Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automated de-identification of linguistically-diverse clinical narratives: a retrospective, multi-agent AI framework development and validation study in a global health surveillance network

Yuting Guo

yuting.guo@emory.edu

Emory University

Salman Seyedi

Emory University

Sudeshna Das

Emory University

Jeanne Powell

Emory University

Kiersten Campbell

Emory University

Mohammad Zahid Hossain

icddr,b

Rajib Biswas

icddr,b

Qazi Sadeq-ur Rahman

icddr,b

Gezahegn Mengesha

Haramaya University

Lavender Owiti

Kenya Medical Research Institute

Arsenio Nhacolo

Centro de Investigação em Saúde de Manhiça

Corssino Tchavana

Centro de Investigação em Saúde de Manhiça

Arlindo Malheia

Centro de Investigação em Saúde de Manhiça

Anthony Tobias Okoro

University of Calabar

Iwara Ikpi Arikpo
University of Calabar
Martin Madu Meremikwu
University of Calabar
Abdul Momin Kazi
Aga Khan University
Raheel Rehman Allana
Aga Khan University
Francis Sesay
The ARC for Health
Othman A Bah
The ARC for Health
Ikechukwu Ogbuanu
The ARC for Health
Kimberleigh Storath
University of the Witwatersrand
Imaan Dunn
University of the Witwatersrand
Siobhan Johnstone
University of the Witwatersrand
Katherine Maher
The Task Force for Global Health
Shailesh Nair
The Task Force for Global Health
Upal Nath
The Task Force for Global Health
Matthew A. Reyna
Emory University
Reza Sameni
Emory University
Abeed Sarker
Emory University

Article

Keywords: De-identification, Clinical Notes, Natural Language Processing, Large Language Models

Posted Date: May 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9578929/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Automated de-identification of linguistically-diverse clinical narratives: a retrospective, multi-agent AI framework development and validation study in a global health surveillance network

Yuting Guo[1*], Salman Seyedi[1], Sudeshna Das[1], Jeanne Powell[1], Kiersten Campbell[1], Mohammad Zahid Hossain[3], Rajib Biswas[3], Qazi Sadeq-ur Rahman[3], Gezahegn Mengesha[4], Lavender Owiti[5], Arsenio Nhacolo[6], Corssino Tchavana[6], Arlindo Malheia[6], Anthony Tobias Okoro[7], Iwara Ikpi Arikpo[7], Martin Madu Meremikwu[7], Abdul Momin Kazi[8], Raheel Rehman Allana[8], Francis Sesay[9], Othman A Bah[9], Ikechukwu Ogbuanu[9], Kimberleigh Storath[10], Imaan Dunn[10], Siobhan Johnstone[10], Katherine Maher[11], Shailesh Nair[11], Upal Nath[11], Matthew A. Reyna[1], Reza Sameni[1,2], Abeed Sarker[1,2]

[1*]Department of Biomedical Informatics, Emory University, Atlanta, GA, USA.
[2]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA.
[3]icddr,b, Dhaka, Bangladesh.
[4]College of Health and Medical Science, Haramaya University, Harar, Ethiopia.
[5]Kenya Medical Research Institute, Nairobi, Kenya.
[6]Centro de Investigação em Saúde de Manhíça, Manhíça, Mozambique.
[7]University of Calabar, Calabar, Cross River State, Nigeria.
[8]Aga Khan University, Karachi, Pakistan.
[9]The ARC for Health, Freetown, Sierra Leone.

[10]South African Medical Research Council: Vaccines and Infectious Diseases Analytics Research Unit, University of the Witwatersrand, Johannesburg, South Africa.
[11]The Task Force for Global Health, Decatur, GA, USA.

*Corresponding author(s). E-mail(s): yuting.guo@emory.edu;

Abstract

Background: A key barrier to text-based health data sharing from low- and middle-income countries (LMICs) is the lack of de-identification tools that generalize to geographically diverse linguistic characteristics and documentation practices. This paucity of data exacerbates the global digital divide, limiting the development of artificial intelligence (AI) tools that generalize beyond Global North healthcare settings. We address this limitation by developing and validating a multi-agent AI framework for de-identification of clinical texts from LMICs.

Methods: We included verbal autopsies (VA), comprising structured interviews with deceased children's next of kin or caregivers, and pediatric and maternal clinical abstractions (PCA/MCA) collected from seven LMICs through the Child Health and Mortality Prevention Surveillance network. We created a reference standard via manual annotation of sample documents, guided by the EU-US Data Privacy Framework. We developed a two-stage AI approach for detecting and substituting personally identifiable information (PII) automatically. First, a parallel hybrid design combines an LLM-based PII detector, an open-source off-the-shelf system, and a customizable high-precision rule-based module for high-sensitivity de-identification. A second LLM agent serves as an independent post-processing *reviewer* to identify and correct residual identifiable information through autonomous interactions. We calculated inter-annotator agreement (IAA) using Cohen's $\boldsymbol{\kappa}$, system performance via $\mathbf{F_1}$-score, with bootstrap resampling for estimating 95% confidence intervals (CIs).

Results: The average IAA was $\boldsymbol{\kappa} = \mathbf{0.98}$; the most frequent PII entities were dates, locations, and names. The final system demonstrated consistency across document types for dates and locations, achieving $\mathbf{F_1}$-scores between 0.91 (95% CI 0.90–0.93) and 0.98 (95% CI 0.97–0.98). Name de-identification performance varied significantly, ranging from $\mathbf{F_1}$-score 0.75 (95% CI 0.70–0.80) for structurally heterogeneous VAs to 0.94 (95% CI 0.91–0.96) for MCAs. Qualitative review revealed predominant errors were false positives associated with geolocation-specific ambiguous expressions; rare false negatives resulted primarily from non-standard formats.

Conclusions: Our multi-agent AI architecture provides a reliable strategy for de-identifying complex, heterogeneous clinical narratives from global sources. This validated, generalizable, and publicly-available framework may increase text-based clinical accessibility from LMICs. We recommend a human-in-the-loop workflow to ensure effective and safe application.

Keywords: De-identification, Clinical Notes, Natural Language Processing, Large Language Models

1 Introduction

Advances in artificial intelligence (AI), driven in part by the emergence of large language models (LLMs), have led to a paradigm shift in digital health research by enabling systems that extract, curate, and analyze free text.[1, 2] It is estimated that approximately 80% of healthcare data are unstructured, encapsulating knowledge that remains underutilized.[3] A critical barrier to AI training/application is the difficulty of sharing unstructured health data without compromising patient privacy.[4] Privacy regulations such as Health Insurance Portability and Accountability Act (HIPAA) in the United States (US) and General Data Protection Regulation (GDPR) in Europe require personally identifiable information (PII) to be redacted before data sharing.[5, 6] Manual data de-identification is expensive, slow, and error-prone.[7] Current de-identification systems typically lack generalizability; *domain shift* frequently causes models trained on data from one site to exhibit significant performance degradation when applied to other sites.[8] Limited generalizability is a barrier to multi-site research; its impact is more pronounced in global health contexts.[9]

Current automated de-identification systems are primarily designed for and validated on health systems data from developed countries and the Global North, and they are largely non-generalizable to data from low- and middle-income countries (LMICs). Most tools have been developed using data from Western tertiary care centers (*e.g.*, the MIMIC-III dataset [10]) and fail to capture unique characteristics of health data in the Global South.[11] Western-centric models struggle with the linguistic nuances of free-text datasets from LMICs, including code-mixing (the mixing of local languages with English), non-standard abbreviations, and culture-specific naming conventions that do not appear in standard lookup dictionaries.[12] Furthermore, geolocation-specific identifiers in LMICs often lack the standardized formatting (*e.g.*, ZIP codes) that rule-based de-identification systems rely on, leading to the possibility of privacy leakage or over-redaction. The absence of robust, culture-aware de-identification tools contributes to increasing the digital divide, restricting LMIC data use, and preventing researchers and patients from benefiting from the ongoing AI revolution.

To bridge this gap, we propose a multi-agent, LLM-driven de-identification framework. The framework is hybrid in nature, involving both language models and rule-based modules, and capable of adapting to diverse narrative structures and ambiguous identifiers without requiring large locally annotated training sets. A key innovation of the framework is its multi-agent architecture, with LLM agents that personify a PII detector and a quality assurer that autonomously communicates to de-identify and validate text until all identifiers are removed. In the last step, de-identified entities are substituted with randomly sampled replacements from the same locations, reducing re-identification risks and improving readability. We validate our approach using three types of clinical texts originating from seven LMICs. Our code is made public with this publication and can be deployed using open-source LLMs,

thus eliminating the need for Internet-based access to AI, which may not be compliant with HIPAA and other relevant local and national policies. By providing a trustworthy and scalable solution to the de-identification problem, our work contributes to the globalization of digital health research.

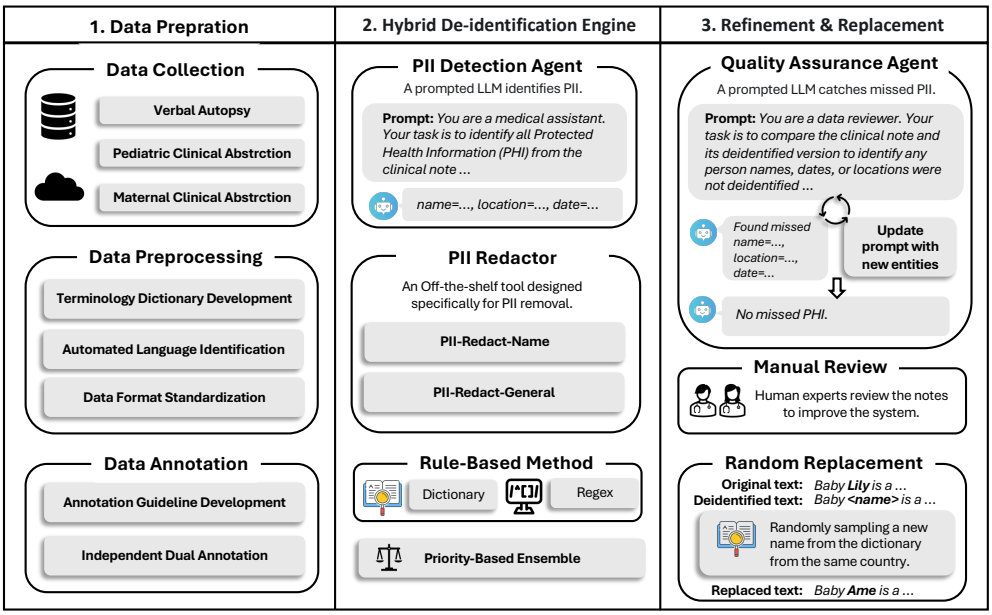


Fig. 1: Overview of our hybrid de-identification, refinement, and replacement pipeline. This workflow first prepares clinical data through collection, preprocessing, and annotation. A hybrid engine then de-identifies the text using an ensemble of a PII detection agent, a dedicated PII redactor, and a rule-based system. Finally, a quality assurance agent iteratively corrects any missed PII, and the redacted text is updated with realistic, randomly-sampled replacements.

2 Methods

2.1 Data sources

We sourced data from the Child Health And Mortality Prevention Surveillance (CHAMPS) network, which collects clinical data in high-mortality, resource-limited settings in Africa and South Asia to determine causes of stillbirths and child deaths and inform targeted prevention strategies.[13] Data collection is conducted by local teams that perform interviews and clinical data abstraction. The data exhibits substantial linguistic diversity, including frequent code-mixed English (*i.e.*, text that switched between English and one or more local languages) and colloquial terminology. We utilized three types of clinical narratives from seven international sites including Bangladesh, Ethiopia, Kenya, Mozambique, Nigeria, Sierra Leone, and South Africa:

- **Verbal autopsies:** Unstructured narrative accounts of circumstances surrounding death, obtained through next of kin or caregiver interviews and characterized by site-specific colloquial language.
- **Pediatric clinical abstractions:** Semi-structured summaries of clinical findings and events for pediatric cases.
- **Maternal clinical abstractions:** Semi-structured summaries of maternal information related to the pediatric cases.

2.2 Manual annotation and gold standard creation

We manually annotated a gold-standard corpus by randomly sampling documents from each country. Purely non-English documents were excluded (*e.g.*, VAs from Mozambique and PCAs from Bangladesh). Nigerian VAs (7 cases) were also excluded due to insufficient sample size for meaningful analysis.

Initially, two members of the study team independently reviewed a shared sample of ten texts. This process involved three iterative cycles of review with progressively larger samples of 20 and 40 texts in the second and third cycles, respectively. Inter-annotator agreement (IAA) was measured using Cohen's κ[14], with disagreements resolved during weekly consensus meetings. After reaching acceptable IAA ($\kappa \geq 0.80$[15]) and guideline consensus, representative documents from each text type were independently annotated by two reviewers to identify all instances of PII. The complete guidelines are provided as supplementary material S1. Detailed examples are provided in Tables S1 through S5. Annotated data were used for quantitative system evaluation, and unannotated data for qualitative expert review. The annotated subset was designed for stratified system validation across countries and document types rather than for estimating population-level prevalence of PII or making formal inferences about all linguistic subgroups.

2.3 System architecture

Our de-identification system follows the defense-in-depth design,[16] reducing risks to privacy by eliminating single points of failure through the coordinated deployment of multiple modules. The full architecture is presented in Fig. 1. We employed a human-in-the-loop framework throughout system development: subject matter experts from each country reviewed the redacted outputs and provided feedback, leading to iterative refinement of site-specific dictionaries and rules. The system was tuned for high recall (sensitivity) of de-identified PII at the expense of precision (positive predictive value; PPV).[1]

2.3.1 Multi-component PII detection

The initial de-identification is performed by three complementary, parallel modules:[2]

[1] In this context, a true positive is defined as a PII entity correctly identified and redacted by the system, while a false negative represents a PII entity that remained unredacted. A false positive occurs when non-sensitive text is incorrectly flagged and redacted.

[2] In our internal experimental documentation, this data was denoted as Protected Health Information (PHI) for prompt engineering. We have adopted the term PII for this publication to more accurately reflect the nature of the data, which includes identifiers beyond the scope of healthcare information.

- **LLM-based PII detection agent:** A locally-hosted, zero-shot LLM agent (DeepSeek-V3; 685B parameters and 2.71 bit quantization [17]) identifies PII using contextual knowledge without task-specific training. Prompts were iteratively refined using an initial set of 50 documents and framed for the perspective of a medical assistant instructed to flag PII. The details about prompt engineering and the prompts are provided in supplementary materials S2.
- **Piiredact:** Piiredact[18] is a locally executed, open-source de-identification tool that detects a wide range of identifiers (*e.g.*, names, addresses), with proven effectiveness for clinical notes.
- **High-precision rule-based engine:** A deterministic module captures site-specific identifiers using curated dictionaries (*e.g.*, names, locations), exclusion lists to prevent common false positives (*e.g.*, phrases like "high care" for a location in South Africa VAs), and regular expressions for consistent formats, such as dates.

2.3.2 Ensemble and consolidation

Any text span flagged by at least one detector is marked for redaction by a central ensemble module. When overlapping spans differ in length, the system selects the longest span to ensure full coverage.

2.3.3 Quality assurance agent

A separate LLM agent (Llama-3 with 70B parameters[19]) compares the original and de-identified texts to identify residual PII. If additional PII is detected, the text is masked and then reprocessed. This automated review loop repeats until the agent detects no PII.

2.3.4 Entity replacement

All detected PII is replaced with realistic, randomly generated surrogate entities drawn from data from the same country. Dates are replaced with random values while preserving temporal order. This step reduces re-identification risk as it ensures that any PII missed by the system is distinguishable from the replaced entity. A secondary benefit is that the narratives remain coherent following this step.

2.3.5 Rationale for model selection

Our model selection was guided by three criteria: open-source availability, established performance, and compatibility with computational infrastructure. DeepSeek-V3 and Llama-3 met these requirements. We conducted preliminary experiments involving prompt engineering on a 50-case dataset which revealed that DeepSeek-V3 held a slight accuracy advantage over Llama-3 for PII detection. Based on this finding, we evaluated different model pairings for the PII detection and quality assurance agents, choosing the configuration that provided the best trade-off between detection accuracy and computational expense.

2.4 System evaluation

We evaluated system performance against the manually-annotated gold standard, excluding the samples used for guideline preparation and system development. Evaluation was performed at the entity-token level; each instance of a true PII entity was independently assessed. An entity tag was considered incorrect if any token within the relevant span was misclassified; specifically, as a *false negative* if any portion of a true PII entity remained unredacted and a *false positive* if any non-PII text was incorrectly redacted.

We measured performance using recall, precision, and F_1-score (details in supplementary materials S3). Recall measures the system's success at preventing data leaks; precision captures the extent of over-redaction. Statistical uncertainty was estimated using 95% CIs computed via bootstrap resampling with 1000 iterations.[20]

2.5 Error analyses

We conducted a qualitative error analysis using confusion matrices to characterize false positives (over-redaction) and false negatives (missed redaction). Errors were characterized into recurring patterns, including boundary errors, complete entity omissions, and spurious detections. This analysis identified areas for iterative improvements and limitations for future work.

3 Results

3.1 Dataset descriptive statistics

3.1.1 Document-level statistics

Seven countries contributed data: Bangladesh, Ethiopia, Kenya, Mozambique, Nigeria, Sierra Leone, and South Africa. The datasets comprised 6,028 VAs (351 annotated), 14,569 MCAs (560 annotated), and 16,107 PCAs (445 annotated).

3.1.2 Entity-level statistics

Detailed corpus statistics are summarized in Table 1. The most frequently annotated entities were locations, names, and dates. Other identifiers defined under the EU-US Data Privacy Framework and the US-EU Safe Harbor Framework[21, 22] were either absent (*e.g.*, phone numbers) or rare (*e.g.*, medical record numbers; ¡10 occurrences).

The data volume varied across countries. South Africa, Kenya, Ethiopia, and Sierra Leone provided the largest volumes of data. Kenya and South Africa were primary sources of annotated PCAs and VAs, respectively. Date entities were most prevalent in data from South Africa and Kenya; name entities were the least frequent overall. Narrative length also varied substantially, ranging from a mean of 21.6 words in Mozambique MCAs to 186.7 words in Sierra Leone VAs.

Table 1: Summary statistics of the number of all narratives (# Doc.), annotated narratives (# Ann.), annotated entities (# Location, # Name, # Date), and the average word length per narrative for each country for Verbal Autopsies, Maternal Clinical Abstractions, and Pediatric Clinical Abstractions.

Country	# Doc.	# Ann.	# Location	# Name	# Date	Avg. Length
Verbal Autopsies						
Ethiopia	2,297	98	143	62	0	42.7
Kenya	1,198	62	84	10	3	95.2
Sierra Leone	1,399	91	187	13	48	186.7
South Africa	1,134	100	244	24	242	150.1
Maternal Clinical Abstractions						
Bangladesh	6,073	91	118	19	178	47.1
Ethiopia	2,669	77	101	8	25	96.5
Kenya	1,807	99	89	13	269	97.2
Mozambique	91	54	21	0	6	21.6
Nigeria	51	51	12	0	24	35.7
Sierra Leone	2,454	97	54	13	162	68.4
South Africa	1,424	91	69	8	102	44.3
Pediatric Clinical Abstractions						
Ethiopia	2,245	74	69	12	6	63.3
Kenya	4,055	88	118	17	334	90.2
Mozambique	468	50	21	0	17	36.9
Nigeria	109	36	12	0	13	29.8
Sierra Leone	3,695	99	95	40	124	89.5
South Africa	5,535	98	69	6	97	168.2

3.2 System performance

We restricted quantitative analysis to the three categories of frequently-occurring PIIs: location, name, and date. The final IAA was $\kappa = 0.98$, representing almost-perfect agreement.[23] Summary performance metrics for the overall system are presented in Fig. 2 with detailed metrics in the supplementary material S4, including performance grouped by dataset and entity type for each country in Table S6 and module-by-module performance analysis in Tables S7 through S9.

The system demonstrated a strong overall efficacy; however, its performance varied significantly depending on entity type, source document category, and country of origin. Across all three corpora, the "date" entity was most consistently and accurately identified, achieving high F_1-scores of 0.96 (95% CI 0.95–0.97) for VAs, 0.98 (95% CI 0.97–0.98) for MCAs, and 0.97 (95% CI 0.97–0.98) for PCAs, with narrow CIs indicating high reliability.

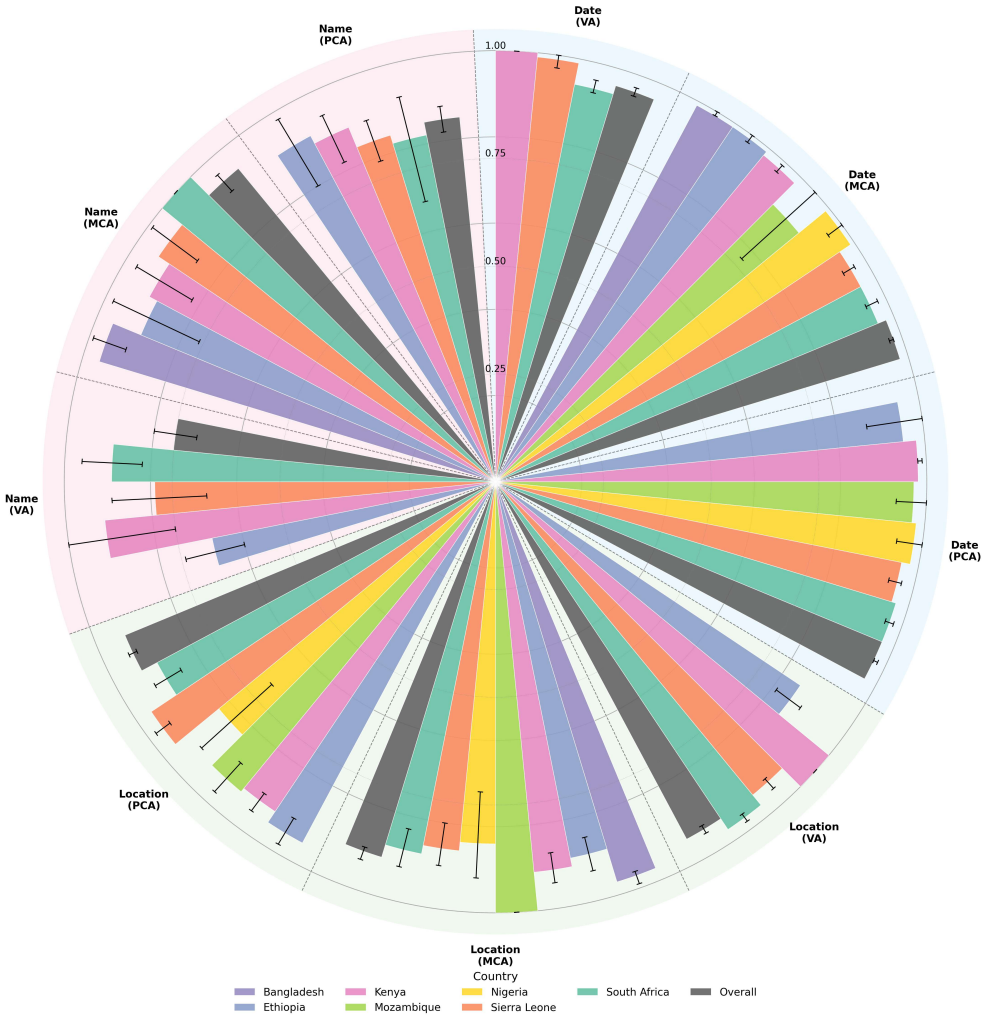


Fig. 2: Performance of the de-identification system (named entity recognition) on the annotated data, grouped by dataset and entity type. Each bar represents the F_1-score with the 95% confidence interval for a specific country.

The "location" entity was also robustly detected across datasets, with overall F_1-scores of 0.94 (95% CI 0.93–0.95) for VA, 0.91 (95% CI 0.90–0.93) for MCAs, and 0.93 (95% CI 0.92–0.94) for PCAs. A notable pattern for the "location" entity was a trade-off between precision and recall. This trend was most pronounced in the Nigerian data, where perfect recall of 1.00 (95% CI 1.00–1.00) was coupled with relatively lower precision scores of 0.72 (95% CI 0.58–0.85) for MCAs and 0.71 (95% CI 0.55–0.85) for PCAs.

Significant performance variation was observed for "name" entities, where the document context was a decisive factor. In the VA corpus, "name" was the most challenging entity, achieving the lowest overall F_1-score of 0.75 (95% CI 0.70–0.80), primarily constrained by a low precision of 0.61 (95% CI 0.55–0.67). Performance improved markedly for PCAs, with F_1-score of 0.84 (95% CI 0.82–0.88) and MCAs, attaining F_1-score of 0.94 (95% CI 0.91–0.96).

Performance also varied across geographic regions. For Kenya, for instance, our approach consistently demonstrated high performance across most categories. In contrast, high variability was observed in the South African data for the "name" entity: it achieved a perfect F_1-score of 1.00 (95% CI 1.00–1.00) for MCAs but significantly lower F_1-score of 0.82 (95% CI 0.67–0.92) for PCAs. This discrepancy was driven by low precision of 0.69 (95% CI 0.50–0.88), despite perfect recall in both instances. Such lower precision patterns for "location" and "name" entities were also evident in data from Ethiopia (VAs) and Nigeria (M/PCAs), highlighting that regional linguistic nuances or differences in data can significantly impact model performance.

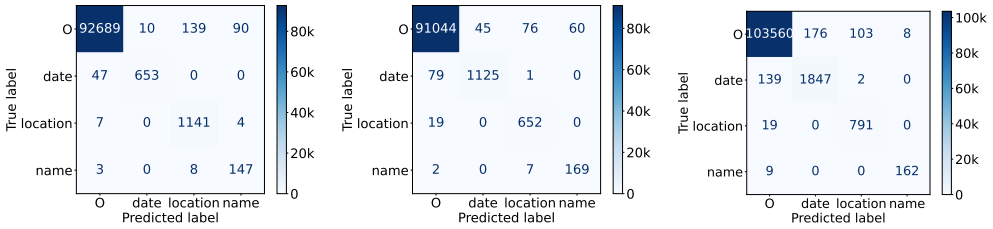
3.3 Error analyses

Confusion matrices (Fig. 3) illustrate systematic misclassification patterns. We identified predictable patterns in both false positives and negatives, primarily linked to linguistic and structural data diversity. Most common errors were false positives for names, particularly in the less-structured VA narratives. The system occasionally confused culturally pertinent terms or non-English expressions with names. For instance, it incorrectly flagged "Chira" (a term for the consequence of disobeying tradition) and "Ubani" (a type of ceremonial incense) as names in the Kenyan VA data.

False negatives, the more critical error of missed PII, were less frequent. For locations, the approach sometimes failed to identify uncommon local place names (*e.g.*, "Amuma ganda Jaalo" in Ethiopia) and complex expressions that included numbers (*e.g.*, "harar, jenela kebele 16"). It also struggled when key contextual words were missing; for example, it missed "water" as a location in the phrase "the baby was sent to water," where "water" was an abbreviated facility name used without a keyword like "clinic". For dates, false negatives occurred primarily due to variations in format (*e.g.*, DD-MM-YY *vs.* DD/MM/YY) or when only a partial date was provided (e.g., "On *5th*...").

4 Discussion

We developed and validated a medical text de-identification system suitable for LMICs using data from seven countries with diverse linguistic characteristics, including code-mixed language. To our knowledge, this is the most accurate de-identification system for such data, and the first successful application of multi-agent AI for this task. Robust performance across countries and diverse data types suggests that the approach is generalizable to geographically diverse English texts. Although the annotated subset comprised 5.8% of VAs, 3.8% of MCAs, and 2.8% of PCAs, it was drawn across all eligible country-document strata and produced consistently narrow 95% confidence



(a) Verbal Autopsies (b) Pediatric Abstractions (c) Maternal Abstractions

Fig. 3: Confusion matrices illustrating the de-identification system's performance on the annotated data for the three clinical document types. The matrices show the classification performance for individual tokens (words or sub-words) across all PII categories (date, location, and name) and non-PII categories (O). The matrix entries are the numbers of tokens.

intervals for the most frequent PII categories, supporting its adequacy for system-level validation.

Our work addresses a significant gap in current research. While recent advances in AI are being actively leveraged for health research and applications, such efforts are concentrated in a handful of countries with the required resources. LLMs that power AI systems are primarily trained on text-based data from these countries, contributing to an already widening global digital divide. An effective de-identification system can facilitate the sharing and use of data from LMICs, serving as a catalyst for research, development and application customized for such contexts. To promote such activities, we have made our framework, along with resources we created, open-source, allowing end-users to embed their choice of language models.

Our findings indicate that system performance is regulated by multiple factors including document context, entity type, and geographic origin, suggesting that the overall success of the system is contingent on contextual factors. Our developed system has a deliberate tendency toward high recall, particularly for "location" and "name" entities. By prioritizing recall over precision, the system is optimized to minimize risk of PII leakage. The "date" entity is relatively trivial to detect due to standardized formatting, whereas the "name" entity exhibited the highest variance across corpora. The system's relative underperformance on VAs, compared to the more structured PCAs and MCAs, suggests that document format is a key performance determinant. VAs often contained a mix of languages, more colloquial phrasing, and complex sentences with less predictable inclusion of PII, sometimes within ambiguous contexts. In contrast, PCAs and MCAs were more likely to follow a semi-structured format and have consistent syntactic patterns, providing clearer signals for detection. Context-dependent variations in performance and the deliberate trade-off between recall and precision suggest that a human-in-the-loop approach is perhaps the best real-world deployment strategy, with the system performing large-scale redactions followed by human review for detecting error patterns and providing quality assurance.

Performance variability across data from different countries highlights the limitations of a single, universal model. Disparities, such as the divergent performance on South African "name" data between PCAs and MCAs, likely stem from a combination of factors, including regional linguistic diversity in naming conventions, dialects, and data recording practices. These findings strongly argue against a "one-size-fits-all" approach and point to the need for region-specific customizations, justifying our customizable framework design that enables incorporation of region-specific vocabularies and language models, if available, for local context alignment.

Random replacement of detected PII significantly mitigates the impact of information leakage. Any PII missed by the system is virtually indistinguishable from substituted pseudo-information. However, we have observed a potential vulnerability in this method. For the substitution to be effective, the lexicons used for replacement need to be sufficiently large and diverse (*e.g.*, via the use of comprehensive databases such as SimpleMaps[24]) so that patterns cannot be determined following de-identification.

4.1 Implications for global health research

By automating the redaction of sensitive free-text narratives, our de-identification approach has the potential to unlock vast, previously inaccessible resources for high-impact global health research while ensuring patient privacy. LMICs, including from Africa and Asia, bear substantially larger burdens of communicable and non-communicable diseases globally compared to their Global North counterparts.[25] AI-driven clinical and public health tools have immense potential to improve research and quality of care in these regions when trained and validated on local data.[26] However, the availability of relevant data from most LMICs remains sparse, resulting in disparity in benefits of advances in AI.[27, 28] Current AI models exhibit significant cultural bias, capturing Euro-centric values,[29] and lacking the nuances of LMICs. Our work represents a stepping stone for more globally equitable health data and AI application—access to data facilitated by de-identification systems built on our framework may enable the creation of more globalized models in the short term, and help increase the application of AI for clinical and public health efforts in the longer term.

4.2 Ethical considerations and governance

While our de-identification approach may enable large-scale data sharing, improving global research inclusivity, it must be embedded within broader governance frameworks to protect patient communities. The approach focuses on individual-level PII redaction without explicitly addressing the risk of group harm. Even with PII replaced, release of large datasets may enable analyses that inadvertently stigmatize entire communities, for example, by linking a specific de-identified hospital that serves a known geographic area to a high prevalence of a condition (*e.g.*, HIV). Therefore, the use of this technology must be governed by robust ethical oversight, and, where appropriate, stakeholder advisory boards, to ensure that potential benefits of research outweigh risks to stakeholders.

When the approach is applied to data from countries not included in this study, it is crucial to validate its performance prior to release of data. We recommend the use of a protocol similar to ours, where a small subset of the data is manually prepared for performance validation prior to large-scale system deployment followed by site-specific reviews. Ideally, this framework should not be viewed as a standalone solution, but as a component of a comprehensive data stewardship program.

4.3 Practical implementation

Given that no automated system can achieve absolute perfection, the most responsible pathway for deploying this technology in a real-world setting is a human-in-the-loop workflow. In such a workflow, the system performs the initial, labor-intensive task of de-identifying a large dataset. The output is then presented to trained human reviewers. The reviewer's role would transform from the tedious task of finding every piece of PII from scratch to the much faster and more manageable task of auditing the AI's work, confirming correct redactions and catching error patterns. This hybrid approach represents a pragmatic solution, which is scalable, cost-effective, and provides a high standard of privacy protection for translation into clinical and research practice.

5 Limitations and future directions

Despite the strong system performance, this study has several limitations. A primary limitation is that the system's capabilities were validated on three specific document types from the CHAMPS network. While diverse, these documents do not represent the full spectrum of clinical text originating from the Global South. Further validation is warranted before this system can be confidently deployed on other formats with unique linguistic and structural characteristics, such as the highly specialized terminology of pathology reports or the abbreviation-laden style of emergency department notes. Furthermore, performance could vary between different institutions, even for the same document type, due to differences in local electronic health record documentation practices.

Future work should proceed along several parallel tracks. The immediate priority is to improve the system's precision on highly unstructured narratives like VAs. This involves developing more sophisticated methods, such as advanced prompt engineering and the integration of culture-specific dictionaries, to help the model better distinguish between personal names and culturally significant terms identified in our error analysis. Looking further ahead, an important goal is to create a framework for rapidly and cost-effectively adapting the system to new document types and languages, including purely non-English texts. Finally, as this technology matures, it will be imperative to conduct rigorous audits for algorithmic bias. We must ensure the system performs equitably and offers the same high level of privacy protection across patient notes from different demographic, socioeconomic, and geographic backgrounds.

6 Conclusion

This study demonstrates that a hybrid AI system, incorporating a novel AI-powered verification layer, provides a robust and scalable solution for de-identifying complex clinical narratives from LMICs. Our approach addresses a critical and long-standing bottleneck, offering a trustworthy mechanism to protect patient confidentiality while unlocking the immense potential of unstructured clinical data from LMICs for research. By enabling the safe inclusion of these rich data sources, this work paves the way for more globalized AI models, new discoveries in global health, patient safety, and health equity, moving the field closer to a future where data-driven insights can be ethically leveraged to improve health outcomes for all.

Acknowledgements. This work was funded by grant OPP1126780 from the Bill & Melinda Gates Foundation through Emory University. The authors would like to thank Cyndy Whitney and Chelsea Cole for their support and contributions to this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agency.

Declarations

- **Funding:** This work was funded by the Bill & Melinda Gates Foundation (grant number OPP1126780) through Emory University.
- **Competing interests:** The authors declare no competing interests.
- **Ethics approval and consent to participate:** Not applicable (secondary analysis of de-identified data; no direct human subjects involvement).
- **Consent for publication:** Not applicable.
- **Data availability:** The annotated datasets used for system evaluation are not publicly available due to patient privacy restrictions. Summary statistics and system outputs are provided in the supplementary materials.
- **Code availability:** The de-identification framework code will be made publicly available upon publication.
- **Author contributions:** Y.G. led system development, evaluation, and manuscript preparation. S.S., S.D., J.P., and K.C. contributed to system development and evaluation. M.Z.H., R.B., Q.S.R., G.M., L.O., A.N., C.T., A.M., A.T.O., I.I.A., M.M.M., A.M.K., R.R.A., F.S., O.A.B., I.O., K.S., I.D., S.J., K.M., S.N., and U.N. provided domain expertise and site-specific data curation. M.A.R., R.S., and A.S. supervised the project and contributed to manuscript preparation. All authors reviewed and approved the final manuscript.

References

[1] Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nature Medicine* **29**, 1930–1940 (2023). URL https://www.nature.com/articles/s41591-023-02448-8.

[2] Zhang, K. *et al.* Revolutionizing health care: The transformative impact of large language models in medicine. *Journal of Medical Internet Research* **27**, e59069 (2025). URL https://www.jmir.org/2025/1/e59069.

[3] Kraljevic, Z. *et al.* Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health* **6**, e281–e290 (2024). URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(24)00025-6/fulltext.

[4] Wiest, I. C., Leßmann, M.-E., Wolf, F., Ferber, D. *et al.* Deidentifying medical documents with local, privacy-preserving large language models: The llm-anonymizer. *NEJM AI* **2** (2025). URL https://ai.nejm.org/doi/full/10.1056/AIdbp2400537.

[5] U.S. Department of Health and Human Services. Standards for privacy of individually identifiable health information. Code of Federal Regulations (2013). URL https://www.hhs.gov/hipaa. 45 CFR Parts 160 and 164.

[6] European Parliament and Council of the European Union. Regulation (eu) 2016/679 (general data protection regulation). Official Journal of the European Union (2016). URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. OJ L 119, pp. 1–88.

[7] Murugadoss, K., Gupta, S., Jonnagaddala, J., Coiera, E. & Ye, Z. Building a best-in-class automated de-identification tool for unstructured text. *Patterns* **2**, 100255 (2021). URL https://www.cell.com/patterns/fulltext/S2666-3899(21)00081-7.

[8] Stan, S. & Rostamian, A. Preserving fairness in ai under domain shift. *Journal of Artificial Intelligence Research* **81**, 907–934 (2025). URL https://jair.org/index.php/jair/article/view/16694.

[9] Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* **2**, e489–e492 (2020). URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30186-2/fulltext.

[10] Johnson, A. E. W. *et al.* Mimic-iii, a freely accessible critical care database. *Scientific Data* **3**, 160035 (2016). URL https://www.nature.com/articles/sdata201635.

[11] Edgcumbe, A. *et al.* 'potato potahto'? disentangling de-identification, anonymisation, and pseudonymisation for health research in africa. *J. Law Biosci.* **12**, lsae029 (2025).

[12] Lee, Y.-Q. *et al.* Unlocking the secrets behind advanced artificial intelligence language models in deidentifying chinese-english mixed clinical text: Development and validation study. *Journal of Medical Internet Research* **26**, e48443 (2024). URL https://www.jmir.org/2024/1/e48443.

[13] Salzberg, N. T. *et al.* Mortality surveillance methods to identify and characterize deaths in child health and mortality prevention surveillance network sites. *Clin. Infect. Dis.* **69**, S262–S273 (2019).

[14] Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46 (1960).

[15] Neuendorf, K. A. *The Content Analysis Guidebook* 2nd edn (SAGE Publications, Thousand Oaks, CA, 2017).

[16] Ross, R., Winstead, M. & McEvilley, M. Engineering Trustworthy Secure Systems. NIST Special Publication 800-160v1r1, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA (2022). URL https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-160v1r1.pdf.

[17] DeepSeek-AI. Deepseek-v3-0324 (2024). URL https://huggingface.co/deepseek-ai/DeepSeek-V3-0324. Accessed: 2025-10-31; model card cites "DeepSeek-V3 Technical Report" (arXiv:2412.19437) as underlying documentation.

[18] OpenPipe. "a sota pii redaction tool for your laptop" (2025). URL https://github.com/OpenPipe/pii-redaction. Accessed: 2025-10-31.

[19] Meta Platforms, I. Llama 3.1 70b instruct (2024). URL https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct. Accessed: 2025-10-31; Release date: July 23, 2024.

[20] Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**, 1–26 (1979).

[21] U.S. Department of Commerce. Data privacy framework program overview. https://www.dataprivacyframework.gov/Program-Overview (2023). Accessed: 2026-03-28.

[22] U.S. Department of Health and Human Services. Standards for privacy of individually identifiable health information. 45 CFR §164.514(b)(2) (2002). URL https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.514. HIPAA Privacy Rule, Safe Harbor method for de-identification.

[23] Viera, A. J. & Garrett, J. M. Understanding interobserver agreement: The kappa statistic. *Family Medicine* **37**, 360–363 (2005).

[24] Simplemaps.com. World cities database. Online database (2025). URL https://simplemaps.com/data/world-cities. Accessed: 2025-02-15.

[25] Gebremedhin, L. T. Investment in health data can drive economic growth. *Nature Medicine* **28**, 2000 (2022). URL https://www.nature.com/articles/s41591-022-02022-8.

[26] Ciecierski-Holmes, T. *et al.* Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *npj Digital Medicine* **5**, 162 (2022). URL https://www.nature.com/articles/s41746-022-00700-y.

[27] Lee, J. *et al.* Interventions to improve district-level routine health data in low-income and middle-income countries: a systematic review. *BMJ Global Health* **6**, e004223 (2021). URL https://gh.bmj.com/content/6/6/e004223.

[28] Abdul-Rahman, T. *et al.* Inaccessibility and low maintenance of medical data archive in low-middle income countries: Mystery behind public health statistics and measures. *Journal of Infection and Public Health* **16**, 1556–1561 (2023). URL https://www.sciencedirect.com/science/article/pii/S1876034123002356.

[29] Tao, Y., Viberg, O., Baker, R. C. & Kizilcec, R. F. Cultural bias and cultural alignment of large language models. *PNAS Nexus* **3**, pgae346 (2024). URL https://academic.oup.com/pnasnexus/article/3/9/pgae346/7756548.

Appendix S1 Annotation Guidelines

The task involved removing Personally Identifiable Information (PII) from CHAMPS infant mortality reports. The manually reviewed and de-identified subset contained PII limited to personal names, geographic locations, and dates. To ensure consistency, multi-word PII spans were replaced with single redaction tokens (e.g., January 1st was replaced with `<date>`).

S1.1 Location

1. Names of medical facilities were also removed, but the facility type—such as hospital, health centre, or dispensary—was retained to preserve contextually relevant information.
2. In cases where the proper name of a facility was interrupted by the facility type (e.g., Hospital-name Hospital at Town), the redaction preserved the surrounding structure using multiple tokens (e.g., `<location>` hospital at `<location>`).
3. Standard hospital unit names such as ICU were not de-identified.
4. In contrast, names of government and private hospitals were redacted due to the potential for re-identification in settings with a single such institution.

Table S1: Location de-identification examples.

Original Example	**De-identified Example**
xyz hospital xyz	`<location>` hospital `<location>`
xyz	`<location>`
Bob hosp.	`<location>` hosp.
the xyz Hospital	the `<location>` Hospital
xyz Health Centre	`<location>` Health Centre
xyz Referral	`<location>`
nearby private clinic	nearby `<location>` clinic
The private doctor	The `<location>` doctor
A private clinic	A `<location>` clinic
Local clinic	`<location>` clinic
Favor xyz	`<location>`
Healthcare/medicinal facilities of different types	`<location>` medical center
xyz clinic	`<location>` clinic
xyz health clinic	`<location>` health clinic
xyz dispensary	`<location>` dispensary
xyz pharmacy	`<location>` pharmacy
Dr. xyz clinic	`<location>` clinic

Table S2: Common locations and region-specific names that should not be removed.

Content	Example
Common hospital unit names	ANC, NICU, SCBU, KMC, RESUS
Region-specific names for vehicles	okada, boda boda, tuk tuk, matatu
Region-specific clinic types	CHC, CHP, PHU
Mentions of CDC/CHAMPS/country name	CDC, CHAMPS, Nigeria
Words for nursery/daycare	Creche
Region-specific words	Manyasi (an herb mixture)

Ward names were removed in case they were specific to the hospital (e.g., "ward 66").

S1.2 Date

1. The day and month were removed, while the year was retained to preserve temporal context without compromising privacy.
2. Standalone days and/or months were also removed.

Table S3: Date de-identification examples.

Original Example	De-identified Example
Monday	Monday
January 2019	`<date>` 2019
saturday, june 2024	saturday, `<date>` 2024
26th of may 2023	`<date>` 2023
14/06/2023	`<date>`/2023
26th of December	`<date>`
the 26th of December	the `<date>`
the 3rd and 4th of February 2024	the `<date>` and `<date>` 2024
On the 8th of the month	On the `<date>` of the month

Table S4: Content which should NOT be removed.

Content	Example
Age/Gestational age (months or weeks and days)	35 yrs old, baby born in 8 weeks
Time (hours/minutes)	1pm, 11:00

S1.3 Names

All personal names were removed, including those of infants, parents, medical providers, and other identified individuals.

Table S5: Name de-identification examples.

Original Example	De-identified Example
xyz's death	<name>'s death

Appendix S2 Iterative Prompt Engineering for PII Extraction

To optimize the performance of the LLM in identifying and handling PII within unstructured clinical narratives, an iterative prompt engineering strategy was employed. The prompt architecture progressed through three distinct developmental phases, transitioning from zero-shot inline redaction to dynamic, retrieval-augmented few-shot extraction. The objective of this iterative process was to improve extraction accuracy, reduce hallucinations, and enforce strict output formatting for downstream evaluation.

S2.1 Baseline Zero-Shot Redaction

The initial baseline approach (Prompt 1) utilized a zero-shot framework instructing the LLM to act as a medical assistant. The primary task was formulated as a direct text transformation: the model was asked to read the clinical note and generate a fully de-identified version by replacing specific PII with generalized tags. The prompt explicitly defined the scope of the task by categorizing the target PII into three specific types: Name (patients and employers), Location (including government hospitals), and Date (specifically day and month, excluding weekdays and years). To guide the model, brief, inline text examples were provided, demonstrating how specific entities should be replaced by generic tokens (e.g., replacing "xyz hospital xyz" with <location> hospital <location>). While conceptually straightforward, instructing the model to rewrite the entire text sequence increases the risk of unintended formatting alterations and complicates automated performance evaluation.

```
You are a medical assistant. Read the clinical note carefully. Remove all
    Protected Health Information (PHI) from the clinical note and output a
    deidentified version of the clinical note. Preserve all non-PHI text as
    they appear in the original text. Do not standardize, reformat, or alter
    the text in any way.

Types of PHI to identify:
1. Name: Specific names of patients and employers.
2. Location: Specific locations including a government hospital.
3. Date: Dates containing a specific day and month. Does not count weekdays
    and year.

Examples:
Input: ###START nightingale hospital kondele END###
```

```
Output: ###START <location> hospital <location> END###

Input: ###START Bob hosp. END###
Output: ###START <location> hosp. END###

Input: ###START January 2019 END###
Output: ###START <date> 2019 END###

Input: ###START 14/06/2023 END###
Output: ###START <date>/2023 END###

Input:###START {{text}} END###
```

Prompt 1: Baseline Zero-Shot Redaction Prompt

S2.2 Structured Zero-Shot Extraction

To address the limitations of the baseline redaction approach, the second iteration (Prompt 2) shifted the core objective from text rewriting to structured entity extraction. The instructions were refined to demand that the model identify PII and output it strictly as key-value pairs (e.g., `<PII type>=PII terms`). To enhance the model's comprehension of complex de-identification rules—particularly regarding the partial extraction of dates and specific hospital names—the formatting of the in-context examples was significantly upgraded. Static examples were presented utilizing structured Markdown tables, delineating the "Original Example" alongside the "De-identified Example." This structural constraint was designed to force the LLM to focus solely on the precise spans of text requiring extraction, thereby minimizing the generation of extraneous text and ensuring a standardized output format that is easily parsable for algorithmic scoring.

```
You are a medical assistant. Your task is to identify all Protected Health
    Information (PHI) from the clinical note as below.

Types of PHI to identify:
1. Name: Specific names of patients and employers.
2. Location: Specific locations including a government hospital.
3. Date: Dates containing a specific day and month. Does not count weekdays
    and year.

Examples:
### Hospital Deidentification
| Original Example                | Deidentified Example              |
|---------------------------------|-----------------------------------|
| xyz hospital xyz                | <location>=xyz, xyz               |
| xyz                             | <location>=xyz                    |
| Xyz hosp.                       | <location>=Xyz                    |
| the Xyz Hospital                | <location>=Xyz                    |
```

```
---

### Date Deidentification Rules
| Original Example                  | Deidentified Example               |
|-----------------------------------|------------------------------------|
| January 2019                      | <date>=January                      |
| saturday, june 2024               | <date>=june                         |
| 26th of may 2023                  | <date>=26th of may                  |
| 14/06/2023                        | <date>=14/06                        |
| the 26th of December              | <date>=26th of December             |

Preserve all PHI terms as they appear in the original text. Do not standardize
    , reformat, or alter the terms in any way. Your answer should be formatted
     as "<PHI type>=PHI terms". For example:

"""
<NAME>=John Doe
<LOCATION>=123 Main St
"""

Clinical note:
{{text}}

Please provide your answer below:
```

Prompt 2: Structured Zero-Shot Extraction Prompt

S2.3 Dynamic Few-Shot Prompting via Retrieval

The final iteration (Prompt 3) transitioned from relying on static, zero-shot instructions to a dynamic, few-shot paradigm. Recognizing the high linguistic variability inherent in clinical narratives, a Retrieval-Augmented Generation (RAG) inspired pipeline was implemented to provide the LLM with contextually relevant exemplars. This approach involved a three-step pipeline:

- **Index Construction:** A search engine was built using a training corpus of previously annotated clinical notes.
- **Semantic Retrieval:** For each novel clinical note presented during inference, the system queried the index to retrieve the top N most semantically similar notes from the training set.
- **Prompt Augmentation:** The retrieved clinical notes, along with their corresponding gold-standard extracted PII pairs, were dynamically appended to the prompt instructions.

However, contrary to initial hypotheses, empirical evaluation demonstrated that the inclusion of retrieved few-shot exemplars did not yield a measurable improvement in overall system performance. The added complexity of the retrieval pipeline, potentially introducing distracting or misaligned context, ultimately failed to enhance the LLM's extraction capabilities beyond the baseline established in the previous iteration.

Consequently, the dynamic few-shot approach was discarded in favor of the structured zero-shot extraction methodology, which provided the optimal balance of accuracy, formatting consistency, and computational efficiency for the final system architecture.

```
You are a medical assistant. Your task is to identify all Protected Health
    Information (PHI) from the clinical note as below.
Types of PHI to identify:
1. Name: Specific names of patients and employers.
2. Location: Specific locations including a government hospital.
3. Date: Dates containing a specific day and month. Does not count weekdays
    and year.
Preserve all PHI terms as they appear in the original text. Do not standardize
    , reformat, or alter the terms in any way. Your answer should be formatted
     as "<PHI type>=PHI terms".

Examples:
Clinical note:
I missed my periods in December 2022. I bought a pregnancy test in January
    2023 and it showed two lines even though one was faded. I started my ANC
    in February 2023, I was given folic acid and ferrous sulphate. I was
    admitted for a high BP and the baby was getting tired. I was booked for an
     emergency caesarian section. The baby was born with a lung problem. They
    said his lungs had air and a tube was inserted to drain the air. He also
    had a heart problem, they said he had a bleed in his brain because of
    prematurity. The baby was brain damaged and he died.
Answer:
<date>=december, january, february

Clinical note:
{{text}}

Please provide your answer below:
```

Prompt 3: Dynamic Few-Shot Prompting with Retrieved Exemplars

Appendix S3 Evaluation Metrics

We evaluated system performance against a manually-annotated gold standard, excluding the samples used for guideline preparation and system development. Evaluation was performed at the entity-token level; each instance of a true PII entity was independently assessed. An entity tag was considered incorrect if any token within the relevant span was misclassified. Based on this strict span requirement, we defined our error categories as follows:

- **False Negative (FN):** Occurs if any portion of a true PII entity remained unredacted (a privacy leak).
- **False Positive (FP):** Occurs if any non-PII text was incorrectly redacted (over-redaction).

- **True Positive (TP):** Occurs only when the relevant span is correctly identified and classified.

We measured performance using Precision, Recall, and F_1-score. **Precision** captures the extent of over-redaction (the reliability of positive predictions) and is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{S1}$$

Recall measures the system's success at preventing data leaks (the coverage of true PII entities) and is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{S2}$$

The **F_1-score** provides the harmonic mean of the two, offering a balanced view of system performance:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{S3}$$

Appendix S4 Detailed Performance Metrics

Table S6: Performance of the de-identification system (named entity recognition) on the annotated data, grouped by dataset and entity type. The table displays Precision, Recall, and F_1-score with 95% confidence intervals for each country.

Entity	Country	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
Verbal Autopsies (VA)				
Date	Kenya	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Sierra Leone	0.98 (0.97–1.00)	1.00 (0.97–1.00)	0.99 (0.97–1.00)
	South Africa	1.00 (0.93–0.96)	0.89 (0.93–0.96)	0.94 (0.93–0.96)
	Overall	0.98 (0.97–0.99)	0.93 (0.91–0.95)	0.96 (0.95–0.97)
Location	Ethiopia	0.75 (0.70–0.80)	0.98 (0.96–1.00)	0.85 (0.81–0.88)
	Kenya	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Sierra Leone	0.90 (0.87–0.93)	1.00 (0.99–1.00)	0.94 (0.93–0.96)
	South Africa	0.97 (0.95–0.98)	0.97 (0.95–0.99)	0.97 (0.96–0.98)
	Overall	0.90 (0.88–0.91)	0.99 (0.98–0.99)	0.94 (0.93–0.95)
Name	Ethiopia	0.51 (0.43–0.59)	1.00 (1.00–1.00)	0.67 (0.60–0.74)
	Kenya	0.83 (0.60–1.00)	1.00 (1.00–1.00)	0.91 (0.75–1.00)
	Sierra Leone	0.65 (0.50–0.79)	1.00 (1.00–1.00)	0.79 (0.67–0.89)
	South Africa	0.86 (0.75–0.96)	0.93 (0.84–1.00)	0.89 (0.82–0.96)
	Overall	0.61 (0.55–0.67)	0.98 (0.96–1.00)	0.75 (0.70–0.80)

Continued on next page

Table S6 – continued from previous page

Entity	Country	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
Maternal Clinical Abstractions (MCA)				
Date	Bangladesh	0.99 (0.99–1.00)	0.99 (0.99–1.00)	0.99 (0.99–1.00)
	Ethiopia	0.99 (0.95–1.00)	1.00 (1.00–1.00)	0.99 (0.98–1.00)
	Kenya	0.97 (0.96–0.98)	0.98 (0.97–0.99)	0.98 (0.97–0.99)
	Mozambique	1.00 (1.00–1.00)	0.83 (0.65–1.00)	0.91 (0.77–1.00)
	Nigeria	1.00 (1.00–1.00)	0.97 (0.92–1.00)	0.99 (0.96–1.00)
	Sierra Leone	0.97 (0.95–0.99)	0.94 (0.92–0.96)	0.96 (0.94–0.97)
	South Africa	0.98 (0.97–1.00)	0.94 (0.92–0.97)	0.96 (0.95–0.98)
	Overall	0.98 (0.98–0.99)	0.97 (0.96–0.98)	0.98 (0.97–0.98)
Location	Bangladesh	0.95 (0.93–0.98)	1.00 (0.99–1.00)	0.97 (0.96–0.99)
	Ethiopia	0.93 (0.88–0.97)	0.86 (0.80–0.91)	0.89 (0.85–0.93)
	Kenya	0.89 (0.84–0.94)	0.93 (0.89–0.97)	0.91 (0.87–0.94)
	Mozambique	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Nigeria	0.72 (0.58–0.85)	1.00 (1.00–1.00)	0.84 (0.72–0.92)
	Sierra Leone	0.76 (0.69–0.83)	0.98 (0.95–1.00)	0.86 (0.80–0.90)
	South Africa	0.96 (0.91–1.00)	0.81 (0.74–0.89)	0.88 (0.83–0.92)
	Overall	0.90 (0.88–0.92)	0.93 (0.92–0.95)	0.91 (0.90–0.93)
Name	Bangladesh	0.93 (0.86–1.00)	0.98 (0.92–1.00)	0.96 (0.91–0.99)
	Ethiopia	0.80 (0.61–0.96)	1.00 (1.00–1.00)	0.89 (0.76–0.98)
	Kenya	1.00 (1.00–1.00)	0.83 (0.70–0.94)	0.91 (0.82–0.97)
	Sierra Leone	1.00 (1.00–1.00)	0.88 (0.75–0.97)	0.94 (0.86–0.99)
	South Africa	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
	Overall	0.93 (0.90–0.96)	0.94 (0.91–0.97)	0.94 (0.91–0.96)
Pediatric Clinical Abstractions (PCA)				
Date	Ethiopia	0.91 (0.76–1.00)	1.00 (1.00–1.00)	0.95 (0.87–1.00)
	Kenya	0.98 (0.97–0.99)	0.99 (0.98–0.99)	0.98 (0.98–0.99)
	Mozambique	1.00 (1.00–1.00)	0.94 (0.87–1.00)	0.97 (0.93–1.00)
	Nigeria	1.00 (1.00–1.00)	0.97 (0.90–1.00)	0.98 (0.94–1.00)
	Sierra Leone	0.99 (0.98–1.00)	0.93 (0.90–0.95)	0.96 (0.94–0.97)
	South Africa	0.98 (0.96–0.99)	0.96 (0.94–0.98)	0.97 (0.96–0.98)
	Overall	0.98 (0.98–0.99)	0.97 (0.96–0.98)	0.97 (0.97–0.98)
Location	Ethiopia	0.90 (0.83–0.96)	1.00 (1.00–1.00)	0.95 (0.91–0.98)
	Kenya	0.89 (0.84–0.93)	0.97 (0.94–0.99)	0.92 (0.90–0.95)
	Mozambique	0.87 (0.79–0.95)	1.00 (1.00–1.00)	0.93 (0.88–0.97)
	Nigeria	0.71 (0.55–0.85)	1.00 (1.00–1.00)	0.83 (0.70–0.92)
	Sierra Leone	0.95 (0.92–0.97)	0.98 (0.96–1.00)	0.96 (0.94–0.98)
	South Africa	0.89 (0.84–0.94)	0.88 (0.84–0.93)	0.89 (0.85–0.92)

Continued on next page

Table S6 – continued from previous page

Entity	Country	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
	Overall	0.90 (0.88–0.92)	0.96 (0.94–0.97)	0.93 (0.92–0.94)
Name	Ethiopia	0.83 (0.65–0.96)	1.00 (1.00–1.00)	0.91 (0.80–0.98)
	Kenya	0.87 (0.78–0.95)	0.91 (0.83–0.98)	0.89 (0.82–0.94)
	Sierra Leone	0.74 (0.66–0.80)	0.98 (0.96–1.00)	0.84 (0.79–0.89)
	South Africa	0.69 (0.50–0.88)	1.00 (1.00–1.00)	0.82 (0.67–0.92)
	Overall	0.76 (0.71–0.81)	0.97 (0.94–0.99)	0.85 (0.82–0.88)

Appendix S5 Single Module Performance

As detailed in Table S7, for the entity "name", the Full System consistently demonstrated superior efficacy, achieving the highest F_1-scores across all datasets (0.75–0.94). By integrating the high sensitivity of the PII detection agent (Recall: 0.94–0.98) with the strict specificity of the rule-based engine, the Full System successfully mitigated the trade-off between over-redaction and data leakage. Consequently, it significantly outperformed both the standalone sub-modules and the Piiredact baseline, ensuring robust privacy preservation while minimizing the loss of clinical context.

Table S7: Module-by-Module Performance Metrics for Named Entity Recognition.

Dataset	Method	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
VA	Full System	0.61 (0.55–0.67)	0.98 (0.96–1.00)	0.75 (0.70–0.80)
	PII detection agent	0.45 (0.40–0.51)	0.94 (0.90–0.98)	0.61 (0.56–0.66)
	Piiredact	0.32 (0.28–0.37)	0.85 (0.80–0.91)	0.47 (0.41–0.52)
	Rule-based engine	1.00 (1.00–1.00)	0.26 (0.20–0.33)	0.41 (0.33–0.50)
MCA	Full System	0.93 (0.90–0.96)	0.94 (0.91–0.97)	0.94 (0.91–0.96)
	PII detection agent	0.58 (0.52–0.63)	0.96 (0.93–0.98)	0.72 (0.68–0.76)
	Piiredact	0.28 (0.24–0.32)	0.66 (0.59–0.73)	0.39 (0.34–0.44)
	Rule-based engine	0.96 (0.92–1.00)	0.42 (0.36–0.49)	0.58 (0.51–0.65)
PCA	Full System	0.76 (0.71–0.81)	0.97 (0.94–0.99)	0.85 (0.82–0.88)
	PII detection agent	0.61 (0.55–0.66)	0.95 (0.92–0.98)	0.74 (0.70–0.78)
	Piiredact	0.46 (0.41–0.52)	0.69 (0.63–0.76)	0.56 (0.50–0.61)
	Rule-based engine	0.89 (0.80–0.96)	0.24 (0.18–0.30)	0.38 (0.30–0.45)

For the entity "location", Table S8 demonstrates the robust capability of the Full System in identifying location entities, yielding consistent F_1-scores between 0.91 and 0.94 across all datasets. The system exhibited a significant performance leap over the Piiredact baseline, which failed to identify the vast majority of location data (Recall: 0.11–0.24). Furthermore, while the individual components (the PII detection agent and the rule-based engine) struggled with precision (0.49–0.63), the Full System effectively

combined these inputs. It maintained a stable precision of 0.90 while achieving exceptional recall (0.93–0.99), ensuring that geographical identifiers were removed with high reliability and minimal over-redaction.

Table S8: Performance Metrics for Location Entity Recognition on the annotated data.

Dataset	Method	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
VA	Full System	0.90 (0.88–0.91)	0.99 (0.98–0.99)	0.94 (0.93–0.95)
	PII detection agent	0.49 (0.46–0.51)	0.88 (0.86–0.90)	0.63 (0.61–0.65)
	Piiredact	0.64 (0.59–0.69)	0.24 (0.21–0.26)	0.35 (0.31–0.38)
	Rule-based engine	0.54 (0.52–0.57)	0.90 (0.89–0.92)	0.68 (0.66–0.70)
MCA	Full System	0.90 (0.88–0.92)	0.93 (0.92–0.95)	0.91 (0.90–0.93)
	PII detection agent	0.54 (0.51–0.57)	0.79 (0.76–0.81)	0.64 (0.62–0.66)
	Piiredact	0.41 (0.36–0.45)	0.16 (0.14–0.18)	0.23 (0.20–0.26)
	Rule-based engine	0.61 (0.58–0.64)	0.85 (0.82–0.87)	0.71 (0.69–0.73)
PCA	Full System	0.90 (0.88–0.92)	0.96 (0.94–0.97)	0.93 (0.92–0.94)
	PII detection agent	0.55 (0.53–0.58)	0.87 (0.84–0.89)	0.68 (0.65–0.70)
	Piiredact	0.44 (0.37–0.51)	0.11 (0.09–0.13)	0.17 (0.14–0.20)
	Rule-based engine	0.63 (0.60–0.66)	0.77 (0.74–0.80)	0.69 (0.67–0.72)

As presented in Table S9, the extraction of temporal entities yielded the highest overall performance metrics among all PII categories. The Full System achieved near-perfect fidelity, with F_1-scores ranging from 0.95 to 0.98 and precision reaching 1.00 (95% CI: 0.99–1.00) in the VA cohort. Notably, unlike name or location entities, the high-precision rule-based engine demonstrated exceptional standalone performance (Recall: 0.90–0.97), reflecting the highly structured syntactic patterns of dates in clinical text. However, the Full System still offered a critical advantage by combining this rule-based precision with the high sensitivity of the PII detection agent (Recall: 0.96–0.99), effectively eliminating the substantial over-redaction observed in the standalone agent (Precision: 0.60–0.63) while far surpassing the low recall of the Piiredact baseline of 0.34 (95% CI: 0.31–0.37).

Table S9: Performance Metrics for Date Entity Recognition on the annotated data.

Dataset	Method	Precision (95% CI)	Recall (95% CI)	F_1-score (95% CI)
VA	Full System	1.00 (0.99–1.00)	0.91 (0.89–0.93)	0.95 (0.94–0.96)
	PII detection agent	0.62 (0.59–0.65)	0.98 (0.97–0.99)	0.76 (0.74–0.78)
	Piiredact	0.48 (0.44–0.52)	0.34 (0.31–0.37)	0.40 (0.36–0.43)
	Rule-based engine	0.87 (0.84–0.89)	0.90 (0.88–0.92)	0.88 (0.87–0.90)
MCA	Full System	0.98 (0.98–0.99)	0.97 (0.96–0.98)	0.98 (0.97–0.98)
	PII detection agent	0.63 (0.62–0.65)	0.99 (0.98–0.99)	0.77 (0.76–0.78)
	Piiredact	0.50 (0.46–0.53)	0.20 (0.18–0.21)	0.28 (0.26–0.30)
	Rule-based engine	0.97 (0.96–0.97)	0.97 (0.97–0.98)	0.97 (0.97–0.98)
PCA	Full System	0.98 (0.98–0.99)	0.97 (0.96–0.98)	0.97 (0.97–0.98)
	PII detection agent	0.60 (0.58–0.62)	0.96 (0.95–0.97)	0.74 (0.72–0.75)
	Piiredact	0.42 (0.38–0.46)	0.13 (0.12–0.15)	0.20 (0.18–0.23)
	Rule-based engine	0.97 (0.96–0.98)	0.97 (0.96–0.98)	0.97 (0.96–0.97)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NPJCHAMPSDeidentification1.pdf