

Exploring Asymmetric Academic Integrity Signals in Search- and LLM-Assisted Learning

Rahul R. Divekar

Bentley University, Waltham MA, USA
rdivekar@bentley.edu

Abstract

In a within-subject study, students learned via search and LLMs, then wrote essays without them. Analyses show no detectable difference across stylometric and perplexity features; AI detection probability is tool-dependent; plagiarism overlap is higher after search. Results suggest asymmetric detection signals across learning pathways, with implications for interpreting academic integrity tools.

1 Introduction

While a large concern of educators is that students increasingly use large language models (LLMs) to generate text submitted as their own, many students also use them as learning tools and alternatives to traditional web search for studying and understanding dense and layered topics (Divekar et al., 2025a). We explore whether the choice of learning tool (LLM vs. web search) leaves detectable traces in student-written essays, even when the final text is human-authored without tool assistance in the writing stage.

Why differential detectable traces across learning tools matter becomes apparent when considering how student submissions are evaluated. Graders may see a dashboard of all student submissions with the percentage of overlap with existing sources or AI-generated text probability. It is reasonable to assume that submissions with a higher probability or overlap percentage receive a closer or secondary look from the grader thereby increasing their chances of scrutiny. When the chances of scrutiny for a student increase, the chances of findings of academic integrity violation may also similarly increase. However, if detectable traces are found differently across two sources of learning, we hypothesize that systemic differences in violation findings may manifest attributed to the choice of learning tool. In this paper, we explore the first step: whether there are detectable traces of

tool source in student-generated work and whether some popular online detectors vary in finding those traces.

In an in-person moderated preliminary study, 20 university students learned about topics of general interest using two tools (conditions): a web search engine and an LLM, then wrote an essay demonstrating their understanding. We analyzed the resulting essays using stylometric word and sentence level features (e.g., length, lexical diversity) and perplexity-based measures alongside two commercial AI and plagiarism detection tools. We find that the stylometric and perplexity scores across both conditions do not show reliable or systematic differences, the AI detection scores are tool dependent, and the plagiarism scores show statistically significant differences across both external tools.

These preliminary findings carry implications for academic integrity practices: In a landscape where practitioners may be using multiple tools to detect plagiarism and GenAI-generated text (Balalle and Pannilage, 2025), students who use web search for learning may be more likely to produce text flagged for academic integrity review due to overlap with online sources, while students who learn via LLMs may produce fewer overlap-based detection signals, despite both groups writing their own essays; all the while AI detection scores remain inconsistent in detecting the LLM-assisted learning pathway. If so, the implication to the broad student community would be that students who use Search to learn could be at a higher risk of being flagged for review as compared to students who use LLMs, as overlap-based detection signals appear more consistent across tools than AI-detection signals in this sample.

2 Background and Literature

Some practitioner educators are understandably worried about the large-scale availability of Gen-

erative AI (GenAI) applications such as ChatGPT and whether students use it to generate essays, answers, and other submitted material to the effect of students using GenAI as a *homework machine* that allows large-scale circumvention of the learning process (Mamo et al., 2024; Kiryakova and Angelova, 2023; Divekar et al., 2025c). Other practitioners are seeing GenAI as a positive as it can easily summarize, explain, expand on topics in a conversational way that amounts to a personalized scaffolded learning experience where the student is given bite-sized information in a conversational manner by the LLM leading to an arguably better learning experience (Balalle and Pannilage, 2025; Divekar et al., 2025c).

In the absence of a general consensus on the role of GenAI in education (Balalle and Pannilage, 2025), practitioners set the rules for usage of external tool help for their classes. Foltynnek et al. (2023) term the violations of these rules as Unauthorized Content Generation (UCG): students using unauthorized tools to produce academic work for credit, progression, or award.

We discuss two broadly-available technologies that may enable UCG and their detection: Web search and LLMs, and AI and Plagiarism detectors (Hao, 2026). In the past few decades, web search tools have driven plagiarism where students copy-paste text snippets with little paraphrasing from online sources accessed via search. Several practitioner-oriented tools help detect this UCG source (e.g., Turnitin, Grammarly) by finding textual overlap between student-submitted text and online sources including academic publications among others.

Productized LLMs are a more recent enabler of UCG and present a challenge for the community wanting to detect LLM-assisted UCG. Hao (2026) summarize core paradigms like supervised learning classifiers trained on textual features, end-to-end models, watermarking (designing LLMs that add subtle patterns that can be detected in downstream tasks to ensure provenance), identifying the writing process through document history and other metadata, and similarity matching of texts. Hao (2026) identify that many of these approaches have been operationalized and marketed as online AI-detection tools to practitioners (e.g., GPTZero¹, Originality.ai, Copyleaks², Grammarly

³, Turnitin⁴). Yet, without constrained settings, AI generated writing is difficult to differentiate from human writing (Casal and Kessler, 2023; Fleckenstein et al., 2024; Wu et al., 2023; Pudasaini et al., 2025; Sadasivan et al., 2023; Tufts et al., 2025; Weber-Wulff et al., 2023; Lepp and Smith, 2025; Zu et al., 2025), a challenge significant enough to spur the creation of a dedicated sub-field of Natural Language Processing (Alam et al., 2025).

However, the lack of validity and reliability of off-the-shelf AI-text detection mechanisms has not stopped some practitioners from using the tools that claim to detect GenAI-written text (Gaines, 2025). Given that UCG can stem from AI and search, practitioners draw on the wide availability of plagiarism and AI-text detectors and often deploy both types of tools to uphold academic integrity (Balalle and Pannilage, 2025). While prior published accounts have found that using GenAI detectors can create academic harm to students because of the false positives (Liang et al., 2023; Gaines, 2025), we extend that body of research by showing that when a range of tool types are used in conjunction e.g., plagiarism checkers and GenAI detectors, a different issue may arise. The asymmetry in the consistency between plagiarism detectors and GenAI detectors can increase the likelihood of flagging potential UCG for students who used Search for learning as compared to students who used GenAI for learning; even when either tool is only used for learning and not direct writing. This may occur as the student produced text is influenced by web sources from the Search based learning activity leaving detectable traces of verbatim/near-verbatim overlap with indexed online sources that the plagiarism checkers are likely to find. However, the likelihood of detecting similar GenAI influences in learning can be inconsistent. We qualify this asymmetry in detail in the findings section.

This asymmetry has practical implications for academic integrity. In current educational practice, instructors may rely on similar commercial plagiarism and AI detectors to flag potential violations in conjunction as many commercial platforms provide both features (Balalle and Pannilage, 2025). Our preliminary findings suggest that the conjunctive usage of such tools may introduce a risk of asymmetry: students who use Search, often considered the more traditional and *safer* approach, may face

¹Gptzero.me

²copyleaks.com

³grammarly.com/ai-detector

⁴turnitin.com/solutions/topics/ai-writing

a higher risk of being flagged for UCG because the tools developed to detect it may be resting on more mature development (Foltýnek et al., 2019), while students who learn using LLMs may produce text that does not carry a strong detection signal of either plagiarism or AI generation consistently across tools, despite both groups writing essays by themselves.

We present our exploratory findings through this short paper to the AIME community to discuss potential implications of deploying a suite of off-the-shelf detection tools in everyday educational settings, and to highlight the need for detection frameworks that account for the increasingly diverse ways students access information.

This paper makes three contributions. First, we introduce the notion of post-learning detection asymmetry, showing that the choice of learning tool (search vs LLM) leaves different levels of detectable traces in academic-integrity signals observed in human-authored essays. Second, we highlight that the external academic integrity tools, especially when it comes to GenAI detection, can be inconsistent with each other. Third, we show that while detection tools in some conditions show differences, stylometric features show no reliable differences between the two conditions still making it potentially difficult for practitioners to identify tool influences by stylistic observation alone. Overall, we present an asymmetry in which the conjunctive use of GenAI and plagiarism detection tools, in the absence of stylometric differences, could lead to differential flagging risk for student-generated content.

3 Methodology

We present results from a preliminary study where $N=20$ participants (sample demographics seen in Table 1) were given a task to learn about a topic using Google (search engine) and ChatGPT (LLM product); leading to 40 essays that we analyze. We used off-the-shelf products because despite custom learning applications seen in research, students widely use these products as their go-to in typical educational situations (Divekar et al., 2025a). Product versions used did not contain GenAI summaries in search condition or hyperlinked references in GenAI conditions keeping two core product features isolated. In a similar setup to (Kosmyna et al., 2025; Melumad and Yun, 2025), a counterbalanced within-subject design was used. We conduct a fo-

cused analysis of texts from data collected in (Divekar et al., 2025b,c) where participants from a university were asked to learn two topics of general interest (“how does a powergrid work?” and “how does the internet work”) in counterbalanced order, such that half received each topic first. The tool assigned to the first topic (search or LLM) was randomized, with the alternate tool used for the second topic. They were allowed to take notes and eventually were asked to write an essay to demonstrate their understanding of the topic without either tool support. Access to notes was permitted during the writing phase to approximate academic contexts like open-note assessments where students are allowed a self-created and limited sheet to refer to while producing independent work. All stages were time-boxed sessions separating learning from writing. In-person moderated conditions allowed us to tightly control for tool and note usage yielding valid data. The study was approved by the university IRB.

To characterize the essay writing style, we computed features spanning two linguistic dimensions. Lexical richness was measured using established measures like Corrected Type-Token Ratio (CTTR) and long word ratio (fraction of words greater than 6 characters). Similarly, sentence-level structure was captured through mean sentence length and mean word length, both established stylometric features. To assess whether a certain text exhibited statistical properties associated with LLM-generated text, past research has largely used perplexity and burstiness (Gehrmann et al., 2019; Xu and Sheng, 2024) scores computed using GPT-2 (Hao, 2026). To avoid multi-collinearity, we use perplexity alone. Additionally, to connect our findings with practitioner-oriented implications, we test the essays with two popular online tools that claim to detect plagiarism and AI-generated text: Grammarly⁵ and GPTZero⁶. Both tools provide a plagiarism percentage (overlap with existing online sources) and AI probability percentage.

We use Wilcoxon signed-rank tests to assess whether each feature differs between the two conditions. The features fall into five pre-specified construct families: lexical richness, sentence-level structure, perplexity, AI-detection scores, and plagiarism scores; we apply a Bonferroni correction within each family and report both raw and adjusted

⁵grammarly.com

⁶gptzero.me

Description	Number
Age Range	
18-20	10
21-23	7
24-26	3
Gender	
Female	11
Male	9

Table 1: Participant Demographics

p-values.

4 Results

Table 2 demonstrates first 50 truncated words of the essays produced by Participant (P2) to illustrate the textual features and plagiarism and AI detection scores received by off-the-shelf tools.

Table 3 shows the results from Wilcoxon signed-rank tests. Statistically significant differences across two conditions were observed for AI probability score by GPTZero and plagiarism scores by both GPTZero and Grammarly.

In a follow-up test using a two-way mixed, absolute agreement Intraclass Correlation Coefficient (ICC) showed borderline-poor inter-tool agreement for AI detection percentages ($ICC = .50, p < .001$) and good inter-tool agreement for plagiarism percentages ($ICC = .758, p < .001$) across the two tools.

5 Discussion

We find plagiarism scores are consistent between two detectors and tests show a statistically significant difference in percent of text identified as plagiarized between the two conditions for both detectors. This may reflect that plagiarism detection based on indexing online sources and finding textual overlap is a longer-established measurement paradigm than AI-text detection (Foltýnek et al., 2019). The fact that the Search tool and its web sources were not available during the writing stage yet overlap in the final essay is found with online sources points to a plausible explanation that source text is transferred either from memory or via paraphrased/typed/copied notes into the final essay. However, neither pathways is uncommon in learning especially when many scenarios allow self-created notes during testing.

Even if the text transfers from the source to the final essay when the source is not directly available,

Wilcoxon signed-rank tests do not show differences in the style of writing after learning with either tool in paired tests. That is, despite potentially borrowed language, on the surface, the two essays seem stylometrically similar. The absence of a stylometric difference is unsurprising: the overlapping text is a small fraction of each essay, so aggregate style possibly remains dominated by the participant’s own prose even when localized overlap may differ across conditions. Future work could examine in depth how and what content transfers from source to essay, which might reveal the influence of a source even in its absence and how students integrate it into their conceptual framework without altering their writing style. Another avenue of future research is to more deeply examine writing beyond surface stylometrics used in this paper including automatic and human evaluations.

While essays seem similar and plagiarism scores differ between conditions, AI generated scores show inconsistency between tools: GPTZero’s AI probability shows a statistically significant difference in the percent of text it would flag as AI generated whereas Grammarly’s AI probability score does not show it. Combined with a low inter-tool agreement, results indicate that the two tools differ in their sensitivity to this signal. Prior research by (Choi and Zu, 2025) shows that AI detectors are largely opaque with their output often presenting a probability of the text being AI generated and nothing more. Further, their reliability may depend on whether the model used for generating text is the same as the model used to detect AI-generated text among other factors. However, practitioners who are not linguists or GenAI experts are not expected to have this level of technical information to discern how to interpret this opaque output or which tool to choose or trust. Therefore, the two tools diverging from each other may cause confusion, additional load for practitioners, or unequal outcomes for students.

We contextualize the magnitude of the observed differences. Plagiarism detection tools typically report some degree of textual overlap in most student writing. In the present sample, mean plagiarism scores in the search condition ranged from 12.65% to 22.20% across tools, whereas the LLM condition ranged from 1.90% to 8.80%. Whether either range would prompt closer review depends on institutional policy, as there is no standardized threshold for acceptable overlap across higher education. However, the separation between these

	LLM-assisted learning	Search-assisted learning
Essay Text	Title: Understanding the Power Grid: Structure and Functionality Introduction A power grid, technically known as a electrical grid, is a interconnected network to deliver electricity from the source to the consumer. First, the electricity is generated, at the stations, which produce electricity, and then high-voltage lines are used to carry...	What is the internet? When we talk about the internet, we are basically talking about how two or more devices are connected. These devices are connected together by a device called router, which is the central hub. A router is just a specialized computer whose job is to manage the connectivity..
Grammarly Plagiarism Score	6%	39%
GPTZero Plagiarism Score	24%	49%
Grammarly AI Detection score	17%	0%
GPTZero AI Detection score	1%	0%

Table 2: Excerpt: Truncated verbatim first 50 words of essay produced by participant-2 across two conditions along with plagiarism and AI detection scores.

Measure	LLM Mean	Search Mean	Z	p	Adjusted p
CTTR	6.49	6.14	-1.120 ^a	.263	.526
Long-word Ratio	0.31	0.30	-1.568 ^a	.117	.234
Mean Sentence Length	20.95	18.79	-1.811 ^a	.070	.140
Mean Word Length	5.27	5.11	-1.755 ^a	.079	.158
Document Perplexity	28.56	28.85	-0.187 ^a	.852	.852
GPTZero AI Probability	29.60	8.65	-2.341 ^a	.019	.038*
Grammarly AI Score	16.35	3.25	-1.859 ^a	.063	.126
GPTZero Plagiarism Score	8.80	22.20	-2.294 ^b	.022	.044*
Grammarly Plagiarism Score	1.90	12.65	-2.748 ^b	.006	.012*

Table 3: Within-participant comparisons between LLM- and Search-assisted learning conditions using Wilcoxon signed-rank tests. ^aBased on positive ranks. ^bBased on negative ranks. *Adjusted p significant at <0.05

ranges suggests that essays produced under the two learning conditions could plausibly fall on opposite sides of a review threshold, depending on the policy. AI probability scores showed lower inter-tool agreement and more tool-dependent behavior, with ranges of 16.35–29.60% for the LLM condition and 3.25–8.65% for the search condition. Compared to plagiarism detection, AI detection appears less standardized, and it remains unclear whether widely adopted thresholds or interpretive norms exist for these scores in practice.

Taken together, these patterns point to an asymmetry in how detection tools may respond to essays produced after different learning tools, even when the essays themselves are stylometrically similar and written without direct tool assistance. Plagiarism detectors, which rest on a more mature foundation (Foltýnek et al., 2019), show stronger inter-tool

agreement and appear more likely to consistently flag essays written after search-assisted learning. AI detectors, which are newer and show weaker inter-tool agreement, appear more likely to flag essays written after learning with LLMs with poorer consistency. If practitioners rely on a combination of both types of detectors, as is now a common practice (Balalle and Pannilage, 2025), these patterns suggest that the combination of the choice of learning tool and academic integrity tool, rather than any direct tool usage during the writing stage, could influence which students are flagged for further review; a hypothesis for further investigation.

The findings suggest implications for assessment design, delivery, and scoring. The use of student-written artifacts as evidence of knowledge is undergoing fundamental change as educators determine whether and how GenAI should be incorporated

into learning and evaluated work. When independent authorship is essential, educators should consider moderated or otherwise process-observed assessments that establish how the submitted work was produced rather than relying on post hoc plagiarism or AI-detection scores. When such assessment conditions are not feasible, detector outputs should not determine which students receive additional scrutiny, because even preliminary or uncertain flags may reproduce the learning-pathway asymmetry identified in this study. Further, user interfaces of assessment dashboards that display textual-overlap or AI-probability scores could communicate the uncertainty associated with these measures, discourage ranking submissions by detector scores, or entirely remove aggregate scores and instead support the examination of individual submissions. The continuous emergence of new GenAI tools for learning or producing learning artifacts presents an evolving challenge that requires coordinated attention to both the accuracy of detection algorithms, the interfaces through which their outputs are interpreted, and increasing awareness within the community that uses these tools.

6 Conclusion

While prior findings have determined that LLM-generated text cannot be reliably distinguished from human-generated text, our analysis extends this finding to show that human-written text *after* learning with different tools (LLMs vs. Search) also cannot be reliably distinguished stylometrically. AI detection scores showed mixed and tool-dependent results and did not consistently distinguish the conditions. However, both tested plagiarism detection tools showed statistically significant differences, with higher overlap for essays written after search-assisted learning.

The two types of tools (AI and plagiarism detection) differ in their maturity stage. However, practitioners may use these tools concurrently to uphold academic integrity in everyday settings. In this larger landscape, where plagiarism detectors are arguably more reliable than AI detectors (Foltýnek et al., 2019; Weber-Wulff et al., 2023), our preliminary findings highlight an emerging asymmetry: the mixed usage of these tools may increase integrity flagging risk for search-assisted learning, rather than LLMs, even when neither is directly used for writing. These findings highlight the need to examine potential unintended effects in tool

ecosystems, in addition to isolated tool evaluations.

Limitations

The in-person moderated data collection allows for more experimental control yet reduces the sample size due to logistical challenges and costs associated with in-person research. The null results for stylometric analysis are therefore with caveats. Further, given the lack of literature on detector popularity, our selection of commercial tools was limited and based on search prominence. Therefore, future research should conduct analysis of the scope and popularity of tools and replicate with larger samples and additional tools and topics.

Differential note-taking or memorization behavior across conditions could explain lexical overlap in the Search condition. Furthermore, the task did not instruct students to cite sources. Even if sources are cited, many tools still calculate and display the absolute overlap and leave the process of matching it to cited or found sources to human judgment. The overlap detected in the search condition therefore does not by itself indicate a rule violation — only a higher chance of being flagged for closer review. In many settings, this flag could prompt a follow-up human review in which the overlapping passages and citations are evaluated. Our research finds that essays written after learning with one tool may be at a higher chance of being flagged while leaving subsequent academic integrity adjudication up to policy interpretation.

Conditions tested in this experiment approximate some academic contexts such as open-note assessments; future research should examine other learning pathways, assessment formats, and note-use conditions.

This research is not meant to compare detector performance, as the ground truth for plagiarism and AI-generated text is not compared with the detector output; this paper should not be considered as a criticism or endorsement of the commercial detectors investigated.

Acknowledgments

The author would like to thank student contributors: Sophia Guerra, Lisette Gonzalez, Natasha Boos for their assistance with data collection. The author would like to thank the Faculty Affairs Committee and Valente Research Center at Bentley University for funding this project.

References

- Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, and George Mikros. 2025. Proceedings of the 1st workshop on genai content detection (genaidetect). In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*.
- Himendra Balalle and Sachini Pannilage. 2025. Re-assessing academic integrity in the age of ai: A systematic literature review on ai and academic integrity. *Social Sciences & Humanities Open*, 11:101299.
- J Elliott Casal and Matt Kessler. 2023. Can linguists distinguish between chatgpt/ai and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, 2(3):100068.
- Ikkyu Choi and Jiyun Zu. 2025. Exploring the interpretability of ai-generated response detection with probing. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 99–106.
- Rahul R Divekar, Lisette Gonzalez, Sophia Guerra, Natasha Boos, and Ritika Divekar. 2025a. Can generative ai replace search engines for learning? understanding student preferences, use and perceived proficiency in using ai. *TechTrends*, 69(5):1040–1055.
- Rahul R Divekar, Sophia Guerra, Lisette Gonzalez, and Natasha Boos. 2025b. Exploring the interaction-outcome paradox: Seemingly richer and more self-aware interactions with llms may not yet lead to better learning. *arXiv preprint arXiv:2511.09458*.
- Rahul R Divekar, Sophia Guerra, Lisette Gonzalez, Natasha Boos, and Yalun Zhou. 2025c. Exploring undercurrents of learning tensions in an llm-enhanced landscape: A student-centered qualitative perspective on llm vs search. In *International Conference on Artificial Intelligence in Education*, pages 380–387. Springer.
- Johanna Fleckenstein, Jennifer Meyer, Thorben Jansen, Stefan D Keller, Olaf Köller, and Jens Möller. 2024. Do teachers spot ai? evaluating the detectability of ai-generated texts among student essays. *Computers and Education: Artificial Intelligence*, 6:100209.
- Tomas Foltýnek, Sonja Bjelobaba, Irene Glendinning, Zeenath Reza Khan, Rita Santos, Pegi Pavletic, and Július Kravjar. 2023. [ENAI Recommendations on the ethical use of Artificial Intelligence in Education](#). *International Journal for Educational Integrity*, 19(1):12.
- Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2019. Academic plagiarism detection: a systematic literature review. *ACM computing surveys (CSUR)*, 52(6):1–42.
- Lee Gaines. 2025. [Teachers are using software to see if students used ai. what happens when it’s wrong?](#) NPR. Audio/report for NPR News.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations*, pages 111–116.
- Jiangang Hao. 2026. Detecting ai-generated essays in writing assessment: Responsible use and generalizability across llms. *arXiv preprint arXiv:2603.02353*.
- Gabriela Kiryakova and Nadezhda Angelova. 2023. Chatgpt—a challenging tool for the university professors in their teaching practice. *Education Sciences*, 13(10):1056.
- Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4.
- Haley Lepp and Daniel Scott Smith. 2025. “you cannot sound like gpt”: Signs of language discrimination and resistance in computer science publishing. In *Proceedings of the 2025 ACM conference on fairness, accountability, and transparency*, pages 3162–3181.
- Weixin Liang, Mert Yuksekogonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7).
- Yoseph Mamo, Helen Crompton, Diane Burke, and Christine Nickel. 2024. Higher education faculty perceptions of chatgpt and the influencing factors: A sentiment analysis of x. *TechTrends*, 68(3):520–534.
- Shiri Melumad and Jin Ho Yun. 2025. Experimental evidence of the effects of large language models versus web search on depth of learning. *PNAS nexus*, 4(10):pgaf316.
- Shushanta Pudasaini, Luis Miralles, David Lillis, and Marisa Llorens Salvador. 2025. Benchmarking ai text detection: Assessing detectors against new datasets, evasion tactics, and enhanced llms. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pages 68–77.
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.
- Brian Tufts, Xuandong Zhao, and Lei Li. 2025. A practical examination of ai-generated text detectors for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4824–4841.
- Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Oluhide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1):1–39.

- Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F Wong, and Lidia S Chao. 2023. A survey on llm-generated text detection: Necessity, methods, and future directions. *arXiv preprint arXiv:2310.14724*.
- Zhenyu Xu and Victor S Sheng. 2024. Detecting ai-generated code assignments using perplexity of large language models. In *Proceedings of the aaai conference on artificial intelligence*, volume 38, pages 23155–23162.
- Ji Yun Zu, Michael Fauss, and Chen Li. 2025. Effects of generation model on detecting ai-generated essays in a writing test. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 92–98.