Life of Research · Published 1 November 2026 · DOI 10.5281/zenodo.23849157

A Detector Score Is Not a Release Risk: What PII Detection Benchmarks Certify About Deployment, Why 'Rivals Human Annotators' and 'Fails Under Shift' Are Both Measured, and the Coverage Audit a Release Decision Needs

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Automatic detection of personally identifiable information (PII) is the step that decides what text may leave an organisation, and its published record appears to contradict itself. Clinical de-identification systems have been reported to match or exceed human annotators, with recall above 0.94 on held-out notes and above 0.99 for one rule-based tool; a 2026 stress test reports that encoder, rule-based and language-model detectors all degrade under noisy and informal input, with location recall falling below 0.5, and a unified benchmark reports span F1 below 0.14 for every system it tried. In this analysis the author reviews and synthesises published results to show that most of the contradiction is in what the scores index rather than in the detectors. A PII detection score carries four indices that headlines omit: how the test text relates to the text the detector was built on, the scoring rule, the taxonomy that defines what counts, and the unit over which misses are aggregated. The same shared-task system misses 9.1 percent of protected items under one official rule and 3.6 percent under another; the lowest unified-benchmark scores come partly from entity types the evaluated detectors do not define; and in the 2026 stress test several of the losses attributed to the language model are spans it covered under another label. A residual shift effect survives every index: under fixed scoring, cross-institution transfer cost about ten points of strict F1, and systems run unmodified on new psychiatric records scored a median of 0.63. None of the indices is the quantity a release decision needs, which is the probability that a released document still carries an identifying item someone could find. The paper consolidates the evidence in a table, derives how per-item recall compounds at the document level, and gives a coverage-audit procedure with the sample sizes a release claim requires.

Keywords: personally identifiable information; de-identification; named entity recognition; distribution shift; evaluation methodology; clinical text; privacy risk; text anonymisation

1. Introduction

Every organisation that shares text about people runs some version of the same step. Clinical notes are de-identified before they reach researchers, customer chats are scrubbed before they are logged, court rulings are pseudonymised before publication, and training corpora are filtered before a model sees them (Meystre et al., 2010; Lison et al., 2021; Lukas et al., 2023). The step is almost always automatic, and the evidence that it works is almost always a benchmark score: precision, recall and F1 of a detector against human annotations of the spans that should have been removed.

That evidence has produced two literatures that read as if they describe different technologies. The first is the clinical de-identification record. A rule-based system built at MIT reached an estimated recall of 0.943 on unseen nursing notes, which its authors compare with 0.81 for a single human de-identifier and 0.94 for a consensus of two (Neamatullah et al., 2008). A system evaluated on 3,503 notes spanning 22 note types at a large paediatric hospital was reported as statistically indistinguishable from two additional human annotators (Deleger et al., 2013). Philter reported word-level recall of 99.46 percent on a 2,000-note test corpus from UCSF (Norgeot et al., 2020). A production pipeline is reported to have de-identified over a billion notes without manual review (Kocaman et al., 2023). Recent language-model systems are described in the same register: DeID-GPT is presented as giving the most accurate masking of the methods its authors compared (Liu et al., 2023).

The second literature is about robustness, and its headline numbers are much lower. In the 2016 CEGS N-GRID shared task, nine teams ran their existing de-identification systems without modification on new psychiatric intake records; the best of the 17 scored runs reached F1 0.799 and the median run 0.629 (Stubbs et al., 2017). On radiology reports, none of five public tools exceeded 50 percent F1 on patient names (Steinkamp et al., 2020). A unified benchmark built from ten public PII, NER and financial datasets found that all eight published systems it evaluated scored span F1 below 0.14 (Jha, 2026). OpenAI's Privacy Filter, evaluated independently across 32 benchmarks, scored F1 0.855 on one synthetic PII set and 0.03 to 0.04 on Cyrillic and Arabic-script general entity data (Uppala, 2026). And a 2026 study by Zafar and Nowaczyk, the anchor of this paper, built a stress test of seven natural distribution shifts and reported that encoder, rule-based and generative detectors all lose recall on it, that location recall falls to about 0.46 for all three, and that "aggregate benchmark scores obscure deployment-critical

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

weaknesses" (Zafar and Nowaczyk, 2026).

Read side by side, the two literatures pose a direct question. Does a published PII detection score tell an organisation how much personal information its pipeline will let through? The clinical record suggests that, on the right data, it does, and that the answer is good. The robustness record suggests that the score is a property of the benchmark and that deployment is worse.

In this analysis the author reviews and synthesises published results to show that both literatures are largely right about what they measured, and that they measured different things. A reported PII detection score carries four indices that a headline rarely states. The first is **provenance**, the relation of the test text to the text the detector was developed on, whether the same institution and note type, another institution, another genre or synthetic templates. The second is the **scoring rule**, which decides whether a hit needs exact boundaries, any overlap or a single token, and whether the predicted label has to agree with the gold label. The third is the **taxonomy**, the set of span kinds that count as PII and the annotation convention behind it. The fourth is **aggregation**, the unit over which misses are averaged: mentions, categories, demographic groups, documents, or the people a text describes.

Once the four indices are fixed, much of the apparent conflict dissolves, and what remains is smaller and more specific than either headline. The paper shows three things. First, scoring rule and taxonomy alone move published numbers by amounts comparable to the shift effects in dispute. The winning system of the 2014 i2b2/UTHealth track misses 9.1 percent of protected items under the task's primary rule and 3.6 percent under another rule the same paper reports (Stubbs et al., 2015a; Figure 1a). The lowest unified-benchmark scores arise partly from entity types the evaluated detectors never claimed to find, and partly from exact-span and label mismatches on types they do define (Jha, 2026). Second, a residual shift effect survives when the indices are held fixed: under unchanged strict scoring, a deep-learning model trained on i2b2 data lost about ten points of F1 at another institution (Yang et al., 2019), and the 2016 sight-unseen systems fell well below the 2014 results under the same annotation guidelines (Stubbs et al., 2017). Third, none of the four indices yields the quantity a release decision depends on, which is the probability that a released document still carries an identifying item that someone could find. That quantity compounds per-item recall over the number of items in a document (Figure 1b), and it is reduced by how hard a surviving item is to tell apart from replacement text (Carrell et al., 2013; Carrell et al., 2020).

The paper makes six contributions. It stipulates the four indices and the terms used throughout (Section 2). It re-reads the anchor study at the resolution of its own tables (Section 3) and sets out the opposing record (Section 4). It works through each index with the published results that isolate it (Sections 5 to 8). It separates a detector's miss from a document's exposure and from a person's re-identification (Section 9), and examines the anchor's proposal to combine detectors (Section 10). It consolidates the measured values in Table 1 (Section 11) and gives a coverage-audit procedure, Algorithm 1, with the sample-size arithmetic a release claim needs (Section 12). It closes with the strongest objections, the studies that would settle what remains, and a flat statement of what is not known (Sections 14 to 16).

The argument is not that the robustness critique is wrong. Its central observation, that a clean-benchmark score is not a deployment guarantee, is correct and has been made in the clinical literature for at least fifteen years (Uzuner et al., 2007; Meystre et al., 2010; Johnson et al., 2020). Nor is the argument that the near-human results are inflated; several were measured on real notes from the deploying institution. The argument concerns the inference from "the score fell on a stress set" to "the pipeline leaks this much in deployment", which holds only after all four indices have been fixed and the per-item miss has been converted into an exposure.

2. Scope and terms

The paper concerns detectors that mark spans of text as personal information so that they can be removed, masked, replaced by type tags, or replaced by realistic surrogates. It covers clinical de-identification, general PII detection and text anonymisation tools insofar as they rely on span detection. It does not cover differentially private text rewriting, image redaction except where document images bear on aggregation, or memorisation of PII by trained models except as context.

Detector. Any procedure that outputs a set of character spans with optional labels, whether rule-based, an encoder classifier, a generative model prompted to extract, or a combination.

Item. One gold-standard span annotated as personal information. A person's name mentioned five times is five items.

Typed recall. The fraction of gold items matched by a predicted span carrying the same label, under a stated boundary rule. Most published PII scores are typed.

Coverage. The fraction of gold items, or of gold characters, that fall inside some predicted span regardless of label. Coverage is what a redaction or masking step consumes: a span labelled SSN that is in fact a credit card number is still removed.

Exposure. The event that a released document contains at least one item that the release policy required to be removed. Exposure is a property of documents, not items.

Findability. Given an exposure, the probability that an adversary with stated resources can identify the

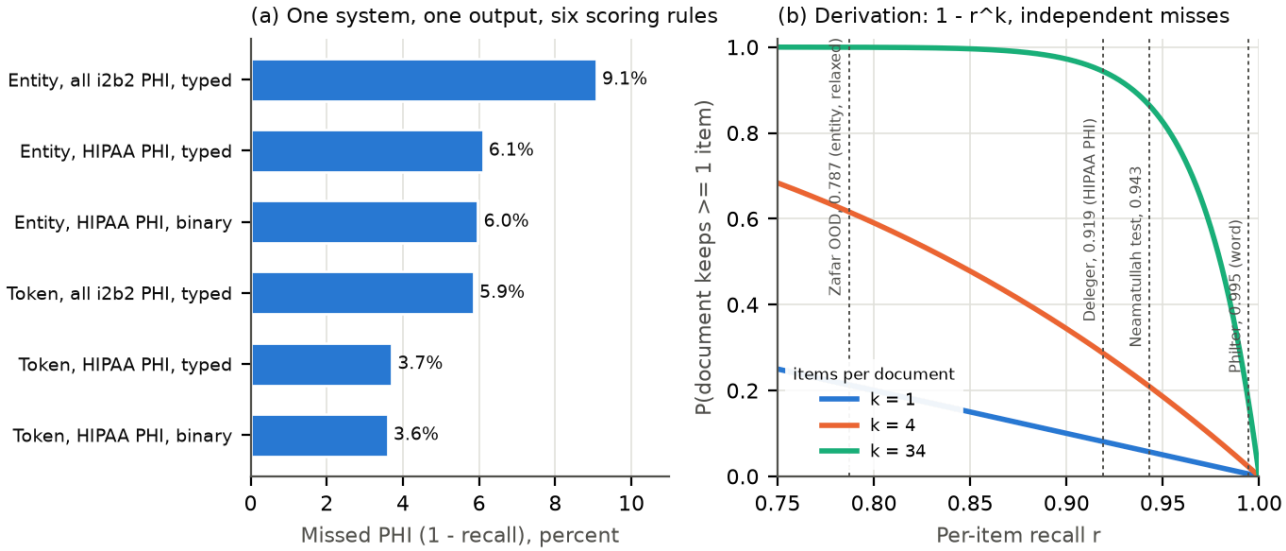


Figure 1. (a) Miss rate (one minus recall) of the winning system of the 2014 i2b2/UTHealth de-identification track under the six scoring rules reported for the same submissions by Stubbs, Kotfila and Uzuner (2015), Tables 2, 4, 5, 6, 7 and 8. (b) Derivation, not measurement: probability that a document keeps at least one of k items when each is missed independently with probability 1 - r. Dashed lines mark published recall values for orientation only; their units differ (typed relaxed entity recall, Zafar and Nowaczyk 2026, OOD set; HIPAA PHI recall on real notes, Deleger et al. 2013; estimated recall on a test corpus, Neamatullah et al. 2008; word-level binary recall, Norgeot et al. 2020). k = 4 is near the anchor's 4.2 entities per stress example; k = 34 is the CEGS N-GRID mean PHI phrases per record (Stubbs et al. 2017).

surviving item as real personal information rather than surrogate or ordinary text.

Residual risk. The probability that a released document leads to a person being identified or to a protected attribute being disclosed. It depends on exposure, findability, the identifying power of what survived, and what the adversary already knows. Residual risk is the quantity regulation and release decisions are about; it is not what a detector benchmark reports.

The four indices of Section 1 are properties of a reported number. Provenance and taxonomy concern what was tested and against what standard; scoring rule and aggregation concern how hits and misses were counted. A number is comparable to another only when all four agree, and transferable to a deployment only when the deployment's text, policy and unit of harm match them.

3. The anchor result, read at the resolution of its tables

Zafar and Nowaczyk (2026) construct two evaluation sets. Set A, the clean baseline, has 100 examples generated from 10 hand-written templates filled with synthetic values from the Python Faker library, giving 350 annotated entities across nine types: person, location, address, phone, email, IP address, social security number, date of birth and credit card. Set B, the stress test, covers seven shift categories (run-on text, conversational register, overlapping entities, abbreviations, mixed PII context, typos and code-switched text), with eight templates per category, each instantiated ten times: 560 examples and 2,330 entities. Gold spans come from the generation process. Three systems are evaluated: spaCy's large English model, Microsoft Presidio in its default configuration (which uses the same spaCy model for named entities), and Qwen2.5-3B-Instruct prompted zero-shot to return JSON. Each system is scored only on the entity types it supports. A prediction counts as correct if it overlaps a gold span and carries the same type; the authors note that this relaxed rule gives "an optimistic upper bound on actual redaction performance".

The headline numbers are in the anchor's Table 4. Overall F1 falls from 0.755 to 0.691 for spaCy, from 0.818 to 0.762 for Presidio, and from 0.921 to 0.848 for Qwen, so the robustness gap is 0.064, 0.057 and 0.073 points of F1. Precision barely moves for any system (Qwen 0.925 to 0.920); the loss is in recall, which falls from 0.917 to 0.787 for Qwen. Per type, location recall falls from 0.725 to 0.457 for both spaCy and Presidio and from 0.938 to 0.471 for Qwen. Qwen's credit-card recall falls from 1.000 to 0.543, and the model returned unparseable output on 14 of the 560 stress examples. Presidio's regular-expression recognisers keep recall at 1.000 for SSNs and IP addresses on both sets, but its date recogniser fires on bare numbers such as "6574", and date-of-birth precision drops to 0.492 on Set B. On typos, the category that separates the systems most, Qwen's F1 is 0.869 against 0.659 for Presidio and 0.522 for spaCy.

Three features of these tables matter for what the result licenses.

The scoring rule is typed, and two of the largest losses are typing losses. The authors state that the

language model "absorbs city names into ADDRESS spans", extracting a full street address including the town as one ADDRESS rather than separately tagging the town as LOCATION, and they call this "an entity granularity problem rather than a detection failure". Under the typed rule, the town counts as a missed location. Under coverage, it was masked. The same holds for credit-card numbers labelled SSN: the anchor reports that the model "frequently labels credit card numbers as SSN", and those digits would have been removed. The paper's risk paragraph nonetheless converts recall directly into exposure: the best model "fails to detect 21.3% of PII entities", and location has "a 54% miss rate". Those figures are one minus typed recall. What fraction of Qwen's 0.471 location recall and 0.543 credit-card recall is uncovered text, rather than covered text with the wrong label, is not reported. The relaxed-overlap rule makes the scores optimistic about boundaries; the type requirement makes them pessimistic about masking. The paper flags the first and not the second.

The clean set is small and templated. One hundred examples from ten templates give a clean-set recall estimate whose sampling uncertainty is not reported, and the per-type clean recalls rest on a few dozen items each. The 0.06 to 0.07 F1 gaps are differences between two template families written by the authors, not between a benchmark and a deployment. Some per-type movements go the other way: Qwen's person recall rises from 0.733 on clean text to 0.777 on the stress set. This is not a criticism of a stress test, whose purpose is to find failure modes. It bounds what the gap means: it measures sensitivity to the seven shifts as the authors wrote them.

Part of the location failure is a property of the generator. The authors report that Faker's invented town names ("Salaston", "Kellieborough", "Figueroahaven") are "almost never detected because they do not appear in any training data or knowledge base", and that town names containing person-like tokens, such as "East David", are tagged as persons. Real deployment text contains real town names, which a general-domain encoder has a better chance of having seen; it also contains rare and foreign ones. Whether the anchor's location recall over-states or under-states deployment behaviour depends on the frequency of names in the deployment population, which no number in the study estimates.

None of this removes the anchor's qualitative findings. Rule-based recognisers are robust to surrounding noise and brittle to format variants (spelled-out emails, phone extensions). Encoders lose unseen surface forms and boundaries in unpunctuated text. A small generative extractor confuses types and sometimes returns nothing. These failure modes are distinct and are documented with examples. What the tables do not establish is the size of any of them in deployment, or that the paper's exposure figures follow from its recall figures.

4. The other side: the near-human record

The opposing position is not that detectors are perfect. It is that, evaluated on real text from the setting in which they are used, good detectors make errors at rates comparable to trained humans, and that human review is not the safe alternative it is assumed to be.

The oldest version of this argument is Neamatullah et al. (2008). Their pattern-matching de-identifier, tuned on a gold-standard corpus of 2,434 nursing notes, was tested on a second corpus of 1,836 notes containing 296,400 words. It produced 90 false negatives, 27 per 100,000 words, for an estimated recall of 0.943; it missed one full date, one age over 89, and no patient names. The authors set this against recall of 0.81 for a single human de-identifier and 0.94 for a consensus of two. The authors nonetheless judged that the system's accuracy was "probably insufficient to be used to publicly disseminate medical data". The near-human result and the not-good-enough verdict sit in the same abstract.

Deleger et al. (2013) ran the largest within-institution evaluation in this literature: 3,503 notes, stratified across 22 note types, drawn from five million documents at one paediatric hospital, with more than 30,000 annotated PHI elements across all 18 HIPAA categories. The best system reached 91.92 percent recall and 95.08 percent precision overall. On a 10 percent subsample annotated by two further humans, the humans scored 92.15 and 94.55 percent recall, and the system 92.91. The study also found that de-identification barely affected downstream medication extraction.

Kayaalp et al. (2014) evaluated name redaction on reports containing more than a million words, 2,388 patient-name tokens and 20,160 provider-name tokens. The NLM Name Scrubber missed 2 of the 2,388 patient-name tokens, a sensitivity of 0.999, and the authors argue that the two misses carried negligible risk because neither person could be distinguished among more than 150,000 people listed in the US Social Security registry. Norgeot et al. (2020) report word-level recall of 99.46 percent for Philter on a 2,000-note UCSF test corpus and 99.92 percent on the 2014 i2b2 test corpus, with precision near 78 percent on both; Philter is explicitly tuned to trade precision for recall, and its secondary metric is F2. Prakash et al. (2025) report token-level F1 of 0.973 for a radiology de-identification model on reports from a second institution, Penn, after large-scale training at Stanford, and 0.960 against 0.632 to 0.754 for commercial cloud services on synthetic Penn reports.

On the general-PII side, the evidence for robustness is thinner but present. Krco et al. (2026) find that an LLM-based anonymiser keeps its re-identification protection across English, Spanish and Chinese (24 to 26 percent of profiles re-identified), while the re-identification rate after Presidio rises from 48 percent in English to 86 to 88 percent in the other two languages. Hsu and Berlin (2026) report that a compact encoder trained on 77

relabelled public datasets achieves the highest recall in 9 of 14 comparisons even under benchmark labels it was never trained on. Trienes et al. (2020) found that a neural de-identifier performed strongly on a varied Dutch corpus drawn from nine institutes and three healthcare domains, even with limited training data, while an existing rule-based system developed for Dutch failed to generalise to it. Acharya and Shrestha (2025) report that lightweight fine-tuned models match frontier LLMs on masking benchmarks derived from AI4Privacy, while noting that the same models degrade on live, informal chat messages.

Two things follow. First, the near-human results are not artifacts of clean benchmarks: several were measured on real notes from the deploying institution, across many note types. Second, they are almost all within-institution or adapted results, scored under rules (word-level, token-level, binary PHI-versus-not) that are among the most lenient in use. The robustness literature is not contradicting these numbers. It is reporting numbers with different indices.

5. Index one: provenance

Provenance is the relation between the text a detector was built on and the text it is tested on. The literature distinguishes at least five relations, and they are not ordered by difficulty in any simple way.

Same source, held-out documents. This is the standard shared-task design. The 2006 i2b2 challenge used discharge summaries from one hospital system, with real PHI replaced by surrogates and deliberately enriched with invented names and with surrogates that were ambiguous with medical terms; the best systems exceeded 98 percent F-measure in every PHI category, and the hard cases were the ambiguous surrogates (Uzuner et al., 2007). In 2014 the best system, a hybrid of conditional random fields with keyword and rule components, reached 0.936 micro-averaged entity-level F1 (Stubbs et al., 2015a; Yang and Garibaldi, 2015). Dernoncourt et al. (2016) report 97.85 binary token-level F1 on the same 2014 data and 99.23 on a MIMIC corpus built by the same team.

Same task, another institution. Yang et al. (2019) trained deep-learning de-identifiers on the 2014 i2b2 corpus and applied them to a 200-note test set drawn from 500 annotated University of Florida Health notes. Strict F1 fell from 0.9547 to 0.8568 and relaxed F1 from 0.9646 to 0.8958; customising with local notes and resources recovered strict F1 to 0.9288. Lee et al. (2021) studied a neural de-identifier across four corpora from two institutions and found that a single external source transferred inconsistently, while pooling several external sources gave F1 around 80 percent; fine-tuning on local data dominated. Posada et al. (2026) evaluated models trained on their SHIELD notes against i2b2 2014 and a Stanford radiology corpus and found an asymmetry: dates, identifiers and phone numbers transferred in both directions, while hospital and location names did not. A DeBERTa model trained on SHIELD recalled 0.39 of hospital names in i2b2 and 0.02 in the radiology corpus. The second figure is partly a taxonomy effect: the authors state that "The AIMI HOSPITAL and LOCATION comparisons are partly confounded by label remapping", and that their conclusion "is strongest in the clean i2b2 direction, where the label taxonomy matches". The i2b2 value is the cleaner measure of provenance.

Same guidelines, new genre, no adaptation. The 2016 CEGS N-GRID sight-unseen track is the cleanest test of this relation in the literature. The records were psychiatric intake notes annotated under the 2014 guidelines, so taxonomy and scoring rule were held fixed. Nine teams ran existing systems unmodified. Under strict matching over all PHI categories, the median system scored 0.629 and the best 0.799, against a median of 0.845 and a maximum of 0.936 in 2014 (Stubbs et al., 2017). The organisers attribute part of the drop to the corpus itself: three times as many tokens per record, PHI about twice as sparse relative to tokens (0.0185 against 0.0358 PHI phrases per token), and more PHI about patients' lives, such as professions, towns and employers. With two months of in-genre training, the best system reached 0.914.

Out-of-the-box tools on a narrow specialty. Steinkamp et al. (2020) annotated 2,503 radiology reports and evaluated five public tools; the best two, both machine-learned, reached micro-averaged F1 of 93.6 and 92.2 percent, yet neither exceeded 50 percent F1 on patient names. Dates made up 78 percent of the PHI.

Synthetic templates. The anchor's sets, Xiao et al.'s (2023) 100 clinical templates with inserted names, Mansfield et al.'s (2022) customer-service templates, and Singh and Narayanan's (2025) 17,000 semi-synthetic sentences all generate gold spans by insertion. Synthetic construction controls the shift precisely and isolates mechanisms. It also imports the generator's distribution of values. The anchor's invented town names are one example. Another comes from Lee et al. (2021): the i2b2 corpora carry surrogate PHI, while the second institution's notes carried authentic PHI, so transfer between them crossed a surrogate-to-authentic boundary as well as an institutional one.

Two general findings from named-entity recognition explain why provenance matters as much as it does. Taille et al. (2020) show that standard English NER benchmarks have unrealistically high lexical overlap between training and test mentions, which makes lexical features look more important than they are; the benefit of contextual embeddings was about 1.2 percent relative micro-F1 in-domain and 13 percent out-of-domain. Lin et al. (2021) replace entities with others of the same class from Wikidata and perturb context words, and find that even the best models drop sharply, consistent with

memorisation of in-domain entity patterns. Fu et al. (2020) reach a similar diagnosis through breakdown analyses that separate performance by dataset bias and annotation errors. A PII detector's recall on held-out documents from its own source therefore over-represents the names, towns and formats it has already seen.

The provenance index cuts in both directions. It explains why the robustness numbers are lower: they are measured further from the development data. It also shows that the robustness numbers are not all one quantity. The anchor's stress set is a shift in register and format, inside a general domain, on synthetic values. CEGS N-GRID is a shift in genre under fixed guidelines. Yang et al. is a shift in institution. These produce drops of different sizes for different reasons, and a single "robustness gap" averages over them.

6. Index two: the scoring rule

A PII score requires a rule for when a predicted span counts as finding a gold span. The rules in use differ along two axes: boundary strictness (exact offsets, relaxed offsets, any overlap, per token) and label strictness (typed, or binary PHI-versus-not).

The 2014 i2b2/UTHealth overview reports the same submissions under six rules: entity-level or token-level, crossed with all i2b2 PHI categories, HIPAA categories only, or binary HIPAA PHI (Stubbs et al., 2015a). For the winning Nottingham system, recall is 0.9092 under the primary rule (entity-level, all i2b2 categories, typed), 0.9390 for entity-level HIPAA, 0.9403 for entity-level binary HIPAA, 0.9414 for token-level i2b2, 0.9629 for token-level HIPAA, and 0.9640 for token-level binary HIPAA. F1 runs from 0.9360 to 0.9768. Figure 1a plots the corresponding miss rates. Choosing between two rules the organisers themselves published changes the fraction of protected items missed from 9.1 percent to 3.6 percent, a factor of about 2.5, with no change to the system or its output. The overview's own explanation is that token-level rules forgive split or partial spans, and that for de-identification it matters less whether a name is captured as one span or two, provided every part of it is removed.

The 2016 overview makes the same point from the other side. Its evaluation script supports strict, relaxed and overlap matching over all PHI or HIPAA PHI; moving from strict all-PHI to token-based HIPAA-only scoring raised most teams' scores by about 0.03 to 0.04 (Stubbs et al., 2017). Yang et al. (2019) show that the rule changes the size of a shift effect as well as the level: their cross-institution drop was 9.8 points of strict F1 and 6.9 points of relaxed F1.

Results reported across studies mix these rules freely. Dernoncourt et al.'s (2016) 97.85 on i2b2 2014 is binary token-level HIPAA F1, under which the shared-task winner scored 0.9768; their improvement over the winner is real but small, while the gap between 97.85 and the commonly cited 0.936 is mostly the rule. Philter's 99.46 percent is word-level binary recall (Norgeot et al., 2020). The anchor's numbers are relaxed-overlap and typed. Uppala (2026) scores the Privacy Filter by exact character offsets and attributes, without an ablation, part of GPT-4o's deficit on highly structured synthetic PII to generative extraction missing exact boundaries; if that reading is right, part of the difference between those two systems is boundary convention rather than detection.

Label strictness interacts with the release action. If spans are deleted or replaced by a generic tag, the label is irrelevant to exposure and binary coverage is the right score. If spans are replaced by type-appropriate surrogates, as in hiding-in-plain-sight, a wrong label produces a wrong-typed surrogate, which may itself draw attention. The surrogate case is where typed scoring has a direct privacy meaning. For deletion pipelines, typed recall understates protection. This is why the anchor's typed location and credit-card recalls cannot be read as exposure rates without knowing the release action and the fraction of typed misses that were covered. That fraction can be small. Kumar et al. (2026) re-score document-image redaction crediting any well-localised box whatever its predicted type, and critical page-level leakage moves only from a range of 0.968 to 1.000 to one of 0.945 to 1.000; the authors conclude that "the high leakage rates are not primarily caused by type-label" errors. How much coverage scoring changes a result is an empirical question for each benchmark, not a guaranteed gain.

7. Index three: what counts as PII

A detector cannot miss what it was never asked to find, but a benchmark can count it as missed.

Clinical de-identification has a legal floor, the 18 HIPAA Safe Harbor categories, and the shared tasks deliberately annotate above it. The 2014 i2b2 guidelines add doctors' names, hospitals, professions, and locations of any size, on the reasoning that longitudinal records combine individually innocuous details into identifying ones (Stubbs et al., 2015a; Stubbs and Uzuner, 2015). Every scoring rule in Section 6 is crossed with the choice between the extended and the HIPAA taxonomy, and the extended categories are harder: profession and location were the most difficult categories in 2014 (Stubbs et al., 2015a). The companion corpus paper reports that the annotators themselves were imperfect against the adjudicated gold standard: their average token-based F1 was 0.927 (Stubbs and Uzuner, 2015).

General PII detection has no shared floor, and the consequences are large. Jha (2026) merges ten public datasets into 48 canonical types and evaluates eight systems by exact span and type. The best system, Presidio, scores F1 0.1385. The per-type results show where the number comes from, and taxonomy explains

only part of it. The most frequent test type is FINANCIAL_ENTITY (873 of the test mentions), derived from XBRL tags in a financial NER corpus; no general PII detector defines it, and both Presidio and spaCy score zero on it, as they do on usernames and companies. The types Presidio does define also score low. It scores zero on credit-card numbers, for which the table notes "Presidio has recognizer but recall = 0" (Jha, 2026), and zero on IP addresses, which the table marks as not covered even though Presidio's regular-expression recogniser for them reaches recall 1.000 in the anchor study (Zafar and Nowaczyk, 2026). It scores 0.02 on SSNs, 0.13 on email and date, 0.26 on person and 0.31 on location, the best-scoring type. Those seven types account for 1,351 test mentions in the per-type table, more than the financial type, which points to exact-span matching and label alignment between source datasets as well as taxonomy. The benchmark's own limitations section concedes that the financial source dominates and that per-entity analysis should be preferred to micro-averaged F1. An F1 below 0.14 therefore measures partly the mismatch between eight detectors' taxonomies and a 48-type union, and partly how the detectors' spans and labels line up with the union's on types they do define. It is evidence about coverage of the union, and weak evidence, mixed with those matching effects, about robustness on the types the detectors define.

Hsu and Berlin (2026) show that agreement on category names does not settle the question. They relabelled two public benchmarks under written contextual definitions of 12 PII categories. On AI4Privacy, in five categories (credit card, driver licence, passport, social security and tax identification numbers), 59 to 86 percent of the originally labelled positives were dropped by the relabelling, and context words naming the category appeared in the text for 0.24 percent of these dropped positives against 54.2 percent of the positives both label sets shared. In an audit of 50 sampled disagreements per category, two LLM judges sided with the relabelling in 68 and 72 percent of AI4Privacy cases. On Nemotron-PII, the disagreement ran the other way: 29 percent of IP-address positives existed only in the new labels, almost all of them IP addresses inside URLs that the original labels had tagged only as URLs. A detector that tags the embedded address is penalised by one label set and credited by the other. Hsu and Berlin also relay a report from the Privacy Filter's developers that AI4Privacy omits spans an LLM reviewer judged to be valid PII, so that the original labels understate that model's performance (de Bourcy et al., 2026, as reported by Hsu and Berlin, 2026). The independent Privacy Filter evaluation by Uppala (2026) restricts one general-NER benchmark, MultiCoNER, to the person type so that predictions outside the model's schema do not count as false positives. Separately, on the PII datasets, where no such restriction applies, it reports precision of 0.31 to 0.54 for most of them.

Taxonomy disagreements reach clinical transfer too. Posada et al. (2026) report near-zero recall on AGE for all models in their i2b2 transfer because of differing label definitions, and note that two of their radiology-corpus comparisons are confounded by label remapping. Lee et al. (2026) argue that existing PII taxonomies are too coarse for Korean court judgments and introduce a three-tier scheme. Zhou et al. (2026) show the problem from the precision side: in mathematics tutoring transcripts, numbers that look like dates or identifiers are redacted by generic detectors, and false redactions cluster in maths-dense passages. Presidio scores F1 0.379 there, against 0.802 to 0.821 for domain-aware prompting of an LLM.

Text anonymisation research has argued for years that the category frame is itself the limitation. Lison et al. (2021) contrast detection of predefined categories with anonymisation that accounts for semantic inference, and Pilan et al. (2022) built the TAB corpus to mark which spans must be masked to conceal a specific person, beyond predefined categories. Krco et al. (2026) find that identifiers written in non-standard forms raise average re-identification across 12 tools from 44 to 69 percent, and that implicit identifiers are barely reduced by any tool. These results do not say that detectors are unrobust. They say that "recall" depends on a definition of the positive class that varies between benchmarks more than detectors vary within one.

8. Index four: aggregation

A miss rate is an average, and the unit it averages over decides what it can be used for.

Over mentions, dominated by frequent categories. Micro-averaged recall weights each category by its frequency. In Steinkamp et al.'s (2020) radiology corpus, dates are 78 percent of PHI; a tool can reach 92 to 94 percent micro-F1 while scoring below 50 percent on patient names, the category most directly identifying. In the anchor, rigid-format types that regular expressions catch perfectly sit beside location recall below 0.5. An aggregate score is a weighted sum whose weights come from the test corpus, not from any judgement of which misses matter.

Over people, by demographic group. Xiao et al. (2023) insert 16 name sets varying by gender, race, popularity and decade into 100 clinical templates and evaluate nine de-identification methods. Most show statistically significant recall differences along most dimensions, and overall name recall ranges from 0.629 for spaCy to 0.974 for flair; Philter, which reports 99.46 percent word-level recall on its own corpus, recalls 0.794 of names at precision 0.227 in these templates. Mansfield et al. (2022) find the highest masking error rates for names associated with Black and Asian or Pacific Islander individuals, and Mishra et al. (2020) document similar gaps in general NER. Xiao et al. trace the lowest-

recall names to polysemy: names that are also English words or places. A mention-averaged recall hides the fact that, for some name groups, the miss rate is up to roughly twice the average.

Over documents. Exposure is a document-level event. If a document contains k items and each is missed independently with probability 1 - r, the probability that the document keeps at least one is 1 - r^k. Figure 1b plots this. At r = 0.919, Deleger et al.'s system recall, a document with four items is exposed with probability 0.29, and one with 34 items, the mean number of PHI phrases per CEGS N-GRID record (Stubbs et al., 2017), with probability 0.94. At Philter's word-level recall of 0.9946 the corresponding values are 0.021 and 0.17. The independence assumption is wrong in both directions: repeated mentions of the same rare name tend to be missed together, which concentrates misses in fewer documents, while correlated failures within difficult note types concentrate them in particular documents. But no published per-item recall can be converted into a document-level rate without either the assumption or a document-level count. Three studies report the count directly. On real radiology reports, Steinkamp et al. (2020) count the reports that kept at least one PHI element after each tool ran: for the two best tools, with token-level recall of 92.6 and 88.2 percent, 91 and 80 of the 1,023 test reports kept PHI, which is 14.1 and 12.4 percent of the reports that contained any. Kumar et al. (2026) score redaction of document images and report that, with tool assistance, a frontier model's localisation F1 rose from 0.090 to 0.249 while page-level critical leakage fell only from 0.990 to 0.968. Krco et al. (2026) report that the chance of re-identifying a synthetic profile rises quickly with the number of identifiers it contains, reaching 94 percent at five direct identifiers for one perturbation-based tool.

Over subjects and conversations. Oh et al. (2026) shift the unit from spans to the people described in legal and online documents and find that even when more than 90 percent of PII spans are masked, subject-level protection against inference can fall to 33 percent, with non-target subjects more exposed than the target. Zhang et al. (2026) annotate recurring identifiers across multi-turn conversations and find that no detector covers every mention of every identifier without substantial false positives on PII-free dialogues, and that reading a long dialogue in one pass loses about a third of the gold characters. If an identity is exposed when any one of its mentions survives, a per-mention recall overstates protection for people mentioned often.

Aggregation explains a feature of the two literatures that otherwise looks like disagreement: the near-human results are mention-level averages over corpora dominated by easy categories, while several of the robustness results are per-category minima or document-level rates. Both are correct averages. They answer different questions.

9. From a miss to an exposure to an identification

A detector benchmark stops at the item. Release decisions are made about people. Two further links separate them, and published evidence bounds both.

Findability. Carrell et al. (2013) proposed hiding in plain sight: replace every detected identifier with a realistic surrogate, so that the identifiers the detector missed are hidden among fake ones. In a pilot on 81 oncology and family-practice notes read by three trained readers, about 90 percent of residual identifiers were concealed, and the authors projected that, if this generalised, systems with recall of 94 to 98 percent could reach effective de-identification above 99 percent. Carrell et al. (2020) tested the idea adversarially at scale: 200 documents from two healthcare settings, four readers including two from the originating institution, explicitly trying to find leaked identifiers. Readers' mean recall of leaked PII was 26.8 percent at precision 37.2 percent, so about 70 percent of leaked identifiers survived an aggressive human attack; internal readers found more (32 percent) than external ones (22 percent). Against human readers, then, surrogate replacement left between about a tenth (2013 pilot) and about a quarter (2020 attack) of leaked identifiers findable. A machine-assisted attack found far more. Carrell et al. (2019) trained a machine-learned "parrot" attacker on text de-identified the same way, and it "correctly hypothesized that 211 (68%) of 310 actual PII leaks in the corpus were leaks", at the cost of wrongly flagging 191 surrogates as leaks. Across the published attacks, the findable share of leaked identifiers therefore runs from about a tenth to about two-thirds, depending on the attacker. Surrogate replacement is not shown to be a defence against an attacker who knows the original or can check candidates against outside records, and it is a weaker one against machine assistance; none of these was tested in the 2020 design. Prakash et al. (2025) report that synthetic surrogates generated this way remain detectable to their model at F1 0.959 across 50 independently surrogated datasets, which concerns the detector's stability rather than an attacker's ability.

Identifying power. Not every surviving item identifies. Kayaalp et al. (2014) argue that their two missed patient-name tokens carried negligible risk because neither person could be distinguished among more than 150,000 people in the US Social Security registry. El Emam et al. (2011) reviewed 14 re-identification attacks on health and other data; on average about a quarter of records were re-identified, but most attacks used data not de-identified to any standard, and the single attack on health data de-identified to current standards succeeded for 0.013 percent of records. On the other side, Sarkar et al. (2024) show that de-identified clinical notes still permit membership inference against models trained on them, and Lukas et al. (2023) show that

scrubbing reduces but does not prevent PII extraction from language models trained on scrubbed text. Krco et al. (2026) find that indirect identifiers, which most detectors do not target, still enable re-identification after the tools they evaluated have run, and that identifiers conveyed only implicitly are barely affected by any of them.

The chain from benchmark to harm thus has at least four factors: per-item miss rate under the deployment's taxonomy and scoring; compounding over items per document; findability given the release format; and identifying power given the adversary. The robustness literature mostly measures the first factor under shifted conditions and reports it as if it were the product. The near-human literature mostly measures the first factor under favourable conditions and reports a comparison with humans, who face the same chain. Neither measures the product. The anchor's sentence that organisations certifying their systems by benchmark F1 are "systematically underestimating their data exposure risk" may be true, but its own numbers cannot establish it, because the second through fourth factors can move the answer in either direction by more than the 0.06 to 0.07 F1 gap it measured.

The precision side belongs in the same chain. Every published operating point trades misses against over-redaction. Philter accepts 78 percent precision to reach 99.5 percent recall (Norgeot et al., 2020). The Privacy Filter is recall-biased, with precision of 0.31 to 0.54 on most PII datasets (Uppala, 2026), and its model card describes it as a data-minimisation component rather than "an anonymization or compliance guarantee" (de Bourcy et al., 2026). Aghakasiri et al. (2025) quantify how much clinically relevant information LLM-based de-identifiers remove inappropriately, and find that an existing automated metric for such removal performed poorly when checked against clinical experts. Meystre et al. (2010) already listed over-scrubbing among the field's unresolved issues. A robustness claim stated only in recall says nothing about what a more robust detector costs the data's users.

10. Combining detectors: what a union buys

The anchor's main recommendation is a hybrid pipeline that runs a regular-expression layer, an encoder and a generative model in parallel and merges their outputs, on the argument that the system's "overall miss rate becomes bounded by the intersection of their failure sets rather than the union" (Zafar and Nowaczyk, 2026).

The set-theoretic claim is correct: if the merged output is the union of detections, an item is missed only if every component misses it. Three qualifications follow from the anchor's own data. First, the bound helps only where failures differ, and the anchor's most important failure is shared: location recall is about 0.46 for all three systems on the stress set. Presidio, in the default configuration evaluated, hands named entities to a spaCy model, and its location recalls (0.725 and 0.457) are identical to the spaCy baseline's. Where two components share a model, their intersection is that model's failure set. Second, a union adds every component's false positives. Presidio's date recogniser alone brings date-of-birth precision down to 0.492 on the stress set, and the generative component's type confusions would add more. Third, the union's benefit is measured by coverage, not by typed recall; merging an ADDRESS span from one system with a LOCATION span from another resolves a typing disagreement that would otherwise be scored as a miss.

Rajgarhia et al. (2025) report large weighted-F1 gains for a hybrid of regular expressions and context-aware LLM filtering across 13 low-resource locales, and Kocaman et al. (2023) report that a hybrid context-based architecture outperforms an NER-only model by 10 percent on i2b2 2014 and makes far fewer errors than three commercial cloud services. Both support combination in principle. Neither reports the overlap of component failure sets, which is the quantity that decides how much a union can help. The anchor's tables could yield it for its three systems, since all were run on the same 560 stress examples; it is not reported.

11. Consolidated evidence

Table 1 collects the measured results the argument rests on, one row per result, with the source named on the row. Values are as reported by the source; "derivation" marks results computed from the definitions in Section 8.

12. A coverage audit for a release decision

Algorithm 1 turns the four indices and the chain of Section 9 into a procedure an organisation can run before trusting a detector on its own text. Its first steps fix the indices; its later steps estimate the quantity the release decision needs, with an explicit confidence bound.

Four design choices need defence.

Step 1 refuses to count undefined categories as misses. This is not leniency: an organisation whose policy requires financial-entity removal and whose detector does not define it has a coverage gap by design, which belongs in the policy decision rather than in a robustness score. It is the step that separates the taxonomy-driven part of Jha's (2026) result from the detection-driven part.

Step 4 scores coverage by default and typed recall only where the release action makes the label matter. For deletion and generic tagging, coverage is the protective quantity. For surrogate replacement, Section 6 explained why the label matters.

Step 7 makes the document the unit of the release decision and states the sample size the claim needs. The rule of three gives an approximate 95 percent upper

Table 1. Measured or derived results used in this paper, with the index each result isolates.

Source	Setting	Result	Index
Zafar 2026	Synthetic clean vs stress, typed relaxed	F1 gap 0.064 / 0.057 / 0.073; Qwen recall 0.917 to 0.787	Provenance
Zafar 2026	Same, per type	LOC recall to 0.457 (encoders), 0.471 (Qwen); CC 0.543	Scoring, taxonomy
Stubbs 2015a	i2b2 2014 winner, six rules	Recall 0.909 to 0.964; F1 0.936 to 0.977	Scoring
Stubbs 2017	Sight unseen, psychiatric	Best 0.799, median 0.629 vs 0.936, 0.845	Provenance
Yang 2019	i2b2 model at UF Health	Strict F1 0.955 to 0.857; relaxed 0.965 to 0.896	Provenance, scoring
Lee 2021	Pooled external sources	F1 about 0.80 without local data	Provenance
Posada 2026	SHIELD-trained to i2b2, radiology	Hospital recall 0.39 (i2b2) and 0.02 (radiology, label-remapped); dates transfer	Provenance, taxonomy
Steinkamp 2020	Five tools, radiology	Best micro-F1 93.6; names below 50 F1; PHI kept in 14.1 and 12.4% of PHI-bearing reports (best two tools)	Aggregation
Neamatullah 2008	Nursing notes test corpus	Recall 0.943; one human 0.81; two 0.94	Baseline
Deleger 2013	22 note types, one hospital	Recall 91.92; on subsample 92.91 vs humans 92.15, 94.55	Baseline
Norgeot 2020	UCSF, i2b2 2014	Word recall 99.46, 99.92; precision 78	Scoring
Xiao 2023	16 name sets, templates	Name recall 0.629 to 0.974; Philter 0.794	Aggregation
Jha 2026	48-type union, exact typed	Best F1 0.1385; most frequent type 0 for all but the financial specialist	Taxonomy
Hsu 2026	AI4Privacy relabelled	59 to 86% of positives dropped in 5 types	Taxonomy
Uppala 2026	Privacy Filter, 32 sets	F1 0.855 to 0.03; precision 0.31 to 0.54	Provenance, taxonomy
Krco 2026	12 anonymisers	Re-id 44% easy to 69% non-standard	Taxonomy
Kumar 2026	Document images	Page-level leakage 0.968	Aggregation
Oh 2026	Subject-level	Over 90% spans masked; protection to 33%	Aggregation
Carrell 2020	HIPS, four readers	Leaked PII found: recall 26.8%	Findability
El Emam 2011	14 attacks reviewed	One on standard-compliant health data: 0.013%	Identifying power
This paper	derivation	Document exposure 1 - r^k (Figure 1b)	Aggregation

Algorithm 1. Coverage audit before releasing detector-processed text

```
Input: detector set D; deployment pool S (unlabelled, from the target stream);
       protection policy P = (categories, weights w_c, release action:
       delete | tag | surrogate); exposure tolerance t; confidence 1 − a
1:  Fix taxonomy: map each detector label to P; categories in P that no
    detector defines are recorded as uncovered by design, not as misses
2:  Stratify S by source, document type, language, format and length;
    draw n documents per stratum; double-annotate items under P;
    adjudicate; record annotator-vs-gold agreement as the human baseline
3:  for each document: run D; if |D| > 1, merge spans by union
4:  Score coverage: item covered if all its characters lie in some
    predicted span, any label (delete/tag); require label match only
    when the release action is surrogate
5:  Report per category: coverage with interval; typed recall separately;
    precision and over-redaction in utility-critical regions
6:  Report per group: coverage by name origin and other declared strata
7:  Report per document: E = share of documents keeping >= 1 item with
    weight above threshold; upper bound U(E) at level 1 − a
    (zero exposures in n documents: U is about 3/n at 95 percent)
8:  if |D| > 1: report pairwise overlap of component miss sets
9:  if release action = surrogate: estimate findability with readers
    blind to which spans are real; report found fraction of leaked items
10: if U(E) x findability <= t: release under P
    (findability = 1 when the release action is delete or tag)
11: else: adapt (local data, rules for failing strata), re-run from step 3
    on fresh documents; never re-audit on the documents used to adapt
```

bound of 3/n on a rate when no events are observed in n trials (Hanley and Lippman-Hand, 1983). To claim a document exposure rate below 1 percent at 95 percent confidence from a clean audit therefore requires about 300 adjudicated documents per stratum in which the claim is to hold; to claim 0.1 percent requires about 3,000. Most of the near-human evaluations in Section 4 annotated hundreds to a few thousand documents in total, and reported item-level rather than document-level rates. Kayaalp et al. (2014) is the exception that shows what the item-level version of such evidence looks like: more than a million words and two missed patient-name tokens.

Step 8 makes the anchor's hybrid argument testable. A union is worth its cost in false positives only if the components' miss sets overlap little on the deployment

stream.

13. The anchor case through the audit

Running the anchor's experiment through Algorithm 1 produces a different reading from both its headline and its critics'.

Step 1 changes little: the anchor already scores each system only on types it supports. Step 4 changes the per-type numbers. Qwen's credit-card recall of 0.543 and location recall of 0.471 include covered-but-mislabelled spans. If the policy action is deletion, coverage for those types could be substantially higher; the paper's examples suggest that many of the location misses are towns absorbed into address spans, though the fraction is not reported. For spaCy and Presidio, part of the location loss is town names tagged as PERSON, which coverage would credit, and part is invented names never detected; the split is not reported. Step 7 converts recall into exposure. At 4.2 entities per stress example, the derivation of Section 8 puts the share of stress examples with at least one typed miss near 0.6 for Qwen ($r = 0.787$) under independence; the real share is not reported, and could be computed exactly from the anchor's outputs. Step 8 asks for the overlap of the three systems' miss sets on those examples, which the hybrid recommendation needs and the paper does not give. Step 9 does not apply to a detection benchmark.

The audit's verdict on the anchor's data is: the shift categories expose real and distinct failure mechanisms; the exposure figures in the risk paragraph are typed-recall figures; the location result partly reflects the synthetic value generator; and the hybrid proposal is plausible but unquantified. The anchor's practical advice, to evaluate on deployment-like inputs and to keep a regression suite of production failures, is the same advice Algorithm 1 formalises. What changes is what an organisation would measure: coverage under its own policy, per document, with a stated bound.

14. The strongest case against this reading

Four objections are serious.

First, the reading may excuse real failures as measurement artifacts. Some of the robustness literature's numbers survive every index: CEGS N-GRID's sight-unseen median of 0.629 was scored under the same guidelines and rule as 2014; Yang et al.'s cross-institution drop was measured under unchanged strict scoring; Steinkamp et al.'s patient-name result is a per-category figure on real reports. The paper does not dispute these. Its claim is narrower: the published range of PII scores, from above 0.99 to below 0.14, owes much of its width to the indices, and the residual shift effect is the roughly 7 to 22 points of F1 seen in the guideline- and rule-matched clinical comparisons (from Yang et al.'s relaxed-F1 drop to the fall in the CEGS N-GRID median over submitted runs). A reader who thinks that residual is the whole story, and that the indices are details, is not being unreasonable; the paper's answer is that the residual cannot be seen until the indices are fixed.

Second, coverage may be too lenient. A mislabelled span is masked, but a wrong-typed surrogate can draw a reader's attention, and a span covered by a too-long prediction may delete useful text. Algorithm 1 handles the first by requiring typed scoring under surrogate release and the second by reporting over-redaction separately. Whether these handle all cases is not known.

Third, the document-level derivation may mislead. Independence is false, and Figure 1b could make modest recall gaps look catastrophic. The figure is a derivation with its assumption stated, and the paper uses it only to show that per-item recall cannot be read as a document-level rate. The direct measurements that exist, page-level leakage of 0.968 in document images (Kumar et al., 2026) and profile-level re-identification that rises with the number of identifiers (Krco et al., 2026), point the same way. In the other direction, findability and identifying power can reduce the harm, by large factors against human readers though by much less against a machine-assisted attacker (Carrell et al., 2019; Carrell et al., 2020; El Emam et al., 2011), so a high exposure rate is not by itself a high re-identification rate.

Fourth, the paper may be re-stating what the clinical literature already knew. The 2014 and 2016 overviews themselves report results under several rules and discuss "how good is good enough" (Stubbs et al., 2015a; Stubbs et al., 2017). Johnson et al. (2020) list unclear annotation guidelines and poor model portability as open challenges, and Aghakasiri et al. (2025) document inconsistent reporting metrics across LLM-based de-identification papers. The paper's contribution is to put those observations to work on the current general-PII robustness debate, where they are mostly absent: the anchor cites none of the clinical shared tasks, and the unified-benchmark and Privacy Filter evaluations report single-rule scores. A reader who regards this as a synthesis rather than a new result has read it correctly.

15. Studies that would settle the open part

Five studies would settle most of what this paper leaves open, and each uses data or outputs that already exist or could be produced without new annotation at scale.

The first is a coverage re-score of the anchor. Using the released benchmark and the three systems' outputs, compute character-level coverage regardless of label, per type and per example, and the overlap of the three miss sets. This would say how much of the reported location and credit-card loss is typing, and what a union would buy.

The second is a rule sweep for general-PII benchmarks. Re-score published systems on AI4Privacy, Nemotron-PII and SPY under exact-typed, relaxed-typed, overlap-

binary and character-coverage rules, as the 2014 i2b2 overview did for clinical systems, and report the spread. If the spread is comparable to the claimed robustness gaps, single-rule comparisons should stop.

The third is a provenance-controlled stress test on real text. Take a corpus with authentic identifiers and institution-level splits, such as the clinical corpora used by Yang et al. (2019) or Lee et al. (2021), and apply the anchor's seven perturbation types to real notes, so that register shift and value-distribution shift can be separated.

The fourth is a document-level audit at the sample sizes of Section 12, on a deployment stream, reporting the share of documents with at least one uncovered high-weight item and its upper confidence bound, alongside item-level recall for the same sample.

The fifth is a findability study for general PII. Replicate Carrell et al.'s (2020) adversarial reading design on conversational and customer-support text with surrogate replacement, and add an LLM-assisted attacker, since the 2020 study found that human readers recovered about a quarter of leaked items, and the one machine-assisted attack measured, a machine-learned parrot attack on clinical notes, exposed about two-thirds (Carrell et al., 2019); neither tested an LLM-assisted attacker or conversational text.

16. What is not known

Five things. The fraction of typed misses in the anchor study, or in any general-PII robustness benchmark, that are covered by a differently labelled span has not been reported. The document-level exposure rate of any widely used general-PII detector on a real deployment stream has not been published with a confidence bound. The overlap of miss sets between rule-based, encoder and generative detectors on the same text has not been reported, so the benefit of combining them is unquantified. How much of the shift effect seen on synthetic stress sets transfers to real text with real identifier distributions has not been measured. And the findability of leaked identifiers under surrogate replacement has been measured for clinical notes with human readers and with one machine-learned parrot attack, not for conversational text or for LLM-assisted attackers.

17. Limitations

In this analysis the author re-reads the published tables of the anchor study and of the cited literature and derives the document-level relation in Section 8, which is applied under an independence assumption stated where it is used. The clinical literature is over-represented because it is where most controlled comparisons exist; general PII detection in customer, legal and code text has fewer such studies, and the paper's provenance and scoring conclusions are extrapolated to it. Several cited 2025 and 2026 sources are preprints, and some report single runs. The four-index frame is a stipulation for organising reported numbers, not an empirical finding; another partition, for example one separating annotation quality from taxonomy, would be defensible. The literature on re-identification risk and formal privacy was sampled for its bearing on detector evaluation rather than surveyed; Lison et al. (2021) and Deusser et al. (2025) cover anonymisation more fully. Finally, Algorithm 1's sample-size arithmetic follows from the rule of three (Hanley and Lippman-Hand, 1983); its practical cost on a deployment stream is an open question that the fourth study of Section 15 would answer.

18. Conclusion

A PII detector's score and an organisation's release risk are linked by a chain whose first link is the only one most benchmarks measure. Between them sit four choices that every reported number makes silently: which text it was tested on relative to its development data, which rule counts a hit, which spans count as personal at all, and what unit the misses are averaged over. Those choices account for much of the spread of published numbers, from above 0.99 to below 0.14, and alone move one system's miss rate by a factor of 2.5. When they are held fixed, a real shift effect remains, of roughly 7 to 22 points of F1 in the guideline- and rule-matched clinical comparisons, concentrated in names, places and institution-specific entities. It is the effect the anchor's stress test detects. Beyond the item, document-level compounding raises exposure well above the item miss rate, while surrogate replacement and the low identifying power of many surviving items lower the harm. Neither the near-human record nor the robustness critique has measured the product. An organisation that wants to know how much its pipeline leaks should fix its own taxonomy, score coverage under its own release action, count exposures per document on its own stream, and audit enough documents for the bound it intends to claim.

References

Acharya, P., & Shrestha, L. (2025). A Comparative Study of Light-weight Language Models for PII Masking and their Deployment for Real Conversational Texts. arXiv:2512.18608. doi:10.48550/arXiv.2512.18608.

Aghakasiri, K., Zambare, N., Thai, J., Ye, C., Mehta, M., Mitchell, J. R., & Abdalla, M. (2025). Not What the Doctor Ordered: Surveying LLM-based De-identification and Quantifying Clinical Information Loss. EMNLP 2025. arXiv:2509.14464. doi:10.48550/arXiv.2509.14464.

Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., & Hirschman, L. (2013). 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). doi:10.1136/amiajnl-2012-001034.

Carrell, D. S., Cronkite, D. J., Li, M. R., Nyemba, S., Malin, B. A., Aberdeen, J. S., & Hirschman, L. (2019). 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. doi:10.1093/jamia/ocz114.

Carrell, D. S., Malin, B. A., Cronkite, D. J., Aberdeen, J. S., Clark, C., Li, M. R., Bastakoty, D., Nyemba, S., & Hirschman, L. (2020). Resilience of clinical text de-identified with "hiding in plain sight" to hostile reidentification attacks by human readers. Journal of the American Medical Informatics Association, 27(9). doi:10.1093/jamia/ocaa095.

de Bourcy, C., Ghalebikesabi, S., Schwarzschild, A., Gorbachev, A., Maruseac, M., Chu, A., et al. (2026). Model Card for OpenAI Privacy Filter. arXiv:2608.18274. doi:10.48550/arXiv.2608.18274.

Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Marsolo, K., Jegga, A., Kaiser, M., Stoutenborough, L., & Solti, I. (2013). Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. Journal of the American Medical Informatics Association, 20(1). doi:10.1136/amiajnl-2012-001012.

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2016). De-identification of Patient Notes with Recurrent Neural Networks. arXiv:1606.03475. doi:10.48550/arXiv.1606.03475.

Deusser, T., Sparrenberg, L., Berger, A., Hahnbueck, M., Bauckhage, C., & Sifa, R. (2025). A Survey on Current Trends and Recent Advances in Text Anonymization. IEEE DSAA 2025. arXiv:2508.21587. doi:10.48550/arXiv.2508.21587.

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. PLoS ONE, 6(12), e28071. doi:10.1371/journal.pone.0028071.

Fu, J., Liu, P., Zhang, Q., & Huang, X. (2020). Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study. AAAI 2020. arXiv:2001.03844. doi:10.48550/arXiv.2001.03844.

Hanley, J. A., & Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA, 249(13). doi:10.1001/jama.1983.03330370053031.

Hsu, A., & Berlin, K. (2026). Broad and Efficient PII Detection with Unified Labels. arXiv:2610.22597. doi:10.48550/arXiv.2610.22597.

Jha, P. (2026). PIIBench: A Unified Multi-Source Benchmark Corpus for Personally Identifiable Information Detection. arXiv:2604.15776. doi:10.48550/arXiv.2604.15776.

Johnson, A. E. W., Bulgarelli, L., & Pollard, T. J. (2020). Deidentification of free-text medical records using pre-trained bidirectional transformers. Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL 2020). doi:10.1145/3368555.3384455.

Kayaalp, M., Browne, A. C., Callaghan, F. M., Dodd, Z. A., Divita, G., Ozturk, S., & McDonald, C. J. (2014). The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. Journal of the American Medical Informatics Association, 21(3). doi:10.1136/amiajnl-2013-001689.

Kocaman, V., Haq, H. U., & Talby, D. (2023). Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets. ML4H 2023. arXiv:2312.08495. doi:10.48550/arXiv.2312.08495.

Krco, N., Yao, Z., Meeus, M., & de Montjoye, Y.-A. (2026). RAT-Bench: A Comprehensive Benchmark for Text Anonymization. arXiv:2602.12806. doi:10.48550/arXiv.2602.12806.

Kumar, V. P. V., Venkatesh, S., & Yamshchikov, I. P. (2026). LeakageBench: Document-Level Leakage Risk for Redacting Personally Identifiable Information in Document Images. arXiv:2609.02207. doi:10.48550/arXiv.2609.02207.

Lee, G., Cho, G., Kim, J., Kim, H., So, Y., Jeon, H., Park, H. M., Kim, S., Park, Y., Pyo, Y., & Lee, J. (2026). Thunder-DeID+: Boundary-Aware Korean De-identification from Court Judgments to Broader Sensitive Documents. arXiv:2610.22733. doi:10.48550/arXiv.2610.22733.

Lee, K., Dobbins, N. J., McInnes, B., Yetisgen, M., & Uzuner, O. (2021). Transferability of Neural Network Clinical De-identification Systems. arXiv:2102.08517. doi:10.48550/arXiv.2102.08517.

Lin, B. Y., Gao, W., Yan, J., Moreno, R., & Ren, X. (2021). RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models. EMNLP 2021. arXiv:2109.05620. doi:10.48550/arXiv.2109.05620.

Lison, P., Pilan, I., Sanchez, D., Batet, M., & Ovrelid, L. (2021). Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. Proceedings of ACL-IJCNLP 2021. doi:10.18653/v1/2021.acl-long.323.

Liu, Z., Huang, Y., Yu, X., Zhang, L., Wu, Z., Cao, C., et al. (2023). DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. arXiv:2303.11032. doi:10.48550/arXiv.2303.11032.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Beguelin, S. (2023). Analyzing Leakage of Personally Identifiable Information in Language Models. IEEE S&P 2023. arXiv:2302.00539. doi:10.48550/arXiv.2302.00539.

Mansfield, C., Paullada, A., & Howell, K. (2022). Behind the Mask: Demographic bias in name detection for PII masking. arXiv:2205.04505. doi:10.48550/arXiv.2205.04505.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology, 10, 70. doi:10.1186/1471-2288-10-70.

Mishra, S., He, S., & Belli, L. (2020). Assessing Demographic Bias in Named Entity Recognition. arXiv:2008.03415. doi:10.48550/arXiv.2008.03415.

Neamatullah, I., Douglass, M. M., Lehman, L. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., & Clifford, G. D. (2008). Automated de-identification of free-text medical records. BMC Medical Informatics and Decision Making, 8, 32. doi:10.1186/1472-6947-8-32.

Norgeot, B., Muenzen, K., Peterson, T. A., Fan, X., Glicksberg, B. S., Schenk, G., et al. (2020). Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. npj Digital Medicine, 3, 57. doi:10.1038/s41746-020-0258-y.

Oh, M. S., Kim, D.-Y., Oh, H., Kang, C., Kang, J., Wang, X., Park, H., Jung, Y. C., & Kim, H. (2026). Subject-level Inference for Realistic Text Anonymization Evaluation. arXiv:2604.21211. doi:10.48550/arXiv.2604.21211.

Pilan, I., Lison, P., Ovrelid, L., Papadopoulou, A., Sanchez, D., & Batet, M. (2022). The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. arXiv:2202.00443. doi:10.48550/arXiv.2202.00443.

Posada, J. D., Love, D., Datta, S., & Desai, P. (2026). SHIELD: A Diverse Clinical Note Dataset and Distilled Small Language Models for Enterprise-Scale De-identification. arXiv:2605.03301. doi:10.48550/arXiv.2605.03301.

Prakash, E., Attias, M., Chambon, P., Xu, J., Truong, S., Delbrouck, J.-B., Cook, T., & Langlotz, C. (2025). Improving the Performance of Radiology Report De-identification with Large-Scale Training and Benchmarking Against Cloud Vendor Methods. arXiv:2511.04079. doi:10.48550/arXiv.2511.04079.

Rajgarhia, H., Gupta, S., Shaik, A., Kumar, G. P., Santhoshraj, Y., Nishitha, S. N. T., & Mukherji, A. (2025). An Evaluation Study of Hybrid Methods for Multilingual PII Detection. arXiv:2510.07551. doi:10.48550/arXiv.2510.07551.

Sarkar, A. R., Chuang, Y.-S., Mohammed, N., & Jiang, X. (2024). De-identification is not enough: a comparison between de-identified and synthetic clinical notes. arXiv:2402.00179. doi:10.48550/arXiv.2402.00179.

Singh, D., & Narayanan, S. (2025). Unmasking the Reality of PII Masking Models: Performance Gaps and the Call for Accountability. arXiv:2504.12308. doi:10.48550/

arXiv.2504.12308.

Steinkamp, J. M., Pomeranz, T., Adleberg, J., Kahn, C. E., & Cook, T. S. (2020). Evaluation of Automated Public De-Identification Tools on a Corpus of Radiology Reports. Radiology: Artificial Intelligence, 2(6). doi:10.1148/ryai.2020190137.

Stubbs, A., Kotfila, C., & Uzuner, O. (2015a). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. Journal of Biomedical Informatics, 58 Suppl. doi:10.1016/j.jbi.2015.06.007.

Stubbs, A., & Uzuner, O. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of Biomedical Informatics, 58 Suppl. doi:10.1016/j.jbi.2015.07.020.

Stubbs, A., Filannino, M., & Uzuner, O. (2017). De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. Journal of Biomedical Informatics, 75S. doi:10.1016/j.jbi.2017.06.011.

Taille, B., Guigue, V., & Gallinari, P. (2020). Contextualized Embeddings in Named-Entity Recognition: An Empirical Study on Generalization. ECIR 2020. arXiv:2001.08053. doi:10.48550/arXiv.2001.08053.

Trienes, J., Trieschnigg, D., Seifert, C., & Hiemstra, D. (2020). Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records. arXiv:2001.05714. doi:10.48550/arXiv.2001.05714.

Uppala, R. (2026). OpenAI Privacy Filter: A Cross-Lingual, Cross-Domain PII Evaluation Across 32 Benchmarks. arXiv:2608.02616. doi:10.48550/arXiv.2608.02616.

Uzuner, O., Luo, Y., & Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association, 14(5). doi:10.1197/jamia.M2444.

Xiao, Y., Lim, S., Pollard, T. J., & Ghassemi, M. (2023). In the Name of Fairness: Assessing the Bias in Clinical Record De-identification. FAccT 2023. arXiv:2305.11348. doi:10.48550/arXiv.2305.11348.

Yang, H., & Garibaldi, J. M. (2015). Automatic detection of protected health information from clinic narratives. Journal of Biomedical Informatics, 58 Suppl. doi:10.1016/j.jbi.2015.06.015.

Yang, X., Lyu, T., Li, Q., Lee, C.-Y., Bian, J., Hogan, W. R., & Wu, Y. (2019). A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Medical Informatics and Decision Making, 19(Suppl 5), 232. doi:10.1186/s12911-019-0935-4.

Zafar, A., & Nowaczyk, S. (2026). Mind the Gap: Robustness Risks in PII Detection Systems. arXiv:2609.03464. doi:10.48550/arXiv.2609.03464.

Zhang, K., Wang, C., Zhong, J., Fryzel, P., Polley, K., Ma, J., & Li, N. (2026). PII-TRACE: A Benchmark for Context-Aware PII Detection in Multi-Turn LLM Conversations. arXiv:2609.22200. doi:10.48550/arXiv.2609.22200.

Zhou, Z., Vanacore, K., Ahtisham, B., Lee, J., Pietrzak, D., Hedley, D., Dias, J., Shaw, C., Schaefer, R., & Kizilcec, R. F. (2026). Utility-Preserving De-Identification for Math Tutoring: Investigating Numeric Ambiguity in the MathEd-PII Benchmark Dataset. arXiv:2602.16571. doi:10.48550/arXiv.2602.16571.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23849157