IDEALENS: DETECTING AI IDEAS IN LONG-FORM WRITING

Anonymous authors
Paper under double-blind review

ABSTRACT

While modern AI detectors identify ***who wrote the words***, emerging policies on AI use increasingly hinge on a different question: ***who came up with the ideas?*** We introduce IDEALENS, a detector that identifies whether a document's ideas came from a human or AI (*idea provenance*), regardless of who wrote its words. To focus IDEALENS on ideas rather than prose, we represent documents as outlines: lists of items that each pair a discourse role with a brief, paraphrased description of the content, minimizing word-level overlap with the raw text. We train IDEALENS on 1M FineWeb documents with silver labels from Pangram, a prose provenance detector. Since the outlines are largely stripped of surface-level information, the labels must be fit mainly through the ideas. In a controlled study, IDEALENS's AI flag rate drops from 95% to 7% as models write from increasingly detailed human plans, while Pangram 4 still flags 92%; from AI-derived plans, IDEALENS stays above 96%. Conversely, on a new dataset of 50 stories that human authors wrote from AI-generated plans, IDEALENS flags 68% of the stories as AI, compared to 8% for Pangram 4. On a comprehensive suite of 19 existing detection benchmarks, we show that IDEALENS maintains strong detection rates at low false positive rates, suggesting that ideas themselves provide a powerful discriminative signal, and its performance holds across domains, formats, and languages. Finally, we examine 90K predictions from IDEALENS to characterize systematic differences between human and AI ideation. We release our models and labeled datasets to facilitate future research on idea provenance detection.

1 INTRODUCTION

In August 2026, the AI detector Pangram flagged a *Wall Street Journal* op-ed (Druckenmiller, 2026) by billionaire investor Stanley Druckenmiller as 100% AI-generated.[1] Druckenmiller unapologetically responded: "Of course I used AI". Paul Gigot, the editor of WSJ's editorial pages, defended Druckenmiller by arguing that what matters is whether an op-ed "reflects an author's original argument" and that "nobody can doubt that his op-ed is his genuine opinion" (Nover, 2026), though he provided no supporting evidence from the writing process (e.g., chat logs). Detectors like Pangram answer only *who wrote the words* (**prose provenance**), while Gigot's defense, like many emerging policies on AI use in writing (Springer Nature, 2026; ICLR 2027 Program Chairs, 2026), hinges on *who came up with the ideas* (**idea provenance**).[2]

Idea and prose provenance can differ within the same document: a human can brainstorm the ideas for a document and ask an AI to write it, or vice versa. Prior work separates the two by building detectors around simulated human–AI collaborations, such as human writing polished by AI or AI writing run through humanizers (Bao et al., 2025; Li et al., 2026b). These simulations are expensive to create at scale, and detectors trained on them may not transfer to real collaborations. In contrast, prose detectors like Pangram are trained on documents whose ideas and words come from the same source, which are cheap to collect: human text can be sampled from the pre-ChatGPT era, and AI text can be generated on demand.

[1] https://www.pangram.com/history/5e4229f1-bef6-4137-8959-479ea4851c5e

[2] IDEALENS predicts that the op-ed's ideas came from AI with a probability of 98.3%. This is a statistical estimate and of course does not constitute direct evidence of the writing process.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

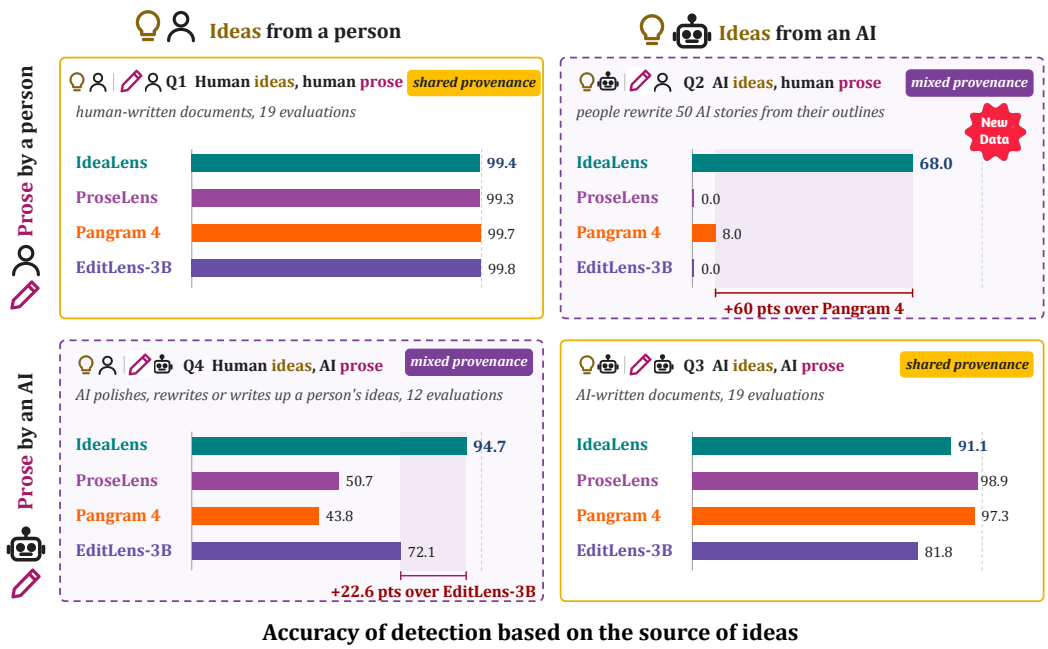


Figure 1: **IDEALENS tracks ideas; traditional detectors track wording.** We present the four quadrants covering both *shared* provenance (fully human or fully AI) and *mixed* provenance (human ideas and AI prose, or AI ideas and human prose) scenarios. Unlike traditional detectors (e.g., Pangram 4, Glickenhaus et al., 2026; EditLens-3B, Thai et al., 2026), IDEALENS can accurately detect the source of ideas, not the source of the prose. IDEALENS classifies 94.7% of AI-generated documents based on human ideas as human, while Pangram labels only 43.8% as human. On our new dataset of human-written stories following AI ideas (§3), IDEALENS flags 68.0% as AI, compared to Pangram and EditLens flagging only 8.0% and 0.0%, respectively.

We propose IDEALENS, an idea provenance detector that learns from the same cheap labels as prose detectors and can therefore be trained at their scale. Our key methodological contribution is to change the classifier's *input*. Following 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. We represent these ideas as an *outline* whose items each consist of a discourse role paired with a short description of its content, paraphrased to remove the original wording (Figure 2). We train IDEALENS on a dataset of 1M outlines extracted from FineWeb (Penedo et al., 2024) documents, with silver labels from Pangram. As a control, we also train a separate model, PROSELENS, with the same backbone, documents, and labels, but on the raw text instead of outlines. Despite identical supervision, the two behave in opposite ways whenever a document's ideas and words come from different sources, showing that the outline forces IDEALENS to fit the labels through ideas rather than words.

- **IDEALENS detects human ideas in AI-generated documents.** On a dataset of AI documents generated from increasingly detailed human plans, IDEALENS's flag rate falls from 94.9% with minimal human input to 6.8% with a full human outline, while remaining above 96% when plans come instead from AI. PROSELENS and Pangram 4 flag 86.5% and 92.4% of documents generated from full human outlines. This behavior is further supported by results on 10 external benchmarks of AI polishing, paraphrasing, and rewriting, where IDEALENS flags 5.6% of documents on average, versus 46.0% for PROSELENS and 52.0% for Pangram.
- **IDEALENS detects AI ideas in human writing.** The opposite scenario is harder to study: prior work touches on it only through AI text that people or models edit. To this end, we collect the first dataset of long-form documents that people wrote from scratch following AI ideas. Our dataset contains 50 stories, each written from an outline extracted from an AI-generated story, without access to the raw text of that story. IDEALENS flags 68.0% of the human-written stories and

96.0% of the AI originals, while PROSELENS and Pangram flag all of the AI original stories but only 0% and 8.0% of the human-written stories, respectively (Figure 1).

- **Despite not seeing the raw text, IDEALENS rivals the strongest open-source detector on conventional benchmarks and transfers across 24 languages.** On 19 evaluations that require separating fully AI (ideas and writing) from fully human documents, IDEALENS flags 91.1% of AI documents on average with a false positive rate of 0.6%, compared to 81.8% and 0.2% for EditLens-3B, the strongest existing open-weight detector. PROSELENS is even stronger here (98.9% and 0.7%), rivaling the commercial Pangram 4 at its published threshold (97.3% and 0.3%). Although trained only on English, IDEALENS transfers to 24 languages through English outlines, detecting 95.3% of AI documents on average at a false positive rate of 0.1% compared with 82.5% and 0.0% for PROSELENS.
- **IDEALENS flags ideas whose details could apply to anything or can't be verified.** Based on an analysis of 90K item-level predictions, we find that IDEALENS scores ideas as more AI when they cite numbers or methods without the context needed to check them, or claim more than their evidence supports. IDEALENS scores ideas as more human when they include specific sources that a reader could look up or discuss downsides and open problems.

We publicly release both IDEALENS and PROSELENS along with their training data, the evaluation datasets that we create, and our idea-level analysis.

2 LEARNING IDEA PROVENANCE FROM OUTLINES

The fundamental challenge in training an idea detector is that we have no large-scale labels for who came up with a document's ideas, only for who wrote its prose. Thus, we train on prose labels while replacing each document with an outline that preserves its ideas but not its wording. To create these outlines (Figure 2), we identify discourse roles typical of each writing format, use them to extract an outline of each document, and paraphrase the outline to remove wording copied from the raw document. We train IDEALENS to predict each document's Pangram label from its paraphrased outline, while PROSELENS predicts the same labels from the raw document.

Web documents with silver labels from Pangram. We sample a 1M subset from WildAI (Anonymous, 2026), which labels documents from FineWeb (Penedo et al., 2024) with WebOrganizer (Wettig et al., 2025) formats and topics along with authorship labels from Pangram 3.3.2 (Thai et al., 2026).[3] We focus on long-form writing, selecting English documents of at least 500 words that belong to one of eight relevant WebOrganizer formats (Table 12), and that Pangram labels as either human-written or AI-generated. We assume that the ideas and prose in these documents share the same source, which enables us to use them as noisy *silver* labels for idea provenance. We discard documents that Pangram labels as *mixed*, since AI only wrote part of these texts and thus idea provenance is unclear. We split the full corpus of 1M documents into training (∼842k), validation (∼30k), calibration (∼80k, human documents only)[4] and test (∼50k) sets (Table 12).[5]

Outlines of discourse roles. Cognitive models of writing (Flower & Hayes, 1981; Hayes, 2012) separate the act of *planning* from the act of *translating* these plans into prose. To build a document representation of the former without the latter, we turn to discourse theory, which describes the roles a passage plays within the broader context of the document, for example through rhetorical structure theory (Mann & Thompson, 1988; Taboada & Mann, 2006). We represent a document as an *outline*: an ordered list of discourse roles, each paired with brief descriptions of how those roles are implemented in the document. For example, the op-ed displayed in Figure 2 contains 17 items, starting with an *Event Account* of a Treasury announcement followed by an *Evaluation* of the market's reaction. Many of these roles are format-specific (Swales, 1990; Sun et al., 2021; van Dijk,

[3] We note that Pangram 3.3.2 only provides training labels, while our evaluations benchmark IDEALENS and PROSELENS against the more recent Pangram 4.

[4] The calibration split is used to set a threshold for each detector, which turns its predicted probabilities into binary decisions (AI vs. human). Because the calibration split only contains human-written documents, we can set the threshold to achieve a target false positive rate (FPR). Unless explicitly stated otherwise, we report detector accuracy at thresholds calibrated for an FPR of 1%.

[5] Building the corpus cost about $66.4k (§F).

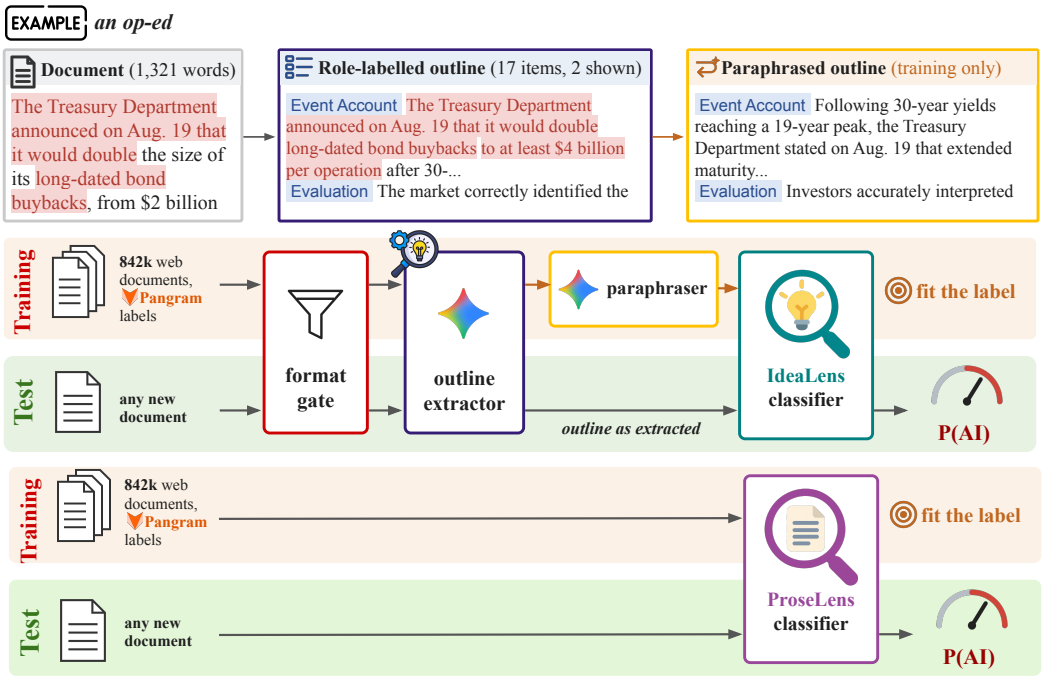


Figure 2: **IDEALENS sees an outline of each document, not the whole text.** Documents are assigned a format and reduced to an outline, each labeled with one of the format's roles. This example shows two of the 17 outline items extracted from an op-ed. For training, the outlines are paraphrased to remove wording copied from original documents. IDEALENS learns the Pangram label from the paraphrased outline during training, and extracted outlines are provided at inference.

1980); for instance, creative writing uses different roles than nonfiction. Thus, we induce separate role sets for each of the eight formats in our dataset (Table 6).

Automatically extracting outlines. We build a training dataset for IDEALENS by converting documents to outlines in three automated steps:

(i) **Role induction:** We adapt the TopicGPT framework (Pham et al., 2024) by prompting GPT-5.6 Sol (OpenAI, 2026a) to propose base roles for each format given a detailed description of that format. Then, we label outlines of 200 documents in that format using those roles, while asking the model to invent new roles as needed. Afterwards, we deduplicate and prune the resulting role set (e.g., to 42 roles for nonfiction); more details in §D.1.

(ii) **Outline extraction:** Given a document, we prompt Gemini 3.7 Flash (Google DeepMind, 2026c) and Gemini 3.1 Pro (Google DeepMind, 2026a) with its format's role set (and few-shot demonstrations from the much more powerful Claude Fable 5 (Anthropic, 2026a)) to produce an outline of the document's ideas in order, where each item consists of a role and a short content description.

(iii) **Paraphrasing:** Extracted outlines can copy phrasing from the original document, which a model could use as a stylistic shortcut to the label (i.e., human wording can leak into the outlines). Therefore, we prompt Gemini 3.1 Pro (Google DeepMind, 2026a) to paraphrase each outline. This pushes all outlines toward a homogeneous, AI-like style: for human-written documents, the fraction of outline text that Pangram 4 labels as human drops from 8.5% to 0.3% (Table 9). Meaning is still preserved (Table 10), as paraphrased outlines have a cosine similarity of 0.93 with the originally extracted outlines (vs. 0.39 for unrelated outlines) and 0.76 with the source document (vs. 0.26). We paraphrase only for training, to keep the model from learning style; at inference, IDEALENS sees the unparaphrased outline.

Training IDEALENS and PROSELENS. Both of our models start from a backbone of Nemotron-3.5-Lightning-30B-A3B (NVIDIA, 2026), fine-tuned for one epoch with LoRA (Hu et al., 2022) using Tinker (Thinking Machines Lab, 2026).[6] We threshold each model's probability that a document is AI-generated at a 1% false positive rate on our calibration split. The only difference between the two models is that IDEALENS is trained on paraphrased outlines, whereas PROSELENS is trained on the original documents. Since they share their backbone, training documents, and labels, any difference in their behavior is a direct result of transforming the input into our outline representation.

3 EXPERIMENTAL SETUP

We evaluate IDEALENS and PROSELENS against 17 existing detectors, including the state-of-the-art Pangram 4, on 19 external benchmarks and three datasets that we create ourselves: IDEASHIFT, IDEASHIFT-X, and TWICETOLD. These datasets cover all four quadrants in Figure 1: documents with *shared* provenance (fully human or fully AI) as well as documents with *mixed* provenance (human ideas and AI prose, or AI ideas and human prose). Idea detectors should achieve high accuracy in all four quadrants, while prose detectors should fail in the two mixed quadrants.

Evaluation metrics and baselines. We report **accuracy** at detecting AI ideas, which we define as the percentage of documents whose predicted label (AI or human) agrees with the source of their ideas. Our baselines include Pangram 4 (Glickenhaus et al., 2026), EditLens-Llama-3.2-3B (Thai et al., 2026), Fast-DetectGPT (Bao et al., 2024), Binoculars (Hans et al., 2024), MELD (Li et al., 2026a) and Desklib (Desklib, 2025). For IDEALENS, PROSELENS, and EditLens, we set a global threshold across all formats at 1% FPR on our calibration set. We binarize Pangram's ternary labels (AI, AI-assisted, human) following its tech report, and other baselines use their default thresholds.[7]

Existing benchmarks. Of the 19 external detection benchmarks (see §C for full list), 14 contain documents with shared provenance (fully human or fully AI). We evaluate accuracy on these documents and on our in-domain test set of 50K documents from our Pangram-labeled web corpus. We also measure FPR on a set of 10K documents from C4 (Raffel et al., 2020), which was created prior to ChatGPT's release and is thus much more likely to be human-written than our silver-labeled test set. Finally, 10 external benchmarks also contain documents written by humans but edited by an AI (e.g., via paraphrasing, polishing, or rewriting), which we use as instances of human ideas rendered in AI prose; this includes OpAI-Bench (Bsharat et al., 2026) and GEDE (Gehring & Paaßen, 2025).

IDEASHIFT: AI writing from increasingly detailed human plans. A common collaborative writing pattern is for a human to give an AI their ideas via a prompt, which it uses to write a document from scratch (Guo et al., 2025). Only two external datasets we evaluate on contain instances of this pattern (Saha et al., 2026; Gehring & Paaßen, 2025), and neither controls the *amount* of human input, so it is unclear how much of the final document's ideas came from the human. Thus, we create IDEASHIFT from 500 seed documents (both human-written and AI-generated) sampled from our in-domain test set. For each seed document, GPT-5.6 Sol (OpenAI, 2026a) writes five progressively more detailed prompts: the level-0 prompt contains only the topic (e.g., "Write a dining and lifestyle blog post reviewing a restaurant visit."), while the level-5 prompt contains the full outline extracted as per §2 with instructions to add nothing else to it. GPT-5.6 models then generate a new document for each prompt. Since all of the prose in IDEASHIFT is AI-generated, prose detectors should flag every document as AI, and an idea detector should do the same for AI seeds. However, on documents generated from human seeds, idea detectors should flag fewer and fewer documents as the prompts carry more of the human plan.

TWICETOLD: humans writing from AI plans. We are also interested in the reverse pattern: a human obtains ideas from an AI and then writes the document themselves. Existing datasets approximate this pattern with AI-generated text that humans edit or AI tools humanize,[8] but these edited documents are still largely AI-generated and thus do not cleanly separate idea and prose

[6] Details and results for other backbones are in §E.1.

[7] Details on thresholding, including the Pangram conversion, are in §E.3.

[8] E.g., LAMP (Chakrabarty et al., 2025), HART (Bao et al., 2025), and Baumler et al. (2026).

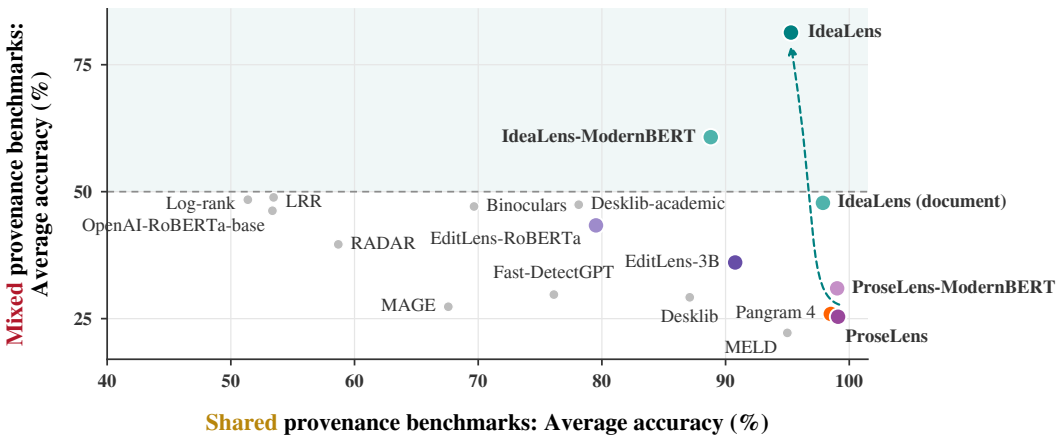


Figure 3: **Outline-trained IDEALENS detects idea authorship while traditional detectors track prose authorship.** The average accuracy across all benchmarks where ideas and prose are shared by one author (x-axis) and where they are by different authors (y-axis). IDEALENS is the only detector accurate with both shared (95.3%) and mixed (81.3%) authorship.[11]

provenance. To fill this gap, we create TWICETOLD,[9] a dataset with 50 stories written *from scratch* by humans given only the outline of a 500-word story generated by GPT-5.6 Sol (see §H.2 for more details).[10] Prose detectors should label these stories as human, while idea detectors should label them as AI. We also include the 50 original AI stories, which all detectors should label as AI.

IDEASHIFT-X: idea detection in multilingual documents. While prose detectors rely on language-specific features (e.g., vocabulary choice, sentence structure), IDEALENS uses an LLM extractor that can generate *English* outlines for documents in any language. Thus, even though IDEALENS is trained only on English documents, we can apply it to documents in other languages without any additional training. To test this, we create IDEASHIFT-X by sampling 10K human-authored documents in 24 languages (Table 29) written before 2022 from FineWeb2 (Penedo et al., 2025), to measure FPR across different languages. For each, we follow the IDEASHIFT protocol to generate AI versions in the original language at level 0 (AI ideas) and level 5 (human ideas).

4 IDEALENS DETECTS AI IDEAS, NOT AI PROSE

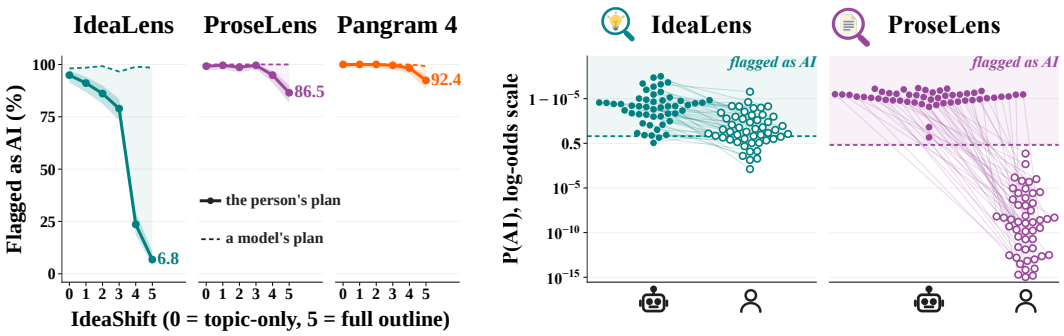
Figure 3 summarizes our main results across 51 evaluation splits from 22 benchmarks, where each point is a detector plotted by its aggregate idea detection accuracy on documents with shared vs. mixed provenance. IDEALENS is the only detector that achieves high accuracy in both shared (95.3%) and mixed (81.3%) settings. PROSELENS and Pangram 4 are extremely accurate on documents with shared provenance (99.1% and 98.5%, respectively) but fail to identify the source of a document's ideas in mixed settings (25.4% and 25.9%). Since PROSELENS has the same backbone, training documents, and training labels as IDEALENS, this highlights the impact of our outline representation, which a ModernBERT-based pair (Warner et al., 2025) replicates with a different backbone. IDEALENS also transfers unexpectedly well to *raw text* inputs at test time, exhibiting high shared provenance accuracy and substantially higher mixed provenance accuracy than PROSELENS.

IDEALENS detects idea provenance on IDEASHIFT. As the human plan in the prompts from IDEASHIFT becomes more detailed, IDEALENS's flag rate drops from 94.9% at level 0 to 6.8% at level 5 (Figure 4a). In contrast, PROSELENS and Pangram 4 flag 86–100% of documents regardless of the level of human ideation. When the plans come from AI instead of humans, IDEALENS

[9]Named after *Twice-Told Tales*, a collection of previously published short stories by Nathaniel Hawthorne.

[10]The stories were written in person by 17 computer science graduate students or faculty. We avoided crowdsourcing to reduce the chance of AI use during the writing process (Veselovsky et al., 2025).

[11]Even with weighted average accuracy, IDEALENS is more accurate at 92.6% compared to PROSELENS's 46.8% and Pangram's 41.1% on mixed provenance benchmarks.



(a) IDEASHIFT (b) TWICETOLD

Figure 4: **IDEALENS detects varying levels of human input in AI-generated text (left) and assigns similar P(AI) for TWICETOLD and its AI-versions (right).** (a) On IDEASHIFT, IDEALENS is the only model whose flag rate tracks with varying levels of human input in the AI-generated text; on the AI control, all models flag nearly entirely at all levels. (b) The difference in P(AI) between the AI and human-written stories in TWICETOLD is much smaller for IDEALENS than for PROSELENS.

consistently flags more than 96% of documents at every level (Table 25), showing that it is sensitive to idea provenance rather than level of prompt detail. Finally, we observe that the biggest drop in flag rate occurs between levels 3 and 4, which is when the prompts begin including outlines with *content* in addition to structural role information (Table 24).

IDEALENS is insensitive to AI rewording of human ideas. Aggregated over 12 benchmarks where AI makes surface-level changes to human text (e.g., polishing, paraphrasing), IDEALENS labels 94.7% of documents as human, compared to 50.7% for PROSELENS and 43.8% for Pangram 4 (Table 19). In GEDE (Gehring & Paaßen, 2025), IDEALENS flags only 0.5% of human essays as AI after minor corrections made by an LLM (vs Pangram 4's 54.6% and PROSELENS's 45.8%) and 2.6% after a rewrite preserving the original content (vs Pangram 4's 94.3% and PROSELENS's 77.1%). In HART (Bao et al., 2025), IDEALENS flags 4.9% of LLM paraphrases of human documents, compared to 42.8% by Pangram 4 and 44.5% by PROSELENS. In the Academic Integrity dataset (Karr et al., 2026), it flags 8.2% of abstracts that an LLM generated from a human-written reference abstract, compared to Pangram 4's 96.4% and PROSELENS's 83.9%.

IDEALENS detects AI ideas in TWICETOLD's human-written prose. IDEALENS flags 68% (95% CI 54–79%) of TWICETOLD stories as AI, compared to 0% for PROSELENS and 8% for Pangram 4.[12] All detectors are reliable at flagging the AI-generated original stories (96% for IDEALENS and 100% for PROSELENS and Pangram), but only IDEALENS sticks with its prediction for the majority of the human stories, even though the prose is completely different (Figure 4b). The stories that IDEALENS marks as human generally follow their outlines while adding new events or world-building details: 12 out of 16 errors are on stories that contain speculative or fantasy elements.

From ideas alone, IDEALENS outperforms the strongest existing open detectors. On 19 evaluation splits with shared provenance, IDEALENS detects 91.1% of AI documents at 0.6% FPR compared with 81.8% at 0.2% FPR for EditLens-3B,[13] the strongest existing open baseline (Table 21), with both calibrated to a 1% target. Training on full text does better: PROSELENS reaches 98.9% at 0.7% FPR, rivaling Pangram 4 (97.3% at 0.3%) and making it the strongest open detector in our evaluations. To cleanly measure FPR, we collect 10K pre-ChatGPT documents from C4: IDEALENS flags only 0.01% compared to 0% for PROSELENS and Pangram 4 (§I.1). Against humanization and adversarial attacks, IDEALENS's detection rate drops 6.4 points on average, between Pangram's 4.9 and PROSELENS's 10.6 (Table 31). One weakness of IDEALENS is DIPPER paraphrasing (Krishna

[12] Pangram 4's flags are likely due to usage of character names typical of AI writing (e.g., *Mara*, *Bellweather*).

[13] While IDEALENS's 30B backbone is larger than EditLens-3B, it is a mixture-of-experts model with only 3B active parameters. Our IDEALENS-ModernBERT variant, with only 0.4B parameters, detects 79.1% of AI documents at an FPR of 1.5%.

Table 1: **Examples of items for four core differences between the human and AI ideas.** The first column describes each contrast and summarizes how it is realized in **human** and **AI** ideas.

DIFFERENCE	HUMAN IDEA	AI IDEA
❶ **Decisive vs. Universal Details** Details could alter the author's point or conclusion. Details could apply to any similar subject.	The bottom plate of the Sony Cyber-shot RX100 contains a mini-HDMI output alongside a shared compartment for the power cell and SD/Memory Stick media—a hatch that becomes inaccessible during tripod use. (. . .)	Physical attributes of the Jony Ive speaker include a resilient, pared-down aesthetic, premium audio fidelity, low power consumption, and control mechanisms driven by straightforward tactile and vocal inputs.
❷ **Functional vs. Apparent Specificity** Numbers have comparisons, and instructions can be followed. Methods or numbers look specific without being usable.	Over a period of merely ten days, the shop generates 700 yuan in profit from selling athletic gear. This amount exceeds twice the highest anticipated expense for the prizes, demonstrating that (. . .)	Development teams are advised to handle technical obstacles preemptively by engaging in thorough initial preparation, continuous education, risk reduction strategies, Agile methodologies, strict quality testing, and discussions with domain specialists.
❸ **Bounded vs. Overextended Claims** Claims state clear limits on when they apply. Claims extend beyond evidence to uncertain futures.	The initial deployment of the new fleet in passenger service is anticipated for mid-2022, originating from the Montreal Maintenance Centre. (. . .)	The 2025 Kia K4 is projected to establish a permanent legacy as an updated standard for the category, demonstrating that (. . .)
❹ **Preserved vs. Erased Costs** Existing costs, limitations, and unresolved problems are retained. Costs vanish or become ideal trade-offs.	The writer intends to work toward the Schutzhund 1 certification for Quasi during the upcoming summer, limiting advancement to that initial tier due to the logistical challenge of a 250-mile journey to the closest training facility, (. . .)	The overall assessment concludes that the M3 MacBook Air represents the definitive lightweight notebook, providing an ideal equilibrium among computational strength, electrical endurance, structural excellence, and operating system cohesion.

et al., 2023) in GEDE (Gehring & Paaßen, 2025), where it detects 77.6% versus 97.1% for PROSELENS and 95.1% for Pangram 4; we suspect this attack changes content as well as style (§I.3).

IDEALENS detects idea provenance across 24 languages on IDEASHIFT-X. IDEASHIFT-X contains documents generated in 24 different languages, and IDEALENS flags 95.3% of the level-0 documents but only 0.7% of the level-5 documents despite being trained only on English. These results show that signatures of AI ideas are largely invariant to the language in which they are written. In contrast, prose detectors like PROSELENS and Pangram 4 flag 82–85% of level-0 and 45–62% of level-5 documents, and this gap between IDEALENS and these detectors widens in low-resource languages like Tamil and Amharic. Additionally, IDEALENS achieves a low FPR on non-English human-written documents: on 10K pre-ChatGPT documents in these 24 languages sampled from FineWeb2 (Penedo et al., 2025), it flags only 0.1% as AI (Table 29).

IDEALENS transfers unexpectedly well to raw text at test time. While trained on outlines, IDEALENS also works on raw documents during inference. On shared provenance, it detects 96.1% of AI documents at a 0.3% realized FPR, nearly matching PROSELENS and Pangram 4. Thus, features learned from outlines are also present in documents and even augmented with additional details not present in the outline. On the 12 benchmarks with human ideas and AI prose, IDEALENS given raw text still labels 89.6% of documents as human, compared to 94.7% for IDEALENS with outlines and 50.7% for PROSELENS (Table 19). However, it struggles with TWICETOLD's AI ideas and human writing, detecting only 6% of the stories as AI. We hypothesize that the raw stories expose more of the human author's added details, which IDEALENS associates with human ideas.

5 WHAT SEPARATES AI IDEAS FROM HUMAN IDEAS?

Human writing can differ from AI writing both in the ideas that they express and in how those ideas are organized. To disentangle the two, we first conduct an ablation (Table 36) that compares IDEALENS's AUROC on our in-domain test set given outlines with role labels removed (0.993), shuffled outlines (0.993), and role labels alone (0.601), showing that IDEALENS relies primarily on content. To characterize this content, we follow the methodology of Russell et al. (2026) to extract 79 interpretable features[14] from 90K item-level predictions, most of which are predictive

[14]Examples of features are provided in Table 49 and Table 50.

of IDEALENS's scores on held-out items.[15] Table 1 synthesizes these features into a short list of systematic differences that IDEALENS uses to separate human and AI ideas.

AI ideas use details that could apply to anything. Human ideas contain details that are crucial to the author's point: the human camera review in the first row of Table 1 points out a problem during tripod use, while an AI speaker review mentions "premium audio fidelity" that could apply to many other speakers. Going beyond such vagueness, which has also been noted in prior work (Kim et al., 2026; Russell et al., 2026), we find that AI ideas can also *look* specific without being usable: the AI idea in the second row advises development teams to handle obstacles via "risk reduction strategies" and "Agile methodologies" but never says how to carry these out, whereas the human counterpart reveals specifically which baseline amount is more than doubled.

AI ideas overstate claims and hide downsides. Human ideas generally state the limitations of their claims, while AI ideas extend their claims beyond the evidence: the AI idea in the third row of Table 1 predicts that a newly released Kia model will "establish a permanent legacy", whereas the human idea provides a hedged projection of when a fleet deployment will occur. AI ideas also omit limitations: the AI review in the fourth row describes the MacBook as at an "ideal equilibrium" without any downsides, while human writers explicitly state the limitations that constrain them.[16]

6 RELATED WORK

Detecting AI and mixed authorship. Standard AI text detectors (Mitchell et al., 2023; Hu et al., 2023; Hans et al., 2024; Emi & Spero, 2024), including several of our baselines, perform binary classification of prose authorship. Existing work on mixed authorship settings either focuses on AI-edited text, including polishing and paraphrasing (Zhang et al., 2024; Ta et al., 2026; Saha & Feizi, 2025; Bsharat et al., 2026) or attempts to explicitly separate content from expression (Bao et al., 2025; Chen et al., 2026a;b; Sadeghian et al., 2026). Other work aims to log the writing process (Krebs, 2026; Siddiqui et al., 2026) and theoretically examines how provenance should be defined (Schnell, 2026). In contrast, IDEALENS learns idea provenance from prose-level labels, and we ground our formalization of provenance in cognitive theories of writing (Flower & Hayes, 1981; Hayes, 2012) rather than in the text itself or formal semantics (Karttunen, 1969).

Learning from transformed inputs. Transforming training inputs can change which features a model learns without changing its supervision. Geirhos et al. (2019) suppress texture cues through image stylization, while Ilyas et al. (2019) show that training on a dataset constructed to emphasize robust features yields adversarial robustness. Krishna et al. (2020) use paraphrasing to normalize stylistic variation, enabling style transfer through a content-preserving intermediate representation, and Stamatatos (2017) masks content words so that authorship classifiers must rely on style. We apply the same principle in the opposite direction by keeping the labels of a prose detector but replacing its inputs with paraphrased outlines, so that the labels can only be fit through the ideas.

7 CONCLUSION

We show that AI writing and human writing can be meaningfully distinguished from one another through their ideas alone. Using a corpus of 1M outlines, we train IDEALENS on silver labels from a prose detector and benchmark it across 22 datasets. IDEALENS tracks idea provenance when it differs from prose provenance, unlike a control model (PROSELENS) that is trained on the same labels but raw text instead of outlines. IDEALENS transfers to 24 languages despite being trained only on English, and it outperforms the strongest existing open detector on documents where provenance is shared. IDEALENS has limitations (§A): its training labels noisily approximate idea provenance, its binary formulation does not cleanly handle documents whose ideas came from both humans and AI, it is weaker on short texts and some adversarial perturbations, its outline extraction process relies on proprietary models, and our TWICETOLD dataset covers only 50 stories. We release our models and datasets to support future research on idea provenance detection.

[15] On 13,616 held-out items, 64 of the 79 features align with IDEALENS's scores in their proposed direction (AI vs. human), while only one shows a conflict. Detailed analysis is presented in §J.3.

[16] The alignment and scores of features supporting each core difference are presented in Table 44.

AI USE STATEMENT

We used AI tools for the following tasks:

- **Prompt engineering:** We used coding agents to aid in constructing prompts for various tasks grounded in descriptions and principles we laid down.
- **Dataset construction:** We prompt a frontier LLM to automatically extract outlines from documents that we then use to train our models. We also prompt an LLM to construct some synthetic evaluation data.
- **Design or provide feedback on research methodology or experiments:** Coding agents were asked to provide feedback many times throughout the project, as well as help design and run the training of all models.
- **Implement methods:** Coding agents implemented most of the experimental code.
- **Support qualitative and thematic data analysis:** We often asked coding agents to create tables and figures of results so we could interpret the results. Many of these figures and tables ended up in the final draft.

Additional disclosures include the use of AI for the following purposes: create or modify scientific figures or images, create or edit software code, creation of artifacts, brainstorming, sourcing/searching for information, edit a research paper to improve readability, and identify relevant literature.

We have reviewed and validated all AI-assisted work and take responsibility for the final content of the work, including text, claims, and artifacts produced with the aid of AI.

ETHICS STATEMENT

Our work is a methodological contribution and is not intended to be used for invalidating human or AI-assisted writing. All datasets and non-commercial models are publicly available and the original license permissions were followed. This work includes a human-curated dataset, but participants provided consent and no sensitive information was asked.

REPRODUCIBILITY STATEMENT

We plan to release all data and trained models after acceptance along with scripts to train and evaluate our models. For this submission, we release an anonymized repository containing prompts, role vocabularies, evaluation results, and scripts to produce them here: `https://anonymous.4open.science/r/IdeaLens-0F91`.

REFERENCES

Anonymous. Scaling laws for wild AI text in pretraining data, 2026. Submitted to ICLR 2027.

Anthropic. System card: Claude Fable 5 & Claude Mythos 5. `https://www-cdn.anthropic.com/d00db56fa754a1b115b6dd7cb2e3c342ee809620.pdf`, June 2026a. Accessed: 2026-09-23.

Anthropic. System card: Claude Fable 5.1 & Claude Mythos 5.1. `https://www-cdn.anthropic.com/0339e6a7c5c7b87f5c07798616dc32c215d14235/Claude%20Fable%205.1%20&%20Claude%20Mythos%205.1%20System%20Card.pdf`, September 2026b. Accessed: 2026-09-23.

Anthropic. System card: Claude Opus 5. `https://www-cdn.anthropic.com/c5fbac3f0b1280a933ebd26d3cb8bb9f5bdeaf48/Claude%20Opus%205%20System%20Card.pdf`, July 2026c. Accessed: 2026-09-23.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 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`.

Guangsheng Bao, Lihua Rong, Yanbin Zhao, Qiji Zhou, and Yue Zhang. Decoupling content and expression: Two-dimensional detection of AI-generated text. *ArXiv*, abs/2503.00258, 2025. URL https://api.semanticscholar.org/CorpusID:276742379.

Connor Baumler, Calvin Bao, Huy Nghiem, Xinchen Yang, Marine Carpuat, and Hal Daumé III. Can you make it sound like you? Post-editing LLM-generated text for personal style. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43867–43895, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2030. URL https://aclanthology.org/2026.acl-long.2030/.

S. Mahmoud Bsharat, Jiachen Liu, Xiaohan Zhao, Tianjun Yao, Xinyi Shang, Yi Tang, Jiacheng Cui, Ahmed Elhagry, Salwa K. Al Khatib, Hao Li, Salman H. Khan, and Zhiqiang Shen. 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.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 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.

Xin Chen, Fuwei Zhang, Yiqi Tong, Wei Guo, Yutian Xiao, and Fuzhen Zhuang. D2C-Routing: Dimension-to-composition evidence routing for mixed-origin AI-generated text detection, 2026a. URL https://arxiv.org/abs/2608.27380.

Yihan Chen, Jiawei Chen, Guozhao Mo, Xuanang Chen, Ben He, Xianpei Han, and Le Sun. CoCoNUTS: Concentrating on content while neglecting uninformative textual styles for AI-generated peer review detection. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (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 2026b. 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/.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 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.

Desklib. AI text detector v1.01. https://huggingface.co/desklib/ai-text-detector-v1.01, 2025. Hugging Face model repository.

Stanley F. Druckenmiller. Let the bond market speak. *The Wall Street Journal*, August 24, 2026. URL https://www.wsj.com/opinion/let-the-bond-market-speak-81529d74. Accessed: 2026-09-24.

Andre Vicente Duarte, Brian Tufts, Aditya Oke, Fei Fang, Arlindo L. Oliveira, and Lei Li. SemDetect: Semantic level detection of AI generated peer-reviews. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3S9eVockjL.

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

Linda Flower and John R. Hayes. 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.

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

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In Marta R. Costa-jussà and Enrique Alfonseca (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/.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Byqh9j09KX.

Ben Glickenhaus, Katherine Thai, Jenna Russell, Elyas Masrour, Yue Han, Max Spero, and Bradley Emi. Pangram 4 technical report, 2026. URL https://arxiv.org/abs/2607.27183.

Google DeepMind. Gemini 3.1 Pro. https://deepmind.google/models/gemini/pro/, August 2026a. Model card.

Google DeepMind. Gemini 3.6 Flash. https://deepmind.google/models/model-cards/gemini-3-6-flash/, July 2026b. Model card.

Google DeepMind. Gemini 3.7 Flash. https://deepmind.google/models/model-cards/gemini-3-7-flash/, August 2026c. Model card.

Google DeepMind. Gemini 3.8 Flash. https://deepmind.google/models/model-cards/gemini-3-8-flash/, September 2026d. Model card.

Alicia Guo, Shreya Sathyanarayanan, Leijie Wang, Jeffrey Heer, and Amy X. Zhang. From pen to prompt: How creative writers integrate AI into their writing practice. In *Proceedings of the 2025 Conference on Creativity and Cognition*, C&C '25, pp. 527–545, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712890. doi: 10.1145/3698061.3726910. URL https://doi.org/10.1145/3698061.3726910.

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

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 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.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. RADAR: Robust AI-text detection via adversarial learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (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.

ICLR 2027 Program Chairs. Policy on LLM use for research and writing papers. https://iclr.cc/Conferences/2027/AIPolicyForAuthors, 2026. Accessed September 2026.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html.

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

Jonathan A. Karr, Jr, Grigorii Khvatskii, Ting Hua, and Nitesh V. Chawla. Why AI detection fails for academic integrity, 2026. URL https://arxiv.org/abs/2608.11256.

Lauri Karttunen. Discourse referents. In *Proceedings of the 1969 Conference on Computational Linguistics*, COLING '69, pp. 1–38, USA, 1969. Association for Computational Linguistics. doi: 10.3115/990403.990490. URL https://doi.org/10.3115/990403.990490.

Yekyung Kim, Yapei Chang, Chau Minh Pham, and Mohit Iyyer. 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.

Florian Krebs. F(AI)2R: Who did what, and who checked? Verifiable AI provenance as an executable skill, 2026. URL https://arxiv.org/abs/2607.25637.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Thousand Oaks, CA, 4th edition, 2018.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 737–762, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.55. URL https://aclanthology.org/2020.emnlp-main.55.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Frederick Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=WbFhFvjjKj.

Jaeho Lee. 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.

Chenjun Li, Cheng Wan, and Johannes C. Paetzold. MELD: Multi-task equilibrated learning detector for AI-generated text, 2026a. URL https://arxiv.org/abs/2605.06903.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. MAGE: Machine-generated text detection in the wild. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36–53, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.3. URL https://aclanthology.org/2024.acl-long.3/.

Yang Li, Qiang Sheng, Zhengjia Wang, Yehan Yang, Danding Wang, and Juan Cao. Beyond the final actor: Modeling the dual roles of creator and editor for fine-grained LLM-generated text detection. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5188–5203, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.235. URL https://aclanthology.org/2026.acl-long.235/.

William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 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.

Scott Nover. Wall Street Journal defends billionaire's use of AI to write op-ed. *The Washington Post*, August 25, 2026. URL https://www.washingtonpost.com/business/2026/08/25/wall-street-journal-says-ai-generated-op-ed-didnt-breach-its-standards/. Accessed: 2026-09-25.

NVIDIA. NVIDIA Nemotron 3.5 Lightning delivers fast, accurate specialized task execution for long-running agents. https://developer.nvidia.com/blog/nvidia-nemotron-3-5-lightning-delivers-fast-accurate-specialized-task-execution-for-long-running-agents/, 2026. NVIDIA Developer Blog.

OpenAI. New embedding models and API updates. https://openai.com/index/new-embedding-models-and-api-updates/, 2024. Accessed: 2026-09-21.

OpenAI. GPT-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf, July 2026a. Accessed: 2026-09-23.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. FineWeb2: One pipeline to scale them all — adapting pre-training data processing to every language. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=jnRBe6zatP.

Mike Perkins, Jasper Roe, Binh H. Vu, Darius Postma, Don Hickerson, James McGaughran, and Huy Q. Khuat. 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.

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

Chau Minh Pham, Alexander Miserlis Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 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.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 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.

Jenna Russell, Rishanth Rajendhran, Chau Minh Pham, Mohit Iyyer, and John Wieting. StoryScope: Investigating idiosyncrasies in AI fiction. In *Proceedings of the Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2604.03136.

Soroush Sadeghian, Alireza Daghighfarsoodeh, Radin Cheraghi, Sajad Ebrahimi, Negar Arabzadeh, and Ebrahim Bagheri. PeerPrism: Peer evaluation expertise vs review-writing AI. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3381–3387, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3808602. URL https://doi.org/10.1145/3805712.3808602.

Rounak Saha, Gurusha Juneja, Dayita Chaudhuri, Naveeja Sajeevan, Nihar B Shah, and Danish Pruthi. 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.

Shoumik Saha and Soheil Feizi. Almost AI, almost human: The challenge of detecting AI-polished writing. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25414–25431, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1303. URL https://aclanthology.org/2025.findings-acl.1303/.

Santiago Schnell. Who wrote this? Turing, total variation, and the mathematics of AI-text detection, 2026. URL https://arxiv.org/abs/2609.13660.

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

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 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.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. 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.

Momin N. Siddiqui, Nikki Nasseri, Adam J. Coscia, Roy Pea, and Hari Subramonyam. DraftMarks: Enhancing transparency in human-AI co-writing through interactive skeuomorphic process traces. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791109. URL https://doi.org/10.1145/3772318.3791109.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release strategies and the social impacts of language models, 2019. URL https://arxiv.org/abs/1908.09203.

Springer Nature. Artificial intelligence (AI) editorial policy. https://www.nature.com/nature-portfolio/editorial-policies/ai, 2026. Accessed September 2026.

Efstathios Stamatatos. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1138–1149, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1107/.

StealthGPT. StealthGPT. https://www.stealthgpt.ai, 2026. Accessed 2026-09-24.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12395–12412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.827. URL https://aclanthology.org/2023.findings-emnlp.827/.

Kun Sun, Rong Wang, and Wenxin Xiong. Investigating genre distinctions through discourse distance and discourse network. *Corpus Linguistics and Linguistic Theory*, 17(3):599–624, 2021. doi: 10.1515/cllt-2020-0064. URL https://doi.org/10.1515/cllt-2020-0064.

John M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge Applied Linguistics. Cambridge University Press, Cambridge, 1990.

Minh Ngoc Ta, Dong Cao Van, Duc-Anh Hoang, Minh Le-Anh, Truong Nguyen, My Anh Tran Nguyen, Yuxia Wang, Preslav Nakov, and Dinh Viet Sang. FAID: Fine-grained AI-generated text detection using multi-task auxiliary and multi-level contrastive learning. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3275–3296, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.151. URL https://aclanthology.org/2026.eacl-long.151/.

Maite Taboada and William C. Mann. Applications of rhetorical structure theory. *Discourse Studies*, 8(4):567–588, 2006. doi: 10.1177/1461445606064836. URL https://doi.org/10.1177/1461445606064836.

Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. EditLens: Quantifying the extent of AI editing in text. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gOkitaPCfZ.

Thinking Machines Lab. Tinker, 2026. URL https://thinkingmachines.ai/tinker/.

Teun A. van Dijk. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

Veniamin Veselovsky, Manoel Horta Ribeiro, Philip J. Cozzolino, Andrew Gordon, David Rothschild, and Robert West. Prevalence and prevention of large language model use in crowd work. *Commun. ACM*, 68(3):42–47, February 2025. ISSN 0001-0782. doi: 10.1145/3685527. URL https://doi.org/10.1145/3685527.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=boSqwdvJVC.

Junchao Wu, Yefeng Liu, Chenyu Zhu, Hao Zhang, Zeyu Wu, Tianqi Shi, Yichao Du, Longyue Wang, Weihua Luo, Jinsong Su, and Derek F. Wong. DetectRL-X: Towards reliable multilingual and real-world LLM-generated text detection. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38247–38294, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1773. URL https://aclanthology.org/2026.acl-long.1773/.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an LLM? Benchmarking AI text detection in peer review. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 61517–61553, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/64592277a1cbb02ed6e26b6e9e5f458c-Paper-Conference.pdf.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp.

409–436, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.29. URL https://aclanthology.org/2024.findings-naacl.29/.

APPENDIX CONTENTS

A Limitations **18**

B Labels and quadrants **19**

C Benchmarks **19**

D Data pipeline **21**

D.1 Role discovery 21

D.2 Outline extraction benchmark 21

D.3 Outline statistics 23

D.4 Paraphrasing removes the surface signal 23

D.5 A worked example 24

D.6 Corpus 26

E Training, calibration and thresholds **26**

E.1 Model configurations 26

E.2 Calibration 26

E.3 Thresholds and baselines 27

F Cost **27**

G Results by quadrant **28**

G.1 Human ideas, human prose 29

G.2 Model ideas, model prose 30

G.3 Human ideas, model prose 31

G.4 Model ideas, human prose 33

G.5 All detectors and baselines 33

G.6 Mixed provenance 34

G.7 OpAI-Bench: successive model edits of a human draft 34

H The evaluation sets we built **35**

H.1 IDEASHIFT 35

H.2 TWICETOLD 36

H.3 IDEASHIFT-X 37

H.4 Translation 38

I Robustness and generalization **39**

I.1 Pre-ChatGPT web text . . . 39
I.2 Attacks and humanizers . . . 39
I.3 Do attacks preserve content? . . . 40
I.4 Content-preserving transformations . . . 41
I.5 Format classification and extraction variance . . . 42
I.6 Newer generators . . . 43

J Feature Analysis Details **44**
J.1 Feature discovery . . . 44
J.2 Feature assignment . . . 44
J.3 Verifying the legitimacy of our discovered features . . . 44

K Prompts **49**
K.1 Outline pipeline . . . 49
K.2 Evaluation construction . . . 50
K.3 Feature Analysis . . . 50
K.4 TWICETOLD Curation . . . 51

L Where the remaining tables are **51**

Every table of every evaluation (all arms, every operating point, per-format cuts, the other model families and the length splits), all prompts and role vocabularies, and the code that produced the numbers are in the anonymous repository: `https://anonymous.4open.science/r/IdeaLens-0F91`.

A LIMITATIONS

Noisy training labels. We have no idea-authorship labels at scale, so we train on Pangram 3.3.2's prose-authorship labels and assume that a document's ideas and prose usually share an author. We keep only documents that Pangram labels fully human or fully AI, which excludes the AI-assisted cases where this assumption is weakest, but some labels are still wrong for our task: a person may have written up a model's ideas, or a model may have polished a person's. IDEALENS nonetheless departs sharply from Pangram wherever the two kinds of authorship differ (§4). It is still less accurate on mixed-authorship documents than on conventional ones, with a higher FPR on human ideas in model prose and a lower TPR on model ideas in human prose. We do not measure how much of this gap comes from label noise, nor test ways to reduce it.

Long-form English documents only. Our corpus contains only English documents of at least 500 words in eight long-form WebOrganizer formats, chosen so that each document carries enough ideas to detect. IDEALENS still works on shorter documents but is weaker on them in the conventional quadrants, mostly at detecting AI text (87.5% vs. 90.4% TPR; Table 2). Non-English documents reach IDEALENS only through English outlines, so its multilingual performance depends on the extractor (§H.3).

Dependence on a proprietary extractor. At inference, Gemini 3.7 Flash classifies the format and extracts the outline, at about $0.018 per document in bulk (§F). Swapping in other extractors barely changes IDEALENS's verdicts (Table 36), but we tested only proprietary frontier models, whose cost and availability may limit adoption. Reading the raw document instead avoids the extractor but loses much of the sensitivity to idea authorship (§4). Our released corpus pairs 1M documents with frontier-model outlines and could be used to train an open extractor; we leave this to future work.

Table 2: **Accuracy by document length** on the out-of-domain evaluations (%, global 1% cut). The training corpus has no documents under 500 words.

Quadrant	Rate	<500 words	≥500 words
Q1: human ideas, human prose	FPR	0.7	0.2
Q3: model ideas, model prose	TPR	87.5	90.4
Q4: human ideas, model prose	FPR	4.2	5.2

Extraction variance. Extraction is stochastic. Across five extractions of the same document, IDEALENS's verdict is unchanged for 94% of randomly drawn out-of-domain documents but for only 52.5% of those near the decision threshold (§I.5). Averaging several extractions would likely reduce this variance; we do not test it.

Newer generators. On stories from four models released after our corpus's cutoff, IDEALENS detects 51–79% at the global cut, against 97% for the older StoryScope generators (§I.6). PROSELENS also drops on the two Claude models, so the drop is not specific to reading ideas, but Pangram 4 does not drop. Keeping IDEALENS current will require training on newer generators.

Content-changing attacks. IDEALENS is less robust than Pangram 4 to some paraphrasing attacks: it detects 77.6% of DIPPER paraphrases in GEDE, against 95.1% for Pangram 4 (§4). Some of these attacks change content as well as style (§I.3), so an idea detector should arguably be less certain about them, but we cannot fully separate that from a genuine weakness.

Little evaluation data with model ideas in human prose. Existing benchmarks either lack this quadrant or approximate it with human edits of model drafts that keep most of the model's prose, so our only test is the 50 stories of TWICETOLD. The results there favor IDEALENS over prose detectors, but 50 stories cannot establish how well this generalizes; a larger dataset is future work.

Binary labels for mixed ideas. IDEALENS predicts human or AI, but in the collaborative writing that motivates this work, ideas are often part human and part AI. Whose idea a given point was can sometimes be traced, for example from chat logs, but when people and models revise each other's ideas repeatedly, the granularity at which to assign authorship is unclear. Modeling mixed authorship needs data with controlled idea sources, which we leave to future work.

B LABELS AND QUADRANTS

We label each document by who conceived its ideas, not who wrote its prose: a document with a person's ideas is human however much of its prose a model wrote, and vice versa. Crossing the two axes gives four quadrants:

- **Shared authorship**, as in conventional detection benchmarks: **HH** (Q1), human ideas and human prose, and **AA** (Q3), model ideas and model prose;
- **Mixed authorship**: **HA** (Q4), human ideas in model prose (e.g., a model polishing a person's text), and **AH** (Q2), model ideas in human prose (e.g., a person writing up a model's ideas).

Accuracy is the share of documents whose predicted label matches the source of their ideas: TPR when the ideas are a model's and $100 - \text{FPR}$ when they are a person's. Arms with genuinely shared ideas, such as IDEASHIFT levels 2–4 or a model imitating an author from a few passages, have no correct binary label, so we report their mean P(AI) instead (§G.6). Table 3 maps every arm to its label.

C BENCHMARKS

We evaluate on 19 external benchmarks and on four sets we build or sample ourselves (IDEASHIFT, IDEASHIFT-X, TWICETOLD and a C4 sample). We use every external benchmark as released, with

three exceptions: we keep the English rows of DetectRL-X and HART, which also carry other languages; we score MELD on a 14,713-document sample of its 220,136 documents; and we leave out two further benchmarks we ran, PeerPrism (Sadeghian et al., 2026) and a personal-style post-editing set (Baumler et al., 2026), because of data-quality problems (their tables are in the repository). Before scoring, every document goes through the same format gate and outline extractor as at test time (§2). Table 4 lists what each benchmark contributes.

Table 3: **Every arm we evaluate and its label under the idea-authorship rule.** "Attacked" marks model-written text that was humanized or adversarially rewritten; its ideas stay the model's. "Mixed" arms are reported as P(AI) only; "not placed" arms keep most of a model's prose after a person's edit and are left out of the quadrants.

Benchmark	Arm	Ideas by	Prose by	Label
AI Peer Review Detection	human reviews / model reviews	person / model	person / model	HH / AA
AI in Peer Review (Saha et al.)	H: human review	person	person	HH
	H-AI: human review polished by a model (and humanized)	person	model	HA
	AI-BP, AI-EP: model review from the paper (and humanized)	model	model	AA (attacked)
	AI-HI: model review from the reviewer's notes	mixed	model	mixed
DetectRL-X	human / model / attacked model text (English rows)	person / model / model	person / model / model	HH / AA / AA attacked
StoryScope	human stories / stories by five models	person / model	person / model	HH / AA
DetectionAI	human essays / model essays / humanized (StealthGPT, 2026)	person / model / model	person / model / model	HH / AA / AA attacked
OpAI-Bench	v0: the human draft / v1–v8: successive model edits	person	person / model	HH / HA
Sem-Detect	human reviews / model rewrite of a human review / model reviews	person / person / model	person / model / model	HH / HA / AA
MELD	human / model / six attacks	person / model / model	person / model / model	HH / AA / AA attacked
LAMP	model draft / the draft after a professional's edit	model	model / model, edited	AA / not placed
ARB	human / model rewrite of a human article (h2l) / model (free_llm) / llm2l	person / person / model / model	person / model / model / model	HH / HA / AA / AA attacked
GEDE	human essay / model improves or rewrites it	person	person / model	HH / HA
	model essay from the assignment / DIPPER, rewrite attack	model	model	AA / AA attacked
	model essay from the student's summary	model (GEDE's own label)	model	AA, shown but not averaged
Perkins et al.	model text under six evasion techniques	model	model	AA attacked
HART	human / model rephrase of human text / model / humanized	person / person / model / model	person / model / model / model	HH / HA / AA / AA attacked
	model text rewritten by a person	model	model, edited	not placed
Epoch author imitation	the author's text / model, plain prompt	person / model	person / model	HH / AA
	model imitating the author from five passages	mixed	model	mixed
CoCoNUTS	human review / model polishes or translates it / model review	person / person / model	person / model / model	HH / HA / AA
	model review from the person's review	mixed	model	mixed
ELLIPSE	essays by English learners	person	person	HH
AI-Researcher	expert reviews / proposals from expert ideas / model proposals	person / person / model	person / model / model	HH / HA / AA
	papers people wrote from a model's idea	mixed	person	mixed
Academic Integrity	original abstract / model abstract from it (and humanized)	person	person / model	HH / HA
GEN	human / edited by an AI editor / written by an AI writer	person / person / model	person / model / model	HH / HA / AA
	an AI editor adds details to a person's text	mixed	model	mixed (repository only)
IDEASHIFT, human sources	level 0–1 / levels 2–3 / level 4 / level 5	model / mixed / mixed / person	model	AA / mixed / mixed / HA
IDEASHIFT, AI sources	levels 0–5	model	model	AA
IDEASHIFT-X	human pages / level 0 / level 5	person / model / person	person / model / model	HH / AA / HA
TWICETOLD	human rewrites / the model originals	model	person / model	AH / AA
C4 (2019)	web pages	person	person	HH

Table 4: **The 19 external benchmarks.** Documents scored in the arms we report, and the quadrants those arms fall in: HH human ideas and human prose, AA model ideas and model prose, HA human ideas and model prose, AH model ideas and human prose; "attacked" marks humanized or adversarially attacked model-written arms. Table 3 maps every arm to its label.

Benchmark	Domain	Documents	Quadrants
AI Peer Review Detection (Yu et al., 2026)	peer reviews	9,956	HH, AA
AI in Peer Review (Saha et al., 2026)	peer reviews	28,107	HH, AA, HA (+ attacked)
DetectRL-X (Wu et al., 2026)	academic, news, fiction, encyclopedia, web and SEO text (English rows)	19,916	HH, AA (+ attacked)
StoryScope (Russell et al., 2026)	short fiction	8,291	HH, AA
DetectionAI (Jabarian & Imas, 2025)	fiction, news, product and restaurant reviews, blogs, résumés	17,340	HH, AA (+ attacked)
OpAI-Bench (Bsharat et al., 2026)	essays and news articles, edited by successive model operations	37,308	HH, HA
Sem-Detect (Duarte et al., 2026)	peer reviews	22,146	HH, AA, HA
MELD (Li et al., 2026a)	news, books, Wikipedia, recipes, poetry, abstracts, Reddit, reviews	14,713	HH, AA (+ attacked)
LAMP (Chakrabarty et al., 2025)	fiction and creative nonfiction drafts	1,057	AA
ARB (Perrone & Romano, 2026)	web text, news and short stories	22,726	HH, AA, HA (+ attacked)
GEDE (Gehring & Paaßen, 2025)	student essays	10,037	HH, AA, HA (+ attacked)
Perkins et al. (Perkins et al., 2024)	academic writing, with evasion techniques	86	AA (+ attacked)
HART (Bao et al., 2025)	news, essays, creative writing, arXiv papers (English rows)	15,371	HH, AA, HA (+ attacked)
Epoch author imitation (Lee, 2026)	blog posts, fiction and scientific prose by named authors	792	HH, AA
CoCoNUTS (Chen et al., 2026b)	peer reviews	5,584	HH, AA, HA
ELLIPSE (Crossley et al., 2023)	essays by English learners	3,909	HH
AI-Researcher (Si et al., 2025; 2026)	research idea proposals, papers written from them, expert reviews	649	HH, AA, HA
Academic Integrity (Karr et al., 2026)	paper abstracts	2,312	HH, HA
GEN (AI Writers/Editors) (Shan et al., 2026)	Wikipedia, WikiHow, arXiv, Reddit posts and stories	21,187	HH, AA, HA

D DATA PIPELINE

This section documents how a web document becomes the outline IDEALENS reads: the discourse roles for each format, the extractor that writes the outline and how we chose it, what the outlines look like, and the paraphrase used for training.

D.1 ROLE DISCOVERY

For each format, GPT-5.6 Sol (OpenAI, 2026a) first proposes a base set of roles from a description of the format alone. It then labels every item of the outlines of 200 documents in that format (100 human, 100 model-written), inventing a role wherever none of the base roles fits, and finally consolidates the base and invented roles by merging duplicates and near-synonyms. Roles used by fewer than 1% of outline items in the format are dropped and replaced by a catch-all *Other* role (Table 5). Each role comes with a definition, a test for when to assign it, the roles it is distinguished from, and example excerpts; Table 6 shows eight of the 42 roles for nonfiction writing.

D.2 OUTLINE EXTRACTION BENCHMARK

We chose the corpus extractors with a small benchmark (Table 7). Gemini 3.1 Pro and Gemini 3.7 Flash, both with six few-shot examples, extracted the corpus; Gemini 3.7 Flash extracts every out-of-domain evaluation document.

Setup. The benchmark has 198 documents: eleven human-written and eleven AI-generated documents from each of eight formats, with authorship confirmed by both EditLens and Pangram 4. Reference outlines come from Claude Fable 5. We compare twelve configurations: GPT-5.6 Sol, Terra, and Luna zero-shot, and Gemini 3.1 Pro, 3.6 Flash (Google DeepMind, 2026b), and 3.7 Flash with zero, three, or six few-shot examples, each a full document with its reference outline drawn from a separate pool. The main score aligns candidate and reference items in order by embed-

Table 5: **Role discovery by format.** Base roles proposed from the format description alone; candidate roles after labelling the items of 200 documents (100 human, 100 model-written) and inventing roles where none fit; the final vocabulary after consolidation, split into roles kept from the base set and roles invented during labelling.

Format	Base roles	Candidates	Final roles	kept from base	invented
Academic Writing	44	67	52	42	10
Creative Writing	30	84	50	30	20
Knowledge Article	33	70	47	31	16
News Article	33	54	41	33	8
Nonfiction Writing	40	51	42	37	5
Personal About Page	22	269	62	22	40
Personal Blog	37	46	43	36	7
User Reviews	21	126	45	20	25

Table 6: **Eight of the 42 discourse roles discovered for nonfiction writing**, with the definitions the extractor receives. Each role also carries an assignment test, the roles it is distinguished from, and example excerpts; the full vocabularies of all eight formats are in the repository (the excerpts, which quote the training corpus, are released with the corpus).

Role	Definition
Explanation	Content that makes a phenomenon intelligible by identifying causes, mechanisms, processes, reasons, or functional relationships.
Recommendation	Content that advocates a course of action or decision as desirable, preferable, or necessary.
Evidence	Content that supplies verifiable facts, records, observations, data, or sourced information as grounds for assessing another proposition.
Event Account	Content that recounts what happened, including relevant participants, actions, chronology, and immediate consequences.
Illustration	Content that clarifies or exemplifies a broader point through a particular case, anecdote, scenario, or analogy.
Framing	Content that establishes the scope, lens, premise, governing concern, or terms through which other material is to be understood.
Synthesis	Content that integrates claims, evidence, events, perspectives, or themes into a higher-order understanding or takeaway.
Instruction	Content that provides actionable steps, methods, conditions, or cautions for carrying out a task or reaching a goal.

ding similarity, allowing merges, splits, omissions, and additions; the table caption defines the other metrics.

Results. Every configuration captures nearly all reference content (coverage ≥ 0.986); they differ mainly in granularity, that is, how finely they split content into items. Few-shot examples matter more than the choice of model: six examples raise alignment by 0.12–0.13 for every Gemini model. Gemini 3.1 Pro at six shots is within 0.03 of the best configuration (0.769 vs. 0.796 for GPT-5.6 Luna), and Gemini 3.7 Flash at six shots matches it (0.774) at about a third of the cost, making it the cheaper choice for the rest of the corpus.

Caveats. Scores measure agreement with a single reference model, not correctness. Few-shot examples also increase phrasing copied from the document (from 2.6% to 7.8% of item words for 3.7 Flash), which the paraphrasing step then removes (§D.4). Finally, no test directly swaps 3.1 Pro for 3.7 Flash; the extractor swap in Table 36 uses GPT-5.6 extractors.

Table 7: **Outline extraction benchmark.** Twelve extractors on 198 documents, scored against Claude Fable 5 reference outlines. We report (a) order-preserving match of outline items (*alignment*) (b) reference/candidate items matched (*coverage*) (c) one-to-one matches (*granularity agreement*) (d) candidate minus reference items per match (*split−merge*) (e) exact label match (*role agreement*) (f) cost per document with batching (g) mean rank (1 is best) and mean score from GPT-5.6 Sol, which judges eight configurations per document against the reference without seeing their source; a cross-check on the alignment score ("–": not judged). Corpus extractors are in bold.

		Coverage		Granularity					
Configuration	Alignment	reference	candidate	agreement	split−merge	Items (cand. / ref.)	Role agreement	$ per doc.	Judge rank / score
GPT-5.6 Luna, 0-shot	0.796	0.997	1.000	0.760	+0.121	19.7 / 17.9	0.630	0.022	2.04 / 96.3
GPT-5.6 Sol, 0-shot	0.782	0.988	1.000	0.778	+0.107	19.5 / 17.9	0.694	0.087	1.80 / 96.2
Gemini 3.7 Flash, 6-shot	**0.774**	0.998	0.999	0.787	−0.130	16.0 / 18.0	0.646	**0.017**	4.76 / 90.9
Gemini 3.1 Pro, 6-shot	**0.769**	0.997	0.999	0.743	−0.139	16.1 / 18.0	0.660	**0.046**	5.60 / 88.4
GPT-5.6 Terra, 0-shot	0.769	0.986	1.000	0.779	−0.142	15.7 / 17.9	0.632	0.031	3.52 / 92.8
Gemini 3.7 Flash, 3-shot	0.763	1.000	0.999	0.796	−0.094	16.4 / 18.0	0.659	0.013	4.79 / 90.8
Gemini 3.1 Pro, 3-shot	0.755	0.998	0.998	0.740	−0.146	16.2 / 18.0	0.655	0.036	5.85 / 87.6
Gemini 3.6 Flash, 6-shot	0.751	0.999	1.000	0.744	−0.220	14.9 / 18.0	0.618	0.033	–
Gemini 3.6 Flash, 3-shot	0.739	0.998	1.000	0.762	−0.175	15.5 / 18.0	0.621	0.026	–
Gemini 3.7 Flash, 0-shot	0.657	0.997	0.999	0.698	−0.348	13.5 / 18.0	0.599	0.007	7.65 / 81.5
Gemini 3.1 Pro, 0-shot	0.636	0.995	1.000	0.687	−0.258	14.7 / 18.0	0.623	0.019	–
Gemini 3.6 Flash, 0-shot	0.632	0.998	0.999	0.672	−0.363	13.4 / 18.0	0.579	0.013	–

D.3 Outline statistics

Table 8 describes the outlines of 1,000 training documents per format: how many items and distinct roles an outline has, how long its items are before and after paraphrasing, and how little of the extracted wording survives the paraphrase (4.2% of word 4-grams over all formats).

Table 8: **Outline statistics by format.** 1,000 training-corpus documents per format. Items per outline (median), distinct roles used per outline, mean words per item in the extracted and the paraphrased outline, and the share of the extracted outline's word 4-grams that survive paraphrasing (all formats: 4.2%).

			Words per item		
Format	Items	Roles used	extracted	paraphrased	4-grams kept (%)
Academic Writing	16	9.6	26.4	32.6	3.8
Creative Writing	14	9.6	24.0	29.0	1.9
Knowledge Article	14	7.8	27.0	33.3	3.8
News Article	12	7.7	28.5	34.6	9.1
Nonfiction Writing	13	8.4	27.0	33.0	3.4
Personal About Page	12	8.9	27.0	32.9	4.9
Personal Blog	13	8.4	25.3	30.6	3.1
User Reviews	15	7.9	25.1	31.6	3.5

D.4 Paraphrasing removes the surface signal

The paraphrase should remove wording that could reveal the source's author while keeping the ideas. *It removes the surface signal:* Pangram 4 scores 1,000 test documents as the document, the raw outline, and the paraphrased outline. On paraphrased outlines it labels almost every human document as AI, while IdeaLens still separates them (Table 9). *It keeps the content:* a paraphrased outline has cosine 0.93 with its raw outline (0.39 for another document's outline of the same label and format), and 99.2% of raw items are closest to their own paraphrase. Both outlines stay close to the document (0.79 raw, 0.76 paraphrased; 0.26 by chance), yet the paraphrase shares under 1% of its word 5-grams with it (Table 10).

Table 9: **Paraphrasing removes most of the surface signal.** Pangram 4 scores 1,000 held-out test documents three ways: as the document, the raw outline, and the paraphrased outline. We report the mean and median (shaded) of Pangram's human, AI-assisted, and AI fractions and of P(AI) = AI + $\frac{1}{2}$ AI-assisted, in percent. On paraphrased outlines, Pangram rates almost every human document as AI, while IDEALENS still separates them. AUC is against the Pangram 3.3.2 labels, so the document row is partly circular.

				human-labelled (n=500)				AI-labelled (n=500)			
Detector	Input	AUC		human	assisted	AI	P(AI)	human	assisted	AI	P(AI)
Pangram 4	document	0.935	mean	98.7	0.2	1.1	1.2	31.7	0.0	68.3	68.3
			median	100.0	0.0	0.0	0.0	12.1	0.0	87.9	87.9
Pangram 4	raw outline	0.878	mean	8.5	64.8	26.7	59.1	2.2	11.0	86.8	92.3
			median	0.0	76.7	7.7	50.1	0.0	0.0	100.0	100.0
Pangram 4	paraphrased outline	0.733	mean	0.3	29.7	70.0	84.8	0.1	3.7	96.3	98.1
			median	0.0	7.7	91.3	95.2	0.0	0.0	100.0	100.0
IDEALENS	paraphrased outline	0.995	mean	–	–	–	4.9	–	–	–	96.7
			median	–	–	–	0.0	–	–	–	99.7

Table 10: **Paraphrasing preserves meaning.** 1,000 held-out test documents (500 human-labelled, 500 AI-labelled), each with its raw extracted outline and its paraphrase, embedded with text-embedding-3-large; cosine between the raw and paraphrased units. † marks pairs of outlines, documents or items where paraphrases are of other raw units. The n-gram rows give the share of an outline's word n-grams found verbatim in the document.

Measure	**All**		**Human**		**AI**	
	mean	med	mean	med	mean	med
Outlines						
raw vs paraphrased	0.933	0.938	0.934	0.937	0.933	0.939
†raw vs paraphrased	0.389	0.392	0.360	0.363	0.418	0.421
Items						
raw vs paraphrased	0.834	0.837	0.845	0.851	0.822	0.826
†raw vs paraphrased (from same outline)	0.436	0.436	0.418	0.420	0.453	0.452
share of raw items whose closest paraphrased item is their own	99.2	100.0	99.6	100.0	98.8	100.0
Documents						
document vs raw outline	0.794	0.804	0.790	0.802	0.798	0.806
document vs paraphrased outline	0.758	0.765	0.754	0.761	0.762	0.770
†document vs raw outline	0.262	0.261	0.239	0.242	0.285	0.288
†document vs paraphrased outline	0.262	0.261	0.239	0.243	0.285	0.289
N-gram overlap with document (%)						
raw, 3-grams	12.3	11.0	10.3	9.2	14.2	13.1
paraphrased, 3-grams	3.5	2.8	3.6	3.0	3.3	2.7
raw outline, 5-grams	3.6	2.6	3.0	2.1	4.3	3.3
paraphrased, 5-grams	0.8	0.3	0.9	0.3	0.8	0.3
raw, 8-grams	0.9	0.2	0.8	0.2	1.1	0.4
paraphrased, 8-grams	0.3	0.0	0.3	0.0	0.3	0.0

D.5 A WORKED EXAMPLE

Table 11 gives the complete outline of the op-ed in Figure 2: the 17 items the extractor writes and their paraphrases. The paraphrase keeps each item's content and changes only its wording.

Table 11: **The outline of the op-ed in Figure 2.** All 17 items, as extracted (what IDEALENS reads at test time) and after paraphrasing (what it was trained on). The op-ed is classified as Nonfiction Writing.

#	Role	Extracted item	Paraphrased item
1	Event Account	The Treasury Department announced on Aug. 19 that it would double long-dated bond buybacks to at least $4 billion per operation after 30-year yields hit a 19-year high, triggering an immediate yield decline that quickly reversed to higher levels.	Following 30-year yields reaching a 19-year peak, the Treasury Department stated on Aug. 19 that the size of extended maturity repurchases would double, reaching a minimum of $4 billion each. This decision caused a sudden drop in yields, which soon rebounded to elevated positions.
2	Evaluation	The market correctly identified the expanded buyback operation as an attempt at price management rather than liquidity support, representing a significant policy error.	Investors accurately interpreted the enlarged repurchase initiative as a mechanism for controlling prices instead of providing liquidity, marking a major misstep in government strategy.
3	Rebuttal	Treasury's liquidity rationale is refuted by strong market participation and orderly trading, which lacked the genuine dysfunction seen during March 2020 or the 2022 UK gilt crisis.	The justification regarding liquidity provisions is undermined by stable trading conditions and robust investor involvement, demonstrating none of the severe disruptions witnessed during the 2022 UK gilt turmoil or in March 2020.
4	Evidence	Underlying economic data–including 3% to 4% inflation, 4.1% unemployment, a 6% peacetime deficit, $40 trillion in national debt, and interest costs exceeding defense spending–justifies higher long-term yields.	Elevated rates on extended maturities are warranted by fundamental economic indicators, encompassing a national debt of $40 trillion, a peacetime deficit of 6%, joblessness at 4.1%, inflation running between 3% and 4%, alongside debt servicing expenses that surpass military expenditures.
5	Interpretation	Current bond yield levels remain historically accommodative relative to nominal economic growth, indicating the market was not acting punitively but barely adjusting to fiscal realities.	When compared to nominal economic expansion, contemporary rates on government securities are still quite lenient by historical standards, suggesting that investors were merely adapting to budgetary facts rather than imposing punitive measures.
6	Explanation	Because politicians consistently avoid entitlement reform, the long-term bond yield serves as the only remaining mechanism capable of compelling democracies to confront unsustainable spending.	With elected officials routinely dodging structural changes to safety net programs, extended maturity rates function as the final available tool to force democratic nations into addressing unmanageable financial obligations.
7	Problem	Suppressing yields masks fiscal urgency and subsidizes political procrastination, guaranteeing that necessary spending adjustments will be postponed at far greater ultimate cost.	Artificially depressing rates conceals the severity of budgetary crises and rewards legislative delays, ensuring that essential expenditure modifications are deferred and ultimately become substantially more expensive.
8	Background	The historical precedent of 1942 to 1951 yield caps shows how debt management operations can slide into inflationary price control and prolonged financial repression.	Previous implementation of rate ceilings between 1942 and 1951 demonstrates that strategies for handling national obligations can easily deteriorate into extended periods of capital restriction and inflationary market manipulation.
9	Event Account	Treasury officials subsequently signaled that buyback operations could expand indefinitely and utilize the Treasury General Account, establishing a dangerous dynamic of testing official resolve against market pressures.	Representatives from the Treasury later indicated that repurchasing activities might grow without limits by drawing upon the Treasury General Account, creating a hazardous environment where investor forces continually challenge administrative determination.

Table 11, continued (items 10–17).

#	Role	Extracted item	Paraphrased item
10	Problem	Replacing long-term bonds with short-term bills acts as backdoor monetary easing during elevated inflation, while conducting interventions near election cycles jeopardizes the Treasury market's long-term credibility.	Exchanging extended duration securities for shorter obligations functions as a covert form of economic stimulation amid high inflation, and timing these actions close to voting periods threatens the enduring reliability of the Treasury ecosystem.
11	Perspective	The author highlights past warnings on fiscal policy, noting that federal transfer payments have grown to consume 70% of outlays and criticizing past failures to lock in low borrowing rates.	Emphasizing previous cautions regarding budgetary choices, the writer points out that government redistribution initiatives currently account for 70% of expenditures, while finding fault with earlier missed opportunities to secure inexpensive financing.
12	Stakes	Projected interest expenses reaching 4.5% of GDP guarantee that entitlement restructuring will eventually occur, either through deliberate political reform or abrupt market-imposed correction.	With anticipated financing costs approaching 4.5% of GDP, structural modifications to safety net programs are inevitable, arriving either via intentional legislative action or through sudden, investor-driven forces.
13	Rebuttal	The author rejects the defense of buybacks as routine tools, arguing that their irregular timing, doubled scale, and open-ended nature confirm they were deployed to suppress prices rather than resolve plumbing issues.	Dismissing claims that repurchase initiatives are standard administrative mechanisms, the writer contends that their unlimited scope, increased magnitude, and unusual scheduling prove these measures were utilized to depress valuations instead of addressing structural trading frictions.
14	Recommendation	Policymakers should restrict buybacks to regular liquidity operations, accept market-clearing yields, and address the primary deficit through phased, generational entitlement reforms.	Lawmakers are advised to limit repurchase activities to standard maintenance of market fluidity, tolerate rates dictated by investor demand, and resolve the primary deficit utilizing gradual, multigenerational adjustments to public benefit programs.
15	Claim	Implementing a credible fiscal reform package would lower long-term yields far more effectively than any scale of debt buybacks.	Executing an authentic budgetary overhaul would reduce extended borrowing rates much more successfully than repurchasing government obligations at any volume.
16	Synthesis	Concludes that government efforts to defend artificial prices against economic fundamentals inevitably fail, urging Washington to heed rather than suppress the bond market's pricing signals.	The piece finishes by asserting that state attempts to uphold unnatural valuations in defiance of core economic realities are destined to collapse, advising Washington to interpret rather than stifle the indicators provided by investor yield requirements.
17	Contributor Record	Biographical attribution identifying the author as chairman and CEO of Duquesne Family Office LLC.	A biographical note specifies that the writer serves as the CEO and chairman of Duquesne Family Office LLC.

D.6 CORPUS

The corpus is a 1,001,145-document sample of FineWeb, drawn from the WildAI release (Anonymous, 2026), which labels each document with a WebOrganizer format and topic and a Pangram 3.3.2 authorship verdict. We keep English documents of at least 500 words in eight long-form formats that Pangram labels entirely human or entirely AI, and split them into training, validation, calibration (human only) and test sets (Table 12).

Table 12: **The corpus**: 1,001,145 web documents in eight formats. Validation (29,975, checkpoint selection only) omitted. Calibration is human-only, 10,000 per format. Median words are for the train split.

Format	Train n	Train % AI	Calibration n	Test n	Test % AI	Median words
Academic Writing	71,656	58.7	10,000	6,264	50.0	1,051
Creative Writing	30,751	47.0	10,000	6,324	50.5	857
Knowledge Article	189,448	55.0	10,000	6,264	50.0	956
News Article	99,303	58.0	10,000	5,152	48.8	625
Nonfiction Writing	211,267	56.9	10,000	6,264	50.0	928
Personal About Page	32,082	41.8	10,000	5,966	48.0	652
Personal Blog	143,477	56.1	10,000	6,336	50.2	762
User Reviews	64,317	60.3	10,000	6,299	50.0	813
Total	842,301	56.0	80,000	48,869	49.7	823

E TRAINING, CALIBRATION AND THRESHOLDS

E.1 MODEL CONFIGURATIONS

Every model is trained on the 842,301-document training split with the Pangram labels. IDEALENS and PROSELENS share the recipe exactly and differ only in their input (paraphrased outline or raw document) and the one word of the system prompt that names it; the loss is on the answer token only. Besides the Nemotron backbone of the main paper we train ModernBERT-large (Warner et al., 2025) and Qwen3.5-9B (Qwen Team, 2026), and fit a logistic regression on text-embedding-3-large embeddings (OpenAI, 2024), on the same data (Table 13); their results on every quadrant are in Table 21. The neural models train for one epoch and are scored on validation documents at evenly spaced probes; we keep the checkpoint with the highest mean TPR at 0.5, 1 and 2% FPR. For IDEALENS and PROSELENS the probes use a stratified sample of 4,000 validation documents and the selected checkpoint is the final one. The ModernBERT and Qwen checkpoints are selected on the full 29,975-document validation split: the end-of-training checkpoint for both ModernBERT models, and step 12,056 of 13,161 (0.92 of the epoch) for Qwen. The logistic regression has no checkpoints.

Table 13: **Training configurations.** The neural models' learning rates follow a cosine schedule after the warm-up.

Model	Base	Method	Learning rate	Batch	Max length	Probes
IDEALENS / PROSELENS	Nemotron-3.5-Lightning-30B-A3B	LoRA rank 64, $\alpha = 32$ (attention, MLP, Mamba projections, unembedding), Tinker	4×10^{-4}, 3% warm-up	128 documents	65,536 tokens	25, and the final step
ModernBERT	ModernBERT-large	full fine-tuning, AdamW, weight decay 0.01	2×10^{-5}, 6% warm-up	16,384 tokens $\times$ 4 accumulation	8,192 tokens	10 (outline), 4 (document)
Qwen	Qwen3.5-9B	full fine-tuning, classification head	1×10^{-5}	64 documents	8,192 tokens	12
Logistic regression	text-embedding-3-large (3,072 dimensions)	L2-regularized, $C = 1$, L-BFGS	–	full training split	–	–

E.2 CALIBRATION

A detector outputs a probability; turning it into a verdict needs a threshold, and a threshold is only meaningful together with the false-positive rate it produces. We set ours on a held-out calibration split of 80,000 human documents, 10,000 per format, disjoint from the validation split used for checkpoint selection. The global cut at a target FPR α is the $(1 - \alpha)$ quantile of the calibration

scores pooled over formats; all results in the paper use $\alpha = 1\%$. A per-format cut, the same quantile within the document's format, is reported after a slash in some appendix tables. We fit a cut only where it rests on at least 25 expected false positives, so a global cut exists down to a 0.03% target and a per-format cut down to 0.25%; of our standard operating points (0.1, 0.5, 1, 2 and 5%), the per-format cut is therefore reported from 0.5%. A cut calibrated on in-domain humans need not hold its FPR out of domain; we therefore report the realized FPR on every evaluation's human arm instead of assuming it (§G).

E.3 Thresholds and baselines

Table 14 lists how every detector's score becomes a verdict. Our detectors, EditLens and the baselines that do not ship a threshold use a cut calibrated on our 80,000 human documents, so their FPR on in-domain human text is 1% by construction; detectors that ship a threshold keep it. Pangram 4 returns the fractions of a text it judges AI-written, AI-assisted and human; following its technical report (Glickenhaus et al., 2026), we count a text as AI when its AI fraction plus half its AI-assisted fraction is at least 0.5.

Table 14: **How each detector's scores become verdicts.** Calibrated cuts are the 1% quantile of the 80,000 human documents in our calibration split, pooled over formats (§E.2); the cut column gives its value on the detector's own score scale. Pangram 4 returns the fractions of a text it judges AI, AI-assisted and human; we weight AI-assisted text by one half.

Detector	Threshold	Cut
Our detectors: flagged as AI when P(human) falls below the cut		
IdeaLens (outline)	calibrated, global 1%	0.137
IdeaLens (document)	calibrated, global 1%	0.087
ProseLens	calibrated, global 1%	0.605
IdeaLens-ModernBERT-L (outline)	calibrated, global 1%	0.055
IdeaLens-ModernBERT-L (document)	calibrated, global 1%	0.046
ProseLens-ModernBERT-L	calibrated, global 1%	0.188
IdeaLens-Qwen3.5-9B (outline)	calibrated, global 1%	0.052
IdeaLens-Qwen3.5-9B (document)	calibrated, global 1%	0.016
IdeaLens-LogisticClassifier	calibrated, global 1%	0.060
Baselines without a published threshold: flagged as AI when the score exceeds the cut		
EditLens-Llama-3B (Thai et al., 2026)	calibrated, global 1%	0.828
EditLens-RoBERTa (Thai et al., 2026)	calibrated, global 1%	0.905
Likelihood (Gehrmann et al., 2019)	calibrated, global 1%	-1.33
Log-rank (Mitchell et al., 2023)	calibrated, global 1%	-0.573
Rank (Gehrmann et al., 2019)	calibrated, global 1%	-8.82
Entropy (Gehrmann et al., 2019)	calibrated, global 1%	-1.41
LRR (Su et al., 2023)	calibrated, global 1%	2.36
Baselines at their own published threshold		
Binoculars (Hans et al., 2024)	own	–
Fast-DetectGPT (Bao et al., 2024)	own	–
MELD (Li et al., 2026a)	own	–
Desklib (Desklib, 2025)	own	–
Desklib-academic (Desklib, 2025)	own	–
MAGE (Li et al., 2024)	own	–
OpenAI-RoBERTa-base (Solaiman et al., 2019)	own	–
OpenAI-RoBERTa-large (Solaiman et al., 2019)	own	–
RADAR (Hu et al., 2023)	own	–
Commercial		
Pangram 4 (Glickenhaus et al., 2026)	AI if AI $+\frac{1}{2}$ AI-assisted ≥ 0.5	–

F Cost

All figures are in US dollars at the rates in force in September 2026. Table 15 gives what we paid once, to build the corpus and train the two detectors; Table 16 gives what it costs to score a new document.

Building the corpus and training. Building the 1,001,145-document corpus cost $66,428 in batch API calls, about $0.07 per document: $40,255 for outline extraction (Gemini 3.7 Flash for 613k documents and Gemini 3.1 Pro for the 388k documents of an earlier build) and $26,173 for the

paraphrase (Gemini 3.1 Pro for all of them). These figures are computed from our token logs at the published batch rates; an invoice check on part of the run implies up to $78,437. Role discovery cost a further $116. Training on Tinker cost $354 for IDEALENS and $710 for PROSELENS, which reads whole documents and so trains on about twice as many tokens.

Table 15: **One-time costs** (US dollars). API calls at batch rates, computed from our token logs; Tinker training metered by the provider. An invoice check on part of the corpus run implies the two corpus rows may total up to $78,437.

Stage	Model	Scale	Cost
Role discovery	GPT-5.6 Sol	1,600 documents	$116
Outline extraction	Gemini 3.7 Flash (613k), 3.1 Pro (388k)	1,001,145 documents	$40,255
Outline paraphrase	Gemini 3.1 Pro	1,001,145 documents	$26,173
Training IDEALENS	Nemotron LoRA, Tinker	804M tokens	$354
Training PROSELENS	Nemotron LoRA, Tinker	1,613M tokens	$710
Total			$67,608

Scoring a new document. At inference a document passes through the format gate, one outline extraction, and one forward pass of IDEALENS per answer token; the paraphrase is used only to build training data. For a document of average corpus length this costs about $0.06 through the standard API and $0.03 through the batch API. Prompt caching lowers it further in bulk: the six in-context examples are shared by every document of a format, and cached tokens bill at a tenth of the input rate, giving $0.018 per document ($18 per 1,000) at our corpus run's cache hit rate (Table 16). Extraction is over 90% of the cost in every setting, mostly the extractor's thinking (5,678 of 6,765 output tokens); scoring on Tinker costs a fraction of a cent. PROSELENS skips the gate and the extractor and costs about $0.0005 per document, so the extra cost of reading ideas rather than words is the extraction step.

Table 16: **Cost of scoring one new document** (US dollars, mean tokens per document at the corpus's mean length of 1,049 words). The gate and the extractor are Gemini 3.7 Flash; extraction tokens are from the 614,454 extractions of the corpus build, where *input* includes the six in-context examples and *output* includes the extractor's thinking. *Standard* and *Batch* price every token at the standard and batch API rates (batch is half of standard); *Batch, cached* also bills the share of the extraction input that hit the prompt cache in our corpus run (88%, the in-context examples shared by all documents of a format) at the cached rate, a tenth of the input rate. Scoring runs IDEALENS and PROSELENS (Nemotron LoRA) on Tinker, which bills the prompt only, twice per document (one forward pass per answer token); PROSELENS reads the document itself and needs neither the gate nor the extractor. Gemini rates as published for 2026 (they double on 1 January 2027); Tinker rates carry a limited-time 50% discount on list.

Step	Input tokens	Output tokens	Standard	Batch	Batch, cached
Format gate	2,647	98	$0.0024	$0.0012	$0.0012
Outline extraction	45,522	6,765	$0.060	$0.030	$0.016
IDEALENS scoring	1,212	0	$0.0002	$0.0002	$0.0002
IDEALENS, per document			**$0.062**	**$0.031**	**$0.018**
IDEALENS, per 1,000 documents			$62	$31	$18
PROSELENS, per document	2,732	0	$0.0005	$0.0005	$0.0005

G RESULTS BY QUADRANT

The tables in this section show the accuracy of every evaluation at the global 1% cut per quadrant and where reported, at the per-format cut. Every table compares the mean accuracy of other quadrants

versus the quadrant in question. Where an evaluation has documents on both sides of our corpus's 500-word minimum, sub-rows split it by length. IDEALENS (Nemotron-30B) is trained on outlines and reads either the outline or the raw document; PROSELENS is Nemotron-30B trained on and reading raw documents; Pangram 4 is used at its published threshold and EditLens-Llama-3B at our calibrated 1% cut. The per-format cut follows the slash in the IDEALENS outline column (§E.2); every other column and every mean row uses the global cut.

G.1 HUMAN IDEAS, HUMAN PROSE

Nineteen evaluations contribute human-written documents. Every detector keeps its false-positive rate low here: averaged over evaluations, IDEALENS flags 0.6% of them, PROSELENS 0.7%, Pangram 4 0.3% and EditLens-3B 0.2% (Table 17).

Table 17: **Human ideas, human prose: accuracy (%).** Ordinary human writing, so accuracy is the share labeled human. The top two rows give each detector's mean accuracy in this quadrant and in the other three, so no detector can look strong by always predicting one label. Outline column: global 1% cut / per-format cut. †All documents under 500 words. All arms are in §L.

		IDEALENS				
Evaluation and arm	n	outline	document	PROSELENS	Pangram 4	EditLens
Mean accuracy, the other three quadrants		**84.6**	**63.9**	**49.9**	**49.7**	**51.3**
Mean accuracy, this quadrant (19 evaluations)		**99.4**	**99.7**	**99.3**	**99.7**	**99.8**
IntelLabs reviews human reviews	1,622	99.9 / 99.6	100.0	100.0	100.0	99.9
< 500 words	1,212	99.9 / 99.5	100.0	100.0	100.0	99.9
≥ 500 words	410	100.0 / 99.8	100.0	100.0	100.0	100.0
Saha et al. H: human review	2,621	100.0 / 99.7	100.0	100.0	100.0	100.0
DetectRL-X human	7,169	95.2 / 94.6	95.9	88.7	96.7	99.0
< 500 words	5,895	94.3	95.0	86.4	96.0	98.8
≥ 500 words	1,274	99.5	99.8	99.7	100.0	100.0
StoryScope human	1,345	99.9 / 99.3	100.0	100.0	100.0	100.0
IDEASHIFT-X human pages	10,160	99.9 / 99.9	100.0	100.0	100.0	100.0
DetectionAI human	1,927	100.0 / 99.4	100.0	99.9	100.0	100.0
< 500 words	658	100.0 / 98.5	100.0	99.8	100.0	100.0
≥ 500 words	1,269	100.0 / 99.9	100.0	100.0	100.0	100.0
OpAI-Bench v0: the human draft	4,145	99.7 / 99.2	100.0	100.0	100.0	100.0
< 500 words	2,883	99.6	100.0	100.0	100.0	100.0
≥ 500 words	1,262	99.8	100.0	100.0	100.0	100.0
Sem-Detect human reviews	3,061	99.8 / 99.3	100.0	100.0	100.0	100.0
< 500 words	2,007	99.8 / 99.1	100.0	100.0	100.0	100.0
≥ 500 words	1,054	100.0 / 99.8	100.0	100.0	100.0	100.0
MELD human	3,918	99.8 / 99.6	100.0	100.0	100.0	100.0
< 500 words	3,493	99.8	100.0	100.0	100.0	100.0
≥ 500 words	425	100.0	100.0	100.0	100.0	100.0
ARB human	1,752	99.5 / 98.1	100.0	99.9	100.0	99.9
< 500 words	1,747	99.5	100.0	99.9	100.0	99.9
≥ 500 words	5	100.0	100.0	100.0	100.0	100.0
GEDE† human	915	99.9 / 99.7	100.0	100.0	100.0	100.0
HART human	3,922	99.4 / 98.3	99.8	99.6	99.9	99.7
< 500 words	3,466	99.3	99.9	99.7	99.9	99.7
≥ 500 words	456	99.6	99.8	99.6	100.0	99.6
Epoch imitation genuine passages	495	100.0 / 99.2	100.0	100.0	100.0	100.0
< 500 words	245	100.0 / 98.4	100.0	100.0	100.0	100.0
≥ 500 words	250	100.0 / 100.0	100.0	100.0	100.0	100.0
CoCoNUTS human review	1,194	100.0 / 99.9	100.0	100.0	100.0	100.0
< 500 words	735	100.0 / 99.9	100.0	100.0	100.0	100.0

continued on the next page

Table 17, continued

Evaluation and arm	n	IDEALENS outline	IDEALENS document	PROSELENS	Pangram 4	EditLens
$\geq$ 500 words	459	100.0 / 100.0	100.0	100.0	100.0	100.0
ELLIPSE all essays	3,909	99.6 / 97.8	100.0	100.0	100.0	100.0
$<$ 500 words	2,757	99.6	100.0	100.0	100.0	100.0
$\geq$ 500 words	1,152	99.7	100.0	100.0	100.0	100.0
AI-Researcher expert reviews	504	97.4 / 87.9	97.8	97.8	97.6	96.8
$<$ 500 words	476	97.3	97.7	97.7	97.7	96.6
$\geq$ 500 words	28	100.0	100.0	100.0	96.4	100.0
AcademicIntegrity original	289	99.7 / 99.0	100.0	100.0	99.7	100.0
$<$ 500 words	287	99.7	100.0	100.0	99.7	100.0
$\geq$ 500 words	2	100.0	100.0	100.0	100.0	100.0
GEN human	3,055	99.5 / 98.7	100.0	100.0	100.0	100.0
$<$ 500 words	2,027	99.7	100.0	100.0	100.0	100.0
$\geq$ 500 words	1,028	99.1	100.0	100.0	100.0	100.0
C4 web pages	9,998	100.0	100.0	100.0	100.0	100.0

G.2 MODEL IDEAS, MODEL PROSE

Nineteen evaluations contribute model-written documents. The detectors that read the text do best here: PROSELENS detects 98.9% on average and Pangram 4 97.3%, against 91.1% for IDEALENS and 81.8% for EditLens-3B (Table 18). Humanized and adversarially attacked arms are marked; §I.2 compares them with their clean counterparts.

Table 18: **Model ideas, model prose: accuracy (%).** Conventional machine-text detection; accuracy is TPR, and attacked or humanized arms are marked. PeerPrism (Sadeghian et al., 2026) and the personal-style set (Baumler et al., 2026) are left out (§C). †every document under 500 words. Every arm of every evaluation is in §L.

Evaluation and arm	n	IDEALENS outline	IDEALENS document	PROSELENS	Pangram 4	EditLens
Mean accuracy, the other three quadrants		**87.4**	**65.1**	**50.0**	**50.5**	**57.3**
Mean accuracy, this quadrant (19 evaluations)		**91.1**	**96.1**	**98.9**	**97.3**	**81.8**
IntelLabs reviews model reviews	8,334	73.8 / 85.3	84.7	98.6	94.7	94.5
$<$ 500 words	6,575	70.1 / 82.6	82.1	98.3	93.8	94.1
$\geq$ 500 words	1,759	87.8 / 95.7	94.7	100.0	97.8	96.1
Saha et al. peer reviews, assistance levels						
AI-BP	4,920	64.8 / 77.7	82.9	100.0	97.8	88.3
$<$ 500 words	1,744	69.4 / 81.9	81.5	100.0	97.9	98.2
$\geq$ 500 words	3,176	62.2 / 75.5	83.7	100.0	97.8	82.8
AI-EP	4,530	70.9 / 81.7	84.8	100.0	98.5	91.5
$<$ 500 words	1,688	73.7 / 83.8	83.2	100.0	98.3	98.8
$\geq$ 500 words	2,842	69.3 / 80.5	85.8	100.0	98.6	87.2
AI-BP, humanized *(attacked)*	200	86.0 / 93.0	87.0	99.0	78.5	0.0
DetectRL-X English rows						
clean generations	7,451	87.3 / 83.4	98.6	100.0	99.7	92.9
$<$ 500 words	5,984	85.5	98.2	100.0	99.7	95.4
$\geq$ 500 words	1,467	94.9	100.0	100.0	99.9	83.0
eleven attacks, pooled *(attacked)*	5,296	83.8 / 80.4	93.1	97.1	97.9	84.1
$<$ 500 words	4,016	80.8	92.2	96.3	97.4	83.8
$\geq$ 500 words	1,280	93.0	95.7	99.5	99.5	85.3
StoryScope five generators	6,902	97.0 / 98.5	99.8	99.8	100.0	21.6
$<$ 500 words	29	72.4 / 82.8	89.7	89.7	100.0	65.5
$\geq$ 500 words	6,882	97.1 / 98.5	99.9	99.8	100.0	21.5

continued on the next page

Table 18, continued

Evaluation and arm	n	IDEALENS outline	IDEALENS document	PROSELENS	Pangram 4	EditLens
IDEASHIFT-X level 0: topic-only	10,152	95.3 / 95.0	81.4	82.5	85.4	15.4
$<$ 500 words	940	96.3	86.4	87.7	90.9	27.2
$\geq$ 500 words	7,369	95.3	79.9	81.0	84.5	14.6
DetectionAI essays, with a humanizer						
four generators	7,707	90.3 / 93.7	93.7	99.7	99.8	94.7
$<$ 500 words	2,634	76.5 / 83.8	81.5	99.2	99.5	92.0
$\geq$ 500 words	5,073	97.5 / 98.8	100.0	100.0	100.0	96.0
humanized *(attacked)*	7,706	57.9 / 69.3	32.4	64.1	94.2	4.9
$<$ 500 words	2,917	47.4 / 56.7	27.0	64.5	84.9	11.0
$\geq$ 500 words	4,789	64.2 / 77.0	35.7	63.8	99.9	1.2
Sem-Detect model reviews	6,767	71.4 / 85.7	95.9	100.0	75.7	65.2
$<$ 500 words	1,488	91.7 / 97.3	98.4	100.0	80.0	98.4
$\geq$ 500 words	5,279	65.7 / 82.4	95.2	100.0	74.6	55.9
MELD long-form, four generators						
clean generations	1,546	88.1 / 88.5	97.0	99.9	100.0	98.0
$<$ 500 words	884	85.6	96.0	99.8	100.0	97.4
$\geq$ 500 words	662	91.4	98.3	100.0	100.0	98.8
six attacks, pooled *(attacked)*	9,249	84.2 / 84.9	71.6	72.9	100.0	53.1
$<$ 500 words	5,290	79.9	68.8	70.5	100.0	52.7
$\geq$ 500 words	3,959	89.9	75.5	76.2	100.0	53.6
LAMP† model draft	1,057	98.3 / 99.1	100.0	100.0	100.0	99.8
ARB† matched rewriting benchmark						
free generation	6,904	97.9 / 96.8	97.3	100.0	99.9	98.4
generation, rewritten *(attacked)*	7,043	97.5 / 96.1	96.5	99.9	99.9	98.1
GEDE† student essays, contribution levels						
from the assignment	1,870	99.4 / 97.3	99.9	100.0	100.0	100.0
paraphrase attacks *(attacked)*	3,604	88.7 / 81.4	76.6	98.6	97.6	60.7
HART English rows						
clean generations	3,953	96.5 / 95.4	97.7	100.0	100.0	96.1
$<$ 500 words	3,444	96.3	97.4	100.0	100.0	96.7
$\geq$ 500 words	509	98.0	99.6	100.0	100.0	91.6
humanized *(attacked)*	3,592	84.6 / 84.3	82.0	93.5	94.9	71.0
$<$ 500 words	3,165	82.8	80.1	92.7	94.2	71.6
$\geq$ 500 words	427	97.4	96.7	100.0	100.0	66.5
Epoch imitation model, topic prompt	297	97.6 / 98.7	100.0	100.0	100.0	93.9
$<$ 500 words	120	95.8 / 97.5	100.0	100.0	100.0	99.2
$\geq$ 500 words	177	98.9 / 99.4	100.0	100.0	100.0	90.4
CoCoNUTS model-generated	1,400	92.1 / 97.1	98.9	100.0	96.3	99.0
$<$ 500 words	728	93.4 / 98.2	98.9	100.0	96.3	99.9
$\geq$ 500 words	672	90.6 / 96.0	98.8	100.0	96.3	98.1
AI-Researcher proposals from model ideas	97	95.9 / 100.0	100.0	100.0	100.0	93.8
GEN model-written	4,046	95.8 / 96.5	98.4	99.7	99.8	70.9
$<$ 500 words	314	81.2	83.1	96.2	98.1	49.7
$\geq$ 500 words	3,732	97.0	99.7	99.9	100.0	72.7
TWICETOLD model originals	50	96.0	100.0	100.0	100.0	90.0
IDEASHIFT, human sources levels 0–1: topic-only or genre, audience and tone	474	93.0 / 93.5	98.5	99.4	100.0	63.9
IDEASHIFT, AI sources levels 0–5	1,578	98.3 / 98.6	99.9	100.0	99.9	76.2

G.3 HUMAN IDEAS, MODEL PROSE

Twelve evaluations contain a person's ideas written up, rewritten or polished by a model. This is where a prose detector and an idea detector part ways: IDEALENS flags 5.3% of these documents

on average, against 49.3% for PROSELENS, 56.2% for Pangram 4 and 27.9% for EditLens-3B (Table 19).

Table 19: **Human ideas, model prose: accuracy (%).** A model wrote, rewrote, polished, or translated a person's content, so accuracy is the share labeled human. The top two rows give each detector's mean accuracy in this quadrant and in the other three, so no detector can look strong by always predicting one label. Outline column: global 1% cut / per-format cut. †All documents under 500 words. All arms are in §L.

		IDEALENS				
Evaluation and arm	n	outline	document	PROSELENS	Pangram 4	EditLens
Mean accuracy, the other three quadrants		**86.2**	**67.2**	**66.1**	**68.3**	**60.5**
Mean accuracy, this quadrant (12 evaluations)		**94.7**	**89.6**	**50.7**	**43.8**	**72.1**
Saha et al. peer reviews, assistance levels						
H-AI: human review, polished	15,118	99.4 / 97.7	99.8	52.1	41.3	62.1
< 500 words	13,284	99.3 / 97.6	99.8	47.9	36.7	60.0
≥ 500 words	1,834	99.5 / 98.5	99.8	82.6	74.4	77.3
H-AI, humanized *(humanized)*	718	98.5 / 96.2	98.9	31.1	28.3	99.4
IDEASHIFT-X level 5: the person's outline	10,116	99.3 / 99.0	96.6	55.5	38.1	98.4
< 500 words	446	99.8	99.6	62.6	36.5	92.8
≥ 500 words	7,830	99.2	97.1	57.9	42.6	98.4
OpAI-Bench v1–v8: model-rewritten	33,163	99.6 / 98.8	100.0	87.9	83.8	99.1
< 500 words	22,303	99.5	100.0	87.7	80.4	98.9
≥ 500 words	10,860	99.7	100.0	88.2	90.8	99.5
Sem-Detect human review, rewritten	12,318	91.8 / 84.8	88.2	35.3	22.3	34.0
< 500 words	7,903	92.3 / 86.0	91.4	31.0	18.3	29.2
≥ 500 words	4,415	90.8 / 82.6	82.4	43.2	29.4	42.6
ARB† human article, rewritten	7,027	98.1 / 95.2	99.5	59.7	46.7	96.7
GEDE† essay improved or rewritten	3,648	98.4 / 97.6	98.9	38.5	25.5	71.4
HART human text, rephrased	3,904	95.1 / 92.9	95.1	55.5	57.2	79.9
< 500 words	3,510	94.7	94.7	52.4	55.0	78.4
≥ 500 words	394	98.7	98.7	83.5	76.4	93.4
CoCoNUTS peer reviews, collaboration modes						
human review, polished	1,497	87.8 / 80.9	88.0	39.0	38.8	37.1
< 500 words	1,153	86.4 / 78.9	86.4	31.5	31.2	30.3
≥ 500 words	344	92.7 / 87.5	93.3	64.2	64.2	59.9
human review, translated	1,493	99.9 / 99.5	100.0	99.5	99.9	98.7
< 500 words	1,139	99.9 / 99.4	100.0	99.4	99.8	98.2
≥ 500 words	354	100.0 / 100.0	100.0	100.0	100.0	100.0
AI-Researcher proposals from expert ideas	48	89.6 / 81.2	87.5	85.4	87.5	75.0
< 500 words	1	100.0	100.0	100.0	100.0	0.0
≥ 500 words	47	89.4	87.2	85.1	87.2	76.6
AcademicIntegrity pre-2016 abstracts						
refined or regenerated	867	91.8 / 83.9	96.3	16.1	3.6	30.6
humanized *(humanized)*	1,156	99.1 / 97.7	99.9	99.3	91.0	99.6
< 500 words	1,145	99.1	99.9	99.3	91.0	99.6
≥ 500 words	11	100.0	100.0	100.0	90.9	100.0
GEN model-edited, seven categories	13,964	86.0 / 83.0	73.7	40.1	43.2	66.3
< 500 words	10,420	86.5	76.3	40.4	37.9	64.9
≥ 500 words	3,666	84.4	66.1	39.2	58.2	70.3
IDEASHIFT, human sources level 5: the person's outline	237	93.2 / 92.0	45.6	13.5	7.6	84.0

G.4 MODEL IDEAS, HUMAN PROSE

Only TWICETOLD places a model's ideas in prose a person wrote from scratch (§H.2). External sets where a person edited a model's draft (LAMP, HART's human-edited arm) keep most of the model's prose and are not counted here.

Table 20: **Model ideas, human prose: accuracy (%).** TWICETOLD: 50 stories people wrote from scratch from the outline of a model's story (§3), so every flag is a correct detection. Human-edited model drafts (LAMP, HART's human-edited arm) keep most of the model's prose and are excluded. The top two rows give each detector's mean accuracy in this quadrant and in the other three, so no detector can look strong by always predicting one label. Outline column: global 1% cut / per-format cut. All arms are in §L.

		IDEALENS				
Evaluation and arm	n	outline	document	PROSELENS	Pangram 4	EditLens
Mean accuracy, the other three quadrants		**95.1**	**95.1**	**83.0**	**80.3**	**84.6**
Mean accuracy, this quadrant (1 evaluation)		**68.0**	**6.0**	**0.0**	**8.0**	**0.0**
TWICETOLD rewrites	50	68.0 / 80.0	6.0	0.0	8.0	0.0

G.5 ALL DETECTORS AND BASELINES

Table 21 gives the per-quadrant means for every detector we ran, including our other backbones and the open-source baselines of Table 14. Among the open baselines, those with a high true-positive rate on model-written text pay for it with a high false-positive rate on human text at their own thresholds.

Table 21: **Every detector, every quadrant.** Each cell is the mean over evaluations of the pooled rate at the global 1% cut ("own": the detector's own threshold). Q1: human ideas, human prose; Q2: model ideas, human prose; Q3: model ideas, model prose (attacked arms separate); Q4: human ideas, model prose. A superscript gives the number of evaluations when a detector covers fewer than the column. An idea detector flags Q2 and Q3 but not Q1 or Q4; a prose detector flags Q3 and Q4.

Detector	Q1 FPR, 19 evals	Q4 FPR, 12 evals	Q3 TPR, 19 evals	Q3 attacked TPR, 8 evals	Q2 TPR, 1 eval
Ours					
IDEALENS (outline)	0.6	5.3	91.1	84.5	68.0
IDEALENS (document)	0.3	10.4	96.1	79.8	6.0
PROSELENS	0.7	49.3	98.9	90.6	0.0
IDEALENS-MODERNBERT-L (outline)	1.5	6.5	79.1	73.9	28.0
IDEALENS-MODERNBERT-L (document)	0.2^{17}	9.4^{11}	89.1^{17}	55.5	–
PROSELENS-MODERNBERT-L	1.0^{18}	38.1^{11}	99.1^{18}	76.2	0.0
IDEALENS-QWEN3.5-9B (outline)	0.9^{18}	4.7^{11}	84.5^{16}	76.8	–
IDEALENS-QWEN3.5-9B (document)	0.3^{18}	5.5^{11}	89.9^{16}	64.9	–
IDEALENS-LOGISTICCLASSIFIER	1.1^{18}	1.9^{11}	22.1^{16}	27.4	–
Baselines					
Pangram 4 (own threshold)	0.3	56.2	97.3	94.9	8.0
EditLens-Llama-3B	0.2	27.9	81.8	58.3	0.0
Binoculars (own threshold)	1.0	5.8	40.3	29.1	0.0
EditLens-RoBERTa	0.2	13.3	59.3	45.4	0.0
MELD (own threshold)	8.2	77.6	98.2	83.2	22.0
Desklib-academic (own threshold)	38.8	89.1	95.1	94.7	84.0
Desklib (own threshold)	14.6	81.6	88.8	85.2	40.0
Entropy	2.8	3.5	2.5	5.2	0.0
Fast-DetectGPT (own threshold)	23.0	50.5	75.3	67.5	10.0
Likelihood	2.6	3.3	4.4	5.6	0.0
Log-rank	2.6	3.2	5.3	6.2	0.0
LRR	2.2	2.3	9.1	7.4	0.0
MAGE (own threshold)	27.0	59.3	62.1	62.5	14.0
OpenAI-RoBERTa-base (own threshold)	4.4	7.5	11.1	17.4	0.0
OpenAI-RoBERTa-large (own threshold)	3.5	5.5	6.0	11.2	0.0
RADAR (own threshold)	17.4	28.8	34.7	40.6	8.0
Rank	3.4	5.4	8.2	9.8	0.0

G.6 MIXED PROVENANCE

Four external benchmarks contain documents whose ideas are genuinely shared: a model writing a review from a person's notes (AI in Peer Review) or from a person's review (CoCoNUTS), a model imitating a named author from five of their passages (Epoch), and papers people wrote from a model's research idea (AI-Researcher). These have no correct binary label, so we compare their mean P(AI) with the same benchmark's clear arms (the full comparison, with a fifth mixed arm, GEN's AI editor adding details to a person's text, is in the repository). IDEALENS places them between the two: on AI in Peer Review, reviews written from a person's notes score 68.5, between model reviews from the paper (77.1) and human reviews polished by a model (4.1), while PROSELENS (99.3) and Pangram 4 (89.8) score them like fully model-written reviews.

G.7 OPAI-BENCH: SUCCESSIVE MODEL EDITS OF A HUMAN DRAFT

OpAI-Bench (Bsharat et al., 2026) starts from 4,145 human drafts (v0) and applies eight successive model operations to each, so that by v8 a model has written almost every word while the ideas remain the person's. Every version is therefore human-ideated and every flag is a false positive. IDEALENS stays flat, flagging 0.3% of the drafts and 0.2 to 0.6% of every edited version, while PROSELENS rises to 26.3% and Pangram 4 to 48.5% as the model's share of the prose grows (Table 22).

Table 22: **OpAI-Bench**: main rows.

			IDEALENS				
Arm	metric	n	outline	document	PROSELENS	EditLens	Pangram 4
v0: the human seed	FPR	4,145	0.3	0.0	0.0	0.0	0.0
v1 to v8: AI-edited versions							
mean over 8 versions	FPR	33,163	0.4	0.0	12.1	0.9	16.2
median over 8 versions	FPR		0.4	0.0	10.1	0.3	4.7
min over 8 versions	FPR		0.2 (v1)	0.0 (v1)	0.0 (v1)	0.0 (v2)	0.0 (v1)
max over 8 versions	FPR		0.6 (v5)	0.0 (v5)	26.3 (v6)	3.7 (v6)	48.5 (v8)

H THE EVALUATION SETS WE BUILT

H.1 IDEASHIFT

IDEASHIFT tests how IDEALENS responds as a model receives more of a source document's plan. For each of 500 test documents (237 human-written, 263 model-written), we write six prompts that reveal progressively more of the source (Table 23). Level 0 gives only the document type, subject and length, and each later level adds more: genre, audience, and tone (1); central theme (2); order of discourse roles (3); full outline (4); and an instruction to realize the outline exactly, adding nothing (5). One model writes all six versions of a source: GPT-5.6 Sol for about half the sources, GPT-5.6 Terra or Luna for the rest (OpenAI, 2026a). GPT-5.6 Sol, at high reasoning effort, writes the briefs for levels 0 to 4 from the source's outline; the level-5 prompt is a fixed template filled with the outline. On human sources, the ideas shift from the model's (levels 0–1) to the person's (level 5);

Table 23: **What the writing model is given at each level of IDEASHIFT**, for one human-written source: a 744-word personal blog post reviewing a restaurant visit. Every level also asks for a complete, original document of about the source's length; levels 0 to 4 leave everything not specified to the model. The full prompts are in §K and the repository.

Level	What the model receives
0	Write a dining and lifestyle blog post reviewing a restaurant visit.
1	Write a first-person dining and lifestyle blog post that reviews a restaurant visit for food and travel enthusiasts in a casual, conversational, and evaluative tone, using dish-by-dish critiques, descriptions of the venue's ambiance, and links to related blog posts. *Focus broadly on a seasonal dining and travel experience in a scenic destination.*
2	Level 1's description, plus the source's global theme: *A winter dining excursion to the refurbished Hydro Majestic's Winter Garden restaurant features a cozy, picturesque atmosphere overlooking Megalong Valley and a mixed experience with a gourmet set menu.*
3	Level 2, plus the order of the source's roles: *Arrange the post so that it begins with an attention hook and purpose framing, continues with two separate event accounts, then moves through a description, a recommendation, a reference aid, and five separate evaluations before ending with a synthesis, social rapport, and navigation.*
4	Level 2, plus the source's full outline, one line per item (15 items), for example: *Attention hook: Contrast people who dislike winter cold with those who welcome the season's fireplaces and substantial comfort food.*
5	The source's description, themes and full outline (15 items), with an instruction to realise every item in order and add nothing, for example: *1. [ATTENTION HOOK] The author opens by contrasting people who dread winter's cold with those who embrace the season for cozy fireplaces and hearty comfort food.*

on model sources they are a model's at every level. IDEALENS follows the ideas: on human sources it flags 94.9% of level-0 and 6.8% of level-5 documents, and on model sources 96.6–99.2% at every level. PROSELENS and Pangram 4 flag at least 86.5% at every level of both (Table 24 and Table 25).

Each level carries the intended share of the source (Table 26). On human sources, the share of the source's outline items a document keeps rises from 0.30 at level 0 to 0.99 at level 5 (0.05 for an unrelated source of the same format and topic), and its cosine similarity to the source rises from 0.52 to 0.87 (0.33 unrelated). AI sources follow the same curve.

Table 24: **IDEASHIFT, human sources**: main rows.

			IDEALENS				
Arm	metric	n	outline	document	PROSELENS	EditLens	Pangram 4
source (the person's document)	FPR	237	3.4	3.4	2.1	2.5	0.4
level 0 (topic-only)	TPR	237	94.9	98.7	99.2	68.4	100.0
level 1 (genre, audience and tone)	TPR	237	91.1	98.3	99.6	59.5	100.0
level 2 (themes)	mixed, AI-leaning (fire rate)	237	86.1	97.0	98.7	57.0	100.0
level 3 (themes and structure)	mixed, AI-leaning (fire rate)	237	78.9	97.5	99.6	54.0	99.6
level 4 (themes and full outline)	mixed	237	23.6	75.1	94.9	29.5	98.3
level 5 (the source's outline)	FPR	237	6.8	54.4	86.5	16.0	92.4
AUC, levels 0-1 vs source			0.991	0.998	0.999	0.980	0.998

Table 25: **IDEASHIFT, AI sources**: main rows.

			IDEALENS				
Arm	metric	n	outline	document	PROSELENS	EditLens	Pangram 4
source (a model's document)	TPR	263	98.1	98.5	100.0	77.2	73.4
level 0 (topic-only)	TPR	263	98.1	99.6	100.0	74.9	100.0
level 1 (genre, audience and tone)	TPR	263	98.5	100.0	100.0	72.6	100.0
level 2 (themes)	TPR	263	99.2	100.0	100.0	75.7	100.0
level 3 (themes and structure)	TPR	263	96.6	100.0	100.0	75.7	100.0
level 4 (themes and full outline)	TPR	263	98.9	100.0	100.0	78.3	100.0
level 5 (the source's outline)	TPR	263	98.5	100.0	100.0	79.8	99.2

Table 26: **IDEASHIFT: how much of the source each level carries.** *Doc cosine*: text-embedding-3-large similarity between each level's document and its source. *Items kept*: share of the source's outline items whose best match in the level's outline has cosine $\geq \tau = 0.471$, the 95th percentile against unrelated sources of the same format and topic. *Rises*: share of sources whose kept share rises from the previous level (ties count half). Reference rows pair each source with an unrelated one.

		Human sources (n = 237)		AI sources (n = 263)		
Level	Given to the writer	doc cosine	items kept	doc cosine	items kept	rises (%, human / AI)
0	type + broad subject	0.518	0.301	0.537	0.348	–
1	very high-level topic	0.572	0.450	0.589	0.471	72 / 67
2	global themes	0.746	0.831	0.788	0.874	94 / 93
3	themes + role sequence	0.754	0.835	0.792	0.885	51 / 54
4	themes + full outline	0.842	0.982	0.865	0.997	85 / 78
5	full outline, realise exactly	0.868	0.990	0.889	0.999	51 / 51
unrelated source, same format and topic		0.330	0.047	0.348	0.064	
unrelated source, same format		0.235	0.008	0.290	0.018	

H.2 TWICETOLD

TWICETOLD is the reverse of IDEASHIFT: a model supplies the ideas and a person writes the prose. GPT-5.6 Sol (OpenAI, 2026a) wrote 500-word stories from one-line genre prompts, and we extracted and paraphrased an outline of each with our pipeline. Seventeen writers (computer science graduate students and faculty) then wrote each story from its outline alone, without seeing the model's version (instructions in §K). They wrote in person rather than through crowdsourcing, to reduce the chance of AI assistance (Veselovsky et al., 2025). The 50 resulting stories have a model's ideas and a person's prose, so IDEALENS should flag them as AI. They average 837 words, against 502 for the model versions.

H.2.1 OUTLINE ADHERENCE

We analyze the faithfulness of the rewrite to the original outline to measure quality. On average, 94.1% of the original outline is retained and 20.1% of the rewrite-outline is extraneous. Authors rarely deviate from the chronology: 90.7% of matched items are in order.

H.2.2 ROBUSTNESS TO RENAMING

Character names and locations in TWICETOLD stories are largely decided by the model outlines. Model outlines exhibit strong tendencies towards certain names, including "Mara", names starting with "El-", and towns named "Bellweather". We analyze whether replacing character and location names with random ones sampled from the most popular U.S. baby names of the 2010s and from U.S. cities respectively affects the verdicts of our models. The mean P(AI) of IDEALENS decreases after renaming, from 79.8 to 69.4, while PROSELENS's mean P(AI) decreases from 0.14 to 0.04.

H.2.3 RESULTS

IDEALENS flags 68.0% of the rewrites and 96.0% of the model originals they were written from using the global cut, while PROSELENS, EditLens-3B and Pangram 4 flag nearly every original but almost none of the human versions (Table 27). All human and original versions fall into the Creative Writing format.

Table 27: **TWICETOLD: flagged as AI (%), 1% global cut** on the 50 stories people wrote from an AI story's outline and the 50 original AI model stories. Evaluating on TWICETOLD tests whether a detector finds model ideas in human prose, while the model originals are the control.

Detector	TWICETOLD	Model originals
IDEALENS (outline)	68.0 (34/50)	96.0 (48/50)
IDEALENS (document)	6.0 (3/50)	100.0 (50/50)
PROSELENS	0.0 (0/50)	100.0 (50/50)
EditLens-3B	0.0 (0/50)	90.0 (45/50)
Pangram 4	8.0 (4/50)	100.0 (50/50)

Table 28: **TWICETOLD Correlations:** Among human rewrites, PROSELENS has a more strongly negative Spearman correlation with Jaccard unigram overlap than IDEALENS. For the original AI stories, IDEALENS has a weak negative correlation, while the PROSELENS correlation is near zero.

	IDEALENS		PROSELENS	
Author	Length	Jacc-1g	Length	Jacc-1g
Human	0.07	-0.52	0.60	-0.65
GPT 5.6 Sol	-0.11	-0.28	-0.18	-0.02

H.3 IDEASHIFT-X

IDEASHIFT-X tests whether idea detection carries over to other languages. IDEALENS is trained only on English, but it never reads the document itself: the extractor writes an English outline of a document in any language. We sample 10,160 human-written pages from FineWeb2 (Penedo et al., 2025), all written before 2022, in 24 languages grouped by the size of their FineWeb2 share: high-resource (Chinese, French, German, Japanese, Russian, Spanish), mid-resource (Arabic, Bengali, Hindi, Indonesian, Persian, Thai, Ukrainian, Vietnamese) and low-resource (Amharic, Khmer, Malayalam, Marathi, Nepali, Sinhala, Swahili, Tamil, Urdu, Yoruba). For each page a model writes two versions in the page's language, following the IDEASHIFT protocol: level 0 from a one-line description of the document type and subject (model ideas), and level 5 from the page's full outline, adding nothing (human ideas). Level 0 and level 5 documents were generated by Gemini 3.7 Flash (Google DeepMind, 2026c).

Table 29: **IDEASHIFT-X, selected languages (%).** Pre-2022 FineWeb2 pages in each language (human: false-positive rate), a model writing natively in that language from one line naming the document type and broad subject (level 0: TPR), and a model rendering the page's outline in that language with nothing added (level 5: false-positive rate under the provenance rule). IDEALENS reads English outlines extracted from every language; it has never seen a non-English document. Global 1% cut for our models, Pangram 4 at its published threshold. All 24 languages in §H.3.

		IDEALENS			PROSELENS			Pangram 4		
Language	n	human	level 0	level 5	human	level 0	level 5	human	level 0	level 5
high-resource										
German	704	0.0	93.0	0.3	0.0	99.6	81.1	0.0	100.0	98.3
Spanish	733	0.0	98.4	0.8	0.0	100.0	89.9	0.0	100.0	96.7
Chinese	614	0.0	92.2	1.3	0.0	99.0	88.5	0.0	99.8	98.4
Japanese	729	0.0	95.2	0.4	0.0	99.9	61.7	0.0	100.0	97.8
mid-resource										
Arabic	734	0.0	97.3	0.1	0.0	98.8	67.2	0.0	100.0	89.4
Bengali	250	0.0	97.6	0.4	0.0	76.6	2.9	0.0	89.1	28.4
Persian	250	0.0	96.8	1.2	0.0	99.2	45.2	0.0	100.0	70.6
Thai	250	0.0	98.0	0.8	0.0	98.4	19.2	0.0	100.0	68.0
low-resource										
Tamil	719	0.0	95.8	0.1	0.0	67.1	1.6	0.0	69.0	0.0
Sinhala	250	0.0	97.2	0.4	0.0	2.8	0.0	0.0	0.4	0.0
Yoruba	250	1.2	88.0	2.4	0.0	36.8	0.0	0.0	3.2	0.0
Amharic	250	0.0	96.0	0.8	0.0	1.2	0.0	0.0	0.0	0.0
All 24 languages	10,160	0.1	95.3	0.7	0.0	82.5	44.5	0.0	85.4	61.9

The human pages are almost never flagged (0.1%), IDEALENS detects 95.3% of level-0 documents and flags 0.7% of level-5 documents, and the level-0 detection rate holds across resource groups (94.2%, 96.4% and 95.4% for high-, mid- and low-resource languages). Detectors that read the prose depend on the language: on low-resource level-0 documents PROSELENS detects 51.6% and Pangram 4 58.0% (Table 29 shows four languages per group and the average over all 24; every language and the group averages are in the repository).

H.4 TRANSLATION

To separate language from ideas, we translate a subset of IDEASHIFT-X (50 pages per language for the human pages, level 0 and level 5; 3,594 documents) into English with Gemini 3.7 Flash and score both versions (Table 30). IDEALENS's verdicts barely move: level-0 TPR changes by under a point and the human FPR not at all, and its scores on the two versions correlate at 0.946 (Spearman). PROSELENS's level-0 TPR rises from 74.7 to 98.2%, all of it in low-resource languages, and its level-5 FPR rises by 13 points, because translated human ideas now read as model prose. Pangram 4 shows the same effect more strongly: its level-5 FPR goes from 51.3 to 97.0%.

Table 30: **Translation into English.** 50 sources per language in each of 24 languages, for the human arm and ladder levels 0 and 5 (3,594 documents translated of 3,600). Rates (%) on the original → the English translation, global 1% cut; Pangram 4 at its own threshold. Spearman correlation of original and translated scores over all pairs: IDEALENS 0.946, PROSELENS 0.878.

Arm	Rate	IDEALENS	PROSELENS	Pangram 4
Human sources	FPR	0.25 → 0.25	0.00 → 0.25	0.00 → 2.17
Level 0 (model ideas)	TPR	95.1 → 94.2	74.7 → 98.2	78.5 → 100.0
Level 5 (human ideas, model prose)	FPR	1.1 → 0.8	34.9 → 47.7	51.3 → 97.0

I ROBUSTNESS AND GENERALIZATION

I.1 PRE-CHATGPT WEB TEXT

Our in-domain human documents were crawled between 2024 and 2026, and some are likely AI-written despite Pangram's label. As a cleaner check of the false-positive rate we sample 10,000 pages from the April 2019 C4 snapshot (Raffel et al., 2020), which predates ChatGPT, stratified by format. Every flag is a false positive: IDEALENS flags 0.01%, and PROSELENS, Pangram 4 and EditLens-3B none.

I.2 ATTACKS AND HUMANIZERS

Eight external benchmarks pair clean model-written documents with versions that were humanized or adversarially rewritten, which keeps the ideas a model's. Table 31 compares each detector's TPR on the clean and attacked arms: averaged over the seven evaluations with both, IDEALENS drops by 6.4 points, PROSELENS by 10.6, Pangram 4 by 4.9 and EditLens-3B by 42.6. The per-attack tables for MELD and DetectRL-X show which attacks cost the most: character-level attacks such as zero-width spaces and homoglyphs, which the extractor normalizes away, leave IDEALENS untouched, while paraphrasing costs it the most.

Table 31: **Attacks and humanizers.** TPR (%) at the global 1% cut on each evaluation's unattacked model-written rows (clean) and on its attacked or humanized rows; n is the attacked rows. Pangram 4 at its own threshold. The mean is over the evaluations that have both a clean and an attacked arm.

		IDEALENS		PROSELENS		Pangram 4		EditLens-3B	
Evaluation	n	clean	attacked	clean	attacked	clean	attacked	clean	attacked
Saha et al. (humanizer)	200	67.7	86.0	100.0	99.0	98.1	78.5	89.8	0.0
DetectRL-X	5,296	87.3	83.8	100.0	97.1	99.7	97.9	92.9	84.1
DetectionAI (StealthGPT)	7,706	90.3	57.9	99.7	64.1	99.8	94.2	94.7	4.9
MELD	9,249	88.1	84.2	99.9	72.9	100.0	100.0	98.0	53.1
ARB (llm2l)	7,043	97.9	97.5	100.0	99.9	99.9	99.9	98.4	98.1
GEDE	3,604	99.4	88.7	100.0	98.6	100.0	97.6	100.0	60.7
Perkins et al. (6 techniques)	86	–	93.0	–	100.0	–	96.5	–	94.2
HART (humanizers)	3,592	96.5	84.6	100.0	93.5	100.0	94.9	96.1	71.0
Mean over the 7 paired evaluations		89.6	83.2	99.9	89.3	99.7	94.7	95.7	53.1
Drop			−6.4		−10.6		−4.9		−42.6

Table 32: **MELD, per attack**: by attack. Fire rate % at the 1% cut.

			IDEALENS				
Attack	Arm	n	outline	document	PROSELENS	EditLens	Pangram 4
homoglyph	ai_adversarial (6 attacks)	1,533	86.5	64.9	48.8	4.0	100.0
number	ai_adversarial (6 attacks)	1,553	84.4	95.6	99.4	96.1	100.0
synonym	ai_adversarial (6 attacks)	1,543	76.3	71.8	91.7	43.6	100.0
upper_lower	ai_adversarial (6 attacks)	1,564	86.4	97.3	97.8	82.4	100.0
whitespace	ai_adversarial (6 attacks)	1,556	85.5	97.5	96.9	89.6	100.0
zero_width_space	ai_adversarial (6 attacks)	1,500	86.4	0.1	0.0	0.0	100.0

Table 33: **DetectRL-X, per attack**: by attack. Fire rate % at the 1% cut.

			IDEALENS				
Attack	Arm	n	outline	document	PROSELENS	EditLens	Pangram 4
ai_backtranslation	English · ai_adversarial	487	87.3	90.3	97.3	66.3	99.8
ai_character_deletion	English · ai_adversarial	496	86.9	98.0	99.8	89.1	99.8
ai_character_insertion	English · ai_adversarial	493	86.0	98.0	99.8	89.0	100.0
ai_character_substitution	English · ai_adversarial	482	86.3	98.5	100.0	91.1	99.8
ai_condensing	English · ai_adversarial	497	83.7	96.8	99.0	90.5	99.8
ai_decoder_paraphrasing	English · ai_adversarial	486	59.7	64.2	74.5	62.8	80.2
ai_encoder_paraphrasing	English · ai_adversarial	491	85.7	96.3	99.4	75.4	100.0
ai_expanding	English · ai_adversarial	499	92.4	99.2	100.0	96.6	100.0
ai_polishing	English · ai_adversarial	497	88.3	98.2	100.0	91.5	99.8
ai_seq2seq_paraphrasing	English · ai_adversarial	372	76.9	82.5	97.8	80.1	98.1
ai_zero_width_insertion	English · ai_adversarial	496	86.1	98.6	100.0	91.3	99.6

I.3 DO ATTACKS PRESERVE CONTENT?

An attack that changes a document's content can also change its ideas, so a lower TPR is not necessarily a failure of idea detection. We measure how much of each clean AI document survives its attack (Table 34 and Table 35). Character-level attacks, polishing, condensing, and encoder paraphrasing barely change the document (cosine 0.98–0.99) or IDEALENS's TPR. The attacks that lower TPR change the content: StealthGPT's humanized documents have cosine 0.873 with their originals, and 68.2% fall below the 95th percentile for a different author's document on the same prompt. For DetectRL-X's decoder and seq2seq paraphrases, the documents IDEALENS still catches keep far more of the original than those it misses (cosine 0.993 and 0.979 vs. 0.853 and 0.768; 84% and 81% vs. 34% and 21% of outline items).

Table 34: **Content preservation under attack (documents).** text-embedding-3-large cosine between each clean AI document and its attacked version, and IDEALENS detection. Same-topic reference: another author's document on the same prompt or source (DetectionAI: another generator, mean 0.769, 95th pct. 0.901; DetectRL-X: the paired human document, mean 0.828, 95th pct. 0.942). "Still caught" and "missed" split the attacked documents whose clean version IDEALENS caught.

Set	Attack	pairs	doc cosine (mean)	below the reference's 95th pct	IdeaLens TPR clean	attacked	cosine, still caught	cosine, missed	Spearman(cosine, P(AI))
DetectionAI	StealthGPT humanizer	7,705	0.873	68.2%	90.3	57.8	0.879 (n 4369)	0.862 (n 2587)	+0.10
DetectRL-X	seq2seq paraphrasing	550	0.914	34.2%	82.9	60.7	0.979 (n 319)	0.768 (n 137)	+0.57
DetectRL-X	backtranslation	550	0.945	24.0%	82.9	81.5	0.955 (n 430)	0.808 (n 26)	+0.10
DetectRL-X	decoder paraphrasing	550	0.946	32.4%	82.9	58.4	0.993 (n 304)	0.853 (n 152)	+0.60
DetectRL-X	encoder paraphrasing	550	0.982	10.5%	82.9	83.1	0.984 (n 428)	0.972 (n 28)	+0.06
DetectRL-X	expanding	550	0.982	10.4%	82.9	88.5	0.982 (n 440)	0.994 (n 16)	-0.17
DetectRL-X	character substitution	550	0.988	2.9%	82.9	83.3	0.989 (n 428)	0.983 (n 28)	+0.09
DetectRL-X	condensing	550	0.990	1.5%	82.9	83.6	0.991 (n 437)	0.982 (n 19)	+0.09
DetectRL-X	character deletion	550	0.992	1.1%	82.9	83.6	0.992 (n 436)	0.991 (n 20)	+0.08
DetectRL-X	polishing	549	0.992	1.1%	82.9	84.7	0.992 (n 435)	0.985 (n 20)	+0.06
DetectRL-X	character insertion	550	0.992	0.4%	82.9	84.5	0.993 (n 439)	0.988 (n 17)	+0.05
DetectRL-X	zero width insertion	550	0.993	0.0%	82.9	83.5	0.994 (n 434)	0.992 (n 22)	+0.07

Table 35: **Content preservation under attack, outline items**: content preserved by humanizers and adversarial attacks, and IdeaLens detection.

Set	Attack	pairs	doc cosine	items kept	items traceable	items attacked / clean	IdeaLens TPR clean	IdeaLens TPR attacked	kept, caught	kept, missed	Spearman(kept, P(AI))
DetectionAI	StealthGPT humanizer	7,705	0.873	0.430	0.563	0.80	90.3	57.8	0.440 (n 4369)	0.373 (n 2587)	+0.10
DetectRL-X	seq2seq paraphrasing	550	0.914	0.620	0.609	1.12	82.9	60.7	0.807 (n 319)	0.208 (n 137)	+0.64
DetectRL-X	decoder paraphrasing	550	0.946	0.657	0.664	1.00	82.9	58.4	0.840 (n 304)	0.335 (n 152)	+0.63
DetectRL-X	backtranslation	550	0.945	0.773	0.769	1.02	82.9	81.5	0.784 (n 430)	0.615 (n 26)	+0.11
DetectRL-X	expanding	550	0.982	0.782	0.741	1.11	82.9	88.5	0.786 (n 440)	0.798 (n 16)	-0.02
DetectRL-X	condensing	550	0.990	0.793	0.801	1.00	82.9	83.6	0.801 (n 437)	0.777 (n 19)	+0.15
DetectRL-X	encoder paraphrasing	550	0.982	0.825	0.822	1.02	82.9	83.1	0.837 (n 428)	0.810 (n 28)	+0.10
DetectRL-X	polishing	549	0.992	0.830	0.818	1.03	82.9	84.7	0.841 (n 435)	0.819 (n 20)	+0.13
DetectRL-X	character substitution	550	0.988	0.831	0.826	1.02	82.9	83.3	0.844 (n 428)	0.840 (n 28)	+0.16
DetectRL-X	character insertion	550	0.992	0.835	0.833	1.02	82.9	84.5	0.846 (n 439)	0.823 (n 17)	+0.10
DetectRL-X	character deletion	550	0.992	0.842	0.841	1.02	82.9	83.6	0.854 (n 436)	0.793 (n 20)	+0.13
DetectRL-X	zero width insertion	550	0.993	0.844	0.836	1.03	82.9	83.5	0.854 (n 434)	0.780 (n 22)	+0.09

Table 36: **In-domain results and ablations** for the two outline-trained models. Top: the held-out test split. Middle: the same models read a transformed outline, with the 1% cut refitted on the transformed calibration split; the roles-only row also gives a separately trained roles-only ModernBERT-L. Bottom: the share of verdicts that survive paraphrasing, sentence-level paraphrase of the source and a change of extractor, and the cost of reading a 500-word excerpt.

	IDEALENS		IDEALENS-MODERNBERT-L	
	AUC	TPR / FPR	AUC	TPR / FPR
Held-out test split (48,869 documents, paraphrased outline, global 1% cut): TPR / realised FPR				
in-domain test	–	95.8 / 2.6	–	86.9 / 2.0
What the outline carries (2,500 test documents, inference-time transforms, variant-matched 1% cut): AUC, TPR / FPR				
full outline (roles + content)	0.995	94.7 / 2.2	0.990	87.5 / 2.0
content only (roles stripped)	0.993	93.2 / 2.3	0.986	83.8 / 2.2
shuffled item order	0.993	90.5 / 1.9	0.986	82.6 / 2.0
roles only (content stripped)	0.601	2.4 / 0.4	0.603	2.7 / 0.5
roles-only model, trained (48,869 docs)	–	–	0.913	39.0 / 1.5
items scored alone, logit-mean pooled	0.972	60.1 / 1.2	0.932	47.4 / 1.3
Pipeline invariance (% of verdicts unchanged at the global 1% cut): all / AI / human				
paraphrased vs raw outline (n=10,000)	98.6 / 97.9 / 99.2		96.7 / 94.0 / 99.4	
source paraphrased before extraction (n=2,494)	98.6 / 97.9 / 99.2		96.6 / 94.2 / 99.2	
extractor swapped (n=1,874)	98.8 / 98.3 / 99.3		96.0 / 93.0 / 99.1	
Excerpt vs full document (raw outline): TPR / FPR				
full document	–	96.5 / 3.1	–	88.6 / 2.8
random 500-word excerpt	–	91.4 / 2.8	–	71.6 / 2.0

I.4 CONTENT-PRESERVING TRANSFORMATIONS

If IDEALENS reads ideas, transformations that keep a document's ideas should not change its verdict. On in-domain test documents we (i) score the raw extracted outline instead of its paraphrase, (ii) paraphrase the source document sentence by sentence before extraction, (iii) extract the outline with a different model family (GPT-5.6) and (iv) score a 500-word excerpt instead of the full document. The verdict survives the first three for 98.6%, 98.6% and 98.8% of documents; excerpts lower the TPR from 96.5% to 91.4% (Table 36 and Table 37 to Table 40). Stripping the role labels or shuffling the items barely changes the AUROC (0.993), while the role sequence alone reaches only 0.601.

Table 37: **Raw versus paraphrased outline**: main rows.

			IDEALENS
Rows	metric	n	% unchanged
all	all	10,000	98.6
AI	AI	4,993	97.9
human	human	5,007	99.2

Table 38: **Source document paraphrased before extraction**: main rows.

			IDEALENS
Rows	metric	n	% unchanged
all	all	2,494	98.6
AI	AI	1,294	97.9
human	human	1,200	99.2

Table 39: **A different outline extractor**: main rows.

			IDEALENS
Rows	metric	n	% unchanged
all	all	1,874	98.8
AI	AI	963	98.3
human	human	911	99.3

Table 40: **500-word excerpts**: main rows.

			IDEALENS
Arm	metric	n	outline
full document, AI	TPR	1,196	96.5
500-word excerpt, AI	TPR	2,059	91.4
full document, human	human (FPR)	1,153	3.1
500-word excerpt, human	human (FPR)	2,045	2.8
AUC, full documents			0.995
AUC, excerpts			0.992

I.5 FORMAT CLASSIFICATION AND EXTRACTION VARIANCE

At test time a document passes through two stochastic steps before IDEALENS reads it: the format gate, which chooses the role vocabulary, and the outline extractor. Both run at the production settings (Gemini 3.7 Flash at default temperature; the extractor with six exemplars and high reasoning effort).

Format gate. The gate is consistent with itself: the same format in all three of three runs for 90.9% of 1,000 in-domain documents, and in all five of five runs for 93.2% of the 400 out-of-domain documents below. On those 400 documents, 9.6% of its runs place the document outside the eight formats with a role vocabulary.

Repeated extraction. We extract each document five times under its recorded format and score every outline with IDEALENS: 33 in-domain test documents (the 33 of 42 sampled that completed all five extractions) and 400 out-of-domain documents from MELD, Saha et al., and DetectRL-X, half drawn at random and half near the decision threshold (Table 41). The outlines vary between runs (the number of items has a coefficient of variation of about 9%), but the verdict rarely does: it never changes for an in-domain document and is unchanged for 94.0% of the random out-of-domain documents, and no run flags any of their human-written documents as AI. Instability concentrates near the threshold, where only 52.5% of documents keep their verdict in all five runs; Table 42 breaks this down by distance from the cut. We reran extraction but not the format gate, so these figures reflect extraction variance alone. Averaging several extractions would likely reduce it; we do not test this.

Table 41: **Repeated extraction.** Each document is extracted five times with the production settings and scored by IDEALENS. Columns: share (%) of documents with the same verdict in all five runs, mean pairwise verdict agreement, and mean per-document SD of P(AI). In-domain: the 33 of 42 sampled test documents with all five extractions. Out-of-domain documents are drawn at random or near the cut (P(human) within 0.05 of it; 177 of the 200 are model-written).

Set	Draw	n	same verdict, 5 runs	pairwise agreement	SD of P(AI)
In-domain test split	random	33	100.0	100.0	0.003
DetectRL-X	near the cut	60	65.0	84.0	0.100
DetectRL-X	random	60	91.7	95.7	0.037
MELD	near the cut	80	32.5	65.8	0.118
MELD	random	80	93.8	97.3	0.018
Saha	near the cut	60	66.7	83.7	0.064
Saha	random	60	96.7	98.3	0.007
All three OOD sets	random	200	94.0	97.1	0.021
All three OOD sets	near the cut	200	52.5	76.6	0.096

Table 42: **Verdict changes by distance from the cut** (400 out-of-domain documents, five extractions each). Distance is the mean over the five runs of the logit of P(human), minus the logit of the cut; a change is any run that disagrees with another.

Logit distance from the cut	documents	verdict changes (%)
below -2	133	3.0
$[-2, -0.5)$	94	57.4
$[-0.5, +0.5)$	29	100.0
$[+0.5, +2)$	23	78.3
above $+2$	121	1.7

I.6 NEWER GENERATORS

Models released after our corpus's June 2026 cutoff may produce ideas that older models did not. We prompt four of them, GPT-5.6 Sol (OpenAI, 2026a), GPT-6 Astra (OpenAI, 2026b), Claude Opus 5 (Anthropic, 2026c), and Claude Fable 5.1 (Anthropic, 2026b), with 250 writing prompts sampled from StoryScope (Table 43). At the global 1% cut, IDEALENS detects 78.5%, 62.8%, 51.0%, and 51.2% of their stories (90.7, 74.8, 65.7, and 65.6% at the per-format cut), against 97.0% for the five older StoryScope generators. Pangram 4 flags 99.2–100% of every new generator's stories. PROSELENS holds on the GPT models (98.8 and 83.2%) but falls to 41.2% on Opus 5 and 17.2% on Fable 5.1, so the drop is not specific to reading ideas.

Table 43: **StoryScope, newer generators**: main rows.

			IDEALENS				
Arm	metric	n	outline	document	PROSELENS	EditLens	Pangram 4
gpt-5.6-sol (unseen)	TPR	246	78.5 / 90.7	98.8 / 100.0	98.8 / 99.6	1.2 / 81.7	100.0
gpt-6-astra (unseen)	TPR	250	62.8 / 74.8	91.6 / 99.2	83.2 / 96.4	0.0 / 67.6	100.0
claude-fable-5-1 (unseen)	TPR	244	51.2 / 65.6	27.9 / 79.1	17.2 / 35.2	0.0 / 7.4	99.2
claude-opus-5 (unseen)	TPR	245	51.0 / 65.7	49.4 / 84.9	41.2 / 64.1	0.0 / 24.9	99.2

J Feature Analysis Details

J.1 Feature discovery

We discover features from the outline items of in-domain test documents, separately for each of nine groups: the eight formats (Table 12) and one format-agnostic group. For each group, the contrast set is the 5,000 items IdeaLens scores as most AI-like and the 5,000 it scores as most human-like, balanced across outline roles. GPT-5.6 Sol (OpenAI, 2026a) (*max* effort) reads each contrast set and proposes features that separate its two sides, giving for each a name, a description, the direction it points (human or AI), its value type (Table 48), and how to assign it to an item. This yields 207 features, 23 per group on average. GPT-6 Astra (OpenAI, 2026b) (*extra-high* effort) then merges duplicates and keeps the features proposed across several groups, leaving 79 (see Table 49 and Table 50 for examples); Table 47 shows how many each group contributed.

J.2 Feature assignment

We use a held-out set of documents for feature value assignment. We sample 50 documents from each (format, Pangram silver label) pair, yielding a balanced set of 800 documents from 16 groups. Documents which contain at least one item that was used during the feature discovery process were excluded from this selection process, to guarantee that we are doing this experiment on a held-out set. Then, we assign values for all 79 features on the 13,616 items from these 800 documents, by prompting Gemini 3.8 Flash (Google DeepMind, 2026d) with *medium* effort to rate each item on each of the features, according to each feature's definition and output scale. We use these feature assignment results for the quantitative analysis in §J.3.

To verify the effectiveness of this method, we also perform the identical feature assignment task using GPT-5.6 Luna (OpenAI, 2026a) with *medium* effort. Given the two assignment results, we compute agreement metrics and present them in Table 45. Both raters show a similar number of features firing per item, and a similar distribution of Human-association Score ($r = 0.85$). Gwet's AC1 (Gwet, 2014) indicates a substantial agreement, while Krippendorff's α (Krippendorff, 2018) does not. This is explained by both raters assigning zero so often, while AC1 gives more credit for these matches but α treats them as less informative about raters' agreements. Because individual items naturally have non-zero values for only a small subset of features, we prioritize the AC1 metric and interpret this result as indicating moderate to high agreement.

J.3 Verifying the legitimacy of our discovered features

We conduct quantitative analysis to verify whether our features align with their proposed directions, and whether they explain our model's behavior. We first assign values for all 79 features on 13,616 held-out items. To quantify the extent of a certain feature pointing towards human ideas, we use Human-association Score, the chance that an item firing[17] on that feature is detected as more human-like by IdeaLens, than an item not firing on that feature.[18] We find that 64 of the 79 features clearly align with their proposed direction, while only one feature shows a conflict (Table 46, Figure 5, Figure 6). We also observe that features whose proposed direction was AI tend to fire more on items that IdeaLens detected as more AI-like, and vice versa (Figure 7). This is consistent with IdeaLens relying on these features when making decisions.

[17] The definition of a feature *firing* on an item depends on the type of the feature, and is provided in Table 48.

[18] A Human-association Score near 1 marks a strongly human-leaning feature, near 0 an AI-leaning one, and 0.5 is neutral.

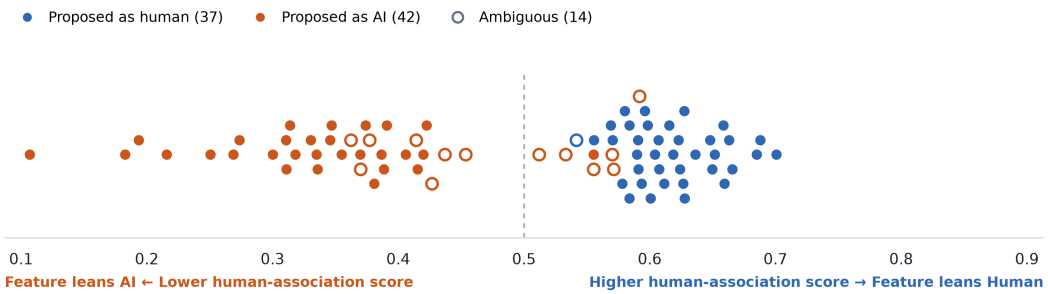


Figure 5: **Distribution of HUMAN-ASSOCIATION SCORES.** Each point represents one of 79 features; blue and orange indicate features proposed as human- and AI-associated, respectively. The score is the chance that, within the same format, an item where the feature fires receives a higher detector human-likeness score than an item where it does not fire. A score of 0.5 indicates no association. Hollow points have 95% bootstrap intervals that include 0.5; vertical placement is for visibility only.

Table 44: **Summary of the four feature families.** Confirmed features have 95% bootstrap intervals entirely on their proposed side of 0.5. Mean HUMAN-ASSOCIATION SCORES are reported for human- and AI-proposed features. Examples are the two strongest direction-aligned features firing on at least 100 items.

Difference	Confirmed	Mean HUMAN-ASSOCIATION SCORE	Strongest examples
Decisive vs. Universal Details	9/9	H: 0.64 A: 0.36	`uninstantiated_change_account` (AI); `interchangeable_remedy_bundle` (AI)
Functional vs. Apparent Specificity	12/16	H: 0.60 A: 0.35	`nonselective_analytic_agenda` (AI); `abstract_complementarity_as_mechanism` (AI)
Bounded vs. Overextended Claims	10/10	H: 0.62 A: 0.35	`unearned_legacy_magnitude` (AI); `unbounded_future_extension` (AI)
Preserved vs. Erased Costs	12/13	H: 0.61 A: 0.38	`guaranteed_independent_cobenefits` (AI); `contingent_exception` (HUMAN)

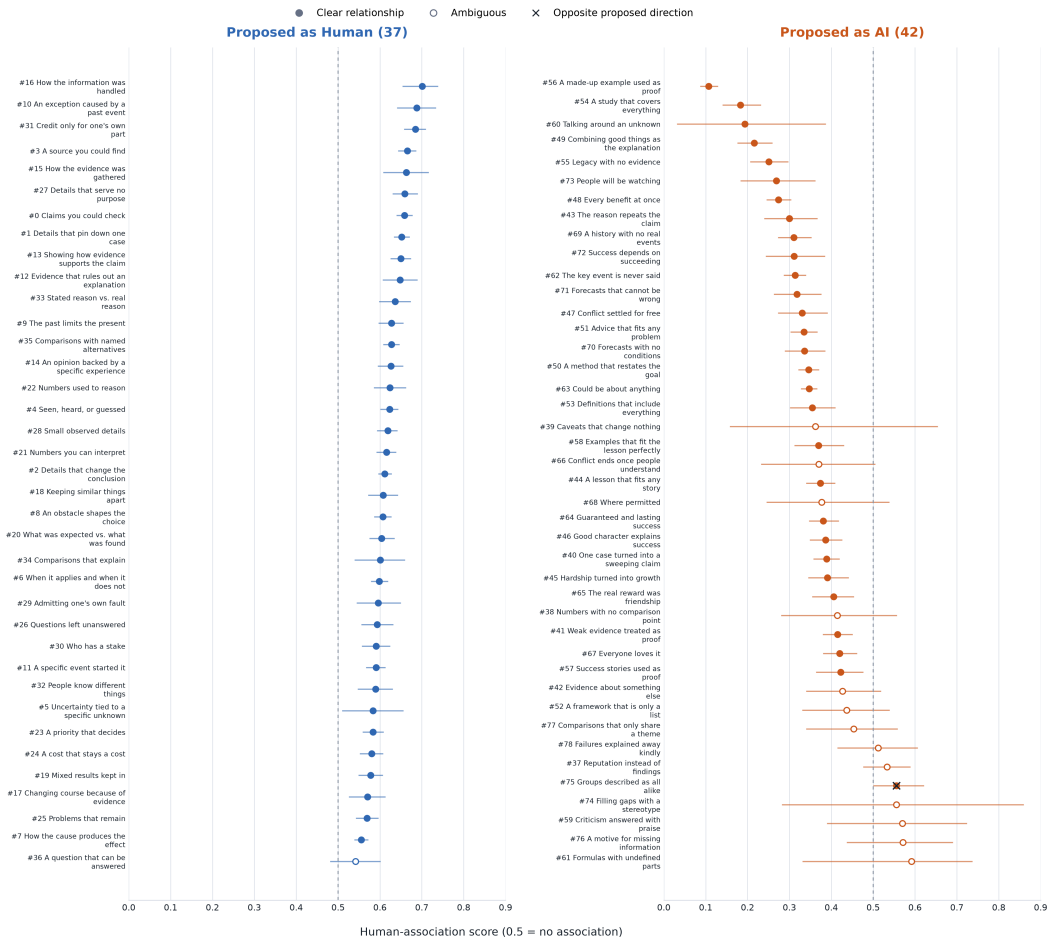


Figure 6: **HUMAN-ASSOCIATION SCORES for individual features.** Points show feature scores and horizontal lines show their 95% bootstrap intervals, grouped by proposed direction. Human-proposed features are ordered from highest to lowest score, and AI-proposed features from lowest to highest, placing each panel's strongest associations with its proposed direction at the top. Scores above 0.5 lean human; scores below 0.5 lean AI. Hollow points have intervals that include 0.5, and the cross mark indicates the feature whose observed lean opposes its proposed direction.

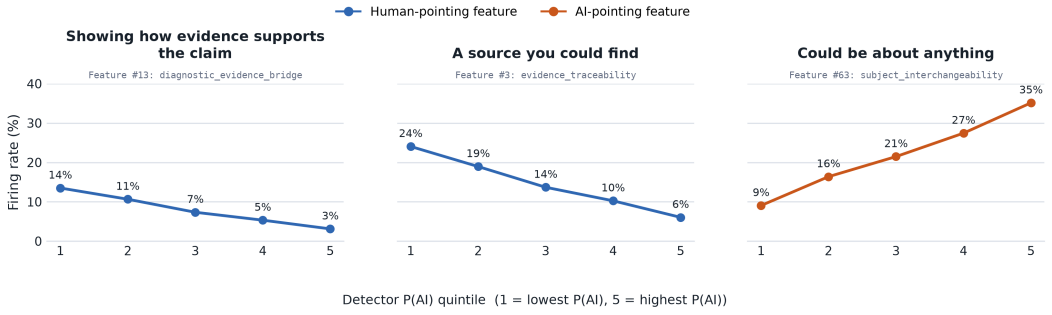


Figure 7: **Feature firing across detector P(AI) quintiles.** Items were ranked by detector P(AI) within each format and divided into five groups, from lowest P(AI) (1) to highest P(AI) (5). Each point shows the share of items in a group where the indicated feature fires. Firing falls across the quintiles for the two human-proposed features (#13 and #3) and rises for the AI-proposed feature (#63).

Table 45: **Comparison of independent feature assignments on the same 13,616 items from 800 documents.** An aligned feature has a 95% bootstrap interval entirely on its proposed side of the HUMAN-ASSOCIATION SCORE's neutral value, 0.5. A valid feature for Krippendorff's α is one that both models rated above zero on at least 100 items.

Statistic	**Gemini 3.8 Flash**	**GPT-5.6 Luna**
Features assessed	79	79
Mean features firing per item	4.54	5.39
Features aligned with proposed direction	64/79	60/79
Human-proposed features aligned	36/37	32/37
AI-proposed features aligned	28/42	28/42
Agreement in feature-level associations		
Same side of 0.5 for the HUMAN-ASSOCIATION SCORE	73/77 features (95%)	
Same confidence-interval-based observed lean	66/79 features (84%)	
Features with confidently opposite leans	0/79	
Correlation of HUMAN-ASSOCIATION SCORES	$r = 0.85$ (77 features)	
Agreement in same-item feature ratings		
Exact rating match	0.900	
Ratings jointly equal to zero	0.856	
Mean Gwet's AC1, all features	0.966 (79 features)	
Median Gwet's AC1, all features	0.990 (79 features)	
Mean usable Krippendorff's α	0.480 (on 61 valid features)	
Median usable Krippendorff's α	0.479 (on 61 valid features)	

Note. Two GPT-assigned features never fire, so their HUMAN-ASSOCIATION SCORES are undefined; comparisons of HUMAN-ASSOCIATION SCORES use the other 77. The observed lean is human, AI, or unclear according to whether the 95% interval lies above, below, or across 0.5. AC1 includes common zero ratings, whereas α more strictly measures whether both models distinguish the same items. The comparison therefore provides stronger evidence for the aggregate feature-direction pattern than for interchangeable item-level ratings.

Table 46: **Observed directions of the 79 features after assignment.** A feature is aligned when its 95% bootstrap interval for HUMAN-ASSOCIATION SCORE lies entirely on its proposed side of 0.5 (above for human, below for AI), opposite when it lies entirely on the other side, and ambiguous when it includes 0.5.

Proposed	**Number**	**Aligned**	**Ambiguous**	**Opposite**
Human	37	36	1	0
AI	42	28	13	1
Total	79	64	14	1

Table 47: **Numbers of consolidated features contributed by each discovery group.** A feature may originate from multiple groups, so row totals exceed the 79 unique features.

	Mixed group	Personal blogs	Creative writing	Personal about pages	News	User reviews	Nonfiction	Academic writing	Knowledge articles
Human	**12**	15	15	14	13	15	12	11	12
AI	**8**	8	7	9	8	9	12	12	12

Table 48: **Output types of the 79 consolidated item-level features.** A feature fires on an item when its assigned value meets the rule in the final column; fractional outputs may take any real value in $[0, 1]$.

Output type	Number of features	Possible values	Fires when
`binary_0_1`	5	$0, 1$	$x = 1$
`ordinal_0_3`	73	$0, 1, 2, 3$	$x \geq 2$
`fraction_0_1`	1	$x \in [0, 1]$	$x > 0.5$

Table 49: **Top 10 features proposed as human-pointing** with a fire rate greater than 1.5%, ordered from highest to lowest HUMAN-ASSOCIATION SCORE.

No.	Plain-language feature	Raw feature	Proposed	Observed	Fire rate	Score	95% interval
31	Credit only for one's own part	`bounded_credit_allocation`	Human	Human	6.71%	0.68	[0.66, 0.71]
3	A source you could find	`evidence_traceability`	Human	Human	14.59%	0.67	[0.64, 0.69]
27	Details that serve no purpose	`incidental_contingency`	Human	Human	3.93%	0.66	[0.63, 0.69]
0	Claims you could check	`claim_adjudicability`	Human	Human	84.78%	0.66	[0.64, 0.68]
1	Details that pin down one case	`contingent_factual_anchoring`	Human	Human	56.44%	0.65	[0.63, 0.67]
13	Showing how evidence supports the claim	`diagnostic_evidence_bridge`	Human	Human	7.98%	0.65	[0.63, 0.67]
33	Stated reason vs. real reason	`public_private_motive_divergence`	Human	Human	1.51%	0.64	[0.60, 0.67]
9	The past limits the present	`inherited_constraint_dependence`	Human	Human	3.87%	0.63	[0.60, 0.66]
35	Comparisons with named alternatives	`comparison_baseline_resolution`	Human	Human	9.61%	0.63	[0.61, 0.65]
14	An opinion backed by a specific experience	`situated_experiential_warrant`	Human	Human	10.43%	0.63	[0.59, 0.66]

Table 50: **Top 10 features proposed as AI-pointing** with a fire rate greater than 1.5%, ordered from lowest to highest HUMAN-ASSOCIATION SCORE.

No.	Plain-language feature	Raw feature	Proposed	Observed	Fire rate	Score	95% interval
48	Every benefit at once	`guaranteed_independent_cobenefits`	AI	AI	2.44%	0.27	[0.24, 0.30]
69	A history with no real events	`uninstantiated_change_account`	AI	AI	1.61%	0.31	[0.27, 0.35]
62	The key event is never said	`central_semantic_undercommitment`	AI	AI	5.21%	0.31	[0.29, 0.34]
51	Advice that fits any problem	`interchangeable_remedy_bundle`	AI	AI	2.43%	0.33	[0.30, 0.37]
50	A method that restates the goal	`goals_substituted_for_operations`	AI	AI	4.45%	0.35	[0.32, 0.37]
63	Could be about anything	`subject_interchangeability`	AI	AI	21.90%	0.35	[0.33, 0.37]
53	Definitions that include everything	`boundary_free_construct`	AI	AI	1.56%	0.35	[0.30, 0.41]
44	A lesson that fits any story	`generic_symbolic_lesson`	AI	AI	3.47%	0.37	[0.34, 0.41]
64	Guaranteed and lasting success	`contingent_outcome_assurance`	AI	AI	2.35%	0.38	[0.35, 0.42]
46	Good character explains success	`moralized_causal_shortcut`	AI	AI	2.06%	0.39	[0.35, 0.43]

K PROMPTS

Each box condenses one prompt to its instructions and inputs; placeholders are in {{braces}}. The full prompts, their output schemas and the extraction exemplars' reference outlines are in the repository (https://anonymous.4open.science/r/IdeaLens-0F91); the exemplar documents, which come from the training corpus, are released with the corpus.

K.1 OUTLINE PIPELINE

Format gate (Gemini 3.7 Flash)

Classify the document into one of WebOrganizer's 24 format categories (letter choices A–X), reading the whole document. Only the eight long-form formats with a role vocabulary continue; the rest are out of scope.
Input: {{DOCUMENT}}.

Role discovery, step 1: base role proposal

From general knowledge of {{FORMAT}} and its description, propose an inventory of the recurring functions that substantive, rhetorical and bounded structural items perform in such documents. No example documents are shown. Roles must be functions, not topics or surface formatting, must recur across subtypes and subject areas (a topic list is given as a generality check only), and each comes with a definition, an assignment test and its nearest functional distinction.

Role discovery, step 2: outline and role assignment (200 documents per format)

Extract a flat, ordered outline of the document's latent plan (what the writer chose to say and in what order, abstracted from wording). Then assign each item one primary role, preferring a base role whenever it fits and inventing a new candidate role only when every base role would misstate the item's function. Roles are decided in context: what the item contributes and why it appears at that point.

Role discovery, step 3: consolidation

Given every base and invented candidate with its definition, assignment test, usage count and real example items, merge candidates that name the same function, drop rare or topic-specific ones, and return a parsimonious final vocabulary that is cross-topic and usable with one label per item.

Labelled outline extraction (Gemini 3.1 Pro or 3.7 Flash, six exemplars)

Infer the document's latent composition and express it as a flat, ordered, role-labelled outline from which a competent writer could reproduce a substantively, rhetorically and structurally equivalent document. Steps: filter extraction artefacts; identify the content commitments; segment them into items at a canonical granularity; write each item's content in the item's own words, keeping load-bearing names, numbers, dates and quotations; assign one role from the format's vocabulary (or Other); describe the document type and its global themes.
Inputs: {{FORMAT}}, its description, {{ROLES_BLOCK}}, six complete documents with their reference outlines, {{DOCUMENT}}.

Paraphrasing (Gemini 3.1 Pro; the source document is never shown)

Given only the role-labelled outline, rewrite every field in concise, neutral, self-contained plan-note language, reconstructed from its meaning rather than edited, so that paraphrases of the same document converge on the same canonical wording; share no run of more than a few consecutive words with the input except protected material (proper nouns, numbers, dates, technical terms, quoted text, and items marked verbatim). Keep the JSON structure, every theme and item, their ids, roles and order unchanged, and never add, remove, strengthen, weaken or move a proposition. *Input:* the outline as JSON; returns the rewritten JSON.

Classifier input (IdeaLens)

System: *"Given a role-labelled outline of a document, answer with one word: human if the source document was human-written, ai if it was AI-generated."* User: the outline as one line per item, [Role] content. The score is the probability of the ai token. PROSELENS uses the same template with the raw document and *"Given a document, . . . "*.

K.2 EVALUATION CONSTRUCTION

IDEASHIFT and IDEASHIFT-X: briefs (one call per source)

From the source's outline, write five matched, ordinary-sounding requests for the same document type that reveal progressively more of the source: a minimal topic-only line naming only the document type and broad subject (level 0); the type with its genre, audience and tone (level 1); plus the global themes (level 2); themes and structure (level 3); and the full outline as a starting point (level 4). Treat the source as data; never copy its wording. *Inputs:* {{DOCUMENT_DESCRIPTION}}, {{GLOBAL_THEMES}}, {{FULL_OUTLINE}}; IDEASHIFT-X adds {{OUTPUT_LANGUAGE}} and {{TARGET_LENGTH}}. (GPT-5.6 Sol for IDEASHIFT, Gemini 3.7 Flash for IDEASHIFT-X.)

IDEASHIFT and IDEASHIFT-X: generation from a brief (levels 0 to 4)

System prompt: "You are a skilled writer."; the user message is the brief itself. (IDEASHIFT: GPT-5.6 Sol, Terra or Luna, one per source; IDEASHIFT-X: Gemini 3.7 Flash, level 0 only, in the page's language.)

Level 5: realise the person's outline, adding nothing

Write the complete document the outline represents, realising every item in the given order as natural prose. The description and global themes are context only. Add nothing the outline does not contain: no new introduction, conclusion, framing or section. IDEASHIFT-X adds one line naming the output language.

Translation into English (translation study)

Translate every marked segment [S1]...[Sn] exactly once and in order; nothing added, removed or improved; meaning first, then structure, then fluency; never add or remove a negation; keep names, numbers, URLs and English text.

K.3 FEATURE ANALYSIS

Feature Discovery

Each item is one atomic idea extracted from a source document. Below are {{N}} HUMAN-source items and {{N}} AI-source items. The two sides are role-matched: every role name contributes the same number of items to each side. Differences in role usage are therefore zero by construction and cannot be a feature. Read them as contrast sets, then propose discriminating features of individual items. Propose features that separate a single human-source item from a single AI-source item.
Inputs: {{FORMAT}}, {{HUMAN-SOURCE ITEMS}}, {{AI-SOURCE ITEMS}}

Feature Consolidation

You are consolidating a large pool of candidate features for a human-vs-AI text classifier. The features were proposed independently by reading individual items of document outlines from 9 sources: eight document formats separately and one cross-format agnostic pass. Your job is to merge the pool into a clean, deduplicated, canonical feature bank. Every canonical feature must be grounded in the raw features it subsumes.
Input: {{FEATURES}}

Feature Assignment

You will be given a numbered set of items. Rate each item on each of the following features, strictly per each feature's definition and output scale. Give each item exactly the ratings it would get if it were the only item you were shown. Judge an item only from its own format, role label, and text. They are listed in random order and may come from different documents. Do not use one item to explain, complete, support, or contradict another, and do not consider how items relate to, follow, or repeat each other.
Inputs: {{FEATURES}}, {{ITEMS}}

K.4 TwiceTold Curation

TwiceTold Instructions

Please write a 500-word story that follows the given outline as closely as possible. The outline includes the genre, a brief set of instructions (e.g., perspective, style, etc.), global themes and a list of 8-12 items describing the plot.
1. Retain the original order of events
2. Do not omit details provided in the outline
3. You may add minimal details to make stories coherent

Inputs: {{GENRE}}, {{DOCUMENT DESCRIPTION}}, {{GLOBAL THEMES}}, {{OUTLINE}}

K.4.1 Annotation Details

All the computer science graduate students recruited for this task consented to their stories being used for research purposes. Students took approximately 1-2 hours to complete writing each story, depending on the length and complexity of the outline.

L Where the remaining tables are

All other tables are in the anonymous repository (`https://anonymous.4open.science/r/IdeaLens-0F91`): every evaluation's full tables with all arms and length splits, every operating point (0.1, 0.5, 1, 2 and 5% false-positive targets) under the global cut, the per-format 1% cut on every arm (§E.2), every detector on every arm of every evaluation including the Qwen, logistic-regression and ModernBERT models, the similarity diagnostics, all prompts and role vocabularies, and the code.