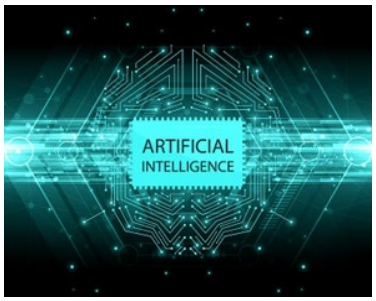
VITALITY

VITAL Teaching Ideas

Detecting AI: Suggestions to Develop Your Detection Skills

Students using AI to generate text and submitting it without proper attribution is a real and valid concern across campus. Addressing this concern requires a multifaceted approach that is fundamentally predicated upon first educating students about AI, its role, or lack thereof, within the course and discipline, as well as providing clear expectations, guidelines, course policies, and grading schema aligned with Villanova University's Code of Academic Integrity and well documented in the syllabus. The communication from the Vice Provost for Teaching and Learning maybe a good resource to get started.



Another facet is AI detection. *Detection is a key element to holding students accountable and maintaining integrity in teaching and learning.*

AI detection presents challenges, in part because there are no good AI language detectors. At best, AI detection tools will be correct 50% of the time, and worse yet, the available tools tend to falsely accuse English language learners—and anyone who may write in ways large language models do—of cheating.

However, research suggests that humans can detect AI generated text quite well and can improve with practice. Colleagues at the University of Pennsylvania created a game called Real or Fake Text, played by presenting participants with the first sentence of the text, which is always human-written, followed by the rest of the essay one sentence at a time. At some point, the text switches from human-written to machine generated text, and the objective of the game is for players to identify what sentence that happens.

The evidence suggests that while human ability to identify the lines of demarcation—where human writing ends and machine generated text begins—may not be better than 25% of the time, humans can detect *any* machine generated language in a given essay almost 75% of the time.

How might we become more adapt at identifying machine generated text? The research suggests a three-step process to develop your skill set to do so.

1. ***Be patient when reading***. While it is overwhelming to think about reading every sentence closely five papers into a 10-paper grading session, the research indicated that the longer the participants took in the game the better able they were to identify AI text. Importantly, participants did not spend more time with longer sentences, suggesting that *it is not so much spending more time reading, but more time between sentences thinking about what was read*.
2. ***Do not look for the errors machines make that humans can least rely upon for detection.*** The researchers recorded the mistakes participants thought alerted them that the text was machine generated, and then quantified how reliable those reasons were (i.e., whether the reason led to the identification of machine generated text).

Even though grammar and identifying "generic" assertions were the 4th and 5th most used by participants, these were also the least reliable among nine distinct reasons given. *Therefore, looking for grammar mistakes or identifying generic assertions to identify machine generated text does not seem to be a particularly good strategy*.

3. ***Look for errors machines make that humans can likely rely upon for detection.*** The four most reliable strategies for detecting machine generated language were violations of common sense, text irrelevant to the narrative, contradictions to the previous sentences, and contradictions of the reader's knowledge. Actively looking for these four errors is likely the best strategy to detect machine generated text.

    a. **Violations of Common Sense**—assertions that violate common sense might fit into the narrative generated by a machine, but those assertions will contain a flaw in logic that is unaccounted for.
    b. **Parts of the text are irrelevant**—these parts may be a narrative off topic, but when a machine generates these errors, it is likely that a sentence will be unrelated to the narrative, especially the previous sentence.
    c. **Contradictions to the previous sentence**—two assertions may be accurate or supported by data and sources yet are presented as equal facts without acknowledgement and/or resolution of their conflict.
    d. **Contradictions of the reader's knowledge**—if what is written contradicts your understanding of people, events, and concepts, it may be machine generated.

The available evidence suggests that professors can train themselves to outperform current AI detection tools. Still, it is important to recognize that developing these skills will take time and practice, so it is important to be patient with ourselves. It is also important to continue dialogue around AI and maintain conversations about our AI detection practices at departmental meetings and contribute insights to departmental AI committees.

Of course, detection is just the beginning; a professor who suspects a student has presented AI generated text as their own will need to engage the student to uncover the facts and look to Villanova University's Code of Academic Integrity for guidance. A future issue of VITALITY will focus upon next steps that can support students' learning and development after suspected detection.

How have you engaged with students around AI? We invite you to share a practice/learning activity/assignment by emailing VITAL so we can continue the conversation in future issues.

Authored by Stefan A. Perun, Associate Director for Digital Learning Pedagogy, November 2023

References

Dugan. L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch., C. (2023). Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. Association for the Advancement of Artificial Intelligence. Retrieved October 30, 2023, from https://www.cis.upenn.edu/~ccb/publications/real-or-fake-text-analysis.pdf

Trust, T. (2023). Essential considerations for addressing the possibility of AI-driven cheating, Part 1. *Faculty Focus*. Retrieved October 30, 2023, from https://www.facultyfocus.com/articles/teaching-with-technology-articles/essential-considerations-for-addressing-the-possibility-of-ai-driven-cheating-part-1/