Online Learning in Educational Research
Volume 6, Issue 3, 253-265.
e_ISSN: 2798-2351
https://www.journal.foundae.com/index.php/oler/index
DOI: 10.58524/oler.v6i3.1284

An Exploratory Analysis of ChatGPT and Gemini Responses on False Positives in AI-Generated Text Detection

Ahmet Erdost Yastıbaş
Gazi University,
TURKEY

Ali Dincer*
Ondokuz Mayıs University,
TURKEY

Tevfik Darıyemez
Atatürk University,
TURKEY

Article Info

Article history:

Received: June 11, 2026
Revised: August 22, 2026
Accepted: September 22, 2026

Keywords:

Artificial intelligence;
AI detection systems;
ChatGPT;
False positives;
Gemini.

Abstract

The increasing use of AI tools has raised concerns about the mislabeling of human-written texts as AI-generated (i.e., false positives), potentially leading to unfair academic decisions in education settings. This study aimed to explore AI-based explanations for this misclassification and develop strategies that L2 teachers can use to reduce them. Adopting a qualitative framework, a textual corpus was created from responses by ChatGPT and Gemini to two prompts about mislabeling texts as AI-generated. The corpus comprised 4 responses totaling 2024 words. It was thematically analyzed. The findings showed that, according to ChatGPT and Gemini, there are five overarching themes regarding possible reasons: linguistic monotony, conventional writing patterns, bias related to non-native English writers, overly polished writing, and limitations of AI detection systems. According to the findings, ChatGPT and Gemini recommended three strategies for L2 teachers to minimize false positives: trace evidence of authors' writing processes, customize their writing assessments, and use verification techniques in decision-making. Overall, this study contributes to the literature by providing a structured synthesis of AI-generated explanations of false positives and identifying pedagogically relevant strategies for L2 writing assessment, while emphasizing the importance of critically interpreting AI detection results.

To cite this article: Yastıbaş, A. E., Dincer, A., & Darıyemez, T. (2026). An exploratory analysis of ChatGPT and Gemini responses on false positives in AI-generated text detection. *Online Learning in Educational Research, 6*(3), 253-265. https://doi.org/10.58524/oler.v6i3.1284

INTRODUCTION

A foreign/second language (L2) teacher asks students to submit a 250-word opinion essay on "What should people do to avoid the worst effects of climate change?" After submission, the teacher uses a free AI-writing detector, which misidentifies one student's essay as entirely AI-generated, resulting in the learner's failure. Although the learner claims authorship, the teacher relies on the detection result. This illustrative example raises two questions: "What might cause such mislabeling?" and "What strategies can L2 teachers apply to minimize the risk of incorrectly labeling authentic L2 students' writing as AI-generated?" For example, Waked et al. (2024) found that the handwritten work of English as a foreign language (EFL) students in an academic writing course could be identified as AI-written by AI detection tools. According to them, such misclassification may negatively affect the academic assessment of EFL students who excel in academic writing. As Waked et al. (2024) and Liang et al. (2023) note, AI detection tools might misidentify texts written by non-native English writers as AI-generated, which could negatively affect those students' academic careers. In addition, the participant L2 writers in Wang and Wang's study (2025) used ChatGPT in different parts of their writing, such as brainstorming and editing. This suggests that L2 writing may be influenced by AI, making it the focus of the present study.

* **Corresponding Author**
Ali Dincer, Ondokuz Mayıs University, TURKEY ✉ ali.dincer@omu.edu.tr

As seen in the example from an L2 classroom, false positives can have negative consequences, especially in formal educational assessments. Identifying the author of a text has become increasingly difficult in the era of GenAI, requiring educators to adopt new approaches to address this problem (Cotton et al., 2024). In the realm of L2 learning and teaching, writing is an indispensable part of language development as a productive skill. L2 teachers frequently ask L2 learners to write in English for assessment and feedback. They may tend to use AI detection systems to assess students' writing (Dao et al., 2025). However, as Liang et al. (2023) and Waked et al. (2024) note, L2 writers may encounter false positives in their L2 classes if AI detection tools are used to identify the author of a text, and L2 writers' assessment may be negatively affected by false positives. Therefore, false positives in L2 writing should be considered a serious issue that needs to be dealt with cautiously.

Research on AI detection tools appears to investigate the performance of those tools in identifying who or what wrote a text in terms of accuracy and/or reliability (Hadra et al., 2026; Van Vlasselaer et al., 2026) and what may affect their performance in identifying who or what wrote a text (Giray et al., 2026). Still, the literature does not seem to focus on how GenAI expresses the potential reasons for false positives and suggest strategies for L2 teachers to use in their L2 classes to reduce false positives. Accordingly, the present study focused on false positives and aimed to investigate their possible causes and how L2 teachers could mitigate them in their L2 classes through AI-based reasoning. Given that GenAI models are continually trained on newly created data when new versions are released and build their answers on large corpora through AI-based reasoning, this study aimed to investigate the responses of two widely accessible AI systems, ChatGPT and Gemini, to develop a comprehensive view of how they mislabel the authors of the text. Two of the main rationales behind choosing these two GenAI tools were their uses in L2 academic writing research (Kurt & Kurt, 2024; Yan, 2024; Derakhshan, 2025; Tran, 2025; Zainurrahman et al., 2026), with ChatGPT being the most visited GenAI in March 2024 by Liu & Wang. (2024) and April 2025 by Liu et al. (2025), and Gemini being the second and third most visited GenAI in March 2024 (Liu & Wang, 2024) and April 2025 (Liu et al., 2025), respectively. Their increased accessibility and practicality as general-purpose tools for gathering information and seeking guidance in everyday contexts were another reason why they were chosen as data collection tools in the present study.

These two prominent AI agents, with over 1 billion active users as of 2026 (OpenAI, 2026; Reid, 2026), are large language models (LLMs) that can generate convincing natural-language responses by mimicking patterns from massive amounts of text data on the Internet (Stokel-Walker, 2023). These LLMs offer a wide array of services and impressive capabilities (AlSagri et al., 2025), but learners might misuse such services and capabilities. For instance, they may use ChatGPT to complete their written assignments (Barrot, 2023) and submit them as if they had prepared them themselves without acknowledging this practice (Perkins, 2023). LLMs' ability to generate convincing natural-language responses (Stokel-Walker, 2023) seems to have improved GenAI's ability to mimic human writing, making it harder for teachers to identify who or what produced the text (Cotton et al., 2024; Fleckenstein et al., 2024). As a solution to this problem, AI detection tools may be used. Because of false positives, relying solely on AI detection tools to identify a text's author may be problematic.

As highlighted in the literature, distinguishing between human- and AI-generated writing remains challenging, and AI detection systems can produce both false positives and false negatives (Alexander et al., 2023; Gao et al., 2023; Giray, 2024; Weber-Wulff et al., 2023). Makiev et al. (2023) similarly found that human reviewers had difficulty distinguishing AI-generated from human-written abstracts, while the accuracy of AI detection tools also varied across systems. These systems may be affected by limitations such as contextual constraints and algorithmic biases, which can lead to human-written texts being classified as AI-generated (Giray et al., 2026; Ibrahim, 2023; Lo, 2023; Weber-Wulff et al., 2023). They may also misclassify AI-generated texts as human-written or struggle with texts containing both human- and AI-generated content (Alexander et al., 2023; Weber-Wulff et al., 2023). Such limitations make AI-based authorship identification unreliable when used as the sole basis for educational decisions.

In the education domain, educators can take several steps to prevent the misuse of AI and ensure the responsible use of AI tools. Before submitting a written document, teachers can educate learners on plagiarism, set clear goals for AI policy in writing assignments, monitor students' writing performance, ask students to submit drafts earlier before final submission, and so forth (Cotton et al., 2024). After submitting the documents, teachers can, in addition to using AI detection tools,

employ several strategies to determine whether the text was written by students or by an AI agent (Cotton et al., 2024). In short, teachers can check the texts by focusing on several aspects of writing submitted by L2 learners, such as the use of language (i.e., grammar), word choice, how cohesive the text is, grammatical errors, and repetition (Alexander et al., 2023; Nguyen & Barrot, 2024), as well as clarity of the text (Nguyen & Barrot, 2024) and paragraph/essay structure (Alexander et al., 2023; Nguyen & Barrot, 2024). Teachers can also check for factual errors and citations in the text, as LLMs tend to generate hallucinations and fabricate references (Cotton et al., 2024; Dincer & Yesilyurt, 2025). In the case of ChatGPT, a learner may even forget to delete its responses, including confirming the learner's input and acknowledging its non-human status. Such expressions also help determine the author of a text (Alam et al., 2024). In short, teachers can take various actions to ensure students' responsible use of AI.

The possible reasons for false positives and strategies for L2 teachers to reduce false positives in L2 classes do not seem to have been studied from the perspective of conversational AI tools in the literature. Therefore, the present study aimed to examine how two conversational AI tools (i.e., ChatGPT and Gemini) explain the possible causes of false positives and what they suggest L2 teachers should use to mitigate them in their classes. In the specific case of L2 writing, L2 writing classes require L2 learners to write and submit various kinds of texts in L2, and L2 learners' misuse of these tools threatens the reliability and accuracy of their writing assessments. L2 teachers may use various strategies and AI detection systems to distinguish human-generated from AI-generated writing. In cases of uncertainty about AI-generated writing and the questionable reliability of AI detection tools in L2 classes, L2 teachers may consult AI chatbots for explanations and practical guidance. By critically examining and synthesizing the responses of two conversational AI tools, this study offers an additional perspective on false-positive classifications and their pedagogical implications for L2 writing assessment.

METHOD

Research Design

This study employed a qualitative research approach based on the analysis of the responses generated by ChatGPT and Gemini (see Figure 1). The study focused on two dimensions: the possible reasons for false positives in labelling authorship in L2 texts and strategies that L2 teachers can use to minimize such misclassification.

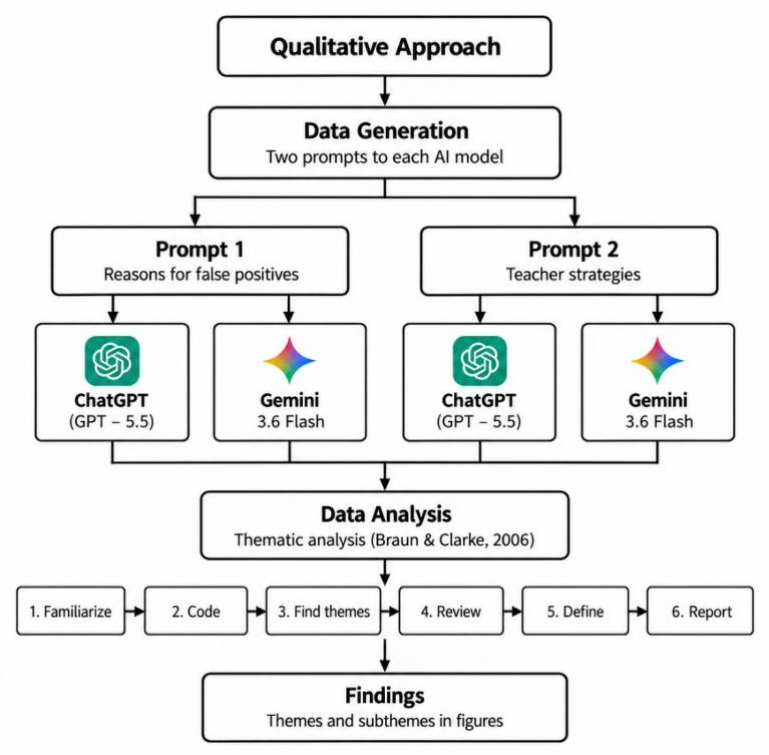


Figure 1. Research Flowchart

The textual data came from ChatGPT and Gemini based on the following assumption: these AI models are trained on newly created corpus data on the Internet, and they respond to human queries; their answers are built on updated data through reasoning; therefore, when they are used as data-collection tools in research, researchers may gain a comprehensible understanding of the modern phenomenon under investigation. It should be noted that, rather than treating AI outputs as empirically established knowledge here, the authors acknowledge that AI-generated corpus data may have drawbacks such as bias and hallucinations, and then critically evaluate the outputs.

Data Collection Procedure and Tool

The data consisted of four textual responses collected through two prompts from ChatGPT and Gemini, which were used for data collection for the present study because both were used in L2 writing research (Kurt & Kurt, 2024; Derakhshan, 2025) and were among the top three most visited GenAIs (Liu & Wang, 2024; Liu et al., 2025). The prompts were "What might cause AI detection systems to label human-written text as AI-generated incorrectly?" and "What strategies can L2 teachers apply to minimize the risk of AI detection systems incorrectly labeling authentic L2 students' writing as AI-generated?" These prompts were entered into the latest free versions of ChatGPT (GPT–5.5) and Gemini (Gemini 3.6 Flash) in July 2026. Each prompt was executed once. This can be considered a design limitation for the present study, but it was preferred because executing prompts multiple times may yield similar responses. Default settings were used, web search was enabled, and both tools were run in a web browser.

Each response was collected in a separate session to minimize the AI tool's conversational memory effects. The responses were saved verbatim as text files and used to create a textual corpus of 4, totaling 2024 words, whose data units were complete and unedited ChatGPT- and Gemini-generated text. The responses of ChatGPT in the corpus included 853 words, and Gemini's ones were composed of 1171 words. Rather than comparing the two AI systems, ChatGPT and Gemini, the responses were treated as a single corpus to identify recurring patterns, ideas, and pedagogical suggestions. Due to the small corpus prepared for this study, it is exploratory and qualitative, focusing only on ChatGPT and Gemini's responses regarding possible reasons for false positives and strategies L2 teachers can use to reduce false positives in their L2 classes. As the study used two publicly accessible, free tools for data collection and involved no human participants, formal human-subjects ethics approval was not required.

Data Analysis

The textual data from the analysis coming from two resources were analyzed following the six stages in Braun and Clarke's thematic analysis framework, which are "familiarizing yourself with your data", "generating initial codes", "searching for themes", "reviewing themes", "defining and naming themes", and "producing the report". In the first stage, all participants in data analysis read the textual corpus several times to get familiar with it. In the second stage, they derived the initial codes from the data. In the third, fourth, and fifth stages, they organized the codes into potential themes, refined them, and developed definitions and names for them. In the last stage, they wrote the report. ChatGPT and Gemini's responses were analyzed separately according to each prompt. When their responses were similar or identical, they were presented together in the results; when they differed, they were presented separately. Thematic analysis was conducted using Microsoft Excel spreadsheets. The researchers and an external expert with a Ph.D. in English language teaching and experience in qualitative data analysis first conducted thematic analyses separately, compared their analyses, addressed differences, and finalized them by consensus to ensure the study's trustworthiness.

RESULTS AND DISCUSSION

Possible Reasons for False Positives

Thematic analyses showed five main possible reasons for false positives. These reasons are presented with the minor-related reasons in Figure 2.

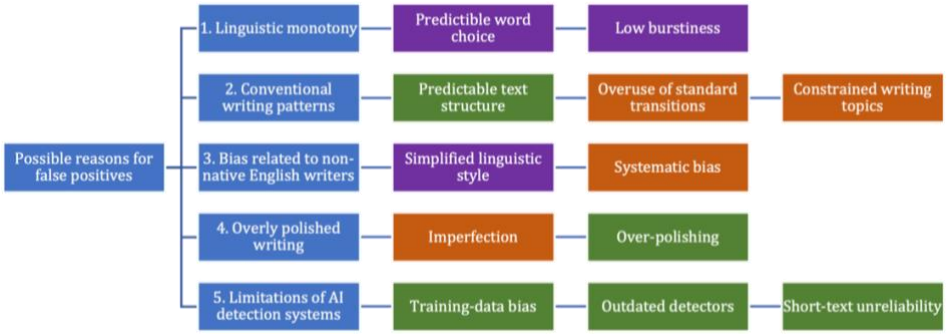


Figure 2. GenAI Explanations for False-Positive Detection of Human-Written L2 Texts
**Note. Blue = overarching categories; purple = both models; green = ChatGPT; orange = Gemini.*

Linguistic monotony: Human writing tends to indicate high linguistic variety in terms of word choice, sentence structure, and sentence length, yet AI-generated writing seems to promote a consistent approach in these linguistic aspects, so if human writing shows low linguistic variety in these aspects, it may be misclassified as AI-generated writing by an AI detection system. That is, if word choice is predictable (i.e., predictable word choice), and sentence lengths and structures are (very) similar in human writing (i.e., low burstiness), like they are in AI writing, that human writing may be misidentified as AI-generated. The following quotes support this. For predictable word choice, ChatGPT writes, "Using common words and straightforward sentence patterns can lower linguistic 'surprise' (often called perplexity), which some detectors associate with AI." Gemini mentions, "AI models are mathematically optimized to choose the most statistically probable next word, resulting in smooth, uniform, and highly predictable sentences." For low burstiness, ChatGPT explains, "A uniform tone, sentence length, and vocabulary throughout a document may appear statistically 'AI-like." Gemini states, "If you write with consistent sentence lengths, standard vocabulary, and a highly structured rhythm, a detector flags the text."

Conventional writing patterns: In specific genres of writing, people are supposed to follow a specific structure to organize their writing. Following a specific writing structure for such a text creates a standardized structure that aligns with the data used to train an AI. Therefore, when a piece of human writing is organized according to a specific textual structure, it may be misidentified as AI-written. The quote from ChatGPT ("Essays, reports, business documents, and technical writing often follow predictable structures that detectors may flag") supports this (i.e., predictable text structure). Specific writing conventions are not limited to textual structures. They also include which words need to be used to connect ideas in a piece of writing. For example, when an L2 learner is asked to write a paragraph, they are expected to use words like first, second, third, or in conclusion. If such words are overused in human writing, it may create a resemblance between human writing and AI writing, so that human writing may be miscategorized as AI-generated. The reply of Gemini ("Human text that frequently uses words like Furthermore, Moreover, in conclusion, It is important to note, Therefore, or Crucially will heavily skew the detector's score toward AI") reveals this (i.e., overuse of standard transitions). In specific genres, people may be asked to write about very specific topics, but those topics may determine how they write strictly because of their textual content. Gemine's explanation, "Some topics leave very little room for stylistic flair," clearly indicates this.

Bias related to non-native English writers: Non-native English writers seem to prefer grammatical, lexical, and syntactical simplicity (i.e., simplified linguistic style) in their writing. As a result, their word choice may become predictable, and the sentence structures and lengths in their writing may be similar. Therefore, simplified linguistic style in non-native writers' texts may cause an AI detector system to misclassify a human text as an AI text, as ChatGPT (whose response is "Some non-native speakers use simpler vocabulary and sentence structures, which can unintentionally resemble patterns detectors associate with AI.") and Gemini (whose reply is "Writers learning English as a second language often rely on simpler, highly correct grammatical structures and standard vocabulary to ensure clarity. Because their writing avoids complex, idiosyncratic native slang or varied syntax, detectors routinely misclassify their original work as AI-generated.") mentioned. Gemini states, "… AI detectors are significantly biased against non-native English speakers." This sentence also implies that there may be systematic bias in AI detection systems.

Overly polished writing: Unlike AI-generated writing, there is a high possibility that human writing may involve mistakes, such as those in writing mechanics and the use of language, so human writing may need to be polished in these aspects, which can be done with proofreading and editing. After being proofread and edited, a human-written text may become flawless. If it is over-polished, this situation may lead to false positives. This can be understood from ChatGPT's excerpt: "Well-edited text with few errors can resemble AI-generated output… If a human revises a piece extensively, removing errors and smoothing transitions, the final result may resemble polished AI text." In addition, a person may have an imperfectionist writing style and write a text free of mistakes. Such human-generated text may cause an AI detection tool to incorrectly flag it as AI-generated. It is clearly understood from the quotation from Gemini: "If a human writer produces incredibly clean, middle-of-the-road prose that lacks personal anecdotes, unique stylistic risks, or regional slang, the detector views the text as 'too perfect' – a trait it associates with machines."

Limitations of AI detection systems: Like AI tools, AI detection systems are trained on data to classify human- and AI-written texts. Therefore, training data sets seem to play a significant role in the quality, accuracy, and reliability of AI detection systems' assessments of text. If there are similarities between a human-created text and a data set used to train an AI detection system, it may lead to training bias against that text because it may be misidentified as AI-written. In contrast, another human-written text that does not share any similarities with the training data set may pass through the detection system. ChatGPT's response ("Many detectors are trained on specific AI models and datasets. Human writing that happens to share those characteristics can be misclassified") supports this finding. In addition, new versions of AI tools are released after they are trained on newly created data, so if an AI detection system is not trained accordingly, it may become outdated and cause false positives. This is clear from the quotation from ChatGPT: "As AI writing evolves, detectors may become less accurate unless they're continually updated." Also, an AI detection system needs long texts to determine the author because long texts provide it with enough data to process, whereas short texts are not as reliable in this regard. ChatGPT's answer ("Detection is generally less reliable on short passages because there isn't enough information for meaningful statistical analysis") reveals this finding.

Strategies for L2 Teachers to Use for Minimizing False Positives

Thematic analyses showed that there are three strategies L2 teachers can follow to mitigate AI detection systems' mislabeling of human-written texts as AI-generated. These strategies were presented in Figure 3.

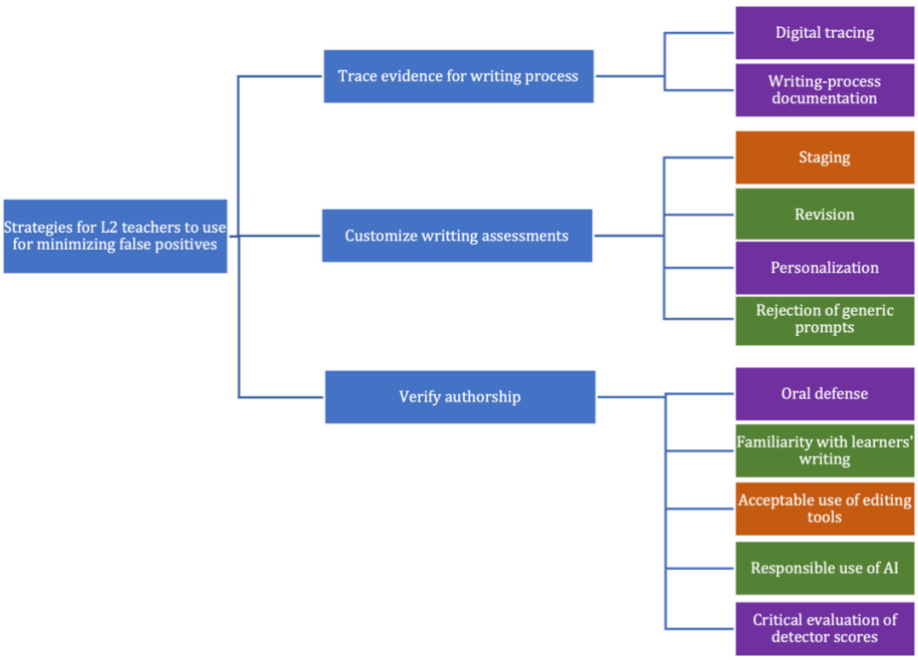


Figure 3. GenAI Strategies to Reduce False-Positive AI Detection of Human-Written L2 Texts
**Note.* Blue = overarching categories; purple = both models; green = ChatGPT; orange = Gemini.

Trace evidence for writing processes: Following the process that L2 learners go through when writing a text can give L2 teachers clues about the text's true authorship in cases of false positives. This can be done in two ways. The first one is digital tracing, in which L2 learners are expected to use online platforms that record the history of their activities as they write a text. The second one is the documentation of the writing process. In two ways, L2 teachers can gradually observe how a text was written by L2 learners so that they can identify the author in the event of a false positive. To illustrate, ChatGPT mentions, "Require brainstorming notes, outlines, multiple drafts, and reflections. Use version histories from collaborative writing platforms (e.g., Google Docs or Microsoft Word." Gemini writes, "Instruct students to compose their essays directly inside cloud-based editors like Google Docs, Microsoft Word Online, or specialized writing platforms. If an AI flag occurs, you can check the document's version history… Have students submit a brief reflection or change log alongside their assignments detailing how they developed the topic, what sources they found first, and how their thesis evolved."

Customize writing assessments: L2 teachers can customize their writing assessments in four ways. Firstly, L2 teachers can ask learners to prepare their writing assessments in stages and check their progression in each stage. This can be understood clearly from the quotation from Gemini: "Breaking a major paper down into sequential checkpoints completely neutralizes the utility of a 'one-click' AI essay generator and provides a clear human paper trail." As part of the writing process, revision should be included to enable L2 learners to enhance their writing based on the feedback they receive and to explain why and how they change their writing in response to that feedback. That ChatGPT writes, "Encourage students to revise based on feedback and explain the changes they made," indicates this. The third way to customize writing assessments is to add personalization, because when L2 learners write about general topics like environmental issues, AI tools can produce texts more easily, since their training datasets may include related content. However, when writing assessments require learners to explain a topic by adding specific, personalized information, AI tools may not produce such texts easily because personalized information may be about something learned in a class at a specific time, and AI tools' training data may not include such information. Therefore, L2 teachers may modify their writing assessments to require learners to explain a topic by integrating personalized information. This can be clearly shown by the excerpts of ChatGPT ("Ask students to connect writing to personal experiences, classroom discussions, local contexts, or recent course materials. These tasks are more difficult to outsource and easier for teachers to authenticate") and Gemini ("AI models struggle to invent specific, real-time authentic context. Instructors can structure assignments to demand hyper-localized or deeply personal elements that naturally break predictive AI patterns."). In parallel to personalization, L2 teachers can tailor their writing assessment topics to be more personalized rather than using generic prompts. To illustrate, Gemini expresses, "Replace open-ended prompts like 'Write an essay analyzing the themes of isolation in Frankenstein' with tailored alternatives: 'Compare the isolation of Frankenstein's monster to a specific, localized community event we discussed in class last Tuesday."

Verify authorship: In the event of false positives, L2 teachers can use several methods to verify whether a text was written by an L2 learner or by AI. The first is oral defense of written work. When L2 learners are asked to talk about their writing, L2 teachers can identify who or what wrote the text by assessing the information learners provide about their writing. The responses of ChatGPT support this ("Ask students to explain their arguments, sources, or word choice. Genuine authors can usually discuss their ideas and revisions…") and Gemini ("If a student's paper returns a high AI score, invite them to an informal, low-stakes chat. Ask them to explain a specific vocabulary choice they made, or have them summarize their core argument in their own words…") The second way is familiarity with learners' individual writing. Each writer has their own writing style, so do L2 learners. If L2 teachers are familiar with their learners' individual writing, they can be aware of what learners can and cannot do in their writing and use this familiarity to identify the writer of a text in case of a false positive. ChatGPT suggests using in-class timed or low-stakes writing for this purpose. ChatGPT's response, "Collect writing samples throughout the semester. Look for consistent patterns in grammar, vocabulary, organization, and common errors rather than relying on a single assignment," implies that writing portfolios can also be used for this purpose indirectly. The third way is to inform L2 writers about editing tools, that is, how to benefit from editing tools. Editing tools can help improve a text by providing feedback on writing mechanics and language use. Yet, using them heavily to edit

a text can be problematic. This is understood clearly from Gemini's excerpt, which is: "L2 students frequently rely on third-party tools to improve their grammar or fluency. However, over-editing can inadvertently trigger AI detection flags." Another way is for L2 learners to use AI responsibly in L2 writing. L2 learners may use AI when they write, but they may not be aware of how to use it responsibly or whether they have to explain their use of AI in their writing.

Therefore, L2 teachers can educate them about these issues and check statements about AI use in their texts to confirm whether a learner wrote the text. For this, ChatGPT expresses, "Clarify when AI assistance is permitted and when it is not. Encourage students to disclose any AI use according to course policy." The last way is about how to evaluate AI detection systems' scores. It is known that AI detection systems may misidentify the author of a text (i.e., false positives and negatives). Therefore, L2 teachers should not make decisions about learners' writing based solely on AI detection systems' scores; they can also use other sources, such as oral defense of the written work or documentation of the writing process, alongside those scores, to determine the author of a text. To illustrate, ChatGPT reports, "Current AI detection systems are not sufficiently reliable to serve as evidence of misconduct on their own. If a detector flags a paper, combine that information with other evidence, such as writing history, drafts, and conversations with students." Gemini explains, "Instructors should treat an AI detection score merely as a soft signal that warrants a conversation, never as standalone proof of academic dishonesty."

Discussion

This study investigated possible reasons for mislabeling student-written texts as AI-generated and provided strategies for L2 teachers to reduce false positives. By integrating the perspectives generated by ChatGPT and Gemini, this study provides a structured understanding of how conversational AI tools frame false-positive classifications and their implications for L2 writing assessment. The findings on possible reasons suggest that AI detection systems may mislabel human-written texts as AI-written, as noted in the literature (Ibrahim, 2023; Cheng et al., 2025; Weber-Wulff et al., 2023). According to the data, author-related factors, such as linguistic monotony, conventional writing patterns, bias against non-native English writers, and overly polished content, as well as AI tool-related factors, including the structural limitations of AI detection systems, appear to contribute to false-positive classifications.

Among the author-related factors, linguistic monotony emerged as a potential contributor to false-positive AI detection. If human writing has low linguistic variety, it may be incorrectly flagged as AI-written by AI detection systems. As the capacity of AI models like ChatGPT and Gemini appears to be strictly dependent on their training data sets, their tendency to produce text seems to follow predictable word choice and consistent sentence-level patterns when analyzed statistically (Grammarly, 2026). Regarding specific writing conventions, different genres have different conventions. Unlike free writing, some writing genres have specific writing requirements for writers to follow. For example, when an L2 learner writes an argumentative essay, they need to follow a specific organizational structure (i.e., an introduction paragraph, two or three body paragraphs, a refutation paragraph, and a conclusion paragraph) and use specific vocabulary, phrases, and transitions, such as "the opponents of this idea claim" and "however" in the refutation paragraph, about a specific debatable topic. Following such specific writing conventions may be problematic and lead to a false positive for human writing if the training data sets for AI models include texts that follow the same conventions. Bias against non-native English writers may also contribute to false positives (Giray et al., 2026).

Humans need language proficiency to read, listen, write, and speak. The language proficiency of non-native speakers is evolving because it takes time to develop the skills needed for reading, listening, writing, and speaking in that language. This may cause non-native speakers to adopt a simplified linguistic style in their writing, which may trigger AI detection systems, as such writing resembles the datasets used to train AI models. When it comes to overly polished writing, AI detection systems may struggle to distinguish human from AI authorship. They may flag human-written text as AI-generated, even for minimally polished text (Saha & Feizi, 2025). The absence of language errors in writing is also associated with AI-generated content, as indicated by the professors in Georgiou (2026). In other words, flawless writing appears to be associated with AI-generated content in both AI-detection-system-based and people-based assessments.

As an AI tool-related factor, the study shows that detection systems have structural limitations: they are not perfect tools and may have flaws (e.g., training data bias, outdatedness, and the need for longer texts to identify the author). That AI detection systems may have flaws is also mentioned in the literature (Giray et al., 2026; Gritsai et al., 2024). When the features of a human-generated text share the features of texts used to train AI detection systems, the evaluation of that text by AI detection systems may be considered biased, unlike another piece of human text that does not share such features, because shared features may trigger AI detection systems. AI tools continue to develop, but if AI detection systems are not updated accordingly, their accuracy in identifying the author tends to decrease. Another structural limitation is the text's length. If texts are short, AI detection systems may not have sufficient data to process, which appears to undermine their ability to identify a text's author (Cheng et al., 2025; Grammarly, 2026).

The findings on possible mitigation strategies indicate that L2 teachers can use three strategies to minimize false positives. As the first strategy, L2 teachers can trace evidence of the authors' writing process. One way to do so is to ask L2 learners to use online writing programs like Google Docs, which record the history of their writing activities. Another way is to require submission of documentation of the writing process and accept writing as a process-based product. "Students' ownership is stressed" (Graham & Sandmel, 2011) in process-based writing, so it requires the active participation of L2 learners in the writing process. If L2 writing classes are based on this approach, L2 learners need to go through multiple stages, including brainstorming, outlining, writing first drafts, proofreading and editing, and writing a final draft. As highlighted by Pramjeeth and Ramgovind (2025), assessment in higher education also needs to be reconceptualized in the AI era, and assignments should be designed to incorporate higher-order thinking skills, creativity, and problem-solving strategies that AI tools cannot easily replicate (Ncube et al., 2026).

Then, as the second strategy, L2 teachers need to customize L2 writing assessments. L2 teachers can succeed in different ways. The first way is to stage L2 writing and use each stage as a checkpoint. This approach prioritizes process over product in L2 writing. Process-based writing finds the process through which a piece of writing is produced more important than the piece of writing as a product (Onozawa, 2010). L2 teachers can include L2 writing revision and ask L2 learners to explain why and how a text is revised. When L2 teachers personalize their writing prompts, L2 learners may not benefit from AI models for writing their texts because these models may not include such personalized information in their training datasets. They can also avoid providing generic, broad writing topics that AI models can author easily.

As the third and final strategy, L2 teachers can verify a text's authorship in several ways. By integrating an oral defense of the written work, L2 teachers can ask L2 learners specific questions about various aspects of the text, such as language use or content, and assess authorship. By becoming familiar with L2 learners' writing styles, L2 teachers can understand how learners in their classes write within time limits by reading and providing feedback on their writing. They can educate L2 learners on the acceptable use of editing tools and AI assistance and highlight how heavy editing can replace the human touch in texts with AI-related features. L2 teachers can also teach L2 learners how to use AI responsibly in their writing and how to explain its use. Responsible use of AI in writing also requires a clear explanation of AI use in writing (i.e., how it was used when the text was written). To verify authorship, L2 teachers should not rely solely on AI detection system scores when they have questions about a learner's text. It may be due to flaws in AI detection tools and false positives (Bendo, 2026; Giray, 2024). Instead, teachers were advised to use detection scores alongside other verification methods, such as oral defense of written work (Ncube et al., 2026).

In addition to the three strategies, the following implications are also drawn from the literature (Dincer & Yesilyurt, 2025; Giray et al., 2026; Griche & Bennis, 2026; Ncube et al., 2026). L2 teachers should consider writing as a process and focus on the processes through which L2 learners write their paragraphs or essays. Monitoring these processes can help L2 teachers see how L2 learners produce a piece of writing. To maintain academic integrity in L2 classrooms, they should also integrate process-based assessment techniques and redesign classes to evaluate authorship. Considering advances in AI technologies, teachers should emphasize human–AI collaboration to promote ethical and responsible L2 learning and incorporate higher-order thinking skills into L2 writing assignments to encourage deeper cognitive engagement, rather than banning learners' use of AI in the classroom or allowing AI to replace human cognition in L2 writing. Also, they should both

improve their own awareness of AI-related issues and support L2 learners' AI literacy as an emerging form of literacy. Furthermore, teachers can promote their learners' responsible, ethical, and transparent engagement with AI tools by encouraging them to acknowledge and disclose their use of AI. Finally, they should emphasize that making mistakes is a natural part of L2 learning and can contribute to their L2 development.

LIMITATIONS

As with most research, this study has limitations that should be considered when evaluating its findings and implications. Given the explanatory nature of the study and the inherent limitations of AI-generated data, the findings should be interpreted with caution, as they reflect a synthesis of data from the textual corpus comprising responses from ChatGPT and Gemini. Despite the researchers' filters and researcher-interpreted syntheses derived through a systematic thematic analysis, readers should be aware that AI outputs are probabilistic and may vary across AI models depending on the prompts and queries. Accordingly, thematic recurrence across the responses also should not be interpreted as the sole evidence of false positives. Therefore, to minimize and overcome these limitations, future studies can examine mislabeling and strategies to reduce false positives in AI use with human participants, integrating L2 learners and teachers to verify the results. Although this study employed detailed data analysis procedures, the researchers suggest integrating multiple data-collection tools and strategies in future research. Adding empirical evidence from AI detector performance data, integrating authentic L2 writing samples, or incorporating L2 teachers' perspectives from semi-structured interviews or focus group discussions might be helpful for a more comprehensive output and for possible suggestions for such cases. Thus, evaluating the findings alongside those of future studies can provide a broader understanding of the phenomenon under investigation for L2 teachers seeking to reduce false positives.

CONCLUSION

This study examined what ChatGPT and Gemini say about the possible reasons AI detection systems misidentify human-written texts as AI-written, and provided strategies that L2 teachers can use to reduce such misidentification. It emphasized the importance of tracing evidence of L2 learners' writing processes, revising writing assessments, and adopting multiple-use verification techniques to minimize the incorrect labeling of authentic L2 students' writing as AI-generated. The findings in the present study extended the existing literature by explaining potential reasons for false positives and suggesting strategies for L2 teachers to use in their L2 classes to reduce them. The findings also indicate that relying solely on AI detection systems to identify a text's author can be unreliable and inaccurate due to their inherent limitations. Therefore, L2 educators and practitioners should be cautious about the results of AI tools and detectors, integrate more human-touch elements into L2 writing tasks, adopt a process-based writing approach, and incorporate personalization into assessments to improve authorship-labelling accuracy for effective L2 instruction.

ACKNOWLEDGEMENT

In this manuscript, the authors collected data using responses from ChatGPT 5.5 and Gemini 3.6 Flash, used ChatGPT 5.5 to draw Figure 1, and used the AI-powered grammar checker Grammarly to improve grammar and spelling. The content was reviewed and edited by the authors, who take full responsibility for its accuracy. The authors gratefully acknowledge the journal editor for constructive feedback during the preliminary review process, the external expert for their role in data analysis, and the anonymous reviewers for their valuable suggestions for improving the study.

REFERENCES

Alam, M. S., Asmawi, A., Haque, M. H., Patwary, M. N., Ullah, M. M., & Fatema, S. (2024). Distinguishing between student-authored and ChatGPT-generated texts: A preliminary exploration of human evaluation techniques. *Iraqi Journal for Computer Science and Mathematics, 5*(3), 287–304. https://doi.org/10.52866/ijcsm.2024.05.03.016

Alexander, K., Savvidou, C., & Alexander, C. (2023). Who wrote this essay? Detecting AI-generated writing in second language education in higher education. *Teaching English with Technology, 23*(2), 25–43. https://doi.org/10.56297/BUKA4060/XHLD5365

Alfarraj, Y. F., & Wardat, Y. (2024). Exploring the impact of ChatGPT on scientific research: Assessing strengths, weaknesses, opportunities, and threats. *Education as Change, 28*(1), 1–27. https://doi.org/10.25159/1947-9417/16006

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus, 15*(2), e35179. https://doi.org/10.7759/cureus.35179

AlSagri, H. S., Farhat, F., Sohail, S. S., & Saudagar, A. K. J. (2025). ChatGPT or Gemini: Who makes the better scientific writing assistant? *Journal of Academic Ethics, 23*(3), 1121–1135. https://doi.org/10.1007/s10805-024-09549-0

Barone, S. M., & Cargile, C. (2020). L1 and L2 writing differences: From understanding to practice. *English Teaching Forum, 58*(3), 10–19.

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Bendo, M. C. D. (2026). False positives in AI writing detection: A small-scale empirical study using authentic Filipino student essays. *ASEAN Journal of Open and Distance Learning, 18*(1), 12–19. https://doi.org/10.64233/VYVI9613

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cheng, A., Lin, Y., Reedy, G., Joseph, C., Wirkowski, S., Mallette, V., Nagesh, V., Krieser, D., & Calhoun, A. (2025). Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content. *Advances in Simulation, 10*, Article 66. https://doi.org/10.1186/s41077-025-00396-6

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). 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

Dao, N. A. D., Nguyen, T. M., Tran, T. T., & Phan, A. (2025). Educational administrators versus teachers' perspectives on the use of AI in language learning: A qualitative study. *Computer-Assisted Language Learning Electronic Journal (CALL-EJ), 26*(4), 418–442.

Derakhshan, A. (2025). Tracking the effects of Gemini as a GenAI tool on L2 learners' writing proficiency and anxiety: Latent growth curve modeling approach. *The Asia-Pacific Education Researcher.* https://doi.org/10.1007/s40299-025-01042-5

Dincer, A. & Yeşilyurt, S. (2025). *Dealing with bias and hallucinations in AI tools*. In İ. Çakır & A. Bakla (Eds.), AI in scholarly research and publishing: Opportunities, challenges and dilemmas (pp. 75–95). Anı Publishing.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence, 6*, 100209. https://doi.org/10.1016/j.caeai.2024.100209

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine, 6*(1), 75. https://doi.org/10.1038/s41746-023-00819-6

Georgiou, G. P. (2026). Key features to distinguish between human- and AI-generated texts: Perspectives from university professors. *AI in Education, 2*(1), 2. https://doi.org/10.3390/aieduc2010002

Giray, L. (2024). The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. *The Serials Librarian, 85*(5-6), 181–189. https://doi.org/10.1080/0361526X.2024.2433256

Giray, L., Roe, J., & Diesta Espiritu, J. (2026). AI writing detectors are ineffective, unreliable, and harmful. *English Teaching: Practice & Critique,* 1–17. https://doi.org/10.1108/ETPC-07-2025-0155

Graham, S., & Sandmel, K. (2011). The process writing approach: A meta-analysis. *The Journal of Educational Research, 104*(6), 396–407. https://doi.org/10.1080/00220671.2010.488703

Grammarly. (2026, March 30). *How do AI detectors work? Methods and limitations.* Grammarly.

Griche, M. S., & Bennis, M. (2026). Generative artificial intelligence integration in foreign language education in higher education: A systematic literature review. *Technology in Language Teaching & Learning, 8*, 102827. https://doi.org/10.29140/tltl.2026.102827

Gritsai, G., Voznyuk, A., Grabovoy, A., & Chekhovich, Y. (2024). Are AI detectors good enough? A survey on quality of datasets with machine-generated texts.

Hadra, M., Cambridge, K., & Mesbah, M. (2026). Evaluating the accuracy and reliability of AI content detectors in academic contexts. *International Journal for Educational Integrity, 22*, Article 4. https://doi.org/10.1007/s40979-026-00213-1

Ibrahim, K. (2023). Using AI-based detectors to control AI-assisted plagiarism in ESL writing: "The Terminator Versus the Machines". *Language Testing in Asia, 13*(1), 46. https://doi.org/10.1186/s40468-023-00260-2

Kurt, G., & Kurt, Y. (2024). Enhancing L2 writing skills: ChatGPT as an automated feedback tool. *Journal of Information Technology Education, 23*, Article 24. https://doi.org/10.28945/5370

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Liu, Y., & Wang, H. (2024). *Who on Earth is using generative AI?* (Policy Research Working Paper No. 10870). World Bank Group Digital Development Global Practice.

Liu, Y., Huang, J., & Wang, H. (2025). *Who on Earth is using generative AI? Global trends and shifts in 2025* (Policy Research Working Paper No. 11231). World Bank Group Digital Transformation Vertical. https://doi.org/10.1596/1813-9450-11231

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

Makiev, K. G., Asimakidou, M., Vasios, I. S., Keskinis, A., Petkidis, G., Tilkeridis, K., … & Iliopoulos, E. (2023). A study on distinguishing ChatGPT-generated and human-written orthopaedic abstracts by reviewers: Decoding the discrepancies. *Cureus, 15*(11), e49166. https://doi.org/10.7759/cureus.49166

Ncube, P. D., Dzvapatsva, G. P., Matobobo, C., & Ranga, M. M. (2026). Redefining student assessment in AI-infused learning environments: a systematic review of challenges and strategies for academic integrity. *AI and Ethics, 6*(1), 68. https://doi.org/10.1007/s43681-025-00871-w

Nguyen, L., & Barrot, J. S. (2024). Detecting and assessing AI-generated and human-produced texts: The case of second language writing teachers. *Assessing Writing, 62*, 100899. https://doi.org/10.1016/j.asw.2024.100899

Onozawa, C. (2010). A study of the process writing approach: A suggestion for an eclectic writing approach. *共愛学園前橋国際大学論集, 10*, 153–163.

OpenAI. (2026, May 11). *How ChatGPT adoption broadened in early 2026*. OpenAI.

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice, 20*(2). https://doi.org/10.53761/1.20.02.07

Pramjeeth, S., & Ramgovind, P. (2025). Reconceptualizing assessment design in the age of AI: Is it genuine or faux? *New Directions for Teaching and Learning, 2025*(182), 29–40. https://doi.org/10.1002/tl.20625

Reid, E. (2026, May 19). *A new era for AI search*. Google.

Saha, S., & Feizi, S. (2025). Almost ai, almost human: The challenge of detecting ai-polished writing. *In Findings of the Association for Computational Linguistics: ACL 2025* (pp. 25414–25431). https://doi.org/10.18653/v1/2025.findings-acl.1303

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature, 613*, 620–621. https://doi.org/10.1038/d41586-023-00107-z.

Tran, T. T. T. (2025). Enhancing EFL writing revision practices: The impact of AI-and teacher-generated feedback and their sequences. *Education Sciences, 15*(2), 232. https://doi.org/10.3390/educsci15020232

Van Vlasselaer, M., Van Droogenbroeck, F., & Spruyt, B. (2026). Who wrote this? Evaluating the reliability of AI detection tools in higher education. *International Journal for Educational Integrity, 22*, Article 16. https://doi.org/10.1007/s40979-026-00226-w

Waked, A. N., Abdelsalam, H., Ashraf, M. W., Pilotti, M., & El Alaoui, K. (2024). 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, Türkiye. ISTES.

Wang, C., & Wang, Z. (2025). Investigating L2 writers' critical AI literacy in AI-assisted writing: An APSE model. *Journal of Second Language Writing, 67*, Article 101187. https://doi.org/10.1016/j.jslw.2025.101187

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

Yan, D. (2024). Comparing individual vs. collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning. *Language Learning & Technology, 28*(1), 1–19. https://doi.org/10.64152/10125/73597

Zainurrahman, Emilia, E., & Rodliyah, R. S. (2026). A genre-based approach with GenAI feedback to teaching written exposition in a tertiary EFL context in Indonesia. *TESL-EJ, 30*(1), 1–33. https://doi.org/10.55593/ej.30117a8