than transcribed writing processes [8] and we expect that these findings will transfer to writing tasks common in a school setting. A transcription detection model could be used to inform teachers and administrators about whether students within a controlled writing system produced an essay using authentic writing processes as compared to potentially transcribing the output of a generative AI system.

1.1 AI Detection

AI detectors provide a quantitative assessment of how likely a document was generated by AI by focusing on the written product. Quantitative models are needed because it is difficult to scale human assessment of AI written text and because human raters are relatively poor at identifying AI generated text. Research indicates that humans perform slightly better than random guessing on early iterations of GPT (~60% with essays generated by GPT-2) but performed at around 50% with later iterations (i.e., GPT-3) [6].

AI detectors do not focus on human generated texts because a poorly written text is almost always produced by a human, but a well-written text is not always written by AI [29]. There are two general approaches to AI detection. The first approach uses AI to detect AI-written texts using large language models (LLMs; e.g., using BERT to predict whether a text was written by a human or an AI). The second approach is a stylometric approach that examines how generative AI differs from humans in terms of linguistic production such as word frequencies, grammar, n-gram occurrences, or syntactic structures [29]. AI detectors rely on machine learning models to scale up detection approaches, but such approaches, while promising in many cases, are not reliable enough for implementation [10, 26].

In the case of using LLMs, Walters [27] found that open-source LLMs like GPT-2 and RoBERTa were able to predict a range of AI writings with accuracies between 60-94% when the writing was produced by GPT-3.5. These numbers were reduced when the writing was produced by GPT-4. In another study, Yan et al. [29] found that fine-tuned LLMs could detect AI writing with accuracies that exceeded 99%, but the dataset used was much more specific than that reported by Walters [27]. Linguistic features can also be used to classify AI-generated texts. For instance, Yan et al. [29] found that AI-generated text contained no spelling or grammar errors and that a model based on linguistic features could detect AI-generated text with an accuracy of 95%. Al Afnan and Mohd Zuki [1] found that academic texts generated by GPT-4 predictably contained declarative sentences that used the active voice and demonstrated a diverse use of present tense indicators. Additionally, academic texts generated by GPT-4 contained high lexical density (high proportion of nouns and lexical verbs) and low lexical diversity (low variety of word types).

Mikros et al. [17] combined LLM and linguistic approaches in an AI detection model and achieved a detection accuracy of 95% using an ensemble method that included ELECTRA and RoBERTa transformer models and linguistic features related to sentiment, type-token ratios, and text readability. Beyond academic research, there are also proprietary plagiarism detection tools like TurnItIn and ZeroGPT. Walters [27] found that these tools reported prediction accuracies between 91-97% (for TurnItIn) and 31-100% (for ZeroGPT).

1.2 The Writing Process

Current AI detection models focus on the written product (i.e., the output of a student or a generative AI). However, writing produced by humans is the result of a complex process that includes planning, generating ideas, initial writing, revision, and final editing. Studying the writing process can help us understand writers' behaviors and cognitive activities while generating text. These activities can be used to assess writer focus, the revision process, content development, and, as we argue, whether a text is written in an authentic process or transcribed.

Systematic research on the cognitive processes involved in text production gained momentum when Hayes and Flower [12] produced a model of the writing process, which was updated in Chenoweth and Hayes [4]. The updated model distinguishes between four subprocesses: the Proposer, the Translator, the Transcriber, and the Evaluator/Reviser. The proposer creates the so-called idea package to be formulated in language. The translator converts the message in the idea package into word strings by resorting to linguistic processes that include lexical access and syntactic frame construction. The transcriber is responsible for turning the word strings into written texts through motor skills. Finally, the evaluator/reviser monitors and evaluates both proposed language that has been transcribed or not yet been transcribed and makes changes when any problems and inadequacies are identified.

These subprocesses of text production are assumed to form a recursive pattern in which the different processes not only occur more or

less simultaneously but also interact with each other during text production. The subprocesses provide unique traces of the writing process in the form of pauses, movements, insertions, and deletions [20], which can be captured using keystroke logging measures. Keystroke logs can be mapped to specific components of the writing process to make inferences about the cognitive demands and processes in text production [4. 25].

Conijn et al. [8] examined the potential for keystroke logs to distinguish between copied texts and free-form texts found in the Villani keystroke dataset [21]. The copied texts were all a single, short fable (652 characters) and the free-form texts were short e-mails of at least 650 characters. The texts were collected from 36 participants, who submitted 338 copied texts and 416 free-form texts. Conijn et al. calculated variables related to pause times, corrections, and word length from the Villani dataset. Using these feature groups within a machine learning model, they were able to classify copied from free-form texts with an accuracy of 78% with the strongest indicator being pause times. Specifically, they found that in free-form texts, pauses before a word were longer and pauses within or after a word were shorter.

In similar work, Trezise et al. [23] found that features such as bursts and pauses could be used to differentiate the writing processes between free writing of non-argumentative essays, self-transcriptions for these essays, and general transcriptions. Trezise et al. analyzed essays written by 62 participants who self-transcribed their essay after writing it and transcribed an essay they did not write (general transcription). They found that free-writing was associated with fewer words typed per minute, shorter writing bursts that included fewer keystrokes, longer pauses, and faster word deletion rates that were more variable.

1.3 Current Study

Previous work has shown that keystroke logs are a potential solution for distinguishing authentic writing from transcribed writing. However, no studies have been conducted on argumentative essay writing. Thus, the current study examines the potential to use keystroke logs to examine differences between authentic and transcribed argumentative essay writing. We hypothesize that the cognitive demands and processes found in authentic essay writing, where students are expected to follow the subprocesses reported in the cognitive models of writing [4], will lead to a unique keystroke profile. This profile can be extracted using machine learning techniques. If a reliable profile of the writing processes found in authentic essay writing can be developed, it could be used to detect when students are transcribing external texts within a controlled writing setting under the presumption that transcribing processes are used in cases of plagiarism. To develop a transcription detection model for essays, we use machine learning approaches to predict whether an essay is authentic or transcribed using keystroke logging variables related to production fluency, pauses, bursts, revisions, and process variances. Our work differs from Conijn et al. [8] and Trezise et al. [23] in the writing genre (argumentative essays), the keystroke logs collected, and the nature of the data (no repeated measures)

2. METHOD

We collected essays and keystroke logs from crowd-sourced workers in two different time periods. Initially, we collected authentic essays from 4,992 participants. From these authentic essays, we selected 500 essays that were highly scored. We then had crowd-sourced workers transcribe these essays while also collecting their keystroke logs.

2.1 Authentic Data Collection

2.1.1 Participants

Participants were hired from Amazon Mechanical Turk (MTurk), an online crowdsourcing platform that enables large-scale participant recruitment in a time-efficient manner. Data was collected prior to November, 2022, when ChatGPT was released. Participants were selected based on the following criteria: 1) be at least 18 years old; 2) be currently living in the United States; 3) have completed at least 50 MTurk tasks with an overall approval rate of at least 98% by requesters on the platform.

2.1.2 Procedure

Participants hired from MTurk were invited to log onto a website built specially for this study. The website housed a demographic survey, a series of typing tests, an argumentative writing task, and a vocabulary knowledge test. Participants were required to use only computers with a keyboard to participate in the study. Their keystroke activities during the typing tests and the argumentative writing task were recorded using a built-in keystroke logging program that captured every keystroke and mouse operation entered by the participants with its timestamp and cursor position information. Participants' role in the study lasted 40-50 minutes. Their participation was compensated by a $0.25 reward and a $11.75 bonus upon completing all the tasks following the instructions.

2.1.3 Argumentative Writing

In the argumentative writing task, participants were asked to write an argumentative essay within 30 minutes in response to a writing prompt adapted from a retired Scholastic Assessment Test (SAT) taken by high school students attempting to enter post-secondary institutions in the United States. To control for potential prompt effects, four SAT-based writing prompts were used, and each participant was randomly assigned one prompt. Prior to the writing task, participants were given instructions about how to write an argumentative essay and suggested that participants should write an essay of at least 200 words in 3 paragraphs and that they should not use any online or offline reference materials.

2.1.4 Apparatus

To collect participants' keystroke information during the typing tests and the argumentative writing task, a keystroke logging program was written in JavaScript and was embedded in the script of the website built for this study. The program unobtrusively recorded every keystroke operation and mouse activity along with relevant timing and cursor position information when participants completed typing tasks and wrote their essays. It also simultaneously identified operation types (e.g., input, delete, paste, replace) and reported text changes in the writing process. Table 1 provides an example output of keystroke logging information reported by the program. *Event ID* indexes the keyboard and mouse operations in chronological order. *Down Time* denotes the time (in milliseconds) when a key or the mouse was pressed while *Up Time* indicates the release time of the event. *Action Time* represents the duration of the operation (i.e., *Up Time - Down Time*). *Position* registers cursor position information to help keep track of the location of the leading edge. *Pause Time* demonstrates the time intervals between consecutive events (i.e., IKIs). *Word Count* displays the accumulated numbers of words typed in. Additionally, *Text Change* shows the exact changes made to the current text while *Activity* indicates the nature of the changes (e.g., Input, Remove/Cut).

Unlike keystroke logging software widely used in writing research (e.g., ScriptLog, Inputlog), this web-based program can be easily deployed in any browser-based writing environment and thus

avoids inconveniences associated with installing or updating any software beyond a standard web browser. It is also highly scalable and well-suited for online keystroke information collection using crowdsourcing. In terms of privacy concerns, the logging capacity of this program is confined to the input fields on the webpages for the typing tests and the argumentative writing task, and hence does not track any other information as participants operate on their computers throughout the experiment.

Event ID	Down Time	Up Time	Action Time	Event	Position	Pause Time	Word Count	Text Change	Activity
1	746	791	45	Leftclick	0	0	0	NoChange	Nonproduction
2	19948	19948	0	Shift	0	19202	0	NoChange	Nonproduction
3	20210	20321	111	I	1	262	1	I	Input
4	20369	20441	72	Space	2	159	1	Space	Input
5	20417	20497	80	b	3	48	2	b	Input
6	20601	20730	129	e	4	184	2	e	Input
7	20658	20809	151	l	5	57	2	l	Input
8	20769	20921	152	i	6	111	2	i	Input
9	20849	21001	152	e	7	80	2	e	Input
10	20937	21097	160	v	8	88	2	v	Input
11	21034	21185	151	Space	9	97	2	Space	Input
12	21105	21257	152	Backspace	8	71	2	Space	Remove/Cut
13	21209	21321	112	e	9	104	2	e	Input
14	21369	21497	128	Space	10	160	2	Space	Input
15	21425	21506	81	t	11	56	3	t	Input
16	21441	21513	72	h	12	16	3	h	Input
17	21585	21745	160	a	13	144	3	a	Input
18	21633	21753	120	t	14	48	3	t	Input
19	21689	21825	136	Space	15	56	3	Space	Input

Figure 1. Example Keystroke Logging Information

2.1.5 Essay Collection and Scoring

We collected 4,992 acceptable essays from the crowd-sourced workers. To ensure the essays were of high quality and, thus, similar to essays that would be produced by generative AI, the essays were scored by trained raters for overall writing quality using a holistic, six-point grading scale commonly used in assessing essays written for college admission in the United States (i.e., the SAT scoring rubric). The SAT rubric evaluates writing quality on multiple dimensions, including test-takers' development of a point of view on the issue, evidence of critical thinking, use of appropriate examples, accurate and adapt use of language, the variety of sentence structures, errors in grammar and mechanics as well as text organization and coherence. Essays were randomly assigned among six raters and each essay was scored by two raters. The raters were graduate students majoring either in English or applied linguistics. All of them had at least two years of experience teaching English composition at the university level. All raters went through at least three rounds of training sessions before they scored the essays independently. The training started off with a briefing about the essay collection methods and the holistic rubric, followed by a discussion of potential biases in essay scoring. For each session, raters were asked to score a batch of essays on their own before they met to discuss the differences in their scores. A total of 60 practice essays were used for training purposes. These practice essays were on the same topics but were sampled from a different dataset. The raters exited the training sessions and started independently scoring the essays only after a satisfying agreement had been reached for the practice essays with a Cohen's Kappa of at least .600 [7]. After scoring the entire data, the Cohen's Kappa = 0.759, $p < .001$, suggesting substantial overall agreement between the raters. If score differences between two raters were two points or greater, the raters adjudicated the scores through discussion. If agreement was not reached, the score was not changed. Average scores between the raters for the adjudicated holistic scores were calculated for each essay.

We selected 500 essays from the larger corpus of 4,992 essays that were of higher quality and thus would better reflect essays produced by generative AI. All essays were scored by at least one rater as 4 (indicating an above average essay). Overall, the essays reported scores of M = 3.94, SD = .40. The essays were written on four prompts that we generally distributed equally. The mean length of the essays was 364.852 (SD = 112.309). On average, the essay contained 4.44 paragraphs (SD = 1.471).

2.2 Transcribed Data Collection

2.2.1 Participants

Participants were hired from Amazon Mechanical Turk (MTurk). Participants were required to meet three threshold qualifications: 1) be at least 18 years old; 2) be currently living in the United States; 3) have an overall approval rate of at least 95% by requesters on the platform.

2.2.2 Procedure

Participants hired from MTurk were invited to log onto a website built specially for this study. The website housed a demographic survey and a transcribing task. Participants were required to use desktop computers with a keyboard to participate in the study. Their keystroke activities during the typing tests and the argumentative writing task were recorded using the same built-in keystroke logging program as in the argumentative essay collection.

For the transcribing task, participants were randomly assigned an essay from the 500 higher quality essays. The assigned essay was presented on the transcribing task page in tandem with a transcribing area. Prior to transcribing, participants were instructed to transcribe the authentic text verbatim following a set of rules. These rules included 1) avoiding the use of any intelligent grammar and spell checkers and 2) refraining from irrelevant activities during transcription. Figure 1 below shows the transcription screen presented to the participants.

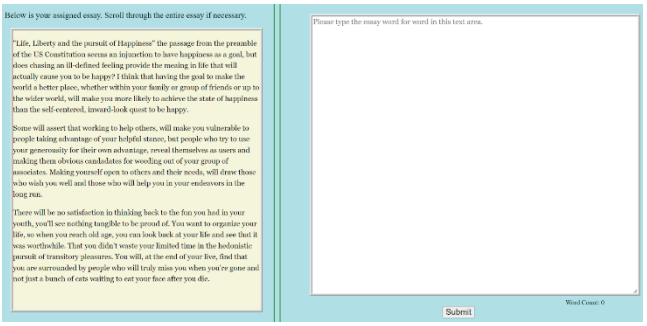

Figure 2. Example Transcription Screen

We conducted checks to ensure the accuracy of the data. First, we examined the keystroke data and removed flawed logs (e.g., logs with record of copying and pasting the original essay or Irrelevant activities during transcribing). Second, we calculated the string similarity between the transcribed essay and the authentic essay using the levenshteinSim function (a similarity function derived using the Levenshtein distance) from the RecordLinkage R package [19]. In order for the transcription to be accepted, a string similarity threshold of .95 and .90 were set. All essays scoring above the .95 threshold were accepted without further inspection. Essays that scored above the .90 threshold but below .95 were inspected manually for missing sentences. If multiple sentences were found to be missing, the data was rejected. These thresholds ensured that the transcribed essay accurately reflected the original. Out of the 63 transcribed essays within this threshold range, five essays were missing more than three sentences from the original essay and were not accepted into the final batch. These essays were collected from

different Mechanical Turk workers so that we had transcribed data for all 500 essays.

2.3 Keystroke Logging Measures

The keystroke logging files were transformed into IDFX files that could be read and analyzed via Inputlog 7.0 [24]. A set of keystroke indices with reference to the participants' production fluency, bursts, pausing behaviors, revision activities, and process variances were generated from the keystroke logs collected from the authentic writers and the transcribed writers. In total, 155 keystroke logging measures were developed.

2.3.1 Production Fluency

We analyzed participants' production fluency during writing or transcribing using Inputlog's summary analysis. The reported measures included means, medians, and standard deviations for the number of different linguistic units (i.e., characters, words, sentences, paragraphs) produced per minute in text production. Note that these measures were calculated based on the writing/transcribing process rather than the final product and thus included any deleted text.

2.3.2 Bursts

Measures of P-bursts (bursts that ended with a pause longer than two or more seconds) were obtained by drawing on the summary analysis provided by Inputlog. In this study, P-bursts were operationalized as continuous text production episodes terminated at pauses of 2 or more seconds, following the rationale that pauses exceeding 2 seconds generally corresponding to higher-level thinking processes such as activities related to planning, and thus eliminating the interruptions caused by transcription-related activities or inefficiencies in motor execution [28]. To obtain measures of R-bursts (bursts that ended with grammatical discontinuities such as revisions), revision analysis was performed to calculate the numbers and lengths of bursts delimited by revision activities.

2.3.3 Pauses

We generated two sets of pause measures using Inputlog's pause analysis based on two different thresholds: 200 milliseconds and 2 seconds. We adopted these two thresholds because the former has the strength of capturing the bulk of language- and planning-related differences in pausing and filtering most inter-key intervals that result solely from the motor constraint of typing [16], while the latter generally reflects higher order cognitive processes, such as planning for new ideas or revising [14]. We also analyzed pauses at different locations (e.g., within words, between words, between sentences) because they are associated with various patterns of pausing behaviors and might provide insights into different underlying cognitive processes in writing [5].

2.3.4 Revisions

Revisions were analyzed by drawing on the revision matrix reported by Inputlog. The revision matrix provides a detailed list of revision events in sequential order. Inputlog distinguishes between two types of revision events: 1) deletions, i.e., characters deleted in the text produced so far, and 2) insertions, which represents the characters inserted into the earlier text. In this study, we calculated the total numbers of deletions and insertions and also developed a set of measures regarding the length of deletions and insertions in characters and words.

2.3.5 Process Variances

To take into account the variability of each participant's entire writing process, variances in process fluency were investigated using Inputlog's fluency analysis which reports production rates (characters produced per minute) at different stages in the writing process. In this study, we analyzed process variances with the number of intervals set at 5 or 10 respectively (i.e., the entire writing time for each essay was divided evenly into 5 or 10 segments). We adopted the standard deviation values reported by Inputlog as a measure of process variance to describe how dispersed the production rate for each interval in relation to the mean value.

2.4 Statistical Analysis

A series of machine learning models were used to predict a binary outcome variable (authentic or transcribed) using the keystroke logging measures. We used four machine learning models: linear discriminant analysis, multi-layer perceptron, random forest, and support vector machine. Prior to running each model, we first removed keystroke logging features that reported NA counts (n = 57). We then divided the data into training and test. The training set comprised 70% of the data and was stratified by condition (authentic or transcribed). The test set comprised the remaining 30% of the data. We then calculated bivariate Pearson correlations using the cor.test() function to identify highly collinear features in the training set. If two or more variables correlated at $r > .899$, the keystroke logging variable(s) with the lowest correlation with a dummy coded outcome variable was removed. This removed an additional 33 keystroke logging variables leaving us with a final number of 65 keystroke logging variables for analysis. These variables were scaled (z-score normalized) prior to analyses using R's scale() function.

We used the CARET package [13] in R to develop machine learning models using the final 65 features. The machine learning models we used were linear discriminant analysis, multilayer perceptron, random forest, and support vector machine. Model training and evaluation were performed using the training and test sets. Within the training process, we used a control parameter that used 10-fold cross-validation for resampling. The model from the training set was then applied to the left-out test set. Estimates of accuracy are reported using overall accuracy, precision, recall, F1, and Kappa values. Variable importance was derived from the varImp() function. The variable importance scores indicate the strength of the contributions that each variable makes in model predictions.

3. RESULTS

3.1 Model Evaluation

Using the 65 variables as features, all models outperformed a baseline model of 50% (see Table 2). The results also indicate that the random forest models performed the best reporting an overall accuracy of 99%. This was followed by the multi-layer perceptron reporting a 98% accuracy and accuracies of 97% and 96% for the support vector machine model and linear discriminant analysis model respectively. Table 3 presents a confusion matrix for the random forest model, which predicted two transcribed essays as authentic and one authentic essay as transcribed. Variable importance for the top 20 variables from the random forest model are reported in Table 4. Table 4 also includes the means and standard deviations for the keystroke logging variables for the authentic and transcribed essays. The variable importance measures indicate that the most predictive variables are related to pause times, insertions, product/process ratio, and deletions. The mean values indicate that authentic essays included longer pauses, more insertions, lower product process ratios, and more deletions.

Table 1. Overall Classification Precision, Recall, F1-score

Model	Precision	Recall	F1	Kappa
LDA	0.953	0.966	0.960	0.920
MLP	0.980	0.974	0.976	0.953
RF	0.987	0.993	0.990	0.980
SVM	0.973	0.967	0.970	0.940

LDA= Linear Discriminant Analysis, MLP = Multi-Layer Perceptron, RF = Random Forest, SVM = Support Vector Machine

Table 2. Confusion matrix for random forest model

		Actual	
		Authentic	**Transcribed**
Predicted	**Authentic**	148	2
	Transcribed	1	149

3.2 Post-hoc Evaluation

Three essays were inaccurately categorized: one transcribed essay was predicted to be authentic and two authentic essays were predicted to be transcribed. We scanned the keystroke log mean responses to better understand what features may have led to the inaccurate predictions. For the transcribed essay that was predicted as authentic, we found that the transcriber had much longer pause lengths and more insertions than the mean scores for transcribed essays. This could indicate that the transcriber was not carefully attending to the task. For the authentic essays that were predicted as transcribed, the writers had fewer insertions (0 and 3) than other authentic writers (M = 344). The writers also had longer pause times, and a lower number of deletions (one writer's data reported 0 deletion versus a mean deletion count of ~130 for authentic writers). An inspection of the content of the two essays indicates that they were on topic and well written. The essays suffered from no grammatical or spelling errors, which could suggest they were written by AI (data collection occurred when GPT-3 was available but before the release of GPT 3.5). It could also be the case that these writers were thoughtful in planning and proposing ideas, leading to longer pause rates. These same writers appear to be highly adept at translating these ideas onto the page leading to few or no deletions and insertions.

4. DISCUSSION AND CONCLUSION

Our goal in this study was to examine the potential for keystroke logs to predict whether an essay was written in an authentic context or transcribed. If such an approach is reliable, a transcription detection model could be integrated into a controlled learning environment like a walled garden to detect when students are transcribing text from outside the garden. Such text is likely to be plagiarized and, considering the availability and reliability of large language models, generated by AI.

To assess differences in keystroke logs between authentic and transcribed texts, we collected essays and keystroke logs from crowdsourced workers in two different time periods. Initially, we collected authentic essays from 4,992 participants. From these authentic essays, we selected 500 essays that were highly scored and thus likely to mimic essays produced by generative AI. We had crowd-sourced workers transcribe these essays while collecting keystroke logs. Our final corpus included 1,000 essays of which half were authentic and half were transcribed. We ran a series of machine learning algorithms using the keystroke logs to predict the classification of the essays.

Table 3. Variable importance, means, and standard deviations for top 20 keystroke features

Variable	Var Imp	Auth Mean (SD)'	Trans Mean (SD)
Pause Time in Secs (T=200, M)	100.000	1.646 (0.783)	0.75 (0.522)
Total Insertions Chars Exclu Space	86.418	307.96 (344.11)	11.006 (73.899)
Total Pause Time in Secs (T=2000)	86.114	1008.714 (311.298)	508.487 (1219.806)
Product Process Ratio	84.304	0.824 (0.111)	0.954 (0.029)
Insertion Length Chars Exclu Space (M)	84.250	8.857 (6.878)	0.977 (3.121)
Total Deletions Words	76.665	115.218 (130.943)	17.034 (17.781)
Pause Time Before Sents (T=200, M)	72.989	12.682 (29.433)	3.895 (16.521)
Deletion Length Chars (M)	62.925	4.355 (3.926)	1.538 (2.83)
Num Of Insertions	58.656	32.95 (36.005)	4.18 (13.007)
Insertion Length Chars Exclu Space (MD)	54.425	4.741 (4.544)	0.665 (2.438)
Num Of Pause Within Words (T=200)	46.291	408.688 (245.747)	764.758 (414.673)
Strokes Per Min 5 Intervals (SD)	41.477	39.496 (21.289)	19.875 (21.367)
Length Rburst Sec (MD)	39.319	6.27 (6.431)	12.21 (14.837)
Pause Between Words (T=200, MD)	37.009	0.731 (0.284)	1.042 (0.491)
Num Of Pause After Words (T=200)	35.664	114.542 (75.965)	213.722 (127.284)
Num Of Revisions	34.798	125.95 (85.112)	52.284 (33.025)
Pause Time bf Words (T=200, SD)	34.136	4.28 (5.35)	2.806 (7.803)
Total Num of Pauses (T=2000)	30.805	79.002 (33.364)	66.796 (71.594)
Pause Time bf Words (T=200, MD)	29.008	0.431 (0.152)	0.566 (0.307)
Pause Time bf Words (T=200, M)	27.500	1.398 (0.825)	0.986 (0.736)

M = mean, MD = median, Num = number, T = threshold, bf = before, sec = seconds, sents = sentence, Exclu = excluding

The results indicated that keystroke logs could accurately predict whether an essay was written in an authentic environment with 99% accuracy using a random forest model. Our other machine learning models produced accuracy results varying from 96% to 98% accuracy with the lowest accuracy reported for a linear discriminant analysis model. Overall, the results indicate that keystroke logs can be used to develop a transcription detection model.

The variable importance metrics from the random forest can be used to interpret which keystroke logs are the most important in the prediction tasks. These metrics indicated that writing in an authentic context showed a greater number of pauses (mean and total) in

general and before sentences and words. In contrast, writers who were transcribing an essay showed a greater number of pauses within words and after words. This is similar to what was reported in Conijn et al. [8] and indicates that authentic writing requires pauses as writers make links between ideas that are proposed and translate those ideas into words and structures. Transcribers, on the other hand seem to process text in the middle or at the end of words, likely because of constraints on working memory. Writers within an authentic context also made a greater number of insertions and longer insertions than transcribers. Authentic writers also deleted more words and characters than transcribers and had a greater number of revisions, similar to what was reported by Trezise et al. [23]. In total, a greater number of insertions, deletions, and revisions likely represent the process by which authentic writers monitor their writing and evaluate whether the writing needs to be changed if problems are identified. Lastly, authentic writers showed a greater deviation in the number of keystrokes they produced, indicating that authentic writing is not a linear process, but one that is much more dynamic that transcribed writing.

Writers that transcribed an essay had a greater product process ratio indicating that the linguistic units (i.e., characters, words, sentences, paragraphs) they produced during text production and those in the product are more similar. This result is likely because the goal of a transcriber is to only produce words that have already been written. There is no process in terms of idea generation, translating ideas into language, or monitoring and revising the output. Transcribers also had longer lengths of writing bursts than authentic writers likely because they were only copying what they had already seen written.

While a preliminary study, the results indicate that authentic and transcribed writing can be strongly predicted based on keystroke logs. Unlike other potential approaches to AI detection, keystroke logging could provide stronger accuracy and may not suffer from known faults. For instance, if a student copies an essay generated by ChatGPT 3.5 or ChatGPT 4, the transcription detection model should still be able to detect that the text is transcribed irrespective to any language characteristics of existing or future AI. This is because the models do not depend on the written product but on the writing process. While generative AI models may improve and become more difficult to detect based on the product alone, the process of authentic writing should remain stable.

The models developed here could be used within a controlled writing space like a walled garden to detect potential plagiarism resulting from copying AI generated text. The models could also be integrated into digital writing platforms like Google Docs or a Microsoft Word plugin with incorporated keystroke logs. Instructions could be given to students to not copy and paste or transcribe writing from another text processor, and the keystroke logging systems could detect if they did copy and paste, alerting teachers. For those that do not copy and paste, the keystroke logging system could be used to detect authentic and copied writing and provide feedback to teachers and students about potential plagiarism concerns.

Beyond plagiarism detection, integrating keystroke logging into walled garden systems or digital writing platforms would afford additional benefits. One of these benefits could be process-based evaluations and feedback to accommodate the writers' real-time needs in text production. For instance, keystroke logging may help these systems and platforms to better detect writers' irregular writing behaviors such as protracted pauses (likely indicating mind-wanderings or difficulties in idea generation) and frequent revisions (signaling struggles in idea formulation or spelling depending on the types of the revisions). Informed by keystroke information, walled garden systems or digital writing platforms may provide useful hints or suggestions in a timely manner to help writers better navigate difficult stages in the writing process.

Our next steps are to integrate the models reported in this paper within a system. Specifically, we plan to incorporate the models into the Intelligent Texts for Enhanced Lifelong Learning framework [18]. The framework produces intelligent texts from any static or digital texts. iTELL texts are monitored, collect telemetry data, including keystroke logs, and disable copying and pasting. At the end of each chapter, students are asked to write a summary of what they have read to demonstrate reading comprehension. The plagiarism detection model will be used within the summary feedback mechanisms to alert both teachers and students to potential plagiarism. We also plan to collect additional authentic and transcribed writing samples from a variety of tasks, prompts, domains, and topics to ensure that the differences between authentic writing and transcribed writing are generalizable. This will prove important because most writing is not linear, like the writing analyzed here, but is rather dynamic with writing happening across time, across spaces, and across platforms.

5. ACKNOWLEDGMENTS

This work was partially funded by the National Science Foundation (Award Number 2112532). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

6. REFERENCES

[1] Al Afnan, M. A., & Mohd Zuki, S. F. (2023). Do artificial intelligence chatbots have a writing style? An investigation into the stylistic features of ChatGPT-4. *Journal of Artificial intelligence and technology*, *3*(3), 85-94.

[2] Catena, E. P., Monea, B., Skeuse, M., Kulkarni, A., & Stornaiuolo, A. (2022). Online Writing Spaces as" Walled Gardens" in English Language Arts Classrooms. English Journal, 112(1), 71-79.

[3] Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning and Teaching, 6(2), Article 2. https://doi.org/10.37074/jalt.2023.6.2.12

[4] Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written communication*, *20*(1), 99-118.

[5] Chukharev-Khudilaynen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research*, 6(1), 61.

[6] 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* (Volume 1: Long Papers), 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

[7] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

[8] Conijn, R., Roeser, J., & Van Zaanen, M. (2019). Understanding the keystroke log: the effect of writing task on keystroke features. *Reading and Writing*, *32*(9), 2353-2374.

[9] Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About detecting deepfake tweets. *Plos One*, 16(5).

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

[11] Graham, S., & Sandmel, K. (2011). The Process Writing Approach: A Meta-analysis. The Journal of Educational Research, 104(6), 396–407. https://doi.org/10.1080/00220671.2010.488703

[12] Hayes, J. R., & Flower, L. (1981). *Uncovering cognitive processes in writing: An introduction to protocol analysis*. ERIC Clearinghouse.

[13] Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, *28*, 1-26.

[14] Limpo, T., & Alvès, R. A. (2017). Tailoring Multicomponent Writing Interventions: Effects of Coupling Self-Regulation and Transcription Training. *Journal of Learning Disabilities*, 51(4), 381-398.

[15] Lyon, C., Barrett, R., & Malcolm, J. (2006). Plagiarism is easy, but also easy to detect. *Plagiary*.

[16] Medimorec, S., & Risko, E. F. (2016). Effects of disfluency in writing. *British Journal of Psychology*, *107*(4), 625-650.

[17] Mikros, G., Koursaris, A., Bilianos, D., & Markopoulos, G. (2023). AI-Writing Detection Using an Ensemble of Transformers and Stylometric Features.

[18] Morris, W., Crossley, S., Holmes, L., Ou, Chaohua, Dascalu, M., & McNamara, D. (in press). Formative Feedback on Student-Authored Summaries in Intelligent Textbooks using Large Language Models. *Journal of Artificial Intelligence in Education*.

[19] Sariyar, M., & Borg, A. (2010). The RecordLinkage Package: Detecting Errors in Data. R J., 2(2), 61.

[20] Spelman Miller, K, Lindgren, E., & Sullivan, K. P. (2008). The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly*, 42(3), 433-454.

[21] Tappert, C. C., Villani, M., & Cha, S.-H. (2009). Keystroke biometric identification and authentication on long-text input. *Behavioral biometrics for human identification: Intelligent applications*, 342-367.

[22] Tay, Y., Bahri, D., Zheng, C., Brunk, C., Metzler, D., & Tomkins, A. (2020). Reverse engineering configurations of neural text generation models. arXiv preprint arXiv: 2004. 06201. https://doi. org/10.48550/arXiv.2004.06201

[23] Trezise, K., Ryan, T., de Barba, P., & Kennedy, G. (2019). Detecting academic misconduct using learning analytics. *Journal of Learning Analytics*, *6*(3), 90-104.

[24] Van Waes, L., & Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition*, *38*, 79-95.

[25] Van Waes, L., Van Weijen, D., & Leijten, M. (2014). Learning to write in an online writing center: The effect of learning styles on the writing process. *Computers & Education*, 73, 60-71.

[26] Varshney, L. R., Keskar, N. S., & Socher, R. (2020, February). Limits of detecting text generated by large-scale language models. In *2020 Information Theory and Applications Workshop (ITA)* (pp. 1-5). IEEE.

[27] Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, *7*(1), 20220158.

[28] Wengelin, Å., Torrance, M., Holmqvist, K., Simpson, S., Galbraith, D., Johansson, V., & Johansson, R. (2009). Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior research methods*, *41*, 337-351.

[29] Yan, D., Fauss, M., Hao, J., & Cui, W. (2023). Detection of AI-generated essays in writing assessment. *Psychological Testing and Assessment Modeling*, *65*(2), 125-144.