



IC01: Large Language Models to support RBQM and Clinical Trial Development

PHUSE / FDA Innovation Challenge

27th February 2024

Laura Trotta, VP Research

Sebastiaan Höppner, Senior Scientist in Statistics



Can you distinguish the human from the computer?

Medical question: can a CT scan detect diverticulitis or diverticulosis?

1



Diverticula are small outpouchings in the inner lining of the large intestine that are common as one gets older. Diverticulosis is the medical term for this condition if the outpouching does not result in symptoms. If the diverticula becomes inflamed or infected, it is called diverticulitis. Constipation, diarrhoea, fever, and stomach discomfort can all be symptoms. Differentiating between diverticulitis and diverticulosis can be done using a CT scan.

CT scans are often used to diagnose diverticulitis and can also detect diverticulosis. CT scans provide detailed images of the inside of the body, and can show the presence of diverticula (small pouches that can form in the lining of the intestine). In diverticulitis, these pouches can become inflamed and infected, and a CT scan can show the extent of the inflammation and infection. CT scans can also be used to rule out other possible causes of the symptoms, such as appendicitis or a bowel obstruction. If you are experiencing symptoms such as abdominal pain, fever, or changes in bowel habits, your doctor may order a CT scan to diagnose the cause.

2



Can you distinguish the human from the computer?

Response 2 was generated by a Large Language Model from Google

1



Clinician answer

Diverticula are small outpouchings in the inner lining of the large intestine that are common as one gets older. Diverticulosis is the medical term for this condition if the outpouching does not result in symptoms. If the diverticula becomes inflamed or infected, it is called diverticulitis. Constipation, diarrhoea, fever, and stomach discomfort can all be symptoms. Differentiating between diverticulitis and diverticulosis can be done using a CT scan.

Med-PaLM 2 answer

CT scans are often used to diagnose diverticulitis and can also detect diverticulosis. CT scans provide detailed images of the inside of the body, and can show the presence of diverticula (small pouches that can form in the lining of the intestine). In diverticulitis, these pouches can become inflamed and infected, and a CT scan can show the extent of the inflammation and infection. CT scans can also be used to rule out other possible causes of the symptoms, such as appendicitis or a bowel obstruction. If you are experiencing symptoms such as abdominal pain, fever, or

2



Challenge objective ?

Can we rely on Large Language Models to better support RBQM and clinical development?



Challenge objective ?

Can we rely on Large Language Models to better support RBQM and clinical development?



Shared knowledge of LLMs

“We share our experience with LLMs acquired in developing ML solutions for RBQM and clinical development and present the results of 3 experiments.

Our objective is to encourage researchers to share openly their results and to build a shared knowledge of LLMs for our industry.”



Today's Agenda

- Our experience with ML for RBQM and clinical development
- LLMs: What & Why?
- 3 Experiments on LLMs
- Conclusions



Our experience with
ML for RBQM and
clinical development

Sharing our experience with 1500+ trials

Track record



10+ years into RBQM



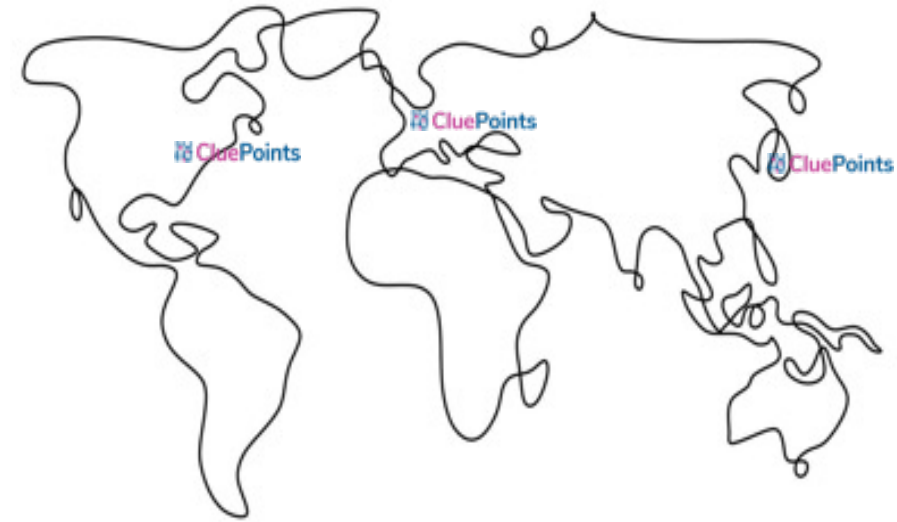
100+ organizations



1200+ studies

therapeutic area agnostic

All phases of development



Unique aggregated data driving AI/ML algorithms



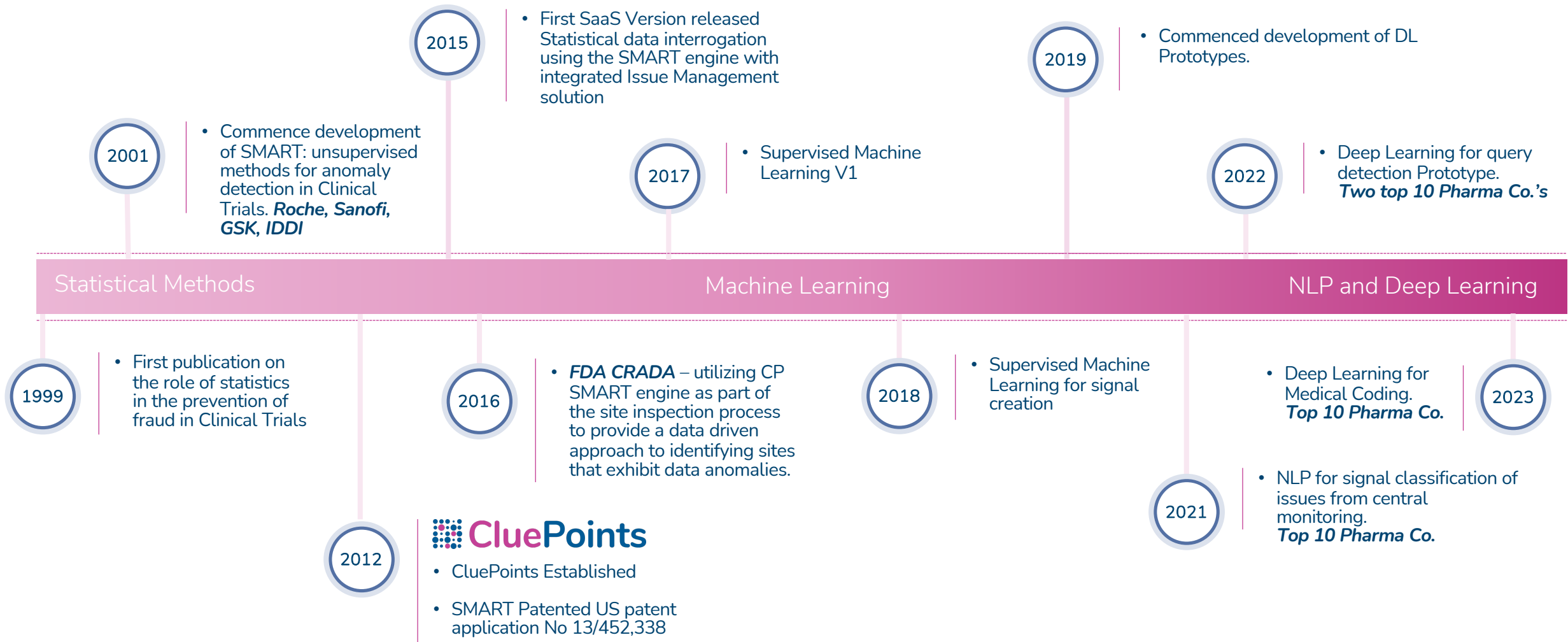
350 risk assessments

120,000+ risk signals

30,000+ resolution actions

© Springer-Verlag Italia 2004

From central statistical monitoring ...to deep learning



Machine Learning to support RBQM – Today

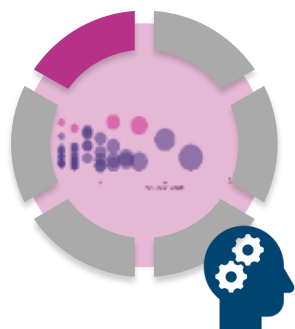
RBQM

- I. Risk detection powered by advanced statistical & ML algorithms (DQA, KRIs & QTLs)
- II. Intelligent signal & actions management system
 - I. ML-driven grouping of data anomalies for automated risk signal creation
 - II. ML-driven detection of risk signals lacking active follow-up & documentation

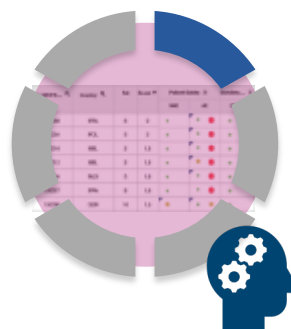
RBQM PLATFORM



**Risk Assessment
Categorization Tool**
Identify & Document
Study Risk Factors



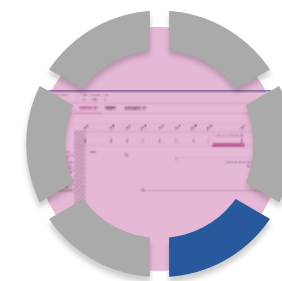
**Data Quality
Assessment**
The Quality Stamp



KRIs & QTLs
Monitor the Study
Conduct using
Operational and
Clinical Data



**Signal & Actions
Management**
Monitor and solve
issues



Patient Profiles
Visualize and Review
Individual Patients



**BEYOND
Data Visualizations**
Design and Explore
Custom Dashboards

Machine Learning to support RBQM – Tomorrow

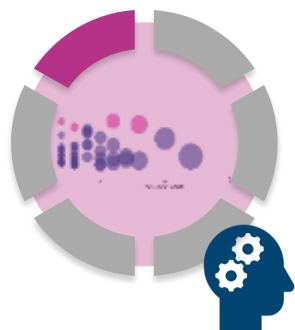
RBQM

- I. Risk detection powered by advanced statistical & ML algorithms (DQA, KRIs & QTLs)
- II. Intelligent signal & actions management system
- III. ML-driven risk assesment (example: enhanced review of protocol)
- IV. Intelligent patient profiles/medical review (example: deep learning for safety detection)
- V. ML-driven data visualizations (example: KRI/QTL dashboards across full portfolio of studies)

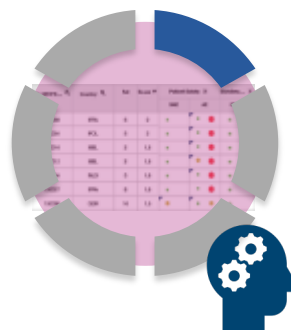
RBQM PLATFORM



**Risk Assessment
Categorization Tool**
Identify & Document
Study Risk Factors



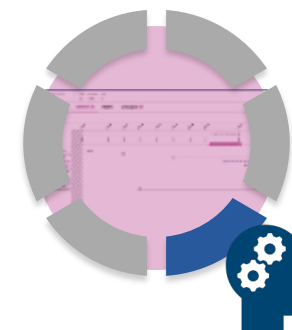
**Data Quality
Assessment**
The Quality Stamp



KRIs & QTLs
Monitor the Study
Conduct using
Operational and
Clinical Data



**Signal & Actions
Management**
Monitor and solve
issues



Patient Profiles
Visualize and Review
Individual Patients



**BEYOND
Data Visualizations**
Design and Explore
Custom Dashboards

Beyond RBQM - Machine Learning for clinical development

Clinical Developments

- I. Medical coding: ML-driven suggestion of MedDRA/WHODrug coded terms
- II. Query detection: ML-driven detection of data points that should be queried



Dictionary: GLOBALB3Mar23					Suggestions	Cluepoints
Code Options	Relevance Score	Drug Name	Preferred Name	Dictionary Release		
	78	CHLORPHENAMIN E	CHLORPHENAMINE	GLOBALB3Mar23		
	76	CHLORPHENAMIN UM	CHLORPHENAMINE	GLOBALB3Mar23		
	74	CHLORMETHINE	CHLORMETHINE	GLOBALB3Mar23		
	74	CHLORPHEDRINE	CHLORPHENAMINE MALEATE;PSEUDOEPHEDRINE HYDROCHLORIDE	GLOBALB3Mar23		
	73	CHLORIMIPRAMIN E	CLOMIPRAMINE	GLOBALB3Mar23		

 Data Anomaly Detection

Act as SME

← Back

< Query 1 / 5373 >

QueryID 1

Study

Patient (subjectkey)

Status Accepted ?

↔ Change status

👤 Assign

💬 Comment

🕒 History

Query information

Form: AE_RD
Category: AE_10
Creation Date: 01/03/2023 12:00 AM
Query variable: AESTDTTM1

Query filters

formrepeatkey: 1
locationoid
studyeventrepeatkey: 1

itemgrouprepeatkey: 2
studyeventname: CommonCRF

Query text

The AE/SAE outcome is Not Recovered/Not Resolved, Recovering/Resolving or Missing and current date is more than 30 days from start date

Query text

AE_RD (AE_RD)

CONC_STUDY_RD (CONC_STUDY_RD) ✕

+

AESTDTTM1	AESSEQ	ICDS_ID	AEEOSI	AEREL	ADACTCD1	AETERM	ADTOXCD3	AETOXCD1	AEOUTCD	AEPV
08/22/2020 UNK:UNK	10002	61507882	No	No	Dose not changed	Tiredness		1	Not recovered/Not resolved	Singl

Reference datapoint

Relevant datapoints



LLMs: what and why?

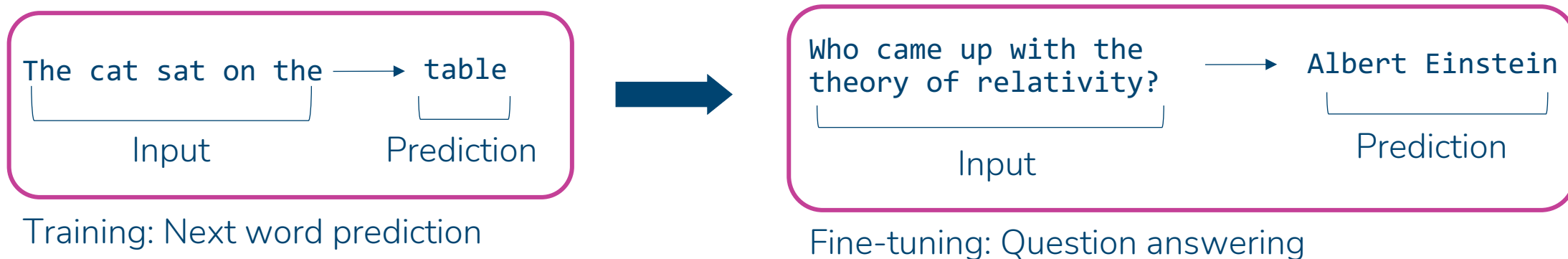
Large Language Models: What?

Large Language Models (LLMs) are deep learning models particularly powerful at performing Natural Language Processing tasks such as language translation, text classification, etc.



Large Language Models: Fine-Tuning

LLMs are **foundation models** trained on broad data that can be adapted to a wide range of downstream tasks. **Fine-tuning** an LLM is the process of tuning the parameters of a pre-trained model **to acquire specific knowledge about the task and the domain.**



Large Language Models: generative versus non-generative

Generative LLMs:

LLMs trained to **generate data**, such as GPT4

(GPT = Generative Pre-trained Transformer)



Non-generative LLMs:

LLMs that do not generate new content but rather **focus on tasks** such as **classification, recognition, and prediction**

The cat jumped on the table.

✓ The cat broke the vase.

✗ The car engine broke.

Is this next sentence likely (Yes/No) ?



3 Experiments on LLM

- Experiment 1: improve documentation of risk signals & promote active risk follow-up
- Experiment 2: Enhance risk detection with information from study protocol (GPT4)
- Experiment 3: Medical coding of AEs and medications

Experiment 1 – Improve documentation of risk signals & promote active risk follow-up



Aim – Enabling more active follow-up and documentation of risk signals



Risk Signal – description of anomalies (i.e., detected risks), grouped and presented to study team



111 organizations > **1417** trials analyzed > **178,270** risk signals generated

Experiment 1 – Improve documentation of risk signals & promote active risk follow-up



Aim – Enabling more active follow-up and documentation of risk signals

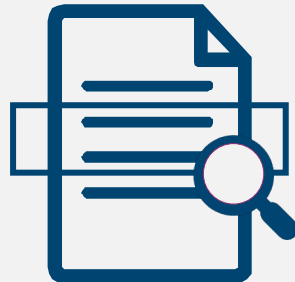


Method – BERT language model, fine-tuned on a database of reviewed risk signals

Risk signal



Screening



Output

Detect

1. Poor documentation
2. Lack of follow-up activities
3. Incorrect classification of risk signals into non-issues

Experiment 1 – Improve documentation of risk signals & promote active risk follow-up



Aim – Enabling more active follow-up and documentation of risk signals



Method – BERT language model, fine-tuned on a database of reviewed risk signals



Results – Promote active follow-up for more reliable risk documentation

Performance: 70% overall accuracy in classification

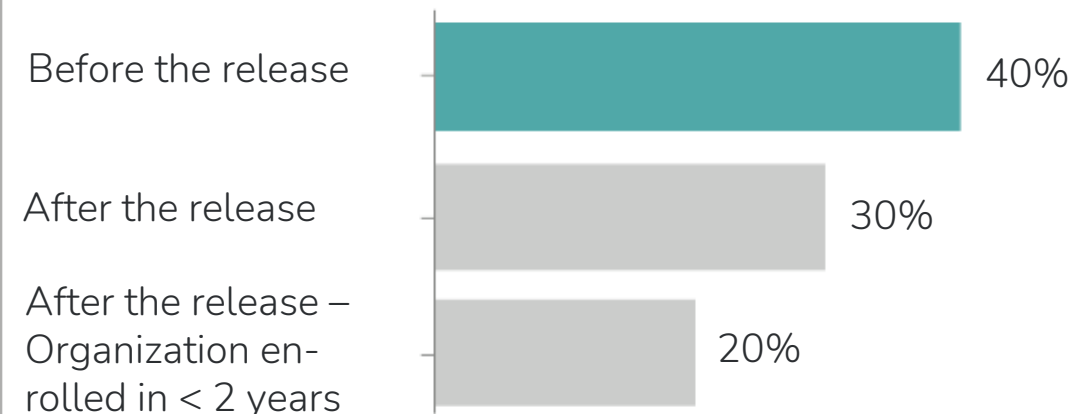
Deployment details:

- Released into production
- Active usage by 35 organizations

Analysis results:

- Studies covered: 279
- Risk signals processed: 19,404
- Decrease in signals without documentation: 25%

Reducing the 'Unclear' Signal Rate



The median rate of signals with unclear documentation by study.

Experiment 2 – Enhance risk detection with information from study protocol



Aim – Enhancing risk detection by accounting for study-specific parameters



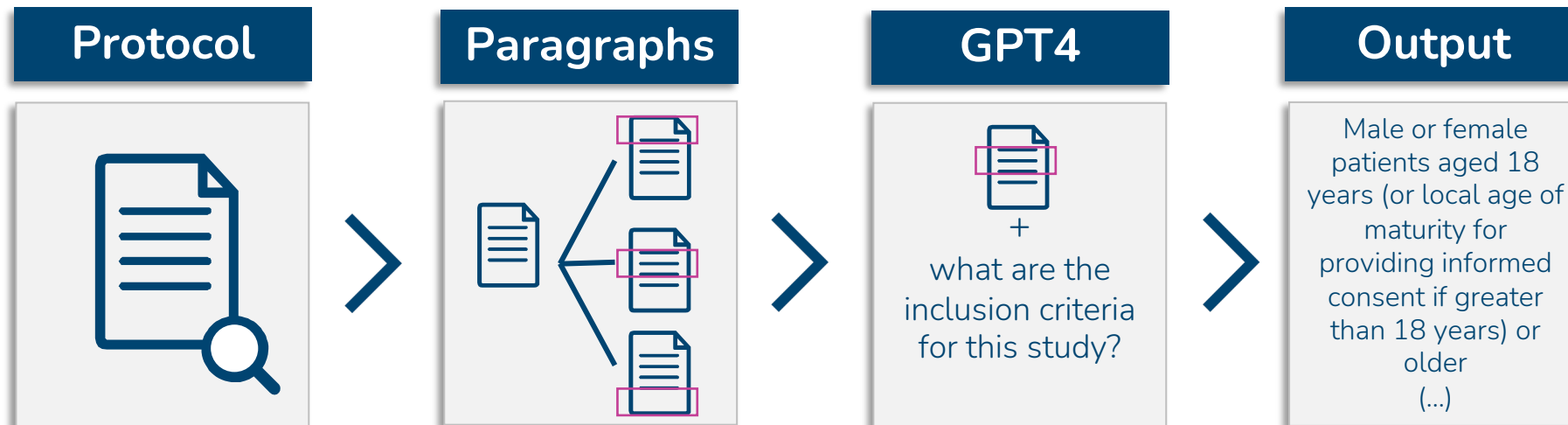
Experiment 2 – Enhance risk detection with information from study protocol (GPT4)



Aim – Enhancing risk detection by better accounting for study parameters



Method – Using GPT4 to extract study specific parameters from protocol



Experiment 2 – Enhance risk detection with information from study protocol (GPT4)



Aim – Enhancing risk detection by better accounting for study parameters



Method – Using GPT4 to extract study information from protocol



Results – Promising technology, need to address the limited entry size

Opportunity

GPT4 is very often right in extracting the relevant information from study protocol.

Limitations

GPT4 has a limited entry size and can only interrogate a few pages at once. This limitation could be overcome with the use of semantic search.

Experiment 3 – Medical Coding of Adverse Events and Medications



Aim – Suggesting coded terms for AE/drug in MedDRA/WHODrug

Medical events	
Input term	Dictionary entry
Aches on arms	Pain in arm
Inflammation bottom of lung	Pneumonitis
Nose bleed	Epistaxis

Drugs	
Input term	Dictionary entry
Oral rehydration salt	Sodium bicarbonate
Hypertonic saline 20%	Sodium chloride
Dafalgan	Paracetamol

Experiment 3 – Medical Coding of Adverse Events and Medications



Aim – Suggesting coded terms for AE/drug in MedDRA/WHODrug



Method – GPT4 versus Bert-like model fined-tuned on a trusted database

GPT4

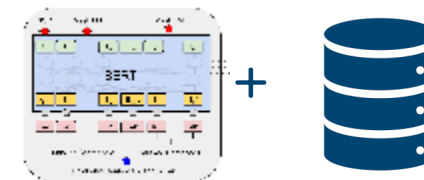


GPT - 4

“Medical coding is the task of coding a medical event to a standard dictionary. Here is an example Right thigh pain: Pain in thigh. Could you code the following terms in MedDRA?”

versus

Fined-tuned LLM



Bert-like model
fined-tuned on a database of
historical verbatim entries
mapped to their dictionary
entry

Experiment 3 – Medical Coding of Adverse Events and Medications



Aim – Enhancing risk detection by better accounting for study parameters



Method – GPT4 versus Bert-like model fined-tuned on a trusted database



Results – Small fined-tuned LLMs produce more accurate predictions

GPT4

Accuracy = 40 %

The model produces **hallucinations**:
coded terms not in MedDRA

Bert-liked + Fine Tuning

Accuracy > 90 %

The model only outputs entries from
MedDRA dictionary

Ex: Inflammation bottom of lungs - pneumonitis
Nose bleed - epistaxis

Conclusions



Key outcomes – 1/3

LLMs are powerful at processing and extracting information from text data

- They can be used to extract knowledge from clinical documents such as a clinical protocols
- LLMs have developed medical knowledge

nature

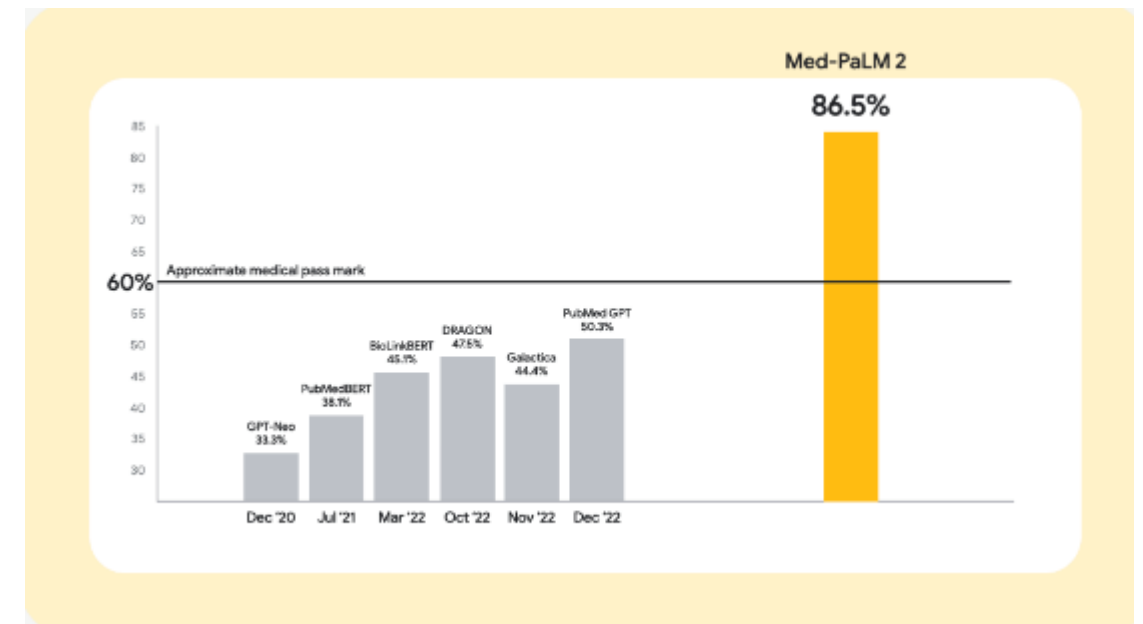
Explore content ▾ About the journal ▾ Publish with us ▾

nature > articles > article

Article | [Open access](#) | Published: 12 July 2023

Large language models encode clinical knowledge

[Karan Singhal](#) , [Shekoofeh Azizi](#) , [Tao Tu](#), [S. Sara Mahdavi](#), [Jason Wei](#), [Hyung Won Chung](#), [Nathan Scales](#), [Ajay Tanwani](#), [Heather Cole-Lewis](#), [Stephen Pfohl](#), [Perry Payne](#), [Martin Seneviratne](#), [Paul Gamble](#), [Chris Kelly](#), [Abubakr Babiker](#), [Nathanael Schärli](#), [Aakanksha Chowdhery](#), [Philip Mansfield](#), [Dina Demner-Fushman](#), [Blaise Agüera y Arcas](#), [Dale Webster](#), [Greg S. Corrado](#), [Yossi Matias](#), [Katherine Chou](#), ... [Vivek Natarajan](#)  + Show authors



Key outcomes – 2/3

Smaller **finetuned** LLMs can produce **accurate clinical predictions**

- General purpose LLMs fail to produce accurate clinical predictions
- Smaller LLMs finetuned on trusted clinical databases produce accurate predictions

npj | digital medicine

Explore content ▾ About the journal ▾ Publish with us ▾

nature > npj.digital.medicine > review.articles > article

Review Article | [Open access](#) | [Published: 29 July 2023](#)

The shaky foundations of large language models and foundation models for electronic health records

[Michael Wornow](#) , [Yizhe Xu](#), [Rahul Thapa](#), [Birju Patel](#), [Ethan Steinberg](#), [Scott Fleming](#), [Michael A. Pfeffer](#), [Jason Fries](#) & [Nigam H. Shah](#)

npj Digital Medicine 6, Article number: 135 (2023) | [Cite this article](#)

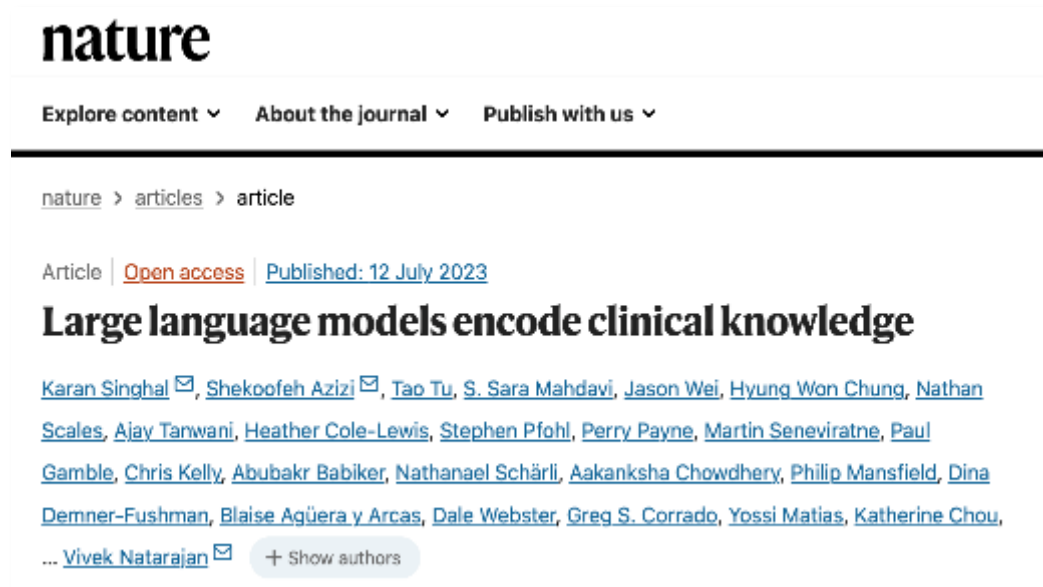
22k Accesses | 12 Citations | 106 Altmetric | [Metrics](#)

While **general-purpose LLMs** (e.g., ChatGPT, Bloom, GPT-4, etc.) trained on text scraped from the Internet can also be useful for clinical tasks, they **tend to underperform Clinical language models on domain-specific tasks** (...)

Key outcomes – 3/3

LLMs used in a **non-generative** mode allow to **overcome the issue of hallucinations**

- Generative LLMs such as GPT4 produce hallucinations
- The potential harm of hallucinations in a clinical context is discussed in the literature



The screenshot shows the top portion of a Nature journal article. The Nature logo is at the top left. Below it are navigation links: 'Explore content', 'About the journal', and 'Publish with us'. The breadcrumb trail reads 'nature > articles > article'. The article title is 'Large language models encode clinical knowledge'. Below the title is a list of authors: Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, and Vivek Natarajan. There is a '+ Show authors' button at the bottom of the author list.

nature

Explore content ▾ About the journal ▾ Publish with us ▾

nature > articles > article

Article | [Open access](#) | Published: 12 July 2023

Large language models encode clinical knowledge

[Karan Singhal](#), [Shekoofeh Azizi](#), [Tao Tu](#), [S. Sara Mahdavi](#), [Jason Wei](#), [Hyung Won Chung](#), [Nathan Scales](#), [Ajay Tanwani](#), [Heather Cole-Lewis](#), [Stephen Pfohl](#), [Perry Payne](#), [Martin Seneviratne](#), [Paul Gamble](#), [Chris Kelly](#), [Abubakr Babiker](#), [Nathanael Schärli](#), [Aakanksha Chowdhery](#), [Philip Mansfield](#), [Dina Demner-Fushman](#), [Blaise Agüera y Arcas](#), [Dale Webster](#), [Greg S. Corrado](#), [Yossi Matias](#), [Katherine Chou](#), ... [Vivek Natarajan](#) + Show authors

The **safety-critical nature** of the domain necessitates thoughtful development of **evaluation frameworks**, enabling researchers to meaningfully measure progress and capture and **mitigate potential harms**. This is especially important for LLMs, since these models may produce text generations (hereafter referred to as 'generations') that are misaligned with clinical and societal values. They may, for instance, hallucinate convincing medical misinformation or incorporate biases that could exacerbate health disparities.

Conclusions

- LLMs show impressive text processing capabilities that can be used to support RBQM and clinical trial developments
- Smaller LLMs fine-tuned on clinical databases generate more accurate predictions than general purpose LLMs
- Generative LLMs produce hallucinations, when possible this can be overcome by using LLMs in a non-generative mode
- A common framework for model validation is needed!



Thank You. Questions?

laura.trotta@cluepoints.com