AI detection systems may incorrectly flag student-created writing as AI-generated, given certain conditions. Such false positives can harm a student's academic record and emotional well-being. This issue particularly affects non-native English speakers (Najarro, 2023), neurodivergent students, and ESL students, highlighting that schools should see detectors as a starting point for discussions rather than as definitive proof.

> "Research shows that English-language-learner (ESL/E-Learner) students are significantly more likely to be incorrectly flagged as having used AI-generated text."[1]

Some reasons for differences between native and non-native speakers include stylistic variations, unusual phrasing, or unconventional word choices. These factors may resemble the AI's training data, and most AI systems are primarily trained on data from native speakers, making it more likely that writing from non-native speakers will be identified as AI-generated. AI-detectors may also penalize ESL essays that show greater differences in vocabulary and grammar.

These false positives may trigger unnecessary academic investigations, which may lead to disciplinary actions, loss of credit, and reputational harm. This can be mitigated by informing ESL students about the limitations of detectors and encouraging them to keep drafts or revision histories as proof of authorship. Institutions should implement safeguards—including manual review, an appeals process which is transparent, and increased educator awareness—to reduce these risks.

The Department of Education report on the use of AI in classrooms noted that AI could bring risks along with its implementation. These risks were delineated in the section of the report labeled "*Three Reasons to Address AI in Education Now" (*Technology, 2023*)*

1. A Stanford-based study (Myers, 2023; Tian, 2023; Woelfel, 2023)

   The study tested seven commercial AI detectors using two datasets: (a) essays written by native-born U.S. students, and (b) TOEFL essays written by non-native speakers. The findings showed that non-native writers are significantly over-represented among false positives, leading to unnecessary academic integrity accusations.

   While the AI detectors performed nearly flawlessly on essays written by native speakers, 61% of the 91 TOEFL essays were incorrectly flagged as AI-generated. All seven detectors unanimously misclassified 18 of the essays (19%). At least one detector flagged 97% of the TOEFL essays as AI-generated. The dataset used was small, so the results are more likely to be challenged (Tian, 2023).

[1] Table 1

2. ResearchGate book chapter (2023) – *AI Detection's High False Positive Rates and the Psychological and Material Impacts on Students* (Hirsch, 2024)

   The study examined the problem of false accusations by conducting interviews with students focusing on neurodivergent and L2 (second-language) writers.

   L2 writers reported greater anxiety, stigma, and extra administrative burden after a false positive ID. The book also notes that "neurodivergent writers, along with L2 writers, are flagged at higher rates."

   In conclusion, a false-positive identification has a higher, measurable psychological cost for ESL learners.

3. The Markup investigative article

   "AI Detection Tools Falsely Accuse International Students of Cheating" by Tara Garcia Mathewson (Mathewson, 2023)

   This study consists of interviews with faculty and students from several U.S. universities, plus replication of the Stanford experiment.

   International/ESL students reported being singled out after detectors flagged their work.

   The article mentions the 61% false-positive rate for non-native essays and highlights real-world disciplinary cases (e.g., a student placed on academic probation after a false flag).

   Conclusion – The bias seen in lab tests translates into actual incidents on campus, harming students' records and trust.

Non-native speaking students' writing is systematically more likely to be falsely labeled as AI-generated, which can result in academic penalties, stress, and unfair treatment. Therefore, institutions should implement safeguards in a policy statement, including manual review, a transparent appeals process, and increased educator awareness, to minimize these risks. Using AI-detectors voluntarily without a policy may cause unforeseen problems. In a survey by the Center for Democracy & Technology (Prothero, 2024), 68% of instructors reported using an AI detector on student work without a policy. To prevent unexpected outcomes, establishing a policy to regulate AI-detectors is crucial.

This policy will include training for ESL instructors, which will be conducted through college-level teacher training and workshops. Topics covered would consist of why, when, and how to use AI-detection tools, including guidance on selecting the best tools. Baseline sampling, initial review scanning, and informing students in advance will also be explained.

Although a policy created and revised by those involved is probably the best solution, alternative options include using technology to reduce AI detector bias toward ESL students (*Pangram Labs AI*

Detection; Tian, 2023) and actions similar to Vanderbilt University's (Coley, 2023) decision to disable the Turnitin AI detection module campus-wide.

Another study (Jiang et al., 2024) finds that this is not an issue; the authors used a larger dataset (https://github.com/Weixin-Liang/ChatGPT-Detector-Bias/) to conclude that there is no bias in AI detectors against ESL students' essays. This analysis appears solid. However, it does not fully address the potential harm of relying solely on the AI detector to assess a student's work. As previously stated, a policy or behavioral change is necessary to resolve the issue entirely.

There is a need to change how teachers of ESL students process those students' work. AI detection tools should only be used as a first filter. Final judgment must always involve human review. AI-Detection may be best suited to highlight those items that require closer examination.

A policy surrounding AI-detector use with non-native speaking students would address the issues of

1. Whether to use such a tool
2. How to use the tool
3. What tools can be used for the best outcomes
4. Why and when tools should be used

Although one could advocate a more technical solution (see Appendix III), this begs the question of how accurate such a solution would be (if one could even procure the mostly proprietary algorithms to evaluate).

There are also ways in which teachers and students can better understand AI use and be guided to use "best practices," but these are mostly optional. The ad hoc use of these would be detrimental to students/classes that did not opt to use them.

Appendix I – AI usage

The following AI tools were used in the creation of this document

1. Grammarly (https://app.grammarly.com/)
    a. Throughout the paper, both the spelling and grammar were checked using Grammarly.
    b. There were no “exchanges”.
    c. Grammarly suggested changes to wording and corrected spelling. Changes were not always accepted.
    d. As an example, I just used Grammarly to change the capitalization in the Lumo prompts (#ii, #iii, and #iv were all lowercase)
2. Google summaries
    a. Primarily used for finding relevant case studies and sources
    b. Some summaries generated by searches were used as a basis for some of the written material
3. Lumo from Proton (https://lumo.proton.me/)
    a. Lumo was used in web search mode, and the following prompts were given:
        i. What are the consequences of having one’s writing incorrectly labelled as AI-generated
        ii. Are ESL students more likely to be incorrectly found to have used AI? Please give all sources.
        iii. What are some case studies dealing specifically with this issue
        iv. Is there a study regarding the negative impact of teachers voluntarily using AI detectors
4. EndNote 2025 (https://endnote.com/)
    a. EndNote was used to create the bibliography
    b. EndNote was used to create in-text citations

The various AI tools were used to find different sources and case studies (Google summaries and searches) to support the content I already had and the various ideas I put forth (Lumo explanations and web searches). These tools also aided in annotating this document (EndNote) and checking spelling and grammar throughout the document (Grammarly). Although this paper could have been finished without the use of any AI, using AI made it easier to locate and identify the resources and relevant case studies I could have eventually located. The error checking by Grammarly and MS Word built-in checker made writing much simpler.

In terms of annotation, EndNote 2025 was used; extra work was added to get all citations into an EndNote library and then cite them in this document. I still would use EndNote again as a citation tool.

During the writing of this paper, I have expressly looked for and found instances where the AI in use either did not understand what was being said in the document, incorrectly tried to correct it, or both.

APPENDIX II – Tables

Table 1. Studies on AI Detection and ESL Writers

Study / Report	**Key Findings on ESL / L2 Writers**	**False Positive Rate / Metric**
Center for Democracy & Technology brief (based on Stanford study) – "Disproportionate Effects of Generative AI Detectors on English Learners"	AI detectors that worked almost perfectly on native-speaker essays falsely flagged a majority of TOEFL essays.	61% of non-native essays flagged; 19% unanimously flagged; 97% flagged by at least one detector.
The Markup (Mathewson, 2023)	Replicated Stanford study. ESL writers disproportionately likely to be flagged; real disciplinary cases reported.	Same figures (~60% false positives).
Turnitin blog (internal research)	Acknowledges bias toward ELL writers. Reports <1% overall false positives when ≥20% of text is AI, but higher incidence with low-AI-content texts, especially for ESL writers.	No precise %, but confirms bias exists.

Table 2. Comparison of AI Detection Tools

Checker Name	**Website**	**Offers Humanizer Tool**	**% Likely AI (Before)**	**% Likely AI (After Humanizing)**
JustDone	https://app.justdone.ai	Yes	14%	0%
GPTZero	https://gptzero.me/	No	6% (2% AI + 4% mixed)	1%
Undetectable	https://undetectable.ai/	Yes	1%	1%
Grammarly	https://app.grammarly.com/	Yes	15%	8%
Copyleaks	https://copyleaks.com/	No	59.7% (first 10,000 chars)	11%
HumanizeAI	https://www.humanizeai.pro/	Yes	N/A	N/A
BypassGPT	https://bypassgpt.co/	Yes	N/A	N/A

Appendix III – Alternate Solutions

Technical Solutions

1. There is a need to develop a method for identifying AI-generated texts while minimizing the likelihood of false positives among ESL students. According to Jiang study (*Pangram Labs AI Detection*), "research investigating these detectors' [AI-detectors] fairness and potential bias is relatively rare…"(Jiang et al., 2024).

   Cost: The Pangram solution costs $5.00/student/year.

2. GPTZero also claims to have come up with a technical solution to the problem(Tian, 2023) and even reran the Stanford study's algorithm on its' new product, with a much better result.

   Cost: Educators can receive this product for free for life.

Assignment-design toolkits that are already published and used

- MIT Sloan "Designing AI-Resistant Assignments" ("AI Detectors Don't Work. Here's What to Do Instead.," 2024) – This provides templates to create prompts which require personal reflection, local data, or iterative drafts – making pure AI-generation impractical
- Carnegie Mellon "Generative AI FAQ for Instructors" (University, 2024) - Offers a decision tree: when to use a detector, when to rely on draft histories, and how to communicate expectations to students

Bibliography

1 AI Detectors Don't Work. Here's What to Do Instead. (2024). *MIT Sloan Teaching & Learning Technologies*.

2 Coley, M. (2023). *Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector*. Retrieved 09/09 from

3 Hirsch, A. (2024). AI detectors: An ethical minefield - Center for Innovative Teaching and Learning.

4 Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? *Computers & Education*, *217*, 105070. https://doi.org/10.1016/j.compedu.2024.105070

5 justdone. (2025). *Justdone*.

6 Mathewson, T. G. (2023, 2023–08–14). *AI Detection Tools Falsely Accuse International Students of Cheating – The Markup*.

7 Myers, A. (2023). *AI-Detectors Biased Against Non-Native English Writers | Stanford HAI*. Stanford University. Retrieved 09/09 from

8 Najarro, I. (2023, 2023–10–06T20:31:40.743). What Teachers Should Know Before Using AI With English Learners. *Education Week*.

9 *Pangram Labs AI Detection*.

10 Prothero, A. (2024). More Teachers Are Using AI-Detection Tools. Here's Why That Might Be a Problem. *Education Week*.

11 Technology, U. S. O. o. E. (2023). *Artificial Intelligence and the future of teaching and learning : insights and recommendations*. U.S. Department of Education, Office of Educational Technology.

12 Tian, E. (2023, 2023–10–25T01:43:00.000Z). *ESL Bias in AI Detection is an Outdated Narrative*. Retrieved 09/09 from

13 University, C. M. (2024). *Generative AI Tools FAQ - Eberly Center - Carnegie Mellon University*.

14 Woelfel, K. (2023). Brief – Late Applications: Disproportionate Effects of Generative AI-Detectors on English Learners. *Center for Democracy and Technology*.