



# The Risks of [Not] Adopting AI in HE

**University of Copenhagen** (November, 2023)





# GPTZero

Humans Deserve the Truth



97% Original 3% AI



 Likely both AI and Human!



## Can AI-Generated Text be Reliably Detected?

Vinu Sankar Sadasivan  
vinu@umd.edu

Aounon Kumar  
aounon@umd.edu

Sriram Balasubramanian  
sriramb@umd.edu

Wenxiao Wang  
wxw@umd.edu

Soheil Feizi  
sfeizi@umd.edu

Department of Computer Science  
University of Maryland

### Abstract

The rapid progress of large language models (LLMs) has made them capable of performing astonishingly well on various tasks including document completion and question answering. The unregulated use of these models, however, can potentially lead to malicious consequences such as plagiarism, generating fake news, spamming, etc. Therefore, reliable detection of AI-generated text can be critical to ensure the responsible use of LLMs. Recent works attempt to tackle this problem either using certain model signatures present in the generated text outputs or by applying watermarking techniques that imprint specific patterns onto them. In this paper, both empirically and theoretically, we show that these detectors are not reliable in practical scenarios. Empirically, we show that *paraphrasing attacks*, where a light paraphraser is applied on top of the generative text model, can break a whole range of detectors, including the ones using the watermarking schemes as well as neural network-based detectors and zero-shot classifiers. Our experiments demonstrate that retrieval-based detectors, designed to evade paraphrasing attacks, are still vulnerable against *recursive* paraphrasing. We then provide a theoretical *impossibility result* indicating that as language models become more sophisticated and better at emulating human text, the performance of even the best-possible detector decreases. For a sufficiently advanced language model seeking to imitate human text, even the best-possible detector may only perform marginally better than a random classifier. Our result is general enough to capture specific scenarios such as particular writing styles, clever prompt design, or text paraphrasing. We also extend the impossibility result to include the case where *pseudorandom* number generators are used for AI-text generation instead of true randomness. We show that the same result holds with a negligible correction term for all polynomial-time computable detectors. Finally, we show that even LLMs protected by watermarking schemes can be vulnerable against *spoofing attacks* where *adversarial humans* can infer hidden LLM text signatures and add them to human-generated text to be detected as text generated by the LLMs, potentially causing reputational damage to their developers. We believe these results can open an honest conversation in the community regarding the ethical and reliable use of AI-generated text. Our code is publicly available at <https://github.com/vinusankars/Reliability-of-AI-text-detectors>.

~26% success rate

AI “anti-detection”  
technology

[ud]  
UNDETECTABLE.AI

“GPT detectors unintentionally penalise both non-native English writers & all writers with constrained linguistic expressions”.

**What is the  
purpose of  
higher ed?**

1. To drive economic, social & cultural **growth**
2. To deliver the highest possible **quality** of education for all students







58 million new, AI-related  
jobs by 2025

26% economic growth by  
2030, powered by AI

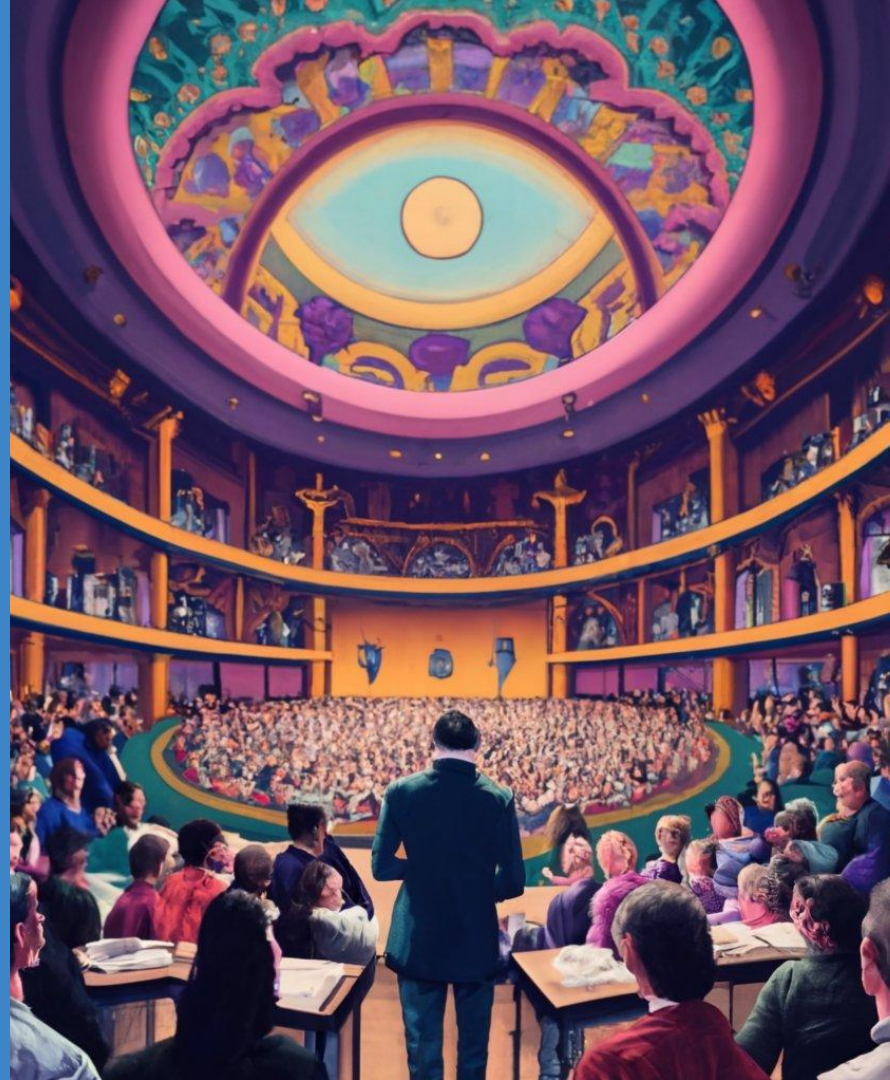
Infrastructural reform



Staff re-training



Curriculum redesign




Average 12%  
“real-world  
transfer”

# Purposeful AI in HE


 **Data-driven needs analysis:**  
who are we teaching?



 **Data-driven learning design:**  
how should we teach & assess them?



# Risks & Questions: AI in HE

 **Data Privacy:** what should & shouldn't we "feed" the machine?

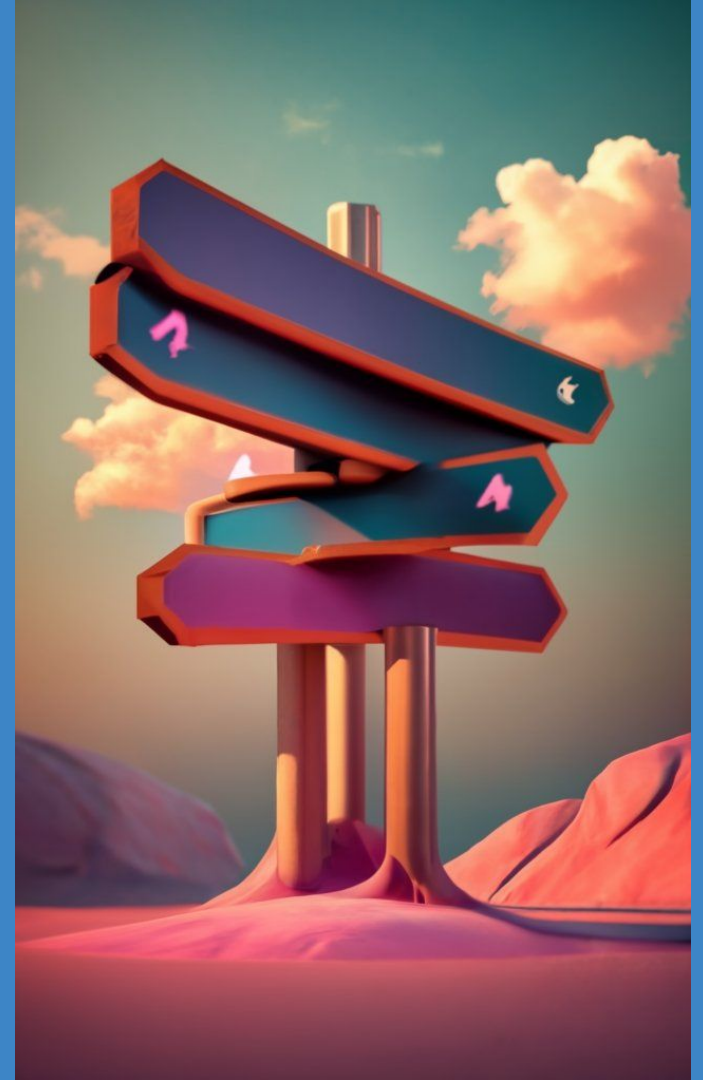
 **Human-Machine Interaction:** if we build it, will they come (and learn)?

Are humans  
& institutions  
ready for an  
AI revolution?





# 2024: A Crossroads for Higher Ed



# Thank You!



**Dr Philippa Hardman**

AI + education. Creator of the DOMSTM  
learning science design engine | TEDX Speake...



*Learn more about my research &  
prototypes*

