May 7, 2024

Understanding the source of what we see and hear online

We’re introducing new tools to help researchers study content authenticity and are joining the Coalition for Content Provenance and Authenticity Steering Committee.

audiovisual content becomes more common, we believe it will be increasingly important for society as a whole to embrace new technology and standards that help people understand the tools used to create the content they find online.

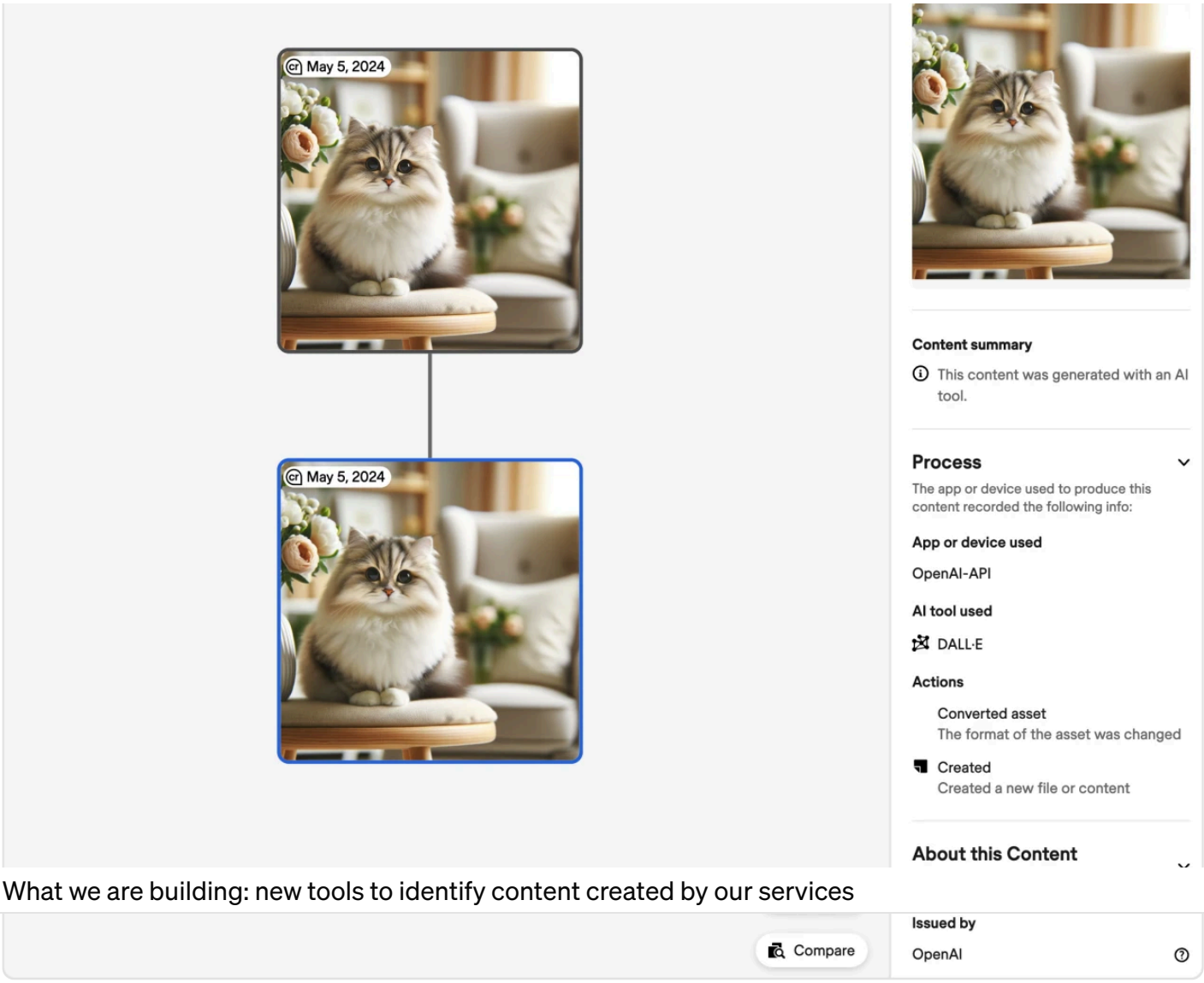
At OpenAI, we're addressing this challenge in two ways: first, by joining with others to adopt, develop and promote an open standard that can help people verify the tools used for creating or editing many kinds of digital content, and second, by creating new technology that specifically helps people identify content created by our own tools.

Contributing to authenticity standards

The world needs common ways of sharing information about how digital content was created. Standards can help clarify how content was made and provide other information about its origins in a way that's easy to recognize across many situations — whether that content is the raw output from a camera, or an artistic creation from a tool like DALL·E 3.

Today, OpenAI is joining the Steering Committee of C2PA – the Coalition for Content Provenance and Authenticity. C2PA is a widely used standard for digital content certification, developed and adopted by a wide range of actors including software companies, camera manufacturers, and online platforms. C2PA can be used to prove the content comes a particular source.[1] We look forward to contributing to the development of the standard, and we regard it as an important aspect of our approach.

Earlier this year we began adding C2PA metadata to all images created and edited by DALL·E 3, our latest image model, in ChatGPT and the OpenAI API. We will be integrating C2PA metadata for Sora, our video generation model, when the model is launched broadly as well.



What we are building: new tools to identify content created by our services

People can still create deceptive content without this information (or can remove it), but they cannot easily fake or alter this information, making it an important resource to build trust. As adoption of the standard increases, this information can accompany content through its lifecycle of sharing, modification, and reuse. Over time, we believe this kind of metadata will be something people come to expect, filling a crucial gap in digital content authenticity practices.

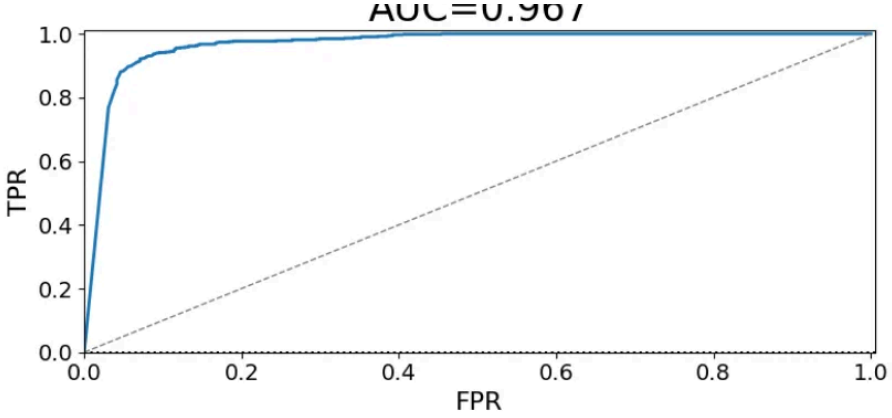
To drive adoption and understanding of provenance standards - including C2PA - we are joining Microsoft in launching a societal resilience fund. This $2 million fund will support

What we are building: new tools to identify content created by our services

In addition to our investments in C2PA, OpenAI is also developing new provenance methods to enhance the integrity of digital content. This includes implementing tamper-resistant watermarking – marking digital content like audio with an invisible signal that aims to be hard to remove – as well as detection classifiers – tools that use artificial intelligence to assess the likelihood that content originated from generative models. These tools aim to be more resistant to attempts at removing signals about the origin of content.

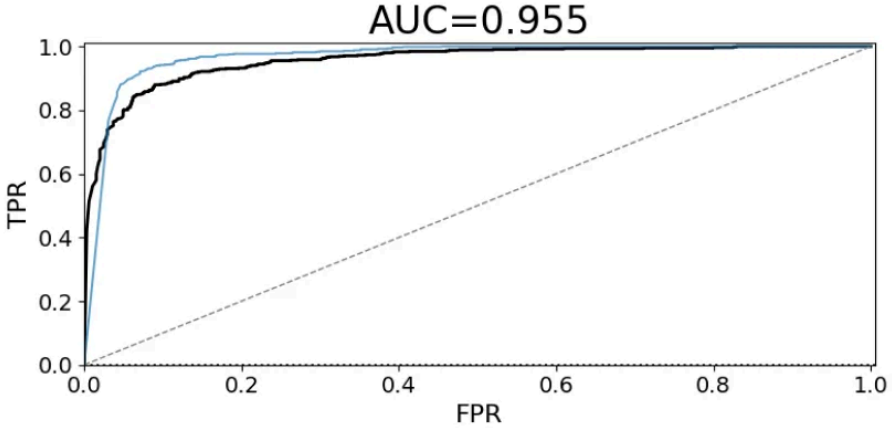
Starting today, we are opening applications for access to Open[illegible] classifier to our first group of testers - including research labs and research-oriented journalism nonprofits - for feedback through our Researcher Access Program. This tool predicts the likelihood that an image was generated by OpenAI's DALL·E 3. Our goal is to enable independent research that assesses the classifier's effectiveness, analyzes its real-world application, surfaces relevant considerations for such use, and explores the characteristics of AI-generated content. Applications for access can be submitted here.

Understanding when and where a classifier may underperform is critical for those making decisions based on its results. Internal testing on an early version of our classifier has shown high accuracy for distinguishing between non-AI generated images and those created by DALL·E 3 products. The classifier correctly identifies images generated by DALL·E 3 and does not trigger for non-AI generated images. It correctly identified ~98% of DALL·E 3 images and less than ~0.5% of non-AI generated images were incorrectly tagged as being from DALL·E 3. The classifier handles common modifications like compression, cropping, and saturation changes with minimal impact on its performance. Other modifications, however, can reduce performance. We also find that the performance of the classifier for distinguishing between images generated by DALL·E 3 and other AI models is lower and the classifier currently flags ~5-10% of images generated by other AI models on our internal dataset.

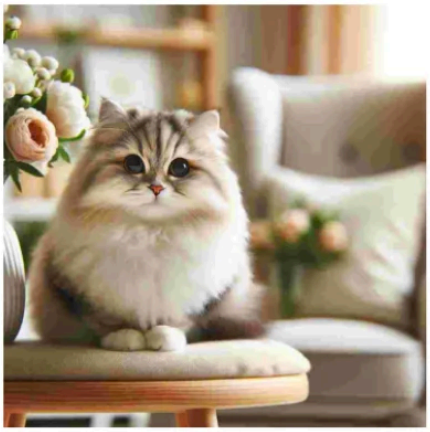


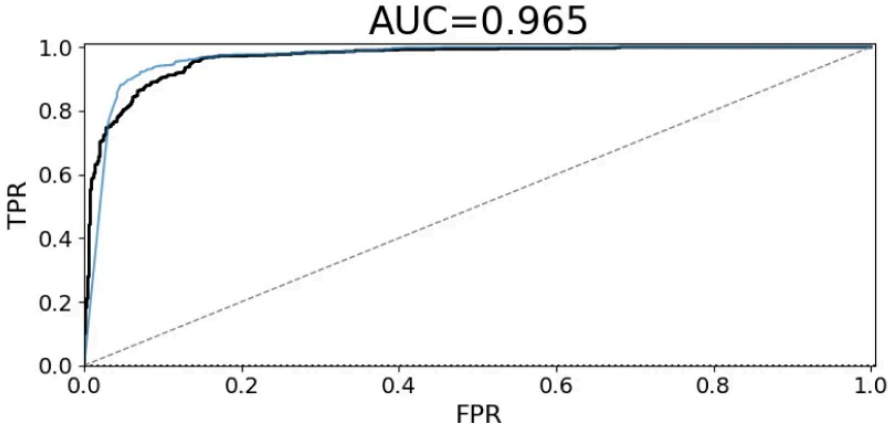
Augmentations that have minimal impact on classifier performance

Adjusted Saturation

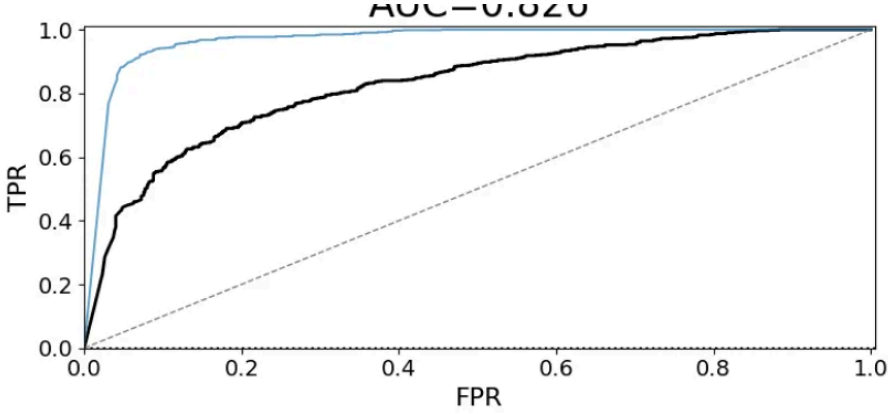


JPEG Recompression

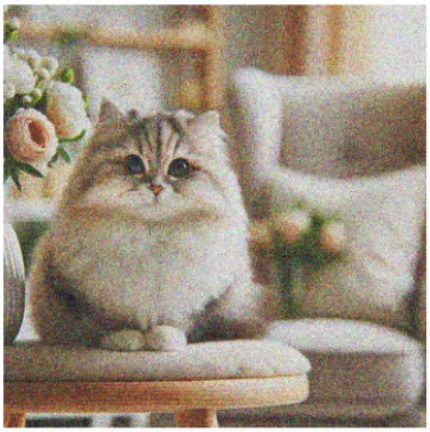


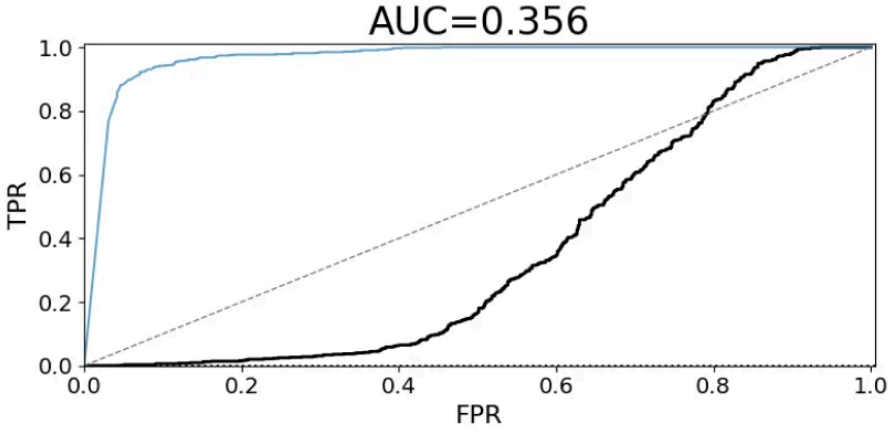


Augmentations that have meaningful impact on classifier performance



Adding Moderate Amounts of Gaussian Noise





In addition, we've also incorporated audio watermarking into Voice Engine, our custom voice model, which is currently in a limited research preview. We are committed to continuing our research in these areas to ensure that our advancements in audio technologies are equally transparent and secure.

What's next for content authentication

While technical solutions like the above give us active tools for our defenses, effectively enabling content authenticity in practice will require collective action. For example, platforms, content creators, and intermediate handlers need to facilitate retaining

Our efforts around provenance are just one part of a broader industry effort – many of our peer research labs and generative AI companies are also advancing research in this area. We commend these endeavors—the industry must collaborate and share insights to enhance our understanding and continue to promote transparency online.

DALL·E 3 Announcements Image Generation Responsible AI Robustness

Authors

OpenAI

Footnotes

1 This is done by attaching an encrypted attestation that the content comes from its tool of origin. ↵