

# **Academic Integrity in the AI Era: Assessing Turnitin's AI Detector**

Milong Zhao

University of California Davis

UWP 001V: Introduction to Academic Literacies

## Introduction

Large Language Model-based Artificial Intelligence (LLM-based AI) refers to artificial intelligence systems built on large-scale neural networks trained to process and generate human-like text. On November 30, 2022, ChatGPT became available to the public, triggering a major shift in both competition among technology companies and the future direction of AI development. It sparked a new wave of LLMs and significantly changed people's lives. AI models are now being used in healthcare to improve medical imaging analysis, in workplace writing to assist with reports and emails, and in investments to help beginners make better financial decisions through algorithmic insights. On January 27, 2025, news broke about the Chinese AI startup's powerful AI model, Deepseek—R1, which led to a sharp decline in technology stocks, including a record-breaking loss for Nvidia, marking one of the biggest one-day selloffs in U.S. stock market history. Within the past two years, AI and the development of it is now woven into our everyday life.

The rapid rise of AI has also brought concerns. When ChatGPT first came out, universities strictly prohibited its use in most assignments, and it remains banned in most college applications. Turnitin, the biggest plagiarism detection software, quickly rolled out an AI detection tool, which many universities now rely on. However, since LLMs are trained on datasets, are they truly ready to function fairly in a diverse academic environment like UC Davis, where international students make up a large part of the student body? Can AI detection tools accurately evaluate ESL (English as a second language) writers' writing? At the same time, while AI is transforming industries, is it reasonable for universities to ban its use in school? This paper explores whether AI detection is reliable and whether banning AI use on campus is reasonable given the technology's current capabilities.

## Literature Review

Turnitin's AI detection works by breaking text into smaller segments of several hundred words, as AI-generated writing tends to follow patterns across paragraphs rather than isolated sentences (Turnitin, 2024). Turnitin runs each segment of sentences separately through a machine learning model that would assign a probability score to each, ranging from zero, very likely written by a human, to one, very likely AI-generated (Turnitin, 2024). Eventually, after evaluating all segments, an overall probability score will be provided. The model would evaluate four characteristics of writings: perplexity, burstiness, repetitiveness, and overly generic or detached writings, to determine the probability that it was generated by AI. (Turnitin, 2024).

Perplexity measures how predictable a sentence is. The idea is that human writing tends to be less predictable, with more variation in word choice and sentence structures, while AI generated texts are consistent through the writing. Burstiness, which refers to the variation in sentence length and structure. This is useful based on the previous research that human writers naturally mix up short and long sentences, while AI tends to produce more uniform sentences.

Turnitin also take into consideration of repetitiveness and overuse of certain phrases. As AI-generated text often relies heavily on common transition words like "additionally," "moreover," and "therefore" (Turnitin, 2024). Lastly, their AI detection system looks for signs of overly generic or detached writing, as AI models struggle to include personal experiences, specific details, or unique perspectives (Turnitin, 2024). According to them, Turnitin AI detector's false detection rate is very low (Turnitin, 2023).

Though this system seems intuitive and efficient, research showed that it did not turn out to be as reliable as the Turnitin team had claimed. Researchers at Stanford University tested how reliable these detectors are when flagging work from ESL writers. A team from Stanford's Department of Computer Science used 91 TOEFL (Test of English as a Foreign Language) essays (Liang et al., 2023). TOEFL, provided by Educational Testing Service (ETS), is a standardized English proficiency test accepted by universities across the world for admission purposes. ETS, at the same time, is the world's largest educational research and assessment organization, also the provider of the Graduate Record Examinations (GRE). The researchers also collected 88 U.S. 8th-grade essays from the Hewlett Foundation's Automated Student Assessment Prize (ASAP) dataset to compare results (Liang et al., 2023). They used seven AI detectors from different company, all of them correctly classified almost all native English essays, but over 61.22% of non-native TOEFL essays were misclassified as AI-generated, with 97.80% flagged at least once (Liang et al., 2023). Researchers believe this happens because of low perplexity scores—non-native writers tend to use simpler vocabulary and sentence structures, which AI detectors associate with machine-generated text (Liang et al., 2023). The high false positive rate in these exam essays just proves how unreliable AI detectors are and how they discriminate against non-native speakers, making them an unfair tool in academic settings (Liang et al., 2023). There are also other research papers that have shown similar results; however, the majority of them was conducted in 2023. Due to updates in detection models and ChatGPT, these results are not discussed here because of their timing.

Aside from the empirical research, personal experiences that graduate students had also contradicted with Turnitin's claim of the reliability of their AI detector. On the Reddit subreddit r/GradSchool, multiple graduate students have complained that paper they have written are being flagged as AI-generated content. Many posts ask for advice on what to do and express concerns about the potential impact on their academic standing. One interesting pattern that has emerged is that students accused of academic dishonesty have run their professors' papers through Turnitin's AI detector, most of them received an AI probability score of around 30% (Reddit, 2023). To further test the reliability of the Turnitin's AI detector, one graduate student ran parts of the novel *A Tale of Two Cities* by Charles Dickens through it, and the result came back as 70% AI-generated (Reddit, 2023). Unless ChatGPT somehow traveled back in time and assisted Charles in the 1800s, the AI written probability percentages should be 0%. The research results from Stanford University and the number of AI detector complains discussion on Reddit both highlight a concern with AI detectors, not just Turnitin's product, but all detectors on the market.

It's possible that Stanford's results lack generalizability due to the sample size they used. It's also possible that with updates, AI detectors are doing a better job now than before. This paper looks at AI detectors as they are right now, comparing data to see if they've improved or if they're still biased and inaccurate.

## Method

To test the reliability of Turnitin's AI detector in different situations, I bought a Turnitin instructor account through Taobao, a Chinese online shopping platform. I set up four different groups of writing for the AI detector to evaluate. By reviewing its performance in each group, we can get a sense of how well it works. AI probability of 20% is considered as a baseline, as Turnitin is unable to provide a probability score for writing that scored under 20%. Still, if a writing is said to have zero AI writing probability, then it would have an AI score of zero.

### Group 1: GPT-Generated Content

I used ChatGPT-4o to generate writing on five different topics. The prompt used was:

*"Generate a semi-essay about how [Topic] works. Make sure to use a human researcher's tone."*

The term semi-essay was used to avoid generating incorrect citations. The topics covered were:

- Two General Science Concepts
  - DNA replication
  - Natural selection
- Two Specific/In-Depth Science Concepts
  - Selective serotonin reuptake inhibitors (SSRIs)
  - fMRI
- One Specific/In-Depth Art Concept
  - Composition of techno music

The purpose of using different concept levels is to evaluate ChatGPT's performance at generating human-like text across varying levels of complexity and generality of topics. After the initial generation, I asked ChatGPT to expand on each topic, adding more depth and detail to assess whether increased depth and length of the topic would influence AI detection. Finally, I used another prompt to humanize the text:

*"Humanize the uploaded research article. Remove all AI-generated content tone."*

I then submitted three versions of the generated content—semi-essay, extended semi-essay, and humanized extended semi-essay—to Turnitin and recorded the AI probability scores for each writing.

### Group 2: ESL Writings

This group included writing from non-native English speakers. I pulled two college

application essays I wrote, plus a short story from high school. I also added two TOEFL essays from my TOEFL teacher. The goal was to see how well the AI detector handles ESL writing.

### Group 3: Grammar Improvement via ChatGPT

This group tested if fixing full human-written writings' grammar with ChatGPT affects the AI detector's score. I included two case studies I wrote, one writing about educational history, one lab report, and two essays. The prompt used was:

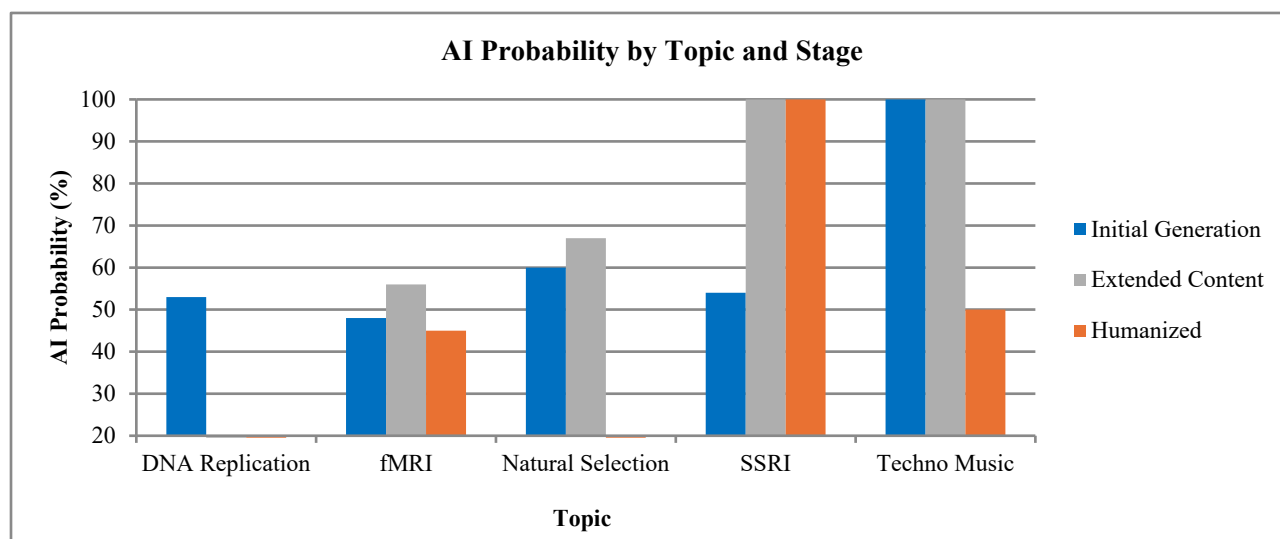
*"Improve the grammar of the uploaded writing. Make sure to keep my writing style."*

### Group 4: Published Papers

To test the AI detector's performance on highly technical writing, I collected three research papers written by UC Davis professors before ChatGPT was available. This group helps check if the detector would mistakenly flag complex academic writing as AI-generated.

## Results

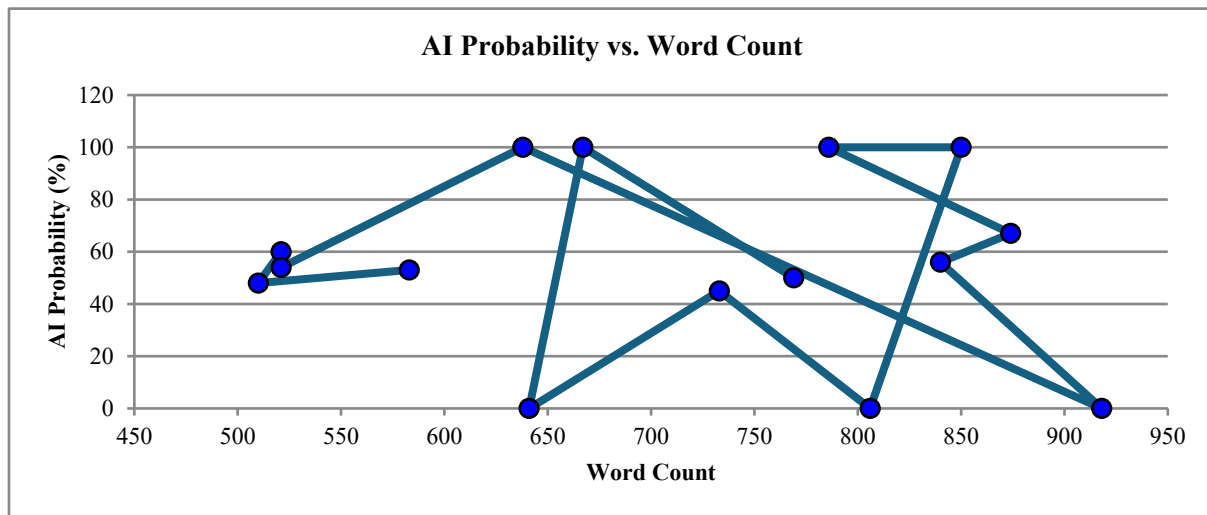
### Group 1: GPT-Generated Content



For the topic of Natural Selection, the AI probability dropped significantly after humanization through ChatGPT. Similarly, for DNA Replication, the AI probability decreased to below 20% after extending the length of the text. This pattern might indicate that general science topics, which are widely discussed, provide the model with extensive training data, allowing it to generate more human-like text naturally.

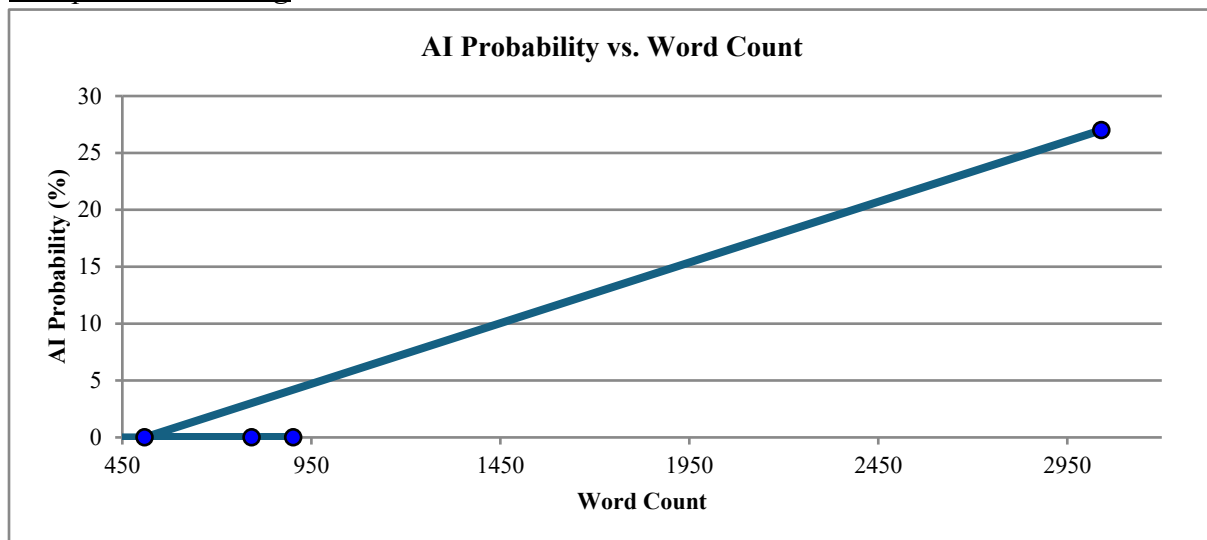
In contrast, more specialized scientific topics such as fMRI and SSRI exhibited different trends. The AI probability scores for fMRI remained relatively consistent, ranging from 45% to 56%, indicating that neither extending the content nor humanization significantly altered the AI-

generated characteristics. For SSRI, the AI probability remained consistently high, particularly for the extended content and humanized versions. A similar pattern was observed in the case of Techno Music. Both the initial generation and the extended content scored 100% in AI probability, while the humanized version dropped to 50%. This drastic shift suggests that ChatGPT's ability to humanize content varies by topic. One possible explanation is that professional topics, such as fMRI and SSRI, or niche topics, such as the composition of techno music, have less general discussion data available compared to broadly discussed scientific concepts, limiting the model's ability to adapt its language to sound more human-like.



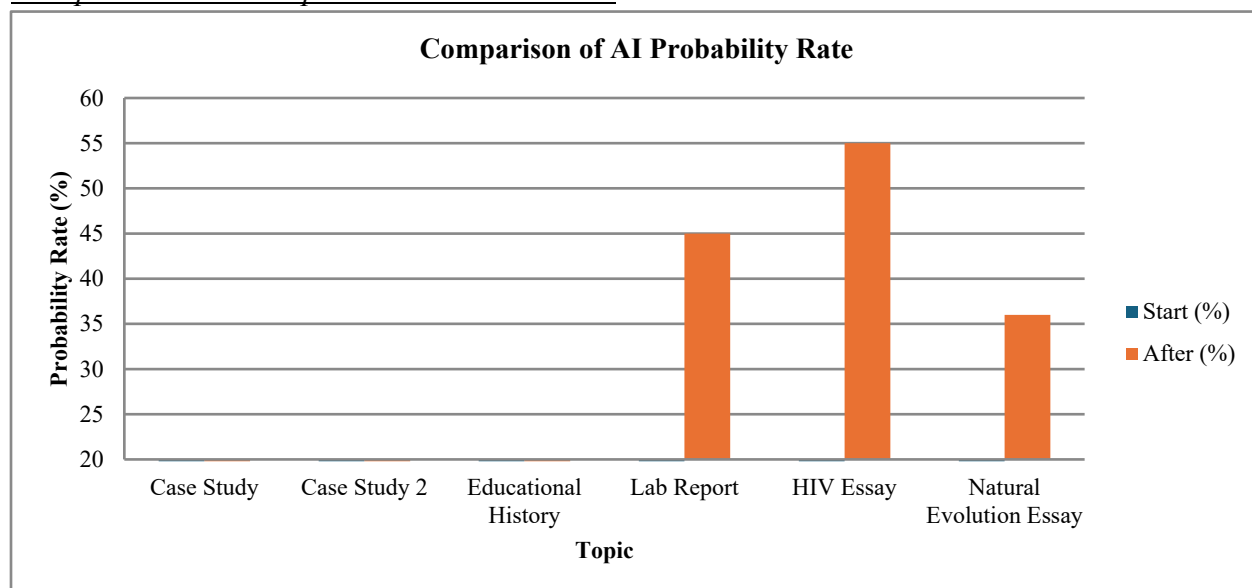
Average word count of initial generation is 552 words, with a standard deviation of 49. The average word count of extended version is 853 words with a standard deviation of 43. The humanized extended version has an average word count of 723 with a standard deviation of 61. From the figure, we could see there is no correlation between word count and the AI probability.

### Group 2: ESL Writing



For ESL writings, I examined five different pieces: a short story I wrote in high school (3040 words), two college application essays (509 and 902 words), and two TOEFL essays (792 and 391 words). Out of these, only the high school short story showed an AI probability score of 27%, while all the other writings registered a 0% AI probability score. It is possible that the short story's length and my choice of less diverse or advanced vocabulary triggered Turnitin's AI detection. However, because the two college application essays and the two TOEFL essays all received a 0% AI probability score, Turnitin's detector did not seem to exhibit any bias against ESL writing. This observation corresponds to the findings of Turnitin's research team, which also indicates that the system does not inherently discriminate based on language background.

### *Group 3: Grammar Improvement via ChatGPT*



For the two case studies and the educational history document, the AI probability score remained at zero both before and after revisions. However, for the lab report and the two essays, the score increased significantly after edits. One interesting observation is that even when I specifically instructed ChatGPT to focus on grammar improvements, structural and wording changes were made. Despite these modifications, the main themes remained intact. ChatGPT primarily enhanced structure and flow, improving transitions and word use to improve clarity. In more personal or cohort-specific writings, such as case studies and educational history, these changes did not increase AI probability scores. This result suggested that Turnitin considers the content of the writing when assessing AI writing probability. In research papers or lab reports, where information is more based on facts, ChatGPT's reorganization of content may have influenced the AI probability score. It could be possible that due to the personal details provided in the case studies and educational history documents, Turnitin ignored the structure changes made by ChatGPT. It is also possible that ChatGPT organizes personal experience-based writing differently from academic writings.

#### Group 4: Published Papers

For the three published papers uploaded, the AI detection rate remains zero for all. The topics covered were cognitive science, neuroscience, and writing developments. This suggested that Turnitin's AI detector does not show bias against highly technical content.

### **Discussion**

Through the data analysis above, there is no clear sign of Turnitin's AI detector being biased against ESL students. However, something worth noticing is that due to the small and non-diverse samples, the results from this analysis have very limited generalizability. It is difficult to conclude whether Turnitin's AI detector is ready to be reliably implemented in a diverse environment like UC Davis—not only due to the limited sample size but also the methodological flaws, including the lack of an equivalent control group, statistical testing, and confounding variables. Writing is a highly individualized task; there should be more research being done about how generalizable the research results from Turnitin or other institutions are before coming to a conclusion.

Still, it is possible that the Turnitin team has made significant improvements to their AI detector in response to complaints raised on online forums like Reddit's r/GradSchool (Reddit, 2023). However, this week, on r/UCDavis, a student reported being accused of using AI to write their essay and is currently working hard to dispute this accusation (Reddit, 2025). According to the post, this isn't the first time the student has faced such an accusation; a previous allegation was successfully disputed when the student provided sufficient evidence to prove they had written the essay independently (Reddit, 2025). For the current dispute, the student has documented evidence from Google Document's version history and was seeking advice on whether this evidence would suffice. This brought concern to how the Office of Student Support and Judicial Affairs (OSSJA) works. If a student is mistakenly accused of using AI to plagiarize but has edited their work locally using Microsoft Word rather than Google Docs, how can they effectively prove their innocence or defend themselves? At the same time, the complains about AI detector in the comments of this post (Reddit, 2025), align closely with those raised on r/GradSchool (Reddit, 2023), with numerous students criticizing the inaccuracies of Turnitin's AI detector. The phenomena observed in these comment sections significantly contradict Turnitin's claims regarding the accuracy of their AI detector, as well as the findings presented in this paper.

Though it is hard to drop any conclusion without knowing the full details of any incident, with multiple complaints stating that Turnitin AI detector was inaccurate, maybe it is time for UC Davis to consider a change. OSSJA functions similarly to a court by adjudicating student conduct cases at the university level, holding hearings where evidence is presented, determining whether a violation occurred, and assigning disciplinary sanctions when necessary (OSSJA, 2018; OSSJA, 2023). The consequences of AI plagiarism could be disciplinary sanctions, ranging from censure to dismissal; grade penalty, receiving an F for the course if the academic misconduct is

confirmed; and academic records, records of academic misconduct maintained by OSSJA (OSSJA, 2023). If OSSJA holds the power to alter a student's future by deciding on cases of academic misconduct, it should transparently define clear criteria for misconduct. What criteria need to be met to be 'guilty'? To what extent is a student's past academic work taken into consideration? Unfortunately, I was unable to find guidelines detailing how AI plagiarism cases are reviewed. I believe it is essential for UC Davis to make an accessible and clear criterion of definition of how each misconduct is being evaluated. It would be helpful for both instructors and students if there are clear criteria of how misconduct on plagiarism of AI is being defined. There could be mock case studies, worksheets, or checklists available to students for us to understand what exactly the moral use of AI on campus is.

### **Conclusion**

Through the data collection, I do not see any bias from Turnitin's AI detector. However, the false positive rate of the detector remains debatable, as there is no clear answer on whether using ChatGPT to improve structure and grammar should be considered plagiarism, especially when the original theme is intact. Despite ongoing uncertainties and improvements needed in regulations regarding AI use in academic settings, instructors could allow students who use AI tools to submit an initial draft created without AI, and that draft will not be used for grading. This original draft could later serve as evidence for both instructors and students if allegations of AI-generated plagiarism arise. Additionally, Turnitin emphasizes on its website that AI probability scores should not be the sole measure or standard for penalizing students; instead, instructors should thoroughly investigate each case before drawing conclusions (Turnitin, 2024). Beyond AI probability scores, instructors should also permit students to demonstrate their innocence through browsing history, screen time records, or witness testimonies/anecdotal evidence. All evidence should be considered to ensure accurate decisions regarding academic integrity.

At the same time, while policies about AI use are still developing, universities should consider external environments beyond the classroom. An experiment conducted by MIT with 444 college-educated professionals showed that the use of ChatGPT increased productivity by 37% (MIT, 2023). Similarly, according to Nature, 38% of researchers believed generative AI could boost productivity in science by helping researchers write papers more quickly (Nature, 2024). For a clearer comparison, I spent at least one hour figuring out how to create each Excel graph used in this article and determining how to organize my data for accurate axes. In contrast, ChatGPT generated the exact same graph in just 15 minutes, and the data provided to it was completely unorganized. With the significant capabilities of AI, universities should consider ways to leverage these tools to enhance productivity and achievement rather than viewing AI as a threat to academic honesty. Classes or instructions could be offered to students on how to use AI to improve their productivity, and what in specific are being considered as plagiarism.

## References

- Callaway, E. (2024, February 26). *Is ChatGPT making scientists hyper-productive? The highs and lows of using AI*. Nature. <https://www.nature.com/articles/d41586-024-00592-w>
- HAI Stanford. (2023, May 15). *AI detectors biased against non-native English writers*. Stanford Institute for Human-Centered Artificial Intelligence. <https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers>
- Office of Student Support and Judicial Affairs. (2023). *UC Davis Code of Academic Conduct: Honesty, Fairness & Integrity*. University of California, Davis. <https://ossja.ucdavis.edu/code-academic-conduct>
- Office of Student Support and Judicial Affairs. (2023). *Guide to formal hearings*. University of California, Davis. <https://ossja.ucdavis.edu/guide-formal-hearings>
- Office of Student Support and Judicial Affairs. (2018). *Judicial FAQs*. University of California, Davis. <https://ossja.ucdavis.edu/judicial-faqs>
- Reddit. (2023, July 3). *My final paper triggered the AI detector of Turnitin* [Online forum post]. Reddit. [https://www.reddit.com/r/GradSchool/comments/14p23x1/my\\_final\\_paper\\_was\\_triggered\\_the\\_ai\\_detector\\_of/](https://www.reddit.com/r/GradSchool/comments/14p23x1/my_final_paper_was_triggered_the_ai_detector_of/)
- Reddit. (2023, October 26). *Help! Turnitin flagged my paper at 97% AI!* [Online forum post]. Reddit. [https://www.reddit.com/r/GradSchool/comments/17d62j9/help\\_turnitin\\_flagged\\_my\\_paper\\_at\\_97\\_ai/](https://www.reddit.com/r/GradSchool/comments/17d62j9/help_turnitin_flagged_my_paper_at_97_ai/)

Reddit. (2025, March 7). *Accused of AI?* [Online forum post].

Reddit. [https://www.reddit.com/r/UCDavis/comments/1j4m3qu/accused\\_of\\_ai/](https://www.reddit.com/r/UCDavis/comments/1j4m3qu/accused_of_ai/)

Turnitin. (2023, September 6). *New research: Turnitin's AI detector shows no statistically significant bias against English-language learners.* Turnitin.

<https://www.turnitin.com/blog/new-research-turnitin-s-ai-detector-shows-no-statistically-significant-bias-against-english-language-learners>

Turnitin. (2024). *AI writing detection model.* Turnitin. [https://guides.turnitin.com/hc/en-us/articles/28294949544717-AI-writing-detection-model#h\\_01J7KP9CQZ6P0BNKDDVN7A190F](https://guides.turnitin.com/hc/en-us/articles/28294949544717-AI-writing-detection-model#h_01J7KP9CQZ6P0BNKDDVN7A190F)

Turnitin. (2024). *Turnitin's AI writing detection capabilities FAQs.* Turnitin.

<https://guides.turnitin.com/hc/en-us/articles/28477544839821-Turnitin-s-AI-writing-detection-capabilities-FAQs>