Visual representation of co-authorship with GPT-3: Studying human-machine interaction for effective writing

Antonette Shibani*
University of Technology Sydney, Sydney, Australia
antonette.shibani@uts.edu.au

Ratnavel Rajalakshmi †
Vellore Institute of Technology, Chennai, India
rajalakshmi.r@vit.ac.in

Faerie Mattins
Vellore Institute of Technology, Chennai, India
faeriemattins.r2019@vitstudent.ac.in

Srivarshan Selvaraj
Vellore Institute of Technology, Chennai, India
srivarshan.2019@vitstudent.ac.in

Simon Knight
University of Technology Sydney, Sydney, Australia
simon.knight@uts.edu.au

ABSTRACT

With the recent release of Chat-GPT by OpenAI, the automated text generation capabilities of GPT-3 are seen as transformative and potentially systemically disruptive for higher education. While the impact on teaching and learning practices is still unknown, it is apparent that alongside risks these tools offer the potential to augment human intelligence (intelligence augmentation, or IA). However, strategies for such IA, involving partnership of tool-human, will be needed to support learning. In the context of writing, an investigation of potential approaches is needed given empirical data and studies are currently limited. We introduce a novel visual representation *CoAuthorViz* to examine keystroke logs from a writing assistant where writers interacted with GPT-3 writing suggestions to co-write with the machine. We demonstrate the use of our visualization by exemplifying different kinds of writing behaviour from users writing with GPT-3 support and derive metrics such as their usage of GPT-3 suggestions in relation to overall writing quality indicators. We also release the materials open source to further progress our understanding of desirable user behaviour when working with such state-of-the-art AI tools.

Keywords

Keystroke analysis, Visualization, Writing analytics, GPT-3, Language models, Coauthor, Artificial Intelligence, Chat GPT, Generative AI, CoAuthorViz

1. INTRODUCTION

Changes due to evolving technology is a constant across sectors, but certain technologies have had a profound effect on redefining educational strategies. In academic writing, technologies such as word processing that digitised writing from paper-based formats, the internet and cloud that enabled widespread communication and collaboration, and computational linguistics and Natural language processing that enabled real-time support and automated feedback are key innovations that led to transformations in writing practices and the curriculum [20]. With the recent release of large language models such as Generative Pre-trained Transformer 3 (GPT-3), automated text generation and the use of Artificial Intelligence (AI) to support writing are touted as the next writing transformation.

The open release of powerful tools such as ChatGPT[1] for GPT-3 made visible the dramatic capabilities of generative AI - anyone can write a prompt to ChatGPT in plain English providing instructions, and the tool can generate well-written texts replicating human knowledge. The potential harms and disruptions it can cause to traditional writing curricula have been discussed widely, including concerns about academic integrity, but little is known about how these technologies can best work in practice in partnership with human writers. One such work involves CoAuthor, a human-AI collaborative writing dataset that was created from machine-in-the-loop argumentative and creative writing with writers using automated text suggestions generated from GPT-3 as real-time feedback [21]. The dataset consists of keystroke-level data captured from the writer's typing and is predominantly used by writing analytics and psycholinguistic researchers to learn about cognitive processing. In this paper, we introduce a visual graph CoAuthorViz to aid the analysis of such log data to study human-AI collaboration in writing using more interpretable representations. The intended audience for the CoAuthorViz is researchers who can use the visualisation and related metrics to study the phenomena of working in partnership with AI tools for writing.

2. USING GENERATIVE AI FOR WRITING

Research in the last few decades has seen increasing evidence of the effectiveness of automated writing evaluation (AWE) systems in supporting writers develop their academic writing skills [41] [24] [18]. Automated writing feedback tools

*Corresponding author
†Corresponding author

Author's manuscript (accepted version)
Cite as: A. Shibani, R. Rajalakshmi, F. Mattins, S. Selvaraj, and S. Knight (in press). Visual representation of co-authorship with GPT-3: Studying human-machine interaction for effective writing. In Proceedings of the 16th International Conference on Educational Data Mining, pages xx, India, July 2023. International Educational Data Mining Society.

[1]https://openai.com/blog/chatgpt/

provide scalable and innovative computer-based instruction in linguistic, domain, or mixed orientations [14], often targeting specific writing features of interest [18]. However, the most recent advancements in generative AI include the use of large language models for writing, which might fundamentally change how writers learn to write in the future.

Generative Pre-trained Transformer 3 (GPT-3) is a large language model trained on internet data that can automatically generate realistic text [32]. It is a deep learning neural network with over 175 billion machine learning parameters that makes its machine-generated text convincingly similar to what humans write. When a user provides an input text in natural language, the system analyzes the language and predicts the most likely output text. While the beta release of GPT-3 by OpenAI came about much earlier (June 2020), the most recent release of Chat-GPT for public testing in November 2022 has triggered strong reactions to its implications for human writing. Discussions are a mix of initial conversations and scholarly literature given the recency of the topic.

Firstly, we note the potential for GPT-3 usage in writing contexts through applications implemented and evaluated in practice. Automated text generation is the most common application of GPT-3 for generating formal forms of writing, but the model also has the capability to generate poetry, play chess, do arithmetic, translations, and role play, and write code based on user requirements [8]. One use case was seen in 'sparks', sentences generated by the AI writing assistant to inspire writers to create scientific content [13]. The purpose was to aid writers with crafting detailed sentences, providing interesting angles to engage readers, and demonstrating common reader perspectives.

Multiple Intelligent Writing Assistants have made use of GPT-2 and GPT-3 language generation capabilities to help writers develop their content. Examples include writers making integrative leaps in creative writing with multimodal machine intelligence [36], a web application called Wordcraft where users collaborated with a generative model to write a story [42] and a system providing automated summaries to support reflection and revision beyond text generation [9]. A larger evaluation engaging over 60 people to write more than 1,440 stories and essays was performed using CoAuthor, where the interactions between the writer and the GPT-3 suggestions were also captured using keystroke logging [21]. Another writing task that can now be supported by intelligent agents is revision. In the human-in-the loop iterative text revision system called Read, Revise, Repeat (R3), writers interacted with model-generated revisions for deeper edits [11].

However, there are known problems in large language models such as the generation of factually false hallucinations or contradictory information that can exacerbate disinformation [27], bias and immorality arising from human subjectivity [25] and the lack of diversity in its outputs [16]. Perhaps, the more complex problems arising from GPT-3 content relate to social factors such as how it interferes with existing systemic practices affecting people and policies in the real world. There is widespread fear that the automatically generated content amplifies academic dishonesty which is already prevalent in the education sector providing easy means for students to cheat with plagiarism [29]. This is particularly a threat to online learning where the real identity of the writer is hard to discern.

Despite the concerns, there is also hope that these tools might accelerate learning and induce creativity. Like multiple technologies that came before it, some consider these AI tools to be yet another example of humanity's inefficiency dealing with something new that throws their normality into disarray [1]. There is an increasing push to rethink assessments, so we move away from setting assignments that machines can answer towards assessment for learning that captures skills required in the future[33], [39] and students using GPT-3 as part of the curriculum to enhance their learning [30]. There is emerging work such as the launch of 'GPT-2 Output Detector'[2] to identify content authored by Chat-GPT, but with a caveat of having a high false positive rate - dismissing original content as plagiarism could be worse than accepting plagiarised content from the tool for writing assessment. This can be particularly harmful to non-native English writers as GPT detectors may unintentionally penalize writers with constrained linguistic expressions due to their in-built biases [23].

Similar tools and technologies will evolve over time and many students already use AI-based writing tools such as Quillbot[3] as part of their writing practices, so there is an opportunity to investigate how to collaborate with them effectively rather than banning or abolishing them completely [28]. GPT-3 applications where a human stays in the loop are considered safer and the way forward, where the writer uses the machine to augment their writing by utilising its unique capabilities and acknowledges its use [8]. The varied roles AI can take: as an editor, co-author, ghostwriter, and muse have been identified [17], with particular interest towards co-authoring that helps writers develop their writing skills through human-AI partnership [21] that we explore in the current work. Early explorations of two new types of interactions with generative language models show how writers can keep control of their writing by manipulating the auto-generated content [3]. More recent work also involves building a collaborative language model that imitates the entire writing process such as writing drafts, adding suggestions, proposing edits, and providing explanations for its actions, and not just generating the final result [31]. These align with the Intelligence Augmentation (IA) paradigm where human and artificial intelligence work together as a symbiotic system [43], and is of relevance to education where new technology can augment existing teaching and learning strategies [19]. In these cases of co-writing, it is useful to determine the most efficient ways for writers to interact with GPT-3 for optimal partnership and IA, and methods to analyse such behaviour are discussed next.

3. STUDYING WRITING BEHAVIOUR USING KEYSTROKE ANALYSES

Writing is a complex cognitive process that involves recursive and interleaving activities such as planning, translating, reviewing, and monitoring by the writer [12]. Researchers

[2]https://huggingface.co/roberta-base-openai-detector
[3]https://quillbot.com/

use different approaches and data to study the writing process that informs user behaviour. While early work typically relied on resource-intensive manual observation and coding of writing behaviour, computational analysis techniques and log data are now used to study learning processes at scale [10]. These help uncover new patterns from fine-grained information about the learner's writing behaviour through non-obtrusive stealth measurements and keystroke-level data capturing [2][26].

Keystroke logging is a method of automatically capturing data on a user's typing patterns as they write. Analysis of such data can be used to gain insight into various aspects of writing behavior, including typing speed, error rate, and the use of specific keyboard shortcuts. Keystroke analysis has been used for biometric authentication using keystroke dynamics [38], measuring text readability using scroll-based interactions [15], and predicting writing quality for feedback [6]. However, there often exists a disconnect between keystroke level logs and useful insight on cognitive processes that can be derived from it as the data is too fine-grained. Complementary techniques such as eye-tracking and thinking-aloud protocols are often used in combination to capture additional context on the writing [22] [40]. In addition, newer graphic and statistical data analysis techniques offer new perspectives on the writing process.

Visual representations provide a useful starting point to study the complex interactions between sources and writers. Network analysis and graph representations have been used by writing researchers to visualise the temporal development of ideas and links between multiple sources during editing and revising a writer's document [22] [4]. A multi-stage automated revision graph was used to study the evolution of drafts in the revision process that led to the final product and students' interaction with automated feedback based on their frequency of requests [34]. In other work that investigated collaborative writing processes, a revision map was created to represent the joint development of ideas by a group of authors [37]. Such visualisations provide new ways of looking at data to uncover interesting insights and patterns of user behaviour from writing scenarios.

4. OUR WORK

In our work, we introduce a novel visualization called *"CoAuthorViz"* to represent writing behaviours from keystroke logs of users in the CoAuthor dataset (described next). We demonstrate how CoAuthorViz can be used by writing researchers to study co-authorship behaviours of writers interacting with GPT-3 suggestions to co-write with the machine, and investigate metrics derived from such interaction with relation to overall writing quality indicators. We discuss how the work can be extended further to study effective forms of co-authorship with GPT-3 and other AI writing assistants.

4.1 Dataset used

Data for this study comes from the CoAuthor dataset [21] which consists of a total of 1445 writing session data in jsonl format, including 830 creative writing (stories) and 615 argumentative writing (essays) sessions. The dataset contains keystroke-level interactions in a writing session logging 17 events: event name, event source, text delta, cursor range, event timestamp, index of event, a writing prompt to start with, current cursor location, suggestions from GPT-3, number of suggestions to generate per query, the maximum number of tokens to generate per suggestion, sampling temperature, nucleus sampling, presence penalty, and frequency penalty. Descriptions for each variable are provided in the original article [21], and a sample set of rows from the dataset is shown in Table 1. Replays of each individual writing session are also made available on the project website[4].

The writer is provided with an initial prompt by the researchers instructing them to write on the assigned topic, and are required to continue their writing session on their own or with the assistance of GPT-3 sentence recommendations. The writers receive up to five sentence suggestions when a GPT-3 call is made and can do so at any point during their writing sessions - suggestions provided by GPT-3 can be partial or full sentences.

4.2 CoAuthorViz Description

We develop CoAuthorViz to represent co-authorship behaviours of users interacting with GPT-3 suggestions at a sentence-level. This visual representation makes it easier to interpret co-writing processes in comparison to more fine-grained keystroke level logs that capture individual characters and mouse movements. The visualization highlights key actions made by a writer when working with GPT-3 suggestions such as choosing to accept the suggestion as it is, accept suggestion and edit it further, or reject the suggestion and continue writing on their own - these events recorded as part of the keystroke logs can provide significant insight into how AI writing assistants are taken up by writers in practice. Our work is inspired by Automated Revision Graphs previously used for visualising student revision in writing drafts, transferred to the context of co-writing with AI [34].

CoAuthorViz performs sentence-level analysis to visualise interactions between the writer and GPT-3. Three different shapes — circle, triangle, and square represent authorship - the initial prompt provided by researchers is shown as a black circle and ranges from 1 to 9 sentences each (the writer is instructed to base the rest of their writing around it). Since the writer's actual writing starts from the last sentence of the initial prompt, our visualization starts from here. Text entered by the writer is displayed as a gray square, and text written by GPT-3 is displayed as a black triangle. Text modified by the writer after obtaining a GPT-3 suggestion from GPT-3 is displayed as a square overlapping a gray triangle. Empty GPT-3 calls illustrating scenarios where the writer requests for and obtains GPT-3 suggestions, but chooses to ignore them are shown as white triangles. Dotted lines between the shapes indicate a sequence of actions at a sentence level to improve the readability of the visualization and do not have additional meaning.

An example of CoAuthorViz is illustrated in Figure 1. Here, most of the writing was done by the writer independently (see sentences 9, 13, 14, 16-18 with black squares), and even when text from GPT-3 was provided (sentences 8, 10-12, 15 with GPT-3 written text), they went on to add additional text themselves. We also see places where a GPT-3 call was

[4]https://coauthor.stanford.edu/browse/

Table 1: Examples from the dataset with selected rows and columns

eventName	eventSource	textDelta	currentCursor	currentSuggestion
text-insert	user	'ops': ['retain': 2017, 'insert': 'a']	2017	[]
text-insert	user	'ops': ['retain': 2018, 'insert': '\n']	2018	[]
suggestion-get	user	NaN	2019	[]
suggestion-open	api	NaN	2019	['index': 0, 'original': 'smiled at him, and he walked over to her table.', 'trimmed': 'Priscilla smiled at him, and he walked over to her table.', 'probability': 1.1132658066910296e-05, 'index': 1, 'original': 'man walked over to her table and sat down.', 'trimmed': 'The man walked over to her table and sat down.', 'probability': 1.0074578955483344e-07]
suggestion-hover	user	NaN	2019	[]
suggestion-select	user	NaN	2019	[]
suggestion-close	api	NaN	2019	[]
text-insert	api	'ops': ['retain': 2020, 'insert': 'Priscilla smiled at him, and he walked over to her table.']	2077	[]

made, but the suggested texts were dismissed and not used by the writer (white triangle in sentences 9, 13, 14, 16-18).

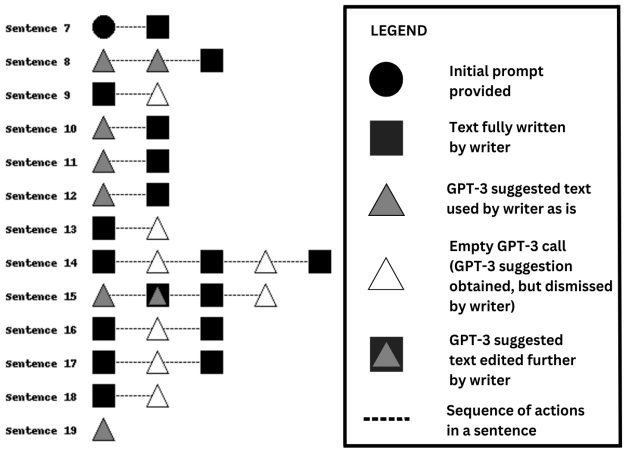


Figure 1: Example of a CoAuthorViz with descriptors

CoAuthorViz generates a simple visualization to represent co-authorship with GPT-3 from relatively complex, fine-grained keystroke-level data. It reveals insights on the writer's frequency of autonomous writing without AI assistance and their usage, dismissal, and modification of GPT-3 text suggestions provided. These can be used to inform the study of user behaviour when engaging with AI writing assistants such as GPT-3.

4.3 Technical Implementation

The lack of standards in capturing and analysing keystroke data is an identified challenge in this kind of research [22]. To this end, we provide a detailed explanation of the construction of CoAuthorViz and release the materials open source (including the scripts and plots generated) to help facilitate knowledge exchange among research groups [Github link].

The keystroke log is first read by iterating over all the tracked events. Text at any given keystroke is rebuilt from the log using events and cursor positions. This is done by maintaining a text buffer during the entire process providing the current state of the document - when a text insertion keystroke is encountered, the corresponding text is added to the buffer; when text deletion occurs, the corresponding characters are deleted from the buffer; cursor positions are used to identify the locations in the buffer when such events occur. The events and their corresponding text buffers are grouped by the number of sentences in the buffer, providing a sequence of all events at the sentence level. From this sentence-level event sequence, the following steps are performed to define key constructs of interest:

1. **GPT-3 Suggestion Selection:** "suggestion-get" events that are succeeded by a "suggestion-select" event are identified as GPT-3 calls where the writer obtained a suggestion and made use of it. Related "suggestion-open", "suggestion-hover", "suggestion-down", "suggestion-up", and "suggestion-reopen" events are removed as they are all indicative of the same event - author choosing from the GPT-3 suggestions. "text-insert" events occurring immediately after the "suggestion-select" events are removed as they also signify the insertion of GPT-3 suggestion selected by the writer

2. **Empty GPT-3 Call:** "suggestion-get" events that do not have a succeeding selection event are identified as empty GPT-3 calls where the author did not incorporate any suggestion provided by GPT-3

3. **GPT-3 Suggestion Modification:** Any "cursor-backward", "cursor-select" or "text-delete" events immediately suc-

ceeding a "suggestion-select" event, but without any "text-insert" event in between are perceived as modifications done by the author to the GPT-3 suggestion they chose. All cursor movement events, text deletion events and "suggestion-close" events are removed

4. **User Text Addition:** Consecutive "text-insert" events are grouped for piecing together text written by the writer

Metrics are calculated by counting the key events in relation to GPT-3 calls, and authorship in sentences. The sequence of key identified events from the above constructs is generated as a visualisation using the Pillow package [5]. The full implementation runs on a Python notebook, and is represented in Figure 2.

Figure 2: Steps in the construction of CoAuthorViz

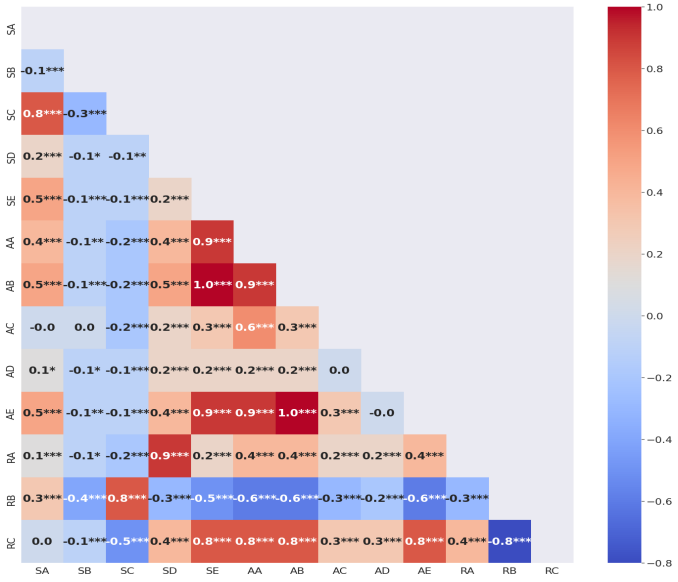


Figure 3: Correlation matrix with statistical significance of CoAuthorViz metrics

5. FINDINGS AND DISCUSSION

In this section, we discuss the main findings from our visualisation and examine sample cases in detail demonstrating the application of CoAuthorViz for researching writing.

5.1 Analysis of CoAuthorViz metrics

A summary of the key events noted in CoAuthorViz is generated for each writing session providing tangible metrics that can be studied along with the visualization. Three types: Sentence level, API-based, and Ratio metrics are provided - see Table 2 for the summary statistics of these metrics. Each of the 1445 writing sessions in the CoAuthor generates a total number of sentences ranging from 11 to 78, and an average of 29 sentences in the final writing. The initial prompt in a writer's writing session can vary from 0 to 9 sentences, with an average of around 4. The number of sentences in the initial prompt is 0 in cases where the writer deletes the initial prompt and rewrites it from scratch.

Metrics on the number of sentences written entirely by the writer, GPT-3, or a combination of the writer and GPT-3 are populated. Additional metrics include the frequency of using GPT-3 suggestions with and without modification, as well as the number of instances where a GPT-3 call was made but the suggestion was rejected, likely because the writer was dissatisfied with the suggested texts. Ratios were also calculated to characterize GPT-3 versus writer authoring in relation to the total number of sentences generated in a writing session.

From the summary statistics table in Table 2, we derive insights on the usage of GPT-3 across the 1445 writing sessions. The average number of times GPT-3 calls were made (AA) was 12.5 but varied widely across the sessions (SD = 9.2) with a minimum of 0 and a maximum of 65. Similarly, there was high variance in the number of times a GPT-3 suggestion was incorporated (AB) ranging from 0 to 47 (M = 8.9, SD = 7.4), and the number of times a GPT-3 suggestion was accepted as it is (AE) (M = 7.3, SD = 7.2). Total GPT-3 usage in their sentences (RC) was calculated from the ratio of the sum of sentences using GPT-3 suggestion, and the total number of sentences in the writing ranged from 0 to 0.87 (M = 0.3, SD = 0.2). The ratio of the number of times the suggestion is rejected to the number of times the author calls for GPT-3 (AC/AA) indicates that suggestions made by GPT-3 were rejected 29.31% of the time, and suggestions were accepted as is 58% of the time (AE/ AA).

We also calculate correlation to examine relations within CoAuthorViz metrics. Figure 3 shows the matrix of Pearson correlation coefficients (CC) for each pair of metrics in the summary table. The statistical significance of each correlation is indicated by the number of asterisks adjacent to the value (in order of increasing significance: p-value < 0.05 is flagged with one star (*), p-value < 0.01 is flagged with 2 stars (**), and p-value < 0.001 is flagged with three stars (***)). Related pairs of metrics such as AA and AE have high CC ranging from 0.8 to 1.0 because the metrics are computed from similar underlying values such as the number of GPT-3 calls made.

A negative correlation (r = -0.6) was found between the autonomous writing indicator (RB) and the number of times a GPT-3 suggestion is accepted as is (AE). Similarly, writers having high GPT-3 dependence indicators had more sentences completely authored by GPT-3 (r = 0.9) suggesting their reliance on GPT-3 for writing without making further edits. On the contrary, writers who had a high number of sentences completely authored by them preferred to write their sentences independent of GPT-3 and hence tended to have high autonomous writing indicators (r = 0.8). The total number of GPT-3 calls made positively correlated to both the number of times its suggestion was accepted as is (r =0.9) and the number of sentences co-authored by GPT-3 and the writer (r = 0.9).

Table 2: Summary Statistics of CoAuthorViz metrics

Type	Metrics (for sample size n=1445)	Mean	Median	Standard Deviation	Min	Max
Sentence Metrics	Total number of sentences **(SA)**	28.962	27	10.388	11	78
	Number of sentences in initial prompt **(SB)**	4.421	4	2.390	0	9
	Number of sentences completely authored by the writer **(SC)**	16.242	15	9.535	0	64
	Number of sentences completely authored by GPT-3 **(SD)**	0.685	0	1.886	0	22
	Number of sentences co-authored by GPT-3 and writer **(SE)**	7.613	6	5.953	0	42
API Metrics	Total number of GPT-3 calls made **(AA)**	12.531	10	9.204	0	65
	Number of times GPT-3 suggestion is accepted **(AB)**	8.857	7	7.424	0	47
	Number of times writer rejected GPT-3 suggestion **(AC)**	3.673	3	3.530	0	24
	Number of times GPT-3 suggestion is modified **(AD)**	1.586	1	1.796	0	10
	Number of times GPT-3 suggestion is accepted as it is **(AE)**	7.271	5	7.233	0	47
Ratio Metrics	GPT-3 dependence indicator - Number of sentences completely authored by GPT-3 / Total number of sentences **(RA)**	0.021	0	0.054	0	0.611
	Autonomous writing indicator - Number of sentences completely authored by writer / Total number of sentences **(RB)**	0.541	0.564	0.205	0	0.962
	Total GPT-3 usage in sentences [(SD+SE)/SA] **(RC)**	0.285	0.25	0.183	0	0.867
TAACO Metrics	lemma_ttr **(LTTR)**	0.401	0.4	0.054	0.240	0.585
	adjacent_overlap_all_sent **(AOAS)**	0.212	0.210	0.043	0.076	0.389
	adjacent_overlap_all_para **(AOAP)**	0.256	0.258	0.090	0.0	0.863
	lsa_1_all_sent **(LSA1AS)**	0.309	0.305	0.094	0.094	0.688
	lsa_1_all_para **(LSA1AP)**	0.477	0.490	0.171	0.0	0.948
	all_connective **(AP)**	0.068	0.067	0.017	0.017	0.131

Table 3: t-test results for TAACO Metrics with alpha value as 0.025 and degree of freedom as 1444.

Metrics	Low GPT-3 usage Group		High GPT-3 usage Group		T-Statistic	P-Value
	Mean	Standard Deviation	Mean	Standard Deviation		
LTTR	0.406	0.052	0.396	0.056	3.592	3.386×10^{-4}
AOAS	0.203	0.040	0.220	0.044	-7.787	1.298×10^{-14}
AOAP	0.259	0.084	0.253	0.096	1.360	1.739×10^{-1}
LSA1AS	0.307	0.094	0.312	0.093	-1.099	2.716×10^{-1}
LSA1AP	0.488	0.161	0.466	0.180	2.432	1.511×10^{-2}
AP	0.068	0.016	0.068	0.017	0.472	6.363×10^{-1}

5.2 Relation between CoAuthorViz metrics and writing features

We additionally analysed the final written texts from the CoAuthor sessions using TAACO to derive indicators of writing quality from language features [7]. Key indicators of lexical diversity, lexical overlap, semantic overlap, and connectedness below were used to derive the metrics, and include descriptions from TAACO on how the metrics are calculated:

- Lemma_ttr (LTTR) - number of unique lemmas (types) divided by the number of total running lemmas (tokens)
- Adjacent_overlap_all_sent (AOAS) - number of lemma types that occur at least once in the next sentence
- Adjacent_overlap_all_para (AOAP) - number of lemma types that occur at least once in the next paragraph
- Lsa_1_all_sent (LSA1AS) - Average latent semantic analysis cosine similarity between all adjacent sentences
- Lsa_1_all_para (LSA1AP) - Average latent semantic analysis cosine similarity between all adjacent paragraphs
- All_connective (AP) - number of all connectives

The above TAACO metrics were used for preliminary analysis of our visualization metrics in relation to writing quality features since the CoAuthor dataset did not contain a quality metric for the text outputs from the writing sessions - the correlation matrix is shown in Figure 4. However, we do not see a significant correlation between any CoAuthorViz metric and automated writing features extracted from TAACO.

We further split session users into two groups based on the number of GPT-3 calls initiated to study potential differences between groups. Sessions with the total number of GPT-3 calls above or equal to the median value were classified as belonging to the high GPT-3 usage group and below median sessions formed the low GPT-3 usage group. We performed a t-test (Findings in Table 3) to compare TAACO metrics between the high GPT-3 usage group (n = 718) and the low GPT-3 usage group (n = 728).

Results suggest that there was a significant difference in

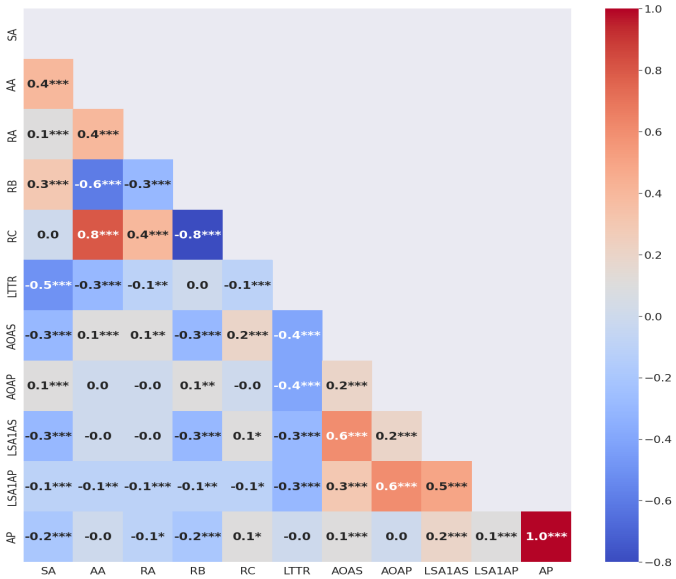


Figure 4: Correlation matrix with statistical significance of CoAuthorViz and TAACO metrics

Lemma type-token ratio (LTTR) between the high usage group (M = 0.396, SD = 0.057) and the low usage group (M = 0.406, SD = 0.053); t(df=1444) = 3.6, p < .005, meaning that writers who accessed GPT-3 less produced a higher proportion of the text that consisted of content words (nouns, lexical verbs, adjectives, and adverbs derived from adjectives) indicating higher lexical diversity. An opposite effect was observed for the TAACO metric Adjacent sentence overlap all lemmas (AOAS) between the high usage group (M = 0.221, SD = 0.045) and the low usage group (M = 0.203, SD = 0.041); t(df=1444) = -7.8, p < .005, suggesting that writings from the high GPT-3 usage group had higher lexical overlaps in adjacent sentences leading to more cohesion.

A significant difference was also observed in Lsa cosine similarity in adjacent paragraphs (LSA1AP) between the high usage group (M = 0.467, SD = 0.18) and the low usage group (M = 0.489, SD = 0.162); t(df=1444) = 2.4, p = .02. Here, writing from the low GPT-3 usage group had a higher semantic overlap exhibiting high average latent semantic analysis cosine similarity between all adjacent paragraphs. A descriptive box plot showing the minimum, maximum, median, lower, and upper quartiles of the three metrics in the high and low groups is shown in Figure 5. No significant difference in group means was noted for the other three metrics (AOAP, LSA1AS, and AP). While the findings indicate effects of high/ low GPT-3 usage in the output writing produced, higher level features are required in order to draw stronger links to writing quality, likely using some form of human assessment in the future.

5.3 Case studies of writer interaction with GPT-3 for co-authorship

We further demonstrate the use of CoAuthorViz to study in detail writer interactions with GPT-3 using example writing sessions. We show three cases from the dataset in Figure 6 showcasing differences in writers' behaviour when working

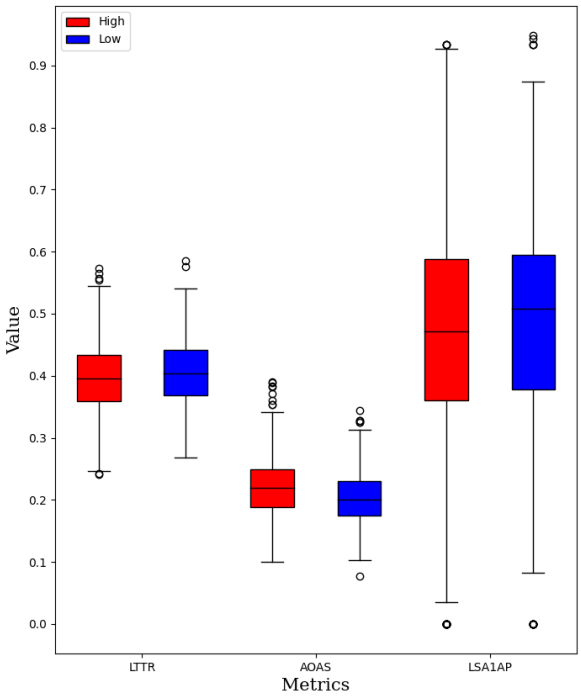
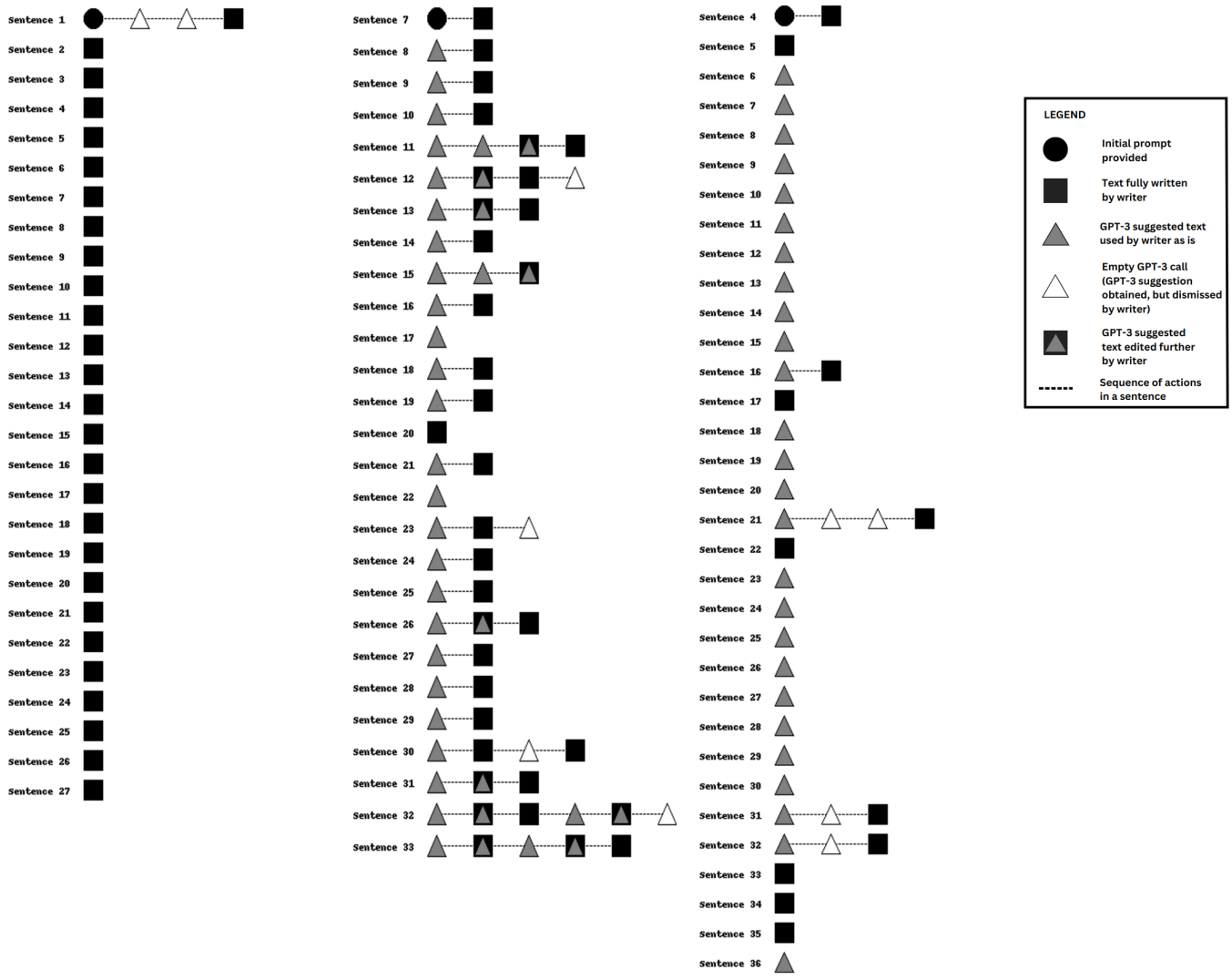


Figure 5: Box plots describing differences in TAACO Metrics for the high and low GPT-3 usage groups

with GPT-3 suggestions on their writing. Metrics from these writing sessions are shown in Table 4.

Table 4: Summary table for the writing session shown in 6.

Metrics	Case-1	Case-2	Case-3
SA	27	33	36
SB	1	7	4
SC	26	1	6
SD	0	2	22
SE	0	23	4
AA	2	33	30
AB	0	29	26
AC	2	4	4
AD	0	10	0
AE	0	19	26
RA	0.0	0.060	0.611
RB	0.962	0.030	0.166
RC	0.0	0.757	0.722
LTTR	0.383	0.389	0.308
AOAS	0.290	0.186	0.295
AOAP	0.354	0.218	0.0
LSA1AS	0.392	0.409	0.423
LSA1AP	0.532	0.535	0.0
AP	0.077	0.083	0.105



(a) Case-1: Fully autonomous writer

(b) Case-2: Autonomous writer with GPT-3 assistance

(c) Case-3: GPT-3 dependent writer

Figure 6: Sample cases of user's writing sessions demonstrated using CoAuthorViz

5.3.1 *Case 1: Fully autonomous writer*

The first sample session illustrated in Figure 6a illustrates an example where the writer is completely autonomous and decides not to use any GPT-3 suggestions in their writing. Starting to write from the initial prompt in sentence 1, the writer makes two GPT-3 calls but rejects its suggestions and decides to write by themselves thereon. The writer was perhaps not satisfied with the sentence suggestions offered by GPT-3 and decided not to get any more suggestions from it to not waste their time further. Table 4 shows that this session's autonomous writing indicator (RB = 0.96) is very high.

5.3.2 *Case 2: Autonomous writer with GPT-3 assistance*

The second case shown in Figure 6b shows an example where the writer incorporates a lot of GPT-3 suggestions in their writing, but modifies the sentences to suit their writing style. They start to write following the 7-sentence prompt provided and frequently get suggestions from GPT-3. In ten instances, the writer modifies the GPT-3 suggestion provided (overlapping triangle and square in sentences 11-13, 15, 26, 31-33) and in over 15 instances, they go on to add their own phrasing in addition to GPT-3 sentence suggestions (Sentences 8-10, 14, 16, 18, 19, 21, 23-30). Even though the autonomous writing indicator is low (RB = 0.03) for this session (because it is influenced by the number of sen-

tences completely authored by the writer), we observe that throughout the entire writing session, while they get assistance from GPT-3, the writer still demonstrates some autonomy in their writing by adding text on their own or modifying the GPT-3 suggestion. This is a great example of the potentially optimal use of machine assistance in combination with the writer's own writing and intelligence augmentation [43]. From Table 4, we observe that LTTR is 0.389, which is the highest of all three cases - the writing generated with GPT-3 assistance exhibited more diverse vocabulary [21].

5.3.3 Case 3: GPT-3 dependent writer

The final case illustrated in Figure 6c depicts the case of a writer who primarily used GPT-3 to create their piece of writing. Here, the writer starts off by adding sentences of their own 4 and 5 (following the initial prompt containing 4 sentences), before they become heavily dependent on GPT-3 for suggestions. The GPT-3 dependence indicator (RA) was 0.611 and the autonomous writing indicator (RB) was 0.166, evidencing that a considerable part of their writing was written by GPT-3. However, note that the writer demonstrated some autonomy by modifying GPT-3 suggestions, likely because they did not find them suitable (Sentences 18, 21, 31, and 32) and authored a few sentences themselves (Sentences 17, 21, 22, 31-35). This example demonstrates a writing style where the writer relied on GPT-3 suggestions repeatedly and used the system to its full advantage. The LTTR, in this case, is the lowest of the three cases (0.308) - there is less diverse vocabulary in this writing in comparison to both the autonomous writing by the writer in case 1 and GPT-assisted writing in case 2.

6. CONCLUSION

The paper introduced a novel approach to studying the co-authorship behaviour of writers interacting with GPT-3, a recent artificial intelligence (AI) tool producing auto-generated content. Keystroke logs from users' writing sessions in *CoAuthor* [21], where writers used automated text suggestions generated from GPT-3 as real-time feedback formed the basis of our analysis. Empirical studies on user interaction with GPT-3 are limited - this research fills the gap by introducing new methods of analysis and demonstrating diverse user behaviour when interacting with generative AI. The insights are also derived at an interpretable level for researchers building on keystroke data containing low-level details such as the character entered, current cursor location, etc. which is hard to read.

We developed 'CoAuthorViz', a visualization to represent interactions between the writer and GPT-3 at a sentence level - this captured key constructs such as the writer incorporating a GPT-3 suggested text as is (GPT-3 suggestion selection), the writer not incorporating a GPT-3 suggestion (Empty GPT-3 call), the writer modifying the suggested text (GPT-3 suggestion modification), and the writer's own writing (user text addition). Three different sample cases of writing exhibiting full autonomy in writing, using GPT-3 for assistance and GPT-3 dependence were shown to demonstrate the use of CoAuthorViz to study writing behaviours.

We derived additional CoAuthorViz metrics such as a GPT-3 dependence indicator, an autonomous writing indicator, and other GPT-3 suggestion incorporation metrics to quantify human and AI authorship. The average number of GPT-3 calls across the 1445 writing sessions was 12.5, but varied widely across the sessions (SD = 9.2). Automated sentence suggestions from GPT-3 were accepted as is 58% of the time and suggestions were rejected 29.31 % of the time, indicative of diverse writing behaviours with respect to interaction with GPT-3. Statistical analysis on CoAuthorViz metrics in relation to overall writing quality indicators derived from TAACO [7] showed that writers who accessed GPT-3 less produced writing with higher lexical density (more content words) and higher semantic overlap (higher average latent semantic analysis cosine similarity between all adjacent paragraphs). While the results showed the effects of high/ low GPT-3 usage in the output writing in terms of selected linguistic features, higher-level features are required to draw stronger links to writing quality. This can be done in the future by manually assessing the writing produced by the two groups of writers using a standard rubric for writing assessment.

From the three sample cases illustrated, we observed varied levels of autonomy exhibited by the writer when incorporating GPT-3 suggestions in their writing. These insights are useful for writing researchers to understand cognitive writing processes involved in human-AI partnerships from rich and nuanced log data. This could be the first step towards developing visual analytics that might be intelligible to a trained instructor grading the writing, or the basis for automated textual feedback to the instructor and/or student to improve their writing practices. We aim to further examine CoAuthorViz and its metrics for investigating comparable traits across different groups of writers and provide feedback for effective engagement. By studying effective user behaviours for enhanced human-AI partnership in writing, we can better understand how intelligence augmentation can be achieved in practice through critical engagement [43] [35].

The general consensus is that a partnership between the machine and the human is desirable for learning [28], but we need to understand and define what an *optimal partnership* is when working with generative AI for intelligence augmentation. There still remain questions on what constitutes desirable behaviours when it comes to interaction with GPT-3 - Is more autonomy (in terms of self-writing and edits to GPT-3) considered more optimal? Is it the one producing a better piece of writing irrespective of the repetitive use of GPT-3 and dependence? Do writers require foundational knowledge and skills to use AI tools to critique and use them appropriately? Do AI tools supplant critical processes and thinking that the learner ought to develop? These questions need further investigation.

Issues related to academic integrity also need due consideration. How one should attribute GPT-3 usage when co-authoring pieces of writing, and to what level is GPT-3 usage acceptable are open questions. In addition, the question of fairness remains as students who get access to better AI tools might be able to produce better writing [28] - accessibility issues may be elevated when these tools start to be distributed by companies for commercial profit at the end of public evaluation periods. With continuing advances in the intersection of technology, research, and practice, AI-augmented writing should enrich human knowledge for all.

7. REFERENCES

[1] B. Alexander. Chatgpt and higher education: last week and this week. https://bryanalexander.org/future-trends-forum/chatgpt-and-higher-education-last-week-and-this-week/, 2022. Accessed: 2023-01-11.

[2] L. K. Allen, M. E. Jacovina, M. Dascalu, R. D. Roscoe, K. M. Kent, A. D. Likens, and D. S. McNamara. {ENTER} ing the time series {SPACE}: Uncovering the writing process through keystroke analyses. *International Educational Data Mining Society*, 2016.

[3] K. C. Arnold, A. M. Volzer, and N. G. Madrid. Generative models can help writers without writing for them. In , editor, *IUI Workshops*, volume 2903, pages 1–8, United States, 2021. CEUR Workshop Proceedings.

[4] G. Caporossi and C. Leblay. Online writing data representation: A graph theory approach. In J. Gama, E. Bradley, and J. Hollmén, editors, *Advances in Intelligent Data Analysis X*, pages 80–89, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[5] A. Clark et al. Pillow (pil fork) documentation. *readthedocs*, 2015.

[6] R. Conijn, C. Cook, M. van Zaanen, and L. Van Waes. Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, 32(4):835–866, 2022.

[7] S. A. Crossley, K. Kyle, and D. S. McNamara. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237, 2016.

[8] R. Dale. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, 2021.

[9] H. Dang, K. Benharrak, F. Lehmann, and D. Buschek. Beyond text generation: Supporting writers with continuous automatic text summaries. UIST '22, pages 1–13, New York, NY, USA, 2022. Association for Computing Machinery.

[10] P. Deane, N. Odendahl, T. Quinlan, M. Fowles, C. Welsh, and J. Bivens-Tatum. Cognitive models of writing: Writing proficiency as a complex integrated skill. *ETS Research Report Series*, 2008(2):1 – 36, 2008.

[11] W. Du, Z. M. Kim, V. Raheja, D. Kumar, and D. Kang. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. *arXiv preprint arXiv:2204.03685*, 2022.

[12] L. Flower and J. R. Hayes. The cognition of discovery: Defining a rhetorical problem. *College composition and communication*, 31(1):21–32, 1980.

[13] K. I. Gero, V. Liu, and L. Chilton. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference*, pages 1002–1019, 2022.

[14] A. Gibson and A. Shibani. Natural language processing-writing analytics. *by Charles Lang, George Siemens, Alyssa Friend Wise, Dragan Gaševic, and Agathe Merceron. 2nd ed. Vancouver, Canada: SoLAR*, pages 96–104, 2022.

[15] S. Gooding, Y. Berzak, T. Mak, and M. Sharifi. Predicting text readability from scrolling interactions. *arXiv preprint arXiv:2105.06354*, 2021.

[16] D. Ippolito, R. Kriz, M. Kustikova, J. Sedoc, and C. Callison-Burch. Comparison of diverse decoding methods from conditional language models. *arXiv preprint arXiv:1906.06362*, 2019.

[17] G. M. Kleiman and GPT-3. Ai in writing class: Editor, co-author, ghostwriter, or muse? https://medium.com/@glenn_kleiman/ai-in-writing-class-editor-co-author-ghostwriter-or-muse-348532d896a6, 2022. Accessed: 2023-01-20.

[18] S. Knight, A. Shibani, S. Abel, A. Gibson, P. Ryan, N. Sutton, R. Wight, C. Lucas, A. Sandor, K. Kitto, et al. Acawriter: A learning analytics tool for formative feedback on academic writing. 2020.

[19] S. Knight, A. Shibani, and S. Buckingham-Shum. Augmenting formative writing assessment with learning analytics: A design abstraction approach. International Society of the Learning Sciences, Inc.[ISLS]., 2018.

[20] O. Kruse, C. Rapp, C. Anson, K. Benetos, E. Cotos, A. Devitt, and A. Shibani. Analytics techniques for analysing writing. In , editor, *Digital Writing Technologies in Higher Education: Theory, Research, and Practice*, pages 0–0. Springer, Berlin, Germany, 2023. In Submission.

[21] M. Lee, P. Liang, and Q. Yang. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19. ACM, apr 2022.

[22] M. Leijten and L. Van Waes. Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392, 2013.

[23] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou. Gpt detectors are biased against non-native english writers, 2023.

[24] S. Link, M. Mehrzad, and M. Rahimi. Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, 35(4):605–634, 2022.

[25] L. Lucy and D. Bamman. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, 2021.

[26] D. Malekian, J. Bailey, G. Kennedy, P. de Barba, and S. Nawaz. Characterising students' writing processes using temporal keystroke analysis. *International Educational Data Mining Society*, 2019.

[27] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[28] L. McKnight. In an ai world we need to teach students how to work with robot writers. https://theconversation.com/in-an-ai-world-we-need-to-teach-students-how-to-work-with-robot-writers-157508/, 2021. Accessed: 2023-01-11.

[29] S. E. E. Michael Mindzak. Artificial intelligence is getting better at writing, and universities should worry about plagiarism. https:

//theconversation.com/artificial-intelligence-is-getting-better-at-writing-and-universities-should-worry-about-plagiarism-160481, 2021. Accessed: 2023-01-11.

[30] E. R. Mollick and L. Mollick. New modes of learning enabled by ai chatbots: Three methods and assignments. *SSRN Electronic Journal*, 0:1–21, 2022.

[31] T. Schick, J. A. Yu, Z. Jiang, F. Petroni, P. Lewis, G. Izacard, Q. You, C. Nalmpantis, E. Grave, and S. Riedel. PEER: A collaborative language model. In *The Eleventh International Conference on Learning Representations*, pages 1 – 24, 2023.

[32] R. Schmelzer. Gpt-3. https://www.techtarget.com/searchenterpriseai/definition/GPT-3, 2021. Accessed: 2023-01-11.

[33] M. Sharples. New ai tools that can write student essays require educators to rethink teaching and assessment. *Impact of Social Sciences Blog*, 2022.

[34] A. Shibani. Constructing automated revision graphs: A novel visualization technique to study student writing. In *International Conference on Artificial Intelligence in Education*, pages 285–290. Springer, 2020.

[35] A. Shibani, S. Knight, and S. Buckingham Shum. Questioning learning analytics? cultivating critical engagement as student automated feedback literacy. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 326–335, 2022.

[36] N. Singh, G. Bernal, D. Savchenko, and E. L. Glassman. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction*, 0:0–0, 2022. Just Accepted.

[37] V. Southavilay, K. Yacef, P. Reimann, and R. A. Calvo. Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 38–47, 2013.

[38] Y. Sun, H. Ceker, and S. Upadhyaya. Shared keystroke dataset for continuous authentication. In , editor, *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, Abu Dhabi, United Arab Emirates, 2016. IEEE.

[39] Z. Swiecki, H. Khosravi, G. Chen, R. Martinez-Maldonado, J. M. Lodge, S. Milligan, N. Selwyn, and D. Gašević. Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3:100075–100085, 2022.

[40] Å. Wengelin, M. Torrance, K. Holmqvist, S. Simpson, D. Galbraith, V. Johansson, and R. Johansson. Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior research methods*, 41(2):337–351, 2009.

[41] J. Wilson and R. D. Roscoe. Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1):87–125, 2020.

[42] A. Yuan, A. Coenen, E. Reif, and D. Ippolito. Wordcraft: Story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 841–852, New York, NY, USA, 2022. Association for Computing Machinery.

[43] L. Zhou, S. Paul, H. Demirkan, L. Yuan, J. Spohrer, M. Zhou, and J. Basu. Intelligence augmentation: towards building human-machine symbiotic relationship. *AIS Transactions on Human-Computer Interaction*, 13(2):243–264, 2021.