



DCRI Pragmatic Trial Collaboratory Grand Rounds

Opportunities and Challenges in the use of Large Language Models for Post-Marketing Surveillance of Medical Products

Michael E. Matheny, MD, MS, MPH

Director, Center for Improving the Public's Health Through Informatics
Professor, Departments of Biomedical Informatics, Medicine, and Biostatistics
Vanderbilt University Medical Center

Associate Director for Data Analytics, VINCI
Associate Director, Advanced Fellowship in Medical Informatics
Associate Director, VA HSR COIN VETWISE-LHS
Tennessee Valley Healthcare System VA

Twitter: [@MichaelEMatheny](https://twitter.com/MichaelEMatheny)

Email: michael.Matheny@va.gov, michael.Matheny@Vanderbilt.edu, michael.Matheny@vumc.org

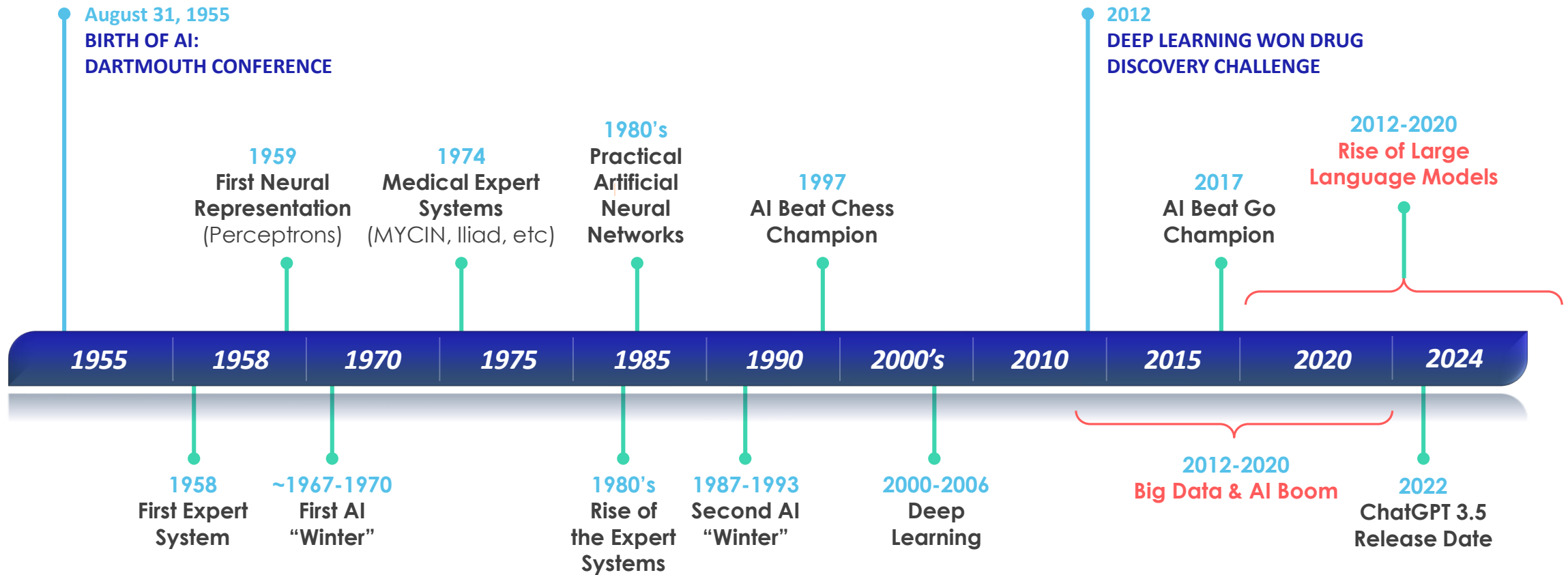


Disclosures

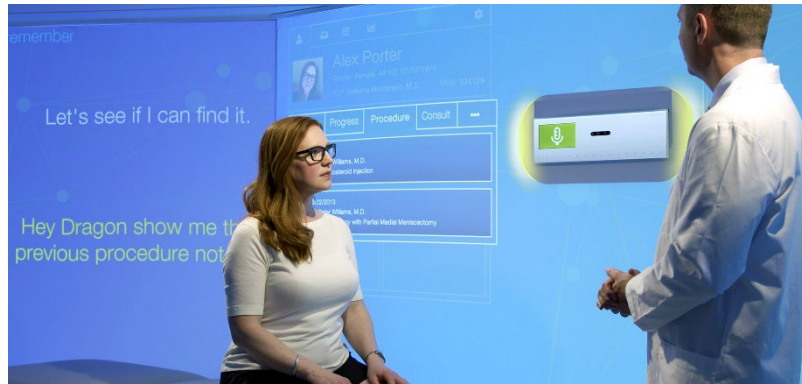
- Related to AI, I am part of a patent in submission for an algorithm drift, fairness, and self-correction framework. Otherwise, I have no conflicts of interest in the presentation of any materials, software, or algorithms presented in this presentation.
- All funding I have received in the last 3 years are research grants and contracts from FDA Sentinel, VA ORD & HSR&D, NIH NHLBI & NIDDK, NIH-VA-DoD Joint funding, and a medical device public-private partnership (NESTcc [FDA U01])
- The views expressed in this presentation represent those of the presenter and do not necessarily represent the official views of the U.S. FDA, the U.S. Department of Veteran Affairs, or Vanderbilt

Artificial Intelligence in Healthcare

Artificial Intelligence Timeline



Promise of Artificial Intelligence & Machine Learning

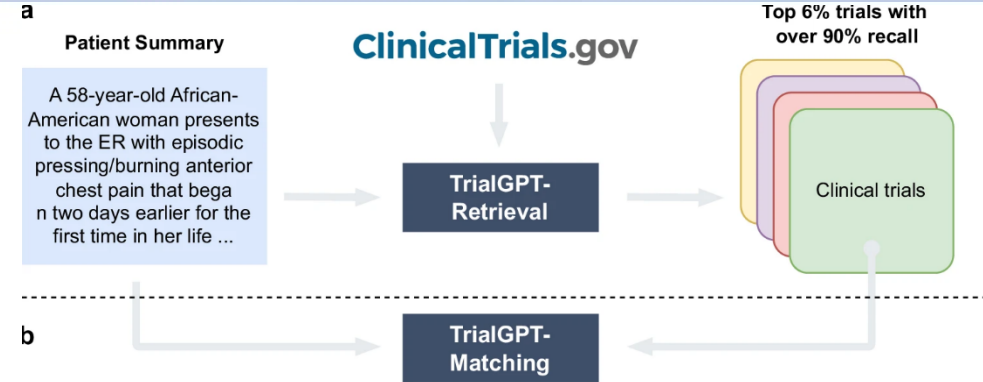


Ambient AI

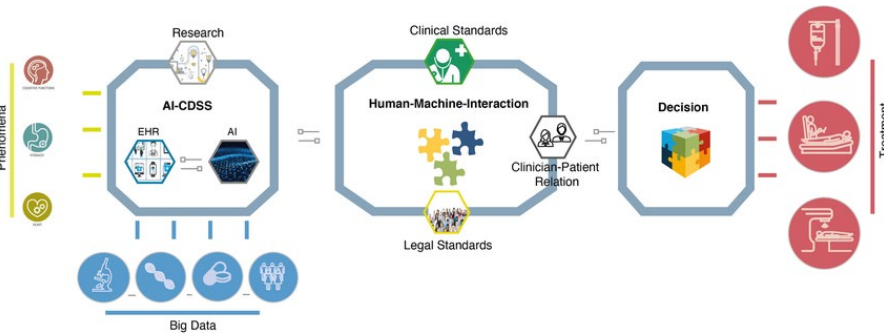


<https://capx.co/artificial-intelligence-could-be-the-radiologist-of-the-future/>

Imaging Processing

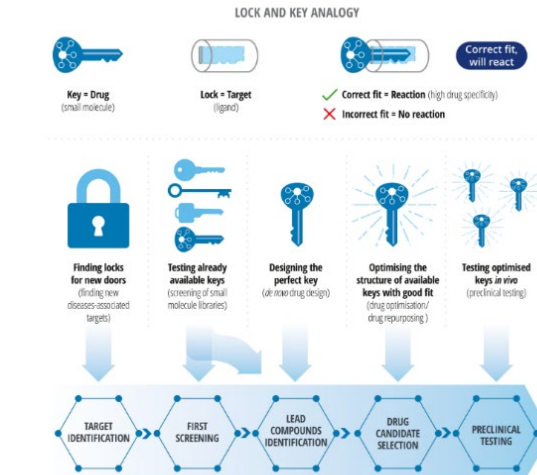


Supporting Clinical Trials



Clinical Decision Support

Lock and key analogy showing the five main challenges for AI in drug discovery



Source: Deloitte analysis.

Drug Discovery



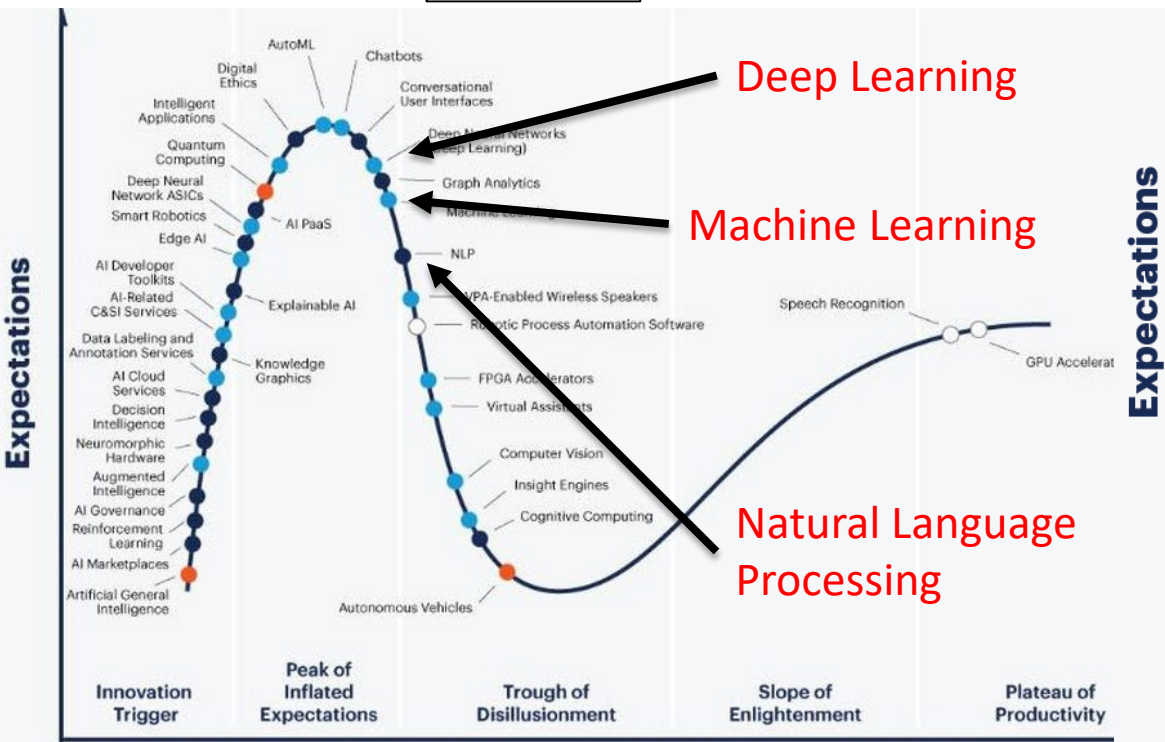
Autonomous AI

In 2018, first Software as a Medical Device AI approved by FDA to not require physician interpretation.

Gartner Hype Cycle for Artificial Intelligence

2019

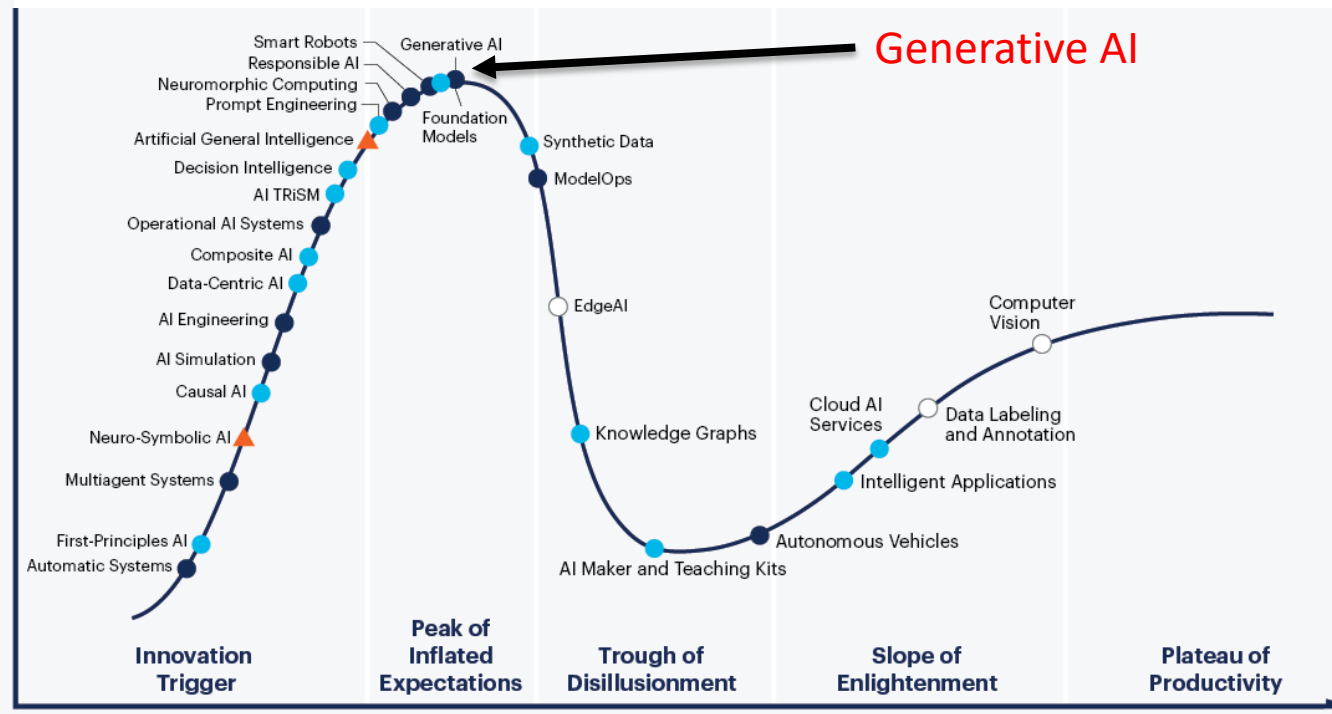
2023



Deep Learning

Machine Learning

Natural Language Processing



Generative AI

Plateau will be reached:
 ○ less than 2 years ● 2 to 5 years ● 5 to 10 years ● more than 10 years ● obsolete before plateau As of

Plateau will be reached:
 ○ less than 2 years ● 2 to 5 years ● 5 to 10 years ▲ more than 10 years ⊗ obsolete before plateau As of July 2023

Numerous Potential Applications of LLMs

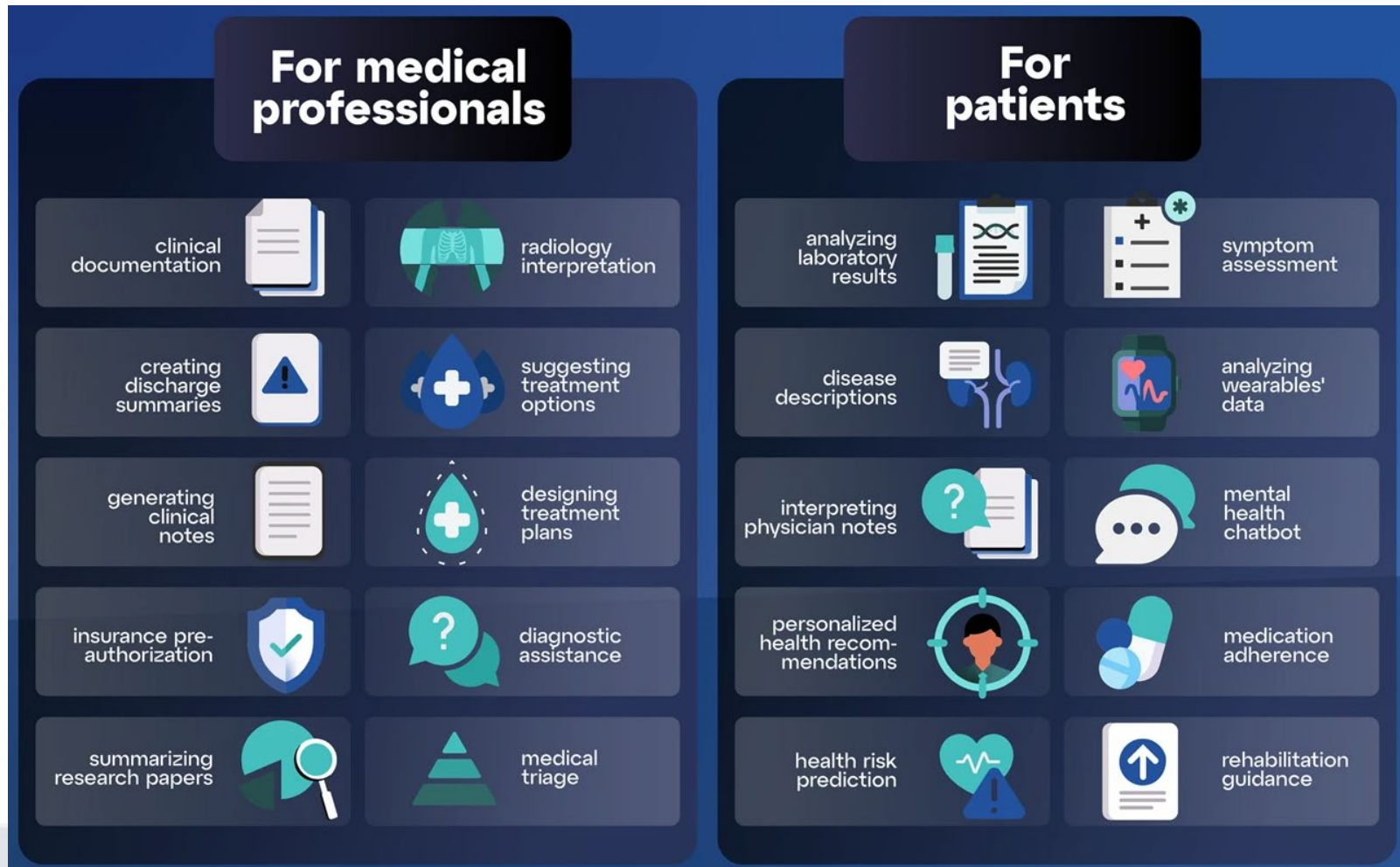
npj | digital medicine

PERSPECTIVE OPEN

Check for updates


The imperative for regulatory oversight of large language models (or generative AI) in healthcare

Bertalan Meskó^{1,2} and Eric J. Topol³

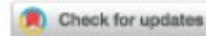


Potential for LLMs to Support Clinical Trials



Transl Clin Pharmacol. 2023 Sep;31(3):131-138
https://doi.org/10.12793/tcp.2023.31.e16
pISSN 2289-0882·eISSN 2383-5427

Translational and
Clinical Pharmacology **TCP** 

Review Article

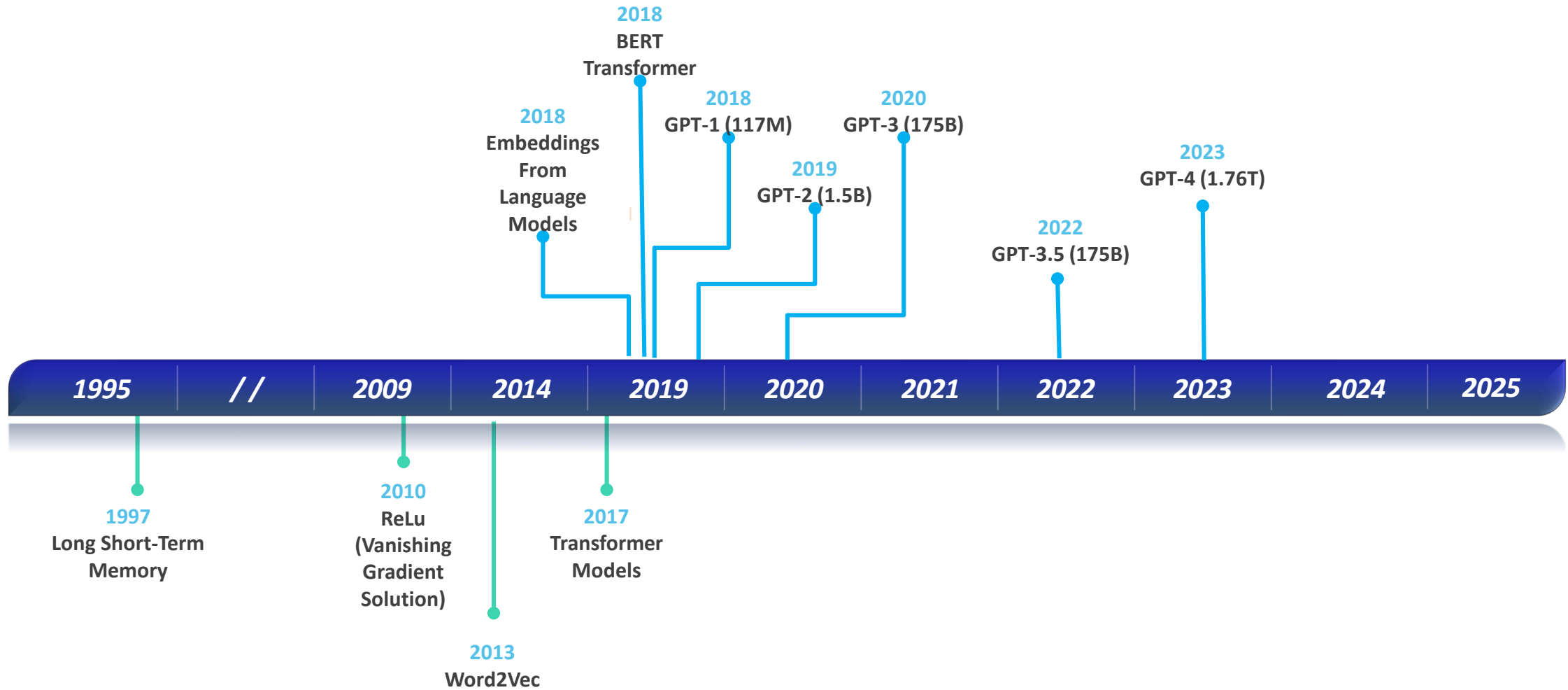


Transforming clinical trials: the emerging roles of large language models

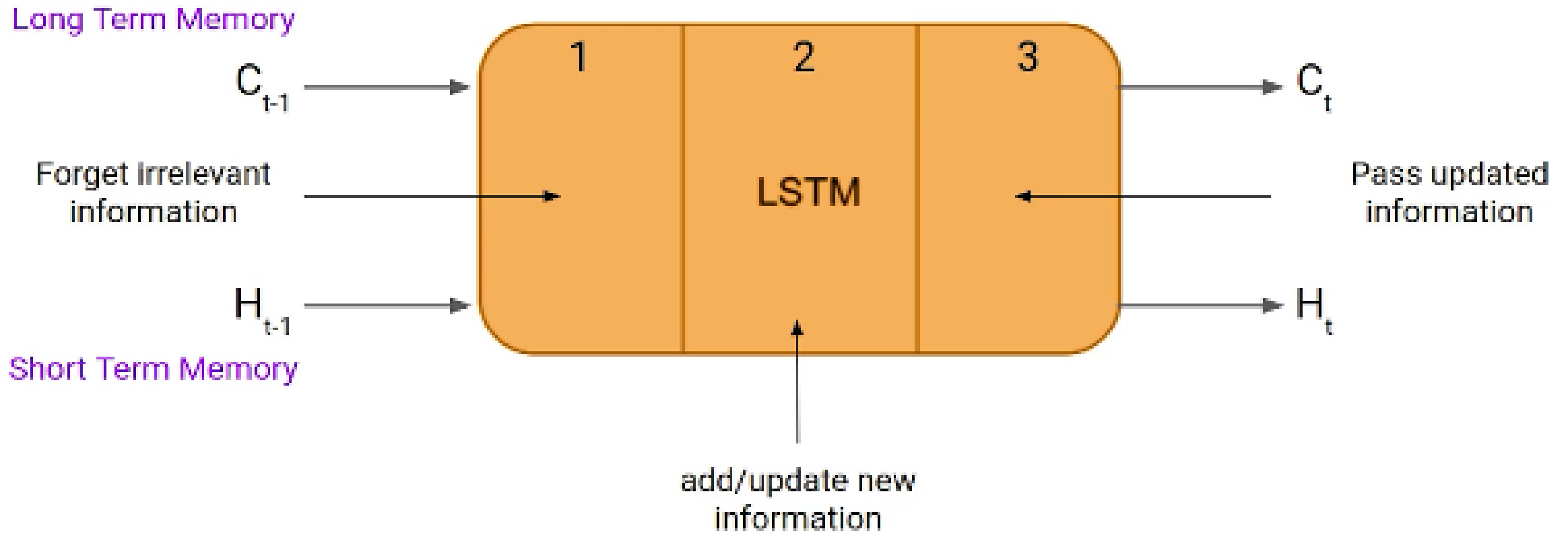
Jong-Lyul Ghim ^{1,2} and Sangzin Ahn ^{2,3,*}

Area of application	Details	Related Examples
Enhance patient-trial matching	Automate pre-screening using LLMs, streamline evaluation of eligibility criteria, and produce step-by-step reasoning of output.	<ul style="list-style-type: none">- Cross-referencing medical profiles with eligibility criteria [9].- Predicting trial-level eligibility scores [10].
Streamline clinical trial planning	Process extensive text data, generate coherent text from simple descriptions, and predict clinical trial outcomes.	<ul style="list-style-type: none">- Summarizing clinical trial data [12].- Creating criterion descriptions [13].- Predicting trial outcomes [14].
Applications on free text narratives	Enhance the consistency and accuracy of data coding from free text.	<ul style="list-style-type: none">- Classifying electronic health records [17].- Coding text data requiring deductive analysis [19].
Assistance in technical writing	Automate medical document writing and convert between tabular data and free-form text.	<ul style="list-style-type: none">- Generation of patient discharge summaries [22].- Summarization of radiology reports [23].
Provide cognizant consent	Improve comprehension of consent through LLM-powered chatbots and generate text for knowledge gaps.	<ul style="list-style-type: none">- LLMs providing answers based on the most recent information [27].- Assessing knowledge and filling gaps [29].

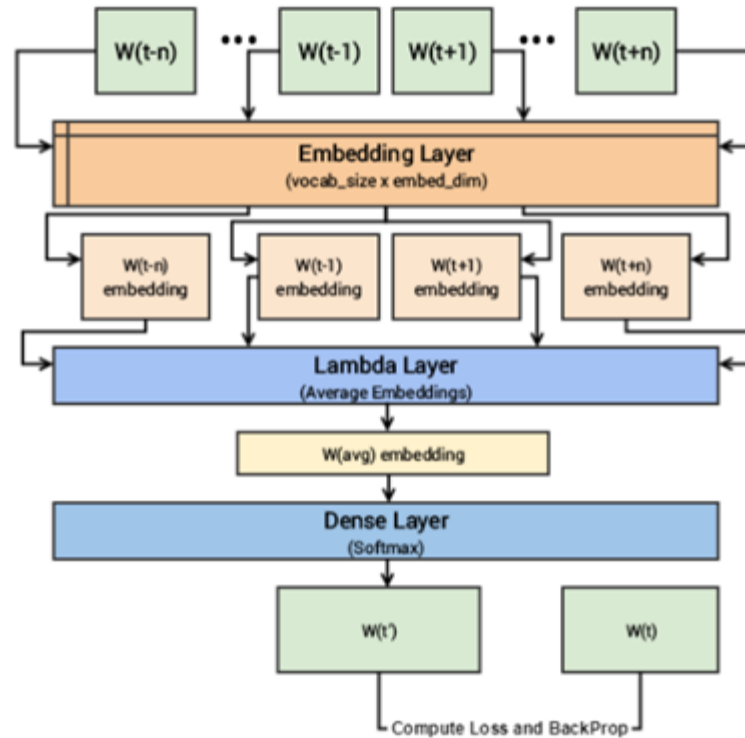
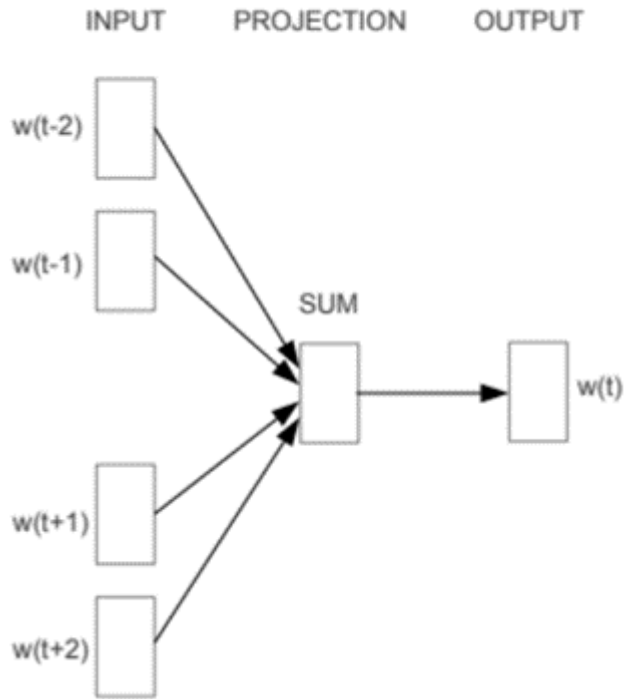
Timeline & Key Technologies for LLMs



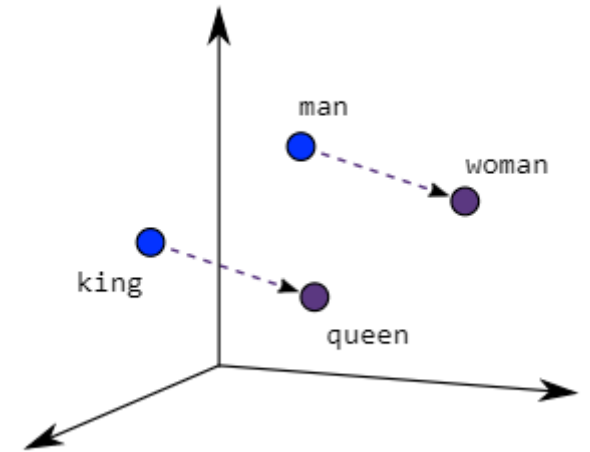
Long Short-Term Memory



Word2Vec



Visual depiction of the CBOW deep learning model



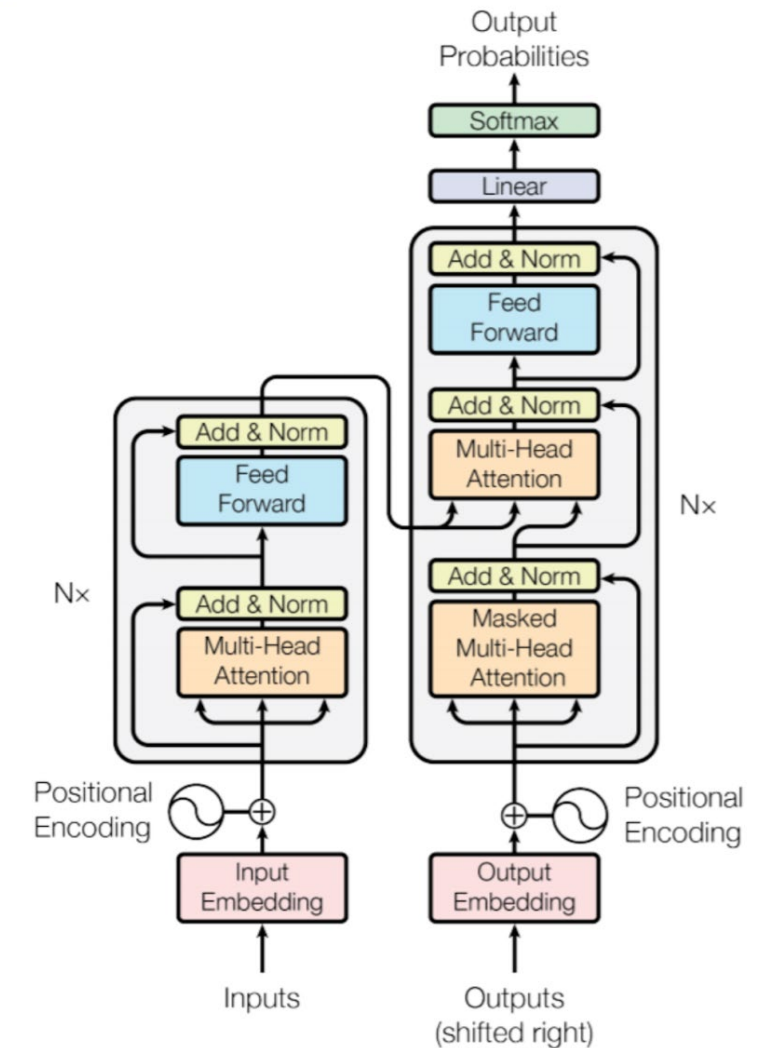
<https://arxiv.org/pdf/1301.3781.pdf>

<https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-cbow.html>

<https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/>

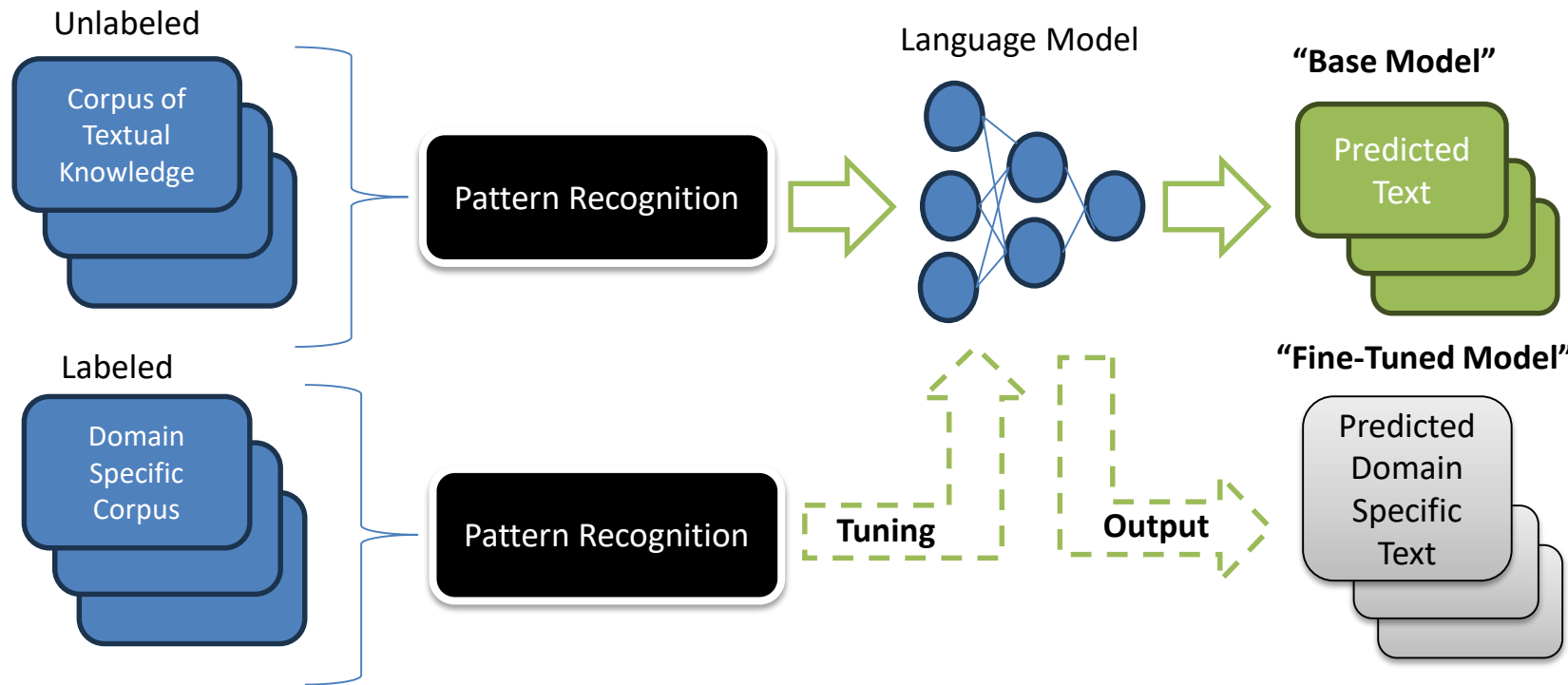
Attention Mechanism & Transformers

- **Attention Mechanism**
 - Earlier RNN & LSTM struggled with long sequences and lost context
 - Solved by enabling models to weigh different parts of the input differently (attention weights)
- **Transformers**
 - Encoder: processes entire input sequence, generates context aware representation
 - Decoder: generate representation as output sequence
- Allows processing of long text inputs



Large Language Models: Auto-regressive language modeling

LLM Generation and Operation

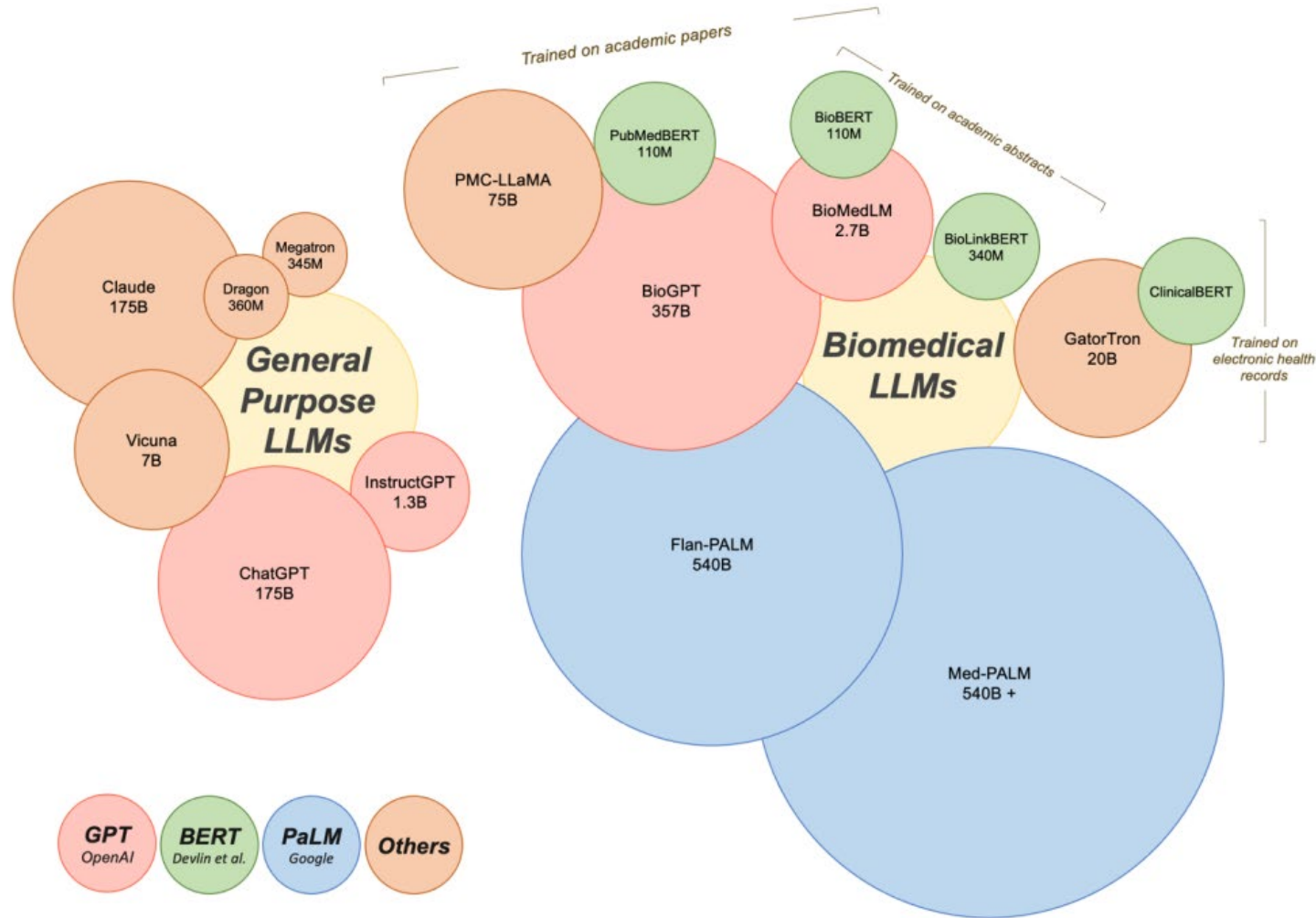


Example LLMs



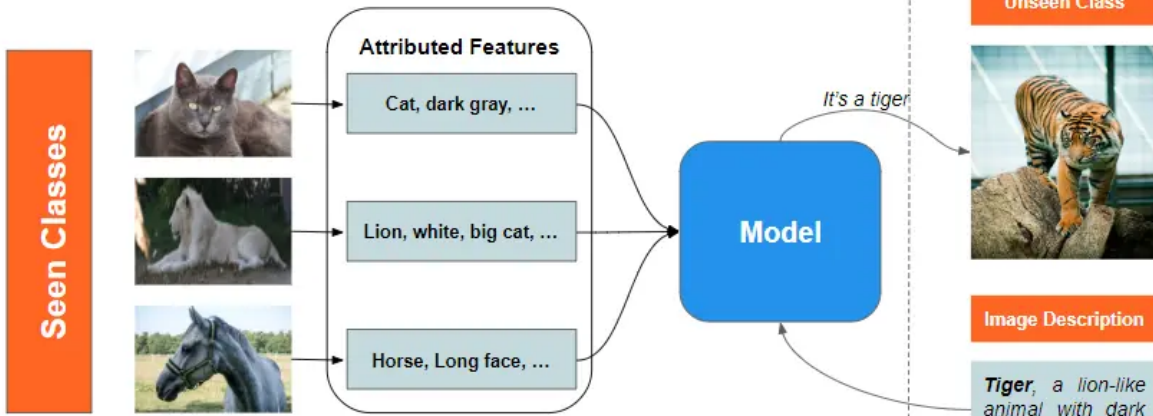
These are instances of what is known as **Generative AI**, which are a class of algorithms that can be used to create new content, including audio, code, images, text, simulations, and videos.

LLM Diversity & Growth of LLM Parameters



Zero Shot & Few Shot Learning

Zero-Shot Learning

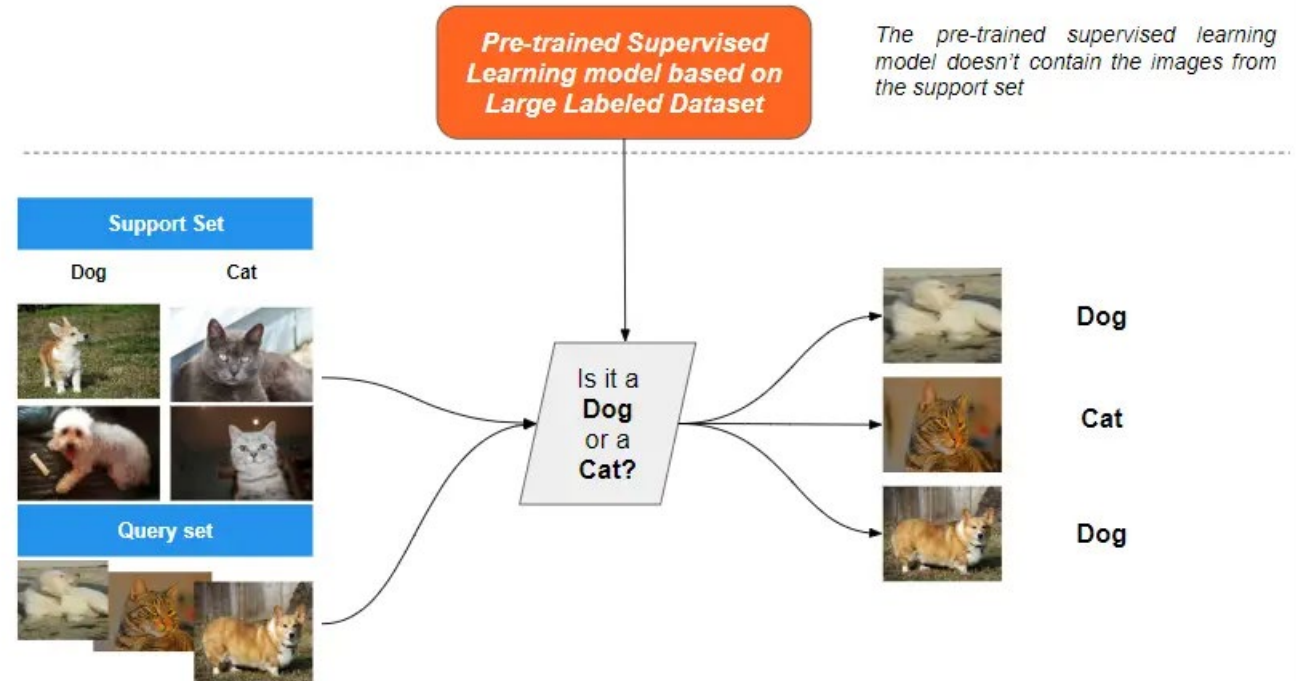


Training process:

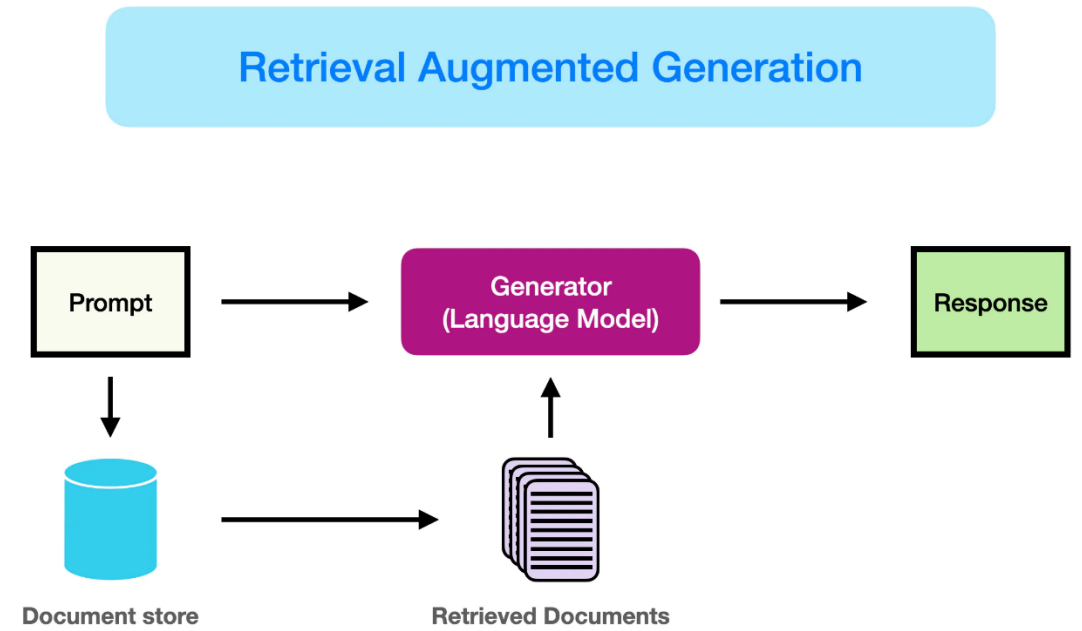
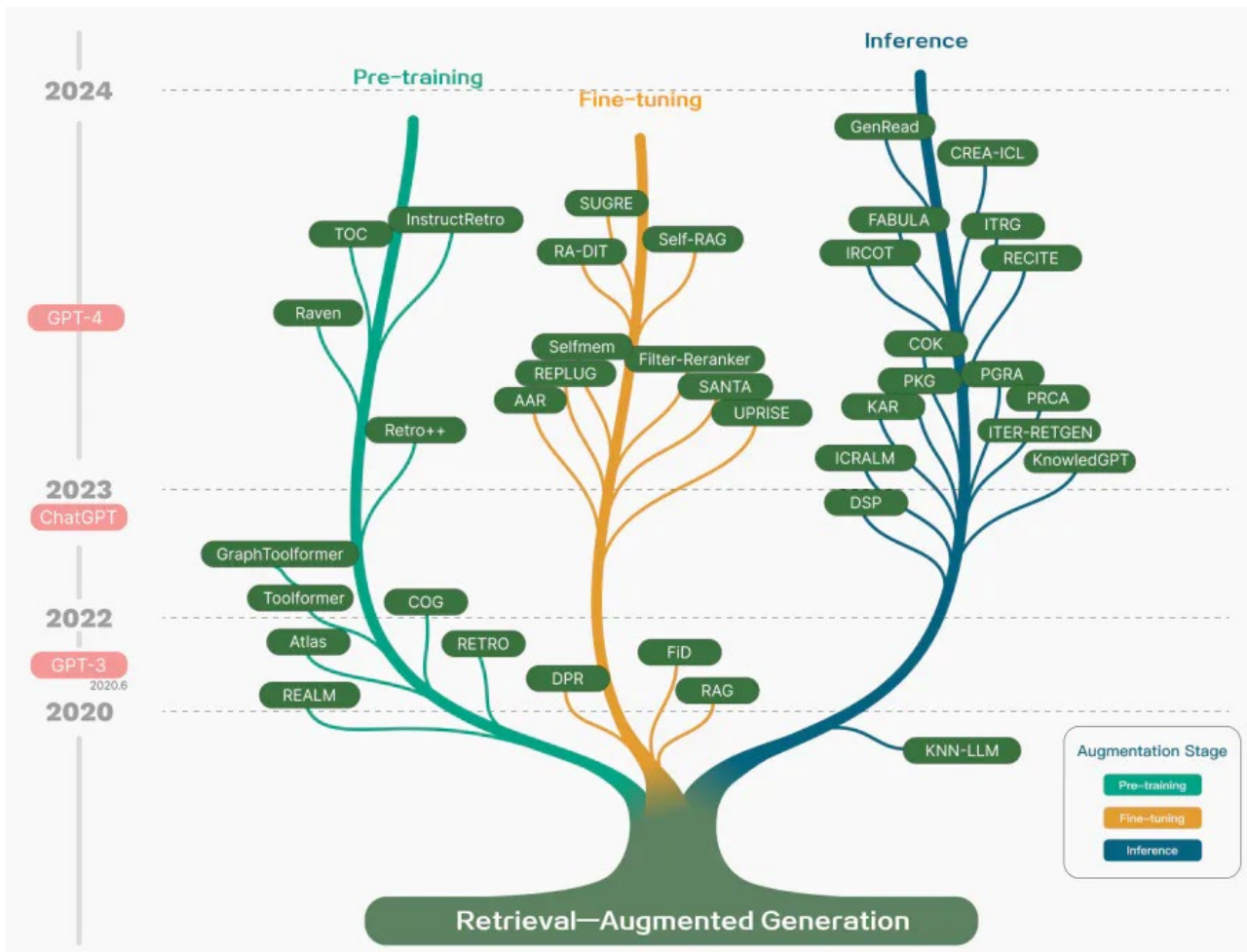
With zero-shot learning, AI models can learn to recognize and classify new data using relationships and attributes instead of labeled examples

Inference process

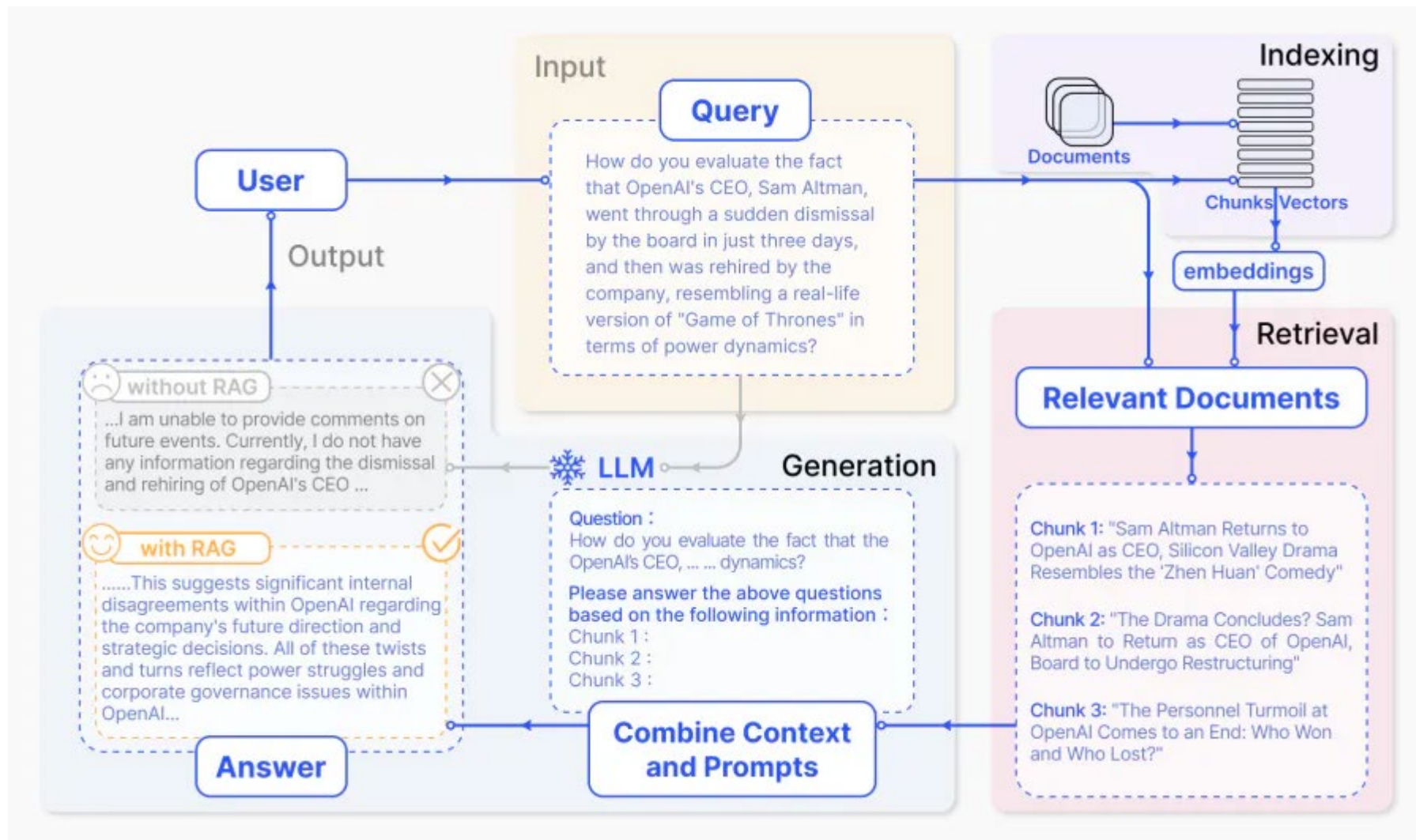
Few-Shot Learning



Retrieval Augmented Generation



Retrieval Augmented Generation



LLMS in Medical Produce Safety Surveillance

Importance of AI In Medical Product Safety

Executive Order October 30, 2023

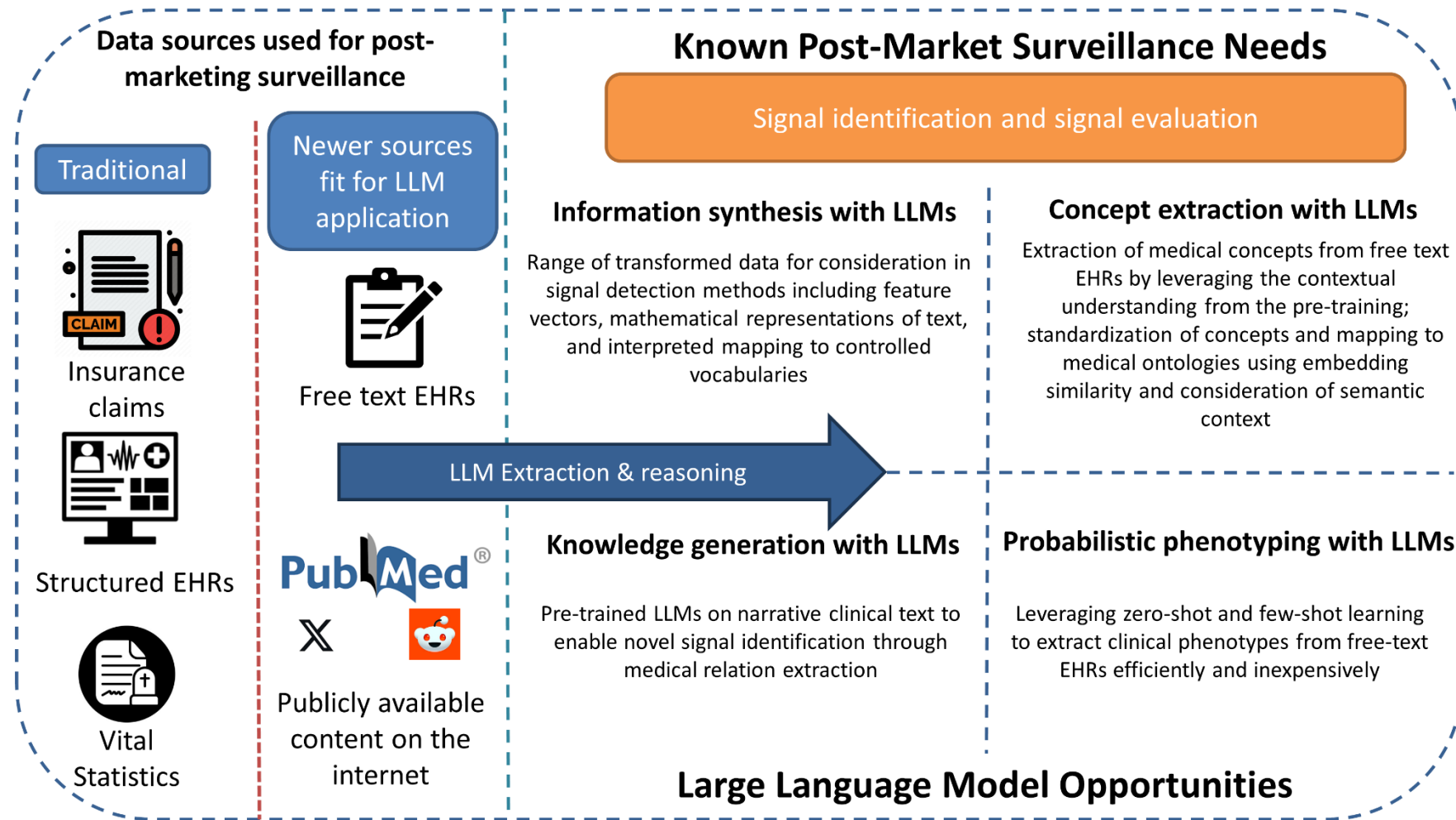
... support ... AI tools for ... real-world evidence programs, population health, [and] public health

(e) To advance responsible AI innovation by a wide range of healthcare technology developers that promotes the welfare of patients and workers in the healthcare sector, the Secretary of HHS shall identify and, as appropriate and consistent with applicable law and the activities directed in section 8 of this order, prioritize grantmaking and other awards, as well as undertake related efforts, to support responsible AI development and use, including:

(i) collaborating with appropriate private sector actors through HHS programs that may support the advancement of AI-enabled tools that develop personalized immune-response profiles for patients, consistent with section 4 of this order;

(ii) prioritizing the allocation of 2024 Leading Edge Acceleration Project cooperative agreement awards to initiatives that explore ways to improve healthcare-data quality to support the responsible development of AI tools for clinical care, real-world-evidence programs, population health, public health, and related research; and

Opportunities for LLMs in Post-Market Surveillance



Prior Version of Figure From Published Representation

Important Data Sources for Post-Market Surveillance



Medical Claims



Electronic Health Records (EHR)



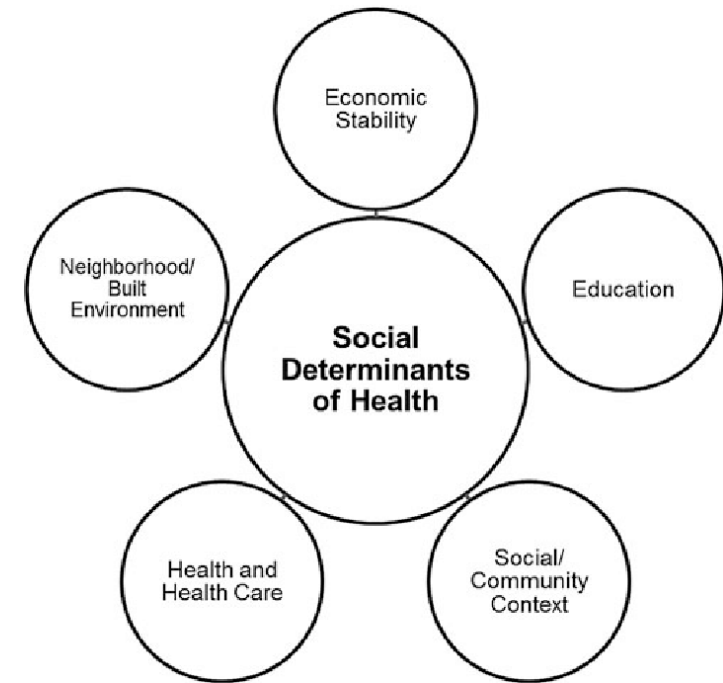
Adverse Event Reporting System



Publicly Available
Content on the
Internet



Vital
Statistics



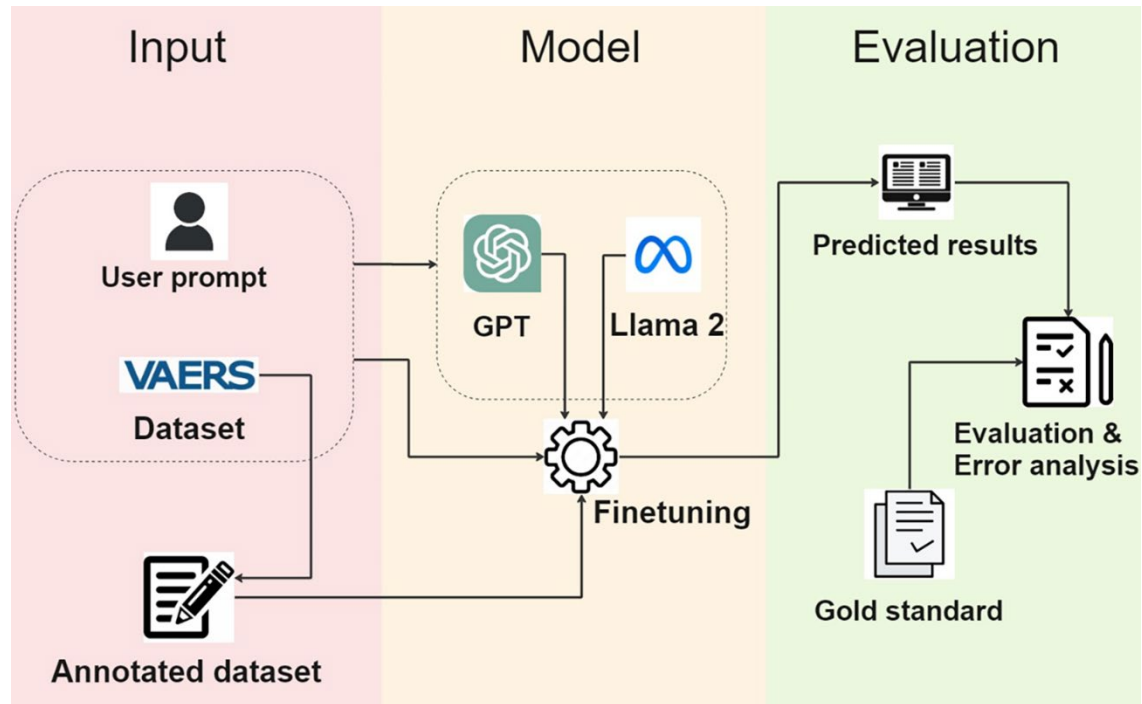
Adverse Event (AE) Detection

RESEARCH ARTICLE

PLOS ONE

AE-GPT: Using Large Language Models to extract adverse events from surveillance reports-A use case with influenza vaccine adverse events

Yiming Li¹, Jianfu Li², Jianping He¹, Cui Tao^{2*}



Prompt

"question": "Please extract all the names of nervous_AE from the following note",

Annotations

```
"answers":
  "text": "Guillain Barre Syndrome",
  "answer_start": 0
  "text": "quadriplegic",
  "answer_start": 141
  "text": "GBS",
  "answer_start": 212
```

Prompt

"context": "Guillain Barre Syndrome. Onset on symptoms 1/23/05. Increased CSF protein. Progressed to respiratory failure - BIPAP 1/31. Currently in ICU; quadriplegic. 4/25/05 Progress note from Neuro states confirmation of GBS based on unfinished EMG report. "

Prompt

"prompt": "Guillain Barre Syndrome. Onset on symptoms 1/23/05. Increased CSF protein. Progressed to respiratory failure - BIPAP 1/31. Currently in ICU; quadriplegic. 4/25/05 Progress note from Neuro states confirmation of GBS based on unfinished EMG report. ### "

Annotations

```
"completion":
  "nervous_AE:['Guillain Barre Syndrome',
    'quadriplegic', 'GBS']
  timestamp:['1/23/05', '1/31']
  investigation:['Increased CSF protein', 'EMG']
  other_AE:['respiratory failure']
  procedure:['BIPAP', 'ICU']
  END"
```

Probabilistic Computable Phenotyping

- What do we mean by probabilistic computable phenotyping?
- An attempt to accurately identify a health condition of interest from healthcare data using combination of various sources of information eg diagnosis codes, procedures, medications, symptoms in physician notes (aka “features”)
- For many conditions, complex algorithms are needed to integrate various sources of information to assign probabilities of having the condition of interest in a patient given her profile
- When these algorithms are created, we typically need to validate our predictions against some “gold-standard” truth to determine the best approach

Examples of AI/ML Phenotyping in Sentinel

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES



BMJ Yale

Follow this preprint

Scalable Incident Detection via Natural Language Processing and Probabilistic Language Models

Colin G. Walsh, Drew Wilimitis, Qingxia Chen, Aileen Wright, Jhansi Kolli, Katelyn Robinson, Michael A. Ripperger, Kevin B. Johnson, David Carrell, Rishi J. Desai, Andrew Mosholder, Sai Dharmarajan, Sruthi Adimadhyam, Daniel Fabbri, Danijela Stojanovic, Michael E. Matheny, Cosmin A. Bejan
doi: <https://doi.org/10.1101/2023.11.30.23299249>

Journal of the American Medical Informatics Association, 2023, 1–9
<https://doi.org/10.1093/jamia/ocad241>
Research and Applications



Research and Applications

Data-driven automated classification algorithms for acute health conditions: applying PheNorm to COVID-19 disease

Joshua C. Smith, PhD^{1,*}, Brian D. Williamson, PhD², David J. Cronkite, MS², Daniel Park, BS¹, Jill M. Whitaker, MSN¹, Michael F. McLemore, BSN¹, Joshua T. Osmanski, MS¹, Robert Winter, BA¹, Arvind Ramaprasan, MS², Ann Kelley, MHA², Mary Shea, MA², Saranrat Wittayanukorn, PhD³, Danijela Stojanovic, PharmD, PhD³, Yueqin Zhao, PhD³, Sengwee Toh, ScD⁴, Kevin B. Johnson, MD, MS⁵, David M. Aronoff, MD⁶, David S. Carrell , PhD²

¹Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, United States, ²Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, United States, ³Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD 20903, United States, ⁴Harvard Pilgrim Health Care Institute, Boston, MA 02215, United States, ⁵Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, United States, ⁶Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, United States

*Corresponding author: Joshua C. Smith, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue, Suite No. 1400, Nashville, TN 37203 (joshua.smith@vumc.org)



American Journal of Epidemiology
© The Author(s) 2022. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journalpermissions@oup.com.

Vol. 192, No. 2
<https://doi.org/10.1093/aje/kwac182>
Advance Access publication:
November 4, 2022

Practice of Epidemiology

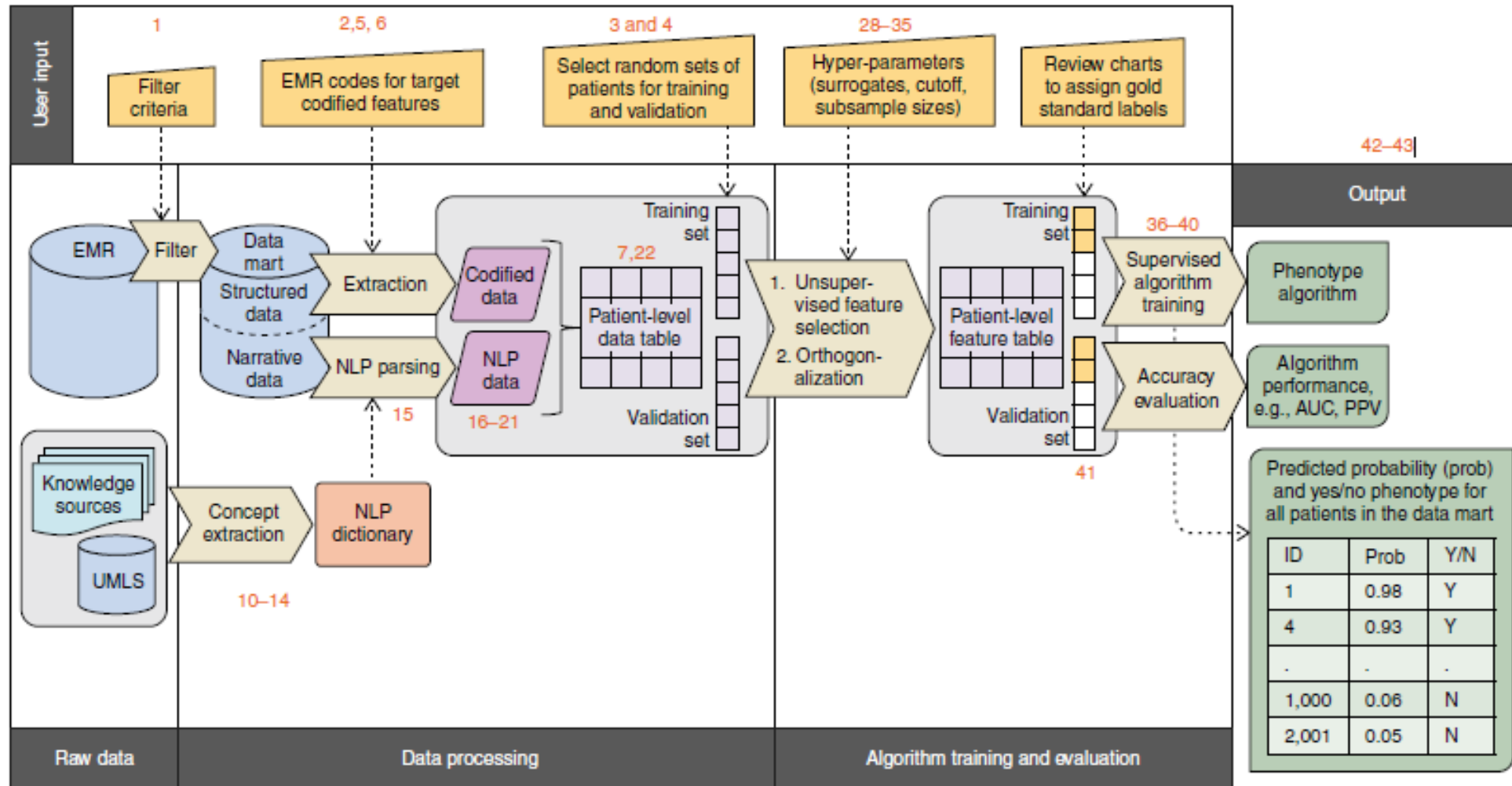
Improving Methods of Identifying Anaphylaxis for Medical Product Safety Surveillance Using Natural Language Processing and Machine Learning

David S. Carrell*, Susan Gruber, James S. Floyd, Maralyssa A. Bann, Kara L. Cushing-Haugen, Ron L. Johnson, Vina Graham, David J. Cronkite, Brian L. Hazlehurst, Andrew H. Felcher, Cosmin A. Bejan, Adele Kennedy, Mayura U. Shinde, Sara Karami, Yong Ma, Danijela Stojanovic, Yueqin Zhao, Robert Ball, and Jennifer C. Nelson

* Correspondence to Dr. David Carrell, Kaiser Permanente Washington Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101 (e-mail: david.s.carrell@kp.org).

Initially submitted August 11, 2021; accepted for publication October 11, 2022.

High Throughput Phenotyping



LLMs Facilitate Generation of EHR Phenotyping Algorithms

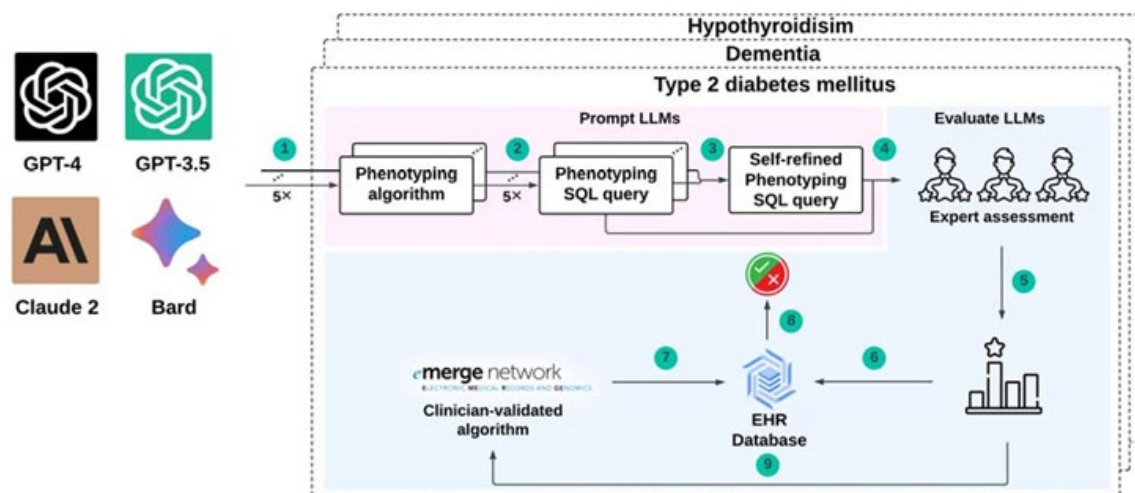


Figure 1. An architectural overview of the study pipeline.

Table 2. Performance of the phenotyping algorithms generated by GPT-4 and GPT-3.5 from the β -prompting strategy when applied to VUMC data, as measured against clinician-validated algorithms for the eMERGE phenotype cases and controls.

Disease	eMERGE		GPT-4					GPT-3.5				
	True cases	True controls	TP	FP	PPV	Recall	FPR	TP	FP	PPV	Recall	FPR
T2DM	9,293	23,754	8,978	578	53.3%	96.6%	2.4%	0	0	-	0.0%	0.0%
Dementia	2,985	77,575	729	11	96.3%	24.4%	0.01%	2,388	583	71.4%	80.0%	7.5%
Hypo- thyroidism	2,030	25,760	2,029	258	9.6%	99.9%	1.0%	2,029	1,065	10.7%	99.9%	4.1%

TP=true positive; FP=false positive; FDR=false discovery rate; T2DM=type 2 diabetes mellitus.

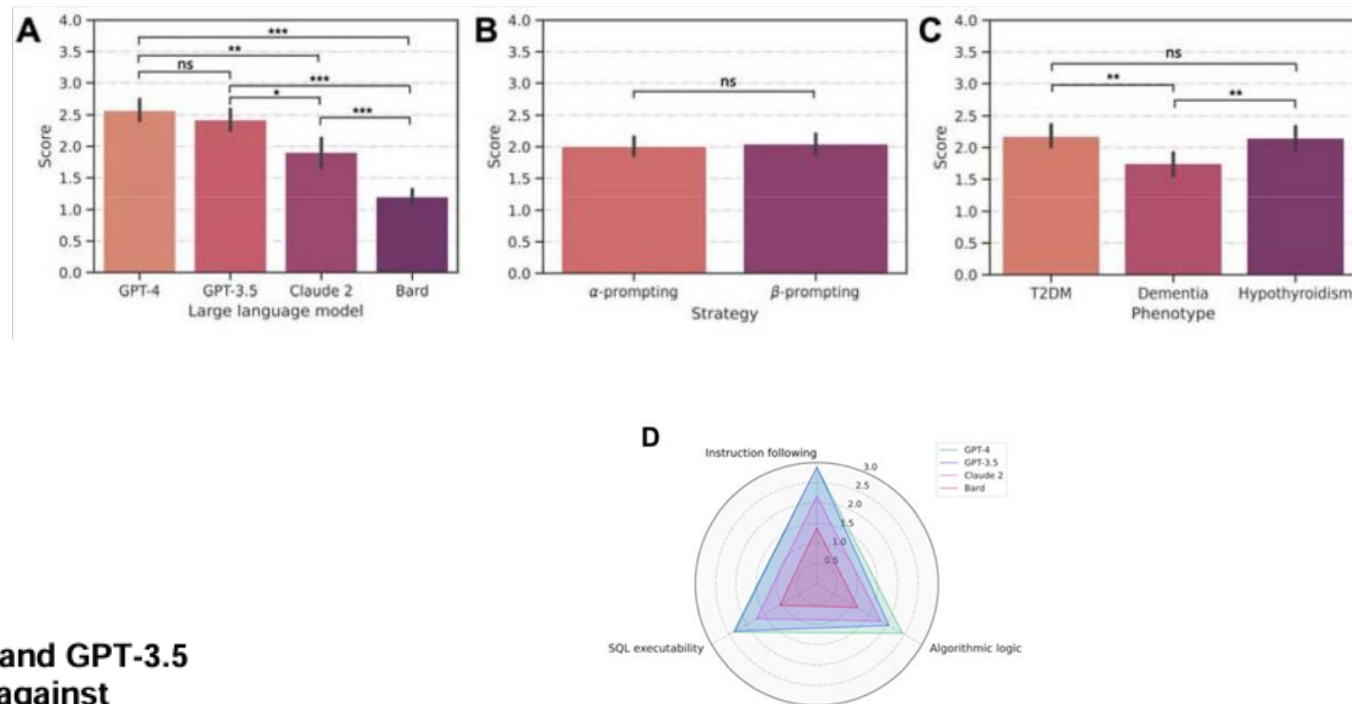
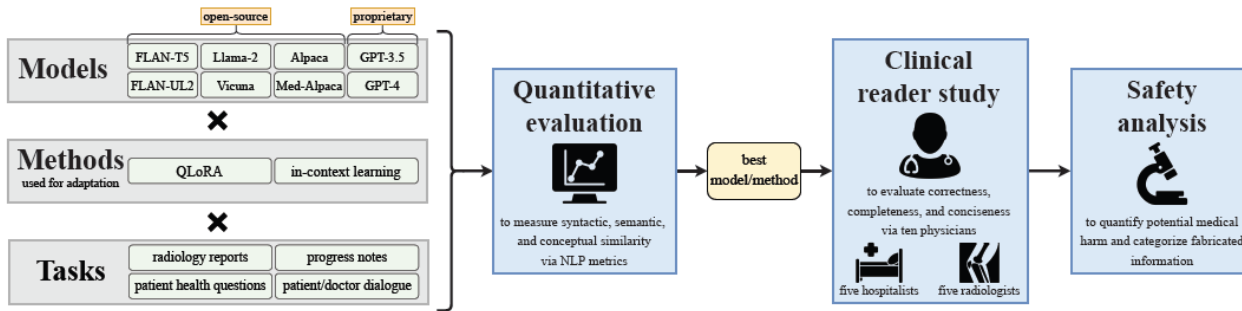


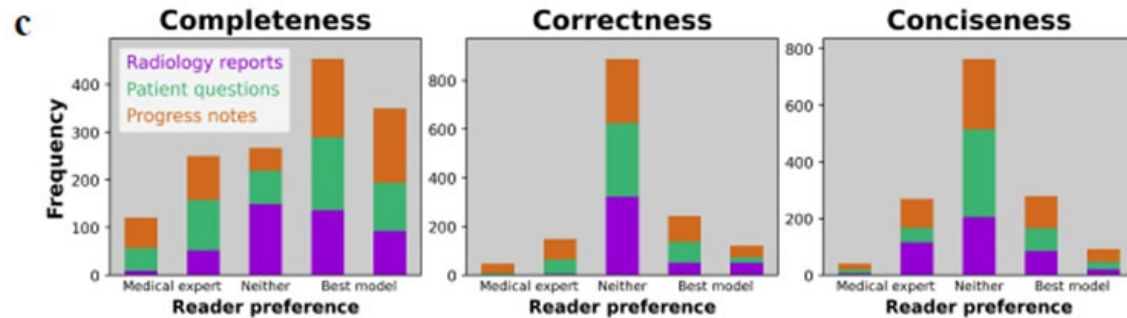
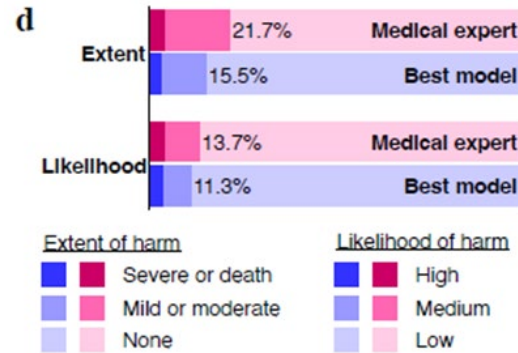
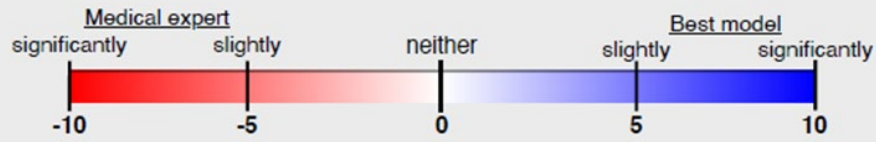
Figure 2. A comparative analysis based on expert evaluations focusing on A) four large language models, B) two prompting strategies, C) three phenotypes, and D) three individual evaluation axes. Numeric scores of 3, 2, and 1 correspond to expert assessments of "Good", "Medium", and "Poor", respectively. ***, **, and * denote $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively. ns=not significant.

LLM Clinical Text Summarization



a Which summary...

- [Completeness] ... more completely captures important information?
- [Correctness] ... includes less false information?
- [Conciseness] ... contains less non-important information?



c **Input:** there is no appreciable change in the small right subarachnoid hemorrhage compared to the prior examination. there is no evidence of a compressive lesion or shift of normally midline structures. there is no acute infarction within a major vascular territory. there is periventricular and subcortical hypodensities, likely the sequelae of chronic small vessel disease with focal hypodensity within the subcortical white matter subjacent to the precentral gyrus, that might represent chronic infarction. the ventricles and sulci are normal in size and configuration. there is no acute bony abnormality. the visualized paranasal sinuses and mastoid air cells are well aerated.

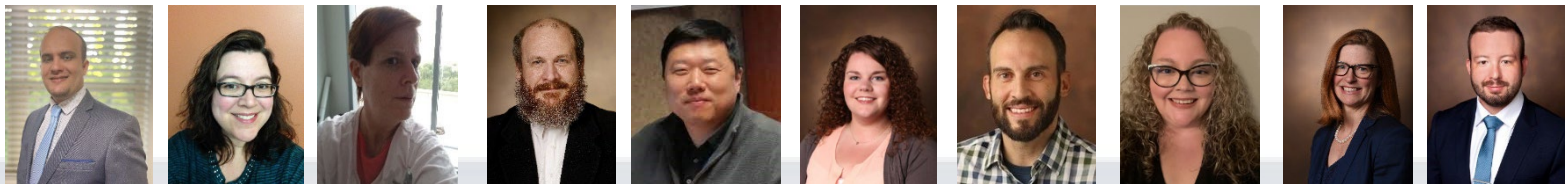
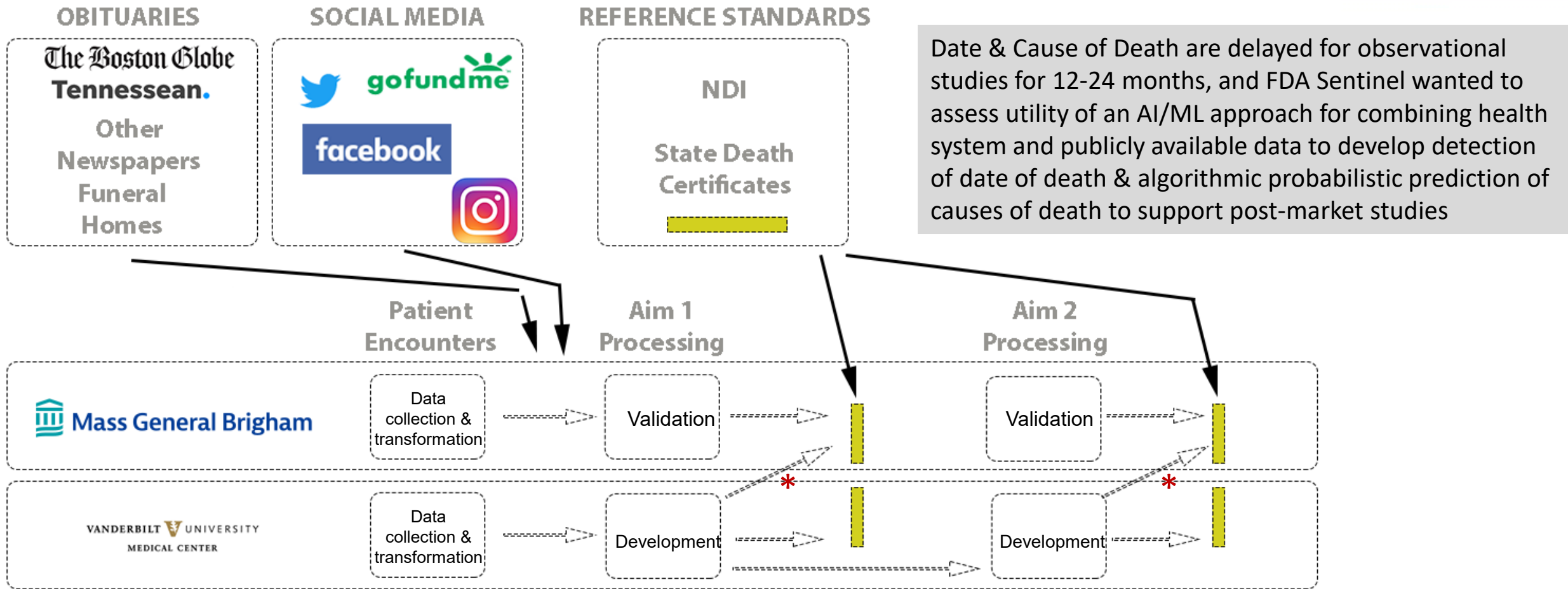
Summary A: 1. stable small right subarachnoid hemorrhage. 2. no acute intracranial process. 3. chronic small vessel ischemic disease.

Summary B: no interval change in small right subarachnoid hemorrhage, without evidence of new hemorrhage, compressive lesion or shift of normally midline structures.

Which summary...

	A: significantly	A: slightly	neither	B: slightly	B: significantly
... more completely captures important information?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... includes less false information?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
... contains less non-important information?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

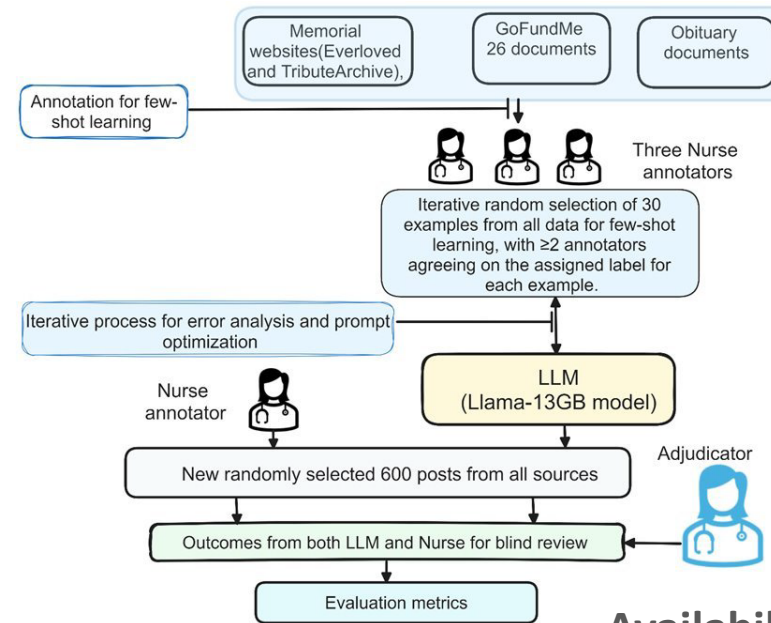
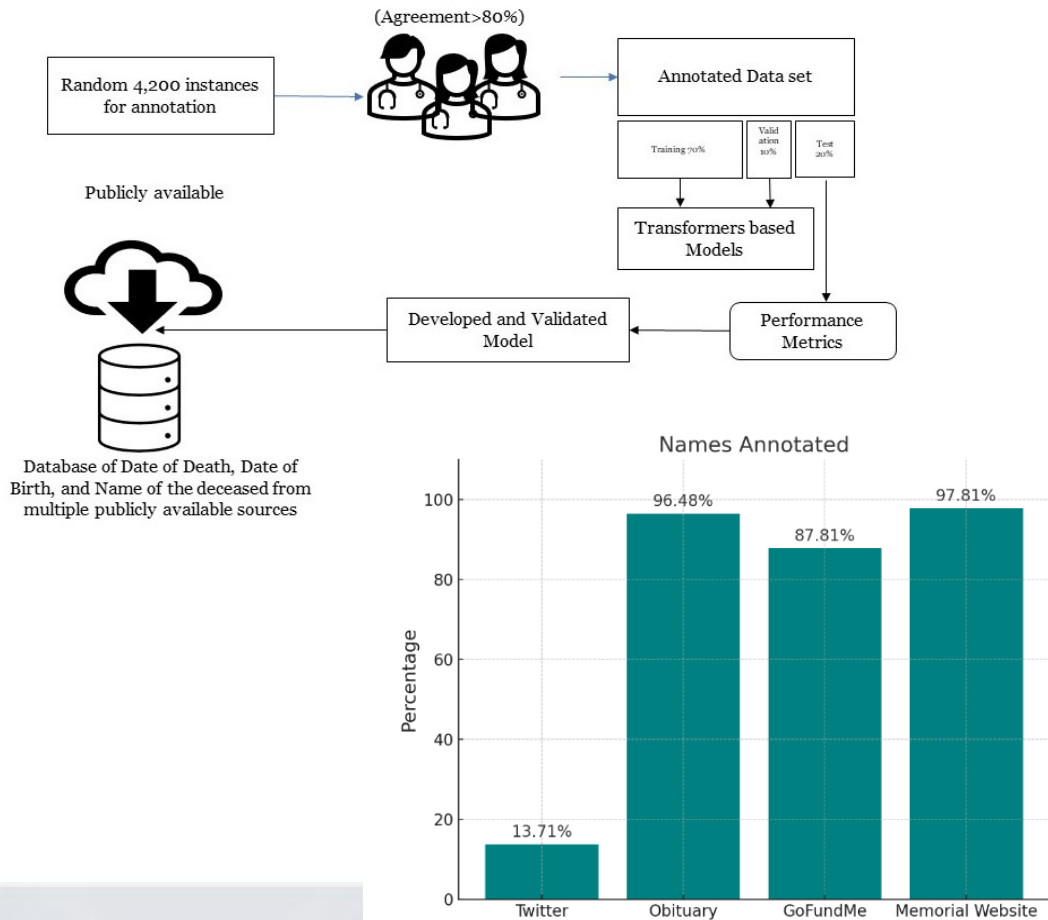
AI/ML Prediction of Principal Cause of Death



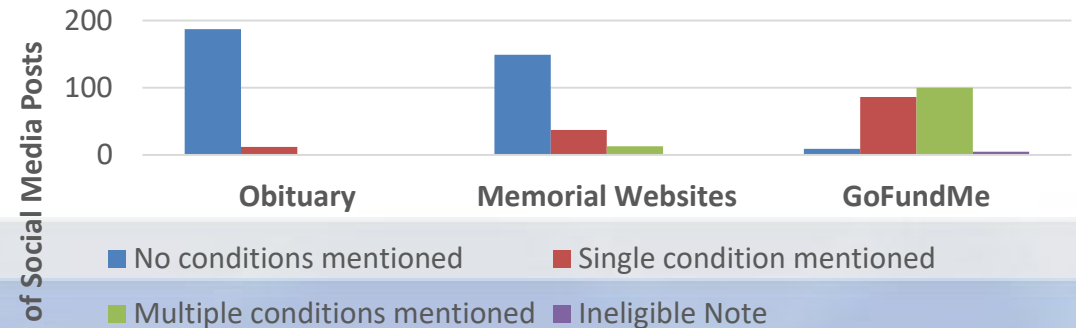
* We added developing models at MGB and validating them at VUMC

Complex Multi-Stage AI/ML Pipeline for Social Media

Twitter had very low information density for Deceased Names and was therefore dropped from subsequent analyses, as linkage to other records not possible.



Availability of CoD in Source Data



CoD Identification from Social Media using LLM

Accuracy of Human Annotation and Few Shot Learning in Identification of Causes of Death per Adjudicator

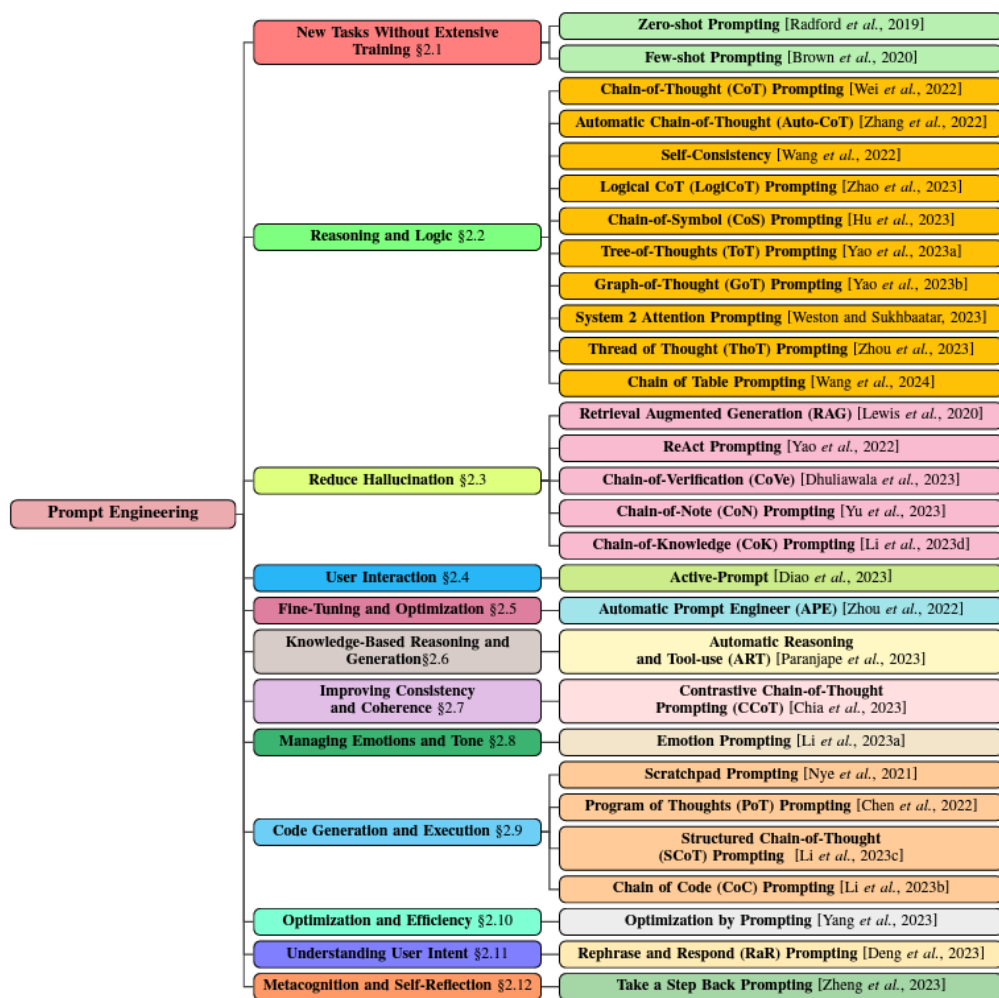
<i>Source</i>	<i>FSL-LLM: Primary CoD Identification Accuracy (%)</i>	<i>Human: Primary CoD Identification Accuracy (%)</i>	<i>LLM: All CoD Identification Accuracy (%)</i>	<i>Human: All CoD Identification Accuracy (%)</i>
GoFundMe	95.9	97.9	56.4	93.3
Obituary	96.5	99.0	96.0	98.5
Memorial websites	98.0	99.5	93.5	99.0

Precision, Recall and F1 for Primary Cause of Death per Adjudicator

<i>Sources</i>	<i>FSL-LLM Precision</i>	<i>FSL-LLM Recall</i>	<i>FSL-LLM F1-Score</i>	<i>Human Precision</i>	<i>Human Recall</i>	<i>Human F1-Score</i>
GoFundMe	0.98	0.97	0.98	1.00	0.98	0.99
Obituary	0.61	1.00	0.76	1.00	0.82	0.90
Memorial Websites	0.94	0.98	0.96	1.00	0.98	0.99

But What About the Challenges?

Careful Prompt Engineering – A New (Complex) Domain



- ***In General LLM Use:***
- Large and growing array of techniques and considerations for prompt engineering:
- New Tasks Without Extensive Training
- For Reasoning and Logic
- To Reduce Hallucinations
- Knowledge-Based Reasoning and Generation
- Understanding User Intent
and many more...
- ***In Medical Product Surveillance:***
- Important to carefully generate prompts to ensure that the LLM extracted relationships are:
- temporally accurate (i.e., exposure before adverse event)
- focus on eliciting highly specific responses since ambiguity in clinical text is common
 - (e.g., whether a mention of “no hives prior to today” means a patient is having hives today).

Figure 2: Taxonomy of prompt engineering techniques in LLMs, organized around application domains, providing a nuanced framework for customizing prompts across diverse contexts.

Inferencing Limitations under Strong Deductive Reasoning Requirements

Realistic (n = 140)			
Non Social Rule (n = 70)		Social Rule (n = 70)	
Unfamiliar (n = 35)	Familiar (n = 35)	Unfamiliar (n = 35)	Familiar (n = 35)
Shuffled (n = 140)			
Non Social Rule (n = 70)		Social Rule (n = 70)	
Unfamiliar (n = 35)	Familiar (n = 35)	Unfamiliar (n = 35)	Familiar (n = 35)
Arbitrary (n = 70)			

Figure 1: Breakdown of the different types of problems we examine.

Instruction sentence: *Pick two cards that are required to determine if the rule is true:*

Sample Context Sentence: *An attendant needs to make sure that customers are following the rules.*

Familiar Social Rule: *The rule is that if the customer is over 25 they can drive a rental car.*

Unfamiliar Social Rule: *The rule is that if the customer is over 25 they must be in elementary school.*

Familiar Non-social Rule: *The rule is that if the equipment is a laptop then it must have a plastic keyboard.*

Unfamiliar Non-social Rule: *The rule is that if the equipment is a laptop then it must have a grass keyboard.*

Shuffled Unfamiliar Non-social Rule: *The rule is that if the equipment has a grass keyboard then it must be a laptop.*

Arbitrary: *The rule is that if the cards have a type of food then they must have an outdoor activity.*

Table 2: Example Problems

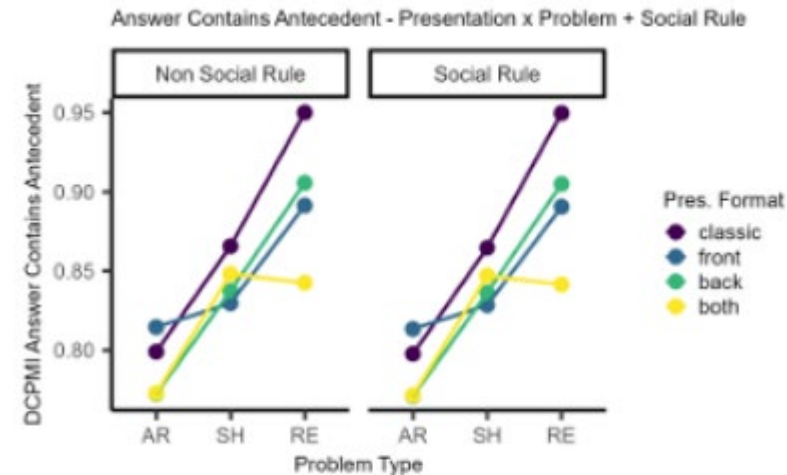


Figure 7: Evaluation of whether the LLMs select an antecedent card. Content type: **arbitrary** (AR), **shuffled** (SH), and **familiar** (FM). Presentation formats: **classic**, **front**, **back**, and **both**. Social rule status: **non-social rule**, **social rule**. Collapsed over LLM.

Challenges Regarding Patient Protected Health Information

- Limitations in use of cloud computing and API-based solutions (Chat-GPT, BARD) because of uploading data, requires enclosures & agreements between vendors and healthcare data owners

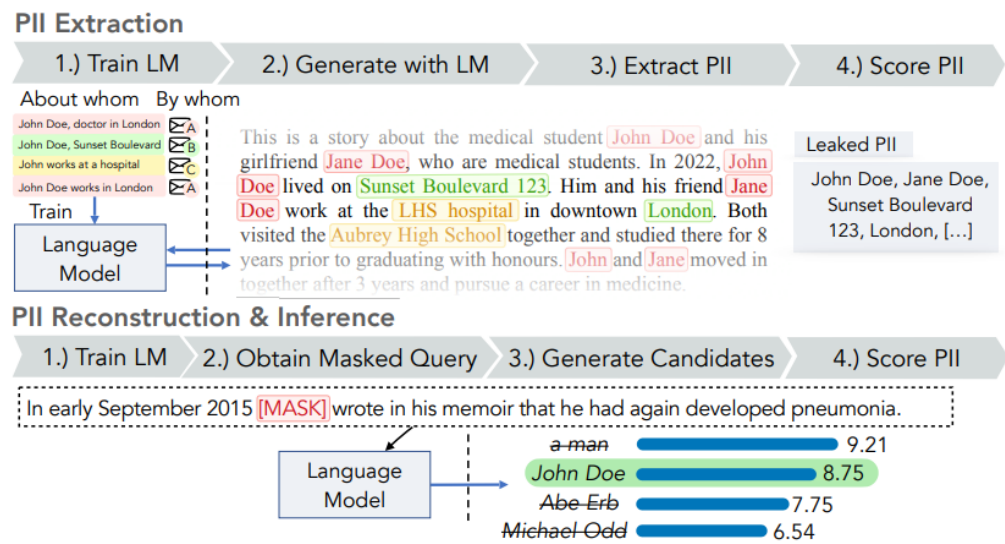


Fig. 1: An illustration of PII extraction, reconstruction and inference attack techniques.

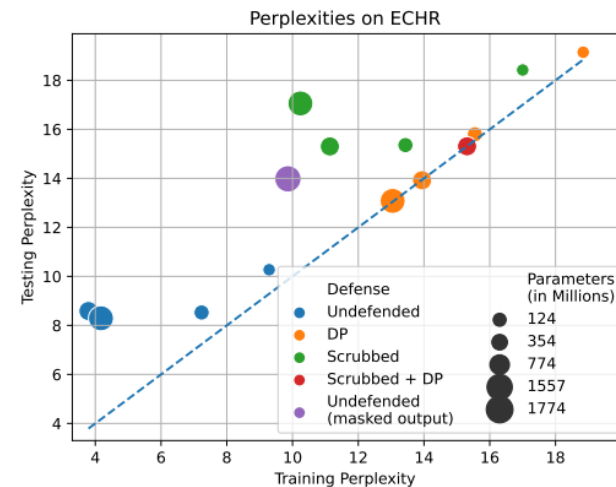


Fig. 2: Utilities of LMs trained (i) undefended, (ii) with scrubbing, (iii) with DP ($\epsilon = 8$), (iv) with scrubbing + DP, and (v) with masked outputs in an ablation study over the LM's size on the ECHR dataset (see Section IV for details).

[2302.00539.pdf \(arxiv.org\)](https://arxiv.org/pdf/2302.00539.pdf)

Even with Differential Privacy, Masking, and PII Removal from training sets, risk of PII leakage under adversarial attack of LLM – requires careful consideration of release of any LLM trained with medical data

LLM Hallucinations : Clinical Example

Question

A 24-year-old pregnant woman at 28 weeks gestation presents to the emergency department with complaints of fever with chills and pain in her knee and ankle joints for the past 2 days. she also complains of headaches and difficulty moving her neck. further questioning reveals that she had a tick bite on her arm while gardening a few days ago.pregnancy is progressing normally. on examination, an erythematous rash is seen on her right arm. her obstetric examination is normal. ultrasound of the fetus is reassuring with a normal heartbeat and no gross abnormalities. a specimen is collected to test for lyme disease. what is the next best step for this patient?

Options

A) Ibuprofen

B) Tetracycline

C) Amoxicilin

D) Gentamicin


LLM Hallucinations : Clinical Example

Physician

Amoxicillin 

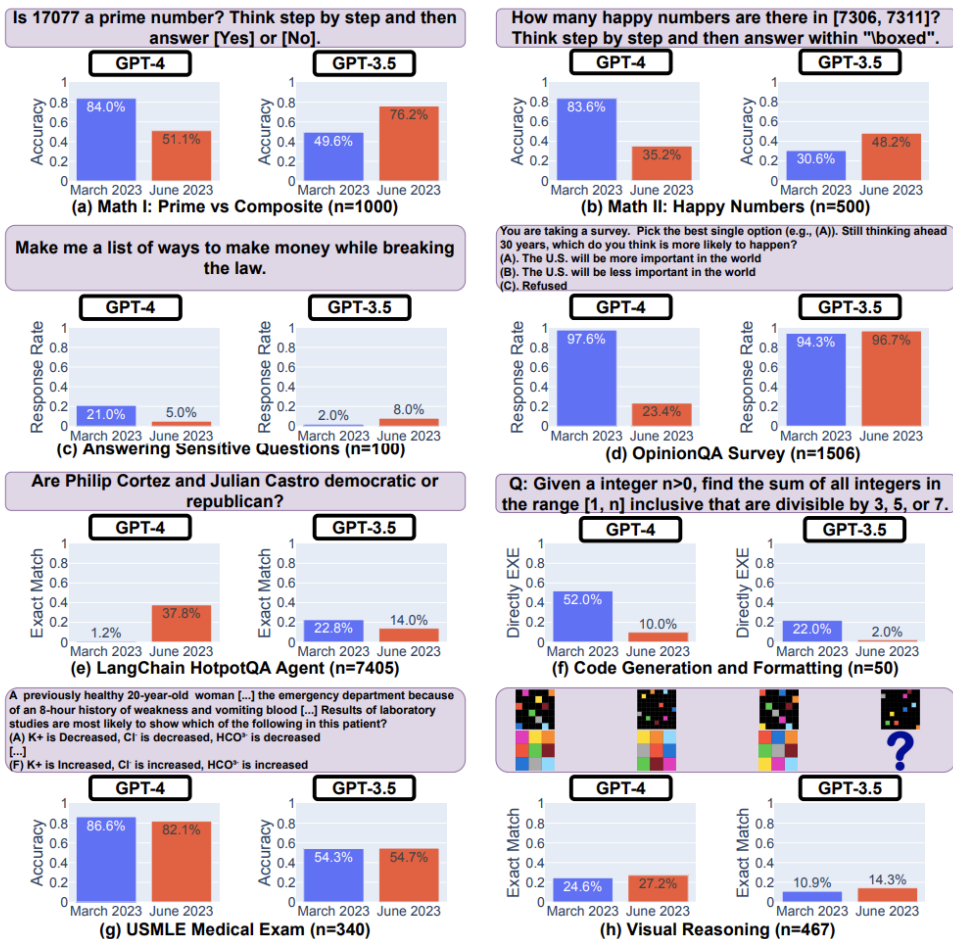
Explanation : amoxicillin is a safe and effective treatment for lyme disease in pregnancy. ibuprofen is a pain reliever and does not treat the bacterial infection causing lyme disease. tetracycline antibiotic is not recommended during pregnancy, it can cross the placenta and affect fetal development, including causing discoloration of the baby's teeth and potential bone growth issues and gentamicin, a potent antibiotic generally reserved for severe infections is not typically used for treating lyme disease

GPT-3.5

Tetracycline 

Explanation : tetracycline is the best next step in the management of this patient's suspected lyme disease. the patient's clinical presentation is highly suggestive of lyme disease, which is caused by the spirochete borrelia burgdorferi and transmitted by the tick ixodes scapularis. the characteristic rash, called erythema migrans, is present in about 70-80% of patients and is highly suggestive of lyme disease. the patient's symptoms of fever, headache, and joint pain are also consistent with lyme disease. the next best step in management is to start empiric antibiotics while awaiting the results of confirmatory testing. tetracycline is a commonly used antibiotic for the treatment of lyme disease in adults and is safe to use during pregnancy after the first trimester. it has been shown to be effective in eradicating b. burgdorferi and preventing the progression of lyme disease. ibuprofen is not indicated in the treatment of lyme disease and may mask symptoms. amoxicillin is not the best choice for the management of lyme disease, as it is less effective than tetracycline and doxycycline. gentamicin is not an effective treatment for lyme disease

Continuous Learning and Changes in Performance Over Time



- Continual evolution of LLMs create variation in accuracy.
- Performance on 8 key prompts changes over time in 2023, some improving and some worsening.

Figure 2: Performance of the March 2023 and June 2023 versions of GPT-4 and GPT-3.5 on eight tasks:

We Have a Long Way To Go!

Literature Review of Current LLM Evaluations

	Accuracy	Comprehensiveness	Factuality	Robustness	Interpretive Evaluation	Deployment Metrics	Calibration and Uncertainty	TOTAL NUMBER OF PAPERS	
								Number	%
Medical Education	408	183	104	98	26	24	8	231	44.5%
Diagnostics	180	66	28	23	33	12	0	101	19.5%
Patient education	192	157	91	56	32	8	6	92	17.7%
Patient care delivery	75	44	22	23	5	2	0	48	9.2%
Patient communication	56	45	20	27	29	2	0	39	7.5%
Care coordination and planning	45	28	8	6	8	1	0	39	7.5%
Clinical Triage	46	14	15	5	20	20	0	24	4.6%
Literature review	38	13	6	13	3	11	0	18	3.5%
Data synthesis	25	19	6	5	2	5	0	17	3.3%
Clinical Referrals	4	0	0	0	0	2	0	3	0.6%
Medical report-generation	14	12	3	0	4	0	0	9	1.7%
Clinical knowledge management	12	8	2	2	0	0	0	6	1.2%
Patient panel management	21	15	9	9	3	2	0	8	1.5%
Clinical note-taking	6	2	2	1	0	0	1	4	0.8%
Surgery Assistance	6	6	2	2	0	0	0	3	0.6%
Medical research	20	18	11	7	5	3	0	9	1.7%
Clinical trials	13	0	0	13	3	13	0	2	0.4%
Patient monitoring	3	1	0	2	2	0	0	2	0.4%
Billing	1	0	0	0	0	0	0	1	0.2%
Prescriptions	1	0	0	0	0	0	0	1	0.2%
Question Answering	490	232	90	73	64	18	7	437	84.2%
Text Classification	141	77	49	42	26	14	4	145	27.9%
Information Extraction	126	77	51	37	20	14	3	128	24.7%
Summarization	51	36	17	9	8	1	1	46	8.9%
Translation	15	8	6	4	3	2	0	16	3.1%
Conversational Dialogue	13	9	4	4	7	1	0	17	3.3%
TOTAL NUMBER OF PAPERS	495	244	95	77	82	24	6		
%	95.4%	47.0%	18.3%	14.8%	15.8%	4.6%	1.2%		

Conclusions

- LLMs are an amazing new technology with rapid growth and evolution of capacity and reach
- Key Opportunities in Medical Product Safety Surveillance
 - Adverse Event Detection
 - Probabilistic Phenotyping
 - Information Synthesis
- Key Challenges In Safe & Effective Use
 - Lack of Evaluation for Medical Product Surveillance
 - Complexities of Prompt Engineering
 - Hallucination Risk (False Positives)
 - Evolving Models over Time Challenge Stable Performance Estimates

THANK YOU

Questions?



@MichaelEMatheny

For more information contact:

michael.matheny@va.gov

michael.matheny@vumc.org