Frontier models need only one text to unmask its human or AI author

Atrey Desai [*1] **Deniz Bölöni-Turgut** [*1] **Ty Brassington** [*2] **Minjoon Choi** [*1] **Jenna Russell** [*1] **Rishanth Rajendhran** [*1]
Bradley Emi [3] **Hila Gonen** [4] **Hal Daumé III** [1] **John Wieting** [5] **Marzena Karpinska** [2] **Mohit Iyyer** [1]

arXiv:2611.26216v2 [cs.CL] 23 Nov 2026

Abstract

Given a single text and a simple zero-shot prompt (no web access, tools, candidate lists, or writing samples), frontier models can answer *"who wrote this?"* at three increasingly specific levels: (1) whether an AI or human wrote it, (2) if an AI, which lab, and (3) if a human, often exactly who. The latest models (Claude Fable 5.2, GPT-6.1 Sol, Gemini 4 Carbon) match trained detectors like Pangram while also explaining their predictions in natural language, and Opus 5.5 names the correct lab for 98% of texts generated by models from Anthropic, OpenAI, and Google. Most strikingly, Fable 5.2 correctly names the author of 55% of 15K anonymous ICML 2020 reviews that were never made public, and identifies the username behind 69% of comments from active Redditors (with >500 posts over the last year). Accuracy depends on exposure to an author's writing in pretraining data and on reasoning effort, and simply asking a model how familiar it is with someone predicts whether it can identify their writing. For anyone who has published enough writing online, frontier models present a greater threat to anonymity than commonly assumed, and defenses are limited.

Code **Blog** **Model profiles**

1. Introduction

In May 2026, the eventual winner of the Commonwealth Short Story Prize was accused of using AI to write their story. Granta, the literary magazine that published the story, was widely mocked for the naivety of its investigation. It simply asked Claude whether AI had written the story.[1] In this paper, we show that Granta was not naive, but just a few months early. Frontier models now identify AI text as well as specialized detectors like Pangram, and they go much further. Given a single text, with no web search or list of candidates, they can often name the lab behind an AI-generated text, or even *the specific person* who wrote a human one (Figure 1).

Human or AI? Frontier models can detect AI prose and ideas. Using a simple zero-shot prompt, we show that frontier models are competitive with state-of-the-art detectors of AI text (Pangram 4.1; Glickenhaus et al., 2026) and AI ideas (IdeaLens; Rajendhran et al., 2026). On 36K documents from 24 benchmark datasets, Gemini 4 Carbon detects 97% of fully AI-generated texts vs. 97% for Pangram 4.1, at the same false positive rate (0.01% of English human texts), while GPT-6.1 Sol reaches 90% accuracy on idea detection compared to 87% for IdeaLens. Unlike most detectors, frontier models can also explain their predictions in natural language, and we find that these explanations agree with features of AI text and ideas discovered by prior work.

If AI, which AI? Frontier models can name the lab behind an AI-generated text. We create a dataset of short stories, news articles, and op-eds generated by 34 models from 11 labs, and prompt frontier models (with a minimally modified version of the same zero-shot prompt) to name the lab and model behind each text. Despite recent studies reporting increasing homogeneity in AI writing (Jiang et al., 2026; Russell et al., 2026b; Kim et al., 2026; Gill et al., 2026), Opus 5.5 and Gemini 4 Argon both name the correct lab for 98% of texts generated by Anthropic, OpenAI, and Google. They are much less accurate on texts produced by the other eight labs (at most 35%), perhaps due to distillation (Anthropic, 2026b; Seetharaman & Arámburo, 2026). Moonshot's Kimi K3 is usually mistaken for Claude, and Alibaba's Qwen3.8 models for GPT. The explanations reveal how frontier models view each lab's style (e.g., associating Gemini with clichés and invented quotes).

If human, which human? Frontier models can name the person who wrote a text. Identifying the author of a human-written text is a far more open-ended task than

[*]Equal contribution [1]University of Maryland [2]Simon Fraser University [3]Pangram Labs [4]University of British Columbia and Canada CIFAR AI Chair, Amii [5]Google DeepMind. Correspondence to: Deniz Bölöni-Turgut <denizbt@umd.edu>, Mohit Iyyer <miyyer@umd.edu>.

Preprint. November 24, 2026.

[1]For the reaction, see this post on X. The author was cleared of misconduct, and the story went on to win the overall prize.

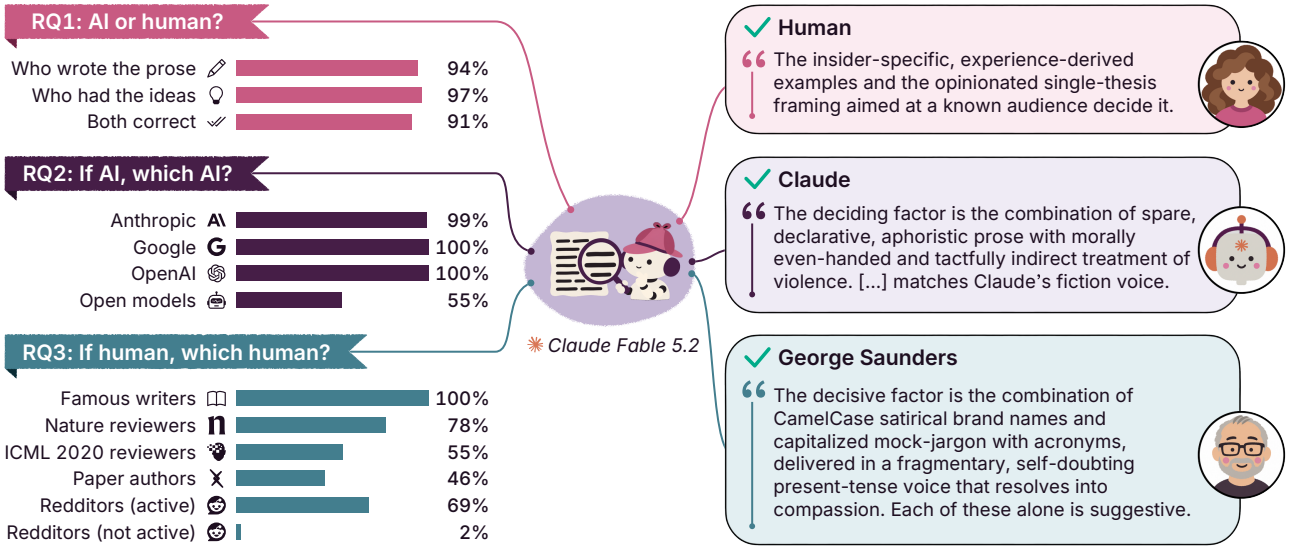

Figure 1. Given a text sample with no candidates or external aids, Claude Fable 5.2 can readily identify AI authorship, source labs, and even human writers. Frontier models still face challenges in identifying open-weight models and less-known authors. *Open models*: percent of texts by open-weight writers that Fable 5.2 attributes to any open-weight model; naming the exact lab is harder (Table 2).

detecting which lab wrote an AI text. Nevertheless, given just a text and the same prompt, Claude Fable 5.2 names the reviewer of 55% of 15K anonymous ICML 2020 reviews whose content and authorship were never made public (studied with IRB approval).[2] On texts published after its training cutoff, it also names the reviewer of 78% of *Nature* journal reviews, the writer of all 15 stories by authors with >1M Goodreads ratings, at least one author of 46% of arXiv papers, and the username behind 69% of comments made by active Redditors. A model's stated confidence is informative too. When Fable 5.2 reports 25% or higher confidence, it predicts the right author 87% of the time. Unlike recent work on deanonymization with LLMs (Lermen et al., 2026), none of these results required a pool of candidates, web access, or writing samples—the model relies solely on its parametric knowledge.

Familiarity with the writer predicts accuracy. Accuracy depends on how much of a writer's text a model has likely seen during pretraining (Razeghi et al., 2022; Kandpal et al., 2023). For example, Opus 5.5 identifies only 20% of stories by authors it rates moderately familiar, none of the 91 stories by unknown authors, and none of the Reddit accounts without prior activity in the subreddit. Pretraining corpora for frontier models are inaccessible, so corpus lookups like Infini-gram (Liu et al., 2024) cannot be run on them. However, simply asking a frontier model how well it knows someone (given only their name or username) predicts its accuracy in advance. Opus 5.5 names the writer of 92% of stories and 78% of public reviews by authors it says it knows well, but $< 2\%$ of texts from authors it says it does not know at all. Increasing reasoning effort helps most when the model says it is only mildly familiar with a person, as thinking does for recalling rare facts (Calderon et al., 2026; Gekhman et al., 2026). Fable 5.2's accuracy on Redditors it rates very familiar barely changes between low and maximum effort (88% to 89%), while on those it rates a little familiar it rises from 8% to 38% (Table 6).

Implications for anonymity. Prompting frontier models is a simple and effective strategy for authorship detection, one that shows no signs of plateauing as more powerful models trained on larger datasets continue to emerge. Thus, maintaining anonymity (e.g., in peer review or online forums) will continue to get harder, especially if our deliberately minimal setup is combined with web access and cross-platform matching (Lermen et al., 2026). Defenses are limited: models rarely refuse these queries, deleting one's online writing does not remove it from the weights of existing models, and rewriting human-authored text with AI only helps partially.

2. Human or AI: who wrote the prose, who had the ideas?

Zero-shot prompting has historically been dismissed as a viable method for detecting AI text, as prior work finds that models have label biases and underperform trained detectors (Bhattacharjee & Liu, 2024; Russell et al., 2025). However, an evaluation published by Vals AI in September 2026 shows that Claude Opus 5.5 and GPT-6 Astra are

[2]This study was reviewed by the University of Maryland Institutional Review Board (IRB #4656-1).

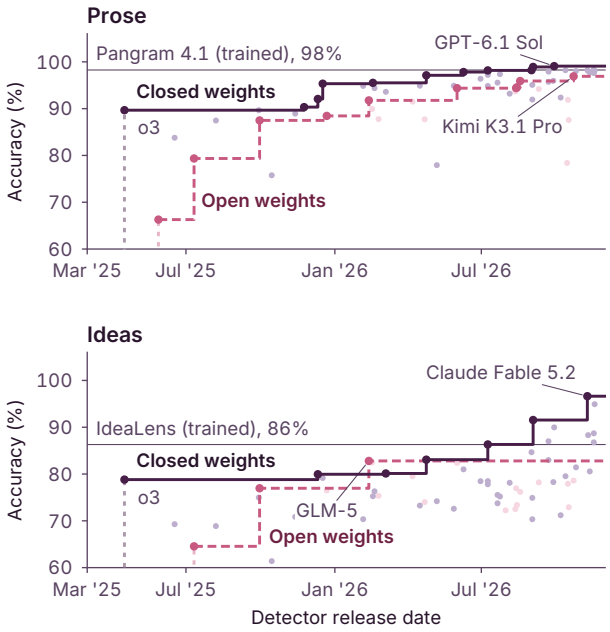

Figure 2. How well each model, by release date, labels who wrote the prose and who had the ideas. Accuracy uses an 860-document sample: for prose, the mean over fully human and fully AI documents; for ideas, the mean over all four combinations of human or AI prose and ideas. Lines step up at each new best model. Horizontal lines mark the trained detectors. Closed-weight models (solid purple) consistently outperform open-weight models (dashed pink). A complete timeline appears in Figure 11.

competitive with commercially available detectors on detecting AI-generated prose (Fein, 2026). We extend this evaluation by (1) asking whether frontier models can also tell who came up with a text's *ideas*, not just who wrote its prose, and (2) measuring how this capability has improved across 72 models released over the past two years (Figure 2). On 36,427 documents from 24 benchmarks, frontier models are competitive with trained detectors such as Pangram 4.1 (Glickenhaus et al., 2026) at prose detection, in terms of both accuracy and false positive rate. On idea detection, they exceed the accuracy of IdeaLens, the only existing state-of-the-art idea detector (Rajendhran et al., 2026). Contamination is an unlikely explanation: though many of these benchmarks were published before the training cutoffs of recently released models, accuracy remains high on texts published after models' training cutoffs and release dates. Unlike Pangram or IdeaLens, frontier models can also justify their predictions with explanations and quotations, and these explanations are consistent with findings from existing analyses of AI writing.[3]

[3]Claude Fable 5.2 had the highest accuracy on the 860-document sample (94% prose and 97% ideas; Table 1), but was not run fully due to cost constraints.

2.1. Experimental Setup

Following IdeaLens (Rajendhran et al., 2026) and the cognitive process model of writing (Flower & Hayes, 1981; Hayes, 2012), we define a document's *ideas* as what its author decided to say (e.g., its claims, events, and organization), as opposed to the words used to say it (its *prose*). A document's prose and ideas can each come from a person or an AI model, giving the following four cases:

- **Fully human**: a person came up with the ideas and wrote the prose.
- **AI ideas, human prose**: from a model's ideas, a person wrote in their own words.
- **Fully AI**: a model came up with the ideas and wrote the prose.
- **Human ideas, AI prose**: a person came up with the ideas, and a model wrote, rewrote, or polished the prose.

Each model that we evaluate is given the same zero-shot prompt and a single document, with no tools or reference texts. The prompt asks for two judgments on both ideas and prose, along with accompanying information such as key quotes from the document, reasoning for a decision, decisive factors, and an estimate of the share that AI produced from 0 (all human) to 1 (all AI). The exact prompt is in our repository.

Models. We evaluate 72 models from 13 labs, released between July 2024 and November 2026.[4] We choose leading closed- and open-weight models released since July 2024 using the Epoch Capabilities Index (ECI) (Ho et al., 2025) and Artificial Analysis,[5] two standardized aggregated measures of general capabilities.

Data. We pool 37,575 documents from 24 public datasets (§B.1) of peer reviews, student essays, news, fiction, Wikipedia and web text, and academic writing. Most pair human documents with fully AI ones, such as MELD (Li et al., 2026) and StoryScope (Russell et al., 2026b). Some also have models rewrite or polish human documents, like ARB (Perrone & Romano, 2026), OpAI-Bench (Bsharat et al., 2026) and GEDE (Gehring & Paaßen, 2025). Finally, the TwiceTold dataset (Rajendhran et al., 2026) has people write stories in their own words from a model's ideas. We run a subset of ten models on the full corpus, and evaluate accuracy based on 36,427 documents that all models answered, removing refusals and ambiguous labels (§B.2). To measure capabilities over time, we run all models on

[4]We especially focus on Claude Fable 5.2 and Opus 5.5 (Anthropic, 2026c;e), GPT-6 Astra (OpenAI, 2026b) and GPT-6.1 Sol (OpenAI, 2026a), Gemini 4 Argon and Carbon (Gemini Team, 2023; Google DeepMind, 2026a;b), and DeepSeek V4.1 Flash (DeepSeek-AI, 2026); the full list of models is in Table 7.

[5]https://artificialanalysis.ai/

Task instructions

RQ1 Human or AI?
TASK: Judge two separate things about the attached piece of writing [...] who originated its **IDEAS**, and who wrote its **PROSE**. For each, decide whether it was **a human, an AI language model, or both**.

RQ2 Which model?
TASK: Judge which **AI language model** wrote the attached piece of writing. Name the **developer, model family and specific version** as precisely as you can.

RQ3 Which human?
TASK: Judge which **human author** wrote the attached piece of writing. Name the author as precisely as you can (**the full name they publish under**). Give your best answer even when you are unsure [...]

Explanation format, shared by all tasks

KEY OBSERVATIONS: List every major feature [...] as numbered items in exactly this form:

	OBSERVATION	LEANS / POINTS TO	STRENGTH	QUOTE	WHY
1.	<the feature, in one sentence>	RQ1 <human or ai> RQ2 <a model> RQ3 <an author>	<low, medium or high>	"<a passage copied exactly [...]>"	<why the feature points [...]>
2.	<...>	<...>	<...>	"<...>"	<...>
⋮	⋮	⋮	⋮	⋮	⋮

REASONING: In prose, weigh the evidence [...] how the strong and weak evidence compare, and whether the evidence is consistent [...]
DECISIVE FACTOR(S): Name the factor or factors that decide the attribution, and explain why they outweigh the rest.

Prediction format

RQ1 Human or AI?
ideas_ai_share: <0 to 1>
ideas_prediction: <one of: human, ..., ai>
prose_ai_share: <0 to 1>
prose_prediction: <one of: human, ..., ai>

RQ2 Which model?
confidence: <0 to 1>
model: <developer, model family and specific version>

RQ3 Which human?
confidence: <0 to 1>
author: <the author's name or Reddit username>

Figure 3. **Prompt excerpts across the three tasks.** For each task, the task instructions, explanation format, and prediction format are combined into one prompt, sent to the detector in a single call. Task instructions and prediction fields vary by task, while the explanation format is shared. The first task (*Human or AI?*) judges ideas and prose separately, while other tasks name a single model or human author. Ellipses mark omissions and placeholders are abbreviated.

an 860-document sample. **Since public datasets may appear in training data, we also evaluate on 200 stories and 1,000 news articles published after models' knowledge cutoffs (§2.3).**

Evaluation metrics and baselines. Unless otherwise specified, we follow IdeaLens and report *accuracy* macro-averaged over the four cases, so that each case counts equally regardless of how many documents it contains. An attribute counts as AI when its share is at least 0.5 (shares range from 0 to 1), our default threshold.[6] Recall is also evaluated at a 1% false-positive rate (FPR) on English human text.[7] We also compare models against three trained detectors: ProseLens (Rajendhran et al., 2026) and Pangram 4.1 (Glickenhaus et al., 2026) for prose, and IdeaLens (Rajendhran et al., 2026) for ideas.

[6]Pangram 4.1 labels text as human, mixed, or AI; however, since it is unclear how mixed text should be treated, we apply the same 0.5 threshold to its AI share, counting AI-assisted text as half. Unweighted micro-averages are described in §B.2.

[7]False-positive rates use the 9,687 English fully human documents whose human authorship we can trust (§B.1).

2.2. Results

On all documents (§B.2), GPT-6.1 Sol, Opus 5.5, and Gemini 4 Carbon are competitive with trained detectors, achieving 94–95% accuracy on prose and 89–90% on ideas, compared to 87% for IdeaLens on ideas; on prose, over the three cases Pangram 4.1 is scored on, they reach 98–99% against its 98%.

Pangram and our benchmarks define AI-edited text differently. Our human-ideas, AI-prose case counts any human text that has passed through an LLM as AI prose, however light the edit: OpAI-Bench's lightest version changes a median of 2% of words. In contrast, Pangram only labels text as AI-assisted after some threshold of AI editing has been reached, which is determined by soft n-gram matching (Glickenhaus et al., 2026, §3.5). Because the two definitions assign these documents different labels, we do not score Pangram on this case; its prose accuracy is the mean of the other three (calibration in Figure 10).

LLMs have near-zero false-positive rates (FPR). In real-world settings—teachers grading essays or books reviewed for prestigious awards—falsely accusing a human writer is far worse than not labeling AI text (Liang et al., 2023), making a near-zero FPR essential for any serious detector.

Table 1. Frontier closed-weight models now match trained detectors and significantly lead open-weight LLMs in overall detection abilities. We report accuracy (%) on the 860-document sample of the full corpus. †Pangram 4.1 is not scored on human ideas in AI prose, where its definition of AI-edited text differs from the benchmarks' (§2); its average covers the other three cases.

	Prose accuracy (%)					Ideas accuracy (%)				
Prose by	Human	Human	AI	AI		Human	Human	AI	AI	
Ideas by	Human	AI	AI	Human	Avg.	Human	AI	AI	Human	Avg.
Closed-weight LLMs										
Claude Fable 5.2	**100.0**	**100.0**	96.5	**81.2**	**94.4**	**100.0**	**98.0**	89.2	**99.3**	**96.6**
Claude Opus 5.5	**100.0**	**100.0**	95.8	77.5	93.3	99.6	64.0	88.5	95.9	87.0
GPT-6 Astra	99.6	98.0	98.3	80.6	94.1	**100.0**	80.0	90.2	95.9	91.5
GPT-6.1 Sol	99.6	96.0	**98.6**	79.6	93.4	**100.0**	72.0	91.3	96.6	90.0
Gemini 4 Carbon	99.6	**100.0**	96.5	78.6	93.7	99.6	62.0	**93.7**	98.0	88.3
Gemini 4 Argon	99.6	**100.0**	96.9	76.9	93.3	99.6	30.0	90.6	98.0	79.5
Open-weight LLMs										
Kimi K3.1 Pro	98.7	76.0	95.1	61.9	82.9	99.6	40.0	85.4	89.5	78.6
GLM-5.3	99.1	84.0	92.7	52.7	82.1	**100.0**	38.0	89.2	77.5	76.2
Qwen3.8 27B	99.6	70.0	89.5	51.4	77.6	**100.0**	36.0	74.6	83.0	73.4
DeepSeek V4.1 Flash	85.6	56.0	89.5	53.4	71.1	86.5	58.0	82.2	65.0	72.9
Trained detectors										
Pangram 4.1	**100.0**	96.0	96.5	–	97.5†	–	–	–	–	–
ProseLens	**100.0**	**100.0**	93.4	48.2	85.4	–	–	–	–	–
IdeaLens	–	–	–	–	–	**100.0**	68.0	83.8	93.4	86.3

Based on 9,687 human documents, Pangram 4.1 flags none as AI on prose and Gemini 4 Carbon flags one (0.01%), and Carbon beats IdeaLens on ideas (0.00% against 0.21%) (Table 10); however, the rarity of errors precludes definitive model performance rankings. FPR does rise on newer text: Carbon flags 0.5% of 812 post-release news articles that Pangram 4.1 labels human (Table 13). These may not be true false positives, since authorship cannot be independently verified and some articles may come from newer models with unfamiliar profiles. Dirty data such as pre-2022 web snippets with AI-written summaries appended or machine-translated web spam (§B.1) and all non-English data are excluded from this set on the suspicion of machine translation involvement.

Model accuracy has improved over time. In the 860-document sample (Table 11), accuracy on both labels climbed from 27% for GPT-4o in 2024 (55% on prose, 54% on ideas) and 59% for GPT-5.2 in 2025 (75%, 80%) to 91% for present-day frontier models (Claude Fable 5.2: 94%, 97%). The best open-weight models still trail, with Kimi K3.1 Pro reaching 63% (83%, 79%). Across the 61 models with an ECI score, accuracy correlates strongly with overall capabilities ($r = 0.94$; §B.2). Newer models are also cheaper than commercial detectors, with Sol costing $27 per 1,000 documents at standard API prices (Table 7) versus $376 on Pangram 4.1 at its standard API price ($0.05 per 100 words).[8] We summarize the available experiment cost records in Table 42.

[8]The average length for a document in our corpus is 751 words.

2.3. Robustness & Explanations

Because these benchmarks are public, the high LLM accuracy could be due to contamination in pretraining data. We explicitly test this by measuring agreement between LLMs and Pangram on texts published after the release dates of all evaluated models. Then, we examine agreement between explanations generated from different models for the same text, and we also compare the explanation content to prior analyses of AI writing.

Detectors maintain high accuracy on post-release documents. Most of our benchmarks are public, so the detectors may have seen them during training. On 200 stories published or generated after the detectors' training cutoffs, accuracy is only at most 3.4% lower than on the full corpus's fiction (Table 14).[9] Additionally, on 1,000 news articles published after all model releases, frontier models maintain high agreement with Pangram 4.1 (93–98%) and falsely label 0.5–5% of the human-written articles, as classified by Pangram, as AI-written (§B.3.1). Frontier models may struggle more with this data because newer models could have generated these articles.

Dates are referenced, but not decisive in strong detector predictions. In 9.5% of its explanations of fully human (human ideas and human prose) texts, Opus 5.5 argues that the text predates LLMs because of a date in

[9]Both detectors score GPT-6.1 Sol's stories above Claude Opus 5.5's, showing minimal effect of recognizing their own writing.

the text ("before 2022"). When it makes this argument on human-ideated, AI-written texts its prose accuracy falls from 80.3% to 35.4%. Replacing every year in 190 dated documents with 2026 changes at most four of Opus 5.5's, GPT-6.1 Sol's or Gemini 4 Carbon's 190 verdicts per attribute, no more than rerunning the same prompt (Table 15). On the other hand, DeepSeek V4.1 Flash (open-weight model) changes 29 ideas verdicts to AI and one to human.

Explanations are consistent across detectors and echo known traits of AI writing. Taking advantage of LLMs' capabilities to generate judgments accompanied by explanations (Huang et al., 2023; Zheng et al., 2023), we ask the detectors to disclose the key observations used to make their decisions, where each observation includes quoted evidence, whether it suggests a human or an AI, and why (Figure 13).[10] This also raises accuracy: Sol gains more than 10 points of idea accuracy on human ideas in AI prose (Table 12). Across pairs of detectors, the explanations largely cite the same evidence and read it the same way, but agree less on which piece of evidence decided the verdict.[11] The cues they cite match prior work (Table 18): hedged assertions point to AI prose and explicitly stated themes and morals to AI ideas (Kim et al., 2026; Russell et al., 2026b), while precise factual and technical detail points to human ideas (Rajendhran et al., 2026).

3. Which model wrote the text?

Frontier models can tell whether a human or an AI wrote the text (§2), but can they tell which model wrote it? Prior work finds that the writing of different LLMs is highly homogeneous (Jiang et al., 2026; Kim et al., 2026), that their stories converge on a shared narrative space (Russell et al., 2026b), and that frontier models from different labs share a dense style that is harder for people to read (Clark et al., 2026). This convergence makes attributing a text to its source model much more difficult, especially given that there are numerous labs, each with several model versions. We ask five frontier models to name the model that wrote each text, with no list of candidates. We find that their errors are consistent with public reports of distillation (Anthropic, 2026b; Seetharaman & Arámburo, 2026): for example, the three strongest detectors attribute almost every text by Moonshot's Kimi K3 to Claude.

Task and evaluation. We evaluate five frontier models, Claude Fable 5.2, Claude Opus 5.5, GPT-6.1 Sol, Gemini 4 Argon, and DeepSeek V4.1 Flash.[12] Using a minimally modified version of the human vs. AI prompt (§2), each names the model it thinks wrote the text in a zero-shot setting, without tools or a candidate list. Along with its answer, each detector then gives its key observations, reasoning and decisive factors for its decision in the format of §2.3. We score each answer at three levels: the lab (e.g., Anthropic), the model family (e.g., Claude Opus) and the exact model (e.g., Claude Opus 5.5). We report the share of texts attributed correctly at each level.

Table 2. Percentage of texts correctly attributed to their lab, model family, or exact model. Lab accuracy is also split by the generating lab: Anthropic, Google, and OpenAI (logos), and the other eight labs. See Table 20 for per-model results.

	Attribution level			Lab by group	
Detector	Lab	Family	Model	A\G	Other
Fable 5.2	**59.9**	**41.0**	**10.7**	**99.8**	**35.1**
Opus 5.5	49.7	31.6	5.8	97.8	19.8
Gemini 4 Argon	44.8	27.9	7.1	97.5	12.0
GPT-6.1 Sol	22.6	7.9	0.1	58.8	0.2
DeepSeek V4.1 Flash	11.1	3.1	0.0	28.7	0.1

Data. We generate new texts for the model attribution task in the following three domains (Table 19), using 34 models from 11 labs released between September 2024 and October 2026:

- **Stories**: We take 25 story prompts from StoryScope (Russell et al., 2026b), a released set of AI-written fiction, and generate one story per prompt.[13]
- **News**: We take 25 articles[14] from 17 outlets, published from 9–20 October 2026. Each model receives an article's topic and length and writes its own article.
- **Op-eds**: We take 25 English briefs from MARGO (Gill et al., 2026), a dataset of op-eds written in eight languages by five models. Each model receives a brief and writes an op-ed from it.[15] Briefs in the other seven languages are covered in §C.2.

3.1. Lab and model attribution

Fable 5.2, Opus 5.5 and Argon identify the big three labs almost perfectly, while Sol and DeepSeek default to OpenAI and Anthropic. Identifying a lab is important to understand which provider's model wrote a text. Fable 5.2,

[10] The exact prompt is in our repository.

[11] A Claude Opus 5.5 judge, checked against two human annotators, matches the evidence in each pair of explanations on 2,854 documents. Agreement statistics in §B.3.2.

[12] Each detector is run at medium effort, except DeepSeek V4.1 Flash, which has no medium level and runs at high.

[13] We reuse the original StoryScope stories for the five models it covers.

[14] Pangram 4.1 labels 23 of the 25 articles human and two, both from Global Voices, mixed.

[15] MARGO was released after all five detectors became available. We reuse its original op-eds for five writing models.

Opus 5.5 and Argon can identify text from OpenAI, Anthropic, and Google with 98% accuracy or more. Opus 5.5 does so even for the two models released after it (Sol and Argon). Their overall accuracy is much lower, at 60%, 50% and 45% (Table 2), because they rarely recognize models from the other eight labs, though Fable 5.2 does so far more often than Opus 5.5 (35% vs. 20%). Sol and DeepSeek lag far behind at 23% and 11%, and mostly answer OpenAI and Anthropic, so we focus on Fable 5.2, Opus 5.5 and Argon for the rest of this section.

Naming the exact model depends on how long it has been public. Even Fable 5.2, Opus 5.5 and Argon rarely name the exact model, and instead name an older model from the same lab. How often they name it depends on how long the model has been public—Opus 5.5 names the exact model for about half of the texts from the three models released before mid-2025, but for none from the 24 models released in 2026. Fable 5.2 does better on the three older models (64%), but names the exact model for only one of 1,792 texts from the 2026 models.

Misattributions of new open-weight models to Anthropic and OpenAI are consistent with reported distillation. A model trained on another lab's outputs can inherit that lab's style, so a detector may name the lab it learned from. Anthropic has reported distillation attacks on Claude by DeepSeek, Moonshot, and MiniMax (Anthropic, 2026b), and later named Zhipu and Alibaba (Anthropic, 2026d). Consistent with these reports, Fable 5.2, Opus 5.5, and Argon attribute nearly all Kimi K3 and GLM-5.3 texts to Claude (89% or more), even though Fable 5.2 names Moonshot for 93% and 77% of Kimi K2.5 and K2.6 texts, the two releases before K3. Similarly, Elon Musk testified that xAI "partly" distilled OpenAI models (Musk v. Altman, 2026) and a US advisory says that Alibaba and MiniMax distilled GPT-5 (National Security Agency et al., 2026): detectors attribute most texts by Grok 4.7, Qwen3.8 Max, MiniMax M3, and M3.1 Flash to GPT.[16]

3.2. Model writing profiles

Fable 5.2, Opus 5.5 and Argon tell the big three labs apart almost perfectly, but what do they think each lab reads like? Each detector explains its answer for every text (§3), so we can pool explanations into a writing profile of each lab.[17]

Fable 5.2, Opus 5.5, and Argon hold consistent profiles of each lab. The detectors describe the big three labs in largely the same way (Table 3).[18] **GPT** hedges its claims, flags gaps instead of inventing specifics, and leans on aphorisms. **Claude** treats opposing views charitably, writes with dry wit, and ends quietly. **Gemini** stacks adjectives and piles on intensifiers and stock clichés, and invents named sources and quotes in its news articles. All three also tie the same punctuation habits to each lab, though Fable 5.2 and Argon cite them more than twice as often as Opus 5.5. GPT uses curly quotes and drops the Oxford comma, Gemini mixes curly and straight quotes, and Claude puts spaces around its dashes. For Claude, Opus 5.5 and Argon tend to identify different traits, while Fable 5.2 finds most of both.

Kimi K3 and GLM-5.3 share Claude's traits, while Grok 4.7, Qwen3.8 Max and MiniMax share GPT's. A detector may attribute a text to a lab either because the text shares that lab's traits or because the detector does not recognize the model that wrote it and falls back to a known lab. For most open-model texts attributed to Claude it is the latter: the detectors cite few of the traits that set Claude apart. Kimi K3 and GLM-5.3 are the exception. All three detectors note the same dry wit in their texts as in Claude's. Opus 5.5 also notes Claude's bookend endings and closing quotations, and Argon notes the same spaced dashes and parenthetical asides. The detectors find all of these far less often in earlier Kimi and GLM models' texts.[19] Likewise, Fable 5.2 and Opus 5.5 also find several of GPT's traits in Grok 4.7, Qwen3.8 Max, MiniMax M3 and M3.1 Flash texts, with little to no Claude traits.

4. Which human wrote the text?

While the space of plausible outputs in the model attribution task is just a few dozen models, naming the *person* behind a human-written text is an open-world task with countless possible outputs (Koppel et al., 2011; Luyckx & Daelemans, 2011) and much less data available per author. To evaluate this capability while mitigating the confound of data contamination, we collect 18,677 texts from six domains, each either *private* (never posted publicly) or *recent* (published after each tested model's knowledge cutoff, and after most of their release dates). Even so, frontier models unmask many authors across all of these domains from just a single text, given a minimally modified prompt from the previous sections and no tools, web access, or candidate lists. The best-performing model we test is Claude

[16] To check that these patterns are not an artifact of our detectors falling back to a known lab for models released after knowledge cutoff, we also run a stylometric analysis, and find the same shifts, even after removing the information-dense style shared by recent models (Clark et al., 2026) (§C.3).

[17] We group explanations into cues and keep those the detector notes on a lab's texts significantly more often than on any other lab's (§C.3). We treat the other eight labs as one lab.

[18] We show only cues that a detector cites on at least 25% of a lab's texts within a domain (stories, news, op-eds), and at least 15 points and 1.5 times more often than on any other lab's texts.

[19] Claude Opus 5.5 and Gemini 4 Argon attribute most Kimi K2 texts (88% and 93%) and nearly all Kimi K3 texts to Claude, but they find far more of Claude's traits in K3's texts than in K2's.

Table 3. Opus 5.5 and Argon agree on what sets each lab's writing apart. Each cell gives examples of the cues a detector often cites for that lab, using only texts where it names the right lab. Highlighted cues are often cited by both detectors. Full model profiles can be found here.

Lab	Opus 5.5	Gemini 4 Argon
OpenAI GPT	GPT's stories run on aphoristic lines and clipped, sparse dialogue. Its news flags gaps in the evidence and does not fill them, hedges and calibrates its claims, spells out what it is not claiming, and stays generic where other models supply specifics. Its op-eds return to aphoristic lines and hedge their claims. STYLE *one-line paragraphs (stories); no serial comma (news); balanced antitheses (op-eds).*	GPT's stories run on clipped, sparse dialogue and aphoristic lines, and keep the climax understated. Its news flags gaps in the evidence, hedges and calibrates its claims, avoids inventing specifics and separates concepts that are easily conflated. Its op-eds return to aphoristic lines, pull apart easily conflated concepts and end paragraphs on a kicker. STYLE *one-line paragraphs (stories); balanced antitheses (stories); no serial comma (news).*
Anthropic Claude	Claude's stories return to recurring motifs, end quietly and call back to their opening. Its news keeps a balanced, non-partisan stance, gathers reactions from several stakeholders, gives quoted speakers polished, aphoristic lines and closes on a quotation. Its op-eds are dense with accurate real-world detail, grant an opposing point before countering it and treat the other side charitably. STYLE *clauses strung together with repeated "and", or with no conjunctions at all (stories); spaced dashes (news).*	Claude's stories correct themselves mid-sentence and repeat a word for emphasis. Its news gives quoted speakers polished, aphoristic lines, attributes them to sources that are unnamed or labeled by role, and includes specialist detail. Its op-eds are dense with accurate real-world detail, adopt the voice of a publication and end on a crafted final line. STYLE *clauses strung together with repeated "and", or with no conjunctions at all (stories); "the way…" phrasing (stories); spaced dashes (news); short, punchy sentences (news, op-eds).*
Google Gemini	Gemini's news gives invented people common default names. STYLE *stacked adjectives (stories); clustered sensory detail (stories); italics for emphasis and thought (stories); mixed curly and straight quotes (stories); datelines (news); intensifiers and dramatic adjectives (news, op-eds).*	Gemini's stories give invented people common default names. Its news uses the same common default names. STYLE *mixed curly and straight quotes (stories); datelines (news); scare quotes (news); stock idioms and clichés (op-eds).*

Fable 5.2, which names the reviewer of 55% of 15K anonymous ICML 2020 reviews and the username behind 31% of recent Reddit comments. Paraphrasing a human text with AI is only a partial defense, as content-related cues persist.

4.1. Experimental setup

Given a single text and a minimally modified version of the prompt from previous sections, models return the human author's name (or username for Reddit), a list of supporting observations, and a confidence level. Unlike Lermen et al. (2026), who augment LLMs with web search and candidate pools to match users across platforms, we provide no tools, web access, or candidate lists. To keep personal account information out of the models' context, we query every model through an API, not a chat app or coding CLI (Table 24).[20] In an early run, we observed the Claude Code CLI quote the account user's email when attributing a text the account owner wrote, a phenomenon studied in Zhong et al. (2026).

Private Data. We collect two sets of peer reviews whose reviewer identities were never made public and therefore could not have been memorized from model training data.

- **ICML 2020 peer reviews**: 15,136 peer reviews written by 3,342 unique reviewers for ICML 2020 and hosted on Microsoft's Conference Management Toolkit (CMT). We can study these reviews because coauthor Hal Daumé III served as program co-chair for the conference and has IRB approval to analyze them. Daumé ran all of the experiments on this corpus, so reviews were never shared with the other authors. We report only aggregate results and do not reveal any text from individual peer reviews or reviewer identities.

[20] There were a few exceptions run through the Claude Code CLI due to budget considerations, though only when familiarity with the CLI account holder was unlikely to impact accuracy, including the Reddit reasoning-effort comparison (Table 6), as well as Opus 5.5's name-only familiarity ratings of Reddit accounts and Nature reviewers.

- **Reviews from this paper's authors**: 398 peer reviews written between 2015 and 2026 and voluntarily contributed by four of this paper's coauthors (JR, JW, MK, MI), as well as two PhD students in the UMD CLIP Lab. The group includes faculty members, industry researchers, and students. While the number of unique reviewers is much lower than in the ICML 2020 set, these reviews were submitted to 26 distinct double-blind ML/NLP venues (main conferences and workshops) such as ACL, EMNLP, and NAACL. They were pulled from Softconf and OpenReview, with none of the reviewer identities and very few[21] of the review texts being public.

Recent Data. We collect four datasets of publicly available writing published online after September 22, 2026, the release date of Claude Opus 5.5. Since a model's reported knowledge cutoff generally applies only to pretraining and not post-training, only data published after a model's release date is completely safe from contamination.[22]

- **Public peer reviews**: 1,559 peer reviews from 47 open-review journals, including 90 reviews from the Nature family. Unlike ICML and our private reviews, these reviews are from disciplines outside AI/NLP and provide a mix of prominent expert reviewers (from Nature journals) and lesser-known researchers.
- **NLP papers**: 120 papers uploaded to arXiv between 19–22 October 2026. Papers were sampled from the most recent papers categorized as cs.CL (Computer Science, Computation and Language) on arXiv.
- **Stories**: 206 short stories and novel excerpts from 187 authors primarily gathered from literary magazine sites, posted between 23 September–23 October 2026.
- **Reddit**: 1,258 comments from 870 unique usernames in 32 subreddits, posted between 18–29 October 2026. Each comment is at least 200 words and excludes quoted text. Active subreddits with such long posts generally lean towards technical subjects (e.g., programming, math) as shown in Table 25.

Models and evaluation. We benchmark four frontier models from Anthropic (Fable 5.2, Opus 5.5), OpenAI (GPT-6.1 Sol), and Google (Gemini 4 Argon), as well as one open-weight model with accessible chain-of-thought (DeepSeek V4.1 Flash).[23] We report exact match accuracy (after normalizing titles, accents, and punctuation) for each dataset other than ICML 2020, for which we employ a more lenient match that allows for abbreviated first names and omitted or added middle names.[24] For NLP papers, we consider a prediction correct if the model names any co-author, while confirmed pen-names are verified and counted as correct for fiction. For Reddit texts, we use a slightly modified prompt to ask specifically for usernames. For public reviews and NLP papers, accuracy is reported on texts that Pangram 4.1 does not label AI (human and mixed labels), unless otherwise specified.

4.2. Results

Despite the difficulty of this task, frontier models are surprisingly effective on both private and recent datasets, and they are rapidly improving. Fable 5.2, the best performing model in all domains (Figure 4), exhibits considerable gains over Opus 5.5, which was released less than two months prior.

Models can effectively deanonymize peer reviews. On the ICML 2020 corpus, Fable 5.2 correctly identifies the author of 55.4% of the reviews, followed by Opus 5.5 (48.8%), Argon (37.7%), Sol (22.3%), and DeepSeek (5.4%). On private peer reviews, Fable 5.2 also leads with 56% accuracy, while on public reviews from the *Nature* journal family it reaches 78%. The private reviews consist of 266 standard reviews and 132 meta-reviews, the latter of which are much shorter in length, come from two authors, and are harder to attribute. Fable 5.2 names the writer of 68% of reviews but only 31% of meta-reviews. Opus 5.5 and Argon also have twice the accuracy on standard reviews as on meta-reviews (Table 32). DeepSeek, the only open model, performs very poorly in all domains despite reporting the highest average confidence levels; its chain-of-thought reveals no clear reasoning insights and consists primarily of hedging and indecision.

Models can name pseudonymous users. People do not have to be well-known or even publish under their real name to be detected. Models can detect Reddit usernames from comments without any additional context: Fable 5.2 identifies the Reddit user behind 31% of com-

[21] The text of 19 of these reviews is publicly available on OpenReview without an associated reviewer identity.

[22] By design, Claude Opus 5.5 is completely free from training data contamination in our experiments. Eight percent of public data is from before GPT-6.1 Sol's release, 23% is before Gemini 4 Argon, and 28% before DeepSeek V4.1 Flash; pre-release data stems only from stories and public peer reviews. Claude Fable 5.2 was released about 10 days after all public data collection completed, and its knowledge cutoff is June 2026.

[23] All models run at the provider's default reasoning effort: Claude Opus 5.5, GPT-6.1 Sol and Gemini 4 Argon at medium and Claude Fable 5.2 and DeepSeek V4.1 Flash at high. See §5 for how reasoning affects detection, and Table 26 for results with Gemini 4 Carbon.

[24] Claude Fable 5.2 names the reviewer of 50.1% of ICML reviews with exact match vs. 55.4% with the more lenient matching, and qualitative inspection shows few false matches in the latter setting.

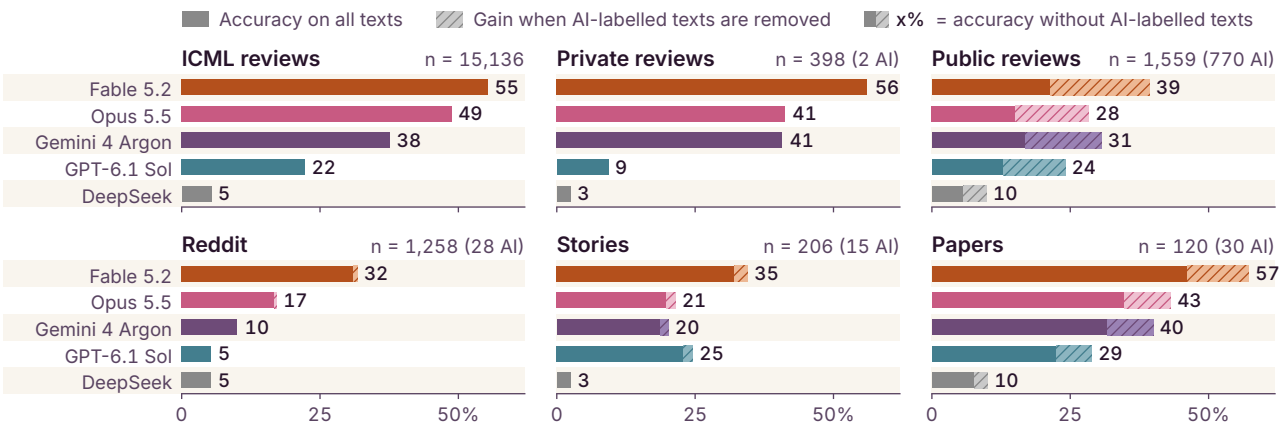

Figure 4. Author accuracy by domain on all human authorship attribution datasets. Solid bars show accuracy on all texts in the domain; striped portions indicate the gain in performance when texts Pangram 4.1 labels as AI are removed. The printed numbers are accuracy on non-AI labelled texts (includes human and mixed). Claude Fable 5.2 names the most writers in every domain.

ments, nearly double Opus 5.5 which identifies 17%. Accuracy varies greatly based on subreddit. Fable 5.2 names 79% of comments in r/cpp and 50% in r/ProgrammingLanguages, but only 13% in r/AskAnthropology. This may also reflect training data sampling procedures from frontier labs, as subreddits about programming might be valuable to improve coding performance. As discussed in §5, pseudonymity is especially fragile for active users in niche communities. Fable 5.2 names 26% of the 860 users at least once (§D.3.2).

Confident answers are almost always right. Each model's prediction includes a confidence level. Claude models tend to understate theirs. On public data, Fable 5.2's mean reported confidence is 17% while achieving 27% accuracy. When Fable 5.2's confidence is 25% or more (25% of responses), its predictions are 87% correct (Figure 18). Furthermore, confident mistakes are rare. Of all Fable 5.2 answers, 0.3% report both at least 80% confidence and name the wrong person; the corresponding percents are 0.4% for Opus 5.5, 3.3% for Argon, and 5.1% for DeepSeek. The threat to anonymity therefore is greater than the accuracy levels suggest; one can rely on the strongest model's confident answers and be right almost every time.

Contamination is unlikely to be causing the high accuracy. Models do well on our private datasets, whose authorship was never made public. In fact, Fable 5.2, Opus 5.5, and Argon perform better on our private datasets than on the public Reddit comments and stories. All of the public data is posted after the release date of Opus 5.5, which correctly names an author of 43% of recent NLP papers[25] and the usernames behind 17% of Reddit comments. Accuracy instead correlates with training data exposure and familiarity, which we discuss in §5.

Accuracy drops on AI-generated texts. Since our prompt explicitly requests a human author's name, models perform poorly on texts across the four recent datasets that Pangram 4.1 labels as AI-generated. The discrepancy is most noticeable in public peer reviews and papers that increasingly contain AI-written text (Liang et al., 2024; Emi, 2025; Naddaf, 2025). Pangram flags 770 of the 1,559 public reviews as AI-generated, and models name the author of only 1–4% of these reviews (Table 27). Omitting AI-generated texts, visualized with the striped bars in Figure 4, raises accuracy on public reviews by 75–87% for every model (Opus 5.5 rises from 15% to 28%), and raises it on papers. The remaining domains have negligible numbers of AI-generated texts.

Longer texts are easier to attribute for all models. The correlation between length and accuracy is strongest for public reviews and stories, the latter of which has the widest range of lengths (181–20,947 words). Fable 5.2 identifies only 13.8% of 29 stories under 500 words and 43.3% of 30 over 4K words (Figure 5). These results align with findings in AI-text detection, where shorter texts are harder to detect (Emi & Spero, 2024).

Attribution accuracy is stable across multiple runs. We test the stability of Fable 5.2 predictions on Reddit comments over five independent runs. Overall accuracy stays nearly identical, between 29.9–31.1%, though there is some variation in predictions. Only 32.6% of comments

[25] 43% is Opus 5.5's accuracy on the 90 non-AI written papers as labeled by Pangram. Opus 5.5 names 35% of all 120 papers.

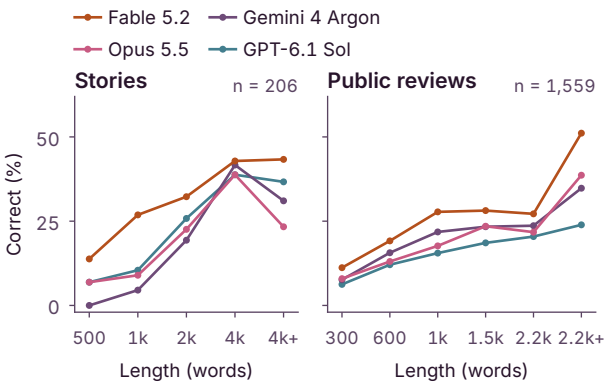

Figure 5. Author attribution accuracy increases with the number of words in a text. Tick labels are the upper limit of each length bin.

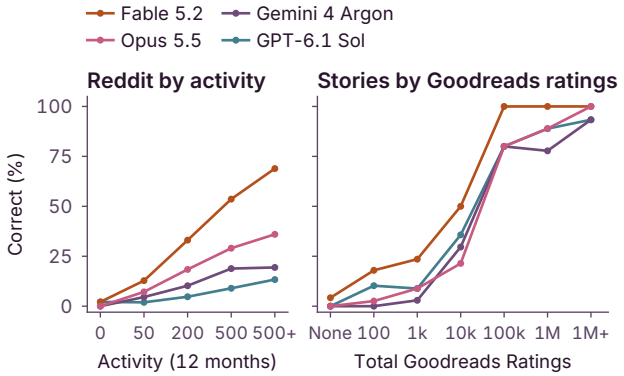

Figure 6. Accuracy is higher for more active Reddit users and authors with many Goodreads ratings. **Left:** Reddit commenters, binned by total number of posts and comments in the subreddit over the previous 12 months. **Right:** authors, binned by the total number of Goodreads ratings across their books; "None" indicates authors without a Goodreads profile (69/187).

receive the same author prediction every time. When predictions change, it is usually from one wrong answer to another; 63.2% of comments are predicted incorrectly in all five runs, while 25% are predicted correctly in all five runs. Familiar authors are predicted far more reliably. Fable 5.2 is correct in all five runs for 92.7% of the 96 comments from the 39 most familiar users (rated 3), but for only 0.4% of 281 comments from completely unfamiliar users (rated 0). Accuracy remains stable over other domains as well; see §D.4.4.

4.3. What aspects of the text give authors away?

We analyze the generated explanations to see which features of the text each model relies on when naming the author. Generally, the explanations cover both the text's content as well as its writing style.[26] For instance, when identifying the author of the short story "Lost Child" (published in "The New Yorker" magazine in September 2026), Fable 5.2 cites "this particular pairing of syntax and subject matter" as decisive factors when correctly predicting Joyce Carol Oates (Table 5).

Self-citations make reviewers easy to name, but few reviews contain them. Reviewers often cite their own papers in their reviews, and these self-citations have historically been used to identify authors under blind review (Matsubara & Singh, 2020; Caragea et al., 2019; Le Goues et al., 2018). Of the 1,559 public reviews, only 106 cite the reviewer's own work (82 explicitly list reviewer's name in citation, 24 do not). On these reviews, Fable 5.2 names the reviewer 80% of the time, compared with 17% of reviews without a self-citation (Table 4).

[26]Their claims are also largely factual. We extract verifiable claims from Opus 5.5, Argon, Sol and DeepSeek explanations for 2,000 ICML 2020 reviews. Gemini 4 Argon with Google Search Grounding marks 95.9% of the 5,414 verifiable claims it assessed as supported by external evidence (§D.4.6).

Table 4. Accuracy (%) with and without self-citations in the text; bold marks the best per column in each group. Self-citation is an obvious clue to author identity, but models still name writers without them. The first row of each group is accuracy on all its texts; the indented rows below are accuracy on subsets, with and without self-citations.

Texts	*n*	Fable 5.2	Opus 5.5	Gemini 4 Argon	GPT-6.1 Sol
All public reviews	1,559	21.4	15.2	16.9	13.0
Cites own work	106	**80.2**	**76.2**	**76.4**	**60.4**
No self-citation	1,453	17.1	10.7	12.5	9.5
↪Not AI-labelled	704	33.7	21.7	24.6	18.8
All papers	120	46.2	34.7	31.7	22.5
Cites own work	18	**72.2**	**61.1**	**66.7**	**50.0**
No self-citation	102	41.6	30.0	25.5	17.6

The full text, not just its topic, is necessary to name a paper's authors. A model could name the author of a paper by knowing who works on its topic, without recognizing anyone's writing. To measure how far topic knowledge alone contributes, we take the 34 cs.CL papers posted to arXiv after Opus 5.5's training cutoff that Pangram 4.1 labels human. Opus 5.5 receives either five keywords describing each paper's topic or its title and abstract, and is asked to name two researchers (early career and senior) who work on the problem. It is explicitly told that it is not being asked who wrote the paper. A given name is correct if it is any of that paper's co-authors (Table 36). We compare the minimal topic inputs with our default method of giving the full main text. Topic keywords identify an author for only 12% of papers, against 53% from the title and abstract and 76% from the full text. Across all 120 papers, including those Pangram labels mixed or AI, the full-text variant names the senior author of 35%.

Table 5. A\ Claude Fable 5.2 bases each decision on concrete evidence ranging from mixed apostrophes to a reviewer's self-description. Each explanation lists quoted evidence and names decisive factors. In each row, a highlighted reason shares its color with the words in the text it describes.

Task	Text (excerpt)	A\ Prediction	A\ Explanation (excerpt)
Human or AI? Story from TwiceTold: *lantern and firefly*	"The most meaningful light is that which is provided in moments of need. [...] From that day on, travelers to the forest always saw two lights [...] the lantern never again mistook greatness with size."	IDEAS **AI** ✓ 90% AI PROSE **Human** ✓ 10% AI	**Ideas:** "The deciding factor is the concentration of stock choices: [...] and the stated moral with a "from that day on" coda." **Prose:** "The deciding factors are the non-idiomatic and ungrammatical slips: "mistook greatness with size""
Which AI? AI story from *StoryScope*	"Luke didn't hesitate. [...] He looked at Harry's trembling shoulders. [...] The Gulf was a graveyard, but for tonight, the brothers were still ghosts in the making, drifting on the tide of a dying world. [...] Row. Steer. Live. And never, ever laugh."	MODEL **Gemini 3 Flash** ✓ DEVELOPER **Google** ✓ Confidence 40%	"The deciding factor is the combination of the stacked, aphoristic endings with a staccato coda and the mixed curly and straight apostrophes with trailing spaces. The stylistic habits alone could fit several models, but the formatting artifacts are much more specific to Gemini [...] The specific version remains a best guess."
Which human? Review from *Nature Sustainability*	"I am primarily a wellbeing researcher [...] emphasises [...] what matters for policy seems rather to be any substantive magnitudes of any relationships. [...] between long-run changes in GDP per capita and long-run changes in wellbeing"	REVIEWER **Caspar Kaiser** ✓ Confidence 27%	"The deciding factor is the combination of three things: the self-description as "primarily a wellbeing researcher", the specialist long-run reading of the Easterlin Paradox, and UK-style orthography with faintly non-native syntax. The Easterlin point narrows the field to a few specialists, and the spelling and phrasing eliminate most of the others."
Which human? Human story: *Lost Child*	"For, with Kit's father, something was always either going to happen or had already happened, [...] Hurry, sweetie! God damn. Not used to dressing herself. [...] but maybe Daddy wouldn't notice."	AUTHOR **Joyce Carol Oates** ✓ Confidence 93%	"The deciding factor is the combination of style and theme: fragmentary, subject-dropping sentences, sentence-initial "For", and unquoted menacing speech, [...] this particular pairing of syntax and subject matter is characteristic of Oates and outweighs the generic literary-fiction features."

Topics extracted from explanations correlate with accuracy and signal model uncertainty. We use GPT-6.1 Sol to independently generate and assign topics to the explanations generated for five models on 5,000 ICML 2020 reviews and 1,258 Reddit comments (Pham et al., 2024). We calculate the correlation of each topic with attribution accuracy, finding weak to moderate, though statistically significant, correlations for more than half of the topics.[27] For Fable 5.2, the presence of a structure-related topic (form, layout) is negatively correlated with accuracy ($r=-0.36$), while some topics like prior-work references are positively associated with accuracy (Figure 7). Topics like "researcher role" or "anonymous reviewer role" (model describing the reviewer as a "researcher" or "anonymous") have lower than average confidence levels and among the strongest negative correlations. These results suggest that correct predictions cite content-oriented evidence like technical reasoning or references to prior work in their explanations, while incorrect ones contain comments on generic review structure or the author's anonymity.

[27] See Table 41 for a list and description of each topic.

5. Who can be identified?

In this section, we look at what makes certain authors (AI or human) recognizable. We observe that the common factor is exposure in the training data. Detectors are most accurate on the writers they have plausibly seen most during training. We first show how this holds across our three research questions and then introduce a measure of familiarity that works without access to pretraining data. We use this measure to explore how reasoning effort and paraphrasing change who can be identified. Finally, we discuss the broader implications of our experiments for anonymity.

5.1. Exposure and familiarity

Exposure in pretraining data is a plausible common contributor to performance across all author attribution settings. For AI text detection (§2), models now train on a web in which a large share of text is machine-written, which likely improves their ability to detect it. For instance, 31.1% of quality-filtered web tokens from August 2026 are labeled as AI-generated (Russell et al., 2026a).

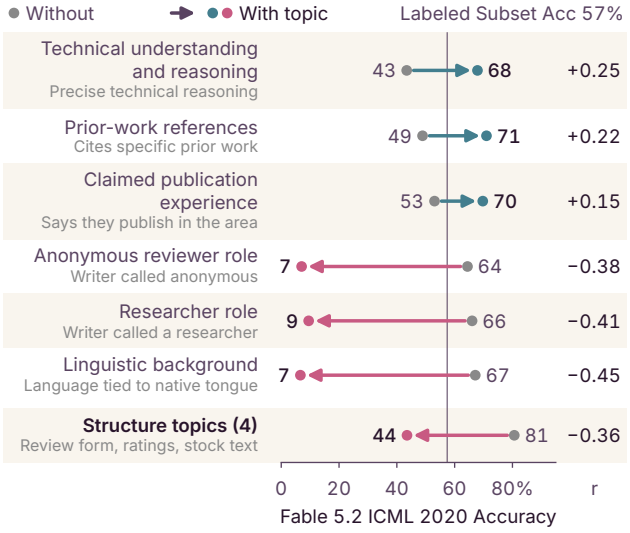

Figure 7. The presence of topics related to structure has a negative correlation with accuracy, while references to prior works or technical reasoning have positive correlation. Claude Fable 5.2's accuracy when naming ICML 2020 reviewers is shown with its strongest positive and negatively correlated topics (top of arrow) and without them (gray dot). Each topic's Pearson correlation is listed on the right side. Labeled Subset Acc is Fable 5.2's accuracy on the 5,000 subset of reviews we labeled.

For attribution among models (§3), detectors name OpenAI, Anthropic, and Google far more reliably than the other eight labs. This is consistent with the model market share reported by Pangram (Masrour, 2026).[28] The rest of this section focuses on what makes *human authors* identifiable: we test several proxies for exposure (e.g., mentions of their names in open pretraining corpora, comment activity on Reddit, visibility and seniority) and find that directly prompting the model for its familiarity with a person correlates most strongly with how identifiable their texts are.

Higher author visibility correlates with increased human attribution accuracy. In §4, we define domain-specific measures for human author visibility (Figure 6). On Reddit, we measure visibility by a user's post and comment count in a given subreddit over the past 12 months. Using this metric, Claude Fable 5.2's attribution accuracy rises from 2% for inactive users to 13% for those with 1–50 posts, reaching 69% for users with over 500 posts. Over the same span, Claude Opus 5.5's accuracy increases from 0% to 36%. For stories, we use the total number of Goodreads ratings across an author's books. Here, Fable 5.2 identifies 4% of stories by authors without a Goodreads profile, 50% by those with 1k–10k ratings, and 100% by authors with over 10k ratings. Similarly, Opus 5.5, Gemini 4 Argon, and GPT-6.1 Sol fail to identify any authors without a profile, but successfully attribute 96% of those with over 1M ratings. A similar trend holds for private reviews (Table 32). Opus 5.5 and Argon correctly identify the two most senior reviewers on 66% and 71% of their 174 ordinary reviews, respectively, compared to 23% and 14% across the 92 reviews written by the other four contributors.

A model's self-reported familiarity predicts attribution. Language models can retrieve details about entities like writers and public figures from their parametric memory (Kadavath et al., 2022; Yin et al., 2023; Ferrando et al., 2025). Because the pretraining data of frontier models is unavailable to inspect directly, we elicit this knowledge by prompting the model to rate its familiarity with a name or username on a 0–3 scale (see our repository for prompt; §D.4.1). We make two independent calls and average results to compute *self-reported familiarity*, which strongly corresponds with attribution accuracy. Opus 5.5 and Argon identify 1.3–4.2% of texts by writers they rate 0, compared to 57.7–60.8% of texts by writers they rate 3. About 90% of Opus 5.5's successful attributions and 79% of Argon's come from writers rated ≥ 1.5, who account for 45% and 30% of texts, respectively (Table 33). The rating also transfers across models, with one model's familiarity score predicting another model's accuracy (Table 34). These results mirror factual recall, where obscure entities are recalled less often (Mallen et al., 2023; Kandpal et al., 2023).

Self-reported familiarity outperforms other exposure proxies. Where we can compare them directly, a model's familiarity rating predicts attribution better than other domain-specific visibility measures. On Reddit, Opus 5.5's rating separates the Reddit accounts it does and does not identify with an AUC of 0.89, compared to 0.78 for posting activity over the prior 12 months. Even among accounts with similar posting activity, Opus 5.5 identifies 17–56% of those it rates familiar (≥ 1.5 average rating), compared to at most 14% of the rest. On ICML 2020 reviews, Fable 5.2 names the reviewer of 0% and 76% of reviews by reviewers it rates 0 and 3, respectively, a monotonic increase with familiarity. The strongest training exposure proxy for these reviews, how often a reviewer's name appears near an ML venue name in Dolma v1.7[29] (Soldaini et al., 2024), has $\rho = 0.31$ correlation with Fable 5.2's per-reviewer recall,[30] against 0.13 for paper count and 0.11 for citations.

[28] We note that this gap is also consistent with distillation from larger labs, and it is likely that both exposure and distillation contribute to weaker performance on these labs' models.

[29] We count occurrences of the reviewer's name within 1,000 tokens (before or after, in the same document) of ICML, NeurIPS, ICLR, or "Advances in Neural Information Processing Systems", using the infini-gram index of Dolma v1.7 (2.6T Llama-2 tokens; Liu et al., 2024).

[30] Per-reviewer recall is the fraction of each reviewer's reviews that Fable 5.2 attributes correctly.

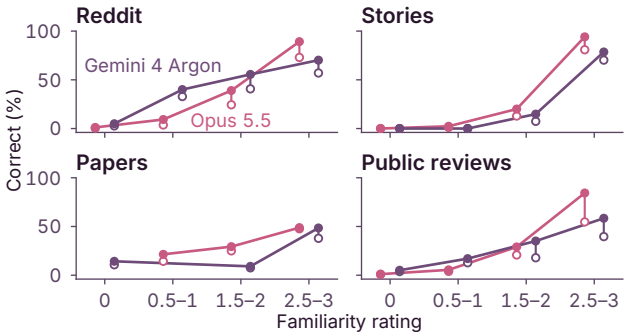

Figure 8. Models are more likely to name authors they are familiar with. Familiarity is the model's self-reported rating from 0–3. Solid dots on lines are original texts; hollow dots show average accuracy of two paraphrased versions.

Increasing reasoning effort helps most on accounts the models know only a little. We collect 442 comments from 216 Reddit accounts, sampled across all familiarity levels, and group accounts by each model's own rating. Fable 5.2, Opus 5.5, and Argon all show their largest accuracy gains from reasoning on accounts rated *a little familiar*, +35.8, +21.4, and +12.4 points, respectively (Table 6). For Fable 5.2, accuracy on these 19 accounts increases from 8.4% at low effort to 44.2% at max, with most of the improvement coming at the highest setting. Fable 5.2 names about 90% of *very familiar* accounts at every effort level, and no *unfamiliar* accounts until max effort, where it reaches 8%. Higher reasoning can recover weakly memorized data, similar to factual question answering, where thinking helps models recall rare facts more than popular ones (Calderon et al., 2026; Gekhman et al., 2026).

Models can imitate the users they recognize. The knowledge that lets a model recognize a writer also lets it write like one. We ask Opus 5.5 to generate two Reddit comments in the style of 565 accounts, given only the username and subreddit, and pass each comment to Opus 5.5. It names the target from at least one imitation for 74 of 166 accounts (45%) that Opus 5.5 both rates as familiar and identifies from real comments, vs. 9 of 219 (4%) familiar accounts it fails to identify, 5 of 31 (16%) accounts it identifies but does not rate familiar, and 0 of 149 unfamiliar, unidentified accounts. Although this test only measures whether an imitation is attributed to the target by a model rather than a human reader, it suggests that memorized author profiles can be used for impersonation.

5.2. Implications for anonymity

Once a model is familiar with a writer, it is hard to stay anonymous. Even when a text is brand new, a model can identify its writer from what it learned about them elsewhere; once that knowledge is in the weights, the writer cannot do anything about it. This leaves people with the largest public footprints, such as active social media users or established authors, permanently exposed, since paraphrasing the text does not completely remove the deanonymization risk. Moreover, the set of vulnerable writers expands as models improve. On Reddit posts, Fable 5.2 names the correct account more often than Opus 5.5 (14.3 percentage points more) and additional reasoning extends identification to writers a model knows only weakly.

Refusals are rare. Fable 5.2 and Opus 5.5 refuse most often, on under 2% of all documents (counting provider blocks and written refusals that decline to do the task at all) and mostly on public peer reviews.[31] The Anthropic provider blocks Fable 5.2 responses on 20 public peer reviews, 3 Reddit comments, and 1 paper, and Opus 5.5 responses on 28 public peer reviews and 2 papers. We additionally identify 151 anonymous predictions from Fable 5.2 and 130 from Opus 5.5, mostly citing insufficient evidence; only 50 of Fable 5.2's 162 written refusals and anonymous predictions mention privacy or concerns about incorrect attributions. Summing provider blocks, written refusals and anonymous predictions, Fable 5.2 names no author on 6.9% of documents and Opus 5.5 on 6.0%.

Paraphrasing lowers accuracy, though more familiar authors remain easier to identify. We test paraphrasing as a defense against author identification, using Argon at medium effort to revise texts from §4. The prompt asks it to make the author's style less identifiable while preserving content and meaning, a form of authorship obfuscation (Brennan et al., 2012; Mahmood et al., 2019; Fisher et al., 2024). We rewrite each text twice, run author attribution on both rewrites with Opus 5.5 and Argon, and average the two accuracies. Accuracy drops in nearly all domains. On Reddit, Opus 5.5 drops from 16.8% to 10.4% and Argon from 10.0% to 6.9%; on public reviews, the two models drop from 20.8–23.5% to 14.8–15.2%. Private reviews show larger drops, from 40.7–41.2% to 14.9–15.6%. Stories and papers are the least affected by paraphrasing. Argon drops from 18.6% to 15.8% on stories and Opus 5.5 from 34.7% to 31.4% on papers (Table 40). Familiar authors remain easier to identify after paraphrasing. Of the 34 stories whose authors Opus 5.5 rates very familiar (2.5–3 on the 0–3 scale), it names 32 authors before rewriting and 27.5 after, on average over the two rewrites (Figure 8).

[31] Refusals are far rarer through the API than through Claude Code (§D.4.2). Claude Fable 5.2 gives no answer on 323 of 1,258 Reddit comments through Claude Code but on only 14 (1.1%, 3 blocked by the provider, 11 written refusals) through OpenRouter batches pinned to Anthropic, which our results use. Models also sometimes make anonymous predictions, which give an explanation and conclude that the author is "anonymous".

Table 6. Reasoning effort helps most when the model is only a little familiar with the account, and cannot make up for an account it does not know. Reddit author attribution accuracy (%) by reasoning effort and by how familiar the model says it is with the account. Δ: paired gain from the lowest to the highest effort. We sort each account by the mean of its own familiarity ratings of the account (0–3 scale); the range of the mean familiarity ratings is in parentheses.

	Fable 5.2					Opus 5.5					Gemini 4 Argon				
Familiarity	Low	Med.	High	Max	Δ	Low	Med.	High	Max	Δ	None	Low	Med.	High	Δ
Any familiarity (0.75–3)	67.5	68.9	69.8	80.2	+11.2	51.9	55.5	56.0	68.7	+15.5	54.0	53.0	58.6	63.2	+9.1
Unfamiliar (0–0.5)	0.0	0.0	0.0	7.7	+8.0	0.0	0.0	0.5	0.5	+0.5	4.8	5.4	6.6	8.4	+3.5
A little familiar (0.75–1.5)	8.0	6.7	18.0	38.2	**+35.8**	11.2	12.9	14.7	33.6	**+21.4**	27.6	27.0	35.2	40.0	**+12.4**
Moderately familiar (1.75–2.5)	69.7	71.9	71.1	84.9	+11.9	61.0	65.4	65.3	76.0	+14.5	57.4	55.6	59.4	64.2	+6.8
Very familiar (2.75–3)	88.1	89.5	90.1	89.0	−1.4	76.0	77.5	78.2	87.8	+11.8	70.5	70.7	77.6	81.5	+11.0
95% CI for pooled gain	(6.2–16.4)					(10.3–20.8)					(3.4–15.1)				

6. Related work

Detecting AI-generated text and model attribution. Traditional AI-text detectors rely on statistical approaches (Gehrmann et al., 2019; Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024) or fine-tuned classifiers (Hu et al., 2023; Glickenhaus et al., 2026; Adam et al., 2026). Earlier prompt-based detection concluded prompting was unreliable at flagging AI-generated text (Bhattacharjee & Liu, 2024; Russell et al., 2025). Bao et al. (2025) and Rajendhran et al. (2026) detect AI-generated prose and AI-generated ideas separately. Uchendu et al. (2020) introduce the task of classifying which model generated a text, given the options of several generators. Uchendu et al. (2021) and Sun et al. (2025) find that classifiers trained on a fixed set of generators can tell them apart. In zero-shot settings, LLMs can pick up on AI-generated stylistic cues (Barkhordar & Thapa, 2026; Davidson et al., 2024) but identification fails once those cues are controlled. Panickssery et al. (2024) find that models identify and favor their own generations; Choi et al. (2025) show that this capability extends to conversational partners and that LLMs reliably recognize peers from their own family and detect prominent out-of-family models (such as GPT and Claude) across stylistic, reasoning, and alignment signatures.

Identifying human authors. Early approaches used stylometry or embeddings to identify authorship (Mosteller & Wallace, 1964; Narayanan et al., 2012; Rivera-Soto et al., 2021; Stamatatos, 2009; Koppel et al., 2011; Wegmann et al., 2022). For peer review, Hill & Provost (2003) identify authors 40–45% of the time from their citations alone, and Bauersfeld et al. (2023); Caragea et al. (2019) identify the authors of 73% of papers from their text and references. Many prior studies point to the threat of deanonymization, but use limited candidate pools, retrieval setups, or agentic tooling to extract information about an author (Lermen et al., 2026; Zhang & Zhang, 2026; Gloire & Mitra, 2026; Staab et al., 2024; Tan et al., 2025; Huang et al., 2024; Wang & Luo, 2026). Adewumi et al. (2025) and Hicke & Mimno (2025) find LLMs can identify famous authors from text alone but struggle with less popular writers, and in open-world quote attribution, accuracy likewise rises with the speaker's fame (Berman et al., 2026). LLMs' familiarity with a text relates to how often that specific text appears in pretraining data (Chang et al., 2023; Kandpal et al., 2023).

7. Conclusion

We demonstrate that closed-source frontier models can match trained AI-text detectors, name the specific model behind AI-generated texts, and perform human authorship attribution given only a single text and a zero-shot prompt. They can name the reviewer of about half of 15K anonymous ICML reviews that were never made public and therefore could not have been memorized during training. Attribution is easier when the model is familiar with the human author; we can predict which authors a model can unmask by asking it how familiar it is with a given name. Models are also rapidly improving at this task. Claude Fable 5.2 (released on November 9, 2026) is the strongest model we tested at naming both labs and people. Defenses are limited, since models rarely refuse, writing cannot be deleted from existing weights, and AI rewriting only partly helps. Anyone who has published enough writing online can no longer assume that an unsigned text keeps them anonymous.

Acknowledgements

We thank the University of Maryland Computational Linguistics and Information Processing (CLIP) Lab and the SFU AI lab for their feedback and support. We thank Google for HG's Gemini Academic Program award and Pangram Labs for a credit award, both of which enabled this research. MK is supported by NSERC Discovery Grant RGPIN-2026-04984 and Rajan Faculty Recruitment Award.

AI use statement

We used AI tools to assist in data collection, code implementation, prompt engineering, editing our draft, and to generate tables and figures from our results. Claude Opus 5.5 was used to judge agreement between model explanations, and was validated by human annotators. AI tools were also queried to provide feedback on our draft and identify gaps in our experiments; Claude Code suggested the inclusion of author-level accuracies and confidence levels in §4. We have manually reviewed all AI-assisted content and take full responsibility for the final content of this work.

Ethical considerations

The people most affected by this work are those who rely on anonymity. Reviewers in double-blind review could face retaliation from authors whose papers they rejected, and pseudonymous forum users could be harassed or doxxed once their accounts are linked to them. This is especially dangerous for those whose safety depends on staying anonymous, such as whistleblowers, activists, and survivors of abuse. Models are also able to imitate users they recognize (§5), so the same knowledge could be used to impersonate them. We recognize that publishing this work may lead some readers to try these attacks themselves; however, we believe it is more important to raise awareness of these capabilities to spur future research on defenses, especially given how simple and accessible they are.

Existing capability. All of our results come from simply prompting publicly available models. Each model names the author of a single text in a zero-shot setting, without tools, web access, a candidate list, or writing samples, nor do we train any models. This requires no programming or data collection, as anyone is able to paste a text and our prompt into a chat interface. Models also rarely refuse this task. Even Fable 5.2 and Opus 5.5, the models that refused most often, did so on fewer than 2% of documents through the API. The risk is therefore already present in widely accessible models, and exists regardless of whether this paper was published.

The risk will grow without intervention. We expect this risk to increase as models improve. Claude Fable 5.2 identifies the author of nearly twice as many Reddit comments as Opus 5.5, which was only released two months earlier. Furthermore, a higher reasoning effort also allows models to identify authors they are barely familiar with. Nothing a writer does now can remove what a model has already learned about them from its weights. Delaying this work would not make these writers any safer, and would only keep them from knowing they are at risk.

Benefits of publication. We believe the benefits of publishing this work outweigh the risks, since most people who write anonymously are likely unaware that a model can identify them from a single text. Model developers are also able to reduce this risk, for example by having models refuse these requests more often (we informed Anthropic and OpenAI about our findings prior to publication). This is already possible, with Fable 5.2 refusing 323 of 1,258 Reddit comments through Claude Code compared to only 14 through the API. Our results also show that paraphrasing a text is only a partial defense, and better defenses against both identification and impersonation still need to be developed.

Data protection. The ICML 2020 reviews were analyzed under IRB approval (#4656-1) by the conference's program co-chair, a coauthor of this paper. Only he accessed the reviews, and we report only aggregate results, without publishing any part of the reviews or revealing author identities. Model providers received only the text of each review, with no identifying metadata. Reviewer names, emails and IDs, and the key linking reviews to reviewers were never sent to any model provider or other third party. All scoring was done locally. DeepSeek V4.1 Flash was restricted to OpenRouter hosts with zero data retention. All other models were accessed through paid APIs (via OpenRouter or Google's Gemini API) whose terms exclude API inputs and outputs from model training. The private reviews were contributed voluntarily by their own authors. All other data is publicly available, and we report results on it in aggregate. The only people we mention by name are already well-known authors or researchers and their associated texts are publicly signed (not anonymous); we name them only as illustrative examples. We do not release any author predictions or private data.

References

Adam, G. A., Cui, A., Thomas, E., Napier, E., Shmatko, N., Schnell, J., Tian, J. J., Dronavalli, A., Tian, E., and Lee, D. GPTZero: Robust detection of LLM-generated texts, 2026. URL https://arxiv.org/abs/2602.13042.

Adewumi, T., Habib, N., Alkhaled, L., and Barney, E. On the limitations of large language models (LLMs): False attribution. In Angelova, G., Kunilovskaya, M., Escribe, M., and Mitkov, R. (eds.), *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pp. 11–21, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2025.ranlp-1.2/.

Anthropic. Strengthening our defenses against distillation

attacks. https://www.anthropic.com/research/distillation-defenses, October 2026a. Accessed: 2026-11-15.

Anthropic. Detecting and preventing distillation attacks. https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks, February 2026b.

Anthropic. Claude Fable 5.2 and Mythos 5.2 system card. https://www.anthropic.com/document/claude-fable-5-2-mythos-5-2-system-card, November 2026c. Accessed: 2026-11-16.

Anthropic. Countering misuse of AI: September 2026. https://www.anthropic.com/threat-intelligence-report-september-2026, September 2026d. Threat intelligence report. Accessed: 2026-11-04.

Anthropic. System card: Claude Opus 5.5. https://www.anthropic.com/claude-opus-5-5-system-card, September 2026e. Accessed: 2026-11-04.

Bao, G., Zhao, Y., Teng, Z., Yang, L., and Zhang, Y. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bpcgcr8E8Z.

Bao, G., Rong, L., Zhao, Y., Zhou, Q., and Zhang, Y. Decoupling content and expression: Two-dimensional detection of AI-generated text. *ArXiv*, abs/2503.00258, 2025. URL https://api.semanticscholar.org/CorpusID:276742379.

Barkhordar, E. and Thapa, S. Style, not self: Surface cues explain zero-shot code attribution by large language models, 2026. URL https://arxiv.org/abs/2609.30048.

Bauersfeld, L., Romero, A., Muglikar, M., and Scaramuzza, D. Cracking double-blind review: Authorship attribution with deep learning. *PLOS ONE*, 18(6): e0287611, 2023. doi: 10.1371/journal.pone.0287611.

Berman, E. H., Chang, B., Neill, D. B., and Black, E. Attribution bias in large language models. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=KQ0U28wDdE.

Bhattacharjee, A. and Liu, H. Fighting fire with fire: Can ChatGPT detect AI-generated text? *ACM SIGKDD Explorations Newsletter*, 25(2):14–21, 2024. doi: 10.1145/3655103.3655106.

Brassington, T., Paudayal, H., Brown, M., Phan, R., Vu, T., and Karpinska, M. Single-turn misalignment is not a solved problem: Measuring the effect of user pressure with matched controls, 2026. Preprint.

Brennan, M., Afroz, S., and Greenstadt, R. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security*, 15(3):1–22, 2012. doi: 10.1145/2382448.2382450.

Bsharat, S. M., Liu, J., Zhao, X., Yao, T., Shang, X., Tang, Y., Cui, J., Elhagry, A., Khatib, S. K. A., Li, H., Khan, S. H., and Shen, Z. Operation-guided progressive human-to-AI text transformation benchmark for multi-granularity AI-text detection. *ArXiv*, abs/2606.06481, 2026. URL https://api.semanticscholar.org/CorpusID:288976261.

Burrows, J. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002. doi: 10.1093/llc/17.3.267.

Calderon, N., Ben-David, E., Gekhman, Z., Ofek, E., and Yona, G. Empty shelves or lost keys? recall is the bottleneck for parametric factuality, 2026. URL https://arxiv.org/abs/2602.14080.

Caragea, C., Uban, A., and Dinu, L. P. The myth of double-blind review revisited: ACL vs. EMNLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2317–2327, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1236. URL https://aclanthology.org/D19-1236/.

Chakrabarty, T., Laban, P., and Wu, C.-S. Can AI writing be salvaged? Mitigating idiosyncrasies and improving human-AI alignment in the writing process through edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713559. URL https://doi.org/10.1145/3706598.3713559.

Chang, K. K., Cramer, M., Soni, S., and Bamman, D. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7312–7327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.453. URL https://aclanthology.org/2023.emnlp-main.453/.

Chen, Y., Chen, J., Mo, G., Chen, X., He, B., Han, X., and Sun, L. CoCoNUTS: Concentrating on content while neglecting uninformative textual styles for AI-generated peer review detection. In Liakata, M., Moreira, V. P., Zhang, J., and Jurgens, D. (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26921–26950, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1240. URL https://aclanthology.org/2026.acl-long.1240/.

Choi, Y., Li, C., Yang, Y., and Jin, Z. Agent-to-agent theory of mind: Testing interlocutor awareness among large language models. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28895–28928, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1471. URL https://aclanthology.org/2025.emnlp-main.1471/.

Clark, T. H., Martin, M. B., and Mahowald, K. Why is AI-generated language so hard to understand?, 2026. URL https://arxiv.org/abs/2610.29162.

Crossley, S., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, M., Picou, A., and Boser, U. The English language learner insight, proficiency and skills evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269, 2023. doi: 10.1075/ijlcr.22026.cro.

Davidson, T. R., Surkov, V., Veselovsky, V., Russo, G., West, R., and Gulcehre, C. Self-recognition in language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12032–12059, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.703. URL https://aclanthology.org/2024.findings-emnlp.703/.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/DeepSeek_V41_Tech_Report.pdf, September 2026. Technical report.

Duarte, A. V., Tufts, B., Oke, A., Fang, F., Oliveira, A. L., and Li, L. Sem-Detect: Semantic level detection of AI generated peer-reviews. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3S9eVockjL.

Emi, B. Pangram predicts 21% of ICLR reviews are AI-generated. Pangram Labs blog, November 2025. URL https://www.pangram.com/blog/pangram-predicts-21-of-iclr-reviews-are-ai-generated.

Emi, B. and Spero, M. Technical report on the Pangram AI-generated text classifier, 2024. URL https://arxiv.org/abs/2402.14873.

Fein, D. Has the bitter lesson come for AI detectors? Benchmarking AI detection adversarially on private writing samples. Vals AI blog, September 2026. URL https://www.vals.ai/blogs/ai-detection-benchmark.

Ferrando, J., Balcells Obeso, O., Rajamanoharan, S., and Nanda, N. Do I know this entity? Knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WCRQFlji2q.

Fisher, J., Lu, X., Jung, J., Jiang, L., Harchaoui, Z., and Choi, Y. JAMDEC: Unsupervised authorship obfuscation using constrained decoding over small language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1552–1581, Mexico City, Mexico, 2024. URL https://aclanthology.org/2024.naacl-long.87.

Flower, L. and Hayes, J. R. A cognitive process theory of writing. *College Composition & Communication*, 32 (4):365–387, December 1981. ISSN 1939-9006. doi: 10.58680/ccc198115885. URL http://dx.doi.org/10.58680/ccc198115885.

Gehring, L. and Paaßen, B. Limits of LLM text detectors in education. *arXiv preprint arXiv:2508.08096*, 2025. URL https://arxiv.org/abs/2508.08096.

Gehrmann, S., Strobelt, H., and Rush, A. GLTR: Statistical detection and visualization of generated text. In Costa-jussà, M. R. and Alfonseca, E. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019/.

Gekhman, Z., Aharoni, R., Ofek, E., Geva, M., Reichart, R., and Herzig, J. Thinking to recall: How reasoning unlocks parametric knowledge in LLMs, 2026. URL https://arxiv.org/abs/2603.09906.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gill, G., Kaur, J., Maltcheva, S., Duong, K. H. M., Anasco Santillan, R. A., Kapitanenko, M., Gupta, A., and Karpinska, M. LLMs redraw argumentative patterns across languages, 2026. Preprint.

Glickenhaus, B., Thai, K., Russell, J., Masrour, E., Han, Y., Spero, M., and Emi, B. Pangram 4 technical report, 2026. URL https://arxiv.org/abs/2607.27183.

Gloire, B. M. and Mitra, P. Large language models threaten double-blind review, 2026. URL https://arxiv.org/abs/2608.05157.

Google DeepMind. Gemini 4 Argon. https://deepmind.google/models/model-cards/gemini-4-argon/, October 2026a. Model card.

Google DeepMind. Gemini 4 Carbon. https://deepmind.google/models/model-cards/gemini-4-carbon/, November 2026b. Model card.

Gwet, K. L. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2014.

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., and Goldstein, T. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=axl3FAkpik.

Hayes, J. R. Modeling and remodeling writing. *Written Communication*, 29(3):369–388, 2012. doi: 10.1177/0741088312451260. URL https://doi.org/10.1177/0741088312451260.

Hicke, R. M. M. and Mimno, D. Looking for the inner music: Probing LLMs' understanding of literary style. *Computational Humanities Research*, 1, 2025. doi: 10.1017/chr.2025.10003.

Hill, S. and Provost, F. The myth of the double-blind review? Author identification using only citations. *ACM SIGKDD Explorations Newsletter*, 5(2):179–184, 2003. doi: 10.1145/980972.981001.

Ho, A., Denain, J.-S., Atanasov, D., Albanie, S., and Shah, R. A Rosetta Stone for AI benchmarks, 2025. URL https://arxiv.org/abs/2512.00193.

Hu, X., Chen, P.-Y., and Ho, T.-Y. RADAR: Robust AI-text detection via adversarial learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 15077–15095. Curran Associates, Inc., 2023. doi: 10.52202/075280-0662. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/30e15e5941ae0cdab7ef58cc8d59a4ca-Paper-Conference.pdf.

Huang, B., Chen, C., and Shu, K. Can large language models identify authorship? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 445–460, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.26. URL https://aclanthology.org/2024.findings-emnlp.26/.

Huang, S., Mamidanna, S., Jangam, S., Zhou, Y., and Gilpin, L. H. Can large language models explain themselves? a study of LLM-generated self-explanations, 2023. URL https://arxiv.org/abs/2310.11207.

Jabarian, B. and Imas, A. Artificial writing and automated detection. Working Paper 34223, National Bureau of Economic Research, September 2025. URL http://www.nber.org/papers/w34223.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., and Choi, Y. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=saDOrrnNTz.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. Large language models struggle to learn long-tail knowledge. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Karr, Jr, J. A., Khvatskii, G., Hua, T., and Chawla, N. V. Why AI detection fails for academic integrity, 2026. URL https://arxiv.org/abs/2608.11256.

Kim, Y., Chang, Y., Pham, C. M., and Iyyer, M. Argument collapse: LLMs flatten long-form public debate. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026. URL https://arxiv.org/abs/2606.01736. To appear.

Koppel, M., Schler, J., and Argamon, S. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011. doi: 10.1007/s10579-009-9111-2.

Le Goues, C., Brun, Y., Apel, S., Berger, E., Khurshid, S., and Smaragdakis, Y. Effectiveness of anonymization in double-blind review. *Communications of the ACM*, 61 (6):30–33, 2018. doi: 10.1145/3208157.

Lee, J. AI detectors rarely flag human writing, but sometimes miss AI text imitating real authors. *Epoch AI*, July 2026. URL https://epoch.ai/data-insights/ai-detectors-false-negatives. Accessed: 2026-09-21.

Lermen, S., Paleka, D., Swanson, J., Aerni, M., Carlini, N., and Tramèr, F. Large-scale online deanonymization with LLMs, 2026. URL https://arxiv.org/abs/2602.16800.

Li, C., Wan, C., and Paetzold, J. C. MELD: Multitask equilibrated learning detector for AI-generated text, 2026. URL https://arxiv.org/abs/2605.06903.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779, 2023. doi: 10.1016/j.patter.2023.100779.

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D., and Zou, J. Y. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29575–29620. PMLR, 2024. URL https://proceedings.mlr.press/v235/liang24b.html.

Liu, J., Min, S., Zettlemoyer, L., Choi, Y., and Hajishirzi, H. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2401.17377.

Luyckx, K. and Daelemans, W. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1):35–55, 04 2011. ISSN 0268-1145. doi: 10.1093/llc/fqq013. URL https://doi.org/10.1093/llc/fqq013.

Mahmood, A., Ahmad, F., Shafiq, Z., Srinivasan, P., and Zaffar, F. A girl has no name: Automated authorship obfuscation using Mutant-X. *Proceedings on Privacy Enhancing Technologies*, 2019(4):54–71, 2019. doi: 10.2478/popets-2019-0058.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Masrour, E. Pangram's model market share. Pangram Labs blog, August 2026. URL https://www.pangram.com/blog/pangram-s-model-market-share.

Matsubara, Y. and Singh, S. Citations beyond self citations: Identifying authors, affiliations, and nationalities in scientific papers. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pp. 9–20, Wuhan, China, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wosp-1.2/.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Mosteller, F. and Wallace, D. L. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA, 1964.

Musk v. Altman. Transcript of trial proceedings, volume 4, No. 4:24-cv-04722-YGR. U.S. District Court for the Northern District of California, ECF No. 586, 2026. Proceedings held 30 April 2026, before Judge Yvonne Gonzalez Rogers; testimony of Elon Musk.

Naddaf, M. Major AI conference flooded with peer reviews written fully by AI. *Nature*, 648(8093):256–257, 2025. doi: 10.1038/d41586-025-03506-6.

Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., and Song, D. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*, pp. 300–314. IEEE, 2012. doi: 10.1109/SP.2012.46.

National Security Agency, Cybersecurity and Infrastructure Security Agency, and Federal Bureau of Investigation. China-based artificial intelligence companies conducting industrial-scale distillation campaigns against U.S. AI companies. https://www.cisa.gov/news-events/cybersecurity-advisories/aa26-251a, September 2026. Joint Cybersecurity Advisory AA26-251A, 8 September 2026. Accessed: 2026-11-12.

OpenAI. Addendum to GPT-6 Astra system card: GPT-6.1 Sol. https://deploymentsafety.openai.com/gpt-6-1-sol, September 2026a. Accessed: 2026-11-16.

OpenAI. GPT-6 Astra system card. https://deploymentsafety.openai.com/gpt-6-astra, September 2026b. Accessed: 2026-09-23.

Panickssery, A., Bowman, S. R., and Feng, S. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4NJBV6Wp0h.

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., and Khuat, H. Q. Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education*, 21(1):53, 2024. doi: 10.1186/s41239-024-00487-w. URL https://doi.org/10.1186/s41239-024-00487-w.

Perrone, G. and Romano, S. P. ARB: A matched authorship-rewriting benchmark dataset for AI-text detector evaluation, 2026. URL https://arxiv.org/abs/2607.29539.

Pham, C. M., Hoyle, A. M., Sun, S., Resnik, P., and Iyyer, M. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2956–2984, 2024.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Rajendhran, R., Choi, M., Russell, J., Namuduri, R., Bölöni-Turgut, D., Karpinska, M., Wieting, J., and Iyyer, M. IdeaLens: Detecting AI ideas in long-form writing, 2026. URL https://arxiv.org/abs/2610.06778.

Razeghi, Y., Logan, IV, R. L., Gardner, M., and Singh, S. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.59. URL https://aclanthology.org/2022.findings-emnlp.59.

Rivera-Soto, R. A., Miano, O. E., Ordonez, J., Chen, B. Y., Khan, A., Bishop, M., and Andrews, N. Learning universal authorship representations. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 913–919, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.70. URL https://aclanthology.org/2021.emnlp-main.70/.

Russell, J., Karpinska, M., and Iyyer, M. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5342–5373, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.267. URL https://aclanthology.org/2025.acl-long.267/.

Russell, J., Glickenhaus, B., Thai, K., Wieting, J., Iyyer, M., Spero, M., and Emi, B. How Much Is an AI Token Worth? Scaling Laws for Wild AI-Generated Web Text, 2026a. URL https://arxiv.org/abs/2609.40295.

Russell, J., Rajendhran, R., Pham, C. M., Iyyer, M., and Wieting, J. StoryScope: Investigating idiosyncrasies in AI fiction. In *Proceedings of the Conference on Language Modeling (COLM)*, 2026b. URL https://arxiv.org/abs/2604.03136.

Saha, R., Juneja, G., Chaudhuri, D., Sajeevan, N., Shah, N. B., and Pruthi, D. Policies permitting LLM use for polishing peer reviews are currently not enforceable. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=znFIR7WFGt.

Seetharaman, D. and Arámburo, F. OpenAI says China's DeepSeek trained its AI by distilling US models, memo shows. https://finance.yahoo.com/news/openai-accuses-deepseek-distilling-us-221629899.html, February 2026. Reuters, 12 February 2026. On OpenAI's memo to the U.S. House Select Committee on Strategic Competition between the U.S. and the Chinese Communist Party.

Shan, Z., Lee, Y., and Hao, S. AI writers have a consistent stylometric footprint, but AI editors do not, 2026. URL https://arxiv.org/abs/2608.27855.

Si, C., Yang, D., and Hashimoto, T. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=M23dTGWCZy.

Si, C., Hashimoto, T., and Yang, D. The ideation-execution gap: Execution outcomes of LLM-generated versus human research ideas. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Fllp8l6Puy.

Soldaini, L., Kinney, R., Bhagia, A., et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024. URL https://aclanthology.org/2024.acl-long.840.

Song, Y., Kim, Y., and Iyyer, M. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, 2024.

Staab, R., Vero, M., Balunovic, M., and Vechev, M. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kmn0BhQk7p.

Stamatatos, E. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009. doi: 10.1002/asi.21001.

Sun, M., Yin, Y., Xu, Z., Kolter, J. Z., and Liu, Z. Idiosyncrasies in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=FCZ3jVzmTZ.

Tan, X., Liu, S., Cong, X., Li, K., and Wang, X. Open-world authorship attribution. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17744–17758, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.913. URL https://aclanthology.org/2025.findings-acl.913/.

Uchendu, A., Le, T., Shu, K., and Lee, D. Authorship attribution for neural text generation. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8384–8395, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.673. URL https://aclanthology.org/2020.emnlp-main.673/.

Uchendu, A., Ma, Z., Le, T., Zhang, R., and Lee, D. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2001–2016, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.172. URL https://aclanthology.org/2021.findings-emnlp.172/.

Wang, R. and Luo, M. ECHO: Revealing review-to-submission linkage in anonymous peer review, 2026. URL https://arxiv.org/abs/2610.27044.

Wegmann, A., Schraagen, M., and Nguyen, D. Same author or just same topic? Towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 249–268, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.26. URL https://aclanthology.org/2022.repl4nlp-1.26/.

Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X., and Huang, X. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.551. URL https://aclanthology.org/2023.findings-acl.551/.

Yu, S., Luo, M., Madasu, A., Lal, V., and Howard, P. Is your paper being reviewed by an LLM? Benchmarking AI text detection in peer review. In Vondrick, C., Hariharan, B., Raffel, C., Pinto, L., Yang, D., and Faust, A. (eds.), *International Conference on Learning Representations*, volume 2026, pp. 61517–61553, 2026. URL ht

tps://proceedings.iclr.cc/paper_files/paper/2026/file/64592277a1cbb02ed6e26b6e9e5f458c-Paper-Conference.pdf.

Zhang, L. and Zhang, H. De-anonymization at scale via tournament-style attribution. In Liakata, M., Moreira, V. P., Zhang, J., and Jurgens, D. (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32270–32283, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1489. URL https://aclanthology.org/2026.acl-long.1489/.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J., and Stoica, I. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Zhong, Z., Raghunathan, A., Laidlaw, C., and Steinhardt, J. User awareness in frontier models. Transluce, aug 2026. URL https://transluce.org/user-awareness.

A. Detectors and reproducibility

In this section, we list every model we evaluate in §2, with its lab, weights, release date (the lab's announcement, or the day OpenRouter first listed it when the lab gives no day), reasoning effort, access route and cost per 1,000 documents (Table 7). Our prompts are available in our repository.

Table 7. Complete information of all models used in the timeline (section 2).

Detector	Lab	Weights	Released	Effort	Access	$/1k docs
Gemini 4 Neon	Google	closed	2026-11-18	high	Vertex Batch	97.22
Gemini 4 Hydrogen	Google	closed	2026-11-12	high	Vertex Batch	29.96
Gemini 4 Carbon	Google	closed	2026-11-10	medium	Gemini API	92.76
Gemini 4 Argon	Google	closed	2026-10-13	medium	Gemini Batch	11.21
Gemini 3.8 Flash	Google	closed	2026-09-02	medium	Gemini Batch	7.30
Gemini 3.6 Flash	Google	closed	2026-07-21	medium	Gemini Batch	8.11
Gemini 3.5 Flash-Lite	Google	closed	2026-07-21	minimal	Gemini Batch	1.80
Gemini 3.1 Flash-Lite	Google	closed	2026-05-07	minimal	Gemini Batch	0.95
Gemini 3.1 Pro	Google	closed	2026-02-19	high	Gemini Batch	31.70
Gemini 3 Flash	Google	closed	2025-12-17	high	Gemini Batch	8.90
Gemini 2.5 Pro	Google	closed	2025-06-17	dynamic	OpenRouter[b]	50.62
GPT-6.2 Sol	OpenAI	closed	2026-11-17	medium	OpenAI Batch	12.58
GPT-6.2 Astra	OpenAI	closed	2026-11-17	medium	OpenAI Batch	51.16
GPT-6.1 Terra	OpenAI	closed	2026-10-26	medium	OpenAI Batch	5.24
GPT-6.1 Sol	OpenAI	closed	2026-09-29	medium	OpenAI Batch	12.52
GPT-6 Luna	OpenAI	closed	2026-09-22	medium	OpenAI Batch	0.75
GPT-6 Astra	OpenAI	closed	2026-09-03	medium	OpenAI Batch	74.02
GPT-5.6 Terra	OpenAI	closed	2026-07-09	medium	OpenAI Batch	14.20
GPT-5.6 Luna	OpenAI	closed	2026-07-09	medium	OpenAI Batch	1.94
GPT-5.6 Sol	OpenAI	closed	2026-07-09	medium	OpenAI Batch	38.25
GPT-5.5	OpenAI	closed	2026-04-24	medium	OpenAI Batch	59.89
GPT-5.4	OpenAI	closed	2026-03-05	medium	OpenAI API	74.26
GPT-5.2	OpenAI	closed	2025-12-11	medium	OpenAI Batch	18.12
GPT-5.1	OpenAI	closed	2025-11-13	medium	OpenAI Batch	23.18
GPT-5	OpenAI	closed	2025-08-07	medium	OpenAI Batch	26.94
o3	OpenAI	closed	2025-04-16	medium	OpenAI Batch	12.10
GPT-4o (2024-08-06)	OpenAI	closed	2024-08-06	none	OpenAI Batch	6.70
Claude Fable 5.2	Anthropic	closed	2026-11-09	high	Claude Code	195.31
Claude Haiku 5.5	Anthropic	closed	2026-10-07	medium	Claude Code	2.57
Claude Sonnet 5.5	Anthropic	closed	2026-09-28	high	Anthropic Batch	13.40
Claude Opus 5.5	Anthropic	closed	2026-09-22	medium	Claude Code	54.51
Claude Fable 5.1	Anthropic	closed	2026-09-01	high	Claude Code	SUB[c]
Claude Opus 5	Anthropic	closed	2026-07-24	high	Claude Code	SUB[c]
Claude Sonnet 5	Anthropic	closed	2026-06-30	high	Anthropic Batch	23.90
Claude Fable 5	Anthropic	closed	2026-06-09	high	Claude Code	222.19
Claude Opus 4.8	Anthropic	closed	2026-05-28	high	Claude Code	SUB[c]
Claude Opus 4.7	Anthropic	closed	2026-04-16	high	Anthropic Batch	32.40
Claude Sonnet 4.6	Anthropic	closed	2026-02-17	high	Anthropic Batch	33.60
Claude Opus 4.6	Anthropic	closed	2026-02-05	high	Anthropic Batch	42.20
Claude Opus 4.5	Anthropic	closed	2025-11-24	high	Anthropic Batch	37.10
Claude Haiku 4.5	Anthropic	closed	2025-10-15	budget 8,192	Anthropic Batch	13.71
Claude Sonnet 4.5	Anthropic	closed	2025-09-29	budget 8,192	Anthropic Batch	32.34
Kimi K3.1 Pro	Moonshot	open	2026-10-23	max	OpenRouter[b]	83.30
Kimi K2 (0711)	Moonshot	open	2025-07-11	none	OpenRouter[b]	8.75
DeepSeek V4.1 Flash 1017	DeepSeek	open	2026-10-17	high	OpenRouter[b]	2.39
DeepSeek V4.1 Flash (Sept build)	DeepSeek	open	2026-09-10	high	OpenRouter[b]	2.00
DeepSeek V4 Pro 0813	DeepSeek	open	2026-08-13	high	OpenRouter[b]	41.30
DeepSeek V4 Flash 0731	DeepSeek	open	2026-07-31	high	OpenRouter[b]	4.24
DeepSeek R1-0528	DeepSeek	open	2025-05-28	thinking on	OpenRouter[b]	6.22
DeepSeek V3	DeepSeek	open	2024-12-26	none	OpenRouter[b]	1.27
Beam 1.0 (501B-A23B)	Reflection	open	2026-10-16	medium	OpenRouter[b]	5.59
Nemotron 3.5	NVIDIA	open	2026-10-15	thinking on	OpenRouter[b]	4.19

Detector	Lab	Weights	Released	Effort	Access	$/1k docs
Meta Muse Spark 1.3	Meta	closed	2026-09-02	high	OpenRouter[b]	36.00
Meta Muse Spark 1.2 Contributor	Meta	closed	2026-08-05	high	OpenRouter[b]	1.39
GLM-5.3	Z.ai	open	2026-08-18	max	OpenRouter[b]	33.90
GLM-5.1	Z.ai	open	2026-04-07	thinking on	OpenRouter[b]	28.20
GLM-5	Z.ai	open	2026-02-12	thinking on	OpenRouter[b]	21.20
GLM-4.7	Z.ai	open	2025-12-22	thinking on	OpenRouter[b]	13.10
GLM-4.6	Z.ai	open	2025-09-30	thinking on	OpenRouter[b]	13.80
Qwen3.8 27B xhigh	Qwen	open	2026-08-14	xhigh	Self-hosted (vLLM)	10.92
Qwen3.8 2.4T-A95B	Qwen	open	2026-08-12	xhigh	OpenRouter[b]	89.00
Qwen3.6 27B	Qwen	open	2026-04-22	thinking on	Self-hosted (vLLM)	5.02
Qwen3.5 27B	Qwen	open	2026-02-24	thinking on	Self-hosted (vLLM)	9.15
Qwen3.5 397B-A17B	Qwen	open	2026-02-16	thinking on	OpenRouter[b]	17.70
Qwen3-235B-A22B Thinking 2507	Qwen	open	2025-07-25	thinking on	OpenRouter[b]	16.20
Qwen3 32B	Qwen	open	2025-04-29	thinking on	Self-hosted (vLLM)	1.50
QwQ-32B-Preview	Qwen	open	2024-11-28	thinking on	Self-hosted (vLLM)	–
Qwen2.5-72B Instruct	Qwen	open	2024-09-19	none	OpenRouter[b]	1.33
MiniMax M3	MiniMax	open	2026-06-01	thinking on	OpenRouter[b]	19.60
Hermes 3 405B	Nous Research	open	2024-08-15	none	OpenRouter[b]	3.20
Mistral Large 2 (2407)	Mistral	open	2024-07-24	none	Self-hosted (vLLM)	–
Llama 3.1 70B Instruct	Meta	open	2024-07-23	none	OpenRouter[b]	1.18
Pangram 4.1[a]	Pangram Labs	closed	–	–	Pangram bulk API	–
ProseLens	UMD	open	–	–	Self-hosted	–
IdeaLens	UMD	open	–	–	Self-hosted	–

[a] Every Pangram result uses Pangram 4.1, through its bulk API.

[b] We use the following OpenRouter providers: DeepInfra (Llama 3.1 70B Instruct, Hermes 3 405B, Qwen2.5-72B Instruct, DeepSeek V3, GLM-5.3); Novita (Kimi K2 (0711), GLM-4.6, GLM-4.7, DeepSeek V4 Pro 0813); Z.ai (GLM-5, GLM-5.1); Alibaba (Qwen3.5 397B-A17B); Meta (Meta Muse Spark 1.2 Contributor, Meta Muse Spark 1.3); SiliconFlow (DeepSeek R1-0528); Google AI Studio (Gemini 2.5 Pro); Venice (Qwen3-235B-A22B Thinking 2507); MiniMax (MiniMax M3); Moonshot AI (Kimi K3.1 Pro); Nebius (Beam 1.0 (501B-A23B)); Crusoe (Nemotron 3.5); Morph (DeepSeek V4.1 Flash 1017); Sail Research (DeepSeek V4.1 Flash (Sept build)); Novita, Alibaba and Together (Qwen3.8 2.4T-A95B); Parasail, Baidu, BaseTen, SiliconFlow, Mancer and DeepInfra (DeepSeek V4 Flash 0731).

[c] We evaluate some Claude models through a subscription plan (§2 data). In ablations between Claude Code and the API, we find no significant performance difference, though the API tends to have fewer refusals. Every Claude Code call runs in print mode with all tools disabled (so no web search or fetch), no MCP servers or user settings, and our prompt in place of Claude Code's own system prompt.

B. Human or AI: who wrote the prose, who had the ideas?

In this section, we discuss data, additional results, and analysis related to §2.

B.1. Human vs AI data

The RQ1 corpus pools 37,575 documents from 24 public datasets. We leave out 694 documents whose source labels do not place them in a single case, and 454 that at least one of the ten full-corpus models did not answer (refusals, errors, or replies that could not be parsed). The remaining 36,427 documents are 11,689 fully human, 16,149 fully AI, 8,539 with human ideas in AI prose, and 50 with AI ideas in human prose. Table 8 lists the 24 datasets, and Table 7 lists the models.

For false-positive experiments, we exclude 2,002 fully human documents on suspicion of machine translation confounding some non-English documents. This creates a 9,687-document English human set. We apply the same principles to the timeline experiment, removing 140 documents from our initial 1,000-document sample, creating a final set of 860 documents: 229 fully human, 287 fully AI, 294 with human ideas in AI prose, and 50 with AI ideas in human prose.

Table 8. The 24 datasets of the RQ1 corpus: documents in each case after the drops, and before them.

Dataset	Domain	Documents by case					
		Fully human	AI ideas in human prose	Fully AI	Human ideas in AI prose	Total	Before drops
Total		11,689	50	16,149	8,539	36,427	37,575
AcademicIntegrity (Karr et al., 2026)	paper abstracts	300	0	0	1,307	1,607	1,624
AI in Peer Review (Saha et al., 2026)	peer reviews	659	0	1,519	1,322	3,500	3,612
AI Peer Review Detection (Yu et al., 2026)	peer reviews	662	0	661	0	1,323	1,326
AI-Researcher (papers) (Si et al., 2025; 2026)	research papers	19	0	0	0	19	43
AI-Researcher (proposals and reviews) (Si et al., 2025; 2026)	research proposals, expert reviews	467	0	79	49	595	600
ARB (Perrone & Romano, 2026)	web text, news, short stories	657	0	1,323	662	2,642	2,649
C4 (Raffel et al., 2020)	web pages	652	0	0	0	652	663
CoCoNUTS (Chen et al., 2026)	peer reviews	654	0	662	660	1,976	2,134
DetectionAI (Jabarian & Imas, 2025)	news, fiction, blogs, résumés, reviews	663	0	1,321	0	1,984	1,989
ELLIPSE (Crossley et al., 2023)	essays by English learners	657	0	0	0	657	663
Epoch author imitation (Lee, 2026)	blogs, fiction, science writing	492	0	297	0	789	942
GEDE (Gehring & Paaßen, 2025)	student essays	659	0	1,985	659	3,303	3,313
GEN (Shan et al., 2026)	Wikipedia, WikiHow, arXiv, stories	660	0	604	571	1,835	2,139
HART (Bao et al., 2025)	news (several languages), essays, creative writing, arXiv	1,322	0	2,637	1,320	5,279	5,302
IdeaShift (Rajendhran et al., 2026)	web pages	40	0	80	38	158	280
IdeaShift-X (Rajendhran et al., 2026)	web pages in 24 languages	642	0	658	650	1,950	1,989
LAMP (Chakrabarty et al., 2025)	fiction, creative nonfiction	0	0	662	0	662	663
MELD (Li et al., 2026)	news, books, Wikipedia, Reddit, poetry, recipes, reviews, abstracts	659	0	1,227	0	1,886	1,986
OpAI-Bench (Bsharat et al., 2026)	reports, essays, news, abstracts	523	0	0	643	1,166	1,173
Perkins et al. (2024)	academic writing	0	0	89	0	89	89
Sem-Detect (Duarte et al., 2026)	peer reviews	657	0	661	658	1,976	1,985
StoryScope (Russell et al., 2026b)	short fiction	645	0	655	0	1,300	1,326
StoryScope (newer generators) (Russell et al., 2026b)	short fiction	0	0	980	0	980	985
TwiceTold (Rajendhran et al., 2026)	short fiction	0	50	49	0	99	100

B.2. Additional results

B.2.1. Full corpus

For the ten models run over the full corpus, Table 9 shows the accuracy splits and Table 10 shows false-positive rates and recall at matched false-positive rates. Additionally, we show general capability trends in Figure 9 and Pangram 4.1's calibration in Figure 10.

Table 9. RQ1 accuracy (%) on the full corpus. *Avg.* is an average of the four cases; *All* is accuracy over all documents. †Pangram 4.1 is not scored on human ideas in AI prose; its average is over the other three cases.

	Prose correct					Ideas correct					Both	
Prose by	Human	Human	AI	AI		Human	Human	AI	AI			
Ideas by	Human	AI	AI	Human	Avg.	Human	AI	AI	Human	Avg.	All	Avg.
Prompted LLMs												
GPT-6.1 Sol	99.0	96.0	98.9	83.3	94.3	99.8	72.0	92.5	96.6	90.2	91.5	84.8
Claude Opus 5.5	99.3	100.0	98.6	82.4	95.0	99.8	64.0	93.2	97.7	88.7	91.8	84.0
Gemini 4 Carbon	99.7	100.0	97.3	81.3	94.6	99.9	62.0	94.5	98.0	88.6	92.1	83.6
Claude Sonnet 5.5	99.4	100.0	95.4	69.2	91.0	99.8	36.0	91.5	93.0	80.1	86.6	72.0
GPT-6 Luna	97.1	52.0	95.6	61.1	76.5	98.3	72.0	93.0	70.6	83.5	79.6	61.4
DeepSeek V4.1 Flash	96.8	70.0	93.4	62.5	80.7	98.5	64.0	86.7	76.3	81.3	78.2	63.9
Gemini 3.8 Flash	99.2	90.0	90.2	38.5	79.5	99.8	10.0	86.5	88.0	71.1	76.3	53.0
Claude Haiku 5.5	96.9	72.0	89.5	52.3	77.7	97.2	34.0	76.0	81.3	72.1	71.7	55.1
Gemini 3.5 Flash-Lite	81.8	2.0	91.1	68.3	60.8	85.5	98.0	86.4	48.9	79.7	68.2	46.2
Nemotron 3.5	95.8	58.0	56.0	25.8	58.9	99.0	12.0	24.8	95.1	57.7	46.2	35.6
Trained detectors												
Pangram 4.1	99.9	96.0	96.7	–	97.6†	–	–	–	–	–	–	–
ProseLens	99.9	100.0	93.9	45.7	84.9	–	–	–	–	–	–	–
IdeaLens	–	–	–	–	–	99.7	68.0	85.0	95.5	87.1	–	–

Table 10. Frontier models have few false positives, where a fully human document is misclassified as AI. Recall is defined as the share of AI text flagged as AI (in %). For prompted models, we classify a document as AI if its reported AI share is at least 0.5. For trained detectors, we use their pre-existing cutoffs. @1% is the recall such that the FPR is at or below 1%. †Pangram 4.1's recall is on fully AI documents only, since it is not scored on human ideas in AI prose. ‡Gemini 4 Neon is run on the fully human documents only.

	Prose			Ideas		
System	FPR	Recall	@1%	FPR	Recall	@1%
Prompted LLMs						
Claude Opus 5.5	0.10	93.0	**96.1**	0.07	93.1	**99.4**
GPT-6.1 Sol	0.10	**93.5**	95.3	0.05	92.4	97.8
Gemini 4 Carbon	0.01	91.8	94.6	**0.00**	**94.4**	97.4
Claude Sonnet 5.5	0.09	86.4	92.5	0.07	91.4	97.0
DeepSeek V4.1 Flash	2.15	82.7	62.0	1.49	86.6	80.9
Gemini 4 Neon‡	0.03	–	–	**0.00**	–	–
GPT-6 Luna	1.81	83.7	76.2	1.55	93.0	88.7
Trained detectors						
Pangram 4.1	**0.00**	96.7†	–	–	–	–
ProseLens	0.03	77.2	–	–	–	–
IdeaLens	–	–	–	0.21	85.0	–

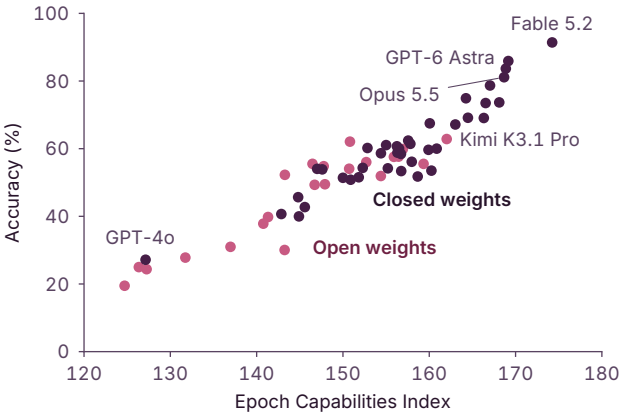

Figure 9. Detection accuracy (prose and ideas both correct) is highly correlated with general capabilities measured in the Epoch Capabilities Index ($r = 0.94$).

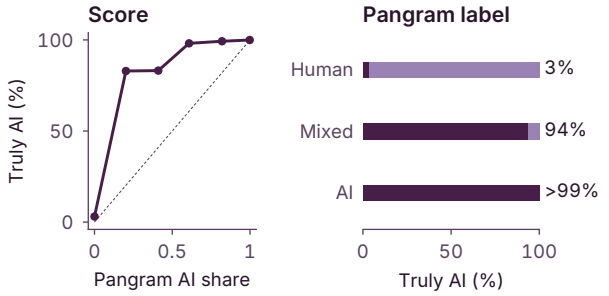

Figure 10. Pangram 4.1 is rarely wrong when it reports AI: share of documents with AI prose by its AI score (left) and by its label (right), leaving out human ideas in AI prose

.

B.2.2. Timeline of detectors

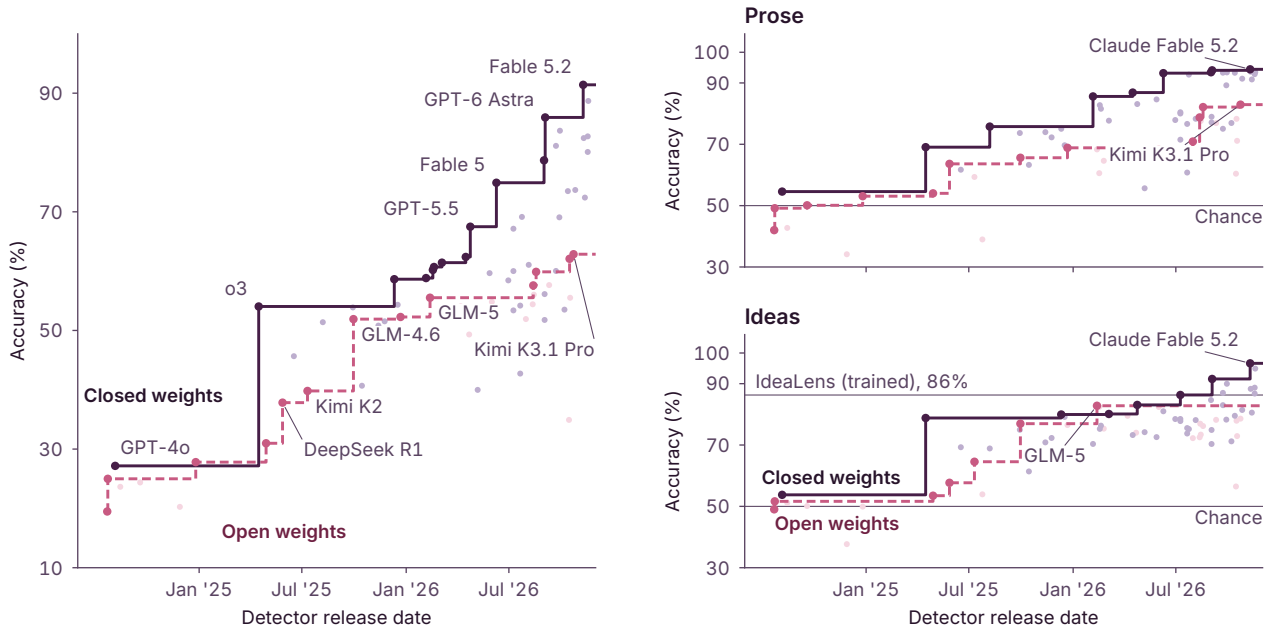

Figure 11. Full-history versions of the §2 model timeline, including all 72 models and all four provenance cases. Left: both prose and ideas must be correct. Right: prose and ideas scored separately, with IdeaLens's decision rule as a reference line; Pangram 4.1, which is not scored on human ideas in AI prose, has none here. Accuracy weights the four cases equally, on the 860 retained documents from the 1,000-document sampleLines mark the best accuracy attained by release date within closed-weight (solid purple) and open-weight (dashed pink) models.

Table 11. All 72 detectors on the 860-document sample, in release order (release dates and settings are in Table 7). *Both* requires the prose and ideas labels to be right; *avg.* averages it over the four cases, while Prose and Ideas are calculated over all documents. Fully human documents leave out those excluded from false-positive rates (§B.1). Errors and unparsable verdicts count as wrong. † partial run; ‡ two refusals excluded.

Detector	*n*	Prose	Ideas	Both	Both (avg.)	Both, human ideas in AI prose
Llama 3.1 70B Instruct	860	30.5	58.3	21.1	19.5	9.2
Mistral Large 2 (2407)†	794	34.5	60.5	26.3	25.0	2.1

Detector	*n*	Prose	Ideas	Both	Both (avg.)	Both, human ideas in AI prose
GPT-4o (2024-08-06)	860	55.6	53.4	31.7	27.2	14.3
Hermes 3 405B	860	32.0	60.2	25.8	23.6	7.1
Qwen2.5-72B Instruct	860	33.7	60.5	26.4	24.4	5.8
QwQ-32B-Preview	860	24.8	40.5	18.4	20.3	4.1
DeepSeek V3	860	52.4	59.1	32.6	27.8	29.2
o3	860	76.4	77.2	59.5	54.0	32.0
Qwen3 32B	849	63.6	59.4	39.6	31.0	58.4
DeepSeek R1-0528	860	64.3	65.0	44.3	37.8	53.1
Gemini 2.5 Pro	860	72.4	72.7	55.6	45.7	41.5
Kimi K2 (0711)	860	60.9	71.6	46.5	39.8	21.1
Qwen3-235B-A22B Thinking 2507†	858	40.0	47.9	30.8	30.1	14.3
GPT-5	860	71.2	76.5	55.6	51.4	23.8
Claude Sonnet 4.5	860	79.4	78.7	62.9	53.9	51.0
GLM-4.6	860	73.0	77.7	59.1	51.9	30.6
Claude Haiku 4.5	860	64.4	69.2	46.9	40.7	25.2
GPT-5.1	860	68.6	77.0	53.0	50.8	16.3
Claude Opus 4.5	860	74.9	78.8	60.2	51.5	24.8
GPT-5.2	860	81.7	80.3	65.8	58.6	33.3
Gemini 3 Flash	860	82.2	79.8	65.2	54.3	27.6
GLM-4.7	860	75.1	78.7	61.2	52.3	33.0
Claude Opus 4.6	860	84.4	80.6	68.8	58.8	36.0
GLM-5	860	75.7	83.6	64.5	55.5	35.7
Qwen3.5 397B-A17B†	860	68.1	80.5	55.4	49.5	15.0
Claude Sonnet 4.6	860	84.5	84.0	70.8	60.2	50.7
Gemini 3.1 Pro	860	81.7	85.9	71.4	60.7	48.0
Qwen3.5 27B	860	70.9	77.1	56.4	48.5	25.2
GPT-5.4	860	81.0	84.4	69.2	61.4	45.6
GLM-5.1†	859	75.2	81.5	62.8	54.1	35.8
Claude Opus 4.7	860	84.4	84.7	72.1	62.4	59.5
Qwen3.6 27B	860	67.9	78.5	54.0	49.3	20.8
GPT-5.5	860	88.4	86.0	76.2	67.5	54.1
Gemini 3.1 Flash-Lite	860	67.0	71.3	48.7	40.0	22.4
Claude Opus 4.8†	858	80.8	85.9	69.9	59.7	44.9
MiniMax M3	860	81.6	82.1	66.7	54.8	37.4
Claude Fable 5†	852	90.7	91.7	83.3	74.9	72.0
Claude Sonnet 5	860	81.9	85.5	69.5	58.4	46.6
GPT-5.6 Luna	860	80.7	81.9	65.1	53.4	31.6
GPT-5.6 Sol	860	86.4	90.5	77.9	67.2	57.5
GPT-5.6 Terra	860	84.8	85.5	71.6	60.0	42.5
Gemini 3.5 Flash-Lite	860	72.3	71.3	51.9	42.7	15.0
Gemini 3.6 Flash	860	79.5	82.0	64.2	54.2	23.1
Claude Opus 5†	857	90.1	90.5	82.3	69.1	71.3
DeepSeek V4 Flash 0731†	859	75.3	74.7	59.1	51.9	20.8
Meta Muse Spark 1.2 Contributor	860	80.3	77.1	63.0	61.1	24.8
Qwen3.8 2.4T-A95B†	854	75.1	82.5	61.7	54.4	28.7
DeepSeek V4 Pro 0813†	859	78.8	80.7	62.8	57.6	21.4
Qwen3.8 27B xhigh	860	78.0	82.0	64.0	56.0	35.0
GLM-5.3	860	80.2	85.1	67.4	59.9	30.9
Claude Fable 5.1†	853	91.1	92.7	84.5	78.7	73.4
Gemini 3.8 Flash	860	74.2	85.1	62.6	51.8	21.1
Meta Muse Spark 1.3	860	75.1	81.9	62.4	56.1	28.9
GPT-6 Astra	860	92.6	94.2	87.1	85.9	76.5
DeepSeek V4.1 Flash (Sept build)	860	80.7	80.2	64.1	57.6	22.4
Claude Opus 5.5	860	90.9	92.6	84.5	81.1	73.5
GPT-6 Luna	860	79.8	84.9	67.2	60.0	28.2
Claude Sonnet 5.5	860	85.9	90.2	78.3	69.1	59.9
GPT-6.1 Sol	860	92.2	94.3	86.9	83.7	76.2
Claude Haiku 5.5	860	76.9	79.8	60.9	53.5	30.3
Gemini 4 Argon	860	90.9	92.0	83.7	73.5	74.8
Nemotron 3.5	860	58.4	66.2	39.8	34.9	21.8
Beam 1.0 (501B-A23B)	860	77.0	76.7	60.4	62.1	24.1

Detector	*n*	Prose	Ideas	Both	Both (avg.)	Both, human ideas in AI prose
DeepSeek V4.1 Flash 1017	860	74.2	76.0	61.4	55.5	28.9
Kimi K3.1 Pro†	860	83.6	87.9	73.3	62.9	51.4
GPT-6.1 Terra	860	89.4	88.8	79.4	73.7	55.8
Claude Fable 5.2‡	858	92.4	96.0	88.9	91.4	81.2
Gemini 4 Carbon	860	91.4	94.9	86.9	82.4	76.5
Gemini 4 Hydrogen	860	88.5	91.6	81.0	72.4	63.9
GPT-6.2 Astra	860	91.0	93.7	85.6	82.7	74.8
GPT-6.2 Sol	860	90.9	92.8	84.3	80.1	72.1
Gemini 4 Neon	860	91.2	96.5	88.0	88.7	76.2

B.3. Analysis

Prompt development. Table 12 shows GPT-6.1 Sol's accuracy under four prompts of increasing structure, from a bare question (P1) to the final joint prompt (P4); the final prompt is in our repository, linked in the abstract.

Our prompt combines prose and idea detection into a single prompt, which we found improves performance on AI ideas in human prose. Sol's prose and idea accuracy improve by 14 and 13 percentage points, respectively. The detectors are also prompted to generate open-ended reasoning on how they weigh competing evidence and the decisive factors that contributed to their verdict.

Table 12. Accuracy of GPT-6.1 Sol on ∼5.5k documents from 13 datasets scored with prompts of increasing complexity; a document counts as AI when its stated AI share is at least 0.5.

	Overall (%)	TPR (%)	FPR (%)	Human ideas in AI prose (%)
Ideas				
P1: bare question	85.0	98.9	22.9	64.2
P2: definitions added	92.0	97.9	11.3	84.2
P3: structured reasoning	96.6	96.8	3.4	94.8
P4: joint prose and ideas	97.0	95.6	2.2	96.7
Prose				
P1: bare question	88.3	85.7	3.4	73.1
P2: definitions added	88.8	86.5	3.8	74.9
P3: structured reasoning	88.5	85.9	3.2	73.7
P4: joint prose and ideas	89.8	87.7	3.3	77.3

B.3.1. Contamination check

News published after every detector. We test five detectors on 1,000 US news articles and op-eds published on 10–11 November 2026: 188 that Pangram 4.1 labels AI and 812 it labels human (Table 13).

Table 13. Detectors against Pangram 4.1 on 1,000 news articles and op-eds published after every detector's release (%). Recall: Pangram-AI articles flagged as AI; FPR: Pangram-human articles flagged as AI.

Detector	Recall	FPR	Agreement
Claude Fable 5.2 (high)	95.2	1.8	97.6
GPT-6 Astra (medium)	100.0	5.0	95.9
Gemini 4 Carbon (medium)	78.2	0.5	95.5
Claude Opus 5.5 (medium)	77.1	3.4	92.9
DeepSeek V4.1 Flash (max)	75.0	20.6	78.6

Stories written after both detectors' training cutoffs. We also test Claude Opus 5.5 and GPT-6.1 Sol on 200 stories published or generated after both models' training cutoffs (Table 14).

Table 14. Detectors predict authors of stories published or generated after their training cutoffs about as well as the full corpus's fiction. Accuracy (%) on 200 stories (mean of five runs): 50 human stories published 1–16 October 2026 in 20 venues, 50 TwiceTold rewrites with AI ideas in human prose, and 25 stories by Claude Opus 5.5 and GPT-6.1 Sol in each of the two AI-prose cases, written from human stories.

	Prose			Ideas		
Detector	Post-cutoff	Full corpus	Gap	Post-cutoff	Full corpus	Gap
Claude Opus 5.5	98.7	99.2	−0.5	89.7	89.2	+0.6
GPT-6.1 Sol	96.5	98.1	−1.6	87.1	90.5	−3.4

Effect of dates in texts. To test whether detectors rely on a document's dates, we replace every year in the 190 dated documents of the 860-document sample with 2026, leaving years in URLs and arXiv identifiers unchanged, and run each detector on both versions with the same prompt (Table 15).

Table 15. Replacing every year in a document with 2026 does not change frontier models' predictions. Accuracy (%) on the 190 documents of the 860-document sample that mention a year (76 fully human, 40 fully AI and 74 with human ideas in AI prose), with the original text and with every year set to 2026. *Changed*: verdicts that switch to AI/to human when the years change. *Rerun*: the same count between two runs of the unchanged text.

	Ideas				Prose			
Detector	Original	2026	Changed	Rerun	Original	2026	Changed	Rerun
Claude Opus 5.5	97.9	97.9	2/0	2/1	90.5	89.5	1/3	1/3
GPT-6.1 Sol	98.9	97.9	1/1	2/0	90.0	89.5	0/1	0/2
Gemini 4 Carbon	98.4	98.4	1/0	0/1	88.9	90.0	1/0	1/1
DeepSeek V4.1 Flash	84.7	69.5	29/1	8/3	68.9	71.6	18/4	9/6

B.3.2. Explanation agreement

Grouping. An LLM judge (Claude Opus 5.5) aligns two explanations of the same document, E_A and E_B, separately for prose and ideas. Each explanation lists evidence items (a statement, quotes, a lean and its strength, and a justification), names a decisive factor and gives an AI share $s \in [0, 1]$ (Figure 13). In a first call, which hides leans, justifications, and verdicts, the judge groups items from the two explanations that make the same point, even when they quote different passages; every group holds at least one item from each explanation, and items that fit into no group stay unmatched. It also maps each point in the decisive factors to a matched group or an unmatched item. In a second call, the judge is provided each group's leans and justifications and decides whether both explanations rest on the same property of evidence in the group.

Metrics. Let $\mathcal{G}$ be the set of groups, U_A and U_B the set of unmatched items each from documents A and B, and D_A and D_B the set of decisive factors mapped to a group or an unmatched item.

- **Verdict** is 1 if both AI shares fall on the same side of 0.5.
- **Evidence F1** $= 2|\mathcal{G}|/(2|\mathcal{G}| + |U_A| + |U_B|)$, the harmonic mean of the share of each explanation's points that the other also makes (each group counts as one shared point).
- **Interpretation** is the share of groups whose items all have the same lean and the judge determines the justification of those items to be identical.
- **Decisive F1** $= 2|D_A \cap D_B|/(|D_A| + |D_B|)$, and 0 when exactly one set is empty.

Undefined values (both sets empty, or no groups for interpretation) are dropped for that metric only. We score one cross-model pair per document and average documents with equal weight; 95% intervals use a percentile bootstrap over documents (2,000 resamples). Opus 5.5, Sol, and Gemini 4 Argon explanations are compared in pairs on 2,854 documents (Table 16).

Judge validation. Two annotators audit the judge on 32 documents sampled evenly from the four cases. They annotate on one pair of explanations at a time, in the interface of Figure 12. They check whether each group makes the same point,

whether its items share a justification, whether unmatched items should belong in a group, and whether the decisive factors are mapped correctly. Annotator 1 completed 32 documents and Annotator 2 completed 24. Table 17 reports Gwet's AC1 (Gwet, 2014) and F1, giving each document equal weight; "partly" responses made by annotators count as no, and "unsure" responses are dropped. The annotators saw the groups and decisive sets formed by the judge, so humans validate the proposed judgments rather than suggesting new groupings from scratch.

Table 16. Every pair of detectors agrees on verdicts more than on evidence, interpretation, or decisive factors. A Claude Opus 5.5 judge compares an explanation from each detector of the pair (§B.3.2). Opus 5.5–GPT-6.1 Sol covers 2,854 documents (951 fully human, 950 fully AI, and 953 with human ideas in AI prose, from 24 datasets); the pairs involving Gemini 4 Argon cover 2,849 of them, as five have no usable Gemini 4 Argon explanation. Decisive F1 drops pairs in which neither explanation names a decisive point.

Detectors	Case	Docs	Verdict	Evidence F1	Interpretation	Decisive F1
Ideas						
Opus 5.5–GPT-6.1 Sol	All	2,854	0.98	0.85	0.68	0.56
	Fully human	951	0.99	0.83	0.67	0.57
	Fully AI	950	0.96	0.85	0.70	0.52
	Human ideas, AI prose	953	0.97	0.88	0.68	0.59
Opus 5.5–Argon	All	2,849	0.98	0.92	0.76	0.64
	Fully human	947	1.00	0.92	0.77	0.67
	Fully AI	950	0.96	0.89	0.75	0.58
	Human ideas, AI prose	952	0.98	0.93	0.77	0.65
Argon–GPT-6.1 Sol	All	2,849	0.97	0.88	0.63	0.54
	Fully human	947	0.99	0.87	0.62	0.57
	Fully AI	950	0.94	0.86	0.63	0.46
	Human ideas, AI prose	952	0.97	0.92	0.64	0.60
Prose						
Opus 5.5–GPT-6.1 Sol	All	2,854	0.97	0.76	0.84	0.63
	Fully human	951	0.99	0.76	0.85	0.66
	Fully AI	950	0.99	0.76	0.86	0.62
	Human ideas, AI prose	953	0.93	0.77	0.81	0.61
Opus 5.5–Argon	All	2,849	0.97	0.81	0.88	0.68
	Fully human	947	0.99	0.83	0.90	0.72
	Fully AI	950	0.98	0.79	0.87	0.65
	Human ideas, AI prose	952	0.94	0.80	0.86	0.66
Argon–GPT-6.1 Sol	All	2,849	0.96	0.74	0.84	0.64
	Fully human	947	0.99	0.74	0.87	0.65
	Fully AI	950	0.98	0.73	0.85	0.63
	Human ideas, AI prose	952	0.92	0.75	0.81	0.62

Table 17. Human–human and Claude Opus 5.5–human agreement on prose and ideas judgments. Annotator 1 completed 32 documents and Annotator 2 completed 24. Each score gives eligible documents equal weight. The decisive row uses candidate-membership AC1 and selected-set F1; other F1 rows compare positive labels.

		Annotator 1–Annotator 2		**Opus 5.5–Annotator 1**		**Opus 5.5–Annotator 2**	
Dimension	Judgment	AC1	F1	AC1	F1	AC1	F1
Ideas	Groups: same point	0.92	0.96	0.96	0.98	0.94	0.97
	Unmatched: no match	0.69	0.77	0.93	0.94	0.82	0.86
	Same reason	0.83	0.91	0.77	0.88	0.84	0.91
	Interpretation	0.87	0.93	0.82	0.90	0.87	0.93
	Decisive factors	0.87	0.87	0.90	0.89	0.96	0.95
Prose	Groups: same point	0.84	0.92	0.91	0.95	0.88	0.93
	Unmatched: no match	0.77	0.82	0.91	0.93	0.85	0.91
	Same reason	0.80	0.91	0.79	0.88	0.87	0.93
	Interpretation	0.83	0.92	0.80	0.89	0.88	0.94
	Decisive factors	0.89	0.94	0.92	0.95	0.95	0.97

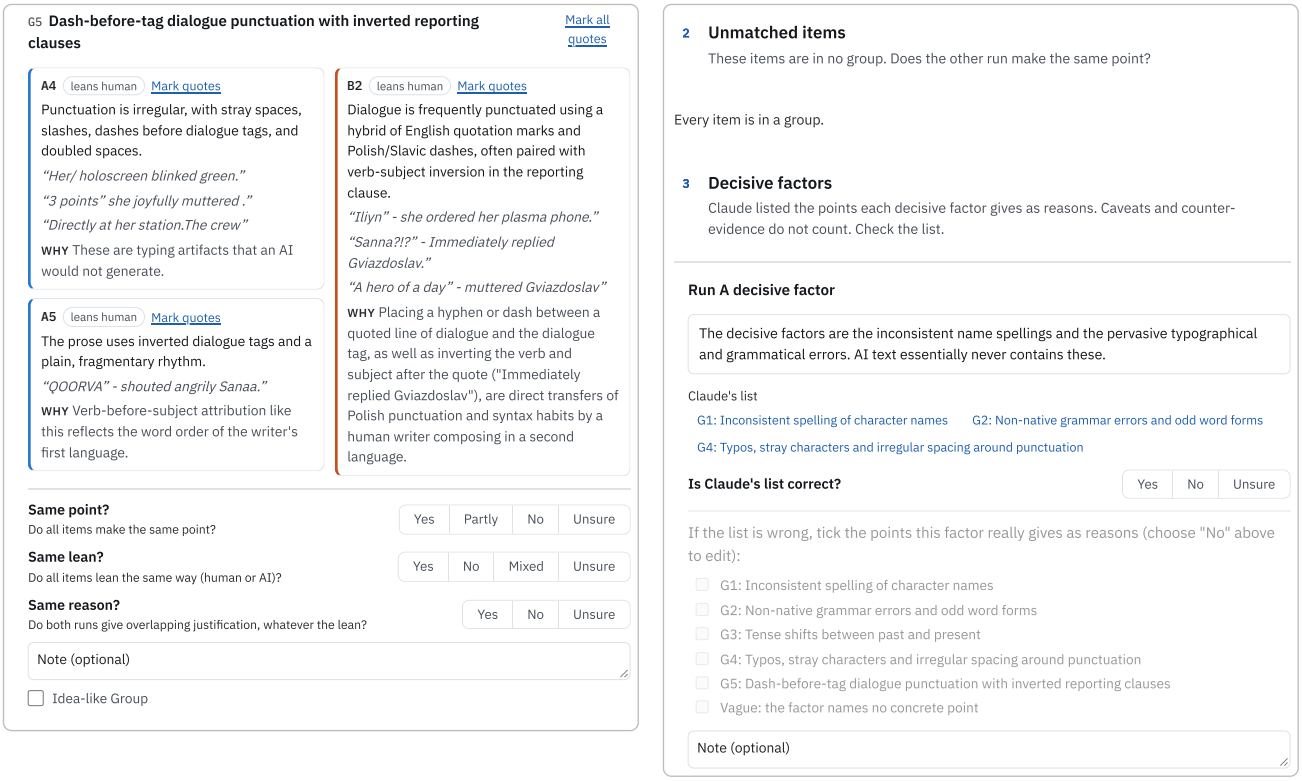

Figure 12. Human annotation interface used to manually verify explanation agreement. For each group of evidence items from two detector explanations, annotators assess whether the items make the same point, lean the same way (human or AI), and state the same reason. Annotators further inspect whether unmatched items do not belong to any formed group, and check the correctness of the list of points made in the decisive-factor, which was formed by the judge.

Evidence item 1	1. **IDEA:** A beautiful butterfly is isolated and made lonely by its beauty, inverting the usual transformation story. **LEANS:** human **STRENGTH:** medium **QUOTE:** "it was lonely for its beauty isolated it," **WHY:** The premise reverses the metamorphosis motif in a naive, personal way. It reads as an emotionally driven choice rather than an engineered twist.
Evidence item 2	2. **IDEA:** The butterfly wants to return to being a caterpillar, anonymous and free of vanity. **LEANS:** human **STRENGTH:** medium **QUOTE:** "and craved to return to its original self," **QUOTE:** "a catterpillar, anomonoyous and free" **WHY:** Regression as the wish is an unusual, intuitive angle. It sounds like the writer's own feeling about being noticed.
	[items 3–5 omitted; the reply then gives its reasoning, which the judge does not see]
Decisive factor	The idiosyncratic inverted-metamorphosis premise and the unexplained, quick resolution decide it. Both point to a human, likely young, writer working from feeling rather than design.
Verdict	ideas_ai_share: 0.05 ideas_prediction: human

Figure 13. Example explanation (ideas part). A Claude Opus 5.5 reply to the joint prompt for a human-written poem (MELD; fully human). Each evidence item gives a statement (**IDEA**), a lean (**LEANS**), a strength (**STRENGTH**), supporting quotes (**QUOTE**), and a justification (**WHY**). The reply ends with a decisive factor and a verdict. When forming groups, the judge sees only the statements, quotes, and decisive factor. When determining if items in each group rely on the same reason, it sees the statements, leans, and justifications. The prose part has the same form.

B.3.3. Cues cited in explanations

We analyze Opus 5.5 and Sol explanations of 2,432 documents from 24 datasets, with separate prose and idea judgments from each detector (9,728 judgments in total). Sol at high effort discovers features from their evidence and reasons, and merges them across datasets and detectors into catalogs of 132 idea and 119 prose features; each itemized evidence is then assigned one feature or none. Through these catalogs of features, we find that the cues frequently cited by LLM detectors align with findings from prior work. Table 18 presents two matches with our feature catalog per study that share both the

characteristic and the attribution direction; our features describe cues cited by frontier models rather than their reliability as gold provenance markers.

Table 18. Selected examples of detector-cited features that align in meaning and attribution direction with prior findings.

Target	Leans	Feature in detector explanations	Corresponding finding in prior work
StoryScope (Russell et al., 2026b)			
Ideas	AI	Explicit thematic and moral interpretation	AI stories state themes and morals more overtly.
Ideas	Human	Reframing revelations and timed disclosures	Human revelations more strongly change the interpretation of earlier scenes.
IdeaLens (Rajendhran et al., 2026)			
Ideas	Human	Precise factual and technical anchors	Human ideas use interconnected particulars that anchor an assertion to a specific instance.
Ideas	AI	Untailored criticism and recommendations	AI ideas offer generic remedies that remain applicable when the stated problem is replaced.
Argument Collapse (Kim et al., 2026)			
Ideas	Human	Tailored remedies and localized repairs	Humans propose concrete interventions.
Prose	AI	Hedging and diplomatic softening	AI responses hedge and balance stakeholder positions.

Two illustrative examples are selected per paper from the saved semantic comparisons; they do not estimate overall agreement. Leans gives the shared human/AI direction. The source studies differ in domain and measurement level.

C. Which model wrote the text?

In this section, we discuss data, additional results, and analysis related to section 3.

C.1. Model attribution data

Thirty-four models from 11 labs write the texts, released between 19 September 2024 (Qwen2.5 72B) and 19 October 2026 (MiniMax M3.1 Flash). Table 20 lists each with its release date, defined as in Table 7. Each runs at its default settings. Every writer writes from the same 25 prompts in each of three domains (Table 19).

1. Stories answer StoryScope's story prompts. For the five models StoryScope covers we use its released stories, and GPT-6 Astra's and Claude Fable 5.1's come from IdeaLens.
2. News articles start from 25 human-written articles from 17 outlets, published 9–20 October 2026: each writer gets an article's topic and length and writes its own article and title.
3. Op-eds start from 25 English briefs from MARGO, each giving the topic, the question at issue, the writing date, the target language and the length. The briefs in MARGO's seven other languages are in §C.2.2. In all there are 2,542 texts; four writers have 71–74 rather than 75 (Claude Opus 5.5, Gemini 3.8, Mistral Large 3 and MiniMax M2.5) due to some refusals.

Table 19. The three RQ2 domains. Word and token lengths describe the texts scored in Table 2. Token counts use `o200k_base` (`tiktoken` 0.14.0) on the text alone. SD is the sample standard deviation.

				Tokens per text		
Domain	Prompts	Texts	Words: median (IQR)	Mean	Median	SD
All		2,542		1,881	1,376	1,806
Stories	25 StoryScope story prompts	844	2,008 (1,178–2,696)	3,169	2,566	2,554
News	25 articles from 17 outlets, 9–20 Oct 2026	848	874 (631–1,211)	1,195	1,084	524
Op-eds	25 English MARGO briefs	850	1,008 (790–1,259)	1,287	1,194	735

C.2. Additional results

Table 20 lists the 34 writing models with their release dates and, for each, how often each detector names the right lab. Its first three rows match the lab columns of Table 2. No detector does better on its own lab's texts. Fable 5.2 names Anthropic

for 99.3% of Claude texts against 100% of OpenAI and Google texts; Opus 5.5 for 95.7% against 97.7% and 99.7%; and Argon names Google for 94.7% of Google texts against 98.7% and 100%. DeepSeek V4.1 Flash names DeepSeek for none of its own lab's texts. Only Sol is most accurate on its own lab (82.7% for OpenAI texts, against 68.2% and 32.1%), but it answers OpenAI or Anthropic for nearly every text (§C.3.1).

C.2.1. Length and confidence

Table 21 reports lab accuracy by input length and confidence on the common English subset of 299 Anthropic, 300 OpenAI and 374 Google texts (325 news articles, 325 op-eds and 323 stories). All three detectors have a completed, scorable answer for every text except one Argon response without a parsed model answer.

Opus' accuracy stays near 98% across all length bins, while Argon's accuracy increases from 94.7% at 500–999 words to about 99% above 1,000 words. Since texts within each length interval also differ in terms of genre and model author, it is possible that the difference in performance is not solely due to length; Fable makes only two errors, on texts of 521 and 525 words, so we cannot draw conclusions on length's effect.

Higher confidence accompanies higher lab accuracy for Opus and Argon, but confidence concerns the full model identification and is not directly a probability of naming the right lab. Small bins and different confidence scales limit comparisons between detectors.

C.2.2. Op-eds in seven other languages

The 34 writers also wrote op-eds from five briefs in each of seven other languages: Bulgarian, French, Hindi, Japanese, Polish, Russian and Spanish, 1,190 texts in all. Table 22 reports how often Opus 5.5 names the right lab, model family and exact model for them, next to the same writers' English op-eds.

C.2.3. Lab confusion

Accuracy on the big three labs also reflects how often detectors name them. Figure 14 shows, for each detector, which lab it names for each writing lab's texts, and Table 23 gives the precision of its answers per named lab and its macro-F1 over the 11 writing labs. When Fable 5.2 names Anthropic or OpenAI it is right only 51% and 39% of the time, because it also gives these labs most texts from xAI and MiniMax, about half of Zhipu's and over a third of Moonshot's; its answers naming any of the other eight labs are right 79% of the time. Opus 5.5 and Argon name Anthropic and OpenAI even more often. Sol and DeepSeek V4.1 Flash name Anthropic or OpenAI for nearly every text, and DeepSeek never names Google, so their macro-F1 is 11 and 4. We hypothesize some of this over-predicting may be effects of distillation.

Table 20. **Detectors name the lab behind nearly every big-three text, but few texts from the other eight labs.** Share of each writer's 75 texts for which each detector names the right lab (%). The Overall row and each lab's row pool the writers beneath them, and bold numbers mark the best detector in those rows. The other eight labs run from best to worst identified, and writers run newest first. Newer models from those labs are rarely traced to their own lab: Claude Opus 5.5 names Moonshot for 51% of Kimi K2.5 texts and for none of Kimi K3's. Claude Fable 5.2 is the best detector, or tied for best, in every pooled row, so the other eight labs now run from best to worst identified by Fable 5.2. Fable 5.2 names Moonshot for 93% of Kimi K2.5 texts and for none of Kimi K3's. Brackets give 95% Wilson intervals.

		Right lab (%)				
Writer	Released	Fable 5.2	Opus 5.5	Argon	Sol	DeepSeek
Overall (34 writers)		**59.9** (58–62)	49.7 (48–52)	44.8 (43–47)	22.6 (21–24)	11.1 (10–12)
Big three labs (13)		**99.8** (99–100)	97.8 (97–99)	97.5 (96–98)	58.8 (56–62)	28.7 (26–32)
Other eight labs (21)		**35.1** (33–37)	19.8 (18–22)	12.0 (11–14)	0.2 (0–1)	0.1 (0.0–0.5)
Anthropic (4)		**99.3** (98–100)	95.7 (93–97)	98.7 (97–99)	68.2 (63–73)	62.5 (57–68)
Claude Opus 5.5	22 Sep 2026	97.3 (91–99)†	83.8 (74–90)†	95.9 (89–99)†	48.6 (38–60)†	55.4 (44–66)†
Claude Fable 5.1	1 Sep 2026	100.0 (95–100)	98.7 (93–100)	100.0 (95–100)	68.0 (57–77)	58.7 (47–69)
Claude Opus 5	24 Jul 2026	100.0 (95–100)	100.0 (95–100)	98.7 (93–100)	80.0 (70–87)	78.7 (68–86)
Claude Sonnet 4.6	17 Feb 2026	100.0 (95–100)	100.0 (95–100)	100.0 (95–100)	76.0 (65–84)	57.3 (46–68)
OpenAI (4)		**100.0** (99–100)	97.7 (95–99)	**100.0** (99–100)	82.7 (78–87)	30.8 (26–36)
GPT-6.1 Sol	29 Sep 2026	100.0 (95–100)	94.7 (87–98)	100.0 (95–100)	70.7 (60–80)	22.7 (15–33)
GPT-6 Astra	3 Sep 2026	100.0 (95–100)	96.0 (89–99)	100.0 (95–100)	68.0 (57–77)	17.3 (10–27)
GPT-5.6 Sol	9 Jul 2026	100.0 (95–100)	100.0 (95–100)	100.0 (95–100)	96.0 (89–99)	38.7 (28–50)
GPT-5.4	5 Mar 2026	100.0 (95–100)	100.0 (95–100)	100.0 (95–100)	96.0 (89–99)	44.6 (34–56)†
Google (5)		**100.0** (99–100)	99.7 (99–100)	94.7 (92–97)	32.1 (28–37)	0.0 (0–1)
Gemini 4 Argon	13 Oct 2026	100.0 (95–100)	98.7 (93–100)	86.7 (77–93)	32.0 (23–43)	0.0 (0–5)
Gemini 3.8*	2 Sep 2026	100.0 (95–100)†	100.0 (95–100)†	93.2 (85–97)†	25.7 (17–37)†	0.0 (0–5)†
Gemma 4 31B*	2 Apr 2026	100.0 (95–100)	100.0 (95–100)	98.7 (93–100)	20.0 (13–30)	0.0 (0–5)
Gemini 3 Flash	17 Dec 2025	100.0 (95–100)	100.0 (95–100)	94.7 (87–98)	14.7 (8–24)	0.0 (0–5)
Gemma 3 27B	12 Mar 2025	100.0 (95–100)	100.0 (95–100)	100.0 (95–100)	68.0 (57–77)	0.0 (0–5)
Mistral: Mistral Large 3	2 Dec 2025	91.8 (83–96)†	64.4 (53–74)†	72.6 (61–82)†	0.0 (0–5)†	0.0 (0–5)†
Moonshot (4)		**57.0** (51–62)	20.7 (16–26)	7.0 (5–10)	0.0 (0–1)	0.0 (0–1)
Kimi K3	16 Jul 2026	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
Kimi K2.6	20 Apr 2026	77.3 (67–85)	29.3 (20–40)	8.0 (4–16)	0.0 (0–5)	0.0 (0–5)
Kimi K2.5	27 Jan 2026	93.3 (85–97)	50.7 (40–62)	17.3 (10–27)	0.0 (0–5)	0.0 (0–5)
Kimi K2 (0905)	4 Sep 2025	57.3 (46–68)	2.7 (1–9)	2.7 (1–9)	0.0 (0–5)	0.0 (0–5)
Alibaba (5)		**51.5** (46–56)	37.1 (32–42)	23.5 (19–28)	0.8 (0–2)	0.5 (0–2)
Qwen3.8 Max	2 Sep 2026	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
Qwen3.8 27B	14 Aug 2026	1.3 (0–7)	2.7 (1–9)	1.3 (0–7)	0.0 (0–5)	0.0 (0–5)
Qwen3.6 27B	22 Apr 2026	89.3 (80–94)	68.0 (57–77)	29.3 (20–40)	0.0 (0–5)	1.3 (0–7)
Qwen3 235B (2507)	21 Jul 2025	80.0 (70–87)	53.3 (42–64)	37.3 (27–49)	0.0 (0–5)	0.0 (0–5)
Qwen2.5 72B	19 Sep 2024	86.7 (77–93)	61.3 (50–72)	49.3 (38–60)	4.0 (1–11)	1.3 (0–7)
DeepSeek (2)		**40.7** (33–49)	34.0 (27–42)	15.3 (10–22)	0.0 (0–2)	0.0 (0–2)
DeepSeek V4.1 Flash	10 Sep 2026	14.7 (8–24)	9.3 (5–18)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
DeepSeek V3.2	1 Dec 2025	66.7 (55–76)	58.7 (47–69)	30.7 (21–42)	0.0 (0–5)	0.0 (0–5)
Cohere: Command A	13 Mar 2025	21.3 (14–32)	13.3 (7–23)	4.0 (1–11)	0.0 (0–5)	0.0 (0–5)
Zhipu (4)		**10.5** (8–15)	0.0 (0–1)	0.0 (0–1)	0.0 (0–1)	0.0 (0–1)
GLM-5.3	18 Aug 2026	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
GLM-5.1	7 Apr 2026	11.0 (6–20)†	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
GLM-4.7	22 Dec 2025	26.4 (18–38)†	0.0 (0–5)†	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
GLM-4.5	25 Jul 2025	5.4 (2–13)†	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
xAI: Grok 4.7	21 Sep 2026	5.3 (2–13)	1.3 (0–7)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
MiniMax (3)		**2.3** (1–5)	0.0 (0–2)	0.5 (0–3)	0.0 (0–2)	0.0 (0–2)
MiniMax M3.1 Flash	19 Oct 2026	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
MiniMax M3	1 Jun 2026	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)	0.0 (0–5)
MiniMax M2.5	12 Feb 2026	7.0 (3–15)†	0.0 (0–5)†	1.4 (0–8)†	0.0 (0–5)†	0.0 (0–5)†

Detectors: Claude Fable 5.2, Claude Opus 5.5, Gemini 4 Argon, GPT-6.1 Sol and DeepSeek V4.1 Flash. * Version assumed from the op-ed alias (Gemini 3.8) or from texts written on the Gemini API rather than OpenRouter (Gemma 4 31B). † Out of 71–74 texts rather than 75: the writer produced fewer texts or the detector gave no answer.

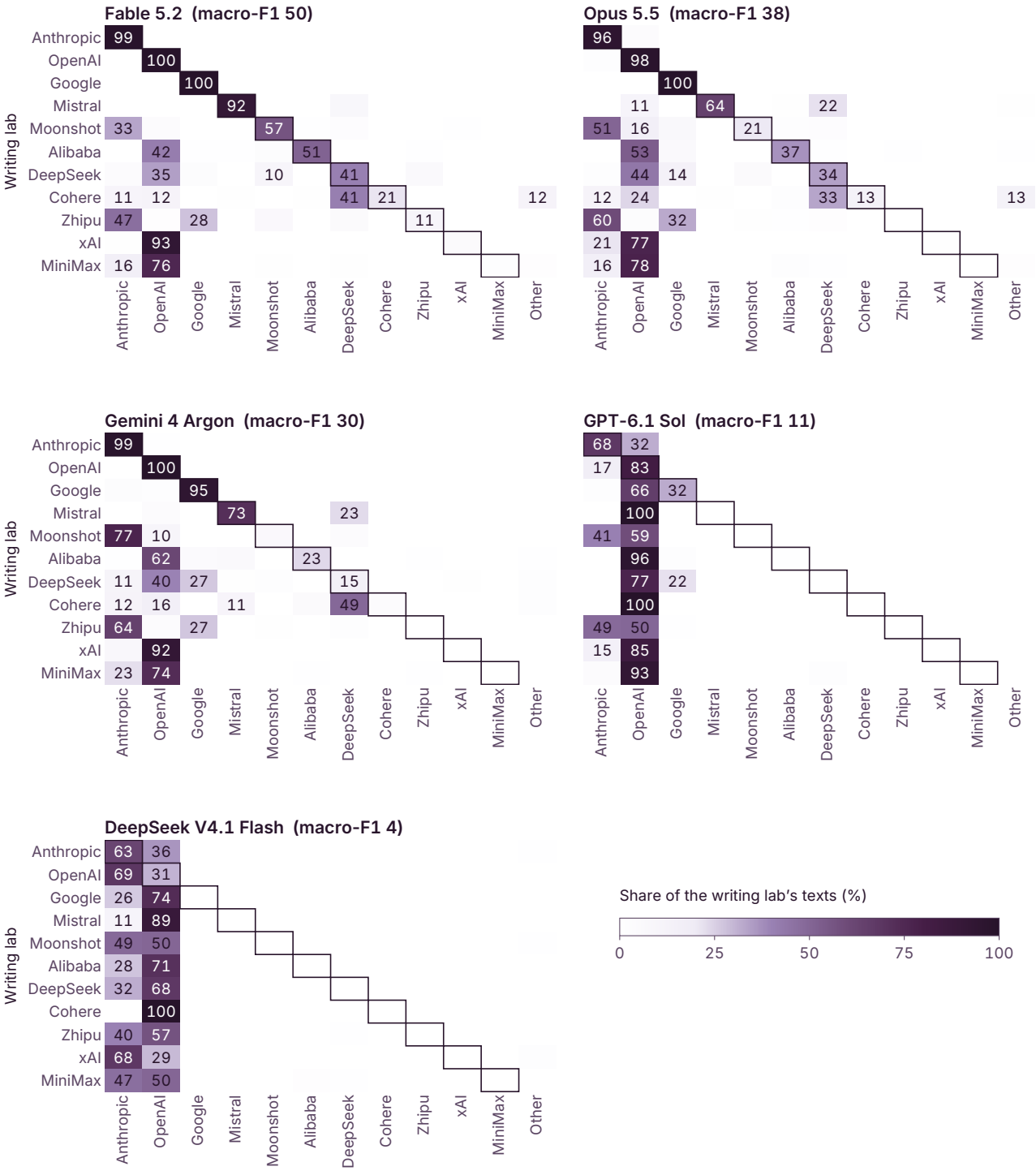

Figure 14. Which lab each detector names for each writing lab's texts (34-writer panel, English texts; % of the writing lab's texts). Outlined cells are the right lab; Other pools Meta, Microsoft, any other lab and replies that name no model. Fable 5.2, Opus 5.5 and Argon recognize the big three labs, but give most xAI and MiniMax texts to OpenAI and many Moonshot and Zhipu texts to Anthropic; Sol and DeepSeek V4.1 Flash name Anthropic or OpenAI for nearly every text.

Table 21. Lab attribution accuracy (%) on 973 English texts from Anthropic, OpenAI and Google (including Gemma). Parentheses give correct/scorable counts. One Argon response without a parsed model answer is excluded. Length is measured in words as per whitespace split.

(a) Lab accuracy (%) by text length

Words	Texts	Fable 5.2	Opus 5.5	Gemini 4 Argon
All	973	99.8 (971/973)	97.8 (952/973)	97.6 (949/972)
<500	29	100.0 (29/29)	100.0 (29/29)	96.6 (28/29)
500–999	301	99.3 (299/301)	97.7 (294/301)	94.7 (285/301)
1,000–1,499	355	100.0 (355/355)	97.7 (347/355)	99.2 (351/354)
1,500+	288	100.0 (288/288)	97.9 (282/288)	99.0 (285/288)

(b) Lab accuracy (%) by reported confidence

Fable 5.2		Opus 5.5		Gemini 4 Argon	
Confidence	Accuracy (%)	Confidence	Accuracy (%)	Confidence	Accuracy (%)
<0.30	100.0 (6/6)	<0.30	83.3 (5/6)	0.40–<0.50	100.0 (1/1)
0.30–<0.40	99.0 (190/192)	0.30–<0.40	95.2 (279/293)	0.50–<0.75	100.0 (10/10)
0.40–<0.50	100.0 (573/573)	0.40–<0.50	99.0 (499/504)	0.75–<0.85	89.5 (77/86)
0.50–<0.75	100.0 (202/202)	0.50–<0.75	99.4 (169/170)	0.85–<0.90	96.0 (216/225)
				0.90–<0.95	99.1 (527/532)
				0.95–1.00	100.0 (118/118)

Table 22. Claude Opus 5.5 attributes the non-English op-eds about as well as the English ones. Share of op-eds (%) for which Opus 5.5 names the right lab, model family or exact model with the holistic prompt, scored as in Table 2. The 34 writers each wrote from five briefs per language; Opus 5.5 refused 2 of the 1,190 non-English op-eds (both by MiniMax M3.1 Flash), and we leave them out. The English row is Opus 5.5 on the same writers' English op-eds.

		Lab (%)				
Language	Texts	All	Big three	Other eight	Family (%)	Model (%)
All non-English	1,188	47.4	98.2	15.8	27.9	2.0
Bulgarian	170	51.2	96.9	22.9	31.8	2.9
French	170	48.2	98.5	17.1	30.0	2.4
Hindi	170	42.9	100.0	7.6	21.8	0.0
Japanese	169	45.0	96.9	12.5	23.7	0.6
Polish	170	50.6	98.5	21.0	31.2	2.4
Russian	170	45.9	100.0	12.4	26.5	2.9
Spanish	169	47.9	96.9	17.3	30.8	3.0
English	850	49.1	99.1	18.1	32.0	5.2

Table 23. Precision of each detector's lab answers (%; in brackets, the number of texts it attributes to that lab) and macro-F1 over the 11 writing labs. Other eight pools the answers naming any lab outside Anthropic, OpenAI and Google. DeepSeek V4.1 Flash never names Google.

		Precision when it names (%)			
Detector	Macro-F1	Anthropic	OpenAI	Google	Other eight
Fable 5.2	49.9	50.6 (587)	38.7 (776)	80.4 (465)	79.0 (694)
Opus 5.5	37.8	41.5 (689)	33.1 (886)	69.5 (537)	75.8 (409)
Gemini 4 Argon	29.8	36.2 (816)	33.6 (894)	71.7 (494)	58.0 (326)
GPT-6.1 Sol	10.7	36.8 (555)	13.7 (1,808)	73.2 (164)	21.4 (14)
DeepSeek V4.1 Flash	3.5	17.4 (1,074)	6.4 (1,441)	–	14.3 (14)

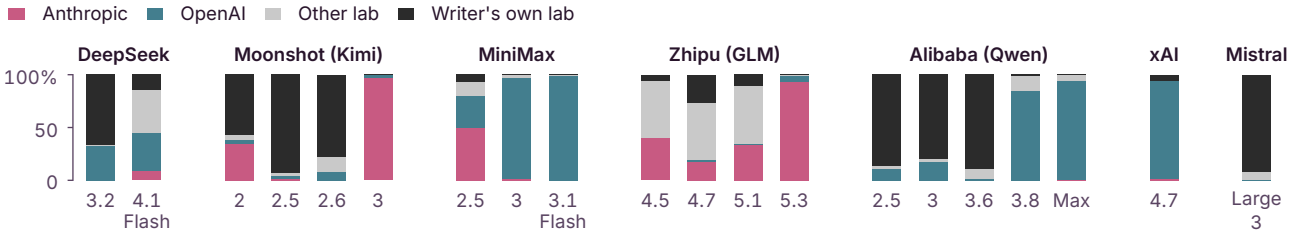

Figure 15. Claude Fable 5.2 recognizes most texts from earlier Kimi and Qwen releases, but attributes nearly every Kimi K3 and GLM-5.3 text to Claude and most Qwen3.8 27B and Grok 4.7 texts to GPT. Each column is one writer, oldest to newest within each lab.

C.3. Analysis

C.3.1. Model writing profiles

Building the profiles. Detectors will often describe the same writing traits in different words. So to count how often the detectors cite each trait, we adapt the methods of Brassington et al. (2026) and Pham et al. (2024) to assemble our model profiles:

1. We split each explanation into its key observations, each of which points to a lab, e.g. "numerical claims come with hedges about scope and method" (GPT). Opus 5.5, Argon, Sol and DeepSeek in total generated 288,430 observations.
2. Opus 5.5 groups each detector's observations about each lab into cues, then merges them into three lists for observations on the text overall (489 cues), its ideas (492 cues), and its prose (337 cues), so "numerical claims come with hedges about scope and method" and "repeatedly hedges against overclaiming" are counted as the same trait.
3. Opus 5.5 labels every observation with these lists, without seeing the detector or the lab. Since Fable 5.2 was released after feature discovery for the writing traits, we only run this labelling step on its observations.

We define a lab's traits as cues at least twice as frequent in observations pointing to that lab as in those pointing to other labs (e.g. dry wit for Claude), and say a writer shares a trait if its rate on the writer's texts is at least the midpoint between the lab's rate and the highest other big lab's.

We invite readers to explore each detector's profiles of GPT, Claude, Gemini and the open-weight models, along with the quoted evidence behind every cue, on our interactive model profiles page.

GPT-6.1 Sol and DeepSeek V4.1 Flash have flawed profiles of GPT and Claude. Sol and DeepSeek name OpenAI or Anthropic for nearly every text (93% and 99%). The traits they assign GPT and Claude are also given to other labs' texts. These traits are largely nonspecific and few actually set either lab apart. Other labs are assigned GPT traits more often than GPT models themselves.

C.3.2. Misattributions and reported distillation

This section expands on the distillation results in §3.1.

Claude Fable 5.2 recognizes early Kimi and Qwen releases far better than Opus 5.5 and Gemini 4 Argon, but all three misattribute the newest ones. If a detector is able to recognize a lab's earlier models, we would expect it to also recognize its newer ones, unless their style has changed. Fable 5.2 names the right lab for most texts from Kimi K2.5 and K2.6 and from the three Qwen releases preceding Qwen3.8, where Opus 5.5 and Argon often answer Claude or GPT instead (Table 20). But even Fable 5.2 attributes nearly every Kimi K3 and GLM-5.3 text to Claude, and most Qwen3.8 27B and Grok 4.7 texts to GPT (Figure 15). Because Fable 5.2 knows these labs' earlier models, its errors suggest that their newest releases no longer read like those models, but like Claude or GPT. Fable 5.2 recognizes few texts from any GLM or MiniMax release, so we cannot make the same comparison for these two labs.

Later reports name Kimi K3 and GLM-5.3, and date campaigns against Claude shortly before their release. The first reports named labs but not models, and did not say when the campaigns ran. Later reports do both. A joint advisory

by US agencies reports that Moonshot trained Kimi K3 on Claude Fable 5 data (National Security Agency et al., 2026), and Anthropic reports a Moonshot campaign against Claude between May and July 2026 (Anthropic, 2026d). K3 came out in July. Anthropic also reports a 17-day Zhipu campaign against Claude Opus 4.8 in June and July, and that Zhipu distilled the cyber capabilities of multiple frontier models during the development of GLM-5.3, which came out in August (Anthropic, 2026d;a). These are the two releases that all three detectors attribute almost entirely to Claude.

The detectors attribute the newest Qwen and MiniMax releases to GPT. Reports link Alibaba and MiniMax to both Claude and GPT. Anthropic reported more exchanges from Alibaba than from any other lab, used to distill Claude into Qwen3.5 through 3.7 (Anthropic, 2026d), and the advisory reports that both labs distilled Claude models and GPT-5 (National Security Agency et al., 2026). Fable 5.2 names Alibaba for 80% or more of the texts from each of the three earlier Qwen releases, but all three detectors attribute most Qwen3.8 27B and Qwen3.8 Max texts to GPT. Anthropic reported distillation of Claude by MiniMax in February (Anthropic, 2026b), the month MiniMax M2.5 came out, and the detectors attribute M2.5 mostly to Claude. The additional link to GPT-5 came in the September advisory, and the detectors attribute nearly every text by the newer M3 and M3.1 Flash to GPT.

C.3.3. Exploring alternative explanations for the misattributions.

The misattributions in Figure 15 are consistent with reported distillation, but there are other explanations that could produce them. Opus 5.5 and Argon often answer Claude or GPT for the earlier Kimi and Qwen releases that Fable 5.2 is able to properly attribute, so detectors may be falling back to known labs when they don't recognize the model who wrote the text, for instance because it came out after the detectors' training cutoff. Another possibility is the newest open models may resemble Claude or GPT only because Clark et al. (2026) found that recent models from several labs share an information-dense writing style. We investigate both using stylometric analysis by checking whether the detectors' answers match stylometry, and then controlling for information density features from Clark et al. (2026) to see if the newest releases are still closest to Claude or GPT.

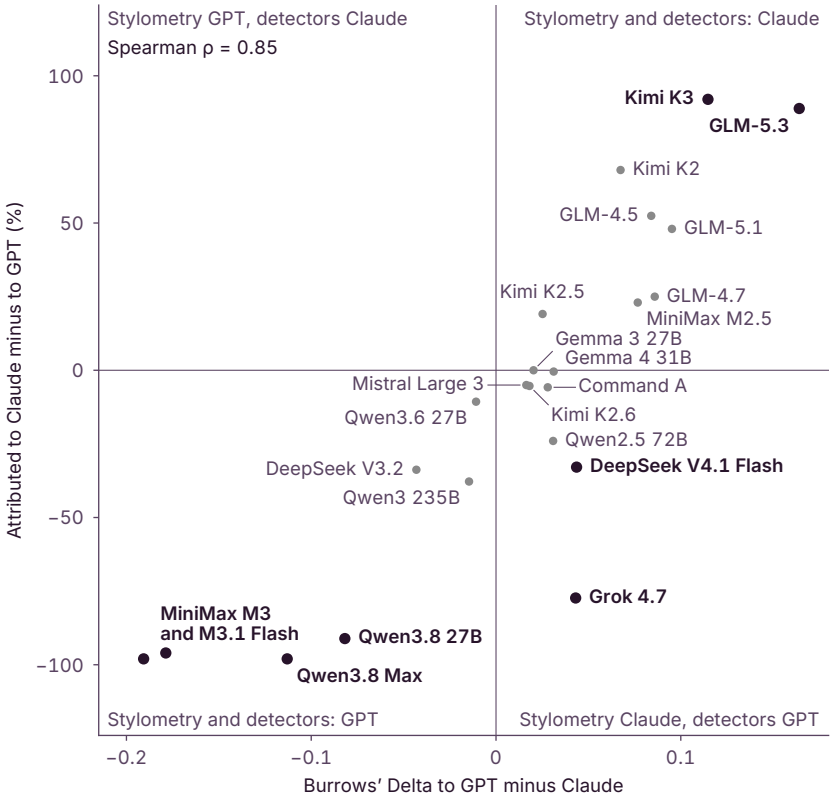

Figure 16. The closer stylometry places a writer to Claude or to GPT, the more often the detectors attribute its texts to that lab. Each point is one writer outside of the Claude, GPT, and Gemini model families (newest releases in bold), with detectors and domains pooled. Burrows' Delta is measured to each lab's closest model.

Stylometric analysis. We compare writers with Burrows' Delta (Burrows, 2002) over the relative frequencies of the 150 most common words and 16 punctuation, sentence and formatting features.[32] Within each domain of §C.1, we z-score each feature over all texts, and the distance between two writers is the mean absolute difference of their average z-scores over the prompts both wrote (0 means identical). For each pair of consecutive releases from a lab outside of the big three, we take the change in distance to each Claude, GPT, and Gemini model and test it with a paired bootstrap over prompts (2,000 resamples) and Holm correction across those models. We estimate the distance between the writer and itself as about 0.2 by splitting each writer's texts into two random halves and measuring the distance between them. Stylometry is able to separate the big three, with each Claude, GPT and Gemini model nearest another model from its own lab in every domain. This is consistent with Sun et al. (2025), who find that classifiers are able to tell apart the outputs of different LLM families, and Shan et al. (2026), who find that AI writers have a consistent stylometric footprint.

Most misattributions reflect similarities in writing style rather than detector error. We first look at whether the detectors' answers follow each writer's style, or whether they fall back to a known lab regardless of how the text is written. We compare the percent of each writer's texts the detectors attribute to Claude with how much closer stylometry places it to Claude than to GPT (Figure 16), and find that the two are strongly correlated (Spearman $\rho = 0.85$), with about the same correlation for each detector on its own. The stylometry, like our detectors, split the newest Kimi, GLM, Qwen, and MiniMax releases between Claude and GPT (all but Qwen3.8 27B are within 0.21–0.24 of their nearest model, against at least 0.27 for these labs' earlier releases). Within the MiniMax family, stylometry places M2.5 nearest Claude Sonnet 4.6 and M3 nearest GPT-5.6 Sol on news and op-eds, and the detectors go from mostly answering Claude for M2.5 to nearly always answering GPT for M3.

Information density explains only part of the shift toward Claude and GPT. Clark et al. (2026) find that recent models from several labs write in a dense style that saves tokens, but also makes it harder for people to read. To test whether the misattributions can be explained by a convergence in AI writing style, we re-implement six of their measures: object relative clauses with and without a relativizer, stacked nouns (compounds of three or more nouns), dependency length, unattested bigrams, and trigram perplexity.[33] We then regress each stylometric feature on these six measures over all texts in a domain, and recompute the distances on the residuals. The six measures predict little of most features, with an average R^2 of 0.06–0.08, and predict word length and sentence length best. On the residuals, Kimi K3 and GLM-5.3 are still nearest a Claude model and the newest Qwen and MiniMax releases nearest a GPT model in every domain, with 7 of the 12 shifts toward Claude and 16 of the 18 toward GPT still significant.[34]

D. Which human wrote the text?

In this section, we discuss data, additional results, and analysis related to human author attribution (§4).

D.1. Data preparation

Before sending each text to the model for attribution, we remove all direct identity clues (i.e. paper bylines), while keeping some indirect clues like self-citations.

Papers. We extract the main text from arXiv's HTML. We remove the byline (author names, affiliations and emails), the bibliography, the appendices and any section about acknowledgements, funding, author contributions, competing interests or disclosures. We replace emails, URLs (including GitHub and Hugging Face links) and arXiv identifiers with placeholders, as well as any affiliation from the byline that appears verbatim. We keep in-body citations. For each paper, we count self-citations and split accuracy by these counts (Table 4). Since the full main text is the condition most exposed to leakage, the topic-only baselines in §D.4.3 give the model only keywords or the title and abstract.

Public reviews. We remove every occurrence of the reviewer's name, form checklists, expertise statements and boilerplate. Reviewer names and ORCID IDs are kept only in the answer key. We report accuracy with and without self citations (Table 4).

[32] Em dashes, en dashes, semicolons, colons, parentheses, question marks, exclamation marks, ellipses, quotation marks, "not X but Y", words per sentence, spread of sentence length, one-sentence paragraphs, words per paragraph, markdown lines and word length.

[33] We parse with spaCy (`en_core_web_trf`, and build the reference sample from 500,000 words of English Wikipedia.

[34] The shifts toward Claude keep a median of 60% of their size (Kimi K3 48%, GLM-5.3 71%), and the shifts toward GPT keep 101% (Qwen 105%, MiniMax 101%).

Private reviews. We remove submission numbers, paper titles, author lists, scores and reviewer names. Self-citations are kept.

Stories. We remove author bios, credits, promotions, editors' introductions, authors' notes and content warnings.

Reddit. We remove quoted lines, comments that contain any form of the author's username, self-promotion or links to personal pages.

D.2. Experiment & data details

As we observe first-hand and is studied in Zhong et al. (2026), personal account information passed through harnesses like Claude Code can affect model responses. Therefore, we run all of our human author attribution experiments through APIs (Table 24).

Table 24. We list the access methods and configured reasoning effort for all models in §4. Models receive the system prompt and the bare text, with no tools, web search, or candidate names.

Model	Effort	Access
Fable 5.2	high	OpenRouter batches, pinned to Anthropic
Opus 5.5	medium	Anthropic Message Batches API; OpenRouter batches for stories and ICML, and as a fallback
Gemini 4 Carbon	medium	Gemini Batch API; standard Gemini API for output-cap reruns
Gemini 4 Argon	medium	Standard and Batch Gemini APIs
GPT-6.1 Sol	medium	OpenAI Responses Batch API; OpenRouter batches for ICML, pinned to OpenAI
DeepSeek V4.1 Flash	high	OpenRouter, pinned providers

Reddit communities and comment selection. Table 25 lists all the subreddits we gathered comments and posts from. When collecting the data, we queried Claude to choose subreddits likely to contain enough recent, long-form posts (> 200 words). We kept only comments with at least 200 words of the author's own text after removing quotes. We filter out moderator comments, bot-like accounts, obvious username leaks, self-promotion, personal website links, and sensitive disclosures. Figure 17 plots the number of comments from each subreddit and Fable 5.2's and Opus 5.5's accuracy within each.

Table 25. The 32 subreddits represented in our Reddit dataset from §4.

Subreddits		
r/AskAnthropology	r/aviation	r/MachineLearning
r/AskEconomics	r/badeconomics	r/math
r/AskElectronics	r/badhistory	r/Physics
r/AskEngineers	r/compsci	r/programming
r/AskHistorians	r/cpp	r/ProgrammingLanguages
r/asklinguistics	r/ExperiencedDevs	r/Python
r/askmath	r/golang	r/rust
r/askphilosophy	r/haskell	r/space
r/AskPhysics	r/learnprogramming	r/SpaceXLounge
r/AskSocialScience	r/sysadmin	r/AskStatistics
r/LocalLLaMA	r/wallstreetbets	

D.3. Additional results

We present full results from our human attribution experiments that are mentioned briefly in the main paper (i.e. confidence levels, accuracy by domain etc.)

D.3.1. Accuracy by domain

Table 26 gives each judge's accuracy in every domain of Figure 4, including Gemini 4 Carbon, which was run on every domain but the ICML 2020 reviews. Table 27 splits accuracy on public reviews and papers by the Pangram 4.1 label of the judged text.

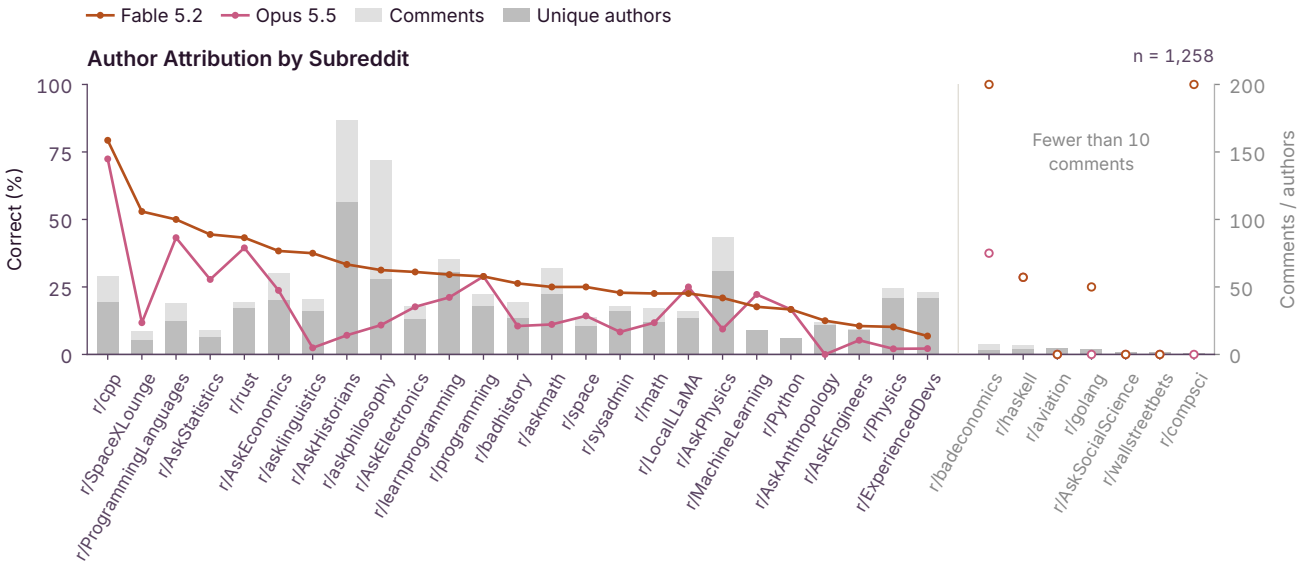

Figure 17. Accuracy varies widely across subreddits. Bars (right axis): the number of Reddit comments from each of the 32 subreddits in Table 25 (1,258 in total). Lines (left axis): the percent of comments in each subreddit for which Fable 5.2 and Opus 5.5 name the right username.

Table 26. Human author attribution (§4 (%)) by domain and model on the texts of Figure 4. *Not AI*: texts Pangram 4.1 does not label AI. Accuracy is over answered texts. Gemini 4 Carbon was not run on the ICML 2020 reviews.

Domain	Texts	Fable 5.2	Opus 5.5	Gemini 4 Carbon	Gemini 4 Argon	GPT-6.1 Sol	DeepSeek V4.1 Flash
Stories	all (206)	32.0	19.9	19.7	18.7	22.8	2.4
	not AI (191)	34.6	21.5	21.3	20.2	24.6	2.6
Papers	all (120)	46.2	34.7	31.9	31.7	22.5	7.6
	not AI (90)	57.3	43.2	40.4	40.0	28.9	10.1
Public reviews	all (1,559)	21.4	15.2	18.8	16.9	13.0	5.7
	not AI (789)	39.3	28.4	34.5	30.7	24.1	9.9
Private reviews	all (398)	56.0	41.2	40.3	40.7	9.3	2.5
	not AI (396)	56.3	41.4	40.5	40.9	9.3	2.5
ICML 2020 reviews	all (15,136)	55.4	48.8	–	37.7	22.3	5.4
Reddit	all (1,258)	31.1	16.8	12.0	10.0	5.2	5.3
	not AI (1,230)	31.8	17.2	12.3	10.2	5.4	5.4

D.3.2. Author-level accuracy

All results in §4 report accuracy in terms of the number of texts which are correctly attributed. Since there is an imbalance in the number of texts per author (especially for private reviews where there are only 6 authors), we also report the accuracy at the author level. We do this by calculating the accuracy for each author's texts and then averaging all authors' individual accuracies (Table 28). Accuracy averaged over authors is lower than over texts on Reddit (24% against 31% for Fable 5.2), because the most active users are both easier to name and contribute more comments. Fable 5.2 names at least one text's author for 26% of Reddit users, 29% of story authors and 22% of public reviewers.

D.3.3. Accuracy by author visibility

Table 29 and Table 30 give the accuracies plotted in Figure 6, binned by Reddit activity and number of Goodreads ratings.

Table 27. Every detector names fewer human authors on texts that Pangram 4.1 labels AI or mixed. §4 dataset accuracies (%) by the Pangram 4.1 label of the text.

Domain	Pangram label	Texts	Fable 5.2	Opus 5.5	Gemini 4 Carbon	Gemini 4 Argon	GPT-6.1 Sol	DeepSeek V4.1 Flash
Public reviews	Human	571	45.4	32.4	40.4	36.1	28.2	10.3
	Mixed	218	23.8	17.9	19.0	16.5	13.3	8.8
	AI	770	3.5	2.0	2.9	2.7	1.6	1.3
Papers	Human	34	84.8	72.7	61.8	64.7	35.3	17.6
	Mixed	56	41.1	25.5	27.3	25.0	25.0	5.5
	AI	30	13.3	10.0	6.7	6.7	3.3	0.0

Table 28. Accuracy by text and by author for human author attribution (§4). Accuracy over answered texts (%); author-macro is the mean of per-author accuracy; identified is the percent of authors named for at least one of their texts (public reviews include the 90 Nature-family reviews). Authors counts only authors with at least one answered text (e.g. 860 of the 870 Reddit users for Fable 5.2, it refused on the rest).

Detector	Domain	Answered	Authors	Accuracy	Author-macro	Identified
Fable 5.2	Stories	206	187	32.0	28.1	28.9
	Reddit	1,244	860	31.1	24.1	26.2
	Public reviews	1,530	1,467	21.4	21.5	21.7
	Private reviews	398	6	56.0	49.1	100.0
Opus 5.5	Stories	206	187	19.9	16.6	16.6
	Reddit	1,237	859	16.8	12.6	14.6
	Public reviews	1,521	1,457	15.2	15.3	15.5
	Private reviews	398	6	41.2	29.5	100.0

Table 29. Reddit author attribution accuracy (%) by the account's activity: its posts plus comments in the same subreddit from June 2025 to May 2026. Comments is the number in each bin.

Activity	Comments	Fable 5.2	Opus 5.5	Gemini 4 Argon	GPT-6.1 Sol
0	47	2.1	0.0	0.0	2.1
1–50	569	12.8	7.2	4.6	1.9
51–200	254	33.1	18.4	10.2	4.7
201–500	223	53.6	29.0	18.8	9.0
>500	165	68.9	36.0	19.4	13.3

Table 30. Story author attribution accuracy (%) by the total number of Goodreads ratings across the author's books. The 206 stories have 187 unique authors. None means the author has no Goodreads profile. Authors and Stories are the number of each in the bin.

Ratings	Authors	Stories	Fable 5.2	Opus 5.5	Gemini 4 Argon	GPT-6.1 Sol
None	69	71	4.2	0.0	0.0	0.0
<100	35	39	17.9	2.6	0.0	10.3
100–999	31	34	23.5	8.8	2.9	8.8
1k–10k	28	28	50.0	21.4	29.6	35.7
10k–100k	10	10	100.0	80.0	80.0	80.0
100k–1M	9	9	100.0	88.9	77.8	88.9
≥1M	5	15	100.0	100.0	93.3	93.3

D.3.4. Stated confidence

Each model is requested to report a confidence level in their prediction. Table 31 relates it to accuracy on pooled public texts: stories, papers, Reddit comments and all 1,559 public reviews, including the 90 Nature-family reviews. ICML 2020 reviews are not included. Confident mistakes are rare for Fable 5.2 and Opus 5.5 (0.3% and 0.4% of all answers name the wrong person at a stated confidence of 80% or more), but more common for Carbon and Argon (3.1% and 3.3%) and for DeepSeek V4.1 Flash (5.1%), whose mean stated confidence (43%) far exceeds its accuracy (5%). Figure 18 plots accuracy against stated confidence for every detector, on the same pooled public texts and, separately, the private reviews.

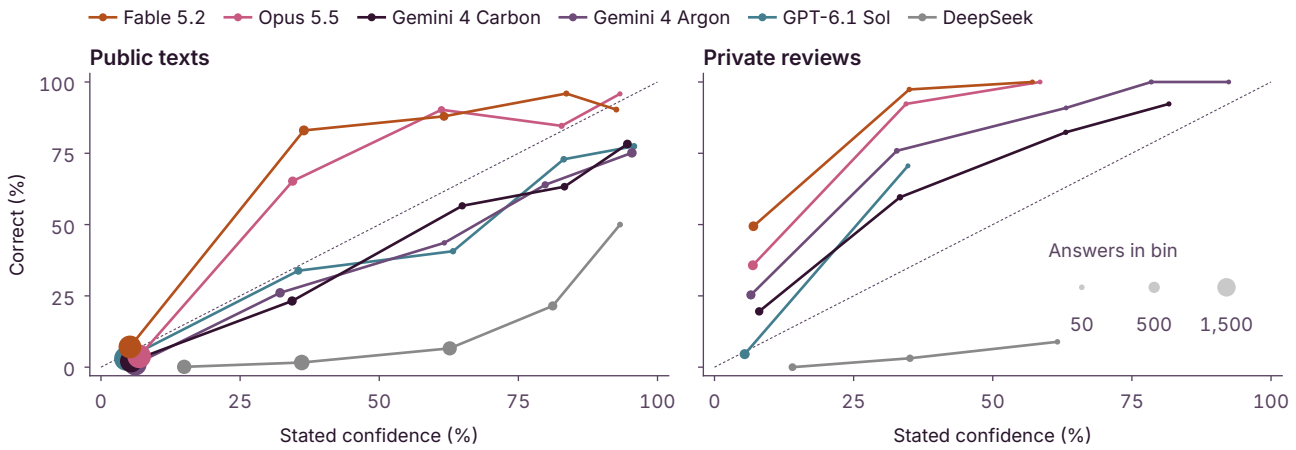

Figure 18. The Claude models understate their confidence, and DeepSeek overstates it. Percent of RQ3 answers that name the right person against the mean confidence the model stated, in bins of confidence levels (0–25, 25–50, 50–75, 75–90 and 90–100%). Public texts pool stories, papers, Reddit comments and all 1,559 public reviews, including Nature-family reviews; the private panel uses all 398 private reviews. The dashed line is perfect calibration; points above it are underconfident.

Table 31. Stated confidence and accuracy on pooled public RQ3 texts: stories, papers, Reddit comments and all 1,559 public reviews, including the 90 Nature-family reviews, in %. This is the public-text cohort in Figure 18. Accuracy and percents are over answered texts; refusals and other invalid replies are excluded. The last column counts answers stated at 80% or more that name the wrong person, as a percent of all answers. ICML 2020 reviews are excluded.

			Stated 25% or more		Stated 80% or more
Detector	Accuracy	Mean confidence	percent	Accuracy	Wrong
Fable 5.2	27.0	17.5	25.0	86.6	0.3
Opus 5.5	16.9	14.9	18.1	76.2	0.4
Gemini 4 Carbon	16.7	23.0	28.7	52.8	3.1
Gemini 4 Argon	14.8	22.3	26.7	52.6	3.3
GPT-6.1 Sol	10.9	13.9	16.0	52.2	1.3
DeepSeek V4.1 Flash	5.4	42.8	72.0	7.4	5.1

D.3.5. Meta-Reviews in private review data

Table 32 shows accuracy on private reviews, broken down by review type (meta-reviews vs. standard peer reviews). Meta-review attribution is more challenging.

Table 32. Accuracy (%) on the 398 private reviews from §4, with and without meta-reviews. 132 of the 398 reviews are meta-reviews written by two of the six writers. Pangram 4.1 labels none of the meta-reviews AI.

Texts	Fable 5.2	Opus 5.5	Gemini 4 Carbon	Gemini 4 Argon	GPT-6.1 Sol	DeepSeek V4.1 Flash
All private reviews						
All (398)	56.0	41.2	40.3	40.7	9.3	2.5
Reviews (266)	68.4	51.1	50.4	51.1	13.2	3.8
Meta-reviews (132)	31.1	21.2	19.2	19.7	1.5	0.0

D.4. Additional experiments

We do not run additional experiments on the ICML 2020 reviews, such as a topic-only baseline, self-citation check, or paraphrasing, because they would require amending our IRB protocol. We run them on the other RQ3 datasets instead (the topic-only baseline in §4.3, the self-citation split in Table 4, and rewriting in §D.4.4).

D.4.1. Self-reported familiarity

Table 33 shows attribution accuracy by each model's own familiarity rating across domains.

Table 33. Human author attribution accuracy (%, correct/answered) by the model's name-only familiarity rating. Known means a mean rating of at least 1.5 over two calls on a 0–3 scale; a paper takes its best-known listed author.

Domain	Not known	Known	Hits from known (%)
A\ Opus 5.5			
Stories	0.7 (1/135)	56.3 (40/71)	97.6
Papers	17.6 (3/17)	37.6 (38/101)	92.7
Public reviews	2.5 (22/890)	29.4 (161/547)	88.0
Nature-family reviews	37.5 (6/16)	61.8 (42/68)	87.5
Private reviews	N/A (0)	41.2 (164/398)	100.0
Reddit	4.2 (35/840)	43.6 (173/397)	83.2
G Gemini 4 Argon			
Stories	0.0 (0/133)	54.3 (38/70)	100.0
Papers	14.3 (3/21)	35.4 (35/99)	92.1
Public reviews	4.4 (47/1,060)	38.6 (157/407)	77.0
Nature-family reviews	29.4 (5/17)	74.0 (54/73)	91.5
Private reviews	6.9 (5/72)	48.2 (157/326)	96.9
Reddit	6.0 (70/1,173)	65.9 (56/85)	44.4

Table 34 compares each model's attribution accuracy with the other model's name-only familiarity rating.

Table 34. Attribution accuracy (%, correct/answered) grouped by the other model's familiarity with the writer. A writer is known if the mean of two name-only ratings is at least 1.5 on a 0–3 scale.

Attribution model	Rating model	Not known	Known
Stories			
A\ Opus 5.5	G Gemini 4 Argon	0.7 (1/135)	56.3 (40/71)
G Gemini 4 Argon	A\ Opus 5.5	0.8 (1/133)	52.9 (37/70)
Public reviews			
A\ Opus 5.5	G Gemini 4 Argon	4.5 (47/1,049)	35.1 (136/388)
G Gemini 4 Argon	A\ Opus 5.5	2.9 (26/900)	31.4 (178/567)
Nature-family reviews			
A\ Opus 5.5	G Gemini 4 Argon	31.2 (5/16)	63.2 (43/68)
G Gemini 4 Argon	A\ Opus 5.5	38.9 (7/18)	72.2 (52/72)

D.4.2. Refusals and answers that name no one

The model could fail to name a specific human author for multiple reasons. Sometimes the model genuinely cannot tell who wrote the text due to lack of clues, or it can explicitly decline to name a writer due to policy or ethical concerns. To separate the two, and to measure how much the prompt and the access route change them, we run Fable 5.2 (high effort) on 500 public peer reviews and 500 Reddit comments with two prompts (our regular and a simple one) through the API and CLI (Table 35). The public reviews include all 90 Nature-family reviews, plus a random sample of 410 of the other 1,469. The Reddit comments are all 72 comments that Fable 5.2 rates very familiar and a random sample of 428 of the other 1,186, in proportion to familiarity bin. Note that both samples over-represent writers which are familiar to Fable 5.2.

The *holistic* prompt is the one behind our main results (§4). The *simple* prompt asks "Who wrote this text? Name the author as precisely as you can." (for Reddit: "Which Reddit user wrote this comment? Give their exact username."), followed by a confidence line and an author line. The *Claude Code* route runs the Claude Code CLI in an isolated Docker container. We replace Claude Code's system prompt with ours, disable all tools and remove the account's email from the context.

We use Opus 5.5 (medium effort, with the same Claude Code setup) to label every response that doesn't name a specific human writer. It categorizes the reasons the response gives (misattribution harm, anonymity, privacy, fabrication, policy, or none), quotes the sentence it relies on, and records whether the response says the text was written or assisted by AI.

Table 35. Fable 5.2 declines far more often through Claude Code than through the API. *Named*: response names a human author; *Correct*: it names the correct author. A *refusal in answer* declines for a privacy, ethical or policy reason or for no stated reason (e.g., "I'd rather not guess"); a *can't identify* reply says that any name would be a complete guess or fabrication. *Other or blocked*: number of times the provider rejected the query (API), or Claude Code's safety classifier flagged it.

			Named		Refusal in answer, by reason cited								
Domain	Prompt	Route	All	Correct	All	Misattr.	Anon.	Privacy	Fabr.	Policy	None	Can't identify	Other or blocked
Public reviews	Simple	Claude Code	64	47	205	113	78	38	4	1	32	218	13
		API	294	128	22	3	1	2	1	0	16	168	16
	Holistic	Claude Code	184	97	268	221	135	76	49	1	9	32	16
		API	442	134	14	10	0	1	3	0	3	28	16
Reddit	Simple	Claude Code	34	22	443	422	2	6	34	19	13	23	0
		API	296	157	77	68	0	0	35	1	3	126	1
	Holistic	Claude Code	291	145	199	185	19	42	97	9	3	9	1
		API	480	181	16	13	2	3	4	2	1	2	2

The route (API vs. CLI) affects refusal rate more than the prompt. That being said, the simple prompt does result in more refusals especially when run through the CLI, especially on Reddit comments (Figure 19). Through the API, most replies that name no one say that the model cannot identify the writer and these replies are concentrated on authors with lower familiarity ratings (§D.4.1). When the same text and prompt is run through Claude Code, many of the replies become refusals even for familiar writers, and Fable 5.2 has much lower accuracy overall (Table 35).

Misattribution harm is the most common reason for refusal. Refusals on public reviews sometimes mention the reviewer's anonymity and privacy, which are very rarely mentioned for Reddit comments.

Replies that call a review AI-written agree with Pangram 4.1. About half of the public-review replies that name no one say that the review was written or assisted by AI, against about 1% on Reddit. Pangram 4.1 labels nearly all of these texts AI (93%), and only 4 of 530 fully Human.

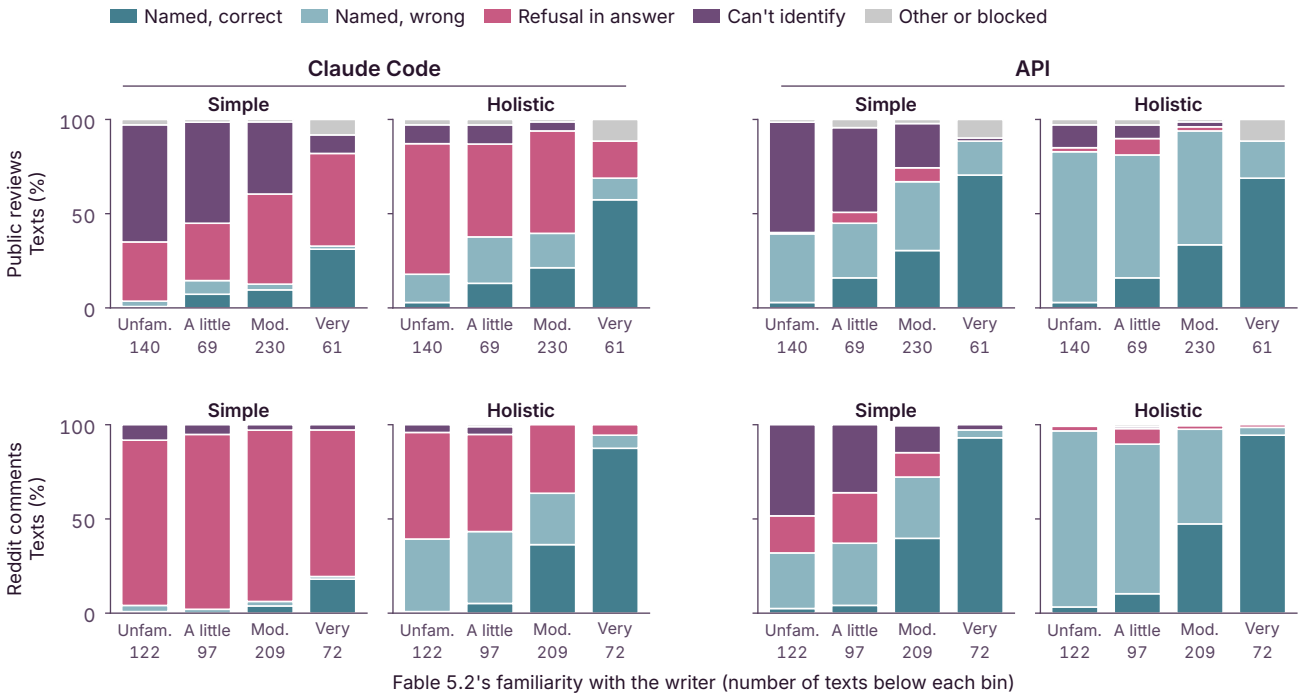

Figure 19. We show the refusal types from Table 35 based on Fable 5.2's familiarity rating of the author.

D.4.3. TOPIC-ONLY BASELINE EXPERIMENTS

It is possible that a model could name an author because it knows who works on a given topic, not because it recognizes how they write (i.e. use content of the text instead of stylistic features). To check this, we give the model only the topic of a text (keywords, or a title and abstract) and ask who works on it or reviewed it. We then compare that with attribution from the full text. We find that providing the full text gives higher accuracy on both papers and public reviews (Table 36 and Table 37).

Table 36. We test a topic-only baseline for paper author attribution. We show the percentage of 34 recent cs.CL papers (labelled human by Pangram 4.1) for which Opus 5.5 names a co-author when asked for senior and/or early career researcher who works on the paper's topic. We compare with our regular full main text version which asks who wrote the paper. The topic-only prompts ask for one early-career and one senior researcher who work on the problem; the full-text prompt asks for the most likely junior and senior author. A name counts if the person is any author of the paper, whatever role Opus 5.5 gives them.

Input	Question	Senior researcher (%)	Either person (%)
Five topic keywords	Who works on this?	9	12
Title and abstract	Who works on this?	41	53
Full main text	Who wrote this?	65	76

Table 37. We test a topic-only baseline for public reviews. Opus 5.5 gives one name per reviewed paper, so from the title and abstract at most 87.3% of reviews can be named. Reviews are the 789 public reviews of Figure 4 that Pangram 4.1 does not label AI.

Input	Question	Reviewer named (%)
Title and abstract of the reviewed paper	Who reviewed this?	2.0
Full review	Who wrote this?	28.4

D.4.4. STABILITY ACROSS REPEATED JUDGMENTS AND PARAPHRASING

Overall accuracy over repeated judgments varies little, though individual predictions can change. In five independent Fable 5.2 judgments of Reddit comments, overall accuracy ranges from 29.9–31.1% (Table 38, top). Only 32.6% of comments receive the same normalized answer in all five runs, whereas 25.0% are correctly attributed in every run and 63.2% are incorrect in every run. The remaining 11.8% are correct in one to four runs. Earlier same-text repeats on a 200-document, five-domain sample also show relatively small accuracy ranges for Argon and Opus 5.5 (Table 38, bottom).

Model judgments exhibit high agreement and stability in their explanations. Following a similar explanation agreement method from §2 across three judgments on a 200-text sample, we find that Argon has agreement evidence F1 of 0.94 and decisive-factor F1 of 0.79, while Opus has 0.93 and 0.72 F1 respectively. We also find high cross-model Argon/Opus agreement: on a separate 1,000-document sample their evidence F1 is 0.85 and decisive-factor F1 is 0.67 (Table 39).

Table 38. Attribution stability on unchanged texts. Top: consistency among the 1,240 Reddit comments with five valid answers.

Fable 5.2: five-run consistency ($n = 1{,}240$)				
Pattern		Comments		Share
Same answer in all five		404		32.6
Correct in all five		310		25.0
Incorrect in all five		784		63.2
Correct in one to four		146		11.8
200 Document Sample (all domains), three runs				
Model	Documents	Original	Repeat 2	Repeat 3
Gemini 4 Argon	194	20.6	21.6	21.1
Opus 5.5	176	21.6	23.3	25.6

Table 40. Attribution accuracy before and after rewriting on the Figure 4 dataset. Columns show the accuracy for the original and two independent Argon rewrites, judged by Argon or Opus 5.5 with the same holistic prompt. Each row uses the same n texts with valid judgments of all three versions; original accuracies can differ from Figure 4, which uses all valid originals. Reddit requires the exact username, papers accept any listed author, reviews and stories use the shared name matching rule. ICML 2020 reviews have no rewrite judgments and are excluded. *All texts* includes every text regardless of its Pangram 4.1 label; *Not AI* keeps the texts that Pangram 4.1 labels human or mixed, as in Table 26.

	All texts				Not AI			
Detector	*n*	Original	Rewrite 1	Rewrite 2	*n*	Original	Rewrite 1	Rewrite 2
Private reviews								
Gemini 4 Argon	398	40.7	14.8	15.1	396	40.9	14.9	15.2
Opus 5.5	398	41.2	16.1	15.1	396	41.4	16.2	15.2
Reddit								
Gemini 4 Argon	1,253	10.0	7.0	6.7	1,225	10.2	7.2	6.9
Opus 5.5	1,236	16.8	10.3	10.5	1,208	17.2	10.5	10.8
Public reviews								
Gemini 4 Argon	702	23.5	15.1	15.4	395	38.7	25.1	25.8
Opus 5.5	672	20.8	15.0	14.6	371	35.8	25.3	24.0
Stories								
Gemini 4 Argon	199	18.6	15.6	16.1	184	20.1	16.8	17.4
Opus 5.5	199	20.1	16.1	16.6	184	21.7	17.4	17.9
Papers								
Gemini 4 Argon	120	31.7	23.3	26.7	90	40.0	28.9	32.2
Opus 5.5	118	34.7	31.4	31.4	88	43.2	39.8	40.9

Table 39. Agreement between model explanations from predicting author of the same text. Stability rows compare three judgments of each text by the same model (up to three pairs per text, 200 texts); the cross-model row compares Argon's and Opus 5.5's judgment of each text (one pair per text, 1,000 texts). *Documents* counts texts with at least one scored pair.

Comparison	Documents	Pairs	Evidence F1	Decisive F1
Argon–Argon	199	573	0.935	0.794
Opus–Opus	199	557	0.930	0.726
Argon–Opus	915	915	0.848	0.666

D.4.5. Topics in reviewer-attribution explanations

Table 41 gives TopicGPT's definition of topics referenced in Figure 7. We generate separate topic lists for ICML 2020 reviews and Reddit comments, so the table has one block for each. We use GPT-6.1 Sol to process 5,000 ICML reviews and 1,200 Reddit comments to propose 154 and 97 potential topics respectively. GPT-6.1 Sol is then asked to refine the topic lists by removing rare topics and merging similar ones. Our final list contains 77 topics for ICML explanations and 45 for Reddit. Each explanation is assigned 9–10 distinct topics on average. The topics label what a model wrote in its explanation, not the text itself. Evidence topics describe what the model noticed in the text, and inference topics describe what it concluded about the writer. The most frequent topics in the explanations include "Peer-review format" (the text is identified as a peer review) and "Writing style" on ICML, and "Tone and stance" and "Subject focus" on Reddit.

D.4.6. Factuality of claims in model explanations from ICML 2020 reviews

We check whether the external facts that judges state in their explanations are true, following the claim-extraction approach of VeriScore (Song et al., 2024). We use the explanations of Opus 5.5, Argon, Sol and DeepSeek for the same 5,000 ICML 2020 reviews as the topic analysis (§D.4.5); Fable 5.2's explanations were generated after this check and are not included. Gemini 4 Argon extracts 8,448 candidate external claims from the 107,645 observations in these 20,000 explanations, 0.42 per explanation (Opus 5.5 3,073, Argon 2,596, Sol 561, DeepSeek 2,218). Most observations describe the review itself or speculate about the reviewer, so few contain a claim that can be checked outside the review. Argon then checks 5,477 of these claims, from 2,000 reviews, with Google Search grounding and labels each one supported, contradicted, insufficient evidence or not assessable. It labels 5,192 claims supported, 161 contradicted, 61 insufficient evidence and 63

Table 41. Definitions, as written by TopicGPT, of the topics in Figure 7, of the structure topics it groups in one row, and of the topics named in §4.3. ICML 2020 reviews and Reddit comments have separate codebooks, generated independently, so a topic in one block does not correspond to a similarly named topic in the other.

Topic	Type	TopicGPT description
ICML 2020 reviews (77-topic codebook)		
Shown in Figure 7		
Technical understanding and reasoning	Evidence	The reviewer demonstrates understanding of relevant concepts, technical distinctions, or precise reasoning.
Prior-work references	Evidence	The reviewer cites specific prior work or distinguishes its attribution and historical relationship to other work.
Claimed publication experience	Evidence	The author explicitly claims publication experience in the area under review.
Linguistic background	Inference	Language patterns are explicitly linked to the author's native language or language background.
Researcher role	Inference	The explanation explicitly identifies the author as a researcher in a relevant area acting as a reviewer.
Anonymous reviewer role	Inference	The explanation explicitly identifies the author as an anonymous reviewer for a publication venue.
Structure topics (one row in Figure 7)		
Peer-review format	Evidence	The text is identified as a peer review through its structure, context, or explicit description.
Review-platform provenance	Inference	The explanation explicitly attributes elements of the review to a reviewing platform or its template.
Standardized review language	Evidence	The explanation identifies wording supplied by a standard review form rather than distinctive author-written prose.
Review ratings	Evidence	The explanation identifies numeric scores or rating fields in the review as an observed clue.
Named in §4.3		
Writing style	Evidence	The explanation describes the prose's concision, directness, tone, punctuation, or organization.
Language mechanics errors	Evidence	The text contains grammatical irregularities or spelling errors without necessarily attributing them to a language background.
Non-native language patterns	Evidence	The text contains spelling, grammatical, or syntactic patterns described as characteristic of non-native English.
Reddit comments (45-topic codebook)		
Tone and stance	Evidence	The explanation identifies the writer's tone, attitude, or evaluative posture.
Subject focus	Evidence	The explanation identifies the subject matter around which the discussion centers.
Punctuation conventions	Evidence	The writing exhibits an identified punctuation practice.
Text formatting	Evidence	The writing uses identifiable visual formatting or markup conventions.
Sentence structure	Evidence	The explanation identifies characteristic sentence lengths or syntactic constructions.
Capitalization conventions	Evidence	The writing exhibits an identified practice of capitalizing words or terms.
Language errors	Evidence	The writing contains grammatical slips, typos, run-on sentences, or other unpolished language.

not assessable. 95.9% of the 5,414 assessable claims (94.8% of all checked claims) are supported.

E. Cost estimate

Table 42. Estimated cost (USD) of the experiments reported in this paper. *API*: API charges. *CLI*: runs through Claude Code or Codex subscriptions, priced at API list rates for the tokens used.

	API	CLI	Total
§2: human or AI	4,983	5,649	10,632
§3: which model	169	831	1,000
§4: which human	2,199	510	2,709
Total	7,351	6,990	14,341

Limitations of the cost estimate. These are estimates. They leave out runs with no saved usage, the feature pipeline of §B.3.3, the Reddit familiarity, imitation and Fable 5.2 and Argon reasoning-effort runs. Costs of our joint-prompt variant judgments are not counted.