Black Boxes Revisited: Understanding GenAI Responses to Student Writing Across the Curriculum

Meghan Velez, Zackery Reed, Darryl Chamberlain, & Cihan Aydiner

Abstract

In fewer than two years, generative artificial intelligence (GenAI) has transformed the educational experience for both students and faculty. Writing feedback and evaluation tools like MyEssayFeedback, EssayGrader, and Markr have been released with the promise that faculty will be able to focus more on teaching than simply grading. However, the proprietary training models of these tools make it difficult to discern precisely what criteria these AI-powered tools might use to evaluate and respond to student writing. Our study implemented ChatGPT3.0 to generate feedback and evaluative scores on short assignments in a broad range of disciplines. We took a descriptive approach to better understand the features and characteristics of student writing highlighted in AI evaluation and feedback responses. Project members engaged in iterations of open and axial coding resulting in two major themes, each with a handful of specific codes. The themes were Content and Tone of AI-Generated Feedback (Criteria Invention, Summarization of Student Response, and Encouragement Hedging Criticism) and Accuracy and Logistical Issues with AI Feedback Generation (Scoring, Inaccuracy, Context Window, and Attention to Purpose of Task). Based on a perceived relationship between student response length and ChatGPT score, we also performed an ad hoc hypothesis test to determine how likely the correlation (r=0.211; weak positive) we observed would occur under different random grading conditions. Results suggested statistically significant evidence that ChatGPT used response length as one, but not only, metric for providing an evaluation score. The somewhat mysterious nature of AI-generated feedback is one of many limitations to its use; yet our findings suggest that those limitations can be characterized, documented, and potentially addressed or compensated for through instructor interventions.

Keywords: *Generative AI Feedback; Automated Assessment Feedback, Writing Assessment, Generative AI Grading*

In fewer than two years, generative artificial intelligence (GenAI) has transformed the educational experience for both students and faculty. While much of higher education's initial response to GenAI focused on students' use of tools like ChatGPT to generate written content (Cotten et al., 2023; Caulfield, 2023; Mao et al., 2023;), more recent scholarship has turned to how faculty might utilize these same tools for teaching and research tasks (Dwivedi et al., 2023). Now, a variety of Generative AI-powered tools are marketed directly to faculty to assist with completing literature reviews (e.g., Research Rabbit; Scite), conducting qualitative data analysis (e.g., Atlas.ti), creating lesson plans and multimodal course content (e.g., TeacherMatic), and responding to student content and questions (e.g., PackBack). Finally, writing feedback and evaluation tools like MyEssayFeedback, EssayGrader, and Markr have been released with the promise that faculty will be able to “cut grading time in half...so [they] can focus on teaching students” (Mozaic Education, 2023).

Setting aside the arbitrary bifurcation of grading and teaching present in these marketing materials, these tools are often presented to faculty with little guidance or context as to how they work, the kinds of feedback they offer, or how they might be utilized in ethical, productive ways to support students’ writing processes (Cutler, 2024; Mao et al., 2023). This lack of guidance is exacerbated in courses outside of composition and communication, where faculty may already feel unsure how to provide direct writing instruction and feedback or may hold widely differing opinions on the ideal role of AI in writing (Del Pilar Gallego Castaño et al., 2016; Clughen & Connell, 2011; Dwivedi et al., 2023). Perhaps most importantly, the proprietary training models of these tools make it difficult to discern precisely what criteria these AI-powered tools might use to evaluate and respond to student writing.

To better understand how AI tools evaluate and respond to student-produced writing, the authors, as part of a larger research team, have been conducting a study of AI-enhanced writing feedback in courses across disciplines. As the pilot phase of the study, during the 2023-2024 academic year, a group of eight faculty members from a broad range of disciplines used ChatGPT to generate feedback and evaluative scores on short writing assignments in their courses. This article presents preliminary results from the pilot study, which was driven by the following questions:

1. How do GenAI tools evaluate and respond to student writing across disciplines?
2. How might the use of GenAI feedback interrupt and/or enhance assignment design and assessment in courses across disciplines?

To respond to these questions, this article focuses on our analysis of the instructors’ chat transcripts with ChatGPT. Because GenAI tools analyze text inputs and generate responses based on predefined patterns and linguistic models, the researchers hoped to learn more about the models themselves by examining how GenAI tools interpret and respond to texts. Our findings indicate that while ChatGPT interpreted and utilized instructor-provided grading criteria in its feedback and scoring, the open-ended feedback it generated also seemed to incorporate criteria not found in the instructor prompts. This suggests that some factors other than the instructors’ own criteria were utilized to evaluate the effectiveness of the student writing samples. Given the unseen “black box” nature of the proprietary models running tools like ChatGPT, our study ultimately asks: what does GenAI think is good writing?

In the remainder of this article, we situate the use of GenAI tools for scoring and responding to student writing within the context of broader conversations about automated essay scoring, writing assessment, and GenAI grading. Then, we outline our method for collecting and analyzing the ChatGPT feedback, highlighting our decision to use ChatGPT over grading-specific platforms like those described above and discussing our process for coding the transcripts. Our Results and Discussion sections emphasize the emergent criteria in ChatGPT feedback and suggest next steps for a factor analysis of AI-generated writing feedback. In providing these initial results, we aim to advance an analytical framework to categorize generative AI evaluations of student writing. By working to identify characteristics of AI-generated writing feedback, we neither advocate for nor advise against its use, as we acknowledge that many faculty and students are already receiving feedback from AI tools throughout their composing processes, whether they seek out feedback from a conversational LLM like ChatGPT or Gemini or simply receive ongoing feedback as a byproduct of using existing composing tools like Microsoft Word that now feature AI assistant integrations. Rather, we argue that rich descriptive data is needed to better understand the nature of AI-generated feedback before considering its use in the classroom.

Literature Review: Automated Essay Scoring and AI Feedback Tools

The use of machine-produced feedback or evaluation tools for student writing has historically been decried in the field of writing studies. For nearly thirty years, scholars in writing assessment have criticized the use of externally produced essay scoring software in the writing classroom or in large-scale assessment settings like standardized testing and placement; these criticisms include that machine scoring undermines the social nature of writing (Cheville, 2004; Dreschel, 1999; Herrington & Moran, 2001; Yancey et al., 2004), that the programs overemphasize surface features of writing in their evaluation and responses (Brock, 1995; Byrne et al., 2010), that the programs of the early 2000s in particular could not discern grammatically correct nonsense from meaningful prose (Elliott, 2010), and that the claims to validity and reliability made by software companies were found lacking when compared to trained human raters (Herrington & Moran, 2009; James, 2007; McCurry, 2010).

Writing studies scholars have also called attention to the ambiguous nature of machine scoring algorithms. Nearly two decades before the release of ChatGPT, Richard Haswell critiqued the “black boxes” (2006, p. 57) of automated essay scoring software, expressing concern over the unknown and unseeable processes contributing to student writing scores within programs like e-rater and WritePlacer. Yancey et al. (2004) drew a direct connection between this lack of transparency and potential bias in machine scoring, contradicting the software companies’ claims of objectivity: “since we can not know the criteria by which the computer scores the writing, we can not know whether particular kinds of bias may have been built into the scoring” (p. 4). Now, although the proprietary training corpora of large language models (LLMs) like ChatGPT have only amplified the mystery of and contention over how machines interpret, evaluate, and respond to written texts, a variety of GenAI tools are now being marketed to educators for use in grading, and an even greater variety of free, non-specialized programs can be used by individual faculty to achieve similar ends.

As mentioned above, since the public release of ChatGPT in 2022, a variety of proprietary tools have been developed and marketed to faculty to assist with a variety of tasks, including grading, and it can be helpful to understand how these tools function in comparison to the machine scoring programs critiqued by Haswell and others. Text-based automated feedback and grading systems inevitably draw from the same collection of machine learning techniques to perform assessment and feedback tasks. For instance, any generative feedback system relies on fine-tuning a pre-trained model, such as GPT3 or GPT4, to suit domain-specific or application-specific needs. While fine-tuning techniques vary widely, the underlying systemic processes are the same. Non-generative feedback and grading systems similarly draw from established natural language processing (NLP) techniques, such as reliance on neural network-based vector embeddings for meaning interpretations, structural analyses, or semantic analysis (such as described by Zhu, Shi, and Zhang, 2021). Similarly, there exist other previously established simpler machine learning methods for specific NLP tasks such as grammar checking; these rely on models not requiring neural network constructions, such as employing measured distances of strings from known spelled words, or more complex probabilistic models such as BaySpell (Golding, 1995) or WinSpell

(Golding & Roth, 1999). While there is a vast literature on the generative and non-generative approaches to NLP techniques used in automating essay feedback and evaluation, the functionalities of these models provide the user either with pre-determined feedback from specialized models, or more flexibly provide generative feedback from more generalist models.

As such, essay grading and feedback systems might employ smaller, more specialized, models for sub-tasks within the overall grading and feedback task. PackBack, for instance, offers automated "AI-Augmented Grading" in the form of a "Digital TA." This system provides a rubric with set categories (e.g., grammar, content, formatting, etc.) for scoring student submissions. The instructors view (and can edit) the score as well as the explanation given by the AI system for determining final grades and evaluation. This suggests that PackBack's system might employ a smaller topic-specific (such as grammar-specific) NLP or LLM-based models for some or each of the pre-determined grading criteria. At the same time, programs like EssayGrader advertise the option to import or customize rubrics, suggesting that a larger, more open model may be used to allow for wide variation in instructors' grading criteria.

Differing opinions exist regarding the virtue of specialized vs. generalist models for educational purposes. Latif & Zhai (2024), for instance, argued that "generic AI models frequently lack the precise contextual awareness required for successful educational interactions" but note that "a fine-tuned ChatGPT may offer individualized, engaging, and successful learning experiences" (p. 2). Likewise, Awidi (2024) employed ChatGPT in a study focused on scoring reflective essays in STEM, arguing that ChatGPT's more open and flexible model allowed for the researcher to train the AI on the specific assignment guidelines using the assignment prompt and grading rubric, which in turn "allowed for a fair comparison with human marking" (p. 4). In that study, researchers focused on instructor-developed rubric criteria, including "'Depth,' 'Analysis,' and 'Clarity/Logic of Writing'" to assess ChatGPT's scoring accuracy relative to the rubric (Awidi, 2024, p. 4). As outlined in the Methodology section below, we also opted to use ChatGPT for our pilot study, both to account for the wide variety of disciplines and writing assignment types present in our data set and to mimic the route that many classroom instructors might take to seek out AI-enhanced feedback: through a free program that at least appears to require little skill or formal training to use.

It is tempting to view the current landscape of AI grading tools (and more general AI tools used for grading) as history repeating itself. Certainly, the marketing claims made by tools like PackBack, Markr, and EssayGrader sound nearly identical to those of Criterion and e-rater, promising faculty the trifecta of "efficiency, objectivity, and freedom from drudgery" (Haswell, 2006, p. 64). Similarly, as was the case in the past, the vast majority of research in the use of the latest machine grading technologies comes from outside of writing studies. Nor is writing studies or writing assessment scholarship cited in the current literature, suggesting that Haswell's (2006) charge may be repeating itself:

> The pattern is that automated scoring of essays emerged during the 1990s out of the kinds of computer linguistic analysis and information retrieval that writing teachers had showed little interest in or had flirted with and then abandoned: machine translation, automatic summary and index generation, corpora building, vocabulary and syntax and text analysis. Researchers and teachers in other disciplines filled the gap because the gap was there, unfilled by us researchers and teachers in writing. (p. 63)

As a result, perhaps, of the lack of writing studies scholarship in AI-aided assessment, much of the current research aligns with the marketing materials in its focus on accuracy and efficiency. For example, Mizumoto & Eguchi's (2023) analysis of GPT-3 automated essay scoring for TOEFL examinations was primarily concerned with the AI's "accuracy and reliability" (p. 1), and Awidi's (2024) study prioritized accuracy, consistency, and objectivity as compared to human essay markers. Both these studies and others suggest efficiency as a main benefit to utilizing ChatGPT and similar systems for essay grading.

As we will outline below, while we did consider accuracy when analyzing our chat transcript data, it was not our primary goal. Instead, we opted to take a more descriptive approach to better understand the features and characteristics of student writing highlighted in AI evaluation and feedback responses. A baseline description of AI generated feedback is missing from much of the scholarship thus far produced about AI grading of student writing, and the authors believe that this descriptive data has much to tell us about what the values embedded in AI tools' production and assessment of writing.

Methodology

The results reported in this paper are taken from a larger project with two major foci: 1) analyzing students' perspectives on the use of LLMs for automated writing assignment feedback, and 2) identifying commonalities and differences among instances of AI-automated feedback on writing assignments across various academic disciplines. A primary methodological goal of the project, in line with the first area of focus, was the use of automated feedback and grading by instructors in classrooms in order to solicit feedback from students (subsequently collected via written responses and focus groups) about that use. This goal motivated various experimental decisions, which we discuss briefly below. However, because the main goal of this article is to report on the characteristics of the AI feedback itself, we will focus primarily on the methodological choices we made in generating and analyzing the AI feedback. As such, other than briefly describing the steps taken to protect student data privacy below, we will then refrain from referencing any methodological processes pertaining to the collection and analysis of student-facing data.

Data Collection—Courses and Assignments

In line with the second area of focus for the larger study, the members of the research team collected instances of automated feedback and grading from courses in a variety of disciplines and course levels. While the particular courses from which data was collected were dependent on the teaching loads of the project members, thus constituting a convenience sample, the varied disciplines and teaching levels of the project members ensured that the convenience sample was quite diverse. In total, data was collected from nine courses at the 100, 200, 400, and 500-levels at a large private, STEM-focused university. These courses spanned the following disciplines: 1) Emergency Services, 2) Computer Science, 3) Mathematics, 4) Homeland Security and Human Resilience, 5) Humanities, and 6) Communication.

In each course from which data was collected, the project member chose a single, low-stakes written assignment for which to obtain automated feedback and evaluation that would be shared with students but not used for actual course grading. The project member then collected and anonymized all submissions from the assignment in text form. In line with our institutional

IRB guidelines, students could opt out of this data collection, choosing not to receive automated feedback or evaluation.

In line with the broad goal of reflecting the kind of non-specialized AI feedback and evaluation in which many classroom instructors might choose to engage, the project team used ChatGPT3.5 (OpenAI, 2023) to process the data.

Instructors in all disciplines except mathematics[1] entered the text of the student submissions into ChatGPT after providing rubric and assignment information. Again, in pursuit of non-specialized instructor simulation, there was not a predetermined unified method for use of ChatGPT other than provision of rubric, assignment, and submission information. As such, there was some variety in the process for use of ChatGPT, such as the level of detail in the instructor-created prompts and the frequency with which ChatGPT was re-prompted with the rubric and assignment information. To protect students' data privacy, no student information was included in the prompts given to ChatGPT.

The choice of ChatGPT was strategic for multiple reasons, including its simulation of open-access use by instructors and its notoriety amongst students (in lieu of other lesser-known products such as PackBack or Lex). Moreover, ChatGPT is a generalist model, as opposed to other systems such as EssayGrader or Markr that might employ multiple more targeted machine learning algorithms to accomplish different tasks such as grading or feedback. Similar to Awidi (2024), our generalist approach also allowed the project team to engage in more task-specific training and control over the ways in which the model is prompted and utilized. A further discussion of the comparison between our use of ChatGPT and other possible systems is given below.

Comparison of Our Method with Other Systems

The specialization of AI-powered graders, writing tutors, and teaching assistants contrasts the generalist model of ChatGPT; however, the generalist approach allowed the project team to be more flexible with the rubric criteria being used for the automated feedback and grading. Similarly, while the Lex system might employ smaller models for different writing feedback sub-tasks, any current feedback tasks within the writing assistance feature reduces to pre-set prompts written by the Lex team for the LLM, or more generally any prompts given by the user. Despite these differences in specialization, the underlying mechanics of the evaluation and feedback system are ultimately very similar between our use of GPT and another instructor's use of Lex or PackBack. Moreover, our focus of reporting is not on accuracy of set criteria, but rather to descriptively and explanatorily report on emergent themes and patterns within the use of generative AI across disciplines, rather than specific to set rubric criteria such as is found in other paid systems.

Our particular use of generative AI for this grading and feedback task simulates an instructor engaging with free and generalist software. Generalist LLMs such as ChatGPT are designed to handle a variety of tasks, and the chat feature at the time of the study offered no extra domain-specific features concerning grading and evaluation. We particularly focus on application of such systems in which the user utilizes rubric information and text-based student submissions, and requires the AI system to evaluate the student submissions according to the rubric as well as providing generative feedback. Whereas paid systems such as PackBack or Lex handle issues of consistency, context window, specialization of sub-tasks, and organization of information, our use of

[1] The mathematics data was processed twice, once with ChatGPT 3.5 as described and then again later in ChatGPT 4-o, with the submission html entered into ChatGPT for better accuracy with processing mathematics equations.

ChatGPT required the user to navigate such considerations themselves. To our knowledge, no free or open-source system allows for grading-specialized subtasks such as those of the mentioned paid systems, meaning that our use of ChatGPT is likely to mirror how instructors who lacking paid subscriptions to the tools outlined above would be most likely to use AI feedback.

As such, our use of ChatGPT potentially differs from the implementations within paid systems as we required the generalist model to handle all potential sub-tasks of grading and evaluation from a single generative-only model and in one pass. One might augment this method by utilizing multiple chat instances to handle specific sub-tasks (e.g. one chat instance for grammar, one instance for content, one instance for structure, one instance for final grading based on the feedback from the other instances); however, we favored the simulation of a user's intention to efficiently leverage the models for general grading, choosing the realism of the simulated grading context over a thorough breakdown of the procedures maximally ensuring proficient and accurate utilization of the models. Moreover, while some of the instructors' rubrics may have contained writing-based evaluation criteria (such as organizational structure, support for ideas, or grammar/mechanics) other rubrics more broadly centered on subject-specific considerations of content. Any other major differences, therefore, between systems of automated feedback reduce to provided user interface options that set the prompts given to the system's LLM, and then the subsequent UI presentations of the model output.

Data Analysis

Analysis of the data followed Strauss and Corbin's (1990) method of grounded theory. Grounded theory is an "iterative process by which the analyst becomes more and more 'grounded' in the data and develops increasingly richer concepts and models of how the phenomenon being studied really works" (Ryan & Bernard, 2000). The project members coded data from one course at a time and were assigned data from courses that were somewhat related to their disciplinary specializations but not did not represent courses that they had taught (for instance, a rhetoric and composition faculty member analyzed transcripts from an organizational communication course). This was to prevent the analysis from becoming overly focused on "accuracy" in terms of the AI feedback's alignment with course content goals: rather than invite comparisons between the instructor's own grading/feedback preferences and the AI output, we strived to focus on identifying possible emergent themes and meaningful explanations of the data.

Project members engaged in iterations of open and axial coding (Strauss and Corbin, 1990). This involved reflexively cycling between coding at finer (open) and coarser-relational (axial) grain sizes of data, breaking down, examining, and conceptualizing (open) the data and then finding relationships between codes and categories (axial) of data to reconstruct the data meaningfully. The project members constantly tested the viability of codes against other coding instances within the data, in line with the constant comparative method (Strauss and Corbin, 1994). This viability testing was in an effort to detect and report on theoretical explanations of ChatGPT's evaluation of the data overall, rather than attempting to describe and report on the minutia of possibly outlying single-instance evaluations. After the project members engaged in isolated coding, the team met to discuss codes and categories that descriptively spanned the various content areas of the data. The team attended to both common themes as well as any themes that meaningfully (or starkly) differed across domains. It is on this collection of domain-spanning themes that we report.

Results

Axial categories for initial themes included the following, which have been grouped into larger categories reflecting the nature and scope of the themes:

Themes Related to Content and Tone of AI-Generated Feedback

- Criteria Invention—AI creates criteria for quality in student responses beyond those specified in the provided rubric;
- Summarization of Student Response—AI summarizes or repeats what students wrote rather than evaluating the quality of the response;
- Encouragement—AI provides non-evaluation comments meant to encourage student affect;
- Hedging Criticism—AI utilizes hesitant language, including passive voice, when criticizing student responses or providing suggestions for improvement.

Themes Related to Accuracy and Logistical Issues with AI Feedback Generation

- Scoring—perceived variability, lack of variability, or harshness of AI-provided grades for student responses, and the apparent relationship or lack thereof between feedback and scores;
- Inaccuracy—AI making an inaccurate statement about the student submission, whether by claiming that the submission contained or lacked a required element of the assignment or making contradictory statements about the content of the submission within the same response;
- Context Window—AI makes changes in feedback when reaching the end of a context window, such as making a new rubric or changing the format of its responses;
- Attention to Purpose of Task—perceived matches and mismatches between AI evaluation and purpose of task, such as attempts to grade student opinions as facts.

Below, we provide two examples of ChatGPT-generated feedback to a student submission and outline the presence of the above themes across the examples. For the purposes of this article, we will focus primarily on the themes that contribute to a descriptive picture of how ChatGPT responds to student writing. The first example is a brief reflection assignment from an undergraduate course focused on technology and communication. Students were asked to reflect on their learning within a unit focused on digital communication technologies, including GenAI.

> **Student Response:** The most surprising thing I learned in this module was how Chat GPT provides theories to the person instead of telling them that they don't have an answer. I am surprised because Chat GPT is a fact based generator that takes the information from websites. I just din't expect AI to give me opinions of what other people think the meaning of an artifact is. I also believe that the use of AI is the most significant to human history. AI increases efficiency but also allows humans to focus on more complex, creative, and strategic aspects of their work. The adoption of AI technologies has led to significant economic changes. Industries that leverage AI experience increased productivity, cost savings, and new business opportunities, contributing to economic growth. I believe that technologies

are more of less reimagined versions of their analog predecessors. This is because before social media many people actively engaged with each other and share their experiences. With digital communication platforms often emphasize interactivity and user participation. Social media, forums, and collaborative online spaces allow users to engage, share, and contribute to conversations and content creation. Although, there are some new aspects like digital data. Digital communication involves the representation of information in discrete, binary code, allowing for precise and reliable storage, transmission, and processing of data. This is a retreat from the continuous signals of analog communication. An unanswered question I have is how Chat GPT is allowed to give people opinions and theories from websites? Is it because AI acts as a human and thats how a person would respond if they didn't have a factual answer?

Table 1: *ChatGPT Feedback and Score on Writing Assignment in a Technology & Communication Course*

Instructor-Provided Domain	**ChatGPT Feedback and Score**
Discussion of Learning (answers all prompt questions, connects module concepts to overall learning) (50 pts)	The reflection effectively addresses the questions posed in the prompt and demonstrates a thoughtful engagement with the concepts learned in the module. The student articulates their surprise regarding ChatGPT's ability to provide theories and opinions, which indicates critical thinking and reflection on the module content. Additionally, the student identifies AI as the most significant communication technology and provides a clear rationale supported by examples of its impact on human history and economic changes. The analysis of digital technologies as reimagined versions of their analog predecessors is well-developed, with examples provided to support the argument. The student also poses an insightful unanswered question about ChatGPT's functioning, which shows curiosity and a desire for deeper understanding. (48/50)
Supporting Evidence from Module Activities and/or Readings (30 pts)	The reflection provides specific evidence from the module readings and activities to support the responses. The student references ChatGPT's function and its surprising ability to provide opinions and theories, which demonstrates an understanding of the course material. Additionally, examples are provided to support the argument regarding AI's significance to human history and the characteristics of digital communication technologies. The inclusion of these specific examples enhances the credibility and depth of the reflection. (28/30)

Conventions for Writing (grammar, mechanics, word count, etc.) (20 pts)	The reflection demonstrates a strong command of grammar and mechanics conventions, with few errors and evidence of careful proofreading. Sentences are clear and well-structured, contributing to the overall coherence of the writing. However, there are some minor issues with punctuation and capitalization that could be addressed for improved clarity and professionalism. Additionally, while the reflection includes references to concepts from the module readings, there are no citations provided in APA format, which is a requirement for academic writing. (18/20)
Overall Feedback	Overall, this reflection provides a thorough and insightful analysis of the module content, supported by specific examples and clear writing. Encouraging the student to include proper citations and address minor issues with punctuation and capitalization would further enhance the quality of their reflection.

First, ChatGPT does not have access to content provided in the course and thus cannot speak to “connections to module content”, which we coded as *Attention to Purpose of Task*. This limitation is evidenced throughout the response, but most clearly in the “Supporting Evidence” and “Conventions” categories when ChatGPT states that the student has provided specific examples and referenced module readings. These examples were also coded with the “Inaccuracy” theme as no actual references or mentions of readings or module assignments/activities were present in the student submission. Within both the “Discussion of Learning” and “Supporting Evidence” rubric categories, we note ChatGPT restating and summarizing the student’s response, which we coded as *Summarization of Student Response*. This summarization was a common occurrence in nearly every instance of ChatGPT feedback produced in the study. By contrast, suggestions for improvement are slim in this response, and are all phrased in passive voice (e.g., the issues with punctuation—which are not specified— “*could* be addressed”), an instance of “Hedging Criticism.” Finally, comments throughout the AI feedback provide insight into evaluative criteria not specified in the assignment requirements or instructor-provided rubric, including professionalism, clarity, depth, and the use of citations as a “requirement for academic writing.” These were all coded as “Criteria Invention.”

The following student response comes from a College Algebra course activity that required students to correct provided “student” work to a typical angular-to-linear velocity question and explain the mistakes made in attempting to solve the question.

> **Student Response:** Hello class, I will be doing Q.2. To start off the problem the student made the mistake of not plugging in the correct things for the formula that they used. This is the correct formula, but there is other stuff to figure in with it. Even if the formula would have been correct in the very last step of his problem they did not figure the math correct. $V = R\, x\, W$, $W = 2r\pi f$. Which can be substituted as $W = d\pi f$ with d=diameter and f= revolutions per minute. $W = (11in)(3300\ \frac{rev}{min})\pi$
>
> $$W = 36300\pi\ \frac{in}{min}$$

To figure out the velocity we have to convert and figure

$$V = 36300\pi\ \frac{in}{min}\ (\frac{1\ mi}{63360\ in})(\frac{60\ min}{1\ hr})$$

$$V = \frac{(36300)(60)\pi}{63360}\ mph$$

$$V \approx 108\ mph$$

The student did not do enough research into the problem to correctly complete it. If the student would have spent more time studying up on a problem like this he would have been much closer to the correct answer. The student did not provide the corrected work. It was incorrect from the very beginning and the student was not even figuring the math correctly with the numbers he did use.

Table 2: *ChatGPT Feedback and Score on Assignment in a College Algebra Course*

Instructor-Provided Domain	**ChatGPT Feedback and Score**
Student posts the errors included in the original problem-solving, explains why a student may have made each error, and provided corrected work (40 pts)	- **Correct Identification of Errors:** - Identified the mistake in not using the correct values in the formula. - Recognized that the formula was not applied correctly in the final step. - Noted the absence of the correct calculations. - **Explanation of Errors:** - Clearly explained the correct formula for linear velocity. - Identified the correct substitutions for the formula. - Articulated the steps for converting units and finding the linear velocity. - **Evaluation of Student Effort:** - Criticized the lack of research and understanding demonstrated by the student. - Suggested that more time spent studying would have led to a better answer. - **Omission of Corrected Work:** - Noted that the student did not provide corrected work. ### Grade: 8/10

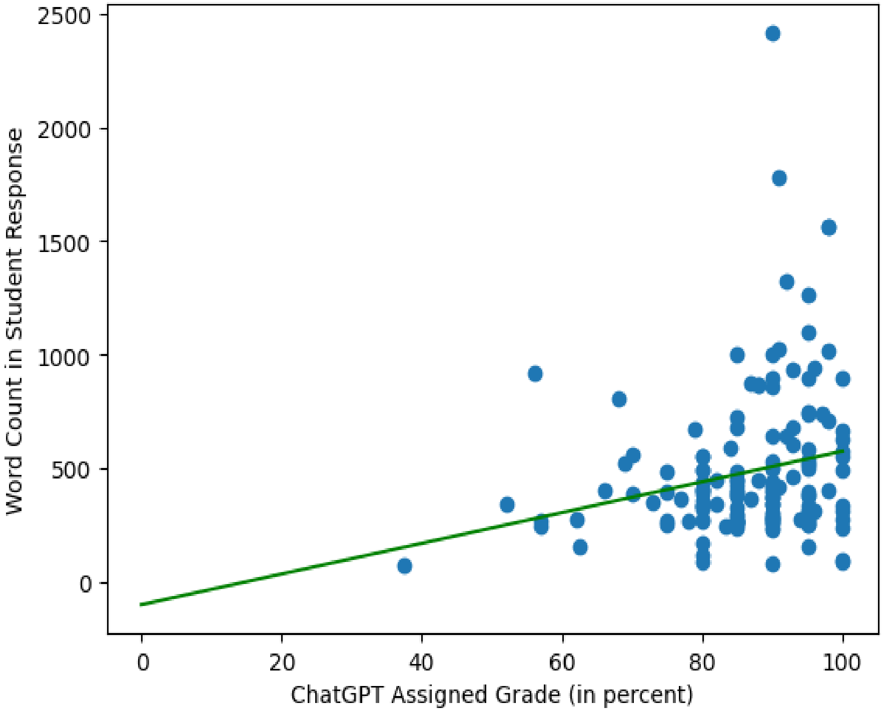
In this example, ChatGPT provides a different scoring metric (out of 10) than the rubric calls for (out of 40), which we noted with the *Scoring* code. Moreover, we again see the summarization of the student's response rather than any notes on what the student did or did not do correctly, which we coded as *Summarization of Student Response*. Here we also see an example of *Inaccuracy* where ChatGPT states the student did not provide something (corrected work) even though the student did provide corrected work. We also see the inclusion of a new category in the rubric, Evaluation of Student Effort, that is not present in the provided rubric, which we coded as *Criteria Invention.*

Ad-Hoc Quantitative Analysis

During discussion of the open and axial coding, the research group questioned the presence of non-instructor provided criteria in ChatGPT's feedback as well as an apparent lack of relationship between qualitative feedback and numerical scores in some transcripts. We also noted a potential proclivity for longer responses to be graded more highly by ChatGPT. Knowing that response length has been found to correlate with higher scores in large-scale writing assessment by human scorers as well as automated essay scoring software (Fleckenstein et al., 2020), the team decided to empirically test whether response length was being used as a factor in ChatGPT's grading response. We performed an ad-hoc correlation analysis between student submission word count and ChatGPT score. To avoid research bias on ChatGPT's grade generation, grades provided after ChatGPT was prompted to provide or revise an initial grade were omitted. ChatGPT produced grades with a mean of 86.8% and standard deviation of 10.3%.

We calculated the correlation between these two sets of values to determine if there was a relation between word count and ChatGPT score. Figure 1 presents the data and correlation. A weak positive correlation (r=0.211) was identified, suggesting word count is correlated with higher ChatGPT provided grades but that it was not the only factor used.

Figure 1: *Correlation Coefficient (r=0.211) Between ChatGPT Assigned Grade & Word Count in Student Response*



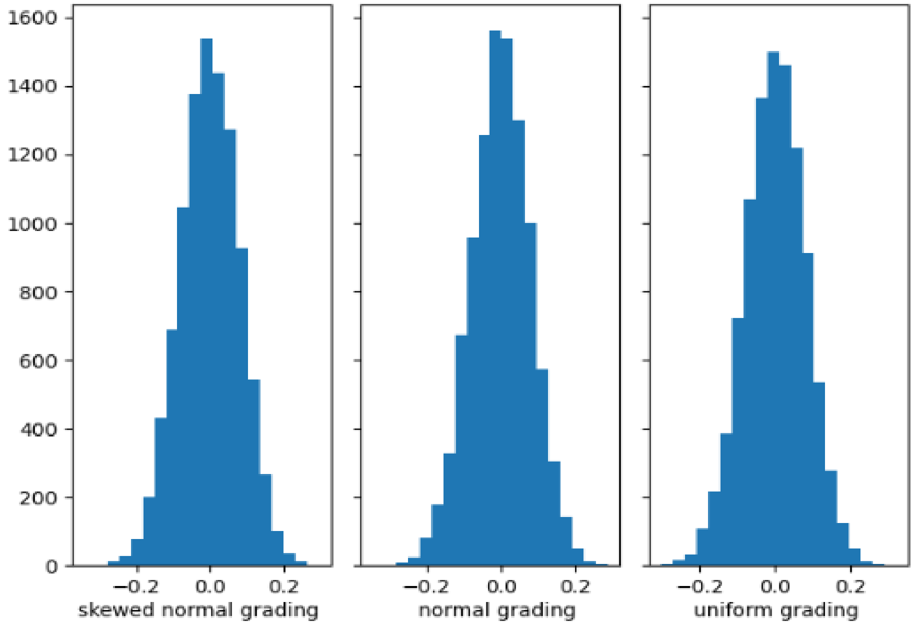
To consider whether this level of correlation could be achieved via random grading, we tested the word count for student submissions against 3 different random grading conditions:

- **Skew-Normal:** Skew-normally distributed grades fit from ChatGPT provided grades (a=0.000299, mean=0.868, stdev=0.103).
- **Normal:** Normally distributed grades (with a cap at 1) with center and standard deviation matching ChatGPT provided grades (mean=0.868, stdev=0.103).
- **Uniform:** Uniformly distributed grades between 0.662 and 1 (two standard deviations from ChatGPT-provided grade mean).

All calculations were performed in Python utilizing the scipy.stats package (Virtanen et al., 2020). The data and python scripts used can be found at https://github.com/Darryl-Chamberlain-Jr/genAI_feedback_project.

For each random grading condition: (a) 10000 sets of random grades were created, (b) each set was aligned to student submission word counts, and (c) each set's correlation coefficient was calculated. The distribution for these correlation coefficients is shown in Figure 2 below. Note for all three random grading conditions, the average correlation was approximately 0 (-0.001, 0.000, and 0.002 for skewed, normal, and uniform conditions, respectively) and the standard deviation of correlation was 0.081. The ChatGPT correlation was beyond two standard deviations correlation in each of the 3 random grading conditions, suggesting it is statistically likely ChatGPT was not providing a random grade for each student submission. Specifically, correlations above 0.2 (or below -0.2) occurred in approximately 1% of the trials (1.21%, 1.35%, and 1.19% for skewed, normal, and uniform conditions, respectively). This suggests ChatGPT grading was likely not completely random and thus the correlation coefficient of r = 0.211 represents 'response length' as a minor factor of ChatGPT's grade generation.

Figure 2: *Distribution of Random Scoring Correlations Using Skewed Normal, Normal, & Uniform Random Grading Conditions*



Discussion

The qualitative and ad hoc quantitative data analysis revealed several significant insights into how GenAI tools evaluate and respond to student writing. The themes identified in our data highlight both consistent and variable elements of AI feedback and warrant further exploration in order to uncover the values and biases about writing that are embedded within the black box of GenAI models. Though we recognize that, as Latour (1987) argues, attempting to open up one black box of GenAI evaluation will likely present us with "a new and seemingly incontrovertible black box" (p. 80), we view the task of characterizing GenAI feedback as a worthwhile precursor to assessing its usefulness, efficiency, or reliability.

Notably, we found that while ChatGPT often adhered to the rubric and assignment prompts provided by the instructors, sometimes to the point of restating them nearly word-for-word in written feedback to students, it also introduced criteria which were not explicitly outlined in the assignment instructions and varied in their degree of relevance to the assignment's purpose or task. This finding, coupled with the (albeit weak) positive correlation between word count and ChatGPT-provided assignment scores, suggests that ChatGPT did not provide entirely random evaluations and that factors other than those specified in the assignment and rubric (such as length) may have influenced the evaluation and feedback ChatGPT provided. Common criteria that appeared in the ChatGPT transcripts included the inclusion of examples, use of sources and citations, linking of ideas through transitions, consideration of opposing viewpoints, as well as broader concepts like "depth" and "clarity." Rarely (if ever) did ChatGPT suggest that student submissions were too wordy, contained irrelevant details, or were too reliant on sources, suggesting a "more is more" approach to writing. In addition, both criteria that were specified by instructors and that were supplied by ChatGPT were applied inconsistently within and across submissions, suggesting an inability of the LLM to recognize actions like "providing examples" or "making connections to readings" with any regularity. Even rudimentary mechanical conventions like "punctuation" and "capitalization" frequently mentioned in AI feedback did not often reflect the actual content of student submissions. For instance, in the first example provided in the Results section, there are no errors in capitalization, yet ChatGPT cited this as a reason for lost points in the "Conventions" section of the rubric and mentioned it again in the overall feedback. This suggests that the content of AI-generated feedback is informed by a combination of instructor-provided criteria, the content of the student submission, and the blanket application of other linguistic concepts stemming from somewhere within the AI's training data and protocols.

The somewhat mysterious nature of AI-generated feedback is one of many limitations to its use; yet our findings suggest that those limitations can be characterized, documented, and potentially addressed or compensated for through instructor interventions. For instance, the lack of attention to the genre and purpose of students' submissions is an obvious red flag for writing instructors who emphasize rhetorical awareness as a marker of effective communication. ChatGPT's difficulty in adapting feedback criteria to the nature of the specific task (e.g., rigidly requiring citations in a reflective journal entry) suggests that despite the conversational tone and structure of GenAI chatbots, they remain as limited as their automated essay scoring predecessors in constructing writing as a social phenomenon that is responsive to context (Cheville, 2004). In other words, without specific training or prompting by an instructor, AI-generated feedback is unlikely to attend much to rhetorical awareness.

On the other end of the spectrum, ChatGPT also did not provide meaningful insights about surface-level features of writing: none of the chat transcripts supplied specific examples of grammar and mechanical errors found in the student samples. While AI tools like Grammarly and Lex may be more adept in this area, this kind of specific feedback may prove difficult to obtain from a generalist AI. These findings point to the need for educators who use AI (and in particular, tools like ChatGPT) to carefully supervise and interpret AI-generated feedback to ensure it aligns with the intended learning objectives and does not mislead students.

Limitations

We identified several limitations that can be addressed in future research. With the rapid development and improvement of GenAI models, it is important to note the grading and feedback ChatGPT provided in this study are artifacts in time and cannot be reliably replicated. Insights gained from this study are specific to ChatGPT 3.5 in October 2023 to May 2024. Furthermore, prompts were not uniform between courses and could have introduced unintended variation in the grades and responses ChatGPT provided. Statistical analysis of student response length and ChatGPT provided grades would have been improved by collecting multiple sets of ChatGPT provided grades over a series of different context windows.

Moreover, as discussed above, we strategically chose ChatGPT for a generalist approach, allowing us to generate feedback and evaluation outputs for a variety of domains. Unlike Awidi (2024), our use of ChatGPT did not result in effective feedback overall, and our approach thus came with the potential limitation of less precise interpretation and evaluation from the LLM as described above. Despite this, our emergent themes still provide important insights into extant human and societal patterns in training data for these LLMs that go beyond issues of specialization.

Future Directions

Future research should focus on refining AI tools to better align with educational goals, exploring ways to mitigate the impact of irrelevant or unintended AI feedback, and examining the long-term effects of AI-assisted grading on student learning outcomes. For instance, further training of generalist LLMs like ChatGPT, as in Awidi's (2024) study, could mitigate the lack of course context awareness that led to some overly vague or inaccurate feedback in our study. Follow-up studies using a more fine-tuned ChatGPT, perhaps compared against more domain-specific LLMs such as EssayGrader, could provide a more robust picture of the merits and limitations of each tool type. Additionally, further studies could investigate characteristics of AI-generated feedback using a larger corpus of a single assignment type, leading to more easily comparable data across responses and further investigation of the role of genre and purpose in the content of AI feedback.

Moreover, our emergent themes give some insight into the extant norms, values, and conventions in LLM training data that were subsequently embedded in LLMs during model training. Given the limited interpretability of latent space features in machine learning models, work that empirically uncovers domain-relevant themes in the outputs of these models is important for developing actionable guidance for non-specialized users. Even the base insight that there are pre-established interpretations of language used in rubric criteria that the model will tend to favor is helpful for new users of GenAI for feedback and evaluative purposes. Similarly, the themes of *Criteria Invention, Summarization* and *Attention to Purpose* offer non-technical ways to communicate aspects of LLM feedback mechanisms that matter for educational use cases. Future research can build on this, explicitly articulating themes such as ours into novice-interpretable insights and even guidance for non-specialist educational users of GenAI.

In addition to educational research, future machine learning efforts can also focus on quality improvement and targeted model training or fine-tuning in order to develop models that adequately perform evaluation tasks aligned with ethical and quality standards for student assessment. Our initial question of "What does GenAI think is good writing?" is somewhat provocative, as one might more accurately describe the actions of GenAI as mimicking than thinking. Our results could therefore be phrased as "When given evaluation-oriented prompts, ChatGPT 3.5 often responds with summarization actions rather than effective evaluative actions and (occasionally) effectively mimics some genre conventions of an evaluative response." This rephrasing provides actionable steps forward, however. Those wanting to develop effective and ethically-oriented automated feedback systems must first amass data instances of effective and ethically-oriented human feedback–in other words, provide observable instances of the phenomena that we desire the LLM to mimic. Then, training techniques such as fine-tuning, reinforcement learning, or increasing inference computation could help the models favor those desired instances of feedback.

Finally, these insights also suggest the need for higher levels of transparency from companies offering automatic feedback tools. This transparency should include awareness of the data on which the employed models were trained or fine-tuned as well as reports of internal studies similar to our own that critically examine the observable preferences of feedback instances produced by these models.

Conclusion

The increasingly marketed and incentivized engagement with GenAI for educational use underscores the need for research highlighting understanding, transparency, and the ethics of GenAI. We have provided preliminary answers to the question *How do GenAI tools evaluate and respond to student writing across disciplines?* and have further investigated the potential for GenAI to enhance or interrupt assessment and feedback practices.

We have provided insights into some emergent themes from analysis of a diverse set of ChatGPT-produced feedback and assessment outputs. In particular, we noted that ChatGPT consistently introduced extra evaluation parameters beyond what was provided in assignment rubrics and prompts, highlighting the existence of human-generated norms and values within the training data for ChatGPT around evaluation of written work. This and other insights from our emergent themes will be important considerations for future work on the use of GenAI in writing assessment.

Moreover, we have continued the tradition of critically analyzing automated feedback and assessment of writing, noting that there is still a need for caution and understanding when using GenAI for evaluation and feedback. In particular, we have suggested that despite the conversational capacity of LLMs, the tools are not grounded in social-rhetorical constructions of communication. The limitations of GenAI in the areas that we identified further suggest that more work is needed to enable researchers and practitioners to understand and explain the mechanisms and boundaries of GenAI technology for educational use.

The ethical implications of employing AI to evaluate student work are particularly nuanced when considering the recursive interplay between AI-generated writing and AI-based grading. As students increasingly turn to AI tools to compose their assignments, and educators subsequently utilize AI to assess those outputs, a closed feedback loop may emerge in which AI systems effectively evaluate their own outputs. This phenomenon raises critical questions about the validity and reliability of such evaluations and the risk of perpetuating systemic biases inherent in AI models. Furthermore, these practices challenge foundational pedagogical principles, mainly cultivating students' critical thinking and developing their authentic voices in writing. Addressing these complexities requires careful reflection to ensure that the educational process does not devolve into a mechanized exchange between algorithms, thereby losing the essential human elements of teaching and learning.

References

ATLAS.Ti. (2024). *AI lab: Accelerating innovation for qualitative data analysis.* https://atlasti.com/atlas-ti-ai-lab-accelerating-innovation-for-data-analysis

Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence 6.* https://doi.org/10.1016/j.caeai.2024.100226

Brock, M. N. (1995). Computerized text analysis: Roots and research. *Computer Assisted Language Learning, 8*(2-3), 227-258.

Byrne, R., Tang, M., Truduc, J. & Tang, M. (2010). eGrader, a software application that automatically scores student essays: With a postscript on the ethical complexities. *Journal of Systemics, Cybernetics & Informatics, 8*(6), 30-35.

Caulfield, J. (2023). University policies on AI writing tools: Overview and list. https://www.scribbr.com/ai-tools/chatgpt-university-policies/

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal, 93*(4), 47-52.

Clughen, L., & Connell, M. (2011). Writing and resistance: Reflections on the practice of embedding writing in the curriculum. *Arts and Humanities in Higher Education, 11*(4), 333-345. https://doi.org/10.1177/1474022211429543

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 61*(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

Cutler, S. (2024, March 29). Graduate students went on strike. Then a dean suggested that professors use AI to keep classes going. *Chronicle of Higher Education.* https://www.chronicle.com/article/graduate-students-went-on-strike-then-a-dean-suggested-that-professors-use-ai-to-keep-classes-going

Del Pilar Gallego Castaño, L., Castelló Badia, M., & Badia Garganté, A. (2016). Faculty feelings as writers: Relationship with writing genres, perceived competences, and values associated to writing. *Higher Education, 71*(5), 719-734. https://doi.org/10.1007/s10734-015-9933-3

Drechsel, J. (1999). Writing into silence: Losing voice with writing assessment technology. *Teaching English in the Two-Year College, 26*(4), 380-387.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Hanaa Albanna, H., Albashrawi, M. A., Al-Busaidi, A.S., Janarthanan Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L.,

Buhalis, D., Wright, R. (2023). "So what if Chat-GPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71,* 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Elliott, S. (2011). Computer-graded essays full of flaws. *Dayton Daily News*.http://www.daytondailynews.com/project/content/project/tests/0524testautoscore.html

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Koller, O. (2020). Is a long essay always a good essay? The effect of text length on writing assessment. *Frontiers in Psychology 11,* 1-10. https://doi.org/10.3389/fpsyg.2020.562462

Golding, A.R., Roth, D. A Winnow-Based Approach to Context-Sensitive Spelling Correction. *Machine Learning 34*, 107–130 (1999). https://doi.org/10.1023/A:1007545901558

Haswell, R.H. (2006). Automatons and automated scoring: Drudges, black boxes, and dei ex machina. In P. F. Ericsson & Haswell, R.H. (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 79-92). Utah State UP.

Herrington, A. & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480-499.

Herrington, Anne, & Moran, Charles. (2009). Writing, assessment, and new technologies. *Assessment in the Disciplines 4*, 159-177.

James, C. L. (2007). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing, 11*(3), 167-178.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence 6*, 1-10. https://doi.org/10.1016/j.caeai.2024.100210

Lex. (2024). *Home*. https://lex.page/

Mao, J., Chen, B., & Liu, J.C. (2023). Generative artificial intelligence in education and its implications for assessment. *TechTrends, 68,* 58-66. https://doi.org/10.1007/s11528-023-00911-4

McCurry, Doug. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing, 15*(2), 118-129.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics 2*, 1-7. https://doi.org/10.1016/j.rmal.2023.100050

Mozaic Education. (2024). *Markr: AI-assisted grading and feedback.* https://www.testmarkr.com/

MyEssayGrader. (2024). *Product direction.* https://www.essaygrader.ai/direction

OpenAI. (2023). *ChatGPT* [Large language model]. https://chat.openai.com/chat

PackBack. (2024). *Writing Lab.* https://www.packback.co/product/writing-lab/

Ryan, G. W., & Bernard, H.R. (2000). Data management and analysis methods. In N. Denzin and Y. Lincoln (Eds.), *Handbook of qualitative research*, 2d ed. (pp. 769–802). Sage.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Sage Publications, Inc.

Strauss, A., & Corbin, J. (1994). Grounded theory methodology: An overview. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 273–285). Sage.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3), 261-272.

Yancey, K., Lunsford, A., McDonald, J., Moran, C., Neal, M., Pryor, C., Roen, D., & Selfe, C. (2004). *CCCC position statement on teaching, learning, and assessing writing in digital environments.* https://dtext.org/f14/505/readings/ncte-CCCC-digital-environments.pdf

Zhu, J., Shi, X., & Zhang, S. (2021). Machine learning-based grammar error detection method in English composition. *Scientific Programming*, 1.