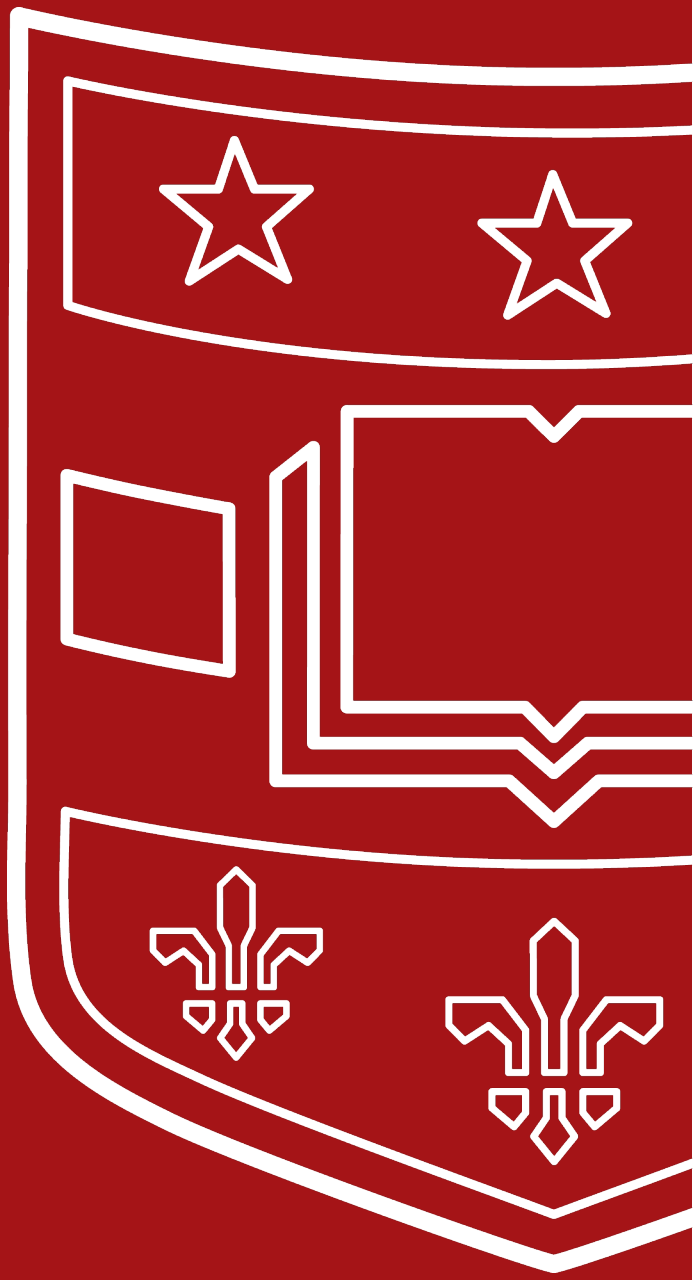
Navigating AI Detection Tools for Academic Integrity

Ruopeng An, PhD, MPP, FACE, FAAHB
Associate Professor
Faculty Lead in Public Health Sciences
Faculty Fellow for AI Innovations in Education
Brown School & Division of Computational & Data Sciences
Washington University in St. Louis



Acknowledgment

- Yuyi Yang, PhD student, DCDS
- Jennifer Smith, Vice Provost
- Sally Wu, Assistant Director for Educational Technology, CTL
- Kat Weir, Program Coordinator, CTL
- Michael Wysession, Executive Director, CTL
- Eric Fournier, Director of Educational Development, CTL

Introduction to AI Writing Detection Tools

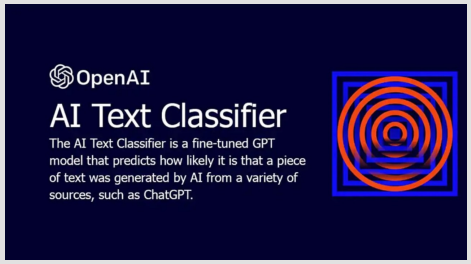
- **Emerging Challenges**: The accessibility of AI writing technologies introduces new hurdles in ensuring the authenticity of academic submissions.
- **Purpose**: Designed to identify text generated by AI, aiding educators in maintaining academic standards.
- **Functionality**: Utilize advanced algorithms to analyze writing patterns, syntax, and linguistic features.
- **Goal**: Differentiate between human-authored and AI-generated content to safeguard academic integrity.

Scribbr

turnitin

grammarly

GPTZero



ORIGINALITY.AI

COPYLEAKS

crossplag

Manual Detection of AI-Generated Writing

- **Monotonous Text**: Look for repetitive sentence structures and lengths.
- **Predictable Language**: Notice generic word choices without unique expressions.
- **Politeness**: AI, like ChatGPT, often uses very polite and formal language.
- **Hedging Phrases**: Look for a lack of bold, original statements and for a tendency to overuse generic hedging phrases: “It’s important to note that …” “X is widely regarded as …” “X is considered …” “Some might say that …”
- **Voice Inconsistency**: If you know the usual writing style and voice of the person whose writing you’re checking (e.g., a student), then you can usually see when they submit something that reads very differently from how they normally write.
- **Incorrect Citations**: AI may incorrectly cite sources or not cite at all.
- **Logical Errors**: Identify contradictions or implausible statements.

Deep Dive into Technology: Perplexity

- Measures the predictability of text.
- Higher perplexity indicates more creative and potentially human writing, while lower perplexity suggests AI generation due to its predictable nature.
- AI vs. Human Writing:
  - AI-Generated Text: Aims for **low perplexity**, producing content that's logical and smooth but predictable.
  - Human-Created Text: Exhibits **higher perplexity** with diverse language choices and occasional errors.

Deep Dive into Technology: Perplexity Example

- Language models work by predicting what word would naturally come next in a sentence and inserting it.
- For example, in the sentence "I couldn't get to sleep last …" there are more and less plausible continuations, as shown in the table below.

Example continuation	Perplexity
I couldn't get to sleep last **night.**	**Low:** Probably the most likely continuation
I couldn't get to sleep last **time I drank coffee in the evening.**	**Low to medium:** Less likely, but it makes grammatical and logical sense
I couldn't get to sleep last **summer on many nights because of how hot it was at that time.**	**Medium:** The sentence is coherent but quite unusually structured and long-winded
I couldn't get to sleep last **pleased to meet you.**	**High:** Grammatically incorrect and illogical

Deep Dive into Technology: Burstiness

- Measures **sentence** structure and length variation. (Perplexity is on the level of words).
- Low burstiness suggests AI authorship, as AI models often opt for the most likely, hence predictable, sentence structures.
- AI vs. Human Writing:
  - AI-Generated Text: Shows **low burstiness**, with uniform sentence lengths (say, 10–20 words) and conventional structures, resulting in monotonous writing.
  - Human-Created Text: Exhibits **high burstiness**, reflecting a wide range of sentence constructions and lengths, signaling creative and diverse writing styles.

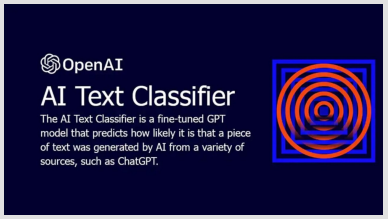
Scribbr GPTZero

Deep Dive into Technology: Burstiness Example

- **High** burstiness:
  - At times, you might see a sentence that's short. Abrupt. Then, suddenly, there emerges a complex and winding sentence, one that weaves through comas, dashes, and semicolons; a sentence that takes you on a journey through vivid imagery and elaborate ideas, winding up in unexpected places and pausing only briefly before the next brief statement. So it goes.
- **Low** burstiness:
  - Every sentence in this text is uniform. The sentences are simple. They are brief. They follow a predictable pattern. There is little variation in length or complexity. The content is straightforward. This consistency continues throughout the paragraph. It maintains a steady rhythm. It's easy to follow.

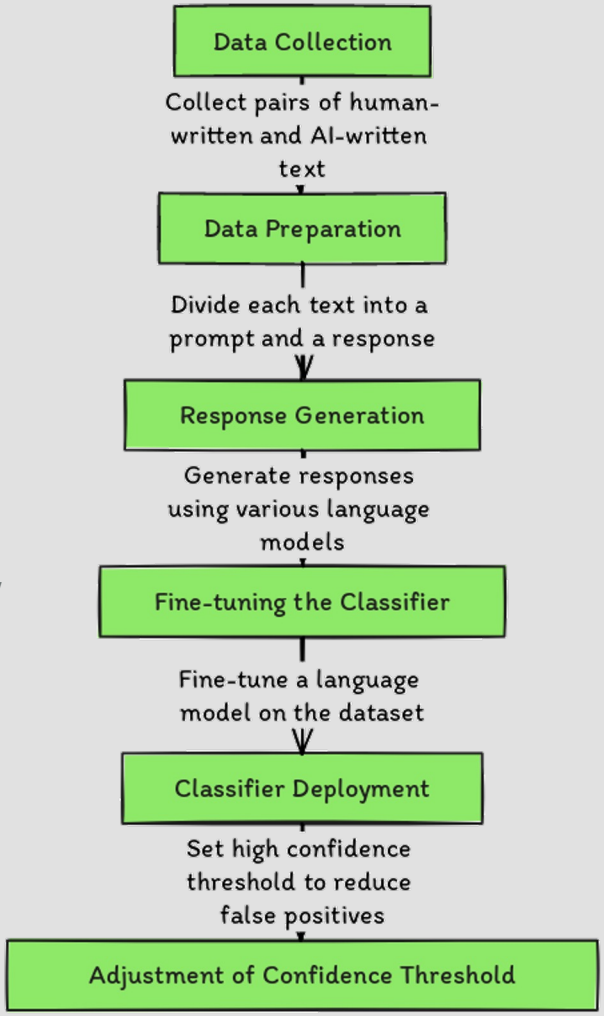
Deep Dive into Technology: Language Model Foundation

- Utilizes a language model fine-tuned on datasets of both human and AI-written texts on the same subjects.
- **Data Collection**: Compiled from human sources including pretraining data and prompts.
- **Text Analysis Approach**:
  - Divides texts into prompts and responses, generating AI responses for comparison.
  - The blend of generator and discriminator models in training
- **Confidence Threshold Adjustment**: Ensures low false positive rates by marking texts as AI-written only when very confident.



crossplag™

ORIGINALITY.AI



Deep Dive into Technology: Text Segments

- Breaks down submissions into overlapped text segments to analyze contextually.
- **AI Detection Process**: Evaluates segments against a model to estimate AI authorship.
- **Scoring System**: Calculates an average AI score for the entire document's likelihood of being AI-written.
- **Underlying Technology**: Utilizes transformer models, similar to predictive text, to recognize AI-generated patterns by comparing the statistical likelihood of word sequences.

turnitin

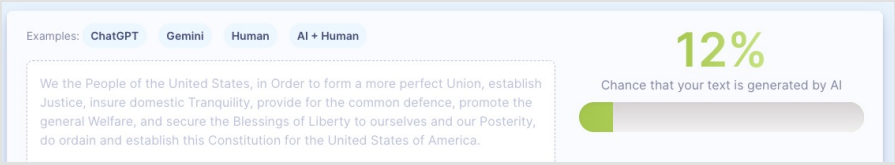
Deep Dive into Technology: Text Segments Example

- **Segmentation**: For instance, if the essay starts with "Artificial Intelligence has revolutionized the way we interact with technology. Its impact on society is profound, offering new opportunities and challenges." The tool might first analyze the sentence "Artificial Intelligence has revolutionized **the way we interact with technology**."
- **Analysis**
- **Overlap and Context**: Next, the tool might analyze a segment that slightly overlaps with the first, such as "**the way we interact with technology**. Its impact on society is profound," to ensure context is maintained and the analysis captures the flow of writing.
- **Scoring Each Segment**
- **Aggregation**

Free AI Writing Detection Tools Link

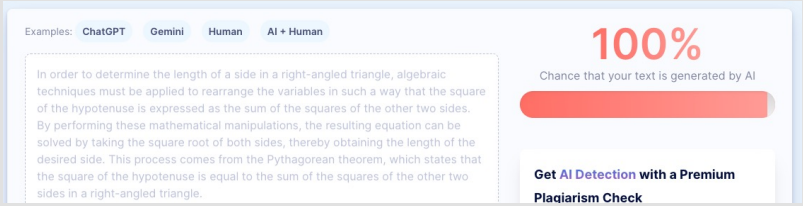
- https://www.scribbr.com/ai-detector/
- https://gptzero.me/
- https://copyleaks.com/ai-content-detector
- https://hivemoderation.com/ai-generated-content-detection
- https://contentatscale.ai/ai-content-detector/
- https://www.zerogpt.com/

Example



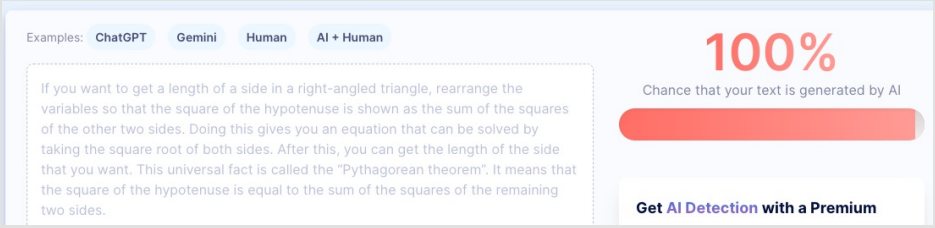
- Preamble to the United States Constitution
  - We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common defence, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our Posterity, do ordain and establish this Constitution for the United States of America.

Edge Case



- Something as simple as stating the Pythagorean Theorem is automatically detected as AI-generated.
  - In order to determine the length of a side in a right-angled triangle, algebraic techniques must be applied to rearrange the variables in such a way that the square of the hypotenuse is expressed as the sum of the squares of the other two sides. By performing these mathematical manipulations, the resulting equation can be solved by taking the square root of both sides, thereby obtaining the length of the desired side. This process comes from the Pythagorean theorem, which states that the square of the hypotenuse is equal to the sum of the squares of the other two sides in a right-angled triangle.

Edge Case



- To verify this, the language was simplified and made to sound casual using the following prompt:
  - If you want to get a length of a side in a right-angled triangle, rearrange the variables so that the square of the hypotenuse is shown as the sum of the squares of the other two sides. Doing this gives you an equation that can be solved by taking the square root of both sides. After this, you can get the length of the side that you want. This universal fact is called the "Pythagorean theorem". It means that the square of the hypotenuse is equal to the sum of the squares of the remaining two sides.

Edge Case

- After asking ChatGPT to say this like a “caveman”:
  - To find side, in big triangle with stick-straight angle, make big line across — hypotenuse — have square equal add-up squares of stick-leg sides. Take square roots for answer. Get side length you look for. Old, always-true rule, "Pythagorean theorem", say big line across square always same as stick-leg sides square put together.

You

Say this like a “caveman”:

If you want to get a length of a side in a right-angled triangle, rearrange the variables so that the square of the hypotenuse is shown as the sum of the squares of the other two sides. Doing this gives you an equation that can be solved by taking the square root of both sides. After this, you can get the length of the side that you want. This universal fact is called the “Pythagorean theorem”. It means that the square of the hypotenuse is equal to the sum of the squares of the remaining two sides.

ChatGPT

To find side, in big triangle with stick-straight angle, make big line across — hypotenuse — have square equal add-up squares of stick-leg sides. Take square roots for answer. Get side length you look for. Old, always-true rule, "Pythagorean theorem", say big line across square always same as stick-leg sides square put together.

Reliability of AI Detectors

- **Performance**: AI detectors work well with longer texts but may struggle with unpredictable AI outputs or edited texts.
- **Accuracy Levels**: Premium tools show up to 84% accuracy, while free tools offer around 68%.
- **Use as Indicators**: They provide helpful indications of AI-generated content *but shouldn't be the sole evidence*.
- **Ongoing Development**: As AI evolves, detection tools continuously adapt, facing challenges in maintaining pace with new language models.

https://www.scribbr.com/ai-tools/how-do-ai-detectors-work/

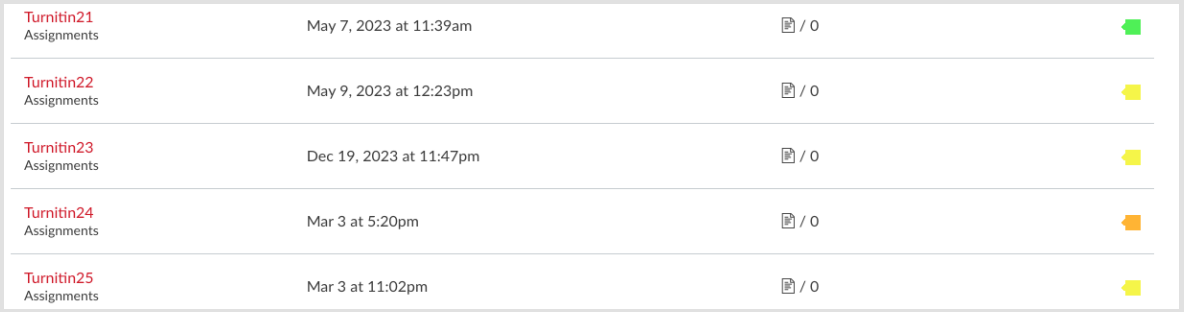
AI Detectors vs. Plagiarism Checkers

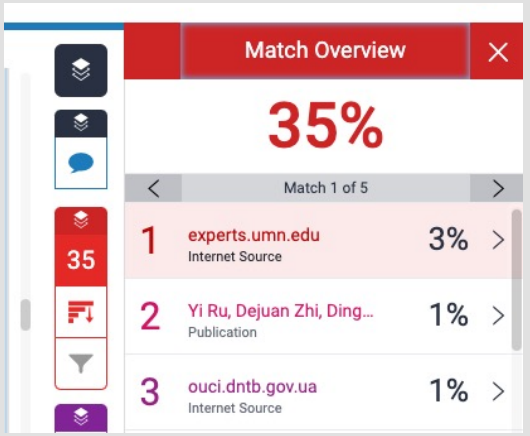
- **AI Detectors**: Identify AI-generated text by analyzing text characteristics without database comparison.
- **Plagiarism Checkers**: Detect copied text by comparing it against a vast database of sources to find similarities.
  - Example: turnitin
    - **Database: Student Papers**: Archives of submitted assignments from various educational institutions; **Academic Publications**: Access to journals, articles, and conference proceedings; **Web Content**: Broad internet sweep, including websites, blogs, news portals, and public repositories.
    - **Fingerprinting Technology**: Advanced method to detect unique content patterns and paraphrasing.
    - **Constant Updates**: Regularly expanded to include new and emerging sources, maintaining relevance and accuracy.

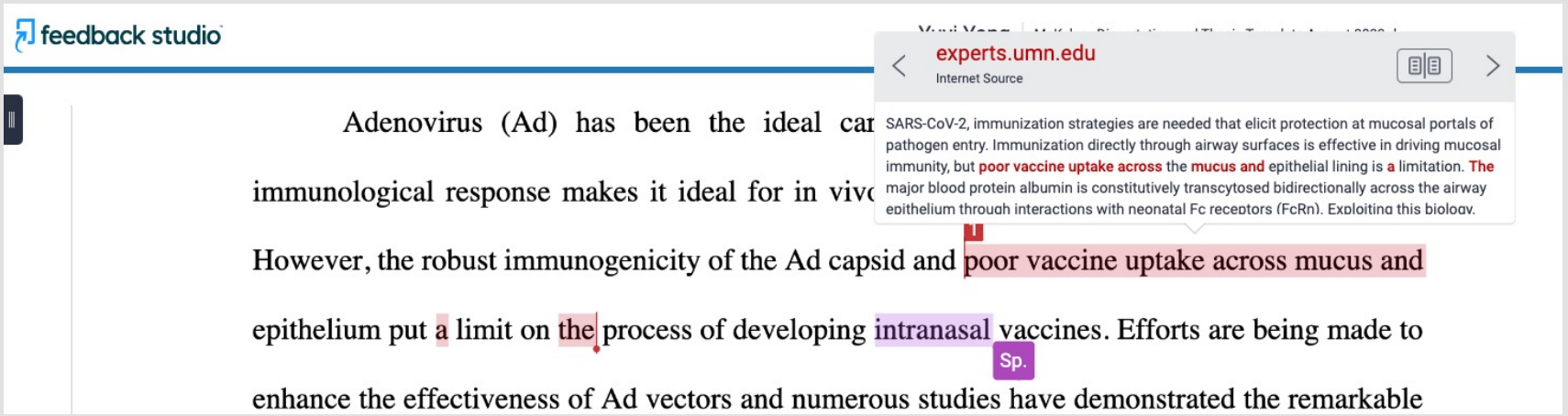
https://www.turnitin.com/products/features/database-content

Example: Identifying Plagiarism with Turnitin

- Submission-Scanning-Similarity Report-Review-Assessment







The Intersection: AI-Generated Text and Plagiarism Detection

- **Plagiarism checkers might flag AI-generated text as it draws from uncredited sources.** While it usually generates original sentences, it may also include sentences directly copied from existing texts, or at least very similar.
  - This is most likely to happen with *popular or general-knowledge topics* and less likely with *more specialized topics* that have been written about less.
- As more AI-generated text appears online, AI writing may become **more likely** to be flagged as plagiarism—simply because other similarly worded AI-generated texts already exist on the same topic.

https://www.scribbr.com/ai-tools/how-do-ai-detectors-work/

Trends: Watermarks Techniques

- **Concept**: OpenAI is exploring a watermarking system for texts generated by AI, like ChatGPT, to invisibly mark content.
- **Current Status**: The technology is under development, with operational details still unclear. (However, images generated by DALL-E 3 already carry the watermark since Feb 2024.)
- **Challenges**: Uncertainty exists on whether watermarks would persist through edits of the generated text.

Detecting AI-Generated Images and Videos (Without Watermark)

- **Emerging Generators**: Tools like DALL-E, Midjourney, and Sora are creating realistic images and videos.
- **Current Giveaways**: Look for anatomical errors, unnatural movements, and nonsensical text in AI creations.
- **Detection Tools**: ***Deepware***, Intel's ***FakeCatcher***, and ***Illuminarty*** offer solutions, though their reliability is yet to be fully assessed.
  - ***Deepware*** scans for AI-specific artifacts (inconsistencies in textures, unusual edges, distortion in images...) and inconsistencies in digital content.
  - ***FakeCatcher*** analyzes the subtle blood flow in video pixels to distinguish between real and synthetic faces.
  - ***Illuminarty*** focuses on identifying irregularities in lighting and shadows that are typical in AI-generated images.
- **Challenge Ahead**: As technology improves, manual detection will become increasingly difficult, underscoring the need for advanced detection tools.

Examples: Detecting AI-Generated Images

"Double Harry" portrait

- If a famous person's name was used as a term in the prompt, sometimes the AI will duplicate them (especially in a blurred background).

https://medium.com/the-generator/15-tips-to-tell-if-an-image-is-ai-art-or-not-2ae703d0ac36

Examples: Detecting AI-Generated Images

- Non-circular pupil: AI can’t do circles. The albedo of the pupil is difficult to replicate.
- The uninterrupted block of teeth: “AI smiles” can appear like the whole row of teeth was been capped in one action.
- The clothing that doesn’t make sense: the random red bow.
- The impossible sails and rigging on the schooner.

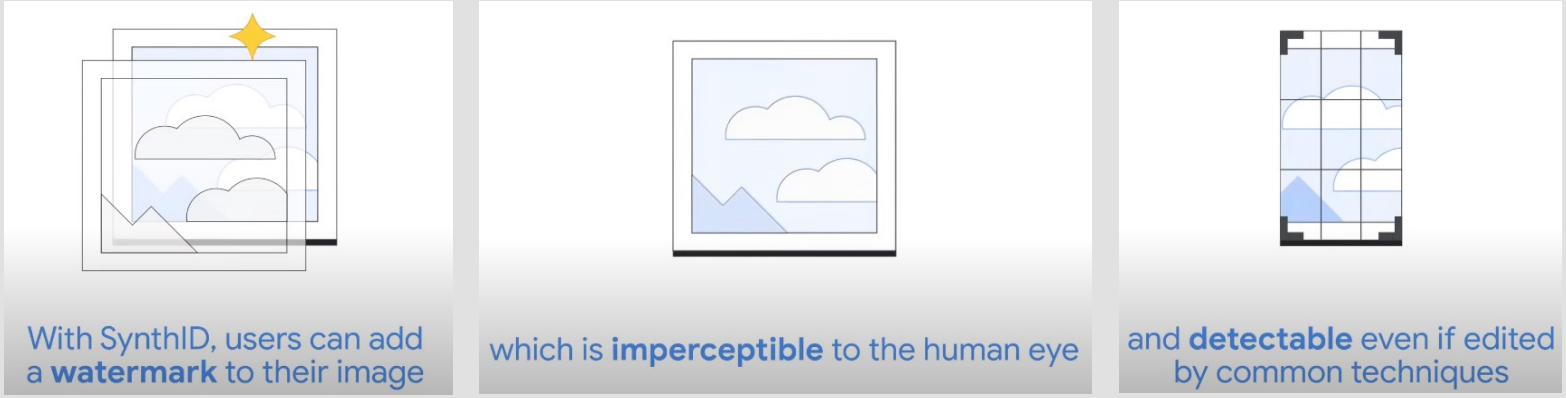
https://medium.com/the-generator/15-tips-to-tell-if-an-image-is-ai-art-or-not-2ae703d0ac36

Examples: Detecting AI-Generated Images

- A realistic image that won the Sony World Photography Awards.
- The fingers: which have mismatched lengths and fingernails.
- The unlikely drape of the costume: which defies fabric bias and no selvage to the sleeve.
- The poorly defined ears.
- The eyes: there's a flat edge to the right iris, and the spark is wrong.

https://medium.com/the-generator/15-tips-to-tell-if-an-image-is-ai-art-or-not-2ae703d0ac36

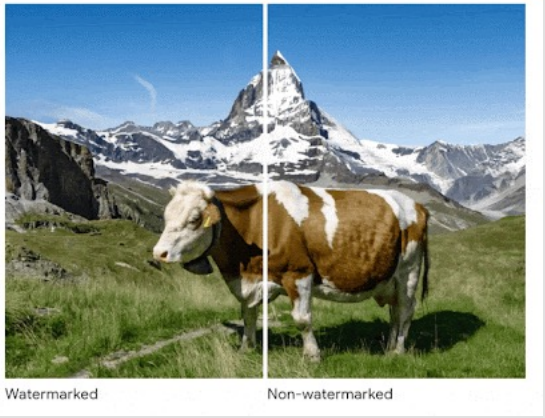
Detecting AI-Generated Images and Videos (With Watermark)

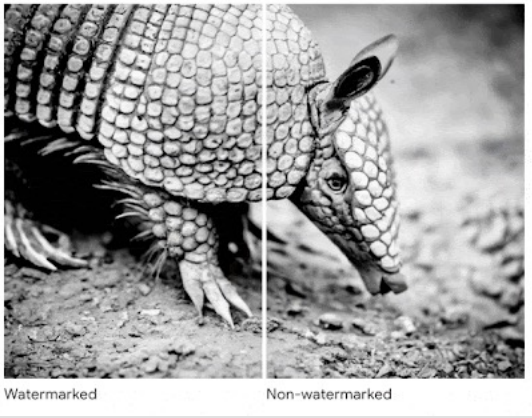


- **Subtle pixel manipulation** involves making minor adjustments to the color values of certain pixels within an image to embed a unique pattern or code.
- This process does not significantly alter the visual content of the image, allowing it to **retain its original appearance to the naked eye**.
- The manipulation is done in such a way that **the pattern can be detected and decoded** by specialized software designed to recognize these watermarks.

https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/

Detecting AI-Generated Images and Videos (With Watermark)

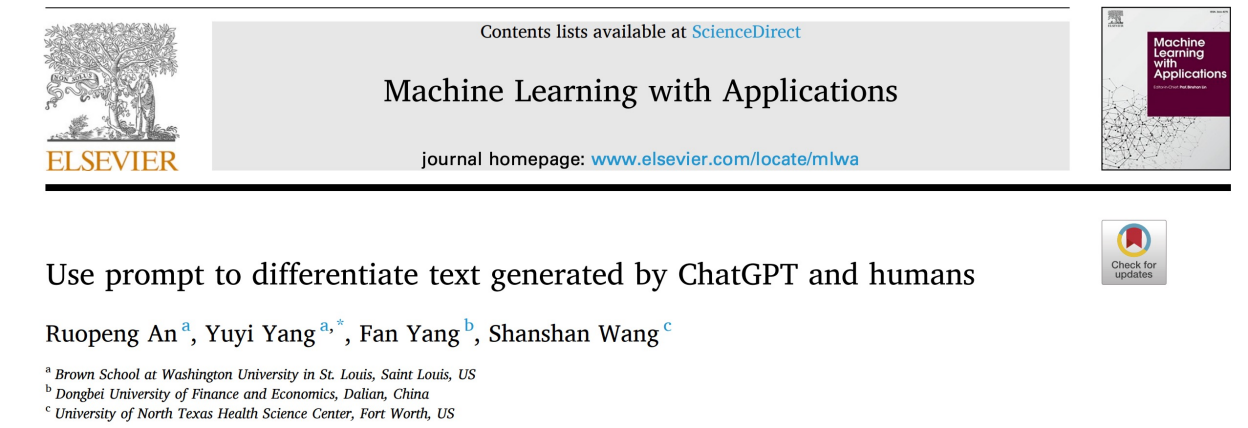




- This method typically leverages the vast number of pixels in digital images to **hide the watermark in the noise level of the image**, which is imperceptible to human viewers but can be extracted algorithmically.
- The pixel changes are usually **spread across the image** to ensure that the watermark remains intact even if the image is cropped or resized.
- This technique ensures the watermark's durability and retrievability without affecting the image's aesthetic and informational value.

https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/

Introducing the AI Detection Tool Developed by Our Team



Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Use prompt to differentiate text generated by ChatGPT and humans

Ruopeng An [a], Yuyi Yang [a,*], Fan Yang [b], Shanshan Wang [c]

[a] Brown School at Washington University in St. Louis, Saint Louis, US
[b] Dongbei University of Finance and Economics, Dalian, China
[c] University of North Texas Health Science Center, Fort Worth, US

- Uses both prompts and essays to distinguish between machine-generated and human-written text.
- Combines similarity scores within machine-generated essays and between human and machine-generated essays to gauge AI authorship likelihood.
- Achieved an AUC score of 0.991 with low false positive and negative rates in tests. Outperforms traditional text passage-based approaches.

https://doi.org/10.1016/j.mlwa.2023.100497

AI Sounds Like a Human…

- AI Humanizer Pro is one of the most popular GPTs in GPT 4.0. They claim to be the “Best AI humanizer to help you get 100% human score. Humanize your AI-generated content to bypass AI detection. Use our advanced humanizer to get past all AI detectors in the market”.

- Contentatscale.ai: “Want Undetectable AI Content? Then AI Humanizer Has You Covered! The AI Humanizer uses a proprietary mix of 3 AI engines, NLP and semantic analysis algorithms to rewrite sentences and paraphrase paragraphs in a way that is so humanlike it bypasses even the toughest and most accurate AI detection tools.”

https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/

Main Limitations of AI Detection Tools

- **Highly Edited Content**: Struggles to detect AI text that's extensively edited or paraphrased.
- **Evolving AI Models**: AI writing tools improve rapidly, outpacing detection methods.
- **Language and Style Specificity**: Detection accuracy varies across languages and writing styles.
- **False Positives/Negatives**: Risk of incorrectly identifying AI-generated or human-written text.
- **Generalization vs. Specificity**: Balancing broad applicability with the need for targeted detection strategies.
- **Data Privacy Concerns**: Ensuring user data privacy while analyzing text.
- **Computational Demands**: High resource requirements for processing large datasets or complex models.
- **Bias in Training Data**: Potential biases in the data used to train detection models.
- **User Trust**: Convincing users of the tool's reliability and necessity.

Future Directions for AI Detection Tool

- **Enhanced NLP Models**: Continuous improvements in natural language processing models for better accuracy in distinguishing AI-written content.
- **Behavioral Analysis**: Using AI to analyze writing patterns and behaviors indicative of automated content generation.
- **Cross-referencing Sources**: Tools might start incorporating cross-referencing functionalities to detect AI content by checking against known AI-generated data repositories.
- **Adaptive Learning Systems**: AI detection tools that continually update their detection algorithms based on new AI content generation methods to stay ahead of evolving technologies.
- Expansion to Different Languages, Integration with Educational Platforms, Improvement in Short Text Detection, and User Interface Development...

Beyond Binary Detection: Embracing Complexity in AI Collaboration

- **Rethinking Detection**: Acknowledging that a 90% AI-generated score does not definitively indicate AI authorship. AI and humans often collaborate, blending creativity and computational power.
- **The Blurred Lines**: As AI integration deepens, distinguishing between human and AI contributions becomes increasingly complex. Cases of joint brainstorming by humans and AI highlight the evolving nature of content creation.
- **Redefining Contribution**: Future tools may need to assess the extent of AI involvement rather than merely detecting AI presence. This involves recognizing the iterative process of human-AI collaboration, where ideas are refined collaboratively.
- **Beyond Percentages**: Moving away from simplistic percentage-based assessments to understand the nuanced role of AI in content generation.
- **Evolving Perspectives**: OpenAI's discontinuation of their detection tool due to accuracy concerns suggests that ***strict detection*** may not be the ultimate goal. The focus should shift towards how AI can enhance educational and creative processes, acknowledging the multifaceted impact of AI integration.

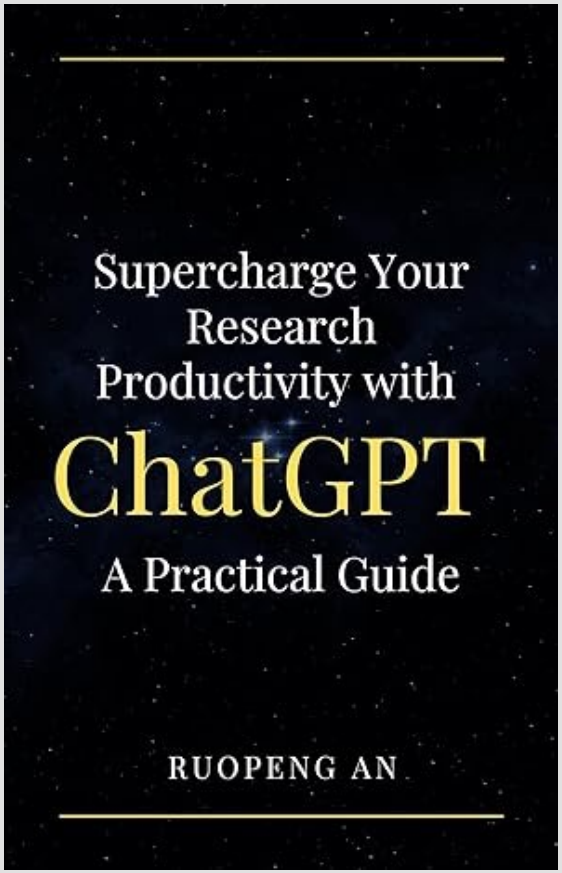
AI Learning Resources

https://aicademe.publish.library.wustl.edu

- AI Certificate Programs
- AI Workshop Series
- Open Classroom Talks
- AI Interest Group

…

Supercharge Your Research Productivity With ChatGPT: A Practical Guide



Learn AI in Fall 2024: Zero to Hero

Program Content (Weeks 1 and 2)

- An overview of AI
- Learn how to code in Python
- Use NumPy and Pandas to do data wrangling
- Use Matplotlib to do data visualization

Program Content (Weeks 3-7: Machine learning)

- Classification tasks
- Regression tasks
- Model training and validation
- Support vector machines
- Decision trees
- Ensemble methods (e.g., random forest, XGBoost)
- Dimensionality reduction
- Unsupervised learning

Program Content (Weeks 8-15: Deep learning)

- Neural network basics
- Computer vision: Convolutional neural network, Image classification, Object detection, Image segmentation
- Natural language processing: Transformers, Sentiment analysis, Named entity recognition, Natural language generation, Text summarization, Question-answering, Text translation, Chatbot
- Recommender system
- Time series forecasting
- Synthetic data generation

Case Study Examples

Participants will use AI tools to solve many challenging real-world problems, for example:

- Diagnosing pneumonia from chest x-rays
- Monitoring proper mask-wearing in pubic spaces
- Designing a question-answer chatbot for COVID-19 symptoms
- Identifying tweets about natural or human-made disasters
- Creating sharable synthetic datasets from confidential medical records

and Many More!