2025
ECIR Text2Story

CAN ZERO-SHOT COMMERCIAL API'S DELIVER REGULATORY-GRADE CLINICAL TEXT DE-IDENTIFICATION?

Veysel Kocaman, Muhammet Santas, Yigit Gul, Mehmet Butgul and David Talby

John Snow Labs Inc., Delaware, USA

4

John Snow LABS

Abstract

Objective: Evaluate Azure Health Data Services, AWS Comprehend Medical, Claude 3.7 Sonnet and Open AI GPT-4o against Healthcare NLP for PHI de-identification.
Dataset: 48 clinical documents annotated by medical experts.
Evaluation Metrics: Assessed at both entity-level and token-level performance.
Results: Healthcare NLP achieved the highest F1-score (96%) vs. Azure (91%), AWS (83%), and GPT-4o (79%).
Cost Efficiency: Healthcare NLP reduces processing costs by over 80% compared to Azure and GPT-4o, thanks to its fixed-cost, local deployment model.
Key Findings: Commercial zero-shot APIs fall short in accuracy, adaptability, and cost-efficiency. Healthcare NLP offers superior performance, customization, and scalability.

Original Text

Harbor Hospital
----
36 Park Avenue, 95108, San Diego, CA, USA
Email: medunites@harborhospital.com,
Phone: (818) 342-7353.

TSICU MRN# 1482928 on 24/06/2019 by ambulance VIN: 1HGBH41JXMN109186.

John Davies is a 62 y.o. patient admitted to ICU after an MVA on 22 Hoyt Street, at 23:00 hours. He works as a driver, and long hours of work reported. He reports dizziness, drowsiness, headache in the frontotemporal region with skin lacerations on his right occipital auricular area. Mr. Davies was seen at 23:12 minutes by attending physician Dr. Meyer Lorand and was scheduled for emergency head and neck CT with further neurological assessment. At 23:18 he was neurologically assessed by Dr. Frank M and was HD stable with normal vital signs and therefore and transferred (ID num 184378) for further radiological investigations.

Masked

<HOSPITAL>

<STREET>, <ZIP>, <CITY>, <STATE>, <COUNTRY>
Email: <EMAIL>,
Phone: <PHONE>.
TSICU MRN# <MEDICALRECORD> on <DATE> by ambulance VIN: <VIN>.
<PATIENT> is a <AGE> y.o. patient admitted to ICU after an MVA on <STREET>, at 23:00 hours. He works as a <PROFESSION>, and long hours of work reported.
He reports dizziness, drowsiness, headache in the frontotemporal region with skin lacerations on his right occipital auricular area.
Mr. <PATIENT> was seen at 23:12 minutes by attending physician Dr. <DOCTOR> and was scheduled for emergency head and neck CT with further neurological assessment.
At 23:18 he was neurologically assessed by Dr. <DOCTOR> and was HD stable with normal vital signs and therefore and transferred (ID num <IDNUM>) for further radiological investigations.

Obfuscated

MERCY HOSPITAL ARDMORE
----
474 North Yellow Springs Street, 14235, Salt Lake City, Utah, US
Email: dalton@mercyhospital.com,
Phone: (765) 896 92 86.
TSICU MRN# US:3025146 on 15/08/2019 by ambulance VIN: 1AAAA00AAAA111000.
Meldon Lemon is a 58 y.o. patient admitted to ICU after an MVA on 390 40th street, at 23:00 hours.
He works as a special educational needs teacher, and long hours of work reported.
He reports dizziness, drowsiness, headache in the frontotemporal region with skin lacerations on his right occipital auricular area.
Mr. Lemon was seen at 23:12 minutes by attending physician Dr. Evangeline Kelly and was scheduled for emergency head and neck CT with further neurological assessment.
At 23:18 he was neurologically assessed by Dr. Lara Courier and was HD stable with normal vital signs and therefore and transferred (ID num 453267) for further radiological investigations.

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.

Methodology

Comprehensive Healthcare NLP Solution: Healthcare NLP (Spark NLP) offers 2,500+ pre-trained models for medical text processing, including NER, information extraction, and clinical text analysis.
The library provides advanced PHI de-identification using NER models, ensuring compliance with privacy regulations while maintaining data utility for research.

Azure Health Data Services: Uses NLP to identify, label, redact, and surrogate PHI in clinical texts, ensuring compliance with HIPAA, GDPR, and CCPA.

Amazon Comprehend Medical: HIPAA-compliant NLP service that extracts medical entities and PHI from clinical texts using machine learning, aiding data automation and workflow optimization.

Open AI GPT-4o: Multimodal model with improved classification accuracy and response times compared to GPT-4, potentially enhancing PHI identification and redaction via prompting. While previous GPT models have been studied for medical text de-identification, GPT-4o's effectiveness in this area remains unverified due to the lack of empirical evaluations.

Anthropic Claude 3.7 Sonnet : Mid-tier model from Anthropic, balances speed and accuracy, making it a strong candidate for healthcare AI. It demonstrates high contextual awareness in PHI extraction, effectively identifying and categorizing patient data.

Entity-Level Evaluation

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.

Ground Truth Sample:
critical result hand delivered to rn charlena conner at 1309 29.9.23 by as significant value called to and read back by adella agee

Healthcare NLP Prediction:
critical result hand delivered to rn <NAME> at <CONTACT> <DATE> by as significant value called to and read back by <NAME>
full_match full_match full_match full_match

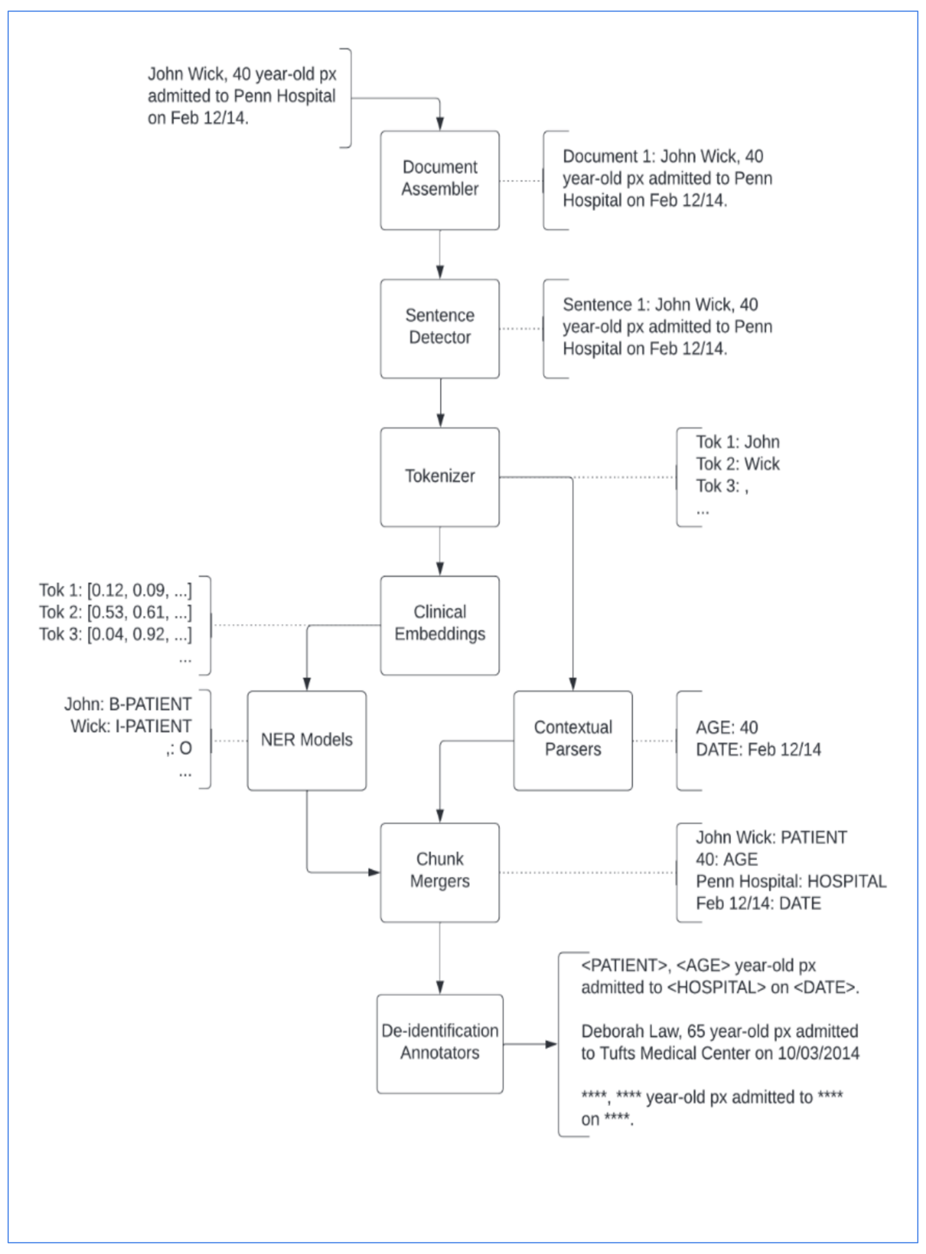
Azure Deidentification Prediction:
critical result hand delivered to rn <NAME> at 1309 29.9.23 by as significant value called to and read back by <NAME>
full_match not_matched full_match

AWS Deidentification Prediction:
critical result hand delivered to rn <NAME> at 1309 29.9.23 by as significant value called to and read back by <NAME> agee
full_match not_matched partial_match

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

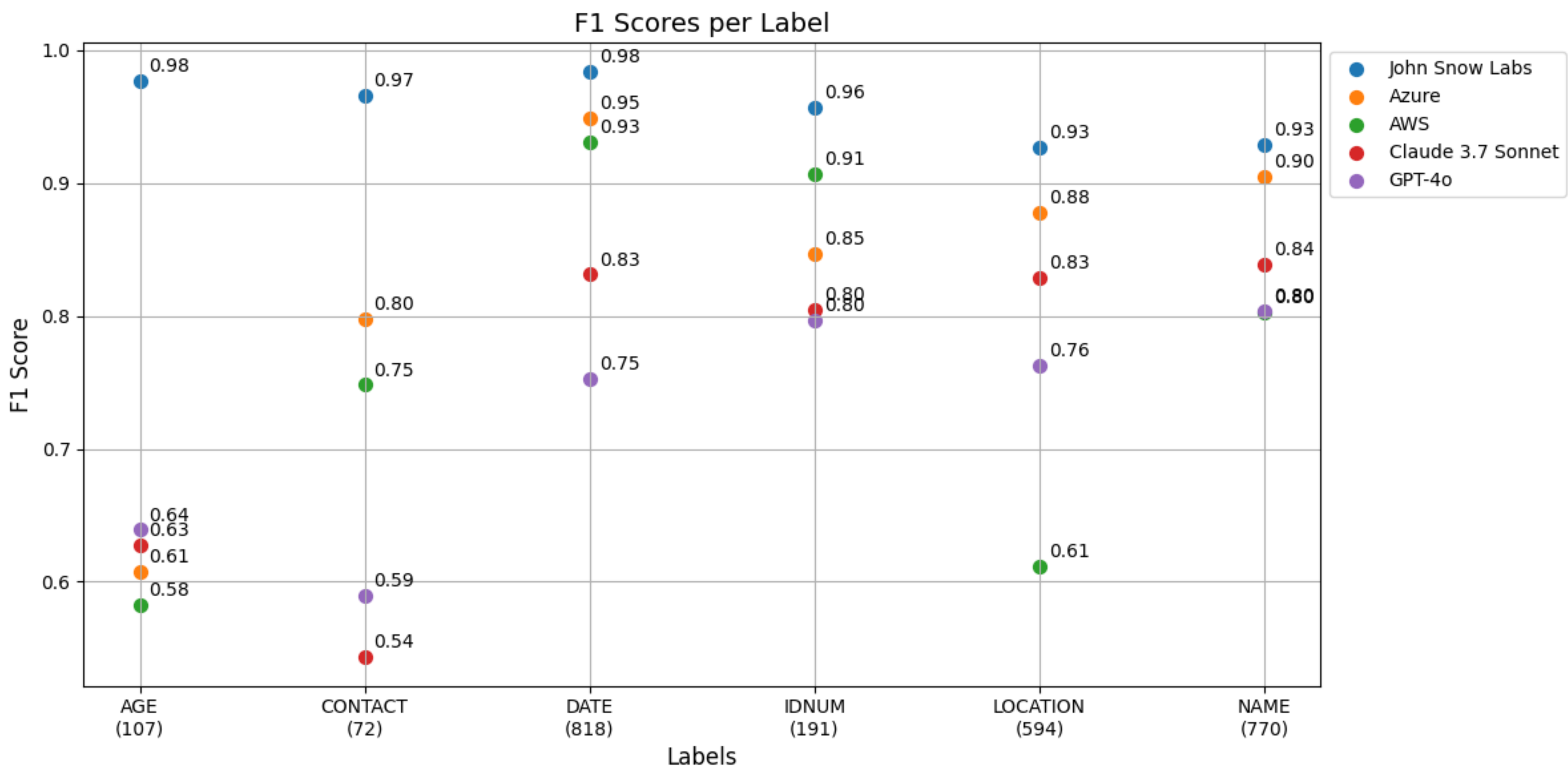
token	token_label	healthcare_nlp_token_label	azure_token_label	aws_token_label
957770228	IDNUM	IDNUM	IDNUM	IDNUM
FIH	LOCATION	LOCATION	LOCATION	O
0408267	IDNUM	IDNUM	IDNUM	IDNUM
46769/5v7d	IDNUM	NAME	IDNUM	IDNUM
237890	IDNUM	IDNUM	IDNUM	IDNUM
2/5/1994	DATE	DATE	DATE	DATE
12:00:00	O	O	O	O
AM	O	O	O	O
TRACHEOESOPHAGEAL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



Results

Healthcare NLP, Azure, Amazon GPT-4o and Claude 3.7 Sonnet PHI Recognition and Benchmark Comparison (Sample size: 45172 PHI entities).

Metric / Entity	Healthcare NLP			Azure			Amazon			GPT-4o			Claude 3.7 Sonnet		
	Precision	Recall	F1-score	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	0.73	0.55	0.63
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	0.74	0.43	0.54
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	0.84	0.83	0.83
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	0.70	0.95	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	0.79	0.87	0.83
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	0.82	0.86	0.84
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	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	0.80	0.78	0.78
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	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	0.81	0.84	0.83
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	0.90	0.92	0.91
cost per 1M doc		**$2,418**			$13,125			$14,525			$21,400			$23,330	



Conclusion

Performance Comparison: Healthcare NLP achieved the highest accuracy, followed by Azure Health Data Services, Amazon Comprehend Medical, Claude 3.7 Sonnet and GPT-4o.
Evaluation Metrics: Healthcare NLP ranked highest in precision, recall, and F1-score across both entity-level and token-level evaluations.
Adaptability & Customization: Healthcare NLP allows pipeline modifications, while other solutions are black-box APIs with no customization.
Cost-Effectiveness: Healthcare NLP enables fixed-cost, local deployment, whereas cloud-based solutions have per-request pricing that scales with data volume.
Final Verdict: Healthcare NLP outperformed alternatives by 5-10%, offering superior accuracy, flexibility, and cost-efficiency.