Rethinking ChatGPT Detection

April 11, 2025

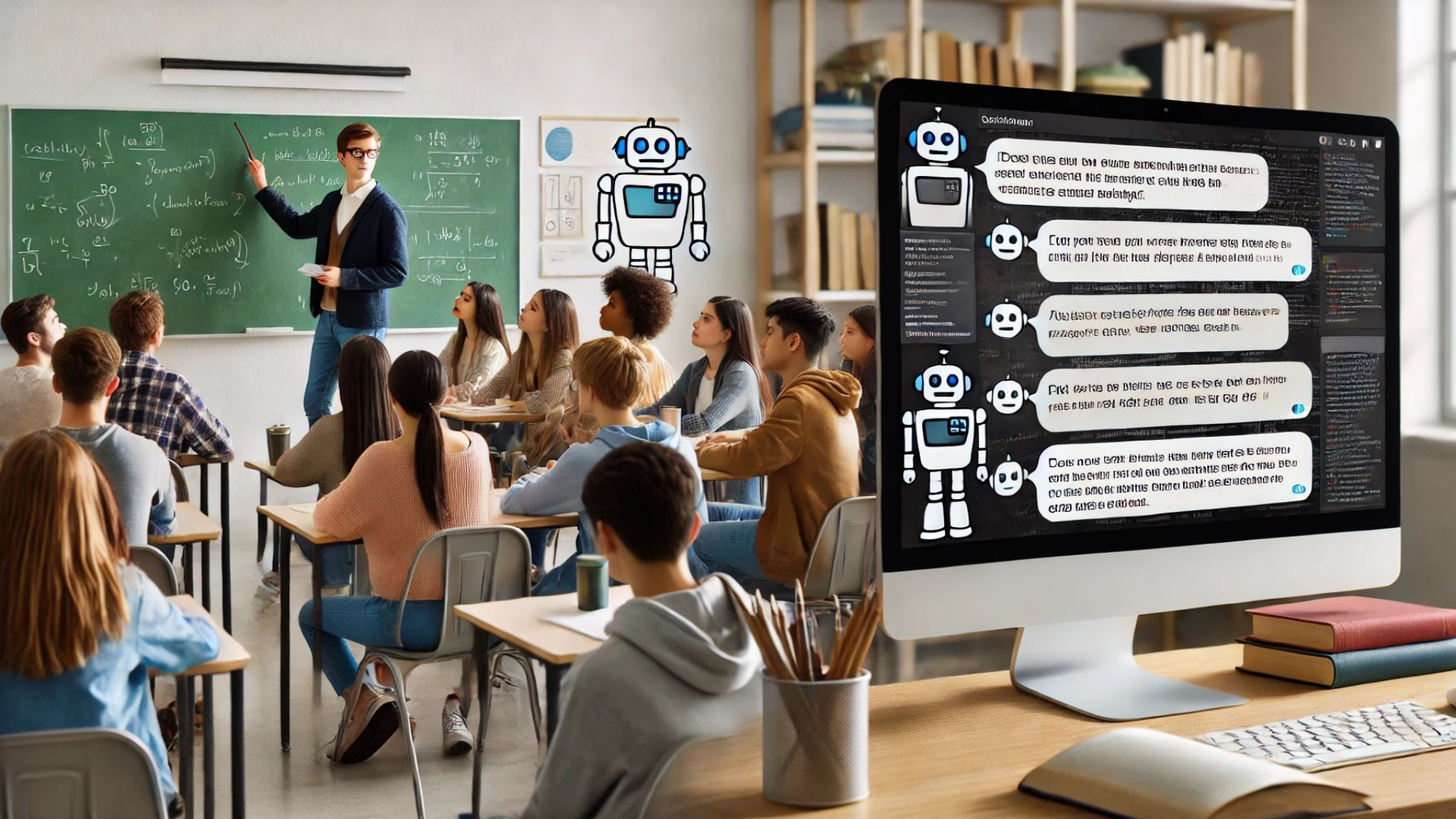
Presented by
Fadi Muheidat & Mandy Taylor

ChatGPT Edu workshops Facilitation Team:

Tracy Medrano, *IDAT*, **Mandy Taylor**, *IDAT*, **Elizabeth Viramontes-Merino**, *IDAT*, and **Fadi Muheidat**, *TRC*

Objective

- Understand the limitations of AI detection tools.
- Explore why false positives/negatives occur.
- Learn alternative strategies to address academic integrity.
- Discuss how to align classroom policies with CSUSB FAM 803.5.
- Discover ways to incorporate GenAI into teaching effectively.



“What’s one challenge you’ve faced related to AI-generated content in your teaching?”

How AI Detection Tools Work

How AI Detectors Work - Step 1

Step 1: Input Text

- The user submits a piece of text (e.g., an essay, paragraph, or response).
- Example: *"The history of ancient Rome spans over 1,000 years, beginning with the founding of the city in 753 BCE."*

How AI Detectors Work - Step 2

Step 2: Tokenization

- The text is broken down into smaller units called tokens (words, subwords, or characters).
- Example:
  - Tokens: ["The", "history", "of", "ancient", "Rome", ...]

How AI Detectors Work - Step 3

Step 3: Statistical Analysis

- The detector analyzes the perplexity and burstiness of the text:
  - **Perplexity**: Measures how predictable the text is. Low perplexity suggests AI-generated text (e.g., repetitive patterns).
  - **Burstiness**: Measures variation in sentence length and complexity. AI tends to produce more uniform sentences.

How AI Detectors Work - Step 4

Step 4: Pattern Matching

- The detector compares the text against a database of known AI-generated samples.
- Example:
  - If the text matches patterns commonly found in AI outputs (e.g., certain word combinations), it's flagged as AI-generated

How AI Detectors Work - Step 5

Step 5: Output Decision

- Based on the analysis, the tool assigns a confidence score (e.g., 80% likely AI-generated).
- Example:
  - Confidence Score: *80% AI-generated.*
  - Verdict: *Flagged as AI-generated.*

***Possible Watermarking** , still not reliable due to paraphrasing*

Why AI Detection Tools Are Unreliable

Challenges

- False Positives: Human writing flagged as AI-generated.
- False Negatives: AI-generated text passing as human-written.
- Rapid evolution of GenAI models makes detection harder.
- Overlap between AI and human writing styles.
- Lack of transparency in detection algorithms.

Rethinking Academic Integrity: Alternative Strategies

How confident are you in relying on AI detection tools?

Discussion:

What's one assignment you could redesign to reduce reliance on AI?

Creating a Clear Classroom Policy

Considerations when making your policy

Essential Parts:

- Clear lines on what is/isn't allowed; examples are helpful
- How students should communicate AI use, if required
- Clear consequences for breaches of policy

Recommendations::

- Explanations for your policy (Context)
- Work towards the goal of shared language and understanding of these tools
- Create a culture of communication
- Disclose your own AI use, such as if you use an AI detector or any similar tool

Policy Example - by Chris Ostro

AI Policy: We live in a brave new world of LLMs and AI. Tools are coming out every day that can do everything from creating art to writing papers. These tools are powerful and can be incredibly helpful. I firmly believe that my role here should be to help you learn to use these tools ethically, not just ban them outright. If you want to use ChatGPT (or another LLM) to help you brainstorm or understand concepts better, I think that's a great idea! When you submit an assignment you are also claiming you created that assignment and that simply isn't the case if you ChatGPT'd it up. Therefore, little of that content, if any, should make it verbatim into any final draft you submit to me. If any of it does, that needs to be cited. **If you use Generative AI in this class in any way,** please include it in your references, making sure to include which prompt was used. If you used an LLM in a way that didn't have text enter your paper, you should still write a little paragraph after your references explaining what you used and in what way. **If you are transparent on your AI use, you will never be reported to the Honor Code, even if you rely on the tools too much (though you may be asked to revise your paper)**. As time has went on, AI detection software has gotten more reliable and, while it's not reliable on its own, it does a great job at highlighting lazy AI use. I encourage you to run any paper through GPTZero before submitting it. If it throws up any erroneous flags, reach out to me and we'll figure it out. If I suspect you submitted AI-written content ***without proper citations***, I will reach out to you to discuss and if you are in breach of policy you WILL be reported to the Honor Code, which makes life much more complicated. **Using verbatim text from LLMs and failing to disclose your use of LLMs will be treated as a form of academic dishonesty akin to plagiarism or cheating, and will be reported to the Honor Board and punishable with a zero on the assignment.**

For more information on how to disclose AI usage, as well as formatting examples, see the "Generative AI Policy" page on Canvas

Leveraging GenAI to Enhance Teaching and Learning

“The future of education isn’t about fighting technology ; it’s about harnessing it **responsibly**.

Some Takeaways

- AI detection tools are **unreliable**; focus on prevention and transparency instead.
- **Redesign** assignments to **encourage** original thinking and collaboration.
- Develop **clear policies** aligned with institutional guidelines.
- Embrace GenAI as a tool to enhance „" **not replace**", teaching and learning.

Q&A

Next workshops:

- **ChatGPT and Instruction: Rethinking Teaching Practices (M 4/14)**
- ChatGPT and Instruction: Rethinking **Student Engagement** (T 4/15)
- ChatGPT and Instruction: Rethinking **Assessment** (W 4/16)
- ChatGPT and Instructions: **Inclusive & Equity-Minded Teaching** (Th 4/17)
- Creating **Custom GPTs** (F 4/18)

All workshops are from 11am-12pm and by Zoom

Thank you