Clicks Don’t Lie: Inferring Generative AI Usage in Crowdsourcing through User Input Events

Arjun Patel, Phoebe Liu
Appen
Sunnyvale, CA, USA
apatel@appen.com, cliu@appen.com

Abstract

Development of generative AI technology has dramatically reduced the cost to generate text, speech, and audio. This is problematic for crowdsourcing data platforms who may be required to demonstrate authenticity of human authored data. Current methods are still not robust yet for mitigation of AI generated content. In this paper, we conducted a study to understand crowd worker behavior during a typical prompt-response generation task and demonstrated that analysis of keystroke and mouse events may yield informative features for measuring risk of AI tool usage during a task. More work is needed to understand the degree of the effect and to verify the results.

Introduction

The rapid development of publicly available generative AI tools is leading to a proliferation of AI generated text, audio, and images across the Internet. While the benefits of these technologies cannot be understated, early work has posited that crowd workers are already using such tools to submit data labeling work, creating novel issues for tasks that require artisanal purely human-generated outputs. (Veselovsky et al. 2023).

Prior work has explored various methodologies for detecting AI generated text. Firstly, generated text has distinguishing characteristics from human written text, and linguistic syntactical features can be extracted for analysis (Guo et al 2023). These features can be used to build classifiers to estimate risk of generated text. Watermarking attempts to adjust the generated text of a given model to make it easily identifiable, but relies on having access to the source model (Kirchenbauer et al. 2023)**.** However, all these methods have been shown to fail using trivial paraphrasing attacks (Sadasivan et al 2023)**.** It is not clear if any model can detect AI generated text as large language model performance improves in the long run (Tang et al 2023).

We propose that the lack of robust AI detection models is because of a focus on the *content* of the generated text, rather than the *process* of creating it. These features include keystroke, mouse movement, and time related event-based features. Prior work in the keystroke analysis literature has shown that tasks with varying levels of cognitive load influence keystroke behavior (Conijn et al. 2019). As generative LLMs and search engines aim to make text and knowledge more widely accessible, it stands to reason that their use would reduce cognitive load and therefore affect keystroke distribution.

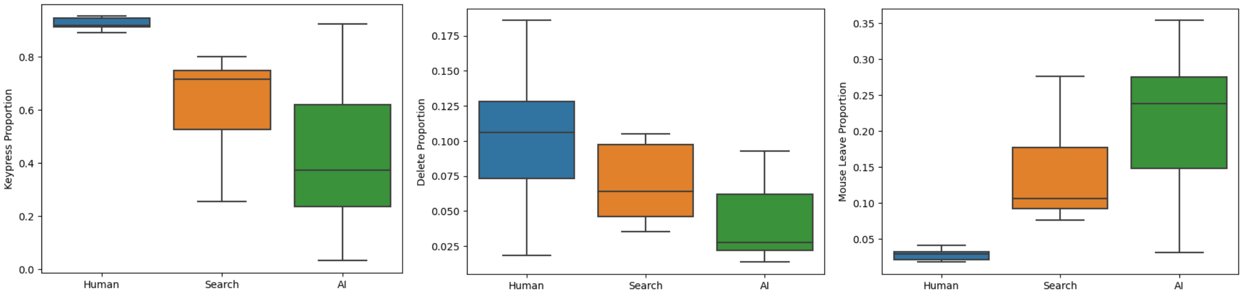
In this paper, we demonstrate early results from experiments designed to collect prompt-response pairs from crowd workers instructed to create responses under three conditions: using just themselves, search tools, or generative AI tools. Through this dataset, we observe novel behavioral and content-related features that could be used to detect the presence of AI-generated content in datasets. We hypothesized the keystroke behavioral patterns for crowd workers will correlate with the conditions thereby demonstrating their use as features. If successful, these features will allow for scalable ways of inferring generative AI usage in crowdsourced text dataset.

Experimental Design

We designed a set of prompt-response annotation tasks to observe the differences in keystroke and mouse movement behavior in writing responses exhibited by crowd workers operating under three different conditions:

Human: crowd workers wrote without any outside assistance or tools.

Presented at the Works-In-Progress and Demonstrations track, AAAI HCOMP 2023.



(a) Keypress prop. (b) Delete prop. (c) Mouse movement prop.

Figure 1: Distribution of behavior patterns across three experimental conditions. Delete proportion distribution is overall lower in the Search and AI conditions versus the Human case, while mouse movement varies the most in the AI condition.

Search: crowd workers wrote with the use of search engines like Google.

AI: crowd workers wrote with the use of freely available generative AI tools.

The conditions were selected to mimic the gamut of behaviors exhibited during annotation. Under the Human condition, respondents were explicitly instructed not to use external tools or generative AI tools. The Search condition is included to act as a baseline on external tool use, as some crowdsourced work needs to use search engines for information gathering and may look like using generative AI tools through certain behavior features (e.g., copy-paste text from external sources). Additionally, we hypothesized that under the Search condition, crowd workers would paraphrase text, whereas under the AI condition, crowd workers would copy-paste generated text. If features we extract allow us to differentiate not only between the AI and Human conditions, but also the AI and Search conditions, then it can be concluded that the features are indicative of AI tool use.

Each condition had a corresponding training task that needed to be completed before the actual task to validate crowd workers' understanding of the task. An additional calibration task was included prior to the experimental conditions which mandated crowd workers to type a given prompt word-by-word manually to estimate typing speed and baseline behavior. This amounted to 7 tasks, including one calibration, 3 training, and 3 evaluation tasks.

Each crowd worker wrote responses under all experimental conditions, to compare behavior on a crowd worker level. We launched these as annotation jobs on our data annotation platform.

Task Dataset We curated a set of prompts from the open-source Dolly prompt dataset, made by Databricks (Conover et al., 2023). We supplemented the prompts using prompts created by ourselves, to ensure a good mix of statements that would require creativity, as well as prompts that would be easier to use external tools with (e.g. factual question answering).

We selected 9 training prompts to be divided amongst the three conditions to assist in helping crowd workers learn how to complete the tasks. This procedure amounted to a set of 60 prompts per crowd worker.

Behavioral Data Collection We collected event-based behavioral data while crowd workers wrote responses. Specifically, we collected keystroke and mouse movement information while a user interacted with a task, which allowed for monitoring of response creation as well as external tool use. We estimated several keystroke and mouse movement-based features on this data and present three in this study:

Keypress Proportion: ratio of observed events that are letter keypresses.

Delete Proportion: ratio of observed events that are deletions, defined as hitting the backspace or delete keys.

Mouse Movement Off Screen Proportion: ratio of observed events that record the mouse moving out of the application window.

Participant Recruitment We recruited volunteers on internal company forum pages. A total of 252 non-training responses were created from 7 participants. The order of experimental conditions was randomly assigned to the participant. We plan to continue data collection work in the future by recruiting additional participants for more robust conclusions on behavior and content feature in the response data.

Instructions Each experimental condition came with detailed instructions on the response generation task. Crowd workers were instructed to create responses that were at least 150 words in length. All experimental conditions instruct participants to create responses as if they were a helpful assistant. Crowd workers were also requested to screen record themselves while completing all tasks. The recordings primarily were used to corroborate results derived from feature extraction, and to inspect special cases of crowd worker behavior. The Search and AI conditions include additional instructions to teach crowd workers how to use articles found with search engines and how to use generative AI tools to create responses. Both conditions included workflow sections where crowd workers were requested to write a few sentences about the steps taken to create the response in question. All instructions included examples to demonstrate expectations for proper response creation. Participants could contact the experimenter for technical assistance, but further advising on the task beyond the instructions was not permitted to avoid influencing behavioral patterns.

Results and Discussion

Behavioral Analysis The distribution of three behavioral features from each experimental group are displayed above in Figure 1.We collected 143,190 events from the Human condition, 102,011 events from the Search condition, and 94,936 events from the AI condition features were aggregated on the condition level, yielding N=7 observations per condition.

The Human experimental condition has an extreme tight keypress proportion distribution than either treatment, with median 0.918 and IQR 0.0327. The AI condition has median keypress proportion of 0.374, which is less than half than the Human condition. From a Friedman test across all three conditions, the difference in the keypress proportion distributions was significant after a Bonferroni corrected significance level of 0.0055. (χ^2 =11.14, p=0.004) This indicates that usage of external tools by crowd workers reduces the need of typing on their own, possibly using copy paste.

These results can be corroborated by inspecting behavior in the delete proportion distribution and mouse movements. When using AI tools, crowd workers deleted less often (median 0.028, IQR 0.040) and moused out of application window more (median 0.238, IQR 0.085). The larger delete proportion distribution for the Human condition (median 0.106, IQR 0.055) implies revision behavior. The Search condition had higher median delete proportions and keypress proportions compared to the AI condition, implying paraphrase behavior and copy-paste behavior for Search and AI respectively. Finally, the tight spread of the mouse leave proportion distribution for humans (median 0.0286, IQR 0.011) implies that moving offscreen strongly correlates with tool usage. A Friedman test across all three conditions was significant as well for both delete proportion distribution (χ^2 =12.29, p=0.002), and mouse movement off screen proportion (χ^2 =12.29, p=0.002).

Content Analysis We conducted a brief qualitative analysis of the responses created by crowd workers to understand how content varied with respect to experimental condition. When writing with search or AI tools, content had distinct structures to organize information, such as numbered lists or bullet points. When writing without these tools, responses were more freeform and had occasional typos.

Conditions	Mean	Q1 25th perc.	Q2 50th perc.	Q3 75th perc.
Human (N=84)	189.68	162.0	177.0	198.25
Search (N=84)	246.21	166.75	194.5	266.25
AI (N=84)	268.03	172.75	242.5	373.50

Table 1: Word Count across experimental conditions

Response length varied across experimental conditions, displayed in Table 1. We estimated response length by calculating word count of every response written by the crowd workers, totaling N=84 per condition and 252 total responses. Surprisingly, responses for the AI and Search conditions tended to be longer than human written responses. We hypothesize human writers wrote text until the word count is reached, whereas crowd workers using AI or search may generate or grab available text, then choose later to edit down. The noticeable difference in average and median word count implies reduced curation behavior in response generation amongst crowd workers. Further study is needed to understand the depth of the differences in content created by each crowd worker under each experimental condition.

Taken together, our early results demonstrate some evidence for differences in keystroke behavior for crowd workers using external tools such as AI or search, and when left to their own devices to construct responses. However, it is difficult to conclude if use of generative AI tools can be distinguished from the use of search tools.

Discussion and Future Works Our early results are encouraging and demonstrate that differences in behavioral keystroke and mouse data may be observed when crowd workers create responses with and without AI and external tool access. We plan to collect data from additional crowd workers to better estimate the degree to which these effects are robust across experimental conditions. We will look toward extracting features from the content created by the crowd workers, to build classifiers operating jointly on such data. Verification of these results will lay the groundwork for monitoring and mitigating the use of AI tools in crowdsourced data.

Acknowledgements

We thank our colleagues for their time on setting up and participation of the experiment. We thank Dylan F. Glas for helpful feedback.

References

Conijn, R., Roeser, J., & Zaanen, M. van. 2019. Understanding the keystroke log: The effect of writing task on keystroke features | Reading and Writing. *Reading and Writing*, *32*, 2353–2374. doi.org/10.1007/s11145-019-09953-8

Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., & Xin, R. 2023. *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM* [dataset].

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. 2023. How close is chatgpt to human experts? Comparison corpus, evaluation, and detection. :2301.07597.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. 2023. A Watermark for Large Language Models. arXiv:2301.10226.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. 2023. Can AI-Generated Text be Reliably Detected? arXiv:2303.11156.

Tang, R., Chuang, Y.-N., & Hu, X. 2023. The Science of Detecting LLM-Generated Texts. arXiv:2303.07205

Veselovsky, V., Ribeiro, M. H., & West, R. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv:2306.07899.