VANIER
CÉGEP/COLLEGE

AI Detector Tools: Teachers Guidelines

This page will be updated regularly.

Overview

These guidelines briefly define generative artificial intelligence (Gen AI) and outline the College's decision not to endorse or recommend the use of AI detection tools due to concerns about their reliability, as well as the ethical and legal implications of their use. They also offer suggestions on how to approach suspected cases of academic misconduct involving Gen AI.

What is Gen AI?

Gen AI refers to machine learning models that are trained on large sets of internet data and refined with human feedback. When prompted, these models can create various types of output such as text, code, image, video, and sound.

The term “artificial intelligence” can be misleading: while Gen AI may appear to reason or think, these systems are mainly predictive in nature. For instance, large language models (LLMs) like ChatGPT or Copilot are text-based Gen AI models that predict the most probable word to follow in a sequence of words, creating a semblance of meaning. While LLMs can produce accurate responses, they may also “hallucinate,” generating inaccurate or convincingly plausible yet fabricated outputs. Finally, because LLMs are trained on internet data, they reflect human biases which they can amplify.

The quality of a model’s output relies on the volume and quality of their training data. As an example, as of the drafting of these guidelines, models such as ChatGPT, Copilot, and Gemini tend to generate more accurate results when queried in English (Ta & Turner Lee, 2023). These models excel at text synthesis and revision. Also, performance can be notably enhanced when they are provided with precise prompts, full context, and relevant documents for the model to analyze.

Because of the predictive nature of LLMs, good prompt design[1] may reduce hallucinations, but it cannot eliminate them. For example, when LLMs are asked to produce a list of

[1] Also called prompt engineering.

references, they may invent titles and URLs. This occurs because LLMs construct responses based on probability, rather than by creating and or verifying content. Despite hallucinations, the outputs produced by more recent models are impressive. LLMs, when provided with samples of a user's writing, can be prompted to mimic their writing style, including any errors they may typically make. This makes it increasingly difficult for teachers to distinguish between human and AI generated content (Casal & Kessler, 2023; Fleckenstein et al., 2024; Scarfe et al., 2024).

AI Detection Tools

Despite the need to identify AI-generated text, current AI detection tools are unreliable. They are not merely inadequate (Fleckenstein et al., 2024; Elkhatat et al., 2023) but their use can be problematic. They generate false positives and false negatives (an issue exacerbated by the ability of LLMs to mimic student writing styles). Moreover, research has demonstrated that these tools tend to produce results biased against non-native English writers (Liang et al, 2023).

For these reasons, like many institutions of higher learning,[2] Vanier has chosen not to enable Turnitin's AI detection features and does not endorse or recommend the use of AI detectors. Teachers should be aware of issues related to sharing student work. Even if a text is anonymized, submitting it to a tool that is not protected behind a college portal may inadvertently breach Quebec's Personal Information Protection Law (Law 25).

The College will continue to monitor the progress of AI detection tools, hoping that an innovative solution may be found.

What to do if you suspect the unpermitted use of AI

- **Review the work**
  Not all students are adept at prompting LLMs to imitate their writing style. Begin by identifying inconsistencies between the text and work the student has previously submitted, especially work done in-class.

[2] Cambridge, Oxford, Yale, MIT, UBC, U of Nottingham, Concordia, Dalhousie, McMaster, USC San Diego, to name a few.

- **Request drafts**
  Emphasize process over product to ascertain the effort invested in writing drafts. Platforms like M365 Word online or Google Docs provide version histories. Student-generated work may include drafts with noticeable, incremental changes, whereas AI-generated text may appear in one or a few drafts with large sections added into a single version.

- **Verify Source Accuracy**
  Hallucinated titles and URLs may be a tip-off that a Gen AI model was used.

- **Test LLMs as a discipline expert**
  While some AI models are trained on data particular to a field of study, most students have access to more generic models. Lacking the expertise to assess the quality of the responses, some students may believe the well-formatted and expressed outputs include deeper or more accurate content than was generated. Using the same evaluation instructions given to students, prompt a few LLMs to see if you notice similar generalities or error patterns in the responses. Instances of similar types of generalities or error patterns or a lack of discipline-specific details covered in class or required by the assignment could be clues that Gen AI was used to generate the content.

- **Assess via interview or conversation**
  Discuss their work process and the final product. Ask about brainstorming, idea development, choice of examples, source selheection, and, if relevant, request digital copies of cited sources. A poor explanation of their work process or a poor description of their work may indicate that they did not produce the text without assistance.

- **Request an in-class (or in-office) writing sample**
  You might consider asking students to produce an in-class paragraph or two about their assignment (for potential topics, see the previous bullet point). In-class writing will not be of the same caliber as at-home writing, but it might reveal something of the student's ability to express their understanding of their own work in written form.

VANIER
CÉGEP/COLLEGE

Helping to minimize unpermitted AI use in the future

For strategies to help discourage unpermitted uses of Gen AI, please see this guide on enhancing student learning and minimizing temptations to cut corners using Gen AI.

References

Casal, JE, Kessler, M. Can linguists distinguish between ChatGPT/AI and human writing?: a study of research ethics and academic publishing. Res. Methods Appl. Linguist. 2023;2(3). https://doi.org/10.1016/j.rmal.2023.100068

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal of Educational Technology in Higher Education, 19*(1), 17. https://doi.org/10.1007/s40979-023-00140-5

Else, H. (2023). Abstracts written by ChatGPT fool scientists. Nature 613, 423. https://www.nature.com/articles/d41586-023-00056-7

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024, June). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence*, 6, 100209. https://doi.org/10.1016/j.caeai.2023.100209

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing Test" case study. PLOS ONE, 19(6), e0305354. https://doi.org/10.1371/journal.pone.0305354

Ta, R and Turner Lee (2023, October). How language gaps constrain generative AI development. *Brookings*. https://www.brookings.edu/articles/how-language-gaps-constrain-generative-ai-development/