Comment on NIST Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)

Background: We are Checkfor.ai, two former Stanford/Google/Tesla researchers with deep experience in AI and machine learning.

On reducing the risk of synthetic content:

While there are many proposals to reduce the risk of AI-generated text content, we believe there is most merit in funding academic research into AI text detection and establishing high-quality benchmarks to establish the efficacy and limitations of available AI detectors.

Text-based watermarks have been shown to be ineffective for short text and reduce LLM output quality and easy to bypass using simple paraphrasing methods [1] [2].

We believe that AI can be a powerful assistive tool for writers, but it also enables malicious actors to produce content that is able to bypass traditional spam filters but is ultimately low-quality, inauthentic, deceptive, or even outright fraudulent. Due to this, we believe detection methods are the only way to keep the Internet free of AI spam in the long term.

Checkfor.ai has produced a classification model with 99.99% accuracy on public datasets of business reviews with an industry-leading false positive rate of 0.001%. We believe that we can reach this accuracy on other domains as well with more training data and investment in our R&D efforts.

We worked with The Transparency Company to identify fake reviews and found that ~40% of fake reviews in 2023 were written by AI (up from near 0% in 2021). These are preliminary results and we will publish a larger scale study on 100 million reviews in the coming months. Another promising approach to mitigating harm created by AI-generated text is to bring in behavioral signals such as reviewer usage patterns and history, and we believe this line of research can also be applied to other spam and fraud related areas of risk.

These are promising results that show, with more research and funding, the risk of low-effort AI-generated spam or misinformation can be minimized. Our classification methods are cost-effective and can be deployed widely across the web - one of the main hurdles is getting large companies to care about the quality of their user generated content enough to filter out AI reviews.

Another blocker to widespread adoption is the existence of low-quality AI detectors posting accuracy numbers on flawed benchmarks (Originality.ai, ZeroGPT), as well as AI detectors with high false positive rate being used in education (Turnitin) [3]. Anyone can train a detector and make claims at accuracy but if their claims at accuracy don't match real world performance, people lose trust in the industry as a whole. This exhibits the need for a large, clean benchmark to provide accuracy and false positive numbers across a number of important domains (student writing, reviews, SEO spam, etc.).

[1] https://arxiv.org/abs/2303.13408
[2] https://arxiv.org/abs/2311.04378
[3]
https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/