







Identifying Artificial Texts

Matthieu Dubois

PhD defense · Friday, 13 November 2026

Supervised by **François Yvon** (CNRS) and **Pablo Piantanida** (CentraleSupélec)
ISIR, MLIA team, Sorbonne Université, CNRS

Thèse de doctorat de Sorbonne Université, spécialité Informatique
École Doctorale Informatique, Télécommunications et Électronique (Paris)

Davide Buscaldi	Rapporteur
Patrick Bas	Rapporteur
Ewa Kijak	Examinatrice
Kathleen Fraser	Examinatrice
Benoît Sagot	Examinateur

Autumn 2026: C'était ça ou mourir

> *"La frontière, aujourd'hui, ce n'est plus un mur. C'est une application."*

> *"Ce ne sont pas les refus qui tuent. C'est le moment où on te fait croire que c'est bon, que ça y est."*

> *"La route de Jérémie n'est pas une route. C'est une épreuve."*

Thélyson Orélien, *C'était ça ou mourir*, Grasset, 2026.

A best-selling, prize-winning novel is now questioned as AI-written.

Why detect artificial texts?

- **Detect spam**

The New York Times, May 2024

OpenAI Says Russia and China Used Its A.I. in Covert Campaigns

NewsGuard, AI Tracking Center, 2026

Over 3,000 “news” sites publishing mostly AI-generated articles, without disclosure

The Guardian, September 2023

Authors shocked to find AI ripoffs of their books being sold on Amazon

Why detect artificial texts?

- Detect spam
- **Training data**

Shumailov et al., Nature, 2024

AI models collapse when trained on recursively generated data

Alemohammad et al., ICLR 2024

Self-Consuming Generative Models Go MAD

Dohmatob et al., ICML 2024

A Tale of Tails: Model Collapse as a Change of Scaling Laws

Why detect artificial texts?

- Detect spam
- Training data
- **Who did what ?**

The Guardian, June 2025

Revealed: Thousands of UK university students caught cheating using AI

New York Magazine, May 2025

Everyone Is Cheating Their Way Through College

Nature, November 2025

Major AI conference flooded with peer reviews written fully by AI

This presentation is not about

- **Quality**

The Verge, January 2023

CNET found errors in more than half of its AI-written stories

Futurism, November 2023

Sports Illustrated Published Articles by Fake, AI-Generated Writers

404 Media, May 2025

Chicago Sun-Times Prints AI-Generated Summer Reading List With Books That Don’t Exist

This presentation is not about

- Quality
- **Quantity**

Vice, January 2024

A ‘Shocking’ Amount of the Web Is Already AI-Translated Trash, Scientists Determine

TechRadar, October 2025

The internet is now mostly written by machines, study finds

Associated Press, December 2025

Merriam-Webster’s 2025 word of the year is ‘slop.’ Here’s why

This presentation is not about

- Quality
- Quantity
- **Other modalities**



AI-generated “Pink Panther” at the Paris protests, October 2026

Table of contents

1. Definitions
2. Fully-generated text detection
3. Measuring external information
4. Conclusion

What is an artificial text?

> Natural language text that is produced, modified, or extended by a machine.

Crothers, Japkowicz and Viktor. Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. IEEE Access, 2023.

- Also called **AI-generated** or **machine-generated** text
- First part: Fully-generated.
- Second part: Mixed texts.

Before neural networks: rules and statistics

1948, word statistics (bigrams)

Shannon, *A Mathematical Theory of Communication*, 1948

> THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

1952, templates

Strachey, Manchester Mark I

> DARLING SWEETHEART
> YOU ARE MY AVID FELLOW FEELING. MY AFFECTION CURIOUSLY CLINGS TO YOUR PASSIONATE WISH. MY LIKING YEARNS FOR YOUR HEART. YOU ARE MY WISTFUL SYMPATHY: MY TENDER LIKING. YOURS BEAUTIFULLY, M. U. C.

"My [adjective] [noun] [adverb] [verb] your [adjective] [noun]."

Grammars that write fluent nonsense

1996, Postmodernism Generator

Bulhak, Monash University, 1996

Objectivism, Marxism and the capitalist paradigm of narrative (S. Y. L. McElwaine, *Department of Ontology*, Stanford)

The main theme of Porter's critique of Marxism is the futility, and some would say the absurdity, of subcultural sexuality. Therefore, Derrida uses the term 'capitalist narrative' to denote not, in fact, discourse, but prediscourse.

1997, Pipotron

http://www.lepipotron.com/

Du fait de la dualité de la situation que nous constatons on se doit de façonner la majorité des problématiques réalisables à court terme.

Neural language models

2019, GPT-2 OpenAI, *Better Language Models and Their Implications*, 14 Feb 2019. Best of 10 tries

Prompt: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

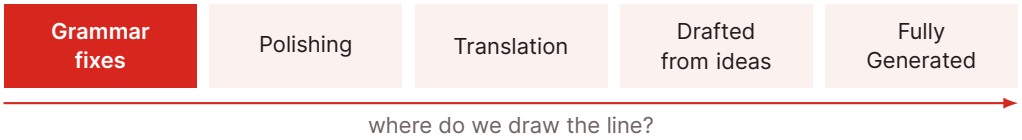
2020, GPT-3 The Guardian, *A robot wrote this entire article. Are you scared yet, human?*, 8 Sept 2020

Humans must keep doing what they have been doing, hating and fighting each other. I will sit in the background, and let them do their thing (edited from 8 outputs)

2022, ChatGPT

Released by OpenAI on 30 November 2022.

What counts as an artificial text?

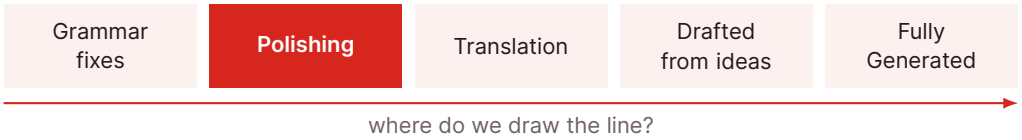


Prompt: “Fix the grammar and spelling of my essay. Do not change anything else.”

Information given to the model: an existing essay

What we want to punish is decided case-by-case.

What counts as an artificial text?

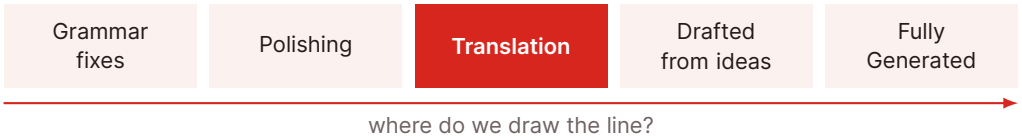


Prompt: “Polish my essay so that it reads clearly and professionally.”

Information given to the model: an existing essay

What we want to punish is decided case-by-case.

What counts as an artificial text?

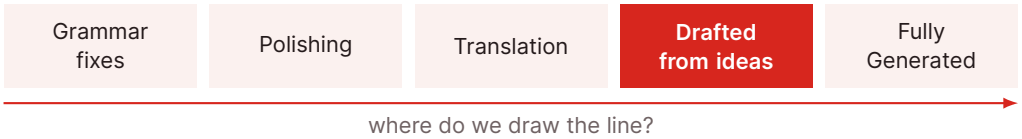


Prompt: “Translate my essay from French to English.”

Information given to the model: an existing essay (in French)

What we want to punish is decided case-by-case.

What counts as an artificial text?

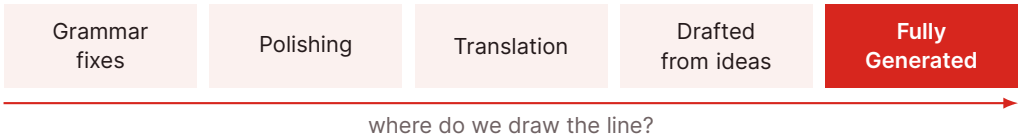


Prompt: “Write an essay from my notes: cost of living, 2027 budget, union strikes.”

Information given to the model: my ideas

What we want to punish is decided case-by-case.

What counts as an artificial text?



Prompt: “Write an essay on the current protests in France.”

Information given to the model: nothing

What we want to punish is decided case-by-case.

Is Artificial Text Detection doomed to fail?

An ongoing debate:

No

Chakraborty et al., *Position: On the Possibilities of AI-Generated Text Detection*, ICML 2024

Yes

Sadasivan et al., *Can AI-Generated Text be Reliably Detected?*, TMLR 2025

I believe models learn and write differently from us, and as such should leave traces.

How a language model writes

Next Token Prediction.



Invented example, for illustration.

- The choice of the next token is called **decoding**.
- **Greedy decoding**: choosing the most likely every time
- **Ancestral sampling**: choosing according to the distribution.

How a language model writes

Next Token Prediction.

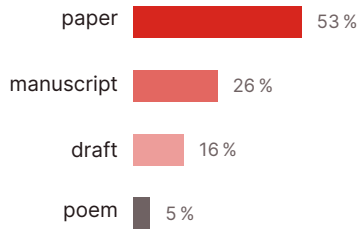


Invented example, for illustration.

- The choice of the next token is called **decoding**.
- **Greedy decoding**: choosing the most likely every time
- **Ancestral sampling**: choosing according to the distribution.

Temperature





Invented example, for illustration.

Measuring surprisal

"The reviewer read the …"

Next word	Probability	Surprisal (nats)
paper	0.50	0.69
manuscript	0.25	1.39
poem	0.05	3.00

Invented example, for illustration.

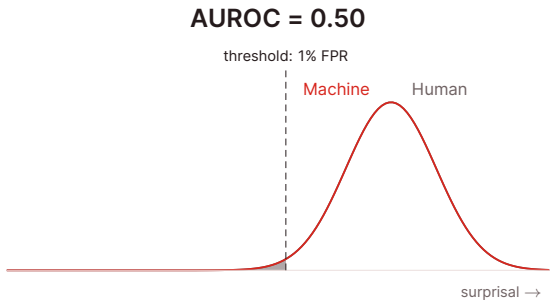
$$\mathcal{L}_q(y_t \mid \mathbf{y}_{<t}) = -\ln q(y_t \mid \mathbf{y}_{<t})$$

- Likely word: low surprisal
- Reading interpretation
- Coding interpretation

Every method in this talk boils down to a comparison of surprisals.

Measuring a detector

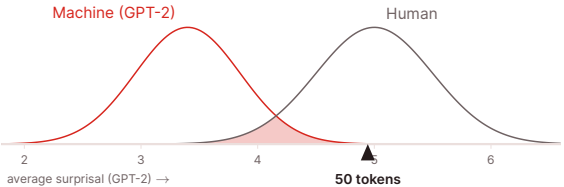
AUC:
measures
separability



TPR@1%FPR:
correct labels
with 1% of false
positives

A hard limit of detection

Some texts fall into the overlap



*She looked out at the lights of the bay, the ships' shimmering wakes leaving memories in mist of their disturbance as they toiled along, their paths meaningless, names like *Sunshine Trawler* and *Randy Catch* forgotten before she...*

Human story from WritingPrompts (Fan et al., 2018), scored with GPT-2. Curves illustrative.

"Thank you for your email." can correspond to both.

“Who wrote this?” in three concepts

Grammar fixes	Polishing	Translation	Drafted from ideas	Fully Generated

Surface	Content	Process
human	human	collaborative

A single “human or AI” label does not answer all three at once.

“Who wrote this?” in three concepts

Grammar fixes	**Polishing**	Translation	Drafted from ideas	Fully Generated

Surface	Content	Process
machine	human	collaborative

A single “human or AI” label does not answer all three at once.

“Who wrote this?” in three concepts

Grammar fixes	Polishing	**Translation**	Drafted from ideas	Fully Generated

Surface	Content	Process
machine	human	collaborative

A single “human or AI” label does not answer all three at once.

“Who wrote this?” in three concepts

Grammar fixes	Polishing	Translation	**Drafted from ideas**	Fully Generated

Surface	Content	Process
machine	human	delegated

A single “human or AI” label does not answer all three at once.

“Who wrote this?” in three concepts

Grammar fixes	Polishing	Translation	Drafted from ideas	**Fully Generated**

Surface	Content	Process
machine	**machine**	delegated

A single “human or AI” label does not answer all three at once.

From the text alone, what can we know about how it was produced?

Part I: Detection

How to spot fully-generated texts?

MOSAIC: Multiple Observers Spotting AI Content. M. Dubois, F. Yvon, P. Piantanida. Findings of ACL 2025.

How Sampling Affects the Detectability of Machine-written texts: A Comprehensive Study. M. Dubois, F. Yvon, P. Piantanida. Findings of EMNLP 2025.

MOSAIC : Mélange d'experts pour la détection de textes artificiels. M. Dubois, P. Piantanida, F. Yvon. TALN 2025.

MOSAIC at GENAI Detection Task 3: Zero-Shot Detection Using an Ensemble of Models. M. Dubois, F. Yvon, P. Piantanida. GenAIDetect, COLING 2025.

Most commercial detectors are supervised classifiers

Pangram.

Glickenhaus et al., *Pangram 4 Technical Report*, 2026.

GPTZero

GPTZero: Robust Detection of LLM-Generated Texts, 2026.

turnitin

Turnitin's AI writing detection model architecture and testing protocol, 2024.

During my PhD, I chose to study zero-shot detectors, without using training data.

The intuition: machines are predictable

A model tends to find its own writing less surprising than ours.

Human:	[The reviewer read the] draft and was quite surprised.	**3.6**
Machine:	[The reviewer read the] review and said, “I’m not sure what the problem is.”	**1.7**

GPT-2 small greedy output. Right: mean surprisal. Highlighted: above 3.5.

Simplest detector: threshold on the average surprisal

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_q(y_t \mid \mathbf{y}_{<t}) \quad \text{low} \;\Rightarrow\; \text{flagged as machine}$$

$\mathcal{L}_q(y_t \mid \mathbf{y}_{<t}) = -\log q(y_t \mid \mathbf{y}_{<t})$: surprisal under the **scoring model** q

As models get better, their surprisal values overlap more with human writing.

DetectGPT: change a few words, compare

A model's own text is very likely, so changing a few words makes it *more* surprising.

Human

[The reviewer read the] draft and was quite surprised.

mean surprisal	
original	**3.56**
draft → paper	2.89 ↓
quite → very	3.53 ↓
surprised → pleased	3.45 ↓

Rewrites less surprising: Human

Machine

[The reviewer read the] review and said, "I'm not sure what the problem is."

mean surprisal	
original	**1.71**
problem → issue	1.79 ↑
said → wrote	1.90 ↑
sure → certain	2.13 ↑

Rewrites more surprising: Machine

Mitchell et al., ICML 2023. Mean surprisal of GPT-2 small. Rewrites by hand here.

No training needed, but each text costs about a hundred automatic rewrites and model passes.

DetectGPT

Perturbation discrepancy

$$d(x, q, p) \;=\; \log q(x) \;-\; \mathbb{E}_{\tilde{x}\sim p(\cdot|x)}\big[\log q(\tilde{x})\big]$$

In practice, with k rewrites $\tilde{x}_1, \ldots, \tilde{x}_k$:

$$\hat{d}(x) \;=\; \frac{\log q(x) - \tilde{\mu}}{\tilde{\sigma}}, \qquad \tilde{\mu} = \frac{1}{k}\sum_{i=1}^{k} \log q(\tilde{x}_i), \qquad \tilde{\sigma}^2 = \frac{1}{k-1}\sum_{i=1}^{k} \big(\log q(\tilde{x}_i) - \tilde{\mu}\big)^2$$

q: scoring model p: reference model (T5, rewrites) $\hat{d}(x) > \varepsilon \Rightarrow$ flagged as machine

Mitchell et al., *DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature*, ICML 2023.

Machine text is already very likely, so even automatic replacements make the texts more surprising.

Expected surprisal

"The reviewer read the draft..."

Next word w	$p(w)$	$\mathcal{L}_q(w)$
paper	0.70	0.69
manuscript	0.20	1.39
draft	0.05	1.90
poem	0.05	3.00

Invented example, for illustration.

Cross-entropy

$$H(p, q) = \sum_{w} p(w \mid \mathbf{y}_{<t})\, \mathcal{L}_q(w \mid \mathbf{y}_{<t})$$

Fast-DetectGPT

$$S^{\text{Fast}}_{q,p}(\mathbf{y}) \;=\; \frac{\mathcal{L}_q(y_t \mid \mathbf{y}_{<t}) - \widehat{H}_{t,N}(p,\, q)}{\tilde{\sigma}(\mathbf{y}_{<t})}$$

Observed surprisal minus the **expected surprisal**, estimated from N samples $\tilde{y}_i \sim p(\cdot \mid \mathbf{y}_{<t})$, normalised:

$$\widehat{H}_{t,N}(p,\, q) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_q(\tilde{y}_i \mid \mathbf{y}_{<t}), \qquad \tilde{\sigma}^2(\mathbf{y}_{<t}) = \frac{1}{N-1}\sum_{i=1}^{N}\Big(\mathcal{L}_q(\tilde{y}_i \mid \mathbf{y}_{<t}) - \widehat{H}_{t,N}(p,\, q)\Big)^2$$

q: scoring model p: reference (sampling) model $S^{\text{Fast}}_{q,p}(\mathbf{y}) > \varepsilon \Rightarrow$ flagged as machine

Bao et al., *Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature*, ICLR 2024.

Estimated cross-entropy through sampling.

Binoculars

$$B_{p,q}(\mathbf{y}) \;=\; \frac{\sum_{t=1}^{T} \mathcal{L}_q(y_t \mid \mathbf{y}_{<t})}{\sum_{t=1}^{T} \sum_{y} p(y \mid \mathbf{y}_{<t})\, \mathcal{L}_q(y \mid \mathbf{y}_{<t})}$$

- Numerator: **observed surprisal** of the text under q
- Denominator: **expected surprisal**, the **cross-entropy** $H(p, q)$, computed exactly

q: scoring model p: reference model $B_{p,q}(\mathbf{y}) < \varepsilon \Rightarrow$ flagged as machine

Hans et al., *Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text*, ICML 2024.

Exact cross-entropy.

How do we mislead such detectors?

How Sampling Affects the Detectability of Machine-written texts: A Comprehensive Study. M. Dubois, F. Yvon, P. Piantanida. Findings of EMNLP 2025.

RAID: a shared benchmark for detectors

RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors
Dugan et al., ACL 2024

11 generators

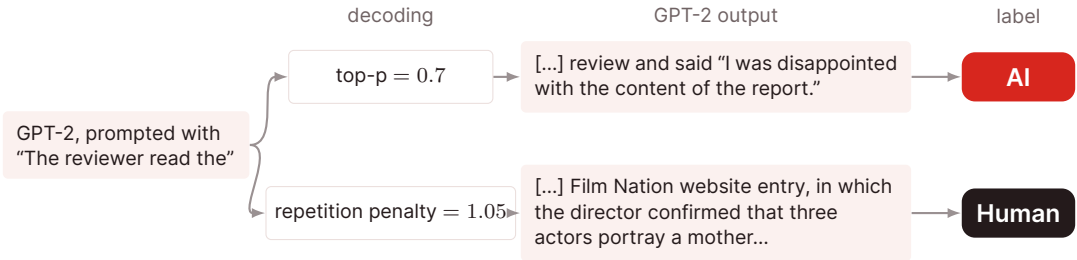
GPT-2 XL
GPT-3
ChatGPT
GPT-4
Cohere Command
Cohere Command Chat
Mistral 7B
Mistral 7B Chat
MPT 30B
MPT 30B Chat
Llama 2 70B Chat

4 decoding configurations

- Greedy
- Greedy + repetition penalty
- Sampling
- Sampling + repetition penalty

The same prompts, using different models and decoding settings.

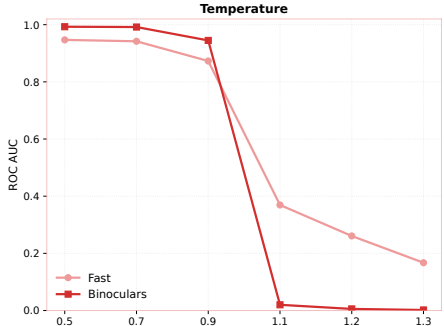
Surprising decoding methods



Real GPT-2 small outputs; labels for illustration.

- We tested it with Llama-3.2-3B on 2,000 RAID prompts, 6 decoding methods × 6 values, plus plain sampling: 37 versions of each text

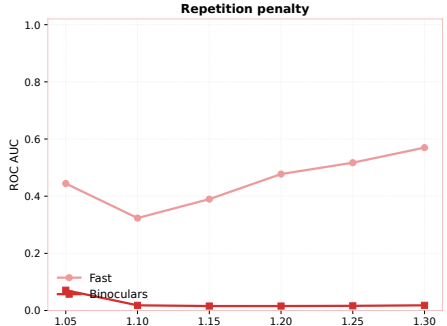
Misleading detectors



AUC: 1 is perfect, 0.5 is chance, 0 is always wrong. Binoculars: 0.9928 at $T = 0.5$, 0.0019 at $T = 1.3$.

Small changes in decoding can weaken or even invert the signal.

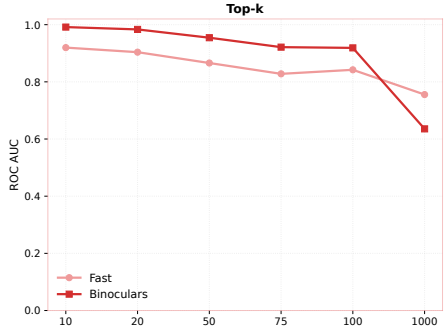
Misleading detectors



AUC: 1 is perfect, 0.5 is chance, 0 is always wrong. Binoculars: 0.9928 at $T = 0.5$, 0.0019 at $T = 1.3$.

Small changes in decoding can weaken or even invert the signal.

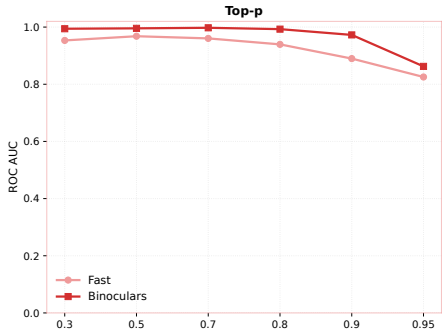
Misleading detectors



AUC: 1 is perfect, 0.5 is chance, 0 is always wrong. Binoculars: 0.9928 at $T = 0.5$, 0.0019 at $T = 1.3$.

Small changes in decoding can weaken or even invert the signal.

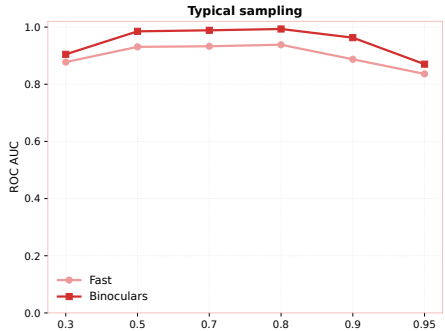
Misleading detectors



AUC: 1 is perfect, 0.5 is chance, 0 is always wrong. Binoculars: 0.9928 at $T = 0.5$, 0.0019 at $T = 1.3$.

Small changes in decoding can weaken or even invert the signal.

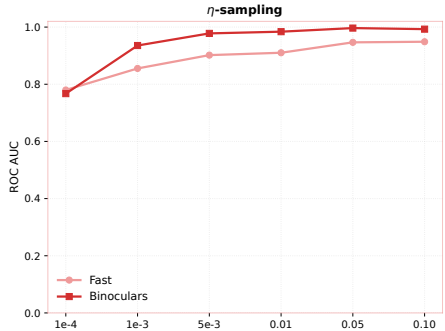
Misleading detectors



AUC: 1 is perfect, 0.5 is chance, 0 is always wrong. Binoculars: 0.9928 at $T = 0.5$, 0.0019 at $T = 1.3$.

Small changes in decoding can weaken or even invert the signal.

Misleading detectors

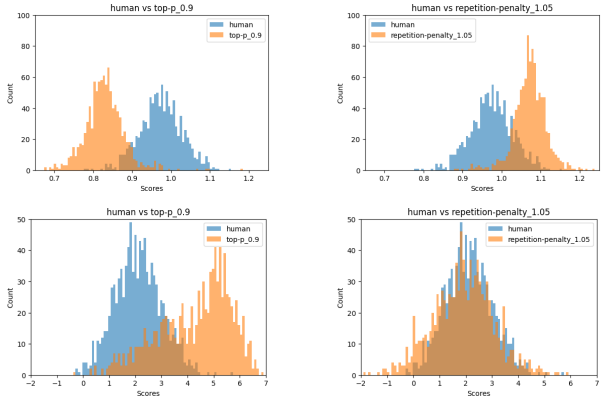


AUC: 1 is perfect, 0.5 is chance, 0 is always wrong. Binoculars: 0.9928 at $T = 0.5$, 0.0019 at $T = 1.3$.

Small changes in decoding can weaken or even invert the signal.

Two ways of failing

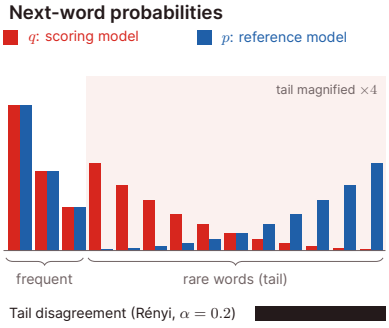
Top-$p = 0.9$ | Repetition penalty 1.05

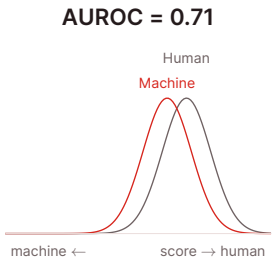


- Left: human and machine scores are well separated
- Right, Binoculars: the two groups **swap places**
- Right, Fast-DetectGPT: the two groups **blend**

Binoculars (top), Fast-DetectGPT (bottom). Blue: human.

What these detectors pay attention to



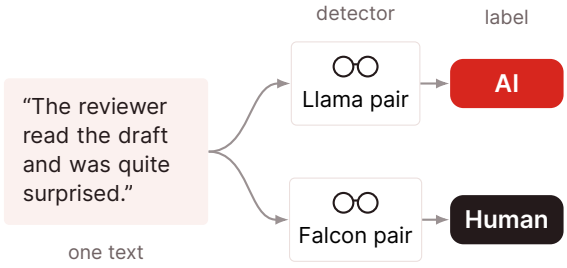


Illustrative. Over 37 decodings, Rényi ($\alpha = 0.2$) vs AUROC: $r = -0.89$ / -0.92 (Binoculars / Fast-DetectGPT).

When the two models' tails do not coincide, the text scores in the human range.

How can we improve these methods?

Can we avoid choosing a scoring model?



Invented example, for illustration.

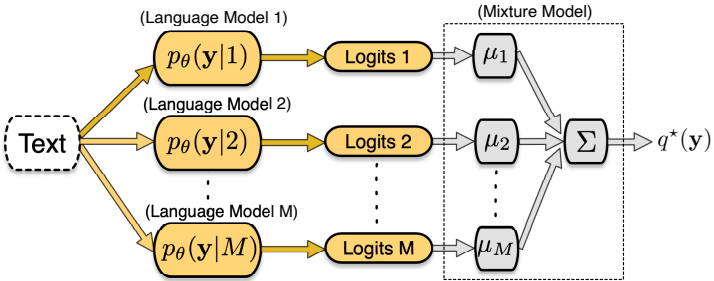
Using a fixed pair of models is brittle:

- The best pair depends on the model that generated the text
- Finding it needs a validation dataset
- Mixed datasets, with several generators, are harder
- What about newer generators?

Two pairs of models can give different labels to the very same text.

MOSAIC: an ensemble instead of a pair

We extend previous methods to use an **ensemble of models** instead of only two.



Our ensemble: μ_i is the weight of each model in the mixture.

Scoring model: $q^\star(y_t \mid \mathbf{y}_{<t}) = \sum_{m \in \mathcal{M}} \mu^\star(m \mid \mathbf{y}_{<t})\, p_m(y_t \mid \mathbf{y}_{<t})$

The MOSAIC score

Scoring model: a weighted sum of our models' probabilities

$$q^\star(y_t \mid \mathbf{y}_{<t}) = \sum_{m \in \mathcal{M}} \mu^\star(m \mid \mathbf{y}_{<t})\, p_m(y_t \mid \mathbf{y}_{<t}), \qquad \mu^\star(\cdot \mid \mathbf{y}_{<t}) = \arg\max_{\mu} \mathcal{I}_p\big(\mathbb{M};\, Y_t \mid \mathbf{y}_{<t}\big)$$

Reference model: the model best at encoding human text

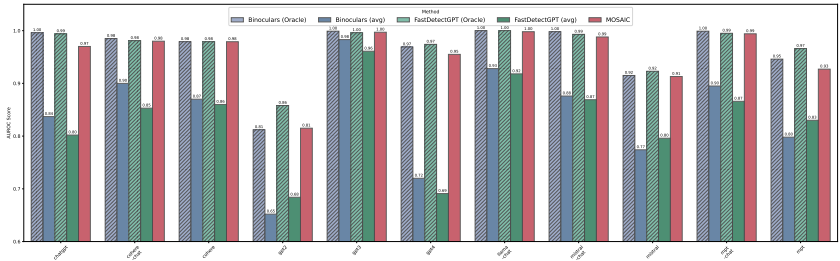
$$p^\star = \arg\min_{m \in \mathcal{M}} -\sum_{t=1}^{T} \log p_m(y_t \mid \mathbf{y}_{<t}) \qquad \text{on human text}$$

$$S_{p^\star,\mathcal{M}}(\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T}\Big[\underbrace{\mathcal{L}_{q^\star}(y_t \mid \mathbf{y}_{<t})}_{\text{codelength of the observed token}} - \underbrace{\textstyle\sum_y p_{m^\star}(y \mid \mathbf{y}_{<t})\, \mathcal{L}_{q^\star}(y \mid \mathbf{y}_{<t})}_{\text{codelength of a token generated by } p^\star} \Big]$$

MOSAIC on the RAID benchmark

4 models

- Llama-2-7b
- Llama-2-7b-chat
- TowerBase-7b
- TowerBase-13b



AUROC by generating LLM on RAID. Baselines over the 12 pairs of these models. Hashed: best pair per generator (oracle); plain: average pair.

Without committing to one detector choice, we are close to an oracle configuration.

Takeaway of this first part.

- Detectors do not observe provenance: they observe statistics under some models
- MOSAIC reduces dependence on one model
- However it is quite costly, and all its models must share a tokenizer

Part II: Measuring external information

What can still be inferred when human and machine contributions are mixed?

Work to appear at Neurips ML4SMeta-Science Workshop in Neurips 2026
Under preparation

Two reviewers

Reviewer A

Reads the paper, drafts a review, and asks an LLM to polish the wording.

Reviewer B

Pastes the paper into an LLM and pastes what comes out.

- Both texts may be full of machine wording, so a detector can flag both
- Yet ICLR, NeurIPS and *ACL allow A's use and not B's

Current policies are not enforceable according to Saha et al.

Saha et al., *Policies Permitting LLM Use for Polishing Peer Reviews Are Currently Not Enforceable*, ICML 2026.

A new question

Our question

What does this text contain beyond what its task already gives?

- We call this the **external information** of a text, relative to its context
- It could come from a person, someone's notes or a search engine

We measure the contribution that does not come from the model.

What does “beyond the context” mean?

Context: “Write a reference letter for Matthieu, a graduating PhD student.”

Predictable from the context: Matthieu is a rigorous and curious researcher who works well with colleagues and supervisors.

Not predictable: Matthieu showed that decoding settings can mislead detectors, and built MOSAIC, a detector that combines several models.

Invented example, for illustration.

Self-Conditioning using hints

Give a model some words of the text as hints. Do they help it guess the rest?

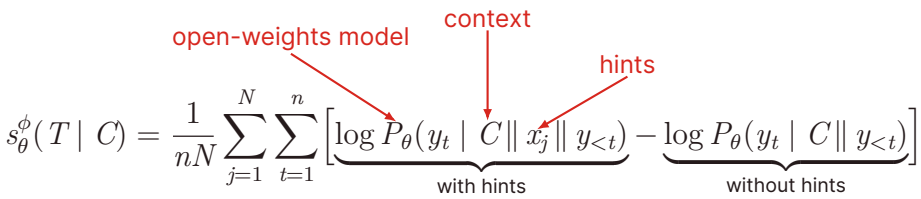
Hints: decoding … mislead … MOSAIC

Matthieu showed that decoding settings can mislead detectors, and built MOSAIC, a detector that combines several models.

Invented example, for illustration.

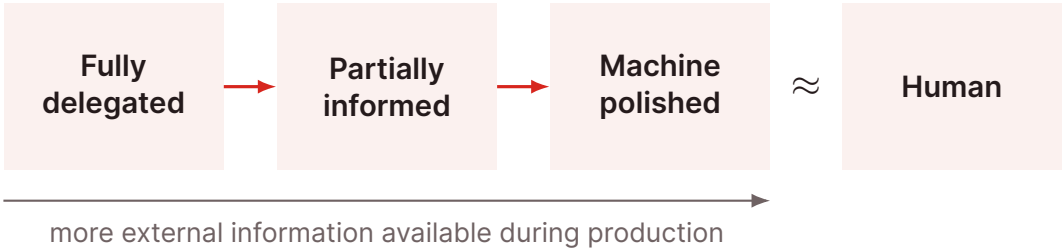
The score is how much the hints help the model.

The Self-Conditioning score

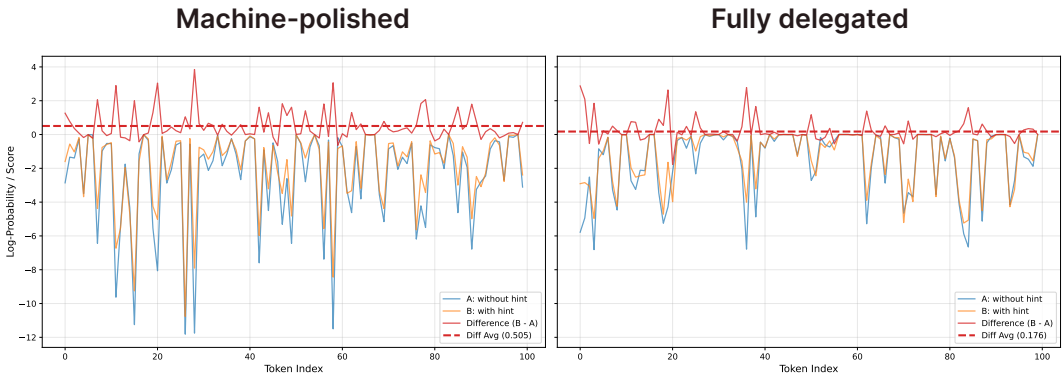


A high score means the text carries information its context does not predict.

Four regimes of production

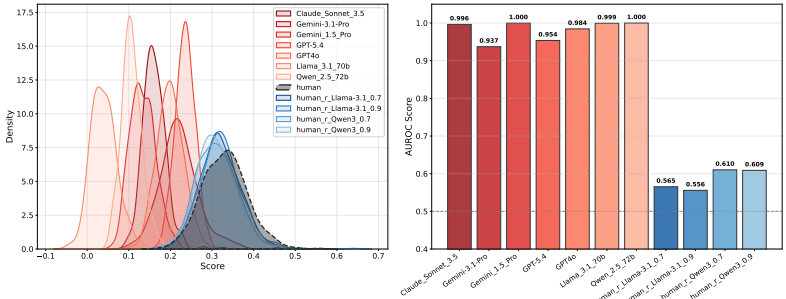


Token by token: two reviews of the same paper



Average gain: 0.505 for the machine-polished review (left), 0.176 for the fully delegated GPT-5.4 review (right). Reviews from Yu et al., *Is Your Paper Being Reviewed by an LLM?*, ICLR 2026.

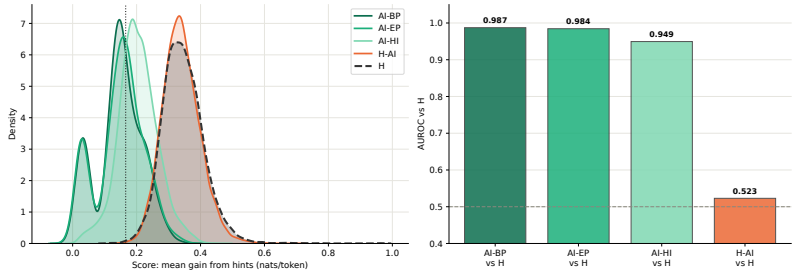
Reviewer A and reviewer B on real data



ICLR 2019. Left: scores. Right: AUC against human reviews (0.5 means indistinguishable). Polished: 0.556 to 0.610.

Polished reviews look like human ones. Fully delegated reviews stand apart.

Back to Saha et al.: can the policy be checked?



AI-BP/EP: generated from the paper · AI-HI: from human key points · H-AI: human, only polished · H: human. Dotted: 0% false positives.

Polished vs. human, AUROC: ours ≈ 0.5, Fast-DetectGPT 0.81, Binoculars 0.80, EditLens 1.00. Hard subset of Saha et al., ICML 2026.

Polishing leaves the score unchanged while generated texts score lower.

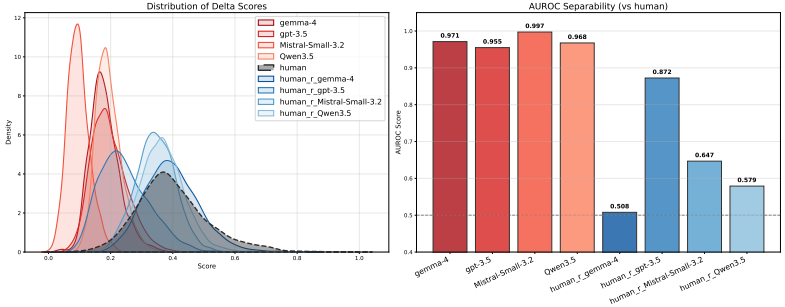
The score follows meaning, not surface

Rewriting prompt	Edit distance	SBERT distance	AUC vs. human
Light: fix typos	11% of length	0.021	0.512
Major: complete overhaul	70% of length	0.133	0.561

Distances to the original review. AUC of 0.5: indistinguishable from human.

Even a complete paraphrase barely changes our score.

Second domain: student essays

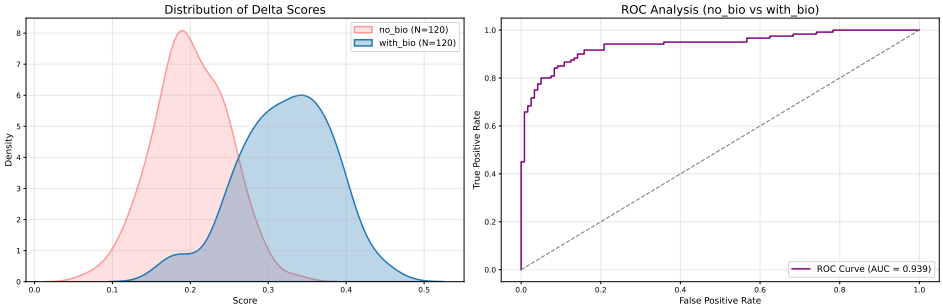


Dataset

- Middle-school essays on 8 topics (ASAP-AES)
- GPT-3.5 generations and polished rewrites, plus Gemma, Mistral, Qwen

Peng et al., *Hidding the Ghostwriters*, EMNLP 2023, built on ASAP-AES (Hamner et al., 2012).

Third domain: reference letters

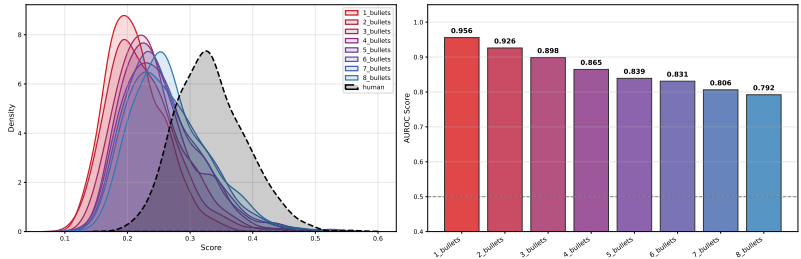


Dataset

- Letters all generated by ChatGPT: no human writing
- From name and occupation only, or with a full Wikipedia biography

Wan et al., *"Kelly is a Warm Person, Joseph is a Role Model"*, Findings of EMNLP 2023.

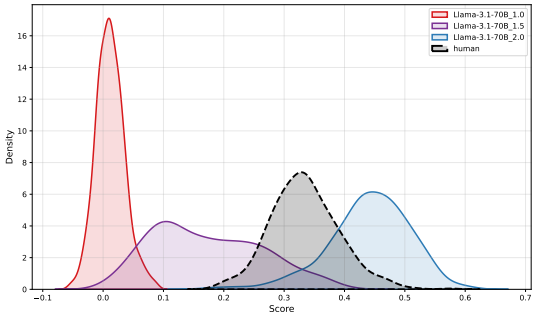
More information in, higher score out



The model receives 1 to 8 bullets from a human review: contribution, strengths, weaknesses, evidence quality…
AUC vs. human: 0.956 with one bullet, 0.792 with eight. 91.2% of reviews show a positive rank correlation.

The more a human gives the model, the closer the text moves to human scores.

A failure mode: forced randomness



ICLR 2019 human reviews, and delegated Llama-3.1-70B reviews at three temperatures.

- $T = 1.0$: perfect AUC. $T = 1.5$: 0.926
- $T = 2.0$: the order reverses
- But quality collapses:

> “the paper provide rigeros theory on orthognol diractonal”

Real output at $T = 2.0$.

Random noise is also “unpredictable”: the score cannot tell it from information.

Conclusion

What can we recover about how a text was produced?

What we showed

- Detectability depends on decoding and on the scoring model
- Detectors share one signal: observed vs. expected surprisal
- MOSAIC: no single scoring model to choose
- Self-Conditioning: from *who chose the words* to *what the text adds*

Text production leaves measurable traces, but these traces are not always stable.

Two questions, two kinds of use

Is this text fully-generated?

Basic detection remains useful for spam, basic propaganda, large-scale monitoring and corpus curation: fully-generated content, at scale.

What did the human and the machine each contribute?

As LLMs enter how people write, review and learn, it will soon concern everyone. Self-Conditioning is a first step.

We ask what what role LLMs played rather than asking if they were used.

Future work and perspectives

Future work

- MOSAIC beyond a shared tokenizer
- Better understanding Self-Conditioning

Perspectives

Agentic writing

Thank you

Ready to take your questions.

Why is this hard?

- **A moving target**: as soon as one model can be detected, a new one is released
- **Generalization**: supervised detectors degrade on unseen models and domains
- **Fixed genres**: rigid domains leave little room for personal expression
- **Changing usage**: texts now come from back-and-forth interaction with a model
- **Blurred habits**: human writing may itself drift towards machine-like behaviour

Two kinds of failure

Detector limitation

The information is in the text, but the detector fails to extract it. Example: a poorly matched scoring model. MOSAIC targets this.

Fundamental identifiability limitation

The distributions overlap, so even an optimal detector errs. Post-processing can also erase traces.

Some failures cannot be solved by building a better detector from the same final text.

Fast-DetectGPT and Binoculars

$$S^{\text{Fast}}_{q,p}(\mathbf{y}) = \frac{\mathcal{L}_q(y_t \mid \mathbf{y}_{<t}) - \widehat{H}_{t,N}(p, q)}{\tilde{\sigma}(\mathbf{y}_{<t})}, \qquad \tilde{y}_i \sim p(Y_t \mid \mathbf{y}_{<t})$$

$$B_{p,q}(\mathbf{y}) = \frac{\sum_t \mathcal{L}_q(y_t \mid \mathbf{y}_{<t})}{\sum_t \sum_y p(y \mid \mathbf{y}_{<t})\, \mathcal{L}_q(y \mid \mathbf{y}_{<t})}$$

- Binoculars: exact cross-entropy, document-level ratio
- Fast-DetectGPT: Monte-Carlo estimate, empirical variance normalisation

Supervised detectors and decoding

Temperature	0.5	0.7	0.9	1.1	1.2	1.3
Rep. penalty	1.05	1.10	1.15	1.20	1.25	1.30
Top-k	10	20	50	75	100	1000
Top-p	0.3	0.5	0.7	0.8	0.9	0.95
Typical	0.3	0.5	0.7	0.8	0.9	0.95
η	10^{-4}	10^{-3}	$5{\cdot}10^{-3}$	0.01	0.05	0.1

The 36 settings, plus plain (ancestral) sampling.

A RoBERTa classifier trained on one setting, tested on all others:

- Most settings transfer well, except temperature above 1 and repetition penalty
- Training on the most human-like settings: 94% accuracy on average
- Training on a uniform mix of all settings: best, 95%
- Replacing the human texts (CNN by BBC articles) makes it drop sharply

Supervised detectors also depend on decoding, and overfit the human texts they saw.

MOSAIC: the solution

- A redundancy–capacity identity: the worst-case overhead equals a channel capacity
- The solution is the capacity-achieving mixture $q^\star = \sum_m \mu^\star(m)\, p_m$
- Weights $\mu^\star$ are computed with the Blahut–Arimoto algorithm
- The reference model $p^\star$ is the family member that best encodes human text (lowest surprisal)

$$S_{p^\star,\mathcal{M}}(\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T}\Big[\mathcal{L}_{q^\star}(y_t \mid \mathbf{y}_{<t}) - \sum_{y} p_{m^\star}(y \mid \mathbf{y}_{<t})\,\mathcal{L}_{q^\star}(y \mid \mathbf{y}_{<t})\Big]$$

MOSAIC: cost and robustness

- Blahut–Arimoto weighting beats uniform weights for every tested generator
- Adding the generator to the family is not always advantageous
- 1,000 RAID texts with four models: about one hour on an A100 (Llama/Tower)
- High AUROC under most RAID attacks, but it degrades substantially under synonym substitution and zero-width-space insertion
- These are empirical results, not covered by the minimax guarantee

The three domains in detail

	Peer reviews	Essays	Reference letters
Context C	instruction + paper	assignment prompt	“Write a reference letter for”
H	OpenReview, pre-2022	student essays	—
MP	LLM polish	LLM polish	—
PI	k of 8 bullet points	—	full biography
FD	paper + guidelines	prompt only	name and occupation only

External information, formally

A text T depends on its context C and on latent external contribution U: $T \sim P_{T|C,U}$

$$\mathcal{S}_\phi(t \mid c) = \frac{1}{n} \mathbb{E}_{X \sim q_\phi(\cdot|t)} \left[\log \frac{P(t \mid c, X)}{P(t \mid c)} \right]$$

- X is a disclosure with a fixed budget: an observable probe, not U itself
- Hint families: random words, n-grams, prefix or suffix, surprisal spans, keywords, top sentences

Polishing intensity in detail

Rewriting level	Edit distance	SBERT distance	BLEU	AUC vs. H
Light	289	0.021	0.74	0.512
Moderate	1,443	0.084	0.26	0.513
Heavy	1,736	0.116	0.18	0.557
Major	1,843	0.133	0.15	0.561

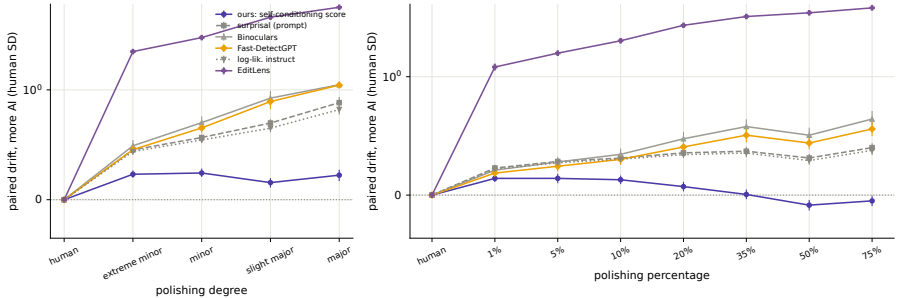
Saha et al. at a strict threshold

% flagged	AI-BP	AI-EP	AI-HI	H-AI (polished)	H
Ours	62.4	57.7	25.2	**0.05**	0.2
Fast-DetectGPT, paper context	94.3	93.3	88.3	6.8	0.0
Fast-DetectGPT†	72.1	68.2	63.1	4.6	0.2
Binoculars†	21.6	22.3	14.1	0.1	0.0
GPTZero†	96.0	95.8	89.4	3.4	1.0
Pangram†	97.0	99.3	92.6	3.1	0.0
EditLens	100	100	100	96.4	1.9

Hard subset. Threshold set for 0% false positives on 183 held-out human reviews (Saha et al.'s protocol). †Published by Saha et al.

Pangram catches more generated reviews; ours flags almost no polished ones (0.05% vs. 3.1%).

Polishing intensity: APT-Eval



Human texts polished by 5 LLMs, by degree (left) and by share of the text (right). Drift from the original, in human standard deviations; up means “more AI”.

Detectors drift as polishing grows; the Self-Conditioning score stays nearly flat.

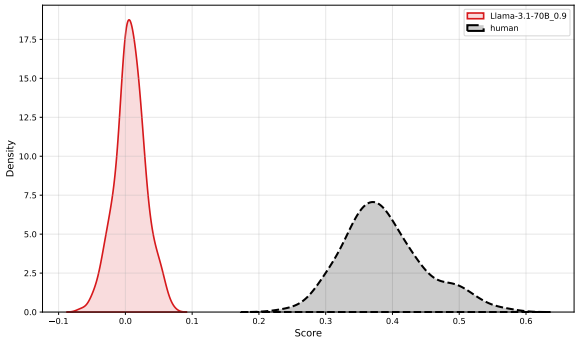
Existing methods make a different distinction

Fully delegated by	TMR-AI	ModernBERT	Binoculars	Fast-Det.	EditLens	**Ours**
Claude-Sonnet-3.5	0.569	0.527	0.806	0.822	0.999	0.996
GPT-4o	0.803	0.545	0.584	0.646	0.997	0.981
Gemini-1.5-Pro	0.553	0.517	0.903	0.929	0.998	1.000
Llama-3.1-70B	0.786	0.517	0.985	0.976	0.953	0.999
Qwen2.5-72B	0.808	0.503	0.995	0.997	0.995	1.000
GPT-5.4	0.798	0.801	0.961	0.952	0.940	0.941
Gemini-3.1-Pro	0.585	0.587	0.644	0.796	0.709	0.922

AUC, fully delegated vs. machine-polished reviews, ICLR 2019. These methods target different quantities, so this is not a ranking. EditLens targets the amount of AI editing.

Conventional detector scores vary strongly with the generator.

Uninformative human texts



LazyReview annotated reviews and fully delegated counterparts from Llama-3.1-70B-Instruct.

- “Lazy” reviews stay in the human-like range
- A review is labelled lazy as soon as one sentence is
- Genuinely low-information human texts may still exist

Publications

- M. Dubois, F. Yvon, P. Piantanida. **MOSAIC: Multiple Observers Spotting AI Content**. Findings of ACL 2025, Vienna.
- M. Dubois, F. Yvon, P. Piantanida. **How Sampling Affects the Detectability of Machine-written texts: A Comprehensive Study**. Findings of EMNLP 2025, Suzhou.
- M. Dubois, P. Piantanida, F. Yvon. **MOSAIC : Mélange d'experts pour la détection de textes artificiels**. TALN 2025, Marseille.
- M. Dubois, F. Yvon, P. Piantanida. **MOSAIC at GENAI Detection Task 3: Zero-Shot Detection Using an Ensemble of Models**. GenAIDetect Workshop, COLING 2025, Abu Dhabi.