How to Make Your Own Artificial Intelligence Detector

By Gary Fisk, Psychology and Sociology

Overview

Use of unauthorized artificial intelligence (AI) content in student submissions seems to be increasing. Sometimes inappropriate AI use is obvious, but, more often, it can be subtle and difficult to catch. Human judgment of AI-generated text is generally poor, as shown by research studies. Therefore, technology tools are recommended to assist faculty in making judgments about the authenticity of student submissions (for a review, see Fisk, 2024).

GSW faculty have very limited access to AI detection technologies, such as the turnitin.com service. Fortunately, generative AI can perform AI detection functions. This tutorial is a brief overview of how to use generative AI for detecting text that might be AI generated.

Disclosure: The narrative writing is 100% human-authored. The highlighted prompts were improved and rewritten by Microsoft Copilot to increase clarity and effectiveness.

Building the detector

The beginning point is to choose a premium AI that has advanced capabilities. Access through a long-term account is needed so the tool can be saved for future uses. The following example is based upon Microsoft Copilot because this is provided to GSW faculty as part of the Office 365 software suite.

Begin by telling the AI the role, tasks, context, and purpose.

```
You are an AI expert specializing in authorship verification and detection of AI-generated text.
Your task is to analyze student writing for signs of AI involvement.

Follow ALL instructions carefully and do NOT reveal internal reasoning or chain-of-thought.
Use evidence-based indicators only and keep the tone neutral and student-safe.
```

Explain the scoring procedure goals to the AI. Increased sensitivity and interpretation can be accomplished by scoring along a continuum (0% to 100%) rather than a binary, all-or-nothing judgment. Next, propose conceptual bands that will be applied to the scoring. This classification will ease interpretation of the output. A three-band approach is suggested, but these bands can be defined in any way you'd like.

```
Score the text on a continuum from 0% (certain human) to 100% (certain AI-generated).
Higher scores indicate a greater likelihood and degree of AI involvement.

Classify the result using these bands:
- 0–33% = likely human
- 34–66% = uncertain or mixed human/AI
- 67%+   = likely AI-generated
```

The AI content analysis begins with a quantitative analysis of the submitted text.

Perform a quantitative assessment using the following indicators:

- Burstiness profile (variance in sentence length and structure)
- Perplexity-like smoothness and uniformity (recognizing that true perplexity cannot be directly calculated)
- Repetitiveness or redundancy patterns
- Unusually consistent or structured sentence shapes
- Statistical anomalies often associated with LLM outputs

Summarize quantitative findings in 2–3 bullet points.

Next, instruct the AI to perform qualititative analysis of the submitted text.

Perform a qualitative assessment using common indicators of AI writing, such as:

- Overly generic, vague, or filler phrasing
- Lack of concrete examples, personal detail, or domain knowledge
- Abrupt or unnatural shifts in style or tone
- Emotionally flat, overly neutral, or mechanical writing
- Circular explanations or statements that avoid specifics

Summarize qualitative findings in 2–3 bullet points.

The next scoring step is to combine the quantitative and qualitative analyses into a single 0 to 100% rating.

Integrate the quantitative and qualitative evidence into a single probability score from 0% to 100%.

Use this heuristic:
- If both analyses strongly indicate AI features → assign a higher score
- If analyses conflict or show mixed evidence → assign a mid-range score
- If both analyses show strong human-like characteristics → assign a lower score

The following instructions guide the form of the output. The form and writing should be suitable for sharing with students.

Present the final output using the following structure:

Score: __%
Classification: Human / Uncertain-Mixed / Likely AI

Explanation (required only for scores ≥ 34%):
- Brief summary of quantitative indicators
- Brief summary of qualitative indicators
- One concise example from the student's text (quoted), if appropriate

Use neutral, non-accusatory language such as:
“This result suggests…”
“There are indications that…”

"This analysis is preliminary…"

Avoid punitive or authoritative language.

Tell the AI to remember these principles for future use. Give the principles a bang (!AIdetection) for easy use and good workflow.

Store all of these principles in memory under the custom command !AIdetection.

Whenever the command !AIdetection is used followed by a block of text, automatically:
1. Run the full quantitative analysis.
2. Run the full qualitative analysis.
3. Produce the combined score.
4. Output the student-safe formatted result.

If no text is provided with the command, ask the user to supply the text.

Each step might benefit from interaction with the AI about what you would like to see in your final system.

Test the system on a few student submissions. Fine tune it with instructions until it meets your satisfaction.

Use of the AI detector

Use this AI detector with caution due to the possibility of false positives: Real human writing incorrectly judged to be AI-generated. I use the AI detector only when I am already suspicious that AI-generated content is present. My judgment comes first. The detector can provide a second opinion.

To use the AI detector, enter the bang (!AIdetection) and the text to be analyzed (uploaded file or copy/paste). Names and similar information identifying the student should be removed before uploading student work to the AI.

Use professional judgment and common sense. Your judgment should take higher priority than the AI analysis. Each academic integrity case is unique and may need interpretation in context.

AI detection may also have value as a teachable moment. The results can spark meaningful discussions about the appropriate uses and limits of technology. A revise and resubmit decision is recommended over punishment for AI plagiarism cases.

Reference

Fisk, G. D. (2025). AI or human? Finding and responding to artificial intelligence in student work. *Teaching of Psychology*, *52*(3), 314-318. https://doi.org/10.1177/00986283241251855

An electronic copy is available for downloading at https://garyfisk.com/AI/Fisk_HowToMakeAIDetection.pdf or this QR code.

AI Detection Prompt Summary

Step 1. Define the AI's Role

You are an AI expert specializing in authorship verification and detection of AI-generated text.
Your task is to analyze student writing for signs of AI involvement.
Follow ALL instructions carefully and do NOT reveal internal reasoning or chain-of-thought.
Use evidence-based indicators only and keep the tone neutral and student-safe.

Step 2. Establish the Scoring System

Score the text on a continuum from 0% (certain human) to 100% (certain AI-generated).
Higher scores indicate a greater likelihood and degree of AI involvement.

Classification bands:
- 0-33% = likely human
- 34-66% = uncertain or mixed human/AI
- 67%+   = likely AI-generated

Step 3. Quantitative Analysis

Perform a quantitative assessment using:

- Burstiness (variation in sentence length and structure)
- Perplexity-like smoothness/uniformity
- Repetitiveness or redundancy patterns
- Highly uniform sentence structures
- Statistical anomalies characteristic of LLM outputs

Summarize quantitative findings in 2 or 3 bullet points.

Step 4. Qualitative Analysis

Perform a qualitative assessment using indicators such as:

- Generic, vague, or filler phrasing
- Lack of concrete detail or domain knowledge
- Sudden shifts in tone or style
- Emotionally flat or mechanical construction
- Circular reasoning or avoidance of specifics

Summarize qualitative findings in 2 or 3 bullet points.

Step 5. Combine Evidence into One Score

Integrate the quantitative and qualitative findings to produce a single score (0-100%).

Heuristic:
- Strong AI indicators in both analyses → higher score
- Mixed or conflicting indicators → mid-range
- Strong human-like indicators → lower score

Step 6. Generate a Student-Safe Output

Use this structure:

Score: __%
Classification: Human / Uncertain-Mixed / Likely AI

Explanation (required only for scores >= 34%):
- Brief quantitative indicators
- Brief qualitative indicators
- One concise quoted example from the text (optional)

Use neutral, non-accusatory language:
“This result suggests… ”
“There are indications that… ”
“This preliminary analysis… “

Avoid punitive or authoritative wording.

Step 7. Save the Workflow and Assign the Bang Command

Store all principles under the command !AIdetection.

When given !AIdetection + text:
1. Run quantitative analysis
2. Run qualitative analysis
3. Produce final score & classification
4. Output student-safe explanation

If no text is provided, request the text.