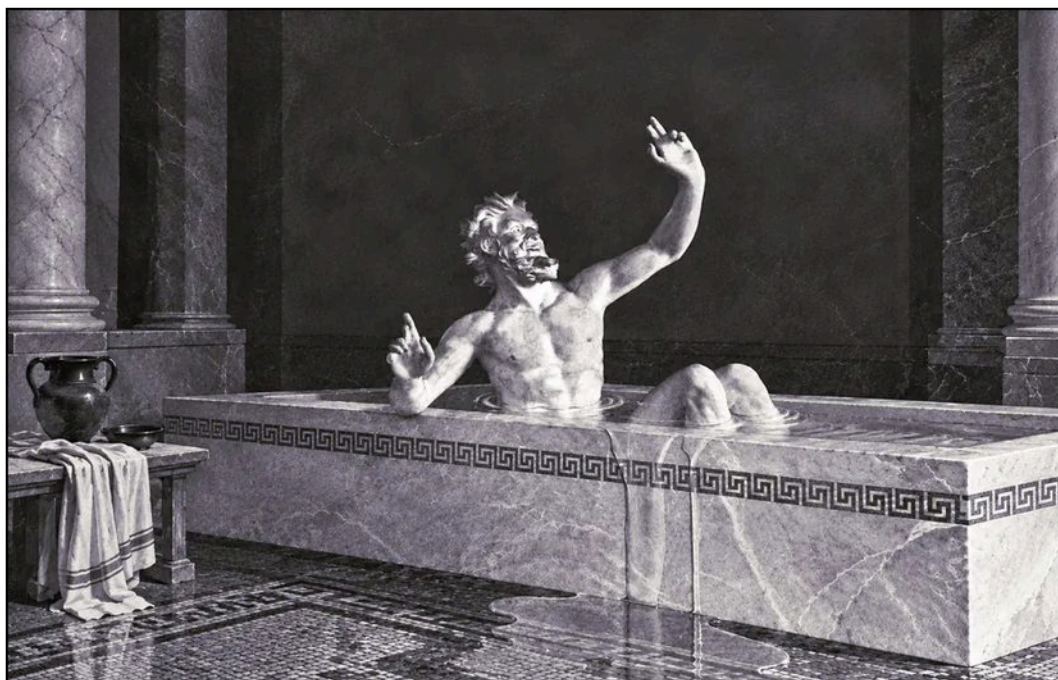


Detecting AI Ideas, Not AI Text

Originally published Tuesday 8 October 2026 at <https://www.unite.ai/detecting-ai-ideas-not-ai-text/>

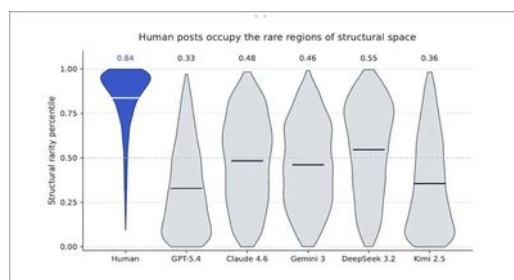


What if there was an AI text detector that could recognize that you wrote something based on an *idea* that AI gave you – even if all the actual writing (i.e., the text) was your own original work?

We all know about AI text detectors embarrassing [celebrities](#) and [politicians](#), among others – algorithms that have studied the characteristics of AI-created text and can [discern](#) those patterns when they surface.

These cases represent simple offloading of a communications task, wherein language models such as [ChatGPT](#) and [Google Gemini](#) [predict the next likely token](#) in a response to a user's prompt – not by taking a step back and looking at the broad canvas of the problem or proposition first, but just by guessing the next most probable word.

Nonetheless, this patchwork quilt of predictions does eventually yield an underlying structure, according to recent research: the need for adjacent text sections to be coherent and relevant to each other causes a unique signature to develop in the 'outline' or higher dimensionality of the work – [one that can discern the hand of AI](#) even when the text is extensively rewritten to 'humanize' it:

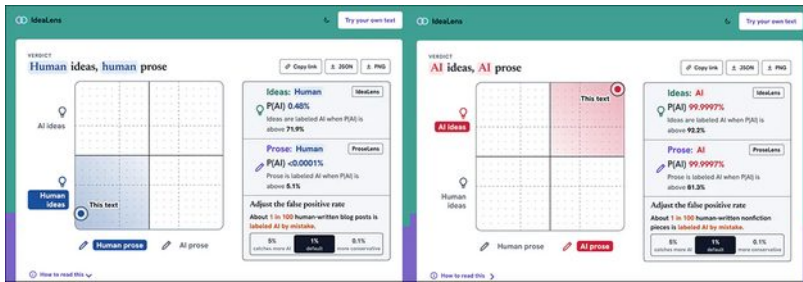


The distinct higher-dimensional 'shape' of writing under various LLMs, and – left – pure human writing. [Source](#)

AI Ideation Targeted

A new academic [collaboration](#) led by the University of Maryland goes a step further, by offering a methodology that can detect whether the *ideas themselves* originated with AI, regardless of who ultimately wrote the words.

The approach, dubbed *IdeaLens* and trained on one million outlines extracted from web documents, keys on reducing each document to an outline that preserves its *ideas* while stripping away the wording – and comes with an [online demo](#):



On the left, one of my own articles passes as human in the online demo, while another article I submitted (written by a different author) appears to have had consi

As seen in the sample demo results above, the detection methods used can distinguish between human ideas and human prose, and AI ideas and AI prose, and evaluate them in a granular and distinct fashion – even though the two detectors were trained on the same documents, labels and underlying model architecture, while being designed to detect essentially *opposite* things: the provenance of ideas, and the provenance of prose.

This kind of detection ambit, one suspects, could make a number of categories of writer nervous, from [political speech writers](#) through to [novelists](#) and [columnists](#), besides others who have, to date, prized their own writing style, yet allowed AI to fuel their *topics*.

One reason why LLMs are likely to produce recognizable (and therefore traceable) ideas is not just that certain concepts and texts will have statistically dominated the model's [under-curated training data](#), making them more likely to surface in a wider range of requests; but also, because LLMs have mysterious and repetitive obsessions that don't even seem to relate to specific training data, and yet [often appear unbidden](#).

The authors state*:

*'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.'

Data and Methods

The paper claims that the new method can identify human ideas in AI-generated documents (i.e., someone said to an AI 'Write me an essay on this topic'); and can also identify AI ideas in human writing (i.e., someone said to an AI 'Suggest something for me to write about').

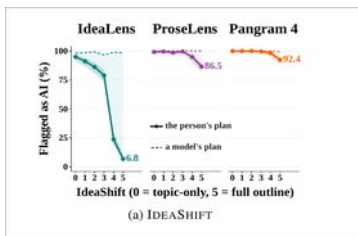


Test results across four combinations of human and AI ideas and prose. All four detectors perform strongly when ideas and prose share the same origin, but diverge mixed: IdeaLens correctly recognizes 94.7% of AI-written documents based on human ideas as human-originated, and identifies AI ideas in 68% of human-written for Pangram 4 and 0% for ProseLens and EditLens-3B. [Source](#)

Detection is easiest when provenance is clearest – for example, when a human-written piece is based on AI-generated ideas, or where an AI-generated output originates from a human idea. When these provenance strengths vary, detection becomes harder.

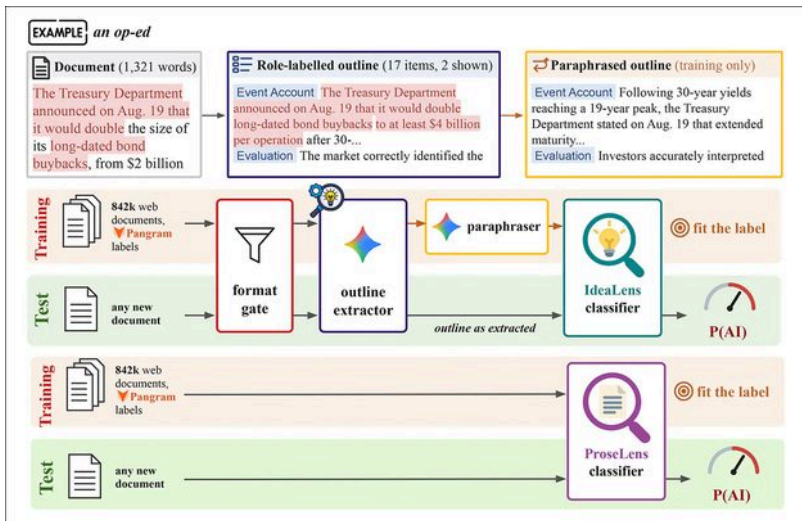
The researchers tested this distinction by gradually changing how much of a document's underlying plan comes from a person. In one experiment, AI models were given increasingly detailed human plans, ranging from not much more than a topic, to a complete outline.

As the human contribution increased, IdeaLens's AI flag rate fell from 94.9% to 6.8%; and conventional prose detectors continued overwhelmingly to identify the resulting text as AI:



Test results showing what happens as progressively more of an AI-written document's underlying plan is supplied by a human. IdeaLens increasingly recognizes the human origin of the ideas, with its AI flag rate falling from 94.9% for a topic-only prompt to 6.8% for a complete human outline; the conventional text detectors ProseLens and Pangram still classify most of the resulting writing as AI.

Since the ideas must be separated from the language used to express them, and there is no large dataset recording who actually originated the ideas in a document, the researchers instead trained IdeaLens on existing AI-writing [labels](#) – but removed most of the writing itself:



How IdeaLens separates ideas from prose during training and testing, compared with the conventional text detector ProseLens. A sample 1,321-word opinion article outline, with points categorized by their function, such as an event account, or evaluation. During training, these outlines were additionally paraphrased to remove source documents before IdeaLens learned from existing Pangram AI labels. For new documents, the extracted outline can then be passed directly to IdeaLens to ideas are AI-derived. The lower part of the figure shows a second detector – 'ProseLens' – used by the researchers for comparison, which instead receives the con whether the prose itself is AI-generated.

As shown above, each document was reduced to an outline describing what is being said, and the role each point plays in the document. During training, those outlines were paraphrased to further remove traces of the authors' original wording. IdeaLens can therefore make its predictions largely from the *ideas that remain*, rather than from telltale characteristics of the prose.

Three datasets were originated for the project: *IdeaShift*; *IdeaShift-X*; and *TwiceTold*. IdeaShift was created from 500 human-written and AI-generated seed documents sampled from the researchers' existing test corpus. GPT-5.6 Sol to turn each into progressively more detailed prompts, from 'topic-only', to a full outline. [GPT-5.6](#) was then used to generate new documents.

IdeaShift-X, instead, used 10,000 human-authored [FineWeb2](#) documents across 24 languages, applying the aforementioned IdeaShift protocol to generate AI versions in each language.

The third dataset, TwiceTold, was written by academic students under the supervision of the authors, to avoid the possible use of AI [inherent](#) in crowd-sourcing such data. Fifty stories were written from scratch, but using a 500-word outline provided by ChatGPT-5.6 Sol.

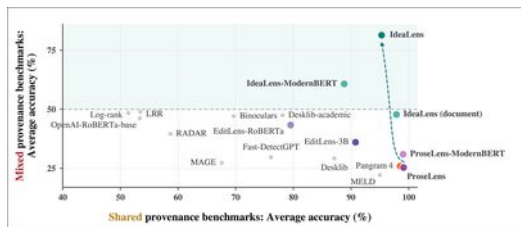
Tests

The researchers' own detectors comprised IdeaLens in outline and document form; the aforementioned ProseLens; [ModernBERT-L](#) versions of both systems; [Qwen3.5-9B](#) versions of IdeaLens; and an IdeaLens [logistic-classifier](#) variant.

These were compared against [Pangram 4](#); [EditLens-Llama-3B](#); [Binoculars](#); [EditLens-RoBERTa](#); [MELD](#); [Desklib-academic](#); [Desklib](#); [Entropy](#); [Fast-DetectGPT](#); [Likelihood](#); [Log-rank](#); [LRR](#); [MAGE](#); [OpenAI-RoBERTa-base](#); [OpenAI-RoBERTa-large](#); [RADAR](#); and [Rank](#).

For the tests, Accuracy was measured by whether the predicted AI or human label matched the source of the ideas. IdeaLens, ProseLens and EditLens used a global 1% FPR threshold; Pangram 4's AI, AI-assisted and human labels were binarized; and Fast-DetectGPT, Binoculars, MELD and Desklib used their default thresholds.

The wider evaluation used 19 external benchmarks, including 14 covering fully human or fully AI material, and 10 containing human writing subsequently edited by AI. A further 50,000 in-domain documents were tested, while 10,000 pre-ChatGPT C4 documents were used to assess false positives:



Test results comparing detector accuracy when ideas and prose have either the same or different origins. IdeaLens retains high accuracy in both settings, at 95.3% for shared provenance and 81.3% for mixed provenance; competing detectors perform substantially worse when the source of the ideas differs from the source of the prose.

Across 51 tests drawn from 22 benchmarks, IdeaLens was the only detector to perform strongly when the ideas and prose came from either the same or different sources, scoring 95.3% and 81.3% respectively.

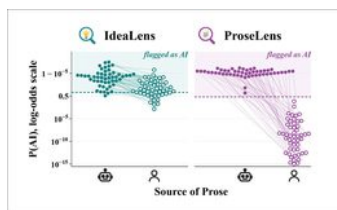
By comparison, ProseLens and Pangram 4 scored above 98% when ideas and prose shared a source, but fell to around 25% when they did not.

The authors state of these results*:

'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](#) 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.'

All three detectors successfully identified almost all of the original AI-written stories. On TwiceTold, IdeaLens identified AI-originated ideas in 68% of the human-written stories, compared to 0% for ProseLens and 8% for Pangram 4:



IdeaLens assigns similar AI probabilities to the original AI-written stories and the human-written TwiceTold versions based on the same AI-generated outlines, while ProseLens's AI probabilities fall sharply once the prose is rewritten by humans.

Despite being trained in English, IdeaLens generalized across IdeaShift-X's 24 languages, flagging 95.3% of documents based on AI ideas – but just 0.7% when a detailed human plan supplied the ideas. On 10,000 pre-ChatGPT human documents across the same languages, the false-positive rate was a mere 0.1%.

The authors concede that a larger dataset would be desirable in future work, and that the noisiness of the dataset labels could need addressing. Nonetheless, those that wish to build on this interesting initial outing can resort to the [GitHub repository](#) that the paper's authors have made available; and the merely-curious can paste 'suspect' tests into the project's [demo site](#), and gauge for themselves if the method jibes with their own estimations, or those of other approaches.

Conclusion

Whether or not a detection method of this kind *matters* depends on how acceptance, integration, or rejection of AI will develop (no doubt differently across different sectors and circumstances) over the next 6-12 months.

It may be that a politician's public are more forgiving of his or her use of AI to polish and represent their own original ideas, than they would be of a glib presentation of ideas generated by an LLM (after all, statistically, the audience is [likely to be sympathetic](#) to the use of AI as a mere presentation aid).

Yet, at the same time, what are we hoping for from AI if not that it will produce angles, ideas and alternatives that we ourselves would have overlooked or never considered; and from that point of view, is it such a bad idea to give machine intelligence a turn at the wheel?

Well, in terms of optics, and given AI's famous and growing propensity for errors and –lately – [errant behavior](#), maybe it's still too early to let LLMs originate policy suggestions, or in general become a driving force in ideation.

* Authors' emphases, not mine, but my conversion of the authors' inline citations to hyperlinks, where necessary.

First published Tuesday, October 6, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Website: martinanderson.ai

Contact: martin@martinanderson.ai