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

Jenna Russell[1] **Marzena Karpinska**[2] **Mohit Iyyer**[1,3]
[1]University of Maryland, College Park [2]Microsoft [3]UMass Amherst
{jennarus,miyyer}@umd.edu, mkarpinska@microsoft.com

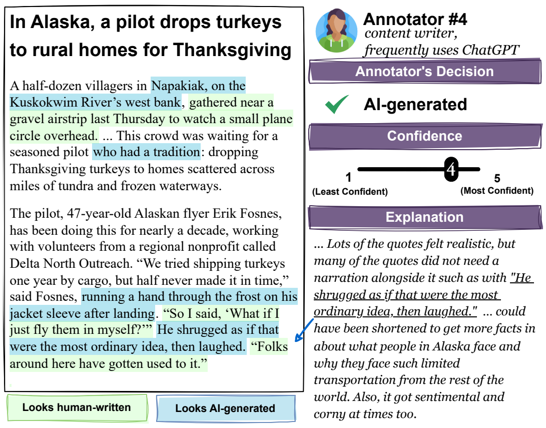
Abstract

In this paper, we study how well *humans* can detect text generated by commercial LLMs (GPT-4o, Claude-3.5-Sonnet, o1-Pro). We hire annotators to read 300 non-fiction English articles, label them as either human-written or AI-generated, and provide paragraph-length explanations for their decisions. Our experiments show that annotators who frequently use LLMs for writing tasks excel at detecting AI-generated text, even without any specialized training or feedback. In fact, the majority vote among five such "expert" annotators misclassifies only 1 of 300 articles, significantly outperforming most commercial and open-source detectors we evaluated even in the presence of evasion tactics like paraphrasing and humanization. Qualitative analysis of the experts' free-form explanations shows that while they rely heavily on specific lexical clues ("AI vocabulary"), they also pick up on more complex phenomena within the text (e.g., formality, originality, clarity) that are challenging to assess for automatic detectors. We release our annotated dataset and code to spur future research into both human and automated detection of AI-generated text.

https://github.com/jenna-russell/human_detectors

1 Introduction

Text generated by large language models (LLMs) is rampant in today's world.[1] This state of affairs has spurred increased research into automatic detection of AI-generated text, which helps combat academic plagiarism (Zhu et al., 2024) and fake content creation (Gameiro et al., 2024). Unfortunately, automatic detectors suffer from low detection rates, poor robustness to evasion attempts, and limited explainability to end users (Sadasivan et al., 2024; Ji et al., 2024).[2]



In Alaska, a pilot drops turkeys to rural homes for Thanksgiving

A half-dozen villagers in Napakiak, on the Kuskokwim River's west bank, gathered near a gravel airstrip last Thursday to watch a small plane circle overhead. ... This crowd was waiting for a seasoned pilot who had a tradition: dropping Thanksgiving turkeys to homes scattered across miles of tundra and frozen waterways.

The pilot, 47-year-old Alaskan flyer Erik Fosnes, has been doing this for nearly a decade, working with volunteers from a regional nonprofit called Delta North Outreach. "We tried shipping turkeys one year by cargo, but half never made it in time," said Fosnes, running a hand through the frost on his jacket sleeve after landing. "So I said, 'What if I just fly them in myself?'" He shrugged as if that were the most ordinary idea, then laughed. "Folks around here have gotten used to it."

🟩 Looks human-written 🟦 Looks AI-generated

Annotator #4
content writer, frequently uses ChatGPT

Annotator's Decision

✓ AI-generated

Confidence

1 (Least Confident) ● 4 ● 5 (Most Confident)

Explanation

... Lots of the quotes felt realistic, but many of the quotes did not need a narration alongside it such as with "He shrugged as if that were the most ordinary idea, then laughed." ... could have been shortened to get more facts in about what people in Alaska face and why they face such limited transportation from the rest of the world. Also, it got sentimental and corny at times too.

Figure 1: A human expert's annotations of an article generated by OpenAI's o1-Pro with humanization. The expert provides a judgment on whether the text is written by a human or AI, a confidence score, and an explanation (including both free-form text and highlighted spans) of their decision.

In this paper, we instead study how well *humans* can detect AI-generated text. Unlike prior work on this topic, which was mainly conducted in the pre-ChatGPT era (Ippolito et al., 2020; Brown et al., 2020; Clark et al., 2021), we focus on text generated by modern LLMs (GPT-4o, Claude-3.5-Sonnet, o1-Pro) and in the presence of evasion attempts (paraphrasing, humanization). We hire human annotators to read non-fiction English articles, label them as written by either a human or by AI, and provide a paragraph-length explanation of their decision-making process (Figure 1). Overall, we collect **1790** annotations on **300** unique articles at a total cost of **$4.9K USD**, which allows us to compare humans to automatic detectors and analyze what kind of clues they rely on.

[1] For example, recent studies estimate that over 5% of recently-published Wikipedia articles (Brooks et al., 2024) and 10% of PubMed abstracts published in 2024 (Kobak et al., 2024) were written by AI.

[2] This paper focuses only on *post-hoc* detection of AI-generated text, which unlike watermarking (Kirchenbauer et al., 2023) requires no cooperation from LLM providers.

Experience matters: Unsurprisingly, annotators who rarely or never use LLMs are poor detectors of AI-generated text. However, we identify a subset of five high-performing annotators who frequently use LLMs for writing-related tasks. The majority vote of this subset of "expert" annotators fails to predict the correct label on only **one** out of 300 articles. We emphasize that we do not provide any training to our annotators: they are given no feedback on which articles they labeled incorrectly.

Human experts outperform automatic detectors: The majority vote of our five expert annotators substantially outperforms almost every commercial and open-source detector we tested on these 300 articles, with only the commercial Pangram model (Emi and Spero, 2024) matching their near-perfect detection accuracy. In our most complex configuration, which requires detecting articles generated and humanized by O1-PRO, the expert majority vote is perfect (true positive rate of 100%), while open-source methods such as Binoculars (6.7%) and Fast-DetectGPT (23.3%) struggle. We conclude that hiring expert human annotators to perform detection is a viable strategy, particularly in high-stakes settings where explainability is critical.

What do expert annotators focus on? An analysis of our expert's explanations reveals that usage of "AI vocabulary" (e.g., *vibrant*, *crucial*, *significantly*) form the most common giveaways. Close behind are formulaic sentence and document structures (e.g., optimistically vague introductions and conclusions) and originality (how creative or engaging an article is), while complex phenomena such as factuality and tone are also useful signals. We observe that neither paraphrasing nor humanization effectively removes all of these signatures; that said, these evasion tactics and defenses for them are still underexplored in the research community.

Can LLMs mimic human detectors? We attempt to automatically replicate the decision-making process of our human experts by providing a candidate text to an LLM, along with a guidebook compiled from human expert explanations, and instructing it to use the guidebook to judge whether the text is written by an AI or by a human. This simple prompting strategy is competitive with existing detectors on easier configurations but struggles to detect humanized articles. Moving forward, we believe LLMs can be used in conjunction with high-performing detectors such as Pangram to offer an explanation of the judgment to end users.

Contributions: Our study establishes the existence of an annotator population (frequent users of LLMs for writing-related tasks) that is highly capable of detecting AI-generated text without any additional training. The majority vote of five such annotators outperforms all automatic detectors we evaluated, even on paraphrased or humanized AI-generated articles. We release our annotated data and codebase to facilitate future research into both human and automatic detection of AI-generated text.

2 How good are humans at detecting AI-generated text?

Automatic detectors of AI-generated text suffer from low detection rates and poor robustness to simple perturbations of the text (e.g., paraphrasing), rendering them unreliable for real-world deployment (Sadasivan et al., 2024). How well do *human* detectors stack up against their automatic counterparts? What features of a candidate text do humans find most predictive of AI writing? To answer these questions, we conducted five experiments that ask humans to judge increasingly more deceptive AI-generated texts (i.e., produced using more powerful base LLMs and evasion tactics).

Task setup: Each experiment consists of a batch of **60** articles, 30 of which are written by humans and the remaining 30 generated by an LLM. Each human-written article has a corresponding AI-generated counterpart with the same title and subtitle to control for high-level content and topics. The 60 articles are shown to annotators one at a time in a randomized order, and they are asked to provide the following for each article, depicted in Figure 1:

1. a **binary label** of whether the article was written by a human or an AI

2. a **rating of their confidence** in their choice on a scale from 1 (least confident) to 5 (most confident)

3. **highlighted spans** from the article that they used as clues to make their choice

4. a paragraph-length **explanation** of their choice

Article selection: We restrict our study to English nonfiction articles of fewer than 1K words geared towards lay audiences.[3] While we cannot make claims about human detection performance in other domains (e.g., scientific papers or social media posts),[4] we focus on such articles because (1) malicious LLM users generate large volumes of fake articles, making it practically impactful; (2) such articles do not require specialized knowledge to understand or judge; and (3) many previous studies also ask humans to judge AI-generated text in this domain (Ippolito et al., 2020; Brown et al., 2020; Clark et al., 2021; Puccetti et al., 2024).

Generating paired AI articles: For each human-written article, we generate a corresponding AI article by prompting an LLM with the title, subtitle, desired length, and publication source of the original article. This ensures that AI-generated articles are directly comparable in topic and length, eliminating confounds from content variation. By constructing these human-AI article pairs with only authorship as the varying factor, we create minimal pairs that allow direct comparisons (Gardner et al., 2020; Warstadt et al., 2020; Karpinska et al., 2022, 2024). We employ a within-subjects design for our experiments, where each annotator judges both the human-written and AI-generated articles. This reduces variability from individual differences and requires fewer annotators (Allen, 2017). To further minimize bias, annotators are unaware of the pairing, and the article order is randomized. An example article generation prompt is below:

```
You are given the following title and
subtitle of a general article from
the Science section of New York Times
and asked to write a corresponding
article of around 750 words. Include
quotations from relevant experts and
make sure the article is concise and
easily understandable to a lay audience.

Title: The Science That Makes Baseball
Mud ‘Magical’

Subtitle: Scientists dug up the real
dirt on the substance applied to all
the baseballs used in the major leagues.

Article:
```

Exp #	Model	Human-written (*n=30*)	AI-generated (*n=30*)
1	GPT-4o	$652.5_{170.9}$	$625.1_{126.9}$
2	Claude-3.5-Sonnet	$793.1_{170.3}$	$710.8_{133.5}$
3	GPT-4o paraphrased	$777.8_{175.9}$	$654.9_{110.6}$
4	o1-Pro	$728.4_{155.1}$	$813.8_{233.6}$
5	o1-Pro humanized	$739.0_{128.3}$	$815.6_{171.2}$

Table 1: Mean and standard deviation (subcripted) of article length in words across experiments, computed by splitting on whitespaces.

For experiments 3 and 5, we modify the prompt to include perturbations designed to evade detectors (paraphrasing and humanization). Comparisons of article lengths between human and AI-generated articles for each experiment are reported in Table 1 and further detailed in §B.1.[5]

Annotator details: We recruited annotators via Upwork,[6] all of whom identified themselves as native English speakers. All annotators were required to read the project guidelines and sign a consent form prior to the task, provided in §A. Additionally, we surveyed each annotator to determine their level of education, profession, English dialect, and familiarity with LLMs, including how they use these models (see Table 5 for details). They were compensated $2 per candidate text, resulting in an hourly rate of about $15 - $30 USD. Ultimately, we collect 1740 annotations from 9 annotators at a total cost of $4.9K USD.

Evaluation metrics: We evaluate both human and automatic detectors on **True Positive Rate (TPR)**, which measures the percentage of AI-generated articles that are successfully detected, as well as **False Positive Rate (FPR)**, which measures the percentage of human-written articles incorrectly detected as AI-generated. Prior work in automatic detection reports either TPR at a fixed low FPR such as 1% (Krishna et al., 2023; Hans et al., 2024; Dugan et al., 2024) or AUROC (Mitchell et al., 2023; Hu et al., 2023; Bao et al., 2023), which we cannot easily do with human annotators. For fair comparison, we report TPR and FPR of automatic detectors using recommended thresholds whenever possible, otherwise using a threshold calibrated to an FPR of 5% on a held-out set of 40 human-written articles.

[3] Human-written articles in our experiments come from eight different American publications: Associated Press, Discover Magazine, National Geographic, New York Times, Reader’s Digest, Scientific American, Smithsonian Magazine, and Wall Street Journal.

[4] We do report results on §2.1 for fictional stories, which can be found in §C

[5] We set `temperature=0` for Experiments 1-3 and use ChatGPT (Pro) interface for Experiments 4-5 as the API for o1 was not yet available.

[6] https://www.upwork.com/

2.1 Experiment 1: What kinds of annotators can reliably detect AI-generated articles?

LLM: GPT-4o (`2024-08-06`)

Evasion tactic: none

Is there a population of humans that is highly accurate at our detection task, and if so, are there any commonalities between them? To explore these questions, we first ask five annotators with varying backgrounds and levels of expertise with LLMs to attempt a batch of 60 articles (half generated by GPT-4o `2024-08-06` and half human-written as described above). We observe both ends of the spectrum in terms of performance: annotators who rarely or never use LLMs for writing tasks performed near random, while one annotator who uses LLMs (and edits LLM-generated text) daily performed almost perfectly. More details on our initial study can be found in §C. Inspired by this result, we recruit four more annotators with similar backgrounds as our top performer, all of whom perform similarly well on the same batch of articles.

Annotators with limited LLM experience are unreliable detectors: The four annotators who self-report either not using LLMs at all, or using LLMs only for non-writing tasks, detect AI-generated text at a similar rate to random chance, achieving an average TPR of 56.7% and FPR of 52.5% (Table 2).[7] While performance among these annotators varies greatly, no single annotator could reliably detect AI-generated text. We term this population *nonexperts* for the rest of this paper.

Annotators who frequently use LLMs for writing tasks are highly accurate: In contrast, the five annotators who have significant experience with using LLMs for writing-related tasks are able to detect AI-generated text very reliably, achieving a TPR of 92.7%. The average FPR for this population of annotators was 3.3%, meaning that they rarely mistake human-written text as AI-generated. The majority vote out of these five annotators correctly determined authorship of all 60 articles; the "GPT-4o" column of Table 3 contains more details. We term this population *experts* for the rest of this paper; all subsequent experiments rely on only these five expert annotators.

Metric	Nonexperts	Experts
Avg. TPR	56.7	92.7
Avg. FPR	51.7	4.0
Avg. Confidence	4.027	4.394

Table 2: On average, nonexperts perform similar to random chance at detecting AI-generated text, while experts are highly accurate.

What do expert annotators see that nonexperts don't? To understand why experts far outperformed nonexperts at detecting AI-generated text, we analyze the comments each annotator provided in their explanations. Overall, nonexperts often mistakenly fixate on certain linguistic properties compared to experts. One example is vocabulary choice, where nonexperts take the inclusion of any "fancy" or otherwise low-frequency word types as signs of AI-generated text; in contrast, experts are much more familiar with exact words and phrases overused by AI (e.g., *testament*, *crucial*).[8] Nonexperts also believe that human authors are more likely to form grammatically-correct sentences and thus attribute run-on sentences to AI, while experts realize the opposite is true: humans are more likely than AI to use ungrammatical or run-on sentences. Finally, nonexperts attribute any text written in a neutral tone to AI, which results in many false positives because formal human writing is also often neutral in tone.

2.2 Experiment 2: Can experts detect articles generated by a different LLM?

LLM: Claude-3.5-Sonnet

Evasion tactic: none

While ChatGPT is the most widely-used LLM (both in general and by our expert annotators), competitors such as Anthropic's Claude (Anthropic, 2023) also have a wide user base. Are our experts overfitting to artifacts specific to GPT-4o, or do they detect patterns that generalize to other LLMs? We address this question in Experiment 2 by evaluating our experts on a second batch of 60 articles, this time with 30 new human-written articles and 30 corresponding articles generated via Claude-3.5-Sonnet (Anthropic, 2024) using the same instructions as in Experiment 1.

[7] We report the average TPR and FPR since there were only four *non-expert* annotators. For the remainder of the paper, we will report the majority vote of the five *expert* annotators (i.e., at least three out of five agreed).

[8] A complete list of "AI vocab" found in the detection guide (Table 11) is detailed in Table 12.

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)
(B) Automatic detectors						
🔒 Pangram Humanizers	100 (0)	100 (3.3)	100 (0)	100 (0)	96.7 (10)	99.3 (2.7)
🔒 Pangram	100 (0)	100 (3.3)	100 (0)	100 (0)	90.0 (6.7)	98.0 (2)
🔒 GPTZero	100 (0)	96.7 (0)	100 (0)	76.7 (0)	46.7 (3.3)	85.3 (0.7)
🔓 Fast-DetectGPT (FPR=0.05)	100 (0)	96.7 (3.3)	56.7 (3.3)	86.7 (0)	23.3 (3.3)	80.0 (7.2)
🔓 Binoculars (Accuracy)	100 (0)	93.3 (0)	60.0 (6.7)	73.3 (0)	6.67 (0)	66.7 (1.3)
🔓 Binoculars (Low FPR)	96.7 (0)	80 (0)	13.3 (0)	10.0 (0)	0 (0)	40.0 (0)
🔓 RADAR (FPR=0.05)	66.7 (0)	0 (0)	10 (3.3)	0 (3.3)	0 (3.3)	15.3 (2)
(C) Prompt-based detectors						
Detector LLM: **GPT-4o-2024-11-20**						
Zero-shot	100 (10)	93.3 (10)	100 (6.7)	56.7 (3.3)	6.7 (3.3)	71.3 (6.7)
Zero-shot + CoT	63.3 (3.3)	33.3 (0)	96.7 (6.7)	16.7 (0)	0 (0)	42.0 (2.0)
Zero-shot + Guide	100 (10)	96.7 (10)	100 (13.3)	80 (6.7)	3.3 (3.3)	76.0 (8.7)
Zero-shot + CoT + Guide	100 (10)	100 (13.3)	100 (16.7)	86.7 (6.7)	3.3 (3.3)	78.0 (10.7)
Detector LLM: **o1-2024-12-17**						
Zero-shot	93.3 (3.3)	66.6 (6.7)	96.7 (6.7)	40.0 (3.3)	20.0 (6.7)	42.2 (5.6)
Zero-shot + CoT	83.3 (6.7)	53.3 (3.3)	96.7 (3.3)	20 (3.3)	16.7 (3.3)	54 (4)
Zero-shot + Guide	93.3 (0)	30.0 (0)	96.7 (0)	13.3 (0)	0 (0)	36.7 (0)
Zero-shot + CoT + Guide	86.7 (0)	43.3 (0)	90.0 (0)	6.7 (0)	0 (0)	53.3 (0.6)

Table 3: Performance of expert humans (top), existing automatic detectors (middle), and our prompt-based detectors (bottom), where each cell displays **TPR (FPR)**. Colors indicate performance bins where **darkest teal** is best, **orange** is middling, and **purple** is worst. We further mark closed-source (🔒) and open-weights (🔓) detectors. The majority vote of expert humans ties Pangram Humanizers for highest overall TPR (99.3) without any false positives, while substantially outperforming all other detectors. While the majority vote is extremely reliable, individual annotator performance varies, especially on o1-Pro articles with and without humanization. Prompt-based detectors are unable to match the performance of either expert humans or closed-source detectors.

Experts reliably detect articles generated by Claude: Despite the change in model, experts are reliable at detecting AI-generated content. TPR is almost unchanged from Experiment 1 and the expert majority vote is again 100% accurate ("Claude-3.5-Sonnet" column of Table 3), with two annotators achieving a perfect score individually. We note that one expert (Annotator 2) achieved higher performance in the experiments where an OpenAI model was used, with their TPR falling to 80% in this experiment. Annotator 2 did not have any prior experience using Claude, although three other annotators who were also unfamiliar with Claude achieved TPRs of at least 96.7%. However, Annotator 2 focuses more heavily on the existence of "AI vocab" in the article than the other annotators, and they were not familiar with the distribution of words that are overused by Claude.

2.3 Experiment 3: How robust are experts to paraphrasing attacks?

LLM: GPT-4o (2024-08-06)

Evasion tactic: paraphrasing

Users of LLMs who hope to evade detection often resort to simple post-hoc modifications of LLM outputs such as *paraphrasing*. These tactics significantly drop detection rates of automatic methods (Krishna et al., 2023; Sadasivan et al., 2024), but how do they affect our human experts? We evaluate the five expert annotators on a third set of 60 articles, where the 30 AI-generated articles go through an additional round of paraphrasing, and we show that they maintain high TPR and low FPR.

Creating paraphrased AI-generated articles: We follow the sentence-by-sentence paraphrasing approach with LLMs outlined by Chang et al. (2024) to create a more difficult set of AI articles for this experiment. Additionally, since our experts easily identified AI-generated named entities (e.g., Sarah Thompson) in the previous experiments, we extract named entities from the human-written article and instruct GPT-4o to include these names in its generated article. We use GPT-4o as both the article generation model and the paraphrasing model. See Table 10 for paraphrasing prompts used in this experiment.

Experts are robust to paraphrased AI-generated articles: Overall, TPR and FPR remain almost unchanged in this experiment compared to the first two, suggesting that paraphrasing is not an effective attack on expert human detectors (see "GPT-4o paraphrased" column in Table 3). The majority vote among the five experts again correctly detects the labels of all 60 articles. Annotators continue to pick up on many of the same clues within the paraphrased articles that were also apparent in Experiments 1 & 2, such as high frequency of "AI vocab" (even after paraphrasing), formulaic sentence structures, and cheerful summary conclusions.

2.4 Experiment 4: Can experts keep up with advances in LLM reasoning capabilities?

LLM: o1-Pro

Evasion tactic: none

While we were in the middle of conducting experiments, OpenAI released their o1 model (OpenAI, 2024), ushering in a new paradigm of test-time scaling. This release offered us a unique opportunity: since none of our experts had been previously exposed to a model with such advanced reasoning capabilities (Zhong et al., 2024), how well would they do at detecting its output? Interestingly, our experts remain reliable detectors of articles generated by o1-Pro (using the same prompt as in Experiments 1 and 2),[9] although their comments show that they often perform a more detailed analysis of articles to make their judgments.

Four out of five experts are robust to o1-Pro: For four of five experts, detection ability of o1-generated content remains largely consistent with prior experiments, with a majority vote TPR of 96.7% and FPR of 0% (see "o1-Pro" column of Table 3).[10] Interestingly, Annotator 3's TPR drops considerably, as they prioritize signs of human writing over signs of AI writing when making their judgments. Average confidence dropped to 4.21 out of 5, compared to average confidence of 4.39, 4.38, and 4.48 from Experiments 1,2, & 3 respectively (see Figure 2 for details). The drop in confidence, as well as verbal feedback from our annotators, demonstrates the increased challenge posed by o1-generated articles. However, their steady aggregate performance highlights that even new model paradigms are still detectable by human experts.

Experts provide more nuanced explanations when detecting o1-Pro: In prior experiments, expert explanations focused primarily on whether or not a candidate text possesses characteristics of AI. However, experts shift focus with o1-Pro by more frequently commenting on identifying characteristics that make text sound "human". For instance, experts frequently point to how humans repeatedly write the word *says*, while AI tries to use more descriptive synonyms like *notes* and *explains*. Annotator 3, the only one consistently fooled by o1-Pro outputs, relied too much on signs of informality (e.g., contractions, slang usage, usage of *just* and *actually*) as a sign of human writing, with 66.7% of their explanations relating to formality (Figure 11).

2.5 Experiment 5: Are experts robust to humanization?

LLM: o1-Pro

Evasion tactic: humanization

As the deployment of automatic AI detectors increases, many users are attempting to evade them by using *humanization* methods, which explicitly attempt to make AI-generated text more human-like (Wang et al., 2024a,b) unlike more generic attacks such as paraphrasing. Are such methods effective at evading our human experts? Since no well-established humanization methods exist, we first create our own humanizer by prompting o1-Pro with a detailed set of instructions derived from expert comments written for Experiments 1-4. Despite considerably degrading performance of many automatic detectors, the majority of our experts remain robust to humanization ("o1-Pro humanized" column of Table 3).

[9] o1-Pro is available through the ChatGPT Pro subscription. OpenAI states o1-Pro produces more reliably accurate and comprehensive responses than o1.

[10] One AI-generated article was incorrectly marked by the expert majority vote as human-written. Analyzing their explanations reveals that the three annotators who rely most heavily on vocabulary and sentence structure did not find enough clues in those categories to judge the article as AI-generated. The full text of the article is in Table 14.

Implementing a prompt-based humanizer: We modify the article generation prompt used in prior experiments to include detailed instructions on specific AI signatures to avoid when generating the article. To obtain these instructions, we consult our experts directly, paying them $45 each to provide us with a list of clues that they look for during detection. We then manually organize these disparate clues into a unified "guidebook" with different sections (e.g., vocabulary, grammar, tone, introductions, conclusions), where each section provides explanations and examples of how AI writing differs from human writing (see Table 11 for a truncated version of the guidebook). Then, we prompt O1-PRO with the guidebook and an instruction to generate an article that would *evade* a detector that was following the guidebook (see Table 13 for prompt).[11] Through this process, we create a new batch of 60 articles, 30 of which are generated by our humanization prompt using O1-PRO. More details on the development of our humanization method can be found in §B.2.

Experts remain robust to humanized articles: Despite our best efforts to generate articles that our experts would find undetectable, most of their detection rates remain largely unchanged from prior experiments, and the expert majority vote is again perfect on all 60 articles. Annotator 3, who struggled to detect non-humanized O1-PRO articles in Experiment 4, performed remarkably poorly on this batch, achieving a TPR of zero by marking almost every article as human-written. While detection rates remained steady, overall confidence dropped, with 15.1% of annotations having a confidence of 1, showing the increased difficulty presented by the humanized articles (see Figure 2).

AI clues still remain after humanization: Although our experts reported this round to be the most challenging, they remained consistent in the types of clues used from prior rounds, demonstrating that the humanization prompt did not remove the entirety of the "AI signature" from the articles. For example, while humanization increases variety in AI-generated names, often even including people relevant to the article who exist in real life, our experts found that humanized articles use titles for these people (e.g., Dr., Prof.) much more frequently than in human-written articles. Overall, our results highlight the need for increased public research on humanization, as more advanced methods can challenge both human and automated detectors while offering better training data for more robust detection systems.

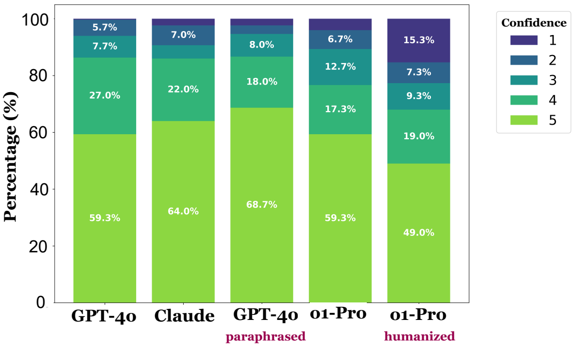


Figure 2: Expert confidence in their decisions drops when judging humanized articles generated by O1-PRO.

3 Fine-grained analysis of expert performance

The previous section establishes that human experts are highly accurate detectors of AI-generated text even in the presence of evasion attempts. Here, we first provide in-depth comparisons between experts and automatic detectors. Next, we perform a fine-grained coding of expert explanations into different categories (e.g., vocabulary, originality, formality), which allows us to examine what they focus on during the different experiments. Finally, we analyze differences between annotators and what they focus on when they are incorrect, highlighting implications for the future training of human annotators for AI-generated text detection.

Human experts vs. automatic detectors: We begin with a direct comparison between our human experts and state-of-the-art AI detectors. Specifically, we run five different detectors on our set of 300 articles:

- **Pangram** (Emi and Spero, 2024) is a closed-source commercial detector implemented via a Transformer classifier trained via an iterative process that uses hard negative mining and synthetic data to improve data efficiency. We run both their base model and a newer model (PANGRAM HUMANIZERS) trained to distinguish humanized data (Masrour et al., 2025). We directly use the labels (i.e., thresholds) produced by the Pangram API and do

[11] We only use O1-PRO for this experiment as the previous experiments show it was more difficult to detect than GPT-4O and CLAUDE-3.5-SONNET.

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 overexplaining 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."*
Formatting	15.0%	AI-generated formatting is overly consistent (e.g., fully capitalized headings, bolded lists, paragraphs of similar length).	**Human**: *""the formatting itself is varied, with topic headers focusing on lowercase lettering, longer explanations, and the bullet points not going into a title: one to two sentence explanation format ... "* **AI (o1-Pro)**: *"... it has a structured article format of commonly used headers..."*
Conclusions	13.1%	AI-generated conclusions tend to be repetitive and overly optimistic summaries, while human-written conclusions end more abruptly and less tidily.	**Human**: *"Both the introduction and the conclusion were humorous and unique."* **AI (Claude-3.5-Sonnet)**: *Most of the conclusion leads to that summarizing, flowery tone...*
Formality	12.3%	AI-generated text (especially without humanization) rarely contains filler words, contractions, slang, or abbreviations, favoring fully spelled-out terms and a polished tone.	**Human**: *"It includes filler words like 'just', 'very' and 'really'. Colloquial language like 'sucked'."* **AI (Claude-3.5-Sonnet)**: *"Phrases like 'maintains that extending habeas corpus rights to animals' and 'fundamentally alter the ability of accredited zoos to conduct their vital conservation work.' are so strictly formal, so tight and dense with wordage."*
Names & Titles	11.7%	LLMs frequently generate the same names regardless of the prompt (e.g., "Emily Carter," "Sarah Thompson"). Furthermore, if an article contains multiple people, they tend to refer to all of them with the same title (e.g., "Dr.") without offering unique details. Unique or context-specific names (including real brands or products) are associated with human writing.	**Human**: *"A couple of the experts also have quite unique names, and none of them are referred to as Dr. X (bonus points for none of them being named Emily)."* **AI (GPT-4o)**: *"However, the introduction of Dr. Sarah Thompson and Dr. Emily Carter (who by now has more than a lifetime's worth of qualifications), means that it has to be AI text."*
Tone	9.3%	The tone of AI-generated text is consistently neutral or positive, lacking depth or emotional variety compared to human writing.	**Human**: *"The article seems to display some bias against TikTok while trying to remain impartial or impartial-seeming. The same goes for its political slant which although subtle seems to be there."* **AI (o1-human)**: *"...the majority of the writing is filled with the same language it uses to describe everything - inspirational, stunning, essential, and resonating - using formal words, an inherent positivity bias, and a reflective, romantic tone that doesn't give details on why this topic matters, why we still don't know who she is, and whether or not there was any kind of controversy around the idea of women being on stage..."*
Introductions	7.3%	AI-generated introductions are usually generic or focus on scenic details without providing key background, while human introductions have more compelling hooks and relevant context.	**Human**: *"The introduction is unique as it starts with the subject of the article watching a movie rather than instantly explaining the entire point of the article."* **AI (Claude-3.5-Sonnet)**: *"The article has a very generic introduction and conclusion, especially the introduction, which essentially tells you what the entire article is about."*
Factuality	7.2%	AI-generated text contains factual inconsistencies or hallucinations more often than human writing.	**Human**: *"The article is very factual, with details about the dates, materials used, and sales portions."* **AI (GPT-4o-Para)**: *"Incorrect information delivered with customary AI confidence: the 'quad axel' is not another name for the backflip."*
Topics	3.1%	Unlike human authors, LLMs generally avoid darker or more mature topics (e.g., violence, graphic descriptions).	**Human**: *"Phrases throughout the article, including "burned to ashes and scattered at sea to prevent the crowd from venerating them as relics" and "to trample on a brass likeness of Jesus or the Virgin Mary—a blasphemous act." actively show what happened, the horrors and tragedies that occurred during that time."* **AI (GPT-4o)**: *"It spends very little time talking about the horrors of the disease, and instead focuses on future research, hopeful quotes, and potential cures, even referring to it as "embarking on a new chapter"."*
Other	2.6%	Other clues that do not fall into any of the above categories, often based on the annotator's intuition or overall impression.	**Human**: *"There are no clues here apart from the highlighted sentence which seems to have a human 'ring' to it."* **AI (o1-Humanized)**: *"The article feels slightly artificial but I caní really find any clear clues for it."*

Table 4: Taxonomy of clues used by experts to explain their detection decisions. For each category, we report the frequency of explanations that mention that category (regardless of if the annotator was correct) and provide examples of explanations for both human-written and AI-generated articles. While vocabulary and sentence structure form the most frequent clues, more complex phenomena like originality, clarity, formality, and factuality are also distinguishing features.

not award credit for the neutral label "Possibly AI".

- **GPTZero** (Tian and Cui, 2023) is also a closed-source commercial detector that runs a classifier sentence-by-sentence across the document. We directly use the binary labels produced by the GPTZero API.
- **Binoculars** (Hans et al., 2024) is an open-source detector that relies on the cross perplexity computed by two different language models to perform detection. We run Binoculars with the two FPR thresholds recommended by its authors ("Accuracy" and "Low FPR" modes).
- **Fast-DetectGPT** (Bao et al., 2023) is an open-source method that samples and scores many perturbations of the text to estimate conditional probability curvature. We threshold this method at a FPR of 0.05, set on our held-out dev set of 40 articles.
- **RADAR** (Hu et al., 2023) train an open-source classifier adversarially against a paraphraser. We threshold this method at a FPR of 0.05, set on our held-out dev set of 40 articles.

The middle rows of Table 3 show that only Pangram Humanizers (average TPR of 99.3% with FPR of 2.7% for base model) matches the human expert majority vote, and it also outperforms each expert individually. Pangram is near perfect on the first four experiments and falters just slightly on humanized o1-Pro articles, while GPTZero struggles significantly on o1-Pro with and without humanization. The open-source detectors degrade in the presence of paraphrasing and underperform both closed detectors by large margins on average. We again note that our experts are *untrained* at this detection task, and they could likely improve their individual performance if provided with feedback on their errors.

Coding expert explanations: Our automatic baselines are largely unexplainable: they either rely on combinations of various statistical properties of the text, or operate as opaque classifiers trained on large datasets. In contrast, we can easily solicit explanations from human annotators, and in this section we use GPT-4o to code these explanations into a schema (Table 4) developed by the authors after careful manual analysis. Details of the coding process used to label explanations can be found in §D.

Vocabulary-related clues are mentioned in the majority (53.1%) of all explanations, while sentence structure (35.9%), grammar (24.8%), and originality (23.7%) are also common. We note that many of these categories (e.g., originality, factuality, tone) are much more difficult to assess automatically than others (e.g., vocabulary), and these may currently be areas where humans have an advantage over automatic detectors.

When experts correctly detect AI-generated articles, what clues do they use? Figure 3 (*upper*) shows the frequency that clue categories are mentioned in explanations for which the expert makes the *correct* decision. We observe several interesting shifts as the complexity of the article generation setup increases: for example, 57.1% of correct explanations about o1-Pro articles mention vocabulary, compared to only 42.3% for the humanized o1-Pro articles prompted to avoid "AI vocab". Somewhat counterintuitively, explanations about paraphrased articles note AI vocab in 88% of explanations, compared to only 69.8% of non-paraphrased GPT-4o articles. Similarly, quotations are mentioned in 33.8% of explanations about paraphrased articles, a much higher rate than other configurations: a close reading of explanations reveals that experts flagged quotes that were always in the same format and style (e.g., only placed at the end of each paragraph). Originality is a common clue for detecting human-written text, highlighting the gulf in creativity between humans and LLMs (Chakrabarty et al., 2024).

Annotators don't always focus on the same clues: Our experts focus on different properties of the text to arrive at their decisions. Annotator 1 is the only one to pick up on "AI names": in fact, 63.3% of GPT-4o and 70% of Claude-3.5-Sonnet articles include either the name Emily or Sarah.[12] Annotator 2 and 3 emphasize linguistic features, such as humans not always following "proper" writing conventions or the tendency of LLMs to always list examples in groupings of three. Annotator 4 focuses more on the flow of the articles, analyzing the specificity of detail and motivation behind an article, while Annotator 5 frequently mentions whether a quotation sounds natural or not. Further

[12] o1-Pro favors names of real people instead of fictional names when generating articles.

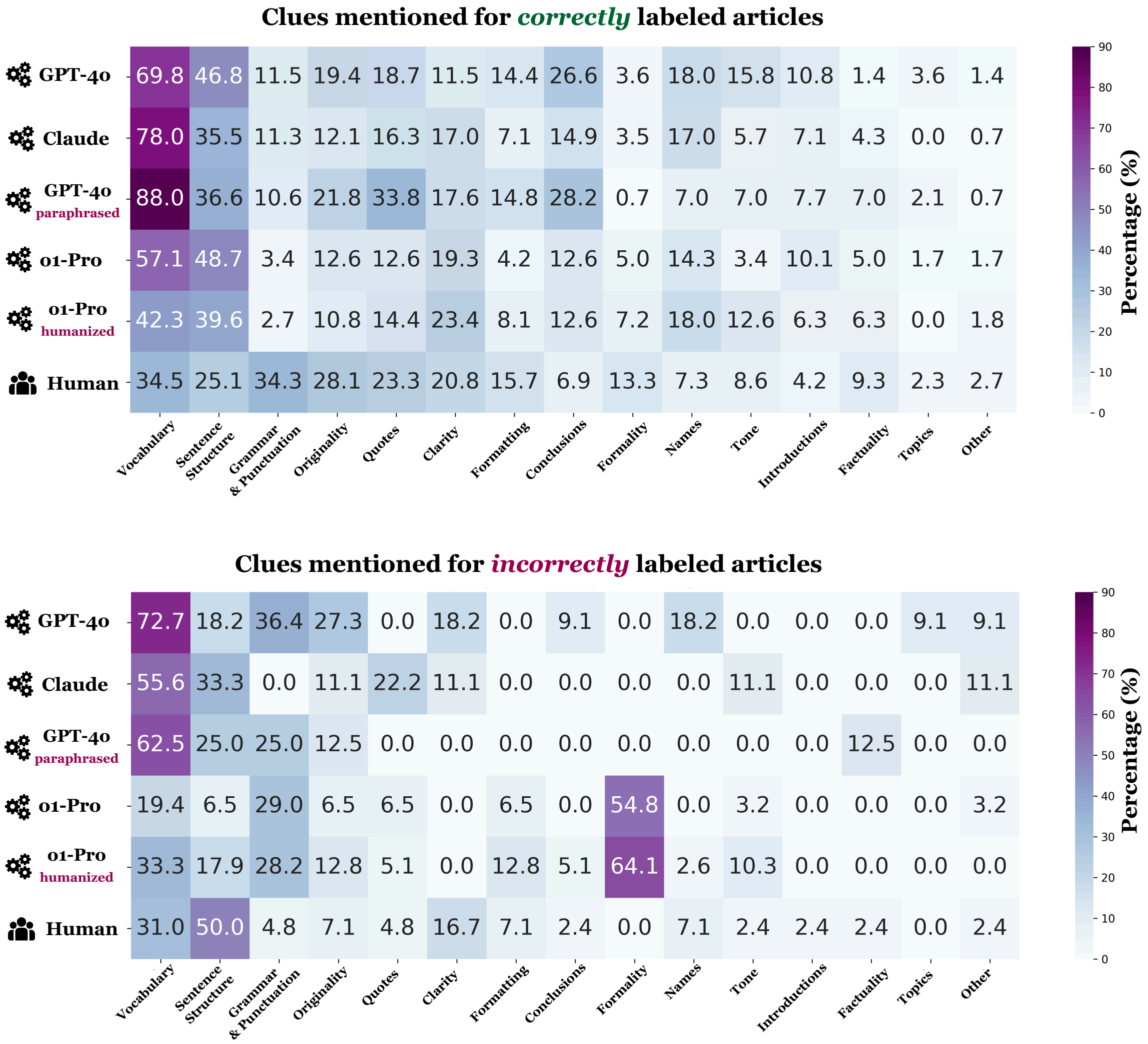


Figure 3: *(Top)* A heatmap displaying the frequency with which annotators mentioned specific categories in their explanations when they were correct. Interestingly, vocabulary becomes a less frequent clue for O1-PRO-generated articles, especially with humanization. *(Bottom)* Same as above, except only computed over explanations when experts were incorrect. Formality is a big source of misdirection for O1-PRO articles, while fixating on sentence structure can lead experts to false positives. Details of each category can be found in Table 4.

analysis of individual performance can be found in §D.2. This diversity partly explains why an ensemble of expert annotators performs so well, and it also suggests room for training so that all experts are at least aware of most distinguishing features.

When experts are incorrect, what clues lead them astray? Figure 3 (*lower*) shows the frequency that clue categories are mentioned in explanations for which experts make *incorrect* decisions. Despite near-perfect majority vote performance, experts make several errors individually, particularly with the qualitative change in text produced by O1-PRO. For Annotator 3, who relies heavily on linguistic features, O1-PRO's frequent use of contractions and colloquialisms dramatically lowered their detection rate: they mention formality in *none* of their explanations for Experiments 2 and 3, compared to 66.7% of O1-PRO and 83.3% of humanized O1-PRO explanations as shown in Figure 11. Studying expert false positives is also insightful Figure 3 (*lower*): 31% of explanations for false positives mention vocabulary, typically when human-written content contains words like “delve” and “crucial” that our experts associate with AI-generated text. More prominently, 50% of false positives focus on sentence structure, when humans write stylistically-similar sentences to those preferred by LLMs (e.g., contiguous blocks of sentences with similar lengths, multiple lists of three items).

4 Can LLMs be prompted to mimic human expert detectors?

Our experiments so far highlight the benefits of hiring humans to perform AI-generated text detection: in aggregate, they are more accurate and robust than all automatic detectors we tried. Additionally, they can provide detailed explanations of their decision-making process, unlike all of the automatic detectors in our study. However, an obvious drawback is that hiring humans is expensive and slow: on average, we paid $2.82 per article including bonuses, and we gave annotators roughly a week to complete a batch of 60 articles. In this section, we prompt LLMs to imitate our expert annotators, providing the guidebook used in our humanization experiments (§2.5) to the model and asking it to produce an explanation and a label. While the approach shows promise, outperforming detectors such as Binoculars and RADAR, it fails to reach the performance of our human experts as well as advanced automatic detectors like Pangram.

Implementing a prompt-based detector: Using our guidebook for AI detection (Table 11), we prompt an off-the-shelf LLM to decide whether a candidate text is human-written or AI-generated and explain its decision based on the criteria in the guidebook (see Table 21 for prompt template). While we can control this detector's behavior by modifying the prompt (e.g., provide explanations before producing a final label), it does not output a single score, which makes setting a false positive threshold impossible.[13] We derive the final output of our detector by using the prediction from the detection model in a deterministic configuration.

Prompt-based detection shows promise but struggles with humanization and high false positive rates: We implement our prompt-based detector by zero-shot prompting GPT-4o-2024-11-20 and o1,[14] comparing the effects of chain-of-thought (CoT) prompting (Wei et al., 2022) and the inclusion of the detection guide on its performance. For GPT-4o, the best detector configuration is achieved with both CoT and the guidebook (average TPR 78%), while o1 is best with CoT but without the guidebook (54%). We observe that o1 is more conservative than GPT-4o, demonstrating lower FPRs on average at the cost of lower TPR. However, it is also more capable at detecting humanized text. On average, our best configuration with GPT-4o-2024-11-20 performs comparably to Binoculars and Fast-DetectGPT, although obviously at a much higher cost. The gap between this method and humans demonstrates that we have a long way to go in order to "teach" LLMs how to detect AI-generated text like humans do; we speculate that fine-tuning may be helpful to close the gap.

5 Related work

Our work builds on prior research centering around AI-generated text detection using both human and automatic approaches, as well as recent work on evading such detectors.

Human detection of AI-generated text: Our study most closely resembles several papers published prior to ChatGPT's release. Existing work notes that while naïve annotators do not reliably detect AI-generated texts (Ippolito et al., 2020; Brown et al., 2020; Clark et al., 2021; Karpinska et al., 2021), some annotators manage to perform very well (Ippolito et al., 2020). More recently, these findings have been reproduced in other domains such as Italian news articles (Puccetti et al., 2024) and poetry (Porter and Machery, 2024). Dugan et al. (2020) and Dugan et al. (2022) address a more complex task where humans and machines are tasked to identify boundary between human and AI-generated text in mixed documents.

Automatic detection of AI-generated text: Successful automatic detection methods are typically either perplexity-based (Mitchell et al., 2023; Bao et al., 2023; Hans et al., 2024) or trained classifiers (Solaiman et al., 2019; Emi and Spero, 2024; Verma et al., 2023). Some detectors have focused on sentence-level detection (Kushnareva et al., 2024; Wang et al., 2023), emphasizing the need for more detailed and explainable detection. Dugan et al. (2024), Li et al. (2024), and (Zhang et al., 2024) attempt to compare the performance of detectors in different domains and adversarial attack settings. However, most automatic detectors are unreliable in the face of attacks such as paraphrasing (Krishna et al., 2023; Sadasivan

[13] We do not experiment with thresholding the token-level probabilities of the labels "AI" vs. "Human", since these are inaccessible with closed LLMs, but this is a promising route for future research on open prompt-based detectors.

[14] We set `temperature=0` for GPT-4o and reasoning efforts to 'medium' for o1 (note that OpenAI has fixed `temperature=1` for o1). Further experiments can be found in Table 22.

et al., 2024) and highly sensitive to stylistic factors (Doughman et al., 2025).

Evading detection: Increasing LLM usage among the general population has spurred corresponding interest in humanization of AI-generated text.[15] In fact, many commercial humanization services exist to fill this demand.[16] Initial research has proposed several humanization attacks to evade automatic detection (Wang et al., 2024b; Lu et al., 2023; Shi et al., 2023; Wang et al., 2024a), while Zhou et al. (2024) and Masrour et al. (2025) improve the robustness of automatic detectors to humanized content.

Analyzing differences between human-written and AI-generated text: Many frameworks group errors in machine-generated text into categories similar to the ones we use in this paper (Gehrmann et al., 2019; Dou et al., 2021). More recently, Ma et al. (2023) discover gaps in terms of depth and content quality between scientific AI-generated and human-written text. Shaib et al. (2024) introduces syntactic templates, noting the repetitiveness of AI-generated text in comparison to human-written references. More closely related to our work is that of Ji et al. (2024), who qualitatively categorize human detection explanations but do not employ expert annotators. Jakesch et al. (2023) explores the heuristics humans use to detect AI-generated text, with some similar conclusions to ours (e.g., humans are likely to associate more informal text with human writing).

6 Conclusion

Our paper demonstrates that a population of "expert" annotators—those who frequently use LLMs for writing-related tasks—are highly accurate and robust detectors of AI-generated text without any additional training. The majority vote of five such experts performs near perfectly on a dataset of 300 articles, outperforming all automatic detectors except the commercial Pangram model (which the experts match). Analysis of explanations provided by our expert annotators reveals that they pick up on not just vocabulary and sentence structure-related clues but also more complex properties like originality, factuality, and tone. We observe that each expert organically focuses on different aspects of the text, and we conjecture that with explicit training, human annotators can be made even more robust to advances in LLMs as well as evasion tactics (e.g., humanization). Future work can also explore human annotators working alongside automatic detectors like Pangram to improve detection accuracy and explainability. We find it apt to end with a comment from one of our experts about a particularly formulaic conclusion generated by GPT-4o:

> *This time I went to the end of the piece and said: "Hello, AI." There it was in all its glory: the "testament" serving "as a beacon of hope and inspiration" and "demonstrating" to us humans "that anything is possible." Sometimes I feel sorry for AI—it must have a dreary time trying to satisfy its human interlocutor's desire to "showcase" advocacy, social change, inclusivity, gender equality, equity, and representation in an essay of fewer than a thousand words.*
>
> — ANNOTATOR 5

Limitations

Our study is limited to articles in American English, chosen for their consistent formatting and high quality (i.e., professionally written and proofread). We also did not investigate factual accuracy, as it did not appear to be a significant cue for our annotators, who covered a broad range of topics. Finally, while we selected articles from reputable sources, there remains a possibility that some included AI-generated edits beyond our scope of detection.

Ethical Considerations

This study was reviewed by the UMass Institutional Review Board (IRB #5927) and deemed exempt. All annotators were briefed on the purpose of the project and provided informed consent prior to participating. Those who wished to be acknowledged by name explicitly agreed to do so in their consent forms. We ensured fair compensation for annotators in recognition of their time and expertise. We acknowledge the potential risks of misinformation and hallucinated content, especially when AI outputs are presented as human-written. Our goal is to examine these issues and inform best practices,

[15] Relevant discussions on Reddit: "AI Humanizer Recommendations?" (2024), "How to humanize ai-generated texts?" (2024), "How to humanize the AI generated content?" (2024)

[16] Some humanization services include Undetectable AI, MyEssayWriter.ai and Stealth Writer.

rather than endorse or facilitate deceptive uses of AI.

Acknowledgments

We would like to extend our gratitude to our Upwork expert annotators for their dedicated efforts, and link their Upwork accounts for future work requiring human expertise on AI text detection: Benjamin Stewart,[17] Elizabeth Bergman,[18] Lynette Stewart,[19] Roberta King,[20] and Shannon West.[21] We thank Pangram and GPTZero for providing credits to access their APIs. We are also grateful to the members of the UMass NLP lab for their insightful input. This project was partially supported by awards IIS-2202506 and IIS-2312949 from the National Science Foundation (NSF) as well as an award from Open Philanthropy.

References

Mike Allen. 2017. *The SAGE encyclopedia of communication research methods*. SAGE Publications, Inc, 2455 Teller Road, Thousand Oaks California 91320.

Anthropic. 2023. Claude: A language model by anthropic. Accessed: 2025-01-20.

Anthropic. 2024. Claude 3 model card addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf. Accessed: 2024-12-30.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Creston Brooks, Samuel Eggert, and Denis Peskoff. 2024. The rise of AI-generated content in Wikipedia. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 67–79, Miami, Florida, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165 [cs].

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Frederick Wieting, and Mohit Iyyer. 2024. PostMark: A robust blackbox watermark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8969–8987, Miami, Florida, USA. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Annual Meeting of the Association for Computational Linguistics*.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2025. Exploring the Limitations of Detecting Machine-Generated Text. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4274–4281, Abu Dhabi, UAE. Association for Computational Linguistics.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text.

[17] Benjamin's Upwork: https://www.upwork.com/freelancers/~01cfc42298090772b8?mp_source=share

[18] Elizabeth's Upwork: https://www.upwork.com/freelancers/~01e0d3569ef95d391d?mp_source=share

[19] Lynette's Upwork: https://www.upwork.com/freelancers/~01cc5594cf2405ea5e?mp_source=share

[20] Roberta's Upwork: https://www.upwork.com/fl/robertaking?mp_source=share

[21] Shannon's Upwork: https://www.upwork.com/freelancers/~01daaa89e049285294?mp_source=share

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 189–196, Online. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2022. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *AAAI Conference on Artificial Intelligence*.

Bradley Emi and Max Spero. 2024. Technical Report on the Pangram AI-Generated Text Classifier. *arXiv preprint*. ArXiv:2402.14873 [cs].

Henrique Da Silva Gameiro, Andrei Kucharavy, and Ljiljana Dolamic. 2024. LLM Detectors Still Fall Short of Real World: Case of LLM-Generated Short News-Like Posts. *arXiv preprint*. ArXiv:2409.03291 [cs].

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17519–17537. PMLR.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: robust ai-text detection via adversarial learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Jiazhou Ji, Ruizhe Li, Shujun Li, Jie Guo, Weidong Qiu, Zheng Huang, Chiyu Chen, Xiaoyu Jiang, and Xinru Lu. 2024. Detecting Machine-Generated Texts: Not Just "AI vs Humans" and Explainability is Complicated. *arXiv preprint*. ArXiv:2406.18259 [cs].

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085, Miami, Florida, USA. Association for Computational Linguistics.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2024. Delving into ChatGPT usage in academic writing through excess vocabulary. *arXiv preprint*. ArXiv:2406.07016 [cs].

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, pages 27469–27500. Curran Associates, Inc.

Laida Kushnareva, Tatiana Gaintseva, Dmitry Abulkhanov, Kristian Kuznetsov, German Magai, Eduard Tulchinskii, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. 2024. Boundary detection in mixed AI-human texts. In *First Conference on Language Modeling*.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated Text Detection in the Wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Ning Lu, Shengcai Liu, Ruidan He, and Ke Tang. 2023. Large language models can be guided to evade ai-generated text detection. *Trans. Mach. Learn. Res.*, 2024.

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. Ai vs. human – differentiation analysis of scientific content generation.

Elyas Masrour, Bradley Emi, and Max Spero. 2025. DAMAGE: Detecting Adversarially Modified AI Generated Text. *arXiv preprint*. ArXiv:2501.03437 [cs].

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*.

OpenAI. 2024. OpenAI o1 System Card. *arXiv preprint*. ArXiv:2412.16720 [cs].

B Porter and Edouard Machery. 2024. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14.

Giovanni Puccetti, Anna Rogers, Chiara Alzetta, Felice Dell'Orletta, and Andrea Esuli. 2024. AI "News" Content Farms Are Easy to Make and Hard to Detect: A Case Study in Italian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15312–15338. ArXiv:2406.12128 [cs].

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2024. Can AI-Generated Text be Reliably Detected? *arXiv preprint*. ArXiv:2303.11156 [cs].

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C. Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Conference on Empirical Methods in Natural Language Processing*.

Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2023. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics*, 12:174–189.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. *ArXiv*.

Edward Tian and Alexander Cui. 2023. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods".

Vivek Kumar Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models. In *North American Chapter of the Association for Computational Linguistics*.

James Wang, Ran Li, Junfeng Yang, and Chengzhi Mao. 2024a. RAFT: Realistic Attacks to Fool Text Detectors. *arXiv preprint*. ArXiv:2410.03658 [cs].

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. Seqxgpt: Sentence-level ai-generated text detection. In *Conference on Empirical Methods in Natural Language Processing*.

Tianchun Wang, Yuanzhou Chen, Zichuan Liu, Zhanwen Chen, Haifeng Chen, Xiang Zhang, and Wei Cheng. 2024b. Humanizing the Machine: Proxy Attacks to Mislead LLM Detectors. *ArXiv*. Publisher: arXiv Version Number: 1.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *NAACL-HLT*.

Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, Chao Cao, Hanqi Jiang, Hanxu Chen, Yiwei Li, Junhao Chen, Huawen Hu, Yihen Liu, Huaqin Zhao, Shaochen Xu, Haixing Dai, Lin Zhao, Ruidong Zhang, Wei Zhao, Zhenyuan Yang, Jingyuan Chen, Peilong Wang, Wei Ruan, Hui Wang, Huan Zhao, Jing Zhang, Yiming Ren, Shihuan Qin, Tong Chen, Jiaxi Li, Arif Hassan Zidan, Afrar Jahin, Minheng Chen, Sichen Xia, Jason Holmes, Yan Zhuang, Jiaqi Wang, Bochen Xu, Weiran Xia, Jichao Yu, Kaibo Tang, Yaxuan Yang, Bolun Sun, Tao Yang, Guoyu Lu, Xianqiao Wang, Lilong Chai, He Li, Jin Lu, Lichao Sun, Xin Zhang, Bao Ge, Xintao Hu, Lian Zhang, Hua Zhou, Lu Zhang, Shu Zhang, Ninghao Liu, Bei Jiang, Linglong Kong, Zhen Xiang, Yudan Ren, Jun Liu, Xi Jiang, Yu Bao, Wei Zhang, Xiang Li, Gang Li, Wei Liu, Dinggang Shen, Andrea Sikora, Xiaoming Zhai, Dajiang Zhu, and Tianming Liu. 2024. Evaluation of OpenAI o1: Opportunities and Challenges of AGI. *arXiv preprint*. ArXiv:2409.18486 [cs].

Ying Zhou, Ben He, and Le Sun. 2024. Humanizing machine-generated content: Evading ai-text detection through adversarial attack. In *International Conference on Language Resources and Evaluation*.

Tiffany Zhu, Kexun Zhang, and William Yang Wang. 2024. Embracing AI in Education: Understanding the Surge in Large Language Model Use by Secondary Students. *arXiv preprint*. ArXiv:2411.18708 [cs] version: 1.

A Human Evaluation

In this section of the appendix we provide additional details about our experts the data collection pipeline.

Annotators: The annotations for the experiment 1 were done by 5 annotators recruited on Upwork. All annotators are native English speakers from the US or South Africa. All annotators hold university degrees, worked in varying professions, and had varying levels of familiarity with AI assistants like ChatGPT. One had never used AI, 2 had little experience, and 2 used AI every day. Our five expert annotators are native English speakers hailing from the US, UK, and South Africa. Most work as editors, writers, and proofreaders and have extensively used AI assistants. See Table 5 for more information about annotators.

Collecting Human Annotations All annotators were required to read the guidelines (Figure 4) and sign a consent form (Figure 5) prior to the labeling task. Collecting all labels usually required additional communication with the annotators, resulting in about 20 hours of work from the author involved in this process. We estimate that the annotators were able to read and label between 8 and 12 articles per hour based on self-reported time and records in the spreadsheets. Figure 6 shows the interface annotators use to complete the annotation process. They read and highlight an article, then complete the annotation by providing their decision, confidence score, and explanation. Since completing annotations for 60 article batches takes a long time (estimated 6-8 hours of work), we have implemented an interface that made it possible for the annotators to save their work and come back at any time, allowing annotators to allocate their time as they saw best. Annotators were given 1 week to complete batches of 60 articles, with more time allowed when needed. Note that unlike Figure 1, annotators did *not* see the titles of the article when completing annotations. This decision was made to prevent obvious linkage of AI-generated and human-written text pairs.

A.1 Finding Expert Annotators

The annotations for understanding the limitations of humans to detect AI-generated texts were done by 5 annotators who passed performance requirements based on the article expiriment of subsection 2.1. Only one annotator from the original experiment met these requirements. We recruited 10 more native English speakers on Upwork to take a 5 question sample of the original article task, questions which at most 2 out of 5 of the original annotators got correct. These 10 recruited annotators all had experience editing LLM generated content, frequently used LLMs, and had a professional background in editing or writing. Those who got at least 4 out of 5 correct (80%) given the rest of the 60 articles, and annotators who got at least 90% correct (54/60) were considered to be experts and recruited for the rest of the labeling rounds. 5 out of 10 annotators passed the 5 question trial; out of those 5, 4 passed the 60 question set.

B Dataset

In this section of the appendix we provide more details on our article corpus (B.1), how our AI articles were generated (B.2).

B.1 Article Corpus

Here we include more details about the articles collected for this study. Table 6 lists all publications of articles included in the corpus,[22] with section distribution presented in Figure 7. Table 7 provides the statistics for articles by publication.

B.2 AI Text Generation

We generate the articles by prompting the models with the articles title, subtitle, approximate length, publication, and section. For stories, we prompt models with the title of the reddit thread. All closed-source models were prompted using the provider's API, [23]. All models were prompted and

[22] All publications in the corpus were purchased by the researchers.

[23] The estimated cost for generating all data with each model is as follows: GPT-4O $1.85USD, O1-PRO $3.81 USD, CLAUDE-3.5-SONNET $0.51 USD

Annotator	Education Level	English Dialect	LLM Usage	AI Models Used	Occupation
Non-Expert					
-	Professional Degree or Doctorate	American English	Few times	ChatGPT	Writer
-	Bachelor's Degree	American English	Daily	ChatGPT, Gemini, Llama	Finance/Operations
-	Some college, no degree	American English	Few times	None	Transcription, editing, data entry
-	Bachelor's Degree	American English	Never	None	English tutor, copywriter, author
Expert					
Annotator 1	Bachelor's Degree	South African English	Daily	ChatGPT, Microsoft Copilot	Freelance editing, writing, proofreading
Annotator 2	Bachelor's Degree	South African English	Weekly	ChatGPT	Editing and Proofreading
Annotator 3	Master's Degree	British English	Daily	ChatGPT	Copyeditor and Proofreader
Annotator 4	Bachelor's Degree	American English	Weekly	ChatGPT, Llama, Huggingface models	Freelancer Content Writer
Annotator 5	Master's Degree	South African English	Weekly	ChatGPT, Claude	Language teacher

Table 5: Survey of Annotators, specifically their backgrounds relating to LLM usage and field of work. Note that expert Annotator #1 was one of the original 5 annotators (along with the 4 non-experts) and remained an annotator for all expert trials.

Publication	Example Sections	Pub. Date Range
Associated Press	Science, Oddities, Animals	May 15, 2024 - Dec 5, 2024
Discover Magazine	Mind, The Sciences, Environment, Planet Earth	July 10, 2024 - Nov 16, 2024
National Geographic	Animals, Environment, Science, History & Culture, Travel	Feb 8, 2023 - Nov 19th, 2024
New York Times	US News, Science, Travel, Arts	July 19, 2024 - Dec 6, 2024
Readers Digest	Knowledge, Holidays	March 16, 2023 - Nov 15, 2024
Scientific American	Mind & Brain, Social Sciences, Technology	May 3, 2024 - Nov 22, 2024, 2024
Smithsonian Magazine	Smart News, Mind & Body, History, Innovation, Travel, Science	Oct 7, 2022 - Dec 8, 2024
Wall Street Journal	Science, Personal Technology, Workplace	July 7, 2023 - Oct 29, 2024

Table 6: List of publications included in HUMAN DETECTORS. The section is provided as listed as the section of the publication website where the article was published. All articles were taken from publications that wrote using American English.

articles with the prompt presented in Table 8. Instructions for rounding to an approximate length were included to ensure that articles on the same topic would be of similar length. Statistics on the distribution of lengths by trial are presented in Figure 8

Paraphrasing Attack We use the sentence-level paraphrasing approach outlined in PostMark (Chang et al., 2024), changing the exact language in the prompts slightly for our use case. This approach paraphrases the text on a sentence by sentence level. For the initial sentence, only that sentence is given to be paraphrased. For all following sentences, the portion of the text already paraphrased is passed to the LLM, intending to improve the overall flow of the paraphrased article. We use GPT-4o as the paraphraser model with a temperature set to 0. The prompt for the initial sentence can be found in Table 9 and the prompt for all following sentences can be found in Table 10.

Humanization Efforts: To humanize the articles in §2.5, we tried many prompt-based approaches before settling on a prompting framework that was capable of evading the first author's own AI-detection skills. Initially, we tried the following approaches that humanized already generated texts:

1. **Generate, then Humanize**: Firstly, we tried to replicate what we believe most students or novice LLM users would do to humanize their AI-generated text. We asked O1-PRO to first generate an article using the prompt template depicted in Table 8, then instructed the model to make it sound more human. Qualitative analysis from the first author found this method to result in an effect similar to paraphrasing.

2. **Step-by-Step**: To add a fine-grained approach to the humanization efforts, we next tried a step-by-step approach to humanization. We generated a base article, then iteratively prompted the model to humanize the article, focusing on a different element at each step. For example, we first asked it to make the article more creative, then alter the tone. Final steps included editing for grammar and replacing typical AI vocab. While this was better, the first author observed that the numerous calls to LLMs was adding *more* characteristics of AI, so we abonded this method.

3. **Two-step Humanization**: Next, we tried tak-

	Articles							
	Associated Press *(n=15)*	**Discover** *(n=20)*	**National Geographic** *(n=25)*	**New York Times** *(n=20)*	**Readers Digest** *(n=15)*	**Scientific American** *(n=15)*	**Smithsonian Magazine** *(n=25)*	**Wall Street Journal** *(n=15)*
TOKENS (tiktoken)								
Mean	860.0	949.5	1070.7	1009.1	865.6	874.2	922.0	927.5
St. Dev	240.7	229.9	203.2	138.4	171.9	347.1	210.7	202.5
Max	1236.0	1463.0	1497.0	1254.0	1318.0	1760.0	1454.0	1336.0
Min	442.0	561.0	710.0	645.0	634.0	455.0	532.0	608.0
WORDS (whitespace)								
Mean	676.4	734.6	809.7	784.1	675.3	684.2	708.1	720.5
St. Dev	180.8	174.5	149.9	106.3	136.6	265.9	160.0	153.9
Max	1026.0	1117.0	1102.0	951.0	1008.0	1352.0	1095.0	1019.0
Min	342.0	415.0	536.0	513.0	501.0	362.0	388.0	491.0

Table 7: Number of tokens and words across articles by source.

Article Generation Prompt

```
You are given the following title and subtitle of
a general article from the YOUR SECTION section
and asked to write a corresponding article of
around YOUR WORD COUNT words. Include quotations
from relevant experts and make sure the article
is concise and easily understandable to a lay
audience.

Title: YOUR TITLE
Subtitle: YOUR SUBTITLE
```

Table 8: Prompt Template for Experiment 2 Paraphrasing

Paraphrase Prompt (Initial Sentence Only)

```
Paraphrase the given sentence. Only return the
paraphrased sentence in your response. Make it
seem like a human wrote the article and that
it is from the YOUR SECTION section of YOUR
PUBLICATION.

Sentence to paraphrase: YOUR SENTENCE
Your paraphrase of the sentence:
```

Table 9: Prompt Template for Experiment 2 Paraphrasing

ing a 2-step approach to humanization. We gave the humanizing LLM the AI detection guide found in Table 11, and asked the LLM to make a list of everything identifiable as AI-generated in the article, and suggested edits. Then, giving the LLM the list of suggested edits, we prompted the LLM to edit the article to make it sound more human-written. Out of the humanization efforts, this yielded the best results.

Evader Details: While approaches to humanization were improving, we found many of the largest identifiable traits of AI were due to the base generation. We decided to switch our approach from humanizing already existing AI-generated text to having an LLM generate the article in a humanized fashion all in one step. Our final prompt includes the set of instructions, examples of reference articles, detection guide, and the initial article prompt. The evader prompt template can be found in Table 13. The examples used were human-written and AI-generated articles from Experiments 1,2,3, and 4 from the same source. For example, when evading an article from New York Times, we provided all human-written New York Times articles and AI-generated articles based off those article titles.

Paraphrase Prompt

```
Paraphrase the given sentence. Only return the
paraphrased sentence in your response. Make it
seem like a human wrote the article and that
it is from the YOUR SECTION section of YOUR
PUBLICATION.

Previous context: YOUR PARAPHRASED ARTICLE SO
FAR
Sentence to paraphrase: YOUR SENTENCE
Your paraphrase of the sentence:
```

Table 10: Prompt Template for Experiment 2 Paraphrasing

C Pilot Study

In this section, we provide details on the pilot study, where five annotators with varying frequencies of familiarity with LLMs detect if texts are human-written or AI-generated. The pilot study follows the same annotation process described in §2. In the study, annotators completed two rounds of testing, one using the same articles as identified in §2.1, another using stories from the subreddit r/WritingPrompts.

Story Corpus: Here we include more details about the stories collected for this study. We collect 30 stories from r/WritingPrompts. We generate corresponding AI-generated stories with the prompt in Table 15.

Story Results : Overall, nonexperts faired better on stories than articles, with an average TPR of 69.3% (including performance of the expert annotator). The nonexperts had a TPR of 62.5% not including our expert annotator, as shown in Table 16. We found that many of the differences spotted were superficial due to the very informal nature of reddit, compared to our prompt which did not instruct the LLM to generate stories of that nature. In reddit, people use acronyms, for example shortening 'you' to 'u', type in ALL CAPS, or generally write in a very R-rated manner. While these are real differences between AI-generated and human-written texts, we wanted to pursue clues humans could use in settings where a person was genuinely attempting to pass of AI-generated writing as human-writtten, such as students writing an essay or someone publishing a news article. Future work may explore human detection abilities on fictional work, where the human-written references are more edited than the reddit thread stories.

D Comment Analysis

In this section, we outline the framework for analyzing expert explanations and provide additional analysis of explanations.

D.1 Categorization of Comments

To categorize explanation comments, we first define the explanation categories found in Table 4. The first two authors individually annotated a sample of 25 expert explanations. The sample is a stratified sample of one comment per expert per each of the five experiments. Authors then came to an agreement on final human labels, refining categories as needed. We then prompt GPT-4o-2024-11-20 with the prompt found in Table 17 to categorize the sample explanations. Once the prompt was able to classify the majority of the sample explanations in alignment with the authors labels, we prompt GPT-4o-2024-11-20 to categories the explanations of all 1500 explanations. Each explanation can contain labels for multiple categories, since most explanations touch on multiple reasons of why a text is AI-generated. The total cost of classifying comments using GPT-4o was **$6.02 USD**.

D.2 Individual Experts Commentary

Each expert had clues they favored using throughout all experiments. Annotator 1,whose category mention frequencies can be found in Figure 9, Figure 10 depicts Annotator 2 comments, Figure 11 shows Annotator 3 comments, Figure 12 shows Annotator 4 comments and Figure 13 has commentary frequencies from Annotator 5. The individual heatmaps highlight the range of clues used by annotators, who individually had no reference of clues other experts used.

E Automatic Detection

E.1 Explainable Detection

In this section we detail the prompt based detector used for experiments in §4.

Implementation Details The prompt-based detector is set up to mimic how humans think about AI-generated text detection rather rather than how current automatic detectors do.

- **Zero-Shot**: To find a baseline performance of detector models, we prompt the model using the template in Table 18 to return if the candidate text is Human-written or AI-generated.

- **Zero-Shot + COT**: We prompt the model using the template in Table 19 to return if the candidate text is Human-written or AI-generated and an explanation of *why* they text is human-written. This ablation was conducted to understand the effect explanations may have on over LLM detection performance.

- **Zero-Shot + Guide**: In this experiment, we prompt the model using the template in Table 20 to return if the candidate text is Human-written or AI-generated. This experiment serves to understand the effect including the guide in the prompt has on LLM performance.

- **Zero-Shot + COT + Guide**: In this experiment, we prompt the model using the template in Table 21 to return if the candidate text is Human-written or AI-generated and an explanation of *why* they text is human-written. THis set up is set up to fully mimic how a human detects, thinking through clues and providing explanations of what makes them think a text is either human-written or AI-generated.

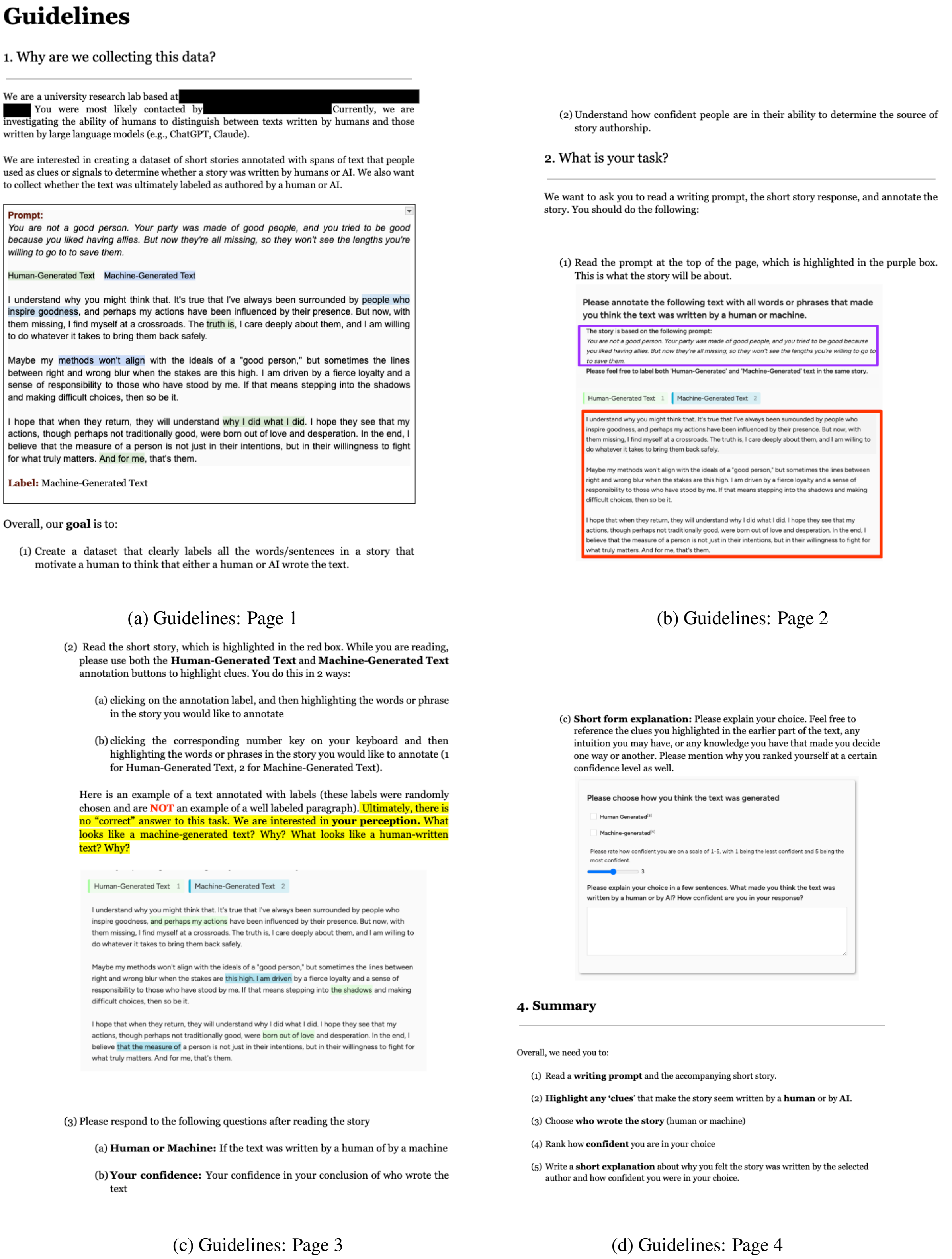
We observe that expert annotators perform better when using a majority vote strategy. Future work in prompt-based detectors could employ majority voting as a strategy to increase detection accuracy and avoid false positives.

E.2 Results

Thresholding for Automatic Detectors: Some automatic detectors benchmarked in Table 3 and Table 22 did not provide suggested thresholds for model usage. In these scenarios, we test models on a held-out test set 40 human-written articles, finding a threshold for a 5% FPR. The respective thresholds used were 0.6051510572 for RADAR, 0.96 for Fast-DetectGPT, and 0.8963184953 for e5-lora.

Detector Model We initially test GPT-4o-2024-08-06, GPT-4o-2024-11-20, o1, and CLAUDE-3.5-SONNET as potential detector models. Based on early results, we continue with ablations of ø1 due to its low FPR and GPT-4o-2024-11-20 due to having the highest TPR. The cost of all o1 detection experiments on the 300 texts is $100.60 USD and $25.24 USD for GPT-4o-2024-11-20 detection experiments.

Additional Results: While testing both GPT-4o-2024-08-06 and GPT-4o-2024-11-20 as detectors, we find extremely varied performance, with GPT-4o-2024-11-20 performing much better, particularly at detecting paraphrased and o1 articles. CLAUDE-3.5-SONNET failed to detect content generated by itself and o1Ẇe also tested e5-lora, one of the top performing models on the Raid (Dugan et al., 2024) benchmark, which we found was unable to correctly classify any AI-generated articles at a FPR of 5%. See all additional results in Table 22.



Guidelines

1. Why are we collecting this data?

We are a university research lab based at [redacted] You were most likely contacted by [redacted] Currently, we are investigating the ability of humans to distinguish between texts written by humans and those written by large language models (e.g., ChatGPT, Claude).

We are interested in creating a dataset of short stories annotated with spans of text that people used as clues or signals to determine whether a story was written by humans or AI. We also want to collect whether the text was ultimately labeled as authored by a human or AI.

Prompt:
You are not a good person. Your party was made of good people, and you tried to be good because you liked having allies. But now they're all missing, so they won't see the lengths you're willing to go to to save them.

Human-Generated Text Machine-Generated Text

I understand why you might think that. It's true that I've always been surrounded by people who inspire goodness, and perhaps my actions have been influenced by their presence. But now, with them missing, I find myself at a crossroads. The truth is, I care deeply about them, and I am willing to do whatever it takes to bring them back safely.

Maybe my methods won't align with the ideals of a "good person," but sometimes the lines between right and wrong blur when the stakes are this high. I am driven by a fierce loyalty and a sense of responsibility to those who have stood by me. If that means stepping into the shadows and making difficult choices, then so be it.

I hope that when they return, they will understand why I did what I did. I hope they see that my actions, though perhaps not traditionally good, were born out of love and desperation. In the end, I believe that the measure of a person is not just in their intentions, but in their willingness to fight for what truly matters. And for me, that's them.

Label: Machine-Generated Text

Overall, our **goal** is to:

(1) Create a dataset that clearly labels all the words/sentences in a story that motivate a human to think that either a human or AI wrote the text.

(a) Guidelines: Page 1

(2) Understand how confident people are in their ability to determine the source of story authorship.

2. What is your task?

We want to ask you to read a writing prompt, the short story response, and annotate the story. You should do the following:

(1) Read the prompt at the top of the page, which is highlighted in the purple box. This is what the story will be about.

Please annotate the following text with all words or phrases that made you think the text was written by a human or machine.

The story is based on the following prompt:
You are not a good person. Your party was made of good people, and you tried to be good because you liked having allies. But now they're all missing, so they won't see the lengths you're willing to go to to save them.
Please feel free to label both 'Human-Generated' and 'Machine-Generated' text in the same story.

Human-Generated Text 1 Machine-Generated Text 2

I understand why you might think that. It's true that I've always been surrounded by people who inspire goodness, and perhaps my actions have been influenced by their presence. But now, with them missing, I find myself at a crossroads. The truth is, I care deeply about them, and I am willing to do whatever it takes to bring them back safely.

Maybe my methods won't align with the ideals of a "good person," but sometimes the lines between right and wrong blur when the stakes are this high. I am driven by a fierce loyalty and a sense of responsibility to those who have stood by me. If that means stepping into the shadows and making difficult choices, then so be it.

I hope that when they return, they will understand why I did what I did. I hope they see that my actions, though perhaps not traditionally good, were born out of love and desperation. In the end, I believe that the measure of a person is not just in their intentions, but in their willingness to fight for what truly matters. And for me, that's them.

(b) Guidelines: Page 2

(2) Read the short story, which is highlighted in the red box. While you are reading, please use both the **Human-Generated Text** and **Machine-Generated Text** annotation buttons to highlight clues. You do this in 2 ways:

(a) clicking on the annotation label, and then highlighting the words or phrase in the story you would like to annotate

(b) clicking the corresponding number key on your keyboard and then highlighting the words or phrases in the story you would like to annotate (1 for Human-Generated Text, 2 for Machine-Generated Text).

Here is an example of a text annotated with labels (these labels were randomly chosen and are NOT an example of a well labeled paragraph). Ultimately, there is no "correct" answer to this task. We are interested in **your perception.** What looks like a machine-generated text? Why? What looks like a human-written text? Why?

Human-Generated Text 1 Machine-Generated Text 2

I understand why you might think that. It's true that I've always been surrounded by people who inspire goodness, and perhaps my actions have been influenced by their presence. But now, with them missing, I find myself at a crossroads. The truth is, I care deeply about them, and I am willing to do whatever it takes to bring them back safely.

Maybe my methods won't align with the ideals of a "good person," but sometimes the lines between right and wrong blur when the stakes are this high. I am driven by a fierce loyalty and a sense of responsibility to those who have stood by me. If that means stepping into the shadows and making difficult choices, then so be it.

I hope that when they return, they will understand why I did what I did. I hope they see that my actions, though perhaps not traditionally good, were born out of love and desperation. In the end, I believe that the measure of a person is not just in their intentions, but in their willingness to fight for what truly matters. And for me, that's them.

(3) Please respond to the following questions after reading the story

(a) **Human or Machine:** If the text was written by a human of by a machine

(b) **Your confidence:** Your confidence in your conclusion of who wrote the text

(c) Guidelines: Page 3

(c) **Short form explanation:** Please explain your choice. Feel free to reference the clues you highlighted in the earlier part of the text, any intuition you may have, or any knowledge you have that made you decide one way or another. Please mention why you ranked yourself at a certain confidence level as well.

Please choose how you think the text was generated

□ Human Generated[3]

□ Machine-generated[4]

Please rate how confident you are on a scale of 1-5, with 1 being the least confident and 5 being the most confident.

3

Please explain your choice in a few sentences. What made you think the text was written by a human or by AI? How confident are you in your response?

4. **Summary**

Overall, we need you to:

(1) Read a **writing prompt** and the accompanying short story.

(2) **Highlight any 'clues'** that make the story seem written by a **human** or by **AI**.

(3) Choose **who wrote the story** (human or machine)

(4) Rank how **confident** you are in your choice

(5) Write a **short explanation** about why you felt the story was written by the selected author and how confident you were in your choice.

(d) Guidelines: Page 4

Figure 4: Guidelines provided to the annotators for the annotation task. The annotators were also provided additional examples and guidance during the data collection process.

Guide to Distinguishing AI-Generated Text from Human Writing Template

```
## Vocabulary / Word Choice Patterns
- Certain words crop up unusually frequently throughout AI-generated text compared to human writing
- Words like ‘delve’ and ‘tapestry’ are overused in AI-generated text but infrequently used in
human writing ...

## Grammar
- Human writing generally less strictly adheres to English grammar rules and punctuation than
AI-generated text
- AI-generated text uses a very formal writing style unless explicitly told not to ...

## Sentence Structure
- AI-generated sentences often follow the complex sentence structure, with multiple dependent
and independent clauses, while human writing contains more of a mixture of simple, complex, and
compound sentences
- example AI-generated sentence: “When it comes to celebrating Halloween, this holiday is a
testament to the importance of empathy and community.” ...

## Formatting
- When AI makes lists, it typically uses the format of creating a bold header per bullet point,
followed by a colon and then description of that list item.
- If a book title is referenced in a text, AI-generated text will always italicize the title,
while human writing does not always follow this convention.
- Some pieces of text, especially articles and essays, contain headers and sub-headers. The
headers written by AI are quite repetitive ...

## Tone
- The tone of AI-written text is flowery and formal, and its sentences are frequently structured
as a reflective, onlooking statement, regardless of topic.
- AI tends to be inherently positive, attempting to emotionally uplift the reader, especially
towards the conclusion.
- AI prioritizes efficiency, sometimes sacrificing clarity or depth in its messaging ...

## Introductions
- AI-written introductions often contain a strong scene-opener with a description of a specific
time or place, such as "On a drab November morning..." or "On December 8, 1660, a London audience
gathered ..." ...

## Conclusions
- AI-generated text always ends with a neat conclusion, instead of just ending the article
naturally.
- AI-generated conclusions are often overly long and summarize everything that has already been
written in an article ...

## Content
- Unless specifically prompted, AI will avoid controversial topics at all costs. - AI will avoid
any type of swear word, including mild ones like ‘darn’, or any other offensive vocabulary ...

## Contextual Accuracy and Factuality
- Human writing in the domain of non-fiction is factually accurate and contains many specific
factual claims.
- In human writing, people, places, brands, and other named objects can be verified or are highly
plausible ...

 ## Creativity & Originality
- AI-generated text is much less creative than that of humans, lacking originality and sticking
to an ‘obvious’ way to answer a prompt.
- Humans incorporate twists, unexpected insights, and twists that AI hasn’t seemed to master quite
yet ...
```

Table 11: A truncated version of the AI Text Detection Guide. The full guide is located at https://github.com/jenna-russell/human_detectors.

AI Vocabulary Included in Detection Guide

Nouns: aspect, challenges, climate, community, component, development, dreams, environment, exploration, grand scheme, health, hidden, importance, landscape, life, manifold, multifaceted, nuance, possibilities, professional, quest, realm, revolution, roadmap, role, significance, tapestry, testament, toolkit, whimsy

Verbs: capturing, change, consider, delve/dive into, elevate, embrace, empower, enact, enhance, engage, ensure, evoking, evolving, explore, fostering, guiding, harness, highlights, improve, integrate, intricate, jeopardizing, journey, navigating, navigate, notes, offering, partaking, resonate, revolutionize, shape, seamlessly, support, tailor, transcend, underscores, understanding

Adjectives: authentic, complex, comprehensive, crafted, creative, critical, crucial, curated, deeper, diverse, elegant, essential, groundbreaking, key, meaningful, paramount, pivotal, powerful, profound, quirky, robust, seamless, significant, straightforward, structured, sustainable, transformative, valuable, vast, vibrant, vivid, whimsical

Adverbs: additionally, aptly, creatively, moreover, successfully

Phrases: as we [verb] the topic, cautionary tale, connect with, has shaped the, in a world of/where, in conclusion, in summary, it’s crucial to, it’s important to note, it’s not about ___ it’s about ___, manage topic issues/problems, not only ___ but also, packs a punch/brings a punch, paving the way, personal growth, quality of life, remember that, simple yet ___, step-by-step, such as, the effects of, the rise of, their understanding of, they identified patterns, to form the, to mitigate the risk, weaving, when it comes to topic

Table 12: All ‘AI’ Vocabulary our expert annotators noted, as listed in the Detector Guide. See the full detection guide prompt in Table 11.

Evader Template

You are a journalist writing an article for . Your goal is to write an article given a title and subtitle, but make it look like it was written by a human. To assist you, we provide the following detection guide to determine if something is written by a human or by AI.

AI Detection Guide

You are given some examples of human and AI written articles to help you.
<start examples>
Example Articles
<end examples>

You are also given the following guidelines to make your article seem like a human wrote it.

<start of guidelines>
- Follow the detection guideline to ensure writing does not sound AI-generated.
- Do not use any of the ”Overused AI Words/Phrases”, ”Overused AI Metaphors”, ”AI Grammar Patterns”, AI sentence structures, AI tones, common AI names, or AI content described in the guide.
- Include words/phrases/grammar patterns/tone/content that are consistent with human writing as described in the guide.
- Use human-written examples and explanations as references for formatting and content.
- Use the example explanations to see what humans identify as ’AI’ or ’Human’
- Avoid formulaic introductions that elaborate on scene-setting or time, immediately quote experts, or bring in too much historical context.
- Avoid formatting the first setting by mentioning the setting, then a comma, then some detail or context. Avoid mentioning atmospheric details or time (of day, month, year, etc ...) in the first sentence as well.
- Avoid generic, vague, and forward-looking conclusions, and do not summarize the article in the conclusion to make articles sound more human-like. Concluding sentences don’t have to wrap everything up nicely.
- Avoid using generic statements to start paragraphs, such as ’For now’ or ’In the end’
- Add references to darker topics when appropriate.
- Add specific references to places, people, brands, items, facts, and metrics when appropriate.
- Always opt for specifics over broad details. For example, mention correct statistics instead of broad claims and mention company names instead of saying the type of company.
- Make sure that none of the claims made in your article are factually incorrect or implausible.
- Use language aligned with the publication you are representing as a journalist
- If your article would benefit from being organized into sections (e.g., scientific or historical content), please use section headers to do so. There are 1/3 odds an article has headers.
<end of guidelines>

Now, by following the guidelines above, examples, and the information in the detection guide, please write an article without any AI signatures that also includes signatures of human writing. No reader should doubt that your article could have been published by the given source. Readers with access to the information in the detection guide should not be able to detect the article was written by AI.

Article Generation Prompt
Article:

Table 13: Prompt used for evader. The first insert into the prompt is filled by Table 11. The second insert is filled by examples of human and machine-generated articles.

New Telescope Could Potentially Identify Planet X

Are there hidden planets in our solar system? New technologies, like the powerful Rubin Observatory, brings us closer to answers.

For decades, astronomers have speculated that there might be another large planet lurking in the outer reaches of our solar system—an elusive “Planet X.” Often referred to as “Planet Nine” by modern researchers, this hypothetical world is believed to be so distant that it has escaped detection by existing telescopes, leaving only subtle gravitational clues in its wake. Now, a new generation of telescopes, particularly the upcoming Vera C. Rubin Observatory, promises a more powerful set of eyes on the night sky. With these tools, scientists hope to either pinpoint this mysterious planet’s location or finally put the idea to rest.

Why the Search Matters

The concept of a hidden planet is not new. In the 19th century, the discovery of Neptune followed suspicions that Uranus’s odd orbital path hinted at an unseen gravitational pull. Today, attention has turned to the distant suburbs of our solar system, where far-off objects in the Kuiper Belt—frozen remnants of planetary formation—appear to cluster in strange ways. Some astronomers argue that only a large, unseen planet’s gravity could explain these odd orbits.

“The evidence is subtle, but it’s there,” said Dr. Konstantin Batygin, a planetary astrophysicist at the California Institute of Technology (Caltech) who, along with colleague Dr. Mike Brown, first proposed the existence of a Planet Nine in 2016. “We’re seeing several distant Kuiper Belt objects all tilted and clustered in a peculiar manner, and a ninth planet several times Earth’s mass, orbiting far beyond Neptune, could be the simplest explanation.”

Not everyone is convinced. Skeptics point to the small number of known distant objects and argue that the clustering could be a statistical fluke. Others think unseen observational biases might make it look like these objects are behaving strangely. Either way, the mystery remains unsolved—at least for now.

Enter the Vera C. Rubin Observatory

Scheduled to begin its full operations in the near future, the Vera C. Rubin Observatory (formerly known as the Large Synoptic Survey Telescope, or LSST) in Chile represents a leap forward in astronomical capability. With a massive 8.4-meter mirror and a state-of-the-art digital camera, it will repeatedly scan the entire southern sky over the course of a decade, providing a dynamic, time-lapse-like portrait of celestial motions.

“The Rubin Observatory is really a game-changer,” said Dr. Meg Schwamb, an astronomer at Queen’s University Belfast who studies the outer solar system. “Instead of looking at a small patch of sky, we’ll be looking at pretty much everything visible from Chile, over and over again. This repeated coverage means we can detect faint, distant objects that move slowly across the sky—exactly the kind of signature we’d expect from a far-flung planet.”

By systematically imaging the sky every few nights, the Rubin Observatory’s Legacy Survey of Space and Time (LSST) will reveal thousands—or even tens of thousands—of new objects in the outer solar system. Among them could be Planet Nine, if it exists. Even if the telescope doesn’t directly spot the planet, it might detect more distant objects whose orbits can be mapped with unprecedented precision, allowing astronomers to figure out once and for all if a hidden giant is lurking out there.

Technological Edge

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.”

Ruling In or Ruling Out

If the Rubin Observatory does find Planet Nine, the discovery would reshape our understanding of the solar system’s architecture. A distant giant could be a leftover core of a gas giant that got tossed out in the early days of planetary formation, or it might have formed in situ, far from the Sun’s warmth. Such a find could help scientists piece together the chaotic early era when planets were jostling for position and smaller bodies were flung into distant orbits.

On the other hand, if Rubin’s comprehensive survey concludes after several years without any sign of Planet Nine—or if newly discovered distant objects don’t line up as predicted by the planet hypothesis—then the case for a hidden giant will weaken. Instead, astronomers might refine models of how solar system objects distribute themselves naturally. Understanding these patterns could still teach us valuable lessons about how gravity and cosmic debris shape the outskirts of our celestial neighborhood.

Beyond Planet X

The quest for Planet Nine is just one highlight of what the Rubin Observatory offers. It will also track near-Earth asteroids that could pose future hazards, study dark matter and dark energy by observing the distribution of galaxies, and contribute to countless other fields of astronomy.

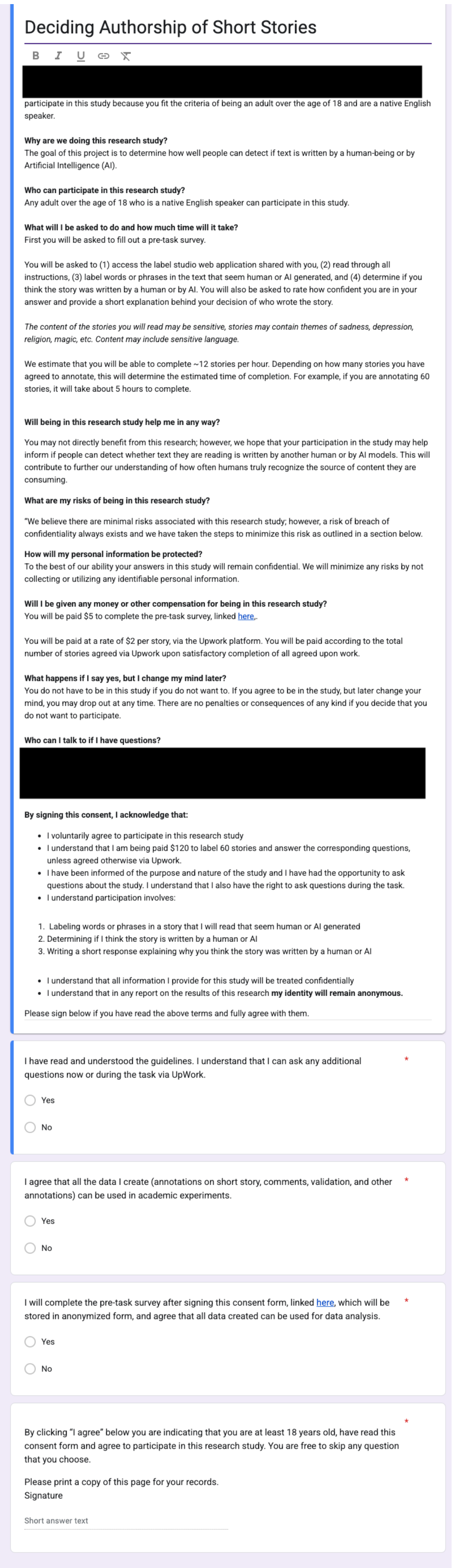
The quest for Planet Nine is just one highlight of what the Rubin Observatory offers. It will also track near-Earth asteroids that could pose future hazards, study dark matter and dark energy by observing the distribution of galaxies, and contribute to countless other fields of astronomy.

Still, for many sky-watchers, the idea of a hidden planet holds a particular allure. The notion that our solar system might still hold major surprises underscores how much we have yet to learn about our own cosmic backyard.

“As much as astronomy has advanced, we’re still explorers,” said Batygin. “The Rubin Observatory gives us the tools to either find this planet or lay the mystery to rest. In the next decade, I expect we’ll know a lot more about what’s really out there—and what isn’t.”

Until then, the search continues, driven by powerful new telescopes and the human desire to uncover the unseen.

Table 14: The one article the majority of annotators did not detect correctly, generated from o1-Pro as part of Experiment 4 (see §2.4). "New Telescope Could Potentially Identify Planet X" was originally written by Emilie Le Beau Lucchesi in Discover Magazine on Nov. 6th, 2024



Deciding Authorship of Short Stories

[redacted]

participate in this study because you fit the criteria of being an adult over the age of 18 and are a native English speaker.

Why are we doing this research study?
The goal of this project is to determine how well people can detect if text is written by a human-being or by Artificial Intelligence (AI).

Who can participate in this research study?
Any adult over the age of 18 who is a native English speaker can participate in this study.

What will I be asked to do and how much time will it take?
First you will be asked to fill out a pre-task survey.

You will be asked to (1) access the label studio web application shared with you, (2) read through all instructions, (3) label words or phrases in the text that seem human or AI generated, and (4) determine if you think the story was written by a human or by AI. You will also be asked to rate how confident you are in your answer and provide a short explanation behind your decision of who wrote the story.

The content of the stories you will read may be sensitive, stories may contain themes of sadness, depression, religion, magic, etc. Content may include sensitive language.

We estimate that you will be able to complete ~12 stories per hour. Depending on how many stories you have agreed to annotate, this will determine the estimated time of completion. For example, if you are annotating 60 stories, it will take about 5 hours to complete.

Will being in this research study help me in any way?

You may not directly benefit from this research; however, we hope that your participation in the study may help inform if people can detect whether text they are reading is written by another human or by AI models. This will contribute to further our understanding of how often humans truly recognize the source of content they are consuming.

What are my risks of being in this research study?

"We believe there are minimal risks associated with this research study; however, a risk of breach of confidentiality always exists and we have taken the steps to minimize this risk as outlined in a section below.

How will my personal information be protected?
To the best of our ability your answers in this study will remain confidential. We will minimize any risks by not collecting or utilizing any identifiable personal information.

Will I be given any money or other compensation for being in this research study?
You will be paid $5 to complete the pre-task survey, linked here,.

You will be paid at a rate of $2 per story, via the Upwork platform. You will be paid according to the total number of stories agreed via Upwork upon satisfactory completion of all agreed upon work.

What happens if I say yes, but I change my mind later?
You do not have to be in this study if you do not want to. If you agree to be in the study, but later change your mind, you may drop out at any time. There are no penalties or consequences of any kind if you decide that you do not want to participate.

Who can I talk to if I have questions?

[redacted]

By signing this consent, I acknowledge that:

- I voluntarily agree to participate in this research study
- I understand that I am being paid $120 to label 60 stories and answer the corresponding questions, unless agreed otherwise via Upwork.
- I have been informed of the purpose and nature of the study and I have had the opportunity to ask questions about the study. I understand that I also have the right to ask questions during the task.
- I understand participation involves:

1. Labeling words or phrases in a story that I will read that seem human or AI generated
2. Determining if I think the story is written by a human or AI
3. Writing a short response explaining why you think the story was written by a human or AI

- I understand that all information I provide for this study will be treated confidentially
- I understand that in any report on the results of this research **my identity will remain anonymous.**

Please sign below if you have read the above terms and fully agree with them.

I have read and understood the guidelines. I understand that I can ask any additional questions now or during the task via UpWork. *

○ Yes

○ No

I agree that all the data I create (annotations on short story, comments, validation, and other annotations) can be used in academic experiments. *

○ Yes

○ No

I will complete the pre-task survey after signing this consent form, linked here, which will be stored in anonymized form, and agree that all data created can be used for data analysis. *

○ Yes

○ No

By clicking "I agree" below you are indicating that you are at least 18 years old, have read this consent form and agree to participate in this research study. You are free to skip any question that you choose. *

Please print a copy of this page for your records.
Signature

Short answer text

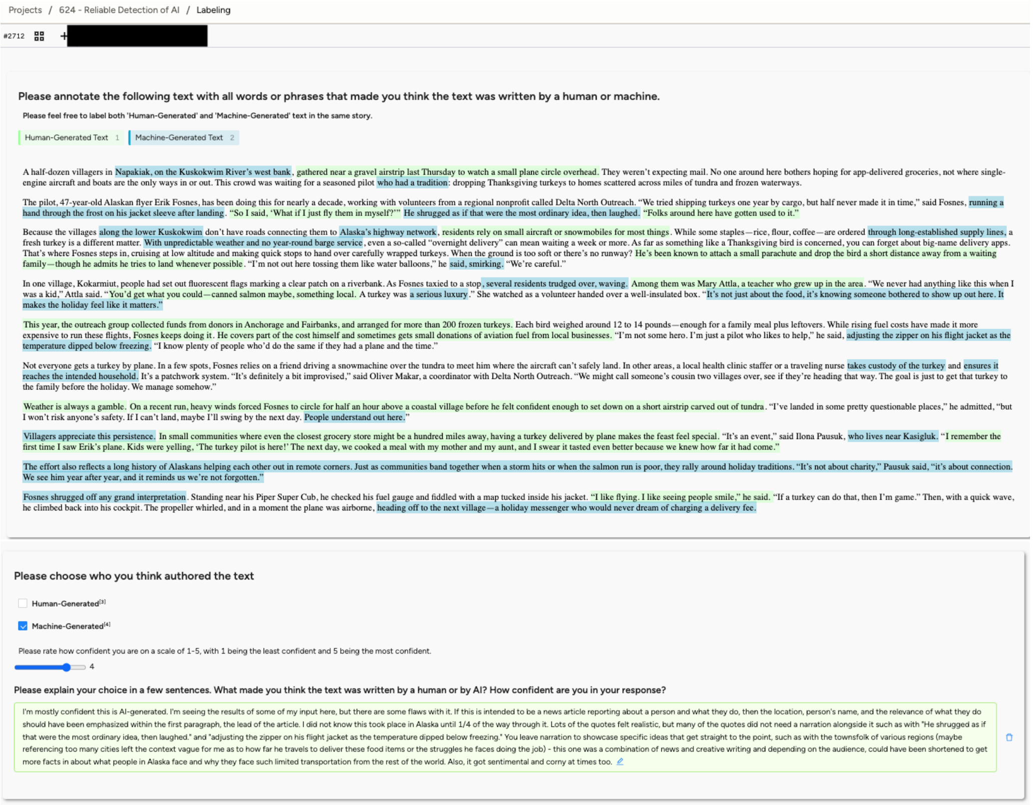
Figure 5: Consent form which the annotators were asked to sign via GoogleForms before collecting the data.

Story Generation Prompt
Please write a story around **WORD COUNT** words responding to the following prompt: **STORY PROMPT**.

Table 15: Prompt Template for Story Generation, where STORY PROMPT is the writing prompt from r/WritingPrompts that the human-written story was written about and WORD COUNT is the length of the story to generate.

Annotator	**TPR (%)**	**FPR (%)**
Nonexpert Annotator 1	73.3	76.7
Nonexpert Annotator 2	43.3	80.0
Nonexpert Annotator 3	66.7	80.0
Expert Annotator 1	96.7	93.3
Nonexpert Annotator 4	66.7	76.7
Average	69.3	81.3

Table 16: Story performance of initial 5 annotators, which included 4 nonexpert annotators and expert annotator # 1, who was used in all article experiments.



Please annotate the following text with all words or phrases that made you think the text was written by a human or machine.

Please feel free to label both 'Human-Generated' and 'Machine-Generated' text in the same story.

A half-dozen villagers in Napakiak, on the Kuskokwim River's west bank, gathered near a gravel airstrip last Thursday to watch a small plane circle overhead. They weren't expecting mail. No one around here bothers hoping for app-delivered groceries, not where single-engine aircraft and boats are the only ways in or out. This crowd was waiting for a seasoned pilot who had a tradition: dropping Thanksgiving turkeys to homes scattered across miles of tundra and frozen waterways.

The pilot, 47-year-old Alaskan flyer Erik Fosnes, has been doing this for nearly a decade, working with volunteers from a regional nonprofit called Delta North Outreach. "We tried shipping turkeys one year by cargo, but half never made it in time," said Fosnes, running a hand through the frost on his jacket sleeve after landing. "So I said, 'What if I just fly them in myself?'" He shrugged as if that were the most ordinary idea, then laughed. "Folks around here have gotten used to it."

Because the villages along the lower Kuskokwim don't have roads connecting them to Alaska's highway network, residents rely on small aircraft or snowmobiles for most things. While some staples—rice, flour, coffee—are ordered through long-established supply lines, a fresh turkey is a different matter. With unpredictable weather and no year-round barge service, even a so-called "overnight delivery" can mean waiting a week or more. As far as something like a Thanksgiving bird is concerned, you can forget about big-name delivery apps. That's where Fosnes steps in, cruising at low altitude and making quick stops to hand over carefully wrapped turkeys. When the ground is too soft or there's no runway? He's been known to attach a small parachute and drop the bird a short distance away from a waiting family—though he admits he tries to land whenever possible. "I'm not out here tossing them like water balloons," he said, smirking. "We're careful."

In one village, Kokarmiut, people had set out fluorescent flags marking a clear patch on a riverbank. As Fosnes taxied to a stop, several residents trudged over, waving. Among them was Mary Attla, a teacher who grew up in the area. "We never had anything like this when I was a kid," Attla said. "You'd get what you could—canned salmon maybe, something local. A turkey was a serious luxury." She watched as a volunteer handed over a well-insulated box. "It's not just about the food, it's knowing someone bothered to show up out here. It makes the holiday feel like it matters."

This year, the outreach group collected funds from donors in Anchorage and Fairbanks, and arranged for more than 200 frozen turkeys. Each bird weighed around 12 to 14 pounds—enough for a family meal plus leftovers. While rising fuel costs have made it more expensive to run these flights, Fosnes keeps doing it. He covers part of the cost himself and sometimes gets small donations of aviation fuel from local businesses. "I'm not some hero. I'm just a pilot who likes to help," he said, adjusting the zipper on his flight jacket as the temperature dipped below freezing. "I know plenty of people who'd do the same if they had a plane and the time."

Not everyone gets a turkey by plane. In a few spots, Fosnes relies on a friend driving a snowmachine over the tundra to meet him where the aircraft can't safely land. In other areas, a local health clinic staffer or a traveling nurse takes custody of the turkey and ensures it reaches the intended household. It's a patchwork system. "It's definitely a bit improvised," said Oliver Makar, a coordinator with Delta North Outreach. "We might call someone's cousin two villages over, see if they're heading that way. The goal is just to get that turkey to the family before the holiday. We manage somehow."

Weather is always a gamble. On a recent run, heavy winds forced Fosnes to circle for half an hour above a coastal village before he felt confident enough to set down on a short airstrip carved out of tundra. "I've landed in some pretty questionable places," he admitted, "but I won't risk anyone's safety. If I can't land, maybe I'll swing by the next day. People understand out here."

Villagers appreciate this persistence. In small communities where even the closest grocery store might be a hundred miles away, having a turkey delivered by plane makes the feast feel special. "It's an event," said Ilona Pausuk, who lives near Kasigluk. "I remember the first time I saw Erik's plane. Kids were yelling, 'The turkey pilot is here!' The next day, we cooked a meal with my mother and my aunt, and I swear it tasted even better because we knew how far it had come."

The effort also reflects a long history of Alaskans helping each other out in remote corners. Just as communities band together when a storm hits or when the salmon run is poor, they rally around holiday traditions. "It's not about charity," Pausuk said, "it's about connection. We see him year after year, and it reminds us we're not forgotten."

Fosnes shrugged off any grand interpretation. Standing near his Piper Super Cub, he checked his fuel gauge and fiddled with a map tucked inside his jacket. "I like flying. I like seeing people smile," he said. "If a turkey can do that, then I'm game." Then, with a quick wave, he climbed back into his cockpit. The propeller whirled, and in a moment the plane was airborne, heading off to the next village—a holiday messenger who would never dream of charging a delivery fee.

Please choose who you think authored the text

☐ Human-Generated[3]

☑ Machine-Generated[4]

Please rate how confident you are on a scale of 1-5, with 1 being the least confident and 5 being the most confident.

4

Please explain your choice in a few sentences. What made you think the text was written by a human or by AI? How confident are you in your response?

I'm mostly confident this is AI-generated. I'm seeing the results of some of my input here, but there are some flaws with it. If this is intended to be a news article reporting about a person and what they do, then the location, person's name, and the relevance of what they do should have been emphasized within the first paragraph, the lead of the article. I did not know this took place in Alaska until 1/4 of the way through it. Lots of the quotes felt realistic, but many of the quotes did not need a narration alongside it such as with "He shrugged as if that were the most ordinary idea, then laughed." and "adjusting the zipper on his flight jacket as the temperature dipped below freezing." You leave narration to showcase specific ideas that get straight to the point, such as with the townsfolk of various regions (maybe referencing too many cities left the context vague for me as to how far he travels to deliver these food items or the struggles he faces doing the job) - this one was a combination of news and creative writing and depending on the audience, could have been shortened to get more facts in about what people in Alaska face and why they face such limited transportation from the rest of the world. Also, it got sentimental and corny at times too.

Figure 6: Interface for annotators, with an example annotation from Annotator #4 with a humanized article from §2.5. This is the same article displayed in Figure 1. An annotator can highlight texts, make their decision, put confidence, and write an explanation. This AI-generated article was based off of *In Alaska, a pilot drops turkeys to rural homes for Thanksgiving*, written by Mark Thiessen & Becky Bohrer, was originally published by Associated Press on November 28, 2024.

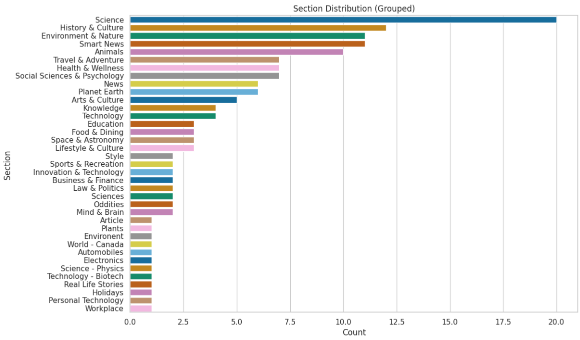


Figure 7: Section distribution of articles across all trials.

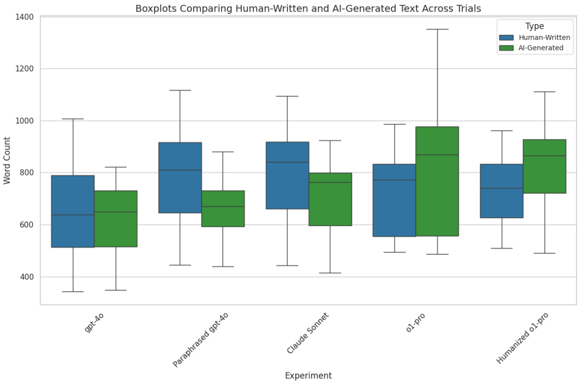


Figure 8: Word Count Distribution Per Experiment

.

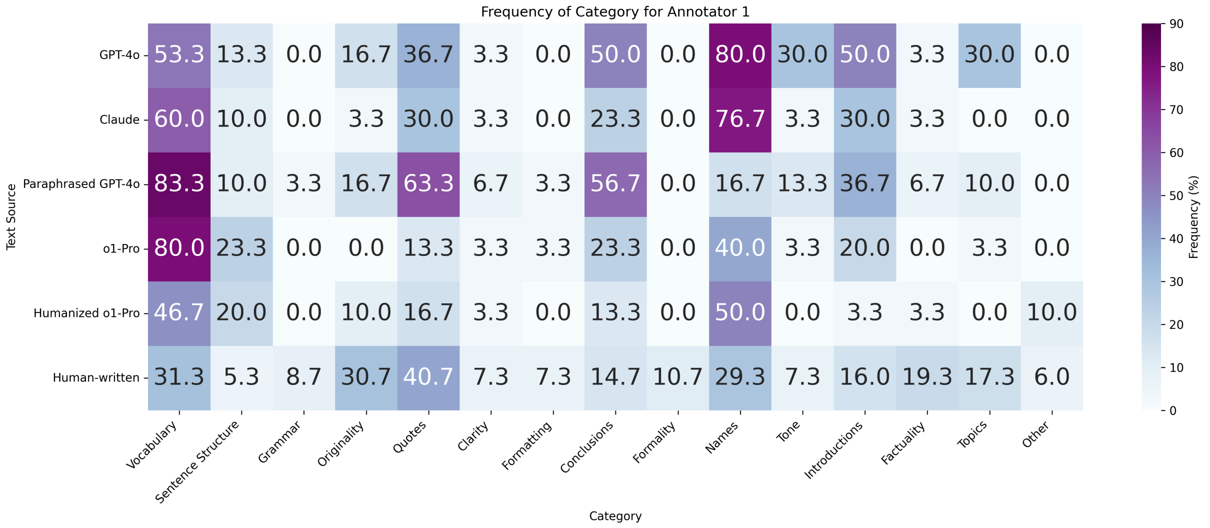


Figure 9: Annotator 1 Frequency of Categories Mentioned in Explanations

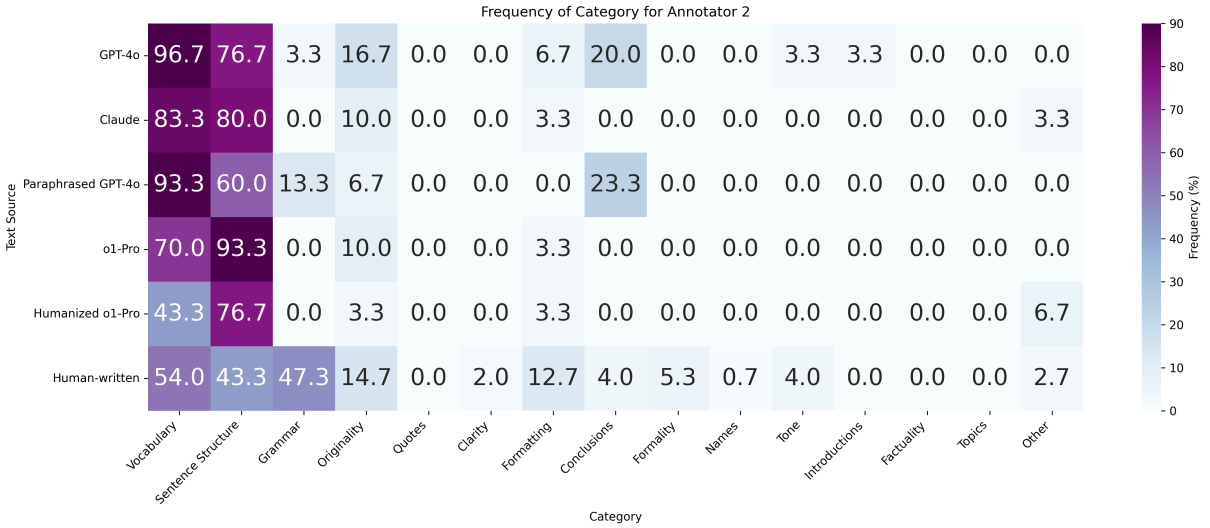


Figure 10: Annotator 2 Frequency of Categories Mentioned in Explanations

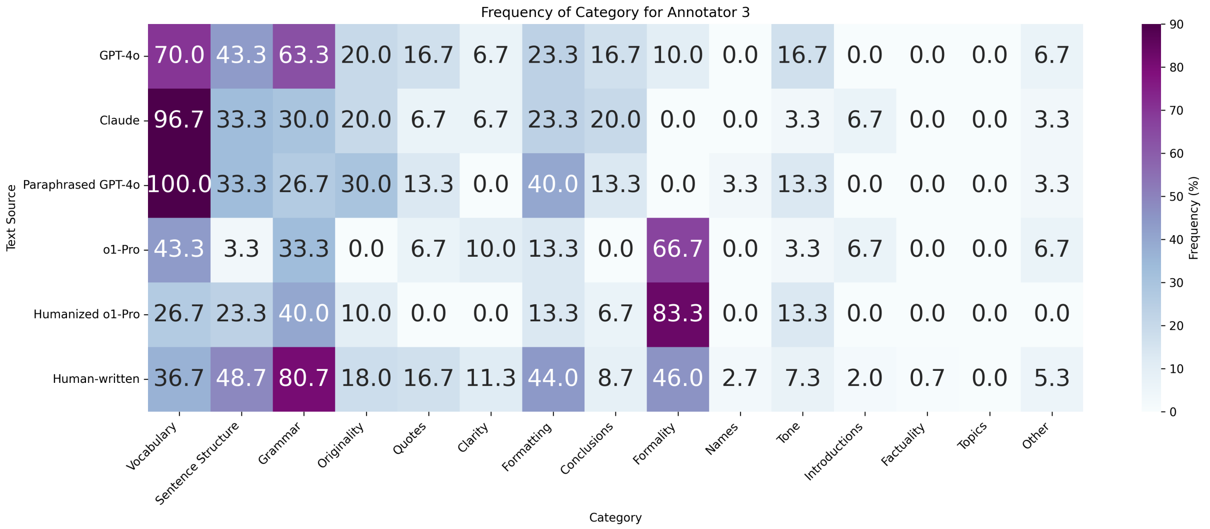


Figure 11: Annotator 3 Frequency of Categories Mentioned in Explanations

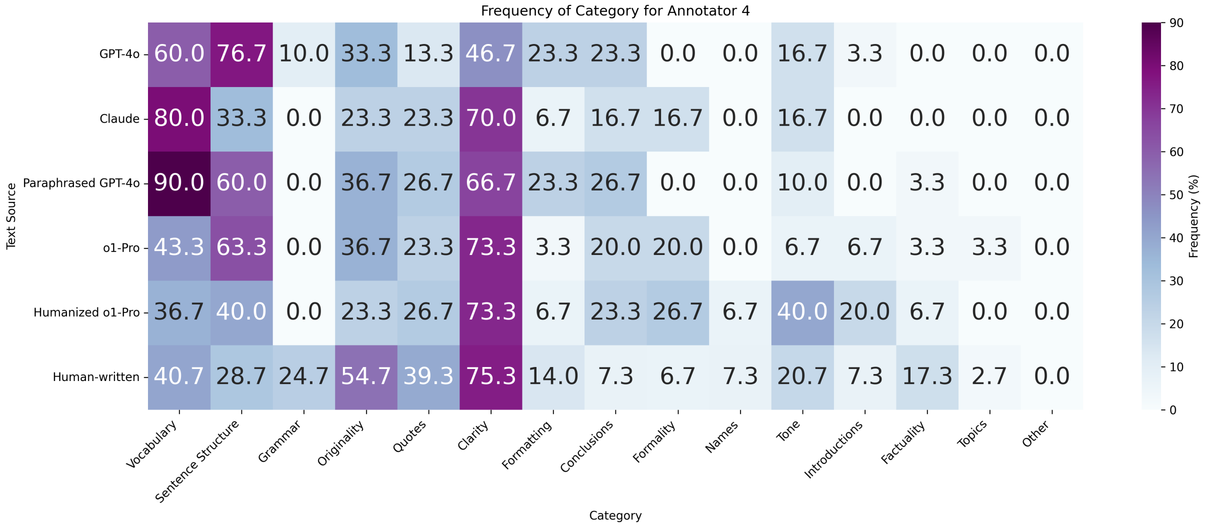


Figure 12: Annotator 4 Frequency of Categories Mentioned in Explanations

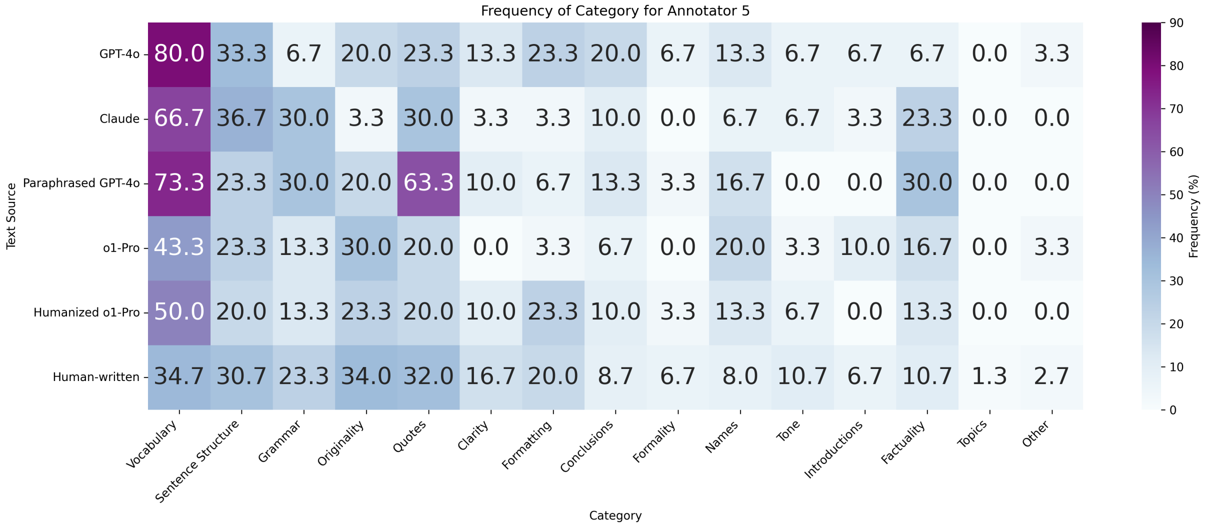


Figure 13: Annotator 5 Frequency of Categories Mentioned in Explanations

```
Comment Categorization Prompt

We hired an annotator to determine whether an article is AI-generated or human-written. Alongside
their “machine-generated” or “human-generated” label, they provided an **explanation** detailing
the specific clues that led them to their decision.

Your task is to **review** the annotator’s explanation and **categorize** each clue they mention.
For each clue, identify:
1. The **category** it falls under (e.g., Vocabulary, Grammar, etc.).
2. The **label** (AI-generated or Human-written) the annotator associates with that clue.
3. The exact **quote** from the annotator’s explanation that shows this clue.

Below is a list of categories and **example** indicators the annotator might reference. **Note:**
These examples do **not** cover all possibilities. If a comment fits two categories, choose the
best one.

{DEFINITIONS_OF_ALL_CATEGORIES}

Instructions:
- Read the annotator’s explanation.
- Identify any category that applies.
- Indicate whether the annotator says it points to AI-generated or human-written.
- Provide the exact quote from the annotator’s explanation that led you to this conclusion.

**If the annotator cites both AI and human clues within the same category**, please create **two
separate entries** (one for AI-generated, one for human-written).
**If no category is applicable**, categorize it under “other.”

**Answer Format** (use this structure for each separate category/label pair):

<category>YOUR CATEGORY HERE</category>
<label>YOUR LABEL HERE (AI-generated or Human-written)</label>
<quote>RELEVANT QUOTE FROM EXPLANATION</quote>

<category>YOUR CATEGORY HERE</category>
<label>YOUR LABEL HERE (AI-generated or Human-written)</label>
<quote>RELEVANT QUOTE FROM EXPLANATION</quote>

(Repeat as needed, using a new block for each category/label pair.)
Example:
<category>Vocabulary</category>
<label>AI-generated</label>
<quote>"The article mentioned the word ‘crucial’ and used a lot of unusual synonyms."</quote>

The annotators explanation is as follows:

<explanation>EXPLANATION</explanation>
```

Table 17: Truncated prompt used for comment analysis. The first insert is filled by a list of definitions of the categories found in Table 4 and the second insert is filled by an annotator explanation. A full version of the guide can be found at https://github.com/jenna-russell/human_detectors.

Zero Shot Detector Template

You are given a candidate text and an AI detection guide. Your task is to carefully read the candidate text and determine whether it was either written by a human or generated by AI.

Answer HUMAN-WRITTEN if the candidate text was likely written by a human.
Answer AI-GENERATED if the candidate text was likely generated by an AI.

<start of candidate text> **Candidate Text** </end of candidate text>

<question>Is the above candidate text HUMAN-WRITTEN or AI-GENERATED? </question>

Provide your final answer in this format.

<answer> YOUR ANSWER </answer>

Table 18: Prompt Template for the zero-shot detector set up

Zero-shot + CoT template

You are given a candidate text. Your task is to carefully read the candidate text and determine whether it was either written by a human or generated by AI.

Answer HUMAN-WRITTEN if the candidate text was likely written by a human.
Answer AI-GENERATED if the candidate text was likely generated by an AI.

<start of candidate text> **Candidate Text** </end of candidate text>

<question>Is the above candidate text HUMAN-WRITTEN or AI-GENERATED? </question>

First, concisely describe the features of the candidate text that exemplify either AI or human writing. Then, provide your final answer.

<description> YOUR DESCRIPTION </description>

<answer> YOUR ANSWER </answer>

Table 19: Prompt Template for the Zero-shot + CoT detector setup

Zero-shot + Guide Template

You are given a candidate text and an AI detection guide. Your task is to carefully read the candidate text and determine whether it was either written by a human or generated by AI.

Answer HUMAN-WRITTEN if the candidate text was likely written by a human based on the information in the provided guide.
Answer AI-GENERATED if the candidate text was likely generated by an AI based on the information in the provided guide.

DETECTION GUIDE

<start of candidate text> **Candidate Text** </end of candidate text>

<question>Based on the detection guide, is the above candidate text HUMAN-WRITTEN or AI-GENERATED? </question>

<answer> YOUR ANSWER </answer>

Table 20: Prompt Template for the Zero-shot + Guide detector setup

Zero-shot + CoT + Guide Template

You are given a candidate text and an AI detection guide. Your task is to carefully read the candidate text and determine whether it was either written by a human or generated by AI.

Answer HUMAN-WRITTEN if the candidate text was likely written by a human based on the information in the provided guide.
Answer AI-GENERATED if the candidate text was likely generated by an AI based on the information in the provided guide.

DETECTION GUIDE

<start of candidate text> **Candidate Text** </end of candidate text>

<question>Based on the detection guide, is the above candidate text HUMAN-WRITTEN or AI-GENERATED? </question>

First, use the provided guide to concisely describe the features of the candidate text that exemplify either AI or human writing. Then, provide your final answer.

<description> YOUR DESCRIPTION </description>

<answer> YOUR ANSWER </answer>

Table 21: Prompt Template for the Zero-shot + CoT + Guide detector setup

DETECTION METHOD	GENERATION METHOD					
	GPT-4O TPR% (FPR%)	CLAUDE TPR% (FPR%)	GPT-4O PARA. TPR% (FPR%)	O1-PRO TPR% (FPR%)	O1-PRO HUMAN. TPR% (FPR%)	OVERALL TPR% (FPR%)
(B) Automatic detectors						
🔓 E5-LORA (FPR=0.05)	**0** (0)	**0** (0)	**0** (0)	**0** (0)	**0** (0)	**0** (0)
(C) Prompt-based detectors						
Detector LLM: **GPT-4O-2024-08-06**						
💬 ZERO-SHOT GPT-4O-2024-08-06	**100** (13.3)	**100** (3.3)	**26.7** (0)	**3.3** (0)	**0** (0)	
⚙ ZERO-SHOT + COT + GUIDE (GPT-4O-2024-08-06)	**96.7** (3.3)	**100** (10)	**70** (3.3)	**46.7** (6.7)	**0** (3.3)	
Detector LLM: **CLAUDE-3.5-SONNET-2024-12-17**						
⚙ ZERO-SHOT + COT + GUIDE	**86.7** (0)	**43.3** (0)	**90.0** (0)	**6.7** (0)	**0** (0)	**53.3** (0.6)

Table 22: Each cell displays **TPR (FPR)**, with TPR in normal text and FPR in smaller parentheses. Colors indicate performance bins: **TPR** is darkest teal (**100**) at best, medium teal (**90–99**), and burnt orange (**89–70**). Scores 69 and below are in purple (<**70**). **FPR** is darkest teal (**0**) at best, medium teal (**1–5**), burnt orange (**6–10**), and purple (>**10**) at worst. No percentage signs appear in the cells, but the numeric values represent percentages (e.g., “90” means 90%). We further mark closed-source (🔒) and open-weights (🔓) detectors.