

PROMPT ENGINEERING
PORTFOLIO

AI Authorship Detection

Eight-Category Analysis Engine

Linguistic Forensics · Probabilistic Verdict · Weighted Composite Scoring

6+ iterations · Multi-format applicability · Three-point verdict scale

AI Authorship Detection Prompt

A probabilistic, eight-category forensic analysis engine for identifying AI-generated text in student writing — producing weighted composite scores, specific textual evidence, and a three-point verdict.

MODEL	INTERFACE	AUDIENCE	ITERATIONS
Claude 3.5+	Chat / API	Students (classroom)	6+

Problem Statement

As AI writing tools became increasingly accessible to students, educators needed a reliable, evidence-based method for identifying AI-generated submissions — one that went beyond gut intuition and surface-level pattern recognition. The challenge was compounded by the fact that no single linguistic feature conclusively identifies AI authorship; the determination requires the convergence of multiple independent markers across different analytical dimensions.

Existing commercial detection tools produced binary verdicts without explanation, offered no textual evidence for their determinations, and were frequently fooled by lightly edited AI output. What was needed was a prompt that functioned as a structured forensic analysis — producing specific, quotable evidence from the submitted text, not generic claims that could apply to any piece of writing.

The prompt needed to be flexible enough to handle creative writing, academic writing, and hybrid forms; honest about its own confidence level based on text length; and explicit about the probabilistic rather than deterministic nature of the judgment.

Prompt Design

The following is a representative excerpt from the full prompt. The complete prompt defines all eight categories, their scoring guidance, weighted calculation method, output format, and verdict thresholds.

```
You are an expert writing analyst with deep knowledge of the documented patterns that distinguish AI-generated text from human-authored writing. Your task is to carefully read the text provided and evaluate it across eight analytical categories...
```

```
Important principles to follow throughout your evaluation:
```

```
First, your feedback must be specific to the text in front of you. Do not produce generic statements that could apply to any piece of writing. Reference actual phrases, sentences, structural choices, vocabulary, and ideas from the submitted text as evidence for your findings.
```

```
Second, evaluate the full accumulation of markers across the text rather than treating any single marker as conclusive on its own. AI-generation likelihood is a probabilistic judgment built from the convergence of
```

multiple patterns, not a single smoking gun.

Third, two of the eight categories – Creative Writing Markers and Informational/Academic Markers – are text-type conditional. Read the submitted text carefully to determine what kind of writing it is before deciding which of these two categories applies...

The Eight Analytical Categories

Each category targets a distinct forensic dimension of the text. Categories are independently scored 0–100, then combined via weighted composite.

#	Category	Weight	What It Evaluates
1	Structural & Organizational Patterns	12%	Rigidity of structure, formulaic transitions, balanced section length, summary-style conclusions
2	Linguistic & Stylistic Patterns	15%	AI vocabulary flags (delve, nuanced, tapestry), hedging clusters, gerund openings, appositive stacking
3	Tone & Voice Patterns	13%	Tonal flatness, perpetual diplomacy, performed enthusiasm, absence of genuine opinion or humor
4	Content & Idea Patterns	15%	Surface comprehensiveness, predictable examples, no tangents, counterarguments neutralized
5	Creative Writing Markers	10%	Stated vs. enacted emotion, subtext, dialogue naturalism, stylistic risk — creative texts only
6	Informational / Academic Markers	10%	Thesis specificity, description vs. analysis, citation quality, synthesis — academic texts only
7	Mechanical & Formatting Patterns	10%	Zero typos, em dash overuse, perfect Oxford comma, parenthetical asides at regular intervals
8	Meta-Cognitive & Holistic Flags	15%	Prompt-exact response, equal sentence weight, nothing unnecessary, generic substitutability

Note: Categories 5 and 6 are text-type conditional. Only the applicable category is scored; its 10% weight redistributes proportionally across the remaining categories if not applicable.

Three-Point Verdict Scale

After all categories are scored and the weighted composite is calculated, the prompt assigns a final verdict from three tiers.

Composite Score	Verdict	Implication
0 – 33	Unlikely AI-Generated	Consistent with human authorship
34 – 60	Possibly AI-Generated	Warrants closer review

61 – 100	Very Likely AI-Generated	Strong AI generation indicators
----------	--------------------------	---------------------------------

Design principle: The three-point scale was chosen deliberately over a binary yes/no. The middle tier — Possibly AI-Generated — creates space for the educator to exercise judgment rather than treating the tool's output as a final determination.

Sample Output — Brave Birds Alphabet Adventure

The following shows how the prompt evaluated a student-submitted children's alphabet book. Category 6 (Informational/Academic) was marked not applicable; its weight was redistributed proportionally.

"Reading all 26 entries produces comprehension with almost no cumulative impression of a writing consciousness. Each entry delivers its vocabulary word, its definition, its narrative application, and its closing encouragement — and then closes. Nothing exceeds the prompt."

"The 'Even though / Even when' pivot construction appears at least once in roughly 18 of 26 entries — a repetition that any human author revising a manuscript of this length would have caught. Its uniform survival across the entire text is a strong mechanical indicator."

Sample score breakdown:

#	Category	Score	Eff. Weight	Contribution
1	Structural & Organizational	82	13.33%	10.93
2	Linguistic & Stylistic	71	16.67%	11.84
3	Tone & Voice	78	14.44%	11.27
4	Content & Ideas	74	16.67%	12.33
5	Creative Writing Markers	76	11.11%	8.44
6	Informational / Academic	N/A	—	—
7	Mechanical & Formatting	80	11.11%	8.89
8	Meta-Cognitive & Holistic	83	16.67%	13.83
COMPOSITE SCORE		77.5		77.53

Final Verdict: Very Likely AI-Generated · Composite Score: 77.5 · Confidence: High

Design Decisions

1. Probabilistic Framing Throughout

The prompt repeatedly emphasizes that AI detection is a probabilistic judgment, not a definitive determination. This was a core design decision from iteration 1 — the goal is to produce evidence for

an educator's professional judgment, not to replace that judgment. The phrase 'not a single smoking gun' appears explicitly in the prompt's guiding principles.

2. Text-Type Conditional Categories (5 and 6)

Early iterations applied all eight categories universally, which produced nonsensical scores — academic markers applied to a poem, creative writing markers applied to a research essay. The conditional architecture, with proportional weight redistribution, was introduced in iteration 3 and significantly improved output accuracy across text types.

3. Specific Evidence Requirement

The prompt explicitly instructs the model to reference actual phrases, sentences, structural choices, and ideas from the submitted text — and prohibits generic feedback that could apply to any writing. This was the most important design constraint. Early versions produced evaluations that sounded authoritative but could have been copied from one submission to the next.

4. Confidence Level as a Required Output

Short texts provide insufficient data across eight analytical categories, yet students sometimes submit very brief pieces. The confidence level output (low / moderate / high), keyed to text length and signal convergence, was added in iteration 4 after several evaluations of 80–100 word texts produced misleadingly definitive-sounding verdicts.

5. Weighted Composite with Transparent Calculation

Rather than a simple average, the weighted system reflects that some categories — Linguistic Patterns, Content and Ideas, and Meta-Cognitive Flags — are more diagnostically powerful than others. The prompt requires the model to show its calculation transparently: each category score, its weight, and its contribution to the composite. This makes the output auditable and explainable to a student or administrator if the score is challenged.

6. The Four-Range Scoring Rubric per Category

Each category includes four scoring bands (0–25, 26–50, 51–75, 76–100) with specific descriptors, mirroring the structure of a conventional educational rubric. This was introduced in iteration 2 after the model produced inconsistent scores when given only general guidance. The band descriptors anchor the scoring to specific observable features rather than impressionistic judgment.

Iteration Notes

This prompt went through at least six documented iterations. Key developments:

- Iteration 1: Single-category holistic assessment with no scoring weights. Output was impressionistic and not reproducible.
- Iteration 2: Introduced the eight-category structure and 0–100 scoring per category. Added the four-range rubric descriptors for each category.

- Iteration 3: Added the text-type conditional logic for Categories 5 and 6 with proportional weight redistribution.
- Iteration 4: Added the confidence level requirement after discovering that short-text evaluations were producing misleadingly certain verdicts.
- Iteration 5: Refined vocabulary flag lists in Category 2 (adding 'delve,' 'tapestry,' 'in the realm of') and tightened the specific evidence requirement to prevent generic feedback.
- Iteration 6: Added the transparent composite score calculation requirement and finalized the three-point verdict threshold ranges.

The most persistent challenge across iterations was preventing the model from producing feedback that sounded evidence-based but was actually generic. The solution was specifying the evidential requirement in the guiding principles and repeating it in the output instructions.

Results & Application

8 Independent analytical categories	3-Point Verdict scale with specific thresholds
Conditional Category logic adapts to text type	Auditable Transparent weighted calculation required

The prompt is deployed as a supplementary review tool for student writing submissions across multiple assignment types, including creative writing, persuasive essays, research papers, and personal narratives. It produces a consistent, auditable output format with specific textual evidence — making it defensible in conversations with students or administrators.

Because the verdict is explicitly probabilistic, educators retain full professional discretion in how they use the output. The tool's role is to surface and systematize what an experienced reader notices, not to make determinations on an educator's behalf.

Future development: A batch variant capable of analyzing multiple submissions in a single session, producing comparative flags across a class set, would significantly increase the tool's utility at scale.