Local PHI Masking in Clinical AI Tools: A Literature Review

Prepared by Perth AI Consulting using Perplexity AI

Version 1.0 (fact-checked) · 2026-05-16

Perth AI Consulting

perthaiconsulting.com.au

> Editor's note: This paper was drafted with AI assistance (Perplexity), fact-checked against every cited source via independent web verification, and corrected before publication. Verification methodology, correction log, limitations, and iteration history are documented in the appendices.

Abstract

Local, on-device, or client-side masking of protected health information (PHI) has become an increasingly important design pattern in clinical AI systems because it changes the privacy boundary of note processing before text reaches downstream models or cloud services. The core promise is architectural rather than merely algorithmic: if PHI can be detected and transformed on the originating device, the attack surface, data-sharing burden, and some compliance risks may be reduced relative to server-side preprocessing, although the masking model itself still becomes part of the regulated system boundary (Uzuner et al., 2007; Microsoft Presidio documentation, n.d.).

The literature shows that local PHI masking is not a single method but a stack of techniques with different error profiles. Across the historical and current literature, the field has progressed from rule-heavy and statistical systems toward deep-learning named entity recognition (NER) and hybrid pipelines, yet the strongest practical systems still retain pattern matching, dictionaries, and document-structure heuristics because many PHI classes remain semi-structured, sparse, or context dependent (Uzuner et al., 2007; Dehghan et al., 2015).

This review examines eight questions relevant to deploying PHI masking in local clinical AI tools: the technical landscape, locally runnable models and libraries, benchmark accuracy, documented failure modes, evidence for layered systems, local-versus-server trade-offs, downstream fidelity loss, and evaluation methodology. The review prioritises retrievable peer-reviewed literature and explicitly distinguishes vendor or product documentation from peer-reviewed findings when vendor material is used to characterise implementation options rather than to substantiate performance claims (Uzuner et al., 2007; Dehghan et al., 2015; Microsoft Presidio documentation, n.d.).

Introduction

Automated de-identification and PHI masking sit at the intersection of privacy engineering and clinical NLP. In the clinical-text literature, de-identification is usually framed as the identification and removal or transformation of PHI spans in free text while preserving as much medically useful content as possible, a tension that has been explicit since early shared-task work on discharge summaries and remains central for modern AI pipelines that feed summarisation, coding, retrieval, and documentation support systems (Uzuner et al., 2007).

The distinctive question for contemporary product design is not only whether PHI can be detected accurately, but whether detection and transformation can occur locally inside a browser tab, desktop application, mobile device, or edge runtime before any raw note text is transmitted. That question matters because the system boundary determines where plaintext PHI exists, which processors see it, what logs can capture it, and which deployment artefacts must be governed as safety- and privacy-relevant components (Microsoft Presidio documentation, n.d.).

A literature review on this topic must therefore bridge two domains that are often discussed separately: the academic de-identification literature and the practical runtime constraints of on-device inference. The former is rich on methods and benchmarks, while the latter is documented more often in software documentation, model cards, and engineering repositories than in formal clinical informatics journals; accordingly, some implementation claims in this review are necessarily sourced from retrievable product or project documentation rather than peer-reviewed evaluations, and these are identified as such inline.

A brief methodology note is warranted. The source set for this review was assembled on May 14, 2026 using retrievable web-accessible academic and technical sources surfaced through searches focused on clinical de-identification, i2b2 2014, MIMIC-linked de-identification, hybrid de-identification methods, Presidio, browser-side inference, and transformer-based clinical note de-identification. Peer-reviewed or academically hosted sources were prioritised wherever available, while vendor documentation and engineering articles were used only to characterise implementation options when peer-reviewed local-deployment evidence was sparse.

1. Approaches to Local PHI Masking

1.1 Rule-based and dictionary approaches

Rule-based de-identification systems use explicit pattern definitions, lexical cues, regular expressions, and curated lists to identify PHI. Early challenge systems performed strongly on highly regular PHI classes such as dates, phone numbers, IDs, ages, postcodes, and some location patterns because these categories have stable surface forms that are amenable to templates and deterministic matching rather than statistical inference (Uzuner et al., 2007; Dehghan et al., 2015).

The continuing importance of rule-based logic is not merely historical. Dehghan et al. (2015) show that rules alone remained the chosen mechanism for several highly structured classes in a top-performing 2014 i2b2 system, including age, street, zip, email, fax, phone, username, identification number, and medical record fields, while machine learning and dictionaries were reserved for more context-dependent entities such as city, hospital, organisation, profession, date, and personal names.

For local deployment, rule-based components have several practical advantages. They are small, interpretable, deterministic, and easy to execute in constrained environments such as browser JavaScript or WebAssembly without loading a large neural model, making them particularly suitable as a first-pass filter or structured pre-pass over forms, headers, and conventional formatting zones (Dehghan et al., 2015; Microsoft Presidio documentation, n.d.).

The main limitation is incomplete coverage. Uzuner et al. (2007) designed the i2b2 challenge corpus specifically to stress non-dictionary-based methods by injecting out-of-vocabulary surrogates and ambiguous PHI terms, showing why pure gazetteer systems are brittle in the face of rare names, lexical novelty, and overlap between PHI and medical terminology.

1.2 Statistical sequence models

Before transformer-based NER, clinical de-identification commonly used statistical sequence models such as conditional random fields (CRFs), hidden Markov models, support vector machines with chunking logic, and decision-tree ensembles. These systems model PHI extraction as token tagging or sequence labelling and can use orthographic, lexical, positional, and contextual features to generalise beyond fixed

patterns (Uzuner et al., 2007).

The 2006 and 2014 i2b2 challenge literature established a stable pattern: statistical methods generally outperform pure rules on ambiguous and context-sensitive PHI, especially when enriched with rule-derived features or post-processing. In their synthesis of the 2006 challenge, Uzuner et al. (2007) concluded that the best-performing systems combined machine learning with task-specific features and regular-expression handling rather than relying on either generic NER or pure rule sets.

CRF-era systems remain relevant to a local-deployment discussion for two reasons. First, they demonstrate that a useful de-identification system need not depend on a very large model. Second, their feature engineering highlights a broader architectural lesson for client-side tools: document position, formatting, headings, and repeated entities within a patient record can be harvested locally as strong signals before any heavier model is invoked (Uzuner et al., 2007; Dehghan et al., 2015).

1.3 Deep-learning NER

Deep-learning systems reframed clinical de-identification as representation learning over token sequences rather than manual feature engineering. The clearest benchmark shift in the retrieved literature is represented by recurrent and transformer-based models trained on clinical de-identification corpora, which substantially improved F1 on established datasets relative to earlier CRF-era baselines (Dernoncourt et al., 2017; Moore et al., 2023).

Dernoncourt et al. (2017) report that a recurrent neural-network de-identification model achieved F1 97.85 on i2b2 2014, with recall 97.38 and precision 98.32, and F1 99.23 on a MIMIC de-identification dataset, with recall 99.25 and precision 99.21. These figures illustrate that, on curated benchmarks, modern learned sequence models can reach very high span-detection performance, but they do not eliminate the need to reason about benchmark design, corpus bias, and out-of-domain drift.

Transformer-based approaches extend this trajectory by fine-tuning BERT-family encoders for PHI extraction. The PhysioNet Transformer-DeID resource (Moore et al., 2023) describes BERT, DistilBERT, and RoBERTa models fine-tuned for the removal of patient identifiers in clinical text through the Hugging Face transformers stack, indicating a practical route for compact local NER where a distilled model is preferred over a larger encoder for runtime reasons; however, this resource is project documentation rather than a standalone peer-reviewed deployment study

and should be read as evidence of technical feasibility, not as definitive proof of browser-grade performance.

1.4 Hybrid and layered systems

The strongest pattern across the literature is that de-identification systems are usually hybrid, whether or not authors foreground that label. Dehghan et al. (2015) explicitly describe a pipeline combining dictionaries, rules, CRF models, a second-pass patient-specific dictionary, and an integration module with priority handling for overlaps, and they attribute measurable gains to both two-pass recognition and the combination of knowledge-driven and data-driven components.

This matters for local masking because an on-device system is often budget-constrained. A layered pipeline lets the device reserve the heavy model for uncertain text while deterministic recognisers handle cheap, high-confidence categories. The same architecture also improves debuggability, because false positives and false negatives can often be traced to a specific stage rather than treated as opaque model behaviour (Dehghan et al., 2015; Microsoft Presidio documentation, n.d.).

1.5 Structured pre-pass walking

A distinct but under-emphasised technique in the de-identification literature is structured pre-pass walking: traversing document zones, headers, forms, semi-structured templates, and metadata before running general free-text NER. This idea appears repeatedly in older systems through sentence-position features, heading-aware models, and the use of "trusted" PHI found in structured header material to propagate labels into narrative text (Uzuner et al., 2007).

For local clinical AI tools, structured pre-pass walking is especially valuable because many real notes are partly templated even when they appear unstructured at the record level. A local system that first normalises line prefixes, field labels, timestamps, signature blocks, patient banners, and imported OCR layout can reduce the search space presented to a neural detector and improve recall on repeated entities that later appear in fragmented form, an effect conceptually consistent with the patient-level two-pass strategy reported by Dehghan et al. (2015).

2. Practical Local Models and Libraries

2.1 Presidio and recogniser frameworks

Microsoft Presidio is one of the most practical open-source frameworks for local de-identification because it is designed as a configurable recogniser framework rather than a single model. According to Microsoft's documentation, Presidio supports predefined or custom recognisers using NER, regular expressions, rule logic, and checksum validation across text, images, and structured data, and it is explicitly presented as configurable rather than guaranteed-complete, with the documentation cautioning that additional protections are still required (Microsoft Presidio documentation, n.d.).

For a local clinical tool, Presidio's significance lies less in out-of-the-box clinical accuracy than in architectural flexibility. It can host regex-heavy recognisers, custom dictionaries, and external NER models, which makes it suitable as the orchestration layer around a hybrid local masking pipeline; however, the documentation retrieved does not establish native browser or WebAssembly deployment, so browser-side use would ordinarily require either porting recogniser logic to JavaScript/WASM or exposing Presidio through a local process rather than running it directly inside the browser sandbox (Microsoft Presidio documentation, n.d.).

The academic literature also shows that framework flexibility matters because no single recogniser family covers all PHI classes equally well. Rule-driven components are often best for highly regular entities, while model-backed recognisers are more useful for person names, hospitals, organisations, and professions, which is exactly the split a framework like Presidio is designed to accommodate (Dehghan et al., 2015; Microsoft Presidio documentation, n.d.).

2.2 BERT-family and clinical fine-tunes

The user-specified examples of bert-base-NER, BioBERT, and ClinicalBERT point to a practical question: which encoders are small enough and domain-specific enough to run locally? The retrieved sources confirm the clinical de-identification use of BERT-family encoders, including BERT, RoBERTa, and DistilBERT in the PhysioNet Transformer-DeID resource (Moore et al., 2023), but the present source set does not provide a retrievable peer-reviewed benchmark directly comparing generic

bert-base-NER against clinical PHI-specific fine-tunes in an on-device setting.

That gap should be stated explicitly rather than papered over with weak evidence. At present, the literature retrieved for this review supports the feasibility of local transformer-based PHI masking and the practical relevance of compact models, but it does not support a strong comparative claim about the local-runtime trade-off between generic bert-base-NER and clinical fine-tunes such as BioBERT or ClinicalBERT. This is better framed as a future-work question than as a conclusion (Moore et al., 2023; Dernoncourt et al., 2017).

Even so, the architecture implications are clear. Distilled encoder variants are more plausible for browser or mobile inference than full-size clinical transformers because local masking is latency-sensitive and often memory-constrained, while domain-adapted clinical encoders are likely to outperform general-purpose NER on abbreviations, note syntax, and specialised entity contexts if they have been fine-tuned on de-identification tasks rather than biomedical concept extraction generally (Moore et al., 2023; Dernoncourt et al., 2017).

The important caveat is that clinical domain adaptation does not automatically mean PHI-specific competence. A model trained for biomedical entity recognition is not equivalent to a model trained to discriminate PHI from clinically meaningful text, and the i2b2 literature repeatedly shows that ambiguity between names and medical terms is one of the central failure modes of the task (Uzuner et al., 2007).

2.3 Browser-side inference and WebAssembly-like deployment

Direct evidence on browser-side PHI masking is sparser than benchmark literature, but the retrieved sources show enough to support a cautious feasibility claim. A 2026 engineering article by Agbo describes a browser pipeline using Transformers.js for NER to strip personally identifying information locally before passing text to a browser-resident language model, which is conceptually aligned with the client-side architecture under review; however, this is industry engineering commentary rather than peer-reviewed evidence and should be treated as a practical example, not an evaluation study (Agbo, 2026).

A more academically grounded signal of local feasibility comes from the existence of compact transformer-based de-identification resources such as DistilBERT variants in the PhysioNet Transformer-DeID package. Running such a model in a browser would generally require conversion to a JavaScript-compatible or WebAssembly/WebGPU-compatible format and aggressive management of model size, sequence length, and memory allocation, but the technical barrier is

engineering adaptation rather than theoretical impossibility (Moore et al., 2023).

The literature retrieved for this review does not provide a peer-reviewed clinical study specifically validating WebAssembly deployment of PHI masking models in production browsers. That absence is itself important: browser-local PHI masking is technologically plausible and increasingly discussed in engineering circles, but its clinical literature base remains less mature than the literature on server-side or offline batch de-identification (Agbo, 2026; Moore et al., 2023).

2.4 Commercial and proprietary libraries

Commercial offerings such as John Snow Labs Healthcare NLP provide clinically targeted de-identification pipelines and pretrained NER components. John Snow Labs' public model documentation describes a de-identification pipeline composed of clinical embeddings and specialised NER models, but this retrieved material is vendor documentation and not a peer-reviewed comparative trial, so it can support a description of available tooling rather than an academic claim that the pipeline is superior to alternatives (John Snow Labs, 2024).

The same caution applies to commercial comparison articles. A John Snow Labs comparison page authored by the company's CEO positions its product against Microsoft Presidio, but because it is vendor-authored, any performance or superiority claims from that source should not be treated as equivalent to shared-task or journal evidence (Talby, 2025).

From a local-deployment perspective, commercial healthcare NLP stacks may be practical on controlled endpoints such as hospital-managed workstations or local servers, but they are less obviously suited to pure browser-only delivery because they often assume a JVM, Python, or server-process runtime. The browser-specific implementation path is therefore currently clearer for lighter open models and custom JavaScript/WASM integration than for heavyweight proprietary stacks, although the latter may still underpin "local" deployments in the broader sense of on-premise or endpoint-resident processing (Talby, 2025; John Snow Labs, 2024).

3. Accuracy on Benchmarks

3.1 i2b2 2006 and 2014 as anchor benchmarks

The i2b2 shared tasks remain the central public benchmarks in clinical text de-identification. Uzuner et al. (2007) report that in the 2006 challenge, the best systems scored above 98% F-measure across PHI categories under the challenge evaluation framework, but they also emphasise that identifying ambiguous PHI remained difficult and that future work needed more heterogeneous datasets.

The 2014 i2b2/UTHealth task expanded the problem to longitudinal narratives covering a broader set of PHI entity types than the earlier discharge-summary benchmark, making it more representative of realistic PHI complexity. Dehghan et al. (2015) report micro-averaged performance on the 2014 test set of 90.65% strict-text F1, 93.06% precision, and 88.36% recall for their best hybrid submission, alongside 94.80% token-level F1 and 93.23% HIPAA strict F1.

These figures should not be compared naively across papers because the task definitions, matching criteria, and PHI label sets differ. Still, they show a broad historical pattern: once the benchmark moved to more entity types and stricter span criteria, performance remained strong but no longer sat uniformly near ceiling, especially for low-frequency and context-sensitive classes such as profession and organisation (Dehghan et al., 2015).

3.2 MIMIC-linked evidence

The retrieved literature provides one strong MIMIC-linked result in Dernoncourt et al.'s recurrent neural-network system. That model achieved F1 99.23 on a MIMIC de-identification dataset, with recall 99.25 and precision 99.21, compared with F1 97.85 on i2b2 2014, indicating that benchmark difficulty and corpus construction materially affect apparent performance (Dernoncourt et al., 2017).

This difference is analytically important. A local PHI masking system should not be judged by a single aggregate F1 reported on whichever dataset yields the highest number, because MIMIC-derived datasets may differ from i2b2 in note style, PHI distribution, surrogate-generation procedure, and prevalence of templated identifiers; high scores on one corpus do not guarantee equal safety on another (Dernoncourt et al., 2017; Uzuner et al., 2007).

The MIMIC project itself is not a de-identification benchmark paper, but its documentation underscores that the dataset is already de-identified and widely used for clinical NLP. That matters because a masking system evaluated only on previously de-identified or surrogate-generated corpora may face hidden generalisation issues when confronted with live institutional text containing local abbreviations, mixed templates, and OCR artefacts not captured in benchmark notes (MIMIC-III Clinical Database, 2016; Uzuner et al., 2007).

3.3 Per-category variation matters more than aggregate F1

The strongest practical lesson from the benchmark literature is that aggregate F1 conceals high-risk class variation. In Dehghan et al. (2015), per-class strict-text F1 figures show pronounced spread: highly structured classes such as dates and zip codes achieved high F1, while context-sensitive categories such as organisation and profession lagged substantially. (Specific per-class numerical figures from Dehghan et al. are pending verification against the source paper — see Appendix C.)

For local PHI masking, this matters more than leaderboard rank. If the system's primary function is to protect raw note text before it leaves the endpoint, a seemingly impressive micro-average can still be operationally unsafe if the residual leak risk clusters in rare names, institution names, or role descriptors that are uncommon in the benchmark but highly identifying in a particular clinic (Dehghan et al., 2015; Uzuner et al., 2007).

3.4 Human comparability claims require caution

Dehghan et al. include a highlight claiming that automated de-identification is comparable to human benchmark performance, but the retrieved article section does not provide a full standalone methodological demonstration of that equivalence in the text captured here. Accordingly, strong claims that a local masking pipeline is "human-level" should be treated cautiously unless the supporting study clearly defines annotator agreement, adjudication, and identical evaluation conditions (Dehghan et al., 2015).

The wider de-identification literature also shows that annotation itself is complex. Uzuner et al. (2007) describe a multi-pass annotation methodology with discussion-based finalisation for their challenge corpus, reflecting the reality that human gold standards are constructed through consensus and may contain residual inconsistency, which limits simplistic appeals to "human performance" as a single stable reference point.

4. Documented Failure Modes

4.1 Ambiguity and lexical overlap

Ambiguity between PHI and medically meaningful text is one of the oldest and most persistent documented failure modes. Uzuner et al. (2007) deliberately injected surrogate names that overlapped with diseases, treatments, and test names in order to test whether systems could preserve clinically important text while still removing PHI, and they found ambiguous PHI substantially more difficult than out-of-vocabulary PHI in general.

This failure mode directly maps to local masking in AI tools. Over-redaction removes informative content and harms downstream synthesis, while under-redaction leaves residual identifiers; ambiguity therefore produces both privacy failures and fidelity failures at once (Uzuner et al., 2007).

4.2 Rare names, professions, and role labels

Dehghan et al. (2015) identify professions and organisations as especially difficult because they are broad, context dependent, and infrequent in the training data. Their error analysis notes examples where role-like language or occupational terms created both false negatives and false positives, and their per-class results show profession and organisation as among the weakest categories.

This is directly relevant to the problem of ambiguous role titles versus surnames. A label such as "Baker," "Porter," or "Justice" can function as a surname, a profession, or part of a non-PHI phrase depending on context, and the literature supports the broader point that context-poor role expressions are among the least stable PHI categories for automated masking, especially when training data are sparse (Dehghan et al., 2015; Uzuner et al., 2007).

4.3 Fragmentation, boundary errors, and propagation gaps

PHI detectors frequently identify part of an entity span but not the entire mention. Dehghan et al. (2015) explicitly attribute a substantial portion of their strict-matching drop to token-level CRF behaviour that captured only subsets of multi-token entities such as doctor names, patient names, city names, and professions, demonstrating that "almost right" boundaries still count as privacy failures under strict evaluation.

Propagation gaps are a related failure mode in longitudinal or repeated text. Because the same patient, clinician, location, or identifier may recur in multiple forms across a note set, systems that rely only on local context can miss later mentions that lack explicit cues. Dehghan et al.'s patient-specific two-pass dictionary was proposed precisely to address these residual mentions, with the authors reporting measurable F1 and recall improvements for patient entities. (Specific numerical gain figures are pending verification — see Appendix C.)

4.4 Data quality noise and OCR-like corruption

The literature clearly documents that poor tokenisation and malformed text damage de-identification performance. In their error analysis, Dehghan et al. (2015) point to missing spaces and concatenated strings such as identifiers glued to hospital abbreviations or personal names, producing both false negatives and false positives because sequence models and rules assumed cleaner boundaries than the text actually contained.

This finding generalises naturally to OCR-derived noise even though the retrieved peer-reviewed sources do not provide a dedicated OCR benchmark for PHI masking. If missing spaces and corrupted token boundaries already hurt performance on textual corpora, OCR artefacts such as split characters, merged lines, and confidence errors should be expected to worsen both recall and boundary accuracy unless an OCR-normalisation layer precedes masking (Dehghan et al., 2015).

4.5 Multilingual and bilingual text

The retrieved literature includes a bilingual clinical-text de-identification study from Korea (Asan Medical Center, Seoul) by Shin et al. (2015) using regular expressions on Korean-English mixed clinical notes. The system achieved high precision and recall on the development dataset, with the authors reporting 99.1% precision, 98.7% recall, and 99.0% F0.5 on 6,039 development clinical notes of 20 types, while noting that some names remained unmasked. This illustrates that multilingual or code-switched text can still expose weaknesses even when pattern-based identifiers dominate the corpus.

This source is useful because it shows that multilinguality is not a purely modern transformer problem. Local masking systems deployed in multilingual clinical environments should not assume that a model or regex set tuned on English i2b2

notes will transfer cleanly to mixed-language records, especially when names, addresses, and contact conventions vary by language and script (Shin et al., 2015; Uzuner et al., 2007).

4.6 Drift and corpus mismatch

The benchmark literature repeatedly warns against over-generalising from a single corpus. Uzuner et al. (2007) argued that future evaluations needed larger and more heterogeneous datasets, and that warning remains salient because a local masking system may confront note genres, abbreviations, and institutional styles absent from training data.

Model drift in this context is not only temporal retraining drift but also operational mismatch between benchmark assumptions and production documents. A detector trained on curated discharge summaries or surrogate-heavy challenge corpora may degrade when exposed to telehealth chat fragments, copied referral text, OCR'd scanned letters, or terse point-form notes, all of which shift tokenisation, context length, and PHI distribution. The retrieved sources support this concern conceptually, even when they do not quantify every deployment scenario (Uzuner et al., 2007; Dehghan et al., 2015).

5. Evidence for Layered Approaches

5.1 Hybrid systems outperform single-family systems

The clearest direct evidence favouring layered approaches comes from Dehghan et al. (2015). Their system combined rules, dictionaries, CRF models, post-processing, second-pass patient-specific dictionaries, and overlap-resolution logic, and the authors report notable gains from integrating knowledge-driven and data-driven methods, including F1 improvements for patient, city, hospital, date, organisation, and profession categories relative to single-component alternatives explored during development.

The 2006 challenge synthesis points in the same direction. Uzuner et al. (2007) describe the best systems as those that paired statistical learning with task-specific patterns and domain features, while pure rule-based systems were generally weaker overall even though they remained useful for regularised categories.

5.2 Marginal gains from second-pass recognition

Layering is not only about mixing rules and NER; ordering and re-use of extracted entities also matter. Dehghan et al.'s two-pass recognition method built a temporary patient-specific dictionary from high-confidence first-pass detections and then re-ran matching to catch residual mentions without local cues, yielding gains across multiple entity types and especially notable improvements for patient entities.

This is strong evidence that propagation-aware ordering matters. In practical local systems, the best sequence is often not "one model over raw text" but rather structured pre-pass, deterministic patterns, first-pass NER, cross-document or intra-note propagation, then conflict resolution and transformation (Dehghan et al., 2015; Uzuner et al., 2007).

5.3 Why ordering matters in local systems

Ordering matters because each layer changes the evidence available to later layers. If a regex stage first captures dates, identifiers, phone numbers, and obvious template fields, the NER model sees a simpler residual problem focused on ambiguous names and institutions. Conversely, if a neural recogniser runs first on noisy raw text, it may waste capacity on classes that could have been

deterministically removed and may create more overlap conflicts downstream. This reasoning is consistent with the architecture choices in both the i2b2-era systems and modern configurable recogniser frameworks such as Presidio (Uzuner et al., 2007; Dehghan et al., 2015; Microsoft Presidio documentation, n.d.).

There is, however, no single universally optimal order demonstrated by the retrieved literature across all corpora. The evidence supports the superiority of layered systems in general more strongly than it supports a one-size-fits-all sequence, which means local deployment should empirically test ordering on the target note mix rather than inheriting a pipeline order uncritically (Dehghan et al., 2015; Uzuner et al., 2007).

6. Local Versus Server-side Masking

6.1 Privacy boundary and data minimisation

The strongest argument for local masking is architectural data minimisation. If PHI is transformed before a note leaves the client device, fewer downstream services handle plaintext identifiers, which narrows exposure through network transit, server logs, third-party processors, and cloud-side debugging surfaces. This is a system-design advantage rather than a claim that masking quality itself is automatically better (Microsoft Presidio documentation, n.d.).

At the same time, local masking does not remove risk, because the local model can still miss PHI and the client environment may itself be insecure. Microsoft's Presidio documentation explicitly warns that automated detection does not guarantee finding all sensitive information and that additional systems and protections are required, a reminder that "local" is not equivalent to "solved."

6.2 Latency, memory, and browser constraints

Local masking can reduce network round trips because preprocessing happens where text is generated. However, that latency advantage competes with runtime constraints: browser tabs and lightweight devices have limited memory for model weights, restricted CPU budgets, and sometimes inconsistent access to hardware acceleration, which makes compact or distilled models more realistic than large clinical transformers for purely client-side use (Moore et al., 2023).

This is why hybrid local architectures are attractive. Deterministic recognisers can cheaply handle obvious PHI, while a smaller model is reserved for the hard residue; the resulting system is more likely to fit browser or desktop constraints than a monolithic deep model that attempts every entity type in one pass (Dehghan et al., 2015; Moore et al., 2023).

6.3 Auditability and update cadence

Server-side masking is often easier to update centrally, monitor, and benchmark because one service can be patched and evaluated consistently. Local masking flips that advantage: it can improve privacy posture by keeping raw text on the endpoint, but it complicates rollout governance because multiple client versions may be in the

field simultaneously, each with potentially different recognisers, thresholds, or model weights. This trade-off follows from general software architecture and is reinforced by the de-identification literature's sensitivity to corpus mismatch and configuration details (Uzuner et al., 2007; Microsoft Presidio documentation, n.d.).

Auditability also becomes more nuanced. Rule-based local components are highly inspectable, but browser-packaged neural models may be harder for compliance teams to inventory and validate unless build provenance, model hashing, and version-locked evaluation practices are formalised as part of the product lifecycle. The retrieved academic sources do not prescribe a single governance framework for this, but they do support the underlying need for precise evaluation and configuration reporting (Uzuner et al., 2007; Dehghan et al., 2015).

6.4 Regulatory interpretation

The retrieved sources do not provide jurisdiction-specific legal advice for Australian health privacy law or detailed HIPAA deployment guidance for client-side masking. What they do support is a narrower and safer conclusion: de-identification quality is evaluated empirically, and both local and server-side systems remain automated recognisers with residual miss risk, so regulatory adequacy cannot be inferred from deployment location alone (Uzuner et al., 2007; Microsoft Presidio documentation, n.d.).

For a white paper aimed at clinical organisations, the defensible position is therefore that local masking may strengthen privacy posture by reducing PHI distribution, but it does not eliminate the need for documented validation, risk analysis, and operational safeguards around the masking system itself. That conclusion is directly consistent with both shared-task literature and framework documentation (Uzuner et al., 2007; Microsoft Presidio documentation, n.d.).

7. Fidelity Loss and Mitigation

7.1 Why masking harms downstream AI quality

The de-identification literature has always recognised that removing PHI must preserve the integrity of clinically useful text as much as possible. Uzuner et al. (2007) explicitly frame de-identification as finding and removing PHI while protecting data integrity, and their challenge design penalised systems that erased medically meaningful terms by introducing ambiguity between PHI and medical concepts.

For downstream AI tools, this trade-off becomes sharper. Generic redaction can destroy referential coherence, timeline continuity, and role attribution, all of which matter for summarisation and reasoning over longitudinal notes; even when privacy is improved, output quality may degrade if the model loses who did what, when events occurred, or whether repeated mentions refer to the same actor (Uzuner et al., 2007; Dehghan et al., 2015).

7.2 Generic redaction versus typed replacement

The retrieved benchmark papers are stronger on detection than on downstream generation quality, so the evidence base here is more limited. Even so, the literature supports an important qualitative distinction: preserving PHI type information and internal consistency is preferable to blunt deletion when the text will be used for further NLP, because type-preserving surrogates and consistent replacement retain grammatical and discourse structure better than blanking tokens entirely (Uzuner et al., 2007).

Uzuner et al.'s surrogate generation process illustrates the principle. They replaced authentic PHI with realistic, format-preserving surrogates, maintained orthographic patterns, and tried to preserve co-reference and relative time offsets across a record, precisely because simple deletion would distort the text too heavily for downstream use. This is not a downstream LLM quality study, but it is a strong methodological precedent for deterministic role-preserving masking over generic black-bar redaction (Uzuner et al., 2007).

7.3 Deterministic and role-preserving masking

A practical implication for local masking systems is that the transformation stage should be treated as a design variable, not an afterthought. Typed placeholders such as `[PATIENT_NAME]`, consistent pseudonyms, or role-preserving deterministic replacements can preserve relationships and note readability better than indiscriminate deletion, especially in clinician-authored prose where who spoke, who examined, and where care occurred are integral to interpretation (Uzuner et al., 2007).

The retrieved sources do not provide a peer-reviewed numeric comparison of generic redaction versus typed tokens versus deterministic role-preserving masking on downstream summarisation fidelity. That gap should remain explicit. At present, the strongest support available is conceptual and methodological rather than numerical, and a direct comparative fidelity study across masking regimes should be treated as a priority for future work rather than inferred from adjacent de-identification benchmarks (Uzuner et al., 2007).

7.4 Mitigating fidelity loss in local pipelines

Three mitigation strategies are well supported by the literature's underlying logic. First, use class-specific transformations rather than one universal redaction operator, because PHI categories differ in how much structure they contribute to the note. Second, preserve within-document consistency through second-pass propagation or deterministic replacement. Third, evaluate masking not only as leak prevention but also as preservation of the clinical task signal needed downstream (Uzuner et al., 2007; Dehghan et al., 2015).

These strategies align especially well with local systems because once the text is transformed on-device, the transformed output may become the only version seen by downstream AI. Any avoidable information destruction at the masking stage therefore becomes irreversible for later components (Uzuner et al., 2007; Dehghan et al., 2015).

8. Evaluation Methodology

8.1 Gold-standard annotation protocols

A valid PHI masking evaluation begins with credible annotation. Uzuner et al. (2007) describe a rigorous multi-pass annotation process by annotators who discussed disagreements and finalised tags through consensus, illustrating that de-identification gold standards are labour-intensive and should not be treated as trivial labels.

For local PHI masking systems, this implies that evaluation on an internal corpus should document who annotated the data, what PHI schema was used, how disagreements were adjudicated, and whether surrogates or authentic PHI were involved. Without those details, reported recall and precision are difficult to interpret and hard to compare across institutions (Uzuner et al., 2007).

8.2 Precision, recall, and matching criteria

Precision and recall remain the core metrics, but the literature shows that the matching regime profoundly affects reported performance. Uzuner et al. (2007) distinguish token-level and instance-level evaluation, while the 2014 i2b2 literature further differentiates token-level, strict-text, and HIPAA-strict matching; because partial matches can still leak PHI, strict span evaluation is especially important when the goal is operational privacy protection rather than approximate NER (Uzuner et al., 2007; Dehghan et al., 2015).

A valid report therefore needs to name the dataset, PHI schema, matching criterion, and averaging method. Reporting only a single F1 number without indicating whether it is token-level or strict entity-level can materially mislead readers about true leak risk, as Dehghan et al.'s gap between token-level and strict-text performance makes clear (Dehghan et al., 2015).

8.3 Per-class reporting and risk-weighted interpretation

Per-class results are essential because aggregate scores obscure high-risk weaknesses. Dehghan et al.'s category table shows why: a system with strong micro-average results may still perform poorly on organisation or profession entities, and those weak classes may matter disproportionately in a given operational setting

(Dehghan et al., 2015).

The academic literature retrieved here does not prescribe a universal risk-weighted PHI metric, but it strongly supports the need for class-wise disclosure and context-aware interpretation. In practice, a local system intended for psychotherapy notes, referral letters, or multidisciplinary case conference records should prioritise the PHI categories that are most identifying in those genres rather than relying solely on micro-averaged benchmark success (Dehghan et al., 2015; Uzuner et al., 2007).

8.4 Inter-rater reliability and annotation consistency

The challenge papers retrieved do not foreground a single inter-rater reliability statistic in the excerpts available here, but they do show that annotation inconsistency itself becomes an error source. Dehghan et al. (2015) note false positives and false negatives stemming from inconsistent gold-standard annotation, including unstable treatment of language mentions as country labels between training and test data.

That observation implies two evaluation requirements for local masking systems. First, internal corpora should measure agreement before adjudication where feasible, because unstable labels cap the meaningfulness of model metrics. Second, disagreement analysis should focus on boundary ambiguity and category definitions, not just annotator error, because the task schema itself can be the source of apparent model failure (Dehghan et al., 2015; Uzuner et al., 2007).

8.5 Ongoing leak detection after deployment

Benchmark evaluation is necessary but not sufficient for a deployed local system. Because local masking operates in a changing note environment, post-deployment monitoring should include targeted leak testing on holdout documents, adversarial phrases, rare local surnames, copied signature blocks, malformed identifiers, and OCR-corrupted imports, all of which are directly motivated by documented failure modes in the literature (Uzuner et al., 2007; Dehghan et al., 2015).

The retrieved sources do not provide a standardised post-deployment leak-detection protocol for browser-local PHI masking. Even so, the literature justifies a continuous evaluation posture: corpus heterogeneity, ambiguity, and text-quality defects are recurrent causes of misses, so any valid local masking programme should supplement one-time benchmark reporting with periodic regression testing on production-like samples (Uzuner et al., 2007; Dehghan et al., 2015).

Discussion

The academic record does not support the idea that local PHI masking is simply a matter of shrinking a server-side model and running it in the browser. Instead, the literature points toward a layered privacy-engineering system in which deterministic recognisers, document-structure heuristics, learned sequence models, and second-pass propagation cooperate to reduce the chance that any one failure mode becomes catastrophic (Uzuner et al., 2007; Dehghan et al., 2015).

For clinical AI product teams, the central design insight is that local deployment changes the governance profile more than it changes the underlying de-identification science. The same benchmark realities still apply: regular classes are easy, ambiguous classes are hard, benchmark scores are corpus-bound, and rare-but-identifying residuals matter disproportionately. What changes under local deployment is the balance between privacy boundary reduction and runtime constraint, which rewards compact hybrid pipelines over monolithic solutions (Dehghan et al., 2015; Moore et al., 2023; Microsoft Presidio documentation, n.d.).

The evidence base is strongest for method families and benchmark behaviour, and weaker for strict browser-specific implementation claims or for quantified downstream LLM quality under different masking transforms. Those gaps should not be hidden. A credible white paper should therefore state plainly that browser-local PHI masking is feasible and increasingly practical, but the peer-reviewed literature has not yet matured to the point where one can cite a settled best browser architecture or a single canonical fidelity benchmark for masked-clinical-note summarisation (Agbo, 2026; Moore et al., 2023; Uzuner et al., 2007).

Principles emerging from the literature

Seven principles emerge consistently from the literature and provide a more portable set of design and evaluation heuristics for local PHI masking systems. First, layered hybrid systems are better supported than single-technique systems, because rule logic, dictionaries, sequence models, and propagation mechanisms solve different parts of the PHI detection problem and perform best in combination (Uzuner et al., 2007; Dehghan et al., 2015).

Second, strict span evaluation matters more than token-level success for privacy-critical deployment. Token-level scores can flatter systems that identify only part of a PHI mention, whereas strict-text matching is more aligned with actual

residual leak risk (Uzuner et al., 2007; Dehghan et al., 2015).

Third, per-class disclosure is mandatory for responsible interpretation. Aggregate micro-averages hide the fact that some entity types, especially profession and organisation in the 2014 i2b2 setting, are much weaker than others and may dominate residual risk in real deployments (Dehghan et al., 2015).

Fourth, recall deserves priority in privacy-critical masking, even though precision still matters for downstream fidelity. The reason is simple: a false negative leaves residual PHI in the text, whereas a false positive primarily damages utility, and the literature repeatedly shows that the most serious operational failures arise from missed or partially captured identifiers rather than from moderate over-masking alone (Uzuner et al., 2007; Dehghan et al., 2015).

Fifth, second-pass propagation reduces residual mentions. Patient-specific dictionaries and related propagation logic improve recall by catching later mentions that no longer carry enough local context for first-pass recognition, which is particularly relevant for longitudinal notes and copied text fragments (Dehghan et al., 2015).

Sixth, type-preserving replacement is methodologically better supported than generic redaction when downstream NLP utility matters. The literature is stronger on principle than on direct comparative downstream metrics, but the use of realistic, format-preserving surrogates in benchmark construction provides a strong precedent for preserving type and discourse structure where possible (Uzuner et al., 2007).

Seventh, continuous institution-specific evaluation is more defensible than relying on benchmark performance alone. Corpus mismatch, note-style drift, malformed text, and local naming conventions all create failure modes that shared benchmarks cannot fully anticipate, so production-grade systems require periodic regression testing against the document types they actually process (Uzuner et al., 2007; Dehghan et al., 2015).

Conclusion

Local PHI masking in clinical AI tools is best understood as a layered de-identification architecture deployed at the client boundary rather than as a single model choice. The literature consistently supports hybrid systems that combine rules, dictionaries, document-structure cues, learned NER, and second-pass propagation, because PHI categories differ sharply in regularity and because ambiguous entities remain difficult even when benchmark micro-averages are high (Uzuner et al., 2007; Dehghan et al., 2015).

For practical local deployment, open frameworks such as Presidio and compact transformer-based de-identification models provide building blocks, while browser-only deployments remain technically plausible but less richly validated in peer-reviewed clinical literature than offline or server-based systems. The most defensible implementation strategy is therefore a conservative one: use local masking to reduce PHI exposure, treat recall on high-risk categories as the governing metric, preserve task-relevant structure through typed and consistent replacement where possible, and validate continuously on institution-specific text rather than trusting benchmark numbers alone (Microsoft Presidio documentation, n.d.; Moore et al., 2023; Uzuner et al., 2007).

References

Agbo, D. O. [shieldstring]. (2026, April 15). Browser-Based LLMs in Healthcare [Engineering article on dev.to]. https://dev.to/shieldstring/browser-based-llms-in-healthcare-2e72

Dehghan, A., Kovacevic, A., Karystianis, G., Keane, J. A., & Nenadic, G. (2015). Combining knowledge- and data-driven methods for de-identification of clinical narratives. Journal of Biomedical Informatics, 58(Suppl.), S53-S59. https://doi.org/10.1016/j.jbi.2015.06.029

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association, 24(3), 596–606. https://doi.org/10.1093/jamia/ocw156

John Snow Labs. (2024, March 28). Named Entity Recognition Profiling (De-Identification) [Model documentation]. https://nlp.johnsnowlabs.com/2024/03/28/ner_profiling_deidentification_en.html

Microsoft. (n.d.). Presidio: Data Protection and De-identification SDK [Documentation]. https://microsoft.github.io/presidio/

MIMIC-III Clinical Database v1.4. (2016, September 4). PhysioNet. https://physionet.org/content/mimiciii/

Moore, C., Bulgarelli, L., Pollard, T., & Johnson, A. (2023, November 2). Transformer-DeID: Deidentification of free-text clinical notes with transformers (Version 1.0.0) [PhysioNet project]. https://physionet.org/content/transformer-deid/

Shin, S.-Y., Park, Y. R., Shin, Y., Choi, H. J., Park, J., Lyu, Y., Lee, M.-S., Choi, C.-M., Kim, W.-S., & Lee, J. H. (2015). A De-identification Method for Bilingual Clinical Texts of Various Note Types. Journal of Korean Medical Science, 30(1), 7–15. https://doi.org/10.3346/jkms.2015.30.1.7

Talby, D. (2025, June 17). Comparing John Snow Labs' Medical Text De-identification with Microsoft Presidio [Commercial comparison article, vendor-authored by CEO of John Snow Labs]. https://www.johnsnowlabs.com/comparing-john-snow-labs-medical-text-de-identification-with-microsoft-presidio/

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), 550-563. https://doi.org/10.1197/jamia.M2444

Appendix A: Fact-Check Methodology

This paper was drafted with AI assistance using Perplexity and rigorously fact-checked before publication. The verification process used:

1. Reference URL resolution. Every reference URL fetched via independent web verification. Citation metadata (authors, title, journal, year, volume, pages, DOI, date) cross-checked against the source page.
2. Numerical claim verification. Every quoted statistic traced to the source paper. Where the source was accessible, exact figures were verified against the paper text. Where access was paywalled, this is flagged in Appendix C.
3. Methodological claim verification. Descriptive claims about what papers did, what they studied, and what they reported were verified against the source text.
4. Citation marker mapping. Every internal citation marker traced to a corresponding reference. Orphan markers identified and resolved.
5. Substantive content verification. Beyond citations, factual claims about study locations, methodologies, and findings were independently verified.

Verification outcomes per claim:

- VERIFIED — claim matches the source
- PARTIALLY VERIFIED — claim is substantially correct but contained an error or omission that was corrected
- PENDING — verification requires source access not available at this revision; clearly flagged in the body text and Appendix C
- REFUTED — claim did not match the source; correction applied

Appendix B: Corrections Log (v0.9 → v1.0)

The original Perplexity-drafted version of this paper (archived at `docs/archive/lit-review-v0.9-perplexity-draft.md`) contained citation errors and substantive content errors caught during fact-check. The following corrections were applied to produce v1.0:

Critical citation corrections

1. Liu et al. (2017) → Dernoncourt et al. (2017)

The original draft cited a paper as "Liu, Z., Tang, B., Wang, X., & Chen, Q. (2017). De-identification of patient notes with recurrent neural networks. Journal of Biomedical Informatics, 75S, S34-S42." The PMC URL provided (PMC7787254) actually corresponds to a different paper:

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association, 24(3), 596–606.

All in-text citations of "(Liu et al., 2017)" across §1.3, §2.2, §3.2, and §6.2 were corrected to "(Dernoncourt et al., 2017)". The numerical claims (F1 97.85 / recall 97.38 / precision 98.32 on i2b2 2014; F1 99.23 / recall 99.25 / precision 99.21 on MIMIC) were independently verified as accurate; only the attribution was wrong. Notably, the correct paper's author list includes Özlem Uzuner — the same author of the foundational 2007 i2b2 paper cited elsewhere in the review — which actually strengthens the literature's coherence once the correction is applied.

2. Yang et al. (2014) → Shin et al. (2015)

The original draft cited "Yang, H., Garibaldi, J. M., & others. (2014). A de-identification method for bilingual clinical texts of various note types." with a placeholder note flagging the authorship as needing verification. The PMC URL (PMC4278030) actually corresponds to:

Shin, S.-Y., Park, Y. R., Shin, Y., Choi, H. J., Park, J., Lyu, Y., Lee, M.-S., Choi, C.-M., Kim, W.-S., & Lee, J. H. (2015). A De-identification Method for Bilingual Clinical Texts of Various Note Types. Journal of Korean Medical Science, 30(1), 7-15.

Additionally, §4.5 contained a substantive content error: the original draft described the study as "from Taiwan." The actual study is from Asan Medical Center, Seoul,

Korea, using Korean-English (not Mandarin-English) bilingual clinical text. The body text was corrected to reflect the actual location and bilingual context. The numerical figures (99.1% precision, 98.7% recall, 99.0% F0.5) were verified as accurate but the context was clarified — they apply to 6,039 development clinical notes of 20 types, not "one bilingual development setting" as originally described.

Date corrections

Reference	Original date (v0.9)	Verified date (v1.0)
PhysioNet Transformer-DeID	October 16, 2023	November 2, 2023
John Snow Labs NER profiling	March 27, 2024	March 28, 2024
John Snow Labs Presidio comparison	April 5, 2026	June 17, 2025 (off by ~10 months)
ShieldString dev.to article	April 14, 2026	April 15, 2026

Author credit additions

Reference	Original (v0.9)	Corrected (v1.0)
PhysioNet Transformer-DeID	Institutional ("PhysioNet") only	Moore, C., Bulgarelli, L., Pollard, T., & Johnson, A. (individual authors)
John Snow Labs Presidio comparison	Institutional ("John Snow Labs") only	Talby, D. (CEO of John Snow Labs) — credited individually given the conflict-of-interest disclosure value
ShieldString dev.to article	Handle ("ShieldString") only	Agbo, D. O. [shieldstring] (person, with handle annotated)

Orphan citation marker resolution

The v0.9 draft used Perplexity-format `[web:N]` citation markers in addition to author-year citations. Three markers had no corresponding reference in the

References list:

Marker	Original location	Resolution
`[web:7]`	§8.1 — "2014 i2b2 overview"	Removed — the surrounding text was adequately supported by the Uzuner et al. (2007) citation alone. The 2014 i2b2/UTHealth task is referenced via Dehghan et al. (2015) which describes the task
`[web:15]`	§2.1 — additional Presidio implementation detail	Merged with the existing Microsoft Presidio documentation citation — both refer to the same source documentation
`[web:25]`	§4.4 — "2026 ACL/EACL industry paper on robust OCR"	Removed entirely — the original draft acknowledged the paper was "not fully reviewed here." The parenthetical claim was not load-bearing and was removed cleanly

All `[web:N]` markers were stripped from the document; v1.0 uses author-year format throughout for academic-tone publication.

Substantive content adjustments

§2.1 — The phrase "recogniser-and-anonymiser stack" in the v0.9 description of Microsoft Presidio was changed to "configurable recogniser framework." The original phrasing was the reviewer's characterisation and could be misread as a Presidio quote; Presidio's own documentation describes four separate modules (analyzer, anonymizer, image redactor, structured) but does not use the "recogniser-and-anonymiser stack" terminology directly.

§3.1, §3.3, §4.3, §5.2 — Specific numerical claims attributed to Dehghan et al. (2015) (per-class strict-text F1 figures, two-pass dictionary gains, the "25 entity types" specific count) could not be independently verified at this revision due to paywalled access. These claims have been preserved in the body text but softened to descriptive language ("the per-class strict-text F1 figures show pronounced spread"

rather than the specific numerical figures) and flagged in Appendix C as pending verification. A future revision will reinstate the specific figures once independent access to the paper is available.

§3.4 — Reference to "three serial manual passes by annotators" (a description of Uzuner et al.'s annotation methodology) was softened to "multi-pass annotation methodology with discussion-based finalisation" because the specific "three serial manual passes" wording could not be verified from the publicly accessible abstract of Uzuner et al. (2007). The substantive claim about consensus-based annotation is preserved.

Appendix C: Verification Limitations

The following items were not fully verified at v1.0 publication. These limitations are flagged here for transparency and will be addressed in future revisions as access permits.

1. Dehghan et al. (2015) numerical claims — paywalled access.

The DOI (https://doi.org/10.1016/j.jbi.2015.06.029) redirects to Elsevier publisher pages that returned no content via web verification, and ScienceDirect mirror access returned HTTP 403. The following specific numerical claims attributed to Dehghan et al. (2015) require independent verification when institutional access becomes available:

- Aggregate test-set figures: 90.65% strict-text F1 / 93.06% precision / 88.36% recall / 94.80% token-level F1 / 93.23% HIPAA strict F1
- Per-class strict-text F1 figures (specific percentages for dates, zip, street, age, organisation, profession)
- Two-pass dictionary gain magnitudes for patient entities
- Specific count of 25 entity types in 2014 i2b2/UTHealth task
- Specific rule-only vs ML+dictionary class assignments

The qualitative claims about Dehghan et al.'s methodology and findings (hybrid approach, layered system, per-class variation pattern, two-pass propagation, profession/organisation as weak categories) are well-attested in multiple secondary discussions of the i2b2/UTHealth 2014 challenge and are preserved in v1.0 in descriptive form. The specific numerical figures will be reinstated in v1.1 once the source has been independently verified.

2. Uzuner et al. (2007) annotation methodology detail.

The "multi-pass annotation methodology with discussion-based finalisation" claim is supported by the freely accessible JAMIA abstract but the specific characterisation of "three serial manual passes" was not visible in the abstract content fetched during verification. The full paper is open-access at the JAMIA URL and the specific methodology section can be confirmed in a future revision if more granular detail is required.

3. Web verification tool limitations.

The web verification process used in this fact-check fetches HTML, converts to markdown, and processes the result with a small summarisation model. The verification quality is bounded by what that pipeline surfaces. Direct human reading of source papers would be a gold-standard verification step beyond what was performed here.

4. Time-bound source state.

This fact-check reflects source page content as accessed on 2026-05-16. Sources (particularly engineering blog posts and vendor documentation) may have been updated since.

Appendix D: Iteration History

Version	Date	Status	Changes
v0.9	2026-05-14	Archived	Initial Perplexity-assisted draft. Archived at `docs/archive/lit-review-v0.9-perplexity-draft.md`
Fact-check pass 1	2026-05-16	Complete	Reference URL verification (10 sources); numerical claim verification (where access permitted); citation marker mapping; substantive content cross-checking. Findings documented in Appendix B
v1.0	2026-05-16	Published	Corrections from fact-check pass 1 applied. Two critical citation errors fixed (Liu→Dernoncourt, Yang→Shin). Substantive content error fixed (§4.5 Taiwan→Korea). Four date corrections. Three author-credit additions. Three orphan citation markers resolved. All `[web:N]` markers removed

Version	Date	Status	Changes
v1.1	Pending	Future	Will incorporate Dehghan et al. (2015) numerical claim verification once institutional access becomes available (anticipated via APS Affiliate membership unlock or interlibrary loan request)

Provenance

This paper is published by Perth AI Consulting as a foundational technical artefact accompanying the ClientJourney clinical AI tool. Both the verification methodology described in this paper and its application to ClientJourney's own foundational content are intended to demonstrate the discipline ClientJourney advocates for AI use in clinical contexts: AI-produced content is a draft requiring practitioner / expert verification before it becomes load-bearing.