

Optimizing AI for Assessing L2 Writing Accuracy: An Exploration of Temperatures and Prompts

Yiran Xu¹, Charlene Polio², and Adam Pfau²

¹The University of California, Merced, CA, USA

²Michigan State University, East Lansing, MI, USA

This study investigates the impact of temperature and prompt settings on ChatGPT-4 in assessing second language (L2) writing accuracy. Building on Pfau et al. (2023), we used a corpus of 100 essays by L2 writers of English and examined how three temperature settings (0, 0.7, 1) and two prompt types (defined, undefined) influenced ChatGPT-4's performance in error detection compared to human coding. Results indicated that ChatGPT-4, while generally underestimating error counts compared to human coders, showed a strong positive correlation with human coding across various settings. Notably, prompts with a detailed definition of errors yielded higher correlation coefficients ($\rho = 0.826$ to 0.859) than those without ($\rho = 0.692$ to 0.702), suggesting that more detailed prompts enhance ChatGPT-4's performance. Descriptive statistics showed that with a less-detailed prompt, the error detection ability of ChatGPT-4 was nearly identical across temperature settings, yet with a more detailed prompt, ChatGPT-4's performance was slightly better at higher temperatures. We discuss the importance of temperature in relation to prompt specificity for reliable L2 writing accuracy assessment and provide suggestions for optimizing AI tools such as ChatGPT-4 for assessing L2 writing accuracy.

Introduction

Over recent decades, writing accuracy has been an important construct in evaluating second language (L2) writing skills. In the last five years specifically, accuracy has been measured in areas such as written corrective feedback (e.g., Barrot, 2021; Benson & DeKeyser, 2019; Cheng & Zhang, 2021; Kim & Emeliyanova, 2021; Nicolas-Conesa et al., 2019; Sun & Qi, 2022; Suzuki et al., 2019), collaborative writing (e.g., Hidalgo & Lázaro-Ibarrola, 2020; Jiang & Eslami, 2022; Phan & Dao, 2023), and writing assessment (e.g., Abrams, 2019; Ahmadi et al., 2019; Ajabshir & Poorebrahim, 2021; Kessler et al., 2021). Nonetheless, assessing L2 writing accuracy presents significant challenges due to the absence of a universally accepted theory-based definition, the labor-intensive nature of accuracy coding, and the lack of reliability in coding (Pfau et al., 2023; Polio & Shea, 2014; Polio & Yoon, 2021).

To address these challenges, recent investigations have explored the efficacy of automated tools for assessing writing accuracy and performing grammatical error corrections (Coyne et al., 2023; Crossley et al., 2019; Jiang et al., 2023; Pfau et al., 2023). In particular, Coyne et al. (2023) analyzed the capabilities of advanced large language models, specifically GPT-3.5 and GPT-4 in English grammatical error correction and showed the models' strong performance in sentence-level revision tasks. However, they also pointed out the models' sensitivity to prompts and the tendency to favor fluency over minimal edits. Furthermore, Pfau et al. (2023) processed 100 L2 English essays across five proficiency levels with ChatGPT-4 and found a strong correlation between error detection by ChatGPT and human coding ($\rho = 0.97$ with one method and 0.94 with another). They noted that the output of ChatGPT-3.5 was relatively inconsistent in its identification of errors and respective error types. In contrast, ChatGPT-4 demonstrated enhanced consistency in its performance, though discrepancies in coding were still present. Different from Coyne et al. (2023) and Pfau et al. (2023), Jiang et al. (2023) compared four large language models (GPT-4, GPT-3.5, iFLYTEK, and Baidu Cloud) in assessing the linguistic accuracy of Chinese learner language, as gauged by error-free T-units (i.e., an independent clause and its associated dependent clauses) and error-free sentences. They found that iFLYTEK achieved the highest accuracy score, matching 81.3% of the human assessment deemed as the benchmark. Although GPT-4 demonstrated the highest precision (i.e., the proportion of identified errors found to be true errors) score of 0.88, it achieved a lower recall (i.e., the proportion of all actual grammatical errors that ChatGPT successfully identified) score of 0.85, an issue also noted in Pfau et al. (2023).

It is important to note that in Jiang et al. (2023) and Pfau et al. (2023) there was a significant reliance on manual labor for parsing T-units and sentences prior to evaluating the language accuracy across these models. This raises questions about the potential of these models to automatically process and evaluate accuracy data with minimal human input. These studies suggest that while these tools offer substantial benefits, careful consideration and customization are necessary to align them with the specific needs and objectives of L2 writing accuracy assessment. In addition, any use of ChatGPT to assess accuracy requires human intervention in terms of formatting the data and checking the accuracy and consistency of the output.

In this study, we expand on Pfau et al. (2023) by taking into consideration two important parameters that may affect ChatGPT-4's processing of grammatical errors: temperature and prompt. The OpenAI Developer Forum (2021) explains that temperature is a crucial parameter that controls the level of creativity and randomness in the responses produced by AI tools. With a lower temperature, such as 0.2 (on a scale of 0 to 1), ChatGPT-4 produces responses that are more "deterministic and focused" (OpenAI Developer Forum) or, according to our interpretation, more predictable and consistent. Conversely, a higher temperature like 0.7 introduces greater creativity and variability, possibly incorporating a broader array of linguistic structures and ideas. At temperature 0, the model should become more predictable, invariably selecting the most probable token, and therefore should result in more identical outcomes for each prompt generation. Meanwhile, a temperature setting of 1 significantly enhances the model's creativity, yielding more varied and diverse outputs.

Temperature is especially pertinent when considering the assessment of L2 writing accuracy because inconsistent output may affect the reliability of coding. For example, Pfau et al. (2023) observed inconsistency in the model's identification of errors when the temperature was set to 0.7 (the default in the user interface version). In attempts to reprocess the errors in L2 text, they found that ChatGPT-4 occasionally consolidated errors into broader categories and suggested stylistic or semantic modifications to sentences that were grammatically accurate. This was empirically tested by Coyne et al. (2023) who compared ChatGPT's performance in grammatical error correction tasks across three temperature settings (0.1, 0.5, 0.9) and found that lower temperatures yielded more accurate detection and consistently outperformed higher settings. They used crowd-sourced workers as the source of their human coding, but their study did not explain or detail this procedure in any way, nor were any statistics on intercoder reliability given.

Another important parameter is prompt, which has been shown to affect ChatGPT's output. Prompts act as the initial input or instruction given to the model, setting the context and direction for its responses. In a grammatical error detection task, the nature of the prompt can greatly influence the model's performance and accuracy. For example, a prompt where the model is asked to detect grammatical errors without specific instructions relies largely on the model's pre-existing knowledge and understanding to identify and correct errors. On the other hand, prompts that provide specific instructions or examples can refine the model's focus and potentially improve its accuracy in detecting errors. Pfau et al. (2023) experimented with different prompts to decide which prompt would be the most effective, but they did not systematically compare the output from the different prompts and used only one in their study. Coyne et al. (2023) tested different prompts in both zero-shot (e.g., prompting the model to correct a sentence without prior examples; if no errors exist, the model should return the original sentence) and few-shot settings (e.g., similar to a zero-shot prompt but accompanied by one or more examples of sentence-correction pairs to guide the model) in eliciting ChatGPT's response to sentence-level error correction tasks. Interestingly, after adding more examples to the best-performing zero-shot prompt, the performance score dropped slightly, indicating that zero-shot prompts performed marginally better than few-shot prompts with more examples. In the current study, we further investigate the role of prompt by comparing two prompts, one that does not

provide a definition of error (henceforth, *undefined*) and one that includes a detailed explanation (with examples) of what should be considered an error (henceforth, *defined*). In addition, we consider the feasibility of using the API-key version of ChatGPT 4, which is a fee-based platform that allows the processing of large amounts of data input using Google Sheets (OpenAI Help Center). It also allows for custom temperature settings.

Research Questions

Building on Pfau et al. (2023), which explored the capacity of ChatGPT-4 in assessing L2 writing accuracy, this study delves deeper into how temperature and prompt can affect the reliability of ChatGPT-4 when processing L2 writing accuracy. This study is guided by the following research questions:

1. How does temperature affect the correlation between human coding and ChatGPT-4's coding of accuracy?
2. How does prompt input affect the correlation between human coding and ChatGPT-4's coding of accuracy?

Methods

Corpus

We used the same dataset from Pfau et al. (2023), a set of 100 essays selected using stratified random sampling from a convenience sample of over 3000 essays written by Greek learners of English (corpus compiled by Kim, 2019). This subset includes essays from L2 writers across five distinct proficiency levels, each responding within a 35-minute timeframe to the same prompt: *What characteristics are most important for a good friend?* The distribution of these essays across proficiency levels is detailed in Table 1.

Table 1*Dataset across proficiency levels*

Grammar Subscore	Number of Essays	Number of Words	Word Count: Mean; Range
1-1.5	20	3,319	163.90;70-254
2	20	3,680	182.10;113-254
3	20	4,584	226.75;148-297
4	20	4,919	243.45;179-323
4.5-5	20	5,554	274.35;193-354
Total	100	22,056	

Notes. 1 (fail), 2 (narrow fail; satisfies some, but not all B2 criteria), 3 (marginal pass; B2), 4 (clear pass; exhibits B2/C1 features), 5 (honors pass; exhibits some C1/C2 features), half-points between each score (e.g., 1.5, 4.5).

Accuracy Measures

We initially selected the two most commonly used accuracy measures in the last five years (2019-2023): the rate of errors per 100 words and the ratio of error-free T-units (see Appendix A for a list of the studies). While clause-based measures like the error-free clause ratio have gained some traction in recent research, we opted not to employ them in our analysis because clause-based measures are typically better suited for academic writing, which is known for its complexity, dense structure, and advanced syntax. Our research, however, focuses on an expository writing task characterized by a more direct style of writing, where the sentence complexity tends to hinge more on the length and structure of the T-units rather than the intricate embedding of multiple clauses. Therefore, we believe the two measures we chose are more representative and effective in the current context, in addition to being the most common.

The ratio of error-free T-units measure, which we had planned to investigate, was excluded in the final analysis because it was immediately clear that ChatGPT-4 could not accurately identify error-free T-units. For most, but not all essays, T-units (which were defined in the prompt) were listed even if they contained errors. This was remarkably consistent across temperatures and happened both with and without prompts that defined an error. Although the issue of ChatGP-4's inability to accurately parse T-units could potentially be addressed by incorporating human coding, as Jiang et al. (2023) had done, we chose not to adopt this approach. Our goal is to assess the feasibility of using ChatGPT-4 for assessing writing accuracy with an emphasis on minimizing human effort. Additionally, when we used the user-interface

version as opposed to the API-key version, ChatGPT-4 was much better at identifying error-free T-units for reasons that are unclear. We did not adopt this approach either because the user-interface version can only process one request at a time, which significantly limits its scalability and practicality for larger datasets. Therefore, in order to develop a streamlined, automated process that could be applied consistently across multiple texts without requiring extensive manual oversight or sequential processing of individual requests, we decided that the rate of errors per 100 words was the most feasible measure.

Temperature Settings

Three temperature settings were selected: 0, 0.7, and 1. At a temperature setting of zero, ChatGPT-4 is designed to generate responses that are highly deterministic and consistent, offering predictable and reliable output. The 0.7 setting represents ChatGPT-4's default mode, balancing reliability with a moderate level of creativity. On the other hand, a setting of one⁹ maximizes ChatGPT-4's creative potential, introducing a higher degree of creativity and variability in its responses. In assessing linguistic accuracy, we hypothesized that the lowest temperature setting (0) would yield the most accurate results, whereas the highest temperature setting (1) would result in the least accurate outcomes due to its increased unpredictability.

Prompts

Two types of prompts were investigated for each accuracy measure. First, we adopted a prompt without providing the model with specific instructions on how to define or code an error, what we are calling the *undefined* prompt. It was: *Identify all of the errors in the essay, and count the total number of errors*. Then, we adapted Polio and Shea's (2014) guidelines and employed a more refined prompt that included a detailed definition of what constitutes an error, what we are calling the *defined* prompt (This prompt is given in Appendix B).

Analyses

Before calculating the correlations between ChatGPT-4 output and human coding, we first checked whether the *total number of errors* that ChatGPT-4 reported were accurate and matched the actual number of errors identified by the system. This was to validate the error counting mechanism of ChatGPT-4. We noticed inconsistencies in ChatGPT-4's ability to accurately tally the errors it identified. Therefore, for our correlation analyses, we relied on the actual number of errors identified by ChatGPT-4, rather than the total number of errors it initially reported. This approach ensured a more accurate basis for comparing ChatGPT-4's performance with human

⁹According to the [OpenAI API documentation](#), the temperature parameter can vary, with ranges of 0-1 or 0-2 depending on the specific task. The documentation indicates a range of 0-2 for generating completions based on a given prompt and parameters. However, in our trials with temperatures above 1, we observed that the output can become erratic with a higher likelihood of producing irrelevant or nonsensical responses. We therefore have decided to use a maximum temperature of 1 for our operations.

coding. While this distinction may seem trivial, it emphasizes the need for human intervention at various steps.

Next, using the actual (not reported) total number of errors identified by ChatGPT-4, we ran the correlations for the numbers identified by ChatGPT-4 with human coding figures. The error counts were normalized to errors per 100 words. The detailed human coding process, documented in Pfau et al. (2023), involves independent calculations conducted without reference to ChatGP-4's output. A stratified random sample (from the different proficiency levels) of 30 essays was used, with four for norming and 26 for assessing intercoder reliability, maintaining a high consistency with a reliability coefficient (r) of 0.97 for errors per 100 words between the two coders.

The assumption of normality was tested prior to conducting the correlation analysis between human and machine coding of grammatical errors. The Shapiro-Wilk test was used to assess the normality of the distribution of scores. Results indicated that the data did not follow a normal distribution (Shapiro-Wilk $W = 0.731$, $p < 0.001$ for undefined prompt output; Shapiro-Wilk $W = 0.785$, $p < 0.001$ for defined prompt output). Given this finding, the non-parametric Spearman's rank correlation was adopted for subsequent analyses, as it does not assume normality of the data.

Results

Descriptive Statistics

The average number of errors identified by human coders was 11.71 errors per 100 words ($SD = 8.26$). Table 2 presents the descriptive statistics for the average number of errors detected by ChatGPT-4 across prompts and temperature settings. We provide sample ChatGPT-4 responses to the two prompts in Appendix C. Overall, ChatGPT-4 consistently reported lower total error counts than human coders across different temperature settings in both undefined and defined prompts. This suggests that ChatGPT-4, while effective in identifying errors, might not capture as many errors as human coders, as shown in Pfau et al. (2023), regardless of prompt or temperature.

Table 2

Descriptive statistics for the number of errors detected by ChatGPT-4 across prompts and temperatures (Mean with standard deviations in the parentheses), $N = 100$

ChatGPT Coding	Temperature ($T = 0$)	Temperature ($T = 0.7$)	Temperature ($T = 1$)
Undefined prompt	7.05 (3.28)	7.11 (3.30)	6.95 (3.88)
Defined prompt	6.81 (4.80)	7.20 (5.01)	7.01 (4.98)

Correlation Analyses

Table 3 shows the Spearman's rank correlation (ρ) for ChatGPT-4 and human coding on errors per 100 words. Our findings demonstrate a strong positive correlation across varying temperature settings of the ChatGPT-4 model, indicating a consistent relationship between the error rankings produced by ChatGPT-4 and those by human coders. Notably, the correlation coefficients were $\rho = 0.702$, 0.692 , and 0.693 ($p < 0.001$) for undefined prompts without a definition at temperatures 0, 0.7, and 1, respectively. This correlation was found to be substantially stronger when ChatGPT-4 was provided with defined prompts including a definition and examples, yielding coefficients of $\rho = 0.826$, 0.838 , and 0.859 ($p < 0.001$) across the same temperature settings. This suggests that while ChatGPT-4 consistently underestimated the number of errors compared to human coders, providing a defined prompt aligns the model's performance more closely with human coding standards.

Table 3

Spearman's rank correlation (ρ) for ChatGPT-4 and human coding (errors/100 words)

Temperature (T)	Undefined prompt		Defined prompt	
	P	p -value	ρ	p -value
$T = 0$	0.702	<0.001	0.826	<0.001
$T = 0.7$	0.692	<0.001	0.838	<0.001
$T = 1$	0.693	<0.001	0.859	<0.001

Additionally, our analysis revealed insights into how the temperature parameter of ChatGPT-4 influences (or not) the correlation with human coding error rates. While the correlation remained robust across different temperature settings, a discernible trend emerged: the correlation coefficients remained largely the same across temperature settings when the undefined prompts were used. However, as the temperature increased from 0 to 1, the correlation coefficients slightly increased when ChatGPT-4 was provided with defined prompts. This suggests that a higher temperature, which typically allows for more variability and creativity in responses, may align more closely with the nuanced decision-making processes of human coders when adequate contextual information is supplied.

Precision and Recall

Pfau et al. (2023) showed that the user-interface version of ChatGPT-4 had excellent precision. Specifically, out of 1301 errors identified, only 10 were not counted as errors by humans (calculated as the number of agreements/the number of agreements plus the number of disagreements = 0.99), so we did not conduct another precision analysis based on temperature setting and prompt. Recall, an indication of how many errors ChatGPT-4 missed, however, was more problematic, at .69 (calculated as the number of manually coded instances identified by the automated approach divided by the number of agreements plus the number of missed errors). Recall was significantly worse at the lowest proficiency levels, likely because of challenges in parsing sentences. Because of the large number of essays processed (2 prompts \times 3 temperatures \times 100 = 600), we did not calculate recall in this study. In addition, for the defined prompts, the description of errors was not presented in sequence as found in the essays, so the output would have been challenging to format. Based on Table 2, however, ChatGPT-4, regardless of prompt and temperature, still underestimates the number of errors. In examining the output, we found that this was particularly a problem at the lowest levels for proficiency where ChatGPT-4 would identify one error in sentences that humans could identify as having multiple errors, presumably because the sentence could not be parsed.

Discussion and Future Directions

Building on our investigation in Pfau et al. (2023), this study further examined ChatGPT-4's capabilities as an automated tool for assessing linguistic accuracy. Despite its potential, our results indicate that human intervention remains crucial, especially when adjusting prompt specificity and temperature settings for optimal performance. Notably, our research reveals that providing a detailed definition of errors in the prompt typically enhances ChatGPT-4's performance. This enhancement may be stronger at higher temperature settings, where increased creativity is allowed.

We note that we are concerned, however, about the lack of consistency across time and platforms for ChatGPT-4. It remains unclear whether this issue is unique to ChatGPT-4 or if it also affects other AI tools. Furthermore, our dataset included a range of proficiency levels from 1.5 to 5 on the CEFR scale. The variation in accuracy scores likely resulted in higher correlations than what would be observed in a truncated sample consisting of, for example, just one level.

Given the ongoing refinement and evolution of AI tools like ChatGPT, recommending them as static tools for research poses a significant challenge. This is particularly true for ChatGPT-4, where we had difficulty getting the same results from the user-interface platform and the API key platform using Google Sheets. Therefore, we suggest that future research focusing on using ChatGPT-4 to assess linguistic accuracy must thoroughly document the coding process, including considering the versioning of AI tools and parameters used because they may change the performance and outcomes. This will help reproducibility and allow for meaningful comparisons across studies. Additionally, acknowledging the limitations inherent to AI tools' current state is essential, particularly regarding its understanding and processing of nuanced linguistic features.

Additionally, both Coyne et al. (2023) and Pfau et al. (2023) used datasets covering a range of CEFR levels (note that Coyne et al. does not provide details on the participants' level, but they refer to Bryant et al., 2019, which describes the learner corpus). We do not know how well ChatGPT-4 would do in detecting nuanced differences within the same proficiency level. In experimental or intervention studies, where outcomes between two groups of similar proficiency are compared, it is uncertain if ChatGPT-4 can effectively identify slight variations in language accuracy or improvement. This highlights a potential further investigation in using AI tools like ChatGPT-4 for detailed linguistic analysis, especially in educational contexts where detecting subtle improvements or changes in language proficiency is crucial.

Furthermore, in Pfau et al. (2023), we provided examples of types of errors that ChatGPT-4 consistently missed including run-on sentences and sentence fragments. For this reason, we did not delve into particular types of errors in detail in the current study. However, we anticipate that with more advanced prompt engineering, ChatGPT's performance can be further enhanced and possibly, at some point, be able to calculate other measures such as error-free T-units. By designing prompts that explicitly target the model's recognition of specific error types, researchers and educators could improve ChatGPT's detection accuracy for nuanced grammatical and syntactic issues.

Finally, we observed a lack of research on evaluating the linguistic accuracy of spoken language using AI tools such as ChatGPT-4, probably because AI models are primarily trained on vast corpora of written texts. This training foundation inherently biases these models towards the prescriptive grammatical norms, syntax, and vocabulary found in written language, which can differ significantly from spoken language in various ways. Given the lack of research on spoken language, we suggest incorporating diverse datasets of spoken language to help improve AI models' understanding and processing of spoken language in future model training.

Conclusion

Our study underscores the current limitations and human oversight necessary when employing AI tools, particularly ChatGPT-4, in analyzing the accuracy of L2 writing data. Despite identifying promising correlations with human coding, our findings reveal that significant human intervention is indispensable at various stages, including initial data setup, error counts, and final

output check. Interestingly, our investigation indicated that temperature settings did not seem to affect the output substantially whereas prompt choice did. Moreover, our exploration of integrating an API key with Google Sheets for processing extensive datasets highlighted a critical limitation. Specifically, for tasks requiring high precision, such as analyzing error-free T-units, this method proved less reliable than anticipated, particularly when compared to the direct use of the online user-interface version of ChatGPT-4. We thus promote a balanced approach that leverages AI capabilities while acknowledging and compensating for its current constraints through careful human intervention.

References

- Abrams, Z. I. (2019). The effects of integrated writing on linguistic complexity in L2 writing and task-complexity. *System*, 81, 110. <https://doi.org/10.1016/j.system.2019.01.009>
- Ahmadi, S., Riasati, M. J., & Bavali, M. (2019). An investigation of Iranian IELTS test takers' performance in bar chart and table prompts of academic writing task 1. *Cogent Education*, 6(1). <https://doi.org/10.1080/2331186X.2019.1640655>
- Ajabshir, Z. F., & Poorebrahim, F. (2021). Assessing EFL learners' written performance: The case of task repetition. *Southern African Linguistics and Applied Language Studies*, 39(3), 295-305. <https://doi.org/10.2989/16073614.2021.1942098>
- Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, 36(4), 584-607. <https://doi.org/10.1080/09588221.2021.1936071>
- Benson, S., & DeKeyser, R. (2019). Effects of written corrective feedback and language aptitude on verb tense accuracy. *Language Teaching Research*, 23(6), 702-726. <https://doi.org/10.1177/1362168818770921>
- Bryant, C., Felice, M., Andersen, O.E., & Briscoe, T. (2019, August). The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 52–75). Florence, Italy. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-4406>
- Cheng, X., & Zhang, L. J. (2021). Sustaining university English as a foreign language learners' writing performance through provision of comprehensive written corrective feedback. *Sustainability*, 13(15), 8192. <https://doi.org/10.3390/su13158192>
- Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research*, 11(2), 251–270. [10.17239/jowr-2019.11.02.01](https://doi.org/10.17239/jowr-2019.11.02.01)

- Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M., & Inui, K. (2023). Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*. <https://doi.org/10.48550/arXiv.2303.14342>
- Hidalgo, M. A., & Lázaro-Ibarrola, A. (2020). Task repetition and collaborative writing by EFL children: Beyond CAF measures. *Studies in Second Language Learning and Teaching*, 10(3), 501-522. <https://doi.org/10.14746/ssllt.2020.10.3.5>
- Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages*, 8(4). <https://doi.org/10.3390/languages8040247>
- Jiang, W., & Eslami, Z. R. (2022). Effects of computer-mediated collaborative writing on individual EFL writing performance. *Computer Assisted Language Learning*, 35(9), 2701-2730. <https://doi.org/10.1080/09588221.2021.1893753>
- Kessler, M., Ma, W., & Solheim, I. (2021). The effects of topic familiarity on text quality, complexity, accuracy, and fluency: A conceptual replication. *TESOL Quarterly*, 56(4), 1163-1190. <https://doi.org/10.1002/tesq.3096>
- Kim, Y., & Emeliyanova, L. (2021). The effects of written corrective feedback on the accuracy of L2 writing: Comparing collaborative and individual revision behavior. *Language Teaching Research*, 25(2), 234-255. <https://doi.org/10.1177/1362168819831406>
- Nicolas-Conesa, F., Manchon, R. M., & Cerezo, L. (2019). The effect of unfocused direct and indirect written corrective feedback on rewritten texts and new texts: Looking into feedback for accuracy and feedback for acquisition. *The Modern Language Journal*, 103(4), 848-873. <https://doi.org/10.1111/modl.12592>
- OpenAI Help Center. How can I access the ChatGPT API? Retrieved from <https://help.openai.com/en/articles/7039783-how-can-i-access-the-chatgpt-api> on February 1, 2024.
- OpenAI Developer Forum (2021). Cheat sheet: Mastering temperature and top_p in ChatGPT API. Retrieved from <https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api-a-few-tips-and-tricks-on-controlling-the-creativity-deterministic-output-of-prompt-responses/172683> on December 21, 2023.
- Phan, H. L. T., & Dao, P. (2023). Engagement in collaborative writing: Exploring learners' control of task content and text quality. *International Journal of Applied Linguistics*, 33(2), 242-259. <https://doi.org/10.1111/ijal.12462>
- Polio, C., & Shea, M. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, 23, 10-27. <https://doi.org/10.1016/J.JSLW.2014.09.003>

- Polio, C., & Yoon, H. J. (2020). Exploring multi-word combinations as measures of linguistic accuracy in second language writing (Eds.). In B. Le Bruyn, & M. Paquot (Eds.), *Learner corpora and second language acquisition research* (pp. 96–121). Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108674577>
- Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, 2(3), 100083. <https://doi.org/10.1016/j.rmal.2023.100083>
- Sun, H., & QI, W. (2022). Effects of written corrective feedback on college EFL students' writing accuracy and linguistic knowledge acquisition. [写作纠正性反馈对大学英语学习者写作准确性及语言知识习得的影响] *Chinese Journal of Applied Linguistics*, 45(3), 445-461. <https://doi.org/10.1515/CJAL-2022-0310>
- Suzuki, W., Nassaji, H., & Sato, K. (2019). The effects of feedback explicitness and type of target structure on accuracy in revision and new pieces of writing. *System*, 81, 135. <https://doi.org/10.1016/j.system.2018.12.017>

Appendix A – Selected Studies

Selected studies using the rate of errors per 100 words or its variations (2019-2023)

- Al-Hawamdeh, B., Hussen, N., & Abdelrasheed, N. S. G. (2023). Portfolio vs. summative assessment: Impacts on EFL learners' writing complexity, accuracy, and fluency (CAF); self-efficacy; learning anxiety; and autonomy. *Language Testing in Asia*, 13(1), 12. <https://doi.org/10.1186/s40468-023-00225-5>
- Alghizzi, T. M., & El Deen, A. A. M. A. (2020). Enhancing Saudi EFL learners' CAF in narrative writing through native speaker kids' reading website. *International Journal of Applied Linguistics & English Literature*, 9(6), 67-81. <https://doi.org/10.7575/aiac.ijalel.v.9n.6p.67>
- Cheng, X., & Zhang, L. J. (2021). Sustaining university English as a foreign language learners' writing performance through provision of comprehensive written corrective feedback. *Sustainability*, 13(15), 8192. <https://doi.org/10.3390/su13158192>
- Cho, M. (2019). The effects of prompts on L2 writing performance and engagement. *Foreign Language Annals*, 52(3), 576-594. <https://doi.org/10.1111/flan.1241>
- Eckstein, G., & Chang, R. (2022). How does the language control of L1 and L2 writers develop over time in first-year composition? *Written Communication*, 39(4), 600-629. <https://doi.org/10.1177/07410883221099474>
- Gao, J., & Min, S. (2021). A comparative study of the effects of L1 and L2 prewriting discussions on L2 writing performance. *System*, 103, 1. <https://doi.org/10.1016/j.system.2021.102654>
- Golparvar, S. E., & Azizsahra, M. (2022). The effect of graph complexity and planning on graph writing performance and descriptive strategies. *Foreign Language Annals*, 56(1), 117-143. <https://doi.org/10.1111/flan.12676>
- Hidalgo, M. A., & Lázaro-Ibarrola, A. (2020). Task repetition and collaborative writing by EFL children: Beyond CAF measures. *Studies in Second Language Learning and Teaching*, 10(3), 501-522. <https://doi.org/10.14746/ssllt.2020.10.3.5>
- Kim, Y., & Emeliyanova, L. (2021). The effects of written corrective feedback on the accuracy of L2 writing: Comparing collaborative and individual revision behavior. *Language Teaching Research*, 25(2), 234-255. <https://doi.org/10.1177/1362168819831406>
- Nicolas-Conesa, F., Manchon, R. M., & Cerezo, L. (2019). The effect of unfocused direct and indirect written corrective feedback on rewritten texts and new texts: Looking into feedback for accuracy and feedback for acquisition. *The Modern Language Journal*, 103(4), 848-873. <https://doi.org/10.1111/modl.12592>

- Nusrat, A., Ashraf, F., & Narcy-Combes, M. (2019). Effect of direct and indirect teacher feedback on accuracy of English writing: A quasi-experimental study among Pakistani undergraduate students. *3L, Language, Linguistics, Literature*, 25(4). <https://doi.org/10.17576/3L-2019-2504-06>
- Plakans, L., Atta, G., & Zeynep, B. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*, 36(2), 161-179. <https://doi.org/10.1177/0265532216669537>
- Sarré, C., Grosbois, M., & Brudermann, C. (2021). Fostering accuracy in L2 writing: Impact of different types of corrective feedback in an experimental blended learning EFL course. *Computer Assisted Language Learning*, 34(5-6), 707-729. <https://doi.org/10.1080/09588221.2019.1635164>
- Vasylets, O., Mellado, M. D., & Plonsky, L. (2022). The role of cognitive individual differences in digital versus pen-and-paper writing. *Studies in Second Language Learning and Teaching*, 12(4), 721-743. <https://doi.org/10.14746/ssllt.2022.12.4.9>
- Selected studies using the ratio of error-free T-units (2019-2023)**
- Abrams, Z. I. (2019). The effects of integrated writing on linguistic complexity in L2 writing and task-complexity. *System*, 81, 110. <https://doi.org/10.1016/j.system.2019.01.009>
- Ajabshir, Z. F., & Ebadi, S. (2023). The effects of automatic writing evaluation and teacher-focused feedback on CALF measures and overall quality of L2 writing across different genres. *Asian-Pacific Journal of Second and Foreign Language Education*, 8(1), 26. <https://doi.org/10.1186/s40862-023-00201-9>
- Ajabshir, Z. F., & Poorebrahim, F. (2021). Assessing EFL learners' written performance: The case of task repetition. *Southern African Linguistics and Applied Language Studies*, 39(3), 295-305. <https://doi.org/10.2989/16073614.2021.1942098>
- Alghizzi, T. M., & El Deen, A. A. M. A. (2020). Enhancing Saudi EFL learners' CAF in narrative writing through native speaker kids' reading website. *International Journal of Applied Linguistics & English Literature*, 9(6), 67-81. <https://doi.org/10.7575/aiac.ijalel.v.9n.6p.67>
- Barrot, J., & Gabinete, M. K., (2021). Complexity, accuracy, and fluency in the argumentative writing of ESL and EFL learners. *IRAL, International Review of Applied Linguistics in Language Teaching*, 59(2), 209-232. <https://doi.org/10.1515/iral-2017-0012>
- Cho, M. (2019). The effects of prompts on L2 writing performance and engagement. *Foreign Language Annals*, 52(3), 576-594. <https://doi.org/10.1111/flan.12411>

- Fathi, J., & Rahimi, M. (2022). Examining the impact of flipped classroom on writing complexity, accuracy, and fluency: A case of EFL students. *Computer Assisted Language Learning*, 35(7), 1668-1706. <https://doi.org/10.1080/09588221.2020.1825097>
- Fathi, J., & Rahimi, M. (2022). Examining the impact of flipped classroom on writing complexity, accuracy, and fluency: A case of EFL students. *Computer Assisted Language Learning*, 35(7), 1668-1706. <https://doi.org/10.1080/09588221.2020.1825097>
- Jiang, W., & Eslami, Z. R. (2022). Effects of computer-mediated collaborative writing on individual EFL writing performance. *Computer Assisted Language Learning*, 35(9), 2701-2730. <https://doi.org/10.1080/09588221.2021.1893753>
- Kessler, M., Ma, W., & Solheim, I. (2021). The effects of topic familiarity on text quality, complexity, accuracy, and fluency: A conceptual replication. *TESOL Quarterly*, 56(4), 1163-1190. <https://doi.org/10.1002/tesq.3096>
- Kim, Y., & Emeliyanova, L. (2021). The effects of written corrective feedback on the accuracy of L2 writing: Comparing collaborative and individual revision behavior. *Language Teaching Research*, 25(2), 234-255. <https://doi.org/10.1177/1362168819831406>
- Meletiadiou, E. (2021). Opening Pandora's box: How does peer assessment affect EFL students' writing quality? *Languages*, 6(3), 115. <https://doi.org/10.3390/languages6030115>
- Mohammadzadeh, A., Ahour, T., & Saeidi, M. (2020). A sociocultural perspective on second language writing: The effect of symmetrical versus asymmetrical scaffolding on intermediate EFL learners' writing accuracy, fluency, and complexity and their attitudes. *Education Research International*, 5292356, <https://doi.org/10.1155/2020/5292356>
- Phan, H. L. T., & Dao, P. (2023). Engagement in collaborative writing: Exploring learners' control of task content and text quality. *International Journal of Applied Linguistics*, 33(2), 242-259. <https://doi.org/10.1111/ijal.12462>
- Rahimi, M. (2019). Effects of increasing the degree of reasoning and the number of elements on L2 argumentative writing. *Language Teaching Research*, 23(5), 633-654. <https://doi.org/10.1177/1362168818761465>
- Rahimi, M., & Zhang, L. J. (2019). Writing task complexity, students' motivational beliefs, anxiety and their writing production in English as a second language. *Reading and Writing*, 32(3), 761-786. <https://doi.org/10.1007/s11145-018-9887-9>
- Şahin Kızıl, A. (2023). Data-driven learning: English as a foreign language writing and complexity, accuracy and fluency measures. *Journal of Computer Assisted Learning*, 39(4), 1382-1395. <https://doi.org/10.1111/jcal.12807>

- Sang, Z., & Zou, W. (2023). The effect of joint production on the accuracy and complexity of second language writing. *Journal of Psycholinguistic Research*, 52(2), 425-443. <https://doi.org/10.1007/s10936-022-09882-8>
- Wu, Q., & Albert, A. (2022). The role of teacher-generated, learner-generated, and creative content in Chinese EFL students' narrative writing: A contextual perspective. *Languages*, 7(3), 212. <https://doi.org/10.3390/languages7030212>
- Xu, J., & Zhang, S. (2022). Understanding AWE feedback and English writing of learners with different proficiency levels in an EFL classroom: A sociocultural perspective. *The Asia-Pacific Education Researcher*, 31(4), 357-367. <https://doi.org/10.1007/s40299-021-00577-7>

Appendix B: Refined Prompt

Identify all of the errors in the essay, and count the total number of errors. Errors are defined as the following. Each type is followed by example(s).

1. Missing subject

But soetmime you might have you own secret that can't tell anybody except one person.

2. Missing verb

When he thinks he have to something, he does it finally even very difficult thing that other people give up.

3. Missing verb complement or object or required prepositional phrase

and I'm missing now.

4. Verb phrase problem: Wrong tense/aspect or misformed tense/aspect. May also include wrong participle in a participle clause or an attempt at something passivelike where it does not belong

I have been studied there for eight months.

It can be reduce the accident rate.

5. Preposition problem: May include missing, incorrect, or additional preposition

And my brother-in-law graduated in MSU 10 year ago.

6. Sentence fragment: A sentence fragment is a sentence missing a main verb.

I have five members. My parents (father, mother), younger sister, younger brother, and me.

7. Run-on sentence: A run-on sentence includes two sentences without a conjunction. They may be attached by a comma.

As time goes by and having more sense of being a part of this campus, I love to enjoy the great service provided by school such as gyms, libraries, labs in departmental building, all of them are well-organized and convenient for faculty and students to have a better living and do academic research.

8. Problem with relative clause formation including wrong relative pronoun, reduced relative clause (use of infinitive instead of participle), or resumptive pronoun

It is the place that we enjoy in it.

9. Wrong modal or modal where not needed

Every day we can get many useful information from him.

10. Incorrect formation of passive voice including get-passive

The building built by the construction company (as opposed to something like: The building which is sat on the hill).

11. Subject-verb agreement error

She has the religion of buddism which mean she is a buddist.

12. Wrong pronoun or possessive determiner (including reflexive) and it/there.

That's the reason why I don't like them. (them refers to father)

13. Quantifier-noun agreement (much/many, this/these) or other quantifier problems (a few/few); not including singular plural

There are little students comparing MSU.

14. Problematic comparative or superlative formation

In recent research, the capital of Korea, Seoul, is the worst clean city in the world.

15. Singular/plural error (including making mass nouns plural)

Because there are all kinds of store around it.

16. Negation problem (including missing do)

So, my father couldn't study no more

17. Wrong, extra, or missing article (for frequent English proper nouns, require appropriate article use but not for foreign words)

From the middle of September to the end of November, it was a very nice scenery.

18. Wrong lexical item (including conjunctions, phrasal verb)

Also, we have many green and colorful flowers in the yeard (meaning on campus).

19. Wrong word form (e.g., adjective for noun) or wrong derivational formation

He has much patient.

It is very crowdy.

20. Word order problem

How did you stay for 13 hours every day in school?

21. Missing or extra word not included above

I was really tired of routine work, stay late evening.

22. Severe punctuation error (be very lenient with comma errors) including possessives such as “My brothers house” or contraction problems such as its/it’s

That why, I have a time to do my work.

23. Gerund/infinitive error

I had such a great experience to study there.

24. Genitive error

My university’s friend (for my university friend)

25. Spelling errors

Other comments

1. Don’t double penalize for subject-verb agreement and singular plural articles. For example, Vistor are pleased with the sight (only a number problem).
2. Code so that the sentence is minimally changed. If unclear, code first error.
3. Generally, don’t count register errors (various informal variants).

Appendix C

Sample Essay:

Friends that they have valued with different characteristics, they must have a lot of friendships. Also, this people can help other people to found their personal identity. For starters, the most import characteristic of a good friend must hear me and said me to take the write choice. A good friend should be charesmetic and a person with a full of life. In addition, a friend should be logical for many reason but I didn't know why they might be. On other hand, it is important to have valuable friend because you can do with your friend everything you could imagine. For example, talk about political subject or for the forological system of my country. Taking everything mentioned above into consideration, I believe that different people maybe you have changed your life. (Essay 29408_b_1.5)

ChatGPT-4's response to the undefined prompt (Temperature = 0.7):

1. "they must have a lot of friendships" - should be "they must have a lot of friends"
2. "Also, this people can help other people to found their personal identity" - should be "Also, these people can help others find their personal identity"
3. "the most import characteristic" - should be "the most important characteristic"
4. "must hear me and said me to take the write choice" - should be "must listen to me and advise me to make the right choice"
5. "should be charesmetic" - should be "should be charismatic"
6. "a person with a full of life" - "should be a person full of life"
7. "a friend should be logical for many reason" - should be "a friend should be logical for many reasons"
8. "but I didn't know why they might be" - unclear statement, needs to be rephrased
9. "On other hand" - should be "On the other hand"
10. "it is important to have valuable friend" - should be "it is important to have valuable friends"
11. "you can do with your friend everything you could imagine" - should be "you can do everything with your friend that you could imagine"

12. "talk about political subject" - should be "talk about political subjects"
13. "for the forological system of my country" - unclear statement, needs to be rephrased
14. "Taking everything mentioned above into consideration" - should be "Taking everything mentioned above into consideration"
15. "different people maybe you have changed your life" - unclear statement, needs to be rephrased

Total number of errors: 15.

ChatGPT-4's response to the defined prompt (Temperature = 0.7):

1. Missing subject: "For starters, the most import characteristic of a good friend must hear me and said me to take the write choice."
2. Verb phrase problem: "For starters, the most import characteristic of a good friend must hear me and said me to take the write choice."
3. Problem with relative clause formation: "Friends that they have valued with different characteristics, they must have a lot of friendships."
4. Wrong or extra preposition: "Also, this people can help other people to found their personal identity."
5. Incorrect formation of passive voice: "Also, this people can help other people to found their personal identity."
6. Wrong pronoun or possessive determiner: "Also, this people can help other people to found their personal identity."
7. Singular/plural error: "Also, this people can help other people to found their personal identity."
8. Subject-verb agreement error: "Friends that they have valued with different characteristics, they must have a lot of friendships."
9. Wrong word form: "A good friend should be charesmetic and a person with a full of life."
10. Wrong lexical item: "A good friend should be charesmetic and a person with a full of life."
11. Missing verb complement: "In addition, a friend should be logical for many reason but I didn't know why they might be."

12. Spelling errors: "A good friend should be charismatic and a person with a full of life."
13. Wrong, extra, or missing article: "On other hand, it is important to have valuable friend because you can do with your friend everything you could imagine."
14. Preposition problem: "For example, talk about political subject or for the forological system of my country."
15. Gerund/infinitive error: "For example, talk about political subject or for the forological system of my country."
16. Run-on sentence: "Taking everything mentioned above into consideration, I believe that different people maybe you have changed your life."
17. Sentence fragment: "Taking everything mentioned above into consideration, I believe that different people maybe you have changed your life."
18. Word order problem: "For starters, the most import characteristic of a good friend must hear me and said me to take the write choice."
19. Missing or extra word not included above: "For starters, the most import characteristic of a good friend must hear me and said me to take the write choice."

Total number of errors: 19.

About the Authors

Yiran Xu is an Assistant Professor of Writing Studies at the University of California, Merced. Her research centers on multilingual writing and second language acquisition, particularly among adult learners in classroom settings. Drawing on usage-based theory, her work seeks to understand how multilingual learners develop academic and disciplinary writing skills through a mixed-methods approach. Her work has been published in the *Journal of Second Language Writing*, *System*, *Linguistics and Education*, *Research Methods in Applied Linguistics*, *Studies in Second Language Learning and Teaching*, and others.

Website: <https://sites.google.com/view/yiranxu>

Charlene Polio is a Professor in the Department of Linguistics, Languages, and Cultures at Michigan State University. Her work sits at the intersection of second language writing and second language acquisition focusing on language development in and through writing. She is also involved in corpus-based approaches to analyzing published research articles. Her most recent work has appeared in *Journal of English for Academic Purposes* and *Research Methods in Applied Linguistics*. She is the current co-editor of *TESOL Quarterly*.

Website: <https://charlenepolio.com/>

Adam Pfau is a PhD candidate in the Second Language Studies program at Michigan State University. His research interests include computer-assisted language learning, automated writing evaluative tools, and data-driven learning approaches utilizing corpus tools. He currently works for Michigan Judicial Institute as an education analyst.

Website: <https://pfauadam2326.wixsite.com/pfauadam>