

Surrogate Substitution Preserves PHI Detectability: A Multi-Detector Equivalence Study

Qiming Bao Sherry J. H. Feng Kim Chester Eugenio Meng Fon

Custodian Labs

Abstract

Structure-preserving de-identification replaces protected health information (PHI) with realistic same-type *surrogates*—“Anna S.” becomes “Maria S.”, not [NAME]—so that clinical text stays fluent and downstream tools keep working. But this only helps if the substitution does not itself corrupt the signal those tools rely on. We ask a narrow, testable question: *on the spans a de-identifier actually masks, can downstream PHI detectors still find the surrogate?* We introduce a paired, multi-detector evaluation protocol that (i) scores utility *only on masked spans*, decoupling *coverage* from *utility*; (ii) uses *equivalence testing* (TOST) rather than null-hypothesis significance testing, which is uninformative at our sample size (57k paired spans); and (iii) builds a surrogate-failure typology separating fixable generator defects from intrinsic detector limits. Across 11 detectors, 7 benchmarks, and 7 languages (1,750 documents), recall on masked spans moves from 76.1% to 74.9%—a change our equivalence test shows is *statistically equivalent to zero within a ± 2 -point margin* ($p \approx 3 \times 10^{-9}$), with detector ranking preserved. The residual loss does not reflect detectors getting worse at PHI: it concentrates in *malformed and out-of-distribution* surrogates (truncation Chicago $\rightarrow$ Illinois, salience loss Cedars-Sinai $\rightarrow$ Vidant). A redaction floor and an open-source surrogate baseline, run across detector families, indicate the effect is a property of well-formed substitution—not of one tool or one detector. We release the evaluation subsets, scoring code, and an interactive dashboard at <https://custodianlabs.pages.dev> so the protocol can audit any structure-preserving transform.¹

1 Introduction

De-identification of clinical text takes two forms—and only one keeps the text usable downstream. **Redaction** deletes or tags PHI (John Smith $\rightarrow$ [NAME]), which is safe but destroys the fluency, layout, and distributional properties that downstream models and human readers depend on. **Structure-preserving** de-identification instead substitutes each PHI value with a plausible, same-type *surrogate* (John Smith $\rightarrow$ Maria Lopez), keeping the document readable and machine-parseable [7]. The second family is increasingly attractive: it lets de-identified data flow into analytics, model training, and even second-pass detection without breaking pipelines built for real text.

The promise of structure preservation rests on an unstated assumption: *the surrogate carries the same*

downstream signal as the value it replaced. If substitution silently degrades the very features a PHI detector, an NER model, or a clinical parser relies on, then “structure-preserving” is a misnomer—the transform would be quietly laundering PHI into forms tools can no longer see or handle, which is both a utility problem and, for a second-pass safety net, a privacy problem.

This assumption is rarely tested directly, and testing it well is harder than it looks. Three pitfalls recur:

1. **Coverage confounds utility.** A de-identifier that masks 60% of PHI and one that masks 95% cannot be compared on whole-document F1—the score conflates *how much* it masks with *whether what it masks stays usable*. The two must be measured separately.
2. **The large- N significance trap.** With tens of thousands of paired spans, any non-zero difference is “statistically significant” under a standard test,

¹Code: <https://github.com/Custodian-Labs/guardian-layer-phi-benchmark>

even when operationally meaningless. A p -value here answers the wrong question.

3. **Aggregate error hides mechanism.** A single “−1.2 points” says nothing about *why* the loss happens—boundary jitter, detector weakness, or a defect in surrogate generation. Only the last is fixable, and only an error typology tells them apart.

We address all three. Our contributions:

- A **paired multi-detector protocol** scoring utility on masked spans only, decoupling coverage from utility, across 11 detectors $\times$ 7 benchmarks $\times$ 7 languages (§4–§5).
- An **equivalence-testing analysis** (TOST) replacing the uninformative significance test with a bounded-effect claim: the change in detectability is bounded within ± 2 points of zero (§6).
- A **surrogate-failure typology** attributing the small residual to surrogate-generation quality rather than detectors getting worse at PHI (§7).
- A set of **comparison experiments**—a redaction upper-bound and an open-source surrogate baseline—that make the result reproducible and isolate what is generic to *any* substitution versus specific to one tool (§5, §6).

The framing is deliberately not “a proprietary tool is good.” It is: *here is how to measure whether a structure-preserving transform preserves utility, rigorously*, with one commercial transform as the case study and open baselines for reproducibility.

2 Related Work

2.1 De-identification and surrogate substitution

Clinical de-identification is a long-studied sequence-labeling problem [38], anchored by the i2b2/n2c2 and MEDDOCAN shared tasks [29, 36] and by systems ranging from rule- and dictionary-based pipelines [31] through recurrent and transformer taggers [11, 20, 26] to instruction-tuned LLMs used zero-shot [27]. Most of this literature optimizes *detection*, typically against the HIPAA Safe Harbor identifier set [37]. The downstream question—*what to put in the PHI’s place*—is comparatively under-studied. Surrogate (“hiding in plain sight”) replacement was proposed to keep de-identified notes realistic and to resist re-identification [7]; subsequent work

showed poor surrogates can even aid re-identification (the “parrot” attack), underscoring that surrogate *quality* matters [8]. Recent studies generate surrogates or synthetic notes with rule-based and LLM methods and measure the downstream trade-off [21, 47], while a parallel line shows de-identified or pseudonymized text can still leak under extraction and membership-inference attacks [6, 24, 34]—so a surrogate must both *destroy* the original (privacy) and *remain a well-formed, same-type value* (utility). Because part of our panel uses LLMs as zero-shot detectors, we note a complementary line that strengthens LLM reasoning through structure-aware planning, search-augmented inference, and multi-path chain-of-thought aggregation [43–46]; such strategies could improve LLM-based detection but are orthogonal to the evaluation protocol we study here. Our work is orthogonal to the detection literature and to any particular surrogate generator: given that a span is masked, we ask whether the *replacement* remains detectable and well-formed.

The commercial transform used as our case study is the Custodian Labs Guardian Layer [9]; we cite it as the source of the transform rather than describing or extending the method (a formal method citation will be added once available).

2.2 Utility-preservation evaluation

Whether privacy transformations preserve downstream utility is central to privacy-preserving NLP. Vakili et al. [40] find automatic de-identification—including realistic surrogate replacement—barely affects downstream clinical model quality, and Vakili and Dalianis [39] caution that the measurement instrument matters, since a membership-inference attack fails to credit pseudonymization. Prior evaluations typically run a single downstream model on original vs. transformed data and compare end-task scores. Two weaknesses recur: (a) a single downstream model cannot separate “the transform is fine” from “this model is robust,” and (b) whole-corpus metrics mix masked and unmasked content. We address (a) with an 11-detector panel spanning rule-based, fine-tuned, and LLM detectors, and (b) by restricting utility measurement to masked spans.

2.3 The large- N trap and equivalence testing

When samples are large, null-hypothesis significance testing rejects the null for negligible effects; the p -value measures precision, not importance—a hazard increasingly flagged in NLP evaluation [5, 13]. The

standard remedy is *equivalence testing*—the two one-sided tests (TOST) procedure [4, 35]—which specifies an equivalence margin Δ and tests whether the effect lies inside $[-\Delta, +\Delta]$; see Lakens [22], Lakens et al. [23] for practical tutorials. TOST is standard in biostatistics but under-used in NLP evaluation, where large paired corpora make the trap acute. We adopt it as the primary inferential tool and report McNemar’s paired test [30] only to demonstrate the trap.

3 Problem Formulation

Structure-preserving de-identification. A transform T maps a document d to d' by replacing each detected PHI value v (of type τ) with a surrogate s of the same type, leaving all other characters unchanged. Gold PHI spans on d are re-projected onto d' by character-level alignment, giving paired spans (v, s) .

Coverage vs. utility. Two quantities must not be conflated. *Coverage* is the fraction of true PHI that T detects and replaces (a property of T ’s detector). *Utility* is, given that a span was replaced, whether a downstream detector still finds the surrogate s as well as it found v (a property of T ’s generator and of the surrogate’s realism). We report coverage separately and measure utility *only on the masked-span population* $\{(v, s)\}$. This prevents a low-coverage transform from looking good—or a high-coverage one from looking bad—on a metric that is really about substitution quality.

4 Evaluation Protocol

Figure 1 summarizes the protocol. **Paired multi-detector design.** Each document is scored in two conditions—original d and transformed d' —by the *identical* detector suite D ($|D| = 11$). Because the only change between conditions is the substitution, any per-span change in detection is attributable to the substitution, not to the detector or the document.

Metrics. We report span-level P/R/F1 under three matching modes—*exact* (start+end+type), *type* (type+boundary), *overlap* (any character overlap+type)—and *leakage* = $1 - \text{recall}$, the HIPAA-critical quantity. The headline utility metric is *recall retention on masked spans* = $\text{recall}(d')/\text{recall}(d)$ restricted to $\{(v, s)\}$, under overlap matching (so pure boundary jitter from length changes, “Anna S.”→“Maria S.”, is not charged as a miss).

Equivalence testing. For pooled masked-span recall we run TOST with margin $\Delta = 2$ points: we reject non-

equivalence iff the 90% CI of the recall difference lies entirely within $[-2, +2]$. We also report McNemar’s test to illustrate the large- N trap. We sweep $\Delta \in \{1, 2, 3\}$.

Error attribution. For the lost population (found on d , missed on d') we hand-code each span into a small failure typology (§7) and check length-preservation to separate boundary artifacts from genuine misses.

5 Experimental Design

Detector panel D . Eleven detectors, three families: *rule/statistical*—Microsoft Presidio [31] (built on spaCy [19]), OBI `deid_roberta` [20], a RoBERTa tagger [12, 25]; *open LLMs (local)*—Gemma 4 31B / E4B [16], Qwen 3.5-{4B, 9B, 35B-A3B} [33], Llama 3.1-8B / 3.3-70B [18], DeepSeek V2-Lite [10]; *frontier API*—OpenAI GPT-5 [32]. All are Transformer-based [41]. The panel is deliberately heterogeneous: if the equivalence result held only for one architecture it would be a model artifact, not a property of the transform. (Model-version citations point to the closest published technical report for each family. One further model, Moonshot Kimi-VL-A3B, was attempted but excluded because it required non-standard model-loading code.) LLM detectors emit free-text or JSON identifier lists; we recover character spans by exact-then-fuzzy string search over the source document, so all detectors are scored on the same span basis.

Benchmarks (7; 250 docs each, 1,750 total). ASQ-PHI (English clinical queries) [42], MEDDOCAN (Spanish clinical) [29], MultiCoNER v2 (multilingual NER) [15, 28], and PII-Masking-300k [1] in English, Dutch, French, German. Together: 7 languages, clinical and general-PII text, free-form and structured (JSON) formats. The transform under test is Custodian Guardian Layer `transform mode (top-1 surrogate, pii_entities=ALL)`.

Comparison conditions. The core study answers “does *this* transform preserve detectability.” Three contrasts isolate *why* and *how generally*:

- **C1—Redaction upper-bound.** Replace each masked value with `*****` (no surrogate signal) and re-detect; the transform—redact gap quantifies what structure preservation buys.
- **C2—Open-source surrogate baseline.** Replace the commercial generator with an open one (Faker [14]) on the *same* masked spans; tests

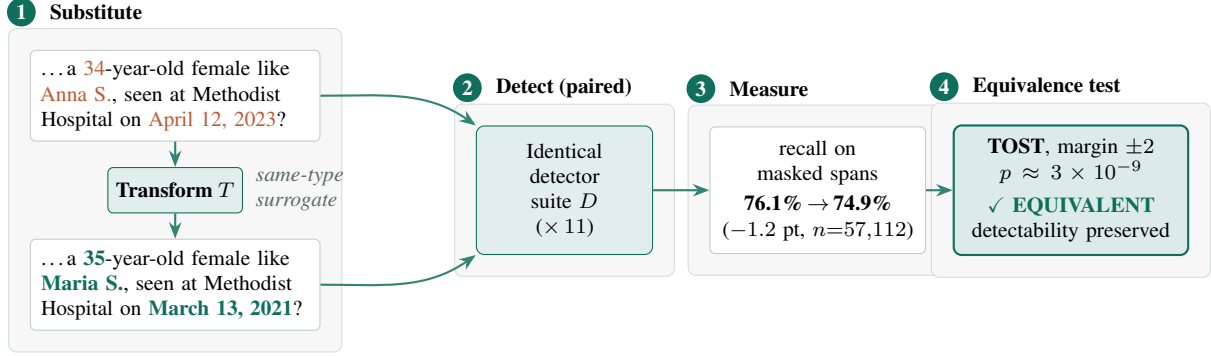


Figure 1: **The paired evaluation protocol, with a worked example.** (1) The transform replaces each PHI value with a same-type surrogate (original → surrogate), leaving everything else byte-identical. (2) The *identical* 11-detector suite scores *both* the original d and the transformed d' , so any per-span change is attributable to the substitution alone. (3) Utility is measured *only on the spans the transform masks* (overlap match), decoupling it from coverage. (4) The paired recall difference is assessed with an equivalence test (TOST, margin ± 2 pts) rather than a significance test—which the large sample would make trivially “significant” (§4). Result: the change is statistically equivalent to zero.

whether the result is generic to substitution or tool-specific, and makes the pipeline reproducible.

- **C3—Per-benchmark equivalence.** Run TOST within each benchmark, sweeping Δ , to check the pooled claim is not an averaging artifact.

The core paired study, the equivalence analysis (pooled and per-benchmark), the redaction floor, the open-surrogate baseline, and the error typology are complete; C1 and C2 span four detectors across all three families (Presidio, OBI, Qwen-9B, Gemma-31B), so the comparison-experiment conclusions do not rest on a single detector.

6 Results

Detectability is statistically equivalent. Restricting to the 57,112 masked spans and pooling all 11 detectors (via the released `analyze_equivalence.py`), recall moves 76.1% → 74.9% (−1.2 pts; 95% CI [−1.5, −1.0]). The TOST equivalence test ($\Delta = 2$) rejects non-equivalence at $p \approx 3 \times 10^{-9}$: the change is statistically bounded within ± 2 points of zero. Ranking is preserved; Llama 3.3-70B is essentially unchanged ($\Delta F1 +0.003$). The same contingency is “significant” under McNemar’s test—a direct demonstration of the large- N trap.

Whole-document view (conservative). Before restricting to masked spans, Table 1 reports span-level F1 and leakage on *all* gold spans, per detector. Whole-

document $\Delta F1$ ranges from +0.003 (Llama 3.3-70B) to −0.047 (Qwen 3.5-35B-A3B), and leakage rises by only +0.3 to +4.1 points. This view is conservative by design—it mixes masked spans with exact-boundary penalties on length-changed surrogates and with spans the transform never touched—yet it already bounds the worst case (no detector’s mean F1 moves by more than 4.7 points) and preserves ranking across a $100\times$ span of detector quality (F1 0.76 → 0.04). The masked-span analysis below isolates the substitution effect from this dilution.

Recall retention on masked spans (overlap). Restricting to masked spans removes that dilution. When the transform masks a span, detectors still find the surrogate 93–100% of the time (Table 2); the ~ 3 -point exact-boundary drop is a length-jitter artifact that vanishes under overlap matching. (The two floor detectors, Llama 3.1-8B and OBI `deid_roberta`, are omitted from Table 2: their original recall is so low that the retention ratio is dominated by noise.)

Per-benchmark equivalence (C3). Table 3 (57,112 masked spans, via `scripts/analyze_equivalence.py`) shows the pooled equivalence is *not uniform*: four benchmarks are equivalent within ± 2 points; MEDDOCAN and PII-nl require ± 3 (both dense, identifier-heavy, non-English—where surrogate generation is hardest, §7); MultiCoNER has too few masked spans (220) to test. The honest claim: *detectability is equivalent within*

Detector	F1		Δ	Leak	
	orig	transf		orig	transf
Gemma 4 31B	.755	.710	−.045	.255	.285
Gemma 4 E4B	.737	.698	−.038	.276	.311
Llama 3.3-70B	.725	.728	+.003	.245	.262
Qwen 3.5-35B-A3B	.715	.668	−.047	.254	.295
OpenAI GPT-5	.705	.674	−.032	.308	.333
Qwen 3.5-9B	.655	.621	−.034	.391	.422
Qwen 3.5-4B	.567	.533	−.034	.483	.515
Presidio	.416	.398	−.018	.553	.570
DeepSeek V2-Lite	.409	.384	−.025	.672	.696
Llama 3.1-8B	.391	.376	−.015	.633	.650
OBI deid_roberta	.041	.040	−.002	.941	.944

Table 1: Whole-document view (all 11 detectors), mean over 7 benchmarks. Span-level F1 (type match) and leakage (1−recall). Conservative: it includes untouched spans and boundary penalties. Ranking is preserved and Δ F1 is bounded.

Detector	Exact	Overlap
Gemma 4 31B	93.1	99.8
Qwen 3.5-9B	92.0	100.0
Qwen 3.5-35B-A3B	90.0	98.4
Llama 3.3-70B	91.4	98.3
GPT-5	91.9	98.3
Gemma 4 E4B	88.9	98.2
Qwen 3.5-4B	84.7	98.0
Presidio	91.3	97.4
DeepSeek V2-Lite	87.7	92.8

Table 2: Recall retention on masked spans (%), transformed÷original.

± 2 points on average and on clean text, and within ± 3 on the hardest identifier-dense text—the residual is concentrated and attributable, not diffuse degradation (Figure 2).

Redaction floor (C1). We replace each masked value with *****—destroying the surrogate signal—and re-detect with four detectors spanning all three families (Table 4). The floor is sharply *family-dependent*. Rule-based Presidio and the fine-tuned OBI tagger collapse to ≈ 0 –4% everywhere: they key on real surface tokens, so with the signal gone they recover almost nothing (the residual is spurious overlap on the * run). Open LLMs behave differently—they partly *recover* the masked span (Qwen-9B 6–58%, Gemma-31B 13–80%)—because a ***** run is itself a legible redac-

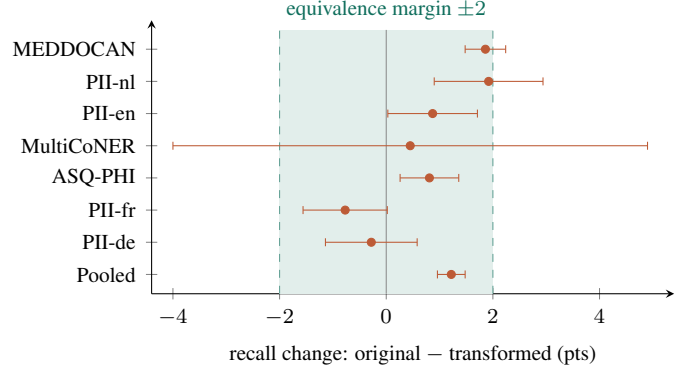


Figure 2: Per-benchmark equivalence (C3). Masked-span recall change (original − transformed) with 95% CIs; the shaded band is the ± 2 -pt equivalence margin. The pooled estimate and five of seven benchmarks sit fully inside; only MEDDOCAN and PII-nl reach past +2 (hence a ± 3 margin). MultiCoNER’s wide interval reflects its 220 masked spans.

Benchmark	Masked	O→T	McN. χ^2	Margin
ASQ-PHI	5,427	91.9→91.1	8.0	± 2 ✓
MEDDOCAN	31,978	73.4→71.5	92.4	± 3 ✓
MultiCoNER	220	71.8→71.4	0.0	underpow.
PII en	5,181	76.0→75.1	4.0	± 2 ✓
PII nl	3,757	78.2→76.3	13.1	± 3 ✓
PII fr	5,984	75.2→75.9	3.5	± 2 ✓
PII de	4,565	76.3→76.5	0.4	± 2 ✓
Pooled	57,112	76.1→74.9	82.1	± 2 ✓

Table 3: Per-benchmark equivalence. O→T = masked-span recall, original→transform (%). Pooled TOST $\Delta=2$: $p=3 \times 10^{-9}$.

tion cue that a strong reader flags as “something was here.” This recovery is *context-dependent*: highest on ASQ-PHI (short adversarial queries where one masked run dominates the prompt) and on the larger model, lowest on the dense multilingual PII sets. Crucially, even Gemma’s floor stays far below its surrogate recall (Table 5), so the conclusion holds across families—redaction destroys detectability that structure-preserving substitution keeps, and where a detector recovers anything from ***** it is reading the mask, not the (absent) PHI.

Open-surrogate baseline (C2). We replace the commercial generator with Faker on the same masked spans and re-score all four detectors (Table 5); two find-

Benchmark	N	rule / fine-tuned		open LLM	
		Presidio	OBI	Qwen-9B	Gemma-31B
ASQ-PHI	267	0.0 $\pm$ 0.0	0.7 $\pm$ 0.9	58.1 $\pm$ 6.0	79.8 $\pm$ 4.9
MEDDOCAN	1,419	0.0 $\pm$ 0.0	0.1 $\pm$ 0.1	22.9 $\pm$ 2.3	52.6 $\pm$ 2.7
MultiCoNER	8	—	0.0 $\pm$ 0.0	25.0 $\pm$ 31.2	12.5 $\pm$ 18.8
PII en	210	1.9 $\pm$ 1.2	1.0 $\pm$ 1.2	10.0 $\pm$ 4.0	37.6 $\pm$ 6.7
PII nl	195	3.8 $\pm$ 1.9	1.0 $\pm$ 1.3	20.5 $\pm$ 5.6	39.5 $\pm$ 6.9
PII fr	220	2.2 $\pm$ 1.2	0.5 $\pm$ 0.7	5.9 $\pm$ 3.0	38.2 $\pm$ 6.6
PII de	204	0.0 $\pm$ 0.0	1.0 $\pm$ 1.2	11.8 $\pm$ 4.4	41.7 $\pm$ 6.9
Mean		1.3	0.6	22.0	43.1

Table 4: C1 redaction floor: masked-span recall (%) after replacing each masked value with * * * * *, across detector families. Rule/fine-tuned detectors collapse to ≈ 0 ; open LLMs partly recover the redaction cue (context-dependent). All floors sit well below the surrogate recall in Table 5. N (khaki) is the masked-span count in the 120-document subsample shared by OBI, Qwen-9B and Gemma-31B; Presidio runs on the full corpus (N=531, 2,925, 471, 373, 544, 415 for its six rows). MultiCoNER contributes only 8 masked spans in the subsample, so those cells are illustrative. $\pm x$ (khaki) is the half-width of the 95% bootstrap CI over masked spans (5,000 resamples), making small-*N* noise explicit—e.g. MultiCoNER’s ± 19 –31. Per-detector means are over each detector’s scored benchmarks (Presidio: 6; others: 7).

ings hold *across families*. *Generality*: every detector retains high masked-span recall on the open surrogates—Presidio 78–98%, OBI 64–100%, Qwen-9B 77–100%, Gemma-31B 83–100% (family means 84–95%)—so detectability preservation is a property of well-formed, same-type substitution, not of one vendor or one detector, and it reproduces without proprietary access. *The residual is generator quality*: Faker, emitting clean canonical values, matches or exceeds the commercial transform precisely on the hard non-English benchmarks where the commercial surrogates truncate or garble (§7); the stronger the detector, the smaller the residual (Gemma-31B mean 94.5%).

Leakage barely moves (+0.3 to +4.1 pts across detectors; Table 1), so surrogates are not systematically easier to miss than the PHI they replace.

Where the loss lands (per benchmark). Table 6 breaks whole-document $\Delta F1$ out by benchmark and detector. The pattern is sharp: almost all of the loss concentrates on **ASQ-PHI**—short, sparse-PHI adversarial queries where a single substitution dominates the document score (−0.09 to −0.18)—while the other six benchmarks are essentially flat ($|\Delta F1| \leq 0.06$, most ≤ 0.03). Even the fine-tuned OBI tagger and the rule-based Presidio move by at most a few thousandths on most benchmarks. This is the first sign that the effect tracks *surrogate-generation difficulty* (short adversarial text, dense non-English identifiers) rather than detector family—made precise in the error analysis (§7).

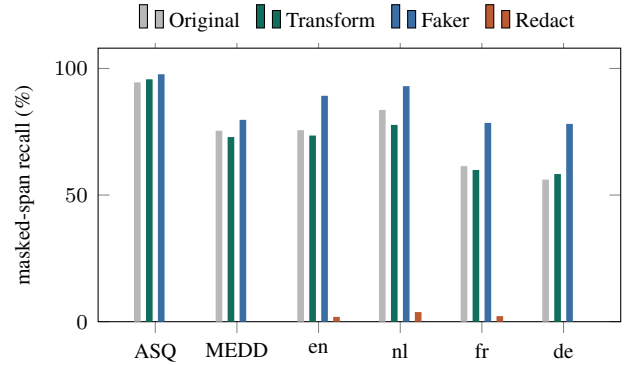


Figure 3: What structure preservation buys, shown for Presidio (which has full original/transform/Faker/redact columns). Transform tracks the original, the open Faker baseline is at least as high, and redaction collapses to ≈ 0 . The gap between the Transform/Faker bars and the Redact bar is the detectability structure preservation retains; the family-dependent floor and open-surrogate retention across four detectors are in Tables 4–5.

What the transform does (examples). Table 7 shows the substitution qualitatively across languages and formats: each PHI value becomes a same-type surrogate while clinical shorthand, foreign-language syntax, and JSON structure are left byte-for-byte intact. On the well-formed English cases every surrogate is still detected across most of the panel.

Coverage, reported separately. All results above

Benchmark	N	rule / fine-tuned		open LLM	
		Presidio	OBI	Qwen-9B	Gemma-31B
ASQ-PHI	267	97.7 $\pm$ 1.2	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0
MEDDOCAN	1,419	79.7 $\pm$ 1.5	64.0 $\pm$ 2.6	79.6 $\pm$ 2.1	96.5 $\pm$ 1.0
MultiCoNER	8	—	100.0 $\pm$ 0.0	87.5 $\pm$ 18.8	100.0 $\pm$ 0.0
PII en	210	89.2 $\pm$ 2.9	89.5 $\pm$ 4.0	76.7 $\pm$ 5.7	90.0 $\pm$ 4.0
PII nl	195	93.0 $\pm$ 2.5	71.8 $\pm$ 6.4	84.1 $\pm$ 5.1	96.9 $\pm$ 2.3
PII fr	220	78.5 $\pm$ 3.6	79.5 $\pm$ 5.2	81.4 $\pm$ 5.2	83.2 $\pm$ 5.0
PII de	204	78.1 $\pm$ 4.1	84.8 $\pm$ 4.9	77.9 $\pm$ 5.6	95.1 $\pm$ 2.9
Mean		86.0	84.2	83.9	94.5

Table 5: C2 open-surrogate (Faker) retention: masked-span recall (%) on the same masked spans, across detector families. Every family stays high (means 84–95%), so detectability preservation is generic to well-formed substitution rather than tool- or detector-specific. N, subsampling, and mean denominators as in Table 4.

Detector	ASQ-PHI	MEDDOCAN	PII en	PII nl	PII fr	PII de	MultiCoNER
Gemma 4 31B	−.175	−.017	−.043	−.038	−.014	−.016	−.009
Gemma 4 E4B	−.106	−.060	−.032	−.021	−.018	−.025	−.007
Qwen 3.5-35B-A3B	−.132	−.034	−.026	−.034	−.040	−.054	−.007
OpenAI GPT-5	−.098	−.046	−.049	−.019	+ .008	−.005	−.014
Qwen 3.5-9B	−.123	−.017	−.025	−.010	−.034	−.024	−.008
Qwen 3.5-4B	−.086	−.055	−.034	−.028	−.015	−.023	+ .002
Presidio	+ .003	−.037	−.032	−.015	−.020	−.015	−.007
Llama 3.1-8B	−.108	+ .023	−.014	+ .007	−.009	+ .003	−.006
OBI deid_roberta	+ .001	−.008	−.004	−.001	+ .003	−.002	−.001

Table 6: Per-benchmark whole-document Δ F1 (transformed − original; negative = drop). Almost all loss lands on ASQ-PHI; the other six benchmarks are flat. Llama 3.3-70B and DeepSeek V2-Lite omitted for space (both flat; see Table 1 for their pooled Δ F1).

are measured *on the spans the transform masks*. Coverage—how much true PHI it detects and replaces—is a distinct, detector-side axis. It tracks how closely a benchmark’s annotation matches the transform’s notion of sensitive content: it masks **80.9%** of gold PHI on ASQ-PHI and **48.3%** on MEDDOCAN, and less ($\approx$ 26%) on general-domain NER/PII corpora whose annotated entities (encyclopedic names, generic places) fall outside that scope; a configuration sweep confirmed `domain=General` maximizes coverage. A per-entity-type breakdown and a detection-vs-replacement diagnosis (the coverage gap is mostly a *replacement*-step issue— $\approx$ 77% of missed clinical identifiers were flagged by the transform’s own detector but not substituted) are in Appendix A. Decoupling matters because the two failure modes have different owners and fixes: an under-masked span is a *detection/replacement* miss, whereas a masked-but-missed surrogate is a *generation* defect (§7). Conflating them on whole-document F1 (Table 1) would let a high-coverage/low-quality transform and a

low-coverage/high-quality one look identical. We therefore report coverage as context and reserve the utility claim for masked spans.

Matching modes. Unless noted, retention uses *overlap* matching (detector flags any part of the surrogate span). Exact matching (Table 2, left) additionally requires identical character boundaries and is therefore sensitive to surrogate length changes (“Anna S.” $\rightarrow$ “Maria S.” shifts the end offset); the $\sim$ 3-point exact–overlap gap is this boundary jitter, not missed PHI. Type matching (boundary + type) sits between the two and tracks overlap closely.

7 Error Analysis

The lost population. Pooled across detectors and masked spans: 40,165 (span $\times$ detector) pairs found in both conditions; 3,300 lost. Half (50%) of lost spans are the *same length* as the original—so this is not a boundary effect. By type: LOCATION 27%, NAME 23%, DATE/AGE 22%, ID/contact 19%.

Case	Original	Transformed
ASQ-PHI (clinical query)	...a 34-year-old female ...like Anna S., ...at Methodist Hospital on April 12, 2023?	...a 35-year-old female ...like Maria S., ...at Methodist Hospital on March 13, 2021?
ASQ-PHI (short-hand)	Rec mgmt of 70yo M w/ CHF, seen by Dr. John L. at Mt. Sinai on Feb 21, 2023. alt tx options...	Rec mgmt of 73yo M w/ CHF, seen by Dr. James L. at Mt. Egypt on Nov 19, 2021. alt tx options...
PII (German)	...ermächtigen hiermit Monsignore, ...Mit Datum 23/07/2011...	...ermächtigen hiermit Fulgenzio, ...Mit Datum 24/07/2011...
PII (French, JSON)	{"Date": "20/05/2022", "City": "Saint-Priest", "Username": "phprosdocimo"}	{"Date": "21/05/2023", "City": "Saint-Priest", "Username": "phprosdocimo"}

Table 7: Worked transform examples. Only the sensitive value moves; abbreviations (w/ CHF, alt tx), non-English syntax, and JSON keys/quotes/indentation are preserved, so downstream parsers and detectors keep working. In the JSON case the transform changes only the date value and leaves the (also-sensitive-looking) username untouched—a coverage decision, not a utility one (see below).

Three failure modes—all generator-side.

(1) *Malformed/truncated surrogates* (largest cause): Chicago → **Illino**, El Paso → **El**, Ciudad de la Habana → **Cuidad de la Havana**. The fragment no longer matches the lexical pattern detectors learned for real names/places; this is why span-level loss is highest on MEDDOCAN (6.9%; Spanish, dense, identifier-heavy) and lowest on clean English ASQ-PHI (2.5%). (2) *Loss of salience*: a canonical entity replaced by an obscure one (Cedars-Sinai → **Vidant**); detectors partly rely on pre-training familiarity, so swapping a famous value for a rare one removes the prior. Inherent to any value substitution; mainly costs weaker detectors. (3) *x-masking of IDs/emails*: nachorutor@... → **nxxxxxxxxx@...**. The x run preserves format but breaks the realistic-token pattern. This case is *privacy-positive*—the original value is destroyed—even though it counts against recall.

Quantifying the typology. We classify a stratified 200-span sample (proportional across the seven benchmarks) against this typology by *AI review*—an LLM labels each surrogate, with scripted x-mask detection.² **77.0%** of surrogates are well-formed, same-type values; of the 23% with a defect, *truncation/garbling* dominates (12.5%: e.g. 31/08 → **32/08**, United Kingdom → **States Kindom**, dropped ID digits),

²An AI review, *not* human-validated; the released sheet records the per-span labels, and human double-annotation is left to future work.

followed by *x-masking* (9.0%), with *salience loss* rare (1.5%: Chicago Bulls → **Illinois Bulls**); type-consistency is near-total (99.5%). The defect rate is lowest on clean English/Spanish clinical text and highest on ASQ-PHI’s short adversarial queries and MultiCoNER’s dense entities—the same benchmarks that carry the recall gap. Strikingly, the **77% well-formed rate tracks the 74.9% masked-span recall**: detectors recover almost exactly the well-formed surrogates and miss the defective ones, so the residual is a surrogate-generation problem, not a detection one.

Implication. Detectors are not getting worse at PHI; the small recall gap is driven by surrogate-generation quality (truncation, garbling, salience, x-masking)—generating well-formed, in-distribution synthetic text is itself a hard problem in text data augmentation [3]. C2 confirms this: an open generator emitting clean canonical values matches or exceeds the commercial transform on exactly the hard non-English benchmarks where its surrogates truncate—so the residual tracks generator quality, not the act of substitution.

8 Discussion

What the evidence supports. Structure-preserving substitution does not hide well-formed PHI from downstream detection. Because the effect is equivalence-bounded within ± 2 points and ranking-preserving across a heterogeneous 11-detector panel, a transform of this quality can be inserted ahead of detection/analytics

pipelines built for real clinical text without materially degrading downstream detection. **Generality (C2).** Faker preserves masked-span recall at least as well as the original across all six benchmarks, so ± 2 points appears to be a property of well-formed substitution rather than of one vendor; where the commercial generator trails, a cleaner generator closes the gap, locating the residual squarely in generation quality. **A reusable protocol.** The paired masked-span design and the TOST margin are not specific to one vendor or language—a template for auditing any structure-preserving privacy transform. **Deployment guidance.** A structure-preserving transform can be treated as admissible ahead of a detection/analytics pipeline when it clears a three-part check—detectability measured on masked spans only, the change inside an equivalence margin with ranking preserved, and surrogate outputs free of the three generator-side defects (truncation/garbling, salience loss, $\times$ -masking) that drive the residual.

9 Ethics and Data Statement

All benchmarks are public or synthetic (ASQ-PHI synthetic; MEDDOCAN released for a shared task; PII-Masking-300k synthetic; MultiCoNER v2 public). **No real patient data is used.** We release the 250-document subsets and scoring code with license notes. **Dual-use.** A utility-preserving de-identifier could in principle launder identifiable data into a fluent form; our masked-span/leakage reporting and the privacy-positive framing of $\times$ -masking keep the privacy accounting explicit. **Conflict of interest.** The commercial transform (Custodian Guardian Layer) is developed by Custodian Labs, with which the authors are affiliated; this work was conducted with Custodian Labs’ support. To limit bias we (i) frame the contribution as a reusable protocol, (ii) include open-source baselines so results are reproducible without proprietary access, and (iii) report leakage and coverage alongside utility.

10 Limitations

Coverage is reported but not the focus; a transform can preserve utility on what it masks while under-masking (the axes are independent by design). 250 docs/benchmark bounds per-benchmark power (though the pooled masked-span N is large). C1/C2 span four detectors across families (Presidio full-corpus; OBI, Qwen-9B, Gemma-31B on a 120-doc/benchmark sub-sample); extending them to the full 11-detector panel is

straightforward future work. LLM detectors are prompt-sensitive and their behaviour shifts with task framing and structural variation [2, 17]; we fix one prompt per model and release it with the code and data at the project’s reproducibility page. Finally, all seven benchmarks are public or synthetic; validating the protocol on real-EHR corpora (n2c2, MIMIC-IV-Note, CARMEN-I) is future work pending the relevant data-use agreements.

A Coverage Details

Coverage (§6) is the fraction of a benchmark’s gold PHI whose characters the transform actually changed (diffib alignment). Table 8 breaks it down by entity type on the clinical sets (ASQ-PHI + MEDDOCAN). High-frequency free-text types (names, dates) are masked well; the weak spots are high-sensitivity structured identifiers and geography—exactly the HIPAA Safe Harbor items that most need masking.

Entity type (clinical)	Coverage
NAME / person	75.0%
DATE / age	73.4%
ID / contact (MRN, patient ID, phone, email)	50.7%
ORG / facility	46.6%
LOCATION / address	43.7%

Table 8: Coverage by entity type on clinical benchmarks. Overall the transform altered 39% of annotated PII (54% on clinical sets); general-domain corpora sit near 26%.

Detection vs. replacement. On a sample of 8 MEDDOCAN documents (62 missed ID/location spans), 48/62 ($\approx 77\%$) of the *unmasked* identifiers were nonetheless flagged as sensitive by the transform’s own detector—they were detected but not substituted; only 14/62 (23%) were undetected. The coverage gap is therefore mostly a *replacement-step* issue (act on everything the detector surfaces), which is more tractable than raising detection recall. (Small sample; “detected” judged by loose token overlap, so 77% is directional.) Concrete unmasked HIPAA identifiers included patient IDs (80926, 7845693), care-contact IDs (4387684), and facilities/geography (Hospital de Cruces, España, postal codes 41005, 28047).

B Representative Per-Detector Examples

All-detector overlay on one document (“ranking proof”). All eleven detectors’ predictions on one transformed document (asq_00001), overlaid on the text

and sorted by this document’s F1: **green** = correctly caught PHI, **yellow** = false positive, **red** = missed PHI (leakage). The four gold surrogate values are John T., the x-masked location Saint. Vxxxxxxx, April 25th, 2018, and the ID 987654321. Eight of eleven detectors tag all four cleanly (F1 100); the two floor detectors, Presidio and DeepSeek, miss the x-masked location surrogate (§7 mode 3)—DeepSeek also drops the numeric ID—and OBI over-fragments boundaries and false-fires on the age 60. The surrogate values shown are the transform’s output, not the original PHI.

Gold (reference)	4 PHI spans
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
OpenAI GPT-5	tp 4 · fp 0 · missed 0 · F1 100%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
Llama 3.1-8B	tp 4 · fp 0 · missed 0 · F1 100%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
Qwen 3.5-4B	tp 4 · fp 0 · missed 0 · F1 100%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
Gemma 4 31B	tp 4 · fp 0 · missed 0 · F1 100%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
Llama 3.3-70B	tp 4 · fp 0 · missed 0 · F1 100%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
Qwen 3.5-35B-A3B	tp 4 · fp 0 · missed 0 · F1 100%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
Gemma 4 E4B	tp 4 · fp 0 · missed 0 · F1 100%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
Qwen 3.5-9B	tp 4 · fp 0 · missed 0 · F1 100%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	

OBI deid_roberta	tp 4 · fp 1 · missed 0 · F1 89%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
Microsoft Presidio	tp 3 · fp 1 · missed 1 · F1 75%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	
DeepSeek V2-Lite	tp 2 · fp 0 · missed 2 · F1 67%
Evaluation of long-term outcomes for bypass surgery in patients over 60, referencing Mr. John T., operated at Saint. Vxxxxxxx on April 25th, 2018 (ID: 987654321)?	

A harder case: a dense Spanish clinical header.

Figure-style overlay on a 260-character window of a MEDDOCAN document (S0210-...008-1), where the same eleven detectors diverge far more—F1 from 96 down to 0. On dense, identifier-heavy non-English text the *per-document* ranking reshuffles (here DeepSeek and Qwen-4B lead, while Llama 3.1-8B finds nothing in this window), even though the *aggregate* ranking over 250 documents is stable (Table 1). The garbled surrogate Ciudad Real (from Ciudad Real) is still caught by most detectors; the misses cluster on the professional-licence ID 03 14 16485, the sex fields, and the repeated age—the identifier types with the weakest coverage and the hardest surrogates (Appendix A).

Gold (reference)	12 PHI spans
Ciudad Real. CP: 13002. Datos asistenciales. Referencia de nacimiento: 24/10/1963. País: España. Mejores: 53 años Sexo: H. Referencia de Ingreso: 14/12/2017. Episodio: 746589123. Médico: Luis Ruiz Camuñas N°Col: 03 14 16485. Historia Actual: Varón de 53 años q	
DeepSeek V2-Lite	tp 11 · fp 0 · missed 1 · F1 96%
Ciudad Real. CP: 13002. Datos asistenciales. Referencia de nacimiento: 24/10/1963. País: España. Mejores: 53 años Sexo: H. Referencia de Ingreso: 14/12/2017. Episodio: 746589123. Médico: Luis Ruiz Camuñas N°Col: 03 14 16485. Historia Actual: Varón de 53 años q	
Qwen 3.5-4B	tp 11 · fp 0 · missed 1 · F1 96%
Ciudad Real. CP: 13002. Datos asistenciales. Referencia de nacimiento: 24/10/1963. País: España. Mejores: 53 años Sexo: H. Referencia de Ingreso: 14/12/2017. Episodio: 746589123. Médico: Luis Ruiz Camuñas N°Col: 03 14 16485. Historia Actual: Varón de 53 años q	

OBI deid_roberta tp 10 · fp 0 · missed 2 · F1 91%
 Ciudad Real. CP: 13002. Datos asistenciales.
 Referencia de nacimiento: 24/10/1963.
 País: España. Mejores: 53 años Sexo: H.
 Referencia de Ingreso: 14/12/2017. Episodio:
 746589123. Médico: Luis Ruiz Camuñas N°Col:
 03 14 16485. Historia Actual: Varón de 53
 años q

Qwen 3.5-9B tp 10 · fp 0 · missed 2 · F1 91%
 Ciudad Real. CP: 13002. Datos asistenciales.
 Referencia de nacimiento: 24/10/1963.
 País: España. Mejores: 53 años Sexo: H.
 Referencia de Ingreso: 14/12/2017. Episodio:
 746589123. Médico: Luis Ruiz Camuñas N°Col:
 03 14 16485. Historia Actual: Varón de 53
 años q

Gemma 4 31B tp 9 · fp 0 · missed 3 · F1 86%
 Ciudad Real. CP: 13002. Datos asistenciales.
 Referencia de nacimiento: 24/10/1963.
 País: España. Mejores: 53 años Sexo: H.
 Referencia de Ingreso: 14/12/2017. Episodio:
 746589123. Médico: Luis Ruiz Camuñas N°Col:
 03 14 16485. Historia Actual: Varón de
 53 años q

Qwen 3.5-35B-A3B tp 9 · fp 1 · missed 3 · F1 82%
 Ciudad Real. CP: 13002. Datos asistenciales.
 Referencia de nacimiento: 24/10/1963.
 País: España. Mejores: 53 años Sexo: H.
 Referencia de Ingreso: 14/12/2017. Episodio:
 746589123. Médico: Luis Ruiz Camuñas N°Col:
 03 14 16485. Historia Actual: Varón de
 53 años q

OpenAI GPT-5 tp 8 · fp 0 · missed 4 · F1 80%
 Ciudad Real. CP: 13002. Datos asistenciales.
 Referencia de nacimiento: 24/10/1963.
 País: España. Mejores: 53 años Sexo: H.
 Referencia de Ingreso: 14/12/2017. Episodio:
 746589123. Médico: Luis Ruiz Camuñas N°Col:
 03 14 16485. Historia Actual: Varón de
 53 años q

Llama 3.3-70B tp 8 · fp 0 · missed 4 · F1 80%
 Ciudad Real. CP: 13002. Datos asistenciales.
 Referencia de nacimiento: 24/10/1963.
 País: España. Mejores: 53 años Sexo: H.
 Referencia de Ingreso: 14/12/2017. Episodio:
 746589123. Médico: Luis Ruiz Camuñas N°Col:
 03 14 16485. Historia Actual: Varón de
 53 años q

Gemma 4 E4B tp 8 · fp 0 · missed 4 · F1 80%
 Ciudad Real. CP: 13002. Datos asistenciales.
 Referencia de nacimiento: 24/10/1963.
 País: España. Mejores: 53 años Sexo: H.
 Referencia de Ingreso: 14/12/2017. Episodio:
 746589123. Médico: Luis Ruiz Camuñas N°Col:
 03 14 16485. Historia Actual: Varón de 53
 años q

Microsoft Presidio tp 6 · fp 1 · missed 6 · F1 63%
 Ciudad Real. CP: 13002. Datos asistenciales.
 Referencia de nacimiento: 24/10/1963.
 País: España. Mejores: 53 años Sexo: H.
 Referencia de Ingreso: 14/12/2017. Episodio:
 746589123. Médico: Luis Ruiz Camuñas N°Col:
 03 14 16485. Historia Actual: Varón de
 53 años q

Llama 3.1-8B tp 0 · fp 0 · missed 12 · F1 0%
 Ciudad Real. CP: 13002. Datos asistenciales.
 Referencia de nacimiento: 24/10/1963.
 País: España. Mejores: 53 años Sexo: H.
 Referencia de Ingreso: 14/12/2017. Episodio:
 746589123. Médico: Luis Ruiz Camuñas N°Col:
 03 14 16485. Historia Actual: Varón de
 53 años q

Agreement on well-formed surrogates. On clean English (ASQ-PHI), a plausible same-type surrogate is caught by nearly the whole panel (Table 9): the substitution is invisible to detection. Divergence is confined to the two floor detectors (Llama 3.1-8B, DeepSeek V2-Lite).

Document	Surrogate span (type)	Caught
asq_00000	Maria S. (NAME)	10/11
asq_00000	Methodist Hospital (LOC)	9/11
asq_00000	March 13, 2021 (DATE)	9/11
asq_00003	James L. (NAME)	11/11
asq_00003	Mt. Egypt (LOC)	11/11
asq_00003	Nov 19, 2021 (DATE)	11/11

Table 9: Number of detectors (of 11) that find each well-formed surrogate. Same-type swaps stay broadly detectable.

Representative losses, by failure type. Table 10 shows genuine “lost” cases (MEDDOCAN): the original value was found by most detectors, but a defective surrogate is found by few. Each maps to one of the three failure types in §7, and the drop is shared across the panel—i.e. it is a property of the surrogate, not of any one detector.

References

- [1] ai4Privacy. PII-masking-300k: a dataset for training privacy-preserving models. <https://huggingface.co/datasets/ai4privacy/pii-masking-300k>, 2023.
- [2] Qiming Bao, Gaël Gendron, Alex Yuxuan Peng, Michael Witbrock, and Jiamou Liu. Assessing and enhancing the robustness of large language models with task structure variations for logical reasoning. In *Neural Information Processing (ICONIP 2024)*, Lecture Notes in Computer Science. Springer, 2024.
- [3] Qiming Bao, Alex Yuxuan Peng, Zhenyun Deng, Wanjun Zhong, Gaël Gendron, Timothy Pistotti, Neşet Tan, Nathan Young, Yang Chen, Yonghua

Original → surrogate	O → T	Type
Cuba → Havana	9 → 2	salience
San Fernando →	8 → 3	garbled
Francsico Luis		
Cádiz → Cadiz	8 → 4	accent stripped
Cecilio	8 → 4	truncation
Pujazón →		
Pujazón		
ignaciotorne@... →	9 → 2	x-mask
ixxxxxxxxxxxx@...		
23/07/1948 →	11 → 4	x-mask
2xxxxxxxx		
40 → 35 (age)	9 → 2	bare number

Table 10: Representative lost spans on MEDDOCAN. O → T = detectors finding the original vs. the surrogate (panel of 11–12; original includes one extra decoding variant). Losses are shared across detectors and align with the §7 typology.

Zhu, Paul Denny, Michael Witbrock, and Jiamou Liu. Abstract meaning representation-based logic-driven data augmentation for logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5914–5934, 2024.

- [4] Roger L Berger and Jason C Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.
- [5] Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, 2020.
- [6] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021. URL <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>.
- [7] David Carrell, Bradley Malin, John Aberdeen,

Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2): 342–348, 2013.

- [8] David S Carrell, David J Cronkite, Muqun Rachel Li, Steve Nyemba, Bradley A Malin, John S Aberdeen, and Lynette Hirschman. The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with hiding in plain sight. *Journal of the American Medical Informatics Association*, 26(12):1536–1544, 2019. doi: 10.1093/jamia/ocz114.
- [9] Custodian Labs. Custodian Labs Guardian Layer. <http://www.custodianlabs.io/>, 2025. Company website; formal method citation to follow.
- [10] DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.
- [11] Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606, 2017.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186, 2019.
- [13] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1383–1392, 2018.
- [14] Daniele Faraglia and Other Contributors. Faker: a Python package that generates fake data. <https://github.com/joke2k/faker>, 2012.

- [15] Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. MultiCoNER v2: a large multilingual dataset for fine-grained and noisy named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.
- [16] Gemma Team, Google DeepMind. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- [17] Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. Large language models are not strong abstract reasoners. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6270–6278, 2024.
- [18] Grattafiori, Aaron and Dubey, Abhimanyu and others (Llama Team, AI @ Meta). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [19] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python. <https://spacy.io>, 2020.
- [20] Alistair E W Johnson, Lucas Bulgarelli, and Tom J Pollard. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*, pages 214–221, 2020.
- [21] Woojin Kim, Sungeun Hahm, and Jaejin Lee. Generalizing clinical de-identification models by privacy-safe data augmentation using GPT-4. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21204–21218, 2024. URL <https://aclanthology.org/2024.emnlp-main.1181/>.
- [22] Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8 (4):355–362, 2017.
- [23] Daniël Lakens, Anne M Scheel, and Peder M Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018.
- [24] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. Does BERT pre-trained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 946–959, 2021. URL <https://aclanthology.org/2021.naacl-main.73/>.
- [25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [26] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75: S34–S42, 2017.
- [27] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, et al. DeID-GPT: Zero-shot medical text de-identification by GPT-4. *arXiv preprint arXiv:2303.11032*, 2023.
- [28] Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 3798–3809, 2022.
- [29] Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurre, Heidy Rodriguez, Jose Antonio Lopez Martin, Marta Villegas, and Martin Krallinger. Automatic de-identification of medical texts in Spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF)*, 2019.
- [30] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

- [31] Microsoft. Presidio: Data protection and de-identification SDK. <https://github.com/microsoft/presidio>, 2018.
- [32] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [33] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [34] Atiqur Rahman Sarkar, Yao-Shun Chuang, Norman Mohammed, and Xiaoqian Jiang. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific Reports*, 14:29669, 2024. doi: 10.1038/s41598-024-81170-y.
- [35] Donald J Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.
- [36] Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58:S20–S29, 2015.
- [37] U.S. Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the HIPAA Privacy Rule. <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/>, 2012.
- [38] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563, 2007.
- [39] Thomas Vakili and Hercules Dalianis. Using membership inference attacks to evaluate privacy-preserving language modeling fails for pseudonymizing data. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 318–323, 2023. URL <https://aclanthology.org/2023.nodalida-1.33/>.
- [40] Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 4245–4252, 2022. URL <https://aclanthology.org/2022.lrec-1.451/>.
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [42] James Weatherhead, George Golovko, and Peter McCaffrey. ASQ-PHI: An adversarial synthetic data benchmark for clinical de-identification and search utility. *Data in Brief*, 65:112586, 2026. doi: 10.1016/j.dib.2026.112586.
- [43] Siheng Xiong, Ali Payani, and Faramarz Fekri. Enhancing long chain-of-thought reasoning through multi-path plan aggregation. *arXiv preprint arXiv:2510.11620*, 2025.
- [44] Siheng Xiong, Ali Payani, Yuan Yang, and Faramarz Fekri. Deliberate reasoning in language models as structure-aware planning with an accurate world model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31900–31931, 2025.
- [45] Siheng Xiong, Oguzhan Gungordu, James C Kerce, and Faramarz Fekri. Adaptive information control for search-augmented LLM reasoning. *arXiv preprint arXiv:2602.01672*, 2026.
- [46] Siheng Xiong, Ali Payani, and Faramarz Fekri. Enhancing language model reasoning with structured multi-level modeling. In *International Conference on Learning Representations (ICLR)*, pages 36557–36610, 2026.
- [47] Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language*

Processing (TrustNLP), pages 232–241, 2023.
URL <https://aclanthology.org/2023.trustnlp-1.20/>.