John Snow LABS

Healthcare Language Models for Information Extraction

Gursev Pirge
08-04-2025

John Snow LABS

Healthcare Language Models

140+ million

Downloads on PyPI.
"Most Widely Used NLP Library in the Enterprise."

O'Reilly Media

54% share

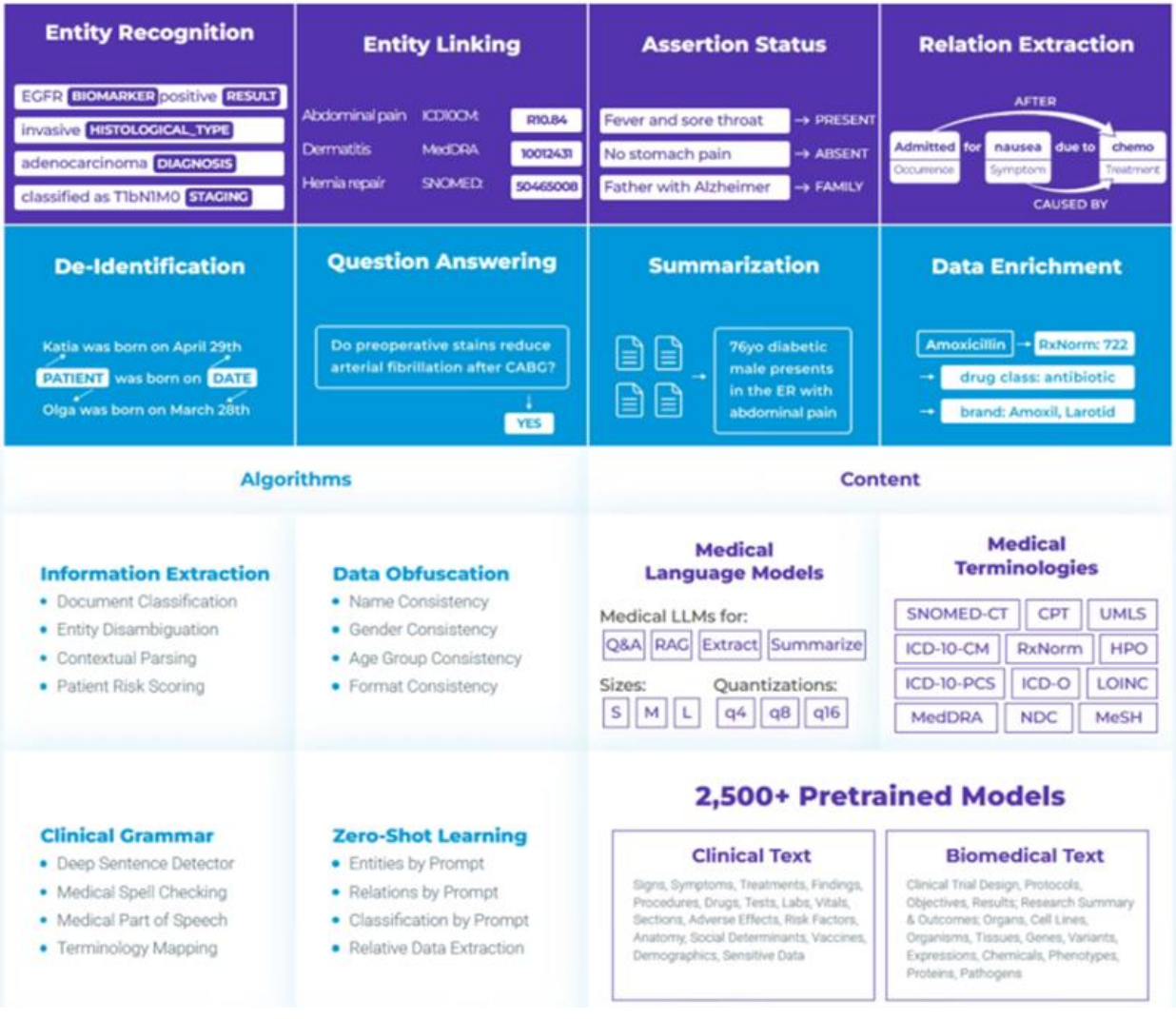
of GenAI in Healthcare projects at large companies use John Snow Labs

Gradient Flow

#1 Accuracy

on 25 benchmarks in peer-reviewed papers

Papers with Code

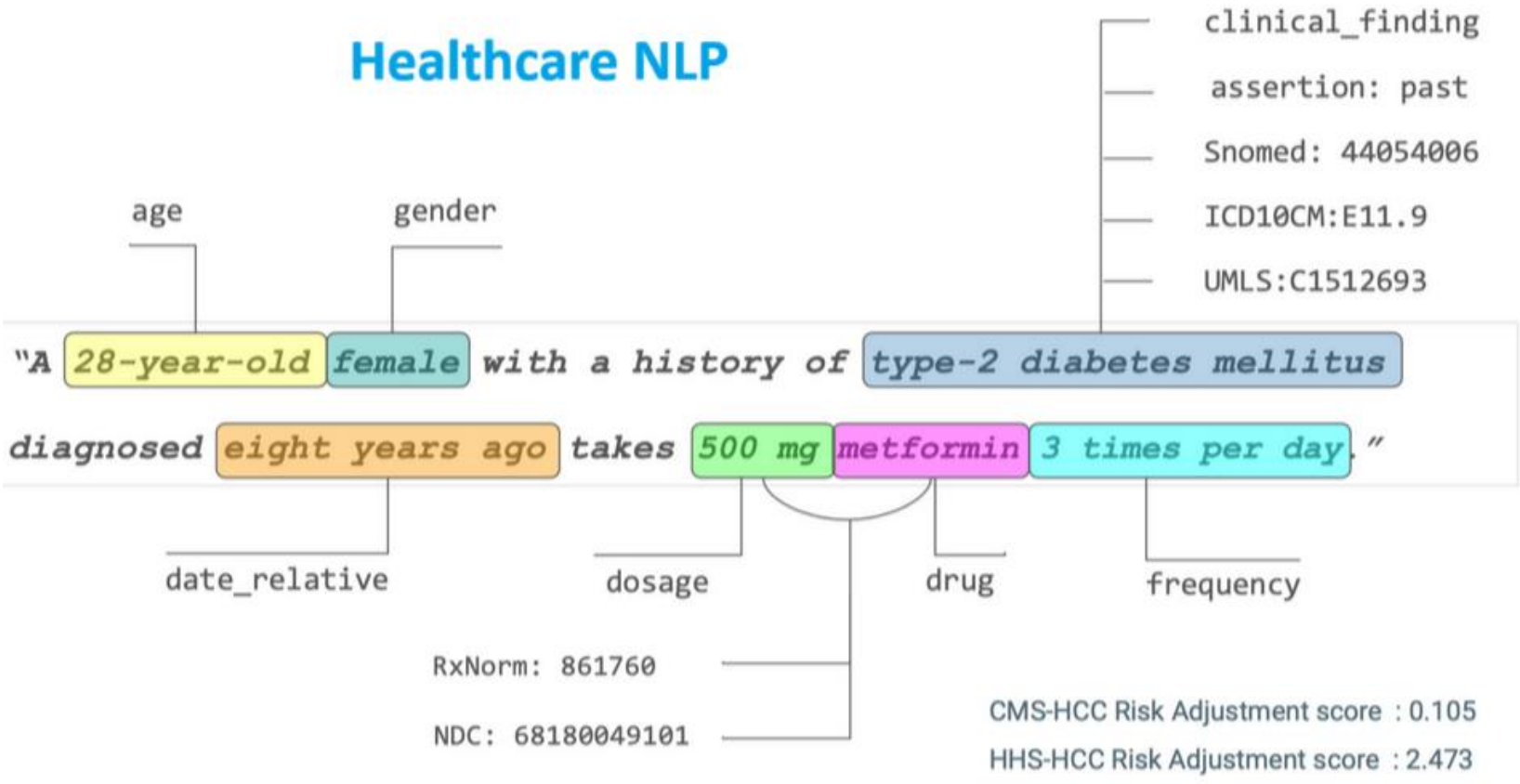


John Snow LABS

Healthcare Language Models

Named Entity Recognition

John Snow LABS

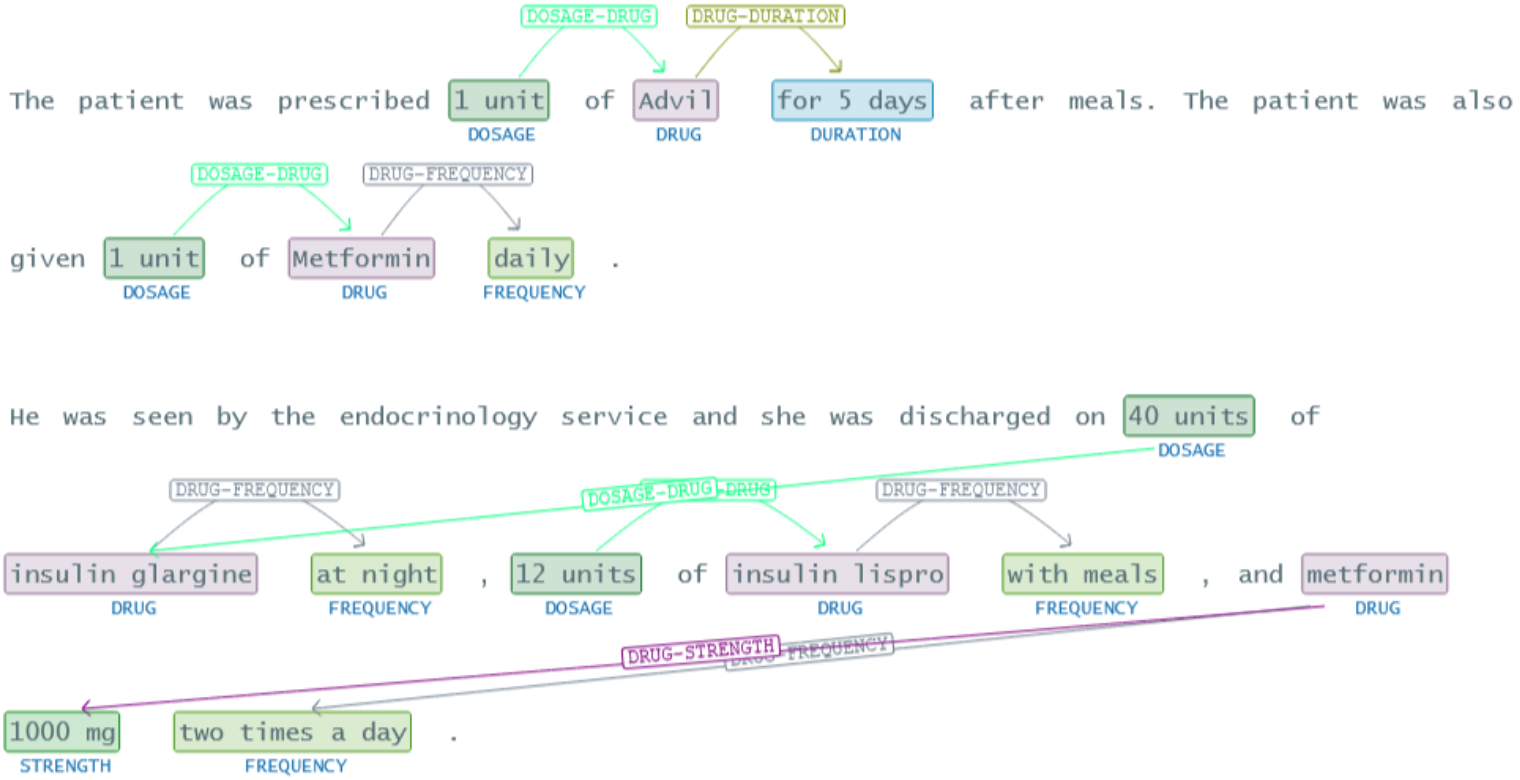


Assertion Status Detection

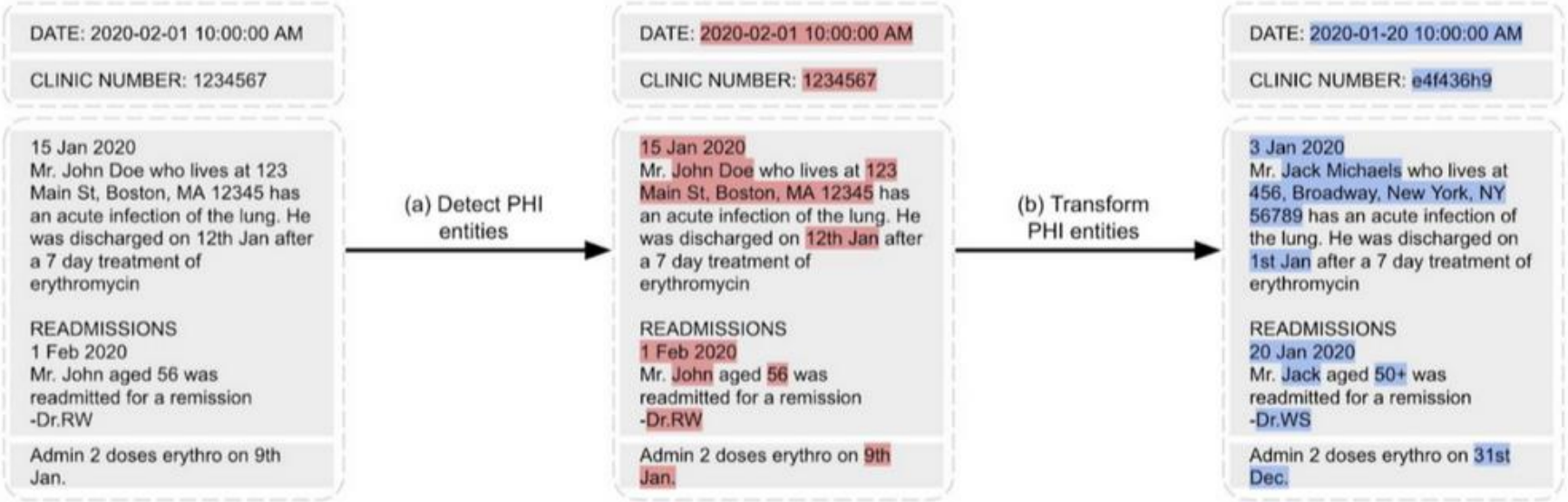
"Mother with a lung cancer, a patient is diagnosed as breast cancer in 1991 and then admitted to Mayo Clinic in Oct 2000, went under chemo for 6 months, discharged in April 2001 with a prescription of 2 mg metformin 3x per day. No sign of gynecological disorder but she suffers from acute cramps if she doesn't take her drug."

Chunk	Entity	Assertion
lung cancer	Oncological	Family
breast cancer	Oncological	Past
chemo	Treatment	Past
gynecological disorder	Disorder	Absent
acute cramps	Disorder	Conditional

Relation Extraction



Deidentification



Clinical Text De-Identification of PHI Data

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	

[Submitted on 21 Mar 2025 (v1), last revised 31 Mar 2025 (this version, v2)]

Can Zero-Shot Commercial APIs Deliver Regulatory-Grade Clinical Text Deidentification?

Veysel Kocaman, Muhammed Santas, Yigit Gul, Mehmet Butgul, David Talby