Chapter 11
AI Detection's High False Positive Rates and the Psychological and Material Impacts on Students

Whitney Gegg-Harrison
https://orcid.org/0009-0008-4258-2834
University of Rochester, USA

Claire Quarterman
Independent Researcher, USA

ABSTRACT

This chapter, per the authors, explains the inherent impossibility of "AI detection," and explores the material and psychological impacts of AI detection false positives on students. A small corpus study is presented demonstrating much higher than advertised rates of false positives across a range of popular "AI detection" tools. Based on this study along with news reports and first-person testimony from affected students, the chapter presents the possibility that neurodivergent writers, along with L2 writers, are more likely to be impacted by false positives. Given the current rates of mental health challenges on college campuses and the likelihood of a disproportionate impact on students who already face marginalization, the use of these AI detection tools is argued to be unethical. The chapter closes with recommendations for writing teachers.

INTRODUCTION

The release of ChatGPT in November, 2022 came towards the end of what was, for most college instructors, the fifth straight semester of disruption due to the COVID-19 pandemic. Instructors saw magazine headlines like "The College Essay is Dead" (Marche, 2022), and were left to wonder whether the work students submitted at the end of the semester was actually their own, and what they should do to best

DOI: 10.4018/979-8-3693-0240-8.ch011

prepare for the next semester, when ChatGPT and other tools would be in their classrooms whether they wanted them there or not. Professors face an increasing challenge: how to know whether a text was actually written by the student who submitted it?

The remote learning that took place during the pandemic had already led many schools, fearing rampant cheating, towards using automated surveillance tools, with predictably terrible impacts on the stress levels of student test-takers (Harwell, 2022). So it is not surprising that the release of ChatGPT also stoked fears about cheating, and alongside those fears, an impulse towards banning the use of ChatGPT and other similar tools (Johnson, 2023; Yang, 2023). Capitalizing on this impulse, developers like Edward Tian of GPTZero and companies like Turnitin raced to create tools that would "detect" the use of tools like ChatGPT. It has always been possible for students to turn in work that is not their own, e.g. through contract-cheating. What AI text-generators have changed is primarily due to their accessibility and speed relative to human writers, which likely changes the calculus for students considering passing off AI-generated text as their own.

In this chapter, we argue that automated detection tools are not the answer to this problem. Large Language Models (LLMs) that power tools like ChatGPT are designed to produce text that is plausibly human, and because of this, there will always be the risk of both false positives and false negatives when tools are used to determine whether the text was produced by a human or by an LLM[1]. Both of the authors of this text have had their own writing falsely flagged as "AI-generated"; this informs the position we take in this chapter.

THE FUNDAMENTAL IMPOSSIBILITY OF "AI-DETECTION" FOR WRITING

We begin by breaking down the "GPT" in ChatGPT, which stands for "Generative Pre-trained Transformer". The word "transformer" refers to a type of neural network architecture first introduced by the famous paper "*Attention is all you need*" (Vaswani et al., 2017). Models using transformer architectures can learn about what words are likely in various contexts by processing sequences of text; the model can attend not just to the immediately preceding word, but to words in other positions in the sequence, and even to intermediate representations in other layers of the neural network.

The learning occurs during "pre-training". In essence, the model learns by playing a "fill in the blank" game: given a sequence of text with a mask in place of one of the words, it predicts what word should replace the mask, and updates its parameters to account for how close or not the guess was to the actual masked word. "Closeness" is based on the representations within the neural network, which are called "embeddings". A word embedding is a vector of values representing information about the contexts in which that word appears, which can be thought of as representing coordinates in a massively multidimensional space. For any pair of words, their "embedding" vectors can be compared to generate a measurement of how "similar" they are. While there is good reason to be skeptical about precisely how much world knowledge is actually captured in these word embeddings (Bender & Koller, 2020), some cognitive scientists argue that word embeddings closely align with human conceptual knowledge (Piantadosi & Hill, 2022). Research suggests that transformer models can predict nearly all of the variance in human neural responses to sentences (Schrimpf et al., 2021), and that alignment with humans at both the neural and behavioral level is possible even with "developmentally-realistic" amounts of training data (Hosseini et al., 2022). To be clear, this does not mean that LLMs are equivalent to human