

TracerYX

Introducing TracerYX PII Redaction & Tokenization

First-party detection. Customer-local processing.
Configurable protection with measured evidence.

95.79%	90.36%	Customer-local
Complete-span coverage	Character F1	Protection before export

Latest primary evaluation: 797 retained English documents from NVIDIA Nemotron-PII and Gretel PII masking; 790 annotation-valid. TracerYX at threshold 0.20. [1]

Executive summary

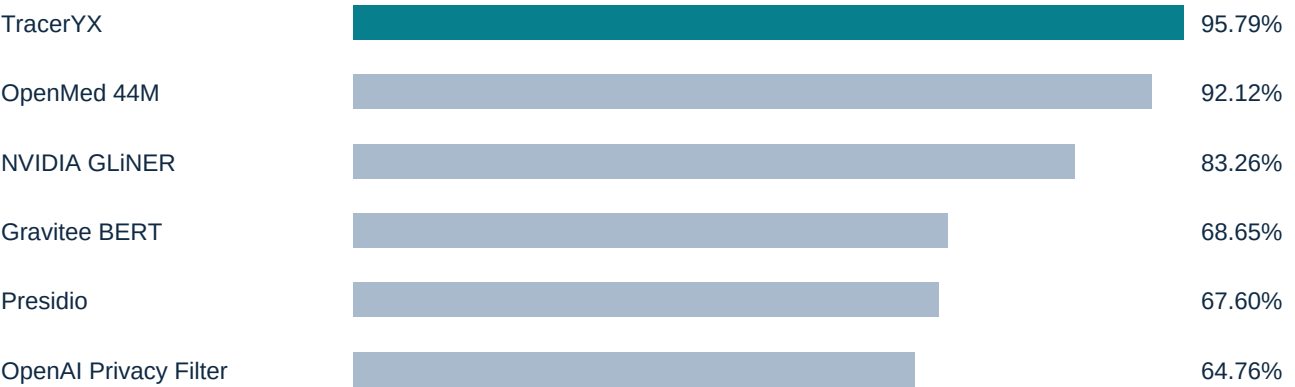
TracerYX combines first-party contextual models and deterministic recognizers with local redaction and opaque-reference tokenization. On the latest primary cohort, it achieved the highest observed complete-span coverage among the six tested system configurations. [1]

Against the tested Presidio setting, coverage was higher by **28.19 percentage points** using displayed table values. TracerYX removed 3.64% of unannotated characters, compared with 3.42% for Presidio. [1]

Review edition | Engine: traceryx-owned-5-rc5 | Local SDK: 0.1.0-preview.2
This is a strengths-led engineering evidence report. No new benchmark was run for this document. Results retain their original cohorts, settings and measurement boundaries. [1, 4]

Strong complete-identifier coverage

Complete-span coverage requires every character of a reference identifier to be protected. This is more demanding than detecting any part of a name, address or identifier. Predicted entity type is scored separately. [1]



System / setting	Full spans	Char. F1	Removal	Errors
TracerYX / .20	95.79%	90.36%	3.64%	0
Presidio / .01	67.60%	77.15%	3.42%	0
Gravitee BERT / .30	68.65%	81.96%	0.86%	0
OpenMed 44M / .10	92.12%	93.45%	2.15%	0
NVIDIA GLiNER / .50	83.26%	87.78%	0.83%	0
OpenAI Privacy Filter	64.76%	75.96%	0.79%	53

797 retained documents; 7 annotation-unknown Nemotron documents excluded from accuracy metrics and explicitly tracked. Completed valid documents include 124 with at least one reference miss. Processing failures remain in coverage denominators. [1]

What customers can take from this

The tested configuration offers strong full-identifier coverage with local inference. OpenMed has the highest character F1 and lower removal in this panel; TracerYX leads the pooled full-span metric. These are distinct strengths. [1]

Comparator thresholds were frozen before test acquisition. OPF uses CPU inference and a 2,048-token context window; its 53 errors materially affect its score. These observations concern tested integrations and hardware, not the best achievable performance of each model. [1]

Choose the coverage-preservation balance

Customers can choose a preset or an administrator-authorized threshold. Lower learned-score thresholds generally protect more text; mandatory deterministic findings remain active. A threshold is an operating control, not a privacy probability. [2]

RC5 preset / threshold	Full spans	Removal	Miss documents
High coverage / .05	95.31%	7.84%	45
Balanced / .20	92.55%	4.57%	76
Preserve text / .40	89.47%	3.51%	96

Separate RC5 cohort: 311 synthetic documents across three sources; structurally valid annotations. This is not the 797-document latest cohort. [2]

Default results by source in the same RC5 study

Source	Documents	Full spans	Removal	Char. F1
Nemotron	100	92.91%	3.48%	87.00%
Gretel PII masking	100	99.52%	7.43%	89.73%
Gretel finance / English	111	88.64%	4.61%	75.51%

Effective policy is recorded

The local gateway supports administrator defaults, bounded authenticated request overrides and receipts recording model/policy version, effective sensitivity, mandatory rules and tokenization mode. The reported controls are API and policy-file surfaces. [2]

Evidence for the retained default

On this cohort, Balanced exceeded Presidio .01 coverage by +30.02 points; the paired descriptive 95% interval was [+27.60, +32.52]. Removal was lower by 0.93 points, with interval [-1.79, -0.06]. These are unadjusted source-stratified document-bootstrap comparisons, not universal ranking tests. [2]

A first-party pipeline, executed locally

Stage	Technique and product behavior
1. Validate	Validate input, policy and resource limits before transformation. Map normalized text back to original offsets.
2. Detect	Two first-party sparse contextual taggers; deterministic identifier/checksum patterns; bounded field and prose context; optional known-value matching.
3. Discriminate	A first-party gradient-boosted corrector: 160 trees over 92 numerical features, including token shape, nearby margins, punctuation and rule membership.
4. Resolve spans	Combine selected token runs and mandatory findings; repair supported identifier boundaries and overlaps.
5. Transform	Redact detected spans or issue opaque references. SECRET/PAYMENT findings follow discard policy.
6. Export	The local OTLP integration preserves validated structural IDs and timestamps. Processing failures do not fall back to exporting raw input.

External models are offline comparators; the product does not invoke Presidio, competitor weights or an external inference API. Generic software libraries are used for training/building. [2-4]

Observed local SDK example

Synthetic input: customer Jordan Patel; primary and backup email jordan.patel@example.com; a delivery address; birth date; customer ID; a secret; a payment card; handling instruction. [4]

Field	Observed behavior (token IDs abbreviated)
customer_name	[PERSON:...]
user_email + backup_email	[EMAIL:same-token] in both fields
delivery_address / birthdate / customer_id	[ADDRESS:...]/ [DATE:...]/ [ACCOUNT:...]
authorization / card_number	[discarded-sensitive]
handling_note	Keep it refrigerated

Within a local SDK call/batch, repeated exact same-class values share a token. A later batch receives a new token. These ephemeral SDK aliases are not a reversible vault. The gateway reference store has separate tenant/owner/scope and transaction controls; do not transfer those guarantees to the browser/SDK alias map. [2, 4]

Conformal prediction: useful research, scoped evidence

TracerYX has evaluated an offline document-level conformal layer. It scores all annotated private characters, including characters beyond the initial proposals, and derives a finite-sample expansion threshold from separate calibration data. Deterministic protections remain included. [1, 2, 7]

Promising historical coverage gain

RC4 policy / 303-document cohort	Full spans	Removal	Held
Primary .20	94.48%	4.33%	0
Conformal / 5% target	99.58%	54.26%	0
Conformal / 1% target	62.79%	93.05%	100

The 5% policy increased complete-span coverage, but at substantial text-removal cost. Held documents receive zero detection credit. [3]

Latest independent risk-check outcome

Latest primary policy / 797 documents	Full spans	Removal	Held	Miss docs
Conformal / 5% target	35.68%	6.90%	400	20
Conformal / 1% target	36.16%	81.83%	400	0

The zero observed released-document miss count at 1% accompanies 400 withheld inputs and heavy removal. It does not establish perfect useful redaction. No evaluated joint policy passed all source-level limits on misses, mean removal, removal-tail risk and withholding. [1]

Relationship to the supplied healthcare manuscript

The work adopts the principle of checking privacy proxies together with utility and withholding. It does not implement the manuscript's population re-identification model, patient-level betting guarantees or review optimizer. Annotation coverage is a measurable proxy, not re-identification risk. [1, 7]

Conformal targets require suitable exchangeability and correct annotations. Customer-adjusted thresholds carry no statistical certificate. The evaluated conformal wrapper is offline; the commercial SDK evidence does not establish a certified conformal deployment. [1, 2, 4]

Historical breadth, with frozen settings

The RC4 study tested a broader panel on 303 synthetic documents from Nemotron, Gretel and English finance; five documents had unresolved annotations. Rows below retain the original reference or development-selected settings. [3]

Configuration	Full spans	Removal	Char. F1
TracerYX RC3	89.97%	6.57%	76.00%
TracerYX RC4 primary (0.20)	94.48%	4.33%	83.54%
TracerYX RC4 utility (0.40)	91.30%	3.26%	85.65%
Gravitee BERT small (0.01, reference)	77.65%	1.27%	85.29%
Gravitee BERT small (0.30, tuned)	77.60%	1.27%	85.27%
OpenMed 44M (0.01, reference)	77.28%	2.57%	83.51%
OpenMed 434M (0.01, reference)	76.11%	2.33%	82.61%
Presidio (0.20, reference)	65.98%	4.93%	66.73%
Presidio (0.01, tuned)	66.99%	5.42%	66.59%
Gretel GLiNER (0.70, reference)	49.04%	0.56%	64.88%
Gretel GLiNER (0.01, tuned)	64.17%	1.66%	73.86%
GLiNER multi PII (0.50, reference)	77.34%	4.27%	73.78%
GLiNER multi PII (0.70, tuned)	73.04%	3.13%	74.24%
GLiNER2 PII (0.50, reference)	69.16%	2.24%	73.92%
GLiNER2 PII (0.20, tuned)	76.75%	3.30%	75.59%
NVIDIA GLiNER (0.50, reference)	69.27%	1.77%	81.81%
NVIDIA GLiNER (0.60, tuned)	68.52%	1.63%	81.42%
Presidio (0.50)	61.04%	4.80%	64.56%
spaCy NER diagnostic	44.64%	4.22%	50.32%
Scrubadub defaults	9.24%	0.67%	25.00%
OpenAI Privacy Filter	63.00%	0.88%	74.47%

Common development threshold search was recorded. Reference settings are benchmark reference points, not necessarily SDK defaults. A Presidio .00 SDK-default audit matched the .01 outputs in the recorded cohorts. GLiNER2 terminal-period mapping was analyzed separately; its integration errors are not presented as clean model superiority. [3]

Evidence across three public dataset families

The frozen v2 study compared eight configurations on 1,650 documents/conversations, with 2,250 separate calibration units. Nemotron/Gretel evaluation cohorts were reused from prior frozen-v2 checks; PII-TRACE was newly introduced. Eighteen Nemotron evaluation annotations were unresolved. [5]

Configuration	Nemotron	Gretel	PII-TRACE	Pooled
TracerYX v2	92.84%	96.49%	95.85%	95.05%
v2 + conformal 5%	99.35%	98.35%	99.04%	98.82%
Presidio .20	66.98%	63.50%	62.56%	64.63%
Presidio .50	62.65%	54.01%	60.93%	58.25%
Scrubadub defaults	14.19%	13.91%	27.28%	16.07%
spaCy NER	43.70%	33.68%	32.69%	37.21%

Full-span coverage. v2 pooled unannotated removal: 4.88%; v2 conformal 5%: 33.17%. Conformal 1% withheld all 400 Nemotron inputs. Different native privacy taxonomies and annotation conversions apply. [5]

Measured progress in the context-v2 work

On a fresh 1,000-document Gretel cohort, coverage rose from 88.79% to 96.49%. Fully covered street-address references rose from 36/161 to 138/161; address references from 24/111 to 102/111; customer IDs from 173/206 to 198/206. Removal increased by 0.55 points. These are historical development milestones, not RC5 test scores. [6]

Instruction-preservation improvement

A later narrow sentence-boundary fix preserved complete coverage on 506 additional documents while slightly reducing unnecessary removal. Larger retraining attempts were rejected when source-level or instruction gates failed. RC4 later preserved 120/120 historical instruction probes and covered 90/90 accompanying identifiers. Those probes are exposed regressions, not fresh population evidence. [3, 6]

The original 42-case smoke test and first 400-document owned-model evaluations are retained as baseline history. Their results must not be represented as current SDK accuracy. TRACE false-positive controls and later benign-text failures remain part of the evidence record. [5-7]

Broader documents, clearly separated

Latest RedactionBench diagnostic: 200 documents

System	Mandatory spans	Diagnostic F1	Outside removal	Errors
TracerYX .20	74.38%	36.43%	22.20%	7
Presidio / selected	64.63%	24.12%	11.78%	1
Gravitee BERT	50.29%	32.36%	3.19%	1
OpenMed 44M	44.18%	53.45%	6.56%	14
NVIDIA GLINER .50	2.48%	3.24%	1.60%	173
OpenAI Privacy Filter	0.34%	0.44%	0.20%	195

A separate annotation policy distinguishes mandatory and contextual spans. Diagnostic F1 is not official R-Score. Seven TracerYX processing errors exceeded existing input limits. The post-hoc within-limit slice reached 93.37% mandatory-span coverage; the full 200-document result remains the headline. Competitor failures materially affect their scores. [1]

Selected category strengths, with sample sizes

Category	Valid documents	Mandatory-span coverage	Outside removal
Emails	16	98.76%	11.77%
Medical	16	99.35%	11.29%
Legal	18	97.69%	5.11%
Operations	15	96.65%	6.33%
Logs	17	59.50%	43.72%

These small exploratory slices illustrate strengths and variability. The medical slice does not establish clinical readiness. [1]

Human-annotated TAB legal check

On RC5's separate 32-document TAB check, the target-sensitive adaptation covered 74.16% of reference spans versus 65.45% for Presidio. Under the all-entity adaptation, RC5 covered 62.36% versus 63.78%, with F1 68.56% versus 77.82%. The adapted metrics differ from official TAB scoring and cannot be pooled into the synthetic headline. [2]

Protection beyond the detector

Evidence	Verified result / qualification
Retained privacy implementation	646 tests passed in the latest droplet rerun; test counts measure specified behaviors, not population accuracy. [1]
Native model parity	1,649 development documents; zero threshold-decision mismatches against the Python reference. [2, 3]
Gateway transaction integration	11 PostgreSQL + actual CLI checks passed in the RC5 study; scoped references and rollback paths tested. [2]
Commercial local SDK build	61 automated checks across SDK, integration, UI and commerce suites; offline package install and signatures checked. [4]
Account-to-protected-trace E2E	14 grouped commercial checks, 16 installed-SDK behavior assertions, 17 focused regressions and 2 browser regressions. Real built app, installed package, exporter, sidecar and collector. [4]
Malformed input / upstream failure	Malformed OTLP rejected without forwarding; upstream failure receives protected payload, with no raw fallback. [4]

Typed protection, measured separately

In the latest 192-reference credential/payment diagnostic, 190 were fully detected and 154 fully discard-typed; 36 detected references lacked complete discard typing and two were missed. Full-span recall cannot substitute for semantic handling. [1]

Release decision and deployment scope

RC6 experimental candidates did not pass all release guards; the existing RC5 engine was retained. The installed preview uses frozen RC5 assets. Production billing was not activated by the reported tests. Stripe responses and release approvals in commerce E2E were synthetic fixtures. [1, 4]

Supportable positioning: **customer-local, first-party redaction with configurable sensitivity, opaque tokenization and strong measured complete-span coverage on named English benchmarks.** A production release requires its separate operator, legal/redistribution and signing approvals. [4]

Residual misses, benign-text false positives, long-input failures and domain differences remain. These qualifications define how to use the evidence; they do not erase the measured wins. No perfect-redaction, clinical-safety, compliance or market-leadership certificate is asserted. [1-4]

Benchmark register and source map

This review consolidates the benchmark stages found in the workspace and retained droplet reports. Cohorts overlap across some historical studies: their sizes must not be summed as independent evidence. Accuracy and integration checks have different denominators.

Stage	Dataset / evidence	Treatment here
Initial smoke and owned model	42 authored cases; initial and revised 400-record Nemotron evaluations	Historical baseline [7]
Initial conformal & Gretel	400 Nemotron test records; 1,000 Gretel test records	Historical calibration/transfer [7]
Context v2 / three-dataset	Nemotron, Gretel, PII-TRACE; 1,650 test units	Historical comparator table [5, 6]
Discrimination experiments	v3/v3b/v3c; sentence-reset; 506-document narrow confirmation	Rejected models / accepted narrow fix [6]
Droplet RC3 and RC4	General study; 303-document RC4 panel and TAB	Historical improvement and broad panel [3, 8]
RC5	311 synthetic, 32 TAB; 90-document correction confirmation	Retained engine operating controls [2]
RC6 research / retained RC5	797 primary; 200 RedactionBench; failed candidates	Latest comparative headline [1]
Installed local SDK	Commercial build and October 2 end-to-end verification	Integration evidence [4]

Internal source references (paths relative to the TracerYX workspace):

[1] Latest retained-engine evaluation
output/droplet-benchmark/industry-v6/delivery/traceryx_redaction_tokenization_rc6_study.md

[2] RC5 threshold and tokenization study
output/droplet-benchmark/industry-v5/traceryx_threshold_tokenization_study.md

[3] RC4 comparator and context-correction study
output/droplet-benchmark/industry-v4/traceryx_context_corrector_study.md

[4] Local SDK build / end-to-end reports
implementation/docs/local-sdk/BUILD_REPORT.md; END_TO_END_REPORT.md

[5] Frozen three-dataset benchmark
implementation/privacy-eval/three-dataset/REPORT.md

[6] Context and discrimination reports
implementation/privacy-eval/context-v2/REPORT.md; discrimination-boundary-fix/REPORT.md

[7] Baseline / early conformal reports
research/pii-comparison/REPORT.md; implementation/privacy-eval/{REPORT.md,final,conformal,gretel-conformal}/REPORT.md

[8] RC3 general-purpose study
output/droplet-benchmark/traceryx_general_pii_study.md

Presentation references: [OpenAI GPT-4o System Card](#); [Anthropic Model System Cards](#). Structure inspired by release documentation; no endorsement or external audit is implied.