

Analyzing AI Hallucinations to Train Critical Thinking and Increase Self-Efficacy in First-Semester EFL University Students

Arifi N. Waked^{1,2} , Hebah Alamr^{1,2}

¹Department of Sciences and Human Studies, Prince Mohammed Bin Fahd University, Saudi Arabia,

²Center for Cognitive Intelligence, Innovation, and AI-Driven Societal Impact, Saudi Arabia.

How to cite: Waked, A. N.; Alamr, H. (2026). Analyzing AI Hallucinations to Train Critical Thinking and Increase Self-Efficacy in First-Semester EFL University Students. In: 12th International Conference on Higher Education Advances (HEAd'26). Valencia, 15-18 June 2026. <https://doi.org/10.4995/HEAd26.2026.21805>

Abstract

This study examined the impact of critical thinking exercises analyzing AI hallucinations on self-efficacy in female native speakers of Arabic enrolled in their first semester at an English-medium university in Saudi Arabia. Participants included 109 students enrolled in one of eight sections of an academic writing course. Four sections engaged in an in-class exercise which students were told was intended to improve both their critical thinking and writing skills. Students uploaded their writing to a chatbot prone to hallucinations and prompted it to correct their grammar prior to analyzing output to determine which suggestions were accurate and which were hallucinations. Self-efficacy was measured in all students at the beginning and end of the semester. Only students in the testing group showed significant improvement in self-efficacy between pre and post-tests, suggesting that AI hallucinations serve as a beneficial tool to teach critical thinking to first-semester university students, thus increasing their self-efficacy.

Keywords: *Self-Efficacy; Critical Thinking; EFL; Writing Education; AI Hallucinations; AI Hallucination Pedagogy*

1. Introduction

When Microsoft's OpenAI released their first version of Chat Generative Pre-Trained Transformer (ChatGPT) in November of 2022, a previously inconceivable world of pedagogical opportunities for English as a foreign language (EFL) learners suddenly appeared quite literally at our fingertips. ChatGPT and similar generative artificial intelligence (AI) programs use Large Language Model (LLM) Functionality to accomplish a range of tasks, including translation, text summarization, and generating conversation. They are generally able to provide responses that are callbacks to previous prompts to provide further clarification and refine previously generated output (Soco, 2025). However, LLMs are most famously known for their ability to

generate written output intended to imitate human-like responses when given prompts generated by the human users of these programs. In this domain, ChatGPT's output in particular is known to be generally accurate and informative, as well as both lexically and semantically coherent in a wide variety of genres (Lo, 2023; Waked et al., 2024a).

For EFL courses, LLMs have primarily been explored as a means to provide individualized feedback to enhance foundational skills of English-language writing. Students are generally tasked with completing short compositional responses to prompts provided by instructors. As studies primarily focus on beginning-to-introductory level EFL education, LLM-based activities in the extant literature often concern the instruction of highly specific aspects of the English language, such as the correct use of a singular grammatical tense. Such exercises have been described as empowering, as they allow students to engage in self-directed learning (SDL) by noting their own errors in the individualized feedback provided. For moderate-to-proficient university-level EFL writers, it has been shown that students are able to use the real-time, individualized feedback of LLM programs to recognize and reflect on patterns of errors in their own writing. This feedback allows students to engage in SDL activities that positively correlate with higher scores on both the grammar and overall compositional competence later exhibited in handwritten essays (Waked, 2025).

While the output of LLMs can serve as an incredible pedagogical tool in a wide variety of ways, the reality of AI hallucinations and what they may mean for SDL has yet to be thoroughly examined. In the context of AI, Adam and Novian (2025; p. 110) define a hallucination as "AI-generated inaccuracies that challenge academic integrity." Thus far, academic discussion of the pedagogical implications of hallucinations centers around the fact that they serve to make LLMs unreliable learning tools, as students may be unable to differentiate between accurate and inaccurate feedback (e.g., Ahmad et al., 2023). However, as of yet, no empirical studies have attempted to make use of the fact that detecting AI hallucinations is an important aspect of technological literacy, requiring critical analysis of and engagement with AI-generated output. It can be argued that the real pedagogical weakness of AI is not the mere existence of hallucinations themselves, but rather the fact that students may not be encouraged to actively engage with AI-generated output in a way that trains them to recognize said hallucinations.

Passive engagement with AI has been hypothesized to lead to decreased critical thinking skills. Critical thinking is an essential life skill that can be related to every academic discipline. As such, it has several overlapping, yet differently nuanced, definitions. For the purposes of general educational skills, critical thinking involves the ability to apply situationally appropriate logic based upon analytical skills and reflective thinking (Altun & Yildirim, 2023). The ability to critically analyze feedback to best select the most appropriate and beneficial means of improvement can be especially true of cognitively demanding tasks, such as writing. Writing is a task that is heavily dependent upon working memory, the function of memory responsible for short-term storage, retrieval, and manipulation of information. Writing itself is also a

cyclical process of drafting, receiving feedback, critically engaging with feedback, and revising earlier drafts. This process becomes even more complex for EFL learners when the concept of negative language transfer is introduced. Negative language transfer can be defined as any time that a language user applies the rules of their native language to the use of their foreign language in a way that does not fit the phonological or morpho-syntactic rules of the foreign language (Waked, 2024b). When facing AI hallucinations in corrective feedback, improving critical thinking skills becomes even more important for the academic success of EFL writers.

The current study differs from those in the extant literature as it intentionally employs an LLM chatbot known to have a high rate of hallucinations (48%; Mauran, 2025). This program was selected to encourage EFL writers to actively engage in the critical thinking process of analyzing the accuracy of suggested grammatical changes to their own written work. While reflection on AI-generated feedback has been shown to be a useful tool for long-term learning in EFL writers (Waked, 2025), students' abilities to learn to identify feedback as either a true learning opportunity or as a hallucination has yet to be explored. This study also differs in that, rather than measuring course outcomes or other academic outcomes, it examines the development of self-efficacy, a skill widely associated with both long-term academic and career success and overall lifelong satisfaction (e.g., Gloudemans et al., 2013). In the current study, critical analysis of AI feedback and detection of hallucinations offers a unique, real-world opportunity to allow students to engage in critical thinking in a way that can be tangibly related to their own academic writing. For first-semester freshmen, the cyclical relationship between self-efficacy and academic success makes it all the more imperative to determine whether incorporating explicit training on a critical thinking exercise can be associated with increased self-efficacy at this crucial point in a student's academic development.

As such, the current study aimed to answer the following research question:

Does completing critical thinking exercises analyzing hallucinations in generative AI output increase students' self-efficacy?

Based on the previously discussed research, the following hypotheses were formed:

H1: Giving students exercises analyzing AI hallucinations throughout the semester will increase students' belief in the development of their own critical thinking abilities (Altun & Yildirim, 2023). This will lead to an increase in self-efficacy between the beginning and end of the semester as students become more confident in their ability to critically analyze AI output.

H2: Students in the testing group who are given critical thinking exercises analyzing AI hallucinations throughout the semester will show greater self-efficacy at the end of the semester than students in the control group who were not given such exercises.

In the current study, 109 first-semester female students enrolled in one of eight sections of an introductory academic writing course were assigned to either the test or control condition.

Students in four sections were given a critical thinking task in which they analyzed the output of an LLM chatbot when given the text of an unsubmitted essay assignment and the prompt, "Correct my grammar. Highlight all changes." Knowing that the chatbot had a high rate of hallucinations, students analyzed feedback to determine which changes reflected true errors in their own composition and which were instead AI hallucinations. Students then submitted a reflection paragraph in which they described the types of errors in their writing identified by the chatbot (e.g., tense inconsistency or errors in capitalization) as well as what "corrections" had been made by the chatbot that did not correspond to true errors in their essays (hallucinations). Students provided justification as to why these suggested corrections were in fact hallucinations. This exercise was repeated throughout the semester for three of the four essays used as assessments in this class. In the control condition, students in the remaining four sections of the same introductory academic writing course were not given any reflective task prior to submitting essay assignments. Bandura's (1997) General Self-Efficacy Scale (GSES) was given to all participants in the third week of the semester prior to beginning the first essay assignment and in the tenth week of the semester following the submission of the third essay assignment.

2. Methods

2.1. Participants

Participants included 109 first-semester female EFL students attending an English-medium university located in Saudi Arabia. Students were native speakers of Arabic. Participants were enrolled in one of eight sections of an introductory academic writing course and ranged in age from 18-21 years with a mean age of 18.8 years. Participants included only students who completed the semester (i.e., did not drop the course).

Students in all sections learned to write essays belonging to four different genres (descriptive, compare and contrast, persuasive, and informative). However, students in four sections ($N = 57$) were given the critical thinking task of analyzing generative AI hallucinations as part of their first three essay assignments. Students in the remaining four sections ($N = 52$) did not complete any critical thinking/analytical tasks.

2.2. Materials

Materials used in this study include the GSES, which was completed by all students in both week three and week ten of the sixteen-week semester. This measure contains eight statements presented on a five-point Likert scale and measures an individual's sense of self-efficacy (Bandura, 1997). For the testing group, students were also required to submit the text of three of their four total essay assignments to the LLM chatbot, OpenAI GPT Mini-4o. While this chatbot presents similar linguistic capabilities as standard OpenAI GPT chatbots, such as

ChatGPT, the mini version used in this study is designed to produce output more quickly and efficiently (OpenAI, 2026). However, its emphasis on speed over precision leads to a high likelihood of hallucination unless prompts are extraordinarily precise and specific (Mauran, 2025).

2.3 Procedure

The study itself is quasi-experimental as students were divided into the testing and control groups based upon the section of the academic writing course for which they personally chose to register. Aside from critically analyzing AI hallucinations, all sections were taught as similarly as possible.

In the third week of the semester, prior to beginning work on the first assignment, students were all given the GSES to evaluate their sense of self-efficacy at the time. Throughout the semester, students were taught to write essays in four genres (descriptive, compare and contrast, persuasive, and informative) using the five-paragraph essay format. Students worked on independent essays, but were placed into groups for workshopping and writing support. Students were assigned to groups based on the sentence formation and sentence construction tasks of the Wechsler Individual Achievement Task-III, increasing the likelihood of relatively balanced writing skill levels across all groups within classes. All students were allowed two weeks to write in class and were forbidden from using any AI program to generate their work. To discourage generative AI use, students were required to submit updated versions of their essays to the Blackboard platform at the end of each class. Thus, students were aware that their instructor had the means with which to compare writing styles in the uploaded drafts to their final submissions should there be any suspicion of AI misuse.

For the testing group, students were explicitly told that they would have the opportunity to use a generative AI program to check the grammar of their essays before submitting their final drafts. Students were made aware that this was intended to be a critical thinking task in which they identified hallucinations as well as a task of recognizing patterns of errors in their own writing. In the class immediately prior to the submission deadline, students were instructed to paste the text of their completed essays into OpenAI Mini-4o with the prompt, "Correct my grammar. Highlight all changes." Students identified patterns of errors corrected by the chatbot, such as consistent corrections of errors in subject-verb agreement. Students also identified "corrections" that did not appear to correspond to a pattern of errors in their own writing. In a reflection paragraph, students described the types of errors in their writing correctly identified by the chatbot as well as a description of each hallucination. Students provided justification for their judgement of "corrections" as in fact hallucinations by using material from their textbook and approved academic websites, or by providing an explanation of connotative meaning misinterpreted by the chatbot. This task was repeated for the second and third essay

assignments; however, as the fourth and final essay assignment was due shortly before the final examination period, this task was not repeated prior to submission.

3. Results

A repeated measures ANOVA was calculated for total scores of the GSES in pre- and post-tests for both the control and test groups. Results of the ANOVA [$F(3,1) = 7.97$, $p < 0.001$] indicate a significant difference between the self-efficacy scores in the four conditions tested. Table 1 displays the descriptive statistics of the test and control groups in both the pre- and post-test conditions of this test of self-efficacy.

Table 1. Number (N), Mean, and Standard Deviation (SD) for Scores on the General Self-Efficacy Scale

	N	Mean	SD
Test Group Pre-Test	57	32.26	3.88
Test Group Post-Test	57	35.47	4.06
Control Group Pre-Test	52	32.79	4.44
Control Group Post-Test	52	33.11	3.76

To determine which scores differed significantly from one another, paired sample t-tests were used. As shown in Table 2, only the test group showed significant differences in self-efficacy between the beginning and end of the semester.

Table 2. Differences between Self-Efficacy Scores Before and After Conducting Critical Thinking Exercises using Analysis of AI Hallucinations in Both the Testing and Control Conditions

	t	df	Significance
Test Start and End	6.26*	56	$p < 0.001$
Control Start and End	0.79	51	$p = 0.43$
Test and Control Start	-1.66	51	$p = 0.22$
Test and Control End	0.53	51	$p = 0.16$

4. Discussion

This study aimed to explore the effect of critical thinking exercises analyzing hallucinations in generative AI output on self-efficacy in first-semester freshman students. Results showed a significant difference between the pre- and post-tests of the testing group. Students in this group completed the AI hallucination analysis exercise three times throughout the semester. However, the results did not show any differences between the pre- and post-tests of the control group, which did not complete any such exercise. This provides support for our first hypothesis,

predicting that critically analyzing AI hallucinations in one's own writing can lead to an increase in self-efficacy.

Our results do not, however, support our second hypothesis that a significant difference in self-efficacy would be found between the test and control groups at the end of the semester. While Table 1 shows that the mean of the testing group at the end of the study is slightly higher than that of the control group, this difference is not significant. This discrepancy could be clarified by testing a larger group of participants over multiple semesters. Additionally, further care should be taken to account for confounding factors, such as the number of credit hours for which a student is enrolled, the time of day at which the class takes place, and the length of students' commutes to campus. Other factors that may lead to increases in self-efficacy between the beginning and end of a semester, such as joining clubs or satisfaction with one's major, should also be considered in future work to more precisely determine the impact of this exercise on self-efficacy.

There are several other areas of further investigation that can be developed from this preliminary study. Future research could examine final course grades and compare first-semester GPAs between members of the testing and control groups. As participants in this study were in the first semester of their freshman year, the long-term academic impacts of providing critical thinking exercises analyzing AI hallucinations may be of interest. Further research could additionally compare students who experienced training in identifying AI hallucinations with those who did not with regard to the way in which they view AI as either a useful tool that assists in the completion of their own work or as a substitute for exerting academic effort. It is also important to note that participants in this study were limited to female students at one university in Saudi Arabia. As such, this study should be replicated among students of differing populations.

References

- Adam, M. A., & Novian, H. (2025). In Praise of AI Hallucinations: Re-imagining Critical Pedagogy and Truth Verification in the Age of Algorithmic Authority. *Indonesian Journal of Education and Youth Development*, 1(3), 110-123.
- Ahmad, Z., Kaiser, W., & Rahim, S. (2023). Hallucinations in ChatGPT: An unreliable tool for learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4), 12.
- Altun, E., & Yildirim, N. (2023). What does critical thinking mean? Examination of pre-service teachers' cognitive structures and definitions for critical thinking. *Thinking Skills and Creativity*, 49, 101367. <https://doi.org/10.1016/j.tsc.2023.101367>
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191-215. <https://doi.org/10.1037/0033-295X.84.2.191>
- Gloudemans, H. A., Schalk, R. M., & Reynaert, W. (2013). The relationship between critical thinking skills and self-efficacy beliefs in mental health nurses. *Nurse education*

- today, 33(3), 275-280. <https://doi.org/10.1016/j.nedt.2012.05.006>
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- Microsoft. (2025). *Know it all: ChatGPT and its capabilities*. Soco. <https://soco.com.au/know-it-all-chatgpt-and-its-key-capabilities-that-will-surprise-yourteam/#:~:text=ChatGPT%20is%20particularly%20well%2Dsuit%2C%20research%20and%20many%20more.>
- Muran, C. (19 April, 2025). *OpenAI's o-3 and o-4 Mini hallucinate way more than previous models: A troubling issue nestled in OpenAI's technical report*. Mashable Middle East. <https://me.mashable.com/tech/54945/openai-o3-and-o4-mini-hallucinate-way-higher-than-previous-models>
- OpenAI. (18 July, 2024). *GPT-4o Mini: Advancing cost-efficient intelligence*. OpenAI. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>
- Waked, A., Abdelsalam, H., Ashraf, M., Pilotti, M., & El Alaoui, K. (2024a). The impact of falsely detecting AI-generated text on academic assessment. In M. Shelley & O. T. Ozturk (Eds.), *Proceedings of ICRES 2024-- International Conference on Research in Education and Science* (pp. 573-586), Antalya, Turkiye. ISTES Organization.
- Waked, A. N., El-Moussa, O., Pilotti, M. A., Al-Mulhem, H., El Alaoui, K., & Ahmed, R. (2024b). Cultural considerations for the second language writing anxiety inventory: Saudi Arabian female university students. In *Frontiers in Education* (Vol. 9, p. 1288611). Frontiers Media SA.
- Waked, A. (2025). Incorporating AI-Administered Personalized Feedback in an Inquiry-Based Learning EFL Writing Class. In: 11th International Conference on Higher Education Advances (HEAd'25). Valencia, 17-20 June 2025. <https://doi.org/10.4995/HEAd25.2025.19871>