Dec 11th, 12:00 AM

Human or AI? Using Digital Behavior to Verify Essay Authorship

David Wilson
Brigham Young University, davidwilsonphd@gmail.com

Parker Burnett
Brigham Young University, keer42@gmail.com

Joseph S. Valacich
University of Arizona, valacich@arizona.edu

Jeff Jenkins
BYU, jeffrey_jenkins@byu.edu

Human or AI? Using Digital Behavior to Verify Essay Authorship

Completed Research Paper

David Wilson
Brigham Young University
davidwilson@byu.edu

Parker Burnett
Brigham Young University
parker.burnett@byu.edu

Joseph S. Valacich
University of Arizona
valacich@arizona.edu

Jeffrey L. Jenkins
Brigham Young University
jjenkins@byu.edu

Abstract

Large language models (LLMs) such as OpenAI's GPT-4 have transformed natural language processing with their ability to understand context and generate human-like text. This has led to considerable debate, especially in the education sector, where LLMs can enhance learning but also pose challenges to academic integrity. Detecting AI-generated content (AIGC) is difficult, as existing methods struggle to keep pace with advancements in generation technology. This research proposes a novel approach to AIGC detection in short essays, using digital behavior capture and follow-up questioning to verify text authorship. We executed a controlled experiment as an initial evaluation to test the prototype system. The results obtained show promise in differentiating between user-authored and AI-generated text. The system design and prototype represent valuable contributions for future research in this area. The solution also provides a novel approach to addressing practical challenges posed by LLMs, particularly in maintaining academic integrity in educational settings.

Keywords: LLMs, ChatGPT, deception, keyboard dynamics, mouse tracking, design science

Introduction

Large language models (LLMs) have recently burst onto the scene as a significant development in the field of natural language processing. These systems—most notably OpenAI's GPT-3.5 and GPT-4 (Brown et al., 2020; OpenAI, 2023)—apply recent innovations in deep learning to very large training datasets, resulting in artificial intelligence (AI) tools with an unprecedented ability to understand context and generate text that is easy to mistake for human (Brown et al., 2020). As such, modern LLMs have triggered an explosion of academic and societal discourse, not least because of the accessibility afforded by the free (for now) web-based interface, ChatGPT.

Although the capabilities of LLMs have sparked significant debate and speculation across many domains (Bommasani et al., 2021), education has emerged as a primary point of focus. Many have noted the potential for innovations in the education sector, with LLMs improving learning with individualized feedback at scale (Malik et al., 2019) or enriching the learning experience for distributed education platforms (Li & Xing, 2021). However, the ability for LLMs to generate accurate responses to several classic forms of assessment (e.g., essays, quizzes, exams) has created significant upheaval, with some predicting "the death of the short-form essay" (Yeadon et al., 2022) or even describing the potential for academic dishonesty among college students in terms of "abject terror" (Mitchell, 2022, para. 18).

Part of the angst among educators results from the challenge of distinguishing AI-generated content (hereafter, AIGC) from the human-generated content it is designed to mimic. With objective quality of AIGC in education settings often surpassing that of human submissions (Herbold et al., 2023), students have a strong incentive to submit AIGC as their own. Research has shown that state-of-the-art methods for

detecting AIGC are not nearly effective enough (Sadasivan et al., 2023) to stem the tide of academic misconduct that some view as inevitable (Floridi, 2020). The lack of reliable methods for detecting AIGC weighs even heavier in the context of short-form essays because they constitute a powerful learning tool that encourages active learning (Hamilton, 1989; Ling & Libby, 2010). As such, essays and other forms of writing have been a staple of education assessment for decades (Nesi & Gardner, 2012; Tynjälä, 1998).

The objective of this research is to develop an alternative approach to AIGC detection. Existing detection solutions have notable limitations and are locked in a "technological arms race" as generation and detection models trade advancements in a persistent effort to both detect and evade detection of AIGC (Cotton et al., 2023; Eaton, 2023). The system proposed in this research approaches the detection problem from a fundamentally different angle. Building on the strength of recent research in the domains of digital behavior and deception detection, we design and prototype an AIGC detection solution that evaluates an essay submission by observing the digital behavior of the author during the writing process and during a brief follow-up survey in which the author answers several probing questions derived from the submitted text.

Following the design science methodology (Hevner et al., 2004), we first provide a summary of the relevant literature and existing AIGC detection solutions that inform the design requirements and the initial prototype of the system. We then report promising results of an initial evaluation of the prototype, validating the key aspects of the system design and laying the groundwork for future refinement of the solution. We conclude with a discussion of our contributions to both research and practice.

Research Background

The rise of LLMs as generative AI tools has been met with a corresponding rise in solutions to identify AIGC using a variety of approaches. Most of these tools attempt to "fight fire with fire," employing machine learning models to differentiate human-written text from AIGC (see Sadasivan et al., 2023 for a more extensive review). Some of these solutions require individual customization to accurately differentiate human-generated text from the output of each specific LLM (e.g., Liu et al., 2019). Other solutions exploit known statistical or generative patterns in AIGC to identify texts that follow those patterns (e.g., Gehrmann et al., 2019; Ippolito et al., 2019). To ensure the responsible use of LLMs, some have even proposed intentionally "watermarking" AIGC with features that betray artificial sources while remaining imperceptible to the average human reader (Kirchenbauer et al., 2023).

Another approach to AIGC detection relies on more explicit features in text that are typical of human language but difficult for LLMs to fully replicate. For example, linguistic patterns (e.g., word or n-gram frequencies) in AI-generated text sometimes differ from those typically found in human writing (e.g., Gallé et al., 2021; Zaitsu & Jin, 2023). Logical structure or syntax may reveal AI authorship as some LLMs struggle to interpret and replicate the complex structures or syntax in human language (McCoy et al., 2020). Moreover, LLMs are limited in their ability to capture the nuances of human writing shaped by social, cultural, and even visual/physical contexts (Bisk et al., 2020). Thus, references to colloquialisms or linguistically ambiguous metaphors in writing may indicate human authorship.

Despite the promising results of existing AIGC detection technologies, they are not without notable limitations. Most of these weaknesses can best be understood as a lack of generalizability along various dimensions. For example, detectors trained using output from an older LLM struggle to identify text from newer, more sophisticated models (Uchendu et al., 2020). Despite the impressive performance of LLMs in producing programming code (Sobania et al., 2023), extant detector models were developed and evaluated on natural language data and exhibit poor performance when detecting AI-generated code (Wang et al., 2023). Many detection algorithms have focused on English language detection and exhibit poor detection performance in other languages (Mitchell et al., 2023). Worse, many detector models demonstrate bias against non-native English authors, classifying their writing as AIGC far more often than that of native English speakers (Liang et al., 2023).

Beyond these generalizability issues, AIGC detection models suffer from the same limitations that afflict other deep learning algorithms. For example, prior research has shown that detection models can be vulnerable to adversarial examples designed to evade detection (Thuraisingham et al., 2017). Adversarial examples can be crafted by exploiting the model's sensitivity to certain input features, such as typos and punctuation changes, which can lead to false negatives or false positives in detection (Alzantot et al., 2018). Such adversarial examples can be effective even without a deep understanding of the model weights or

training data (Papernot et al., 2016). Highlighting another vulnerability, recent research has provided convincing evidence that applying a simple paraphrasing attack (in which AIGC is altered by a simple paraphrasing model) can severely reduce the accuracy of even the most sophisticated detection models (Krishna et al., 2023; Sadasivan et al., 2023).

Although advances in AIGC detection models continue to improve accuracy, increase generalizability, reduce bias, and provide stronger defenses against known threats (e.g., Krishna et al., 2023; Mitchell et al., 2023), those solutions will further contribute to the "technological arms race" and likely be vulnerable to future evasive tactics (Cotton et al., 2023). Without discounting the value of continued innovation in building AI-text detection models, we argue that there is room for other technology solutions that address the limitations of the existing state-of-the-art solutions.

Detecting AI-Generated Text with Cognitive and Behavioral Mechanisms

The novel approach to AIGC detection proposed in this paper addresses the problem from a different perspective by measuring the digital behavior of the human author as a proxy for relevant cognitive states during the writing process and during interrogation about the (allegedly) written text. A growing body of research has established that motor movements are influenced by cognitive and emotional changes (Freeman et al., 2008), demonstrating digital behavior as a viable and scalable methodology for studying a wide range of cognitive and emotional processes. Digital behavior data has been used to study negative emotions (Hibbeln et al., 2017), cognitive load (Thorpe et al., 2021), concealed racial prejudice (Wojnowicz et al., 2009), response certainty (Jenkins et al., 2015), and attention (Gozli & Pratt, 2011), among others. We propose two ways in which digital behavior should be relevant to the AIGC detection context.

First, observing a user's behavior during the writing process could reveal patterns that indicate whether the user wrote the text or used AI tools to do so. For example, anti-plagiarism service providers (e.g., Cadmus) have introduced essay writing environments that identify copy/paste behaviors. These tools would not detect AIGC if, for example, an author is transcribing an AI-generated passage without pasting. Depending instead on granular keystroke data capture may reveal patterns that distinguish original authorship from other forms of import or plagiarism. Writing original text requires significant cognitive resources (McCutchen, 1996) and entails coordination of multiple dynamic processes (Olive, 2014), including the motor movement system when writing using a keyboard (Leijten & Waes, 2013). Keystroke data have been used by writing researchers to understand the dynamic, iterative process of writing (Lindgren & Sullivan, 2019), including hesitations (Medimorec & Risko, 2017) and revisions (Conijn et al., 2021). This prior research implies that keystroke data obtained during a writing activity can be expected to follow identifiable patterns that would indicate whether the author of a passage of text is human; indeed, typing behaviors while writing are consistent enough to allow researchers to predict writing quality and other outcomes from keystroke data (see Conijn et al., 2022 for a review).

Although we could find no prior research applying digital behavior data to the AIGC detection context, others have demonstrated that digital behavior data can be used to differentiate humans from automated tools, most notably in relation to detecting online bots. Most of the work in this area focuses on mousing data (e.g., Acien et al., 2022; Iliou et al., 2021; Wei et al., 2019), though the concept has also been demonstrated using keystroke data (Chu et al., 2018) and with data from touchscreen devices (Acien et al., 2021). Given the successful use of behavioral data in classifying automated browsing activity—and the distinctive behavioral signatures expected during authentic human writing—we see potential for those data to reveal markers of AIGC. Our system design incorporates several mechanisms for generating those markers, as described in more detail below.

Second, we argue that authors who use a passage of AIGC copied from an external source will be less familiar with its content and be forced to maintain the deception when forced to answer follow-up questions about the process of writing the passage or the content of the passage itself. Our system design is thus also informed by the longstanding notion that maintaining a deception is cognitively demanding (Nuñez et al. 2005, Gombos 2006), taxing mental resources such as working memory (Carrión et al., 2010), particularly when the deception is placed under scrutiny (Vrij, 2008).

In developing our system design, we draw from prior deception research in which individuals are interrogated following an activity while their digital behaviors are monitored. For example, mouse movement data revealed an attraction toward truthful response targets for participants concealing

information about a (mock) theft they had committed (Jenkins et al., 2019). Among individuals who cheated on an online test for monetary gain, mouse movements during a follow-up questionnaire deviated significantly more than those who did not cheat (Jenkins et al., 2021). In a fraud context, Weinmann et al. (2022) used mouse movements to distinguish people who cheated to receive an unearned payout. Lastly, Monaro et al. (2018; 2017) used both keystroke and mouse movement data obtained during follow-up questioning to differentiate participants who provided fake identity information from those who provided their real identity information.

The common thread among these digital behavior studies and tools is the notion that deception, theft, or concealment of information increases cognitive demands, creating observable differences in keystroke and mouse movement patterns. Importantly, the generalizability of these relationships across a wide variety of contexts is in stark contrast with the key drawbacks of existing AIGC detection solutions we have noted.

Summary of the Research Gap

Sophisticated LLMs produce text that is increasingly difficult to distinguish from human writing. Given the associated risks for abuse and plagiarism using these AI tools, many solutions have been developed that attempt to classify text passages as either AI- or human-generated. Although some of these tools show strong performance in certain settings, they have numerous limitations. Most of these limitations are variations on the theme of poor generalizability, though detection models are also vulnerable to relatively simple paraphrasing attacks or adversarial examples that affect deep learning algorithms in general.

The system proposed in this research is designed to provide an alternative means to judge whether a user authored a passage of text, capturing digital behavior during the writing process as a means of identifying behavioral markers of human versus automated text generation. Framing AIGC detection as a *deception* problem, we also have at our disposal a rich body of research demonstrating that deception drains cognitive resources, and that those cognitive demands produce observable differences in mousing behavior during follow-up interrogation about the deception. Thus, the system design described in the next section embodies the following high-level design goals: (1) capture behavioral markers to accurately identify AIGC and (2) provide generalizable detection accuracy across different LLMs, output languages (including coding languages), and authors' native language.

System Design

This section provides an overview of the system we designed to address the research gap described in the preceding section. The full system design is summarized in Figure 1. The broader system design is envisioned as a multicomponent system comprised of a JavaScript utility that captures typing and mousing behaviors during both the writing and follow-up questioning process and a web service that derives follow-up questions from the provided text and evaluates the typing and mousing behaviors to produce a predictive score indicating the likelihood that the text was written by the user. In the sections that follow, we detail the full system design and demonstrate how it addresses the design requirements introduced above. We then summarize the results from the first of several planned evaluation studies intended to iteratively refine the system and its capabilities.

Our behavior capture method consists of a JavaScript utility that runs in the background during a web interaction and listens for interaction events. Using JavaScript to capture typing and mousing behaviors is a common approach used in prior research (e.g., Mathur & Reichling, 2019; Valacich et al., 2022) to record a diverse set of interaction events (e.g., clicks, mouse movements, key presses, etc.) with precise timestamps. These rich data can then be processed to produce typing and mousing features that are relevant to a given context. Although the methods for extracting these features from web interaction events is beyond the scope of this paper, we adopted the approach and recommendations of Valacich and colleagues (2022) to develop a simple JavaScript utility that can be embedded on a webpage for behavior capture. The capabilities of the JavaScript utility were determined by our system design goals, as described below.

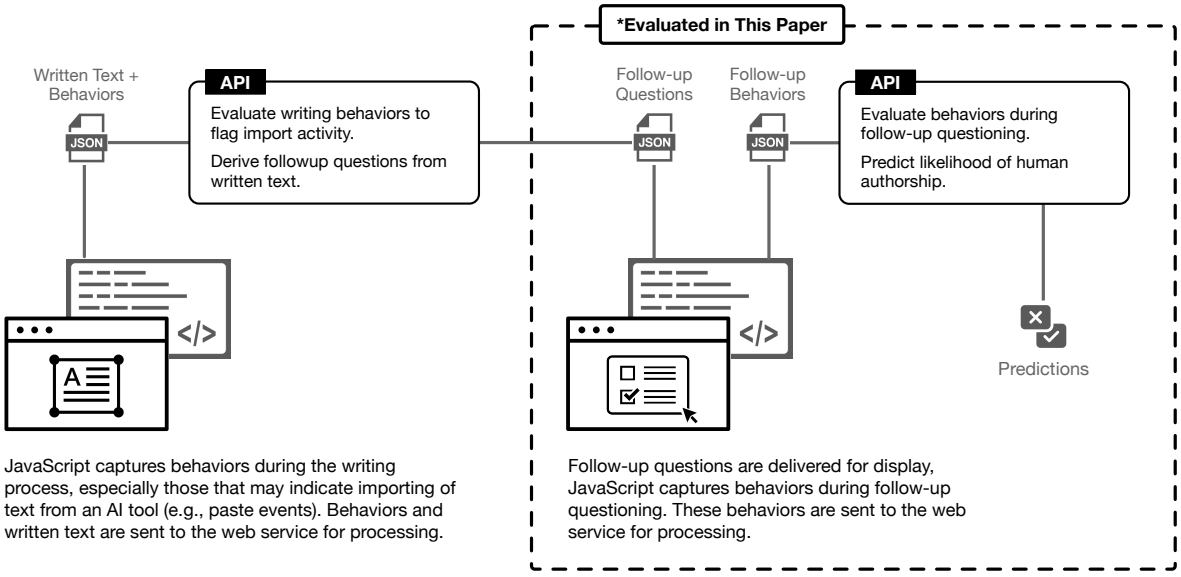

Figure 1. Overview of System Design

Behaviors During the Writing Process

Prior research investigating keystroke data in relation to the writing process has proposed at least five groups of keystroke features relevant to the (human) writing process (see Conijn et al., 2022): (1) duration of key presses and transitions between presses, from which pauses and hesitations can be inferred (Medimorec & Risko, 2017); (2) features that quantify backspaces or deletions, from which revisions can be inferred (Conijn et al., 2021); (3) an indication of verbosity, especially character and word counts (Likens et al., 2017); (4) features related to typing skill or fluency (e.g., frequency of typing mistakes) (Waes & Leijten, 2015); (5) features describing other text manipulation operations, including text selection, paste, and so on (Baaijen & Galbraith, 2018). We followed these recommendations from the literature as a starting set of behavioral features that our JavaScript utility was designed to produce. We expect that list to evolve as we proceed with system evaluation and refinement following the design science process.

Given the lack of prior research investigating whether keystroke data obtained during the writing process can be used to effectively differentiate between human and AI authors, we chose to avoid being prescriptive at this design stage of the system development effort. Some of the features mentioned above should provide clear indications that an author has borrowed the content of a passage from an AI tool—observing a large amount of text pasted from elsewhere without revision, for example. Others—especially in terms of what keystroke patterns truly indicate "normal" human writing behavior—may be more subtle and require learning through the system refinement effort. One important goal during iterative refinement of this system component will be to accumulate data and understanding related to the typical range of typing behaviors that can be expected during free-text writing activities. We anticipate supplementing our accumulated datasets with publicly available keystroke research datasets, some of which contain keystroke samples from similar free-response writing activities (e.g., Killourhy & Maxion, 2012).

Behaviors During a Follow-Up Questioning Process

As we have noted, there is ample evidence that deceptive computer users' interaction patterns are systematically different from those of truthful users. Authors who submit AIGC as their own work are engaging in deception (i.e., plagiarism) that will affect their digital behavior as seen in prior research. Although some researchers have successfully distinguished deceitful from truthful users without overtly challenging the perpetrators about their deception (e.g., Weinmann et al., 2022), most have employed a more direct approach, inducing stronger behavioral reactions using a follow-up questionnaire (Jenkins et

al., 2019; Jenkins et al., 2021; Monaro et al., 2018; Monaro et al., 2017). These follow-up challenges induce additional cognitive load during deception (Vrij, 2008) and typically take the form of unanticipated requests for further detail or elaboration on a portion of the deceiver's story, placing additional cognitive demands on the deceiver (Lancaster et al., 2013).

Our system design adopts this "intervention" paradigm, providing an automated mechanism to derive a set of follow-up questions from the text submitted by the user. As the user responds to the questions, our JavaScript utility captures behaviors used to differentiate those who submitted AIGC from those who authored the text themselves. The design of the follow-up questions was directly informed by the theory and empirical background in the writing domain, in which writing is portrayed as a highly cognitive process (McCutchen, 1996) that includes spurts of creativity intermixed with reconsideration and revision as the text is refined (Medimorec & Risko, 2017). During the writing process, the author engages deeply with the concepts being discussed, resulting in significantly deeper familiarity with and retention of the text and topic (Dunlosky et al., 2013). Thus, a user who passively pastes in AIGC will be less familiar with the text than one who authors the text—just as copy-paste notetaking facilitates poor retention compared with more active notetaking strategies (Igo et al., 2005).

With this logic in mind, the follow-up questions generated by our system are designed to be easy to answer for the author of the submitted text. They consist of multiple choice and true/false questions that are purely factual in nature and directly derived from statements provided in the submitted text. As the system design calls for dynamic follow-up question generation tailored to the submitted text, we explored various commercial NLP question generators, evaluating each in terms of cost, response time, and consistency of questions generated. We ultimately decided to use the ChatGPT service from OpenAI (OpenAI, 2023) as it was the most capable of accurately interpreting the variety of writing samples we tested. After some trial and error, we settled on a set of instructions that produce a set of follow-up questions similar in terms of sentence structure and format but tailored to the content of the submitted text. (The resulting set of instructions is included in the Appendix, Figure A1.) Table 1 provides several examples of questions generated by the system using this procedure.

Although the generated questions are intentionally simple, even users who authored the submitted text may forget how they discussed a certain issue or topic in their writing. Accordingly, the system displays the text of the submitted essay directly above the follow-up questions on the same page so that the user can be reminded of what was written. Although this design decision provides the "correct" answers even to those who plagiarize AIGC from an AI tool, recall that we are more interested in the users' *behavior while providing* those answers to classify text authorship. This approach thus controls for the likely common scenario in which a legitimate author forgets a small detail while retaining the efficacy of the detection approach enabled by the mousing features described below.

Question Type	Question Text	Answer Choices
Multiple choice	According to the author, what is more important than trying to remove personal biases?	Ignoring them, Justifying them, Being aware of them, Accepting them
	Which phrase did the author use to describe their favorite people?	Those who think most like me, Those who look least like me, Those who share my interests, Those who live in my comfort bubble
True/false	The author claims that they have never experienced any biases.	True/false
	The author believes that the IAT is a perfect tool for measuring biases.	True/false

Table 1. Sample System-Generated Follow-up Questions

Using established methods from the mousing–deception literature (e.g., Jenkins et al., 2021), the follow-up questions are displayed as radio buttons in a simple web interface developed by the research team, where the JavaScript utility measures mousing behavior while the user selects answers. The JavaScript utility

records x- and y-positions and timestamps throughout the answering process, from which a variety of mousing features can be extracted.

As with the keystroke features described earlier, we looked to prior mouse tracking research for an approach and initial set of mousing features that would apply to our context. Although much prior research has used tightly controlled interfaces to simplify feature extraction (e.g., Ericson et al., 2021; Freeman et al., 2011), we prefer the more flexible method demonstrated by Jenkins and colleagues (2021) in which the free-browsing movements are broken into submovements that are each evaluated in terms of their movement efficiency. We chose to adopt this method because (1) our follow-up questions are dynamically produced with varying lengths, and (2) we wanted to allow for as much free browsing behavior as possible.

Specifically, we expect substantially more free browsing behavior from users who did not author a submission as they are required to find the answers in the submission displayed on the follow-up page. Extending this logic further, we selected the six features summarized in Table 2 as the starting set of mousing features relevant for our follow-up challenge scenario. Each of the features in Table 2 is defined using the free-browsing paradigm suggested by Jenkins et al. (2021), and each provides a proxy measurement related to the following overarching idea: users who did not author a passage of text will take longer and overall exhibit more free-browsing mouse movements when responding to the follow-up questions displayed by our system.

Feature	Definition	Explanation of Expected Direction (for AIGC Follow-Up)
1. Area under the curve (AUC)	The geometric area (see in-text explanation) between the actual movement path and the idealized response trajectory for each submovement; summed across all submovements during follow-up.	*Positive.* AUC is a standard measure of deviation and indicates searching or uncertainty while mousing. Users who must search for help answering questions will have more deviating mousing patterns.
2. Distance	Total distance traveled while interacting on the follow-up page.	*Positive.* Mousing distance represents searching or uncertainty and will be higher for users who must scroll to find clues in the submitted text.
3. Sub-movements	Count of individual submovements while interacting on the follow-up page.	*Positive.* Submovement count is influenced by longer mousing distances and more pauses (i.e., more searching increases this count).
4. Total time	Time (in milliseconds) spent interacting on the follow-up question page.	*Positive.* Users who are unsure of the questions about their submitted text will be slower to answer. Searching for clues in the unfamiliar text will also take more time.
5. x-flips	Count of direction changes on the x-axis.	*Positive.* More flips indicate more searching or uncertainty.
6. y-flips	Count of direction changes on the y-axis.	*Positive.* More flips indicate more searching or uncertainty.

Table 2. Mousing Features Extracted from Follow-Up Questioning Interaction

The area under the curve (AUC) feature in Table 2 requires more than simple definitional explanation, and is approximated as follows. For each movement point $p_2 \ldots p_n$ in the array of points within a submovement, a triangle can be drawn connecting the current point p_n, to the previous point p_{n-1}, and the beginning point p_0. For example, in Figure 2, the blue triangle connects point A (the beginning point) with points B (the previous point) and C (the current point). Likewise, the red triangle connects points A (the beginning point) with C (the previous point) and D (the current point). For each triangle, the distance of each side can be calculated by the formula $[(x_2 - x_1)^2 + (y_2 - y_1)^2]^{1/2}$ where (x_1, y_1) and (x_2, y_2) are two points on the coordinate plane. Then the area of the triangle can be calculated using Heron's formula. Heron's formula states that

the area of a triangle with sides of length a, b, and c is $[s \times (s - a) \times (s - b) \times (s - c)]^{1/2}$ where s is the semi-perimeter of the triangle, given by $s = (a + b + c)/2$. Finally, the areas of all the triangles are summed together to estimate AUC across the entire movement.

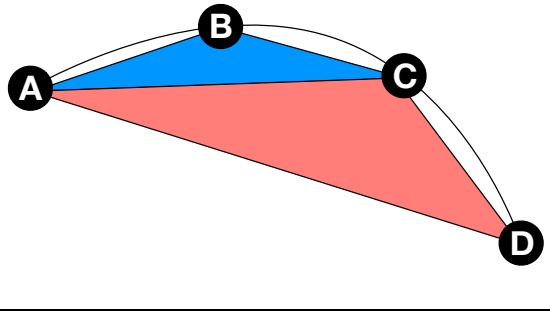

Figure 2. Illustration of (Heron) AUC Calculation

Evaluation

We designed and built our system prototype according to the objectives identified in the Research Background section. Drawing from relevant reference theories, the system provides a novel technological solution to address several notable gaps found in existing AIGC detection solutions. The next stage in the design science process is to evaluate the system to gauge whether it effectively meets the design objectives and iteratively refine the system design and functionality based on our learnings (Hevner et al., 2004).

Study Description

As an extensive evaluation of the full system would strain the page limit of this conference paper, we report here our findings from the first of several planned evaluation studies. Specifically, we conducted a controlled experiment in a real-world setting to test the efficacy of the system-generated follow-up questions—and the mouse movement data gathered while answering them—in distinguishing human-authored writing from AIGC. There were two objectives of the study: (1) provide an initial test of the key components of the system supporting the follow-up questioning process (i.e., the right side of the diagram in Figure 1), and (2) evaluate each of the mousing features summarized in Table 2 for their ability to distinguish answering behaviors surrounding AIGC versus self-authored text.

To accomplish these objectives, we used a within-subjects design in which each participant used our system to answer a set of system-generated follow-up questions about two short essays on the same topic (one written by them and the other written by ChatGPT) while our system monitored their mousing behaviors. Participants were recruited from a graduate business course at a large private university in the US. As a requirement of the course, students were asked to write a short essay reflecting on a recent topic discussed in class. (The essay prompt is included in the Appendix.) Participation in the study was optional and volunteers were rewarded with extra credit equivalent to < 1% of their final grade. The study was conducted with appropriate approval from an institutional review board.

Participants completed the online study remotely but were required to use a laptop or desktop computer to ensure that mousing data could be captured. As approximately one week had passed between the course assignment and the launch of the study, we added an essay review step after the consent and instructions page where participants could optionally read through the essay. (Half were randomly assigned to respond about the AI-generated essay first.) On the next screen, participants answered four follow-up questions derived from the essay they had just viewed. As discussed, the follow-up page also displayed the text of the essay for reference. To ensure the most salient mousing behaviors were registered for participants who needed to consult the essay text, we constrained the essay text to a smaller text box that only displayed a few lines of text and forced scrolling behavior to search further. After answering all 4 questions, participants repeated the essay review and follow-up questions for the other condition. They were then redirected to a brief post-survey containing demographic measures, manipulation checks, and a debrief. (Figure A3 in the Appendix provides a screenshot of the follow-up question page in the prototype system used in the study.)

Participants

A total of 79 students participated in the study. Six participants were removed from the sample for failing to follow instructions or because of incomplete responses, leaving 73 participants, or 146 observations. Approximately 68% of the participants were male, 96% reported English as their native language, and the average age was 23.7.

Analysis and Results

We first performed a manipulation check, comparing the two conditions in terms of participants' perceived difficulty of the follow-up questions they answered for each. ("Please rate the questions you answered about that essay in terms of how easy (=1) or difficult(=7) they were to answer.") A paired samples t-test revealed a significant difference in perceived difficulty of the questions about the self-authored (M = 2.24, SD = 1.30) versus AI-generated essays (M = 3.65, SD = 1.36); $t(71)$ = -7.65, p < .001. These results indicate that answering questions about a passage of text participants had written was easier, as expected.

We then proceeded to evaluate each of the six mousing features obtained on the follow-up question page (see Table 2). We started by calculating within-participant normalized z-scores for each of the six features. This entails calculating a mean and standard deviation of each user's behavior, then normalizing each feature using that user-specific baseline. In the context of our 2 experimental conditions, these z-scores quantify how the features obtained in each condition differs from that individual's baseline behavior. This transformation helps account for natural individual differences in how users behave (e.g., typing speed, reaction times, etc.). The transformation also aids interpretation and facilitates comparison among the features as they are measured on very different scales (e.g., ranging from dozens of x-flips to thousands of milliseconds). Descriptive statistics for all six (normalized) features are summarized in Table 3. As further shown in the box plots comparing each normalized feature across the two conditions (see Figure 3), participants answering follow-up questions about the self-authored versus AI-generated essay produced starkly different mousing behaviors among the six features examined.

Condition	AUC	Distance	Submovements	Total Time	X-Flips	Y-Flips
AI-Generated	1.40(.57)	1.43(.59)	1.39(0.46)	1.12(.65)	1.38(.52)	1.36(.52)
Self-Authored	0.73(.57)	0.65(.60)	0.83(.51)	0.32(.64)	0.77(.55)	0.78(.55)

Table 3. Descriptive Statistics for Normalized Mousing Features: Mean(SD)

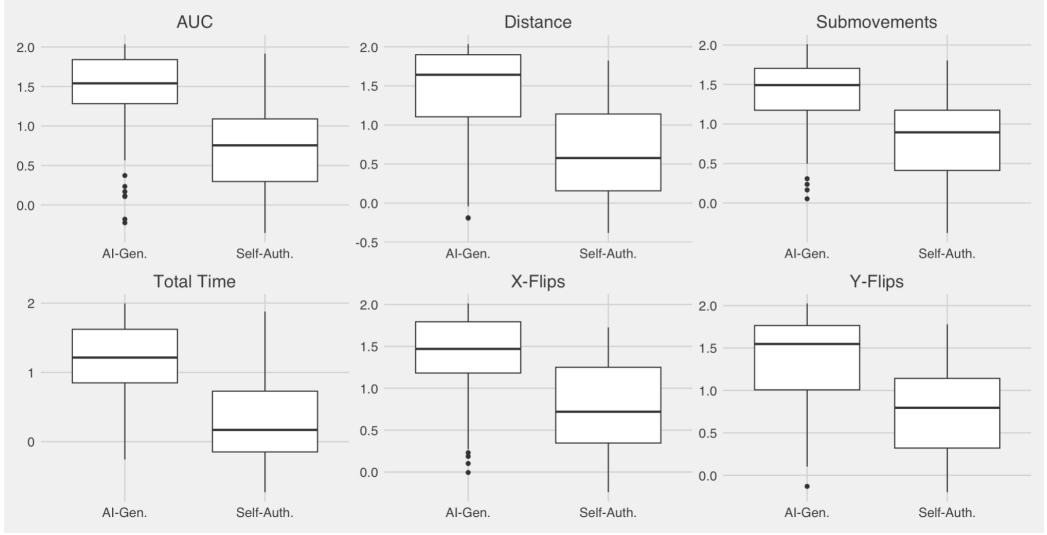

Figure 3. Distributions of Normalized Mousing Features Across Conditions

To statistically quantify these differences in the mousing features across conditions, we completed a mixed-effects linear regression model, which uses random effects to account for individual differences in repeated observations (Pinheiro & Bates, 2000)—for example, differences in memory capacity or demographics—and fixed effects to model the treatment effect. We specified separate models for each of the six normalized mousing features, summarized in Table 4 and Table 5. In all six models, we observe that asking follow-up questions about the AI-generated essay created significantly different mousing behaviors.

	Predicting: AUC		Predicting: Distance		Predicting: Submovements	
Fixed Effect	Estimate(*SE*)	*t* (*df*=144)	Estimate(*SE*)	*t* (*df*=144)	Estimate(*SE*)	*t* (*df*=144)
Intercept	0.73(.07)	10.93***	0.65(.07)	9.33***	0.83(.06)	14.61***
AIGC Cond.	0.67(.09)	7.16***	0.78(.10)	8.00***	0.55(.08)	6.85***

Table 4. Results of the Linear Mixed-Effects Models, Features 1-3

*(Note. ***p <.001.)*

	Predicting: Total Time		Predicting: X-Flips		Predicting: Y-Flips	
Fixed Effect	Estimate(*SE*)	*t* (*df*=144)	Estimate(*SE*)	*t* (*df*=144)	Estimate(*SE*)	*t* (*df*=144)
Intercept	0.32(.08)	4.29***	0.77(.06)	12.42***	0.78(.06)	12.50***
AIGC Cond.	0.80(.11)	7.53***	0.61(.09)	6.92***	0.57(.09)	6.44***

Table 5. Results of the Linear Mixed-Effects Models, Features 4-6

*(Note. ***p <.001.)*

Discussion

The advent of LLMs like ChatGPT has brought about a wealth of opportunities for innovation in various domains, including higher education. However, these models also present significant challenges, particularly in detecting AIGC. Numerous studies have examined various algorithmic methods of identifying AIGC, revealing some promising results that are sure to continue to improve over time. Nevertheless, existing solutions for AIGC detection suffer from limitations—especially in terms of generalizability (Mitchell et al., 2023; Sobania et al., 2023) and biases (Liang et al., 2023)—as well as vulnerabilities to adversarial attacks (Ippolito et al., 2019; Solaiman et al., 2019).

We employed a design science methodology to develop a novel system for AIGC detection designed to address some of the weaknesses of existing solutions. Drawing from relevant literature related to cognitive processes involved in both writing and deception, we derived design objectives to guide our development, producing a comprehensive design and a prototype system that together represent an innovative application of digital behavior in the emerging domain of AIGC detection.

The results from a controlled experiment set within a real-world use case (an essay writing assignment for a graduate business course), demonstrated the successful generation of follow-up questions derived from submitted text. Furthermore, we provided initial evidence that the mousing features extracted during the follow-up interrogation by the system effectively distinguish between user-written and AI-generated text. These results constitute a promising first step in the system evaluation and validation process.

Implications for Research

Our design science methodology produced valuable lessons in the form of system design principles that can be generalized to other related research. One crucial learning is that digital behavior (i.e., *what users do*) might be just as important, if not more important, than other forms of evaluation (usually consisting of *what users know*). As an illustration of this principle, we conducted a post-hoc analysis comparing the accuracy of the users' *answers* to the follow-up questions with their *behavior while answering* those

questions. As the essay text was made available to participants during follow-up questioning, nearly every participant answered every question correctly, regardless of whether they had written the essay themselves or not (on average, 97% correct across both conditions, with no statistically significant difference in answer correctness between the two; $t(78) = 0.276$, *ns*). In contrast, all six of the mousing behaviors displayed significant differences between the two conditions (see Figure 3). This finding underscores the value of evaluating users' actions in addition to their knowledge and highlights the potential for leveraging digital behavior in AIGC detection.

A second learning relates to the value of context-aware follow-up questioning. The approach we developed to accomplish this with our system leveraged ChatGPT to process the submitted text and generate consistent follow-up questions. This worked well in our context, but LLMs like ChatGPT are flexible, powerful tools that can help automate this process in other system development efforts, particularly those in which system design goals call for automatically interpreting context and tailoring subsequent interactions based on that context.

Importantly, our results provide insight into the cognitive process of responding to questions about self-generated content versus the cognitive process of responding to questions about AIGC. Past researchers state that behavior tracking provides "continuous streams of output that can reveal ongoing dynamics of processing" (Freeman et al., 2011, p. 1). In our study, we found that people spend significantly more time and have more direction changes and submovements when answering questions about AIGC, suggesting that people engaged in a greater cognitive search process. Greater deviation was also significant, which has been associated with increased uncertainty and increased cognitive load while making a decision (Jenkins et al., 2019). Future research can use this knowledge of the decision process to develop tools and methods for assessing whether a person generated a passage of text or used an LLM tool.

Practical Contributions

The system described in this research offers several compelling practical contributions. First, it provides an innovative approach for detecting AIGC in educational settings, where maintaining academic integrity is crucial. Using this approach to capture and analyze students' digital behavior during applicable assessment scenarios, educators can gain additional assurance in the integrity of the assessment.

Our approach may also extend beyond educational settings to professional scenarios in which the integrity of a written assessment is important. For example, many hiring procedures for technology professionals require a demonstration of technical knowledge in the form of a coding deliverable. As LLMs like ChatGPT are adept at producing full solutions to coding problems, these technical demonstrations are vulnerable to plagiarism. However, ChatGPT can derive follow-up questions for submitted code with minor adaptations to the design presented in this paper. Mousing behavior could then be recorded while applicants answered the follow-up questions. Although the efficacy of this hypothetical extension remains speculative until empirically validated, we offer the example by way of demonstration that the key assumptions embedded in our system design may well be valuable in several domains.

Limitations and Future Directions

The study reported in this paper does not represent a full system evaluation as recommended by Hevner et al. (2004). Several other follow-up studies are planned, the first of which will address the other components of the system that were not tested here (i.e., the use of keystroke and other behavioral data during the writing process to extract markers relevant for AIGC detection). We also acknowledge that, although prior research pairing mousing and keystroke data with deception detection has been applied broadly in many domains, the present paper does not provide any empirical evaluation of generalizability to the other contexts we highlighted as primary design considerations, including with authors for whom English is a second language. Future research is planned that will replicate and refine these findings in these other important domains, including among more diverse populations. These future studies will help further refine and validate the system's effectiveness in detecting AIGC across various contexts.

Our system design is best suited for essay-length writing scenarios, especially in settings where plagiarism and academic honesty are important. It is not optimized for short-form content like tweets and may not generalize to those use cases. Adapting the system to accommodate various text formats and use cases constitutes a promising opportunity for future research and development efforts.

We used ChatGPT to extract the set of follow-up questions from each essay. This approach allowed for the extraction to be standardized and automated. However, sending essay texts to a third-party service like ChatGPT may raise issues related to privacy and consent. Future system evaluations could explore other extraction methods that avoid sharing text with a third-party service provider.

Our experimental design did not allow for other relevant AIGC import scenarios that are likely common in educational assessment scenarios. There are likely several other real-world uses of ChatGPT that fall in the middle of the spectrum between "full" authorship of an essay and the plagiaristic use of AIGC represented by our two experimental conditions. For example, users may adjust the content after pasting from ChatGPT, transcribe the content from another open window, or use ChatGPT to produce an outline that they then expand. Future iterations that build on this work will need to incorporate some of those scenarios in the system evaluation strategy to ensure that our approach remains effective in the most common real-world applications.

Conclusion

Our research presents a novel system for detecting AIGC by applying digital behavior capture and automated follow-up questioning as a method for verifying the author of a text passage. Following the design science research paradigm, our system design and prototype offer potential solutions to some of the limitations and vulnerabilities found in existing detection solutions. We executed a controlled experiment as an initial evaluation study to test the prototype system in its ability to produce differentiating mousing features during follow-up questioning of the alleged author of an essay. The results reveal significant differences that show promise in differentiating between user-authored and AI-generated text. The system represents a promising step towards addressing the challenges posed by AI-generated text in various domains, particularly in educational settings where maintaining academic integrity is vital.

References

Acien, A., Morales, A., Fierrez, J., & Vera-Rodriguez, R. (2022). BeCAPTCHA-mouse: Synthetic mouse trajectories and improved bot detection. *Pattern Recognition*, *127*(C), 1–13. **https://doi.org/10.1016/j.patcog.2022.108643**

Acien, A., Morales, A., Fierrez, J., Vera-Rodriguez, R., & Delgado-Mohatar, O. (2021). BeCAPTCHA: Behavioral bot detection using touchscreen and mobile sensors benchmarked on humidb. *Engineering Applications of Artificial Intelligence*, *98*(1), 1–11. **https://doi.org/10.1016/j.engappai.2020.104058**

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., & Chang, K.-W. (2018). Generating natural language adversarial examples. *arXiv*. **https://doi.org/10.48550/arxiv.1804.07998**

Baaijen, V. M., & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction*, *36*(3), 199–223. **https://doi.org/10.1080/07370008.2018.1456431**

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience grounds language. *arXiv*. **https://doi.org/10.48550/arxiv.2004.10151**

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. **https://doi.org/10.48550/arxiv.2108.07258**

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *arXiv*. **https://doi.org/10.48550/arxiv.2005.14165**

Carrión, R. E., Keenan, J. P., & Sebanz, N. (2010). A truth that's told with bad intent: An ERP study of deception. *Cognition*, *114*(1), 105–110. **https://doi.org/10.1016/j.cognition.2009.05.014**

Chu, Z., Gianvecchio, S., & Wang, H. (2018). *From database to cyber security* (P. Samarati, I. Ray, & I. Ray, Eds.). Springer Nature.

Conijn, R., Cook, C., Zaanen, M. v., & Waes, L. V. (2022). Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, *32*(4), 835–866. **https://doi.org/10.1007/s40593-021-00268-w**

Conijn, R., Speltz, E. D., & Chukharev-Hudilainen, E. (2021). Automated extraction of revision events from keystroke data. *Reading and Writing*, *34*(2), 1–26. **https://doi.org/10.1007/s11145-021-10222-w**

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *60*(4), 1–12. **https://doi.org/10.1080/14703297.2023.2190148**

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest*, *14*(1), 4–58. **https://doi.org/10.1177/1529100612453266**

Eaton, S. E. (2023). The academic integrity technological arms race and its impact on learning, teaching, and assessment. *Canadian Journal of Learning and Technology*, *48*(2), 1–9. **https://doi.org/10.21432/cjlt28388**

Ericson, J. D., Albert, W. S., & Bernard, B. P. (2021). Investigating the relationship between web object characteristics and cognitive conflict using mouse-tracking. *International Journal of Human-Computer Interaction*, *37*(2), 99–117. **https://doi.org/10.1080/10447318.2020.1808352**

Floridi, L. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds & Machines*, *30*(1), 681–694. **https://doi.org/10.2139/ssrn.3827044**

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology*, *137*(4), 673–690. **https://doi.org/10.1037/a0013875**

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*(59), 1–6. **https://doi.org/10.3389/fpsyg.2011.00059**

Gallé, M., Rozen, J., Kruszewski, G., & Elsahar, H. (2021). Unsupervised and distributional detection of machine-generated text. *arXiv*. **https://doi.org/10.48550/arxiv.2111.02878**

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). Gltr: Statistical detection and visualization of generated text. *arXiv*. **https://doi.org/10.48550/arxiv.1906.04043**

Gozli, D. G., & Pratt, J. (2011). Seeing while acting: Hand movements can modulate attentional capture by motion onset. *Attention, Perception, & Psychophysics*, *73*(8), 2448–2456. **https://doi.org/10.3758/s13414-011-0203-x**

Hamilton, R. (1989). The effects of learner-generated elaborations on concept learning from prose. *The Journal of Experimental Education*, *57*(3), 205–217. **https://doi.org/10.1080/00220973.1989.10806506**

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays. *arXiv*. **https://doi.org/10.48550/arXiv.2304.14276**

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. **https://doi.org/10.2307/25148625**

Hibbeln, M., Jenkins, J. L., Schneider, C., Valacich, J. S., & Weinmann, M. (2017). How is your user feeling? Inferring emotion through human-computer interaction devices. *MIS Quarterly*, *41*(1), 1–21. **https://doi.org/10.25300/misq/2017/41.1.01**

Igo, L. B., Bruning, R., & McCrudden, M. T. (2005). Exploring differences in students' copy-and-paste decision making and processing: A mixed-methods study. *Journal of Educational Psychology*, *97*(1), 103–116. **https://doi.org/10.1037/0022-0663.97.1.103**

Iliou, C., Kostoulas, T., Tsikrika, T., Katos, V., Vrochidis, S., & Kompatsiaris, Y. (2021). Detection of advanced web bots by combining web logs with mouse behavioural biometrics. *Digital Threats: Research and Practice*, *2*(3), 1–26. **https://doi.org/10.1145/3447815**

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2019). Automatic detection of generated text is easiest when humans are fooled. *arXiv*. **https://doi.org/10.48550/arxiv.1911.00650**

Jenkins, J. L., Larsen, R., Bodily, R., Sandberg, D., Williams, P., Stokes, S., Harris, S., & Valacich, J. S. (2015). A multi-experimental examination of analyzing mouse cursor trajectories to gauge subject uncertainty. Twenty-First Americas Conference on Information Systems, San Juan, Puerto Rico.

Jenkins, J. L., Proudfoot, J. G., Valacich, J. S., Grimes, G. M., & Nunamaker, J. F., Jr. (2019). Sleight of hand: Identifying concealed information by monitoring mouse-cursor movements. *Journal of the Association for Information Systems*, *20*(1), 1–32. **https://doi.org/10.17705/1jais.00527**

Jenkins, J. L., Valacich, J. S., Zimbelman, A. F., & Zimbelman, M. F. (2021). Detecting noncompliant behavior in organizations: How online survey responses and behaviors reveal risk. *Journal of Management Information Systems*, *38*(3), 704–731. **https://doi.org/10.1080/07421222.2021.1962600**

Killourhy, K. S., & Maxion, R. A. (2012). Free vs. Transcribed text for keystroke-dynamics evaluations. Workshop on Learning from Authoritative Security Experiment Results, Arlington, VA.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. *arXiv*. **https://doi.org/10.48550/arxiv.2301.10226**

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *arXiv*. **https://doi.org/10.48550/arxiv.2303.13408**

Lancaster, G. L. J., Vrij, A., Hope, L., & Waller, B. (2013). Sorting the liars from the truth tellers: The benefits of asking unanticipated questions on lie detection. *Applied Cognitive Psychology*, *27*(1), 107–114. **https://doi.org/10.1002/acp.2879**

Leijten, M., & Waes, L. V. (2013). Keystroke logging in writing research. *Written Communication*, *30*(3), 358–392. **https://doi.org/10.1177/0741088313491692**

Li, C., & Xing, W. (2021). Natural language generation using deep learning to support mooc learners. *International Journal of Artificial Intelligence in Education*, *31*(2), 186–214. **https://doi.org/10.1007/s40593-020-00235-x**

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native english writers. *arXiv*. **https://doi.org/10.48550/arXiv.2304.02819**

Likens, A. D., Allen, L. K., & McNamara, D. S. (2017). Keystroke dynamics predict essay quality. 39th Annual Meeting of the Cognitive Science Society, London, UK.

Lindgren, E., & Sullivan, K. (2019). *Observing writing : Insights from keystroke logging and handwriting*. Brill.

Ling, C., & Libby, T. (2010). Writing mini-cases: An active learning assignment [Article]. *Issues in Accounting Education*, *25*(2), 245–265. **https://doi.org/10.2308/iace.2010.25.2.245**

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv*. **https://doi.org/10.48550/arxiv.1907.11692**

Malik, A., Wu, M., Vasavada, V., Song, J., Coots, M., Mitchell, J., Goodman, N., & Piech, C. (2019). Generative grading: Near human-level accuracy for automated feedback on richly structured problems. *arXiv*. **https://doi.org/10.48550/arxiv.1905.09916**

Mathur, M. B., & Reichling, D. B. (2019). Open-source software for mouse-tracking in Qualtrics to measure category competition. *Behavior Research Methods*, *51*(5), 1987–1997. **https://doi.org/10.3758/s13428-019-01258-6**

McCoy, R. T., Frank, R., & Linzen, T. (2020). Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *arXiv*. **https://doi.org/10.48550/arxiv.2001.03632**

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, *8*(3), 299–325. **https://doi.org/10.1007/bf01464076**

Medimorec, S., & Risko, E. F. (2017). Pauses in written composition: On the importance of where writers pause. *Reading and Writing*, *30*(6), 1267–1285. **https://doi.org/10.1007/s11145-017-9723-7**

Mitchell, A. (2022). *Professor catches student cheating with ChatGPT: 'I feel abject terror'*. **https://nypost.com/2022/12/26/students-using-chatgpt-to-cheat-professor-warns/**

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. *arXiv*. **https://doi.org/10.48550/arxiv.2301.11305**

Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., & Sartori, G. (2018). Covert lie detection using keyboard dynamics. *Scientific Reports*, *8*(1), 1–10. **https://doi.org/10.1038/s41598-018-20462-6**

Monaro, M., Gamberini, L., & Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PLoS ONE*, *12*(5), 1–19. **https://doi.org/10.1371/journal.pone.0177851**

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge University Press.

Olive, T. (2014). Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research*, *6*(2), 173–194. **https://doi.org/10.17239/jowr-2014.06.02.4**

OpenAI. (2023). *Gpt-4 technical report* **https://cdn.openai.com/papers/gpt-4.pdf**

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical black-box attacks against machine learning. *arXiv*. **https://doi.org/10.48550/arxiv.1602.02697**

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer-Verlag.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv*. **https://doi.org/10.48550/arxiv.2303.11156**

Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). An analysis of the automatic bug fixing performance of ChatGPT. *arXiv*. **https://doi.org/10.48550/arxiv.2301.08653**

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv*. **https://doi.org/10.48550/arxiv.1908.09203**

Thorpe, A., Friedman, J., Evans, S., Nesbitt, K., & Eidels, A. (2021). Mouse movement trajectories as an indicator of cognitive workload. *International Journal of Human-Computer Interaction*, *38*(15), 1464–1479. **https://doi.org/10.1080/10447318.2021.2002054**

Thuraisingham, B., Biggio, B., Freeman, D. M., Miller, B., Sinha, A., Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected. 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX.

Tynjälä, P. (1998). Writing as a tool for constructive learning: Students' learning experiences during an experiment. *Higher Education*, *36*(2), 209–230. **https://doi.org/10.1023/a:1003260402036**

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

Valacich, J. S., Jenkins, J. L., & Čišić, D. (2022). Digital behavioral biometrics and privacy: Methods for improving business processes without compromising customer privacy. International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.

Waes, L. V., & Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to l1 and l2. *Computers and Composition*, *38*(A), 79–95. **https://doi.org/10.1016/j.compcom.2015.09.012**

Wang, J., Liu, S., Xie, X., & Li, Y. (2023). Evaluating AIGC detectors on code content. *arXiv*. **https://doi.org/10.48550/arXiv.2304.05193**

Wei, A., Zhao, Y., & Cai, Z. (2019). A deep learning approach to web bot detection using mouse behavioral biometrics. 14th Chinese Conference, CCBR, Zhuzhou, China.

Weinmann, M., Valacich, J. S., Schneider, C., Jenkins, J. L., & Hibbeln, M. (2022). The path of the righteous: Using trace data to understand fraud decisions in real time. *MIS Quarterly*, *46*(4), 2317–2336. **https://doi.org/10.25300/misq/2022/17038**

Wojnowicz, M. T., Ferguson, M. J., Dale, R., & Spivey, M. J. (2009). The self-organization of explicit attitudes. *Psychological Science*, *20*(11), 1428–1435. **https://doi.org/10.1111/j.1467-9280.2009.02448.x**

Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., & Testrow, C. (2022). The death of the short-form physics essay in the coming AI revolution. *arXiv*. **https://doi.org/10.48550/arxiv.2212.11661**

Zaitsu, W., & Jin, M. (2023). Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through japanese stylometric analysis. *arXiv*. **https://doi.org/10.48550/arXiv.2304.05534**

Appendix

Figure A1 contains the set of instructions used by the system to derive the follow-up questions from ChatGPT (GPT-4, Mar 23 Version). Figure A2 contains the essay prompt used during the experiment, both by the participants for the self-authored essays as well as to prompt ChatGPT for the essays used in the AI-generated condition. Figure A3 contains a screenshot of the follow-up question page in the prototype system used in the study.

I would like your help generating a similar set of 4 questions for a short essay. The 4 questions I would like for each essay are:

1. A multiple-choice question about one main theme in the content of the essay.
2. A multiple-choice question about a certain phrase used in the content of the essay.
3. A true/false question about one of the arguments made in the essay.
4. A true/false question about a phrase used in the content of the essay.

The text of the essay is found below:

[essay text]

Figure A1. ChatGPT Prompt Used for Automated Follow-Up Question Generation

Explore the role of personal biases in shaping your perception and treatment of others in your daily life. Reflect on specific instances when your biases may have influenced your interactions, and discuss the potential consequences of such biases on interpersonal relationships. How can you work towards recognizing and mitigating these biases to foster more inclusive and empathetic connections with others?

Figure A2. Essay Prompt Used in the Study

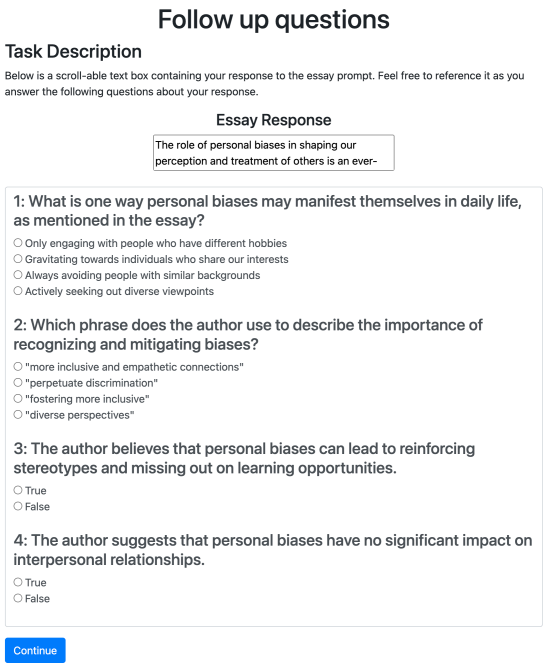

Figure A3. Follow-Up Question Interface