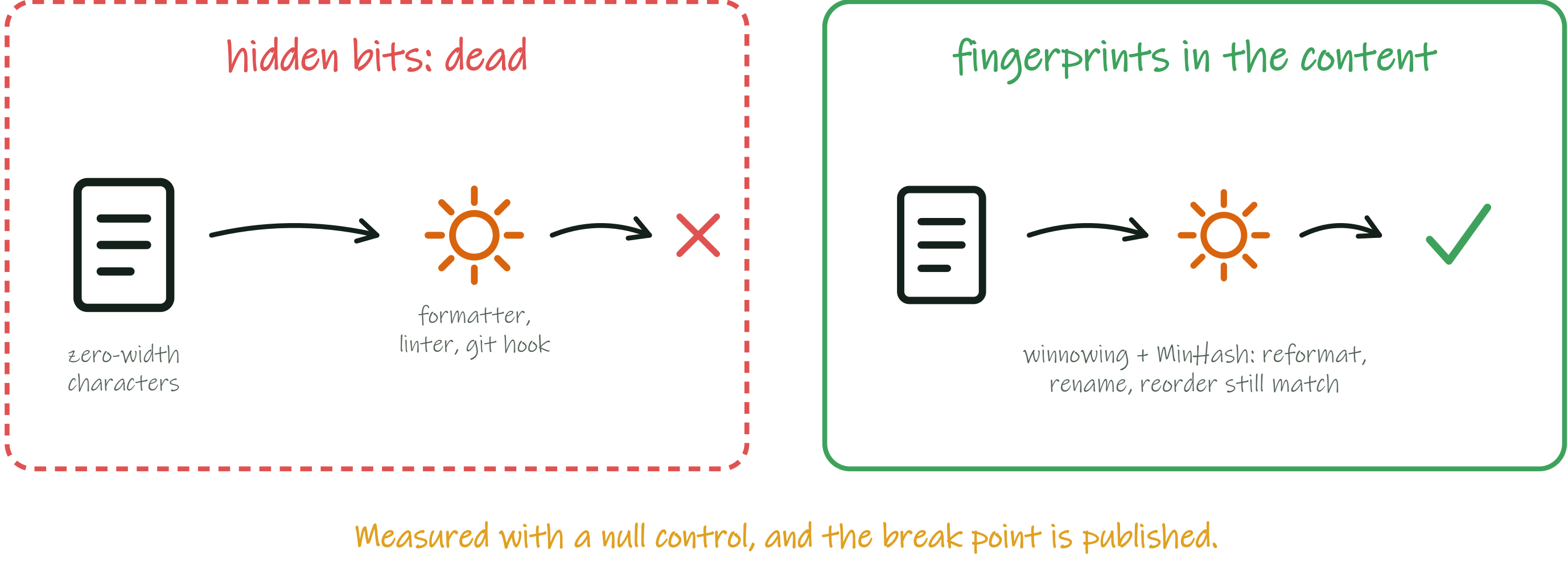


When Steganography Dies: Robust Text and Code Provenance via Winnowing and MinHash Fingerprints

Hagen Schmidt, inventor · research in the association Lebensfluss e.V., Gratkorn near Graz · spin-out in preparation · ip@destill.ai

Any formatter strips a hidden-bit watermark, so we measure fingerprints that live in the content itself.

Jeder Formatierer entfernt versteckte Bits, darum messen wir Fingerabdrücke, die im Inhalt selbst stecken.



Hidden bits do not survive editing

Zero-width spaces and trailing-whitespace bits disappear in one pass of a formatter, linter, Unicode normalizer or git hook. Copy detectors also tend to report one number and leave out how often unrelated documents get flagged.

Versteckte Bits überleben keinen Formatierer, und Kopiererkenner verschweigen meist ihre Falschtreffer.

3/3

code reformat, rename, comments: all detected

0.95

text similarity after sentence reorder (3/3)

0

false positives in both null controls

0.00

heavy paraphrase: missed, the stated boundary

Intrinsic fingerprints, strongest signal, fixed threshold

1

Normalize

Text: lower-case, drop stop words, stem. Code: identifiers to ID, literals to STR/NUM, drop comments.

2

Fingerprint

Winnowing over k-grams (text k=4, code k=5), plus MinHash over 2-shingles for text and a call/literal multiset for code.

3

Compare

Similarity is the strongest signal: winnowing containment in either direction, MinHash agreement, or multiset Jaccard.

4

Decide with a null control

DETECTED at $\tau = 0.40$ (weak 0.25–0.40). Unrelated originals must score below τ .

Normalisieren, Fingerabdruck bilden, stärkstes Signal nehmen, an $\tau = 0.40$ entscheiden, mit Nullkontrolle.

Scorecard (paper, $\tau = 0.40$; mechanical transforms n = 3)

MODALITY	TRANSFORMATION	SIMILARITY	VERDICT
Text	Whitespace + case / fragment in larger doc	1.00 (mean)	DETECTED 3/3
Text	Sentence reorder	0.95 (mean)	DETECTED 3/3
Text	Light / moderate / heavy reword (n=1 each)	0.40 / 0.01 / 0.00	borderline / MISSED / MISSED
Text	Null control (3 unrelated pairs)	max 0.00	0 false positives
Code	Reformat, rename, comments, snippet in file	1.00 (mean)	DETECTED 3/3 each
Code	Extract-method / control-flow refactor (n=1)	1.00 / 1.00	DETECTED
Code	Python port / reimplement with different API (n=1)	0.67 / 0.38	DETECTED / weak
Code	Null control (3 unrelated pairs)	max 0.33	0 false positives

Three originals per modality; the hard cases are hand-made single instances (n=1). Light reword is 0.398 in the manifest, printed as 0.40 in the paper. One command: textcode-scorecard.js, manifest textcode-2026-06-30.json.
Struktur und Reihenfolge werden erkannt, die echte Neufassung entkommt, beide Nullkontrollen ohne Falschtreffer.

Where the fingerprint breaks: share of content words changed by the attacker

Substitution only · k=4 (paper)



+ sentence reordering · k=4



+ reordering · k=2 (since 2026-09-07)



■ all 7 recognised ■ partly ■ none □ not reported · n=7 documents: direction, not magnitude. Back-translation, 7 texts: recognised 4/7 with k=4, 7/7 with k=2.
Wo der Fingerabdruck bricht: Anteil der geänderten Inhaltswörter. Neu seit der Einreichung: k=2 verschiebt den Bruchpunkt von 20,8 % auf 33,7 %.

Honest limits (from the paper)

- ! Rewrite boundary: heavy paraphrase and different-API reimplementation escape by construction; past that point a match would be false.
- ! Hidden-bit layer is dead: kept only as exact-copy tamper evidence. Text MinHash needs paragraph-length input.
- ! Small corpus: three originals per modality, n=1 hard cases, a single $\tau = 0.40$ with no ROC yet.
- ! Code refactor robustness depends on the call/literal multiset; a refactor that also changes the API degrades toward reimplementation.

Grenzen: Neufassung entkommt, kleiner Korpus, eine feste Schwelle, Text braucht Absatzlänge.

Since submission

2026-08-25

Change-rate curve on 7 real English documents (substitution only, k=4 engine): 7/7 recognized up to 18.7% changed content words, 2/7 at 24.1%, 0/7 at 36.1%; false positives 0/4 same-topic, 0/17 unrelated.

C:/Fortress/Documentation/SSOT/evidence/PARAPHRASE_TEXT.md

2026-09-04

Adding sentence reordering, which changes 0.0% of scored tokens, moves the break point from 24.1% to 18.7% (k=4 engine). Total loss stays at 36.1%.

C:/Fortress/Documentation/SSOT/evidence/TEXT_ANGRIFF_31.md

2026-09-05

Blind round-trip back-translation (EN to Arabic/German to EN, n=7): at the same change rate it costs on average 0.151 more similarity than substitution; 3 of 7 fall below τ (k=4 engine).

C:/Fortress/Documentation/SSOT/evidence/TEXT_RUECKUEBERSETZUNG_35_NEUHEIT_36.md

2026-09-07

Engine change: text winnowing k=4 to k=2. Break point under substitution plus reordering 20.8% to 33.7%; back-translation 4/7 to 7/7; false positives still 0/56 same-family and 0/1169 unrelated, but the same-family maximum rose 0.086 to 0.188. n=7: direction, not magnitude.

C:/Fortress/Documentation/SSOT/evidence/TEXT_AUFLÖSUNG_33B.md

Nach der Einreichung gemessen: kürzere Schindel k=2 verschiebt den Bruchpunkt auf 33,7 %, bei weniger Abstand zur Falschtreffer-Schwelle.

Open standard proposal — and four ways to build this with us

We propose this null-controlled, per-transformation scorecard as an open evaluation protocol for text and code provenance, for example as input to CEN-CENELEC JTC 21 or C2PA soft binding, and seek co-proposers.

Wir schlagen diese Bewertungskarte mit Nullkontrolle als offenes Prüfprotokoll vor und suchen Mit Antragsteller.

IP licence or acquisition

Patent applications filed. License or acquire the engine and its reproducible harness.

Co-founder / takeover partner

Run the provenance business with us, from pilot to scale.

Standard co-proposer

Propose the null-controlled scorecard with us as an open protocol.

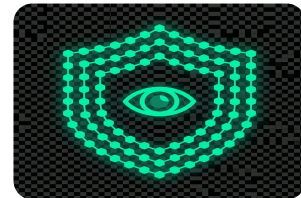
Research / funding consortium

Join a Eurostars, FFG or Horizon consortium for a labelled corpus.



destill.ai/ais26 · ip@destill.ai

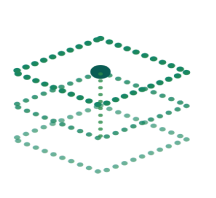
Handout with every table and number · Papers · Gespräch am Poster 13:30–14:30



DESTILL .ai



FORTRESS



NI-Stack