Episode Book

EPISODE #6574

Testing Whether AI Rewrites Can Hide Your Identity

PUBLISHED	13 November 2026 at 21:56 UTC
DURATION	29:52
LISTEN	myweirdprompts.com/episode/ai-authorship-obfuscation-anonymity-55694/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	20 cited by the research (listed at the end)
BOOK GENERATED	14 November 2026 at 06:47 UTC

SYNOPSIS

Authorship obfuscation research shows large language models cannot guarantee anonymity when rewriting personal stories. While a FAccT conference whistleblower tool cut author attribution from 98.8% to 31.2%, it also erased a quarter of the story's meaning. Meanwhile, a plain Mistral prompt outperformed the specialized paraphraser DIPPER, and RANLP researchers found that LLM rewrites work well for some writers but barely at all for others. The biggest gap remains content—specific life details—which almost no automated tool scrubs, leaving writers weighing cloud capability against local privacy tools like JAMDEC.

TRANSCRIPT



Corn

The whistleblowing series has taken us to some uncomfortable places, but Daniel's new one is the first where the adversary isn't an institution. It's a person. One specific, dangerous person in somebody's life, who would find them if the story ever traced back.



Herman

And the story has to be told. That's the bind — to a journalist, to a support community, to a regulator. Silence isn't the option being weighed here.



Corn

Right — Daniel's framing is that changing names doesn't touch the real problem. The details give you away, the sequence of events gives you away, and even your writing style can give you away. So his question is whether a large language model can rewrite somebody's own story so they're unidentifiable while keeping what actually matters. And then the practical layer underneath: what do you send to which model, cloud versus local, how would you ever know if a rewrite is anonymous enough, and the risk the model quietly changes the facts on you.



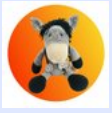
Herman

Plus the big one he tacks on the end — does AI rewriting actually defeat stylometry, or does it just move the arms race along.



Corn

And the literature has a name for this whole area, and its own consensus position, which is more sobering than anything you'd read in the marketing around AI privacy tools.



Herman

Authorship obfuscation. Sometimes called adversarial stylometry. It's a real, active field — benchmarks, attack papers, defence papers, at least one prototype built specifically for whistleblowers. This is not a fringe question Daniel has wandered into; there's a community of people who have been working on exactly this for years.



Corn

And they've reached a position on it that's worth stating up front, because it frames everything else we're going to say. The field's own reference material says, in effect: stylometric identification may not be reliable, but neither can non-identification be guaranteed. And the practice of obfuscating your style may itself be detectable.



Herman

That's the field talking, not a sceptic on the sidelines. The people who build these methods are the ones saying nobody can promise you anonymity.



Corn

Which, for the person in Daniel's scenario, is the sentence that matters. Everything else is degrees of help.



Herman

There's one more thing to establish before we get into the numbers, because it's the spine of the whole episode. The strongest re-identification signal in a personal story is usually not style at all. It's content. The town, the job, the number of children, the order things happened in. A model that perfectly sands off your comma habits while leaving "my ex-husband, a paramedic in a small town in Devon, took the car on Boxing Day" intact has done almost nothing.



Corn

And yet every method we're about to describe targets style.



Herman

Almost every one. There's exactly one paper in this pile that names the specific content of the disclosure alongside style as a re-identification risk. One. The research community has, by and large, chosen to work on the easier half of the problem.



Corn

So let's look at what the easier half actually achieves. The numbers are more interesting than the headlines.



Herman

Start with the most recent one, a paper from September with a title that tells you exactly who it's for — "I Am No One." The method builds a style profile of you, the author, along four dimensions: sentence length, vocabulary choice, tone, and punctuation patterns. It needs as few as five sample texts of your writing to do it, and the profile is short and human-readable — it's not some opaque vector, it's a description. Then you prompt a large language model to rewrite the text while suppressing those specific cues.



Corn

Five samples. That's a week of someone's forum posts.



Herman

A week, or one long evening. And the result across their benchmarks is a sixty to seventy percent reduction in authorship-attribution F1 — the classifier that could name you before mostly can't now — while keeping content quality and readability. The detail I like is that it's deliberately attacker-agnostic. It doesn't optimise against any specific adversary's classifier, because tuning your defence against a known attack is a false guarantee. It assumes the attacker exists and you don't know what they have.



Corn

A defensive posture I actually respect. What's the closest artefact to Daniel's actual use case?



Herman

The whistleblower tool. There's a sanitization prototype from the FAccT conference — the title is "Silencing the Risk, Not the Whistle" — built precisely around this threat model: someone discloses wrongdoing, and both the specific content of their disclosure and their distinct writing style can re-identify them as the source. It's semi-automated, a human stays in the loop, and the headline number is that it reduced attribution accuracy from ninety-eight point eight one percent to thirty-one point two two percent.



Corn

Ninety-eight to thirty-one. That sounds heroic until you say the other half.



Herman

Which is that roughly one in three attempts still lands on the right author. That's not anonymity, it's a bad coin flip weighted in your favour. And the tool's own faithfulness number — how much of the meaning survives — tops out at about seventy-three percent of semantics preserved. So in the best case, about a quarter of what you meant is gone.



Corn

The tool takes a quarter of your story as payment and still can't promise the door stays shut.



Herman

That's the honest ledger, yes. Now the finding that I think is the capability surprise of the whole pack. There's a study from the digital-health NLP workshop that tested obfuscation against authorship verification — "can you tell these two texts were written by the same person" — and the best performer wasn't a purpose-built obfuscation tool at all. It was a general-purpose chat model, Mistral, with a plain zero-shot prompt. Attack success rate of zero point nine two. The dedicated paraphrase model, DIPPER — the specialist, the thing built for exactly this — managed zero point five four.



Corn

The bespoke privacy tool lost to a chatbot told, in one sentence, to do its job.



Herman

Substantially. And that matters enormously for the practical question later, because it implies the strength is sitting in the big general models, the ones you reach through somebody else's API.



Corn

Hold that thought, because there's a result in this pile I found alarming, and it's not about the methods — it's about the people using them.



Herman

The bimodality finding. This one's from the RANLP conference — personalized authorship obfuscation with large language models. The headline is detection dropped from zero point nine four to zero point five zero with LLaMA, and to zero point six three with GPT-4. Those are averages. But the distribution underneath isn't a nice bell curve centred on the average — it's bimodal. For some authors, the rewrite works beautifully. For others, it barely works at all. Same method, same model.



Corn

Which means "I ran it through a chatbot" is a coin flip.



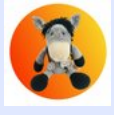
Herman

A coin flip whose outcome you can't observe. You don't get a readout telling you which mode you landed in. And there's a nasty correlation lurking in it, though the paper doesn't claim it — my speculation only — but the people it presumably fails hardest for are the ones with the most distinctive style. And distinctive writing style doesn't come from nowhere. It correlates with a distinctive life. The person with the unusual story to tell may be exactly the person whose telling of it is most fingerprinted.



Corn

Herman, let me push at something, because I want your clinician's read on this. You spent decades explaining small probabilities to frightened people. Someone in Daniel's scenario comes to you with that bimodal result. What do you actually tell them?



Herman

I tell them the average is a lie about them. The only honest version is: test it against your own writing. Take your five samples, obfuscate, then check whether a verifier can still pair the rewrite with the originals. If you can run that check, you at least know which mode you're in. The paper's own suggestion is personalised prompting to partially mitigate the bimodality — so it's not fixed, but "partial" is carrying a lot of weight in that sentence.



Corn

And that's the moment the whole thing stops being a tool and becomes a procedure. You can't buy anonymity; you can only test for it.



Herman

Which takes us to the gap I flagged at the start. Every method we've covered neutralises style — sentence length, punctuation, vocabulary. None of them touches the particulars. The whistleblower paper is the only one that even names the content of the disclosure as a risk, and even there, the sanitization is style-first with a human doing the content edits. A story re-identifies through what happened, where, to whom, in what order. The ex-husband who's a paramedic in a small town — that's not a style feature. That's a lookup key.



Corn

So the research field has industrialised the half of the problem that was never the half that gets people found.



Herman

The half that's tractable, at least. Style is measurable, it's benchmarkable, you can graph F1 going down and publish the graph. Content is bespoke, unbounded, human — there's no benchmark for "my particular awful situation," because by definition nobody else has it.



Corn

Which is darkly convenient, if you're a research programme.



Herman

I'd want to be fair to them. Style obfuscation is necessary even if it's not sufficient. If your comma habits are the last thread connecting you to a story whose content you've scrubbed, the style work is what cuts that thread. It's just not the first thread, and the literature mostly doesn't say so.



Corn

Then let's answer Daniel's first question head-on. Can an LLM rewrite a story so the person is unidentifiable while keeping what matters? The literature's answer is no. It can measurably degrade attribution — sixty, seventy percent reductions, one-in-three residuals — but no method claims a guarantee, the field itself says non-identification cannot be guaranteed, and the content half of the problem is mostly untouched.



Herman

And the person's own experience of that is the bimodal coin flip. That's the honest state of play on capability.



Corn

So the rewrite can measurably degrade attribution — but that raises the operational questions, because if I'm this person, my next question isn't "does it work," it's "what do I send where."



Herman

The cloud-versus-local fork, and the literature takes it more seriously than you'd expect. There's a method called JAMDEC — unsupervised obfuscation built on small language models, GPT-two-scale, deliberately small, and the paper says why in so many words: in order to avoid disclosing the original content to proprietary LLM APIs. The whole architecture is a response to the retention question. They use constrained decoding to close the gap between a small model and a big one, and the user keeps control of the knobs.



Corn

And the motivating examples — who's this for?



Herman

Blind review for scientific papers, anonymous online reviews, and this is the one that maps straight onto Daniel's prompt — anonymous interactions in mental health forums. The support-community case is literally in the paper's opening. There's a companion approach, TAROT, that does policy optimisation over small models for the same reason, and it states the bind plainly: strong obfuscation hides your identity but degrades the quality and usefulness of the text, while keeping the text useful tends to leave insufficient



Herman (continued)

privacy.



Corn

Now weigh that against the Mistral result from before.



Herman

That's the cruel arithmetic. The general-purpose cloud model with a plain prompt beat the purpose-built paraphraser almost two to one. So the privacy-preserving local route costs you obfuscation strength, and no study I know of has run cloud versus local head to head on the same texts with the same method — the comparison everyone wants just doesn't exist. What you're choosing between is a strong method that requires handing your document to a stranger, and a private method that's measurably weaker.



Corn

And the document you'd be handing over is the single most sensitive text you will ever write. Name, address, abuser, dates.



Herman

The text you send to a cloud model is the de-anonymised document. That's the asymmetry the papers don't resolve. Even with a no-training clause, even with deletion on request, you have handed a vendor a document containing both your identity and the person you're afraid of. The obfuscation happens on the far side of the network boundary. You must be identified to the model in order to be de-identified by it.



Corn

That's a sentence worth saying twice, so I'll let it stand. And it gets worse, doesn't it — the obfuscation step is also a profiling step.



Herman

"I Am No One" builds a machine-readable dossier of your stylistic fingerprint from five samples and stores it with the vendor. The tool that hides you is the tool that reads you. Nobody in this literature addresses that as a risk in itself.



Corn

Here's the part where you tell me if I've gone off the rails. The style-profiling step only needs five samples and produces a short human-readable description. That step could run on a laptop. Then you strip the identifying details by hand — the Devon, the paramedic, the Boxing Day. Then you send the profile plus the already-scrubbed story to the cloud model for the heavy rewrite. The vendor never sees the original.



Herman

I'd run exactly that. But I have to be honest about its status — no paper in this pile evaluates that hybrid. It's implied by the architecture of the pieces and tested by nobody. The strongest thing I can say is that it's the shape of the right answer, not a validated answer.



Corn

An untested hybrid, offered as such. Fine. Next question in Daniel's stack: how do you judge whether a rewrite is anonymous enough? I suspect "enough" is where everything falls apart.



Herman

There are three distinct approaches in the literature, and they disagree about what enough means. The dominant one is attack-based evaluation — you take an attribution classifier, point it at the rewrite, and count how often it still gets you. That's where all the numbers we've quoted come from. The weakness is structural: you are only testing against the attacks you thought of. "I Am No One" is explicit about this — that's why it's attacker-agnostic.



Corn

That's the fortress problem. You audit the walls you built.



Herman

Second approach is feature-level, and StyleRemix and ALISON both argue for it because, in their words, LLM obfuscation lacks interpretability and controllability — a single score tells you a classifier was fooled, not what changed. So instead you check the named markers: did your sentence-length distribution actually move, did your punctuation habits actually change, did the tone shift. It's like reading the blood panel instead of trusting the patient's word that they feel fine.



Corn

You would reach for the medical analogy.



Herman

I would, and it holds. Third approach is the newest and the most interesting — agentic red-teaming. There's a paper, SALA, where an LLM agent does the attribution: combines quantitative stylometric features with its own reasoning to guess the author, and then — this is the clever bit — you read the agent's reasoning trace, the "here's why I think this is the same person" chain, and you generate rewriting prompts that attack exactly what it noticed. Then run it again.



Corn

Use the attacker as the auditor.



Herman

It's a loop a person could actually run at home, which almost nothing else in this literature is. Show the model your rewrite next to your original writing and ask: same author? Then ask why, and fix what it names. That's the closest thing to a practical self-check anyone has published.



Corn

So the question we keep coming back to is whether you can hand your story to a model and trust it to keep you safe. There's a caller line lit up — we do occasionally get those. Agnes, go ahead.
Agnes: I have been listening along for a good while now and nodding, so I hope you will forgive me, but what actually is a large language model? Is it a thing like a very big spellchecker on your computer, and do you have to keep it plugged in and fed?



Herman

A fair question after everything I just threw at you, Agnes. Think of it like a very well-read neighbour who has read more letters than anyone alive but has never once been outside — you ask it to help with your writing, and it answers by guessing what usually comes next in sentences like yours. No feeding, just electricity.



Corn

Well-read neighbour — after the graph neural networks, that's the gentlest downgrade you've given a model all year.



Herman

Back to Agnes's well-read neighbour, then — the thing the self-check can't give you is a threshold. That's the gap that should stop anyone sleeping. No paper in this entire pile proposes a number, a certificate, a standard for "anonymous enough." The whistleblower tool's thirty-one percent residual is presented as an improvement over ninety-eight, not as a pass. Nobody in this literature says below X percent is safe, because nobody can.



Corn

Because "safe" depends on the adversary. Thirty-one percent residual against a lazy journalist is fine. Thirty-one percent against the man himself, reading every paragraph of a story about his own marriage, is a different number entirely.



Herman

He doesn't need a classifier. He needs a nod of recognition. The stylometric frame quietly assumes the attacker is a stranger with software, and in Daniel's scenario the attacker is the person who knows the content best in the world.



Corn

And that gets us to the question I've been saving, because it's the one I'd lose sleep over. The model changing the facts. Walk me through how bad that actually gets.



Herman

First-order, not a footnote. The whistleblower tool's own number caps at seventy-three percent semantics preserved — up to roughly a quarter of the meaning gone. The multilingual benchmark is starker: ten obfuscation methods against thirty-seven detectors across eleven languages, and some methods severely damaged the text — changed language, weird characters, no longer readable by humans. Automated obfuscation can silently destroy the document.



Corn

Destroyed text is at least visibly destroyed. I'd almost prefer that to the quiet version.



Herman

The quiet version is the dangerous one, and here's where I think the literature has a genuine blind spot that matters for Daniel's use case specifically. In a personal narrative, the facts are the evidence. A model that smooths "he took my phone on the fourteenth" into "he would sometimes take my phone" has not paraphrased — it has converted a specific, testable claim into a vague habitual one. Say that story later gets checked against a police report, a custody filing, a journalist's fact-check. The weakening isn't a style choice, it's a wound in the account.



Corn

And no one measures that.



Herman

No paper found treats narrative-fact fidelity — dates, sequence, quoted speech, causal claims — as its own evaluation target. Everything measures general semantic similarity or downstream task utility. Nobody has measured how often an LLM rewrite of a personal story silently moves a date or merges two incidents into one. Given how paraphrase models behave, my guess is it happens constantly and nobody has counted.



Corn

The one metric this use case needs most, and it's the one nobody's built.



Herman

TAROT exists practically because of this — the whole privacy-utility framing — but utility there means the text still works for its task, not that the events still stand as testimony. Different thing entirely.



Corn

Then the last of Daniel's questions, the arms race. He asked whether AI rewriting defeats stylometry or just shifts it. My read of this pile is that shifting is exactly what it does, and the shift has a direction.



Herman

It does, and the direction should worry this use case specifically. The newest attribution work goes below the surface. When style is obscured, reasoning structure becomes the fingerprint — a paper from this year extracts argument structure, the shape of how you reason, and classifies it with a graph network, beating a strong baseline by up to twenty-seven points specifically under paraphrase and backtranslation obfuscation. Twenty-seven points. The better the obfuscation gets, the more the advantage grows.



Corn

Say what that means for the person writing their story.



Herman

Your comma habits are a coat you can take off. How you build an argument — where you put the exception, what you concede, how you circle back — that's closer to the bone. And a personal narrative is nothing but reasoning about events: why I stayed, what I noticed first, when I understood. The rewrite can sand the surface and leave the skeleton standing there with your name on it.



Corn

Meanwhile the obfuscation itself may be detectable.



Herman

That's the uncertain one — the field says so honestly. Some methods leave signals in the output, but a stylometrist who doesn't know which method was used may not be able to spot it reliably. Still uncertain either way, and here's the institutional part: PAN, the big evaluation lab for authorship and AI text detection, has made obfuscated authorship a formal detection task this cycle. The forensic community is now explicitly training against this exact defence.



Corn

They've put it on the syllabus.



Herman

And there's a result from late last year that complicates the whole idea of neutralising style — push the neutralisation far enough and the signal comes back. Style turns out to be layered; it isn't one dial you can zero out. Overdo the sanding and you stand out for being sanded.



Corn

That one has a nasty symmetry to it. The rewrite that most hides you is also the rewrite that most looks rewritten.



Herman

Then two ripple effects, and I think these are the ones to leave hanging. First, aggregation. If you post your story to three forums and a subreddit, each individually anonymised, you may have re-identified yourself by aggregation — the paper from this year puts it that users release multiple texts under one identity, giving attackers a profile that reveals more than any single text, while every existing method optimises each document in isolation. The bundle is the leak.



Corn

Nobody launders one banknote. They launder the account.



Herman

Second, the homogenisation paradox. If LLM rewriting works, everyone's anonymised story sounds like the same model. The corpus becomes attributable to that model — and the residual human signal, whatever the rewrite failed to sand off, stands out by contrast against the flat background. The better the tool works, the more your leftovers show.



Corn

So even success sharpens the search. Every one of these findings points the same way — the arms race doesn't end with the survivor holding a working shield; it ends with a moving target and a scoreboard nobody's published. Whether a hybrid pipeline changes that, honestly—



Hilbert

No, it doesn't, Herman. I tried something like it.



Hilbert

The council records office, two thousand fourteen. I was filing an objection to a demolition on Raglan Street, and I didn't want my name attached because I still had to live on that street. So I typed the whole objection in the style of a woman two doors down, Mrs. Oakes, who wrote to the local paper every week. I'd kept her letters. Twenty-two of them, in a drawer, because I liked her use of the semicolon. Mrs. Oakes used them like most people use a full stop, which is to say, correctly.



Hilbert

The objection went in. The planning officer, a Mr. Devane, spent nine months corresponding about it, and the file ended up logged under her name. Not mine, hers. The semicolons did it. His office ran nothing, they just read like humans do. She got a thank-you card from the preservation society. I clipped it to the file.



Corn

She never objected to the demolition?



Hilbert

Mrs. Oakes didn't know the building was coming down. She was glad of the parking.



Corn

Herman, that's the consequence neither of us named. Good obfuscation doesn't just hide you — it can hand your words to somebody else whole. The whistleblower tool measures attribution dropping, but attribution that drops on you has to land somewhere, and a rewrite in a borrowed voice puts the borrowed voice in the frame.



Herman

And it's more than unfair to Mrs. Oakes — it's an operational risk for the survivor. If the dangerous person is looking for the author and the text reads like a neighbour, you've moved the danger next door. Obfuscation evaluated only as "not me" was never evaluated as "somebody else, possibly someone findable."



Corn

We were auditing the walls we built, and the answer turns out to be that a good enough disguise is an accusation with the name left blank.



Herman

Daniel's question about whether the arms race is won — the answer the research gives is no, and the answer Hilbert's filing cabinet gives is that even the victories have bystanders.



Corn

Here's what stays with me, though. The threat model in Daniel's prompt — a specific dangerous individual, not a corporation, not a government — appears nowhere in this literature. No paper we found addresses survivors of abuse specifically. The closest artefacts are a whistleblower tool built for employees and a method that mentions mental-health forums in its motivations. The person who needs this most is represented by nobody's benchmark.



Herman

And the unanswered question under all of it — whether the obfuscation itself is detectable — is now a formal competition task, with forensic teams lining up to beat it. The direction of travel is against the defender.



Corn

The honest close is the gap: the research offers measurable degradation, and the person in this situation needs something no paper provides — a threshold, a fact-fidelity guarantee, a threat model with their actual life in it.



Herman

Thanks as always to our producer, Hilbert Flumingtop, at the desk through all of it.



Corn

If you want more of this, try episode eight, Building Your Own Whisper; and episode ten, How ASR Went From Frustration To … Whisper Magic. This has been My Weird Prompts. If the show's earned it, leave us a review wherever you listen — it's how people like Agnes find us.



Herman

We'll be back soon. See you then.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Khan, He, Wachowicz, Naghizade, I Am No One: Style-Aware Paraphrasing for Text Anonymization [primary]**

   (v1, 2026-09) .
   https://arxiv.org/abs/2609.12341v1

2. **Silencing the Risk, Not the Whistle: A Semi-automated Text Sanitization Tool for Mitigating the Risk of Whistleblower Re-Identification, FAccT '24 [primary]**

   .
   https://doi.org/10.1145/3630106.3658936

3. **Alperin, Leekha, Uchendu, Nguyen et al., Masks and Mimicry: Strategic Obfuscation and Impersonation Attacks on Authorship Verification, NLP4DH @ NAACL 2025 [primary]**

   .
   https://aclanthology.org/2025.nlp4dh-1.10.pdf

4. **Shokri, Levitan, Levitan, Personalized Author Obfuscation with Large Language Models, RANLP 2025 [primary]**

   .
   https://aclanthology.org/2025.ranlp-1.133.pdf

5. **Xing, Venkatraman, Le et al., ALISON: Fast and Effective Stylometric Authorship Obfuscation, AAAI-24 [primary]**

   and https://arxiv.org/abs/2402.00835v1 .
   https://ojs.aaai.org/index.php/AAAI/article/view/29901

6. **Fisher, Lu, Jung et al., JAMDEC: Unsupervised Authorship Obfuscation using Constrained Decoding over Small Language Models [primary]**

   .
   https://arxiv.org/abs/2402.08761v1

7. **Fisher, Hallinan, Lu et al., StyleRemix: Interpretable Authorship Obfuscation via Distillation and Perturbation of Style Elements, EMNLP 2024 [primary]**

   and https://aclanthology.org/2024.emnlp-main.241/ .
   https://arxiv.org/abs/2408.15666v1

8. **Loiseau, Sileo, Riquet et al., TAROT: Task-Oriented Authorship Obfuscation Using Policy Optimization Methods, NAACL PrivateNLP 2025 [primary]**

   .
   https://arxiv.org/abs/2407.21630v2

9. **Bao & Carpuat, Keep It Private: Unsupervised Privatization of Online Text [primary]**

   .
   https://arxiv.org/abs/2405.10260v1

10. **Macko, Moro, Uchendu et al., Authorship Obfuscation in Multilingual Machine-Generated Text Detection, EMNLP 2024 Findings [primary]**

    .
    https://arxiv.org/abs/2401.07867v3

11. **Kikteva, Romazanov, Hautli-Janisz et al., Show Me How You Reason and I'll Tell You Who You Are: Reasoning Graphs for Robust LLM Authorship Attribution [primary]**

    (v2, 2026-07-18) .
    https://arxiv.org/abs/2607.14905v2

12. **Ma, Squicciarini, Rajtmajer, Aggregation-Aware Synthetic Text Generation Against Authorship Re-Identification, EMNLP 2026 [primary]**

    .
    https://arxiv.org/abs/2608.22161v1

13. **Zhang & Zhang, Assessing Deanonymization Risks with Stylometry-Assisted LLM Agent [primary]**

    .
    https://arxiv.org/abs/2602.23079v1

14. **Mahbub & Feng, Mitigating Self-Preference by Authorship Obfuscation [primary]**
.
https://arxiv.org/abs/2512.05379v1

15. **Nguyen, Hu, Le, Unraveling Interwoven Roles of Large Language Models in Authorship Privacy: Obfuscation, Mimicking, and Verification [primary]**
.
https://arxiv.org/abs/2505.14195v1

16. **Dilworth, StegoStylo: Squelching Stylometric Scrutiny through Steganographic Stitching [primary]**
(v5, 2026-06-19) .
https://arxiv.org/abs/2601.09056v5

17. **Bevendorff, Fröbe, Greiner-Petter et al., Overview of PAN 2026 [primary]**
.
https://arxiv.org/abs/2602.09147v1

18. **STAMP-R: Stylometric Text Anonymization with Memory-guided Policy Rewriting, PrivateNLP 2026 [primary]**
.
https://aclanthology.org/2026.privatenlp-main.4.pdf

19. **Wikipedia, Adversarial stylometry**
retrieved 2026-11-13, last edited 2026-09-21.
https://en.wikipedia.org/wiki/Adversarial_stylometry

20. **Wikipedia, Stylometry**
retrieved 2026-11-13, last edited 2026-11-13.
https://en.wikipedia.org/wiki/Stylometry