Can Zero-Shot Commercial API's Deliver Regulatory-Grade Clinical Text De-Identification?

Veysel Kocaman[1], Muhammed Santas[1], Yigit Gul[1], Mehmet Butgul[1] and David Talby[1]

[1]John Snow Labs inc. 16192 Coastal Highway, Lewes, DE 19958, USA

Abstract

We systematically assess the performance of three leading API-based de-identification systems—Azure Health Data Services, AWS Comprehend Medical, and OpenAI GPT-4o—against our de-identification systems on a ground truth dataset of 48 clinical documents annotated by medical experts. Our analysis, conducted at both entity-level and token-level, demonstrates that our solution, Healthcare NLP, achieves the highest accuracy, with a 96% F1-score in protected health information (PHI) detection, significantly outperforming Azure (91%), AWS (83%), and GPT-4o (79%). Beyond accuracy, Healthcare NLP is also the most cost-effective solution, reducing processing costs by over 80% compared to Azure and GPT-4o. Its fixed-cost local deployment model avoids the escalating per-request fees of cloud-based services, making it a scalable and economical choice. Our results underscore a critical limitation: zero-shot commercial APIs fail to meet the accuracy, adaptability, and cost-efficiency required for regulatory-grade clinical de-identification. Healthcare NLP's superior performance, customization capabilities, and economic advantages position it as the more viable solution for healthcare organizations seeking compliance and scalability in clinical NLP workflows.

1. Introduction

Electronic Health Records (EHRs) are now widespread across the United States healthcare system, with adoption rates surpassing 96% in acute care hospitals and 86% among office-based physicians [1]. Although structured data, such as billing and claims information, constitutes a substantial component of EHRs, a significant proportion of clinical information remains in unstructured formats, including progress notes, discharge summaries, radiology reports, and pathology reports. This unstructured data contains valuable contextual details essential for comprehensive patient care. Its secondary use in research has gained increasing importance, with potential benefits in areas such as population health management, real-world evidence generation, patient safety enhancements, and drug discovery. However, processing unstructured data poses substantial ethical and technical challenges. The inherent variability of free-text documentation complicates efforts to preserve privacy, as sensitive patient information is frequently embedded within clinical narratives.

Given the highly sensitive nature of this data, it must undergo a de-identification process before use. De-identification involves removing or obscuring personal health information (PHI) from medical records to protect patient privacy. De-identified data refers to health information that has been stripped of all "direct identifiers"—elements that could uniquely identify an individual. The Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor guidelines define 18 such direct identifiers (U.S. Department of Health & Human Services, 2023) [2], though any additional data points capable of uniquely identifying a patient must also be considered. The federally regulated HIPAA Privacy Rule outlines two primary methods for de-identifying PHI: Expert Determination and Safe Harbor.

Recent studies suggest that deep learning-based automated de-identification models can

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'25 Workshop, Lucca (Italy), 10-April-2025

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3964/paper8.pdf

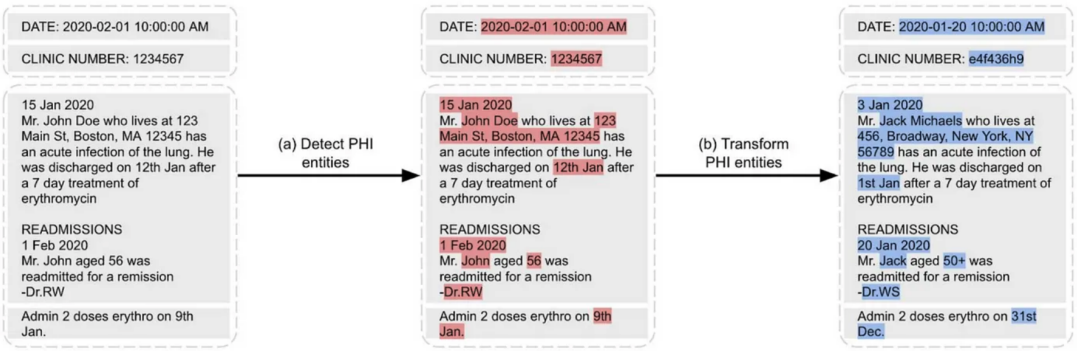

Figure 1: De-Identification process identifies potential pieces of content with personal information about patients and removes them by replacing them with semantic tags or fake entities.

surpass human annotators in identifying PHI, with hybrid approaches demonstrating the greatest potential [3]. Once the de-identification criteria for a specific dataset have been established, advanced technologies can be employed to automate the detection of protected health information (PHI) in both structured and unstructured data. The combination of machine learning techniques and sophisticated Natural Language Processing (NLP) algorithms has markedly enhanced the capacity to identify and flag PHI across various data formats. To streamline the de-identification process, researchers can utilize Large Language Models (LLMs), specialized NLP models, and cloud provider APIs for processing extensive clinical datasets. However, the task of handling ambiguous or novel instances of identifiable information remains challenging, necessitating continuous improvement of these automated tools to strike a balance between efficiency and the nuanced interpretation required in healthcare settings. It is worth noting that while LLMs offer powerful capabilities, their application in de-identifying sensitive data (PHI) may be considered excessive or potentially unreliable for certain use cases, particularly when a high degree of customization is required. The choice of technology should be carefully evaluated based on the specific requirements of the de-identification task and the desired level of precision.

This study examines the performance and compares de-identification services, developed by us and named as Healthcare NLP library, AWS Comprehend Medical, and Azure Health Data Services, with a focus on their accuracy when applied to a dataset annotated by healthcare experts. The comparison of these services provides valuable insights into their respective strengths and limitations, enabling informed decision-making for researchers, developers, and organizations seeking appropriate de-identification tools. Additionally, this comprehensive analysis equips stakeholders with the necessary information to select the most suitable tool based on accuracy, compliance, cost-effectiveness, and scalability for processing sensitive healthcare data.

For researchers, this analysis helps identify the most accurate, reliable, and cost-effective service for processing sensitive data, which is crucial for maintaining data integrity in clinical studies. Developers benefit from understanding the ease of integration and API flexibility of each service, essential factors for building scalable solutions that can handle large volumes of clinical data [4]. Organizations, especially in the healthcare and finance sectors, gain valuable insights into the compliance capabilities and performance of these tools, ensuring that the chosen solution aligns with regulatory requirements while enhancing operational efficiency.

The comparison highlights variations in performance among the evaluated services. Our Healthcare NLP library achieved the highest accuracy, with macro and weighted average F1-scores of 96% and 99%, respectively, followed by Azure Health Data Services with 85% macro

and 99% weighted average F1-scores, and AWS Comprehend Medical with 80% macro and 98% weighted average F1-score. However, performance may vary based on specific use cases and dataset characteristics. Additionally, a cost analysis for processing one million clinical notes (each containing 5,250 characters) revealed that the Healthcare NLP library is the most cost-effective option, followed by Azure Health Data Services and AWS Comprehend Medical.

2. Background and Related Work

The de-identification of unstructured data has been extensively studied, with various Natural Language Processing (NLP) approaches proposed over the years [5, 6]. This process can be divided into two main subtasks: first, identifying Protected Health Information (PHI) within the text, and second, replacing those identifiers through either masking (substituting them with placeholder values) or obfuscation (replacing them with randomly generated values based on their type). Among these, the task of PHI identification has been the primary focus of research [4].

Early de-identification systems in the clinical domain were predominantly rule-based, as seen in the work of Sweeney [7] and Gupta et al. [8]. These systems relied on regular expressions, syntactic rules, and specialized dictionaries to detect PHI in text. While rule-based approaches are effective in identifying structured PHI elements such as phone numbers, email addresses, and license numbers, they struggle with more complex entities, including personal names, professions, and hospital names [9]. Rule-based systems, while effective in specific contexts, often exhibit limited generalizability across diverse datasets. These systems typically require substantial modifications to their underlying dictionaries and rule sets when applied to new environments, hindering their adaptability and scalability in varied clinical settings.

The field of automated PHI detection and de-identification has seen significant advancements in recent years, with several major cloud providers and specialized services offering solutions to address the growing need for secure handling of sensitive healthcare data. The concept of automatic de-identification gained prominence in 2014 through the Informatics for Integrating Biology and the Bedside (i2b2) project, which introduced a pioneering academic NLP challenge focused on automatically detecting PHI identifiers from medical records [10]. This initiative accelerated research and development of Machine Learning and Deep Learning algorithms for robust PHI identification, laying the groundwork for more sophisticated approaches that are now being implemented by major cloud service providers.

Recent research suggests that deep learning-based automated de-identification models can surpass human annotators in PHI identification, with hybrid approaches demonstrating the greatest potential [3]. In the current landscape, several key players have emerged with offerings designed to streamline the process of PHI detection and de-identification. Several studies have conducted performance comparisons of PHI detection systems, providing valuable insights into the effectiveness of various de-identification approaches. These comparisons are crucial for researchers and healthcare organizations seeking to implement efficient and accurate de-identification processes while maintaining data utility for secondary use in research and analytics.

A notable study by Steinkamp et al. [11] evaluated five publicly available de-identification tools on a large corpus of narrative-text radiology reports. The research assessed token-level recall, precision, and F1 scores for each tool across various PHI subcategories. The study found that machine learning systems outperformed rule-based systems, with the best-performing system (NeuroNER) achieving a token-level F1 score of 93.6%. However, this performance was still below the acceptable level for clinical use (95% recall) on sensitive categories of PHI.

Recent advancements in Large Language Models (LLMs) have prompted researchers to

investigate their potential for de-identifying clinical notes. A study by Altalla et al. [12] compared the de-identification performance of GPT-3.5 and GPT-4, revealing GPT-4's superior capabilities in this domain. The study, published on January 31, 2025, reported that GPT-4 achieved remarkable results with a precision of 0.9925, recall of 0.8318, F1 score of 0.8973, and accuracy of 0.9911, significantly outperforming its predecessor, GPT-3.5.

Despite these promising results, the application of LLMs for de-identification presents several challenges. The nascent stage of LLM utilization in this field raises concerns regarding the privacy and security of health data, particularly when employing API-based models [13]. Moreover, LLMs may encounter difficulties in striking a balance between effective de-identification and preserving the clinical utility of notes, potentially altering non-sensitive information crucial for research and analysis [14]. The variation in performance across different datasets highlights the need for continued development to achieve consistent and reliable results across diverse clinical settings.

This study aims to contribute to previous performance comparisons in PHI entity recognition and assist researchers and decision-makers in selecting the most suitable tool for processing large-scale datasets with high accuracy and cost-effectiveness. To achieve this, we compare three widely used and advanced de-identification tools that incorporate state-of-the-art models while ensuring consistency: Our Healthcare NLP library, Azure Health Data Services, AWS Comprehend Medical and GPT4o, a state-of-the-art commercial multi-modal LLM.

3. Experiments and Results

3.1. The Deidentification Solutions

In this section, we will provide brief information for each de-identification solution that supports different set of PHI entities. The list of PHI entities supported by each model is shared in Table A2.

3.1.1. Healthcare NLP & LLM Library

The Healthcare NLP library is a powerful component of Spark NLP platform [15], specifically designed to facilitate NLP tasks within the healthcare domain [16]. This library offers over 2,500 pre-trained models and pipelines tailored for medical data, enabling accurate information extraction, named entity recognition (NER) for clinical and medical concepts, and robust text analysis capabilities. Regularly updated with advanced algorithms, it helps healthcare professionals derive meaningful insights from unstructured medical data sources such as electronic health records, clinical notes, and biomedical literature.

Additionally, the library features custom large language models (LLMs) in various sizes and quantization levels for tasks like medical note summarization, question answering, retrieval-augmented generation (RAG), and healthcare-related conversational interactions. It also provides a robust solution for de-identifying medical records using advanced NER models to automatically detect and remove PHI from clinical notes. This ensures compliance with privacy regulations while preserving data utility for research, enabling secure data sharing, enhancing patient privacy, and promoting innovation in medical research.

The Healthcare NLP library allows users to create custom de-identification pipelines targeting specific labels or to utilize pre-trained pipelines with two lines of code to de-identify a broad range of entities. These entities include AGE, CONTACT, DATE, ID, LOCATION, NAME, PROFESSION, CITY, COUNTRY, DOCTOR, HOSPITAL, IDNUM, MEDICALRECORD, ORGANIZATION, PATIENT, PHONE, STREET, USERNAME, ZIP, ACCOUNT, LICENSE, VIN, SSN, DLN, PLATE, IPADDR, EMAIL, and more. In *Beyond Accuracy: Automated De-Identification of*

Large Real-World Clinical Text Datasets [4], the de-identification process is explained in detail, describing the implementation of a hybrid context-based model architecture for automated clinical note processing.

In this study, a pre-trained de-identification pipeline was utilized, specifically designed to extract and de-identify entities such as NAME, IDNUM, CONTACT, LOCATION, AGE, and DATE. Notably, this pipeline operates independently of any large language model (LLM) components.

3.1.2. Azure Health Data Services

Azure Health Data Services' de-identification service is designed to safeguard sensitive health information while maintaining data utility. This API employs advanced natural language processing techniques to identify, label, redact, or surrogate PHI in unstructured medical texts. The service provides three essential operations: Tag, Redact, and Surrogate, which allow healthcare organizations to process various types of clinical documents securely and efficiently. By utilizing machine learning algorithms, the service can detect HIPAA's 18 identifiers and other PHI entities, ensuring compliance with various regional privacy regulations such as GDPR and CCPA.

3.1.3. Amazon Comprehend Medical

Amazon Comprehend Medical is a HIPAA-eligible natural language processing (NLP) service that leverages machine learning to extract valuable health data from unstructured medical text. This tool quickly and accurately identifies medical entities such as conditions, medications, dosages, tests, treatments, and Protected Health Information (PHI) from various clinical documents including physician's notes, discharge summaries, and test results. With its ability to understand context and relationships between extracted information, AWS Comprehend Medical offers a robust solution for healthcare professionals and researchers looking to automate data extraction, improve patient care, and streamline clinical workflows.

3.1.4. Open AI GPT-4o for Deidentification

GPT-4o is a multi-modal model that offers improvements in response times and classification accuracy compared to GPT-4, which could theoretically enhance the precision of identifying and redacting sensitive information via prompting. While GPT-3.5 and GPT-4 have been extensively studied for their de-identification capabilities, particularly in processing medical text, GPT-4o presents an intriguing option due to its enhanced performance over GPT-4 in various tasks. However, no formal study has yet evaluated GPT-4o's de-identification capabilities. Given the importance of PHI redaction in healthcare AI applications, understanding the model's strengths and limitations in this area remains crucial. Despite these advantages, its effectiveness in de-identification remains speculative without empirical studies directly assessing its performance. While there are cost-effective alternatives for de-identification, we opted for GPT-4o due to its widespread adoption, strong presence in research, and its demonstrated advancements over previous models.

3.2. Dataset

The annotation of patient identifiers within clinical data is a critical process in healthcare research and data management. This study employed a comprehensive annotation methodology utilizing the John Snow Labs' Annotation Lab software, which facilitated a multi-stage approach to entity recognition and labeling. The process began with a pre-annotation step using deep learning models to extract initial entities, followed by human refinement guided by a dynamic

annotation guide. This iterative approach, involving multiple rounds of review and correction, ensured high accuracy and adaptability throughout the fine-tuning and evaluation phases [17].

The dataset employed in this study comprised 48 clinical notes meticulously annotated by our domain experts. The dataset was specifically curated to facilitate the evaluation of de-identification systems in a healthcare context. Expert annotations focused on six key entity types: IDNUM, LOCATION, DATE, AGE, NAME, and CONTACT. These entities represent critical categories of Protected Health Information (PHI) that are commonly subject to de-identification under regulatory frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR).

The selection of these entity types was motivated by their frequent occurrence in clinical narratives and their significance in ensuring patient privacy. Identifiers such as patient names, contact details, and unique ID numbers pose a high risk of re-identification if not properly anonymized. Similarly, location information, age, and date-related details can contribute to indirect re-identification, necessitating robust de-identification strategies. By centering the benchmark on these entities, this study ensures that the performance evaluation remains directly aligned with real-world de-identification challenges in healthcare settings.

To enhance reproducibility, the benchmark dataset utilized in this study has been made publicly available in a dedicated repository[18]. This ensures transparency and facilitates further research in the field of healthcare de-identification.

3.3. Comparison of the Solutions

The most significant difference between these tools lies in their adaptability. Azure Health Data Services, Amazon Comprehend Medical and GPT-4o are API-based, black-box cloud solutions, making modifying or adapting results to specific needs impossible. On the other hand, the Healthcare NLP library's de-identification pipeline can be loaded and utilized with just two lines of code. The pipeline outputs can be customized by adjusting its stages to meet specific needs, and it can also be used locally with no internet connection.

3.3.1. Evaluation Criteria

In this benchmark study, we employed two distinct approaches to compare accuracy:

3.3.2. Entity-Level Evaluation

Since de-identifying PHI data is a critical task, we evaluated how well de-identification tools detected entities present in the annotated dataset, regardless of their specific labels in the ground truth. The detection outcomes were categorized as:

- **full_match:** The entire entity was correctly detected.
- **partial_match:** Only a portion of the entity was detected.
- **not_matched:** The entity was not detected at all.

For example, for the text: "Patient John Doe was admitted to Boston General Hospital on 01/12/2023.", the ground truth entity "John Doe (NAME)" could have the following predicted entities:

- Predicted Entity: "John Doe (NAME)" ==> **full_match**
- Predicted Entity: "John" ==> **partial_match**
- Predicted Entity: "Patient" ==> **not_matched**

For evaluation results, please refer to Figure A1 in the Appendix section.

3.3.3. Token-Level Accuracy

The text in the annotated dataset was tokenized, and the ground truth labels assigned to each token were compared with predictions made by the Healthcare NLP library, Amazon Comprehend Medical, Azure Health Data Services, and GPT-4o model. Classification reports were generated for each tool, comparing their precision, recall, and F1 scores. Token-level evaluation results are presented in Figure A2 in the Appendix section.

3.4. Methodology

In this study, differences were observed between the predictions generated by the de-identification services and the ground truth annotations. The ground truth dataset utilized generic entity labels; for instance, all names were annotated as **NAME**, rather than distinguishing between **PATIENT_NAME** and **DOCTOR_NAME**. To ensure consistency in evaluation, the predicted labels from the de-identification tools were mapped to their corresponding ground truth labels.

To maintain a fair comparison, entities that did not have a direct mapping to the ground truth labels—such as **PROFESSION**, **ORGANIZATION**, and other non-essential entity types—were excluded from the predictions before conducting the performance evaluation. This preprocessing step ensured that the assessment focused solely on the six critical entity types relevant to healthcare de-identification. Entity mapping table showing entity mapping across different providers can be seen at Table A3. After obtaining the model predictions and applying the preprocessing steps, the entity distribution was summarized in Table A5. While evaluating GPT4o, we used a one-shot prompt to provide the model some sample PHI entity extraction tasks (the prompt is shared in the Appendix).The model was configured with a temperature of 1 and executed as a single run, while all other parameters were maintained at their default settings to ensure consistency in evaluation.

4. Experiments and Results

4.1. Performance Evaluation

The final results can be found at Table 1. The entity-level and token-level evaluations including comparative analyses and benchmark scores can be found in the Appendix.

Table 1
Healthcare NLP, Azure, Amazon, and GPT-4o PHI Recognition and Benchmark Comparison (Sample size: 45172 PHI entities).

Metric / Entity	Healthcare NLP			Azure			Amazon			GPT-4o		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
AGE	0.96	1.00	**0.98**	0.94	0.45	0.61	1.00	0.41	0.58	0.87	0.50	0.64
CONTACT	0.96	0.97	**0.97**	0.73	0.88	0.80	0.78	0.72	0.75	0.67	0.53	0.59
DATE	0.97	0.99	**0.98**	0.91	0.99	0.95	0.90	0.97	0.93	0.79	0.72	0.75
IDNUM	0.98	0.94	**0.96**	0.78	0.93	0.85	0.95	0.86	0.91	0.70	0.92	0.80
LOCATION	0.93	0.92	**0.93**	0.89	0.87	0.88	0.52	0.74	0.61	0.82	0.72	0.76
NAME	0.92	0.94	**0.93**	0.92	0.89	0.90	0.85	0.76	0.80	0.79	0.82	0.80
O	1.00	1.00	**1.00**	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Macro Avg	0.96	0.97	**0.96**	0.88	0.86	0.85	0.86	0.78	0.80	0.80	0.74	0.76
Non-PHI	1.00	1.00	**1.00**	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
PHI	0.96	0.97	**0.96**	0.91	0.92	0.91	0.81	0.85	0.83	0.81	0.77	0.79
Macro Avg	0.98	0.98	**0.98**	0.95	0.96	0.95	0.90	0.92	0.91	0.90	0.88	0.89
cost per 1M doc		**$2,418**			$13,125			$14,525			$21,400	

The primary objective of de-identification is to accurately detect PHI entities. In this regard, we also wanted to evaluate binary classification performance in which entities were classified

as either PHI or non-PHI, disregarding specific subcategories. The PHI entity detection results are also summarized in Table 1 and Figure 2.

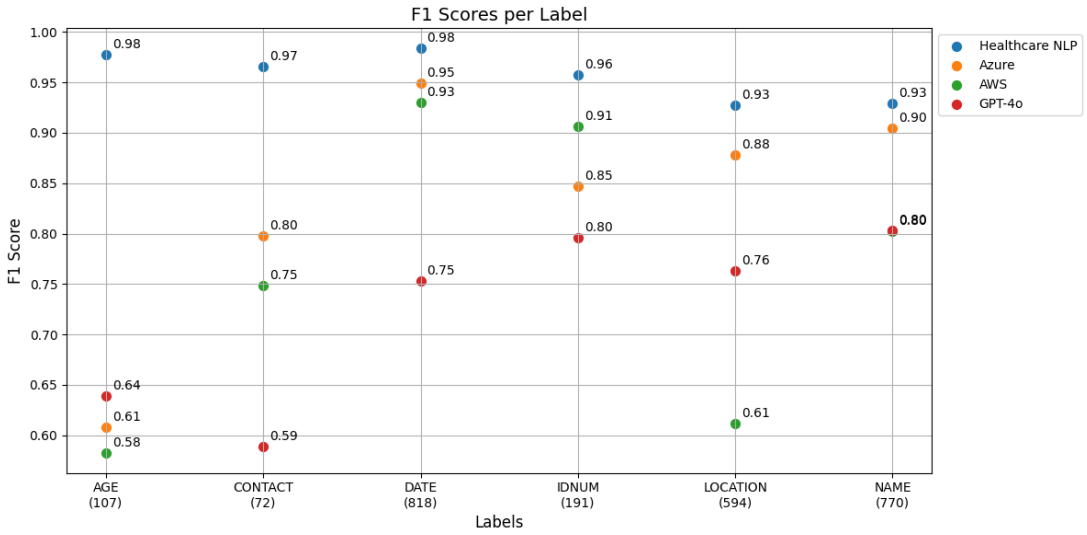

Figure 2: Visualization of the F1-Scores for each label

4.2. Cost Estimation for De-identifying Clinical Data

Cost is a critical factor when processing large-scale clinical datasets. To estimate expenses, we simulated the cost of de-identifying 1 million unstructured clinical notes, each averaging 5,250 characters.

The pricing estimates are as follows:

- **Amazon Comprehend Medical:** Processing 1M documents costs approximately **$14,525**.
- **Azure Health Data Services:** Processing 1M documents costs approximately **$13,125**.
- **Open AI GPT-4o:** Processing 1M documents costs approximately **$21,400**.
- **Healthcare NLP:** Using John Snow Labs' Healthcare NLP Prepaid on an EC2 c6a.8xlarge instance ($1.2/hour), de-identifying PHI from 48 documents took 39.4 seconds. Extrapolating, processing 1M documents would take approximately 228 hours (9.5 days), but with proper scaling, it could be completed in a single day. The total estimated cost:
  - **Infrastructure:** $273
  - **License:** $2,145 (if one-month license cost set to $7,000)
  - **Total: $2,418**

5. Conclusion

In this study, we conducted a comparative analysis of the performance of Healthcare NLP, Amazon Comprehend Medical, Azure Health Data Services, and Open AI GPT-4o model on a ground truth dataset annotated by medical experts. The evaluation was performed at two levels: entity-level and token-level.

The entity-level analysis demonstrated that Healthcare NLP outperformed its counterparts in accurately capturing entities while minimizing missed detections. Azure Health Data Services

exhibited the second-best performance, followed by Amazon Comprehend Medical. The GPT-4o model ranked fourth in this comparative assessment.

The token-level evaluation further reinforced these findings, with Healthcare NLP achieving the highest precision, recall, and F1-score. Azure Health Data Services, Amazon Comprehend Medical and GPT-4o followed in that order, indicating a consistent pattern of superior performance for Healthcare NLP across both evaluation metrics.

A key differentiator among these tools is their adaptability. While Azure Health Data Services, Amazon Comprehend Medical and GPT-4o function as API-based, black-box cloud solutions with no customization capabilities, Healthcare NLP provides a flexible and transparent framework. Its de-identification pipeline can be implemented with minimal coding effort, and users can modify pipeline stages to tailor the output to their specific requirements.

From a cost-effectiveness perspective, Healthcare NLP emerges as the most viable solution for large-scale clinical data processing. Unlike cloud-based services, which impose per-request pricing that escalates with increasing data volumes, Healthcare NLP allows for fixed-cost, local deployment. Even when processing substantial datasets, such as one billion clinical notes, its pricing remains stable over the same time period, providing a significant economic advantage over API-based alternatives.

In summary, Healthcare NLP consistently outperformed Azure Health Data Services, Amazon Comprehend Medical, and GPT-4o across all evaluation metrics by 5-10%, achieving the highest accuracy while minimizing missed detections. Beyond its superior performance, its adaptability offers a crucial advantage over the black-box nature of cloud solutions, enabling users to customize de-identification pipelines to meet specific needs. Furthermore, its cost-effective deployment model presents substantial savings, making it a compelling alternative to API-based solutions.

References

[1] K. Myrick, D. Ogburn, B. Ward, Percentage of office-based physicians using any electronic health record (ehr)/electronic medical record (emr) system and physicians that have a certified ehr/emr system, by us state: National electronic health records survey, 2017, National Center for Health Statistics (2019) 2021–04.

[2] U.S. Department of Health & Human Services, Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule, Web Page, 2023. URL: https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html, accessed on October 24, 2024.

[3] B. Negash, A. Katz, C. J. Neilson, M. Moni, M. Nesca, A. Singer, J. E. Enns, De-identification of free text data containing personal health information: a scoping review of reviews, International Journal of Population Data Science 8 (2023).

[4] V. Kocaman, D. Talby, H. U. Hak, Rwd143 beyond accuracy: Automated de-identification of large real-world clinical text datasets, Value in Health 26 (2023) S532.

[5] P. M. Nadkarni, L. Ohno-Machado, W. W. Chapman, Natural language processing: an introduction, Journal of the American Medical Informatics Association 18 (2011) 544–551.

[6] K. Khin, P. Burckhardt, R. Padman, A deep learning architecture for de-identification of patient notes: Implementation and evaluation, arXiv preprint arXiv:1810.01570 (2018).

[7] L. Sweeney, Replacing personally-identifying information in medical records, the scrub system., in: Proceedings of the AMIA annual fall symposium, American Medical Informatics Association, 1996, p. 333.

[8] D. Gupta, M. Saul, J. Gilbertson, Evaluation of a deidentification (de-id) software engine to share pathology reports and clinical documents for research, American journal of clinical pathology 121 (2004) 176–186.

[9] Z. Liu, B. Tang, X. Wang, Q. Chen, De-identification of clinical notes via recurrent neural network and conditional random field, Journal of biomedical informatics 75 (2017) S34–S42.

[10] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, Journal of the American Medical Informatics Association 14 (2007) 550–563.

[11] J. M. Steinkamp, T. Pomeranz, J. Adleberg, C. E. Kahn Jr, T. S. Cook, Evaluation of automated public de-identification tools on a corpus of radiology reports, Radiology: Artificial Intelligence 2 (2020) e190137.

[12] B. Altalla', S. Abdalla, A. Altamimi, L. Bitar, A. Al Omari, R. Kardan, I. Sultan, Evaluating gpt models for clinical note de-identification, Scientific Reports 15 (2025) 3852.

[13] Z. Liu, Y. Huang, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, Y. Li, P. Shu, et al., Deid-gpt: Zero-shot medical text de-identification by gpt-4, arXiv preprint arXiv:2303.11032 (2023).

[14] A. R. Sarkar, Y.-S. Chuang, N. Mohammed, X. Jiang, De-identification is not enough: a comparison between de-identified and synthetic clinical notes, Scientific Reports 14 (2024) 29669.

[15] V. Kocaman, D. Talby, Spark nlp: natural language understanding at scale, Software Impacts 8 (2021) 100058.

[16] V. Kocaman, D. Talby, Accurate clinical and biomedical named entity recognition at scale, Software Impacts 13 (2022) 100373.

[17] H. A. Xu, V. Loftsson, B. Kulynych, B. Kaabachi, J. L. Raisaro, Accelerating clinical text annotation in underrepresented languages: A case study on text de-identification, in: Digital Health and Informatics Innovations for Sustainable Health Care Systems, IOS Press, 2024, pp. 853–857.

[18] JohnSnowLabs, De-identification benchmark ground truth dataset, 2024. URL: https://github.com/JohnSnowLabs/spark-nlp-workshop/blob/master/tutorials/academic/DeIdentification_Benchmarks_Text2Story2025/deidentification_benchmark_ground_truth_48_doc.csv, accessed: March 9, 2025.

Appendix

Table A1
Example of Original and NER Detection Text

Type	Description	Example Text
Original	The original text with identifiable information	He is a 60-year-old male.
NER Detection and Masking	Text with Named Entity Recognition (NER) applied.	He is a <AGE> male.

Table A2
Comparison of De-identification Solutions

Tool	Entities De-identified	Key Features
Healthcare NLP Library	AGE, CONTACT, DATE, ID, LOCATION, NAME, PROFESSION, CITY, COUNTRY, DOCTOR, HOSPITAL, IDNUM, MEDICALRECORD, ORGANIZATION, PATIENT, PHONE, STREET, USERNAME, ZIP, ACCOUNT, LICENSE, VIN, SSN, DLN, PLATE, IPADDR, EMAIL	Highly flexible; the de-identification pipeline can be easily loaded with two lines of code and customized to meet specific requirements. Additionally, it can be used locally.
Azure Health Data Services	DATE, DOCTOR, HOSPITAL, IDNUM, PATIENT, MEDICALRECORD, PHONE, AGE, STREET, STATE, CITY, HEALTHPLAN, PROFESSION, ZIP, EMAIL, ORGANIZATION, USERNAME, FAX, URL, LOCATIONOTHER, ACCOUNT, COUNTRYORREGION, SOCIALSECURITY	API-based, black-box solution; no direct control over results; suitable for integrated, cloud-based environments but lacks flexibility for task-specific adjustments.
AWS Comprehend Medical	DATE, NAME, ADDRESS, ID, AGE, PHONE_OR_FAX, PROFESSION, URL, EMAIL	API-based, black-box solution; de-identification is limited to specific pre-configured models; lacks customization and flexibility for adapting results to specific needs.
GPT-4o	No pre-built set of entities	API-based, black-box solution; de-identification is run via prompting.

Evaluation Results

The results obtained by comparing the predictions made by Healthcare NLP, AWS Comprehend Medical, and Azure Health Data Services with the ground truth entities are presented below.

chunk	chunk_label	healthcare_nlp_is_matched	healthcare_nlp_prediction	azure_is_matched	azure_prediction	aws_is_matched	aws_prediction
957770228	IDNUM	full_match	IDNUM	full_match	IDNUM	full_match	IDNUM
FIH	LOCATION	full_match	LOCATION	full_match	LOCATION	not_matched	no_prediction
0408267	IDNUM	full_match	IDNUM	full_match	IDNUM	full_match	IDNUM
46769/5v7d	IDNUM	full_match	NAME	partial_match	IDNUM	partial_match	IDNUM
237890	IDNUM	full_match	IDNUM	full_match	IDNUM	full_match	IDNUM
2/5/1994	DATE	full_match	DATE	full_match	DATE	full_match	DATE
2-5-94	DATE	full_match	DATE	full_match	DATE	full_match	DATE
4-2-94	DATE	full_match	DATE	full_match	DATE	full_match	DATE
February 6 , 1994	DATE	full_match	DATE	full_match	DATE	full_match	DATE
February 10 , 1994	DATE	full_match	DATE	full_match	DATE	full_match	DATE

Figure A1: Entity Level Evaluation

To further analyze the performance of each de-identification tool, a token-level evaluation was conducted. This involved tokenizing the ground truth text and associating each token with the corresponding predicted labels from Healthcare NLP, Amazon Comprehend Medical, Azure Health Data Services and GPT-4o.

token	token_label	healthcare_nlp_token_label	azure_token_label	aws_token_label
957770228	IDNUM	IDNUM	IDNUM	IDNUM
FIH	LOCATION	LOCATION	LOCATION	O
0408267	IDNUM	IDNUM	IDNUM	IDNUM
46769/5v7d	IDNUM	NAME	IDNUM	IDNUM
237890	IDNUM	IDNUM	IDNUM	IDNUM
2/5/1994	DATE	DATE	DATE	DATE
12:00:00	O	O	O	O
AM	O	O	O	O
TRACHEOESOPHAGEAL	O	O	O	O
FISTULA	O	O	O	O

Figure A2: Token Level Evaluation

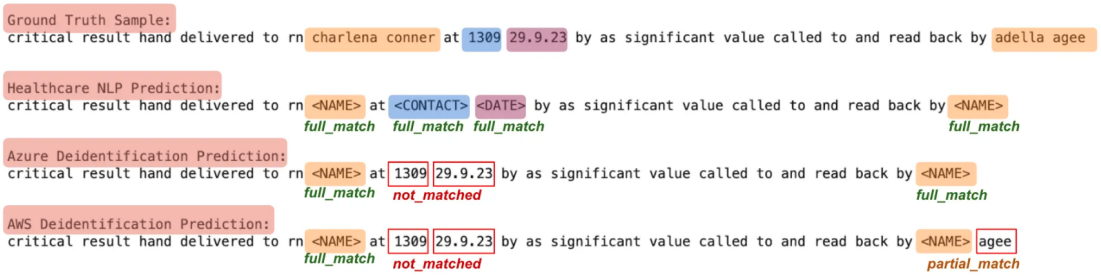

Figure A3: De-identification Results of the Tools on a Sample Text

Table A3
Comparison of De-identified Entities

Ground Truth Label	Healthcare NLP Library	Azure Health Data Services	AWS Medical Comprehend	GPT-4o
AGE	AGE	AGE	AGE	AGE
DATE	DATE	DATE	DATE	DATE
LOCATION	LOCATION, CITY, COUNTRY, HOSPITAL, STREET, ZIP	HOSPITAL, STREET, STATE, CITY, ZIP, LOCATIONOTHER, COUNTRYORREGION	ADDRESS	LOCATION
NAME	NAME, DOCTOR, PATIENT	DOCTOR, PATIENT	NAME	NAME
IDNUM	IDNUM, MEDICALRECORD, VIN, SSN, DLN, PLATE, ACCOUNT, LICENSE	IDNUM, MEDICALRECORD, ACCOUNT, SOCIALSECURITY	IDNUM	IDNUM
CONTACT	CONTACT, PHONE, EMAIL	PHONE, EMAIL, FAX	PHONE_OR_FAX, EMAIL	CONTACT

Table A4
Match Statistics for Healthcare NLP, Azure, AWS, and GPT-4o Predictions. The table shows the number of matches and their corresponding percentages for the different prediction models.

Match Type	Healthcare NLP	Azure	AWS	GPT-4o
Full Match	1342 (90.7%)	1258 (85.0%)	1108 (74.9%)	983 (66.5%)
Partial Match	124 (8.4%)	164 (11.1%)	219 (14.8%)	280 (18.9%)
Not Matched	13 (0.9%)	57 (3.8%)	152 (10.3%)	216 (14.6%)

Table A5
De-identified Chunk Label Counts for Different Tools

Chunk Label	Ground Truth	Healthcare NLP	Azure	AWS	GPT-4o
DATE	582	591	617	571	566
NAME	380	401	393	333	391
LOCATION	236	253	253	310	183
IDNUM	185	178	231	175	234
CONTACT	49	51	63	42	51
AGE	47	52	49	44	63

GPT-4o Prompt

You are an expert medical annotator with extensive experience in labeling medical entities within clinical texts. Your role is to accurately identify and annotate Protected Health Information (PHI) entities in the provided text, following the specified entity types.

Instructions:

1. **Review the Text**: Carefully read the text to understand its medical context.
2. **Identify PHI Entities**: Locate any terms or phrases that represent PHI, based on the following entity types:
   - IDNUM, LOCATION, DATE, AGE, NAME, CONTACT
3. **Annotate Entities**: For each identified PHI, provide the start and end character indices, the entity type, and the exact text (chunk) of the entity.
4. **Response Format**: Return the annotations in a structured JSON format, as demonstrated in the examples below.

Example:
Input Sentence:
"MD Connect Call 11:59pm 2/16/69 from Dr. Hale at Senior Care Clinic Queen Creek, SD regarding Terri Bird."
Annotated Entities:

[
{{'begin': 24, 'end': 30, 'entity_type': 'DATE', 'chunk': '2/16/69'}}
{{'begin': 42, 'end': 45, 'entity_type': 'NAME', 'chunk': 'Hale'}}
{{'begin': 50, 'end': 67, 'entity_type': 'LOCATION', 'chunk': 'Senior Care Clinic'}}
{{'begin': 69, 'end': 79, 'entity_type': 'LOCATION', 'chunk': 'Queen Creek'}}
{{'begin': 83, 'end': 84, 'entity_type': 'LOCATION', 'chunk': 'SD'}}
{{'begin': 96, 'end': 105, 'entity_type': 'NAME', 'chunk': 'Terri Bird'}}
]
—
Task:
Extract all PHI entities from the text below. The entity types to identify are: **IDNUM, LOCATION, DATE, AGE, NAME, CONTACT**.
Expected Output Format:
{entities:[
{'begin': <start_index>, 'end': <end_index>, 'entity_type': '<entity_type>', 'chunk': '<extracted_text>'}
]}
—
Text to Annotate:
{text}
—
Your Response:

Figure A4: Example of GPT-4o prompt for detecting Protected Health Information entities in clinical text