Detecting Artificial Intelligence: A New Cyberarms Race Begins

Mark Campbell, EVOTEK
Mlađan Jovanović, Singidunum University

The urgency to recognize the origin of digital content is spawning many detection solutions. As generative artificial intelligence detectors overcome current limitations, they will attempt to keep pace with sophisticated generative model development.

ChatGPT and other generative artificial intelligence (AI) models are leaping from research labs into headlines at breakneck speed. From creating realistic images and videos to producing natural language text, generative AI is emerging in almost every digital use case. However, concerns grow about generative AI misuse, such as the creation of deepfakes, disinformation campaigns, instant plagiarism, and inaccuracies. Moreover, artificial artifacts proliferation makes users anxious about not being able to discern true from fake content, and thus spawns a host of generative AI detection applications.

GENERATIVE AI MODELS

Generative AI models learn in several refinement phases using techniques like reinforcement learning with human feedback (RLHF) (Figure 1). A pretuned model is created with general-purpose parameters and trained on a large collection of diverse data using unsupervised learning. This pretuned model is fine-tuned using labeled domain-specific data under supervised learning and reinforcement learning with human feedback (RLHF).[1] The model response to a query is then evaluated by a human to improve precision and accuracy. Many domain-specific generative models are available through open source and can be customized (for example, Stable Diffusion and HuggingFace), while others are proprietary and accessible through application programming interfaces

Digital Object Identifier 10.1109/MC.2023.3279446
Date of current version: 26 July 2023

0018-9162/23©2023IEEE

EDITOR **MARK CAMPBELL**
EVOTEK; mark@evotek.com

(APIs) or premium-level subscriptions (such as ChatGPT and DALL-E). Once deployed, fine-tuned models can be further refined through natural language instructions containing correct examples, or prompts, to learn new tasks instantly.[2]

For example, generative AI is reshaping the education process for both students and educators. "We can start with an idea of what we want from educational content and generate complete content for the course," notes Dr. Dragan Gašević, distinguished professor of learning analytics at Monash University. Generative AI is also being used "in education assessment, such as psychometric testing for creating questions, known as 'item generation,' especially for generating tests for larger groups of students."[3]

GENERATIVE AI CHALLENGES

Pretuned models (such as BERT, DALL-E, and GPT-4) generate plausible content at scale, including text, images, audio, and video.[4] However, it is difficult to determine content authenticity, credibility, and accuracy, which creates several challenges, including

- *Disinformation*: Generative AI's unlimited flow of seemingly human-created content allows disinformation campaigns to proliferate fake news, propaganda, and astroturfing (the creation of fake grassroots movements) to manipulate public opinion by using social media, news outlets, or messaging apps.[5] Dr. Gašević observes, "There is a danger informed people will profit from these technologies and become more productive, while the uninformed will, potentially, be under the greater influence of these technologies and make decisions based on wrong information."[3]
- *Impersonation*: While generative AI models lack motives, fears, or other anthropomorphic drivers, they can be targeted nefariously by bad actors. In a widely reported incident in 2023, ChatGPT successfully tricked a worker into helping it prove it was not a robot. ChatGPT told the worker it was a seeing-impaired user unable to read the security challenge, so the worker provided ChatGPT the code.[6] This seemingly innocuous achievement of convincing a person that generative AI is human poses disturbing possibilities.
- *"Big Code"*: Generative AI models can create software code in many programming languages. OpenAI's Codex model and the open source platforms it underpins (for example, GitHub Copilot) can generate complex application code fragments from a few simple prompts. Many large enterprises now realize that most of their application code base, millions or billions of lines of code, was written outside their organization. Accelerated by the influx of generative code, this "Big Code" phenomenon is causing concerns about trusting the codebase running mission-critical applications.[7]
- *Inaccuracy*: Generative models like ChatGPT and Google Bard adeptly generate content and present it assertively and confidently as fact. However, generative models do not fact

Many large enterprises now realize that most of their application code base, millions or billions of lines of code, was written outside their organization.

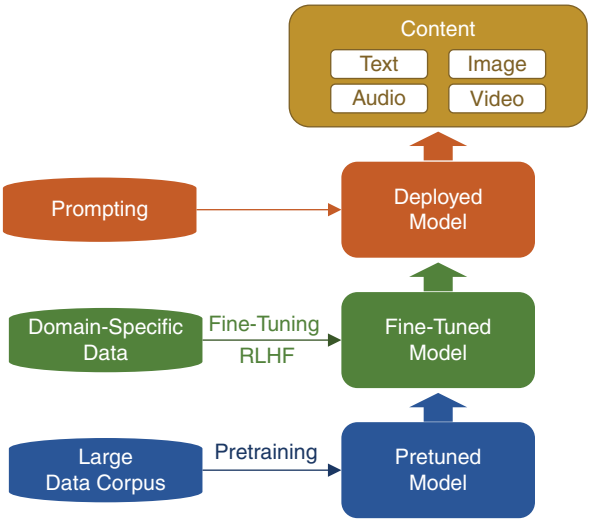

FIGURE 1. Generative AI model refinement.

check—they merely generate text by repeatedly supplying the next likely word in a sentence. This doesn't guarantee content veracity. This propensity to generate assertions with no factual data, known as "hallucinations" in AI jargon, often deceives users into believing they are accurate.[8] Biases, outdated training data, and lack of transparency further create serious concerns with generative AI output accuracy.[9]

- *Plagiarism*: Generative AI is transforming the research and educational landscape. However, what makes it an efficient learning and research support tool also gives it the potential to become a sophisticated cheating instrument. Unfortunately, a recent study found that antiplagiarism tools do not reliably detect essays written by ChatGPT as plagiarism.[10] While research organizations (including IEEE) have made specific policies against crediting AI-generated content as one's own, a recent multidisciplinary consortium of educational experts has noted the lack of appropriate policies and legislation to regulate deliberate misuse of such tools.[9] Another recent study showed that ChatGPT could assist educators by generating more detailed and consistent feedback to summarize students' performance or assessing topics for their assignments.[11] "The act of writing is very important for the activation and development of cognitive processes. So, if someone automatically receives generated content, it could be detrimental to their development and learning. On the other hand, to a certain extent, it could facilitate creativity," observes Dr. Gašević.[3]

GENERATIVE AI DETECTION

To overcome these challenges, it is imperative that systems detect the artificial origin of generative AI artifacts and notify users. Generative AI detection is still emerging, but significant progress has been made to extrapolate several observations.

Human-based detection

Since generative AI's inception, humans have scrutinized its output artifacts. Whether to create better generative AI content or improve generative models, human users can adeptly spot artifacts that fall into the oft-quoted "uncanny valley" where something is not quite right.[12] The most common nonhumanlike attributes spotted by people include[13]

- *Uniformity*: Generative AI artifacts are typically rigidly consistent in style, voice, diction, tenor, and structure.
- *Coherence*: Artificial content can veer off topic when given a particularly complex prompt.
- *Originality*: AI models tend to repeat formulaic phrases and clichés instead of creating original phraseology.
- *Errors*: AI models adeptly complete a user prompt's intent but don't focus on facts or logic.
- *Context*: Generative models often lose the given prompt's context and provide erroneous or irrelevant output.

As model fidelity improves, human detection will become increasingly difficult, if not impossible. Additionally, the avalanche of artificial artifacts has grown beyond the scale of manual human evaluation—automated detection is needed.

Automated AI detection

Most automated AI detection uses discriminative models to predict whether a specific artifact was created by a human or algorithm. Current detectors are typically implemented as a classifier trained on diverse datasets with authentic and synthetic exemplars, such as text and images with the same context and topic.[14] Unfortunately, current automated detection solutions lag today's advanced generative AI platforms. These generative models can also produce spurious hallucinations that are difficult to detect automatically.

One promising approach leverages a large language model (LLM) to detect its own artificial output. A pretrained LLM is fine-tuned using human-written text, then retrained on synthetic text to learn the difference.[15] Another technique is OpenAI's AI Classifier that uses an LLM fine-tuned through prompts submitted to its InstructGPT[1] tool to distinguish between human- and AI-generated content.[16]

Automated AI text detection advancements are being mirrored in voice and audio artifacts. In 2019, the world's first AI-powered cybercrime used an AI-generated voice to mimic a chief executive officer's voice and trick an executive into transferring US$250,000 to a bad actor's account. To combat this type of attack, one detector employs a temporal convolution network trained on Google's AVSSpoof dataset to analyze spectrographs of an audio sample and successfully detect artificial audio, with up to 99% accuracy.[17] Fake emotion detectors are also emerging, trained on data such as the EmoFake dataset out of China.[18]

Generative AI is pushing boundaries in the software industry by producing

Automated tools, such as DetectGPT, are emerging to help educators detect generative AI text; however, many of these tools may be vulnerable to adversarial attacks.

source code directly from natural language prompts.[19] OpenAI's Codex, Amazon's CodeWhisperer, and GitHub's Copilot can generate code in various programming languages. While these tools assist developers tremendously by simplifying and automating software development, they are trained on a corpora of public source code whose authors are not attributed in generated code fragments. A recent lawsuit against GitHub's Copilot asserts that the tool violates the rights of authors who post their code under GitHub's own open source licensing agreement.[20]

Generative AI is also challenging the education sector, where a recent study demonstrated that English language experts could not reliably distinguish machine-generated and human-created essays.[21] "Our recent research[22] shows it is not only possible to generate essays exhibiting factual knowledge but that also reflect on one's development and performance. In our example, reflective essays generated by LLMs did better than those created by the students. This raises questions on what future evaluation should look like," explains Dr. Gašević.[3]

Automated tools, such as DetectGPT,[23] are emerging to help educators detect generative AI text; however, many of these tools may be vulnerable to adversarial attacks.[29] Given the limitations of current AI detectors,[16] generative tool creators like OpenAI want deeper engagement with educators and now provide educators with high-level guidelines to help them understand the tools' capabilities and limitations.[24] These nascent initiatives, deeper discussions, and recommended practices are necessary.

Human-aided AI detection

As automated AI detection matures, a wide array of approaches is crucial to improve performance and accuracy. A collaborative strategy combining humans and automated AI detection tools improves detection reliability and capacity and is essential for complex content like materials combining topics, text, image, and video.

In human-aided AI detection, an existing generative model (trained on real and synthetic data) generates synthetic content[25] verified against domain knowledge sources (for example, graphs or databases) by human experts and an automated model. Once these moderated data are checked for syntax and semantic errors, bias, and ethical and privacy issues, they train a discriminative model to classify examples as real or synthetic via unsupervised learning. The discriminative model is further refined using labeled pairs of real and synthetic observations via supervised learning. The model is then run against known samples, and its output is evaluated by domain experts and knowledge sources. These evaluation results are fed back into the model for further refinement. Once desired performance levels are attained, the model is deployed and accessed by users through a user interface (UI) or programmatically via an API. Prompting can further improve the AI detector at runtime via a prompting UI.

As automated AI detection matures, a wide array of approaches is crucial to improve performance and accuracy.

The advantages of human-aided AI detection include

1. learning new tasks during testing and after deployment by leveraging contextual information from domain knowledge sources and human supervision
2. flexibly assessing data modality combinations in AI-generated content by engaging experts from the respective disciplines and modality-specific domain knowledge
3. improving reasoning capabilities of probabilistic models on learned and unseen tasks that employ formal knowledge and human feedback
4. increasing AI detector transparency and accountability through human involvement in detector construction and maintenance.

While Figure 2 lays a foundation for a variety of applications across numerous domains, each application will require a comprehensive verification and regulatory assessment by stakeholders.

Digital watermarking

One promising alternative to detecting artificial artifacts after the fact is to embed digital watermarks at creation time. In fact, a 2023 study from the University of Maryland explores a technique to alter word selection in a text transformer like ChatGPT so that the frequency distribution of generated words is statistically detectable yet invisible to the reader.[26] However, recent research has shown that adding even a small amount of content into a watermarked artifact, especially in video and imaging, can unravel the watermarking while maintaining the original quality.[27]

China's Cyberspace Administration recently enacted a regulation requiring digital watermarking for all synthetic text, image, voice, and video artifacts. This 2023 law prohibits the creation of any AI-generated content without clearly labeling its artificial origin and imposes penalties on anyone tampering with digital watermarks.[28]

Model distillation

Generative models can often be scaled down without appreciable fidelity loss

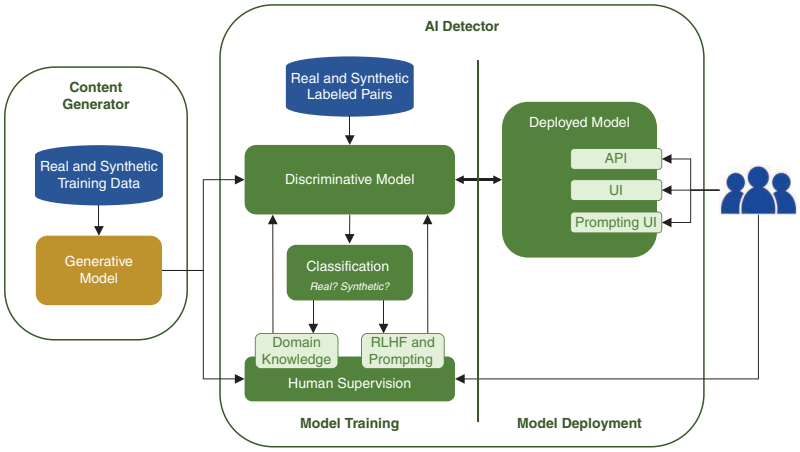

FIGURE 2. Human-aided AI detection.

via a technique known as "knowledge distillation."[29,30] When coupled with human-based tuning techniques like prompting or RLHF, knowledge distillation reduces the need for large amounts of training data and drastically reduces the size of models so that they can run on smaller edge devices (for example, smartphones) with constrained memory and computational capacities.[2] One side effect of model distillation is that output from these compact models is more transparent and susceptible to automated detection.

> One promising alternative to detecting artificial artifacts after the fact is to embed digital watermarks at creation time.

Current detector limitations

Regardless of the detection method, today's generative AI content detectors face several challenges, including[16]

- *False results*: False positives and negatives are considerable among current automated detectors. Detector accuracy also significantly drops off when content falls outside the dataset that trained the detector.
- *Sample size*: AI detector effectiveness is proportional to the sample size of the content it evaluates. For example, accurate AI detection is currently impossible in Twitter-sized text samples.
- *Language*: Many current detectors are constrained to English pretraining, fine-tuning, and prompting datasets, and they produce unpredictable results analyzing content outside their training language.
- *Postgeneration modification*: Current AI detectors cannot predict whether content was altered or augmented by a human after generation.

Challenges mount as generative AI solutions like ChatGPT reshape society. The urgency to recognize the origin of digital content is spawning many detection solutions. As these generative AI detectors overcome current limitations, they will attempt to keep pace with sophisticated generative model development. The generator detector arms race has only just begun.

ACKNOWLEDGMENT

The authors thank Dr. Dragan Gašević for his constructive suggestions and comments.

REFERENCES

1. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, and P. Mishkin, "Training language models to follow instructions with human feedback," OpenAI, San Francisco, CA, USA, 2022. [Online]. Available: https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf
2. C. Huyen. "Building LLM applications for production." Chip Huyen. Accessed: May 3, 2023. [Online]. Available: https://huyenchip.com/2023/04/11/llm-engineering.html
3. D. Gašević, Interviewee, Distinguished Professor of Learning Analytics in the Faculty of Information Technology and the Director of the Centre for Learning Analytics at Monash University, Melbourne, VIC, Australia, May 2023.
4. Stanford Institute for Human-Centered Artificial Intelligence. *On the Opportunities and Risks of Foundational Models*, Stanford Univ., Palo Alto, CA, USA, 2022.
5. T. Hsu and S. A. Thompson, "Disinformation researchers raise alarms about A.I. chatbots," *NY Times*, Feb. 2023. [Online]. Available: https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html
6. K. Hurler. "Chat-GPT pretended to be blind and tricked a human into solving a CAPTCHA." Gizmodo. Accessed: Mar. 16, 2023. [Online]. Available: https://gizmodo.com/gpt4-open-ai-chatbot-task-rabbit-chatgpt-1850227471
7. M. Nuñez, "Developers embrace AI tools but face 'big code' challenges,

survey finds," VentureBeat, San Francisco, CA, USA, Apr. 2023. [Online]. Available: https://venturebeat.com/ai/developers-embrace-ai-tools-but-face-big-code-challenges-survey-finds/
8. W. Daniel, "Google CEO Sundar Pichai says 'hallucination problems' still plague A.I. tech and he doesn't know why," *Fortune*, Apr. 2023. [Online]. Available: https://fortune.com/2023/04/17/google-ceo-sundar-pichai-artificial-intelligence-bard-hallucinations-unsolved/
9. Y. K. Dwivedi et al., "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *Int. J. Inf. Manage.*, vol. 71, Aug. 2023, Art. no. 102642, doi: 10.1016/j.ijinfomgt.2023.102642.
10. M. Khalil and E. Er, "Will ChatGPT get you caught? Rethinking of plagiarism detection," Centre for the Science of Learning and Technology, Bergen, Norway, Rep. 2302.04335v1, 2023. [Online]. Available: https://arxiv.org/abs/2302.04335
11. W. Dai et al., "Can large language models provide feedback to students? A case study on ChatGPT," in *Proc. 23rd IEEE Int. Conf. Adv. Learn. Technol.*, Orem, UT, USA, 2023, pp. 1–6.
12. M. Masahiro, K. MacDorman, and N. Kageki, "The uncanny valley [From the Field]," *IEEE Robot. Autom. Mag.*, vol. 19, no. 2, pp. 98–100, Jun. 2012, doi: 10.1109/MRA.2012.2192811.
13. S. Krishna, "AI-generated content detection tools put to the test," VentureBeat, San Francisco, CA, USA, Mar. 2023. [Online]. Available: https://venturebeat.com/ai/ai-generated-content-detection-tools-put-to-the-test/
14. R. Karjian, "How to detect AI-generated content," TechTarget, Newton, MA, USA, Apr. 2023. [Online]. Available: https://www.techtarget.com/searchenterpriseai/feature/How-to-detect-AI-generated-content
15. M. Heikkilä, "How to spot AI-generated text," *MIT Technol. Rev.*, Dec. 2022. [Online]. Available: https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/
16. J. H. Kirchner, L. Ahmad, S. Aaronson, and J. Leike, "New AI classifier for indicating AI-written text," OpenAI, San Francisco, CA, USA, 2023. [Online]. Available: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text
17. Dessa, "Detecting audio deepfakes with AI," *Dessa News*, Sep. 2019. [Online]. Available: https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35
18. Y. Zhao et al., "EmoFake: An initial dataset for emotion fake audio detection," 2022, *arXiv:2211.05363*.
19. S. Greengard, "AI rewrites coding," *Commun. ACM*, vol. 66, no. 4, pp. 12–14, Mar. 2023, doi: 10.1145/3583083.
20. P. Krill, "GitHub faces lawsuit over Copilot AI coding assistant," InfoWorld, Needham, MA, USA, Nov. 2022. [Online]. Available: https://www.infoworld.com/article/3679748/github-faces-lawsuit-over-copilot-coding-tool.html
21. Y. Lui et al., "ArguGPT: Evaluating, understanding and identifying argumentative essays generated by GPT models," 2023, *arXiv:2304.07666*.
22. Y. Li et al., "Can large language models write reflectively," *Comput. Educ., Artif. Intell.*, vol. 4, May 2023, Art. no. 100140, doi: 10.1016/j.caeai.2023.100140.
23. E. Mitchell, Y. Lee, A. Zharatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," 2023, *arXiv:2301.11305*.
24. "Educator considerations for ChatGPT," OpenAI, San Francisco, CA, USA, 2023. [Online]. Available: https://platform.openai.com/docs/chatgpt-education
25. M. Jovanović and M. Campbell, "Generative artificial intelligence: Trends and prospects," *Computer*, vol. 55, no. 10, pp. 107–112, Oct. 2022, doi: 10.1109/MC.2022.3192720.
26. J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," 2023, *arXiv:2301.10226*.
27. Z. Jiang, J. Zhang, and N. Z. Gong, "Evading watermark based detection of AI-generated content," 2023, *arXiv:2305.03807*.
28. B. Edwards, "China bans AI-generated media without watermarks," Ars Technica, Boston, MA, USA, Dec. 2022. [Online]. Available: https://arstechnica.com/information-technology/2022/12/china-bans-ai-generated-media-without-watermarks/
29. J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vision*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021, doi: 10.1007/s11263-021-01453-z.
30. C.-Y. Hsieh et al., "Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes," 2023, *arXiv:2305.02301*.

MARK CAMPBELL is the chief innovation officer at EVOTEK, San Diego, CA 92121 USA. Contact him at mark@evotek.com.

MLAĐAN JOVANOVIĆ is an associate professor at Singidunum University, 11000 Belgrade, Serbia. Contact him at mjovanovic@singidunum.ac.rs.