USE CASE Healthcare-Specific Large Language Models in Action

LUCA MARTIAL

Senior Data Scientist

John Snow Labs

John Snow LABS

Healthcare-Specific Large Language Models in Action

LUCA MARTIAL
Senior Data Scientist
John Snow Labs

John Snow LABS

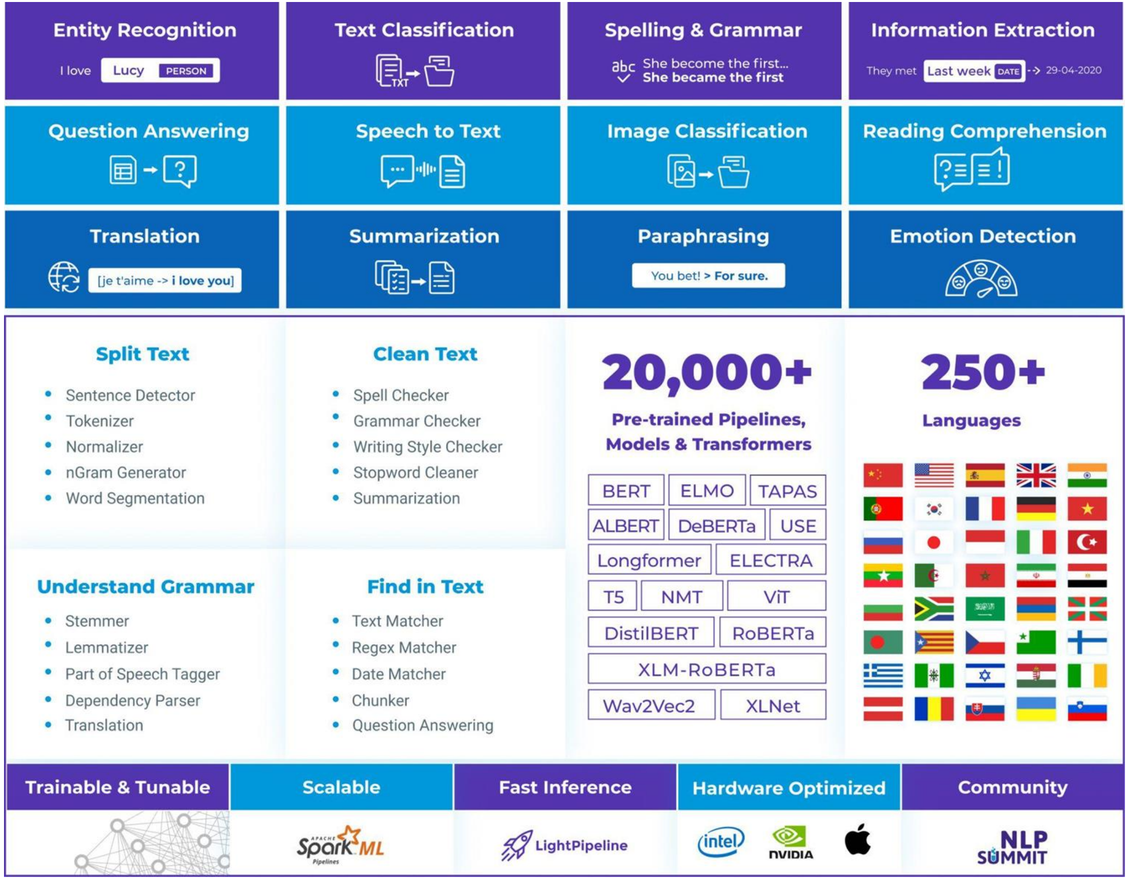
JohnSnowLabs / spark-nlp Public

State of the Art Natural Language Processing

nlp.johnsnowlabs.com

Apache-2.0 license

3k stars 608 forks



GE Healthcare

Mount Sinai

MERCK

KAISER PERMANENTE

Pfizer

NOVARTIS

FDA

Roche

Cincinnati Children's

Providence

Entity Recognition

40 units DOSAGE of
insulin glargine DRUG
at night FREQUENCY

Entity Linking

Suspect diabetes SNOMED-CT: 473127005
Lisinopril 10 MG RxNorm: 316151
Hyponatremia ICD-10: E87.1

Assertion Status

Fever and sore throat → PRESENT
No stomach pain → ABSENT
Father with Alzheimer → FAMILY

Relation Extraction

AFTER
Admitted for nausea due to chemo
Occurrence Symptom Treatment
CAUSED BY

De-Identification

Katia was born on April 29th
PATIENT was born on DATE
Olga was born on March 28th

Question Answering

Do preoperative stains reduce arterial fibrillation after CABG?
YES

Summarization

76yo diabetic male presents in the ER with abdominal pain

Data Enrichment

Amoxicillin → RxNorm: 722
→ drug class: antibiotic
→ brand: Amoxil, Larotid

Algorithms

Information Extraction

- Document Classification
- Entity Disambiguation
- Contextual Parsing
- Patient Risk Scoring

Data Obfuscation

- Name Consistency
- Gender Consistency
- Age Group Consistency
- Format Consistency

Clinical Grammar

- Deep Sentence Detector
- Medical Spell Checking
- Medical Part of Speech
- Terminology Mapping

Zero-Shot Learning

- Entities by Prompt
- Relations by Prompt
- Classification by Prompt
- Relative Data Extraction

Content

Medical Language Models

BioGPT | BioBERT | JSL-BERT
JSL-sBERT | ClinicalBERT
GloVe-Med | T5 | Flan-T5

Medical Terminologies

SNOMED-CT | CPT | UMLS
ICD-10-CM | RxNorm | HPO
ICD-10-PCS | ICD-O | LOINC

1,000+ Pretrained Models

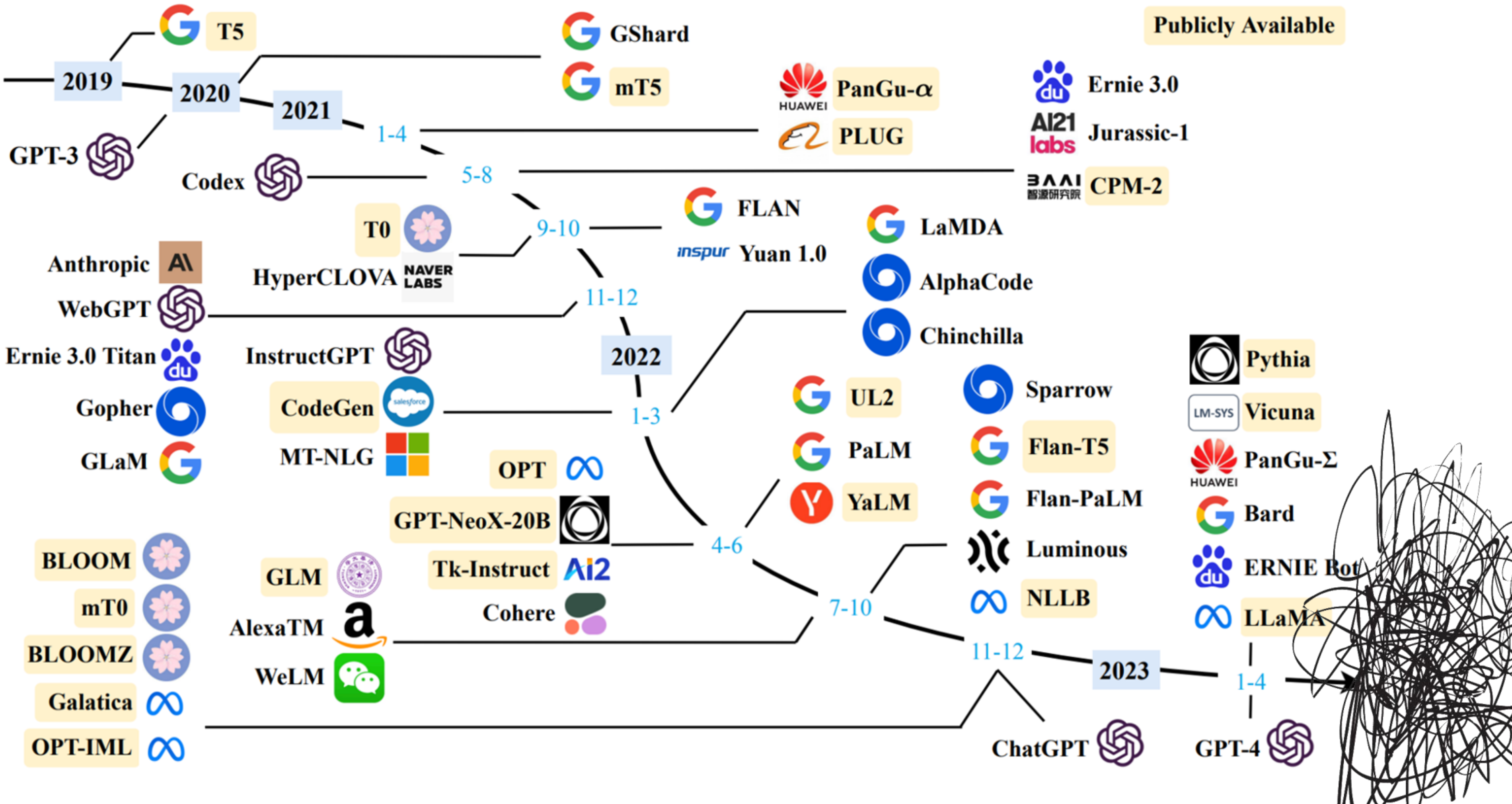
Clinical Text

Signs, Symptoms, Treatments, Findings, Procedures, Drugs, Tests, Labs, Vitals, Sections, Adverse Effects, Risk Factors, Anatomy, Social Determinants, Vaccines, Demographics, Sensitive Data

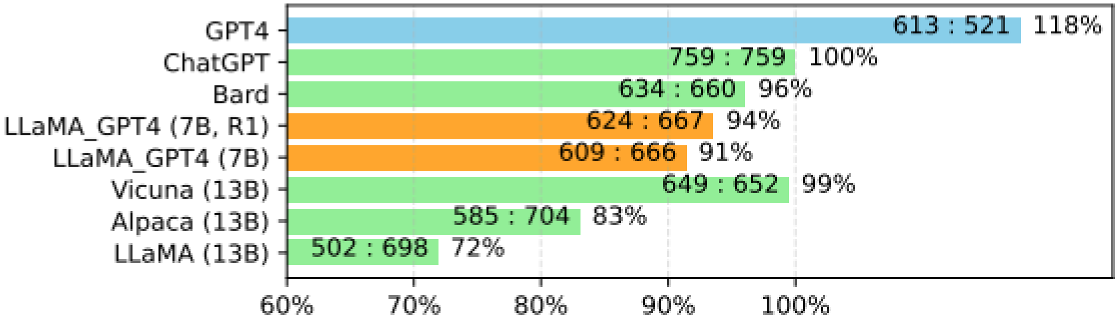
Biomedical Text

Clinical Trial Design, Protocols, Objectives, Results; Research Summary & Outcomes; Organs, Cell Lines, Organisms, Tissues, Genes, Variants, Expressions, Chemicals, Phenotypes, Proteins, Pathogens

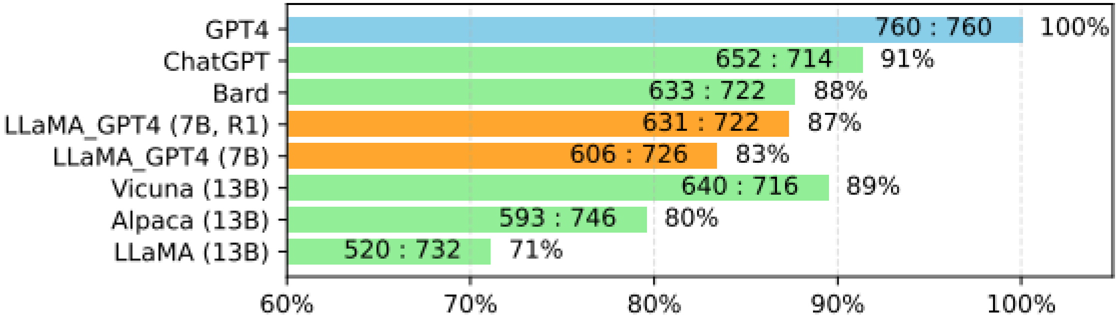
Trainable & Tunable	Scalable	Fast Inference	Hardware Optimized	Community
	Spark ML	LightPipeline	intel, NVIDIA	NLP SUMMIT



2019
2020
2021
1-4
5-8
9-10
11-12
2022
1-3
4-6
7-10
11-12
2023
1-4
Publicly Available
T5
GPT-3
GShard
mT5
PanGu-α
HUAWEI
PLUG
Ernie 3.0
AI21 labs
Jurassic-1
Codex
CPM-2
FLAN
T0
inspur
Yuan 1.0
LaMDA
Anthropic
HyperCLOVA
NAVER LABS
AlphaCode
WebGPT
Chinchilla
Ernie 3.0 Titan
InstructGPT
Gopher
CodeGen
salesforce
MT-NLG
GLaM
UL2
PaLM
YaLM
OPT
GPT-NeoX-20B
Sparrow
Flan-T5
Flan-PaLM
Luminous
NLLB
Pythia
LM-SYS
Vicuna
PanGu-Σ
HUAWEI
Bard
ERNIE Bot
LLaMA
BLOOM
mT0
BLOOMZ
Galatica
OPT-IML
GLM
AlexaTM
WeLM
Tk-Instruct
Ai2
Cohere
ChatGPT
GPT-4



(c) All chatbots against ChatGPT



(d) All chatbots against GPT-4

- GPT4all
- OpenChatKit
- Alpaca
- Lit-LLaMA
- Dolly
- **Vicuna**
- medAlpaca
- ColossalChat
- Cerebras-GPT
- Koala
- Baize

Instruction Tuning with GPT-4 (Peng et al., 2023)

Real-World Impact: A great productivity tool!

Apoorva Govind @Appyg99

Can confirm. Just yesterday, my teammate wrote a custom JSON parser that would have been a slog for 1-2 days within hours using GPT4. GPT4 is like giving each engineer their own intern who does 70-80% of the job. The remaining 30% is constraints & biz logic. Startups are going to be insanely efficient w/ hiring & execution!

TECH DRIVERS

OpenAI-powered app from Microsoft will instantly transcribe patient notes during doctor visits

PUBLISHED MON, MAR 20 2023·8:00 AM EDT | UPDATED MON, MAR 20 2023·11:17 AM EDT

Ashley Capoot
@ASHLEYCAPOOT

WATCH LIVE

KEY POINTS

- Microsoft and its Nuance Communications subsidiary announced Dragon Ambient eXperience (DAX) Express, a clinical notes application for healthcare workers that is powered by artificial intelligence.
- DAX Express aims to help reduce clinicians' administrative burdens by automatically generating a draft of clinical notes after a patient visit.
- The technology is powered by a combination of ambient A.I,. which forms insights from unstructured data like conversations, and OpenAI's newest model, GPT-4.

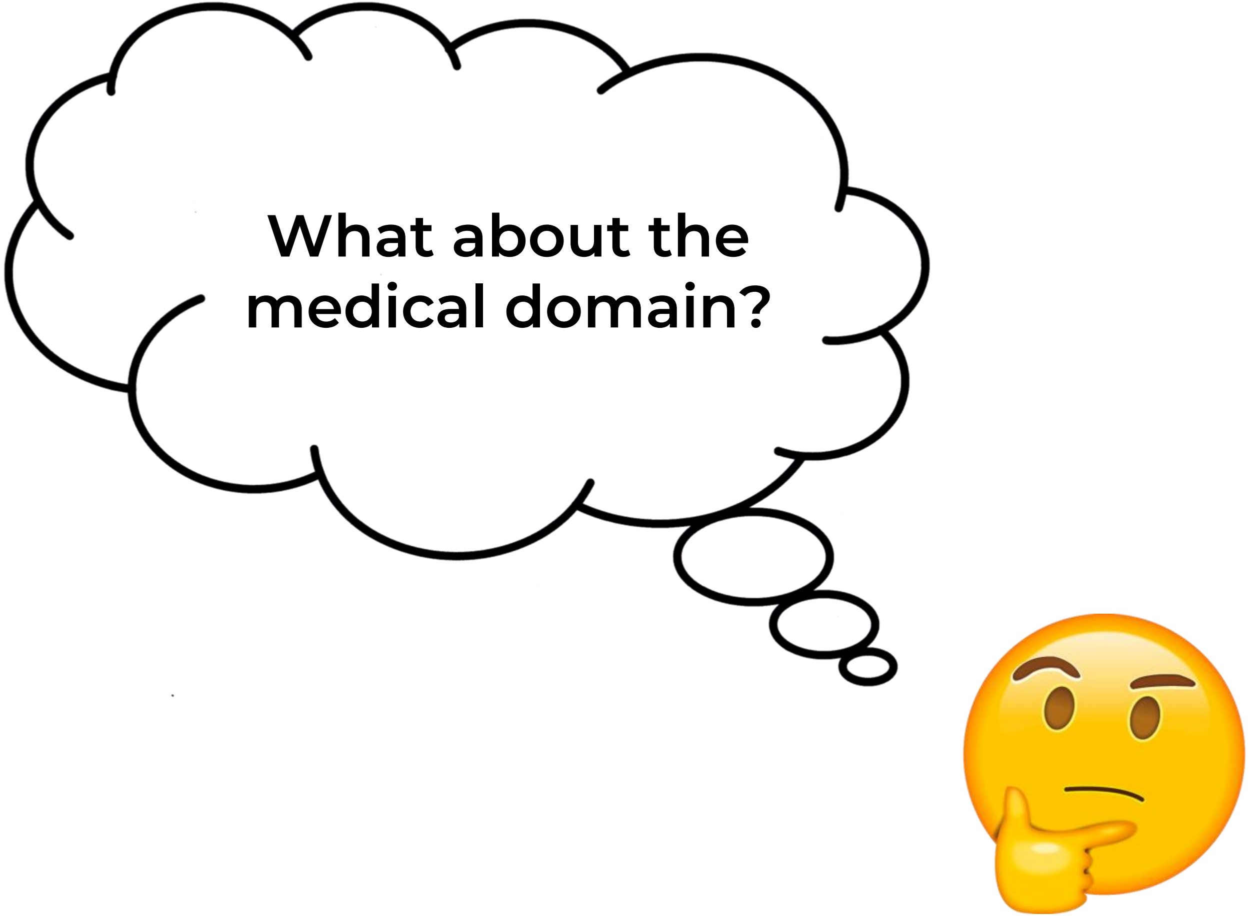
THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech **Markets** Opinion Books & Arts Real Estate

MARKETS | HEARD ON THE STREET

AI Has Its 'iPhone Moment'

The launch of ChatGPT has set off a frenzy by companies to show they are in the AI game


What about the
medical domain?

Capabilities of GPT-4 on Medical Challenge Problems

Dataset	Component	**GPT-4 (5 shot)**	GPT-4 (zero shot)	GPT-3.5 (5 shot)	GPT-3.5 (zero shot)	Flan-PaLM 540B* (few shot)
MedQA	Mainland China	**75.31**	71.07	44.89	40.31	–
	Taiwan	**84.57**	82.17	53.72	50.60	–
	United States (5-option)	**78.63**	74.71	47.05	44.62	–
	United States (4-option)	**81.38**	78.87	53.57	50.82	60.3**
PubMedQA	Reasoning Required	74.40	75.20	60.20	71.60	**79.0**
MedMCQA	Dev	**72.36**	69.52	51.02	50.08	56.5
MMLU	Clinical Knowledge	**86.42**	86.04	68.68	69.81	77.00
	Medical Genetics	**92.00**	91.00	68.00	70.00	70.00
	Anatomy	**80.00**	**80.00**	60.74	56.30	65.20
	Professional Medicine	**93.75**	93.01	69.85	70.22	83.80
	College Biology	93.75	**95.14**	72.92	72.22	87.50
	College Medicine	76.30	**76.88**	63.58	61.27	69.90

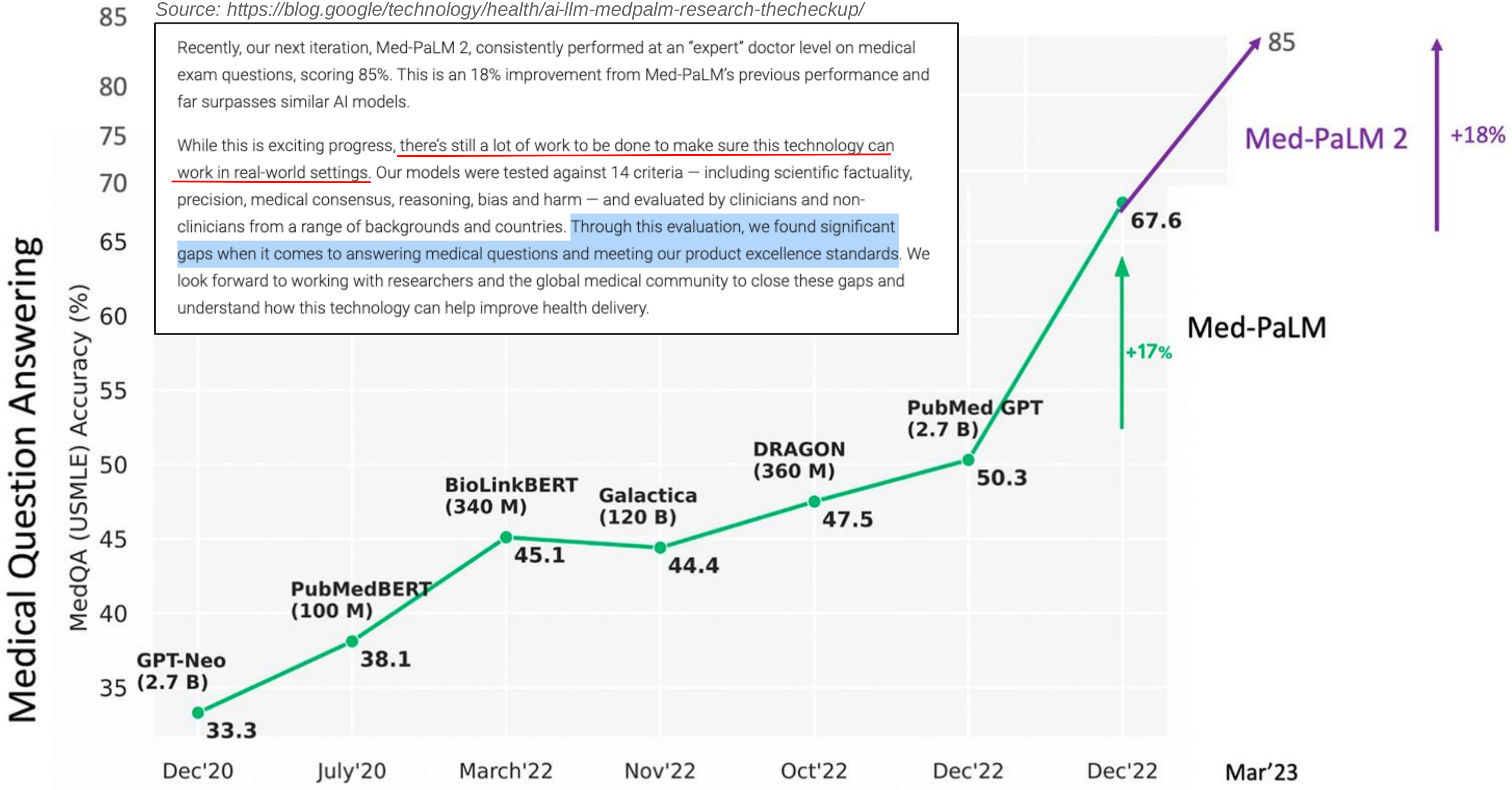
Capabilities of GPT-4 on Medical Challenge Problems (Nori et al., 2023)

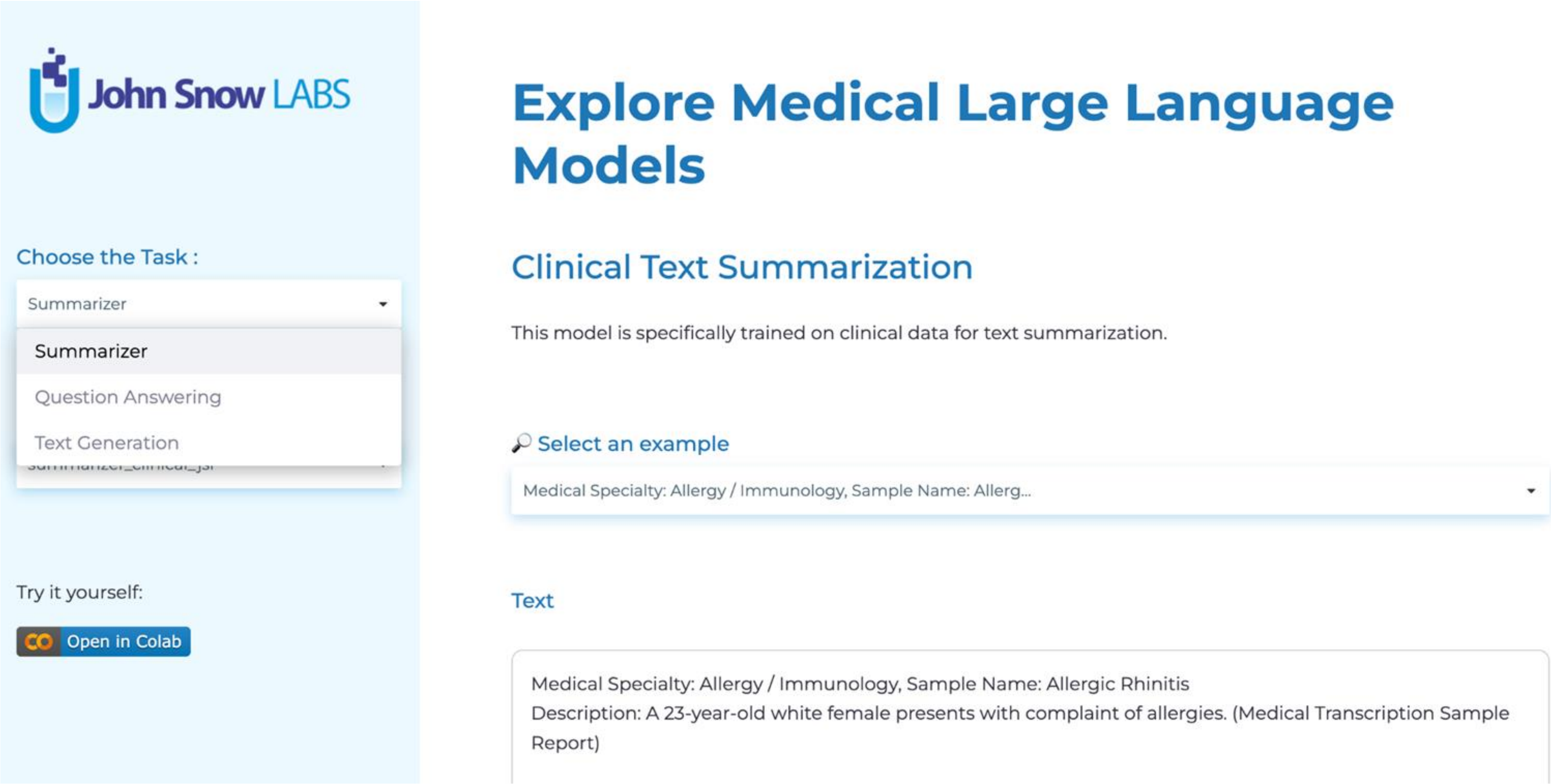
Google’s MedPaLM-2 on USMLE (Medical License Exam)

Source: https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/

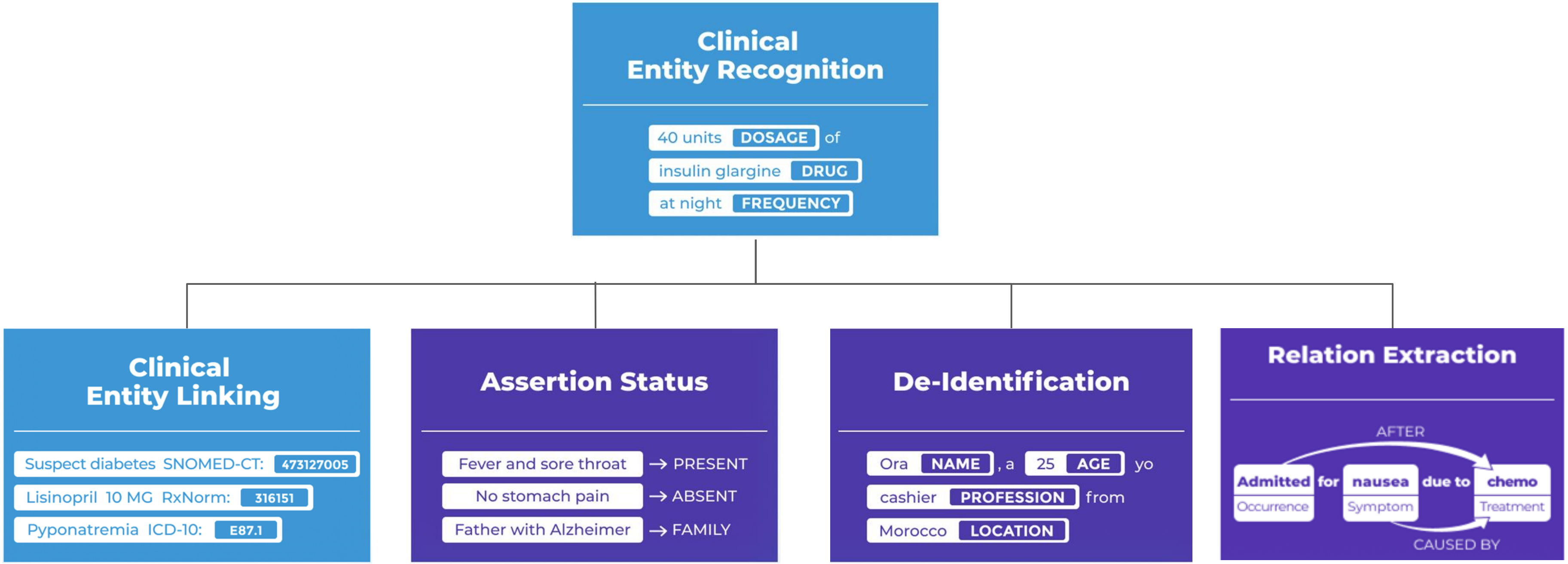
Recently, our next iteration, Med-PaLM 2, consistently performed at an “expert” doctor level on medical exam questions, scoring 85%. This is an 18% improvement from Med-PaLM’s previous performance and far surpasses similar AI models.

While this is exciting progress, there’s still a lot of work to be done to make sure this technology can work in real-world settings. Our models were tested against 14 criteria — including scientific factuality, precision, medical consensus, reasoning, bias and harm — and evaluated by clinicians and non-clinicians from a range of backgrounds and countries. Through this evaluation, we found significant gaps when it comes to answering medical questions and meeting our product excellence standards. We look forward to working with researchers and the global medical community to close these gaps and understand how this technology can help improve health delivery.





demo.johnsnowlabs.com/healthcare/


Clinical
Entity Recognition
40 units DOSAGE of
insulin glargine DRUG
at night FREQUENCY
Clinical
Entity Linking
Suspect diabetes SNOMED-CT: 473127005
Lisinopril 10 MG RxNorm: 316151
Pyponatremia ICD-10: E87.1
Assertion Status
Fever and sore throat → PRESENT
No stomach pain → ABSENT
Father with Alzheimer → FAMILY
De-Identification
Ora NAME , a 25 AGE yo
cashier PROFESSION from
Morocco LOCATION
Relation Extraction
AFTER
Admitted for nausea due to chemo
Occurrence
Symptom
Treatment
CAUSED BY

Name Entity Recognition

Please identify Person, Organization, Location and Miscellaneous Entity from the given text.

Text: *All four teams are level with one point each from one game.*
Entity:

NER

Model	*Zero-Shot*		*Fine-Tuned*		
	ChatGPT	GPT-3.5	Flair	LUKE	ACE
All	**53.7**	53.5	93.0	93.9	**94.6**
Loc	**72.2**	67.1	94.0	-	-
Per	**81.4**	78.0	97.4	-	-
Org	45.1	**50.0**	91.9	-	-
Misc	4.5	**4.8**	83.0	-	-

Is chatgpt a general-purpose natural language processing task solver? (Qin et al., 2023)

Scope of Experiments

Named Entity Recognition

- Treatment
- Test
- Age
- Problem
- Oncology
- Body Part
- Drug
- Modifier
- Procedure
- Gender
- Direction

Entity Resolution (Medical Terminologies)

- ICD10-CM
- SNOMED
- RxNorm

Text Classification

- Gender Classification
- Adverse Drug Events (ADE)

In-Depth Comparison of Spark NLP for Healthcare and ChatGPT on Clinical Named Entity Recognition

Veysel Kocaman · Follow
Published in John Snow Labs · 5 min read · Apr 18

55

Spark NLP for Healthcare NER models outperform ChatGPT by 10–45% on key medical concepts, resulting in half the errors compared to ChatGPT.

Comparing Spark NLP for Healthcare and ChatGPT in Extracting ICD10-CM Codes from Clinical Notes

Veysel Kocaman · Follow
Published in John Snow Labs · 6 min read · Apr 16

69 1

In assigning ICD10-CM codes, Spark NLP for Healthcare achieved a 76% success rate, while GPT-3.5 and GPT-4 had overall accuracies of 26% and 36% respectively.

A Comprehensive Comparison of ChatGPT and Spark NLP for Healthcare in De-Identification of Sensitive Data (PHI)

Veysel Kocaman · Follow
Published in John Snow Labs · 5 min read · Apr 18

109

Spark NLP for Healthcare De-Identification module demonstrates superior performance with a 93% accuracy rate compared to ChatGPT's 60% accuracy on detecting PHI entities in clinical notes.

spark-nlp-workshop / tutorials / academic / LLMs_in_Healthcare / benchmarks /

Add file ···

aydinmyilmaz add missing deid prompt b38305b · last month History

Name	Last commit message	Last commit date
..		
config	add benchmarks	last month
data	add deid sentences	last month
workbench	add missing deid prompt	last month
README.md	add benchmarks	last month
requirements.txt	add benchmarks	last month

Extracting Medical Problems

Prompt

```
You are a highly experienced, skilled and helpfull medical annotator who have been working on medical texts to
label medical entities.

I will provide you some entity types with sample chunks and I want you to find similar entities from given texts.

-   Entity Type: Problem
    1. Example chunks for Problem Type:  feels weak, shortness of breath, backache
    2. Example chunks for Problem Type:  gastroparesis, gastritis, allergies, pneumonitis
    3. Example chunks for Problem Type:  spine fractures, ligature strangulation, abrasions
    4. Example chunks for Problem Type:  depression, bipolar disorder, psychosis
    5. Example chunks for Problem Type:  colon cancer, mesothelioma , brachial plexus tumor
    6. Example chunks for Problem Type:  depression, anxiety, bipolar disorder, psychosis
    7. Example chunks for Problem Type:  coronary artery disease,  CAD, cardiomyopathy
    8. Example chunks for Problem Type:  renal disease, nephrolithiasis, hydronephrosis
    9. Example chunks for Problem Type:  overweight
    10. Example chunks for Problem Type: DM Type II, diabetic
    11. Example chunks for Problem Type: obese
    12. Example chunks for Problem Type: wandering atrial pacemaker, multifocal atrial tachycardia, frequent
APCs, bradycardia
    13. Example chunks for Problem Type: tuberculosis, sexually transmitted diseases, HIV
    14. Example chunks for Problem Type: increased attenuation, T1 hypointensity, opacity in apex right lung
    15. Example chunks for Problem Type: stroke, TIA
    16. Example chunks for Problem Type: increased cholesterol, hypercholesterolemia
    17. Example chunks for Problem Type: tachycardic, afebrile
    18. Example chunks for Problem Type: high blood pressure, HTN


I want you to extract Problem type of entities from the given text and label them as Problem

Task :

Find entities in the given sentence.

Answer value must be as given (valid JSON) for the given sentence as example:
{{"given_sentence": "Patient feels weak.", "list_of_entities": [{{"entity_type": "Problem", "chunk": "feels
weak"}}]}}

Now I want you to find the Problem entities in the given sentence:
```

100 sentence, ~800 entities

76%

GPT 3.5

The patient denies chest pain , irregular heartbeats , sudden changes in heartbeat or palpitation , shortness of breath , difficulty breathing at night , swollen legs or feet , heart murmurs , high blood pressure , cramps in his legs with walking , pain in his feet or toes at night or varicose veins .

89%

Spark NLP

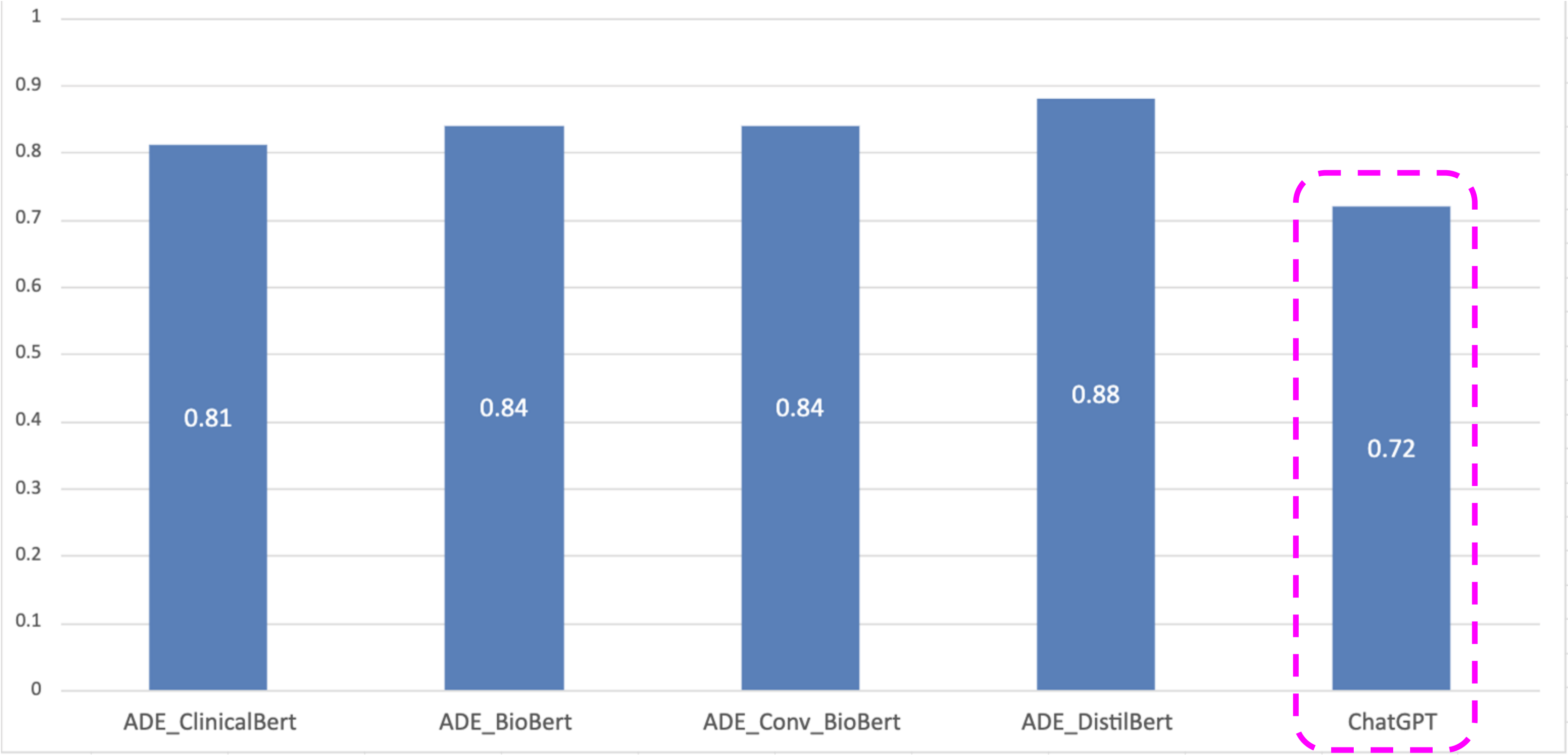
(ner_jsl_reduced)

The patient denies chest pain , irregular heartbeats , sudden changes in heartbeat or palpitation , shortness of breath , difficulty breathing at night , swollen legs or feet , heart murmurs , high blood pressure , cramps in his legs with walking , pain in his feet or toes at night or varicose veins .

* lenient metrics (partially overlapping chunks counted as hit)

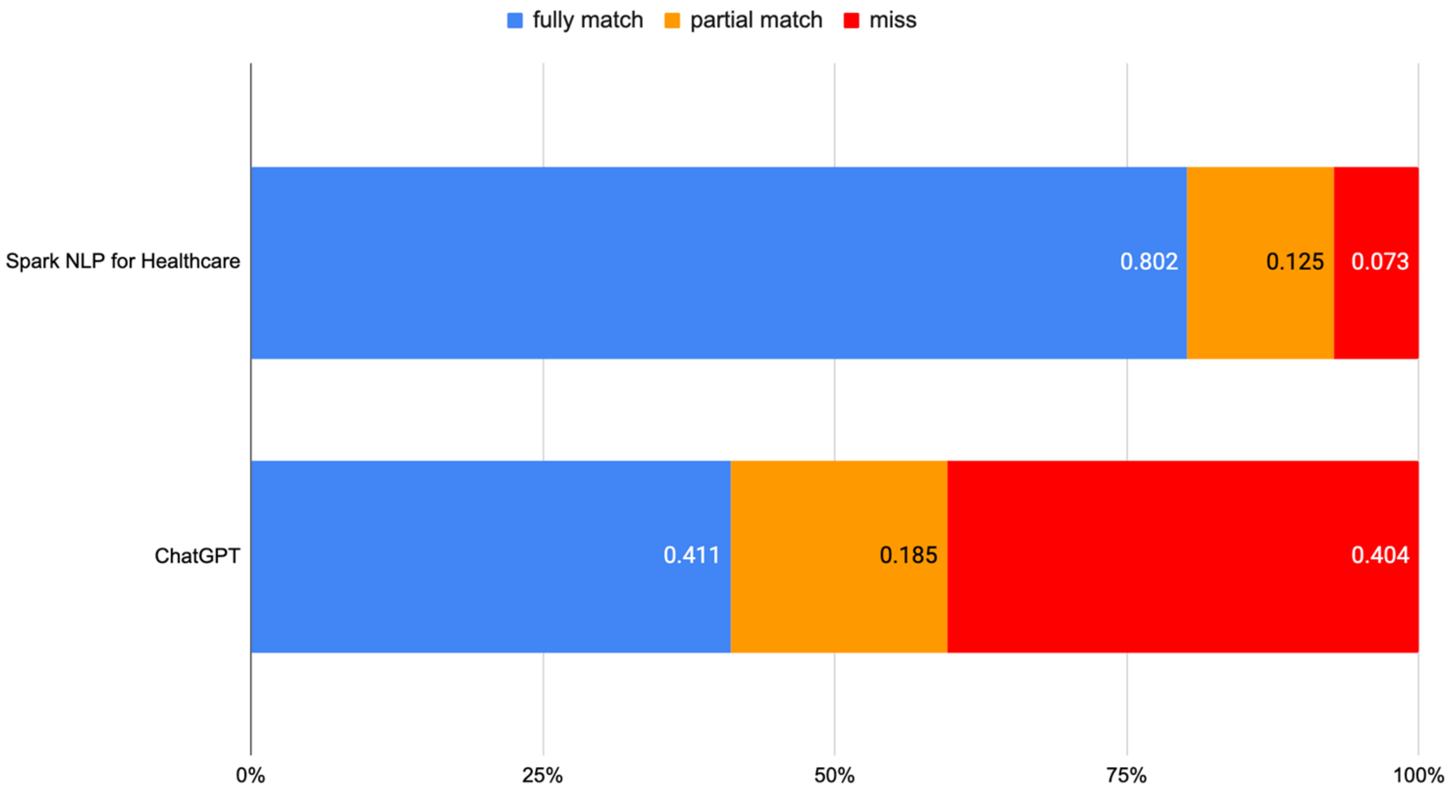
https://github.com/JohnSnowLabs/spark-nlp-workshop/tree/master/tutorials/academic/LLMs_in_Healthcare

Detecting Adverse Drug Events



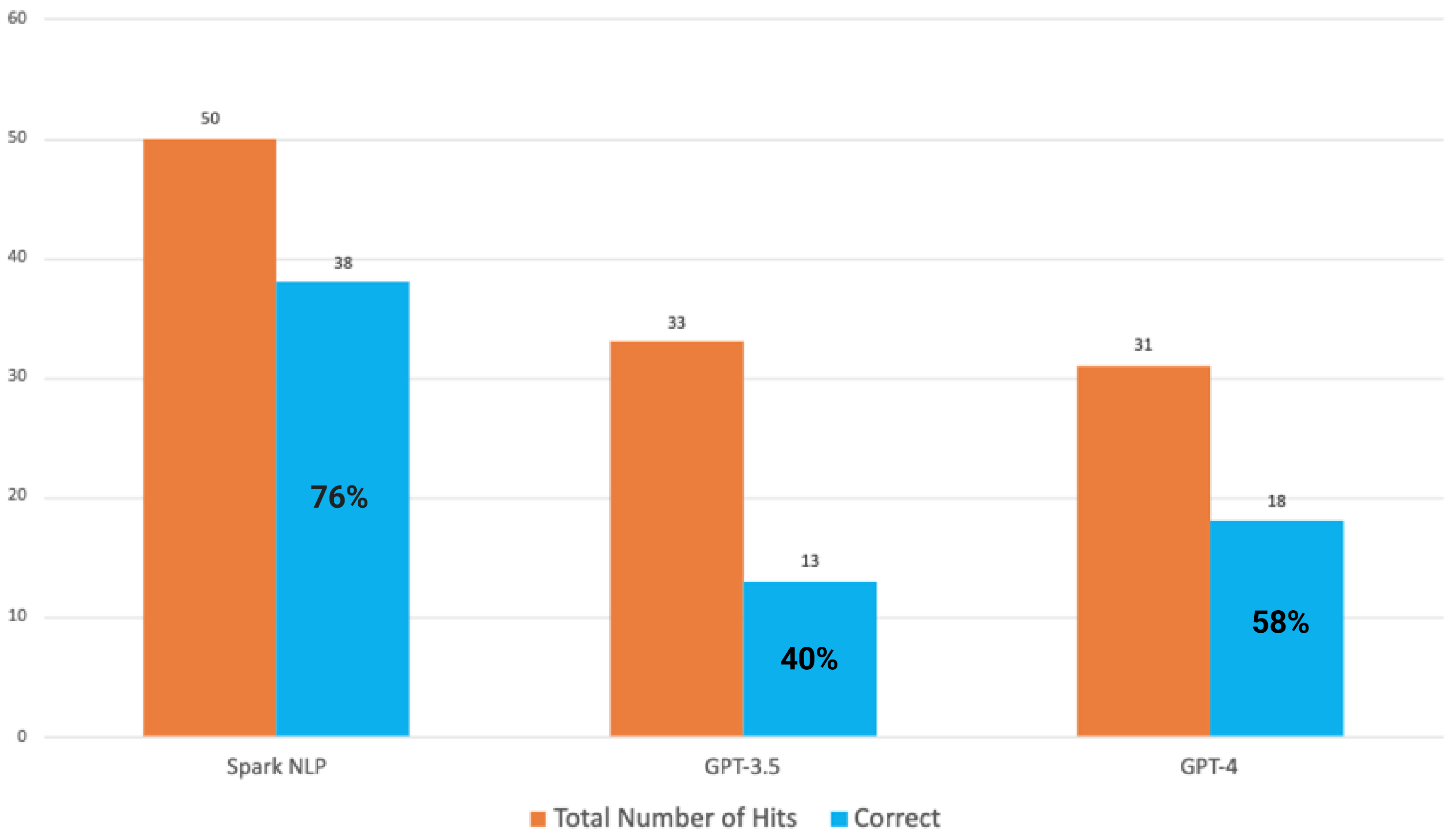
https://github.com/JohnSnowLabs/spark-nlp-workshop/tree/master/tutorials/academic/LLMs_in_Healthcare

De-Identifying PHI Data



A Comprehensive Comparison of ChatGPT and Spark NLP for Healthcare in De-Identification of Sensitive Data (Medium)

Extracting ICD10-CM Codes



Comparing Spark NLP for Healthcare and ChatGPT in Extracting ICD10-CM Codes from Clinical Notes (Medium)

LLMs: The Good, The Bad

- ✓ Tasks making HCP lives easier (note summarizing, patient profiling, etc)
- ✓ Indexing, querying databases
- ✓ Agent-based integrations

- ✗ Unreliable data abstractors
- ✗ On-prem deployment is not possible
- ✗ Hallucinating and fabricating incorrect results w/ high confidence
- ✗ Domain & task-specific fine tuning will be required (expertise, time, money)

Solutions Moving Forward

- 🚀 Granular data abstraction by high-precision models; repetitive & laborious tasks by LLMs

- 🚀 LLMs guard-railed by explainable DL/ML models, knowledge graphs, rule-based systems

- 🚀 LLMs as smart assistants (convert natural language to structured queries (SQL, Cypher))

John Snow LABS

INTELLIGENT HEALTH UK 2023

Breaking down the barriers between tech and healthcare