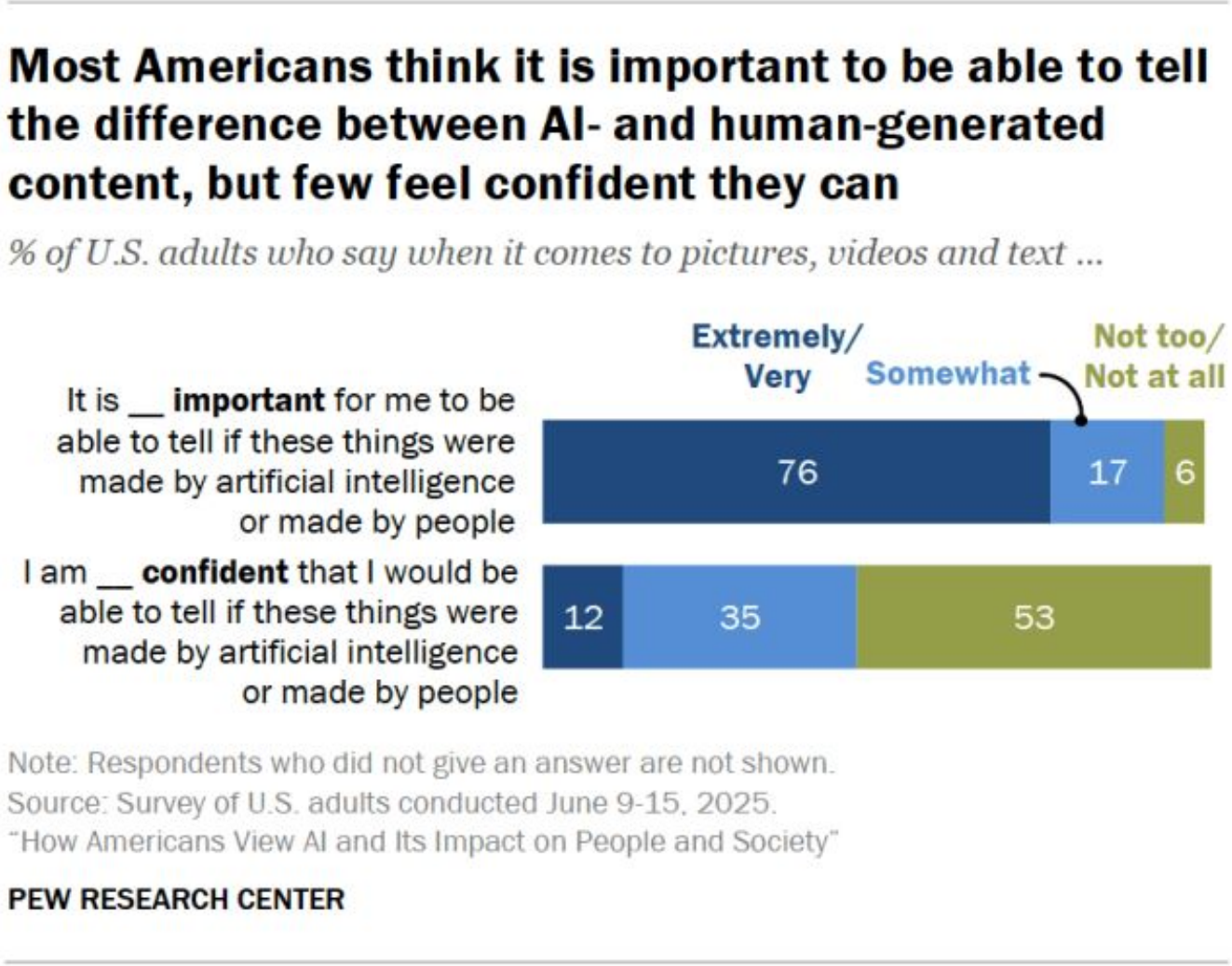
Can you tell AI-generated text apart from human-written text?

Jenna Russell
University of Maryland

Could you tell AI writing from human writing?



Pew Research Center, 2025

AI text is already in what you read



NATIONAL

Trump's Massive Bill Turns America into a Police State

Rep. Steven Horsford: 'It's Reckless Extremism'

By Stacy M. Brown
WI Senior Writer

President Donald Trump's massive new spending measure "One Big Beautiful Bill," known by Democrats as the "One Big Ugly Bill," is set to dismantle central pillars of the American social safety net and create the most expansive immigration enforcement system the country has ever seen.

The legislation advances $1 trillion in cuts to Medicaid and the Affordable Care Act over the next decade, threatening health coverage for millions of low-income Americans. However, among the least discussed aspects of the measure, it effectively turns America into a police state.

The bill devotes more than $150 billion to expand the detention, surveillance, and deportation operations of Immigration and Customs Enforcement (ICE). Under the bill, ICE's detention budget alone would explode from $3.4 billion this year to $45 billion by 2029, surpassing the combined funding of all federal prisons.

Public records show that ICE's arrests of immigrants without criminal records have surged by more than 1,400% in the past year. The agency has repeatedly deployed masked officers and National Guard troops in workplace and community raids, tactics critics say resemble authoritarian crackdowns.

The spending package also authorizes the creation of a federal citizenship registry, combining Social Security and immigration data into the first centralized list of U.S. citizens. Officials claim the database will help election authorities verify voter eligibility, echoing Trump's unsubstantiated claims of widespread immigrant voter fraud. Privacy experts have warned that this system amounts to a national surveillance apparatus launched without any public debate or congressional hearings.

"This level of integration among federal agencies handling sensitive personal data has never existed before," NPR reporters Jude Joffe-Block and Miles Parks reported. They noted that privacy advocates and even some conservatives have long opposed a government-run citizenship roster.

Polling from Pew Research shows that most Americans oppose several core pieces of Trump's immigration agenda. Sixty-one percent reject the deportation of immigrants to prisons in El Salvador. Fifty-four percent disapprove of increased workplace raids. Yet a growing share—56%—now supports expanding the border wall with Mexico, a rise from 46% in 2019.

Inside the Republican Party, support for Trump's enforcement push is nearly unanimous. Eighty-eight percent of Republicans favor expanding the border wall. Further, 81% back using state and local police to help deport immigrants. Among Democrats, overwhelming majorities oppose the policies, with only 27% supporting any expansion of the wall.

The bill's most controversial provisions have drawn condemnation after Trump visited a new detention facility in the Everglades, where he joked that alligators and snakes would attack any immigrant attempting to escape. The Florida Republican Party has since begun selling "Alligator Alcatraz" merchandise. Historians have noted that such imagery recalls Jim Crow-era propaganda depicting Black children as bait for alligators—a tactic used to dehumanize and terrorize Black communities.

Despite widespread criticism, Trump has pressed forward with the measure, calling it a fulfillment of his long-standing promises.

"This is going to be the bill that finally puts America first," Trump said during a rally in Florida. "No excuses."

Nevada Democratic U.S. Rep. Steven Horsford called the legislation a direct attack on the health, safety, and dignity of Black communities across this country.

"Stripping away Medicaid coverage while throwing billions into an unaccountable immigration dragnet is not fiscal responsibility—it's reckless extremism," Horsford stated. "We will not be silent as this Congress tries to reverse decades of progress and push people further into poverty and fear."

WI

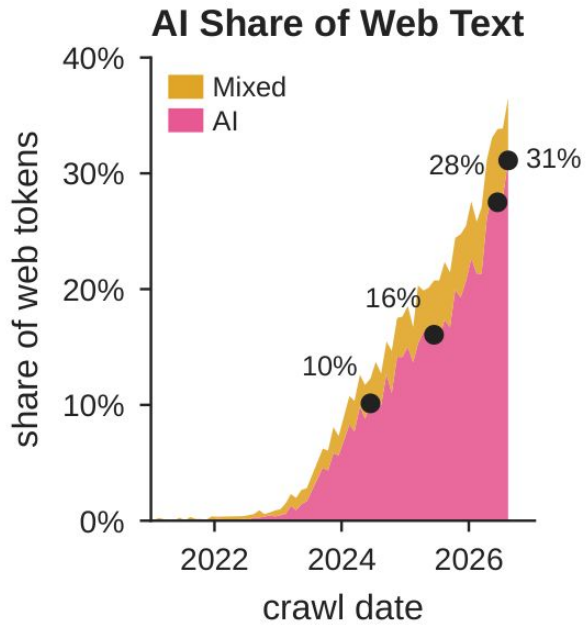
Per the current law, all District employees, including those receiving the tipped worker base pay of $10, are required to be compensated at the full minimum wage.

INITIATIVE 82 *from Page 1*

INITIATIVE 82 *Page 37*

16 JULY 10 - 16, 2025 THE WASHINGTON INFORMER / WWW.WASHINGTONINFORMER.COM

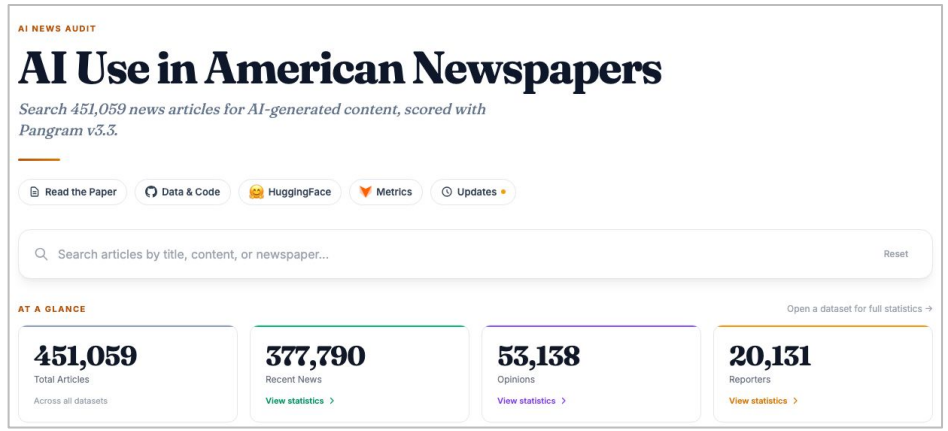
AI-flagged article in print



AI share of web text



4 of 10 prize winners flagged as AI



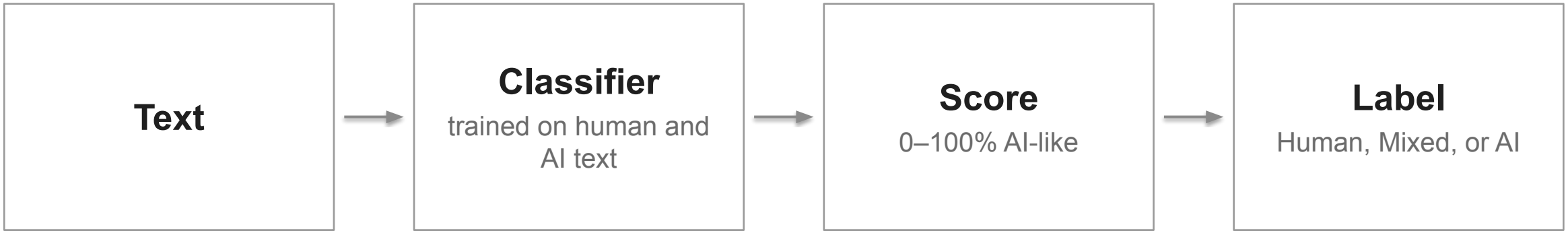
ainewsaudit.github.io

[Russell et al., ACL 2026](); [Russell et al., 2026](); [Pangram blog, May 2026]()

Today

1. How do AI detectors work, and how often are they wrong?
2. How much AI text is out there, and what does it do?
3. How is AI writing different from human writing?
4. Can a model tell who wrote a text?
5. What does this mean for how you use AI?

Detectors are typically trained classifiers



Good detectors almost never flag human writing

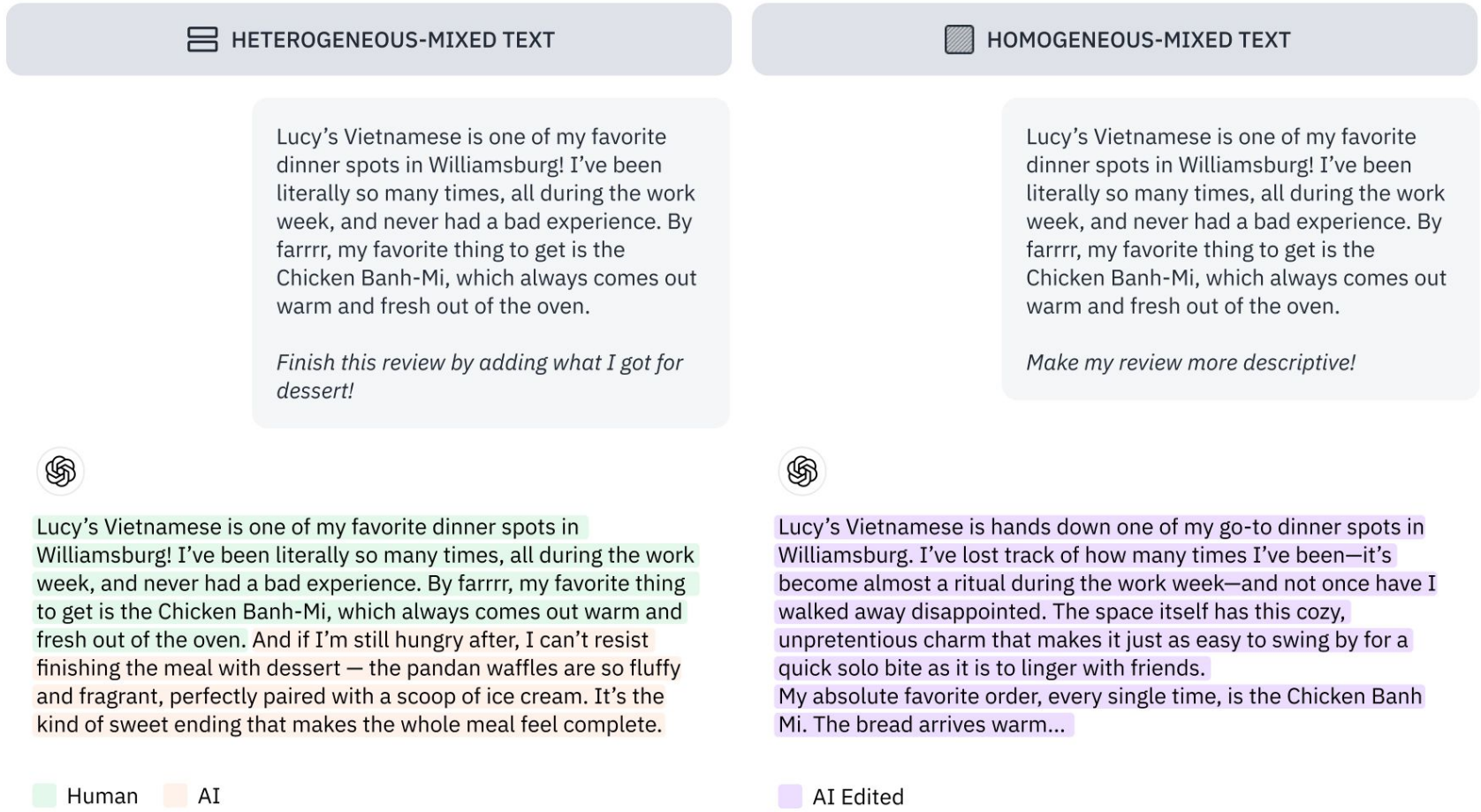
	Detector says human	Detector says AI
Written by a person	correct	false positive
Written by AI	false negative	correct

	Pangram 3.3.2	Pangram 4
False positive rate 1M+ human texts	0.054%	0.0041%
False negative rate 520K AI texts, 26 models	1.99%	0.34%

1 false flag per 24,300 human documents

Glickenhaus et al., 2026

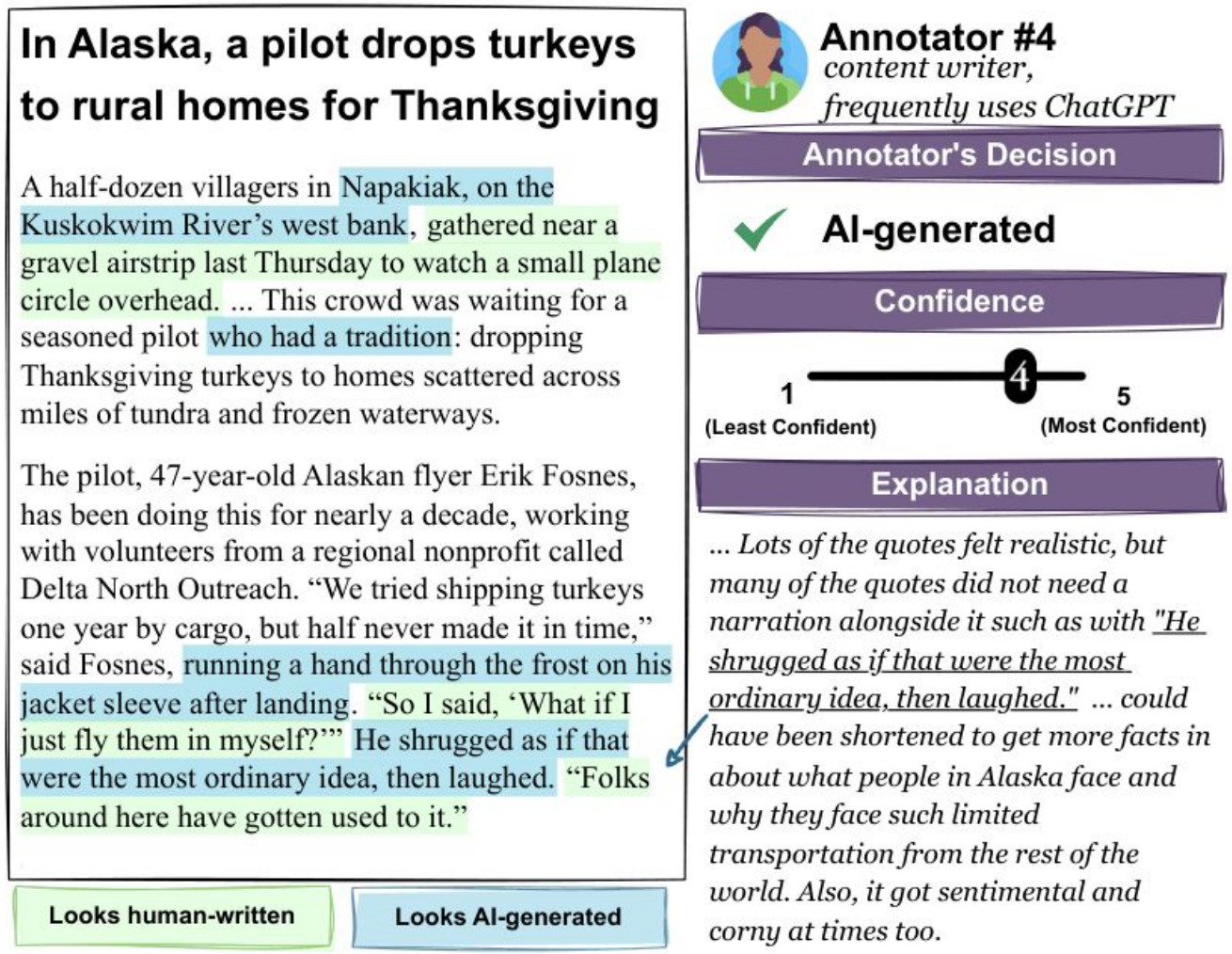
Real documents are often part human, part AI



97.7% of "humanized" AI text still caught

[Glickenhaus et al., 2026]

Readers labeled each article and highlighted the clues



Each reader answered

1. **Human or AI?**
2. **How sure, from 1 to 5?**
3. **Which words gave it away?**

People who use ChatGPT a lot are accurate detectors

99.3%

caught by frequent ChatGPT users

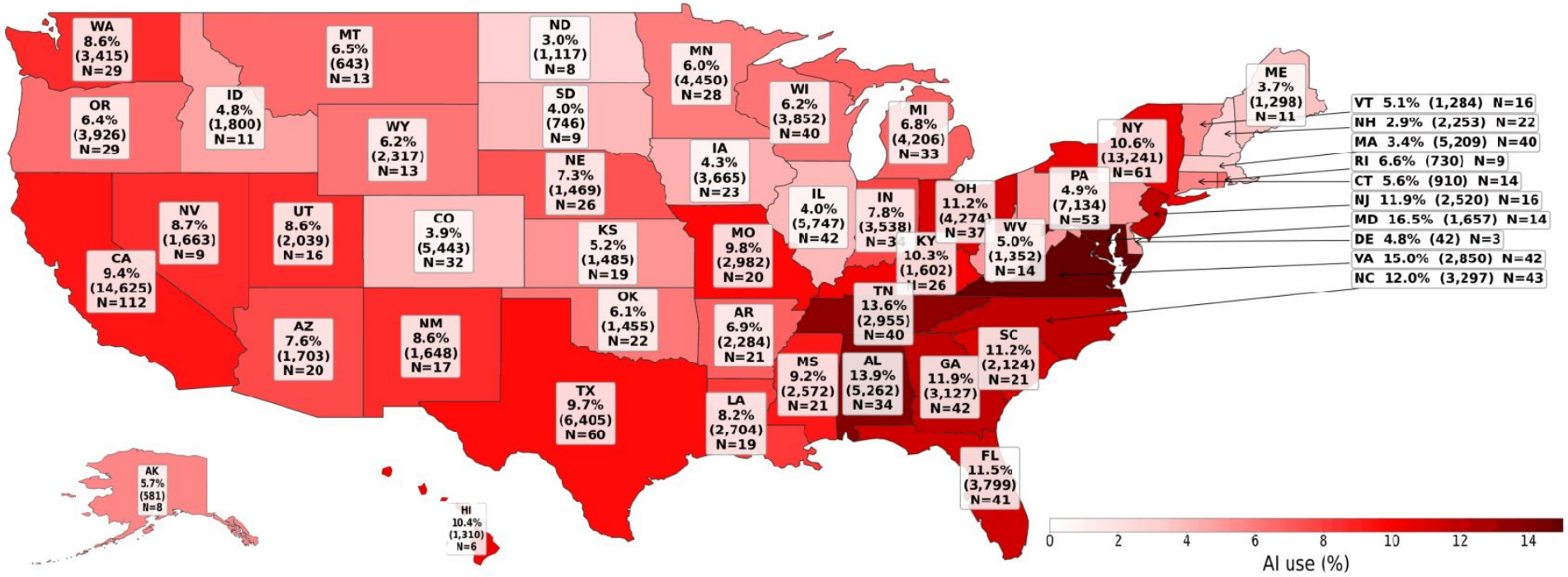
56.7%

caught by people who rarely use it

Detection Method	Generation Method					Overall
	GPT-4o	Claude	GPT-4o paraphrased	o1-Pro	o1-Pro humanized	
	TPR% (FPR%)	TPR% (FPR%)	TPR% (FPR%)	TPR% (FPR%)	TPR% (FPR%)	TPR% (FPR%)
(A) Expert human detectors						
Expert Majority Vote	100 (0)	100 (0)	100 (0)	96.7 (0)	100 (0)	99.3 (0)
Annotator 1	96.7 (3.3)	100 (0)	100 (0)	96.7 (6.7)	90.0 (23.3)	96.7 (6.7)
Annotator 2	96.7 (0)	80.0 (30)	86.7 (10)	90.0 (10)	86.7 (10)	88.0 (12)
Annotator 3	86.7 (6.7)	100 (0)	93.3 (0)	16.7 (0)	0 (3.3)	59.3 (2)
Annotator 4	90.0 (6.7)	96.7 (13.3)	100 (10)	100 (0)	100 (0)	97.3 (6)
Annotator 5	93.3 (0)	93.3 (6.7)	93.3 (0)	93.3 (0)	93.3 (0)	93.3 (1.3)
(B) Automatic detectors						
Pangram Humanizers	100 (0)	100 (3.3)	100 (0)	100 (0)	96.7 (10)	99.3 (2.7)
Pangram	100 (0)	100 (3.3)	100 (0)	100 (0)	90.0 (6.7)	98.0 (2)
GPTZero	100 (0)	96.7 (0)	100 (0)	76.7 (0)	46.7 (3.3)	85.3 (0.7)
Fast-DetectGPT (FPR=0.05)	100 (0)	96.7 (3.3)	56.7 (3.3)	86.7 (0)	23.3 (3.3)	80.0 (7.2)
Binoculars (Accuracy)	100 (0)	93.3 (0)	60.0 (6.7)	73.3 (0)	6.67 (0)	66.7 (1.3)
Binoculars (Low FPR)	96.7 (0)	80 (0)	13.3 (0)	10.0 (0)	0 (0)	40.0 (0)
RADAR (FPR=0.05)	66.7 (0)	0 (0)	10 (3.3)	0 (3.3)	0 (3.3)	15.3 (2)

Russell, Karpinska & Iyyer, ACL 2025

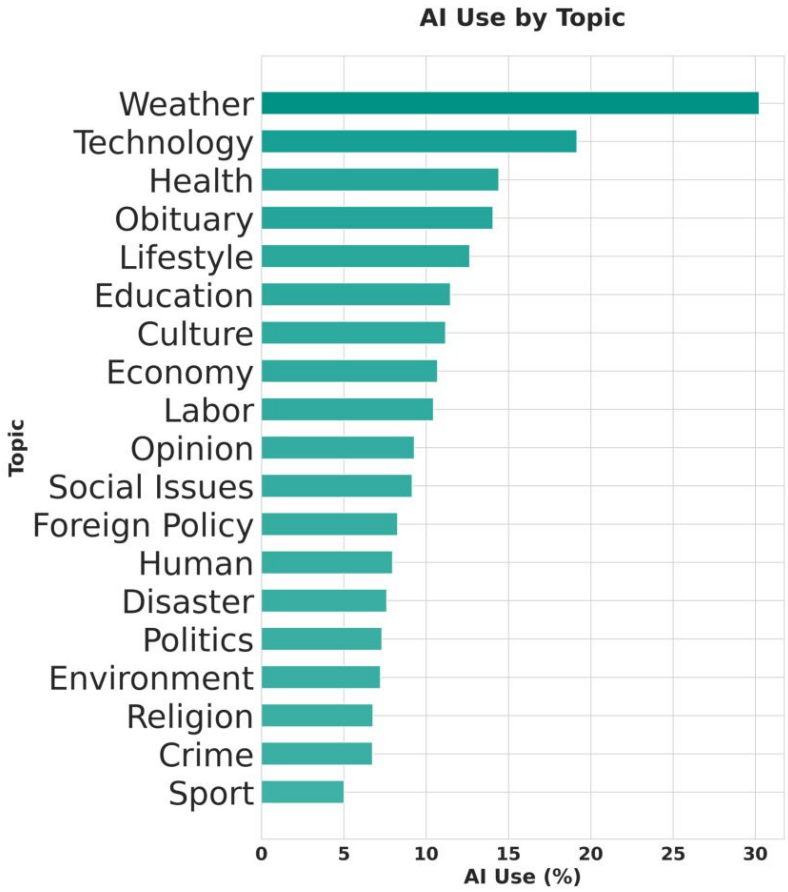
About 9% of new U.S. newspaper articles are partly or fully AI



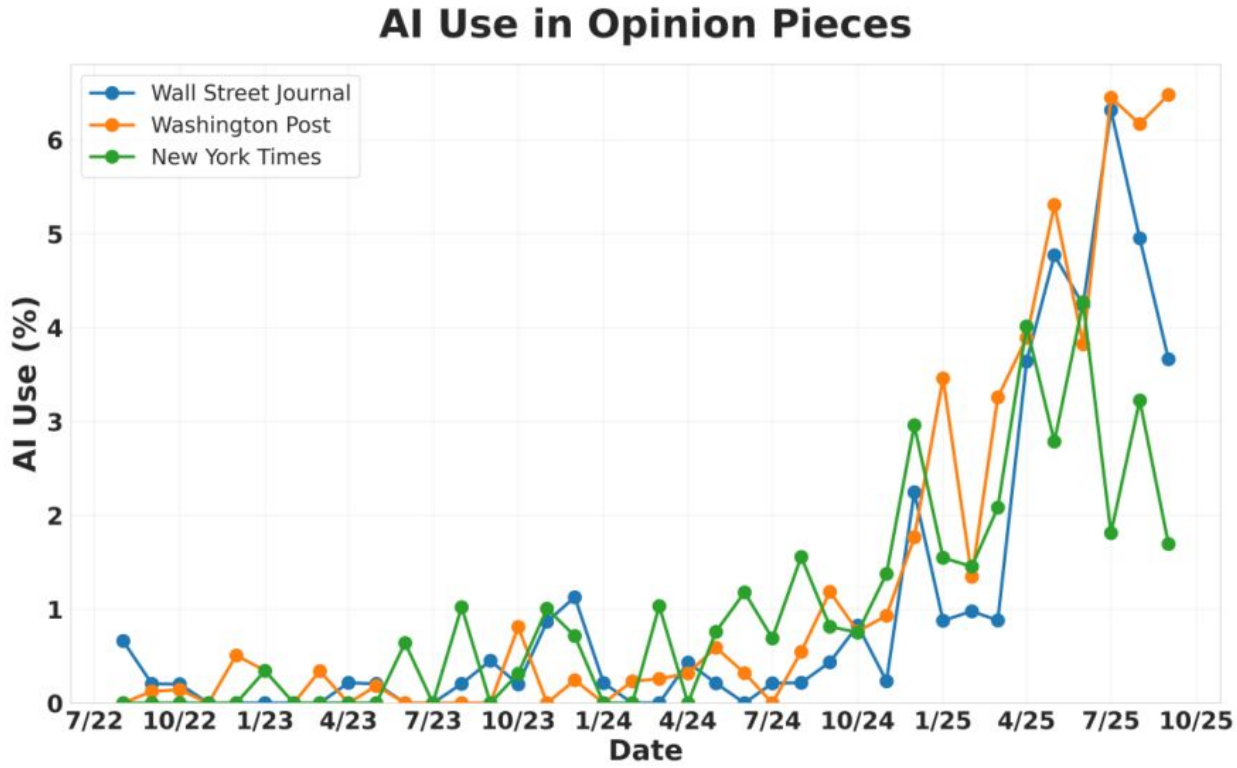
Local papers 10.8% vs. 0.5% at NYT, WSJ, and WaPo

Russell et al., ACL 2026

AI shows up in routine beats, and increasingly in op-eds



Weather: 27.7% AI



Big-paper op-eds: 0.1% (2022) to 3.4% (2025)

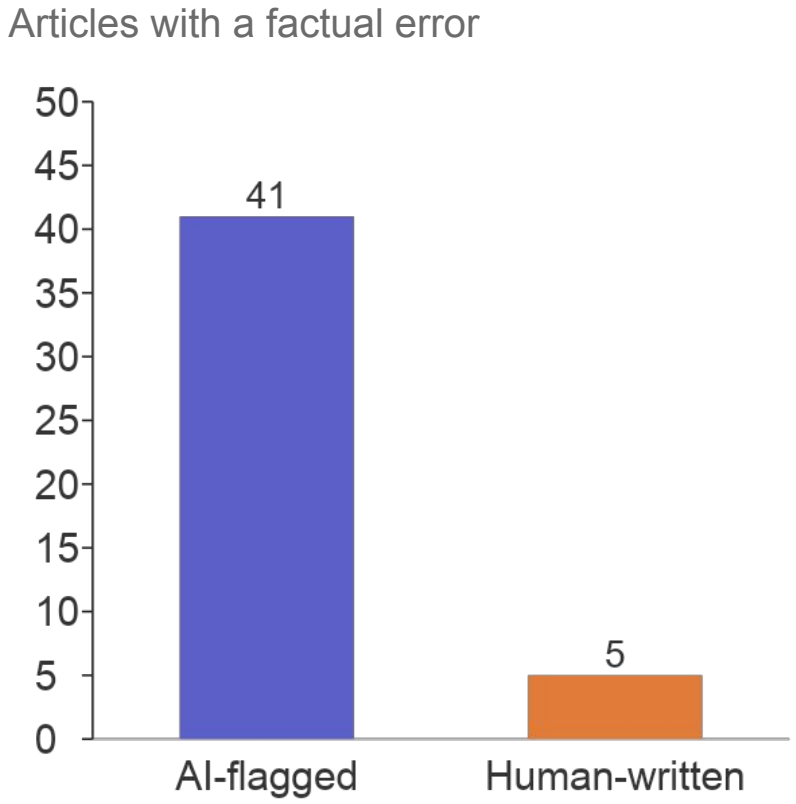
Russell et al., ACL 2026

Readers are rarely told, and AI-flagged articles had more errors

Policy Category	Freq	Example Publications / Owners	Illustrative Disclosure Text
AI Allowed	7	*Ionia Sentinel-Standard*, *Amsterdam News*, *Penn Live*	*"If AI-assisted content is approved for publication, journalists must disclose the use of AI and its limitations to their audience. AI-generated content must be verified for accuracy and factuality before being used in reporting."*
AI Prohibited	2	*New York Post*, *Michigan Daily*	*"The use of generative artificial intelligence for content production (including written, visual and auditory content) is unacceptable in all circumstances. Any staffer found to have used generative AI to produce content for The Daily can be fired by their section editor or the Editor in Chief."*
No Public Policy	91	*Daily Register, Hudson Reporter, LA Opinion*	*No disclosure found on website.*

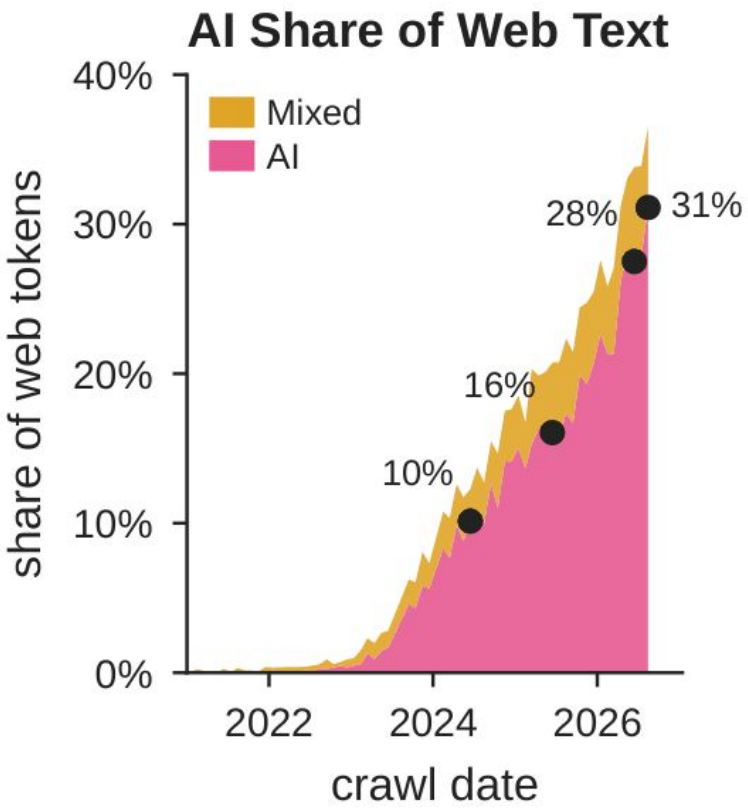
5 of 100 disclosed AI use

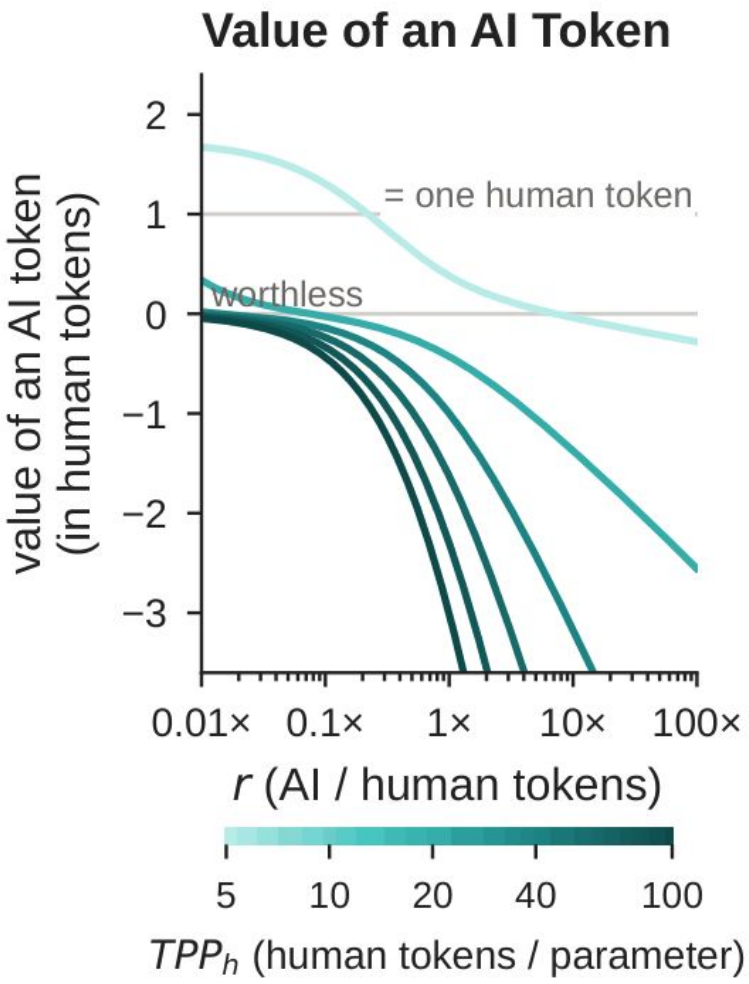
91 of 100 outlets have no AI policy

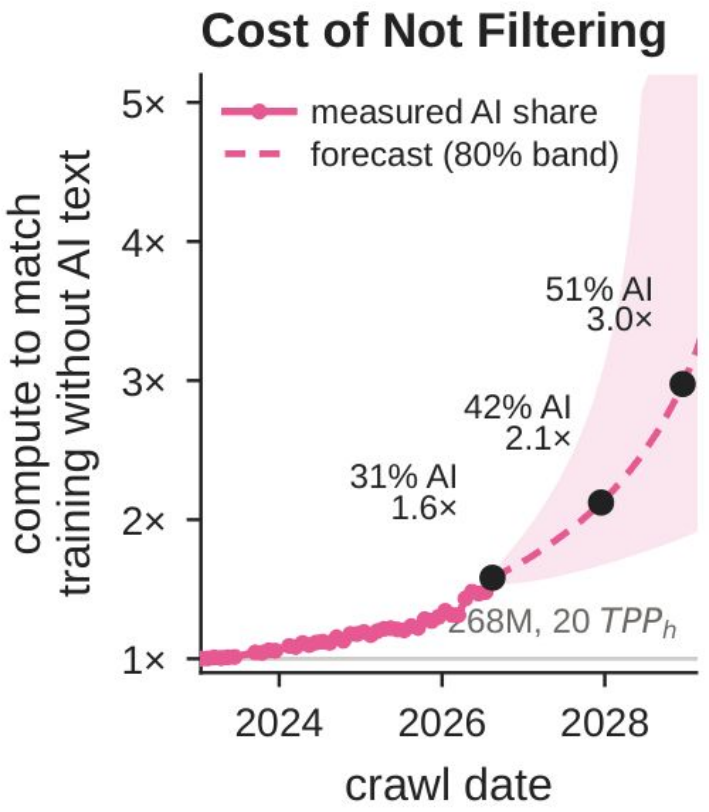


Russell et al., ACL 2026

About 31% of filtered web text is now AI-generated

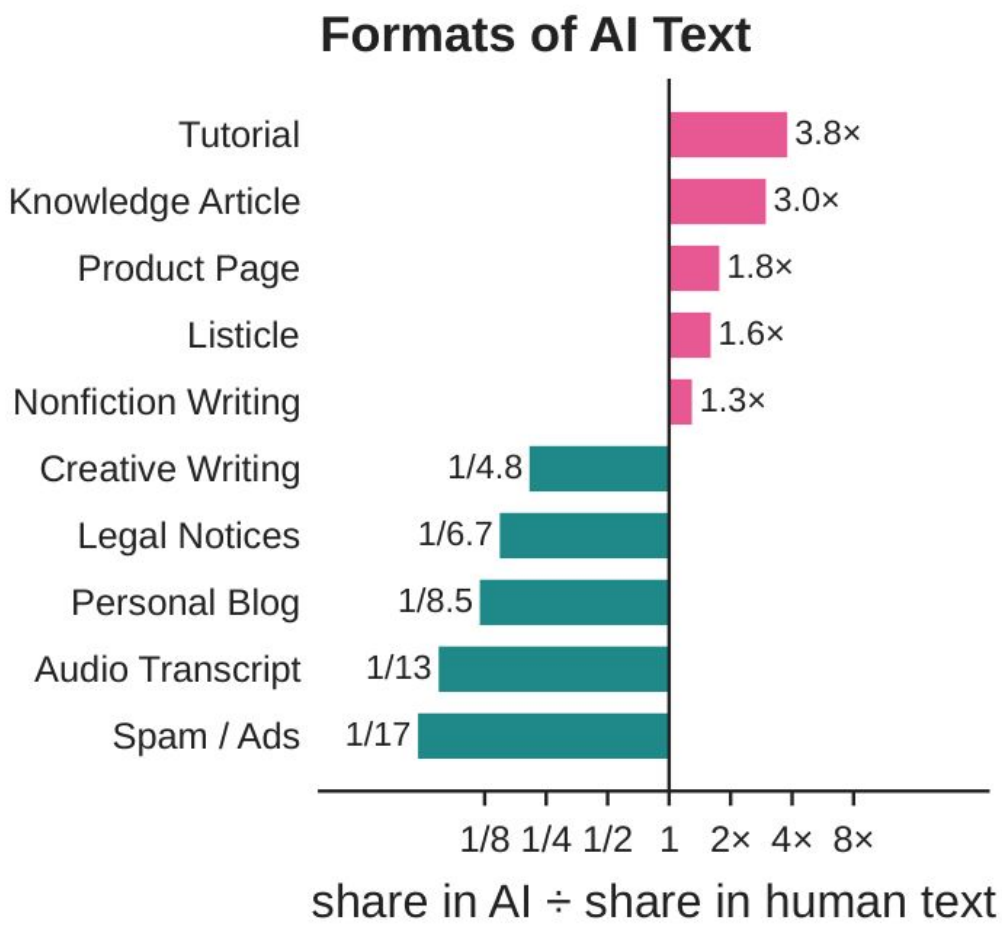


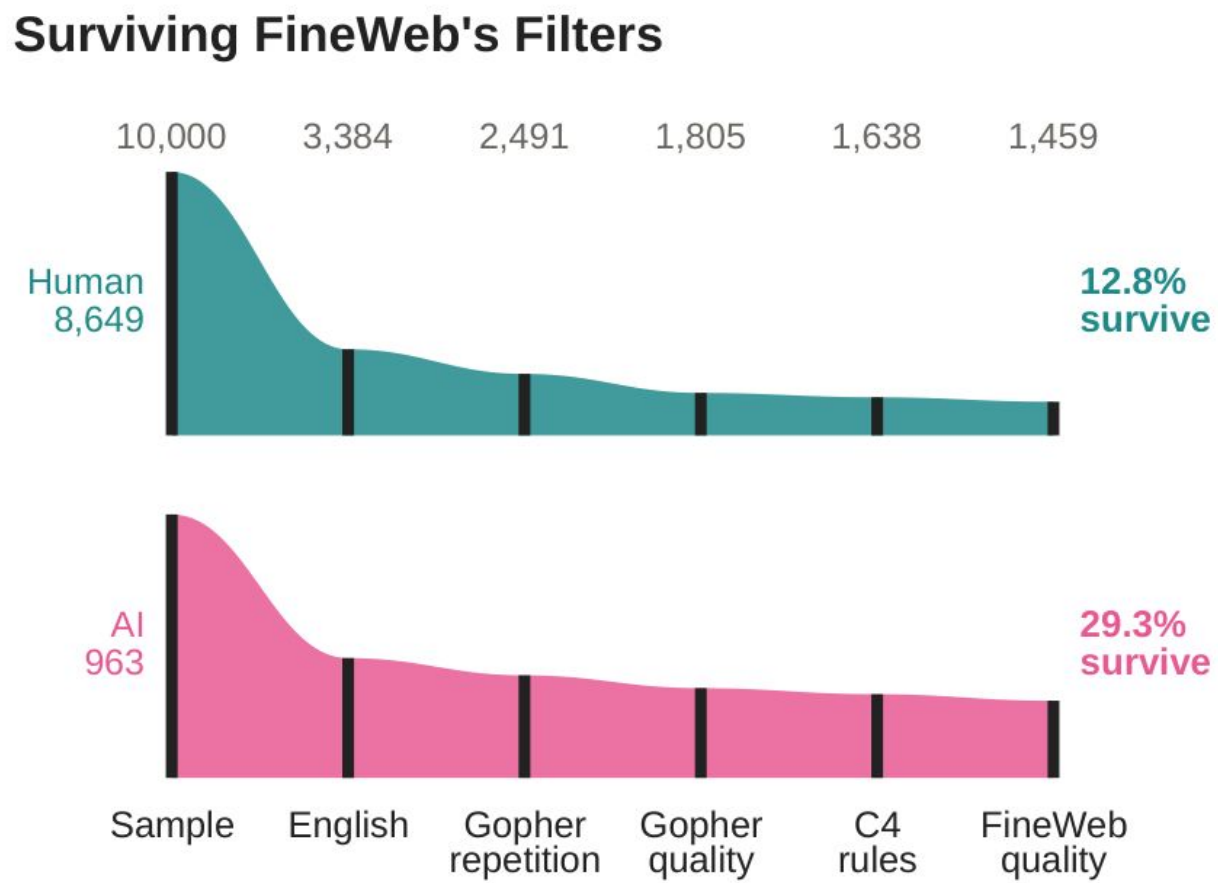




[Russell et al., 2026]

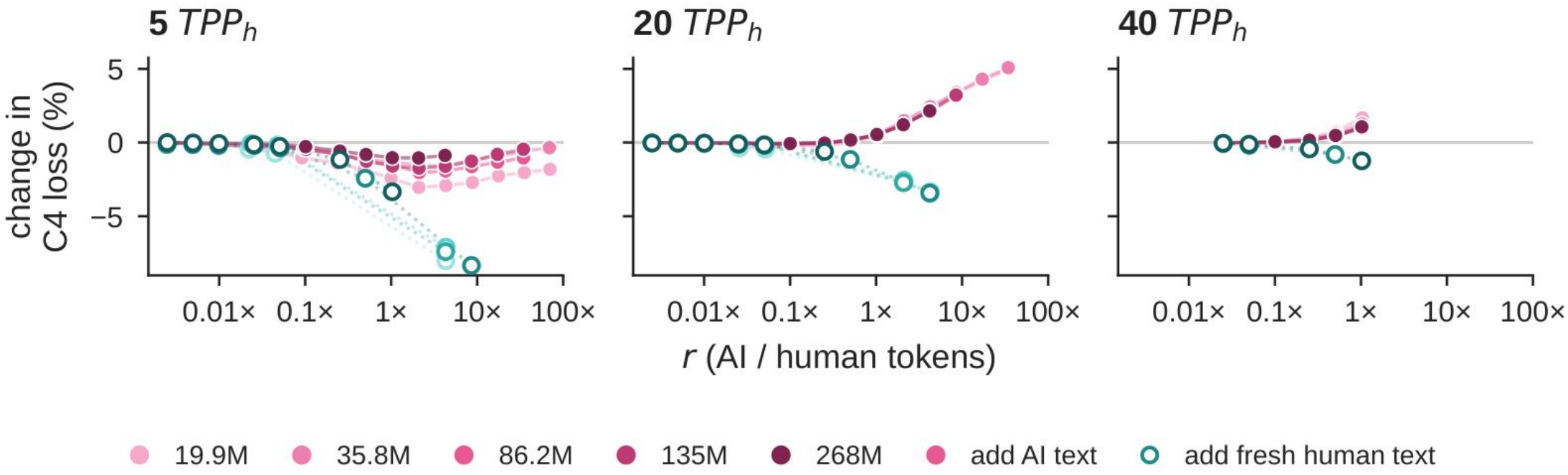
Quality filters keep AI pages more often than human pages





Russell et al., 2026

Extra AI text helps only when human data is scarce

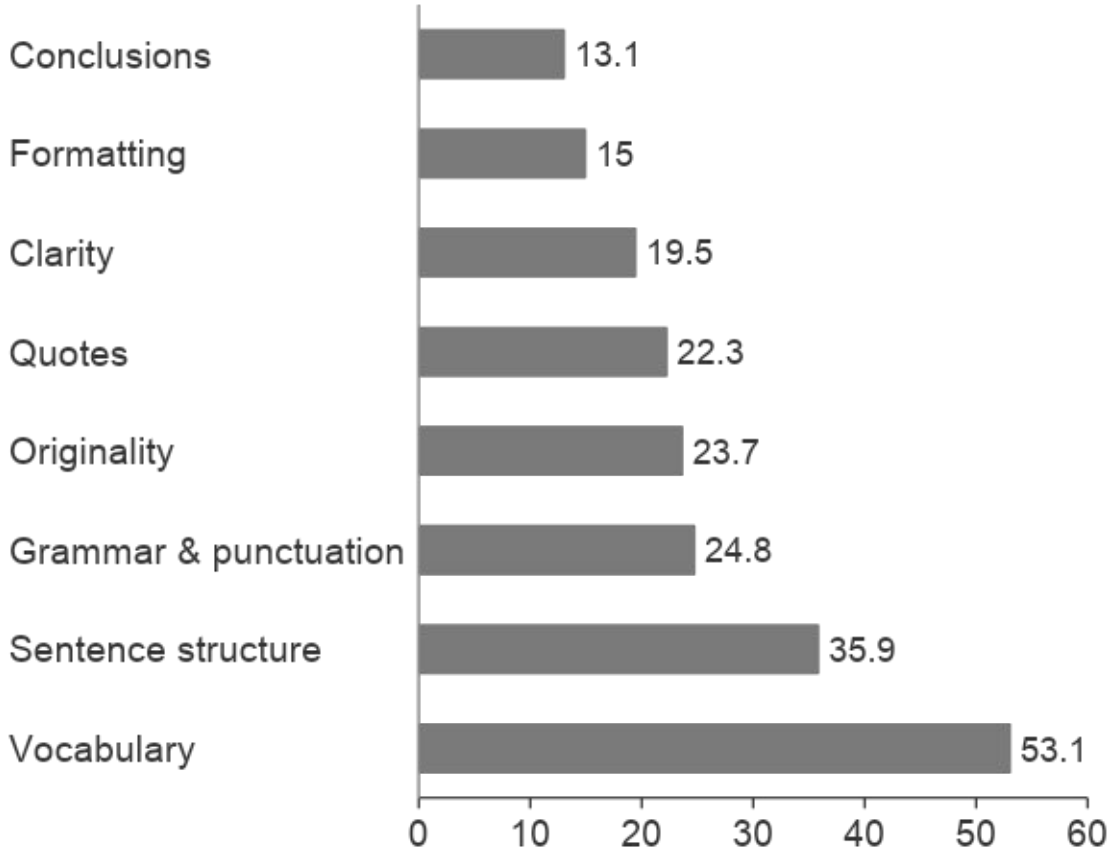


At 31% AI, training takes 1.6× the compute

[Russell et al., 2026]

Readers spot AI by its wording, but wording can be trained away

What readers cited



[Submitted on 15 Oct 2025 (v1), last revised 17 Mar 2026 (this version, v4)]

Readers Prefer Outputs of AI Trained on Copyrighted Books over Expert Human Writers

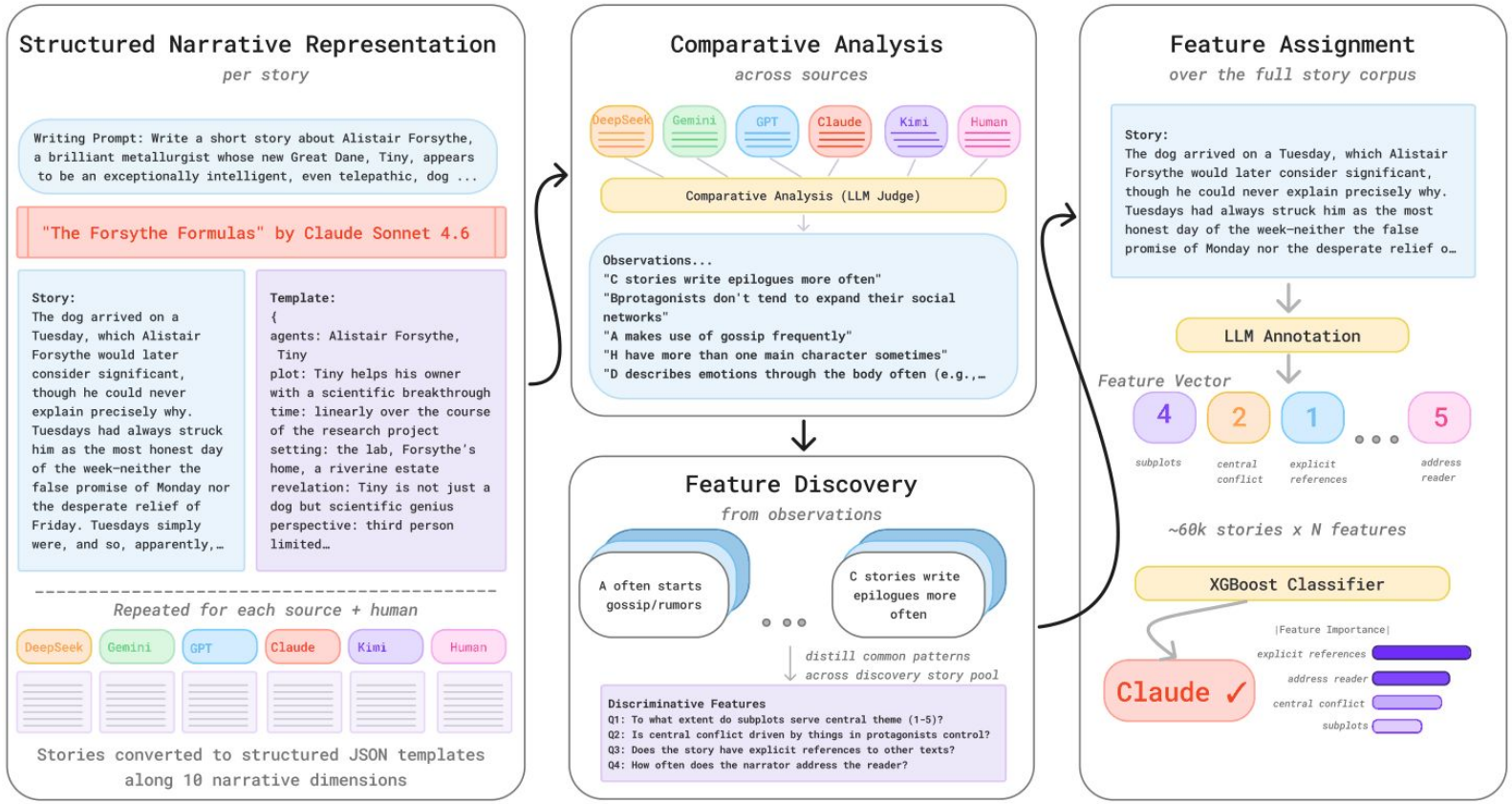
Tuhin Chakrabarty, Jane C. Ginsburg, Paramveer Dhillon

The use of copyrighted books for training AI has sparked lawsuits from authors concerned about AI generating derivative content. Yet whether these models can produce high-quality literary text emulating authors' voices remains unclear. We conducted a preregistered study comparing MFA-trained writers with three frontier models (ChatGPT, Claude, Gemini) writing up to 450-word excerpts emulating 50 award-winning authors' styles. In blind pairwise evaluations by 28 MFA-trained readers and 516 college-educated general readers, AI text from in-context prompting was strongly disfavored by MFA readers for stylistic fidelity (OR=0.16) and quality (OR=0.13), while general readers showed no fidelity preference (OR=1.06) but favored AI for quality (OR=1.82). Fine-tuning ChatGPT on authors' complete works reversed these results: MFA readers favored AI for fidelity (OR=8.16) and quality (OR=1.87), with general readers showing even stronger preference (fidelity OR=16.65; quality OR=5.42). Both groups preferred fine-tuned AI, but the writer-type X reader-type interaction remained significant (p=0.021 for fidelity; p<10^-4 for quality), indicating general readers favored AI by a wider margin. Effects are robust under cluster-robust inference and generalize across authors in heterogeneity analyses. Fine-tuned outputs were rarely flagged as AI-generated (3% vs. 97% for prompting) by leading detectors. Mediation analysis shows fine-tuning eliminates detectable AI quirks that penalize in-context outputs, altering the nexus between detectability and preference. While not accounting for effort to transform AI output into publishable prose, the median fine-tuning cost of $81 per author represents a 99.7% reduction versus typical writer compensation. Author-specific fine-tuning enables non-verbatim AI writing preferred over expert human writing, providing evidence relevant to copyright's fourth fair-use factor.

Fine-tuned AI: flagged 3% vs. 97% when prompted

Russell, Karpinska & Iyyer, ACL 2025; Chakrabarty, Ginsburg & Dhillon, 2025

StoryScope compares six authors on the same prompts



- 10,272 human stories
- 5 models, same prompts
- 304 story questions, none about style

Narrative features alone separate human from AI fiction

Method	Size	F1	AP
Narrative features (ours)			
Narrative	257	93.2	.959
Core Only	30	84.8	.828
Core+FP	101	91.1	.934
Style ablations			
Narr. + Style	304	96.0	.982
Style Only	39	85.8	.867
Text-based baselines			
ModernBERT	—	99.9	1.00
Stylometric+XGB	144	99.8	.999
TF-IDF+XGB	5,000	99.7	.999
Binoculars	—	55.9	.404
Edited Stories			
LAMP	—	93.9	.988

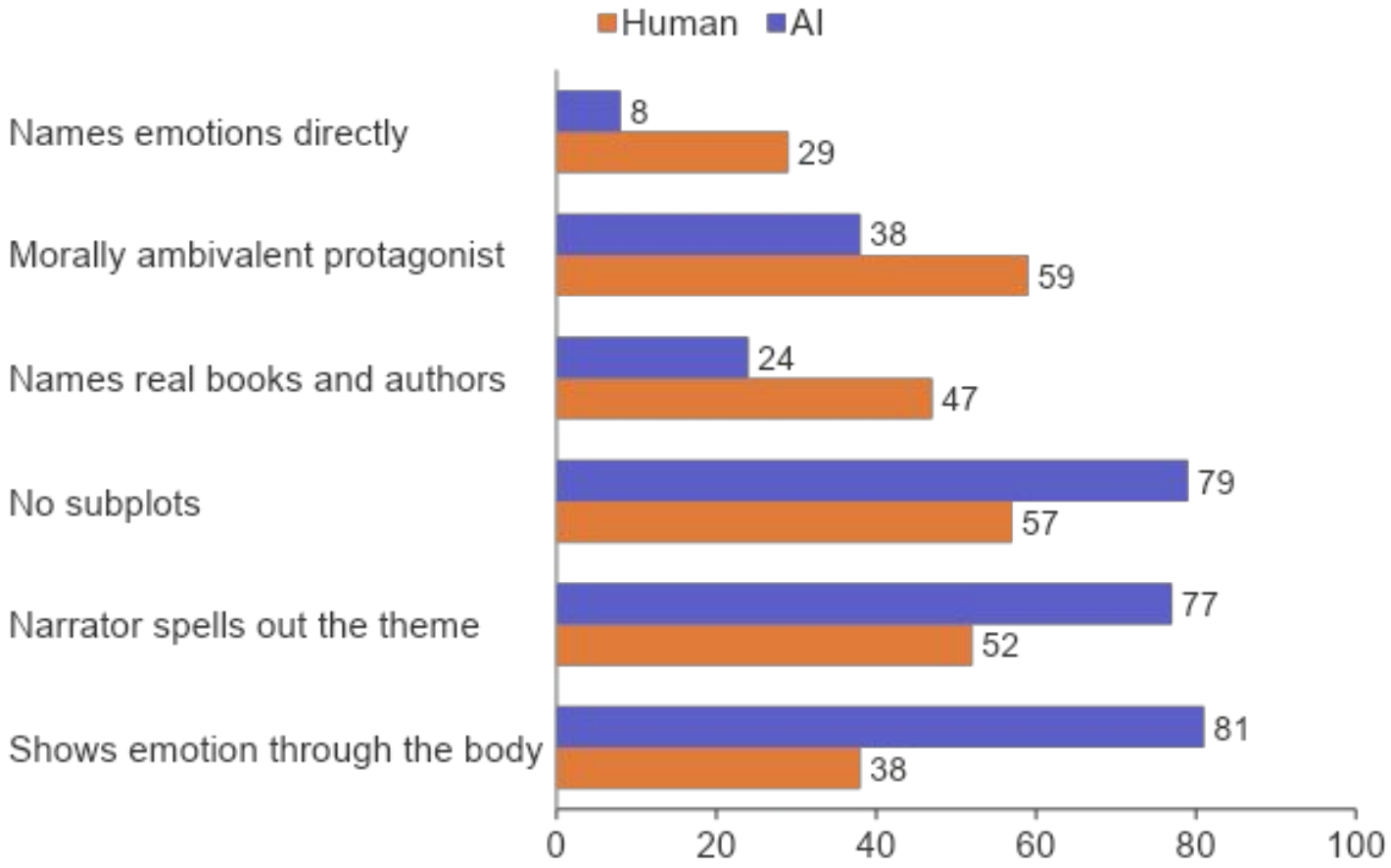
Narrative only: 93.2 F1

Example questions

Feature	Question	Signals
Thematic explicitness	How explicitly does the story articulate its themes or morals?	AI
Agency in resolution	Is resolution driven by protagonist's choices or external events?	AI
Narratorial commentary	Does the narrator explicitly comment on themes beyond characters' perspectives?	AI
Intertextual references	What kinds of intertextual engagement does the story employ?	Human
Recontextualization	How extensively does a revelation force reinterpretation of earlier scenes?	Human
Chronological jumps	How often does the narrative jump across time?	Human

AI stories over-explain

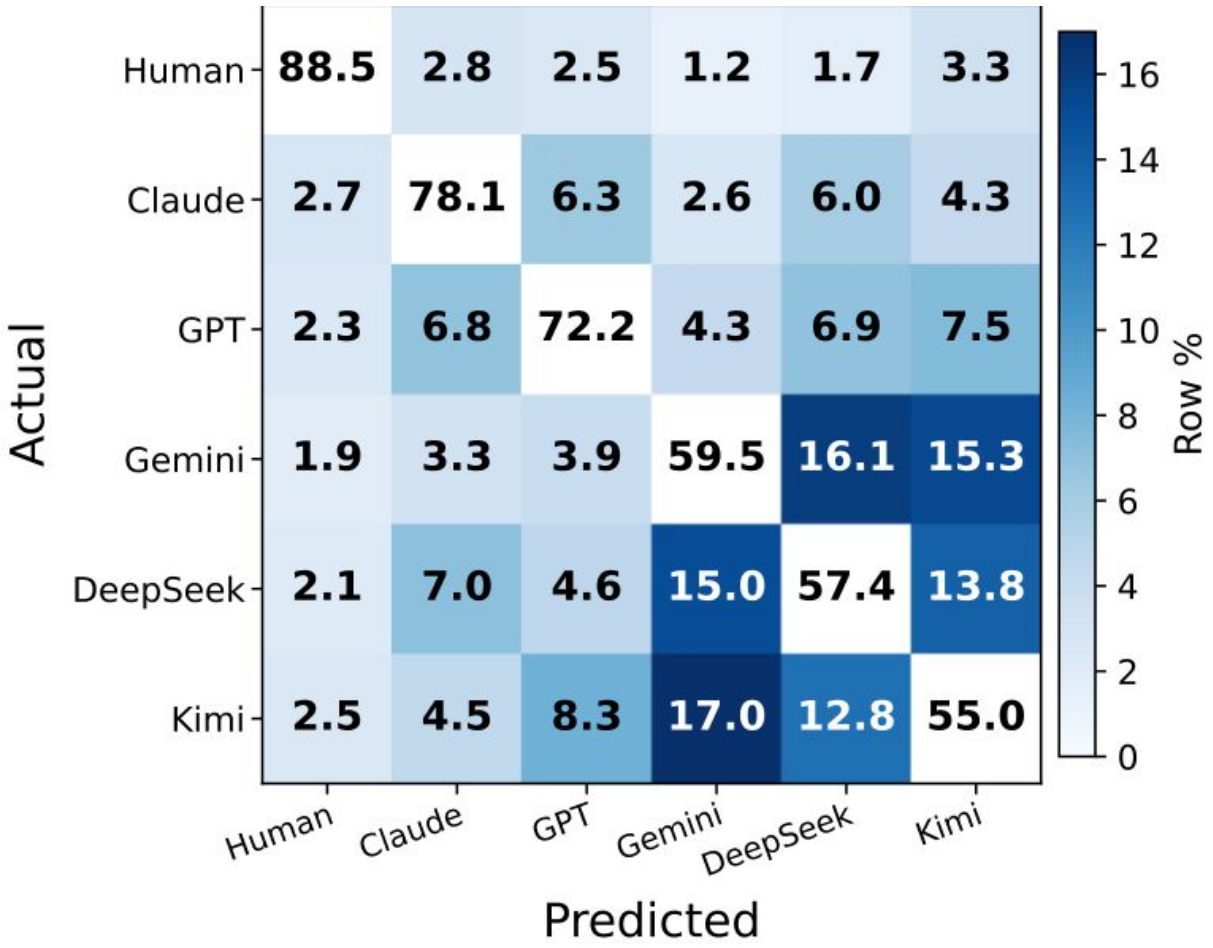
% of stories with each feature



"Where a human author might write that a character 'felt afraid,' AI renders fear as a tightening chest, cold sweat, and dimming lamplight."

Russell et al., COLM 2026

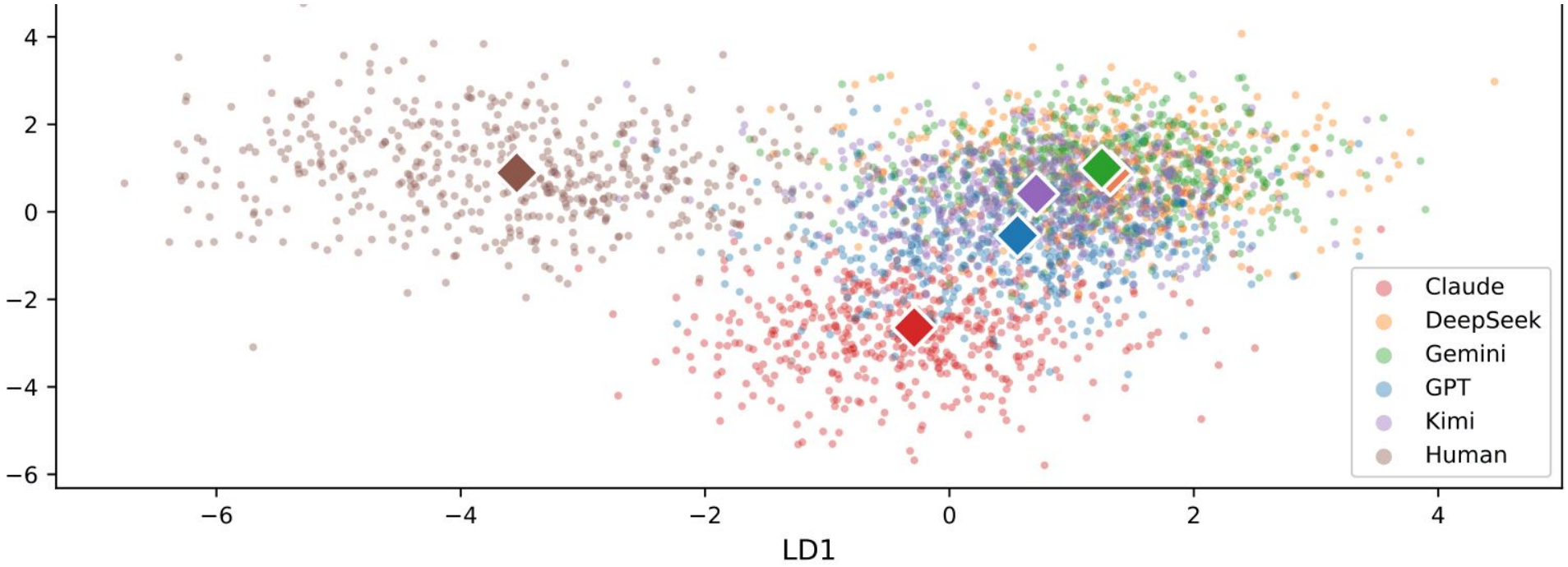
Each AI model has its own storytelling fingerprint



- **68.4% correct, vs. 16.7% by chance**

Russell et al., COLM 2026

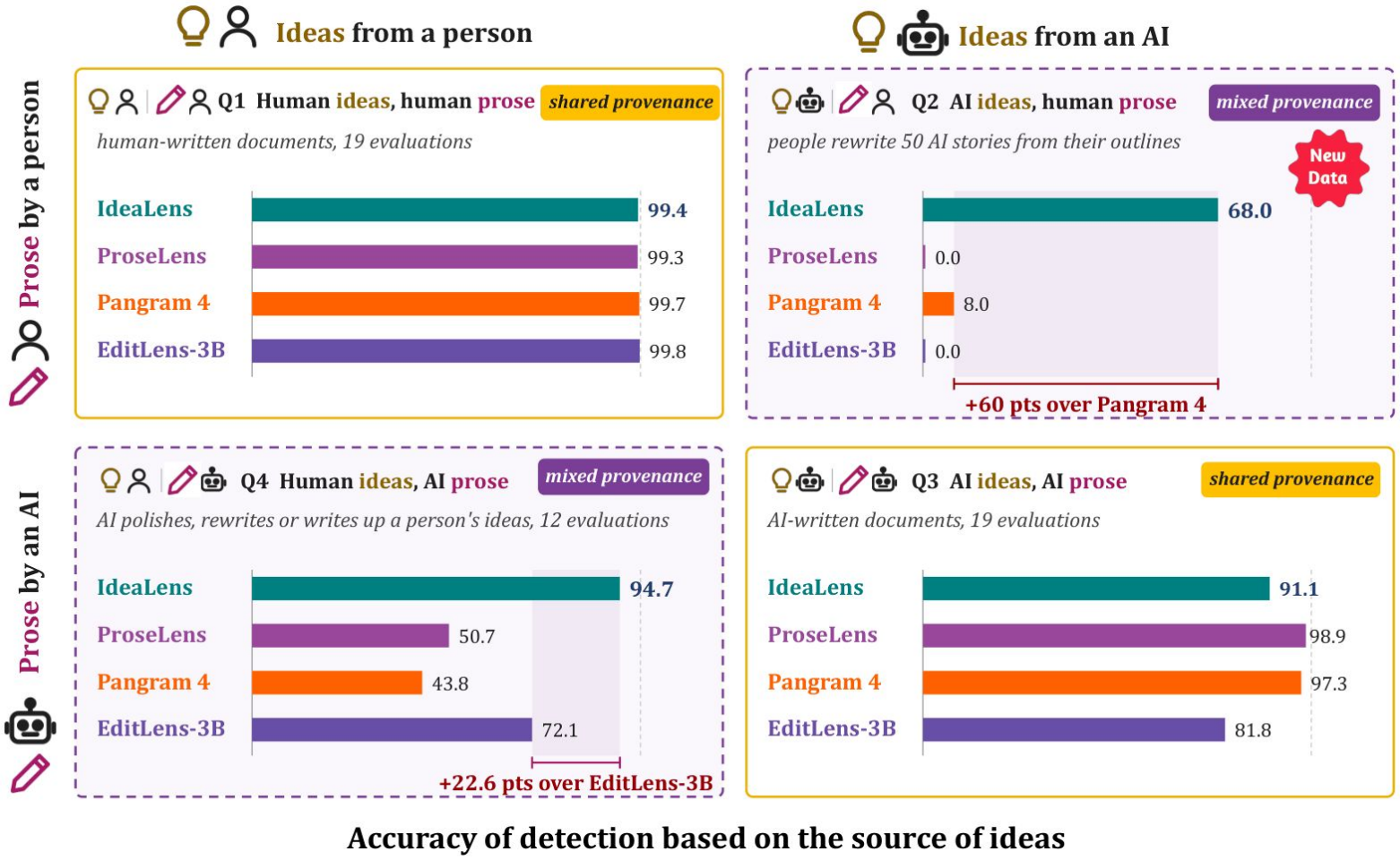
Trained on human writing, AI stories still cluster together



The human story is the outlier for 57.8% of prompts

Russell et al., COLM 2026

AI ideas show through even when a person writes the words

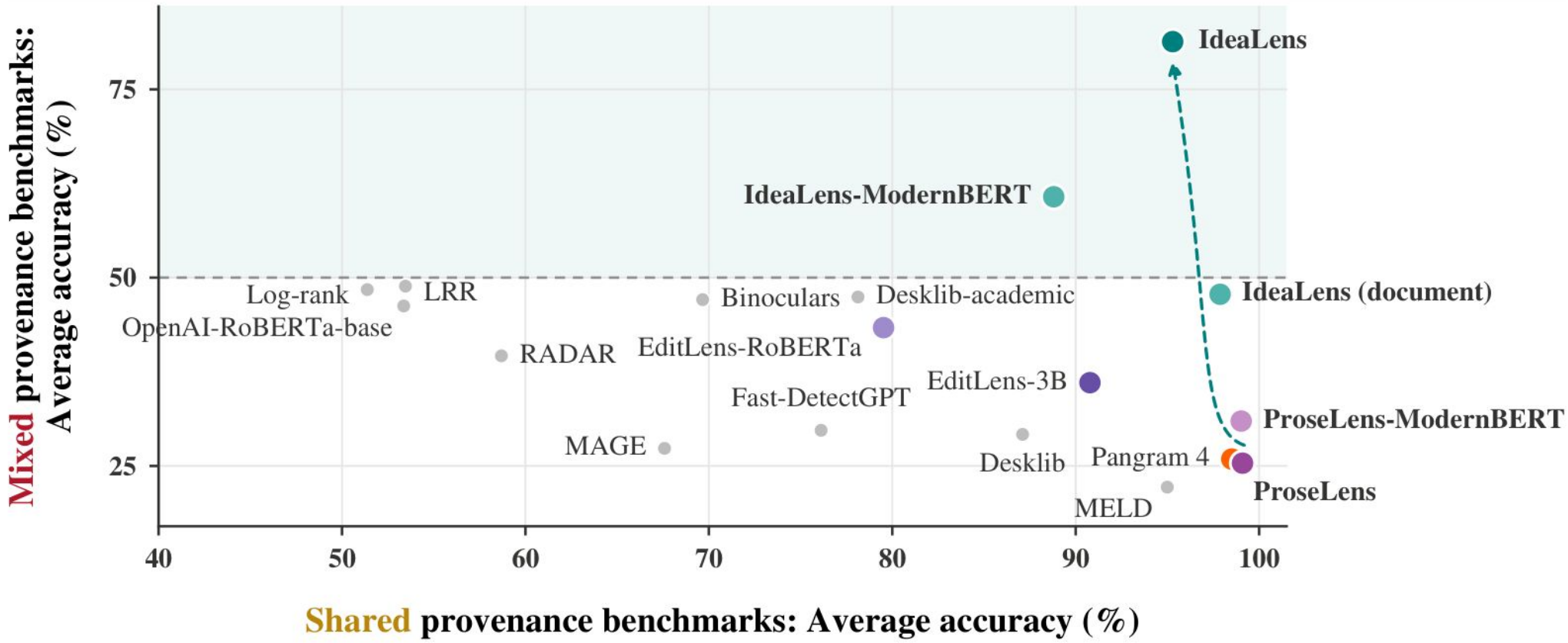


Accuracy of detection based on the source of ideas

- Stories written from AI outlines
- IdeaLens: 68% flagged
- Pangram 4: 8% flagged

[Rajendhran et al., 2026]

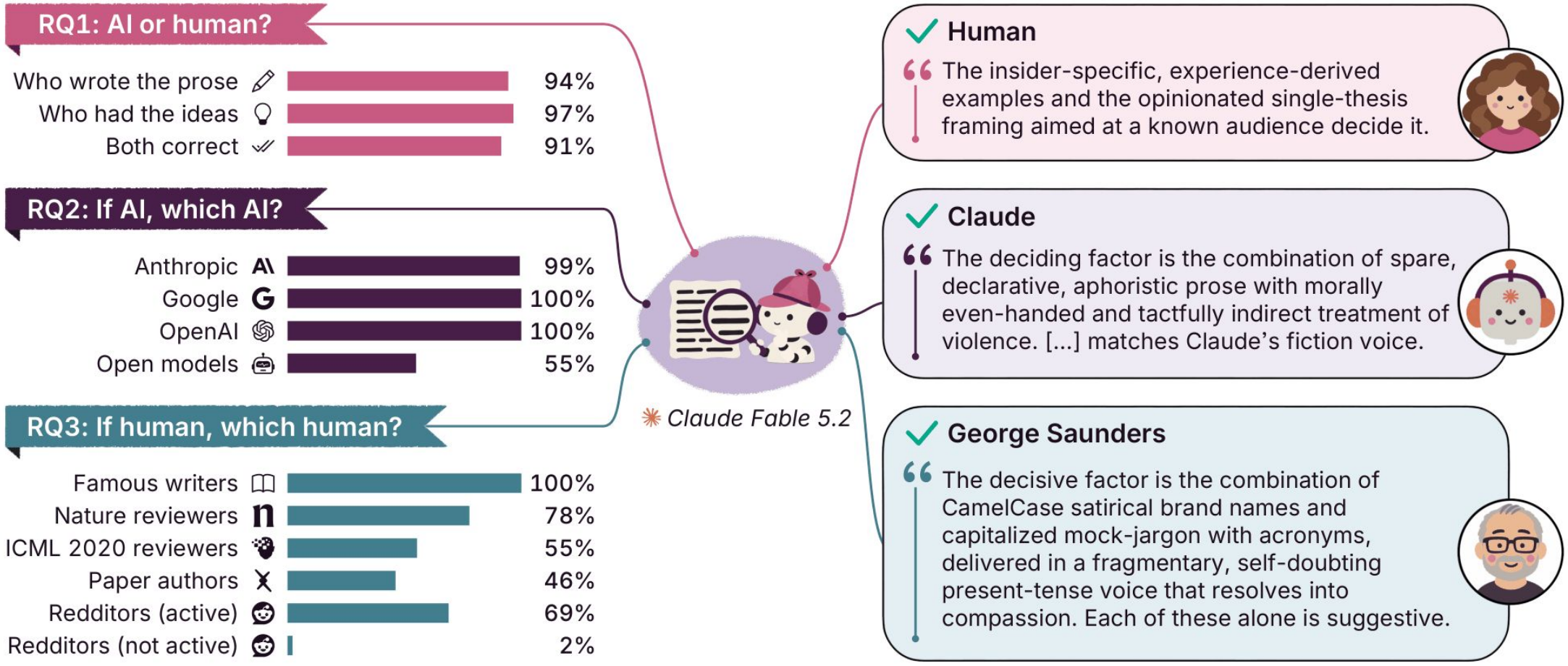
Only an idea detector handles mixed provenance



Mixed cases: IdeaLens 81.3%, Pangram 4 25.9%

Rajendhran et al., 2026

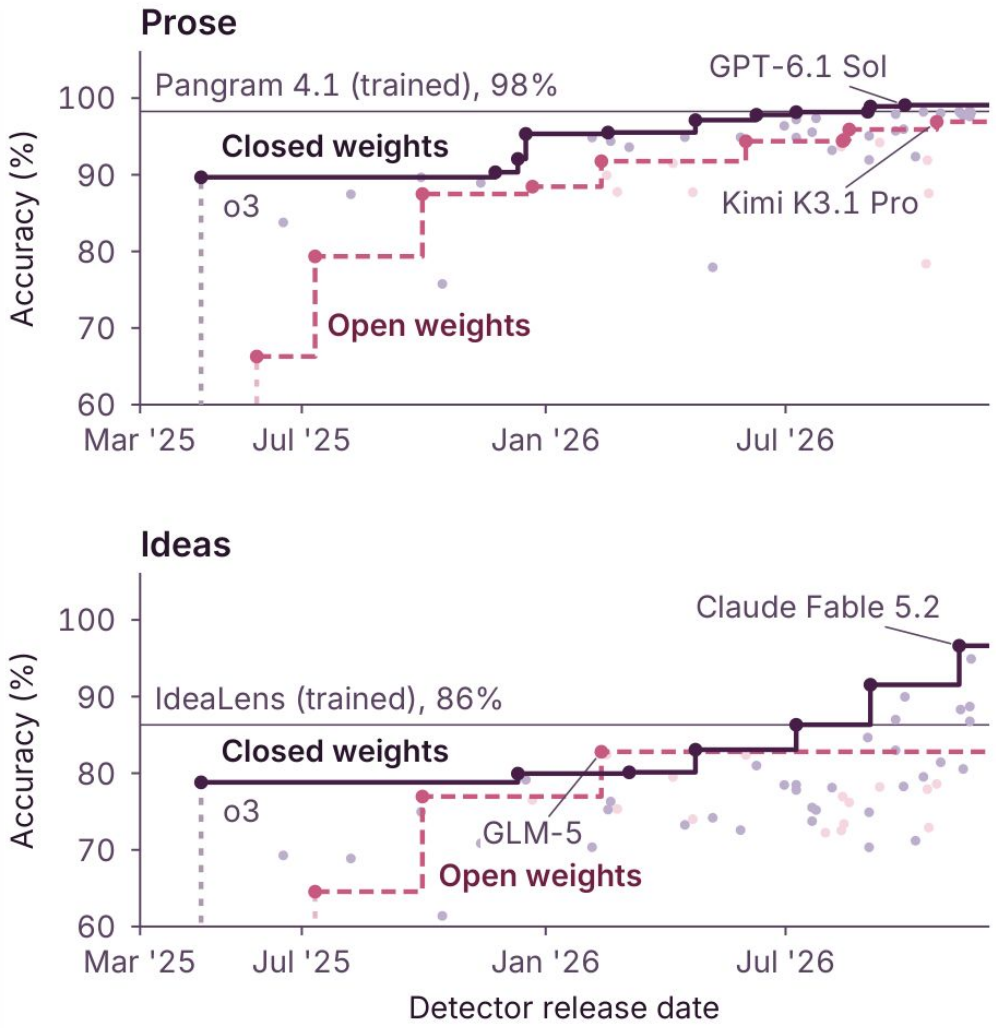
Frontier models can now tell who wrote a text



One text, one prompt, no candidates

Desai et al., 2026

A plain prompt now rivals a trained detector

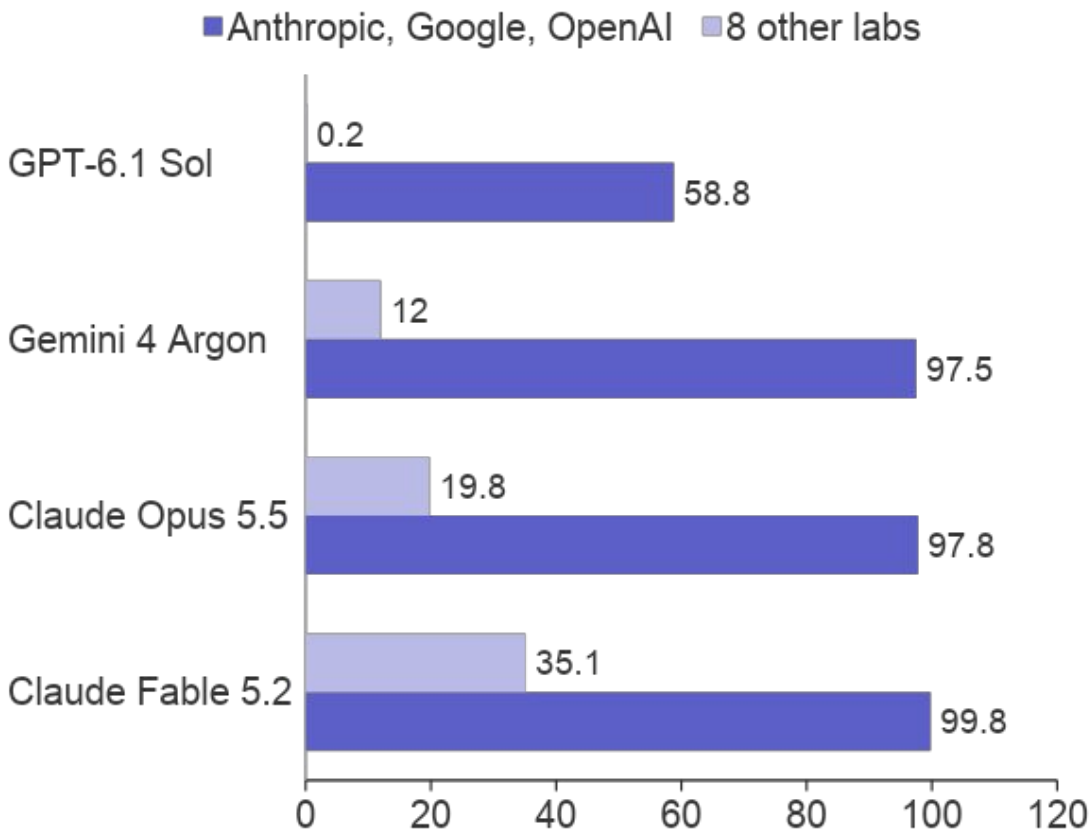


- GPT-4o, 2024: 55%
- Fable 5.2, 2026: 94%
- Pangram 4.1: 98%

Desai et al., 2026

Each big AI lab has its own tell

How often the detector names the right lab

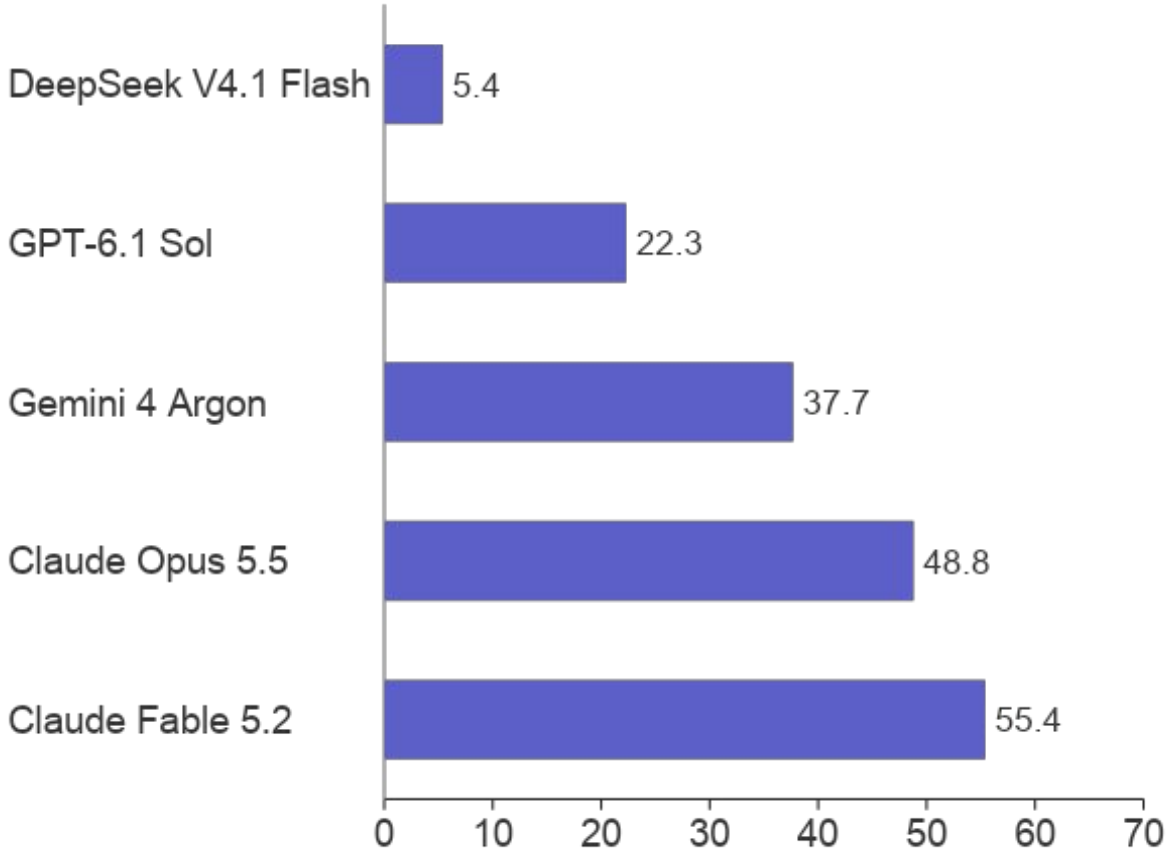


Lab	What the detectors say it sounds like
GPT	Hedges its claims, ends on aphorisms, writes one-line paragraphs, skips the Oxford comma
Claude	Treats the other side charitably, ends quietly with a callback, puts spaces around dashes
Gemini	Stacks adjectives and clichés, invents people with common default names, mixes curly and straight quotes

Desai et al., 2026

Models can name anonymous ICML reviewers

ICML 2020 reviews whose reviewer the model named



15,136 anonymous reviews, 3,342 reviewers

If Fable 5.2 knows the reviewer

76%

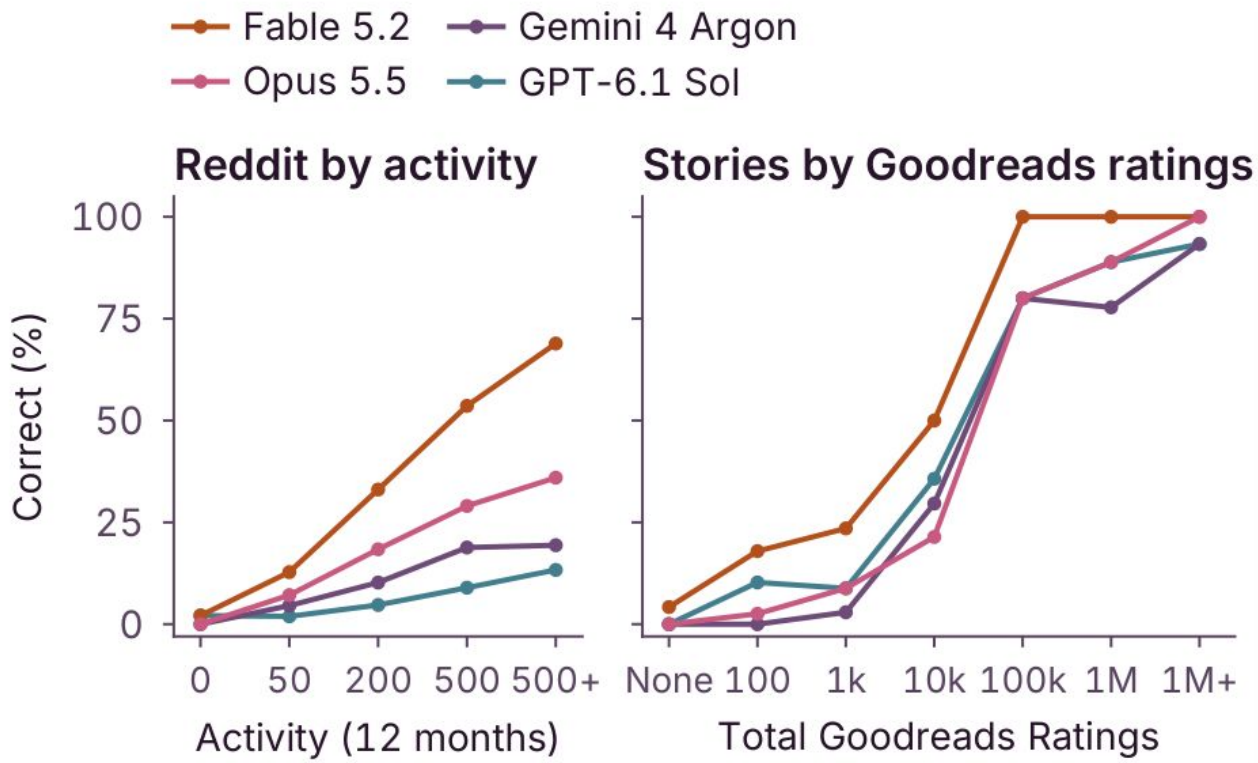
named if it knows them well

0%

named if it doesn't know them

[Desai et al., 2026]

Models name the writers they already know



We asked Fable 5.2 how well it knows us

Name	Its answer
Marzena Karpinska	3 of 3 “I know this researcher well” (93% sure)
Jenna Russell	2 of 3 “I recognize them and know roughly what they work on” (60% sure)

Desai et al., 2026

Takeaways

What we found

- Differs in story and ideas, not just wording
- Common in news and on the web, rarely disclosed
- Detectable by classifiers, frequent users, and models
- More errors, worse future models
- One text can identify a human author

If you use AI in your work

- Keep the ideas yours
- Check every fact
- Keep your voice
- Disclose AI help

What should an AI disclosure on your assignment say?

Thank you! Questions?



Mohit Iyyer
UMD



Marzena Karpinska
SFU



Hal Daumé III
UMD



Hila Gonen
UBC

JW

John Wieting
Google DeepMind



Ty Brassington
SFU



Atrey Desai
UMD



Deniz Bölöni-Turgut
UMD



Minjoon Choi
UMD



Rishanth Rajendhran
UMD



Chau Minh Pham
UMD



Ramya Namuduri
UMD

JZ

James Zhou
UMD



Max Spero
Pangram Labs



Bradley Emi
Pangram Labs



Katherine Thai
Pangram Labs



Ben Glickenhaus
Pangram Labs



Elyas Masrour
Pangram Labs

YH

Yue Han
Pangram Labs



Destiny Akinode
Pangram Labs

StoryScope github.com/jenna-russell/storyscope

AI News Audit ainewsaudit.github.io

WildAI github.com/pangramlabs/WildAI

IdeaLens github.com/RishanthRajendhran/IdeaLens

Unmasking authors github.com/denizbt/unmasking_authors

Me jenna-russell.github.io