People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text



By Jenna Russell, Marzena Karpinska, and Mohit Iyyer

Example text: **Human-written** or **AI-generated**?

One reason Planet Nine (or any Planet X) has been so hard to pin down is that if it exists, it's incredibly faint and slow-moving, possibly hundreds of times farther from the Sun than Earth is. Traditional telescopes rely on painstaking surveys that cover only small portions of the sky at a time. In contrast, Rubin's wide field of view—about 40 times the size of the full Moon—means it will cover the visible sky every few days.

"This isn't just another telescope—it's a new way of doing astronomy," said Dr. Lynne Jones, an LSST researcher at the University of Washington. "We're moving from static snapshots to continuous movies. If there's something out there, no matter how faint, as long as it's moving, we have a good chance of picking it up over time."

The New York Times

How Easy Is It to Fool A.I.-Detection Tools?

THE WALL STREET JOU

English Edition | Print Edition | Video | Audio | Latest Headlines

Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Arts Lifestyle Real E

TECHNOLOGY | PERSONAL TECHNOLOGY | FAMILY & TECH: JULIE JARGON

Students Are Humanizing Their Writing —By Putting It Through AI

It's a battle of the bots: Teachers use AI detection to spot cheating while students use it to maintain innocence

The Washington Post

Democracy Dies in Darkness

Help Desk | Tech in Your Life | Your Data and Privacy | Internet Access | What's New | Ethical Issues | Ask a Question

What to do when you're accused of AI cheating

AI detectors like Turnitin and GPTZero suffer from false positives that can accuse innocent students of cheating. Here's the advice of academics, AI scientists and students on how to deal with it.

August 14, 2023 More than **1 year ago**

8 min | Summary | 288

THE WALL STREET

English Edition | Print Edition | Video | Audio

Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Art

The New York

A New Headache Students: Proving Use A.I.

Students are resorting to extreme me accusations of cheating, including ho their homework sessions.

Listen to this article · 8:52 min Learn more

Share full article

Was a Story That Just Won a Literary Prize A.I.-Generated?

A respected literary magazine has published an award-winning short story many readers believe to be generated by artificial intelligence. Experts aren't all so sure.

Listen · 7:25 min

Share full article

As people have gotten **familiar** with AI-tools like ChatGPT, **Can they detect AI-generated text?**

- Prior studies mainly conducted **pre-ChatGPT**
- Does **not consider evasion** attempts such as paraphrasing
- Prior work had found that **some annotators were significantly better than average, but did not explore the upper bounds of human detections**

… and mostly, we had an inkling it would become quite obvious as we *delved* into the issue…

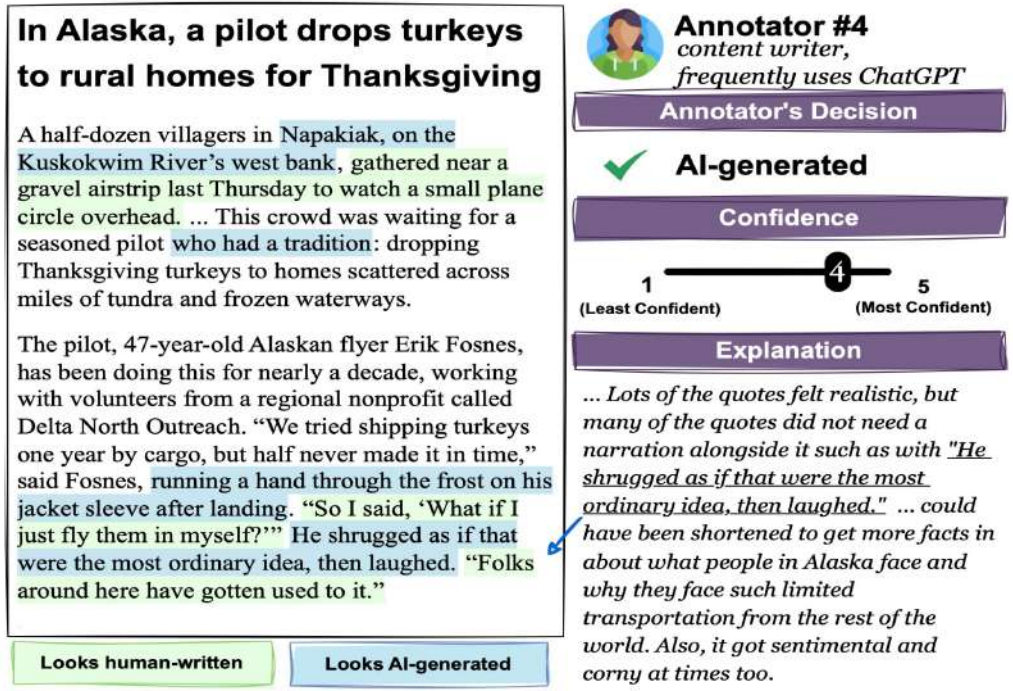
Data

- **30 human-written articles** per round
- **30 AI-generated articles** per round
  - Use the **title & subtitle** from the human-written article

TOTAL:
150 Human-Written &
150 AI-Generated articles

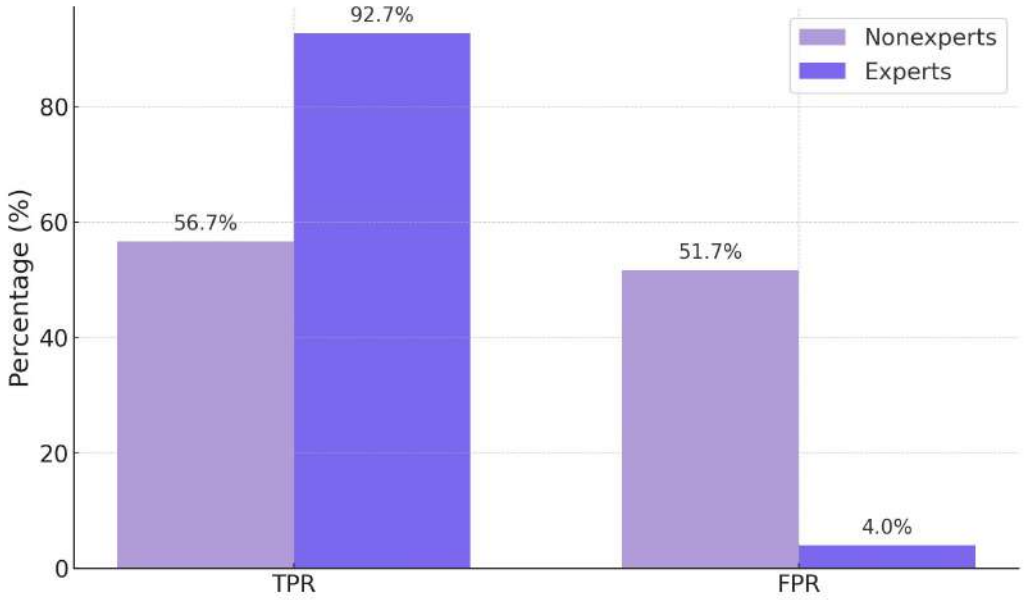
Task Set Up

- Annotators provide
  - Binary Label
  - Confidence Score (1-5)
  - Explanation of Choice
  - Highlighted span used as clues



Who are good detectors?

People who use LLMs frequently for **writing tasks** (e.g., editing, copywriting, creative writing) vastly outperform nonexperts



Who are good detectors?

People who use LLMs frequently for **writing tasks** (e.g., editing, copywriting, creative writing) vastly outperform nonexperts

GPT-4o

GPT-4o-2024-08-06

Experiment 1: What kinds of annotators are good detectors of AI generated text?

- Nonexperts often mistakenly fixate on certain linguistic properties
- Nonexperts believe human authors are more likely than AI to form grammatically-correct sentences
- Nonexperts attribute any text written in a **neutral tone** to AI

Claude

claude-3.5-sonnet

Experiment 2: Are experts robust to AI-output generated from models other than GPT-4o?

- Experts reliably detect articles generated by Claude
- One expert had slightly harder time but same clues were present

O1 Pro

O1 Pro (through ChatGPT interface, no api available)

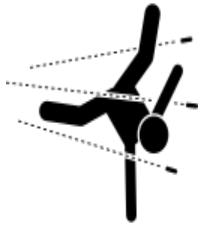
Experiment 3: Do reasoning models still exhibit formulaic outputs?

- ⅘ experts reliably detect o1 pro articles
- One expert is tripped up by the different grammar used by o1Pro

Evasion Attempts

Paraphrasing Attack
(GPT-4o-2024-08-06)

Once story is generated, each sentence is iteratively paraphrased to keep semantic meaning but change exact text.



Humanization Attack
(o1-pro)

Detection Guide and humanization instructions based on annotator feedback from prior rounds given at generation time

Detection Accuracy

	GPT-4o	Claude-3.5-Sonnet	Paraphrased GPT-4o	o1-Pro	o1-Pro Humanized	Overall
Expert Majority	100 (0)	100 (0)	100 (0)	96.7 (0)	100 (0)	99.3 (0)
Pangram	100 (0)	100 (3.3)	100 (0)	100 (0)	96.7 (10)	99.3 (2.7)
GPTZero	100 (0)	96.7 (0)	100 (0)	76.7 (0)	46.7 (3.3)	85.3 (0.7)
Fast-DetectGPT	100 (0)	96.7 (3.3)	56.7 (3.3)	86.7 (0)	23.3 (3.3)	80 (2)
Binoculars	100 (0)	93.3 (0)	60 (6.7)	73.3 (0)	6.67 (0)	66.7 (1.3)
GPT-4o + CoT + Detection Guide	100 (10)	100 (13.3)	100 (16.7)	86.7 (6.7)	3.3 (3.3)	78 (10.7)

Detection Accuracy

	GPT-4o	Claude-3.5-Sonnet	Paraphrased GPT-4o	o1-Pro	o1-Pro Humanized	Overall
Expert Majority	100 (0)	100 (0)	100 (0)	96.7 (0)	100 (0)	99.3 (0)
Pangram	100 (0)	100 (3.3)	100 (0)	100 (0)	96.7 (10)	99.3 (2.7)
GPTZero	100 (0)	96.7 (0)	100 (0)	76.7 (0)	46.7 (3.3)	85.3 (0.7)
Fast-DetectGPT	100 (0)	96.7 (3.3)	56.7 (3.3)	86.7 (0)	23.3 (3.3)	80 (2)
Binoculars	100 (0)	93.3 (0)	60 (6.7)	73.3 (0)	6.67 (0)	66.7 (1.3)
GPT-4o + CoT + Detection Guide	100 (10)	100 (13.3)	100 (16.7)	86.7 (6.7)	3.3 (3.3)	78 (10.7)

Detection Accuracy

	GPT-4o	Claude-3.5-Sonnet	Paraphrased GPT-4o	o1-Pro	o1-Pro Humanized	Overall
Expert Majority	100 (0)	100 (0)	100 (0)	96.7 (0)	100 (0)	99.3 (0)
Pangram	100 (0)	100 (3.3)	100 (0)	100 (0)	96.7 (10)	99.3 (2.7)
GPTZero	100 (0)	96.7 (0)	100 (0)	76.7 (0)	46.7 (3.3)	85.3 (0.7)
Fast-DetectGPT	100 (0)	96.7 (3.3)	56.7 (3.3)	86.7 (0)	23.3 (3.3)	80 (2)
Binoculars	100 (0)	93.3 (0)	60 (6.7)	73.3 (0)	6.67 (0)	66.7 (1.3)
GPT-4o + CoT + Detection Guide	100 (10)	100 (13.3)	100 (16.7)	86.7 (6.7)	3.3 (3.3)	78 (10.7)

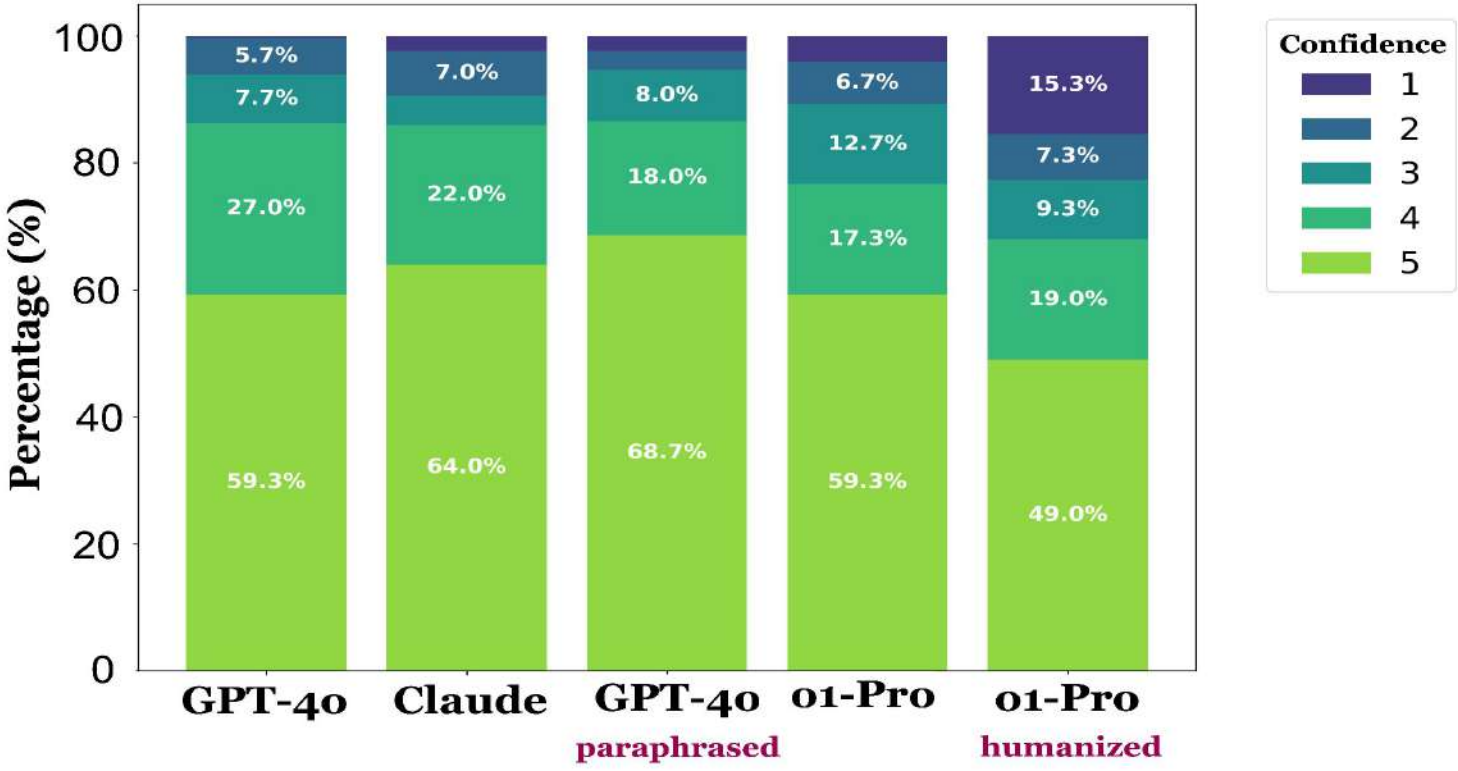
Individual Performance Varied but high across the board

Detection Method	Generation Method					
	GPT-4o TPR% (FPR%)	Claude TPR% (FPR%)	GPT-4o paraphrased TPR% (FPR%)	o1-Pro TPR% (FPR%)	o1-Pro humanized TPR% (FPR%)	Overall TPR% (FPR%)
(A) Expert human detectors						
Expert Majority Vote	100 (0)	100 (0)	100 (0)	96.7 (0)	100 (0)	99.3 (0)
Annotator 1	96.7 (3.3)	100 (0)	100 (0)	96.7 (6.7)	90.0 (23.3)	96.7 (6.7)
Annotator 2	96.7 (0)	80.0 (30)	86.7 (10)	90.0 (10)	86.7 (10)	88.0 (12)
Annotator 3	86.7 (6.7)	100 (0)	93.3 (0)	16.7 (0)	0 (3.3)	59.3 (2)
Annotator 4	90.0 (6.7)	96.7 (13.3)	100 (10)	100 (0)	100 (0)	97.3 (6)
Annotator 5	93.3 (0)	93.3 (6.7)	93.3 (0)	93.3 (0)	93.3 (0)	93.3 (1.3)

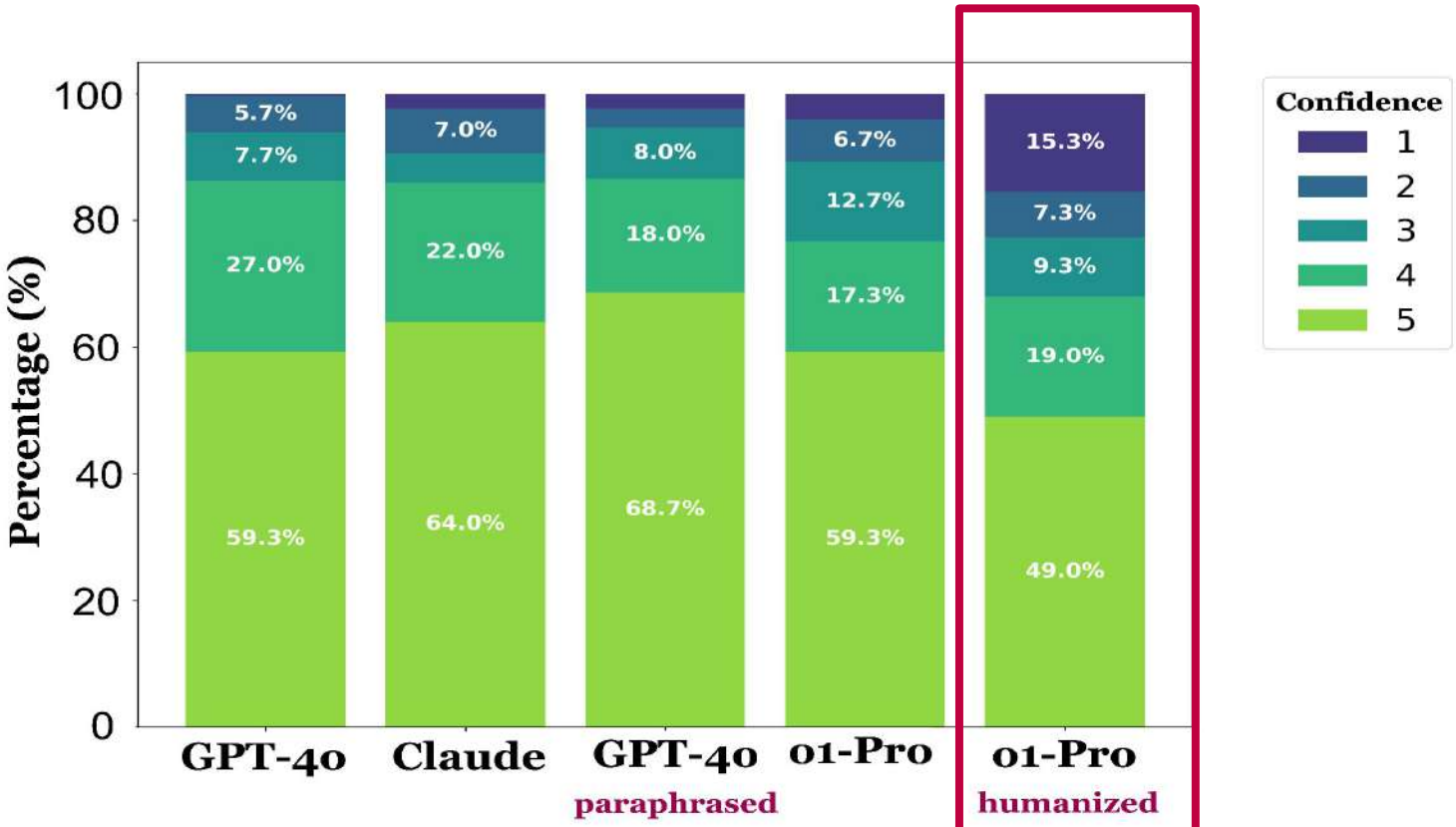
Individual Performance Varied but high across the board

Detection Method	Generation Method					
	GPT-4o TPR% (FPR%)	Claude TPR% (FPR%)	GPT-4o paraphrased TPR% (FPR%)	o1-Pro TPR% (FPR%)	o1-Pro humanized TPR% (FPR%)	Overall TPR% (FPR%)
(A) Expert human detectors						
Expert Majority Vote	100 (0)	100 (0)	100 (0)	96.7 (0)	100 (0)	99.3 (0)
Annotator 1	96.7 (3.3)	100 (0)	100 (0)	96.7 (6.7)	90.0 (23.3)	96.7 (6.7)
Annotator 2	96.7 (0)	80.0 (30)	86.7 (10)	90.0 (10)	86.7 (10)	88.0 (12)
Annotator 3	86.7 (6.7)	100 (0)	93.3 (0)	16.7 (0)	0 (3.3)	59.3 (2)
Annotator 4	90.0 (6.7)	96.7 (13.3)	100 (10)	100 (0)	100 (0)	97.3 (6)
Annotator 5	93.3 (0)	93.3 (6.7)	93.3 (0)	93.3 (0)	93.3 (0)	93.3 (1.3)

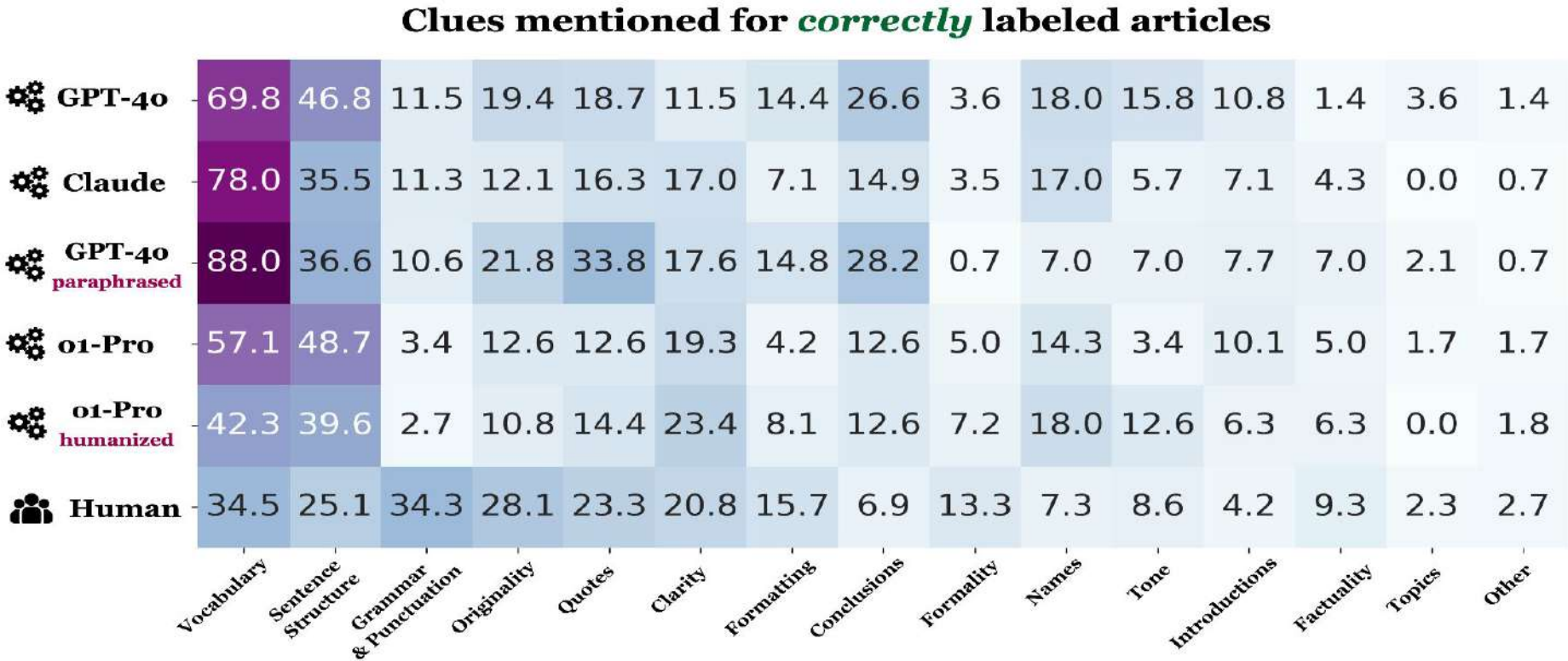
Expert Confidence



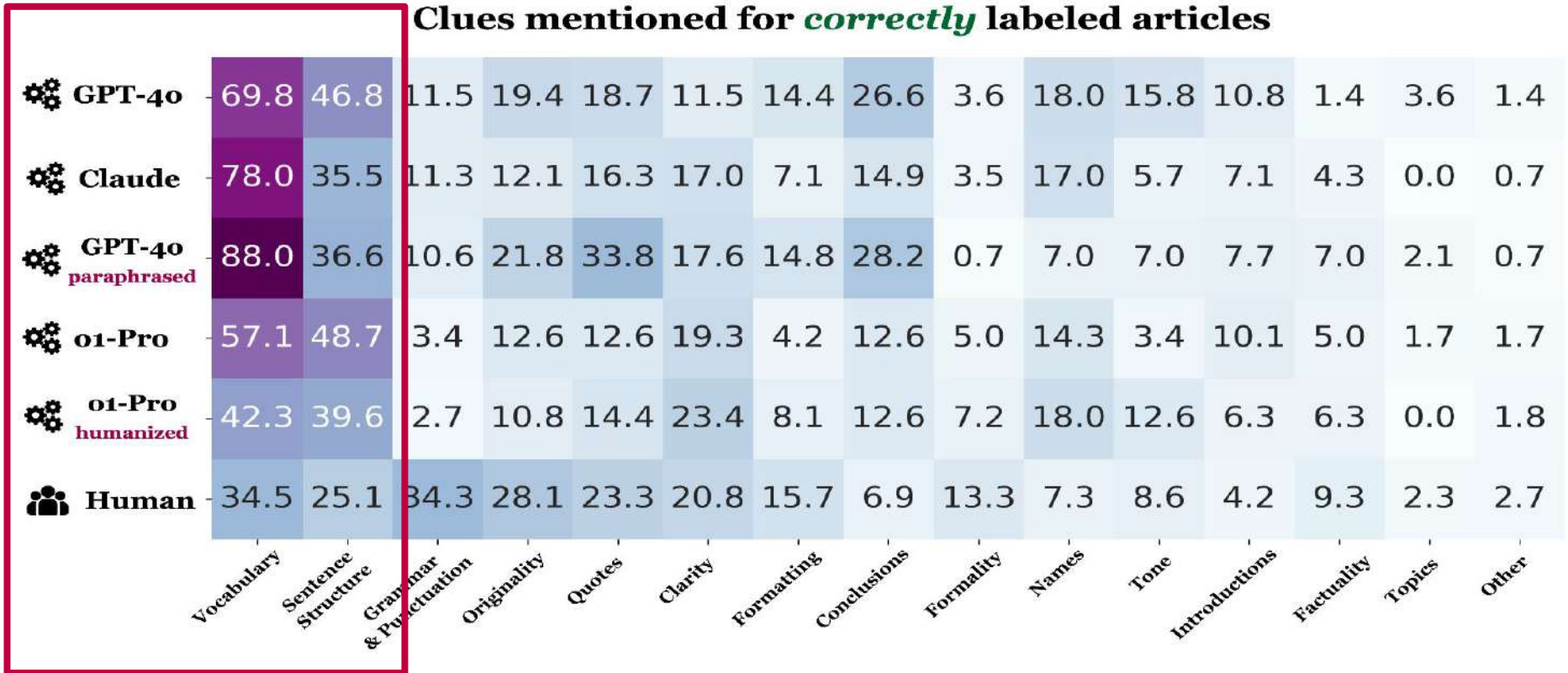
Expert Confidence



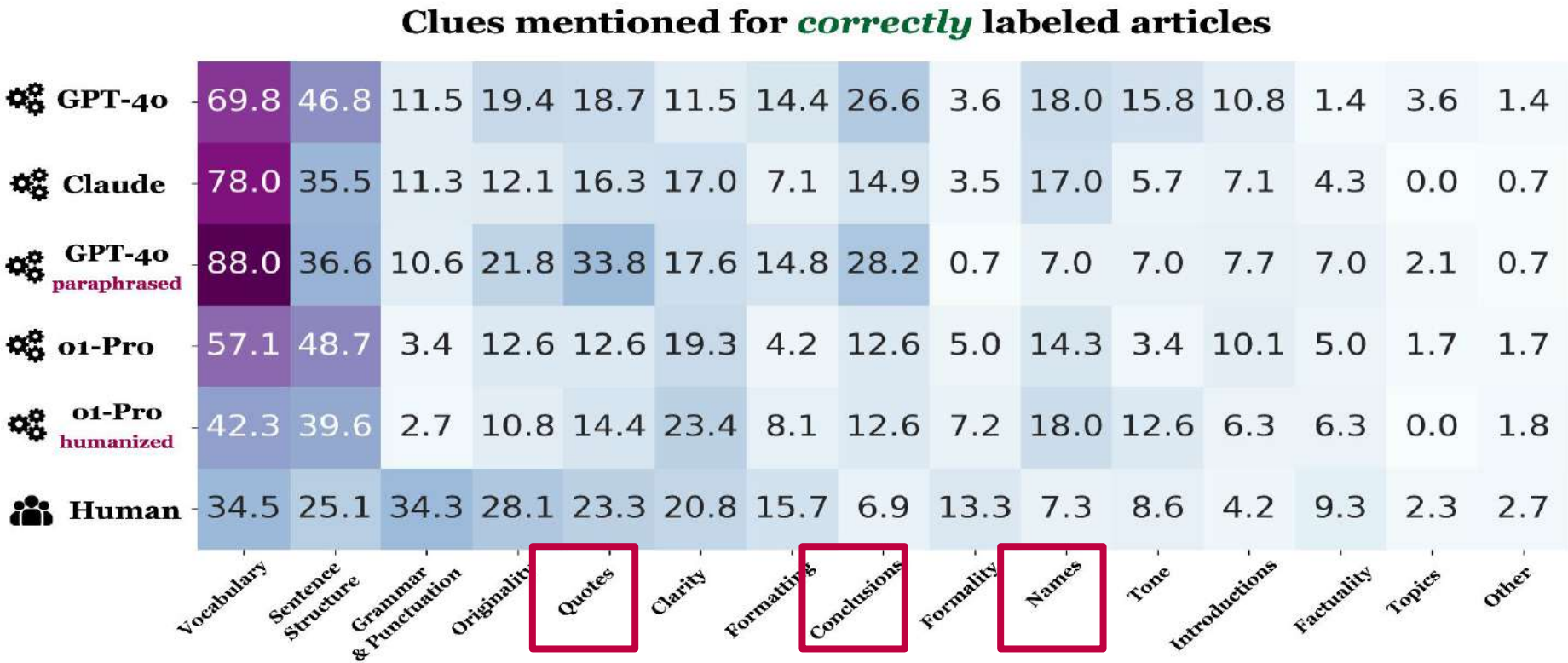
What **clues** do experts use to detect AI-generated content?



What **clues** do experts use to detect AI-generated content?



What **clues** do experts use to detect AI-generated content?



Example text: **Human-written** or **AI-generated**?

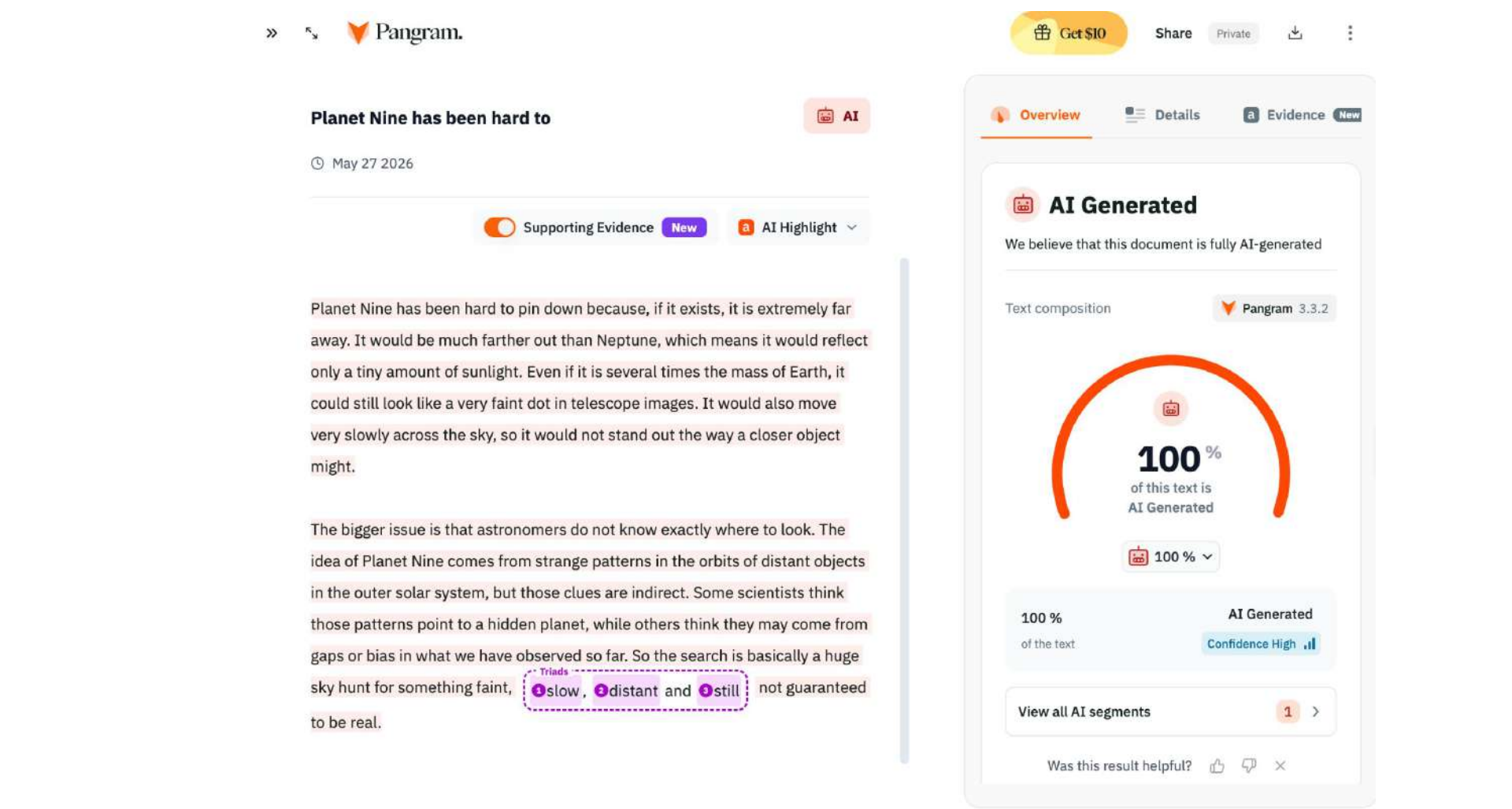
One reason Planet Nine (or any Planet X) has been so hard to pin down is that if it exists, it's incredibly faint and slow-moving, possibly hundreds of times farther from the Sun than Earth is. Traditional telescopes rely on painstaking surveys that cover only small portions of the sky at a time. In contrast, Rubin's wide field of view—about 40 times the size of the full Moon—means it will cover the visible sky every few days.

"This isn't just another telescope—it's a new way of doing astronomy," said Dr. Lynne Jones, an LSST researcher at the University of Washington. "We're moving from static snapshots to continuous movies. If there's something out there, no matter how faint, as long as it's moving, we have a good chance of picking it up over time."

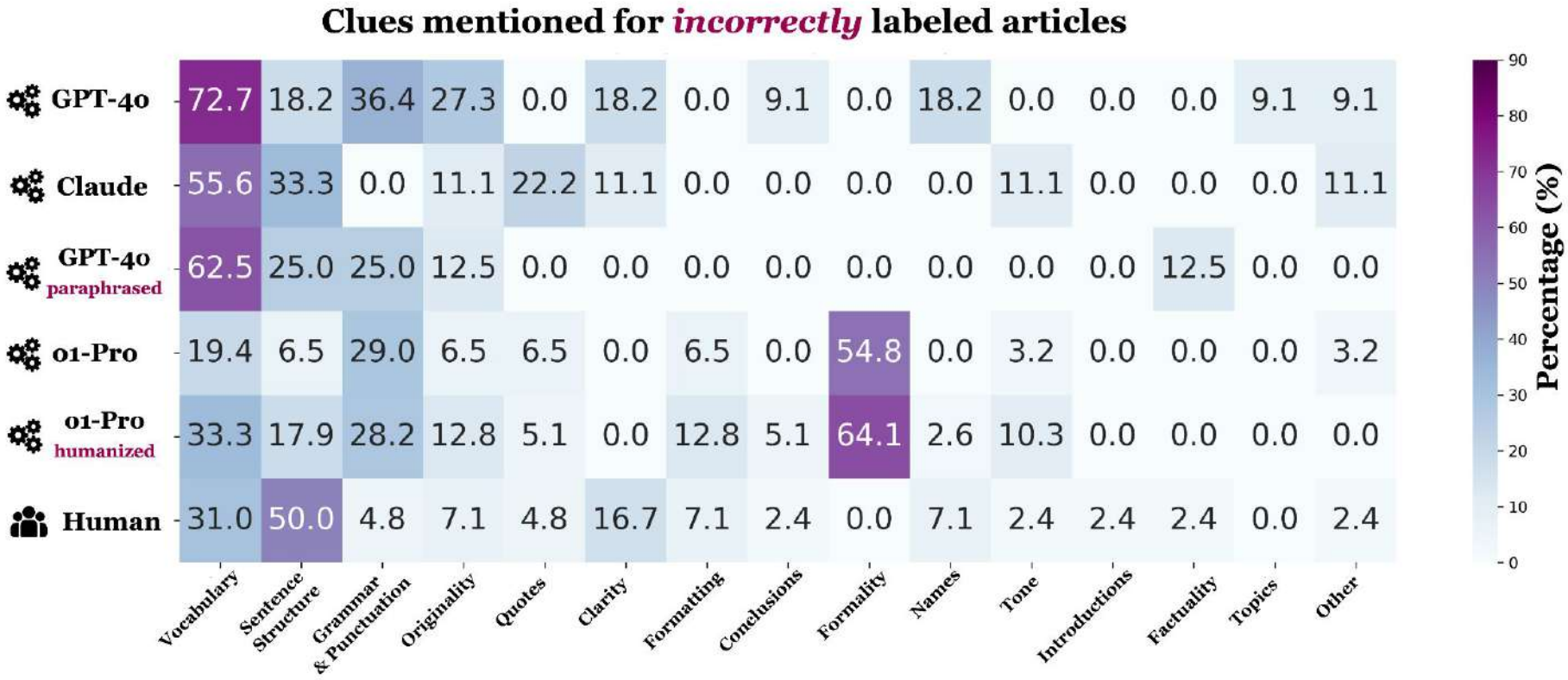
Example text: **Human-written** or **AI-generated**?

Planet Nine has been hard to pin down because, if it exists, it is extremely far away. It would be much farther out than Neptune, which means it would reflect only a tiny amount of sunlight. Even if it is several times the mass of Earth, it could still look like a very faint dot in telescope images. It would also move very slowly across the sky, so it would not stand out the way a closer object might.

The bigger issue is that astronomers do not know exactly where to look. The idea of Planet Nine comes from strange patterns in the orbits of distant objects in the outer solar system, but those clues are indirect. Some scientists think those patterns point to a hidden planet, while others think they may come from gaps or bias in what we have observed so far. So the search is basically a huge sky hunt for something faint, slow, distant, and still not guaranteed to be real.


Pangram.
Get $10
Share
Private
Planet Nine has been hard to
AI
May 27 2026
Supporting Evidence
New
AI Highlight
Planet Nine has been hard to pin down because, if it exists, it is extremely far away. It would be much farther out than Neptune, which means it would reflect only a tiny amount of sunlight. Even if it is several times the mass of Earth, it could still look like a very faint dot in telescope images. It would also move very slowly across the sky, so it would not stand out the way a closer object might.
The bigger issue is that astronomers do not know exactly where to look. The idea of Planet Nine comes from strange patterns in the orbits of distant objects in the outer solar system, but those clues are indirect. Some scientists think those patterns point to a hidden planet, while others think they may come from gaps or bias in what we have observed so far. So the search is basically a huge sky hunt for something faint, Triads 1 slow, 2 distant and 3 still not guaranteed to be real.
Overview
Details
Evidence
New
AI Generated
We believe that this document is fully AI-generated
Text composition
Pangram 3.3.2
100 %
of this text is
AI Generated
100 %
100 %
of the text
AI Generated
Confidence High
View all AI segments
1
Was this result helpful?

What clues mislead experts?



CATEGORY	FREQ	DEFINITION	EXAMPLE EXPLANATIONS
VOCABULARY	53.1%	LLMs use specific words and phrases more often than human writers, which often results in repetitive, unnatural, or overly complex wording.	**Human**: *"Furthermore, I very much doubt AI would have used adventurous adjectives like 'chunky', 'musky' or 'thin' to describe food. Nor would it have used verbs like 'blitzing' or 'bolstering'."* **AI (O1-HUMANIZED)**: *"Odd word choices: wheat that 'stores' a lineage; genes that are 'honed.'"*
SENTENCE STRUCTURE	35.9%	AI-generated sentences follow predictable patterns (e.g., high frequency of "not only ... but also ...", or consistently listing three items), while human-written sentences vary more in terms of length.	**Human**: *"Short choppy sentences and paragraphs."* **AI (O1-PRO)**: *"One pattern I've been noticing with AI, and I think I've stated this before, is the comparison of 'it's not just this, it's this' and I'm seeing it here, along with listings of specifically three ideas."*
GRAMMAR & PUNCTUATION	24.8%	AI-generated text is usually grammatically perfect (also avoiding dashes and ellipses), while human-written text often contains minor errors.	**Human**: *"There's a lot of variety in the article's grammar use, with dashes, brackets, quotes intermixed with sentences, and short spurts of comma sections throughout."* **AI (GPT-4O-PARA)**:*"there's nothing off about the grammar or syntax in this piece..."*
ORIGINALITY	23.7%	AI-generated writing is generally straightforward, "safe," and lacking in surprises or humor, leaving annotators bored or disengaged.	**Human**: *"it's offset by some great analogies and creative phrasing that works well to convey the topic, such as with "amateur sleuths", "catnip for a certain type of Reddit user."* **AI (O1-PRO)**: *"What happens when AI tries to be creative? Penguins "stand on their own flippers"."*
QUOTES	22.3%	AI-generated quotes sound overly formal, lack the varied nuances of real conversation, and often mirror the article's main text too closely in style.	**Human**: *" The quotes being short snippets also makes me think they're real, as the writer had to find a way to fit them into the text, rather than them just perfectly stating either side's views."* **AI (GPT-4O)**: *"The quotes also feel fake, every expert speaks the same way and it's too homogenous with the text."*
CLARITY	19.5%	AI-generated text often lacks concise flow by over-explaining or including irrelevant details, effectively "telling" rather than "showing".	**Human**: *"Words like "meander" are used, but are used sparingly to create better flow of ideas, and its writing style is simplified in the best way possible."* **AI (CLAUDE-3.5-SONNET)**: *"The sentences are condensed to provide the best possible precision with its word choice, but the article lacks flow and clarity."*