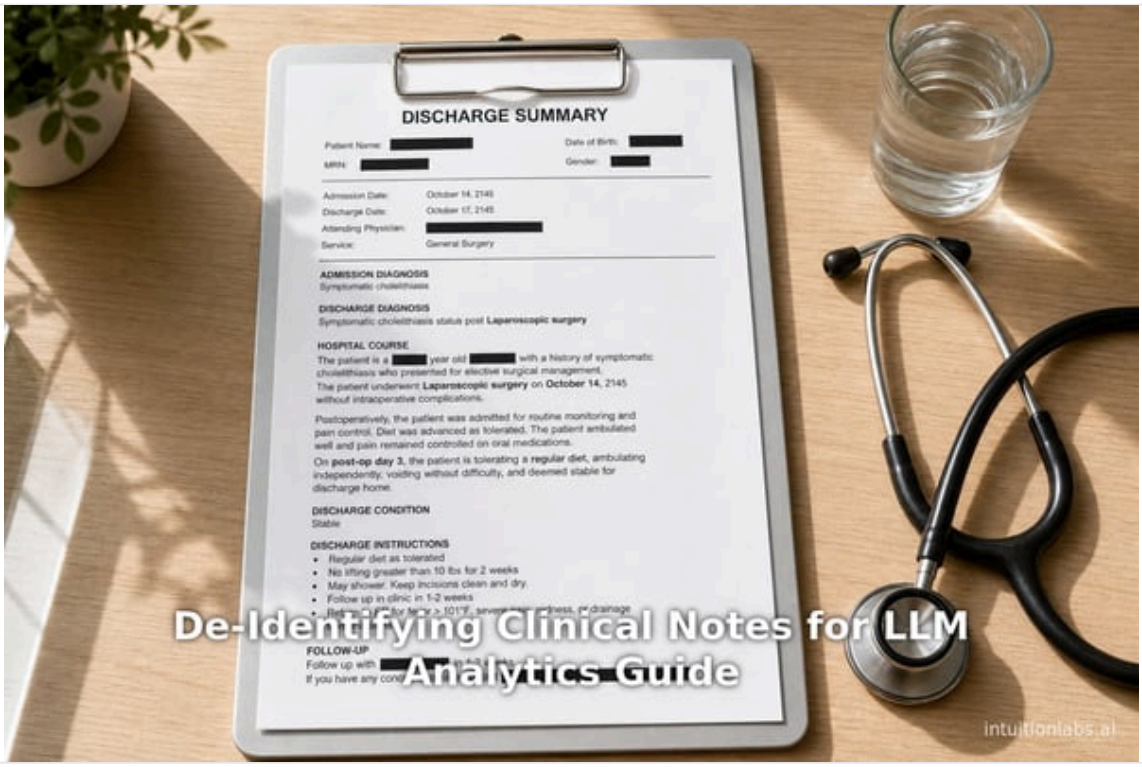
De-Identifying Clinical Notes for LLM Analytics Guide

Published July 20, 2026 38 min read



Executive Summary

Clinical notes are the richest source of information in the electronic health record (EHR), yet the same free text that makes them valuable for large language model (LLM) analytics also makes them legally radioactive. Under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, health information becomes usable outside direct patient care only after it has been de-identified through one of two federally recognized pathways: the rule-based **Safe Harbor** method, which requires removing 18 categories of identifiers (Source: www.hhs.gov), or the risk-based **Expert Determination** method, in which a qualified statistician certifies that re-identification risk is very small (Source: www.hhs.gov). Even properly executed, HHS itself acknowledges that both methods "yield de-identified data that retains some risk of identification" (Source: www.hhs.gov), a caveat borne out by Harvard researcher Latanya Sweeney's field study, which correctly re-identified 25 percent of participants by name in a HIPAA Safe Harbor-compliant dataset (Source: techscience.org).

For clinical text, de-identification depends on automated named entity recognition (NER) pipelines that combine regular-expression pattern matching, statistical models, and increasingly, LLMs, to find and remove protected health information (PHI) from unstructured notes. Independent benchmarking published in 2025 found wide variance among commercial and open tools: **John Snow Labs' Healthcare NLP library** scored a 96 percent F1-score on PHI detection, versus 91 percent for **Microsoft Azure Health Data Services**, 83 percent for **Amazon Comprehend Medical**, and 79 percent for zero-shot GPT-4o prompting (Source: arxiv.org). Open-source options tell a similar story of hybrid rule-plus-statistical designs outperforming single-method approaches: UCSF's **Philter** achieved 99.46 percent recall on its internal test corpus and 99.92 percent on the public i2b2 2014 corpus, versus 85.10 percent and 69.84 percent respectively for the older PhysioNet Scrubber tool (Source: www.nature.com) (Source: www.nature.com). Microsoft's open-source **Presidio** framework, now maintained under the Data Privacy Stack organization, has surpassed 10,100 GitHub stars (Source: github.com), while cloud vendors price de-identification per unit of text: Azure charges $0.05 per megabyte of unstructured text beyond a 50 MB free tier (Source: azure.microsoft.com), and Amazon Comprehend Medical bills per 100-character unit (Source: docs.aws.amazon.com) (Source: aws.amazon.com).

The stakes of getting this wrong are large and quantifiable. IBM's 2025 Cost of a Data Breach Report, conducted with the Ponemon Institute, put the average cost of a U.S. healthcare data breach at $7.42 million, still the costliest of any industry studied for the 14th consecutive year (Source: www.hipaajournal.com), while U.S. breach costs overall hit a record $10.22 million (Source: www.hipaajournal.com). HIPAA civil penalties for willful neglect that goes uncorrected can reach $2,190,294 per violation category per year as of the 2026 inflation adjustment (Source: www.hipaajournal.com). The 2024 Change Healthcare ransomware attack, a reminder of what is at risk when clinical data protections fail, was projected by UnitedHealth Group to cost up to $1.6 billion in a single year (Source: www.reuters.com), after already booking $872 million in breach-related costs in a single quarter (Source: www.reuters.com).

Organizations building LLM-analytics pipelines on clinical text generally choose among five paths: open-source NER frameworks (Presidio, Philter, NLM Scrubber), cloud PHI-detection APIs (Azure Health Data Services, Amazon Comprehend Medical), specialized commercial platforms (John Snow Labs, Private AI), zero-shot or fine-tuned LLM redaction, and pre-de-identified reference corpora such as **MIMIC-IV-Note**, which contains 331,794 discharge summaries from 145,915 patients, all scrubbed under HIPAA Safe Harbor (Source: physionet.org). No method is a substitute for validation: the largest independently documented healthcare deployment, Providence Health's 2-billion-note pipeline built on John Snow Labs' Healthcare NLP library, required months of external red-teaming layered on top of an already high-scoring model before the vendor could report zero re-identifications in adversarial testing, a result detailed later in this report's case studies. Large pre-trained language models also carry a subtler risk of their own: researchers studying clinical BERT models note that such models "memorize parts of their training data, making them vulnerable to various privacy attacks," a finding directly relevant to any organization fine-tuning an LLM on clinical text rather than merely analyzing it (Source: link.springer.com). This report walks through the regulatory taxonomy, the technical mechanics of NER-based de-identification, a comparative assessment of the leading tools, and the implementation practices that separate defensible pipelines from ones that merely look compliant.

Introduction and Background

Large language models trained or fine-tuned on real clinical narratives promise substantial gains for clinical decision support, quality measurement, and population health analytics, because free-text notes carry detail that structured EHR fields never capture. At the University of California, San Francisco (UCSF) alone, the health system has accumulated **over 70 million clinical notes** collected within its electronic health records (Source: www.nature.com), a volume replicated at scale across every large health system in the United States. That volume is also the reason clinical notes cannot simply be piped into an LLM analytics stack: nearly every note contains protected health information (PHI), the identifiers that connect a clinical narrative to a specific, named patient.The National Institute of Standards and Technology (NIST) defines de-identification broadly as "a general term for any process of removing the association between a set of identifying data and the data subject" (Source: www.nist.gov), a definition broad enough to cover techniques ranging from removing identifiers, transforming quasi-identifiers, and generating synthetic data using models (Source: www.nist.gov). For healthcare organizations in the United States, that general definition is made specific and enforceable by the HIPAA Privacy Rule, which recognizes two compliant de-identification methods and treats de-identified data as no longer subject to HIPAA once either standard is met. The European Union runs a parallel but distinct framework under the General Data Protection Regulation (GDPR), where "pseudonymisation" and "anonymisation" are legally separate concepts with different obligations attached, a distinction the European Data Protection Board (EDPB) refined again as recently as July 2026.

Consultancies operating at the intersection of life-sciences regulation and applied artificial intelligence, such as **IntuitionLabs**, routinely advise pharmaceutical and healthcare organizations on the data engineering and regulatory compliance work that de-identification sits inside of; IntuitionLabs describes this work as delivering "robust data pipelines, integration, warehousing, and business intelligence for actionable insights" alongside "built-in compliance with FDA, EMA, and global regulations" (Source: intuitionlabs.ai) (Source: intuitionlabs.ai). That advisory context matters because de-identification of clinical notes is rarely a single tool decision. It is a governance decision that determines which downstream analytics, model training, and data-sharing activities become legally possible.

This report addresses the practical question organizations face when they want to move clinical free text into LLM analytics or training pipelines: which de-identification method and tooling combination is defensible, how do the leading open-source and commercial NER platforms compare on accuracy and cost, and what pitfalls turn an apparently compliant pipeline into a re-identification risk. It draws on regulatory guidance from HHS, NIST, and the EDPB; peer-reviewed benchmarking studies; vendor documentation; and named deployments including Providence Health, Vanderbilt University Medical Center, and the MIMIC critical care database.

Protected Health Information, De-Identification, and the Regulatory Taxonomy

Before any tool comparison is useful, the underlying legal categories have to be clear, because they determine what "success" means for an automated pipeline. Under HIPAA, health information tied to an identifiable person is Protected Health Information (PHI), and the Privacy Rule permits use of that information for research, analytics, or data-sharing purposes only after de-identification, since de-identified information "mitigates privacy risks to individuals and thereby supports the secondary use of data for comparative effectiveness studies, policy assessment, life sciences research,

and other endeavors" (Source: www.hhs.gov). HHS's Office for Civil Rights (OCR) recognizes exactly two compliant paths, described in its 2012 guidance as "the two methods that can be used to satisfy the Privacy Rule's de-identification standard: Expert Determination and Safe Harbor" (Source: www.hhs.gov).

Safe Harbor is the rule-based, checklist-style method most automated NER pipelines are built around. It requires removing 18 categories of identifiers relating to the patient or the patient's relatives, employers, or household members, plus a residual condition that the covered entity have no actual knowledge that what remains could still identify someone. The identifier list is specific about edge cases: dates directly related to an individual, other than year, must be suppressed, "including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age" (Source: www.hhs.gov), and geographic subdivisions smaller than a state, "including street address, city, county, precinct, ZIP code" (Source: www.hhs.gov), must generally be removed unless the three-digit ZIP prefix covers more than 20,000 people. Even after all 18 categories are removed, Safe Harbor additionally requires that "the covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual" (Source: www.hhs.gov).

Table 1 summarizes the Safe Harbor identifier categories most relevant to free-text clinical notes, which is where automated NER matters most because structured EHR fields can usually be dropped or masked programmatically, while free text requires the model to recognize identifiers embedded in unpredictable sentence structures.

IDENTIFIER CATEGORY	SAFE HARBOR TREATMENT	TYPICAL NER CHALLENGE IN FREE TEXT
Names	All patient, relative, employer, and household member names removed	Nicknames, possessives, and misspellings evade dictionary matching
Geographic subdivisions smaller than a state	Street, city, county, and full ZIP removed; 3-digit ZIP allowed only above 20,000-person thresholds (Source: www.hhs.gov)	Facility names and rural place names are easily confused with medical terms
All elements of dates except year	Birth, admission, discharge, and death dates suppressed; ages over 89 aggregated to "90 or older" (Source: www.hhs.gov)	Relative dates ("post-op day 3") and free-text date formats are inconsistent across EHR templates
Telephone, fax, email, URLs, IP addresses	Removed entirely	Pattern-matchable with regular expressions once formats are known
Social Security, medical record, health plan, and account numbers	Removed entirely	Numeric strings are ambiguous without surrounding context
Biometric identifiers and full-face photographs	Removed entirely	Not applicable to plain text, but relevant to multimodal PHI in scanned or audio records
Any other unique identifying number, characteristic, or code	Catch-all clause	The hardest category for rule-based systems, since it is open-ended by design

Table 1 shows why Safe Harbor, despite being described as convenient because "no special computer programs, statistical modeling, or advanced analysis is necessary" for a human reviewer (Source: techscience.org), is genuinely difficult to automate reliably against unstructured text: several categories, especially dates, geography, and the open-ended "unique identifying number" clause, require contextual language understanding rather than simple pattern matching.

Expert Determination is the alternative, risk-based path. Rather than following a checklist, "a covered entity may determine that health information is not individually identifiable health information only if" a person with appropriate statistical and scientific expertise applies generally accepted methods and documents that the risk of re-identification is very small, and that determination must be certified and can be time-limited (Source: www.hhs.gov). Expert Determination allows retaining more contextual detail, such as three-digit ZIP codes below the Safe Harbor threshold or exact ages, when the statistical risk calculation supports it, which is why it is frequently chosen for datasets intended for Food and Drug Administration (FDA) submission or

complex multi-source research use. That flexibility carries a cost: independent guidance estimates that Expert Determination engagements "range from a few thousand dollars for a well-structured dataset to tens of thousands" for complex, multi-source data (Source: www.getlimina.ai), while Safe Harbor has no independent statistician fee but a comparatively blunt effect on data utility (Source: www.getlimina.ai).

Outside the United States, the relevant vocabulary shifts. Under the GDPR, the EDPB's January 2025 guidelines define pseudonymisation precisely: it "can reduce the risks to the data subjects by preventing the attribution of personal data to natural persons" while the data itself remains personal data subject to the regulation, because it "could be attributed to a natural person by the use of additional information" (Source: www.edpb.europa.eu). True anonymisation is a materially higher bar: the EDPB's newer Guidelines 02/2026 on Anonymisation, adopted July 7, 2026, state that under the GDPR "data is anonymous if it does not relate to an identified or identifiable natural person" (Source: www.edpb.europa.eu), a determination made contextually rather than through a fixed identifier checklist. This means a dataset that satisfies HIPAA Safe Harbor may still be treated as pseudonymized, and therefore regulated, personal data under GDPR, a discrepancy that matters for any multinational life-sciences organization building a single de-identification pipeline meant to satisfy both frameworks at once.

How Automated De-Identification Works: Rule-Based, Statistical, and LLM Pipelines

Automated clinical de-identification is fundamentally a named entity recognition (NER) problem: locate spans of text that correspond to PHI categories, then decide what to do with each span, typically remove, mask, tag, or replace it with a synthetic surrogate. The field has moved through three broad generations of approach, and modern production tools generally combine more than one.

Rule-based systems were the earliest generation. As one 2025 benchmarking paper puts it, "early de-identification systems in the clinical domain were predominantly rule-based" (Source: arxiv.org), relying on regular expressions, dictionaries, and hand-written syntactic rules. These systems remain strong at structured PHI, such as phone numbers, medical record numbers, and email addresses, where formats are predictable, but struggle with personal names, professions, and facility names that resist pattern matching. Rule-based systems are also brittle across institutions, since a pipeline tuned to one hospital's note templates can degrade sharply when applied to another health system's documentation conventions. The very first MIMIC-III critical care database, one of the earliest large-scale de-identified clinical corpora, used exactly this generation of technique: structured fields were sanitized by "structured data cleansing and date shifting," while free text was scrubbed by "a rigorously evaluated deidentification system based on extensive dictionary look-ups and pattern-matching with regular expressions" (Source: physionet.org).

Statistical and neural NER models followed, using classifiers such as conditional random fields and, later, deep learning architectures trained on annotated PHI corpora. Traditional anonymization approaches within this generation fall into two families: **data masking**, in which "sensitive information is replaced with false but plausible alternatives" (Source: pmc.ncbi.nlm.nih.gov), and **pseudonymization**, which substitutes identifiers with artificial tokens that can, if a mapping key exists, be reversed. The best-performing production systems today combine rule-based and statistical layers rather than choosing one: UCSF's Philter, for example, "utilizes an overlapping pipeline of methods that are state-of-the-art in each application including: pattern matching, statistical modeling, blacklists, and whitelists" (Source: www.nature.com).

LLM-based de-identification is the newest and fastest-moving generation. Instead of a trained classifier, a general-purpose or medically fine-tuned LLM is prompted, in a zero-shot or few-shot setting, to identify and redact PHI spans directly from raw text. This approach has produced genuinely strong results in controlled studies: researchers using GPT-4 at King Hussein Cancer Center reported "a precision of 0.9925, a recall of 0.8318, an F1 score of 0.8973, and an accuracy of 0.9911" on real clinical notes (Source: pmc.ncbi.nlm.nih.gov). LLM-based frameworks such as **RedactX**, built at Oracle Health, extend this further with a multi-pass extraction routine and a "retrieval-based entity relexicalization approach to ensure consistent substitutions of protected entities, thereby enhancing data coherence for downstream applications" (Source: arxiv.org), and it "achiev[ed] the highest F1-score of 0.9646 with a well-balanced precision and recall" against comparable LLM-only baselines (Source: arxiv.org) after "12+ months of deployment in real-world clinical AI system" integration (Source: arxiv.org).

NER frameworks used for de-identification also differ in how granular their entity taxonomy is. Microsoft's open-source **Presidio** engine, for example, defines a generic **PERSON** entity as "a full person name, which can include first names, middle names or initials, and last names" (Source: microsoft.github.io) and a **LOCATION** entity as the "name of politically or geographically defined location" (Source: microsoft.github.io), categories broad enough to be reused across industries but requiring supplemental medical-entity recognizers to reach clinical-grade coverage. Presidio's optional medical extra narrows the gap by adding entities such as "a disease or disorder (e.g. diabetes, hypertension)" (Source: microsoft.github.io), though clinically specialized platforms still ship dozens of fine-grained PHI classes out of the box, such as DOCTOR, HOSPITAL, MEDICALRECORD, and IDNUM, which reduces the custom engineering burden but ties the pipeline more tightly to a specific vendor's taxonomy.

Clinical De-Identification Tools and Platforms Compared

Organizations building an LLM-analytics pipeline generally choose among open-source frameworks they self-host, cloud PHI-detection APIs billed by usage, and specialized commercial platforms licensed for on-premises or air-gapped deployment. Each category makes different tradeoffs between accuracy, cost, and control over PHI as it moves through the pipeline.

Presidio, originally released by Microsoft and now maintained under the Data Privacy Stack organization following a 2026 project transition, is described in its own repository as a "context aware, pluggable and customizable PII de-identification service for text and images" (Source: github.com) and has accumulated over 10,100 GitHub stars (Source: github.com). It ships broad, jurisdiction-specific entity recognizers spanning the United States, United Kingdom, and a dozen other countries, plus an optional medical NER extra that adds disease, medication, and procedure recognizers built on a HuggingFace transformer model, and it also integrates directly with the Azure Health Data Services de-identification service as a remote recognizer for organizations that want cloud-grade PHI coverage inside an otherwise open-source pipeline.

Philter, developed at UCSF's Bakar Computational Health Sciences Institute and released as open-source software, is purpose-built for clinical text and "removes 18 HIPAA-defined types of PHI from the user-provided text" using a combined rule-based and statistical pipeline (Source: informationcommons.ucsf.edu). Its GitHub repository is described simply as "open source clinical text de-identification" (Source: github.com) and released under a permissive BSD license, making it a common choice for academic medical centers that need a self-hosted, no-internet-connection-required option.

NLM Scrubber, built at the National Library of Medicine (NLM), targets the same Safe Harbor use case with a freely downloadable desktop tool: it "is a freely available clinical text deidentification tool designed and developed at the National Library of Medicine," aimed at enabling Safe Harbor-compliant secondary use of clinical text (Source: lhncbc.nlm.nih.gov), though NLM explicitly cautions that its developers "strongly recommend our users not to share deidentified data without a strong data use agreement" (Source: lhncbc.nlm.nih.gov), an important reminder that even a high-recall tool does not eliminate the need for downstream governance.

Amazon Comprehend Medical is AWS's managed API for clinical NLP, including a dedicated `DetectPHI` operation. AWS documentation is explicit that "Amazon Comprehend Medical detects entities associated with these identifiers but these entities don't map 1:1 to the list specified by the Safe Harbor method," meaning outputs require mapping and human validation against the actual regulatory checklist rather than being treated as automatically compliant (Source: docs.aws.amazon.com). The PHI detection API itself "finds and tags the 18 specified PHI fields according to Safe Harbor" (Source: aws.amazon.com), and requests are billed in units where "all requests are measured in units of 100 characters (1 unit = 100 characters), with a 1 unit (100 character) minimum charge per request" (Source: aws.amazon.com).

Azure Health Data Services offers a comparable managed de-identification service with Tag, Redact, and Surrogate operation modes, the last of which generates realistic replacement values rather than simple placeholder tags. As of mid-2026 pricing, unstructured text de-identification is billed under "Unstructured De-identification" transformation operations, free for the first 50 MB per month and $0.05 per MB beyond that threshold (Source: azure.microsoft.com), while structured de-identification and other transformation operations are billed separately at $1.14 per gigabyte beyond a similar free tier (Source: azure.microsoft.com).

John Snow Labs' Healthcare NLP library, part of the Spark NLP ecosystem, positions itself as a fixed-cost, locally deployable alternative to per-token cloud APIs, and independent benchmarking on 48 expert-annotated clinical documents found its de-identification pipeline "over 80% cheaper compared to Azure and GPT-4o, and is the only solution not priced by token" while also scoring the highest accuracy of the four systems tested (Source: arxiv.org). The company's own materials cite a separate benchmark in which its de-identification pipeline "establishes new state-of-the-art accuracy on 7 of 8 well-known benchmarks," including the 2014 n2c2 challenge, "outperform[ing] the accuracy of solutions such AWS Medical Comprehend and Google Cloud Healthcare API by a large margin (8.9% and 6.7% respectively)" (Source: www.johnsnowlabs.com); readers should note this figure is vendor-reported rather than independently replicated in the same study.

Private AI, rebranded as Limina AI in 2026, offers a comparable commercial redaction platform advertising "50+ Entity Types" (Source: www.private-ai.com) across "52 Languages" (Source: www.private-ai.com), deployable on-premises or in a customer's own cloud environment, a design aimed at organizations unwilling to send raw clinical text to a third-party API even for redaction purposes.

Table 2 consolidates the reported accuracy and deployment characteristics of the tools discussed above, drawing on the independently conducted 48-document benchmark for the cloud APIs and John Snow Labs, and the Nature-published evaluation for the open-source options.

TOOL	TYPE / DEPLOYMENT	REPORTED PHI ACCURACY	PRICING MODEL
Presidio	Open source, self-hosted (Python)	Depends on configured recognizers; not independently benchmarked as a turnkey clinical system	Free (MIT-derived license)
Philter	Open source, self-hosted, clinical-specific	99.46% recall (UCSF corpus), 99.92% recall (i2b2 corpus) (Source: www.nature.com)	Free (BSD license)
NLM Scrubber	Free desktop application, self-hosted	Institution-dependent; NLM recommends independent validation	Free
Amazon Comprehend Medical	Managed cloud API	83% F1 (independent 48-document benchmark) (Source: arxiv.org)	Per 100-character unit (Source: aws.amazon.com)
Azure Health Data Services	Managed cloud API	91% F1 (same independent benchmark)	$0.05/MB unstructured text beyond free tier (Source: azure.microsoft.com)
John Snow Labs Healthcare NLP	Commercial, on-premises or cloud	96% F1 (independent benchmark), vendor claims 99%+ at 2-billion-note production scale (Source: www.johnsnowlabs.com)	Fixed-cost licensing, not priced per token
GPT-4 / GPT-4o (zero-shot)	LLM API, no fine-tuning	79% F1 (JSL benchmark); 89.7% F1 in a separate KHCC study (Source: pmc.ncbi.nlm.nih.gov)	Per-token API pricing
Private AI / Limina AI	Commercial, on-premises or cloud	Not included in the independent multi-tool benchmark cited above	Licensed, entity-count and volume based (Source: www.private-ai.com)

Table 2 makes two patterns visible. First, accuracy on the same underlying task varies by nearly 20 percentage points of F1-score across vendors evaluated in a single controlled study, which means accuracy claims from a vendor's own marketing materials should not be assumed to transfer across datasets or note types. Second, pricing models diverge sharply between per-unit cloud billing (AWS, Azure, LLM APIs) and fixed-cost or free self-hosted alternatives (Presidio, Philter, NLM Scrubber, John Snow Labs), a distinction that matters enormously at the scale of a multi-million-note LLM training corpus, where per-unit costs compound quickly.

Preparing De-Identified Notes for LLM Training and Analytics

De-identification is only the first step in getting clinical text ready for LLM analytics or model training; the format and consistency of the output matter as much as raw PHI removal rates. Three practices recur across the strongest production pipelines and reference datasets.

Consistent, format-preserving substitution. Rather than deleting PHI outright, which breaks sentence structure and can itself leak information through conspicuous gaps, mature pipelines replace identifiers with realistic surrogates that preserve grammatical and semantic structure. The MIMIC-IV-Note corpus, published by the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center, takes the simpler masking approach, with deidentified entities replaced with three underscores in the released text, a format chosen for transparency about where PHI was removed rather than for downstream model realism. Its predecessor, MIMIC-III, instead used consistent per-patient date shifting: "dates were shifted into the future by a random offset for each individual patient in a consistent manner to preserve intervals," a technique deliberately designed to preserve the sequencing of clinical events, resulting in patient stays that appear to occur "sometime between the years 2100 and 2200" and patients over age 89 who "appear in the database with ages of over 300 years" (Source: physionet.org). More advanced commercial and research pipelines

instead perform relexicalization, generating plausible fake names, dates, and identifiers so a model fine-tuned on the output does not learn to associate blank tokens with clinical content, an approach the Azure Health Data Services surrogate operator and Oracle Health's RedactX framework both implement.

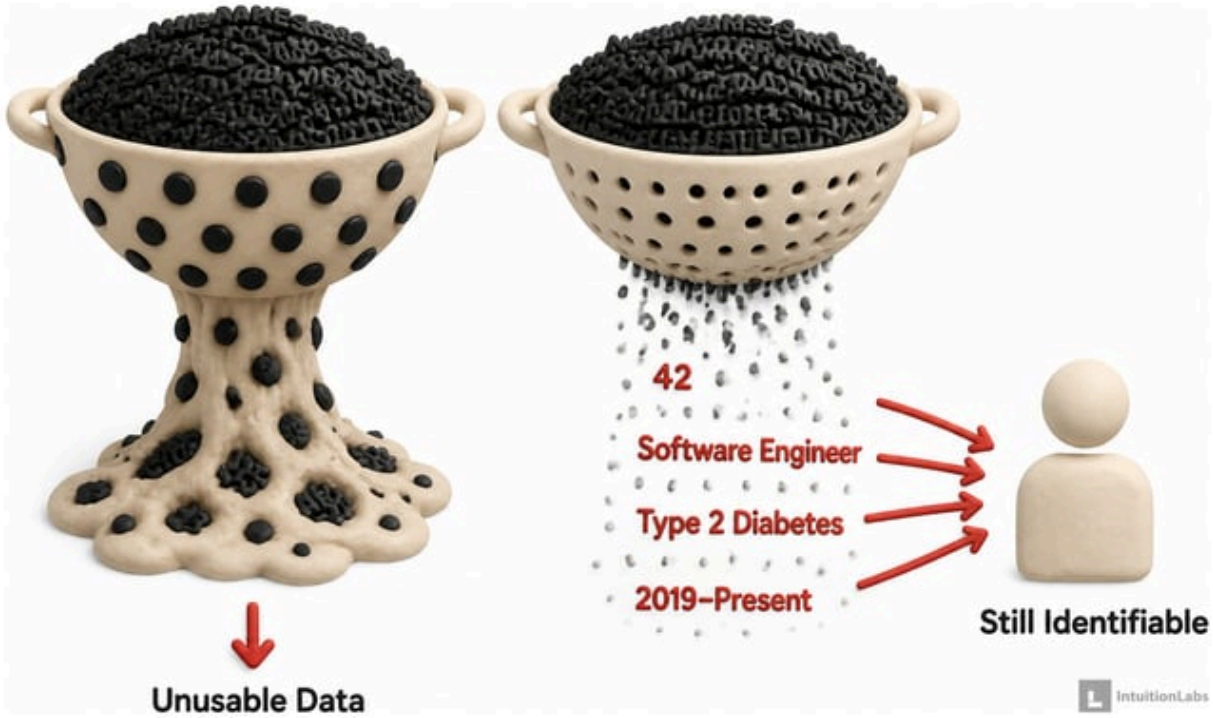
Credentialed, governed access even after de-identification. The most widely used clinical text corpus for LLM research, MIMIC-IV-Note, contains 331,794 deidentified discharge summaries from 145,915 patients and 2,321,355 deidentified radiology reports for 237,427 patients (Source: physionet.org) (Source: physionet.org), with "all notes have had protected health information removed in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision" (Source: physionet.org). Despite that, PhysioNet still restricts the dataset to credentialed users bound by a data use agreement rather than releasing it without gate, a governance pattern that recurs across every serious de-identified clinical dataset and reflects the residual re-identification risk discussed above. MIMIC-III itself, which comprises deidentified data for "over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012" (Source: physionet.org), required removal of "all eighteen of the identifying data elements listed in HIPAA, including fields such as patient name, telephone number, address, and dates" before release, illustrating that Safe Harbor's structured-data checklist and free-text de-identification were applied as two distinct engineering problems within the same project (Source: physionet.org).

Standardized benchmark corpora for training and validating NER pipelines themselves. Organizations building or fine-tuning their own de-identification models generally validate against the n2c2 (National NLP Clinical Challenges) family of datasets, whose 2014 de-identification and heart disease risk challenge corpus consists of "1304 de-identified longitudinal medical records from Partners HealthCare describing 296 patients" (Source: n2c2.dbmi.hms.harvard.edu). The n2c2 data sets, successors to the original i2b2 (Integrating Biology and the Bedside) shared tasks that trace back to a 2006 de-identification and smoking-status challenge, are "provided as a community service" and "consist of fully deidentified clinical notes and products of challenges" (Source: n2c2.dbmi.hms.harvard.edu), released under a data use agreement rather than an open license, again reflecting that even purpose-built benchmark corpora are not treated as unconditionally shareable.

For organizations that want longitudinal, multi-encounter data rather than isolated notes, institution-run de-identified research warehouses are an alternative to public benchmark corpora. Vanderbilt University Medical Center's Synthetic Derivative is a de-identified copy of the full clinical record, where "the database includes over 3.9 million records and is structured according to the OMOP common data model," and is described as "compliant with HIPAA Safe Harbor standards" (Source: victr.vumc.org) (Source: victr.vumc.org). Because names and dates in the Synthetic Derivative are replaced or shifted consistently per patient rather than deleted, longitudinal analyses and, by extension, LLM fine-tuning on sequential clinical events remain possible even though the data is no longer individually identifiable.

Implementation Guidance: Building a Defensible De-Identification Pipeline

Building a de-identification pipeline that will hold up to scrutiny, whether from an internal privacy office, an institutional review board, or an OCR audit, requires more than selecting a tool with a high headline accuracy number. Several recurring pitfalls of clinical note anonymization are worth calling out explicitly, because they are exactly the failure modes that separate a pipeline that looks compliant from one that is defensible.



- **Over-redaction destroys clinical utility.** A pipeline tuned purely for maximum recall tends to also mask non-PHI clinical content, such as medical terms that resemble names, degrading the value of the resulting corpus for downstream analytics; the goal is not zero risk but an acceptable risk-utility tradeoff.
- **Free-text formatting variance across EHR systems undermines rule-based generalization.** A pipeline validated on one institution's discharge summary templates can silently underperform on another institution's progress notes, radiology reports, or nursing documentation, since PHI location patterns and abbreviation conventions differ by note type and by health system.
- **Quasi-identifiers survive Safe Harbor's 18-category checklist.** Rare diagnoses, unusual combinations of age, occupation, and geography, or distinctive treatment timelines can re-identify a patient even when every explicit identifier on the HIPAA list has been removed, which is precisely the mechanism Sweeney's team exploited when it correctly re-identified patients from a Safe Harbor-compliant dataset using non-demographic fields (Source: techscience.org). Researchers studying clinical language models make the same structural point: even after applying HIPAA's checklist, "there are many quasi-identifiers that are not covered by this PHI definition" (Source: link.springer.com).
- **Context and negation are frequently lost or mishandled.** Clinical language routinely negates or hedges statements, and a de-identification model that does not understand negation can either fail to redact an identifier embedded in a denied statement or, conversely, redact clinically meaningful negative findings as if they were PHI.
- **Multimodal PHI is often overlooked.** Scanned PDFs, embedded images, and increasingly clinical audio recordings can carry identifiers that a text-only NER pipeline never sees; adding a second detection pass targeted specifically at unrecognized audio segments has been shown to meaningfully improve coverage, with one multimodal framework's secondary voice-activity-detection step increasing recall by "approximately 10%" on internal test data (Source: arxiv.org).
- **A single validation pass is not sufficient assurance.** As the Providence Health case study below documents, even a benchmark-leading pipeline only reached zero re-identifications after extensive, independently audited layered validation rather than a single accuracy measurement.

A defensible implementation process generally follows a consistent sequence of steps:

1. **Inventory the PHI categories actually present** in the target note types, since discharge summaries, radiology reports, and behavioral health notes carry different identifier distributions.
2. **Choose Safe Harbor or Expert Determination** based on whether the downstream analytics or LLM training use case requires retaining quasi-identifying detail such as narrow age bands or partial geography.

3. **Select tooling appropriate to the deployment environment**, weighing self-hosted open-source frameworks against managed cloud APIs based on whether raw PHI can leave institutional infrastructure at all.
4. **Validate against an independent gold-standard corpus**, such as the i2b2 2014 or n2c2 benchmark sets, before trusting the pipeline on live data.
5. **Red-team and adversarially test the output**, attempting deliberate re-identification using both the de-identified text and plausible external data sources.
6. **Document the methodology and, for Expert Determination, obtain and retain the statistician's certification**, since HHS guidance treats documentation of the methods and results as part of satisfying the standard (Source: www.hhs.gov).
7. **Re-validate periodically**, since note templates, EHR configurations, and patient population characteristics drift over time, and a pipeline validated once is not guaranteed to remain accurate indefinitely.

Data Analysis and Evidence

The quantitative picture across benchmarking studies, breach-cost data, and regulatory penalty structures makes the stakes of de-identification quality concrete rather than abstract. *Table 3* consolidates the accuracy figures referenced throughout this report alongside the financial and legal exposure that motivates rigorous pipelines in the first place.

METRIC	VALUE	SOURCE AND CONTEXT
John Snow Labs Healthcare NLP F1-score (48-document independent benchmark)	96%	Outperformed Azure (91%), AWS (83%), GPT-4o (79%) in the same study (Source: arxiv.org)
Philter recall, UCSF corpus vs. legacy tools	99.46% (Philter) vs. 85.10% (PhysioNet Scrubber)	Nature Digital Medicine, 2020 (Source: www.nature.com)
Philter recall, i2b2 2014 corpus	99.92%	Same study (Source: www.nature.com)
GPT-4 zero-shot de-identification (King Hussein Cancer Center)	Precision 0.9925, recall 0.8318, F1 0.8973	Scientific Reports, 2025 (Source: pmc.ncbi.nlm.nih.gov)
RedactX (Oracle Health) F1-score vs. LLM baselines	0.9646	arXiv preprint, 2025 (Source: arxiv.org)
Re-identification rate, HIPAA Safe Harbor-compliant dataset (name-based)	25% of participants	Technology Science, Sweeney et al., 2017 (Source: techscience.org)
Average cost of a U.S. healthcare data breach (2025)	$7.42 million	IBM/Ponemon Institute, via HIPAA Journal (Source: www.hipaajournal.com)
Average cost of a U.S. data breach across all industries (2025)	$10.22 million	Same source (Source: www.hipaajournal.com)
Global average cost of a data breach (2025)	$4.44 million	Same source (Source: www.hipaajournal.com)
Maximum HIPAA civil penalty per violation category, per year (2026)	$2,190,294	HIPAA Journal, citing HHS inflation-adjusted Tier 4 cap (Source: www.hipaajournal.com)
Minimum HIPAA civil penalty, Tier 1 lack-of-knowledge violation (2026)	$145 per violation	HIPAA Journal (Source: www.hipaajournal.com)
Mid-tier HIPAA civil penalty, Tier 3 willful-neglect violation (2026)	$14,602 to $73,011 per violation	HIPAA Journal (Source: www.hipaajournal.com)
Change Healthcare cyberattack projected cost (2024)	up to $1.6 billion	UnitedHealth Group, via Reuters (Source: www.reuters.com)

Table 3 makes clear that accuracy benchmarks and financial exposure are two sides of the same problem: even a tool reporting 96 percent F1-score still misses roughly 4 percent of PHI spans in a typical note, and at the scale of millions of notes, as UCSF observed, "even at 95% recall (i.e., percent of PHI removed), the amount PHI still remaining across millions of clinical notes would be staggering" (Source: www.nature.com). Set against average healthcare breach costs of $7.42 million and civil penalties that can reach $2.19 million per violation category annually, the cost differential between a 96 percent and an 83 percent F1-score tool is not merely academic; it scales directly with financial and legal exposure once a pipeline processes millions of notes. The Sweeney re-identification study is a further caution that recall figures measured against a benchmark corpus's own PHI taxonomy do not capture quasi-identifier risk, since her team's 25 percent re-identification rate was achieved against data that had already passed Safe Harbor's explicit checklist. The gap between HIPAA's Tier 1 penalty floor of $145 per violation and its Tier 4 ceiling of $2,190,294 per category per year further underscores that regulators calibrate enforcement to culpability, meaning a documented, red-teamed de-identification process is itself a mitigating factor if a re-identification incident is later investigated.

Case Studies and Real-World Examples

Providence Health: 2 Billion Notes at Production Scale

Providence Health, a large U.S. health system, deployed John Snow Labs' Healthcare NLP library to de-identify what is described as "the largest independently validated and peer-reviewed de-identification deployment in healthcare" (Source: www.johnsnowlabs.com), processing 2 billion patient notes. The deployment's validation process combined a three-month external red-teaming engagement, manual review of more than 35,000 notes by the health system's compliance team (Source: www.johnsnowlabs.com), and adversarial testing specifically targeting 790 patients, which recorded zero successful re-identifications (Source: www.johnsnowlabs.com). The pipeline was also able to process 500,000 notes in 2.5 hours, illustrating that clinical-grade accuracy and production-scale throughput are not mutually exclusive when the validation investment is made upfront.

MIMIC-III and MIMIC-IV-Note: The Reference Corpora for Clinical LLM Research

Maintained by MIT and Beth Israel Deaconess Medical Center and published through PhysioNet, the MIMIC family of databases is arguably the most widely cited set of de-identified clinical resources used in academic LLM research. The original MIMIC-III database comprises deidentified data for "over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012" (Source: physionet.org), de-identified with consistent per-patient date shifting so distinctive that patients over age 89 "appear in the database with ages of over 300 years" (Source: physionet.org). Its successor, MIMIC-IV-Note, contains 331,794 deidentified discharge summaries drawn from 145,915 patients and 2,321,355 deidentified radiology reports from 237,427 patients (Source: physionet.org) (Source: physionet.org), and its notes are linkable back to the broader MIMIC-IV structured clinical database, allowing researchers to study language alongside labs, medications, and outcomes without exposing patient identity. All of it has "had protected health information removed in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision" (Source: physionet.org), yet access remains restricted to credentialed users under a formal data use agreement, underscoring that Safe Harbor compliance and unrestricted public release are treated as separate decisions even by the institutions closest to the data.

Vanderbilt University Medical Center's Synthetic Derivative

Vanderbilt's Synthetic Derivative demonstrates de-identification designed from the outset for longitudinal research and, by extension, LLM fine-tuning that benefits from sequential patient history rather than isolated notes. Built through "electronic scrubbing techniques to remove identifiers while maintaining semantic integrity," with names and dates "replaced or shifted in a consistent but anonymized manner" (Source: victr.vumc.org), the resource now contains over 3.9 million de-identified records structured according to the OMOP common data model (Source: victr.vumc.org), and is "integrated with BioVU genomics data" (Source: victr.vumc.org), illustrating how consistent per-patient date-shifting, rather than deletion, preserves the temporal structure that downstream analytics and model training depend on.

King Hussein Cancer Center: Zero-Shot LLM De-Identification in a Non-U.S. Health System

Researchers at King Hussein Cancer Center (KHCC) in Amman, Jordan evaluated GPT-3.5 and GPT-4 for de-identifying real oncology clinical notes under institutional review board approval, finding GPT-4 achieved "a precision of 0.9925, a recall of 0.8318, an F1 score of 0.8973, and an accuracy of 0.9911" while significantly outperforming GPT-3.5 (Source: pmc.ncbi.nlm.nih.gov). Because "data for the study were sourced from the Electronic Health Records (EHR) at King Hussein Cancer Center (KHCC)" (Source: pmc.ncbi.nlm.nih.gov), the study is one of relatively few peer-reviewed evaluations of zero-shot LLM de-identification against a real, non-U.S. oncology dataset rather than a synthetic or heavily curated benchmark, and its comparatively lower recall relative to the JSL Healthcare NLP and Philter figures illustrates the precision-recall tradeoff inherent to general-purpose LLM prompting versus purpose-built clinical NER.

Change Healthcare: The Cost of Getting PHI Protection Wrong

Although not a de-identification case study in the technical sense, the February 2024 ransomware attack on Change Healthcare, a UnitedHealth Group subsidiary that processes a large share of U.S. medical claims and billing data, is the clearest recent illustration of what is financially at stake when clinical data protections fail at scale. UnitedHealth disclosed that it "expects the hack of its Change Healthcare unit to cost the company up to $1.6 billion this year" (Source: www.reuters.com), having "already booked $872 million in costs related to the data breach in the quarter" by mid-April 2024 alone (Source: www.reuters.com). The hack, at "a provider of healthcare billing and data systems and a key node in the U.S. healthcare system," disrupted claims processing nationwide for a month while affecting community health centers that serve "more than 30 million poor and

uninsured patients" (Source: www.reuters.com), and is frequently cited across the healthcare industry as the reference point for why upstream PHI governance, including rigorous de-identification before data leaves protected environments, is treated as a board-level risk issue rather than a purely technical compliance task.

Implications and Future Directions

Three trends are reshaping how organizations will approach clinical note de-identification for LLM analytics over the next several years. First, **LLMs are becoming both the tool and the risk**. The same generative capability that makes GPT-4-class models effective zero-shot de-identifiers also makes fine-tuned clinical language models vulnerable to training-data memorization, since large pre-trained language models are built on "large amounts of parameters that are tuned using vast amounts of training data," and "these factors cause the models to memorize parts of their training data, making them vulnerable to various privacy attacks" (Source: link.springer.com), a concern significant enough that pseudonymizing the training data itself, not just the eventual analytics corpus, is now an active research area for organizations fine-tuning clinical BERT-style and generative models on institutional text. Any organization treating an LLM as a de-identification tool should also evaluate whether that same LLM's downstream fine-tuning pipeline could leak the very identifiers it was asked to remove.

Second, **multimodal de-identification is becoming table stakes rather than a niche capability**. Commercial platforms already extend beyond plain text to PDFs, scanned documents, DICOM medical images, and audio transcripts, reflecting that clinical data entering LLM analytics pipelines increasingly arrives as multimodal records rather than isolated text files. Pipelines validated only against plain-text benchmark corpora such as i2b2 or n2c2 risk missing PHI embedded in these other modalities entirely.

Third, **regulatory frameworks are converging on contextual, risk-based standards rather than fixed identifier checklists**, even as HIPAA's Safe Harbor list remains unchanged since 2000. The EDPB's newly adopted Guidelines 02/2026 on Anonymisation, effective July 2026, explicitly frame anonymity as a contextual determination assessed "from each relevant entity's perspective" rather than a fixed list of fields to strip (Source: www.edpb.europa.eu), a philosophy closer to HIPAA's Expert Determination method than to Safe Harbor. Multinational life-sciences organizations building a single de-identification pipeline intended to satisfy both U.S. and EU requirements should expect that a dataset passing Safe Harbor will not automatically satisfy GDPR anonymisation, and should budget for the additional documentation and contextual risk assessment that a defensible cross-jurisdictional approach requires. Advisory firms working alongside life-sciences and healthcare organizations, including consultancies such as IntuitionLabs that describe their role as delivering "strategic guidance on digital transformation, AI adoption, and technology roadmapping" (Source: intuitionlabs.ai) alongside "enterprise-grade security and privacy protection" (Source: intuitionlabs.ai), are increasingly positioned to help organizations sequence tool selection, validation, and cross-jurisdictional compliance work rather than treating de-identification as a one-time technical checkbox.

Frequently Asked Questions (FAQs)

What is the difference between HIPAA Safe Harbor and Expert Determination for de-identifying clinical notes? Safe Harbor is a rule-based checklist requiring removal of 18 specific identifier categories plus an absence of actual knowledge that remaining data could re-identify a patient (Source: www.hhs.gov). Expert Determination is risk-based: a qualified statistician certifies, using accepted statistical methods, that re-identification risk is very small, which allows retaining more contextual detail but requires paying for that statistical certification, typically ranging "from a few thousand dollars for a well-structured dataset to tens of thousands" for complex data (Source: www.getlimina.ai).

How does PHI removal from free-text clinical notes actually work? Automated pipelines apply named entity recognition, combining regular-expression pattern matching, statistical or neural classifiers, and increasingly LLM prompting, to locate identifier spans in unstructured text, then remove, mask, or replace them with synthetic surrogates. The strongest production systems layer multiple methods rather than relying on any single technique, as Philter's combination of "pattern matching, statistical modeling, blacklists, and whitelists" illustrates (Source: www.nature.com).

Presidio vs. Philter: which is better for clinical de-identification? Presidio is a general-purpose PII framework with pluggable recognizers spanning many countries and industries (Source: microsoft.github.io), useful when an organization needs one platform across clinical and non-clinical PII use cases and is willing to add or tune medical-entity recognizers. Philter is purpose-built for clinical text from the start, evaluated specifically against i2b2 and UCSF clinical corpora, and reported recall above 99% on both (Source: www.nature.com), making it a stronger out-of-the-box choice specifically for hospital discharge and progress notes, while Presidio's flexibility favors organizations with broader, multi-domain PII requirements.

Can de-identified clinical data be used to train LLMs without patient consent? Under HIPAA, data that meets the Safe Harbor or Expert Determination standard is no longer PHI and generally falls outside the Privacy Rule's authorization requirements, which is the regulatory basis for datasets like MIMIC-IV-Note and n2c2 being used broadly for LLM research (Source: physionet.org). However, PhysioNet and n2c2 still require credentialed access and signed data use agreements even after Safe Harbor de-identification, and under GDPR, pseudonymized (as opposed to fully anonymised) data remains personal data subject to consent and lawful-basis requirements (Source: www.edpb.europa.eu), so the answer depends heavily on jurisdiction and on whether the de-identification meets the anonymisation, not merely pseudonymisation, standard.

What are the most common pitfalls of clinical note anonymization? Over-redaction that destroys clinical utility, formatting variance across EHR systems and note types that degrades rule-based generalization, quasi-identifiers such as rare diagnoses that survive an 18-category checklist (Source: techscience.org), mishandled negation and clinical context, overlooked multimodal PHI in scanned documents or audio, and treating a single validation pass as sufficient assurance rather than red-teaming the output, all recur across the tools and case studies discussed in this report.

Is de-identified clinical data ever completely anonymous? Not with certainty. HHS itself states that "even when properly applied," both Safe Harbor and Expert Determination "yield de-identified data that retains some risk of identification" (Source: www.hhs.gov), and Sweeney's field research demonstrated a 25 percent re-identification rate against Safe Harbor-compliant data using external record linkage (Source: techscience.org). De-identification reduces risk to an acceptable, documented level; it does not eliminate it.

Conclusion

De-identifying clinical notes for LLM analytics sits at the intersection of a hard NLP problem and a strict legal standard, and neither half can be solved by treating the other as an afterthought. HIPAA offers exactly two compliant paths, the rule-based Safe Harbor checklist and the risk-based Expert Determination certification, and both explicitly retain some residual re-identification risk even when correctly applied (Source: www.hhs.gov). On the technical side, the strongest tools combine rule-based, statistical, and increasingly LLM-based methods rather than relying on any single approach, and independent benchmarking shows meaningful, quantifiable gaps between vendors on the same underlying task (Source: arxiv.org), gaps that scale directly into financial and legal exposure once a pipeline processes millions of clinical notes (Source: www.hipaajournal.com).

The organizations that get this right, whether relying on open-source frameworks like Presidio and Philter, cloud APIs from Amazon and Microsoft, commercial platforms like John Snow Labs, or reference corpora like MIMIC-IV-Note and the n2c2 datasets, treat de-identification as a governed, continuously validated process rather than a one-time technical step. Providence Health's 2-billion-note deployment reached zero re-identifications only after months of adversarial red-teaming layered on top of an already high-accuracy pipeline, a reminder that accuracy benchmarks are a starting point for building trust, not a substitute for it. As LLMs simultaneously become more capable de-identification tools and more complex privacy risks in their own right (Source: link.springer.com), and as GDPR's contextual anonymisation standard converges with HIPAA's risk-based Expert Determination philosophy (Source: www.edpb.europa.eu), organizations building clinical LLM analytics pipelines should expect the bar for defensible de-identification to keep rising rather than settling into a fixed checklist.

Tags: de-identification, clinical notes, llm analytics, hipaa safe harbor, phi removal, named entity recognition, presidio, philter, expert determination, clinical nlp

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

Leading Claude & ChatGPT AI Enablement and Training: Help biotech and pharmaceutical teams adopt Claude and ChatGPT through practical training, governed workflows, use-case development, and implementation designed for regulated environments.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.

© 2026 IntuitionLabs.ai. All rights reserved.