

Building Healthcare NLP Agents with LLMs

*Certification Trainings,
John Snow Labs
July 17th, 2024*

Veysel Kocaman, PhD

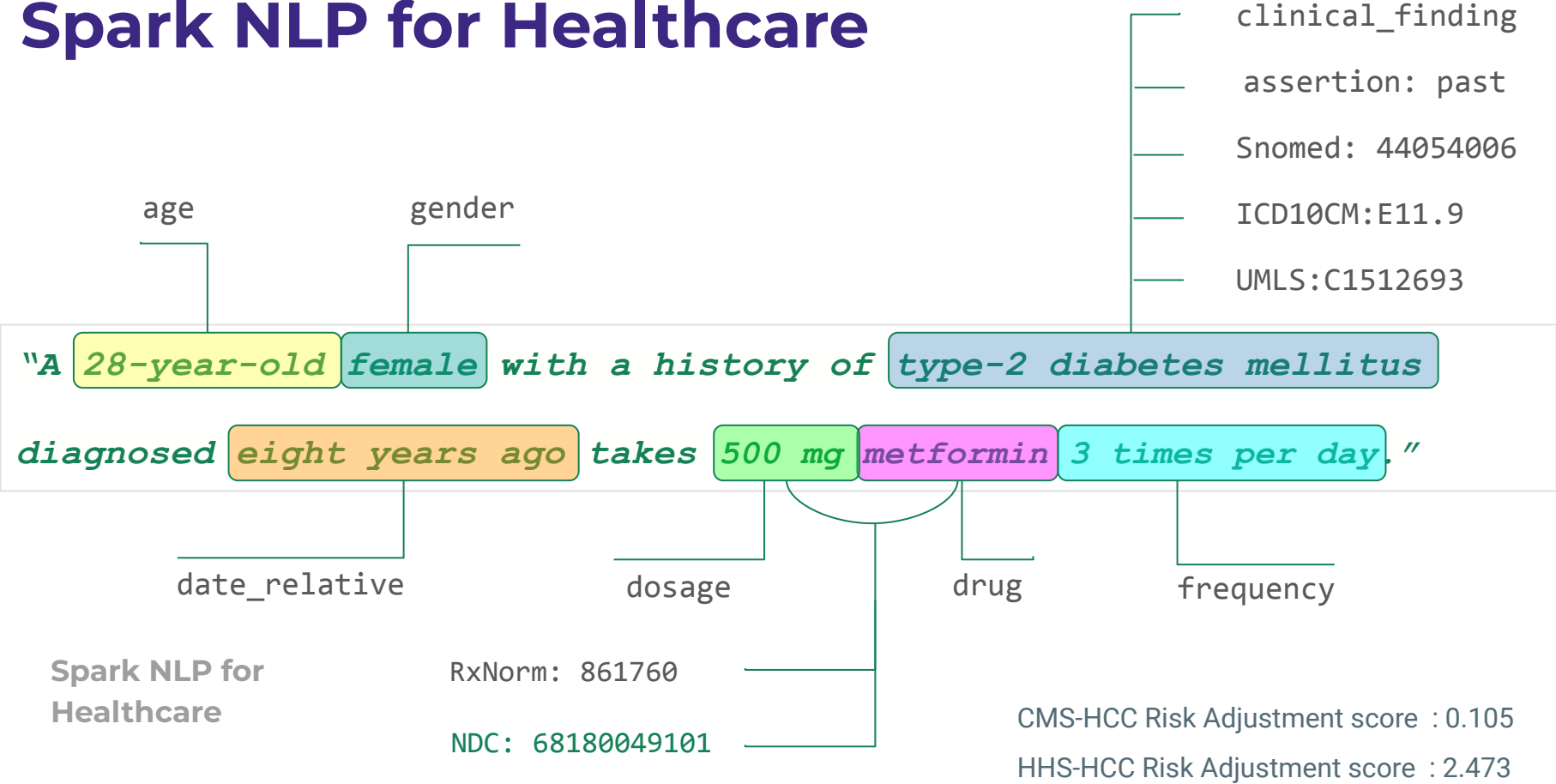
Head of Data Science
John Snow Labs



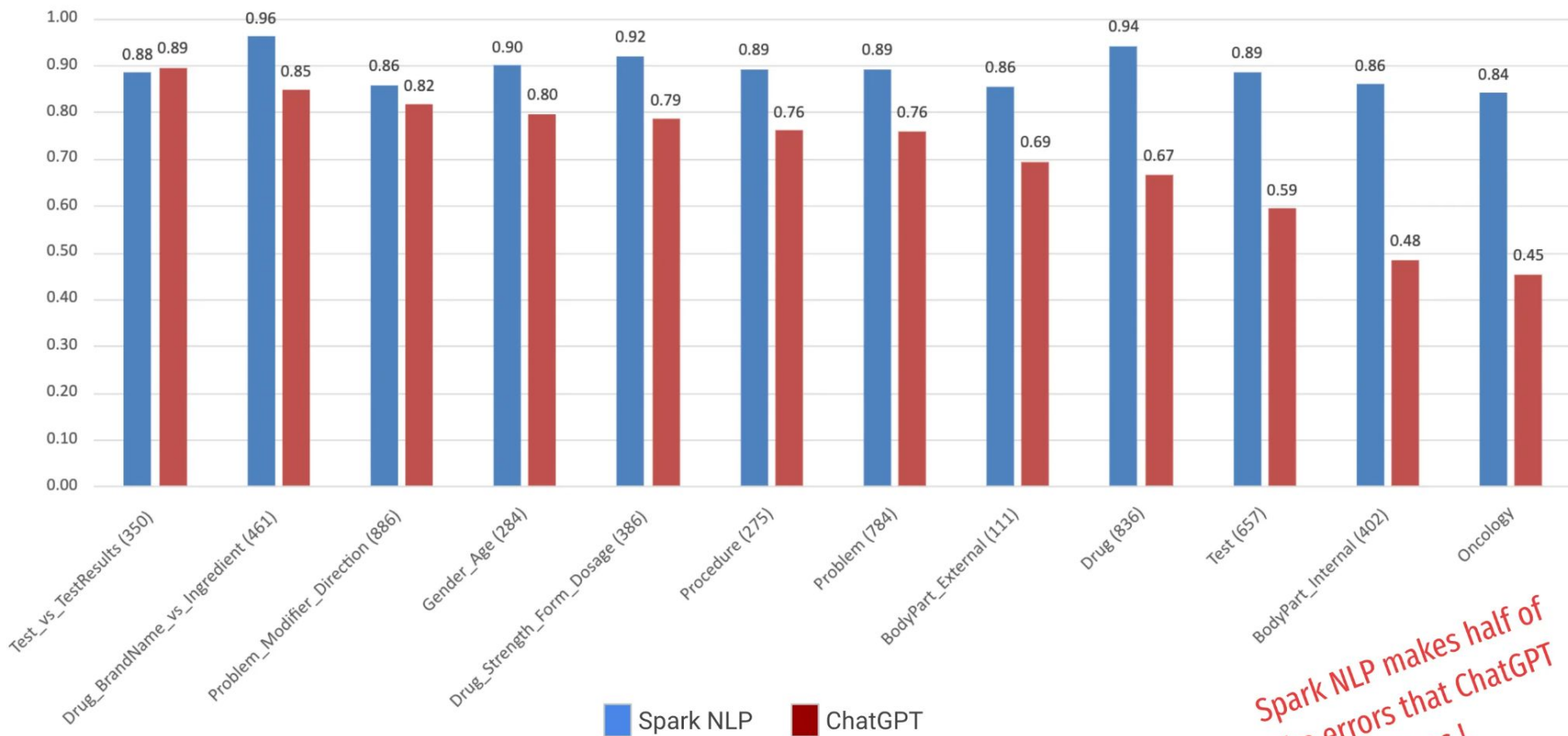
Healthcare NLP by John Snow Labs

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 																		
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 	Data Enrichment Amoxicillin → RxNorm: 722 → drug class: antibiotic → brand: Amoxil, Larotid																		
Algorithms		Content																			
Information Extraction <ul style="list-style-type: none"> Document Classification Entity Disambiguation Contextual Parsing Patient Risk Scoring 	Data Obfuscation <ul style="list-style-type: none"> Name Consistency Gender Consistency Age Group Consistency Format Consistency 	Medical Language Models <table border="1"> <tr> <td>BioGPT</td> <td>BioBERT</td> <td>JSL-BERT</td> </tr> <tr> <td>JSL-sBERT</td> <td colspan="2">ClinicalBERT</td> </tr> <tr> <td>GloVe-Med</td> <td>T5</td> <td>Flan-T5</td> </tr> </table>	BioGPT	BioBERT	JSL-BERT	JSL-sBERT	ClinicalBERT		GloVe-Med	T5	Flan-T5	Medical Terminologies <table border="1"> <tr> <td>SNOMED-CT</td> <td>CPT</td> <td>UMLS</td> </tr> <tr> <td>ICD-10-CM</td> <td>RxNorm</td> <td>HPO</td> </tr> <tr> <td>ICD-10-PCS</td> <td>ICD-O</td> <td>LOINC</td> </tr> </table>	SNOMED-CT	CPT	UMLS	ICD-10-CM	RxNorm	HPO	ICD-10-PCS	ICD-O	LOINC
BioGPT	BioBERT	JSL-BERT																			
JSL-sBERT	ClinicalBERT																				
GloVe-Med	T5	Flan-T5																			
SNOMED-CT	CPT	UMLS																			
ICD-10-CM	RxNorm	HPO																			
ICD-10-PCS	ICD-O	LOINC																			
Clinical Grammar <ul style="list-style-type: none"> Deep Sentence Detector Medical Spell Checking Medical Part of Speech Terminology Mapping 	Zero-Shot Learning <ul style="list-style-type: none"> Entities by Prompt Relations by Prompt Classification by Prompt Relative Data Extraction 	2,000+ Pretrained Models <table border="1"> <tr> <td> Clinical Text Signs, Symptoms, Treatments, Findings, Procedures, Drugs, Tests, Labs, Vitals, Sections, Adverse Effects, Risk Factors, Anatomy, Social Determinants, Vaccines, Demographics, Sensitive Data </td> <td> Biomedical Text Clinical Trial Design, Protocols, Objectives, Results; Research Summary & Outcomes; Organs, Cell Lines, Organisms, Tissues, Genes, Variants, Expressions, Chemicals, Phenotypes, Proteins, Pathogens </td> </tr> </table>		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																
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 NLP for Healthcare

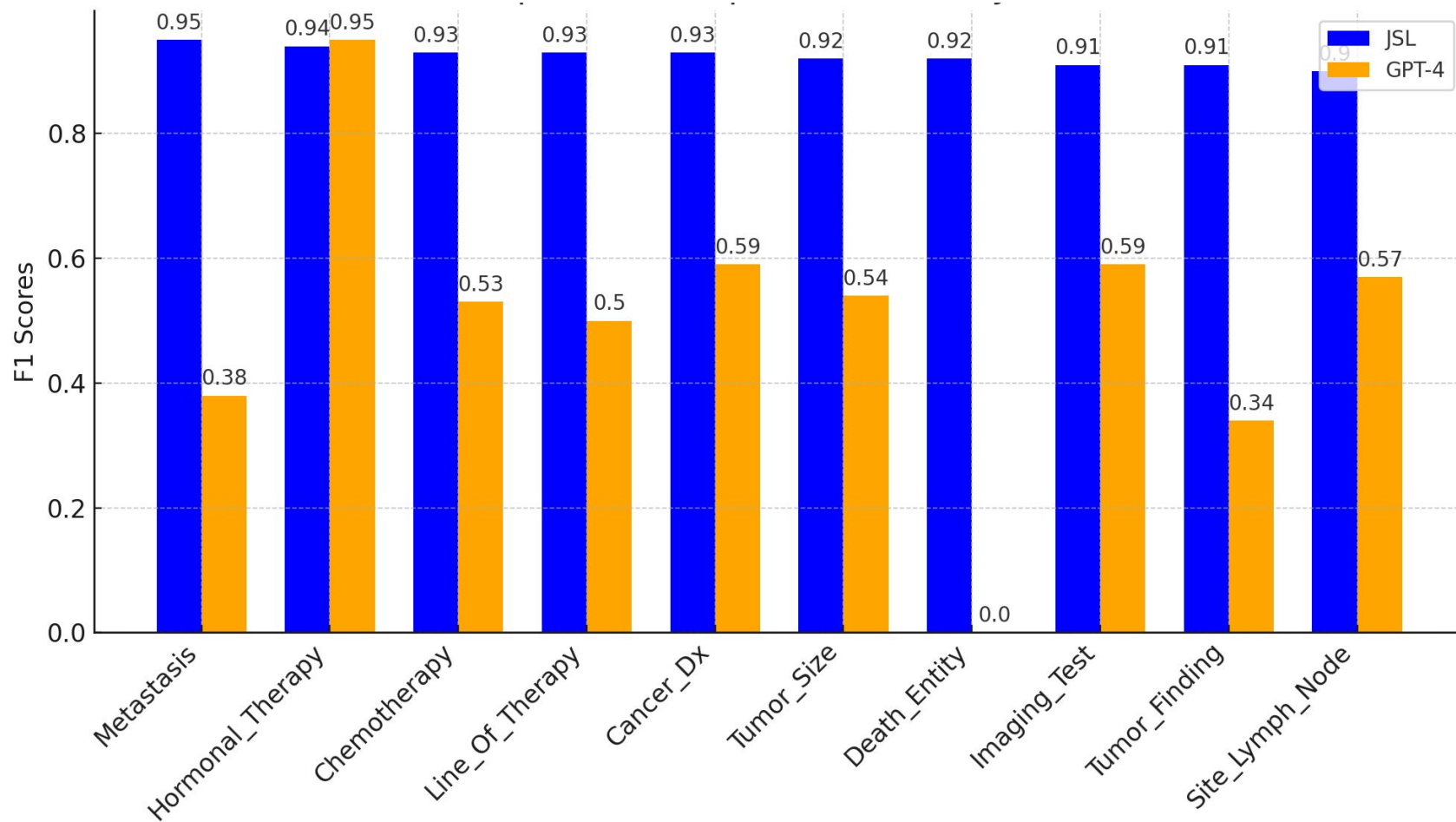


Spark NLP for Healthcare vs ChatGPT (GPT 3.5) on Clinical Entities

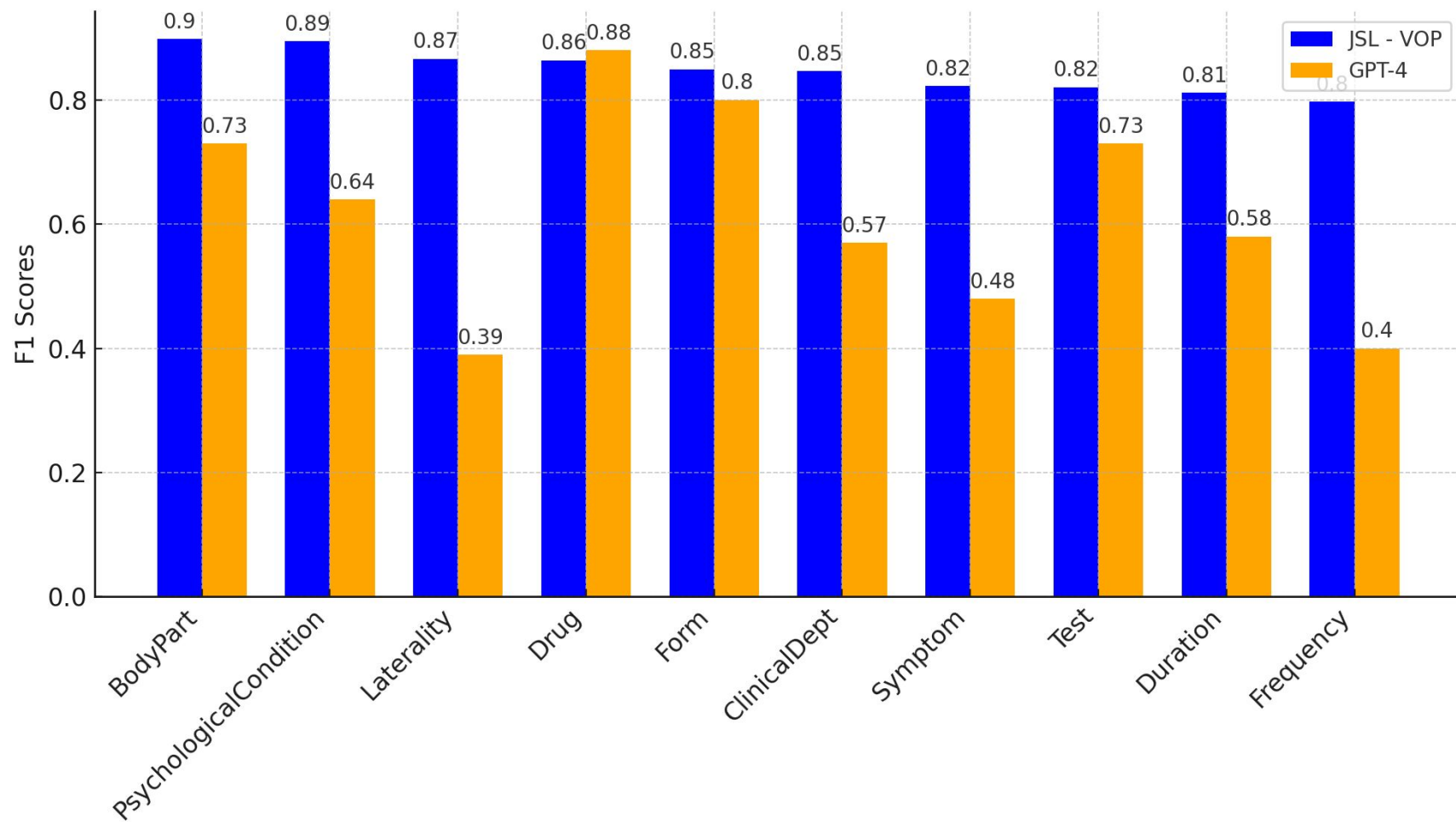


Spark NLP makes half of
the errors that ChatGPT
does!

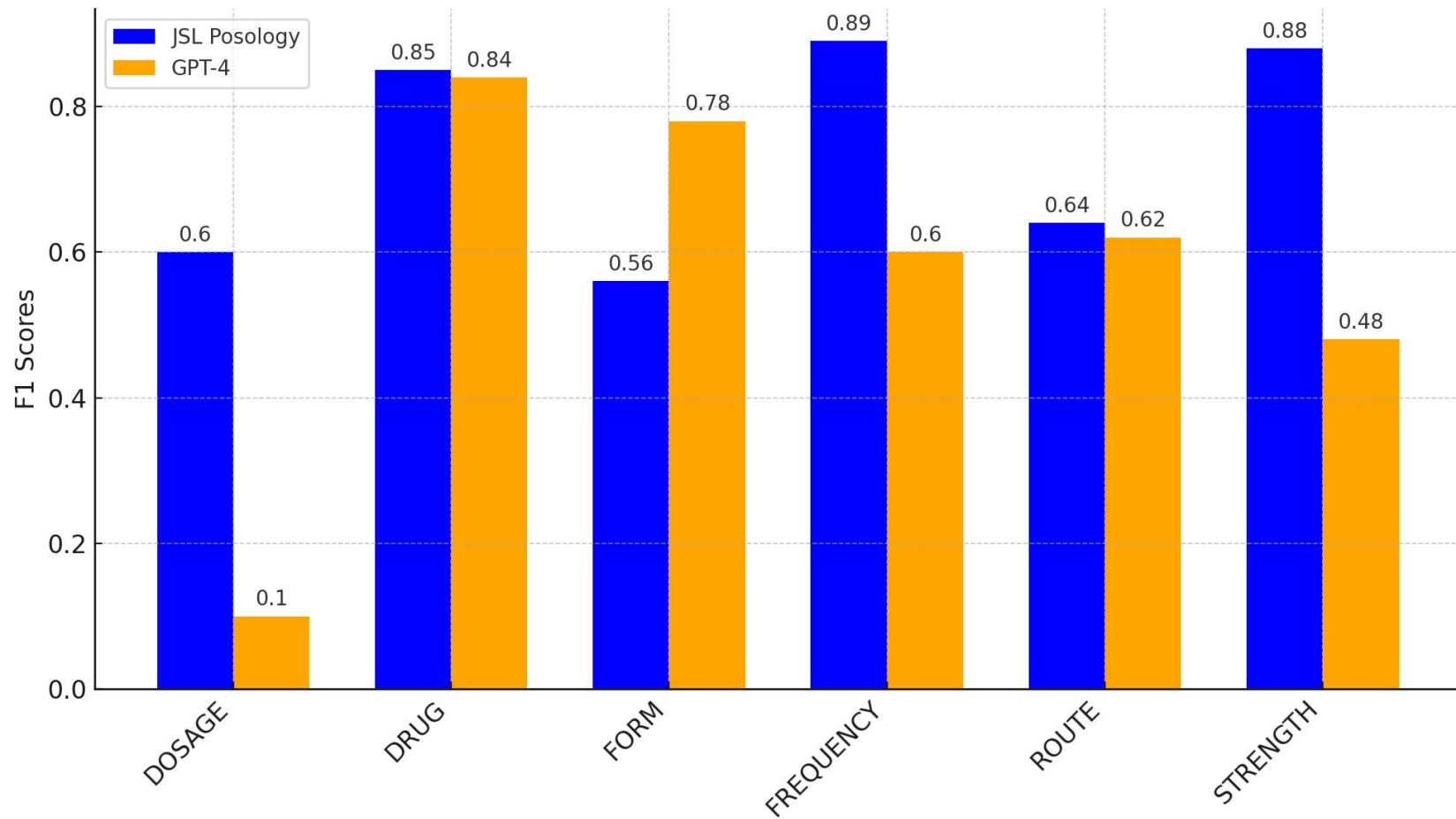
Oncology NER (JSL vs GPT-4)



VOP (Voice of Patient) NER (JSL vs GPT-4)

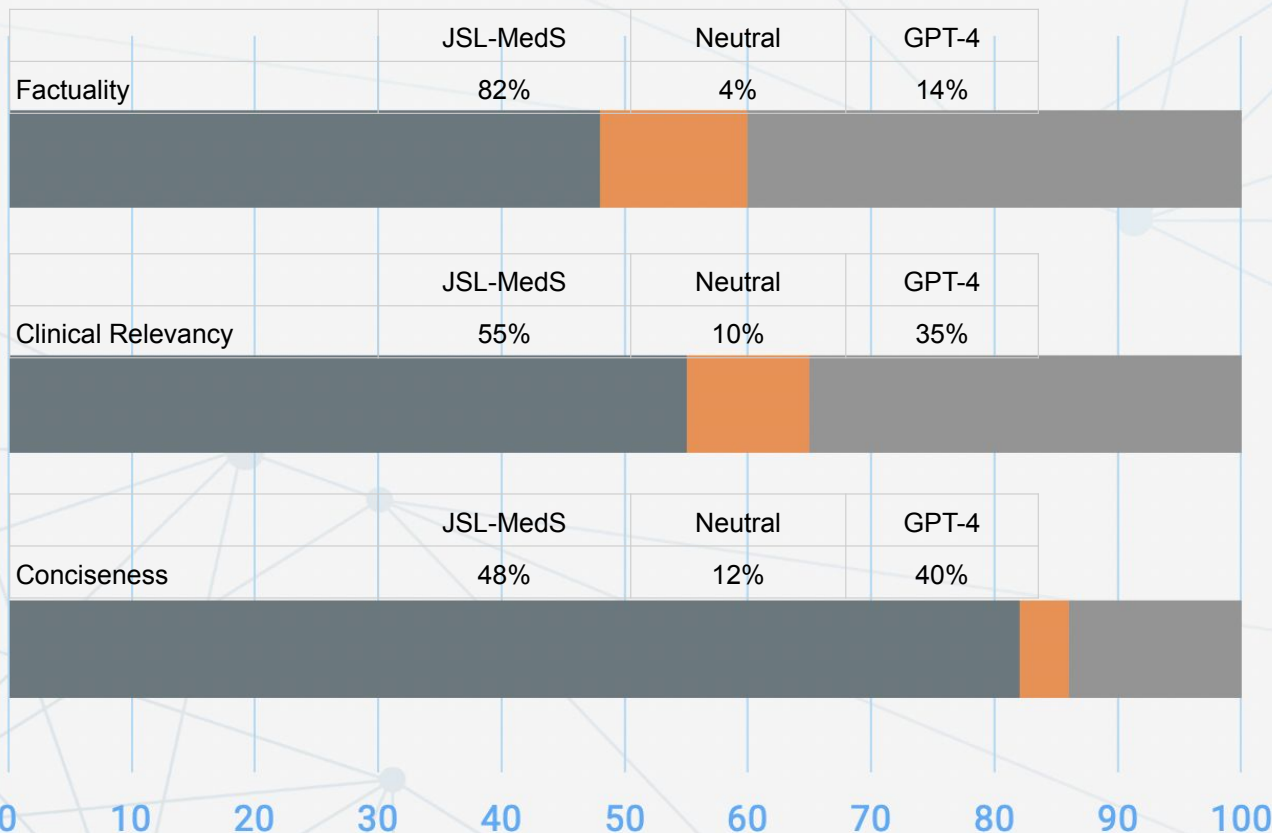


Medication NER (JSL vs GPT-4)



JSL Clinical Summarizer & QA Comparison with GPT-4

Factuality



■ JSL-MedS ■ Neutral ■ GPT-4

John Snow Labs' small language models' (SLMs) preference rate over GPT-4 over 210 questions/tasks/interactions by physicians (randomized blind trials)

Healthcare NLP by John Snow Labs

RxNorm Resolver

ICD10 Resolver

Snomed Resolver

MedDRA Resolver

Assertion Status
Detection

Risk Adj. Module

Named Entity
Recognition

Relationship
Extraction

clinical_finding

Snomed: 44054006

ICD10CM: E11.9

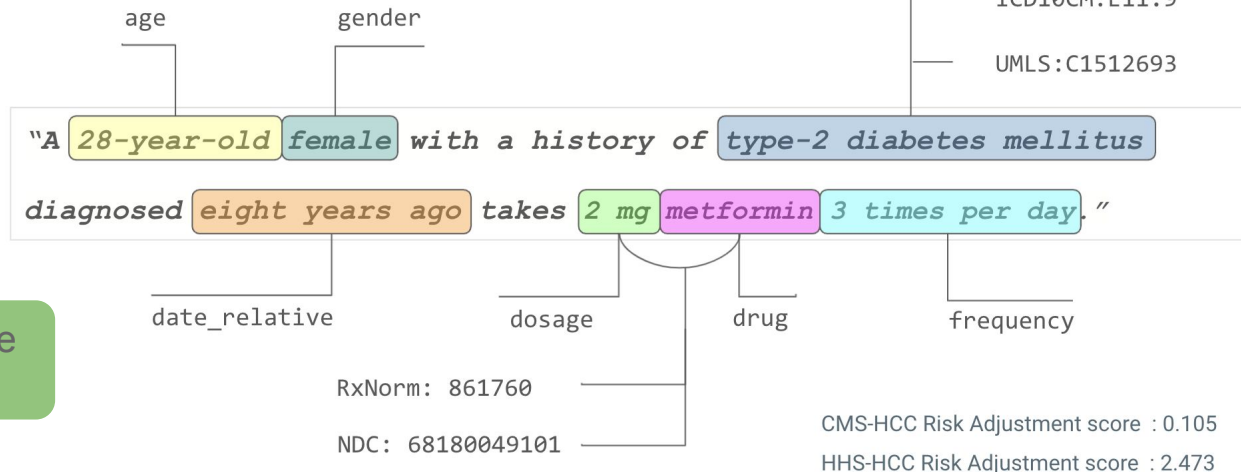
UMLS: C1512693

Sentence Splitter

Tokenizer

Bert Embeddings

Small Language
Models (SLM)





```
# Annotator that transforms a text column from dataframe into an Annotation ready for NLP
documentAssembler = DocumentAssembler()\
    .setInputCol("text")\
    .setOutputCol("document")

sentenceDetector =
SentenceDetectorDLModel.pretrained("sentence_detector_dl_healthcare","en","clinical/models")\
    .setInputCols(["document"])\
    .setOutputCol("sentence")

# Tokenizer splits words in a relevant format for NLP
tokenizer = Tokenizer()\
    .setInputCols(["sentence"])\
    .setOutputCol("token")

# Clinical word embeddings trained on PubMed dataset
word_embeddings = WordEmbeddingsModel.pretrained("embeddings_clinical","en","clinical/models")\
    .setInputCols(["sentence","token"])\
    .setOutputCol("embeddings")

# NER model trained on a medical dataset
clinical_ner = MedicalNerModel.pretrained("ner_clinical_large","en","clinical/models")\
    .setInputCols(["sentence","token","embeddings"])\
    .setOutputCol("ner")\
    .setLabelCasing("upper") #decide if we want to return the tags in upper or lower case

ner_converter = NerConverterInternal()\
    .setInputCols(["sentence","token","ner"])\
    .setOutputCol("ner_chunk")

nlpPipeline = Pipeline(
    stages=[
        documentAssembler,
        sentenceDetector,
        tokenizer,
        word_embeddings,
        clinical_ner,
        ner_converter
    ])

empty_data = spark.createDataFrame([[""]]).toDF("text")

model = nlpPipeline.fit(empty_data)
```


```
text = '''
```

A 28-year-old female with a history of gestational diabetes mellitus diagnosed eight years prior to presentation and subsequent type two diabetes mellitus (T2DM), one prior episode of HTG-induced pancreatitis three years prior to presentation, and associated with an acute hepatitis, presented with a one-week history of polyuria, poor appetite, and vomiting.'''

```
agent_result = process_command_SingleAgent(f"Can you extract Problem, Test and Treatment entities from the following text: {text}")
```

Agent found: SNLP4HC_general_Tool_func

	chunk	begin	end	entity_label	confidence
0	gestational diabetes mellitus	39	67	PROBLEM	0.91976666
1	subsequent type two diabetes mellitus	117	153	PROBLEM	0.75924003
2	T2DM	156	159	PROBLEM	0.9917
3	HTG-induced pancreatitis	184	207	PROBLEM	0.97535
4	an acute hepatitis	264	281	PROBLEM	0.9440667
5	polyuria	321	328	PROBLEM	0.9728
6	poor appetite	331	343	PROBLEM	0.9934
7	vomiting	350	357	PROBLEM	0.9854

 vkocaman Add files via upload

Name

..

data

slides

Domain_Specific_Language_Models.ipynb

Healthcare_NLP_Agents_with_LLMs.ipynb

Integrated Data Science Workflows.ipynb

Medical_Chatbot_API_Example.ipynb

Multimodal_Pipelines_ASR.ipynb

Multimodal_Pipelines_Visual_NLP.ipynb

Oncology_Use_Cases.ipynb

Open_Source_Language_Models.ipynb

Pipelines_LLM.ipynb

README.md

Healthcare_NLP_Agents_with_LLMs.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Table of contents

JohnSnowLabs Healthcare NLP Agents with LLMs (Certification Trainings, July 2024)

Coding an LLM Agent with John Snow Labs Library (Healthcare NLP)

Information Extraction with LLMs

with a simple prompt

with a detailed prompt (instructions and few shot examples)

Information Extraction with Healthcare NLP by John Snow Labs

Building an NER agent with Healthcare NLP

Align with user prompts precisely

Building a multi agent with Healthcare NLP

adding Posology (medication) NER

Align with user prompts precisely

+ Section

John Snow LABS

Open in Colab

JohnSnowLabs Healthcare NLP Agents with LLMs (Certification Trainings, July 2024)

Coding an LLM Agent with John Snow Labs Library (Healthcare NLP)

```
[1] import json
import os

from google.colab import files

if 'spark_jsl.json' not in os.listdir():
    license_keys = files.upload()
    os.rename(list(license_keys.keys())[0], 'spark_jsl.json')

with open('spark_jsl.json') as f:
    license_keys = json.load(f)

# Defining license key-value pairs as local variables
locals().update(license_keys)
os.environ.update(license_keys)

Choose files spark_nlp_f...72_540.json
• spark_nlp_for_healthcare_spark_ocr_9272_540.json(application/json) - 1785 bytes, last modified: 14/07/2024 - 100% done
Saving spark_nlp_for_healthcare_spark_ocr_9272_540.json to spark_nlp_for_healthcare_spark_ocr_9272_540.json

[2] license_keys['JSL_VERSION']

'5.4.0'

[3] license_keys['PUBLIC_VERSION']

'5.4.0'
```

Thank you !

Veysel Kocaman, PhD

Head of Data Science
John Snow Labs



Understanding OMOP CDM

(Observational Medical Outcomes Partnership - Common Data Model)

Enhancing Healthcare through Data

Foundation: Part of the **Observational Health Data Sciences and Informatics (OHDSI)** initiative.

Objective: Utilize open-source data solutions to improve human health via large-scale analysis.

Purpose: Standardize the structure and content of observational healthcare data.

Features:

- Enables efficient, reliable evidence production through analysis.
- Incorporates a common vocabulary and standards for clinical data management.

Focus: Centered on patient outcomes and includes recorded healthcare events.

Community: An open community data standard, fostering collaboration and innovation in healthcare data utilization.



OOP-CDM is a data model that allows clinical information to be presented in a standardized and reusable way for research

