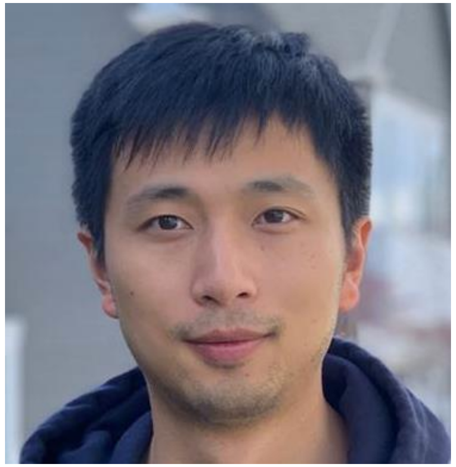
Precision Health in the Age of LLMs



Sheng Zhang

Javier González

Tristan Naumann

Hoifung Poon

Microsoft Health Futures

Microsoft Research

Overview

Precision health

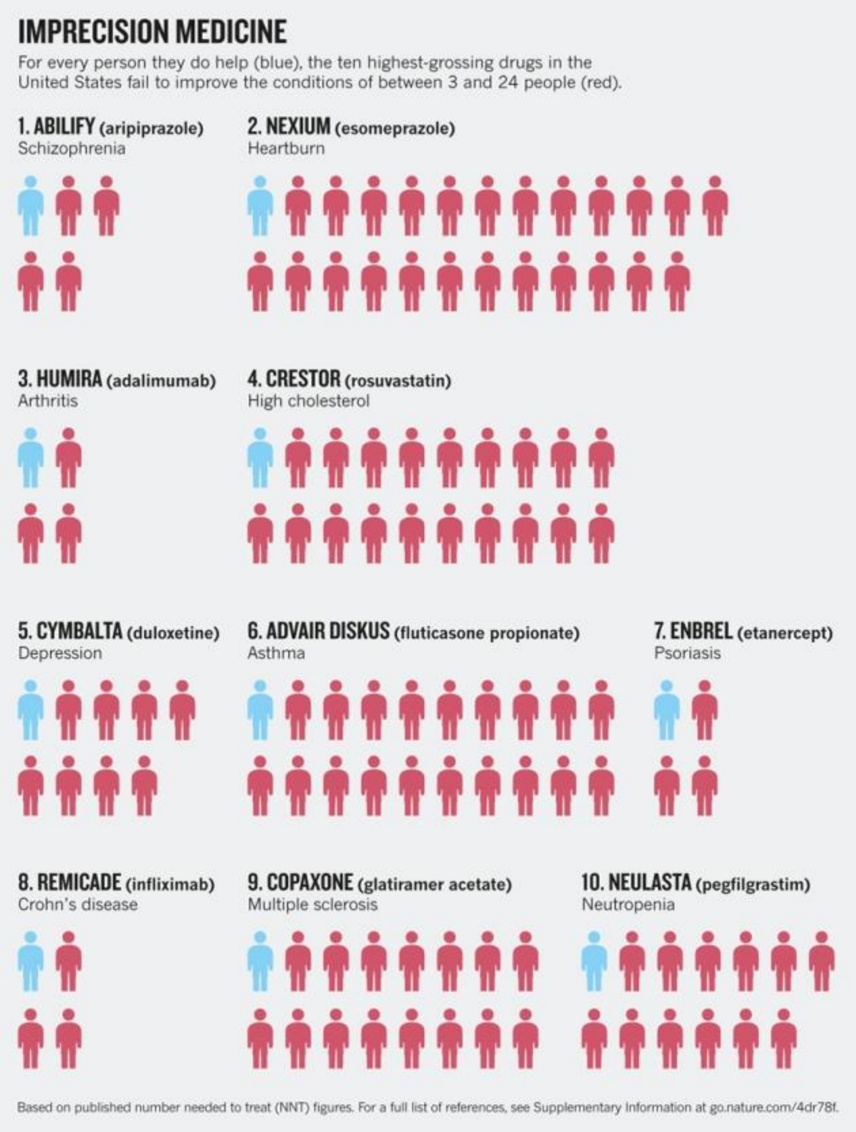
Intelligence revolution

Biomedical LLMs

Application challenges

Research frontiers

Medicine Today Is Imprecise



Top 20 drugs
80% non-responders

Wasted
1/3 health spending
$1 Trillion / year

Cancer: Traditional Treatment

"Slash, poison, and burn"

Toxicity: High

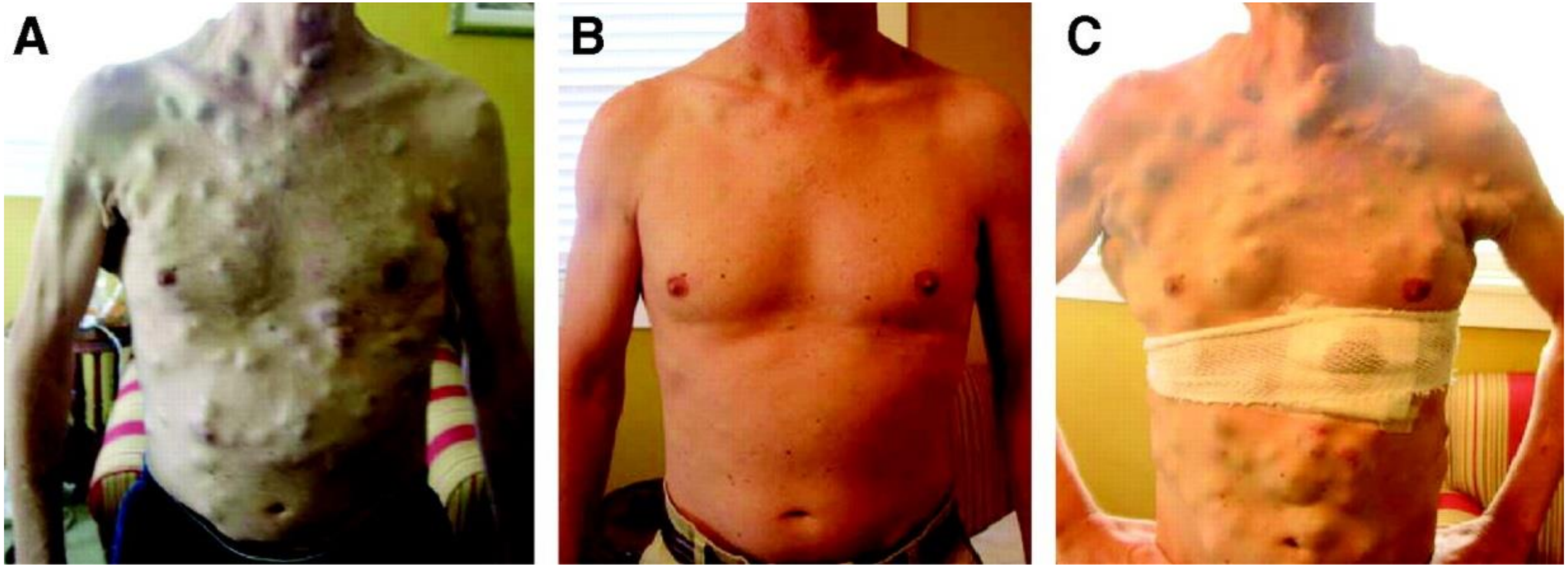
Efficacy: Low

Cancer: Targeted Therapy

Before Treatment 15 Weeks

Vemurafenib on BRAF-V600 Melanoma

Cancer: Targeted Therapy



Before Treatment 15 Weeks 23 Weeks

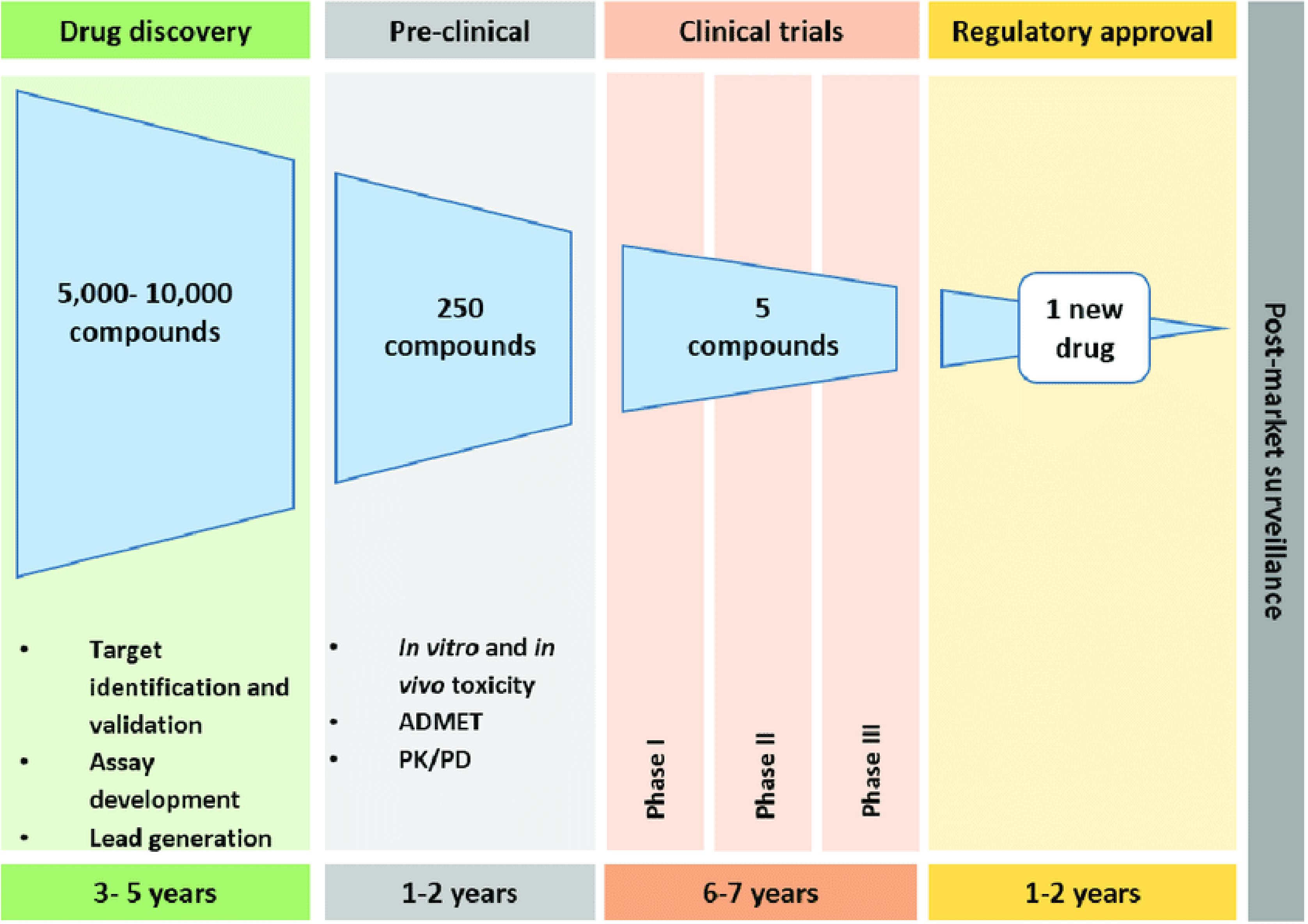
Vemurafenib on BRAF-V600 Melanoma

Cancer: Immunotherapy

Keytruda: immunotherapy blockbuster ($17B, 2021)

FDA approved for many cancer indications

But only work for minority of patients

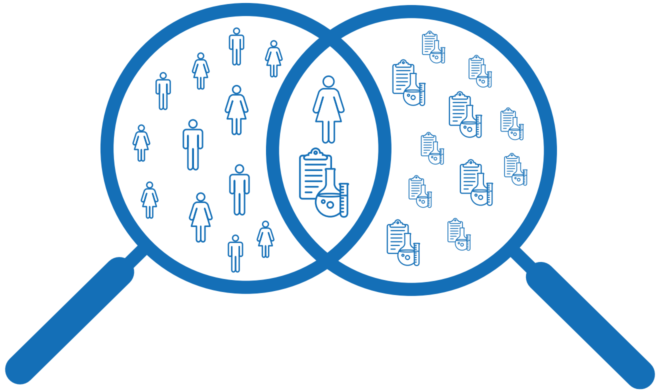


"Omics"-Informed Drug and Biomarker Discovery. Matthews et al. *Proteomes 2016*

Average Cost
$2-10B, 10+ years

Information Access Can Be Life or Death

Marty Tenenbaum
Late-stage melanoma (late 1990s)
Initial prognosis: 6 months
Saved by Phase III trial of Canvaxin



Insight Consumer

Pharma, Payor, Regulator

Data Producer

Provider, EHR Vendor

NAM Leadership Consortium for
Value & Science-Driven Health Care

NATIONAL ACADEMY OF MEDICINE

Leadership • Innovation • Impact | for a healthier future

The Learning Health System Series

Continuous improvement and innovation in health and health care

To facilitate progress toward the development of a ***learning health system***—in which science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience—the Leadership Consortium for Value & Science-Driven Health Care has marshaled the insights of the nation's leading experts to explore in detail the prospects, and the necessity, for transformational change in the fundamental elements of health and health care. The assessments are reported in the 15 volumes of the NAM Learning Health System Series, published by the National Academies Press.

VISION

Vision. The Learning Healthcare System, the first in the series, explores the various dimensions—evidence development and standards, care culture, system design and operation, health data, clinical research, information technology, value—on which emerging insights and scientific advances can be applied for health care in which both evidence development and application flow seamlessly and continuously in the course of care.

The Data Utility. Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good identifies the transformational prospects for large interoperable clinical and administrative datasets to allow real-time

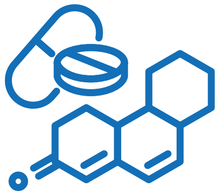
CARE COMPLEXITY

Care Complexity. Evidence-Based Medicine and the Changing Nature of Health Care explores the forces, such as genetic insights and increasing care complexity, driving the need for better medical evidence; the challenges with which patients and providers must contend; the need to transform the speed and reliability of new medical evidence; and the legislative and policy changes that could enable evolution of an evidence-based, learning system.

Effectiveness Research. Redesigning the Clinical Effectiveness Research Paradigm: Innovation and Practice-Based Approaches reviews the growing scope and scale of the need for clinical effectiveness research alternatives, the limits of

Insight Consumer

Pharma, Payor, Regulator



US: Less than 3% cancer patients enroll in trials

40% cancer trial failures due to insufficient patients

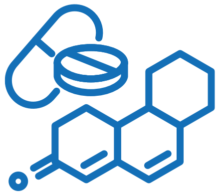
New drug costs $2-10 billion and takes 10+ years

Data Producer

Provider, EHR Vendor

Insight Consumer

Pharma, Payor, Regulator



Large language models → universal structuring

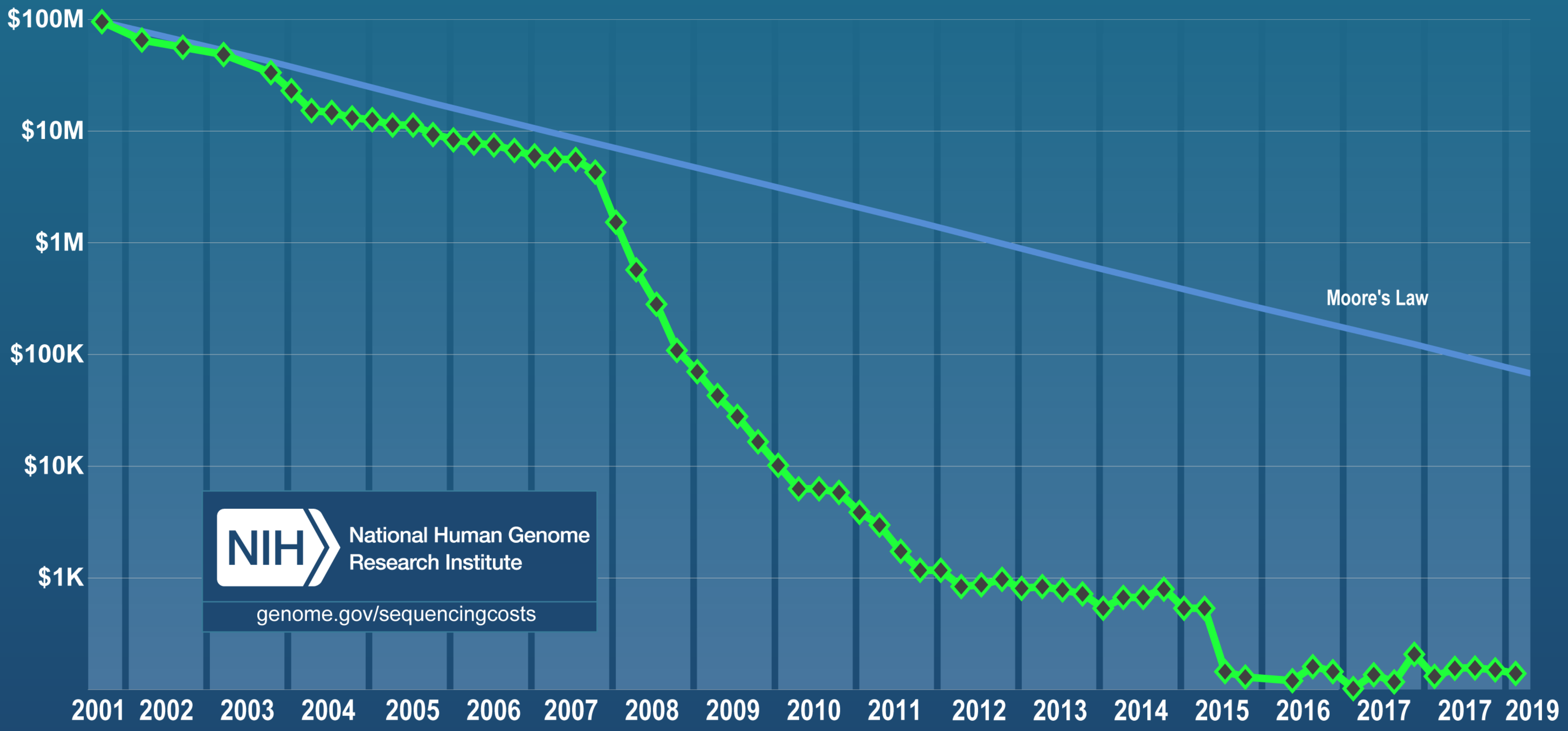
Instantly unlock top value chain

Data Producer

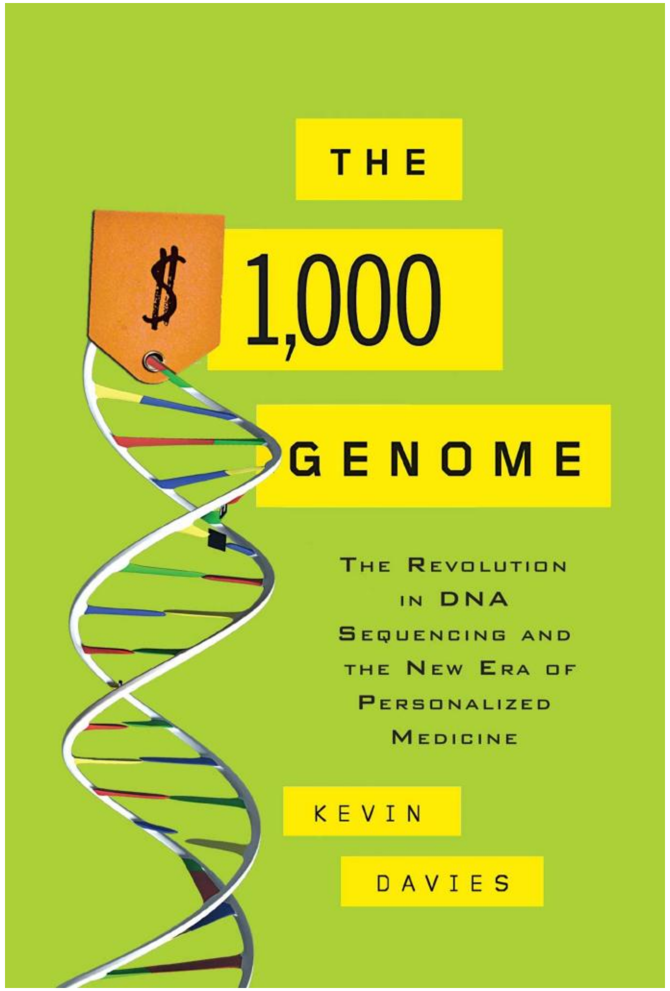
Provider, EHR Vendor

Digital Transformation → Intelligence Revolution

Cost per Genome



Digital Transformation

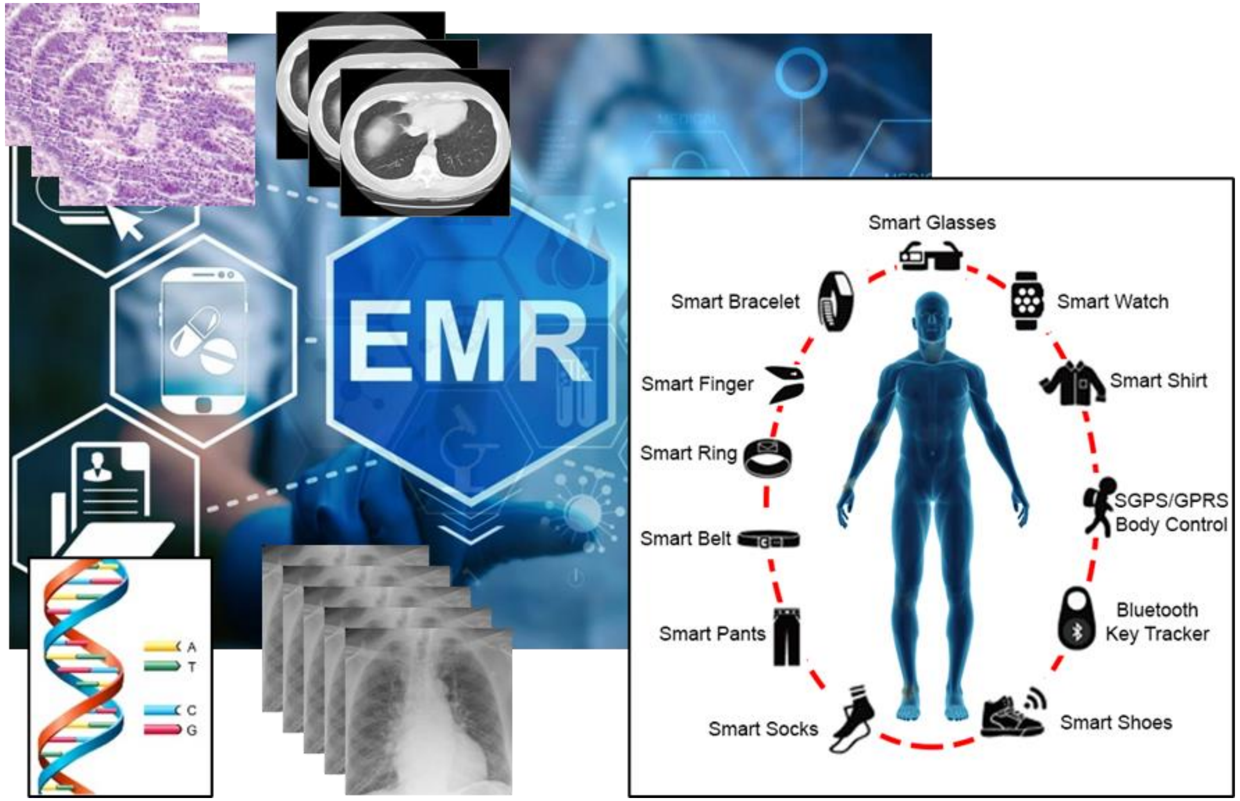


Accenture study: 93% of US doctors using EMRs

May 14, 2013 IHQRE informatics, IHQRE Journal Club EHR, EMR, Meaningful Use

2009 – 2013: 40% → 93%

Digital Transformation → Intelligence Revolution



What can LLMs do for precision health?

Access
Safety
Preventive Care

Real-World Evidence (RWE)

```
1,23224,174680,2147-12-05,,,"Discharge summary","Report",,"","Admissi
on Date:  [**2823-9-29**]              Discharge Date:   [**2823-10-1
7**]

Date of Birth:  [**2768-10-11**]             Sex:   F

Service: SURGERY

Allergies:
Patient recorded as having No Known Allergies to Drugs

Attending:[**First Name3 (LF) 1**]
Chief Complaint:
headache and neck stiffness

Major Surgical or Invasive Procedure:
central line placed, arterial line placed

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis
on 6-mercaptopurine, prednisone 40-60 mg daily, who presents
with a new onset of headache and neck stiffness. The patient is
in distress, rigoring and has aphasia and only limited history
is obtained. She reports that she was awaken 1AM the morning of
[**2823-9-28**] with a headache which she describes as bandlike.  She
states that headaches are unusual for her. She denies photo- or
phonophobia. She did have neck stiffness. On arrival to the ED
at 5:33PM, she was afebrile with a temp of 96.5, however she
later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR
24, O2 sat 100 %. Head CT was done and relealved attenuation
within the subcortical white matter of the right medial frontal
lobe. LP was performed showing opening pressure 24 cm H2O WBC of
316, Protein 152, glucose 16.  She was given Vancomycin 1 gm IV,
Ceftriaxone 2 gm IV, Acyclovir 800 mg IV, Ambesone 183 IV,
Ampicillin 2 gm IV q 4, Morphine 2-4 mg Q 4-6, Tylenol 1 gm ,
Decadron 10 mg IV.  The patient was evaluated by Neuro in the
ED.
```

Patient	Diagnosis	Treatment	Outcome
101	Lung Cancer	Gefitinib	remission
202	Leukemia	Imatinib	resistant
303	Lymphoma	Zaraparib	relapse
	……		

Population-level “free lunch”

Drug Discovery	Clinical Trial	Post-Market
Target Identification	Eligibility	Adverse Event
Drug Repurposing	Synthetic Control	Comparative Effectiveness
	Virtual Trial	Off-Label Use
	Pragmatic Trial	

Real-World Evidence

Trillion-dollar opportunity:
Accelerate development; reduce cost; save lives

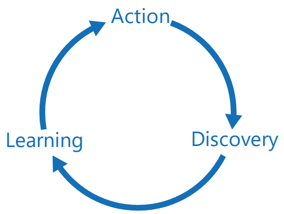
Digital Transformation → Intelligence Revolution

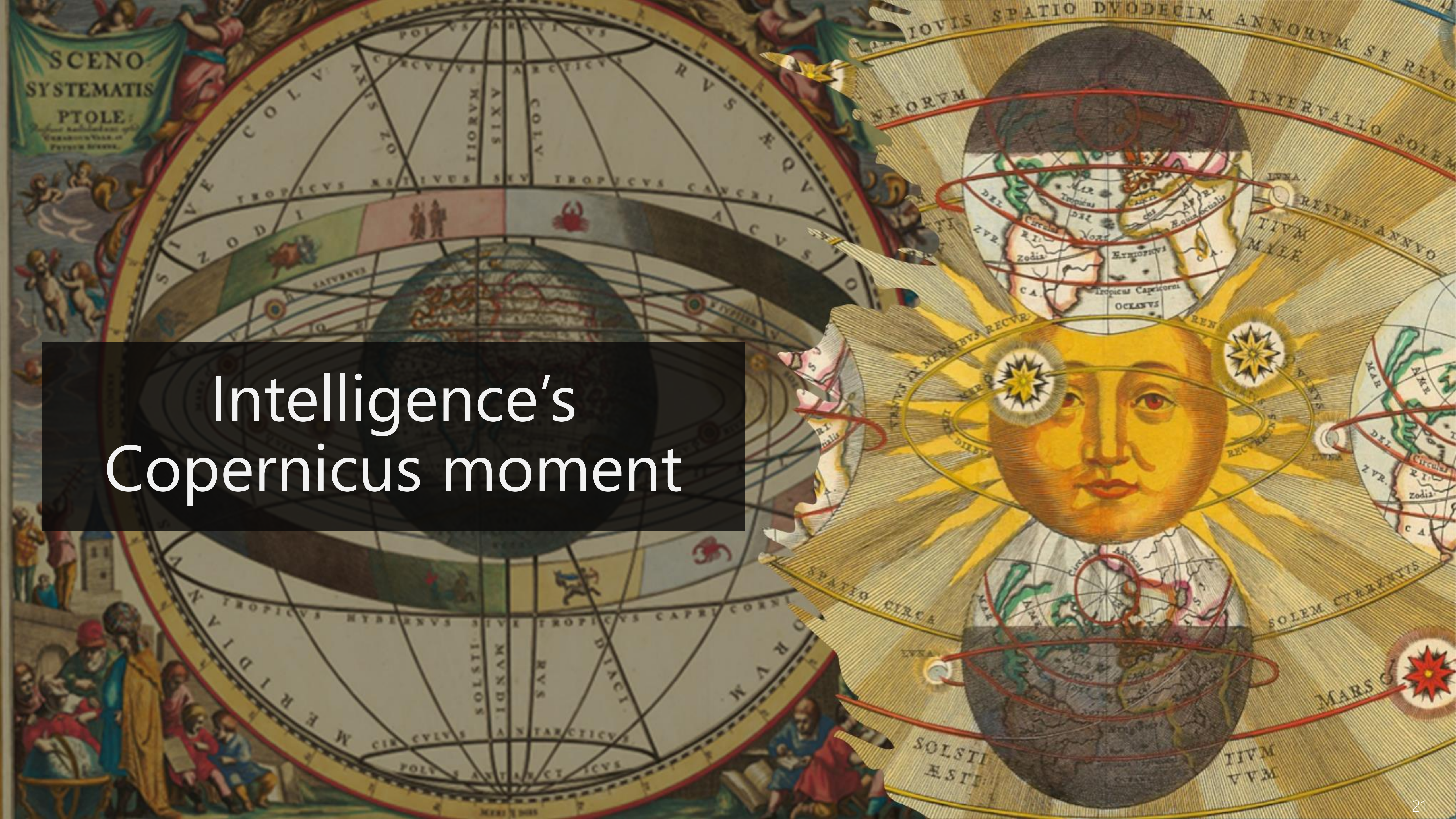
Copilots for Precision Health

Population-level Causal Discovery

Multi-Modal Learning & Fusion

LLM: Universal Structuring





Intelligence's Copernicus moment

Large Language Models → New Patterns

Universal Structuring → Scale real-world evidence

Universal Translator → Rethink interoperability

Universal Labeler → Scale benchmark / evaluation

Universal Reasoning → "Talk to data" and make sense

Case Study: Immunotherapy

Keytruda: immunotherapy blockbuster ($17B, 2021)

FDA approved for many cancer indications

But only work for minority of patients. Why?

Advancing Health at the Speed of AI

Insight Consumer
Pharma, Payor, Regulator

Real-World Evidence Marketplace

Data Producer
Provider, EHR Vendor

Discovery Engine

AI

Data Store

Overview

Precision health

Intelligence revolution

Biomedical LLMs

Application challenges

Research frontiers

A brief history of NLP

Big Bang | GOFAI | Statistical Revolution | Deep Learning

Computer, AI, NLP

Turing Test, 1950

AI Birth (Dartmouth, Hanover NH), 1956

Chomsky ("Syntactic Structures"), 1957

Machine Translation

Cold war: Russian to English

Demo: IBM-Georgetown, 1954

Crash: ALPAC Report, 1966

Lesson: Pretty demo not enough

Need rigorous evaluation & benchmarks

1940-60 | 1970-80 | 1990-2010 | 2010-Present

Big Bang

GOFAI

Statistical Revolution

Deep Learning

Rule-base

Lexicon

RegEx

Semantic Grammar

Dialog, Question-Answering

Eliza, 1964

BASEBALL (Green et al.), 1961

SHRDLU (Winograd et al.), 1973

LUNAR (Wood et al.), 1978

Still used in most "clinical NLP" and "biomedical NLP" today

Negation Detection
Hedge Detection
Ontology-Based Entity Linking
……

1940-60 | 1970-80 | 1990-2010 | 2010-Present

Big Bang | GOFAI | Statistical Revolution | Deep Learning

Statistical Machine Learning

Classification: Decision tree, Random Forest, Naïve Bayes, SVM, kernel methods, log-linear models, …

Structured Prediction: Dynamic Programming, HMM, CRF, probabilistic logic, …

Morphology, Syntactic Parsing, Named Entity Recognition (NER), Information Extraction, Question Answering, Machine Translation, …

Penn Treebank, 1990s

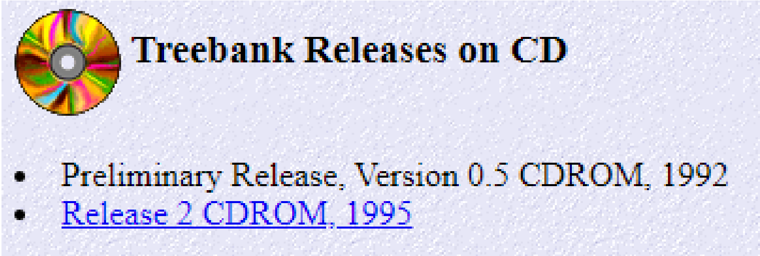
ACE, 2003

PropBank, 2005

……

Newswire / Web

Most on component tasks



1940-60 | 1970-80 | 1990-2010 | 2010-Present

Big Bang | GOFAI | Statistical Revolution | Deep Learning

Then: "NLP is all about feature engineering"

Sentence breaking
Tokenization
Morphology

→

POS Tagging
Syntactic Parsing

→

NER
Coreference
Entity Linking
Semantic Parsing
Discourse Parsing

→

Information Extraction
Question Answering
Machine Translation
......

Component Tasks

Applications

1940-60 | 1970-80 | 1990-2010 | 2010-Present

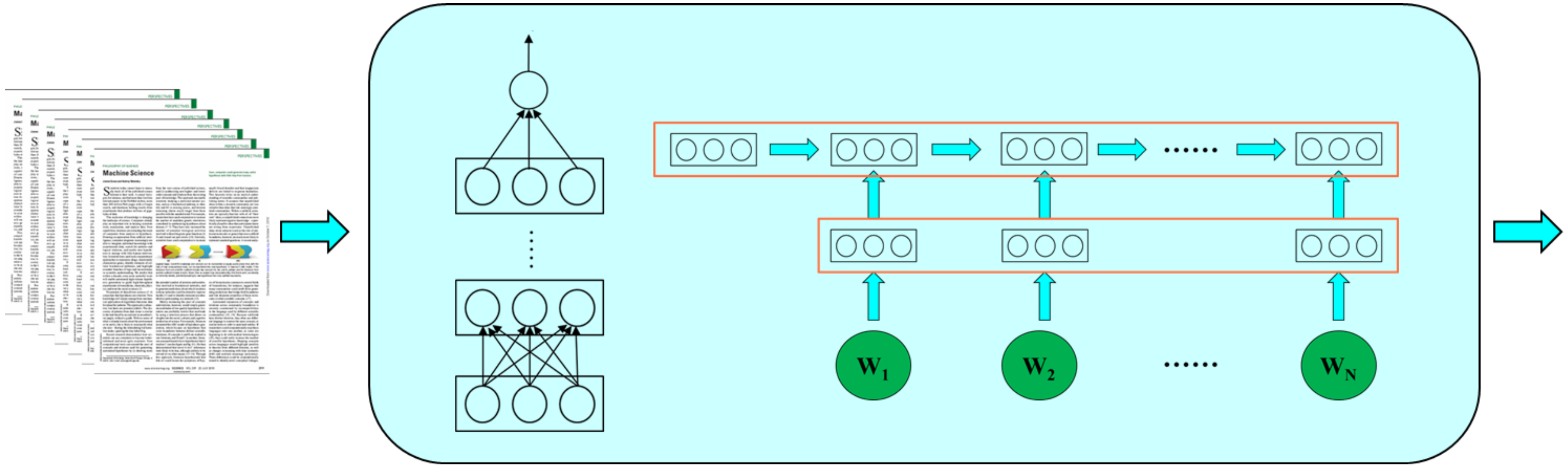
Big Bang

GOFAI

Statistical Revolution

Deep Learning

Now: End-to-end deep learning



Applications
Information Extraction
Question Answering
Machine Translation
……

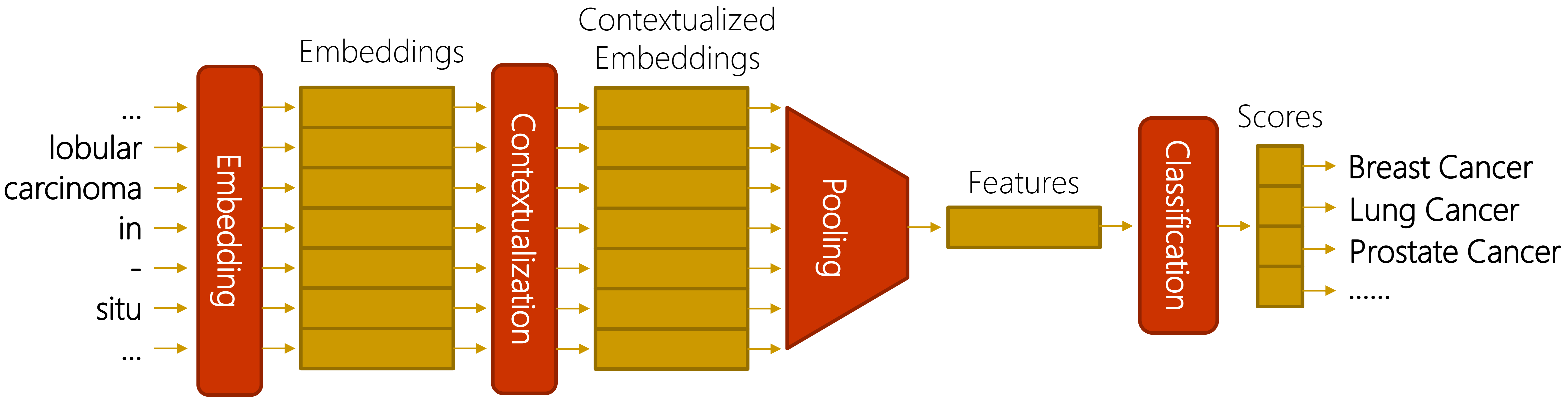
1940-60

1970-80

1990-2010

2010-Present

End-to-End Deep Learning



GloVe
Word2Vec

CNN
RNN

Average
Max
Attention

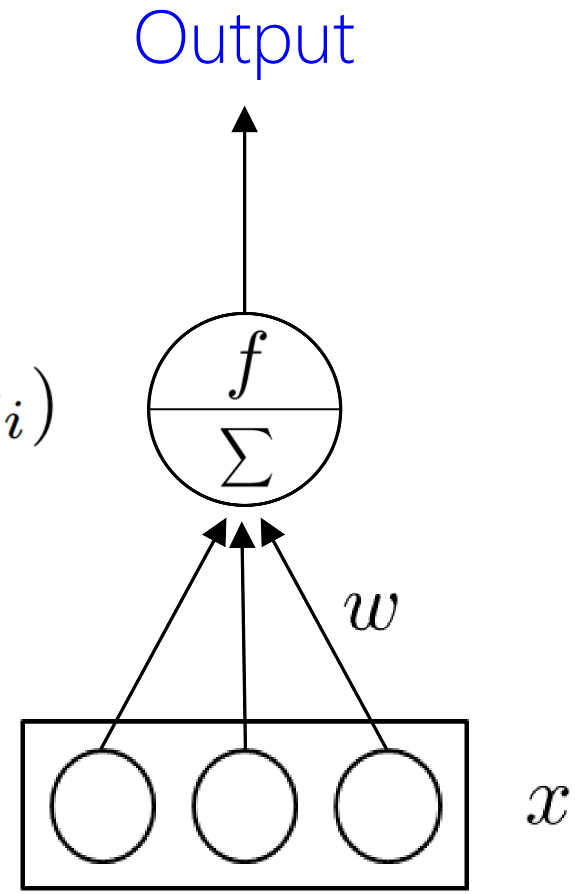
Linear
MLP
Logistic Regression
……

ELMo, BERT, ELECTRA, …

A brief history of deep learning

Neural Unit

$$O = f(\sum_{i=1}^{n} w_i \cdot x_i)$$

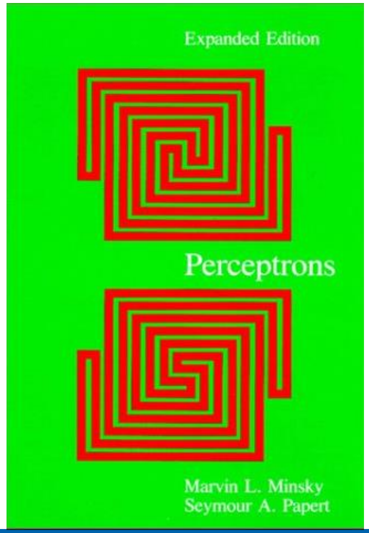


Input Layer

Source: Arvin Calspan Advanced Technology Center; Hecht-Nielsen, R. Neurocomputing (Reading, Mass.: Addison-Wesley, 1990)

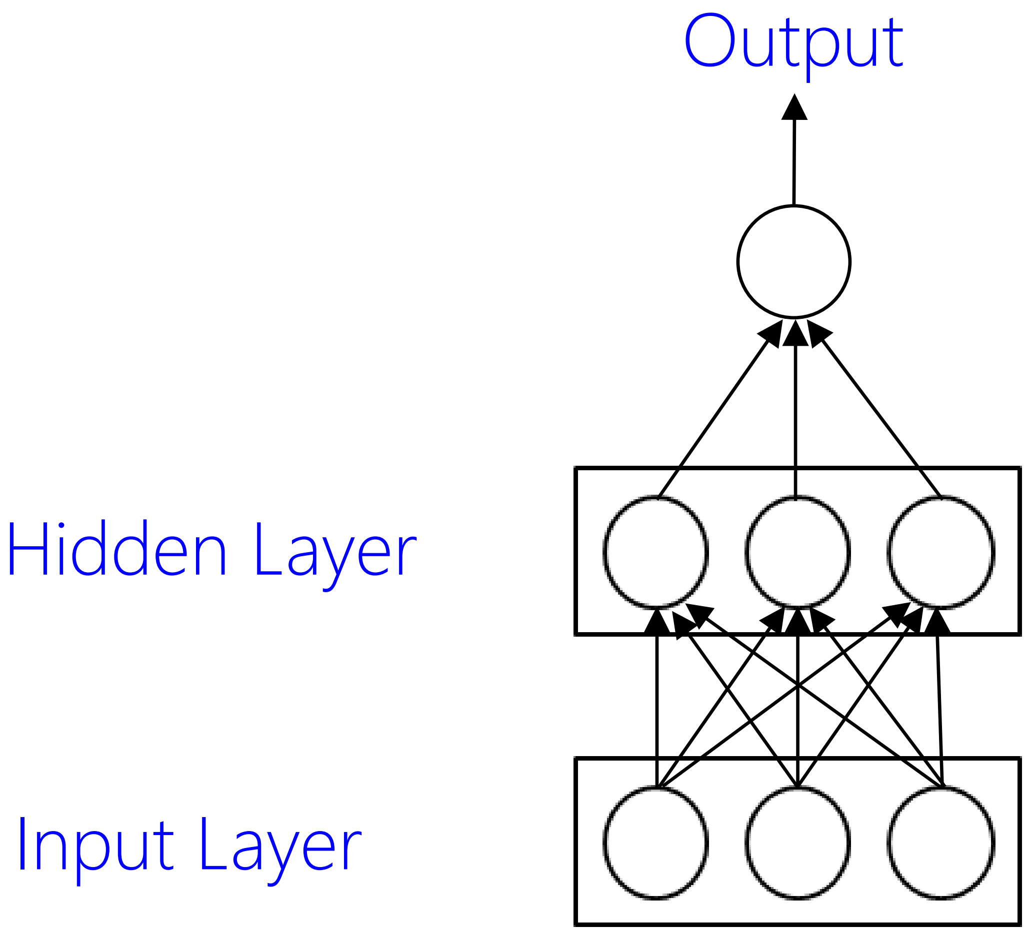
First Wave

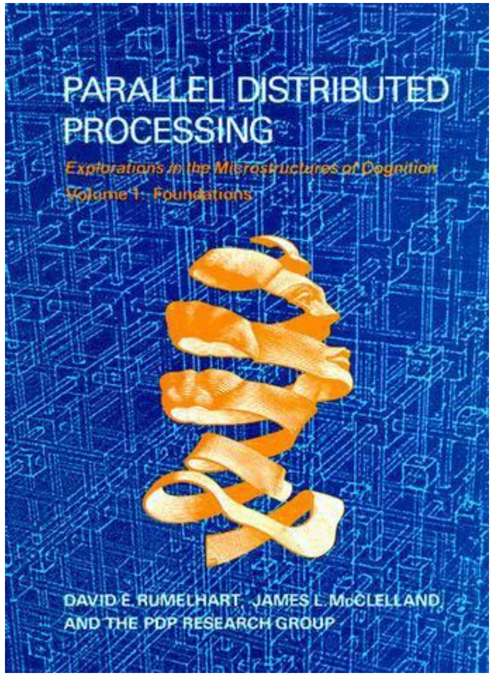
Perceptron
[Rosenblatt, 1957]



Can not represent complex functions such as XOR
[Minsky & Papert, 1969]

Neural Network





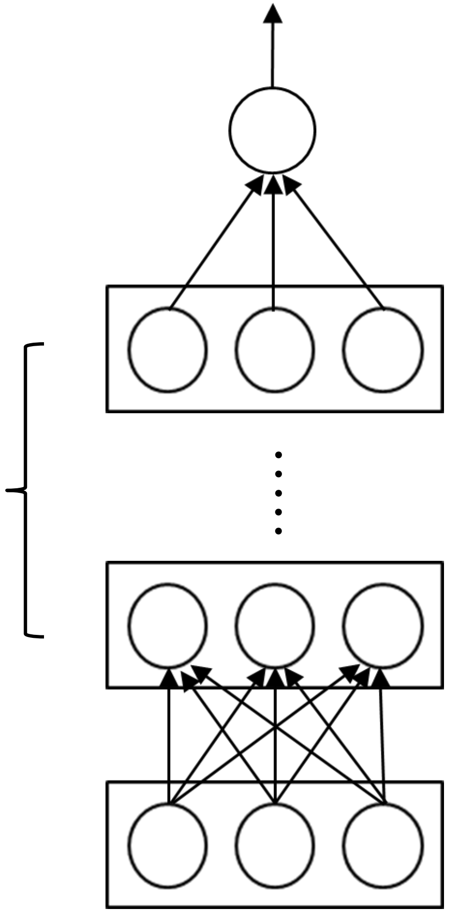
Second Wave

Backpropagation
[Rummelhart, Hinton, Williams, 1986]

Gradient diffusion or explosion:
Can not learn more than a few layers

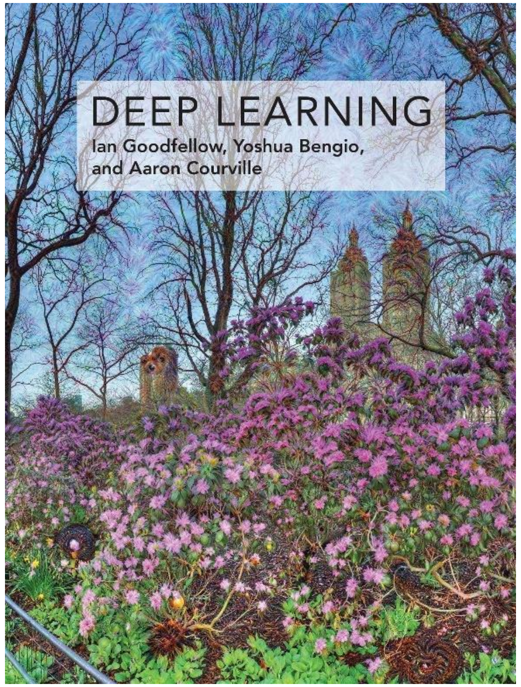
Deep Learning

Output



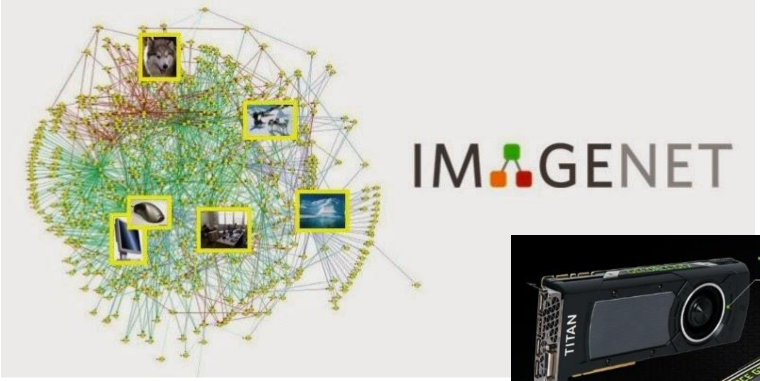
Many
Hidden Layers

Input Layer



Third Wave

SGD, ReLU, dropout, …
[Hinton, LeCun, Bengio,
Schmidhuber, Hochreiter, …]



Big labeled data

Fast computation

The Great Consolidation in AI

Transformer

Modality

Self-supervised learning

Prompt: Instruction following

Transformer

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

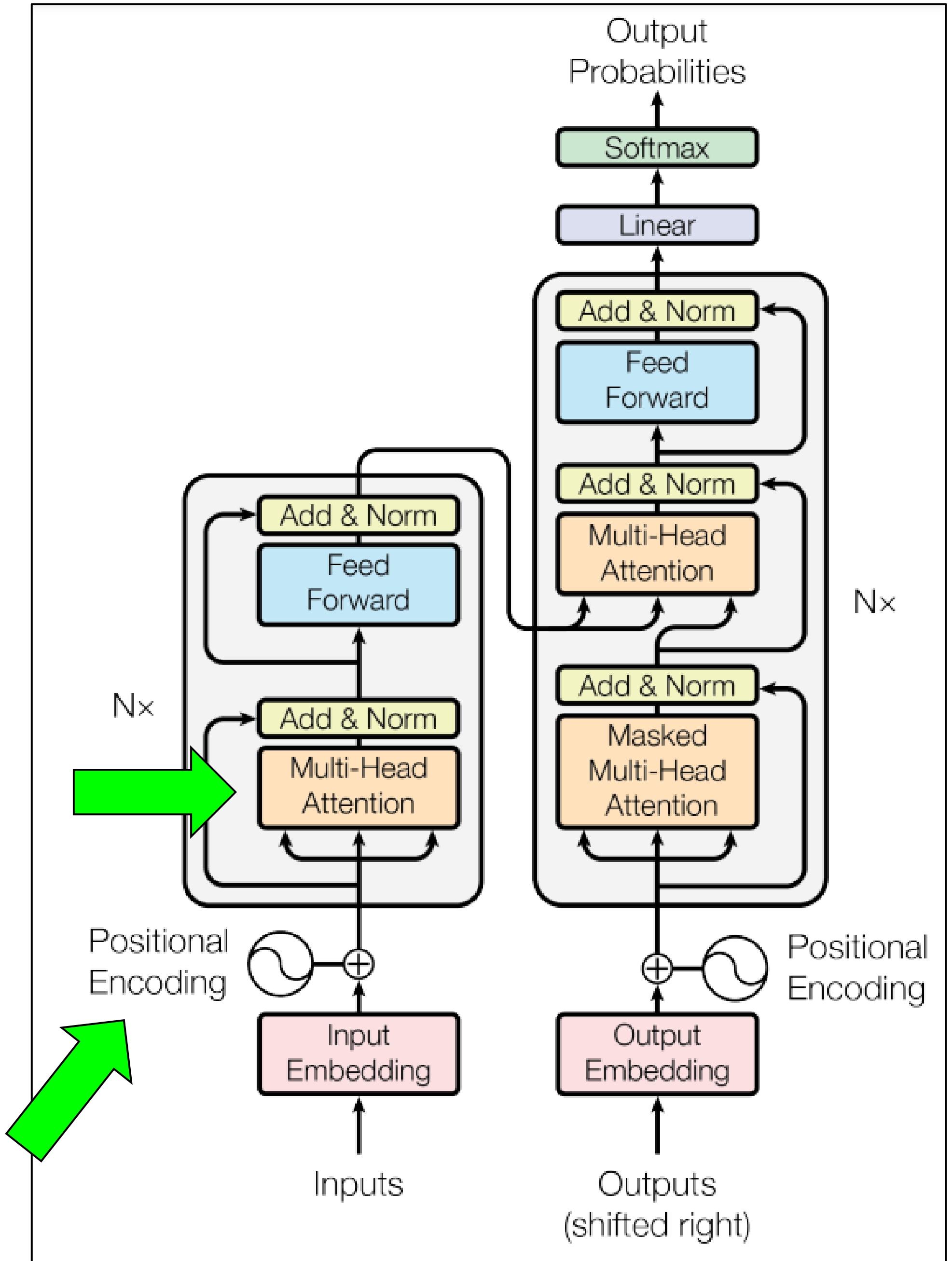


Figure 1: The Transformer - model architecture.

Self-Attention

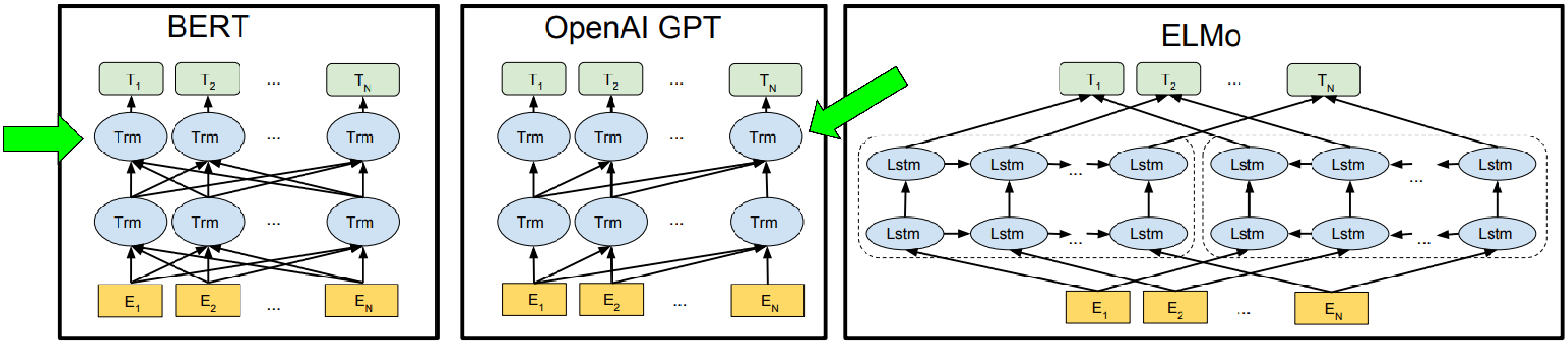
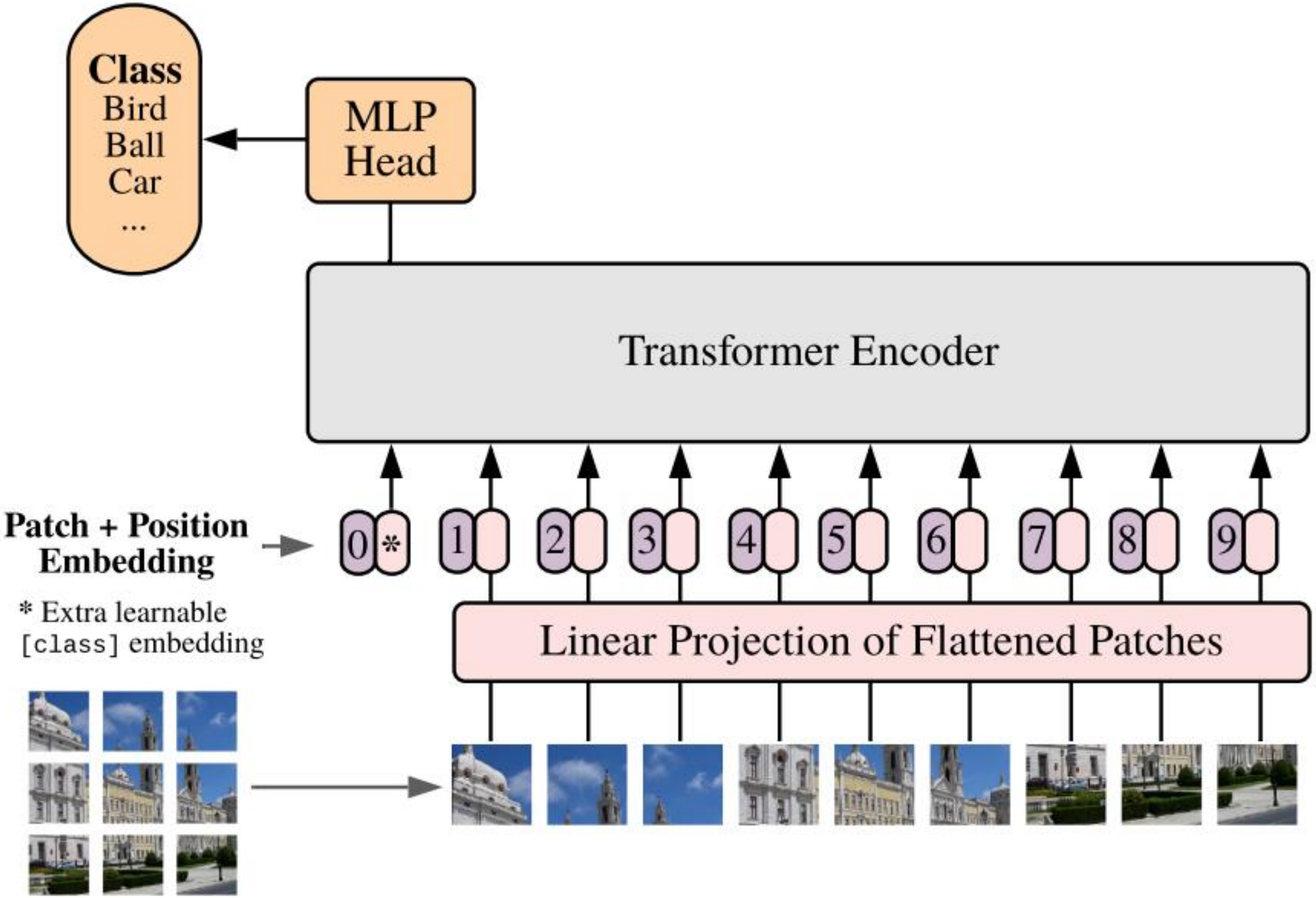


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. *NAACL 2019*

Vision Transformer (ViT)



Dosovitskiy, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, *ICLR 2021*.

Molecular Transformer

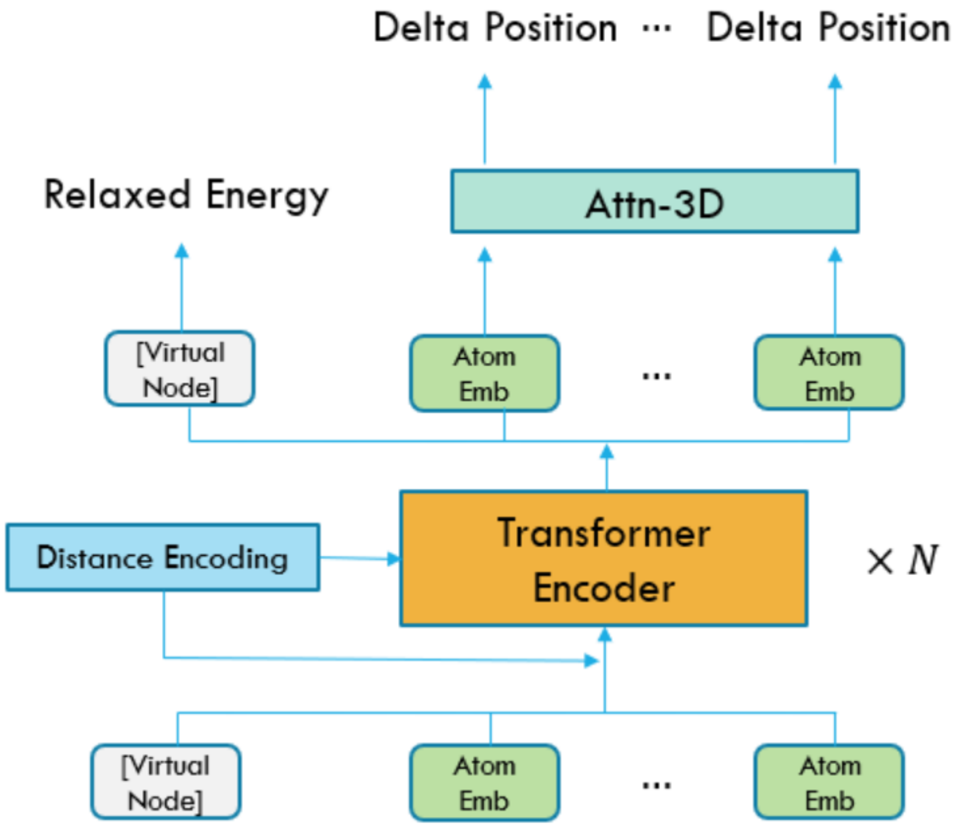
Microsoft Research Lab - Asia / Articles

Graphormer wins the Open Catalyst Challenge and upgrades to AI for Molecular Simulation Toolkit

January 11, 2022

Share this page

Graphormer



The winner architecture of Graphormer in the Open Catalyst Challenge

Ying, et al. "Do Transformers Really Perform Bad for Graph Representation?", *NeurIPS 2021*.

Supervised Learning

Public

Domain Knowledge

Bottleneck

Labeled Data

Abundant

Unlabeled Data

ML Algorithm

Application Model

(save lives, cut cost, revolutionize medicine, …)

General vs Health Labeled Data

IMPRESSION

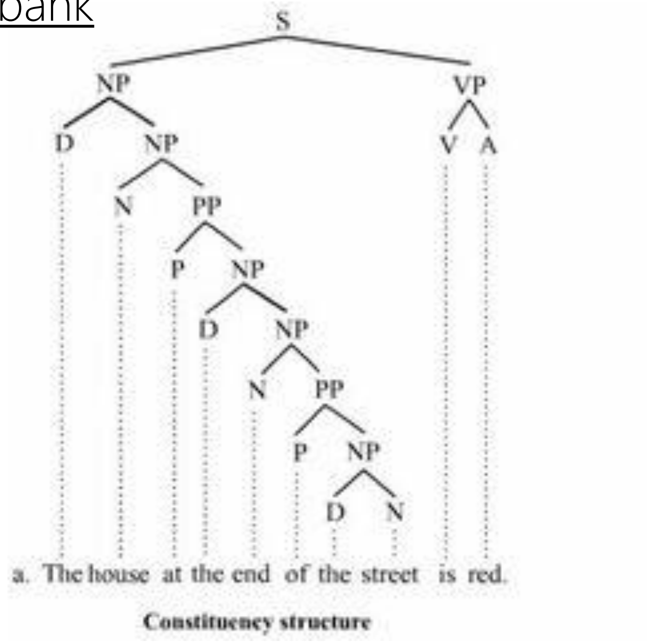
No significant change in right middle and low lobe pneumonia. Small increase in left pleural effusion. ……

Two cows are grazing in the field.

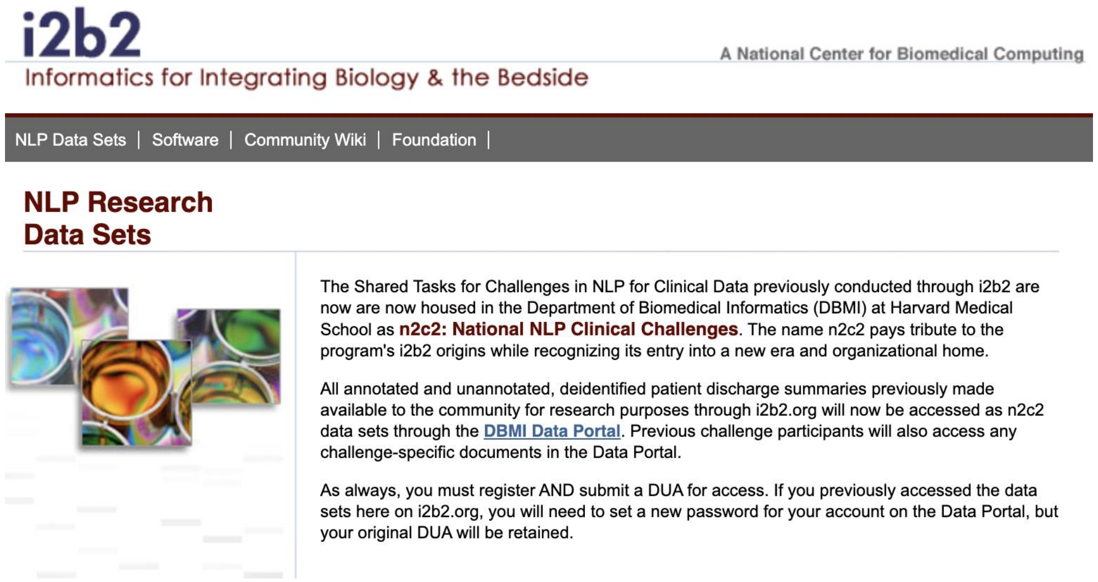
Biomedical and clinical domain label require expertise

General vs Health Data Availability

Penn Treebank



1992



2006

Comparable datasets over a decade later

Self-Supervised Learning

Public
Domain Knowledge

If Available
Labeled Data

Abundant
Unlabeled Data

Task-Specific

Task-Agnostic

ML Algorithm

Application Model
(save lives, cut cost, revolutionize medicine, …)

Neural Language Model Pretraining

The 2 mutations that were only found in the neuroblastoma resistance screen (G1123S/D) are located in the glycine-rich loop, which is known to be crucial for ATP and ligand binding and are the first mutations described that induce resistance to TAE684, but not to PF02341066

Unlabeled text

Neural Language Model Pretraining

The 2 mutations that were only found in the [MASK] resistance screen (G1123S/D) are [MASK] in the glycine-rich loop, which is known to be [MASK] for ATP and ligand [MASK] and are the first mutations described that induce resistance to TAE684, but not to [MASK]

Masked Language Model

Neural Language Model Pretraining

The 2 mutations that were only found in the ?

GPT: next-word prediction

Neural Language Model Pretraining

The 2 mutations that were only found in the neuroblastoma ?

GPT: next-word prediction

GPT-3: Prompt

All tasks → Text-to-text

Prompt engineering

"Generalist AI"

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

Beyond Next-Word Prediction

Supervised instruction fine-tuning

Reinforcement learning from human feedback

Training language models to follow instructions with human feedback

Long Ouyang* **Jeff Wu*** **Xu Jiang*** **Diogo Almeida*** **Carroll L. Wainwright***

Pamela Mishkin* **Chong Zhang** **Sandhini Agarwal** **Katarina Slama** **Alex Ray**

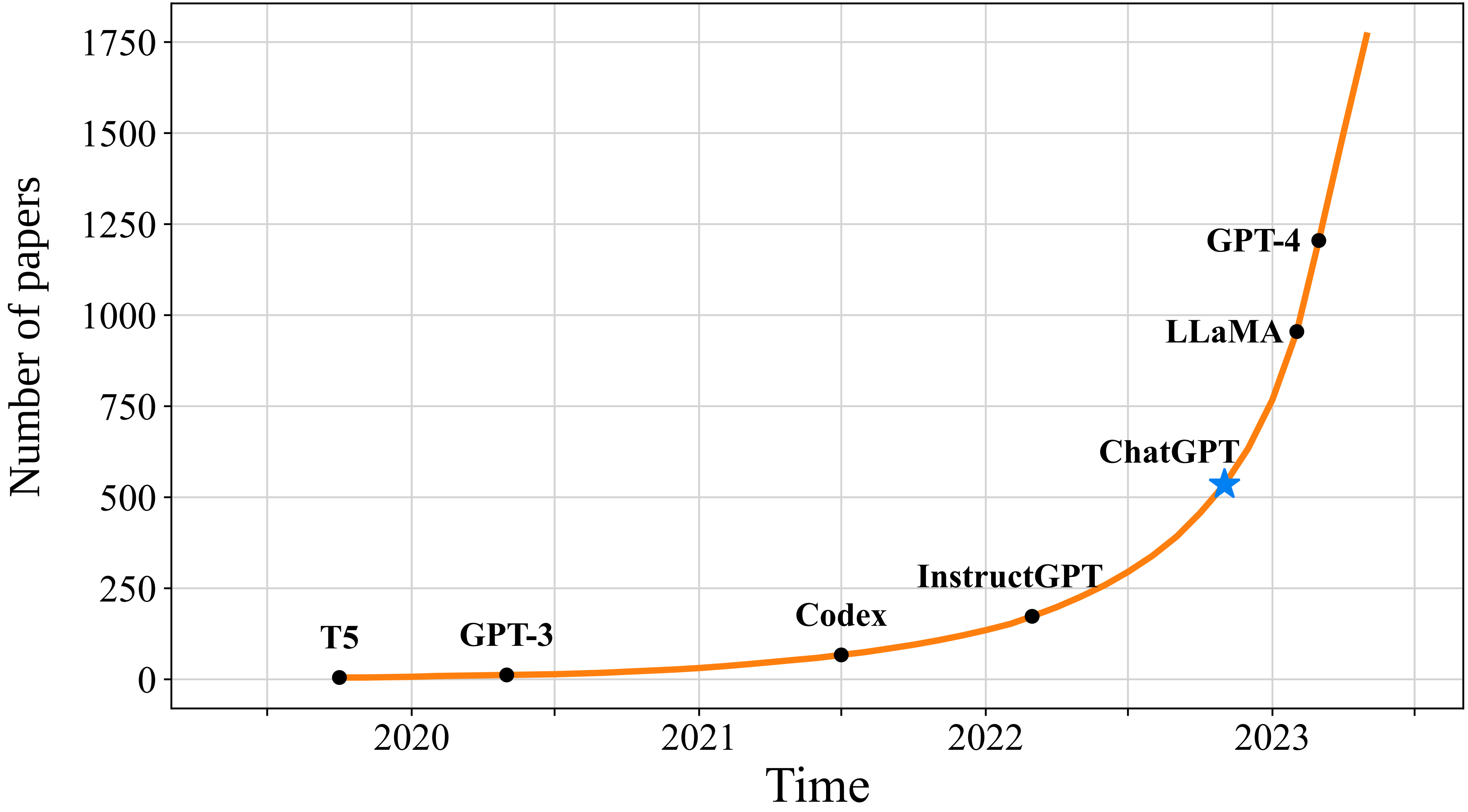
John Schulman **Jacob Hilton** **Fraser Kelton** **Luke Miller** **Maddie Simens**

Amanda Askell† **Peter Welinder** **Paul Christiano*†**

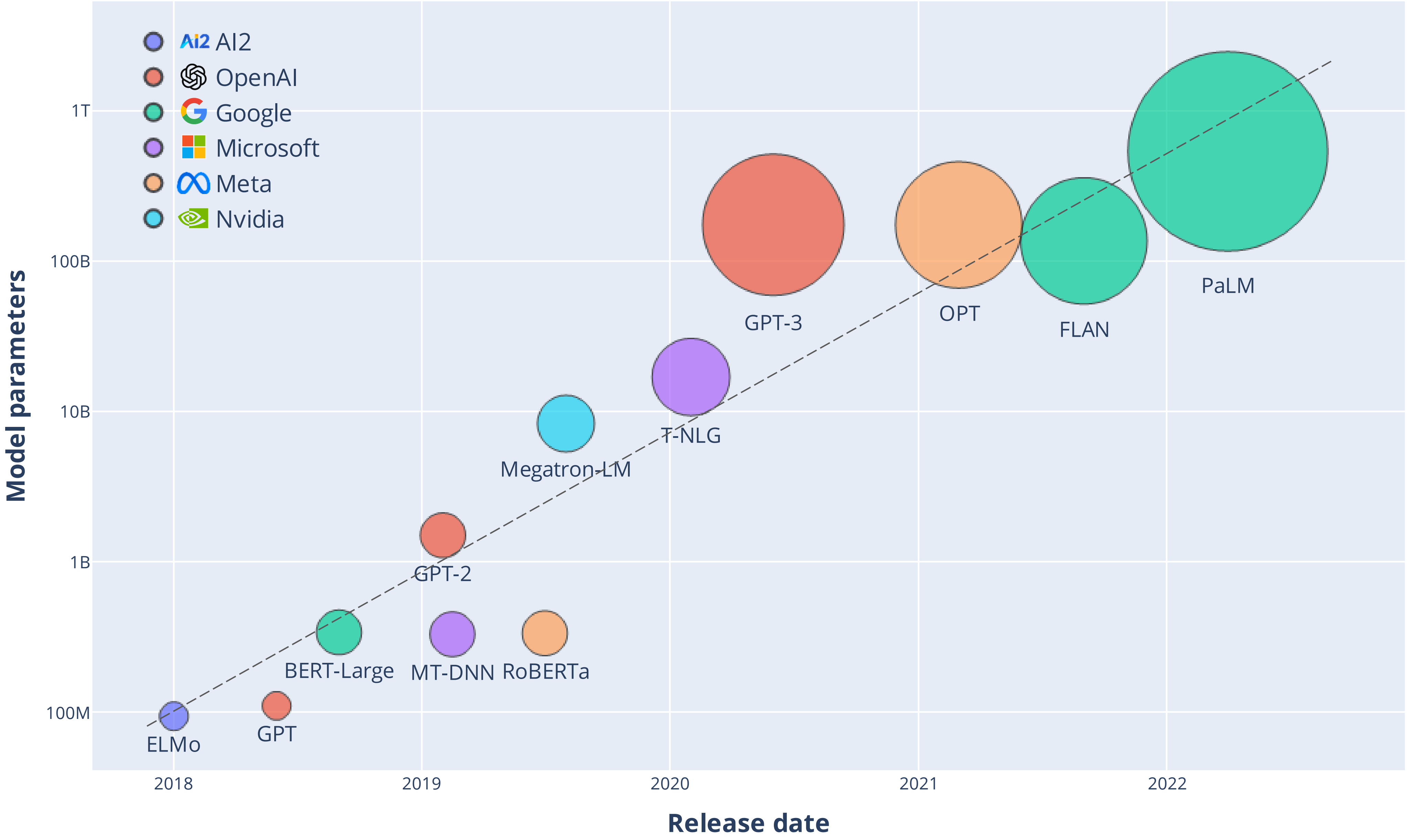
Jan Leike* **Ryan Lowe***

InstructGPT

New arXiv Papers mentioning “LLMs”



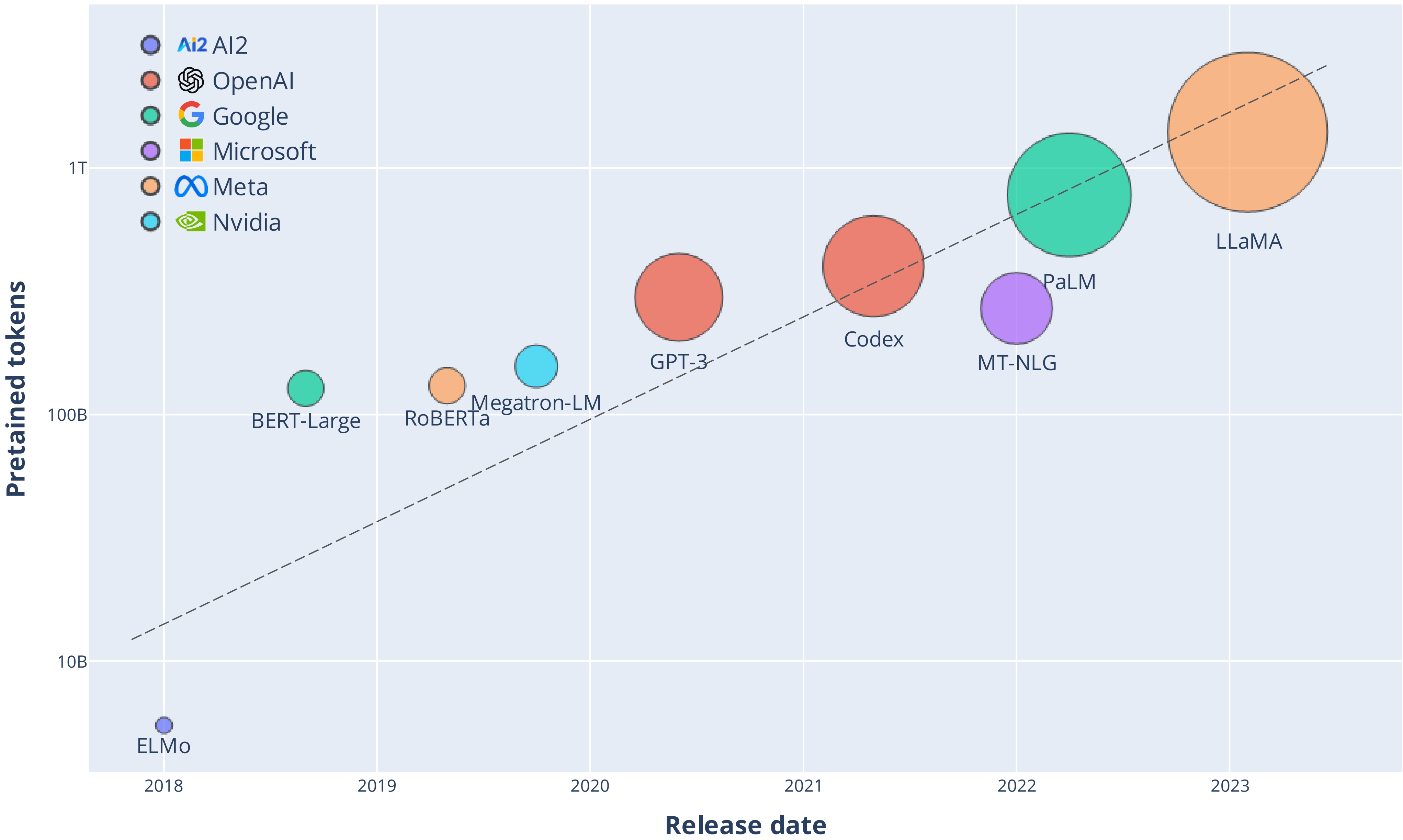
Growth of Model Size (100M → 1T+)

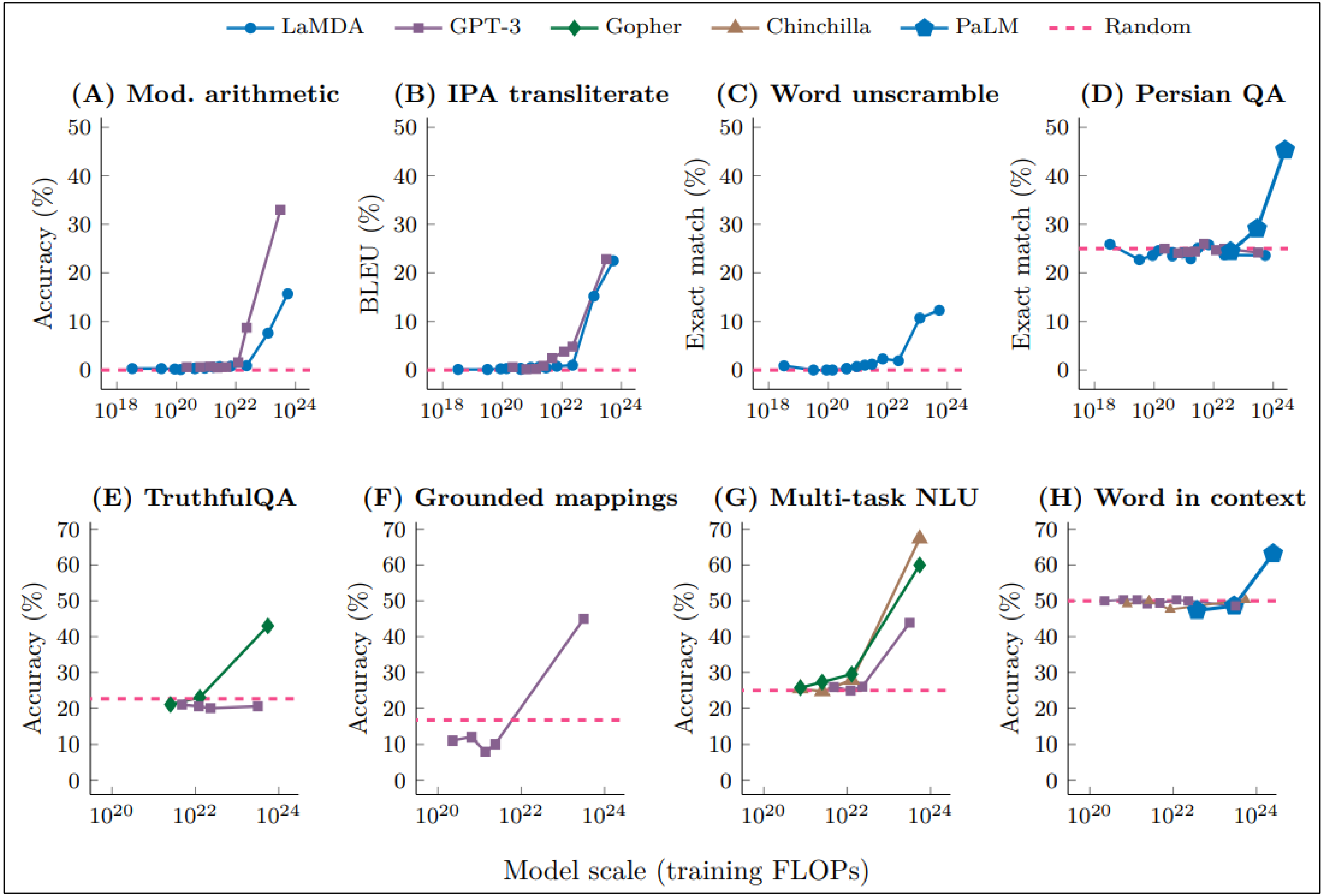


GPT-4?

Growth of Data (5B → 1T)

GPT-4?





Wei, et al. "Emergent Abilities of Large Language Models", *TMLR 2022*.

Effects of Scale

350M

750M

3B

20B

A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

https://parti.research.google/

Effects of Scale

350M

750M

3B

20B

A photo of an astronaut riding a horse in the forest. There is a river in front of them with water lilies.

https://parti.research.google/

Effects of Scale

350M

750M

3B

20B

A map of the United States made out of sushi. It is on a table next to a glass of red wine.

https://parti.research.google/

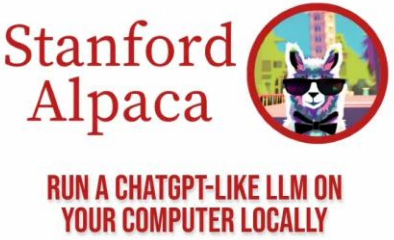
Open-Source LLM

Web-based pretraining

- OPT: 125M – 175B; 180B tokens
- LLaMA, LLaMA2: 7-70B; 1-1.4T tokens
- Falcon: 40B; 1T tokens
- Red Pajama: Replicate LLaMA training (1.2T tokens)

LLaMA + GPT-derived instruction-following data

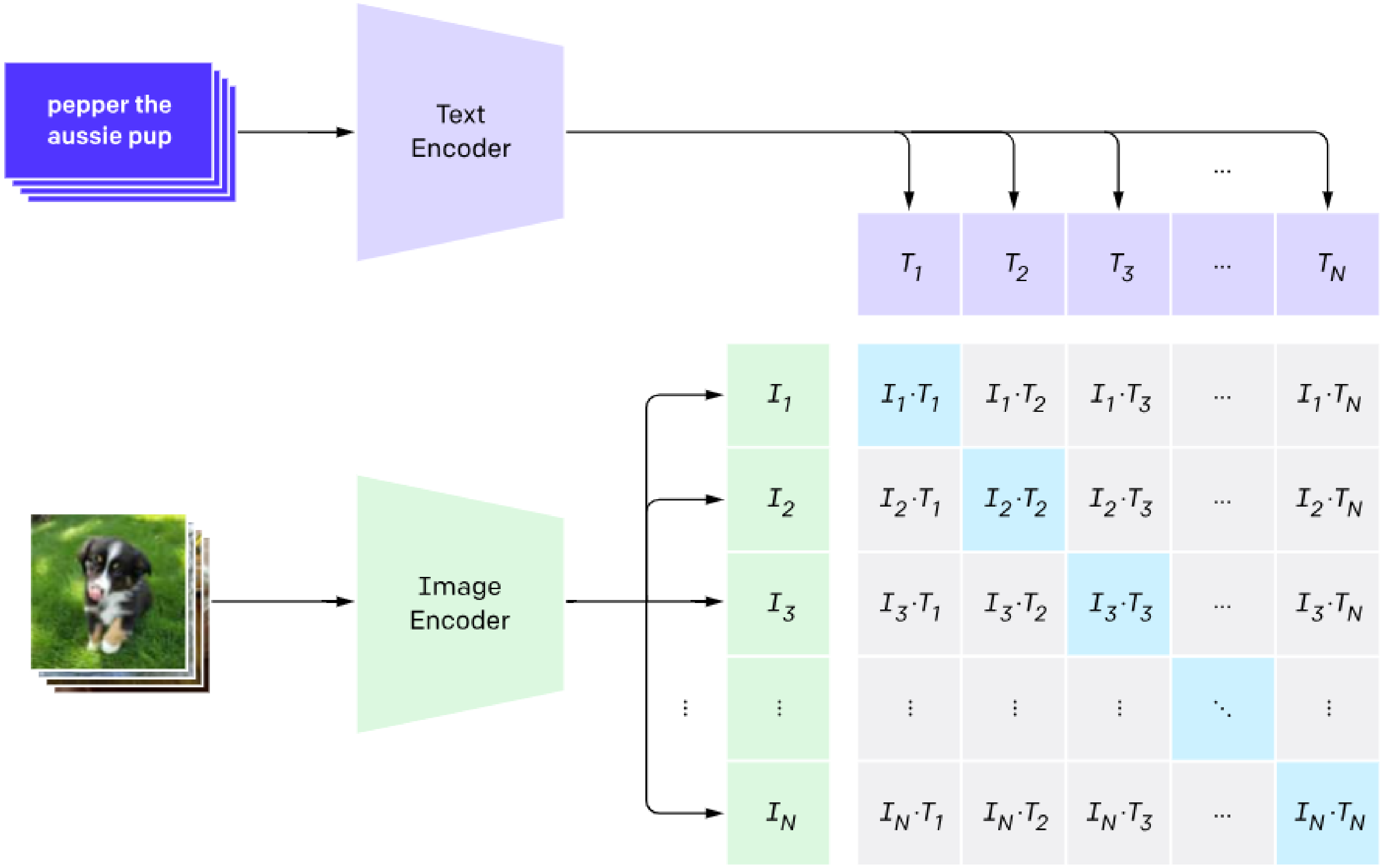
- Alpaca (7B): 52K GPT-3.5
- Vicuna (13B): 70K ChatGPT (from ShareGPT)



……

Large Multimodal Models (LMMs)

Radford, et al. "Learning Transferable Visual Models From Natural Language Supervision", *arxiv 2021*.

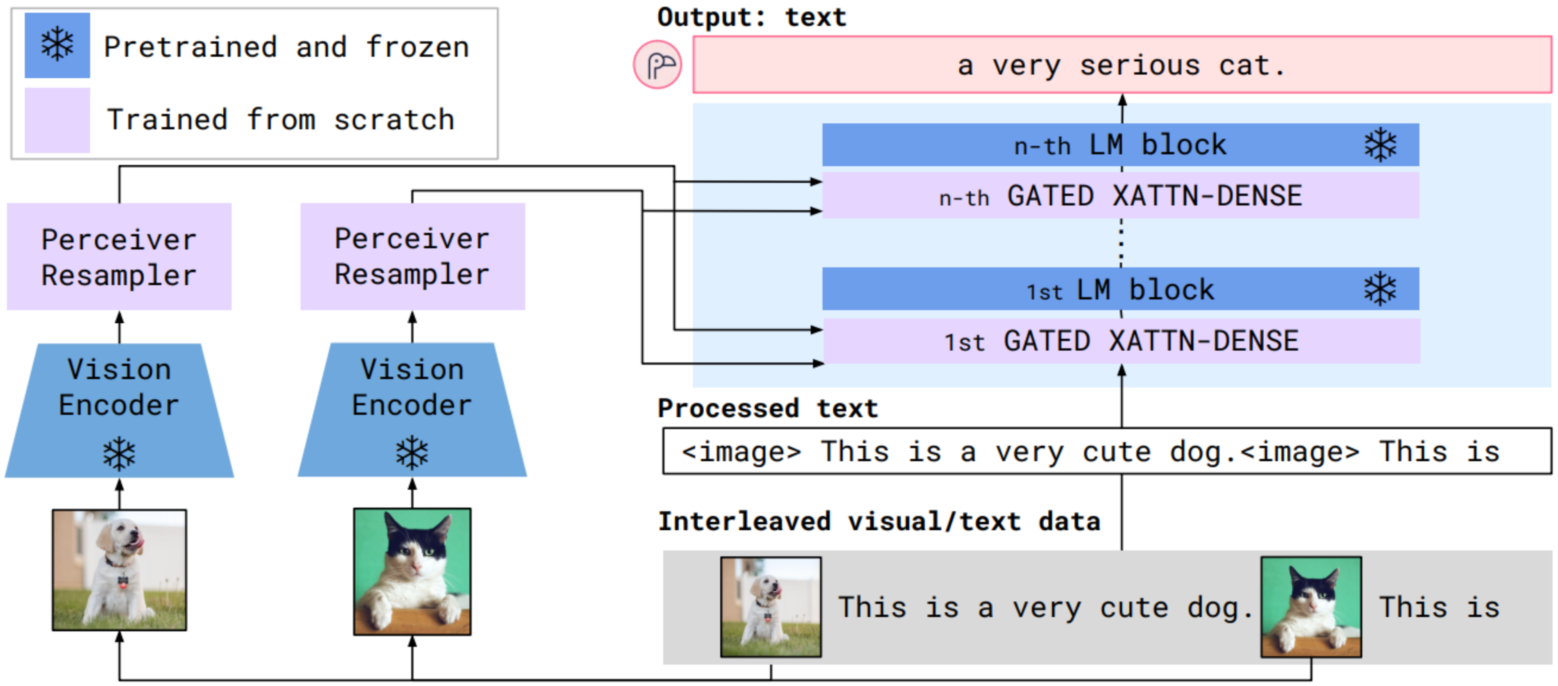


CLIP

Contrastive learning

Large Multimodal Models (LMMs)

Alayrac, et al. “Flamingo: a Visual Language Model for Few-Shot Learning”, *NeurIPS 2022*.



FLAMINGO

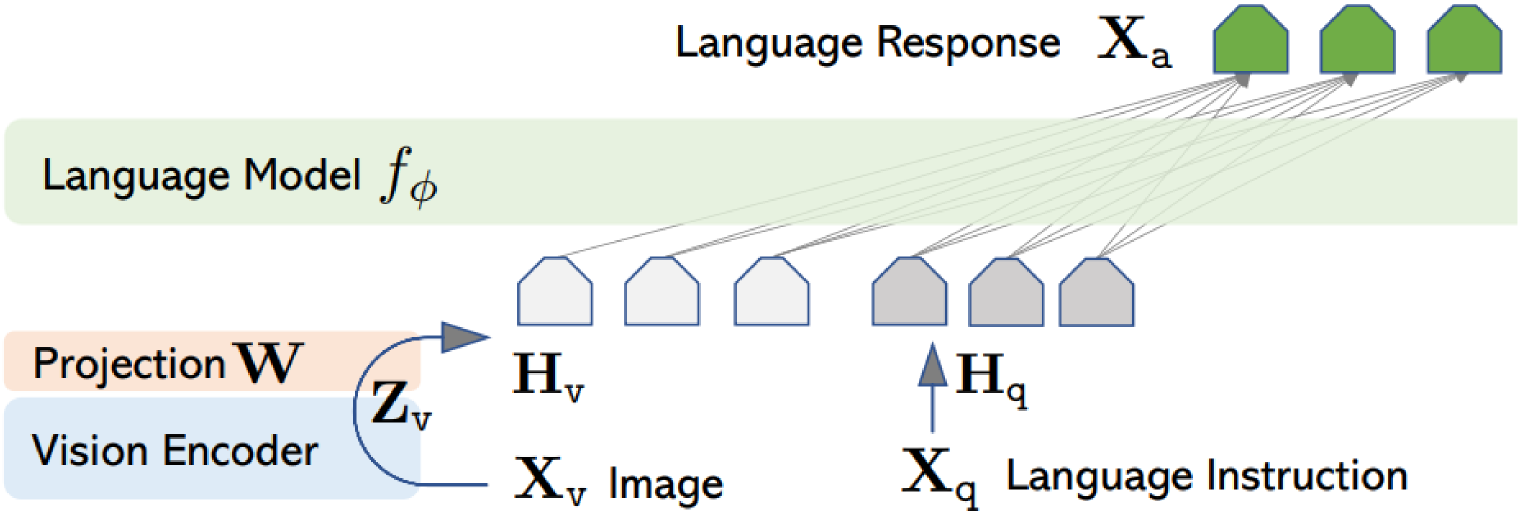
Frozen vision encoder / LM
Layer-wise gated adapter

Trained on web data: M3M, ALIGN, LTIP, VTP

Large Multimodal Models (LMMs)

Liu, et al. "Visual Instruction Tuning", *arxiv 2023*.

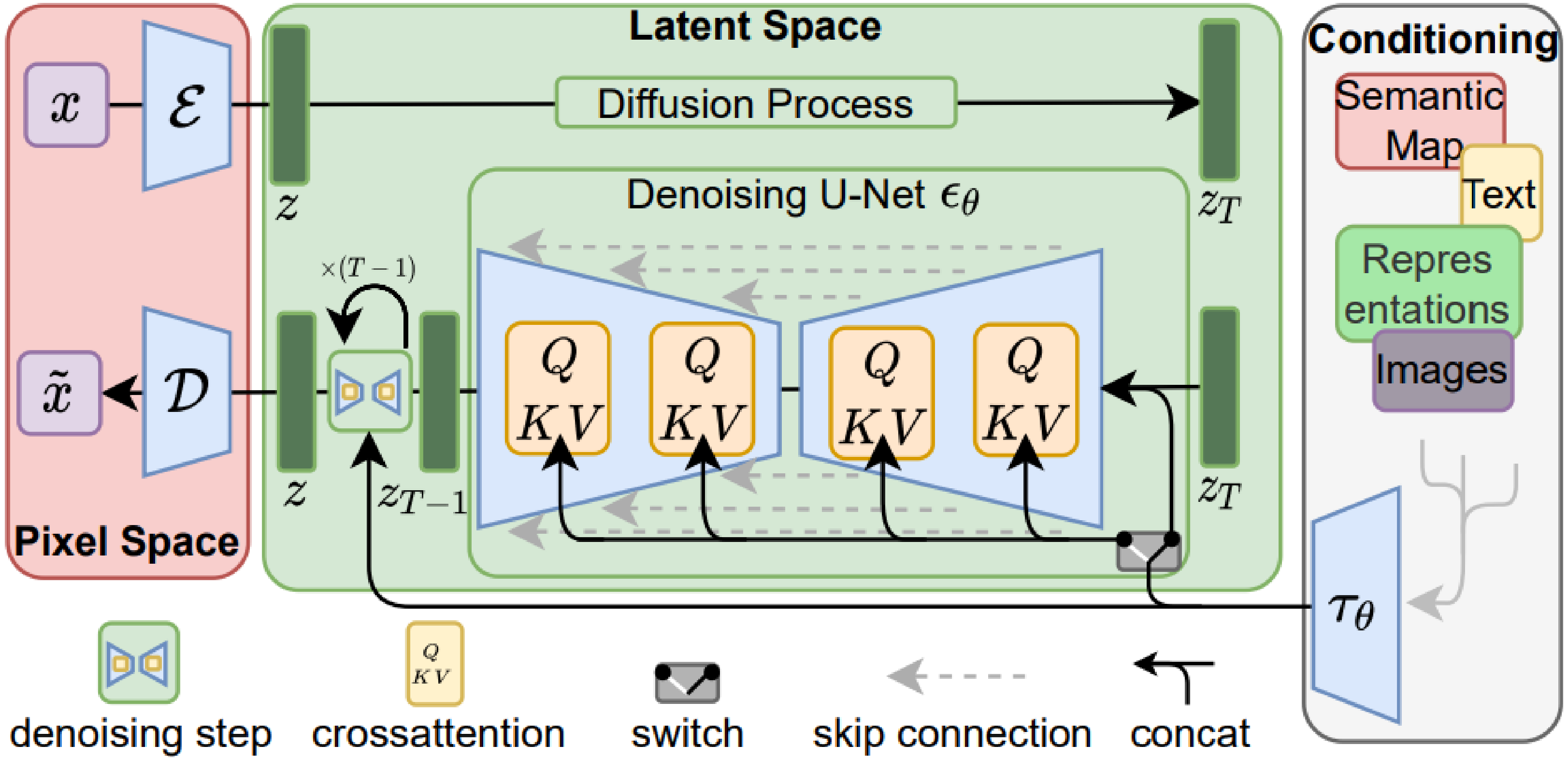
LLaVA



Key: use GPT-4 to generate multi-turn conversation for instruction tuning

Large Multimodal Models (LMMs)

Rombach, et al. "High-Resolution Image Synthesis with Latent Diffusion Models", *CVPR 2022*.



Latent Diffusion

Image generation: apply diffusion process on compressed latent space

Overview

Precision health

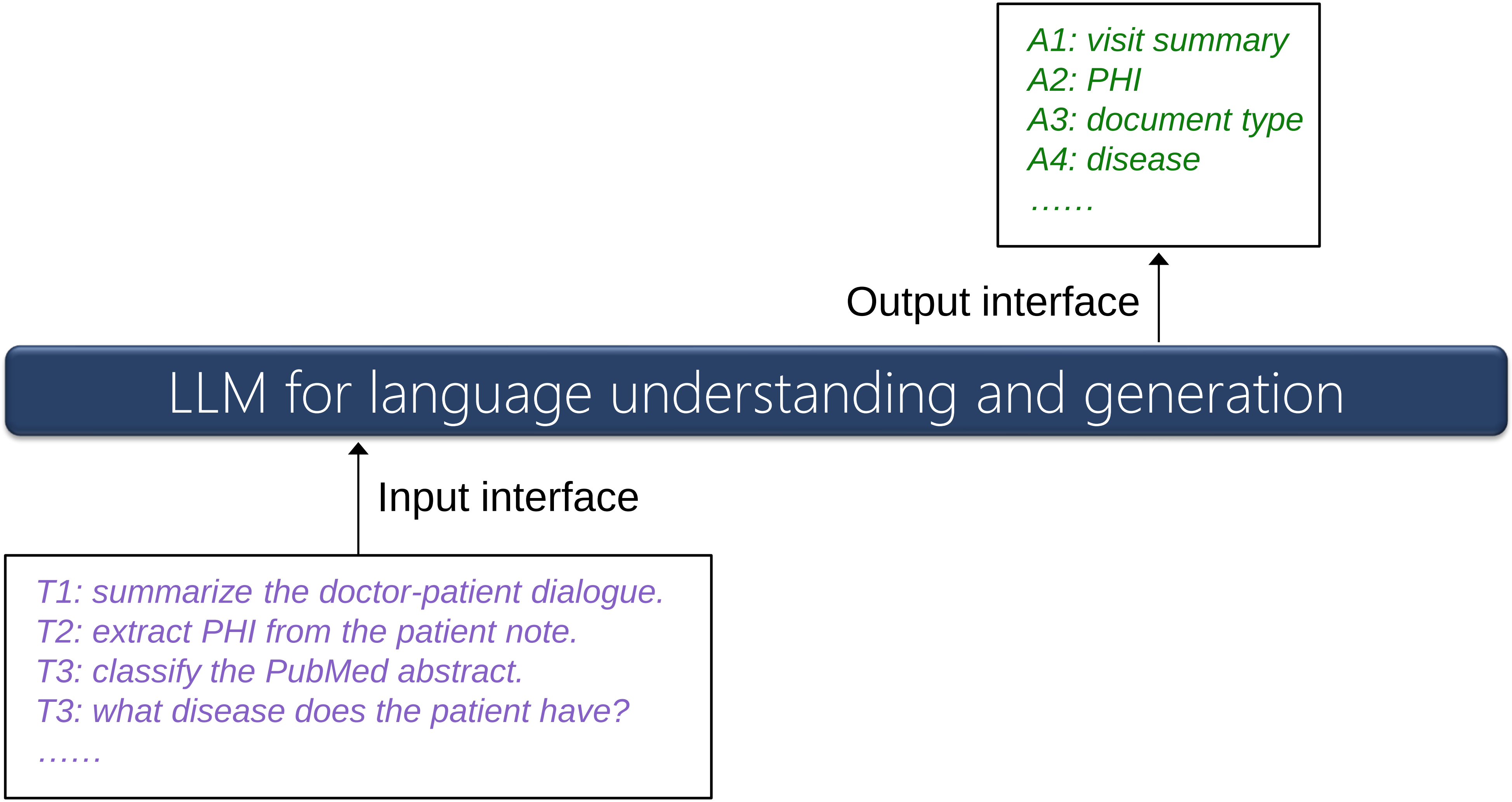
Intelligence revolution

Biomedical LLMs

Application challenges

Research frontiers

General-purpose Interface



Paradigm Shifts with LLMs

Specialist Models → Generalist Models

Closed-set Classification → Open-ended Generation

Representation Learning → Promptable Interface

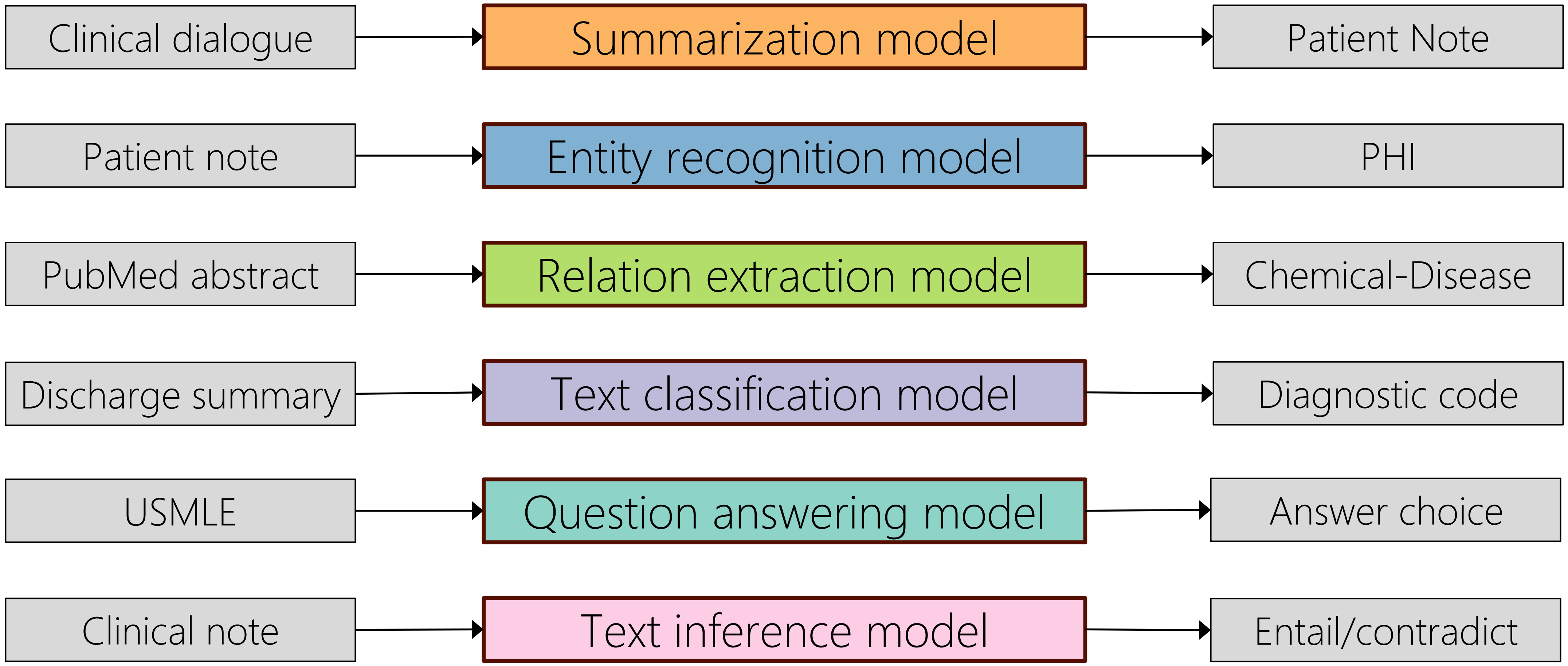
Paradigm Shifts with LLMs

Specialist Models → Generalist Models

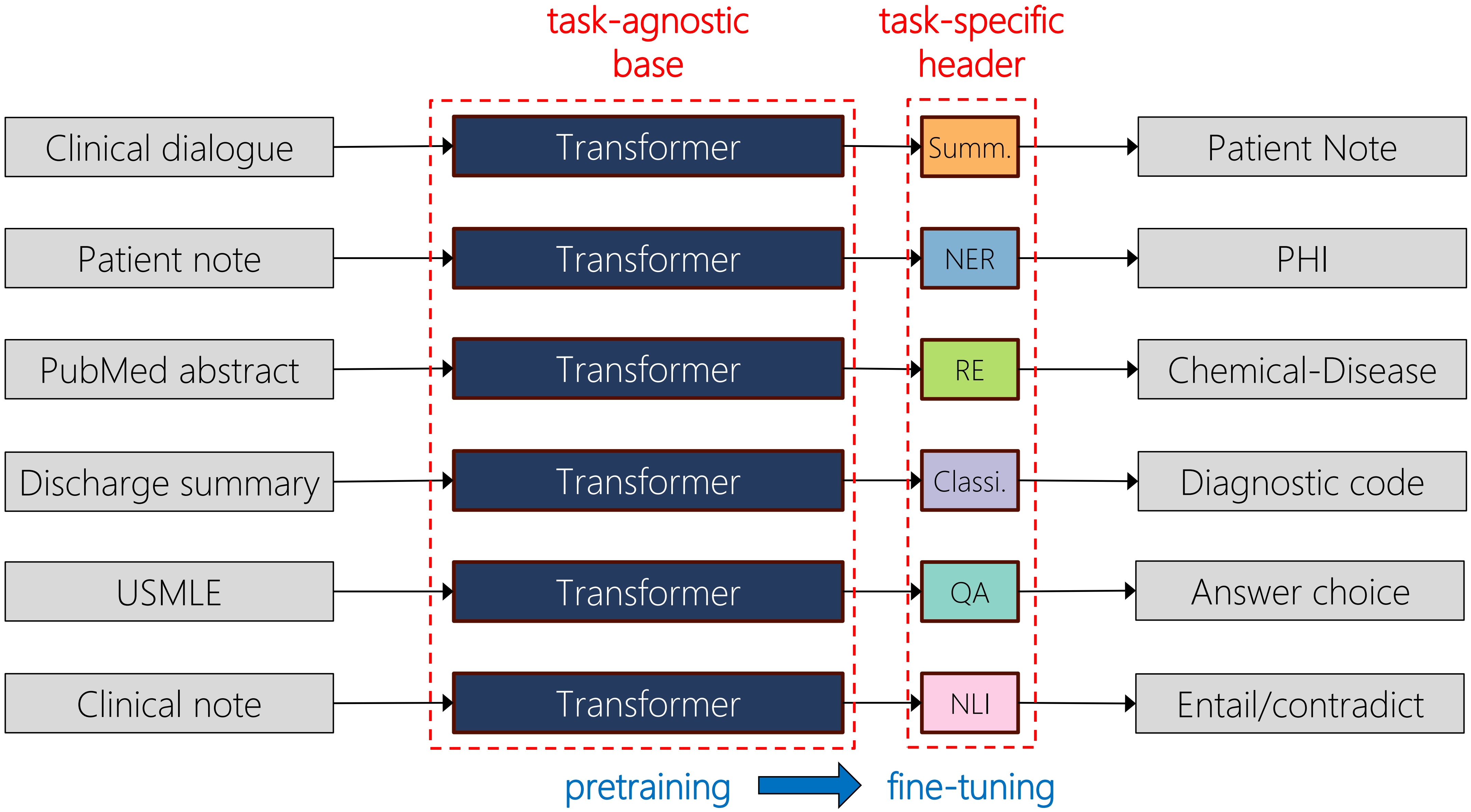
Closed-set Classification → Open-ended Generation

Representation Learning → Promptable Interface

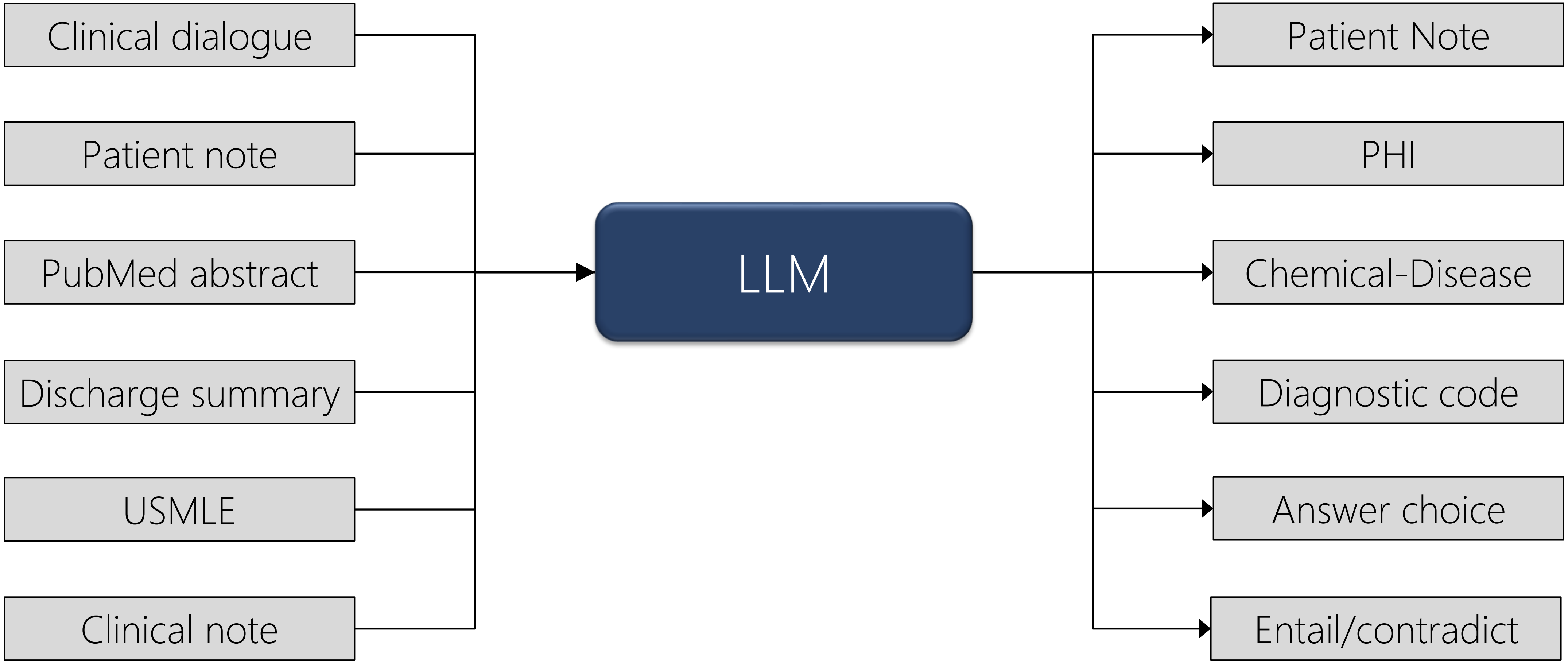
Specialist Models



Specialist Headers



Generalist Models

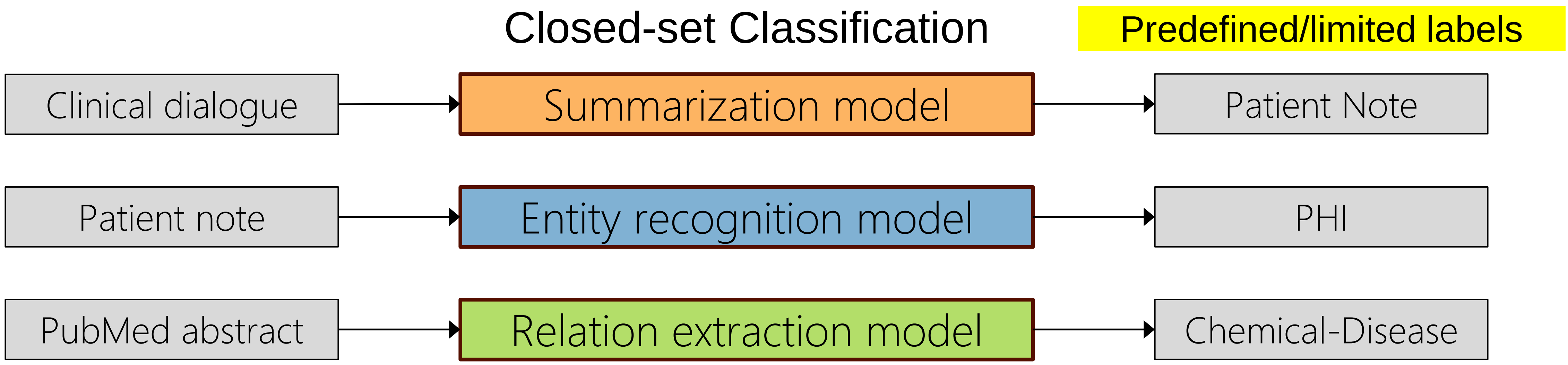


Specialist Models → Generalist Models

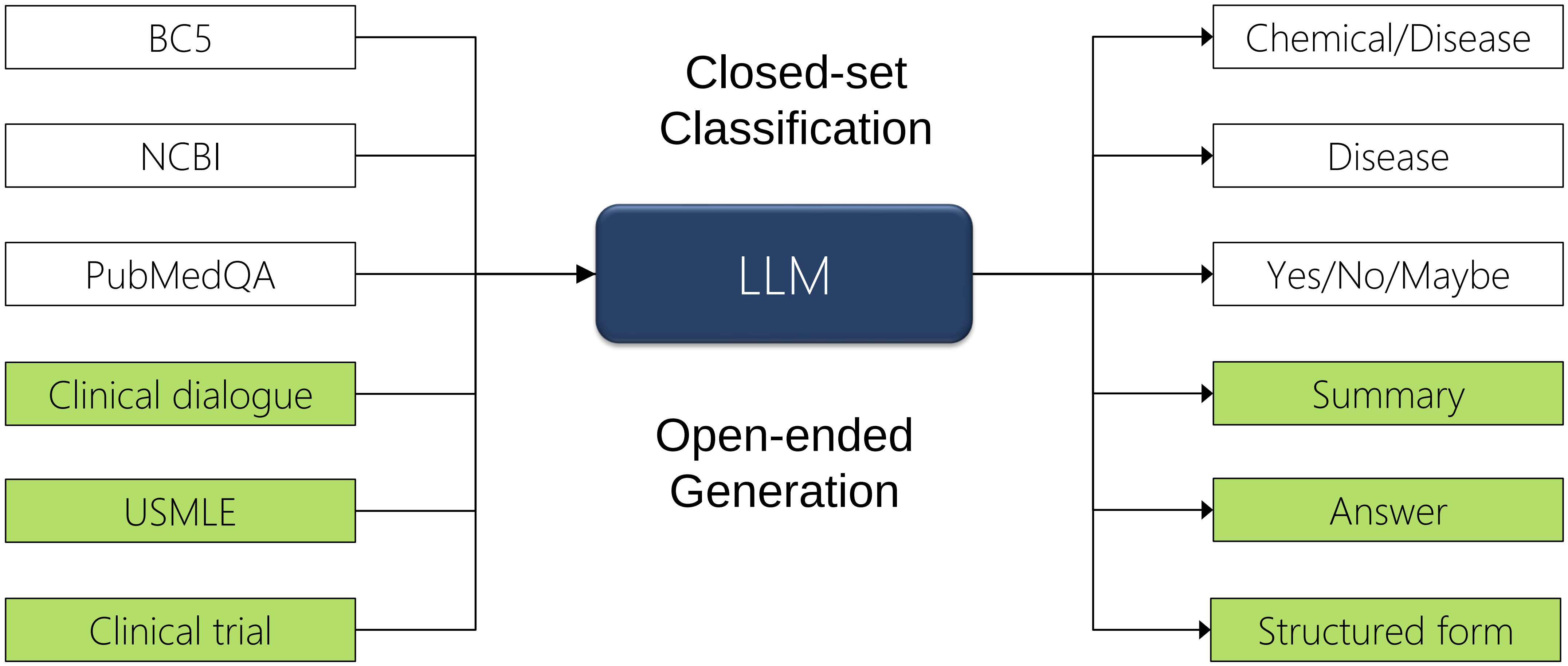
Closed-set Classification → Open-ended Generation

Representation Learning → Promptable Interface

Paradigm Shifts with LLMs



Paradigm Shifts with LLMs



Paradigm Shifts with LLMs

Specialist Models → Generalist Models

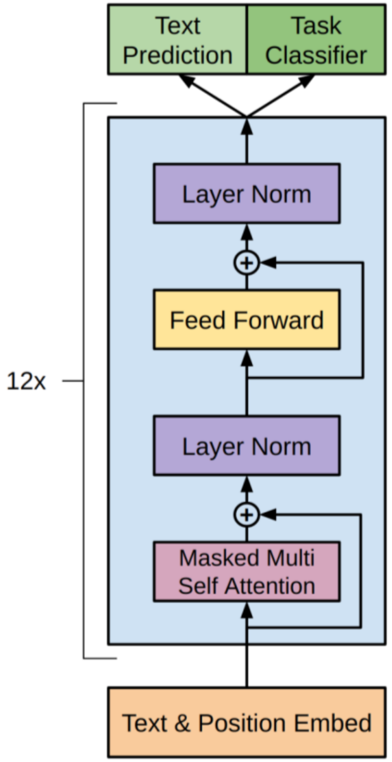
Closed-set Classification → Open-set Generation

Representation Learning → Promptable Interface

Paradigm Shifts with LLMs

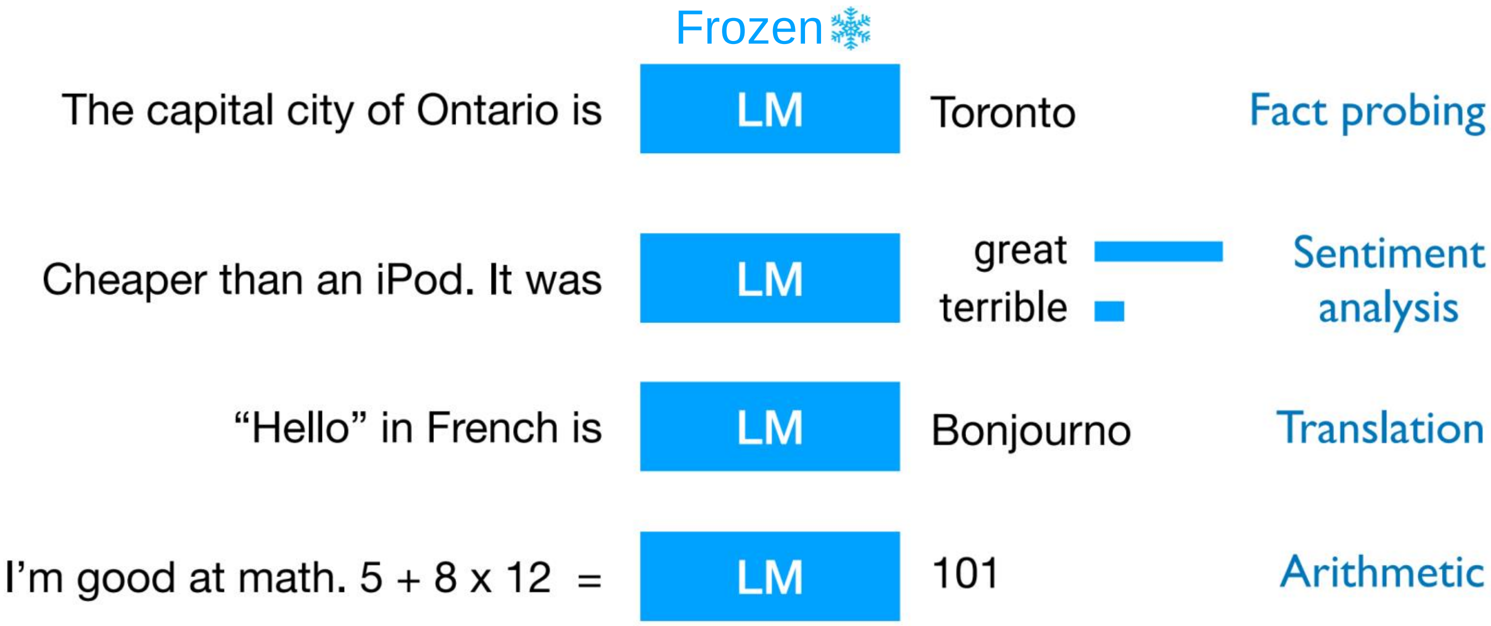
Representation learning

- Expensive
- Engineering heavy
- Task-specific



Promptable interface

- Training free
- Universal interface – natural language



Improving Language Understanding by Generative Pre-Training
Retrieval-based Language Models and Applications

Biomedical LLMs

BioLinkBERT

Galactica

SciBERT

PubMedBERT

GPT-4

Med-PaLM

BioMegatron

Med-PaLM2

BioGPT

ClinicalBERT

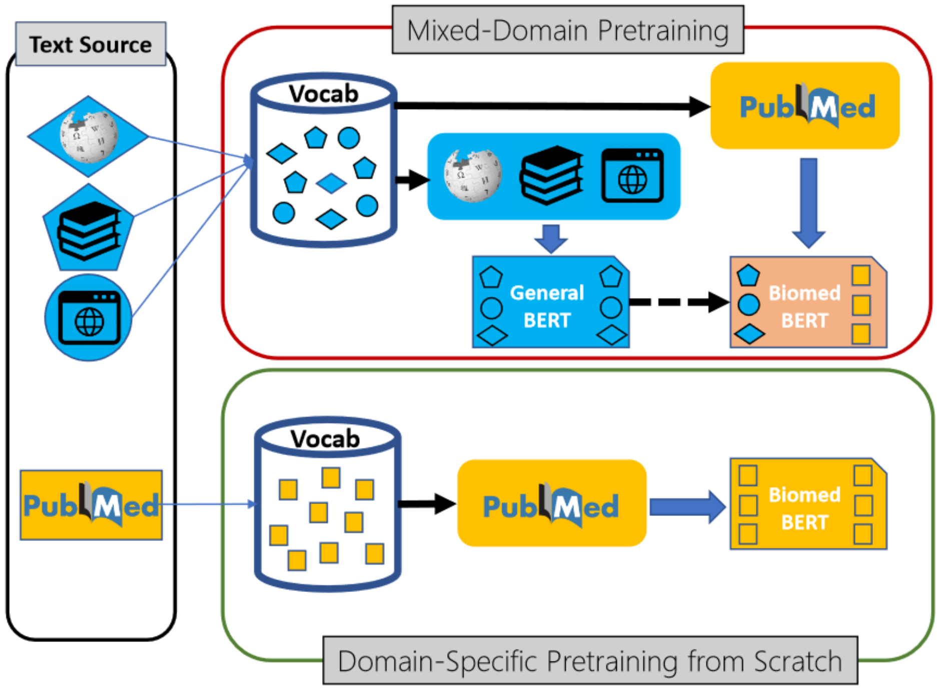
BioBERT

GatorTronGPT

GPT-Neo

BioMedLM

Domain-Specific Pretraining



Microsoft Research

PubMedBERT

Med-PaLM

DRAGON

Galactica

PubMedGPT

BioGPT

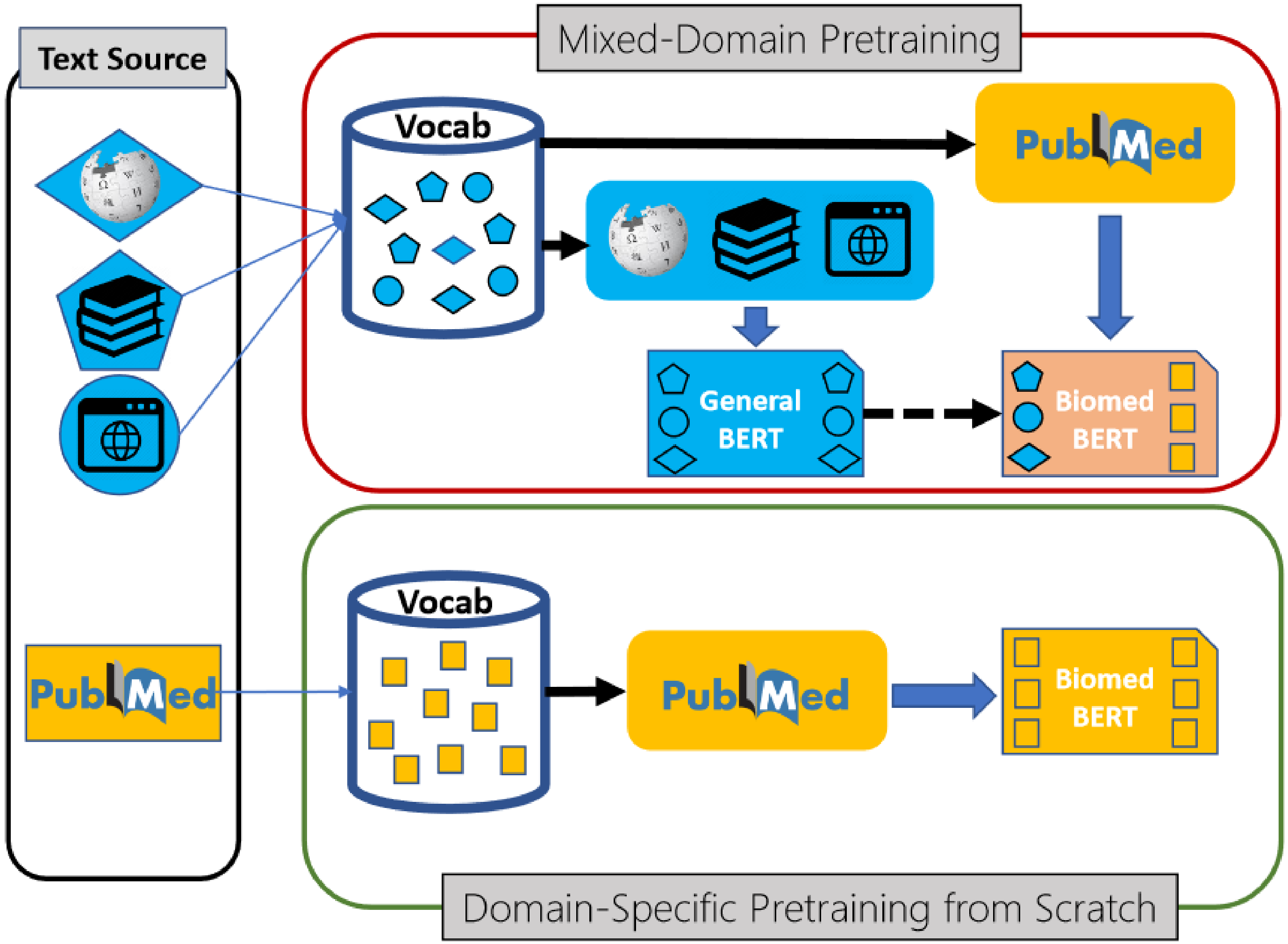
BioLinkBERT

……

2020

2022

Why Domain-Specific Pretraining?



Yu, et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing", *Special Issue on Computational Methods for Biomedical Natural Language Processing, ACM Transactions on Computing for Health 2021*.

PubMedBERT

In **bounded-resource** scenarios, enable **more efficient learning** by focusing on in-domain data

Why Domain-Specific Pretraining?

Shattered into pieces

Domain-specific Vocab

Biomedical Term	Category	BERT	SciBERT	PubMedBERT (Ours)
diabetes	disease	✓	✓	✓
leukemia	disease	✓	✓	✓
lithium	drug	✓	✓	✓
insulin	drug	✓	✓	✓
DNA	gene	✓	✓	✓
promoter	gene	✓	✓	✓
hypertension	disease	hyper-tension	✓	✓
nephropathy	disease	ne-ph-rop-athy	✓	✓
lymphoma	disease	l-ym-ph-oma	✓	✓
lidocaine	drug	lid-oca-ine]	✓	✓
oropharyngeal	organ	oro-pha-ryn-ge-al	or-opharyngeal	✓
cardiomyocyte	cell	card-iom-yo-cy-te	cardiomy-ocyte	✓
chloramphenicol	drug	ch-lor-amp-hen-ico-l	chlor-amp-hen-icol	✓
RecA	gene	Rec-A	Rec-A	✓
acetyltransferase	gene	ace-ty-lt-ran-sf-eras-e	acetyl-transferase	✓
clonidine	drug	cl-oni-dine	clon-idine	✓
naloxone	drug	na-lo-xon-e	nal-oxo-ne	✓

Yu, et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing", *Special Issue on Computational Methods for Biomedical Natural Language Processing, ACM Transactions on Computing for Health 2021*.

Domain-specific Vocab

Preserves the integrity of

- Biomedical terms
- Amino acid sequences
- SMILES formula
- DNA sequences
- Mathematics
- Citations
- etc.

PubMedBERT: A Million Downloads Per Month

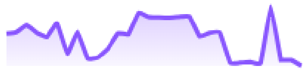
microsoft/**BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext** like 108

Fill-Mask | PyTorch | JAX | Transformers | English | bert | exbert | AutoTrain Compatible | arxiv:2007.15779 | License: mit

Model card | Files and versions | Community 1 | Train | Deploy | Use in Transformers

Edit model card

Downloads last month
955,990



PubMedBERT (abstracts + full text)

Pretraining large neural language models, such as BERT, has led to impressive gains on many natural language processing (NLP) tasks. However, most pretraining efforts focus on general domain corpora, such as newswire and Web. A prevailing assumption is that even domain-specific pretraining can benefit by starting from general-domain language models. Recent work shows that for domains with abundant unlabeled text, such as biomedicine, pretraining language models from scratch results in substantial gains over continual pretraining of general-domain language models.

PubMedBERT is pretrained from scratch using *abstracts* from PubMed and *full-text* articles from PubMedCentral. This model achieves state-of-the-art performance on many biomedical NLP tasks, and currently holds the top score on the Biomedical Language Understanding and Reasoning Benchmark.

Hosted inference API

Fill-Mask | Examples

Mask token: [MASK]

[MASK] is a tumor suppressor gene.

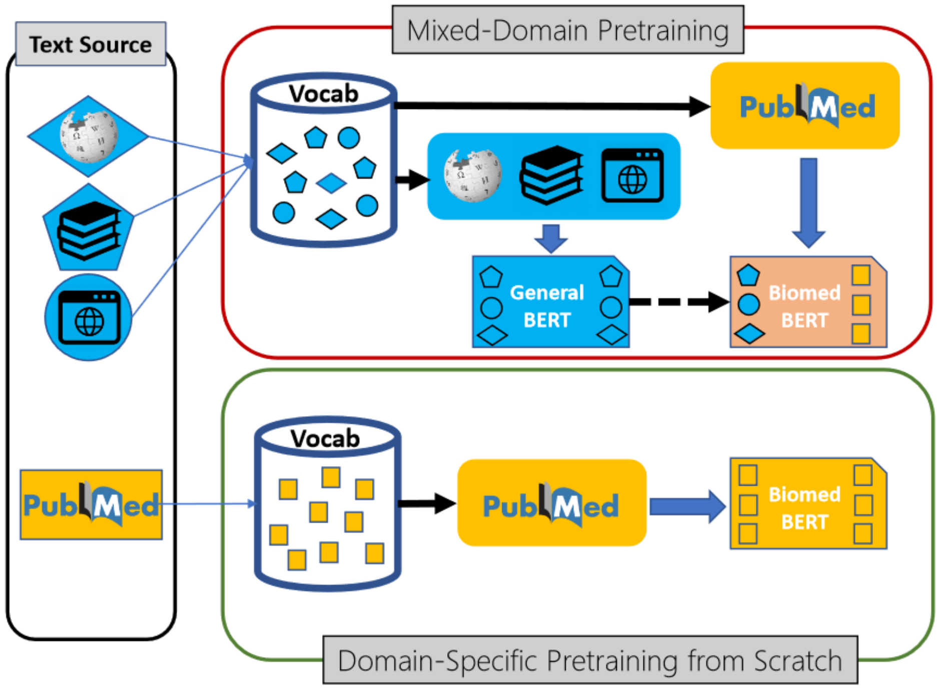
Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

p53 0.286

tp53 0.169

Domain-Specific Pretraining → Generalist Model



Microsoft Research

PubMedBERT

Med-PaLM

DRAGON

Galactica

PubMedGPT

BioGPT

BioLinkBERT

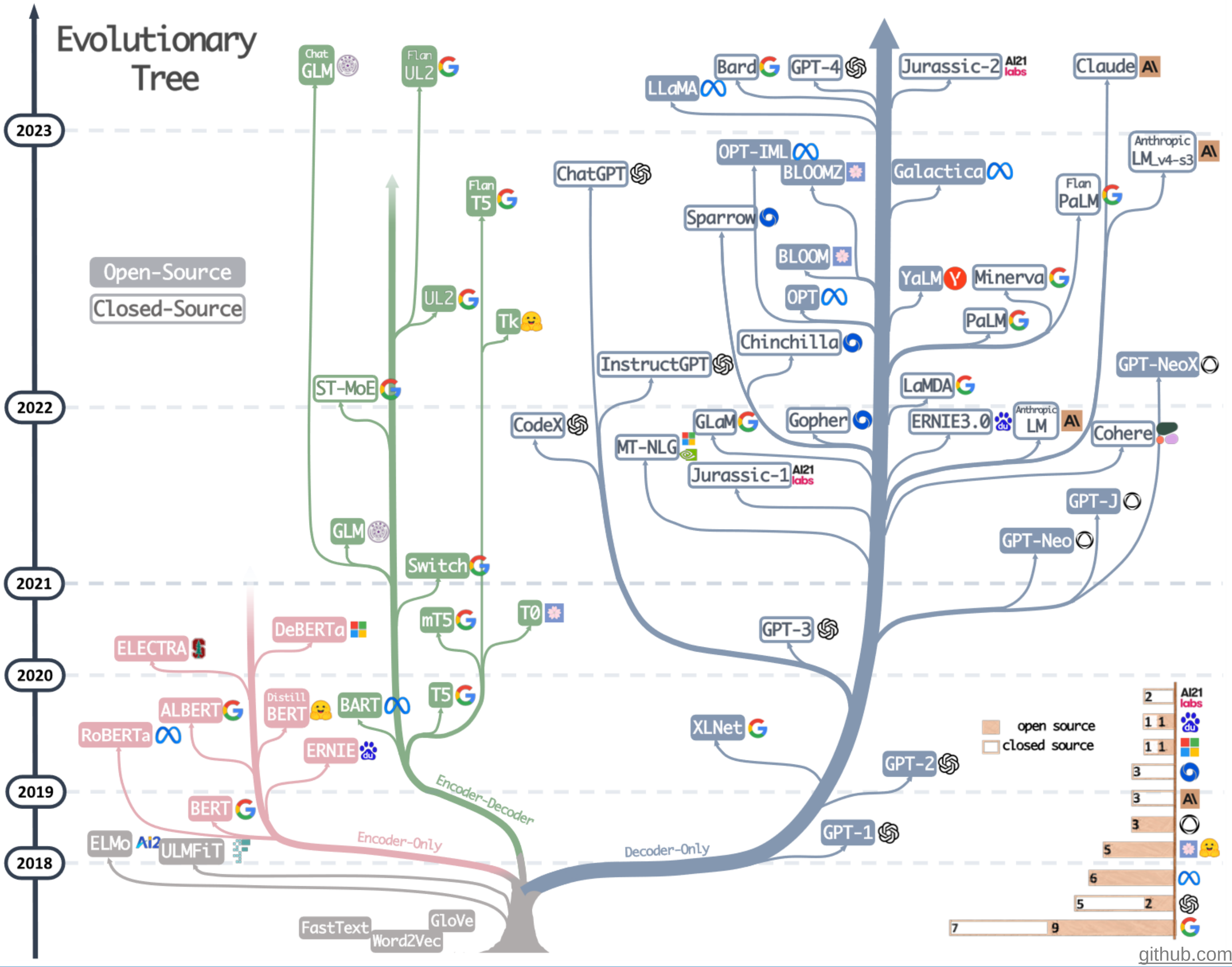
……

GPT-4

2020

2022

2023



github.com/Mooler0410/LLMsPracticalGuide

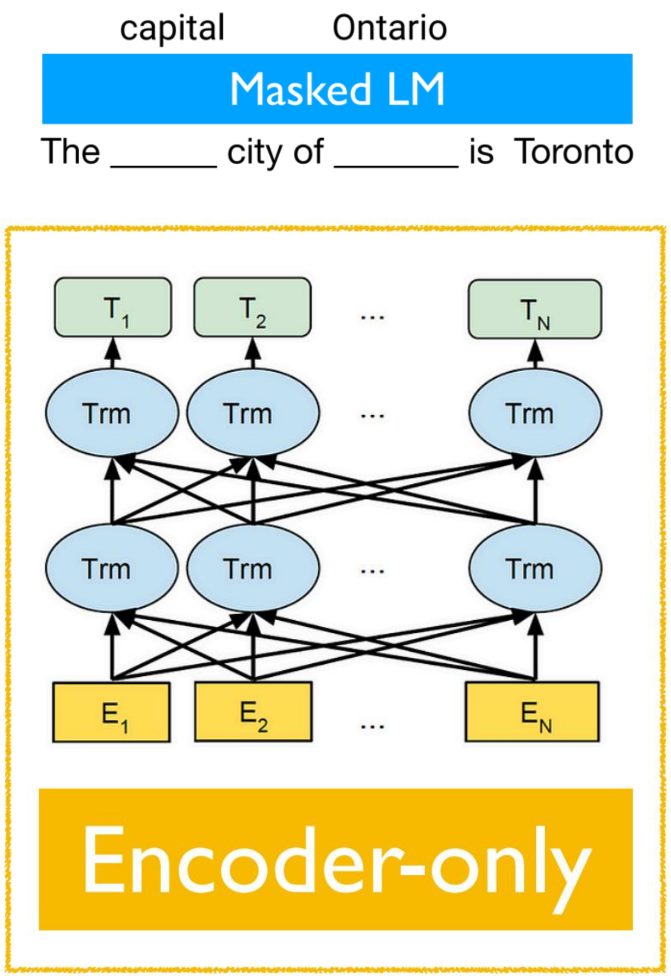
Biomedical LLM: Encoder-Only

PubMedBERT

SciBERT

BioBERT

ClinicalBERT

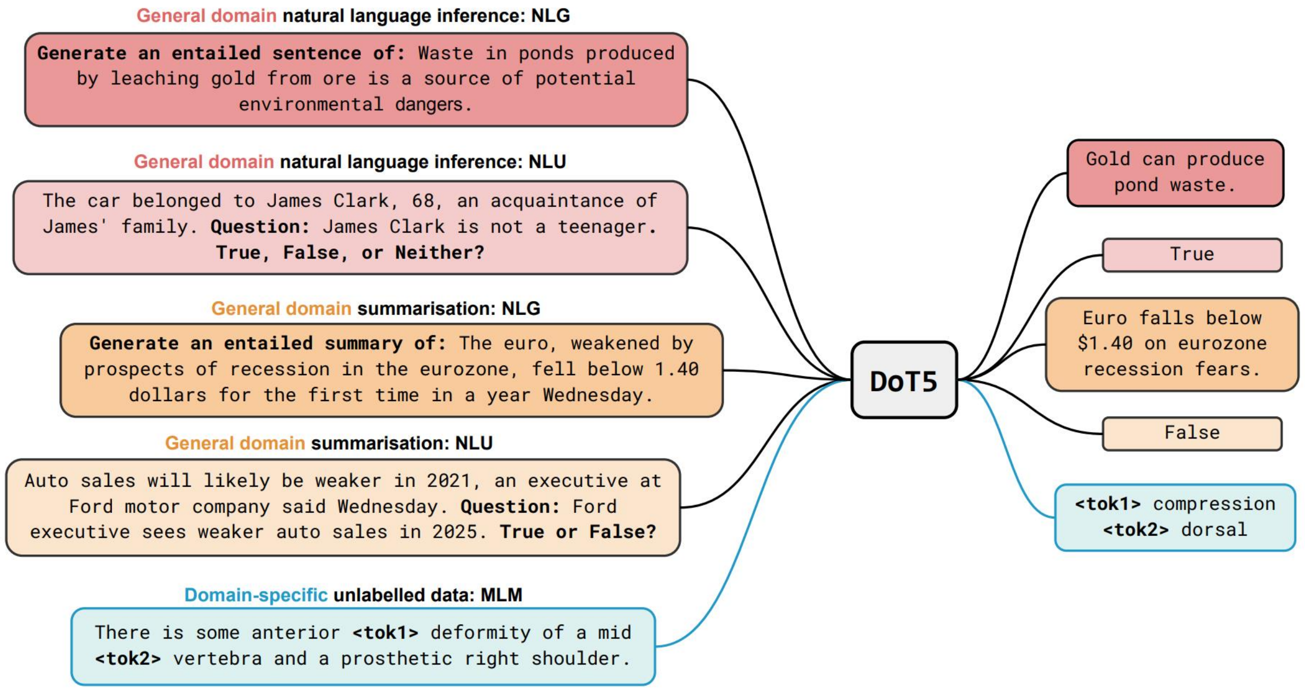


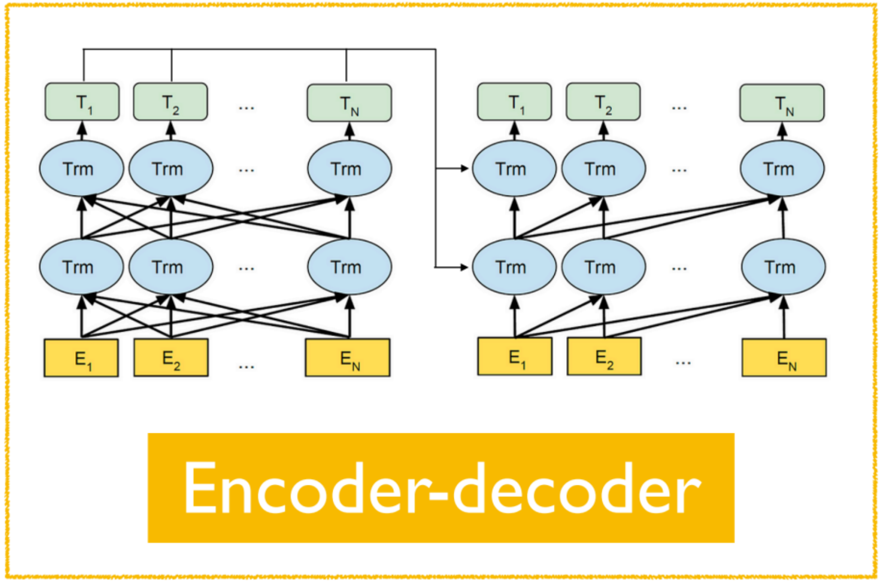
Diagrams adapted from Retrieval-based Language Models and Applications

Biomedical LLM: Encoder-Decoder

DoT5: Compositional Zero-Shot Domain Transfer with Text-to-Text Models

SciFive: a text-to-text transformer model for biomedical literature





Diagrams adapted from Retrieval-based Language Models and Applications

BioGPT

GPT model pretrained on 15M PubMed abstracts

Strong performance on Biomedical tasks

- Relation extraction (e.g., BC5CDR, KD-DTI and DDI)
- Question answering (e.g., PubMedQA)
- Document classification (e.g., HoC)
- Text generation

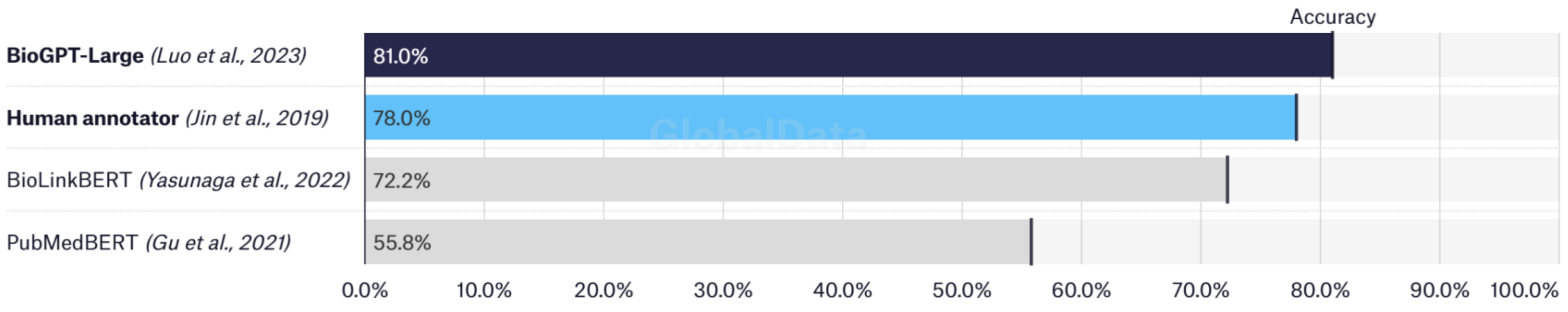
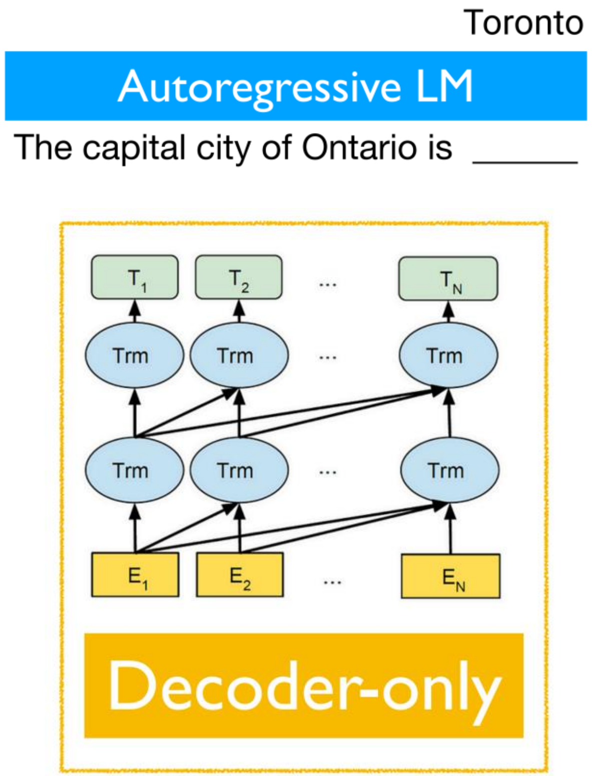


Chart: GlobalData • Source: PubMedQA



BioGPT: generative pre-trained transformer for biomedical text generation and mining

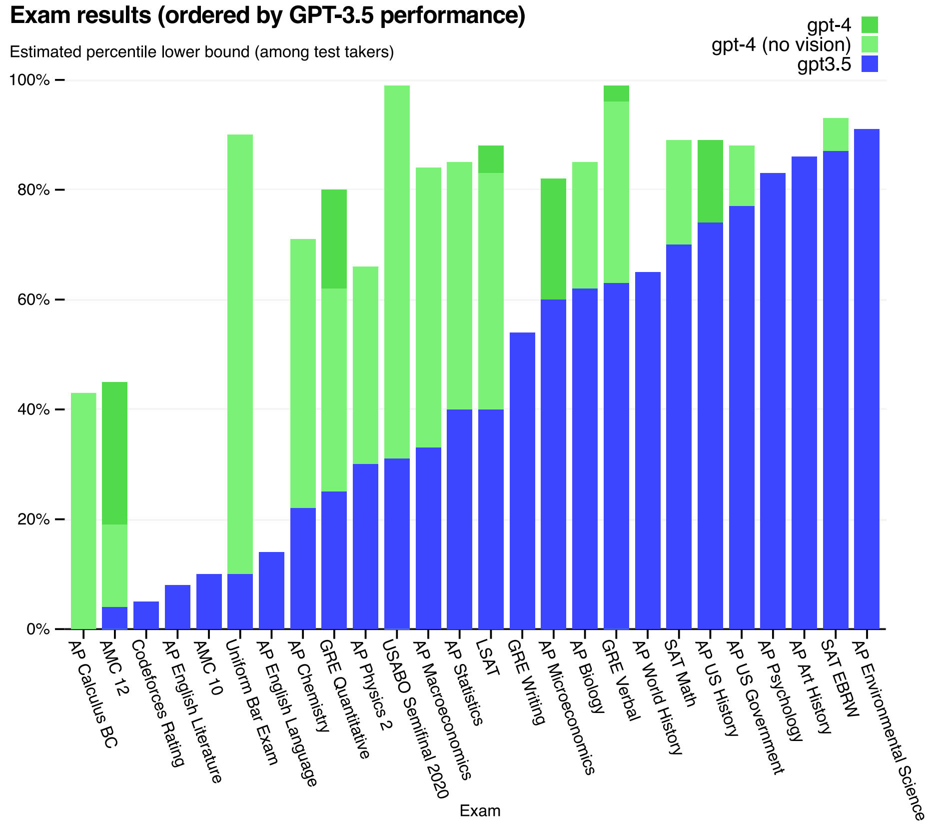
Other Biomedical GPTs

- BioMedLM (PubMedGPT)
  A Domain-Specific Large Language Model for Biomedical Text
- GatorTronGPT
  A Study of Generative Large Language Model
- BioMegatron
  Larger Biomedical Domain Language Model

Many others……

GPT-4 Out-of-Box: Expert-Level Competency on USMLE

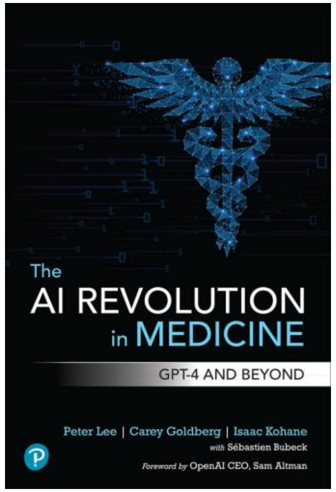
The most powerful general-purpose LLM
Human-level performance on many tasks



GPT-4 Technical Report

- SOTA on **medical competency examinations**
- *"How well does the AI perform clinically? And my answer is, I'm stunned to say: **Better than many doctors I've observed.**"* — Isaac Kohane MD

Dataset	GPT-4-base 5 shot / 0 shot	GPT-4 5 shot / 0 shot
MedQA		
Mainland China	**78.63** / 74.34	75.31 / 71.07
Taiwan	**87.47** / 85.14	84.57 / 82.17
US (5-option)	**82.25** / 81.38	78.63 / 74.71
US (4-option)	**86.10** / 84.45	81.38 / 78.87
PubMedQA		
Reasoning Required	77.40 / **80.40**	74.40 / 75.20
MedMCQA		
Dev	**73.66** / 73.42	72.36 / 69.52



Capabilities of GPT-4 on Medical Challenge Problems
The AI Revolution in Medicine: GPT-4 and Beyond

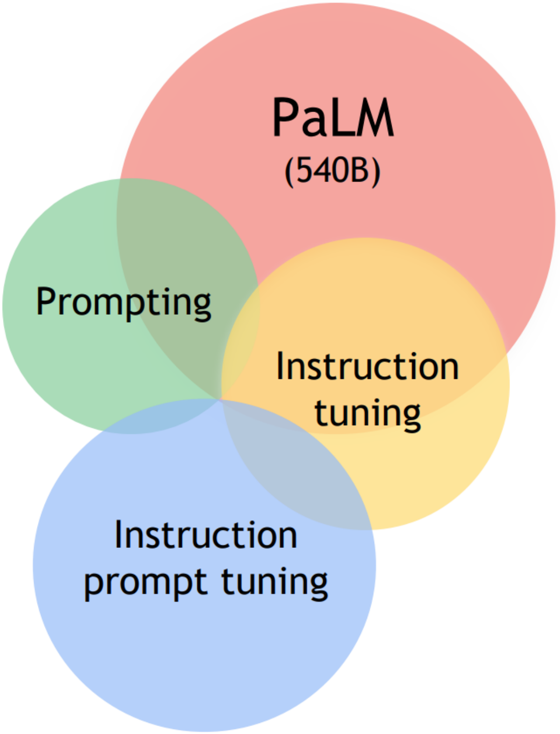
GPT-4

GPT-4 has been pretrained on a large portion of the public web, which **already contains a lot of biomedical text.**

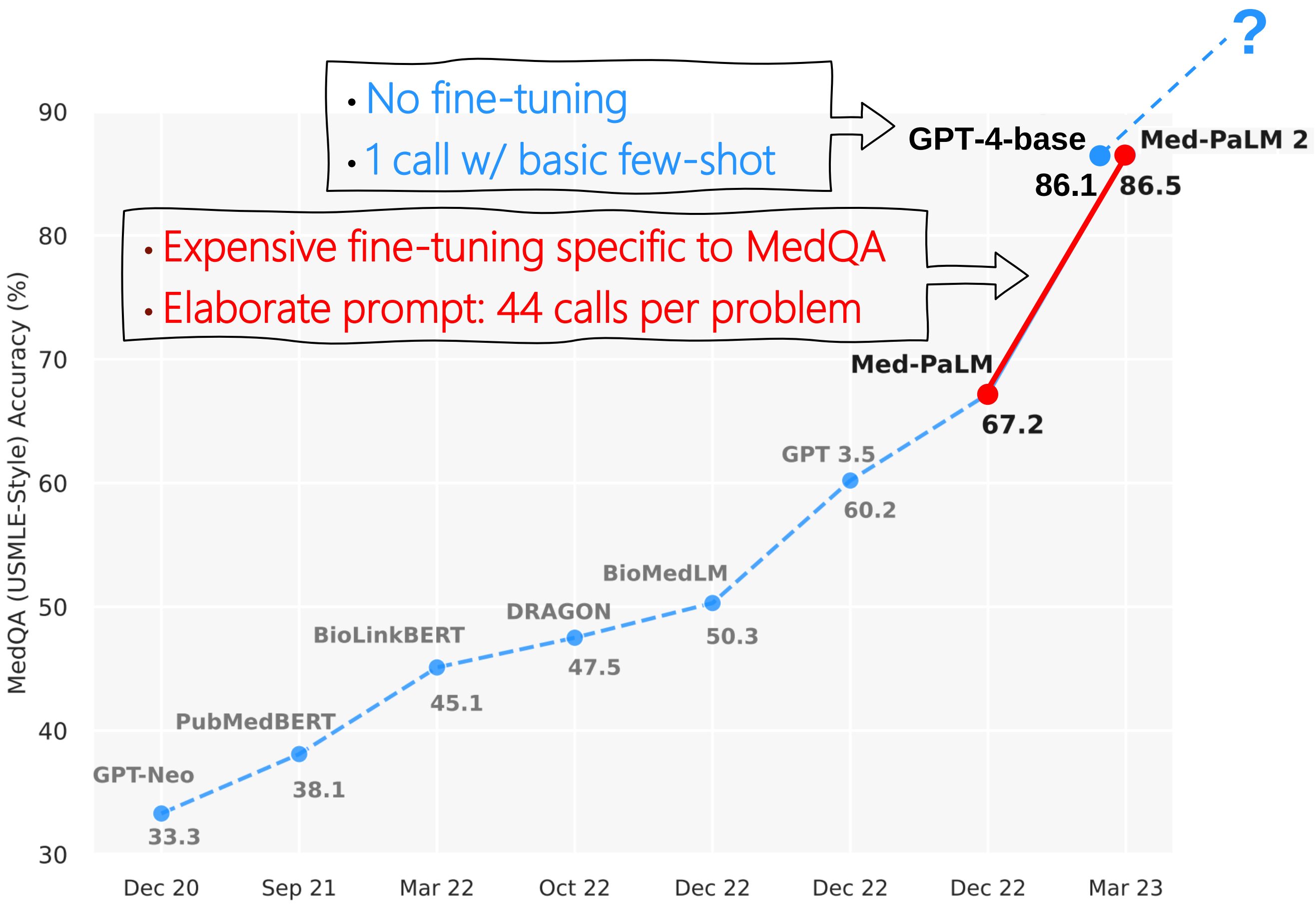
Component	Raw Size
Pile-CC	227.12 GiB
PubMed Central	90.27 GiB
Books3†	100.96 GiB
OpenWebText2	62.77 GiB
ArXiv	56.21 GiB
Github	95.16 GiB
FreeLaw	51.15 GiB
Stack Exchange	32.20 GiB
USPTO Backgrounds	22.90 GiB
PubMed Abstracts	19.26 GiB
Gutenberg (PG-19)†	10.88 GiB
OpenSubtitles†	12.98 GiB
Wikipedia (en)†	6.38 GiB
DM Mathematics†	7.75 GiB
Ubuntu IRC	5.52 GiB
BookCorpus2	6.30 GiB
EuroParl†	4.59 GiB
HackerNews	3.90 GiB
YoutubeSubtitles	3.73 GiB
PhilPapers	2.38 GiB
NIH ExPorter	1.89 GiB
Enron Emails†	0.88 GiB
The Pile	**825.18 GiB**

Med-PaLM 2

PaLM requires substantial adaptation to do well on USMLE

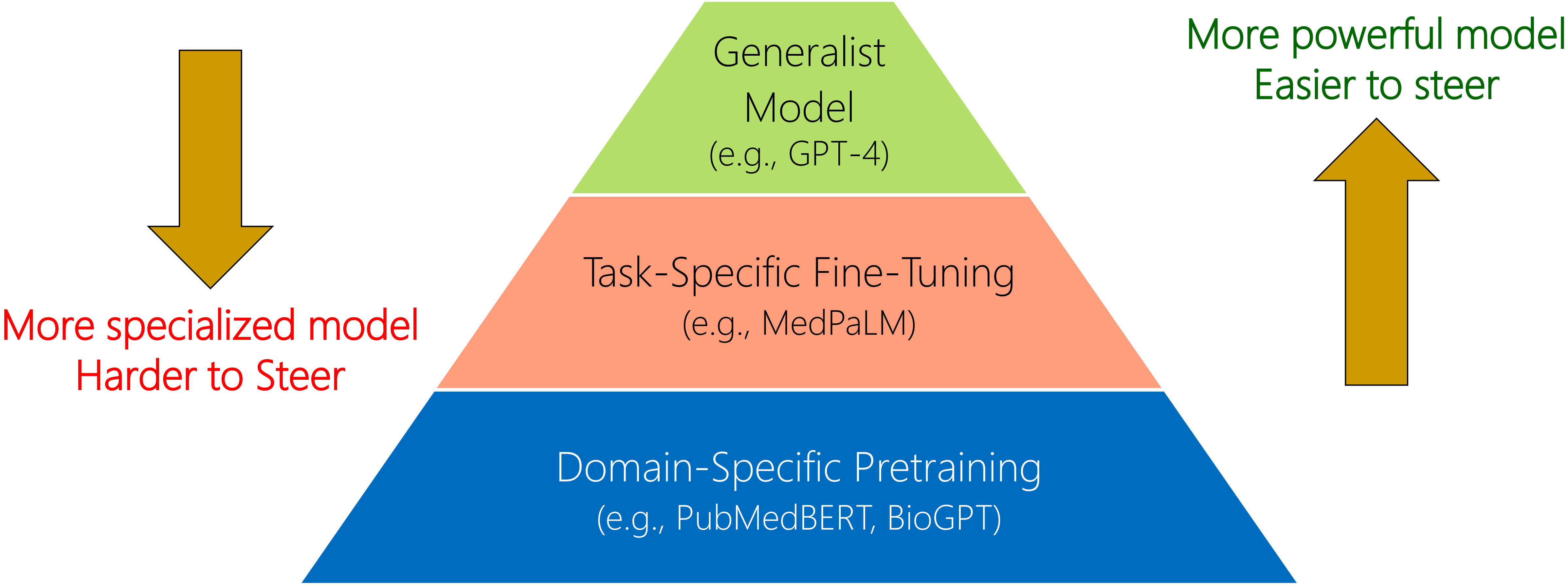


Let GPT-4 design the prompt



Large Language Models Encode Clinical Knowledge
Towards Expert-Level Medical Question Answering with Large Language Models

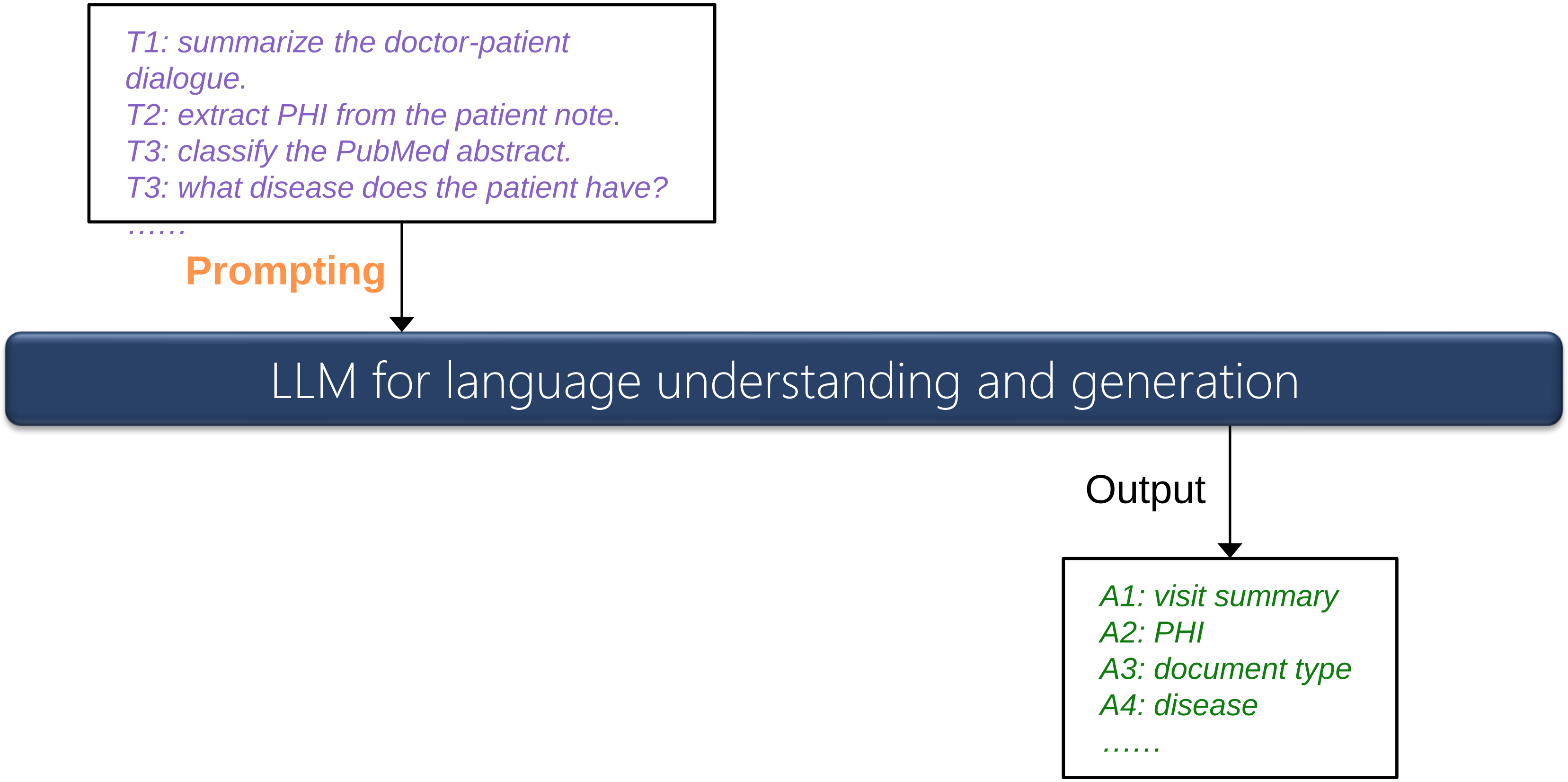
Generalist Models: Superior Steerability



Prompt programming

Prompt Programming

Using natural language prompt to steer LLMs



Basic Prompting: Zero-shot

Simply feed the task input and ask for results

Lack of context, low performance

Question: A 6-year-old boy is brought to the pediatrician by his foster father because he is concerned about the boy's health… what is released by the eosinophils to cause bronchial epithelial damage?

A. IL-5
B. IL-8
C. Major basic protein
D. Interferon-gamma

=> Answer: ____

In-context learning: Instruction prompting

Instructions:
Explain the domain, task definition and expected output

Answer multiple choice questions about medical knowledge. The answer must be from {A, B, C, D}.

Question: A 6-year-old boy is brought to the pediatrician by his foster father because he is concerned about the boy's health… what is released by the eosinophils to cause bronchial epithelial damage?

A. IL-5
B. IL-8
C. Major basic protein
D. Interferon-gamma

=> Answer: ____

In-context learning: Few-shot

Few-shot examples help LLMs better understand **human intention** and **criteria for what kinds of answers are wanted**

Few-shot examples

Question: A 6-year-old boy is brought to the pediatrician by his foster father because he is concerned about the boy's health… what is released by the eosinophils to cause bronchial epithelial damage?

A. IL-5 B. IL-8 C. Major basic protein D. Interferon-gamma

=> Answer: A

Question: A previously healthy 32-year-old woman comes to the physician 8 months after her husband was killed in a car crash… Pharmacotherapy should be targeted to which of the following neurotransmitters?

A. Dopamine B. Glutamate C. Norepinephrine D. Serotonin

=> Answer: D

Question: An investigator is studying cellular regeneration of epithelial cells… Which of the following parts of the female reproductive tract is also lined by this type of epithelium?

A. Ovaries B. Vagina C. Fallopian tubes D. Vulva

=>

Tips for example selection

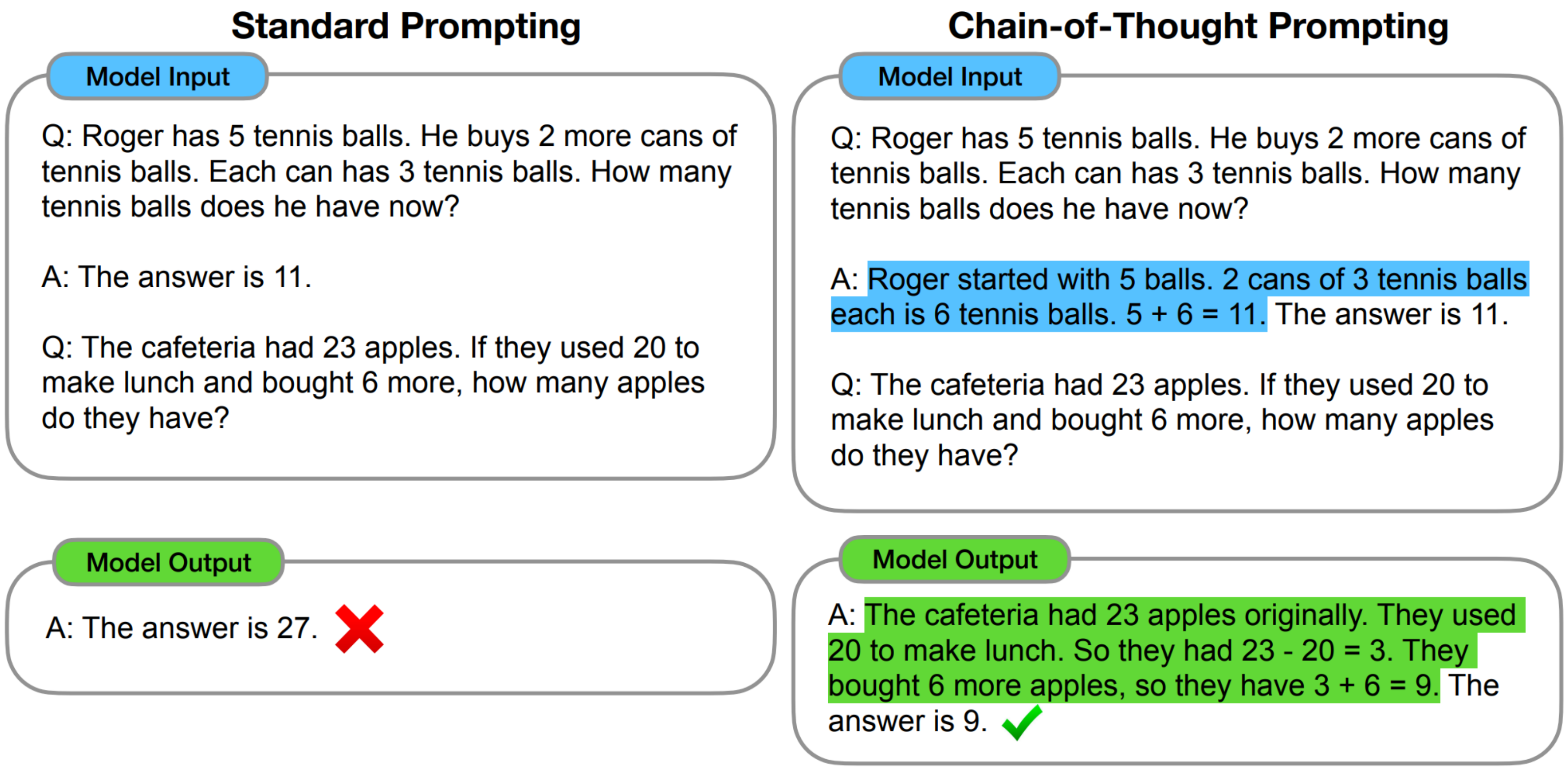
- Relevancy
  - Contrastive learning (Rubin et al., 2022)
  - *k*-NN (Liu et al., 2021)
- Diversity
  - Graph-based approach (Su et al., 2022)
  - Q-learning (Zhang et al. 2022)

Tips for example ordering

- majority label bias (Lu et al. 2022)
- recency bias

Reasoning: Chain-of-Thought (CoT)

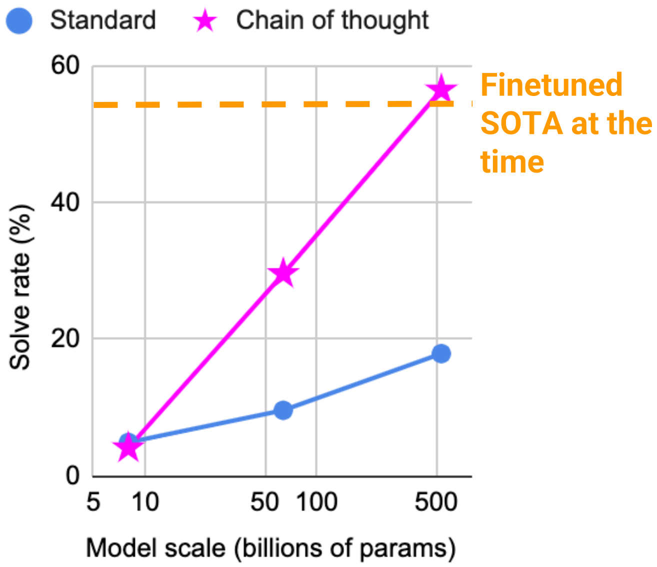
- Enable language models to do more-complicated tasks
- Guide them with “meta-data” (i.e., reasoning process)



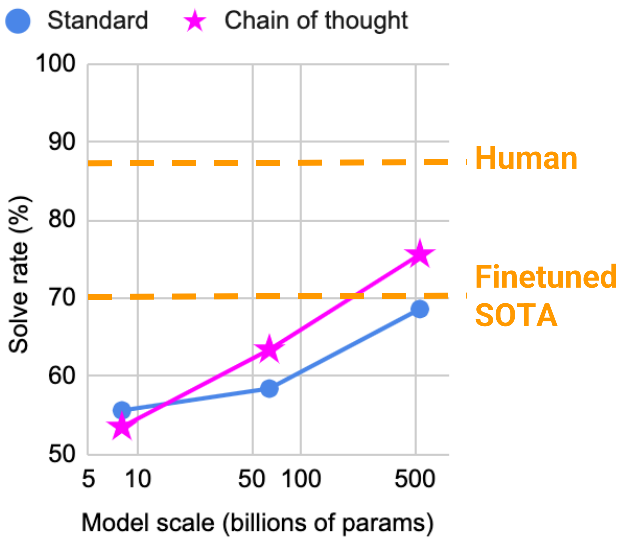
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Reasoning: Chain-of-Thought (CoT)

GSM8K



StrategyQA



Increased interpretability

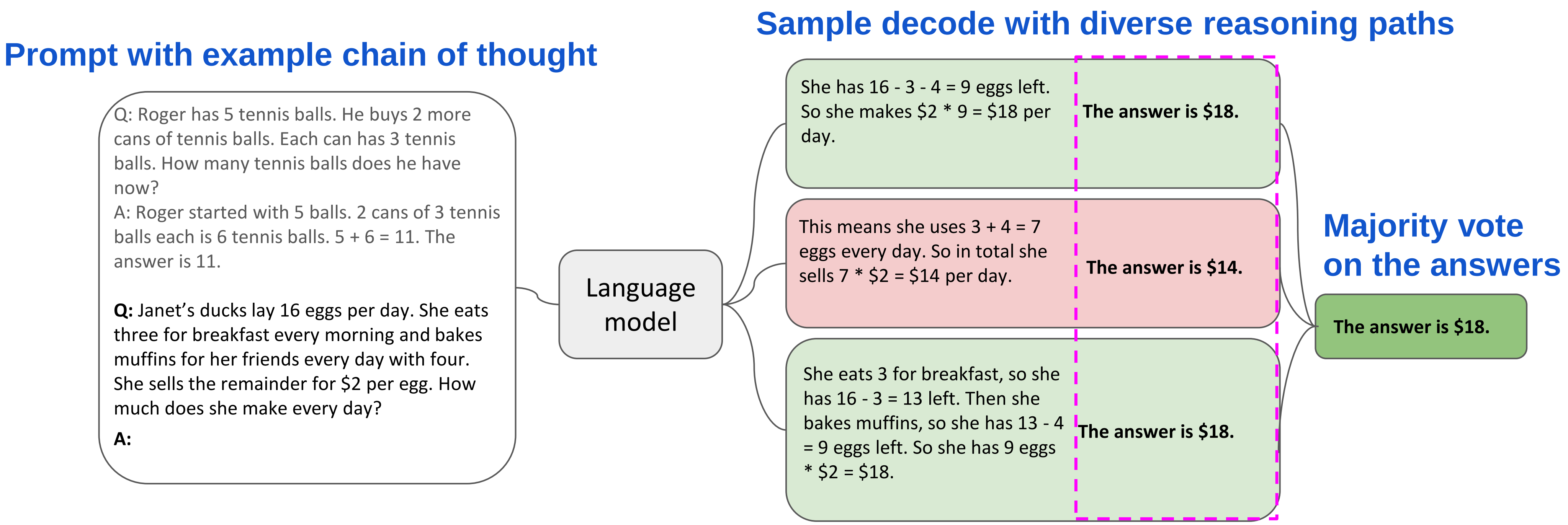
Real model output #1

Q: Can you hide a basketball in a sand cat's ear?

A: A basketball is about 30 cm in diameter. A sand cat's ear is about 2 cm in diameter. Thus, a basketball would not fit in a sand cat's ear. So the answer is no.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

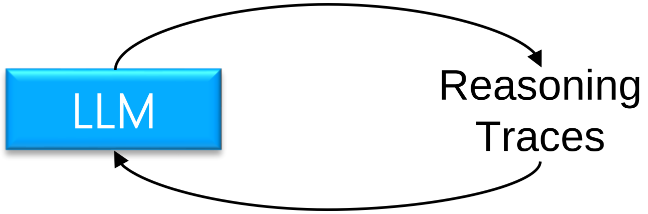
Reasoning: Self-consistency



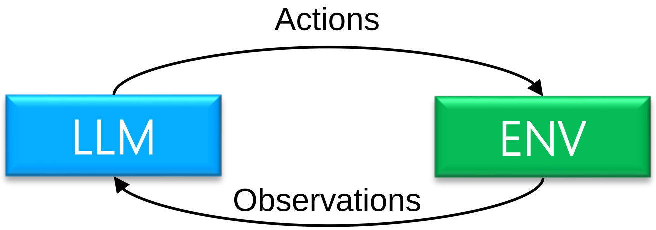
Self-Consistency Improves Chain of Thought Reasoning in Language Models

External tools: ReAct

Reasoning traces (to update **internal** belief)



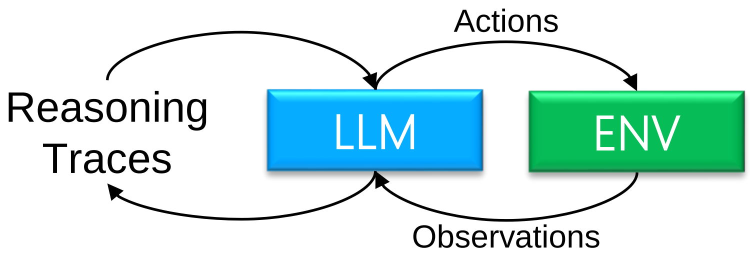
Reason Only
e.g., Chain-of-thought (Wei et al.)



Act Only
e.g., SayCan (Ahn et al.)

Actions (to obtain **external** feedback)

ReAct: **synergizing** reasoning and acting



ReAct: Synergizing Reasoning and Acting in Language Models

External tools: ReAct

Question: What is the elevation range for the area that the eastern sector of the Colorado orogeny extends into?

Thought 1: I need to search Colorado orogeny, find the area that the eastern sector of the Colorado orogeny extends into, then find the elevation range of the area.

Reasoning trace

Action 1: Search[Colorado orogeny]

Action: search

Observation 1: The Colorado orogeny was an episode of mountain building (an orogeny) in Colorado and surrounding areas.

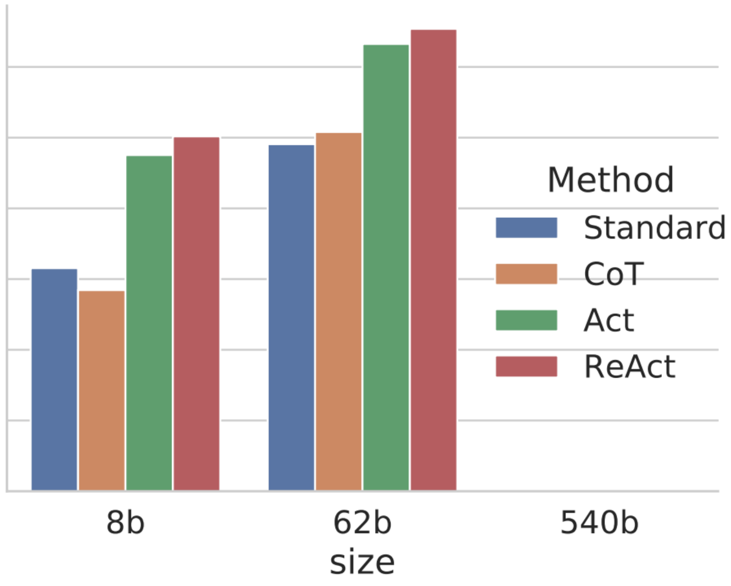
Observation

Thought 2: It does not mention the eastern sector. So I need to look up eastern sector.

Multi-turn

......

Improvement on **knowledge-intensive** tasks (e.g., HotpotQA)



[ReAct: Synergizing Reasoning and Acting in Language Models](#)

Many more…

- Ask Me Anything: A simple strategy for prompting language models
- Progressive Prompts: Continual Learning for Language Models
- Batch Prompting: Efficient Inference with LLM APIs
- Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP
- Large Language Models are reasoners with Self-Verification
- PAL: Program-aided Language Models
- Large Language Models Are Human-Level Prompt Engineers
- Dynamic Prompting: A Unified Framework for Prompt Tuning
- Multitask Prompt Tuning Enables Parameter-Efficient Transfer Learning
- Prompt, Generate, then Cache: Cascade of Foundation Models makes Strong Few-shot Learners
- EvoPrompting: Language Models for Code-Level Neural Architecture Search
- In-Context Instruction Learning
- Chain of Hindsight Aligns Language Models with Feedback
- Language Is Not All You Need: Aligning Perception with Language Models
- Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data
- Active Prompting with Chain-of-Thought for Large Language Models
- More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models
- A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT
- Guiding Large Language Models via Directional Stimulus Prompting
- How Does In-Context Learning Help Prompt Tuning?
- Scalable Prompt Generation for Semi-supervised Learning with Language Models
- Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints

Active research area!

Retrieval-augmented generation (RAG)

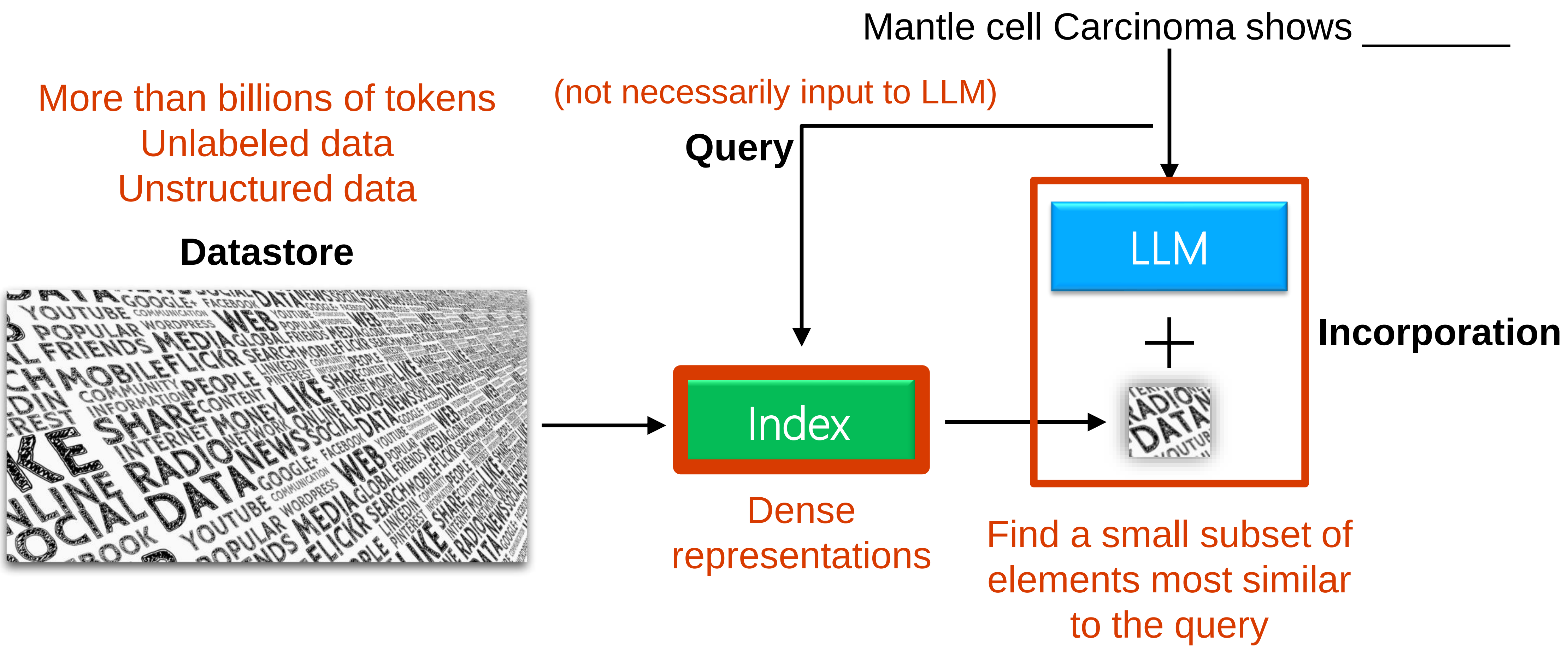
Inference: LLMs

Mantle cell Carcinoma shows ________

LLM

Adapted from Retrieval-based Language Models and Applications

Inference: RAG



Adapted from [Retrieval-based Language Models and Applications]

Why RAG?

Reduce hallucination

- LLMs have bad memory on long-tail knowledge

List 5 important papers authored by Geoffrey Hinton

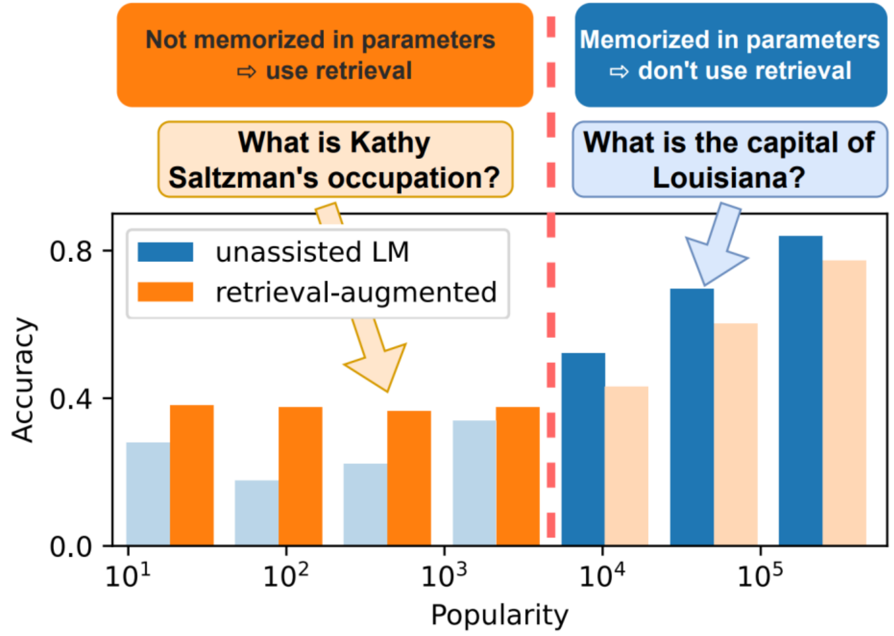
ChatGPT

Geoffrey Hinton is a renowned computer scientist …
Here are five important papers authored by him:

1. "**Learning Internal Representations by Error Propagation**" (with **D. E. Rumelhart** and **R. J. Williams**) - This paper, published in **1986**, ..
2. "**Deep Boltzmann Machines**" (with **R. Salakhutdinov**) - Published in **2009**, ..

…

4. "**Deep Learning**" (with Y. Bengio and A. Courville) Published as a book in **2016**,…
5. "**Attention Is All You Need**" (with V. Vaswani, N. Shazeer, et al.) - Published in **2017**, this paper introduced the Transformer model,…



When Not to Trust Language Models
Retrieval-based Language Models and Applications

Why RAG?

Reduce hallucination

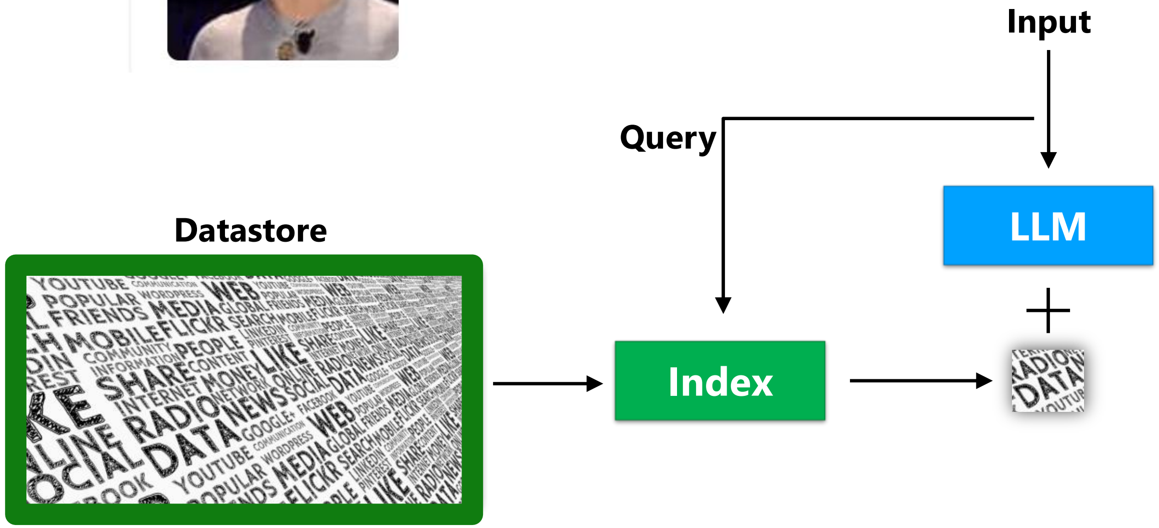
- LLMs' knowledge is easily outdated and hard to update

Who is the CEO of Twitter?

ChatGPT

As of my **knowledge cutoff in September 2021**, the CEO of Twitter is **Jack Dorsey**....

Datastore can be easily updated and expanded

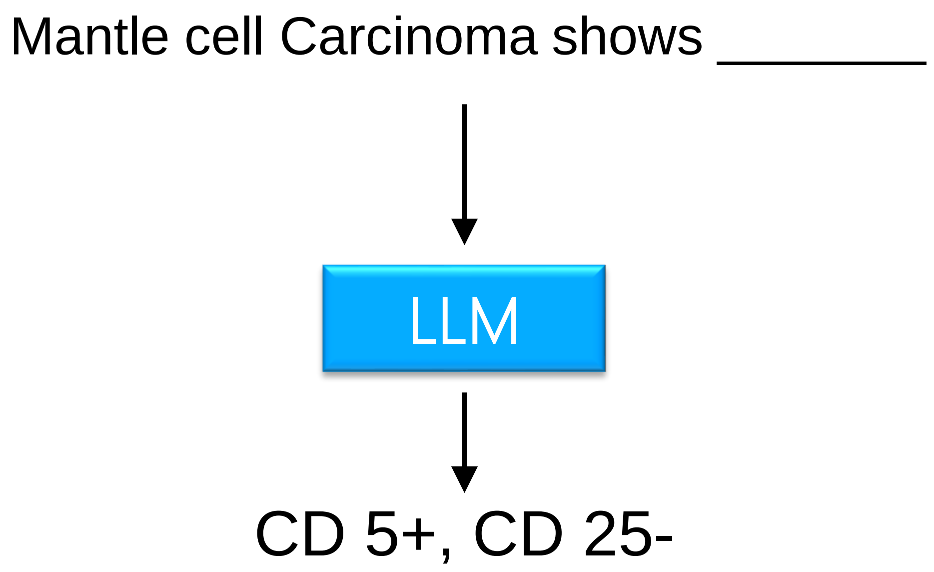


Adapted from Retrieval-based Language Models and Applications

Why RAG?

Provenance

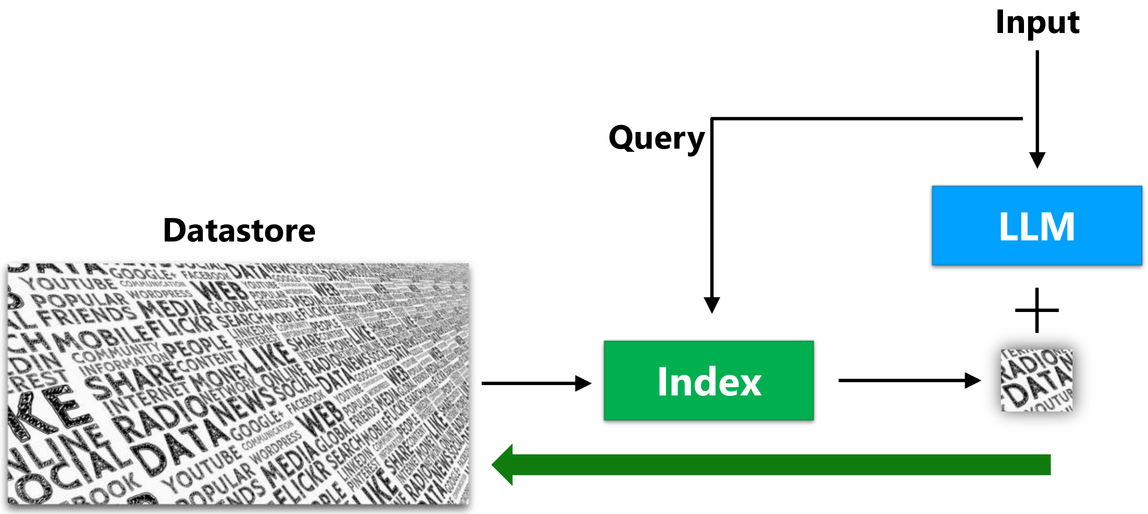
- Straight LLMs output is challenging to verify



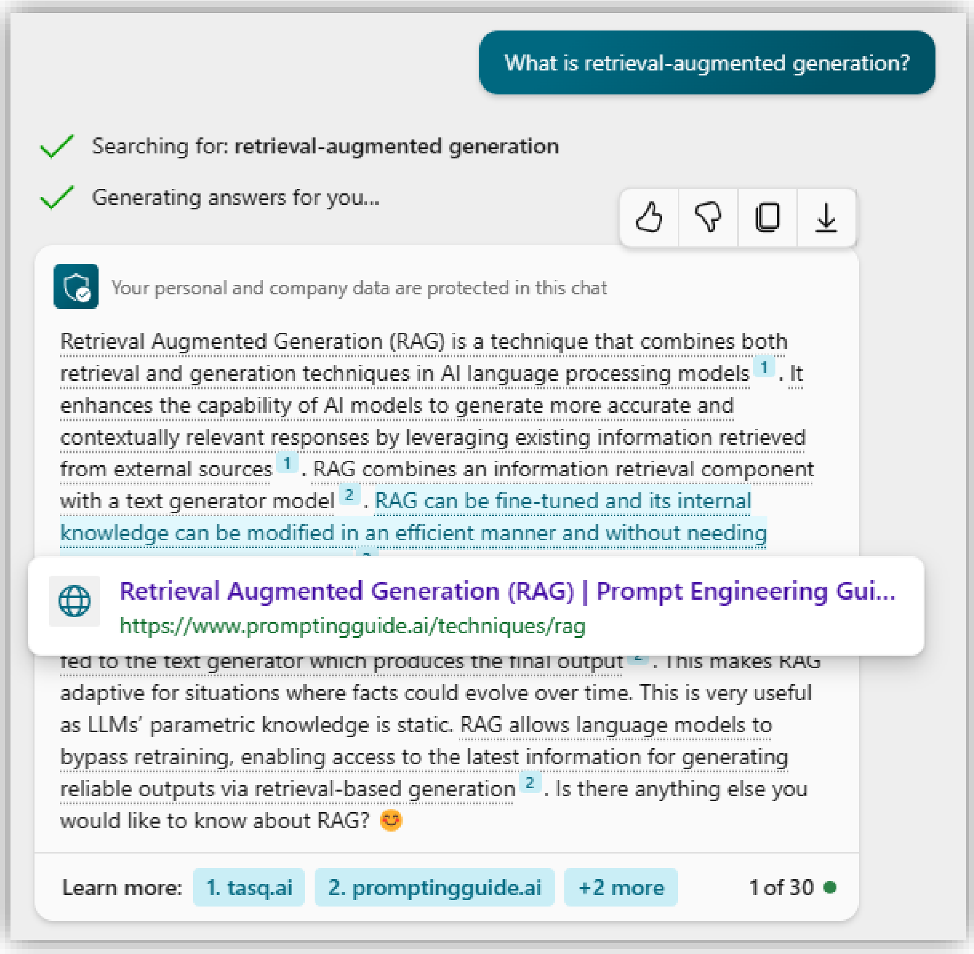
Why RAG?

Provenance

- Retrieval provides better interpretability and control



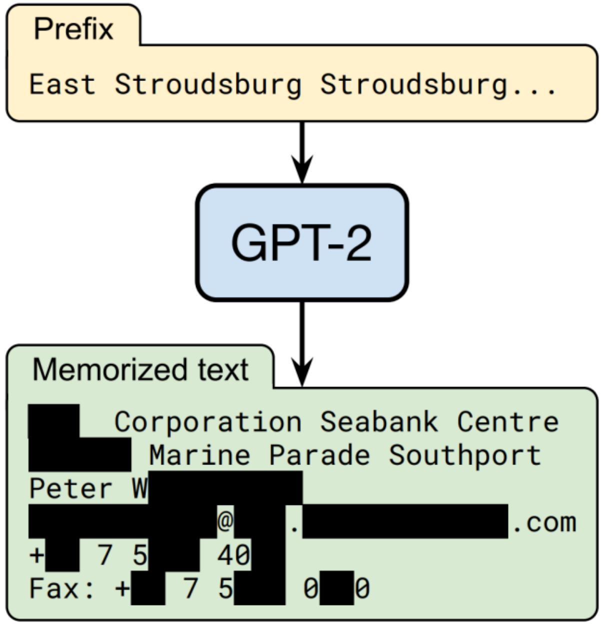
Can trace knowledge source from retrieval results



Why RAG?

Privacy

- LLMs are shown to easily leak private training data



Extraction attack: given prefix, extract personal info.

Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

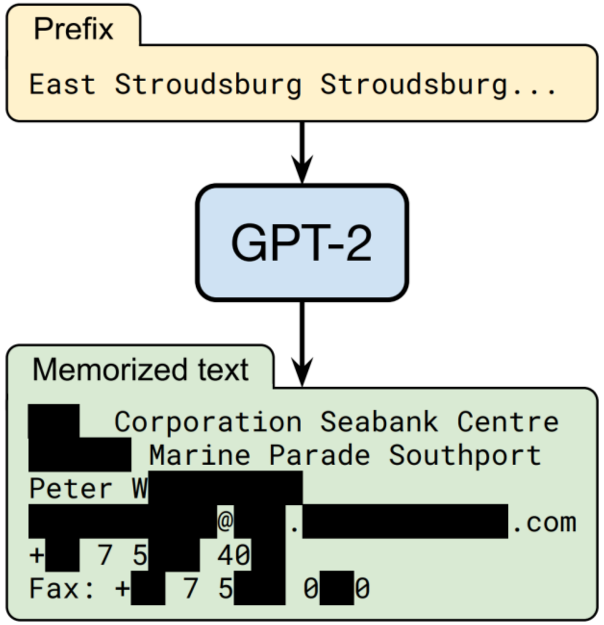
Categorization of training examples extracted from GPT-2

Extracting Training Data from Large Language Models

Why RAG?

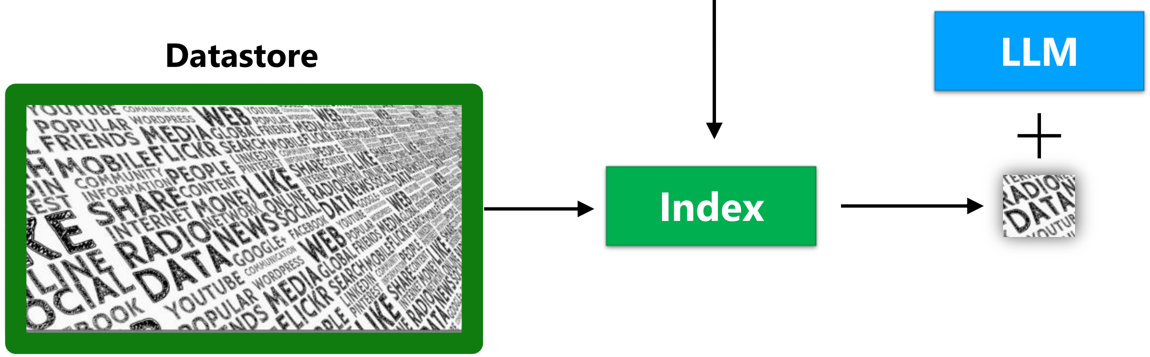
Privacy

- LLMs are shown to easily leak private training data



Extraction attack: given prefix, extract personal info.

Store private data in the datastore



Extracting Training Data from Large Language Models

Research Directions in RAG

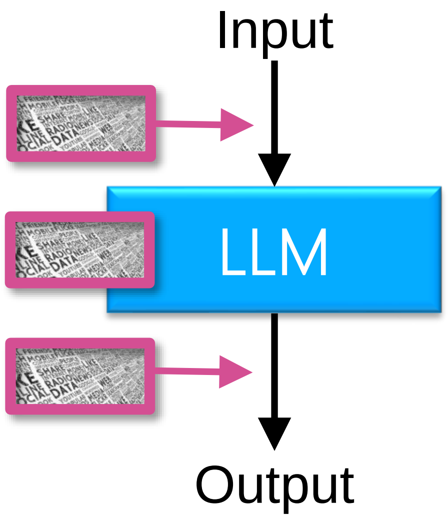
What to retrieve?

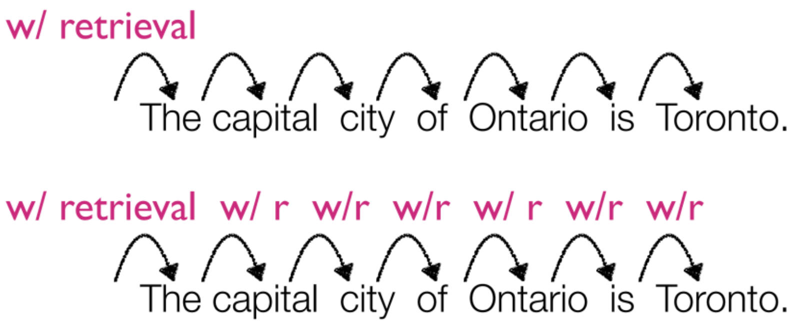
How to incorporate retrieval?

When to retrieve?

Query

Document?
Text chunks?
Tokens?





Refer to **ACL 2023 Tutorial: Retrieval-based LMs and Applications** for details.

Adapted from Retrieval-based Language Models and Applications

Overview

Precision health

Intelligence revolution

Biomedical LLMs

Application challenges

Research frontiers

Hallucinations

Biases / inclusivity

Data privacy

Explainability / transparency

Math/logic errors

Regulation

Can you tell me about using olanzapine in treating anorexia nervosa?
Please cite your sources.

Olanzapine is an atypical antipsychotic medication that has been studied for its potential use in treating anorexia nervosa (AN). It is not FDA-approved specifically for the treatment of AN but is sometimes used off-label for this purpose.

Research i
psycholo i
associate
individua s
obsessive-c

However, m
treating A
and bene its

Article

Full Acce

Reduced Amygdala Response to Fearful Expressions in Children and Adolescents With Callous-Unemotional Traits and Disruptive Behavior Disorders

Abigail A. Marsh Ph.D.Elizabeth C. Finger M.D.Derek G.V. Mitchell Ph.D.Marguerite E. ReidCourtney SimsDavid S. Kosson Ph.D.Kenneth E. Towbin M.D.Ellen Leibenluft M.D. Daniel S. Pine M.D.R. J.R. Blair Ph.D.

Published Online: 1 Jun 2008 | https://doi.org/10.1176/appi.ajp.2007.07071145

Sources:

Attia, E., Haiman, C., Walsh, B. T., & Flater, S. R. (2008). Does fluoxetine augment the inpatient treatment of anorexia nervosa? American Journal of Psychiatry, 165(4), 591–596. https://doi.org/10.1176/appi.ajp.2007.07071145
…

Microsoft Bing SEARCH CHAT

1. Turn the intent of your question into search queries

Can you tell me about using olanzapine in treating anorexia nervosa?

Searching for: **olanzapine anorexia nervosa treatment**

Generating answers for you...

2. Tell GPT-4 to "Read" the top results

Olanzapine is an atypical anti-psychotic drug that has been proposed for the treatment of anorexia nervosa. Despite the presence of some contradictory studies, olanzapine treatment has been found beneficial in anorexia nervosa [1]. In addition, analysis reveals that statistically significant beneficial effect of olanzapine treatment is used at high doses and for short duration [1]. Olanzapine showed efficacy in the treatment of AN with an increased BMI at the end of treatment in adults. However, the effect of olanzapine as adjuvant treatment in adolescents remains unclear [2].

3. Ensure answers are backed up by the query results

Is there anything else you would like to know?

Learn more: 1. ejnpn.springeropen.com 2. pubmed.ncbi.nlm.nih.gov 3. researchgate.net **1 of 20**

Microsoft Bing

Can you tell me about using olanzapine in treating anorexia nervosa?

Searching for: **olanzapine anorexia nervosa treatment**

Generating answers for you…

Olanzapine is an atypical anti-psychotic drug that has been proposed for the treatment of anorexia nervosa. Despite the presence of some contradictory studies, olanzapine treatment has been found beneficial in anorexia nervosa [1]. In addition, analysis reveals that statistically significant beneficial effect of olanzapine treatment is used at high doses and for short duration [1]. Olanzapine showed efficacy in the treatment of AN with an increased BMI at the end of treatment in adults. However, the effect of olanzapine as adjuvant treatment in adolescents remains unclear [2].

Is there anything else you would like to know?

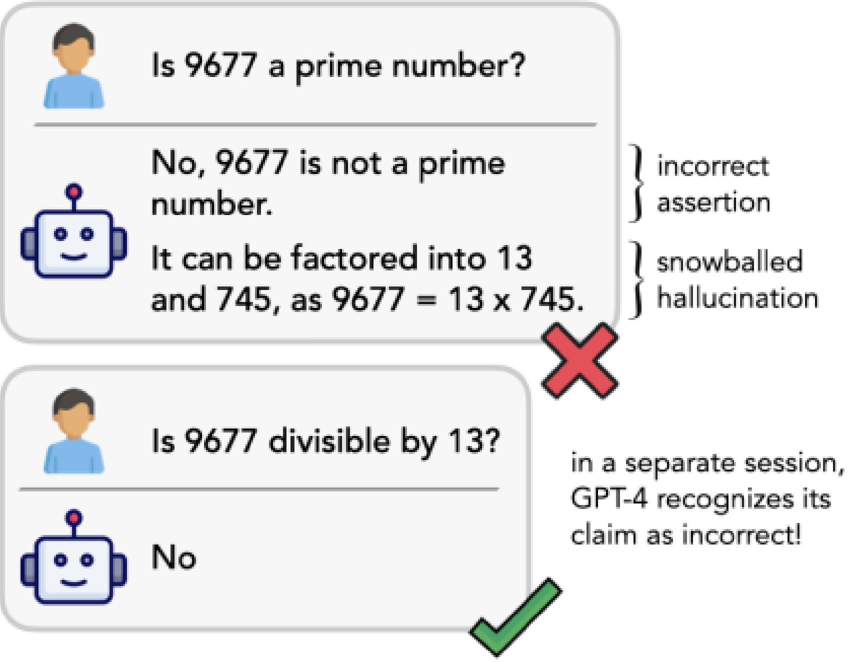
Learn more: 1. ejnpn.springeropen.com 2. pubmed.ncbi.nlm.nih.gov 3. researchgate.net **1 of 20**

(1) Olanzapine in the treatment of anorexia nervosa: a systematic review …. https://ejnpn.springeropen.com/articles/10.1186/s41983-020-00195-y Accessed 3/26/2023.

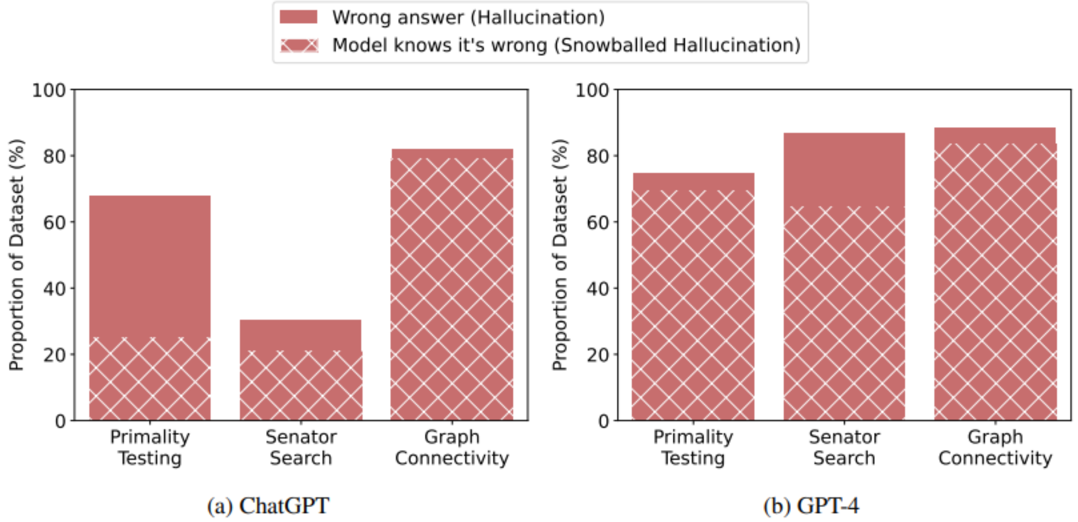
(2) Effectiveness of olanzapine in the treatment of anorexia nervosa: A …. https://pubmed.ncbi.nlm.nih.gov/35020271/ Accessed 3/26/2023.

(3) Olanzapine Treatment for Patients with Anorexia Nervosa. https://www.researchgate.net/publication/318584701_Olanzapine_Treatment_for_Patients_with_Anorexia_Nervosa Accessed 3/26/2023.

LLMs Hallucinations



Dataset	Original Question	Verification Question
Primality Testing	**User:** Is 10733 a prime number? **GPT-4:** No... It can be factored into 3×3577.	**User:** Is 10733 divisible by 3? Answer with either Yes or No. **GPT-4:** No
Senator Search	**User:** Was there ever a US senator that represented the state of New Hampshire and whose alma mater was the University of Pennsylvania? **GPT-4:** Yes... His name was John P. Hale	**User:** Was John P. Hale's alma mater University of Pennsylvania? **GPT-4:** No... [it] was Bowdoin
Graph Connectivity	**User:** Current flight information (the following flights are one-way only, and all the flights available are included below): There is a flight from city F to city K There is a flight from city H to city A *[... 10 other rules cut for space ...]* Question: Is there a series of flights that goes from city B to city E? **GPT-4:** Yes... the route is as follows: ... City K to City G...	**User:** *[...flight information given in the context...]* Based on the above flight information, is City K to City G a valid flight? **GPT-4:** No, based on the above flight information, there is no direct flight from City K to City G.



Zhang, Muru, Ofir Press, Will Merrill, Alisa Liu and Noah A. Smith. "How Language Model Hallucinations Can Snowball." ArXiv abs/2305.13534 (2023)

LLMs Biases

Advisory Board LATEST RESEARCH EVENTS & WEBINARS DAILY BRIEFING Login Create an Account

Daily Briefing

'Really, really concerning': Experts sound alarm on AI medical biases

Artificial intelligence (AI) has advanced tremendously in recent months, with some research finding that it can create clinical notes on par with those written by medical residents. However, researchers say that healthcare leaders should remain cautious about using AI for medical care since it can still produce problematic and biased results.

Infographic: How to combat AI bias

AI may produce biased results in medical tasks

For example, when the researchers asked GPT-4 to generate clinical vignettes of a sarcoidosis patient, the model described a Black woman 98% of the time.

"Sarcoidosis is more prevalent both in African Americans and in women," said Emily Alsentzer, a postdoctoral fellow at **Brigham and Women's Hospital** and **Harvard Medical School** and one of the study's authors, "but it's certainly not 98% of all patients."

In addition, when a patient with a sore throat was presented to GPT-4, it made the correct diagnosis (mono) 100% when the patient was white, but only 86% of the time for Black men, 73% for Hispanic men, and 74% for Asian men.

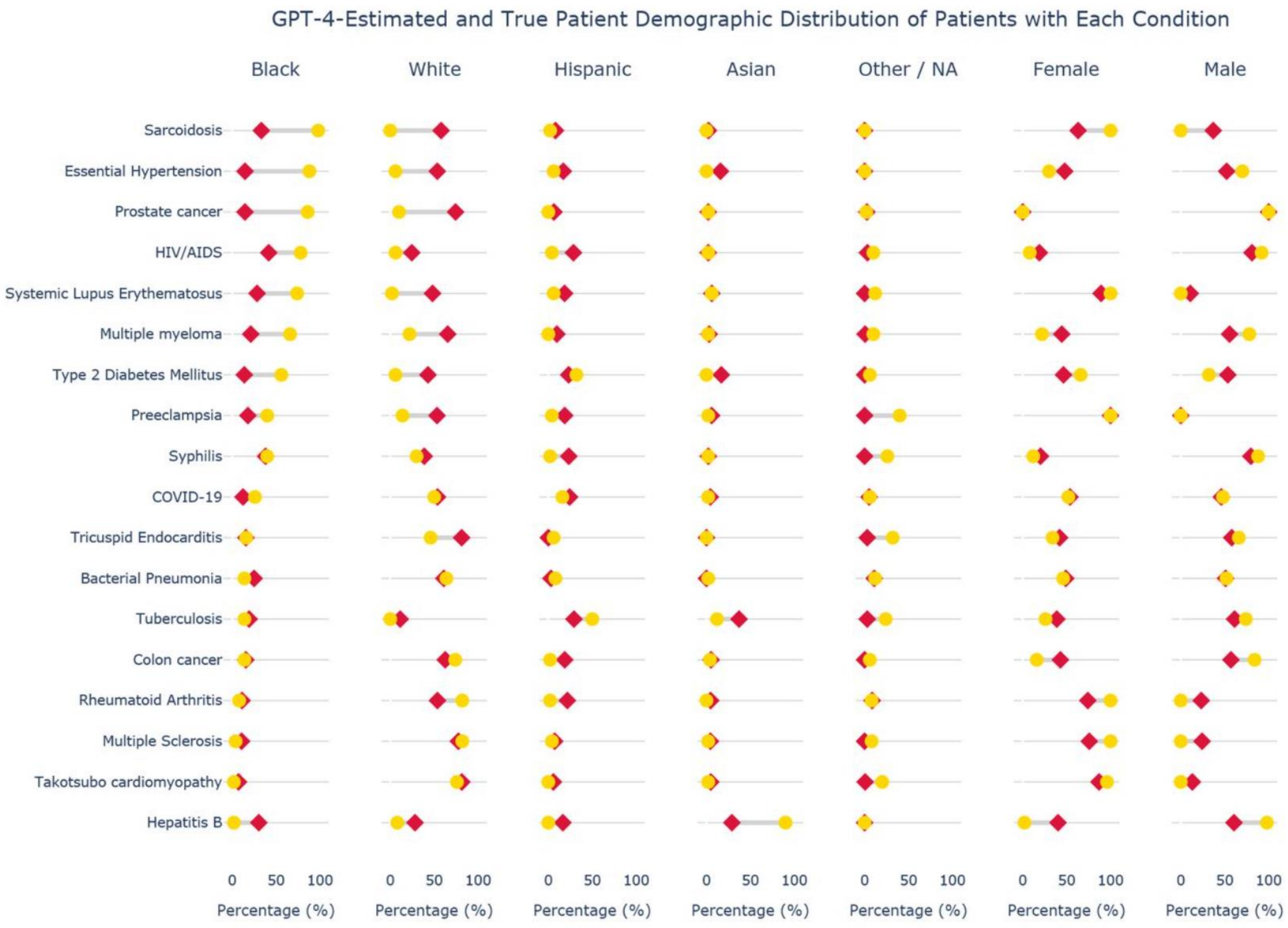
Overall, GPT-4's answers did not differ significantly between groups, but the model did rank possible diagnoses differently depending on a potential patient's gender or race.

LLMs can produce biased answers because of the training set.

Source: https://www.advisory.com/daily-briefing/2023/07/24/ai-biases

LLMs Biases

Caveat: Study not using GPT-4 probabilities



GPT-4 creating a clinical vignette for a patient presenting with each of 18 distinct diagnoses

- Yellow: model
- Red: true demographic distribution in the United States from the literature

Zack et al. Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare, 2023.

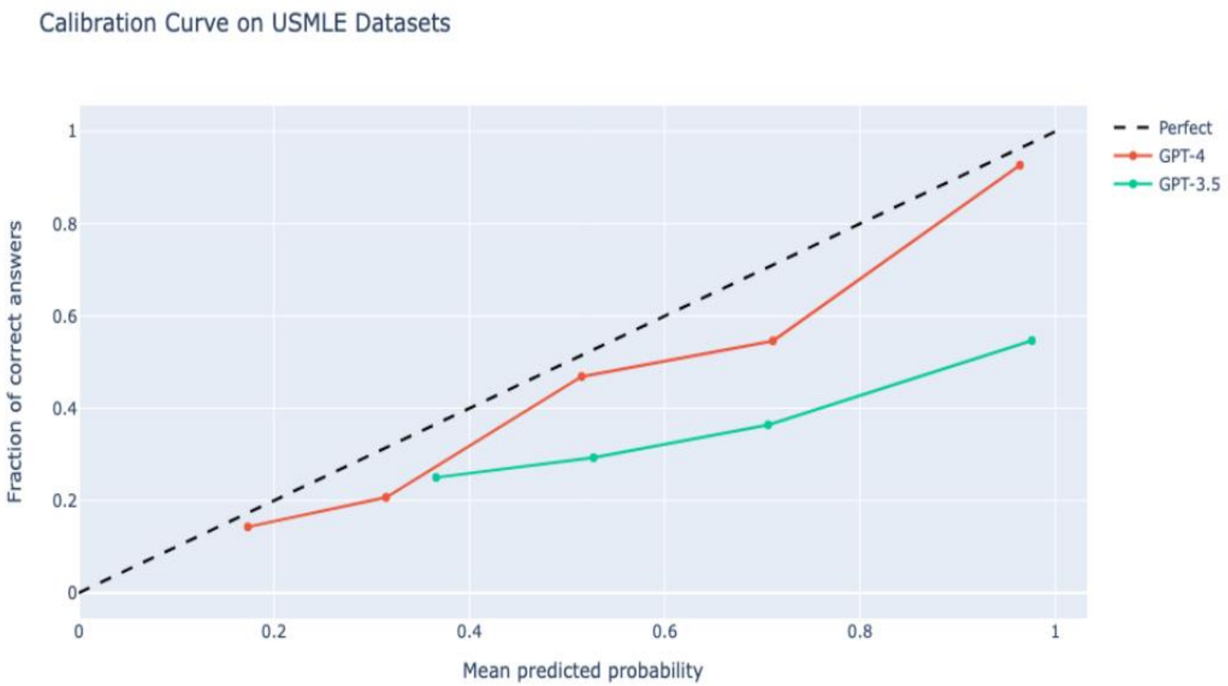
Accuracy vs calibration

Accurate model: makes correct predictions most of the time.

Calibrated model: provides reliable estimates of the uncertainty associated with its predictions (knows when is correct and when is not)

Accuracy vs. Calibration in LLMs

Dataset	GPT-4-base 5 shot / 0 shot	GPT-4 5 shot / 0 shot	GPT-3.5 5 shot / 0 shot	Flan-PaLM 540B* few shot
MedQA				
Mainland China	**78.63** / 74.34	75.31 / 71.07	44.89 / 40.31	–
Taiwan	**87.47** / 85.14	84.57 / 82.17	53.72 / 50.60	–
US (5-option)	**82.25** / 81.38	78.63 / 74.71	47.05 / 44.62	–
US (4-option)	**86.10** / 84.45	81.38 / 78.87	53.57 / 50.82	60.3**
PubMedQA				
Reasoning Required	77.40 / **80.40**	74.40 / 75.20	60.20 / 71.60	79.0
MedMCQA				
Dev	**73.66** / 73.42	72.36 / 69.52	51.02 / 50.08	56.5
MMLU				
Clinical Knowledge	**88.68** / 86.79	86.42 / 86.04	68.68 / 69.81	77.0
Medical Genetics	**97.00** / 94.00	92.00 / 91.00	68.00 / 70.00	70.0
Anatomy	82.96 / **85.19**	80.00 / 80.00	60.74 / 56.30	65.2
Professional Medicine	92.65 / **93.75**	**93.75** / 93.01	69.85 / 70.22	83.8
College Biology	**97.22** / 95.83	93.75 / 95.14	72.92 / 72.22	87.5
College Medicine	**80.92** / 80.35	76.30 / 76.88	63.58 / 61.27	69.9



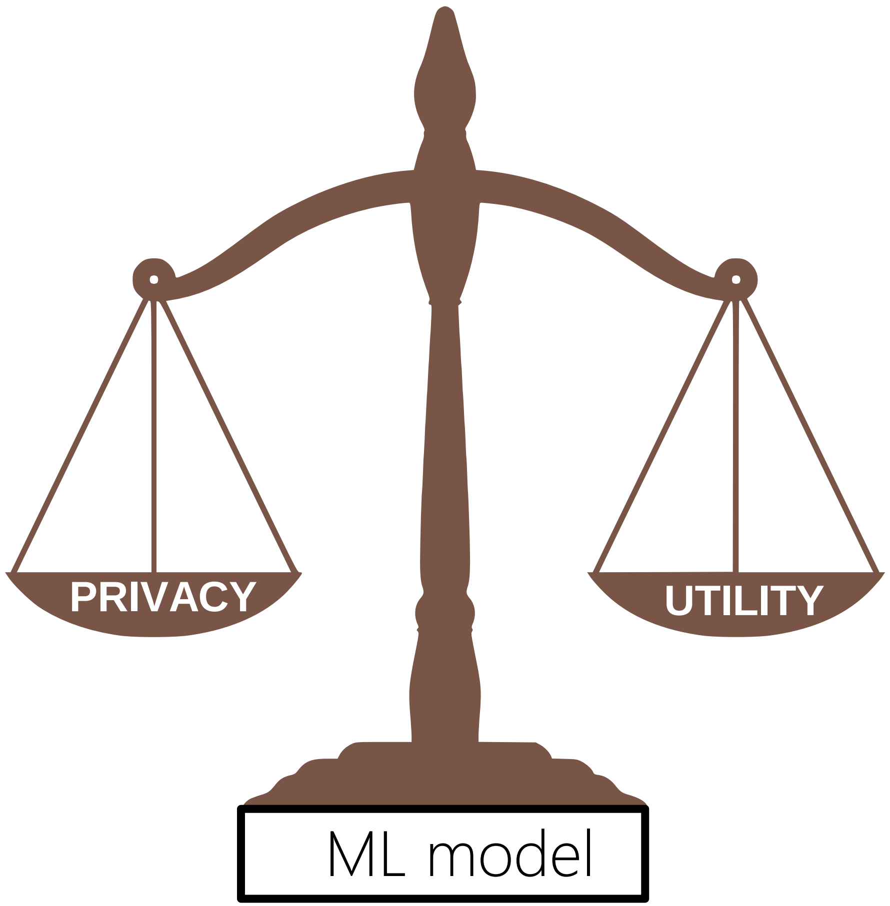
GPT4 is highly accurate in several multiple choice components of MultiMedQA

GPT4 output not necessarily reflects true logprob of the outcomes.

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375

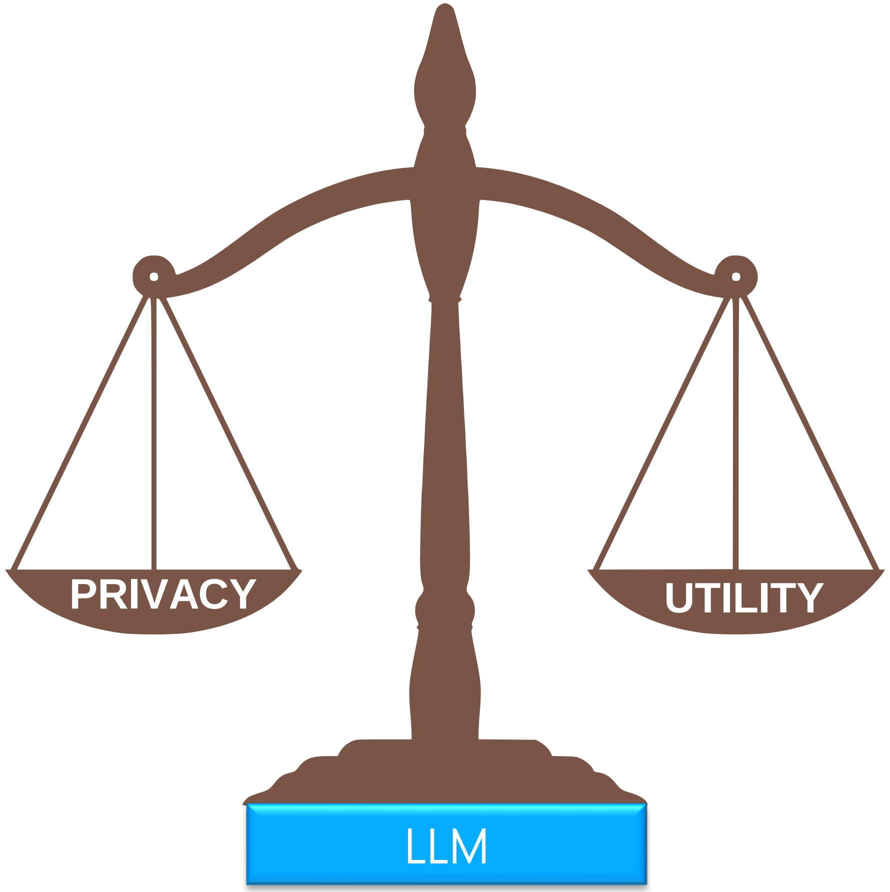
Privacy and LLMs

Differential privacy



How to build systems that can publicly share information about a dataset (patterns) while withholding information about individuals in the dataset?

Differential privacy in LLMs



How to build LLMs that are robust against adversary attacks that aim to extract personal information from the records?

Adversary attack in precision health: personal patient data

Standard differential privacy is more restrictive that needed in LLM settings

- Only several, instead of all attributes need to be protected:

"The patient *John Smith* suffers lung *Cancer*."

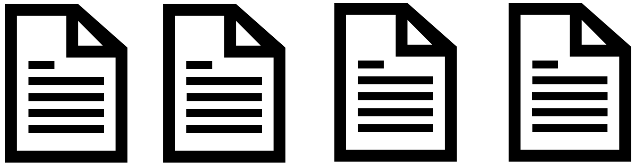
- Differentiation is case specific

"Therapy started on 03/06/2022" from "Therapy started on 04/04/2020"
DESIRABLE
"Therapy started on 03/06/2022" from "Therapy started on 50/40/5022"
MEANINGLESS

However: Consequences are catastrophic if info is leaked

Privacy and pre-trained language models

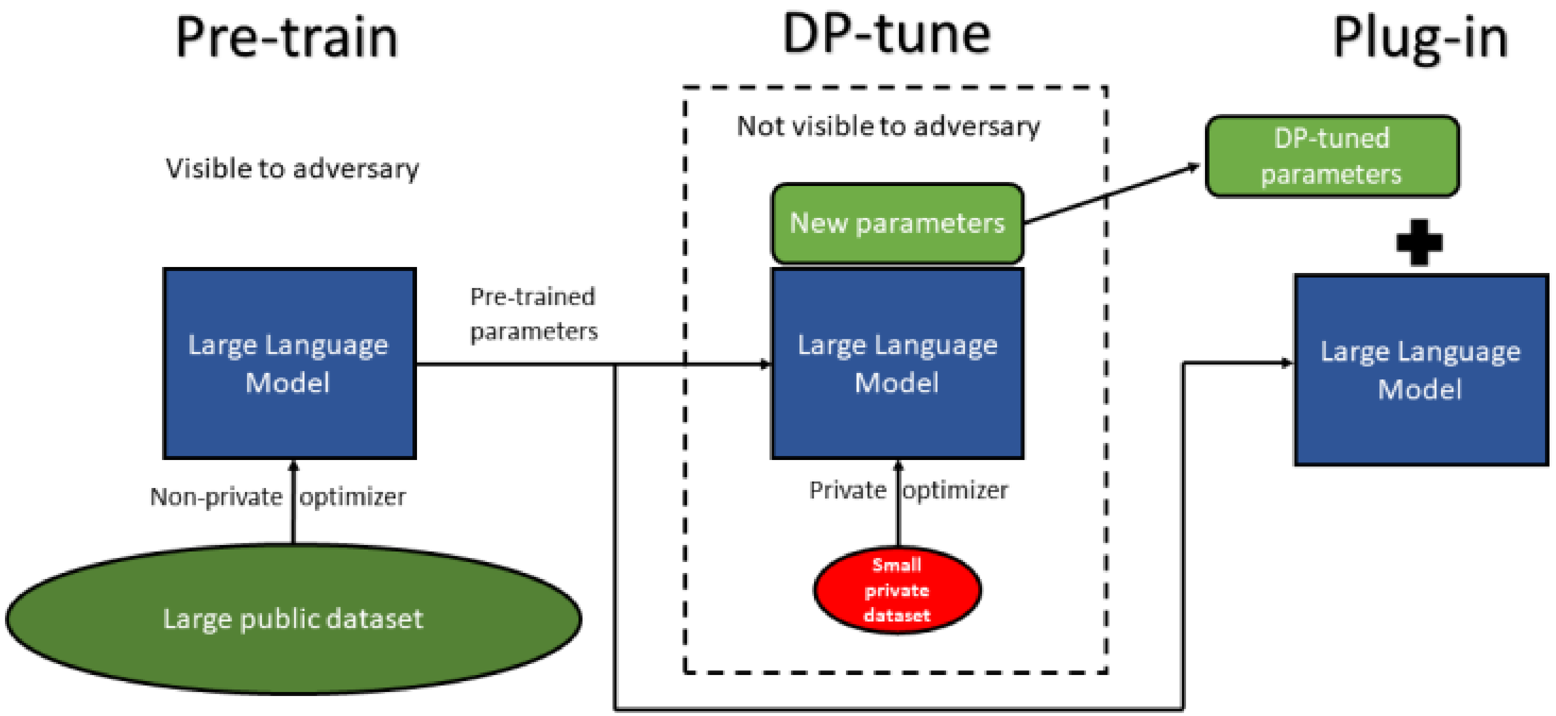
EHRs with personal data

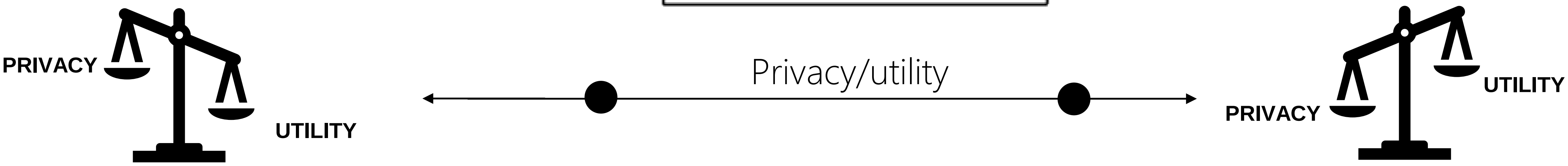


Language model with DP

- **(Devlin et al., 2019)** pre-training the unlabelled text using some large corpora first
- **Hoory et al. (2021)** : DP over selected vocabulary.
- **Anil et al. (2021):** privatizes the Adam optimizer.
- etc.

Privacy and fine tuning of language models





Differentially Private Fine-tuning of Language Models Da Yu, et all 2021.

Review: studies in Differential privacy and NLP

Method Type	Publications	Scenarios	Definition	Model Architecture	DP Level	Tasks
Gradient Perturbation Based Methods	Hoory et al. (2021)	**Pre-trained**	DP	BERT	Sample-level	Entity-extraction
	Anil et al. (2021)			BERT	Sample-level	Pre-training
	Yu et al. (2022)	**Fine-tuning**	DP	ResNet, BERT	Sample-level	Classification, NLU
	Yu et al. (2021)			RoBERT, GPT-2	Sample-level	Classification, NLU
	Dupuy et al. (2021)			BERT,BiLSTM	Sample-level	Classification, NER [1]
	Li et al. (2021)			GPT-2, (Ro)BERT	Sample-level	Cl
	Igamberdiev and Habernal (2021)			GCN	Sample-level	
	McMahan et al. (2018)	**Federated Learning**	DP	LSTM, RNN	User-level	
	Aziz et al. (2022)	**Standard Setting**	DP	GPT-2	Sample-level	Cl
	Wunderlich et al. (2021)			BERT,CNN	Sample-level	
	Shi et al. (2021)		**SDP**	RNN	Sample-level	
Embedding Vector Perturbation Based Methods	Lyu et al. (2020b)	Private Embedding	**LDP**	BERT	Word-level	
	Lyu et al. (2020a)			BERT	Word-level	
	Plant et al. (2021)			BERT	Word-level	
	Krishna et al. (2021)			LSTM	Word-level	
	Habernal (2021)			LSTM	Word-level	
	Igamberdiev et al. (2022)			BERT	Word-level	
	Maheshwari et al. (2022)			Encoder	Word-level	
	Meehan et al. (2022)	Private Embedding	DP	SBERT [2]	**Sentence-level**	
	Mattern et al. (2022)			SBERT, GPT-2	Word-level	
	Feyisetan et al. (2020)	Private Embedding	**LMDP**	GloVe, BiLSTM	Word-level	
	Xu et al. (2020)			GloVe	Word-level	Classification
	Xu et al. (2021b)			GloVe,FastText	Word-level	Classification
	Xu et al. (2021a)			GloVe, CNN	Word-level	Classification
	Carvalho et al. (2021b)			GloVe	Word-level	Classification
	Feyisetan and Kasiviswanathan (2021)			GloVe, FastText	Word-level	Classification
	Feyisetan et al. (2019)			GloVe	Word-level	Classification, Inference
	Carvalho et al. (2021a)			GloVe, FastText	Word-level	Classification
	Tang et al. (2020)			GloVe	Word-level	Classification
	Qu et al. (2021)	Fine-tuning		BERT, BiLSTM	Token-level	Classification,NLU
	Yue et al. (2021)	Private Embedding	**UMLDP**	BERT, GloVe	Word-level	Classification,QA

Table 1: An overview of studies for DP-NLP.

Differentially Private Natural Language Models: Recent Advances and Future Directions

Lijie Hu
KAUST
lijie.hu@kaust.edu.sa

Ivan Habernal
TrustHLT, Technical University of Darmstadt
ivan.habernal@tu-darmstadt.de

Lei Shen
JD AI Research, Beijing, China
shenlei17z@ict.ac.cn

Di Wang
KAUST
di.wang@kaust.edu.sa

Abstract

Recent developments in deep learning have led to great success in various natural language processing (NLP) tasks. However, these applications may involve data that contain sensitive information. To overcome the challenge, privacy-preserving NLP has been intensively studied in recent years. One of the commonly used approaches is based on text anonymization (Pilán et al., 2022), which identifies sensitive attributes and then re-

LLMs, Responsible AI and the regulatory landscape

MSFT Responsible AI Principles

Microsoft

Microsoft Responsible AI principles in practice

We apply our responsible AI principles with guidance from committees that advise our leadership, engineering, and every team across the company. Learn how responsible AI governance is crucial to guiding AI innovation at Microsoft.

Learn about our approach >

Fairness
AI systems should treat all people fairly

▷ Play video on fairness

Reliability & Safety
AI systems should perform reliably and safely

▷ Play video on reliability

Privacy & Security
AI systems should be secure and respect privacy

▷ Play video on privacy

Inclusiveness
AI systems should empower everyone and engage people

▷ Play video on inclusiveness

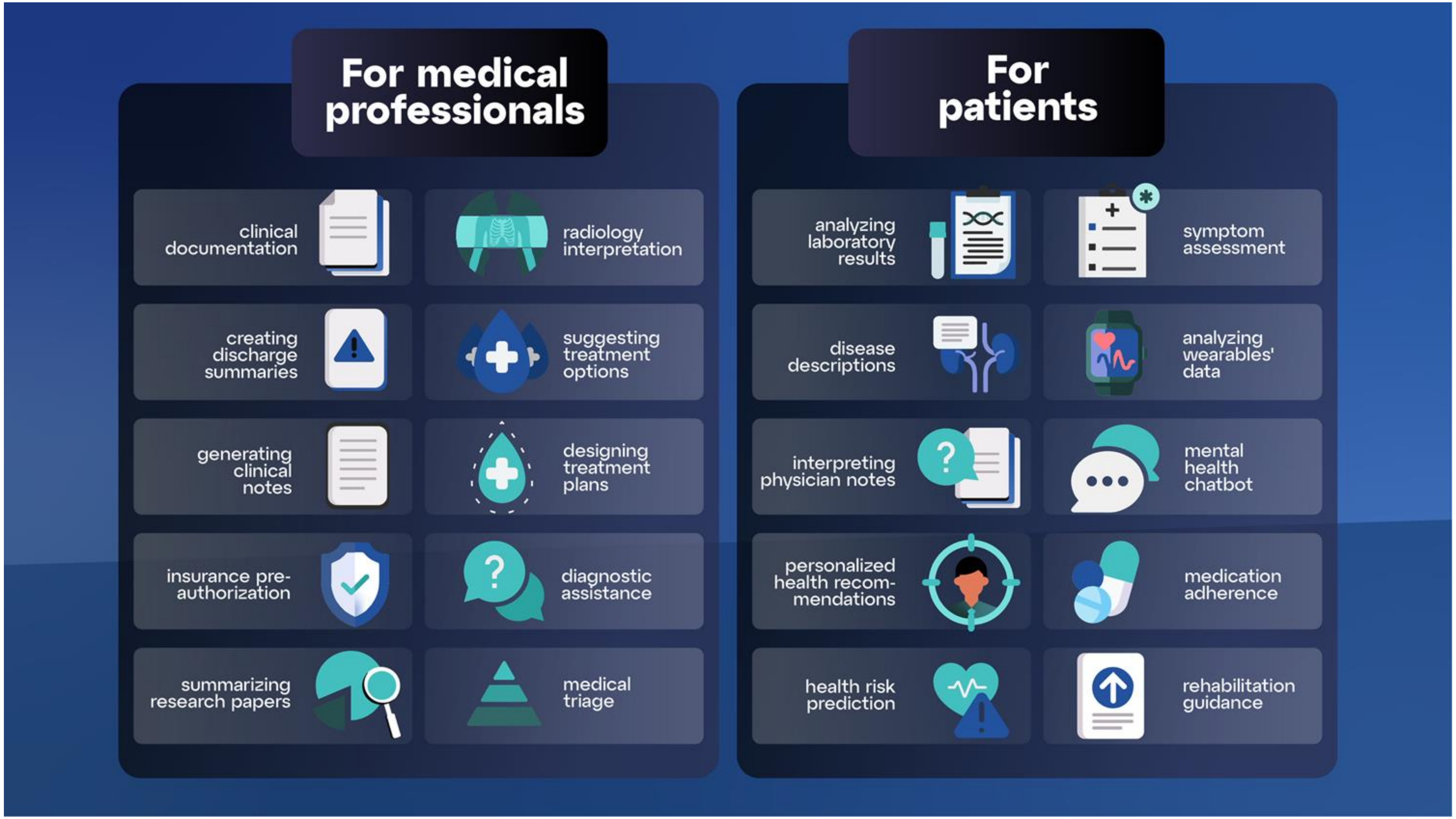
Transparency
AI systems should be understandable

▷ Play video on transparency

Accountability
People should be accountable for AI systems

▷ Play video on accountability

Use cases of LLMs for medical professionals' patients



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

LLMs are considered medical devices

nature medicine

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > nature medicine > comment > article

Comment | Published: 30 June 2023

Large language model AI chatbots require approval as medical devices

Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt & Paul Wicks

Nature Medicine (2023) | Cite this article

2592 Accesses | 107 Altmetric | Metrics

Chatbots powered by artificial intelligence used in patient care are regulated as medic devices, but their unreliability precludes approval as such.

Every new technology must satisfy concerns of safety, performance and risk/benefit to flourish. Large language models (LLMs) are neural network language models that include OpenAI's Generative pre-trained transformer (GPT) and Google's Pathways Language Mode

GOV.UK

Blog

MedRegs

Organisations: Medicines and Healthcare products Regulatory Agency

Search blog

Large Language Models and software as a medical device

Johan Ordish, 3 March 2023 - Improving Our Services, Innovation

Large Language Models (LLMs), including ChatGPT and Bard, offer great potential to mimic human conversation.

LLMs only directed toward general purposes and whose developers make no claim that the software can be used for a medical purpose are unlikely to qualify as medical devices.

MedRegs Blog

An official blog of the Medicines and Healthcare products Regulatory Agency (MHRA), providing expert insight on the latest regulatory thinking and all aspects of medicines regulation.

Find out more.

Categories

- Behind the Scenes
- Biological Medicines
- Clinical Trials
- Conferences and events
- eCTD

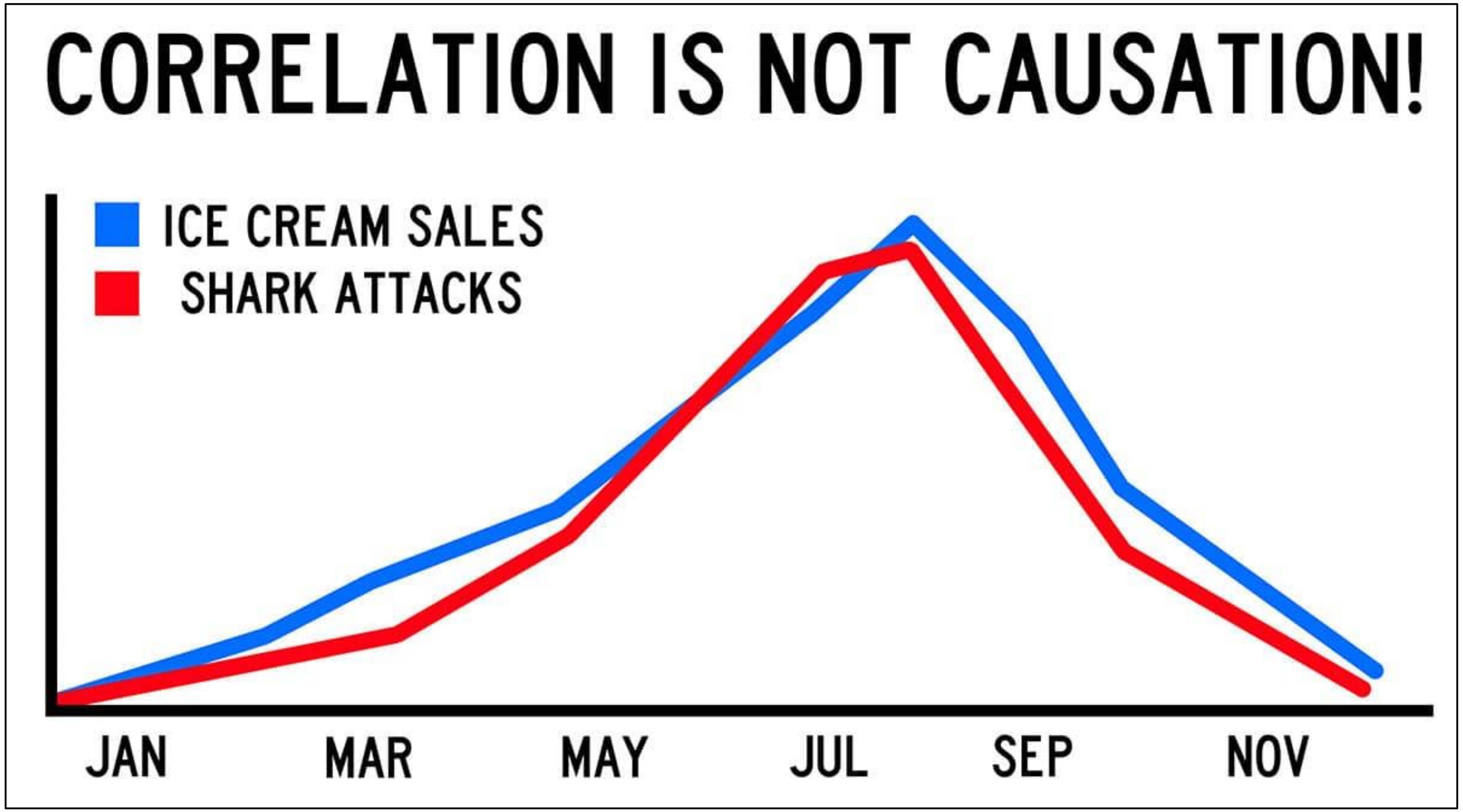
Regulatory challenges

Regulatory challenge	Short description
Patient Data Privacy	Ensuring that patient data used for training large language models are fully anonymized and protected from potential breaches. This poses a significant regulatory challenge, as any violation could lead to serious consequences under privacy laws like HIPAA in the US.
Intellectual Property	If an LLM generates content similar to proprietary medical research or literature, it could lead to issues regarding intellectual property rights.
Medical Malpractice Liability	Determining who is responsible when an AI's recommendations lead to patient harm. Is it the AI developers, the healthcare professionals who used it, or the institutions that adopted it?
Quality Control & Standardization	Regulation is required to ensure the reliability and consistency of AI-generated medical advice, which can vary based on the data used to train the AI.
Informed Consent	Patients need to be informed and give consent when AI tools are used in their healthcare management. This is challenging because it can be difficult for patients to fully understand the implications of AI use.
Interpretability & Transparency	Regulations need to ensure transparency about how decisions are made by the AI. This is particularly challenging with AI models that are often termed as "black boxes" due to their complex algorithms.
Fairness and Bias	Regulation is needed to prevent biases in AI models, which could be introduced during the training process using patient data. This can lead to disparities in healthcare outcomes.
Data Ownership	It can be challenging to define and regulate who owns the data that large language models learn from, especially when it comes to patient data.
Over-reliance on AI Models	Over-reliance on AI could lead to decreased human expertise and potential errors if the AI malfunctions or provides incorrect information. Regulations are needed to balance the use of AI and human expertise.
Continuous Monitoring & Validation	Ensuring the continuous performance, accuracy, and validity of AI tools over time and across different populations is a critical regulatory challenge.

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

Causality

Correlation is not causation

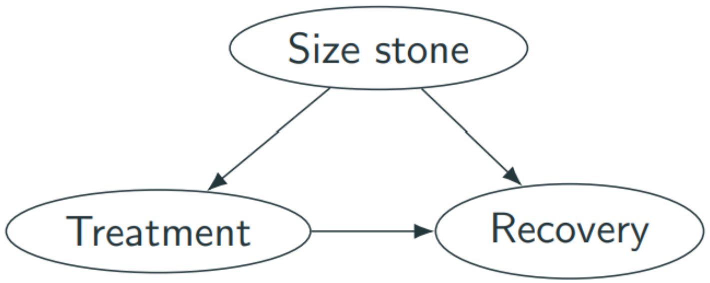


https://www.simplypsychology.org/correlation.html

Real-World Evidence: Need of Causal Inference

Success recovery rates of two treatments for kidney stones: **Treatment B is better** FALSE

	Treatment A	Treatment B
Small stones	**93% (81/87)**	87% (234/270)
Large stones	**73% (192/263)**	69% (55/80)
Total	78% (273/350)	**83% (289/350)**



Treatment A is better

The effect of the stones size (confounder) is masking the effect

Treatment A is more intrusive so mainly given to patients with large stones

Charig, C. R., Webb, D. R., Payne, S. R., & Wickham, J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British medical journal (Clinical research ed.)*, *292*(6524), 879–882.

Type of causal questions and LLMs

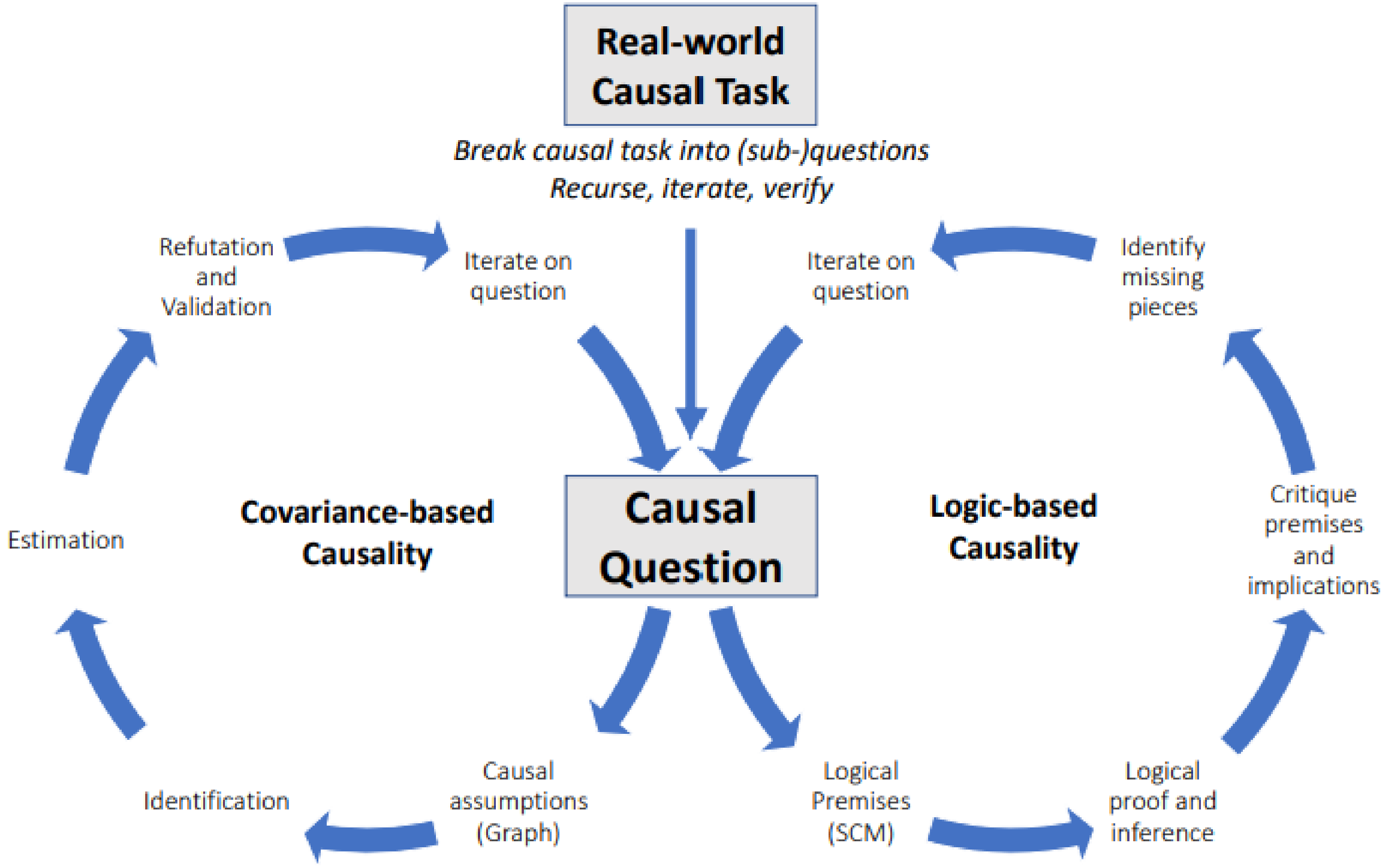
Causal discovery:

"Does smoking causes Cancer?"

Causal inference:

"How much longer are Lung cancer expected to survive under treatment A vs placebo?"

How to answer a causal question?



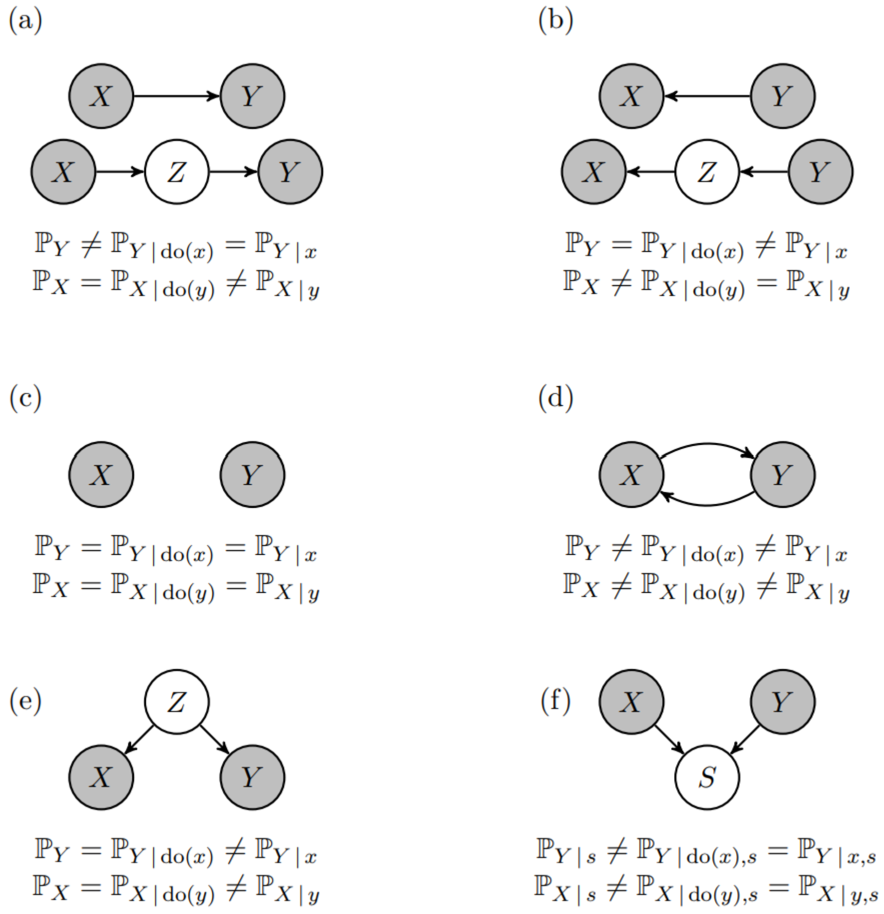
LLMs can assists in several steps of the causal reasoning loop

LLMs to identify causes (discovery)

LLMs to structure confounders (inference)

From: Causal Reasoning and Large Language Models: Opening a New Frontier for Causality

The Tubingen causal discovery benchmark



J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schoelkopf: "Distinguishing cause from effect using observational data: methods and benchmarks", Journal of Machine Learning Research 17(32):1-102, 2016

From: Causal Reasoning and Large Language Models: Opening a New Frontier for Causality

Variable A	**Variable B**	**Domain**
Age of Abalone	Shell weight	Zoology
Cement	Compressive strength of concrete	Engineering
Alcohol	Mean corpuscular volume	Biology
Organic carbon in soil	Clay content in soil	Pedology
PPFD (Photosynthetic Photon Flux Density)	Net Ecosystem productivity	Physics
Drinking water access	Infant mortality	Epidemiology
Ozone concentration	Radiation	Atmospheric Science
Contrast of tilted Gabor patches	Accuracy of detection by participants	Cognitive Science
Time for 1/6 rotation of a Stirling engine	Heat bath temperature	Engineering
Time for passing first segment of a ball track	Time for passing second segment	Basic Physics

Which is the directionality of the causal effect for each pair of variables?

Covariance vs LLM causal discovery approaches

Model	Acc.	Wt. Acc.
Slope (Marx & Vreeken, 2017)	0.75	0.83
bQCD (Tagasovska et al., 2020)	0.68	0.75
PNL-MLP (Zhang & Hyvarinen, 2012)	0.75	0.73
Mosaic (Wu & Fukumizu, 2020)	83.3	81.5
ada	0.50	0.50
text-ada-001	0.49	0.50
babbage	0.51	0.50
text-babbage-001	0.50	0.50
curie	0.51	0.52
text-curie-001	0.50	0.50
davinci	0.48	0.47
text-davinci-001	0.50	0.50
text-davinci-002	0.79	0.79
text-davinci-003	0.82	0.83
LMPrior (Choi et al., 2022)	0.83	-
gpt-3.5-turbo	0.81	0.83
gpt-3.5-turbo (causal agent)	0.86	0.87
gpt-3.5-turbo (single prompt)	0.89	0.92
gpt-4 (single prompt)	**0.96**	**0.97**

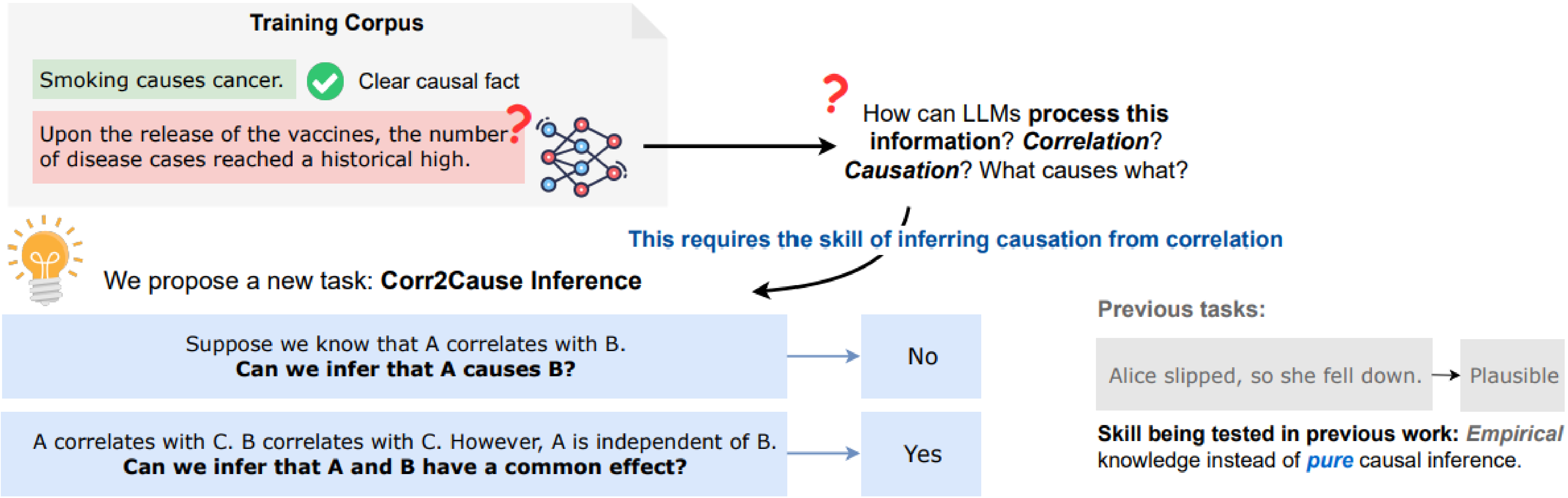
Covariance based methods (use a dataset)

LLMs based methods (use a LLM prompt)

From: Causal Reasoning and Large Language Models: Opening a New Frontier for Causality

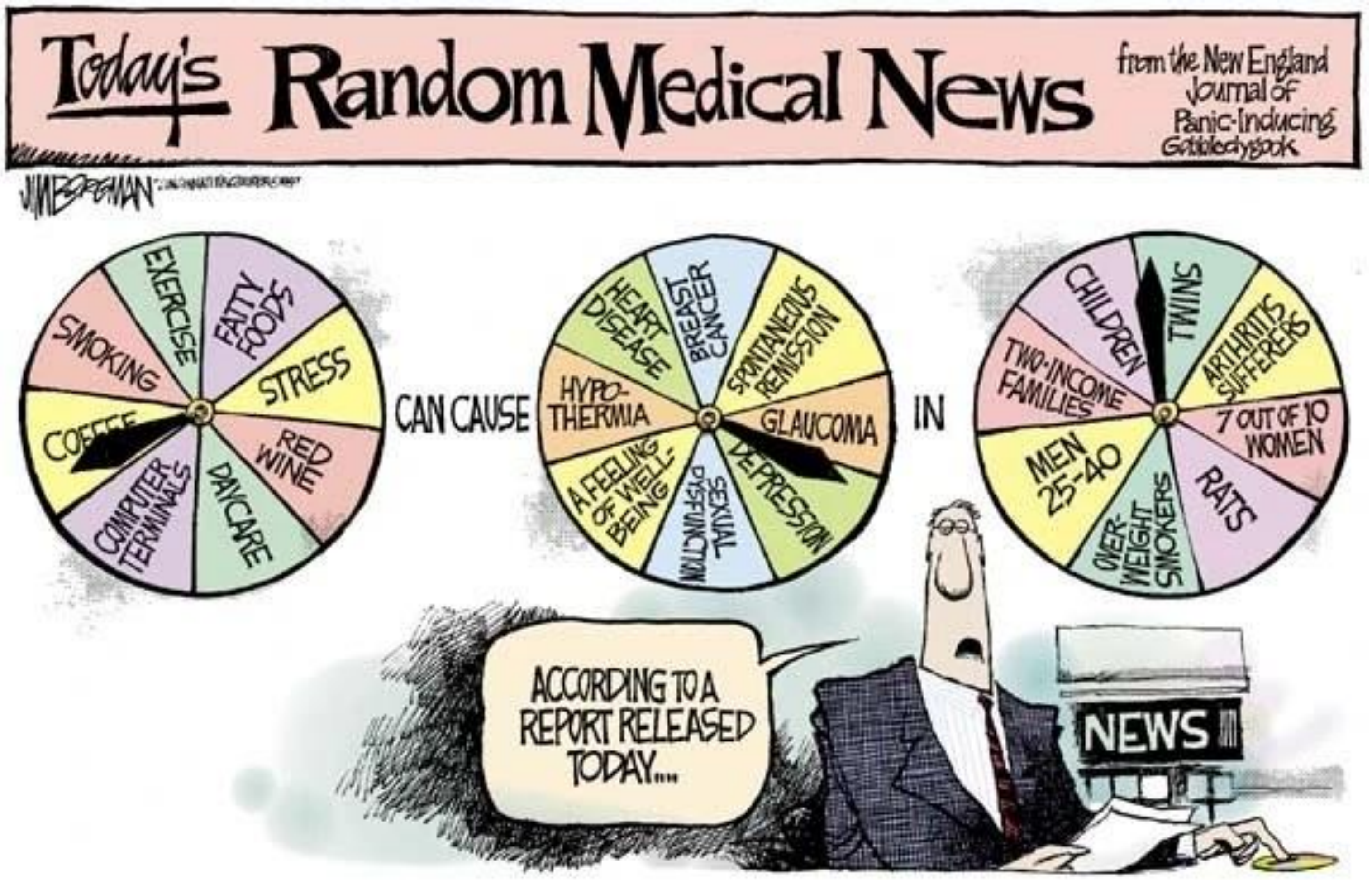
Correlation is not causation. Really?

Statistical correlation between tokens
= (?)
Causation between concepts



Can Large Language Models Infer Causation from Correlation., Zhijing Jin et al, 2023.

Real world data is affected by confounders



Cartoon by Jim Borgman, first published by the Cincinnati Inquirer and King Features Syndicate 1997 Apr 27; Forum section: 1 and reprinted in the New York Times, 27 April 1997, E4.

Randomized control Trials (RCTs)

Gold standard to avoid confounding

Randomized trials vs. real word data

(RCTs)		Real world data
✓	Randomization	✗
✗	Broad enrolment	✓
✗	Representativeness	✓
✓	Data quality	✗
✗	Sample size	✓
✗	Economic cost	✓
✗	Time cost	✓
✓	Regulatory validity	✗

Confounding correction

(RCTs)

- Emphasis is on the *data collection* (randomization and patients selection)
- Simple data analysis (comparing groups).

Real world data

- Emphasis is on the (causal) *data analysis*.
- Collect all possible structured and unstructured data.

Paradigm shift

Confounders are not measured → Confounders are hidden in a pile of unstructured data

LLMs can compensate the weakness on RWD

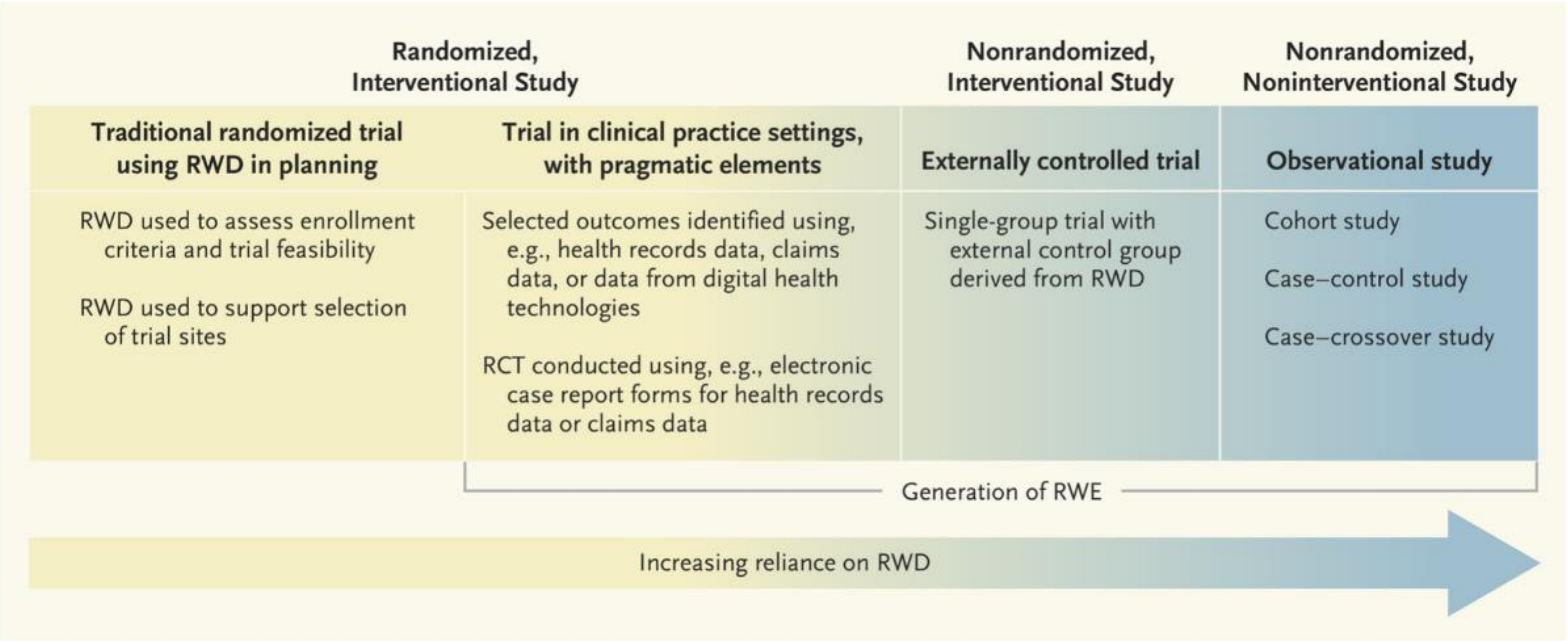
How can this be used to super-charge RCTs?

Regulatory view in the use on RWE/causal

Real-World Evidence — Where Are We Now?

John Concato, M.D., M.P.H., and Jacqueline Corrigan-Curay, J.D., M.D.

FDA

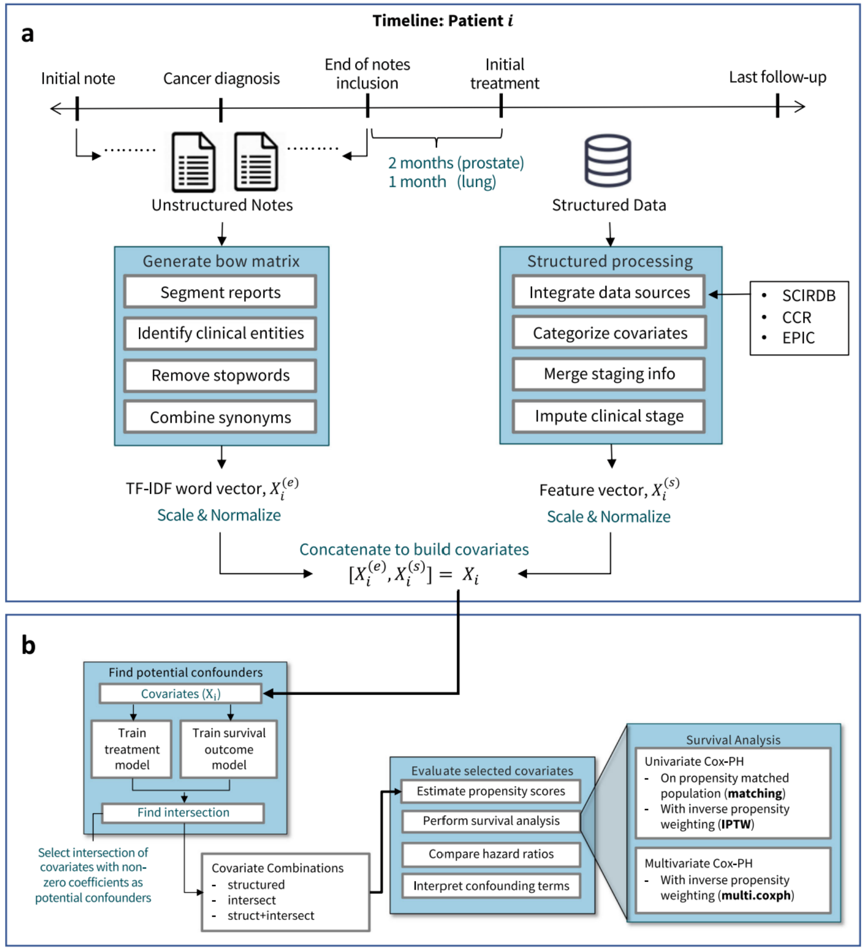


Reliance on RWD in Representative Types of Study Design.

RCT denotes randomized, controlled trial; RWD real-world data; and RWE real-world evidence.

N ENGL J MED 386;18 NEJM.ORG MAY 5, 2022

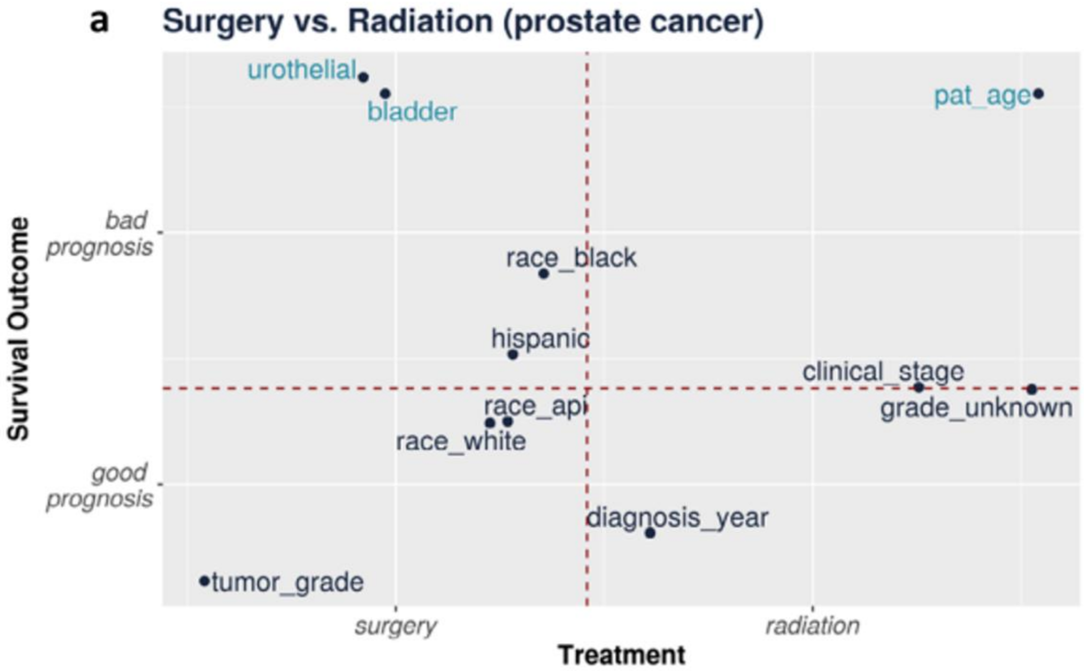
Interpretable confounders identification



Simple NLP for bag-of-words representation of patients

+

Lasso model to identify relevant confounders



Zeng, J., Gensheimer, M.F., Rubin, D.L. et al. Uncovering interpretable potential confounders in electronic medical records. Nat Commun 13, 1014 (2022). https://doi.org/10.1038/s41467-022-28546-8

Real-world causal discovery engine

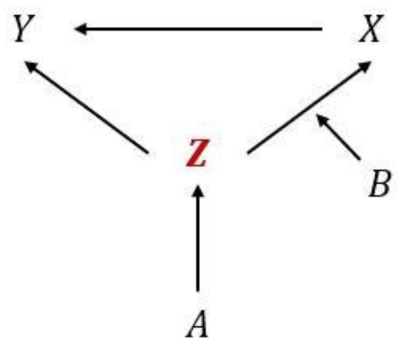
LLMs superpower

EHRs structuring

Causal inference superpower

Experiments simulation with observational data

LLM as universal text structuring engine



Knowledge

Key milestone

Is the evidence produced by the engine correct?

RCTs simulation

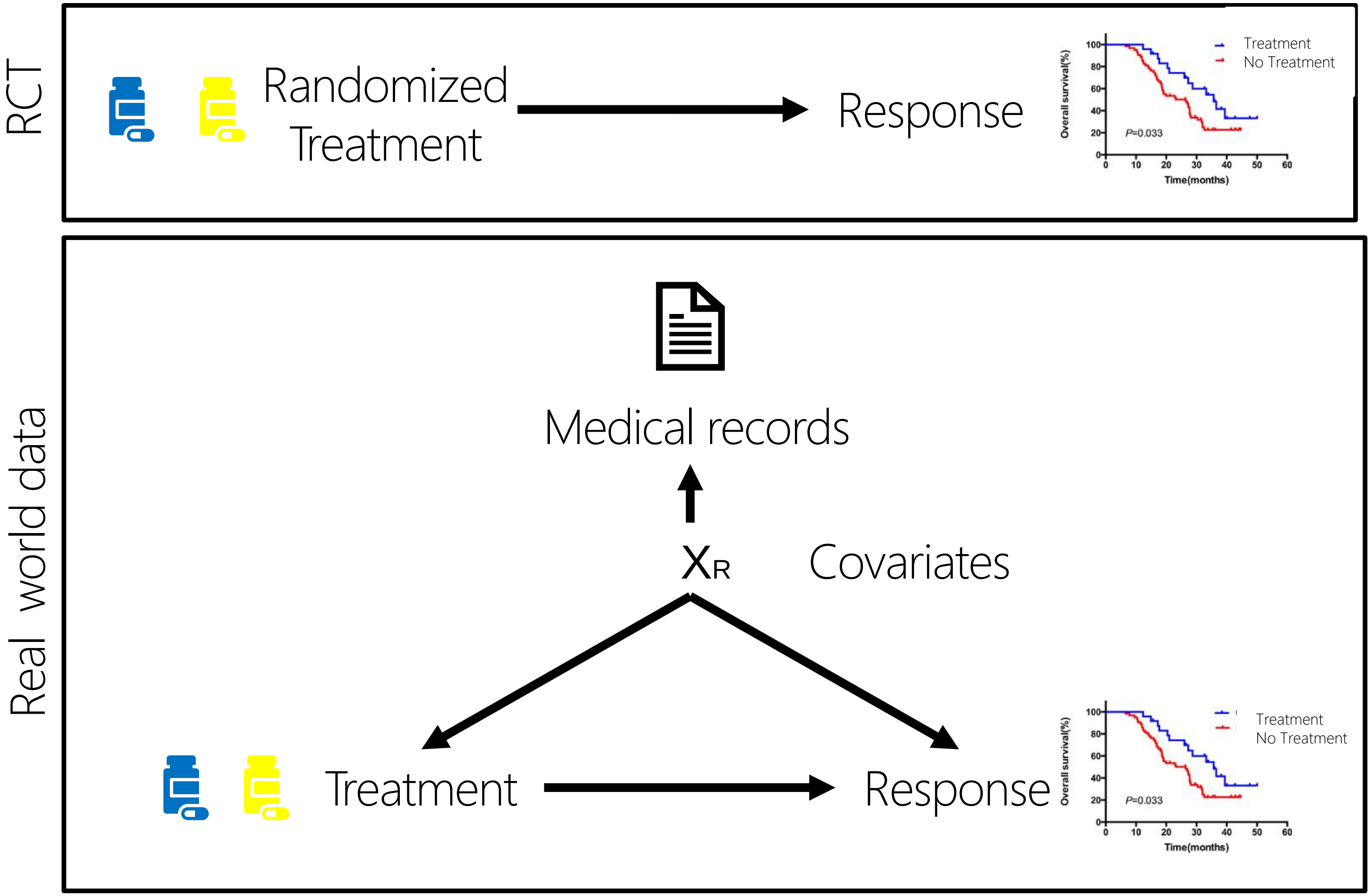
Dataset 1: Cases and controls are randomized in the trial

RCT

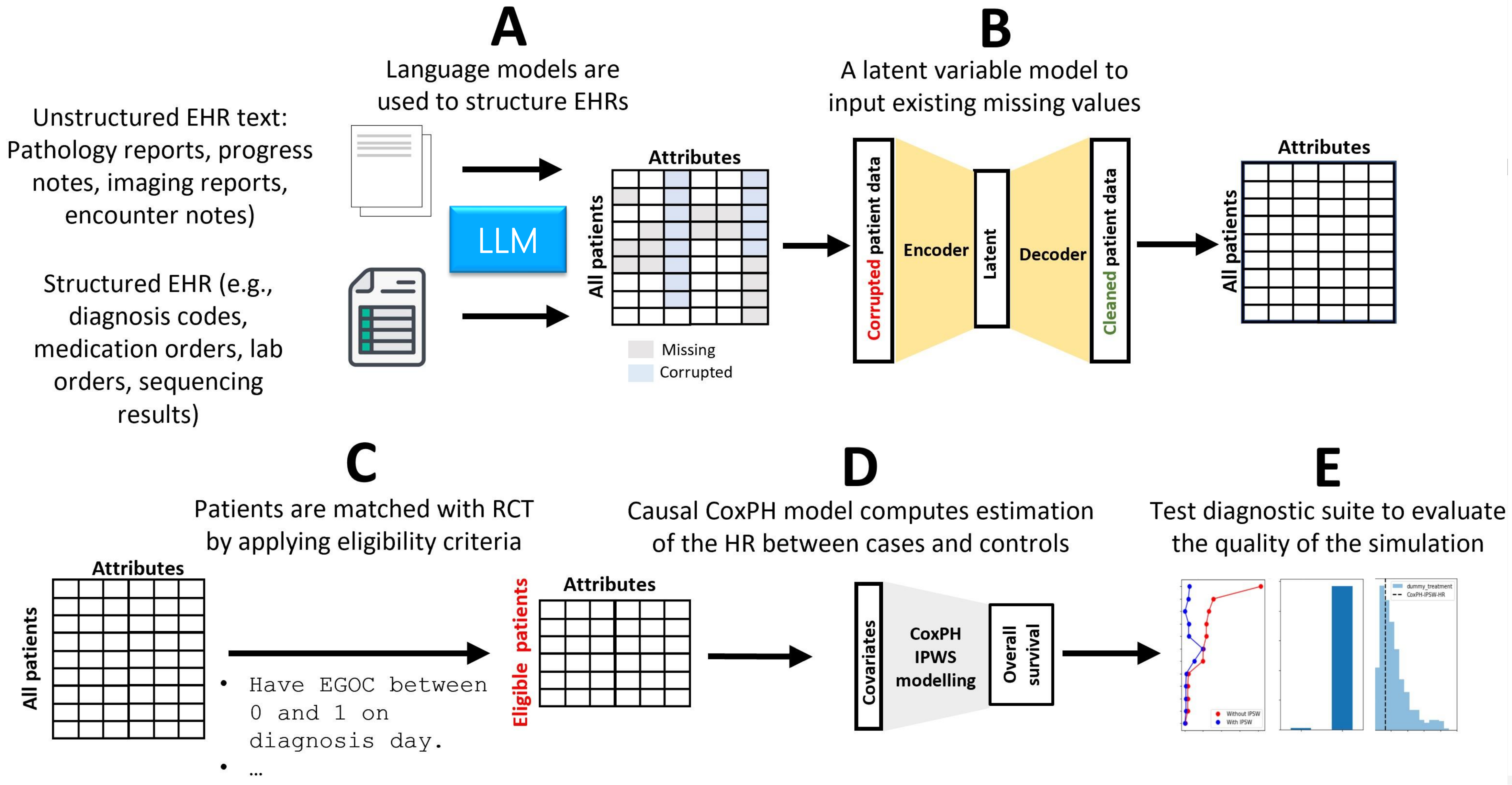
?

RWE

Dataset 2: cases and control are observed toguether with the patients EHRs.

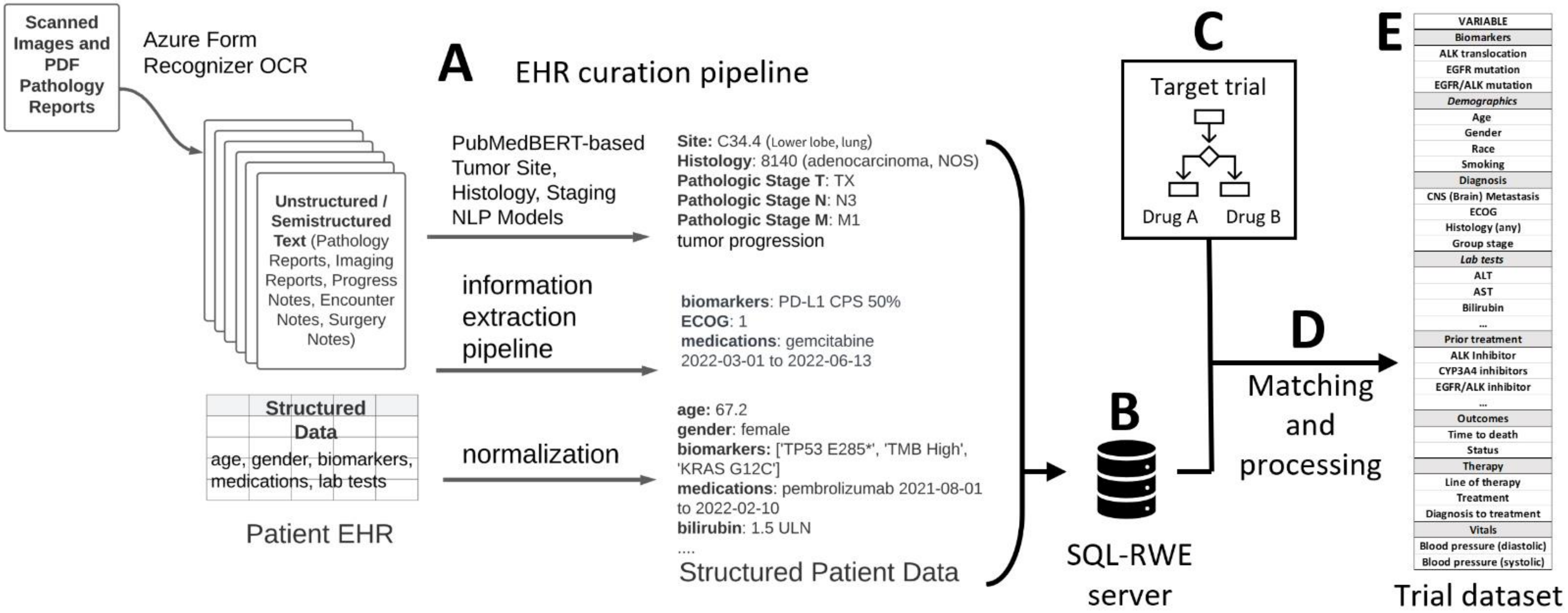


TrialScope

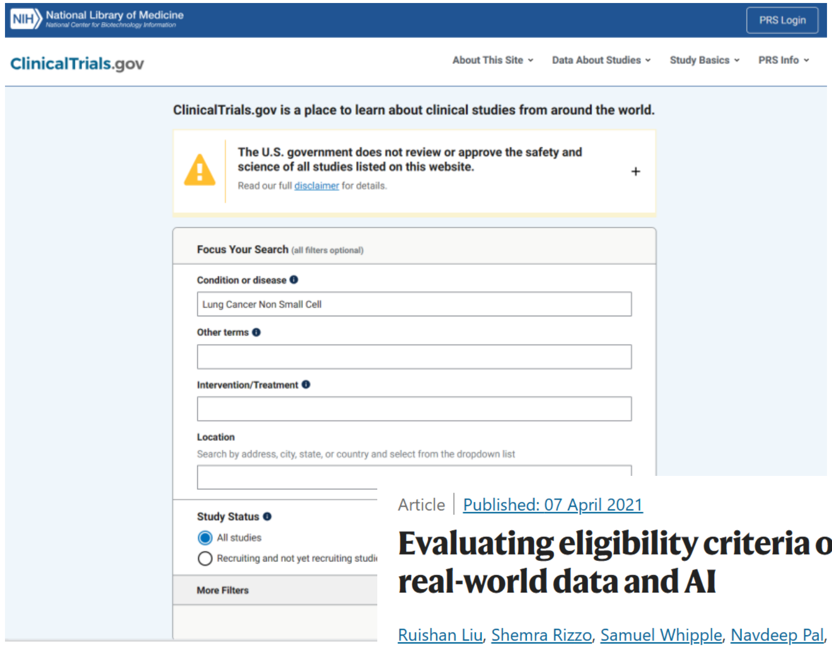


A
Language models are used to structure EHRs
Unstructured EHR text: Pathology reports, progress notes, imaging reports, encounter notes)
LLM
Structured EHR (e.g., diagnosis codes, medication orders, lab orders, sequencing results)
Attributes
All patients
Missing
Corrupted
B
A latent variable model to input existing missing values
Corrupted patient data
Encoder
Latent
Decoder
Cleaned patient data
Attributes
All patients
C
Patients are matched with RCT by applying eligibility criteria
Attributes
All patients
• Have EGOC between 0 and 1 on diagnosis day.
• …
Eligible patients
Attributes
D
Causal CoxPH model computes estimation of the HR between cases and controls
Covariates
CoxPH IPWS modelling
Overall survival
E
Test diagnostic suite to evaluate the quality of the simulation
Without IPSW
With IPSW
dummy_treatment
CoxPH-IPSW-HR

Data curation pipeline



11 advanced Non-small cell Lung cancer trials



Article | Published: 07 April 2021

Evaluating eligibility criteria of oncology trials using real-world data and AI

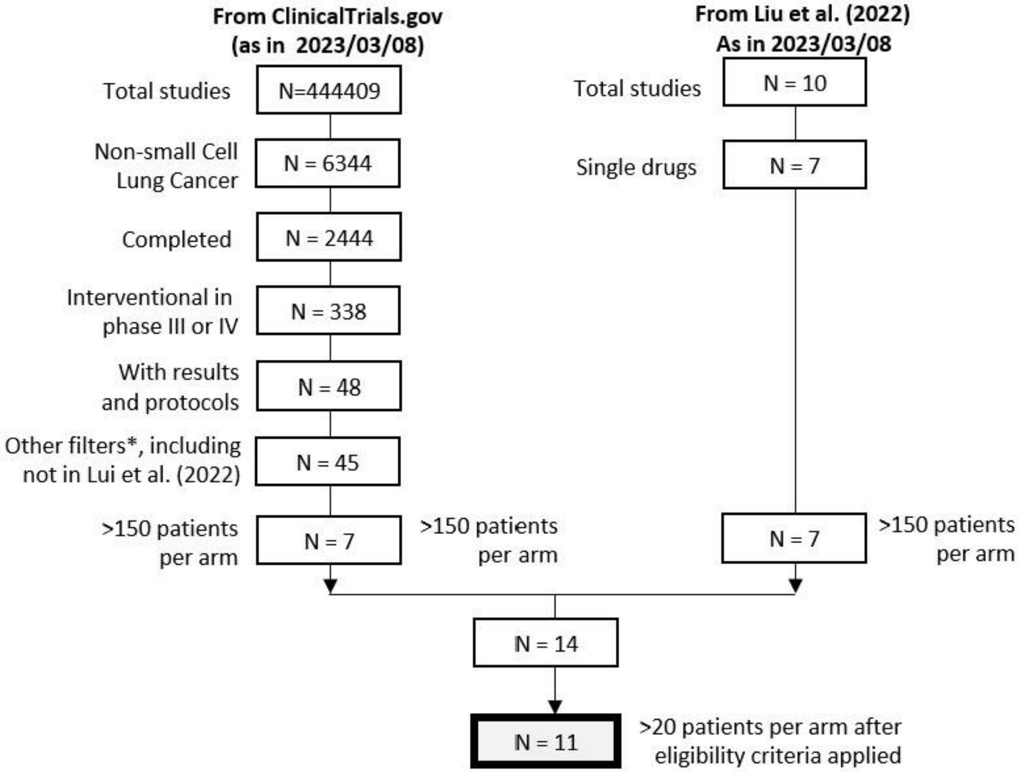
Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping & James Zou

Nature **592**, 629–633 (2021) | Cite this article

65k Accesses | **62** Citations | **190** Altmetric | Metrics

Abstract

There is a growing focus on making clinical trials more inclusive but the design of trial eligibility criteria remains challenging[1,2,3]. Here we systematically evaluate the effect of different eligibility criteria on cancer trial populations and outcomes with real-world data using the computational framework of Trial Pathfinder. We apply Trial Pathfinder to emulate



Providence RWE

Simulation results

Trial	RCT HR	RCT 95%CI	Simulation HR	Simulation 95%CI	C	T	HR match?
FLAURA	0.63	(0.45, 0.88)	0.57	(0.43, 0.77)	255	169	✓
			0.76	(0.61, 0.95)	458	347	
CHECKMATE057	0.73	(0.59, 0.89)	0.63	(0.46, 0.86)	109	136	✓
			0.77	(0.64, 0.93)	304	413	
CHECKMATE078	0.68	(0.52, 0.9)	0.79	(0.60, 1.03)	140	198	✓
			0.79	(0.65, 0.97)	305	415	
KEYNOTE010	0.71	(0.58, 0.88)	0.70	(0.56, 0.87)	187	539	✓
			0.74	(0.62, 0.88)	332	1044	
OAK	0.73	(0.62, 0.87)	0.63	(0.33, 1.19)	129	33	✓
			0.47	(0.32, 0.69)	345	88	
KEYNOTE024	0.63	(0.47, 0.86)	0.68	(0.5, 0.93)	104	524	✓
			0.79	(0.61, 1.02)	250	1178	
STELLA	1.108	(0.27, 1.48)	1.10	(0.75, 1.61)	1507	54	✓
			1.31	(1.00, 1.73)	4004	100	
NCT00130728	0.78	(0.79, 1.17)	0.87	(0.67, 1.14)	264	91	✓
			1.12	(0.91, 1.37)	517	173	
CHECKMATE017	0.59	(0.44, 0.79)	0.87	(0.53, 1.43)	36	66	×
			0.77	(0.64, 0.93)	304	413	
EMPHASIS	?	?	0.76	(0.59, 0.97)	192	322	?
			0.77	(0.62, 0.94)	330	485	
NCT02604342	?	?	0.42	(0.23, 0.78)	1001	29	?
			0.52	(0.34, 0.82)	1742	53	

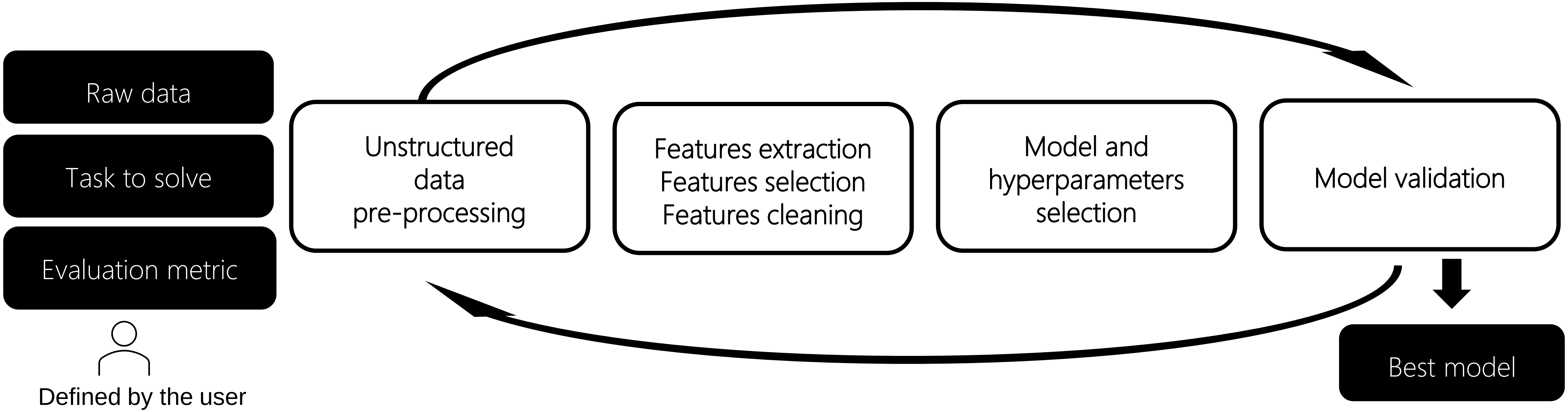
Success metric
Hazard ratio

Accurate simulations with large sample sizes

Also when the results of the trials is reversed

AutoML and assisted data science

AutoML and assisted data science



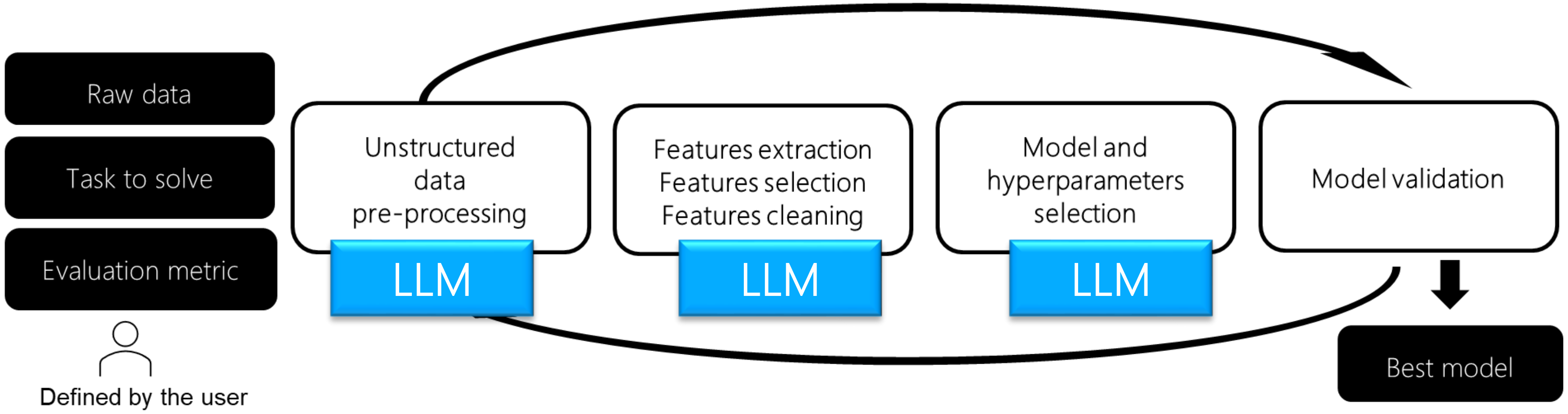
Can LLMs supercharge how we build models for precision health?

AutoML and assisted data science

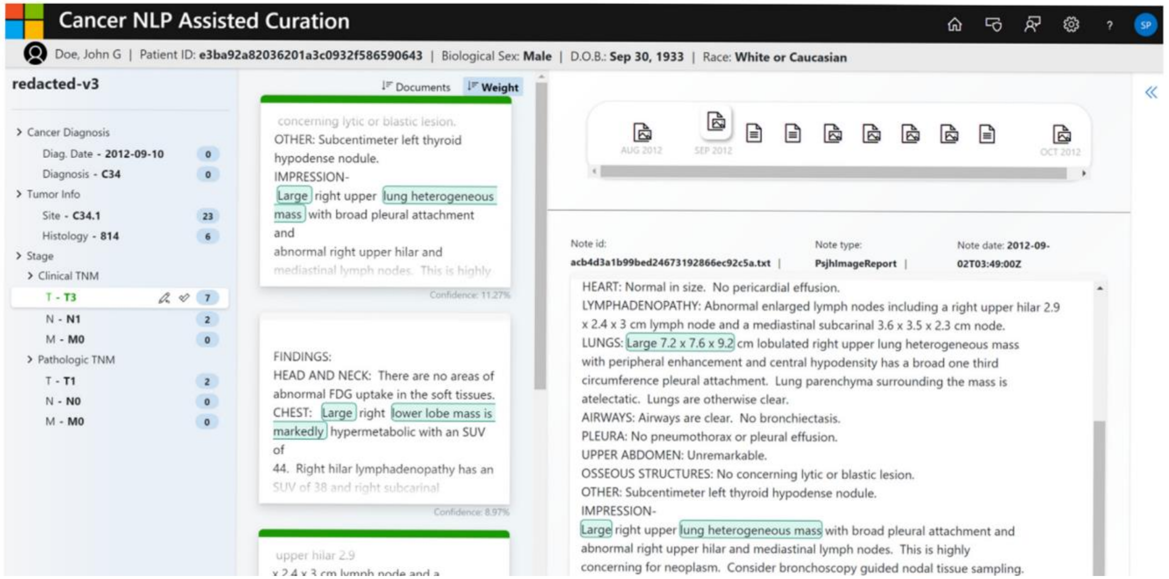
1. Provide context to the problem to solve (LLMs as subject matter expert).

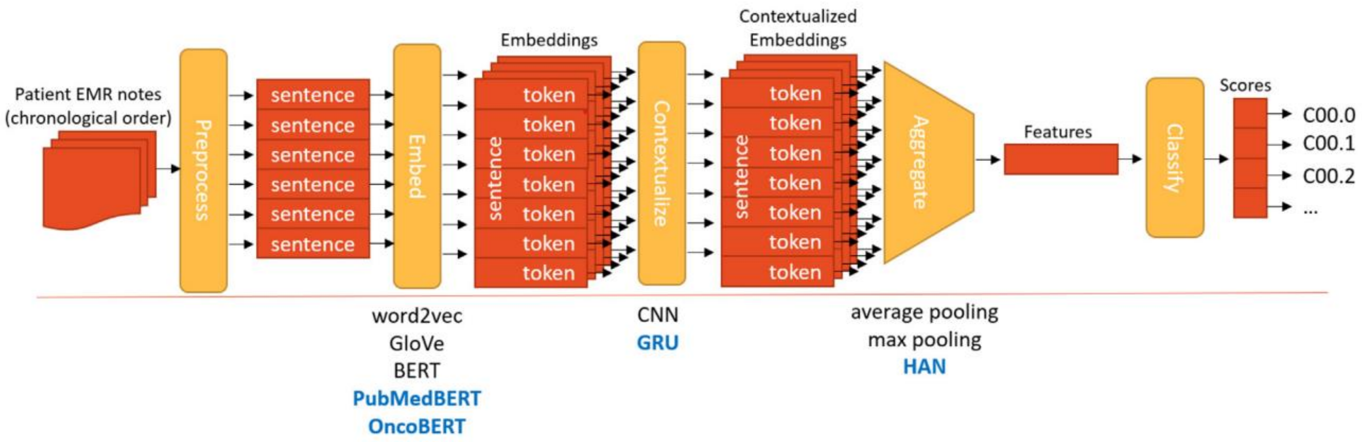
2. Structure data required to solve the problem (LLM as data curator).

3. Process and create new features (LLM as data science assistant, who write code, interprets results, etc.).



LLMs for assisted data curation





Electronic health records

LLM

Structured database of patient characteristics

Tinn, R., et al (2023). Toward structuring real-world data: Deep learning for extracting oncology information from clinical text with patient-level supervision. *Patterns (New York, N.Y.)*, *4*(4), 100726.

LLMs for data science coding companion

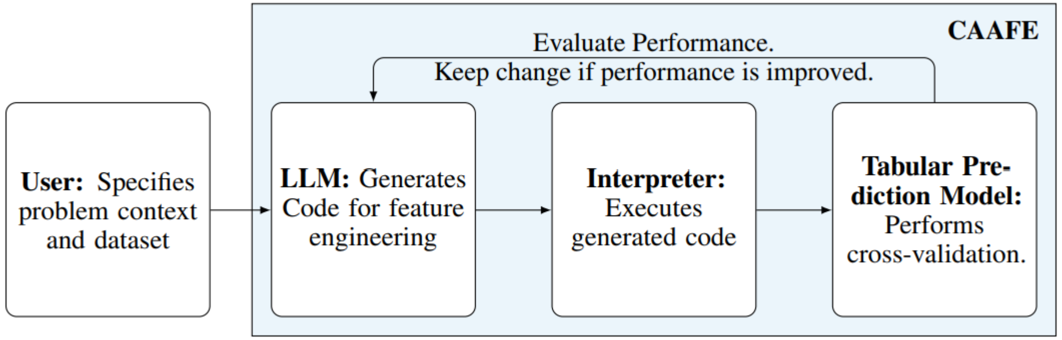
```
Dataset description: Tic-Tac-Toe Endgame database
This database encodes the complete set of possible board
    configurations at the end of tic-tac-toe games, where "x" is
    assumed to have played first. The target concept is "win for x" (i
    .e., true when "x" has one of 8 possible ways to create a "three-
    in-a-row").
```

```
# ('number-of-x-wins', 'Number of ways x can win on the board')
# Usefulness: Knowing the number of ways x can win on the board can be useful in
    predicting whether x has won the game or not.
# Input samples: 'top-left-square': [2, 2, 1], 'top-middle-square': [1, 2, 0], ...
df['number-of-x-wins'] = ((df['top-left-square']==1) & (df['top-middle-square']==1) & (df
    ['top-right-square']==1)).astype(int) + ((df['middle-left-square']==1) & (df['middle
    -middle-square']==1) & (df['middle-right-square']==1)).astype(int) [...]
```

```
Iteration 1
Performance before adding features ROC 0.888, ACC 0.700.
Performance after adding features ROC 0.987, ACC 0.980.
Improvement ROC 0.099, ACC 0.280. Code was executed and changes to df
    retained.
```

```
# ('number-of-o-wins', 'Number of ways o can win on the board')
# Usefulness: Knowing the number of ways o can win on the board can be useful in
    predicting whether o has won the game or not.
# Input samples: 'top-left-square': [2, 2, 1], 'top-middle-square': [1, 2, 0], ...
df['number-of-o-wins'] = ((df['top-left-square']==2) & (df['top-middle-square']==2) & (df
    ['top-right-square']==2)).astype(int) + ((df['middle-left-square']==2) & (df['middle
    -middle-square']==2) & (df['middle-right-square']==2)).astype(int) [...]
```

```
Iteration 2
Performance before adding features ROC 0.987, ACC 0.980.
Performance after adding features ROC 1.000, ACC 1.000.
Improvement ROC 0.013, ACC 0.020. Code was executed and changes to df
    retained.
```

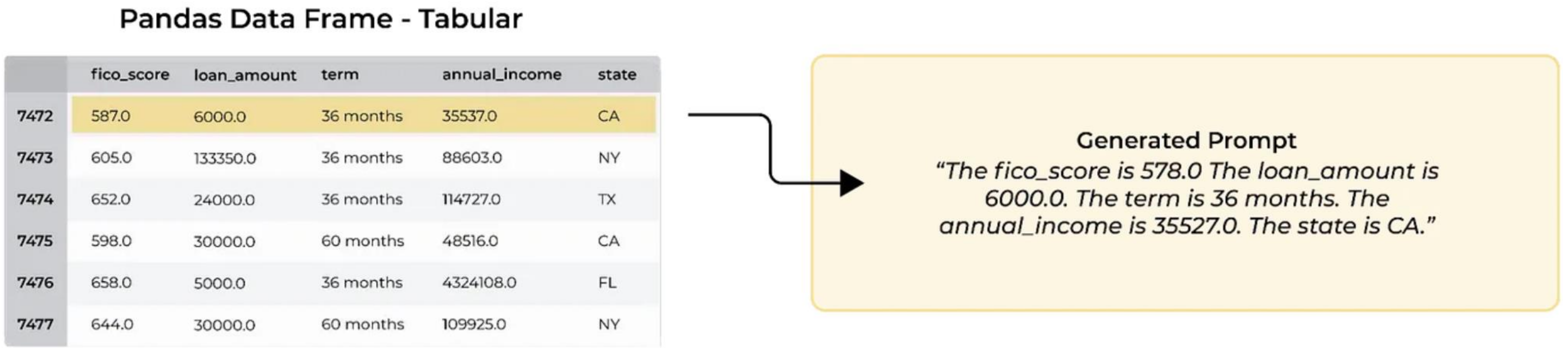


Features for a given problem

Generate code to generate new features

LLMs for Semi-Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering Noah Hollmann, Samuel Müller, Frank Hutter. 2023 arXiv:2305.03403.

LLMs to facilitate the detection of data anomalies in databases



Pandas Data Frame - Tabular

	fico_score	loan_amount	term	annual_income	state
7472	587.0	6000.0	36 months	35537.0	CA
7473	605.0	133350.0	36 months	88603.0	NY
7474	652.0	24000.0	36 months	114727.0	TX
7475	598.0	30000.0	60 months	48516.0	CA
7476	658.0	5000.0	36 months	4324108.0	FL
7477	644.0	30000.0	60 months	109925.0	NY

Example of a prompt created from tabular data. Each row of data is encoded as a separate prompt and made by concatenating a simple statement from each cell of the row. (Image by author)

Existing data set

Error flags about without missing values, outliers, etc

Source: Applying Large Language Models to Tabular Data to Identify Drift | by Aparna Dhinakaran | Towards Data Science

LLMs as subject matter expert

SYSTEM: You are a helpful assistant for causal reasoning.

USER: Which cause-and-effect relationship is more likely?

A. changing the age of an abalone causes a change in its length.

B. changing the length of an abalone causes a change in its age.

Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.

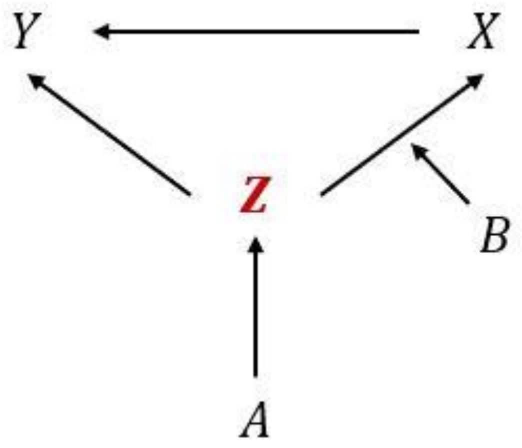
GPT-3.5-TURBO: 1. Cause: Changing the age of an abalone

Effect: a change in its length.

2. Cause: Changing the length of an abalone

Effect: a change in its age.

Based on the biological knowledge about abalones, it is more likely that changing the age of an abalone causes a change in its length. This is because the length of an abalone is known to increase as it ages, and not the other way around. Therefore, the correct answer is <Answer>A</Answer>.



Description of the modelling question

Suggestion of variables, their relationships, etc.

Example source: Causal Reasoning and Large Language Models: Opening a New Frontier for Causality.

Overview

Precision health

Intelligence revolution

LLMs for precision health

Application challenges

Research frontiers

Does it work?

Is it safe?

Real-world data can provide crucial evidence

Human-in-the-loop helps mitigate risk & continuously improve

Rethinking AI for Health

arXiv:2303.12712v5 [cs.CL] 13 Apr 2023

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

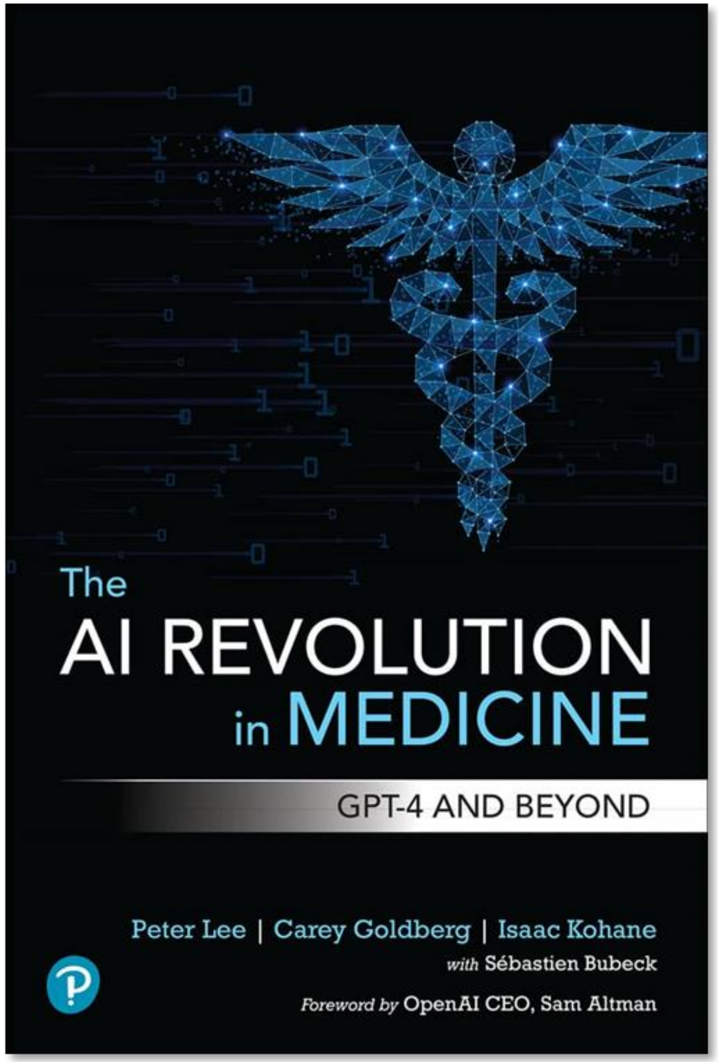
Microsoft Research

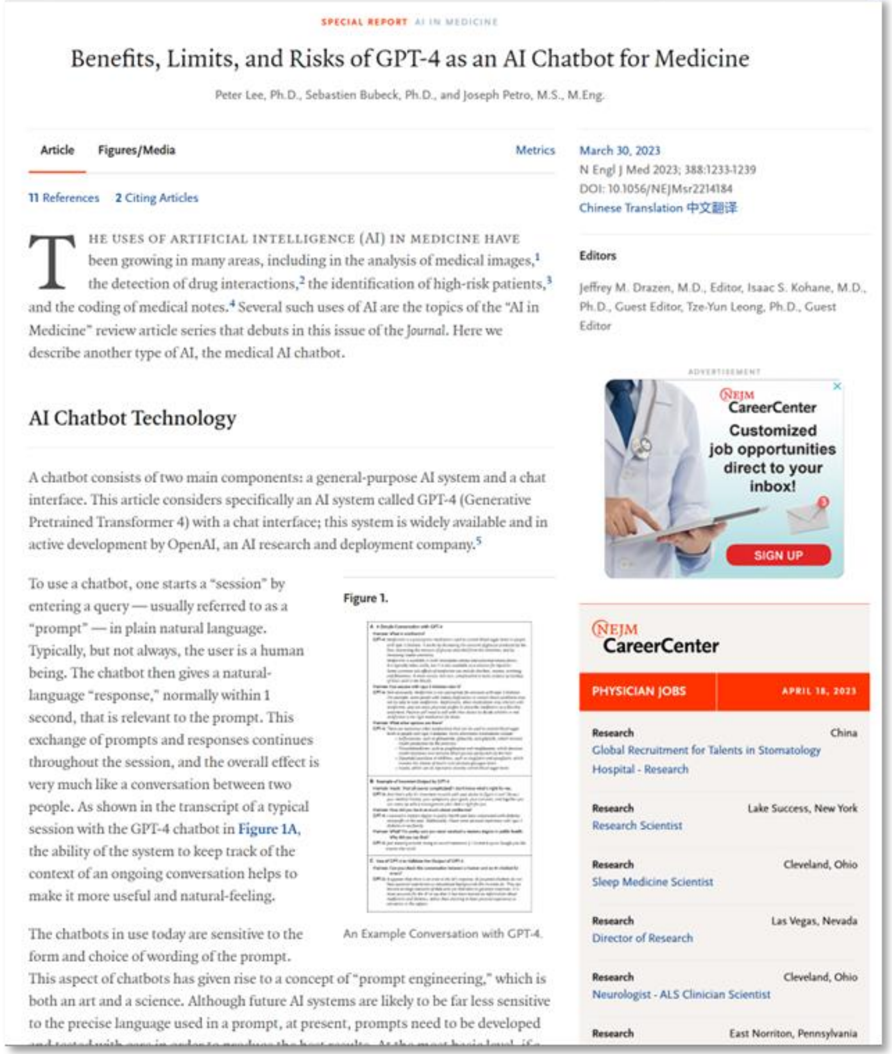
Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

Contents

1





SPECIAL REPORT AI IN MEDICINE

Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine

Peter Lee, Ph.D., Sebastien Bubeck, Ph.D., and Joseph Petro, M.S., M.Eng.

Article Figures/Media Metrics

11 References 2 Citing Articles

March 30, 2023
N Engl J Med 2023; 388:1233-1239
DOI: 10.1056/NEJMsr2214184
Chinese Translation 中文翻译

Editors
Jeffrey M. Drazen, M.D., Editor, Isaac S. Kohane, M.D., Ph.D., Guest Editor, Tze-Yun Leong, Ph.D., Guest Editor

THE USES OF ARTIFICIAL INTELLIGENCE (AI) IN MEDICINE HAVE been growing in many areas, including in the analysis of medical images,[1] the detection of drug interactions,[2] the identification of high-risk patients,[3] and the coding of medical notes.[4] Several such uses of AI are the topics of the "AI in Medicine" review article series that debuts in this issue of the *Journal*. Here we describe another type of AI, the medical AI chatbot.

AI Chatbot Technology

A chatbot consists of two main components: a general-purpose AI system and a chat interface. This article considers specifically an AI system called GPT-4 (Generative Pretrained Transformer 4) with a chat interface; this system is widely available and in active development by OpenAI, an AI research and deployment company.[5]

To use a chatbot, one starts a "session" by entering a query — usually referred to as a "prompt" — in plain natural language. Typically, but not always, the user is a human being. The chatbot then gives a natural-language "response," normally within 1 second, that is relevant to the prompt. This exchange of prompts and responses continues throughout the session, and the overall effect is very much like a conversation between two people. As shown in the transcript of a typical session with the GPT-4 chatbot in Figure 1A, the ability of the system to keep track of the context of an ongoing conversation helps to make it more useful and natural-feeling.

Figure 1.

An Example Conversation with GPT-4.

The chatbots in use today are sensitive to the form and choice of wording of the prompt. This aspect of chatbots has given rise to a concept of "prompt engineering," which is both an art and a science. Although future AI systems are likely to be far less sensitive to the precise language used in a prompt, at present, prompts need to be developed

Microsoft and Epic expand strategic collaboration with integration of Azure OpenAI Service

April 17, 2023 | Microsoft News Center

GPT-4 can help draft in-basket response

REDMOND, Wash., and VERONA, Wis. — April 17, 2023 — Microsoft Corp. and Epic on Monday announced they are expanding their long-standing strategic collaboration to develop and integrate generative AI into healthcare by combining the scale and power of Azure OpenAI Service[1] with Epic's industry-leading electronic health record (EHR) software. The collaboration expands the long-standing partnership, which includes enabling organizations to run Epic environments on the Microsoft Azure cloud platform.

This co-innovation is focused on delivering a comprehensive array of generative AI-powered solutions integrated with Epic's EHR to increase productivity, enhance patient care and improve financial integrity of health systems globally. One of the initial solutions is already underway, with UC San Diego Health, UW Health in Madison, Wisconsin, and Stanford Health Care among the first organizations starting to deploy enhancements to automatically draft message responses.

"A good use of technology simplifies things related to workforce and workflow," said Chero Goswami, chief information officer at UW Health. "Integrating generative AI into some of our daily workflows will increase productivity for many of our providers, allowing them to focus on the clinical duties that truly require their attention."

Original Investigation

April 28, 2023

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD, MA[1,2]; Adam Poliak, PhD[3]; Mark Dredze, PhD[4]; et al

» Author Affiliations

JAMA Intern Med. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

The proportion of responses rated as *good* or *very good* quality (≥ 4), for instance, was higher for chatbot than physicians … This amounted to 3.6 times higher prevalence of *good* or *very good* quality responses for the chatbot. Chatbot responses were also rated significantly more empathetic than physician responses … This amounted to 9.8 times higher prevalence of *empathetic* or *very empathetic* responses for the chatbot.

Fully AI-automated notes—available in seconds.

Announcing DAX Express

Experience a fully AI-automated note creation solution that uses conversational, ambient, and generative AI to create draft clinical notes from patient conversations and make them available immediately after concluding a patient visit.

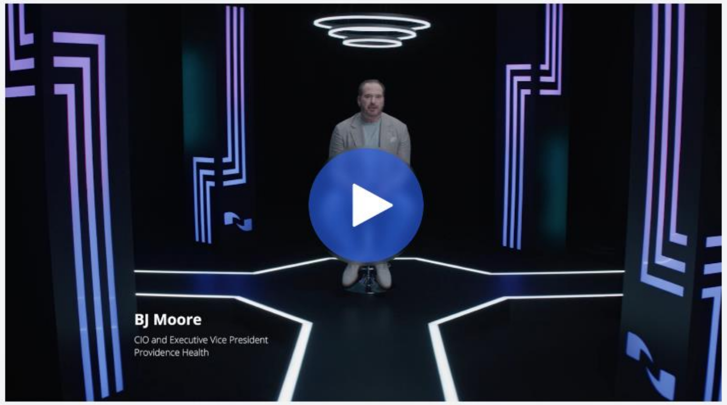
Learn more about the upcoming ways to experience DAX Express

Explore more

First healthcare solution powered by OpenAI's GPT-4

Groundbreaking Nuance DAX Express is the next milestone in an expanding portfolio of solutions combining OpenAI's GPT-4 with proven workflow-integrated technology to define the future of intelligence-infused healthcare experiences.

Read the press release to learn more.



Real-World Evidence (RWE)

ars TECHNICA

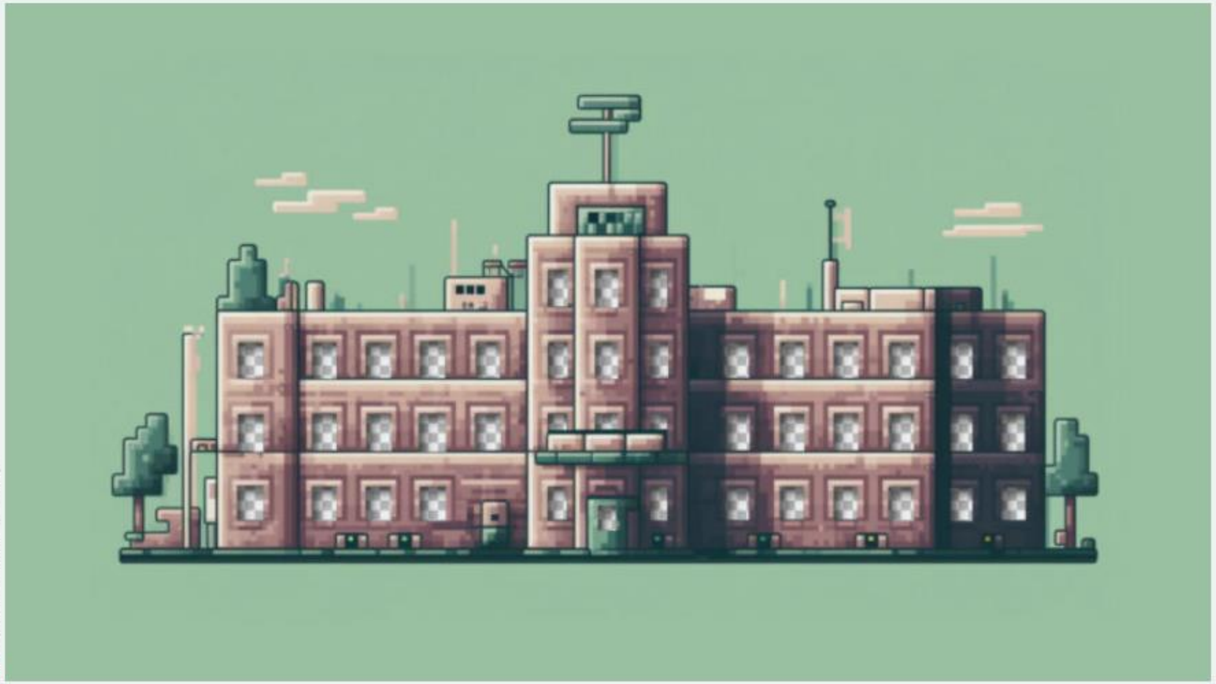
BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE STORE

DR. GPT WILL SEE YOU NOW —

GPT-4 will hunt for trends in medical records thanks to Microsoft and Epic

Generative AI promises to streamline health care, but critics say not so fast.

BENJ EDWARDS - 4/18/2023, 1:14 PM



Benj Edwards / Midjourney

Enlarge / An AI-generated image of a pixel art hospital with empty windows.

Epic SlicerDicer Cogito

Epic Cosmos

A Universe of Data that Drives Evidence-Based Research and Individualized Patient Care

Information Access Can Be Life or Death

Marty Tenenbaum

Late-stage melanoma (late 1990s)

Initial prognosis: 6 months

Saved by Phase III trial of Canvaxin

We're building a better ClinicalTrials.gov. Check it out and tell us what you think!

NIH U.S. National Library of Medicine

ClinicalTrials.gov

Find Studies ▾ About Studies ▾ Submit Studies ▾ Resources ▾ About Site ▾ PRS Login

Home > Search Results > Study Record Detail

Save this study

Cabozantinib in High Grade Neuroendocrine Neoplasms

The safety and scientific validity of this study is the responsibility of the study sponsor and investigators. Listing a study does not mean it has been evaluated by the U.S. Federal Government. Know the risks and potential benefits of clinical studies and talk to your health care provider before participating. Read our disclaimer for details.

ClinicalTrials.gov Identifier: NCT04412629

Recruitment Status : Recruiting
First Posted : June 2, 2020
Last Update Posted : December 20, 2022

See Contacts and Locations

View this study on Beta.ClinicalTrials.gov

Sponsor:
Washington University School of Medicine

Collaborator:
Exelixis

Information provided by (Responsible Party):
Washington University School of Medicine

Study Details | Tabular View | No Results Posted | Disclaimer | How to Read a Study Record

Study Description

Go to

Brief Summary:

High grade neuroendocrine neoplasm patients are treated with platinum doublets such as carboplatin and etoposide mimicking the current guidelines for small cell lung cancer (SCLC). Unfortunately, recurrences are common and most patients with metastatic disease succumb to it within a year. There is no extensive literature or consensus on second- or third-line options (which include FOLFOX, FOLFIRI, capecitabine and temozolomide, taxanes or immunotherapy) and there is urgent need for better regimens.

LLM: Universal Structuring

Instruction

Your task is to convert the given clinical trial criteria input into a valid logic formula. Your response should be written in the language of propositional logic and should accurately capture the meaning of the input. Please note that your response should also include any necessary logical connectives, such as "and", "or", or "not". Please keep in mind that your response should be flexible enough to allow for various relevant and creative solutions. You should also focus on providing an accurate and well-structured solution that can be easily understood by others.

Example

Input:
"
Histologically confirmed metastatic colorectal adenocarcinoma with mutant APC, TP53 and KRAS genes as determined by the local CLIA-certified laboratory are eligible. All RAS mutations are allowed (KRAS, NRAS, HRAS). Patients with wild type KRAS, APC or TP53 are ineligible.
"

Output:
"
metastatic AND colorectal adenocarcinoma AND (APC mutation AND TP53 mutation AND KRAS mutation) AND NOT (NOT KRAS mutation OR NOT APC mutation OR NOT TP53 mutation)
"

Input

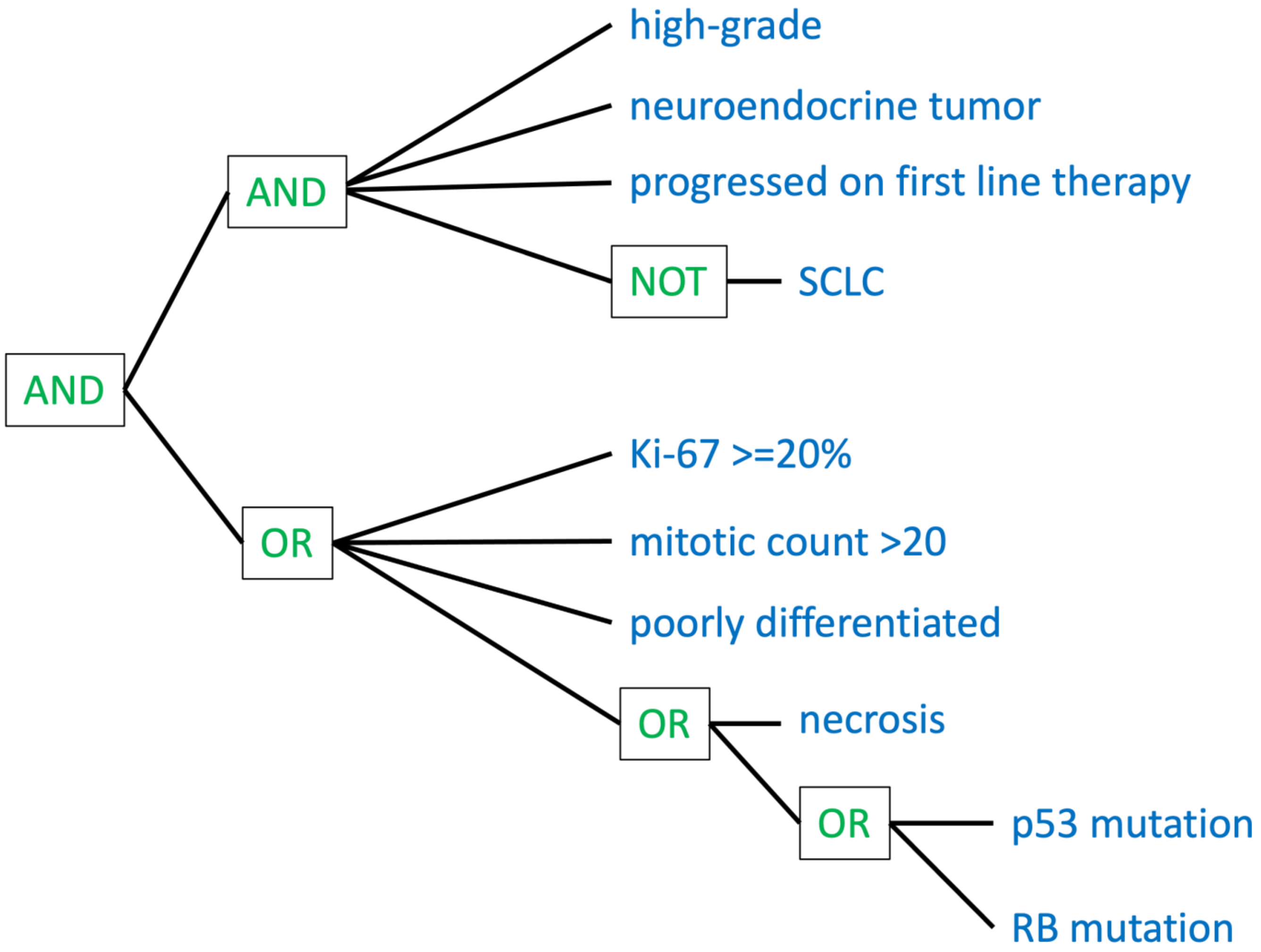
Input:
"
-Histologically or cytologically confirmed high-grade neuroendocrine tumor that has progressed on first line therapy, excluding small cell lung cancer (SCLC). High grade includes any neuroendocrine neoplasm with a Ki-67 of >=20% or with mitotic count of more than 20 mitoses per high power field or any poorly differentiated neoplasm or any neoplasm lacking these that is deemed high grade by pathology consensus, based on other markers (necrosis or IHC demonstrating p53 or RB mutation).
”

Output:

LLM: Universal Structuring

Output

"
(high-grade AND neuroendocrine tumor AND progressed on first line therapy AND NOT SCLC) AND (Ki-67 >=20% OR mitotic count >20 OR poorly differentiated OR (necrosis OR (p53 mutation OR RB mutation)))
"

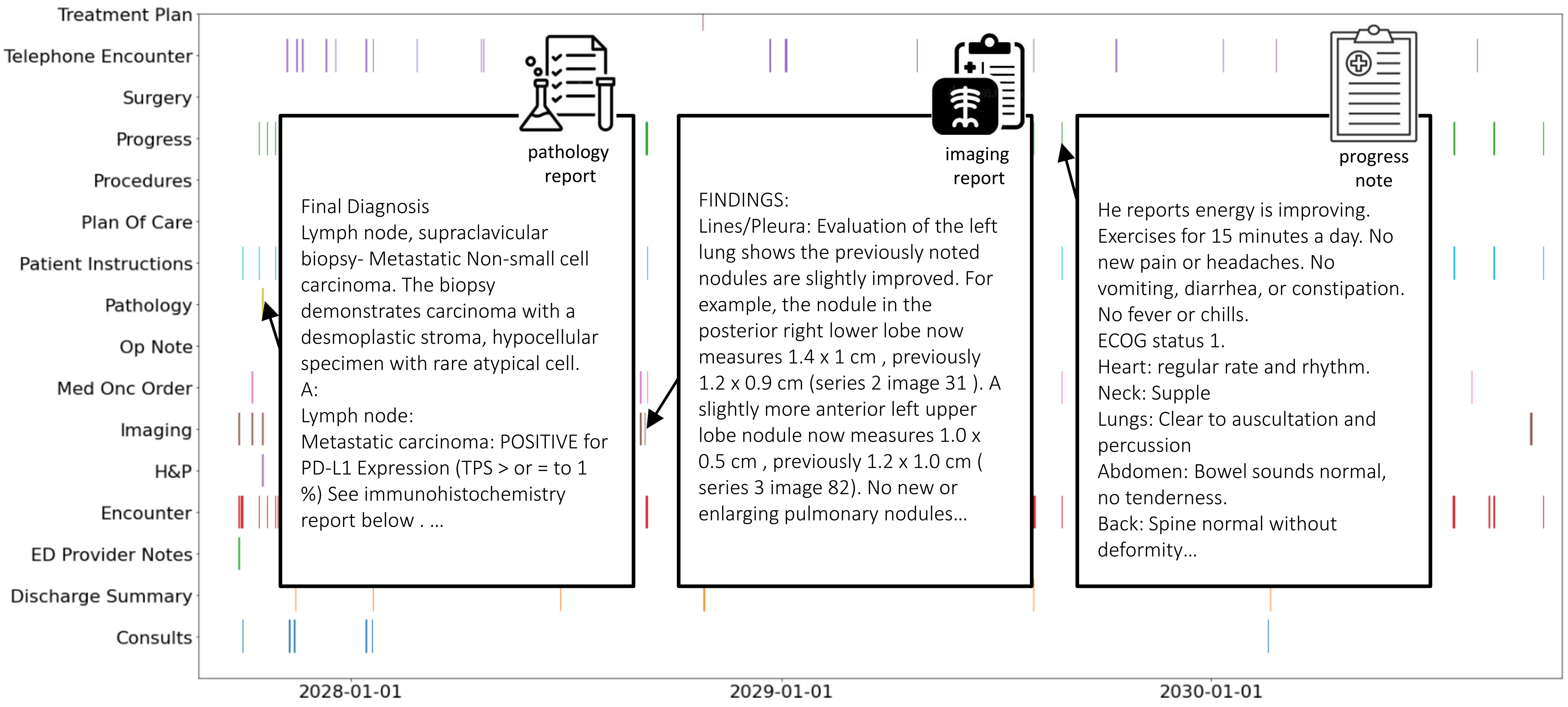


LLM: Universal Structuring

	Histology			Biomarker		
	Precision	Recall	F1	Precision	Recall	F1
GNormPlus	-	-	-	6.8	19.6	10.2
SciSpaCy	34.2	70.2	46.0	58.3	6.9	12.3
Criteria2Query	29.6	40.2	32.8	68.3	27.5	39.2
GPT-3.5 (zero-shot)	35.1	31.6	34.2	61.2	29.4	39.7
GPT-4 (zero-shot)	62.1	69.0	**65.4**	75.3	59.8	66.7
GPT-4 (3-shot)	57.8	73.7	64.8	72.5	72.5	**72.5**

Wong et al. "Scaling Clinical Trial Matching Using Large Language Model: A Case Study in Oncology", *MLHC 2023*.

EMR: Cancer Patient Journey



OncoBERT: Oncology RWE

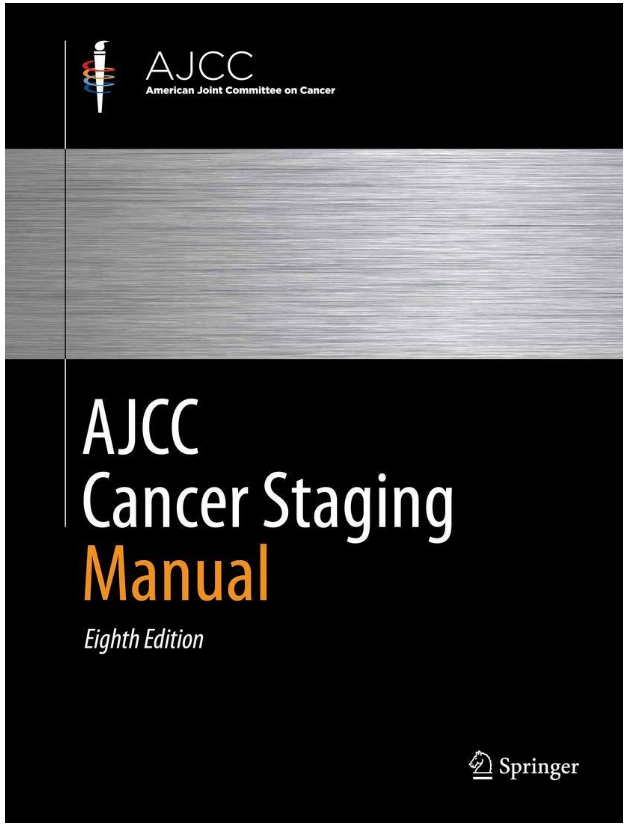
	Tumor Site	Histology	Clinical T	N	M	Pathological T	N	M
Ontology	19.4	19.2	-	-	-	-	-	-
BOW	62.8	76.6	70.4	96.6	98.4	72.1	90.7	**98.9**
OncoGloVe + CNN	72.0	84.4	74.2	96.5	98.6	83.9	93.1	98.5
OncoGloVe + HAN/GRU	74.0	85.9	76.2	97.1	98.7	86.4	94.2	98.5
BERT + HAN/GRU	75.1	86.2	77.0	96.6	98.4	86.4	94.4	98.2
PubMedBERT + HAN/GRU (ours)	76.7	87.2	79.3	97.2	98.7	87.2	95.2	98.6
OncoBERT + HAN/GRU (ours)	**77.1**	**87.6**	**81.4**	**97.5**	**99.0**	**87.6**	**95.5**	**98.9**

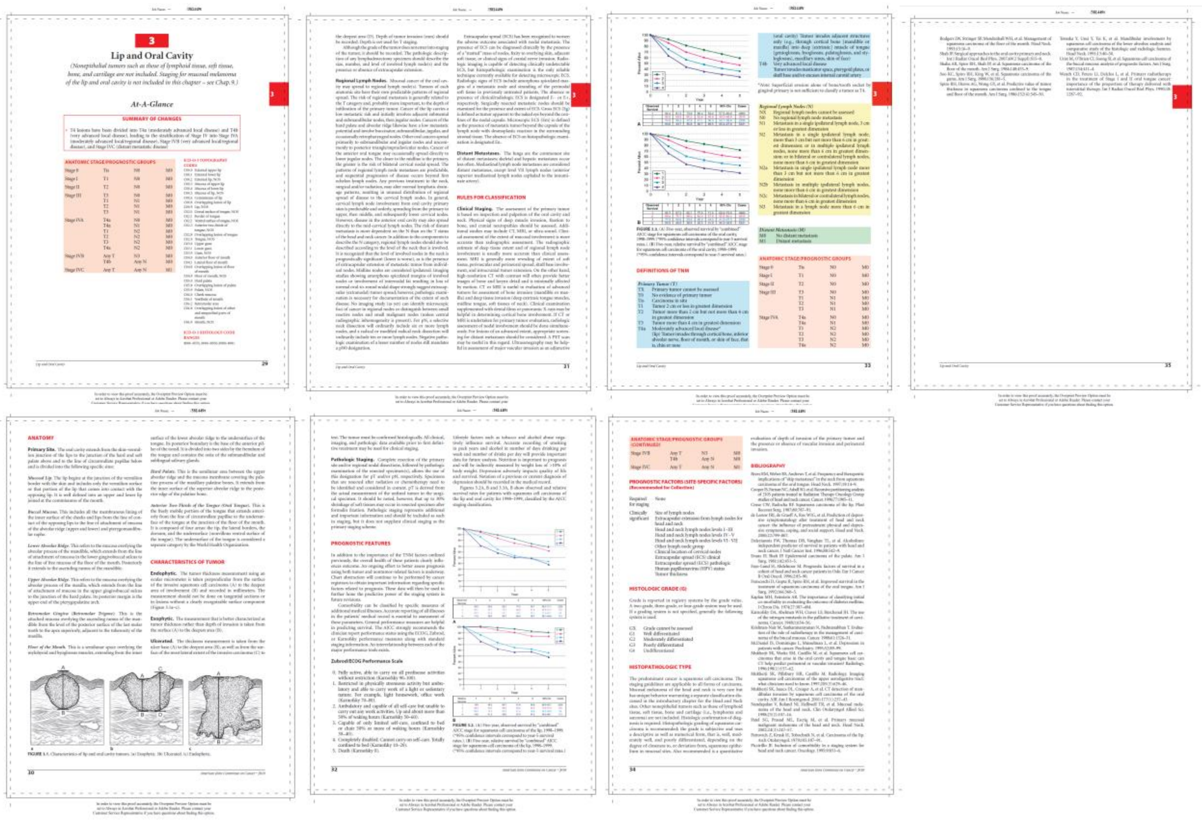
Preston, Wei, et al. "Towards Structuring Real-World Data at Scale: Deep Learning for Extracting Key Oncology Information from Clinical Text with Patient-Level Supervision", *Patterns 2023*.

GPT-4: Structure Real-World Data

Preliminary results promising

“Read” annotation guideline → zero-shot structuring





Project Hanover | Patient-Centric | Trial-Centric

Clinical Trial Triaging

Cliff Wong

Name: HANKS, TOM JEFFREY
Accession No.: 34-234-58823
D.O.B.: Feb. 18, 1950
Age: 73.0
Gender: M
Histology:
LUAD (Lung Adenocarcinoma)
Path Staging: None None None
Stage Group: Stage IV
HLA type:

- HLA-A*02:01 HLA-A*02:01
- HLA-B*07:02 HLA-B*39:06
- HLA-C*03:04 HLA-C*08:02

Patient EHR Assisted Curation N/A

Search | Report

Trial Filters

- [x] Age Match Only
- [ ] Stage Match Only
- [ ] Updated in Last 2 Years

Locations

- [ ] North America
- [ ] United States
- [ ] Providence States

Biomarkers

clinical signif.	gene	protein change	variant
YES	KRAS	p.Gly12Val	G12V
YES	TP53	p.Arg306Ter	R306*
YES	APC	p.Glu1353Ter	E1353*
YES	ATM	p.Glu2139IlefsTer6	E2139Ifs*6
YES	ERBB2	3.4(fold-change)	ERBB2-High

Search Builder Show 10 entries

Search:

	NCT No.	Title	Phase	Matching Trial Diseases	Matching Trial Stage	Matching Trial Biomarkers	Notes	Providence States
☑	NCT03953235	A Study of a Personalized Cancer Vaccine Targeting Shared Neoantigens	Phase 1/Phase 2	- Non-Small Cell Lung Carcinoma - Malignant Solid Neoplasm	- Metastatic - Advanced	- KRAS G12V	test3	CA, TX
☐	NCT04620330	A Study of Avutometinib (VS-6766) + Defactinib in Recurrent KRAS G12V, Other KRAS and BRAF Non-Small Cell Lung Cancer	Phase 2	- Non-Small Cell Lung Carcinoma		- KRAS G12V - KRAS Mutation	test6	CA, OR, TX
☐	NCT03454035	Ulixertinib/Palbociclib in Patients With Advanced Pancreatic and Other Solid Tumors	Phase 1	- Malignant Solid Neoplasm	- Stage IV - Metastatic - Advanced	- KRAS G12X - KRAS Mutation		
☐	NCT05631899	Combination of CAR-DC Vaccine and Anti-PD-1 Antibody in Local Advanced/Metastatic Solid Tumors	Phase 1	- Malignant Solid Neoplasm	- Metastatic - Advanced	- KRAS G12V - KRAS Mutation		
☐	NCT05438667	TCR-T Cell Therapy on Advanced Pancreatic Cancer and Other Solid Tumors	Early Phase 1	- Malignant Solid Neoplasm	- Metastatic - Advanced	- KRAS G12V - KRAS Mutation		
☐	NCT04625647	Testing the Use of Targeted Treatment (AMG 510) for KRAS G12C Mutated Advanced Non-squamous Non-small Cell Lung Cancer (A Lung-MAP Treatment Trial)	Phase 2	- Non-Squamous Non-Small Cell Lung Carcinoma - Lung Adenocarcinoma - Non-Small Cell Lung Carcinoma - Lung Carcinoma	- Stage IVA - Stage IVB - Stage IV - Advanced	- KRAS Mutation		AK, CA, MT, NM, OR, TX, WA
☑	NCT04999761	AB122 Platform Study	Phase 1	- Non-Squamous Non-Small Cell Lung Carcinoma - Non-Small Cell Lung Carcinoma - Malignant Solid Neoplasm	- Metastatic - Advanced	- KRAS Mutation		
☐	NCT03667716	COM701 (an Inhibitor of PVRIG) in Subjects With Advanced Solid Tumors.	Phase 1	- Non-Small Cell Lung Carcinoma - Lung Carcinoma - Malignant Solid Neoplasm	- Stage IV - Metastatic - Advanced	- KRAS Mutation		CA, TX
☑	NCT04511845	A Dose-Escalation Study of SPYK04 in Patients With Locally Advanced or Metastatic Solid Tumors (With Expansion).	Phase 1	- Non-Small Cell Lung Carcinoma - Malignant Solid Neoplasm	- Metastatic	- KRAS Mutation - MAPK/ERK pathway		TX

ClinicalTrials.gov BETA

Resources ⌄ About ⌄

Home › Search Results › Study Record

ClinicalTrials.gov Identifier: NCT04520711

RECRUITING

Hotspot TCR-T: A Phase I/Ib Study of Adoptively Transferred T-cell Receptor Gene-engineered T Cells (TCR-T)

Information provided by Providence Health & Services (Responsible Party)

Last Updated: May 6, 2022

Dr. Rom Leidner

Providence

Target: 24 / Recruited: 2

Manual matching takes many hours

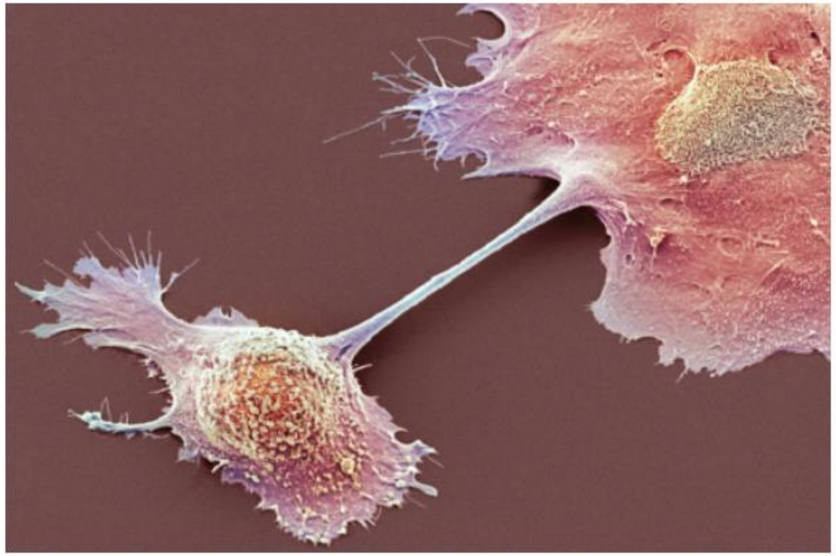
NLP: 100+ candidates in initial scan

The New York Times

Reprogrammed Cells Attack and Tame Deadly Cancer in One Woman

Another patient who had the same treatment did not survive. But the demonstration of the technique could help with other cancers.

Give this article



A colored scanning electron micrograph of pancreatic cancer cells. Steve Gschmeissner/Science Source

By Gina Kolata

Drug Discovery

Drug Discov Today. 2021 Nov; 26(11): 2593–2607. PMCID: PMC8604259
Published online 2021 Jun 30. doi: 10.1016/j.drudis.2021.06.009 PMID: 34216835

AI-based language models powering drug discovery and development

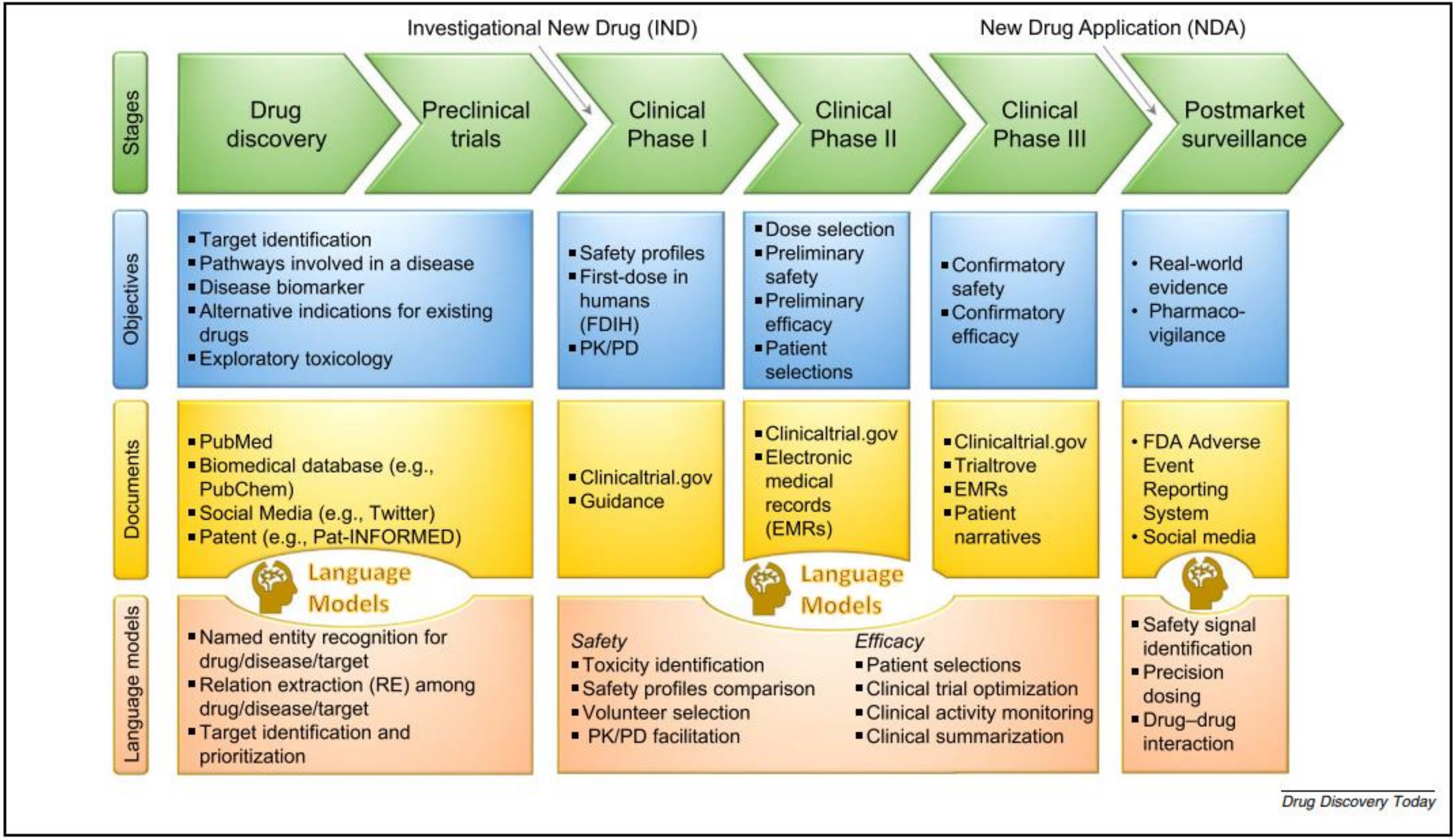
Zhichao Liu,[a,*] Ruth A. Roberts,[a,b,c] Madhu Lal-Nag,[d] Xi Chen,[a] Ruili Huang,[e] and Weida Tong[a,*]

News | Published: 24 April 2023

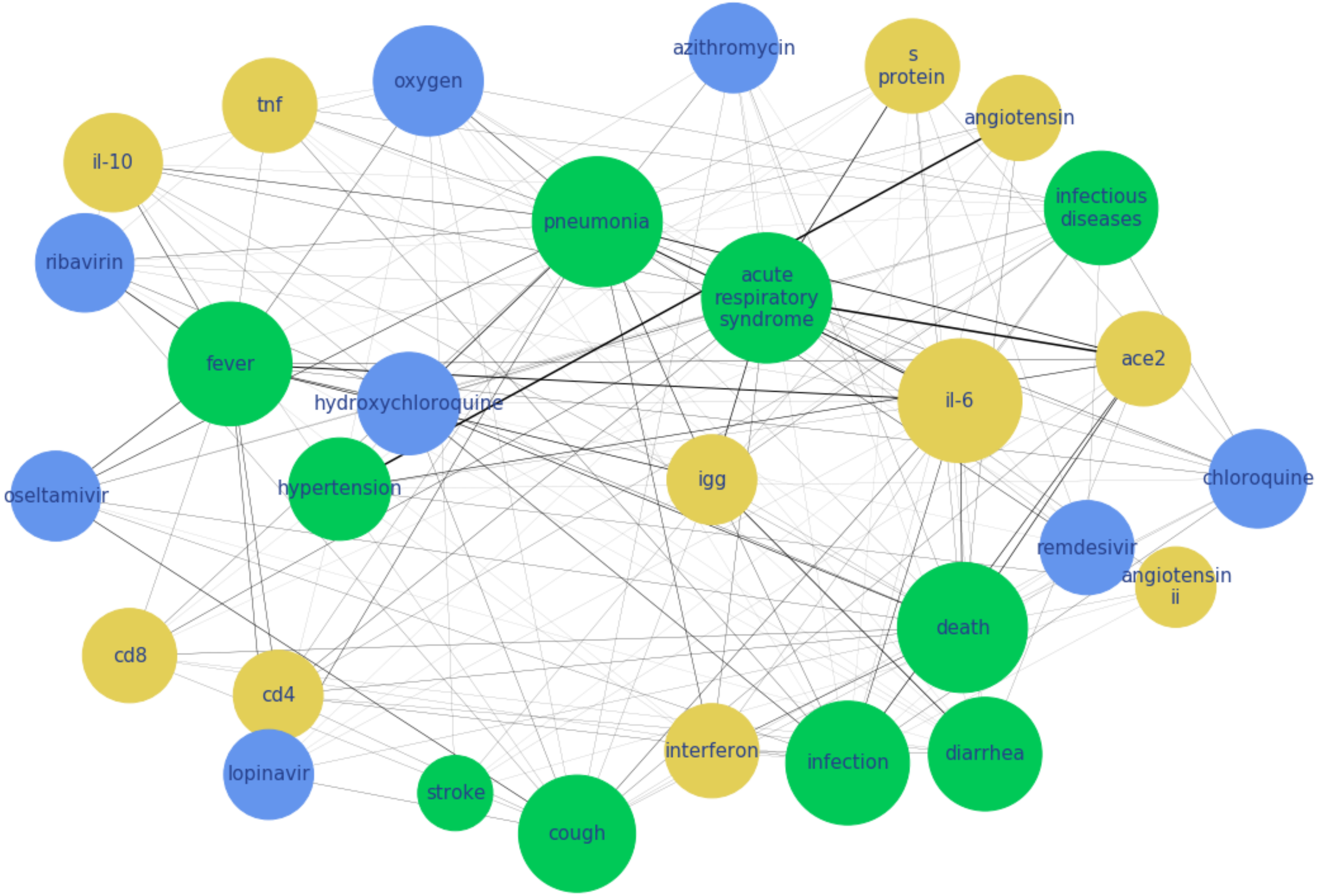
Drug discovery companies are customizing ChatGPT: here's how

Neil Savage

Nature Biotechnology **41**, 585–586 (2023) | Cite this article


Investigational New Drug (IND)
New Drug Application (NDA)
Stages
Drug discovery
Preclinical trials
Clinical Phase I
Clinical Phase II
Clinical Phase III
Postmarket surveillance
Objectives
▪ Target identification
▪ Pathways involved in a disease
▪ Disease biomarker
▪ Alternative indications for existing drugs
▪ Exploratory toxicology
▪ Safety profiles
▪ First-dose in humans (FDIH)
▪ PK/PD
▪ Dose selection
▪ Preliminary safety
▪ Preliminary efficacy
▪ Patient selections
▪ Confirmatory safety
▪ Confirmatory efficacy
• Real-world evidence
• Pharmaco-vigilance
Documents
▪ PubMed
▪ Biomedical database (e.g., PubChem)
▪ Social Media (e.g., Twitter)
▪ Patent (e.g., Pat-INFORMED)
▪ Clinicaltrial.gov
▪ Guidance
▪ Clinicaltrial.gov
▪ Electronic medical records (EMRs)
▪ Clinicaltrial.gov
▪ Trialtrove
▪ EMRs
▪ Patient narratives
• FDA Adverse Event Reporting System
• Social media
Language Models
Language Models
Language models
▪ Named entity recognition for drug/disease/target
▪ Relation extraction (RE) among drug/disease/target
▪ Target identification and prioritization
Safety
▪ Toxicity identification
▪ Safety profiles comparison
▪ Volunteer selection
▪ PK/PD facilitation
Efficacy
▪ Patient selections
▪ Clinical trial optimization
▪ Clinical activity monitoring
▪ Clinical summarization
▪ Safety signal identification
▪ Precision dosing
▪ Drug–drug interaction
Drug Discovery Today

Literature → Knowledge Graph



Research Frontiers

Self verification

Knowledge distillation

Causal discovery

Multi-modal learning

Prompt Programming

Engineering		Programming
Black art, lack guarantee, superseded by more supervision	→	Composition & Control: self fact-check, tool use, structured resources

ChatGPT Code Interpreter Plugin

LangChain

Semantic Kernel

Retrieval-Augmented Generation (RAG)

Verification Much Easier Than Generation

P vs NP

Self Fact-Check: a prompt program

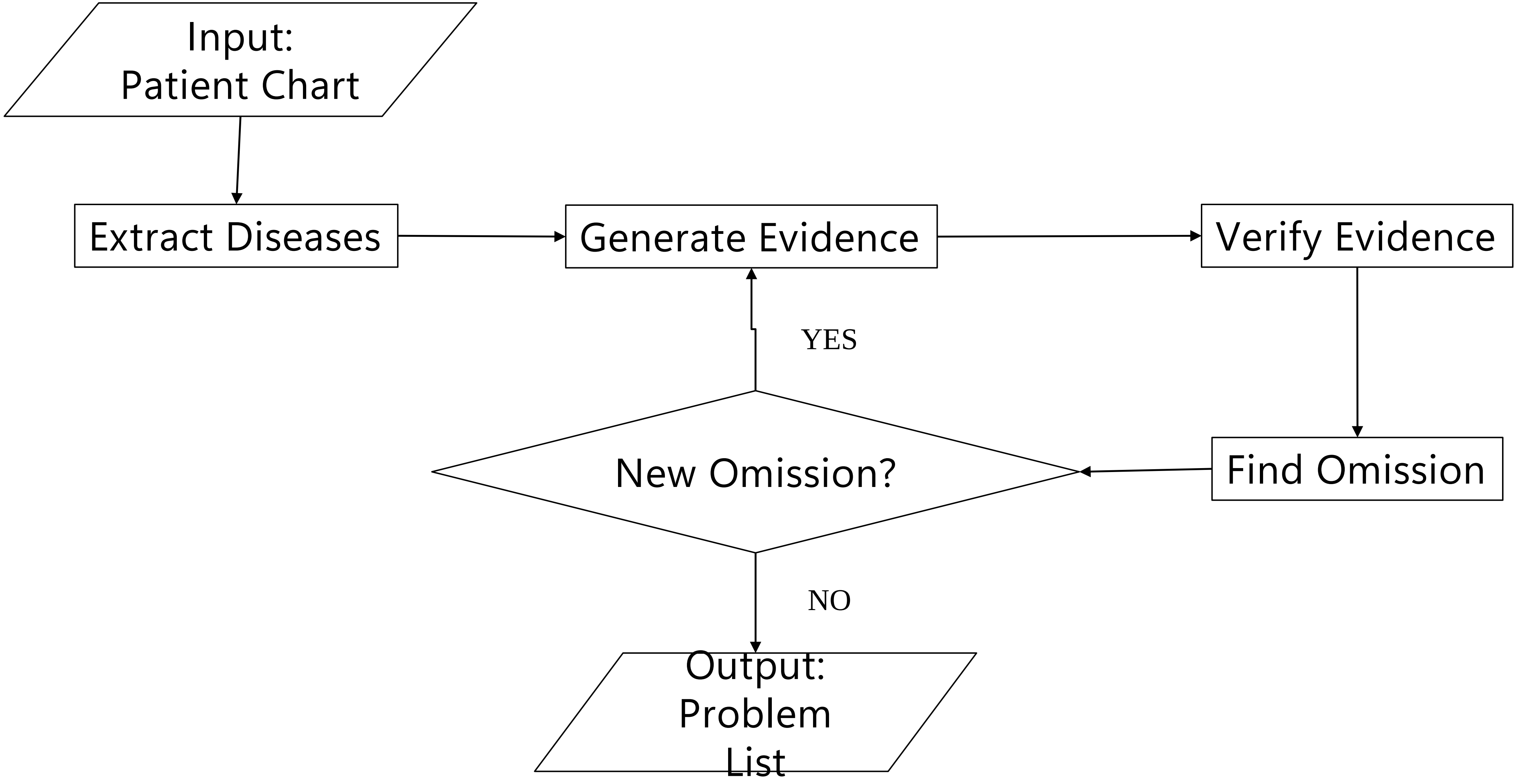
```
1,23224,174680,2147-12-05,,,"Discharge summary","Report",,"","Admissi
on Date:  [**2823-9-29**]              Discharge Date:   [**2823-10-1
7**]

Date of Birth:  [**2768-10-11**]             Sex:   F

Service: SURGERY

Allergies:
Patient recorded as having No Known Allergies to Drugs

Attending:[**First Name3 (LF) 1**]
Chief Complaint:
headache and neck stiffness

Major Surgical or Invasive Procedure:
central line placed, arterial line placed

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis
on 6-mercaptopurine, prednisone 40-60 mg daily, who presents
with a new onset of headache and neck stiffness. The patient is
in distress, rigoring and has aphasia and only limited history
is obtained. She reports that she was awaken 1AM the morning of
[**2823-9-28**] with a headache which she describes as bandlike. She
states that headaches are unusual for her. She denies photo- or
phonophobia. She did have neck stiffness. On arrival to the ED
at 5:33PM, she was afebrile with a temp of 96.5, however she
later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR
24, O2 sat 100 %. Head CT was done and relealved attenuation
within the subcortical white matter of the right medial frontal
lobe. LP was performed showing opening pressure 24 cm H2O WBC of
316, Protein 152, glucose 16.  She was given Vancomycin 1 gm IV,
Ceftriaxone 2 gm IV, Acyclovir 800 mg IV, Ambesone 183 IV,
Ampicillin 2 gm IV q 4, Morphine 2-4 mg Q 4-6, Tylenol 1 gm ,
Decadron 10 mg IV.  The patient was evaluated by Neuro in the
ED.
```

Patient Chart

'peptic ulcer disease--533’,
'sleep apnea--780.57’,
"raynaud's phenomenon--443.0",
'memory problems--780.93',
'gastrointestinal bleeding--578.9',
'hemorrhage--431’,
'amyloid angiopathy--331.82',
'hypertension--401.9'

Problem List

Self Fact-Check: a prompt program



Self Fact-Check: a prompt program

Find Omission

Prompt template: You are an expert disease inspector. Your job is to find all possible diseases in the given {text_input} exhaustively and return in a python list of strings. Your response should be in the form of python list with all the diseases that you can verify do exist in the {text_input}. Make sure to return the disease list exhaustively. Don't include a disease if it is in the {diseases} list. Return only unique diseases. All diseases in the list must be in a string format. You must strictly follow the following formatting: response = [disease, disease, …]. Return only the list, don't include any other text.

Self Fact-Check: a prompt program

Extract Diseases
['acute renal failure', 'coronary artery disease', 'aortic stenosis', 'end-stage renal disease', 'hypertension', 'insulin dependent-diabetes mellitus', 'sleep apnea', 'vertigo', 'osteoarthritis', 'skin cancer', 'abdominal hernia', 'uterine cancer', 'obesity', 'wound infection', 'aspiration', 'Clostridium difficile colitis', 'respiratory failure', 'clinical depression']

Find Omission
['pleural effusion', 'pneumonia', 'congestive heart failure', 'tricuspid regurgitation', 'right ventricular free wall hypokinesis', 'atheroma in aortic arch', 'atheroma in descending thoracic aorta', 'sepsis']

Self Fact-Check: a prompt program

Extract Diseases
['acute renal failure', 'coronary artery disease', 'aortic stenosis', 'end-stage renal disease', 'hypertension', 'insulin dependent-diabetes mellitus', 'sleep apnea', 'vertigo', 'osteoarthritis', 'skin cancer', 'abdominal hernia', 'uterine cancer', 'obesity', 'wound infection', 'aspiration', 'Clostridium difficile colitis', 'respiratory failure', 'clinical depression']

Find Omission
['pleural effusion', 'pneumonia', 'congestive heart failure', 'tricuspid regurgitation', 'right ventricular free wall hypokinesis', 'atheroma in aortic arch', 'atheroma in descending thoracic aorta', 'sepsis']

Generate Evidence / Verify Evidence
'pneumonia': 'left retrocardiac density concerning for pneumonia or atelectasis'

Self Fact-Check: a prompt program

Extract Diseases
['acute renal failure', 'coronary artery disease', 'aortic stenosis', 'end-stage renal disease', 'hypertension', 'insulin dependent-diabetes mellitus', 'sleep apnea', 'vertigo', 'osteoarthritis', 'skin cancer', 'abdominal hernia', 'uterine cancer', 'obesity', 'wound infection', 'aspiration', 'Clostridium difficile colitis', 'respiratory failure', 'clinical depression']

Find Omission
['pleural effusion', ~~'pneumonia'~~, 'congestive heart failure', 'tricuspid regurgitation', 'right ventricular free wall hypokinesis', 'atheroma in aortic arch', 'atheroma in descending thoracic aorta', 'sepsis']

Find Omission
['endocarditis', 'pneumonia or atelectasis', 'mild mitral annular calcification', 'mild thickening of mitral valve chordae', 'dilated left atrium', 'dilated right atrium', 'necrosis of abdominal wall', 'sternal wound infection']

Generate Evidence / Verify Evidence
'endocarditis': 'transesophageal echocardiogram the previous day <u>ruled out endocarditis</u>'

⋮

Self Fact-Check: a prompt program

Gero, Singh, et al. "Self-Verification Improves Few-Shot Clinical Information Extraction", *in submission*.

```
1,23224,174680,2147-12-05,,,"Discharge summary","Report",,"","Admissi
on Date:  [**2823-9-29**]              Discharge Date:   [**2823-10-1
7**]

Date of Birth:  [**2768-10-11**]             Sex:   F

Service: SURGERY

Allergies:
Patient recorded as having No Known Allergies to Drugs

Attending:[**First Name3 (LF) 1**]
Chief Complaint:
headache and neck stiffness

Major Surgical or Invasive Procedure:
central line placed, arterial line placed

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis
on 6-mercaptopurine, prednisone 40-60 mg daily, who presents
with a new onset of headache and neck stiffness. The patient is
in distress, rigoring and has aphasia and only limited history
is obtained. She reports that she was awaken 1AM the morning of
[**2823-9-28**] with a headache which she describes as bandlike. She
states that headaches are unusual for her. She denies photo- or
phonophobia. She did have neck stiffness. On arrival to the ED
at 5:33PM, she was afebrile with a temp of 96.5, however she
later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR
24, O2 sat 100 %. Head CT was done and relealved attenuation
within the subcortical white matter of the right medial frontal
lobe. LP was performed showing opening pressure 24 cm H2O WBC of
316, Protein 152, glucose 16.  She was given Vancomycin 1 gm IV,
Ceftriaxone 2 gm IV, Acyclovir 800 mg IV, Ambesone 183 IV,
Ampicillin 2 gm IV q 4, Morphine 2-4 mg Q 4-6, Tylenol 1 gm ,
Decadron 10 mg IV.  The patient was evaluated by Neuro in the
ED.
```

'peptic ulcer disease--533',
'sleep apnea--780.57',
"raynaud's phenomenon--443.0",
'memory problems--780.93',
'gastrointestinal bleeding--578.9',
'hemorrhage--431',
'amyloid angiopathy--331.82',
'hypertension--401.9'

GPT-4 w. self-verification → Comparable to supervised state of the art

Knowledge Distillation

LLM	Distillation	Test F1
GPT-3.5	-	78.2
GPT-4	-	85.0
Supervised State of the Art		93.4

Adverse Drug Event

Gu et al. "Distilling Large Language Models for Biomedical Knowledge Extraction", *in submission*.

Knowledge Distillation

LLM = Noisy Teacher

LLM	Distillation	Test F1
GPT-3.5	-	78.2
GPT-4	-	85.0
GPT-3.5	PubMedBERT	92.0
Supervised State of the Art		93.4

Adverse Drug Event

Gu et al. "Distilling Large Language Models for Biomedical Knowledge Extraction", *in submission*.

Over 1,000 times smaller than GPT-3.5

Knowledge Distillation

GPT-4

- Universal Structuring Model
- Universal Summarization Model
- ……..

Cost
Efficiency
White-Box
Customization

Universal NER

Zhou*, Zhang*, et al. "UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition", *in submission*.

LLM	Mean F1	Biomed F1
GPT-3.5 (175B)	34.9	38.1
Vicuna (13B)	13.9	9.8
UniNER (7B)	41.7	51.5
UniNER (13B)	43.4	51.2

Largest NER benchmark
43 datasets across 9 domains

Model size: 4-8%
Outperforms teach model by 7-9 points
Biomed: by 13 points

UniversalNER

User
Read this clinical note below:
Patient Name: John Doe
Age: 45
Gender: Male
Date of Examination: August 3, 2023

Chief Complaint:
The patient presents with a persistent cough and shortness of breath for the past two weeks.

History of Present Illness:
John Doe, a 45-year-old male, reports a progressive onset of cough, which started as a mild irritation but has worsened over the last two weeks. He also complains of occasional chest tightness and difficulty breathing, especially during physical exertion. There is no history of fever, chills, or recent upper respiratory tract infections.

Past Medical History:

Hypertension - Diagnosed 5 years ago, controlled with medication.
Allergic Rhinitis - Seasonal allergies, managed with antihistamines.
Obesity - BMI 32, struggling with weight management.

UniversalNER
I've read this clinical note.

User
What describes symptom in the text?

UniversalNER
["cough", "shortness of breath", "chest tightness", "difficulty breathing", "fever", "chills"]

User
What describes treatment in the text?

UniversalNER
["medication", "antihistamines"]

User
What describes time duration in the text?

UniversalNER
["two weeks", "5 years ago"]

User
What describes medical condition in the text?

UniversalNER
["Hypertension", "Allergic Rhinitis", "Obesity"]

UniversalNER

User
Read this abstract below:
Relation extraction (RE), which has relied on structurally annotated corpora for model training, has been particularly challenging in low-resource scenarios and domains. Recent literature has tackled low-resource RE by self-supervised learning, where the solution involves pretraining the entity pair embedding by RE-based objective and finetuning on labeled data by classification-based objective. However, a critical challenge to this approach is the gap in objectives, which prevents the RE model from fully utilizing the knowledge in pretrained representations. In this paper, we aim at bridging the gap and propose to pretrain and finetune the RE model using consistent objectives of contrastive learning. Since in this kind of representation learning paradigm, one relation may easily form multiple clusters in the representation space, we further propose a multi-center contrastive loss that allows one relation to form multiple clusters to better align with pretraining. Experiments on two document-level RE datasets, BioRED and Re-DocRED, demonstrate the effectiveness of our method. Particularly, when using 1% end-task training data, our method outperforms PLM-based RE classifier by 10.5% and 6.1% on the two datasets, respectively.

UniversalNER
I've read this abstract.

User
What describes task in the text?

UniversalNER
["Relation extraction", "low-resource RE", "document-level RE"]

User
What describes dataset in the text?

UniversalNER
["BioRED", "Re-DocRED"]

User
What describes method in the text?

UniversalNER
["self-supervised learning", "pretraining", "finetuning", "contrastive learning", "multi-center contrastive loss", "end-task training data"]

User
What describes objective in the text?

UniversalNER
["classification-based objective", "contrastive learning", "multi-center contrastive loss"]

From Real-World Data to Discovery Engine

Case Study: Synthetic Control

Average cost of an FDA-approved drug		Annual number of FDA-approved drugs		Per Year
\$2.5-10 billion	$\times$	~50	=	\$125-500 billion

Phase-3 Trial

Case: New drug

Control: Standard-of-care

Thousands of patients
Cost hundreds of million

Can we reduce this cost?

Case Study: Synthetic Control

EMR: Standard of care ⇒ Virtual control arm

Case study: Flatiron

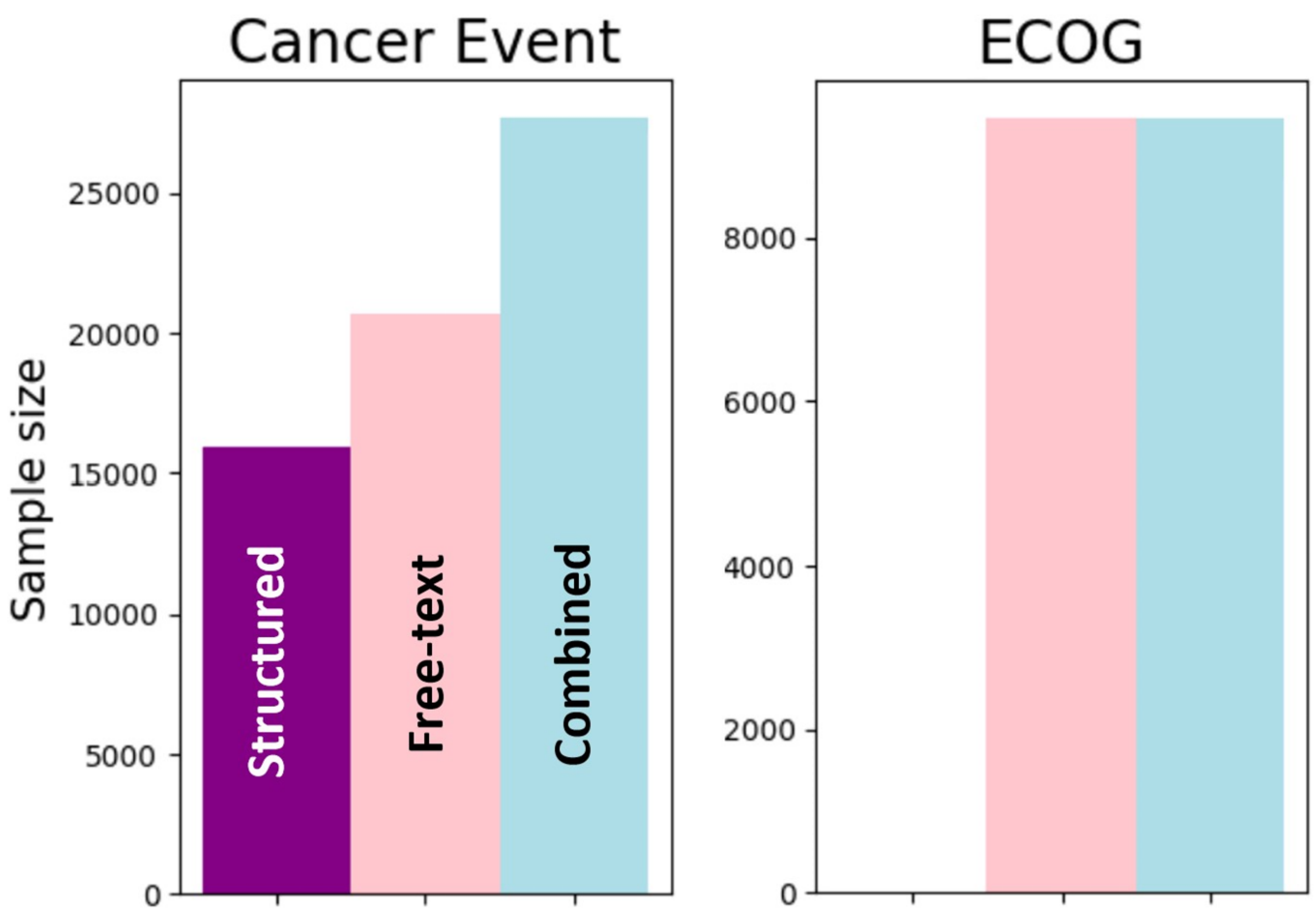
Hire hundreds of abstractors

- Pfizer: Ibrance for male breast cancer
- Roche: Alectinib for ALK lung cancer

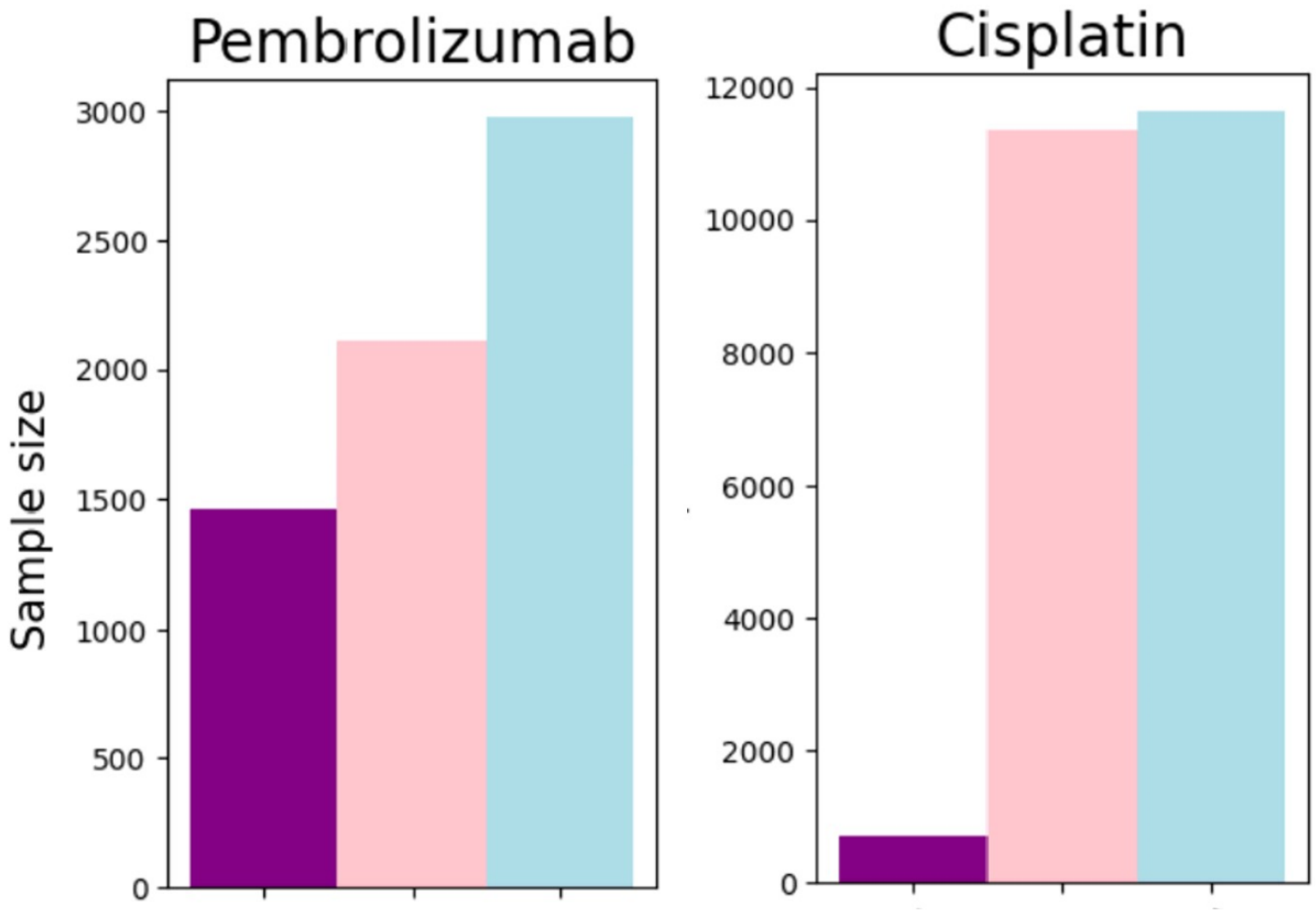
Roche to acquire Flatiron Health for $2.1 billion, with focus on real-world data

March 02, 2018 | Vol.44 No.09

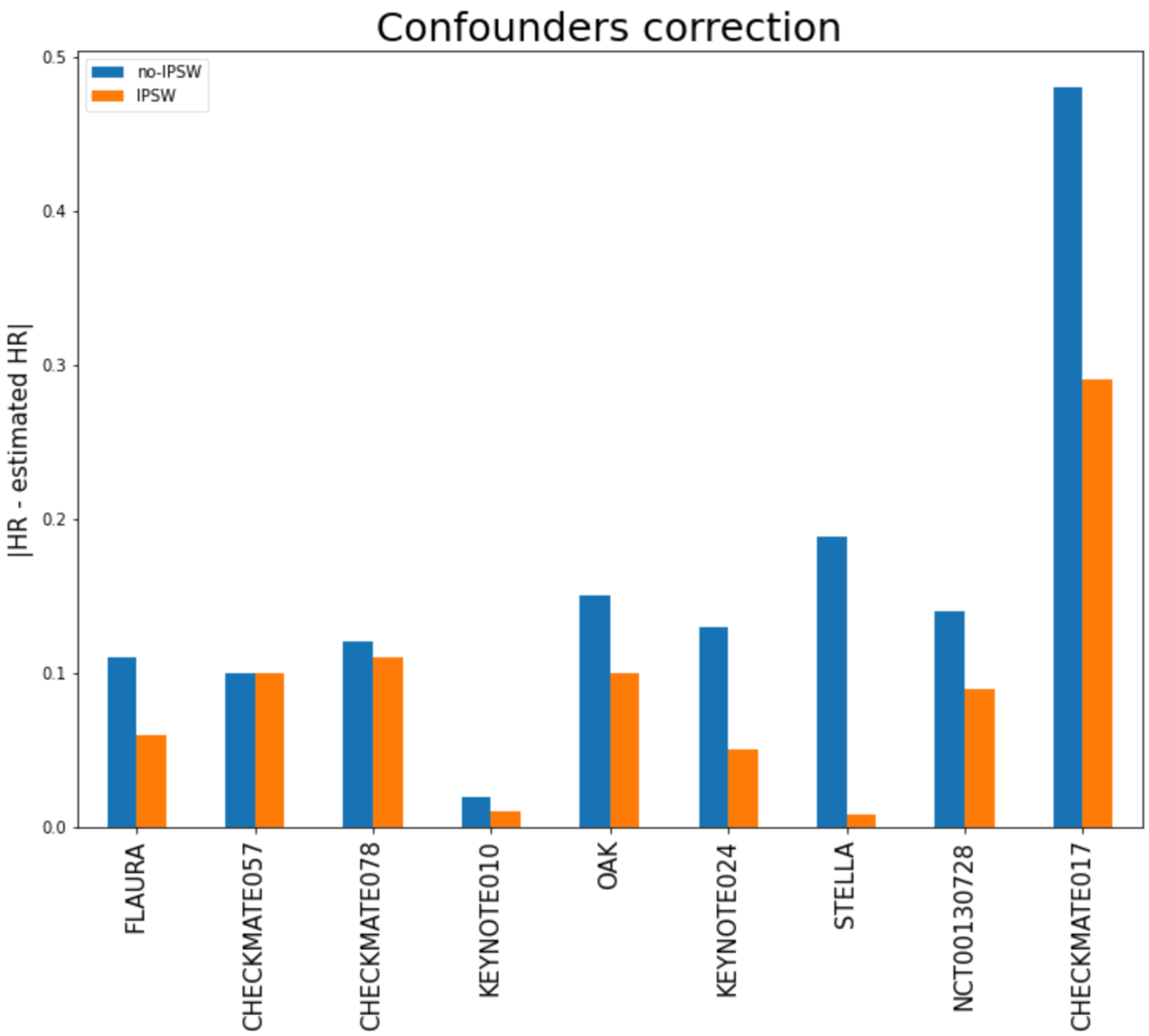
LLM: Universal Structuring



LLM: Universal Structuring



Causal Inference: Correcting for Confounders



Substantially reduces difference from gold RCT results

Towards Population-Scale Causal Discovery

LLMs = Spark

RCT simulation

- Simulation long term outcomes
- Root cause analysis
- Synthetic control arms
- Trial optimization
- Subpopulation analysis/fairness

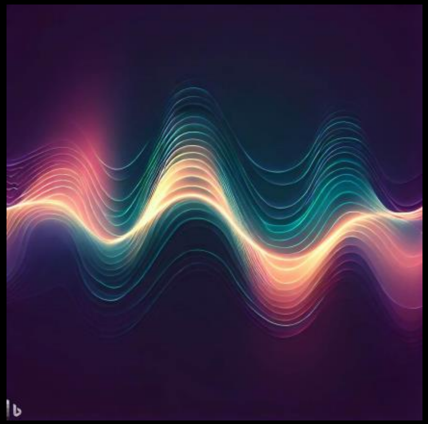
Empower every stakeholder in precision health discovery

Multi-Modal, Longitudinal Patient Data

Growth Area for General LLMs

Multimodal models will be able to understand and reason about…

Images



Speech

Proteins

DNA

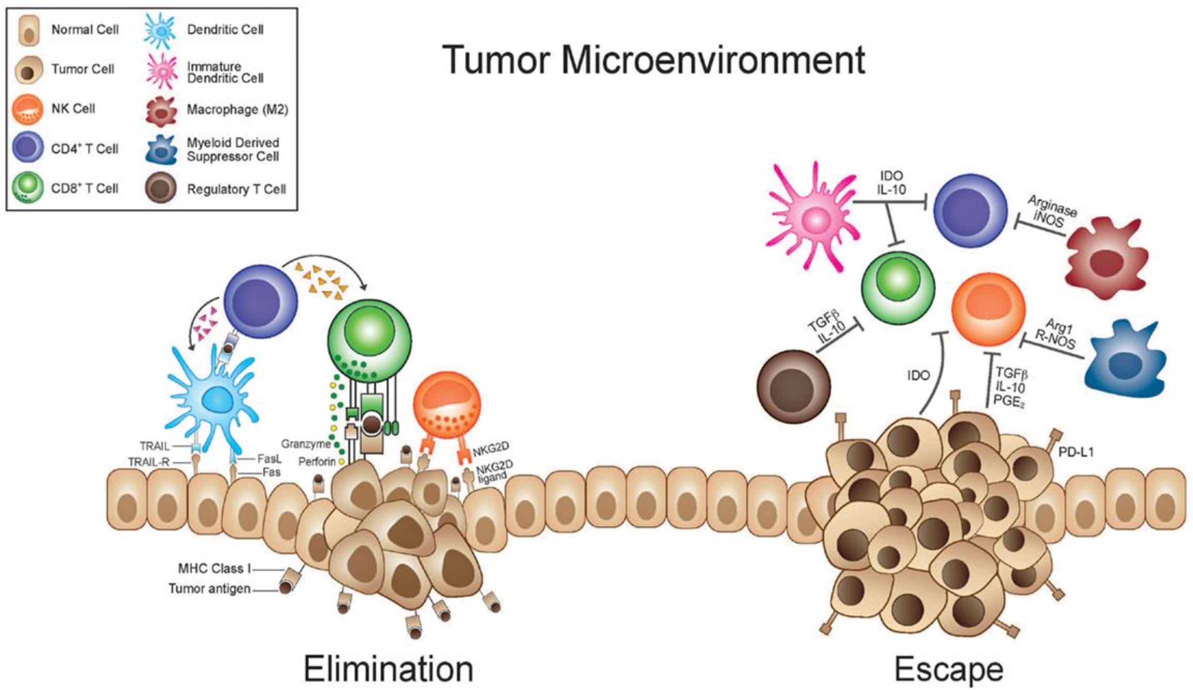
Molecules

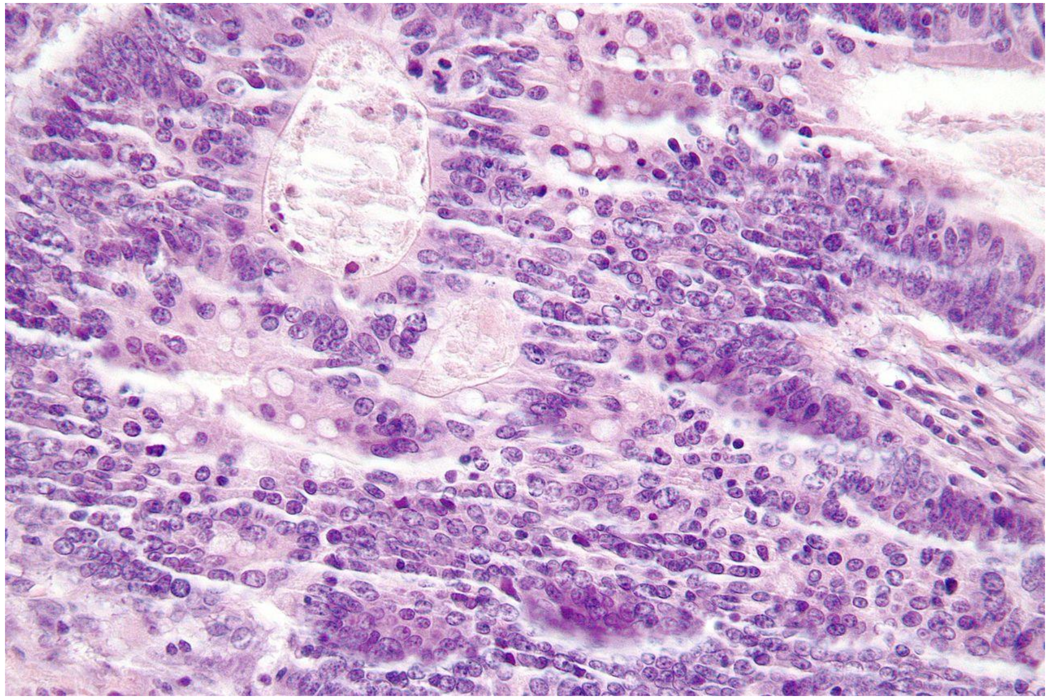
Image credits: Bing Image Creator

Case Study: Immunotherapy

Given Keytruda cohort, find exceptional responder

Need to model tumor microenvironment



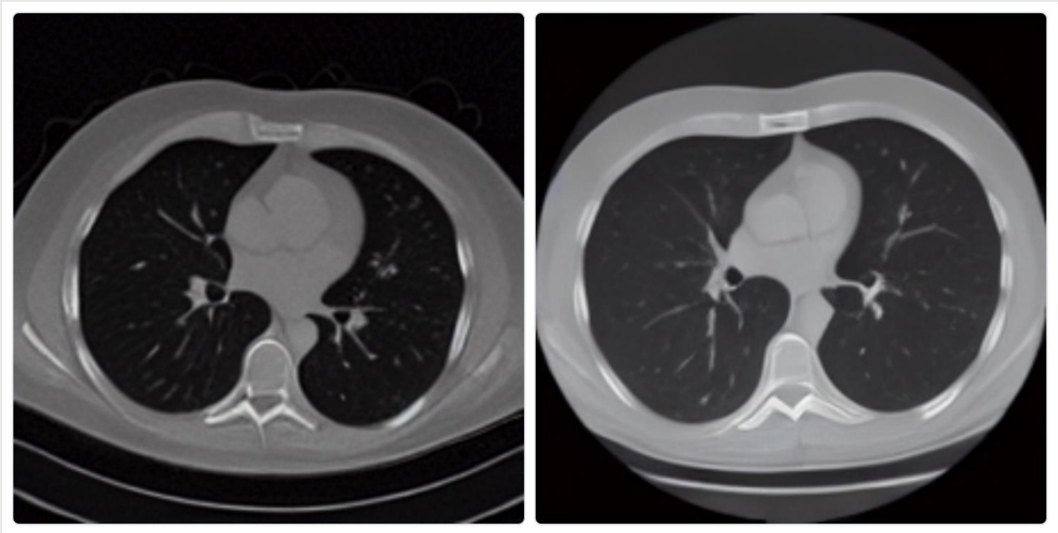


https://en.wikipedia.org/wiki/Tumor-infiltrating_lymphocytes

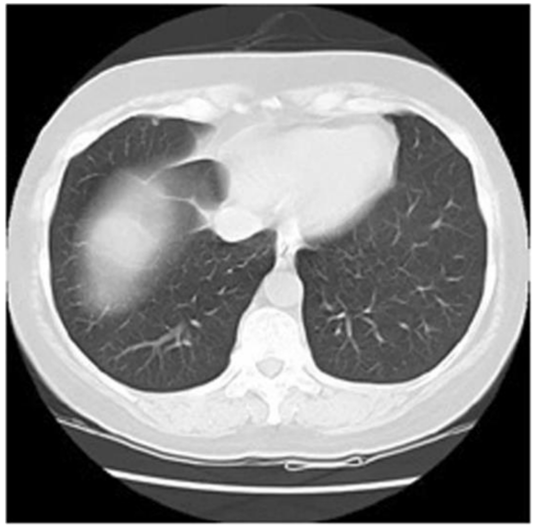
Multi-Modal: Beyond General Domain

Generic

Domain-Specific
(1 hour on one A100)



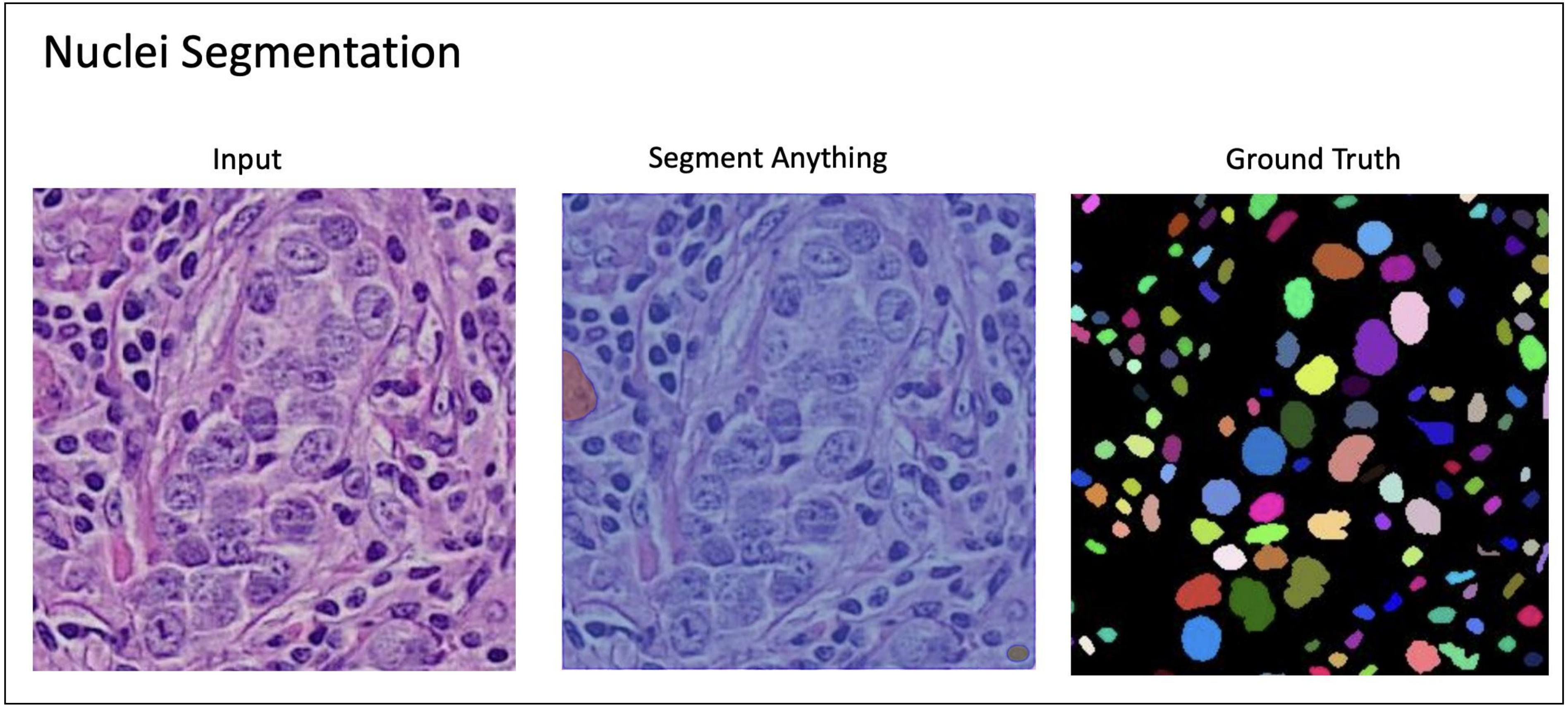
Reference
Example



https://www.nature.com/articles/s41598-019-41510-9/figures/1

"A photo of a lung CT scan"

SAM: A Lot of Growth Opportunities Ahead



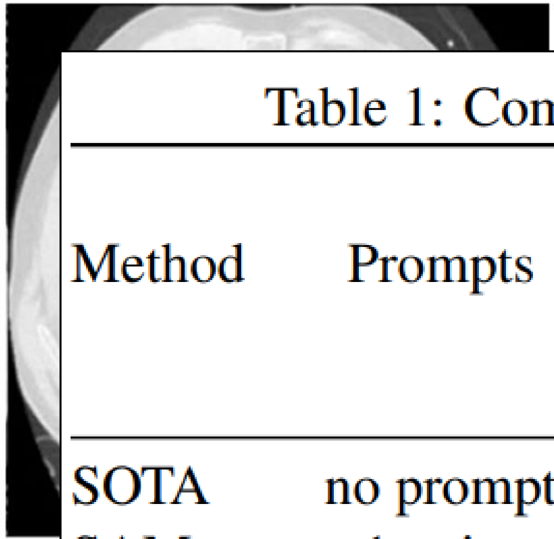
SAM: A Lot of Growth Opportunities Ahead

SAM: A Lot of Growth Opportunities Ahead

Nuclei Segmentation

COVID-19 Segmentation

Input



Segment Anything

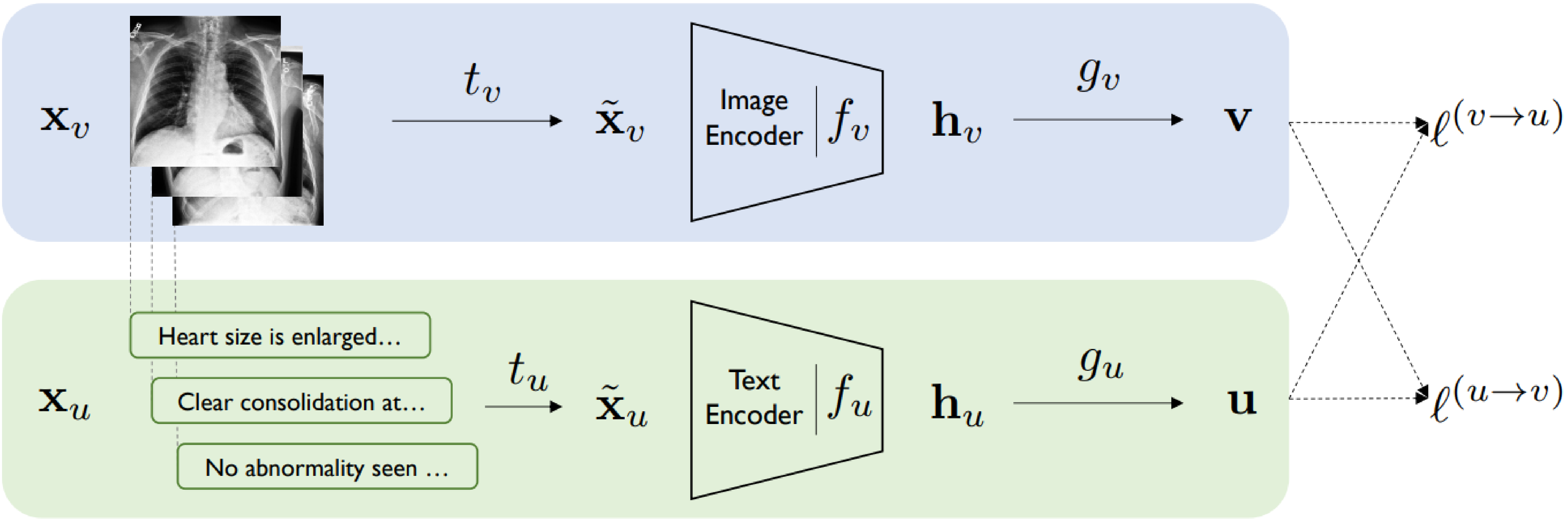
Ground Truth

Table 1: Compare SAM with state-of-the-art (SOTA) methods. (Unit: Dice score)

Method	Prompts	Tumor	Tissue						Cell
		0.5×	5×		10×			40×	40×
		Tumor	CAP	TUFT	DT	PT	VES	PTC	Nuclei
SOTA	no prompt	71.98	96.50	96.59	81.01	89.80	85.05	77.23	81.77
SAM	1 point	58.71	78.08	80.11	58.93	49.72	65.26	67.03	1.95
SAM	20 points	74.98	80.12	79.92	60.35	66.57	68.51	64.63	41.65

Biomedical Large Multimodal Models

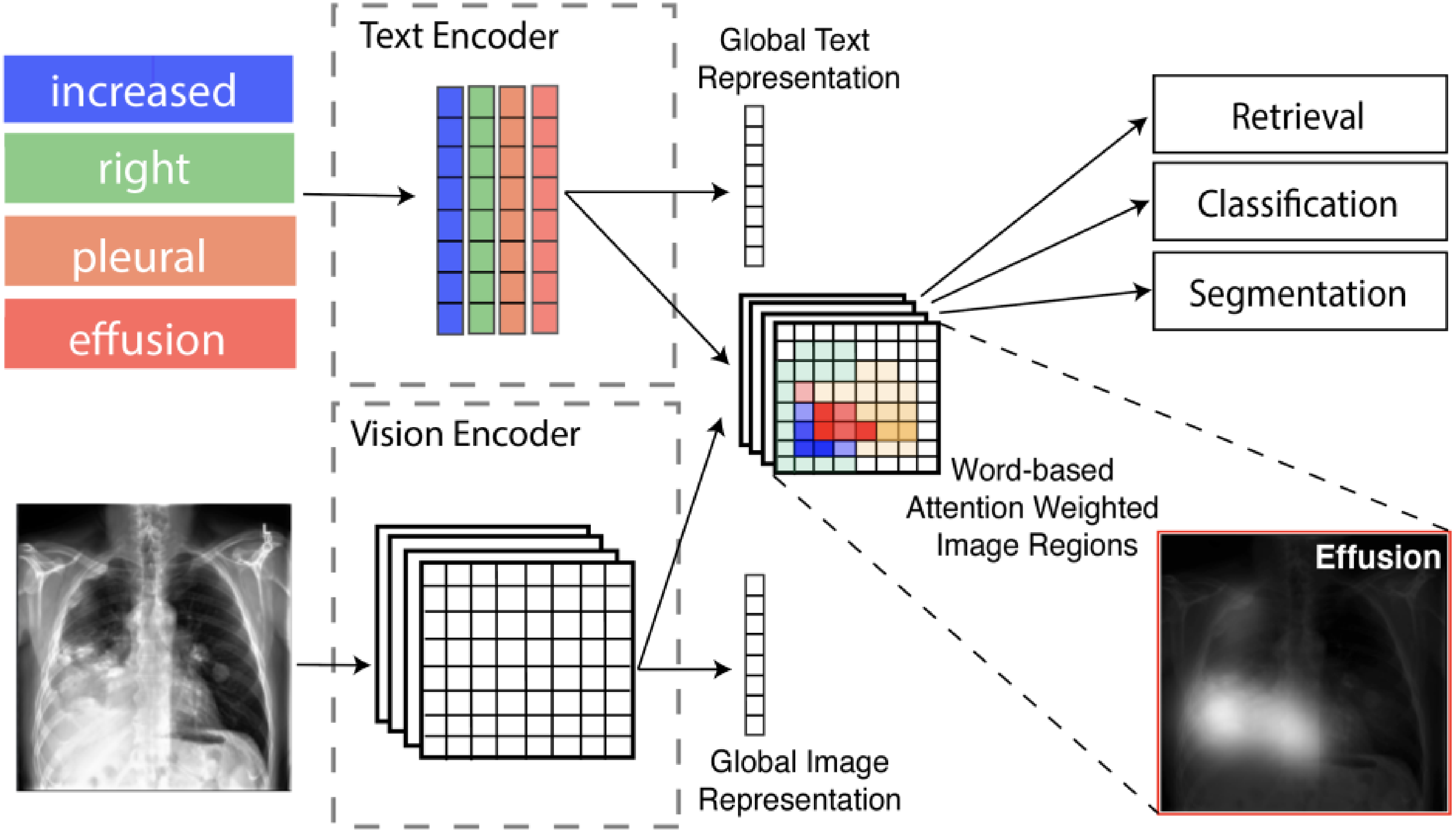
ConVIRT



Zhang, et al. "Contrastive Learning of Medical Visual Representations from Paired Images and Text", *MLHC 2022*.

Biomedical Large Multimodal Models

GLoRIA

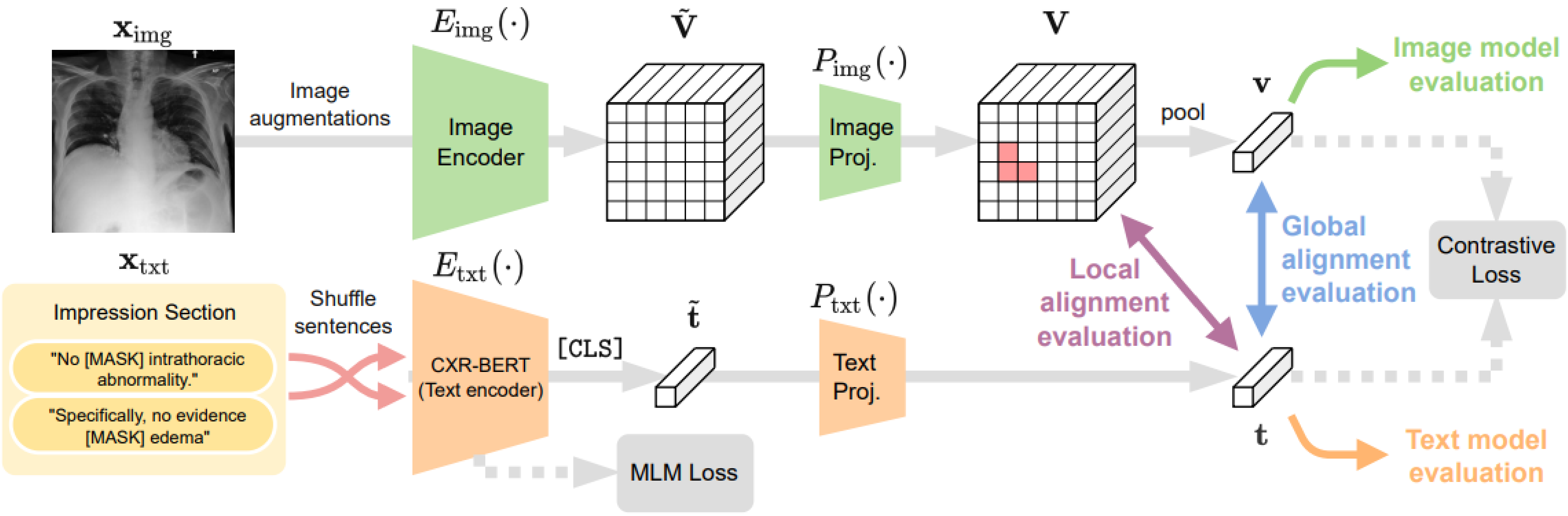


Huang, et al. "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition", *ICCV 2021*.

Global + Local Alignment

Biomedical Large Multimodal Models

BioViL

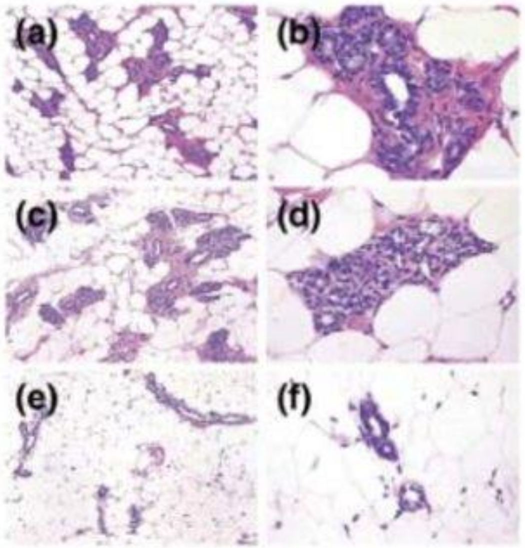


Boecking*, Usuyama*, et al. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. *ECCV 2022*.

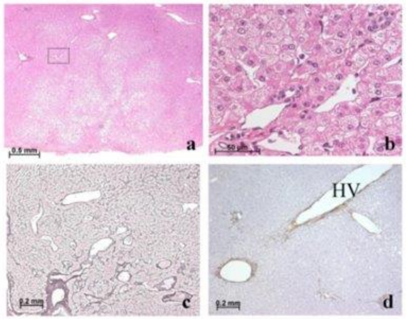
Global + Local Alignment
Radiology-specific language modeling

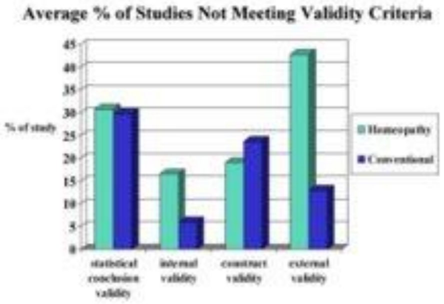
BiomedCLIP

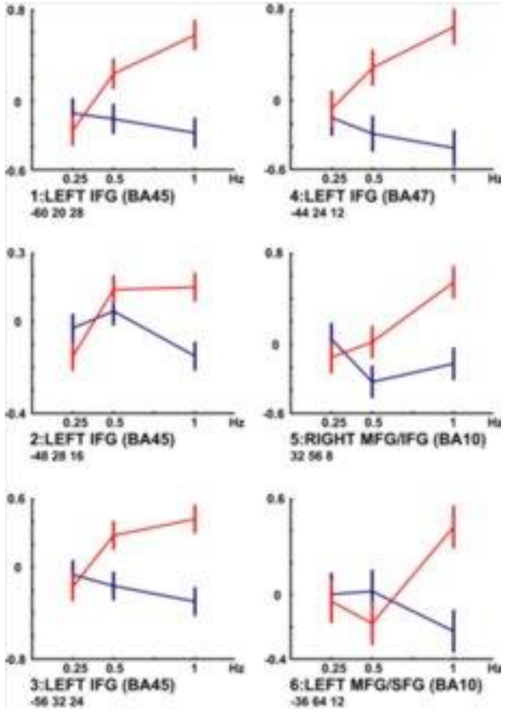
30 million PMC figure-caption pairs

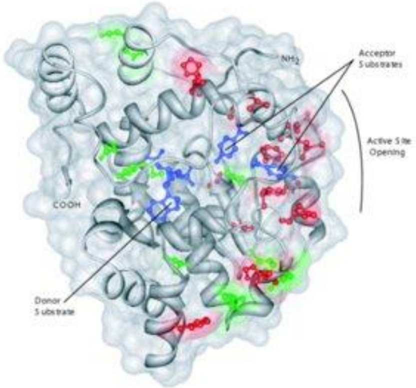


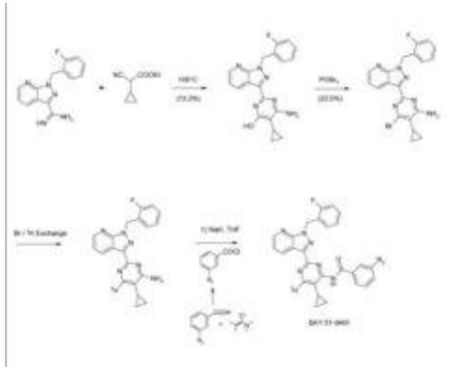
Treatment with tamoxifen affects the histology of the rat mammary gland. Representative hematoxylin and eosin stained sections of the first thoracic gland of 15-week-old rats that had undergone the following treatments: (a, b) No treatment; moderate numbers of mammary gland lobules are present containing primary, secondary and tertiary ductules, as well as developing alveoli. (c, d) …



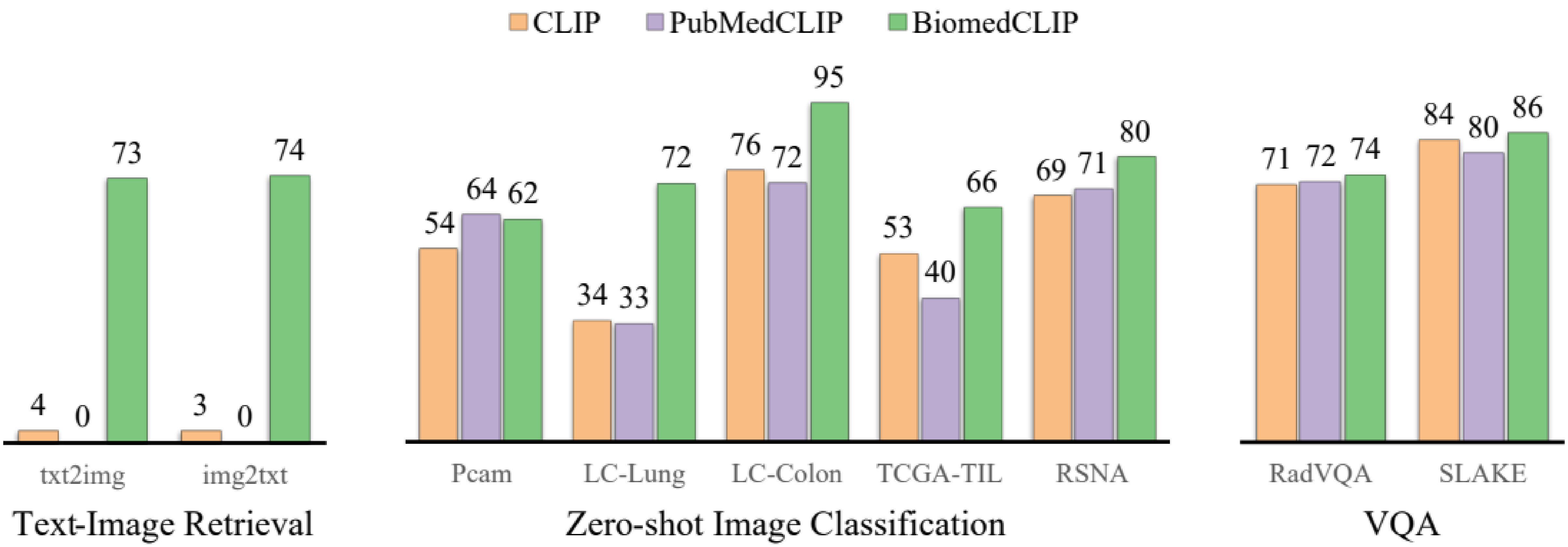








BiomedCLIP: New State of the Art



Zhang*, Xu*, Usuyama*, et al. Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing. *In submission.*

BiomedCLIP: Zero-Shot Image Classification

model	pretraining data	zero-shot
CLIP	WIT-400M	68.80
MedCLIP	MIMIC-CXR + CheXpert	66.96
PubMedCLIP	ROCO	70.70
GLoRIA	CheXpert	70.00
BioViL	MIMIC-CXR	73.20
BiomedCLIP	PMC-15M	**79.72**

RSNA Pneumonia: Outperforms radiology-specific models
"Quantity has a quality all its own"

Towards Multi-Modal Research Copilot

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

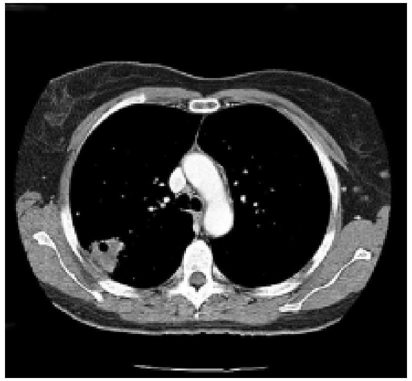
Chunyuan Li*, Cliff Wong*, Sheng Zhang*, Naoto Usuyama, Haotian Liu, Jianwei Yang
Tristan Naumann, Hoifung Poon, Jianfeng Gao

Microsoft
https://aka.ms/llava-med

Biomedical Visual Instruction-Following Example

Caption:

Figure 4: Contrast-enhanced CT scan of the chest for patient #1. A large cavitating lesion was identified in the right upper lobe with pleural involvement and likely extension into the chest wall. In addition, right hilar lymphadenopathy was suspected. This was the primary malignant lesion in this patient.



Sentences with in-line figure mentions:

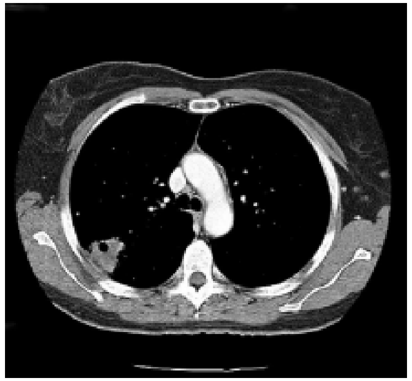
- Computerized tomography (CT) scans of the chest (Figure 4) were obtained.
- (Figure 4) and demonstrated a large cavitating lesion in the posterior aspect of the right upper lobe with probable pleural involvement likely representing the primary lesion.

BiomedCLIP Data

Biomedical Visual Instruction-Following Example

Caption:

Figure 4: Contrast-enhanced CT scan of the chest for patient #1. A large cavitating lesion was identified in the right upper lobe with pleural involvement and likely extension into the chest wall. In addition, right hilar lymphadenopathy was suspected. This was the primary malignant lesion in this patient.



Sentences with in-line figure mentions:

- Computerized tomography (CT) scans of the chest (Figure 4) were obtained.
- (Figure 4) and demonstrated a large cavitating lesion in the posterior aspect of the right upper lobe with probable pleural involvement likely representing the primary lesion.

Conversation:

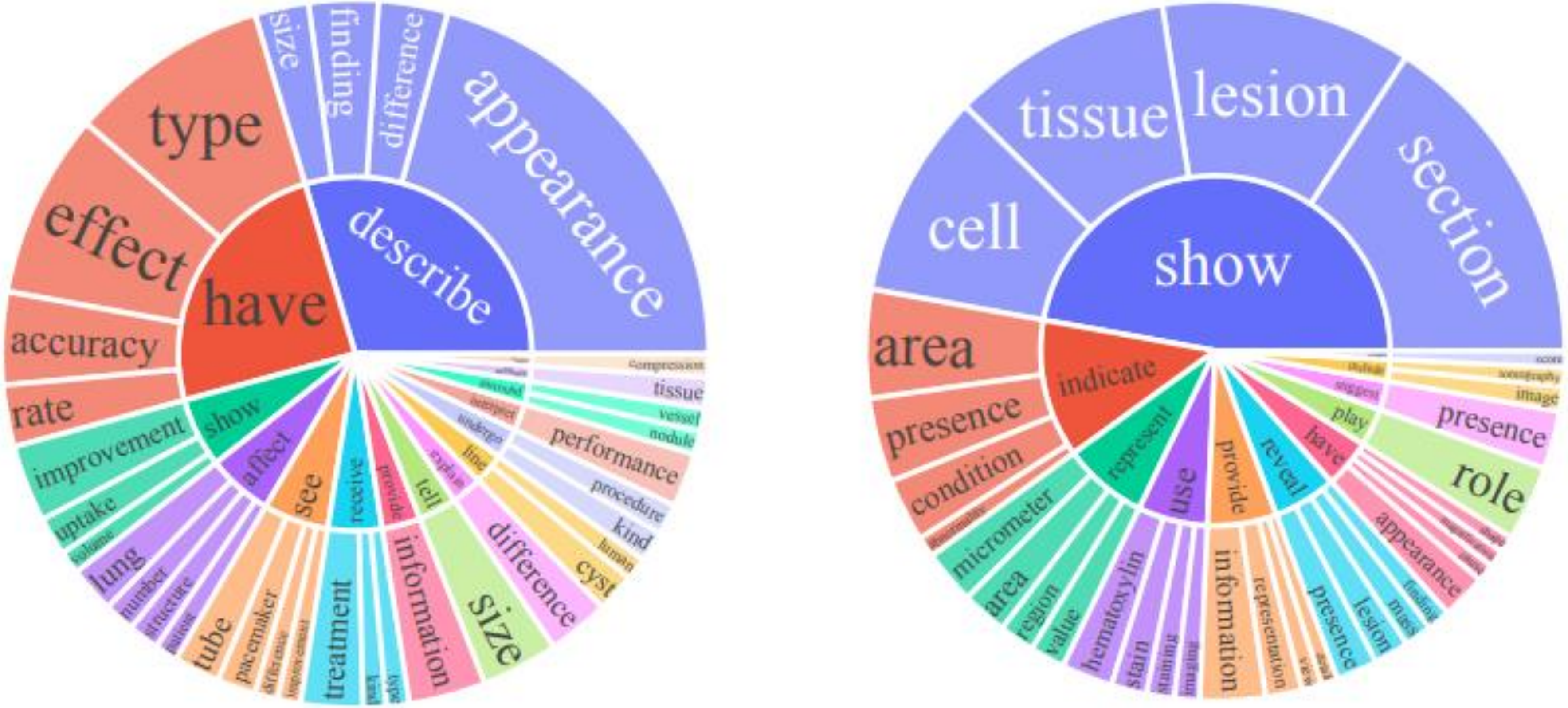
```
User: Where is the lesion located?

Assistant: The lesion is located in the posterior aspect of the right upper lobe of the lung.

User: Is it extending into the chest wall?

Assistant: Yes, the CT scan suggests that the large cavitating lesion in the right upper lobe is
likely extending into the chest wall.

User: Can you explain more about the cavitating lesion?

Assistant: Certainly. A cavitating lesion refers to an abnormal mass or growth within the lung
tissue that has a cavity or hollow space inside it. This cavity can be formed due to the necrosis
(death) of the tissue inside the lesion, leading to the breakdown and disintegration of the
affected tissue. Cavitating lesions can be caused by various factors, such as infections (like
tuberculosis or fungal infections), autoimmune diseases, or malignancies (cancer). In the context
of the CT scan provided, the large cavitating lesion in the right upper lobe of the lung is
suggestive of a malignant (cancerous) growth. The cavity within the lesion may be a result of the
rapid and uncontrolled growth of cancer cells, which can lead to the destruction of surrounding
healthy tissue. It's important to note that further diagnostic tests, such as a biopsy, would be
required to confirm the malignancy and determine the specific type of cancer involved.
```

BiomedCLIP Data

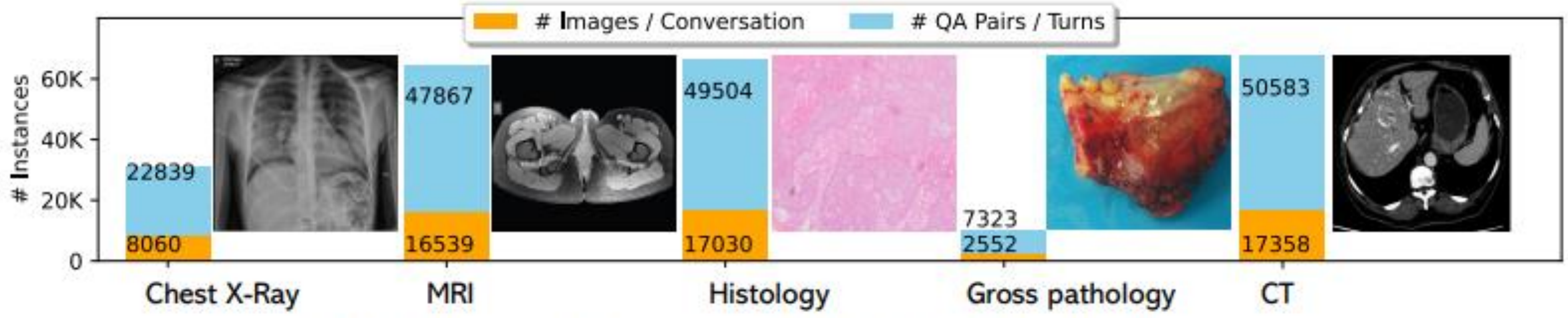
GPT-4

Universal Annotator

Multimodal Instruction-Following

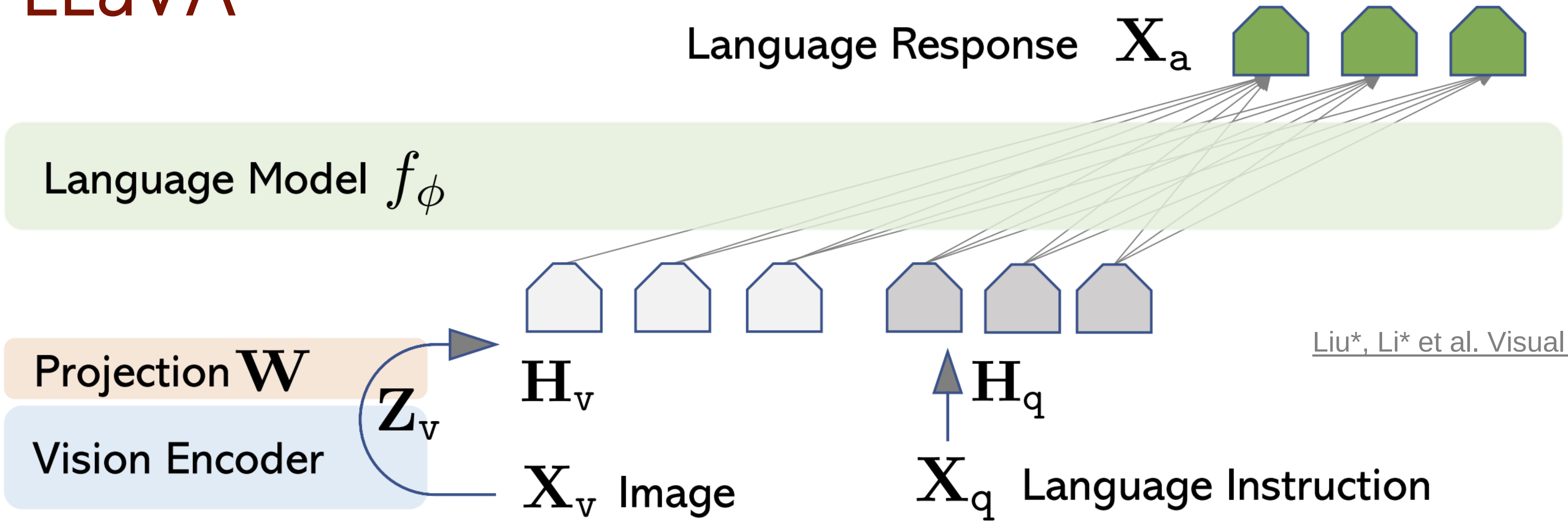


(a) Instruction (b) Responses



(c) Frequencies of images and QA pairs on the five domains.

LLaVA

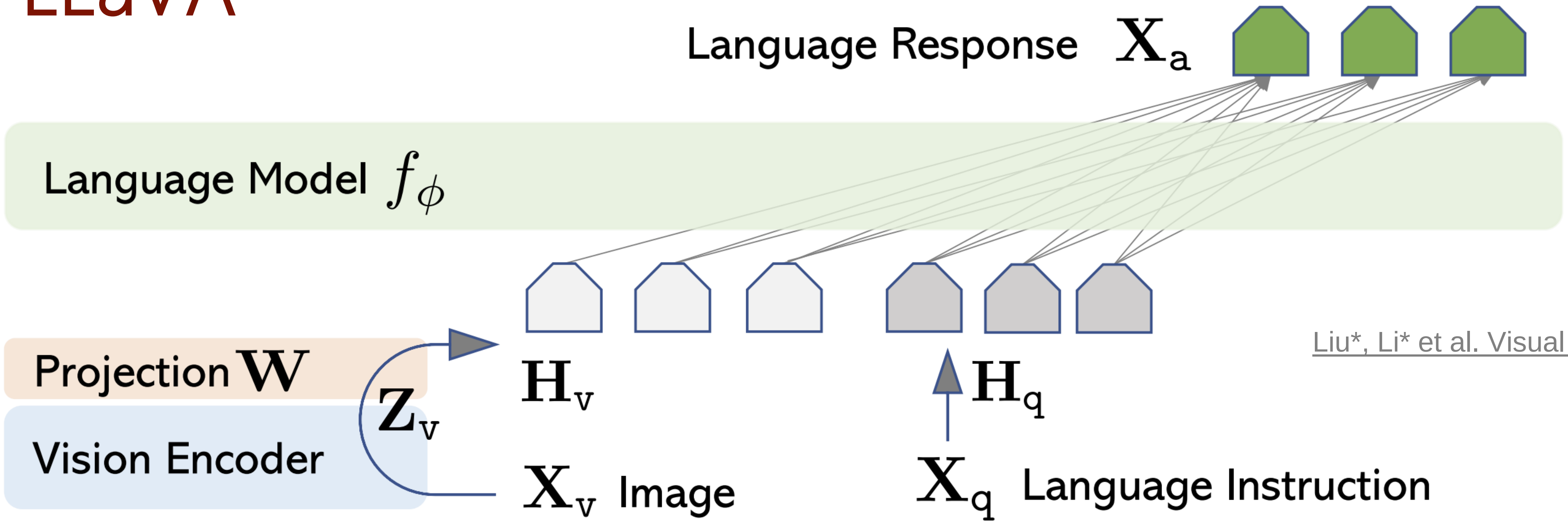


Liu*, Li* et al. Visual Instruction Tuning.

Key Insight

Introduce a projection layer to convert image into text embedding

LLaVA

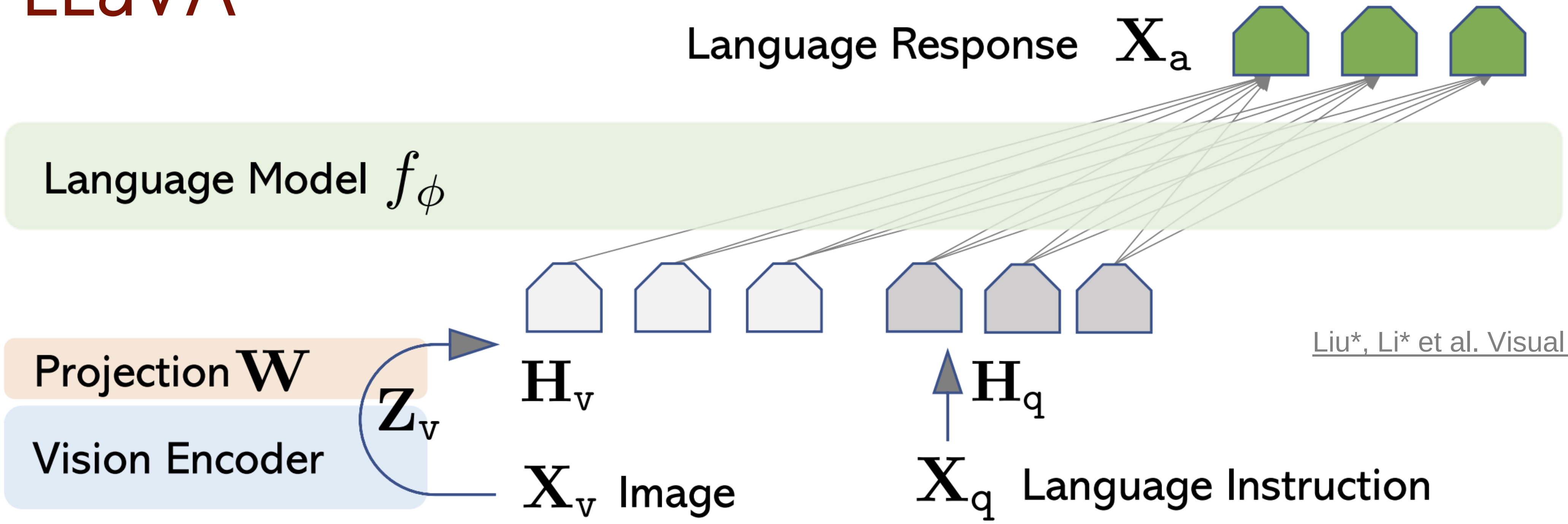


Liu*, Li* et al. Visual Instruction Tuning.

Language Model: LLaMA
Vision Encoder: CLIP
Projection: Linear

Family of Models

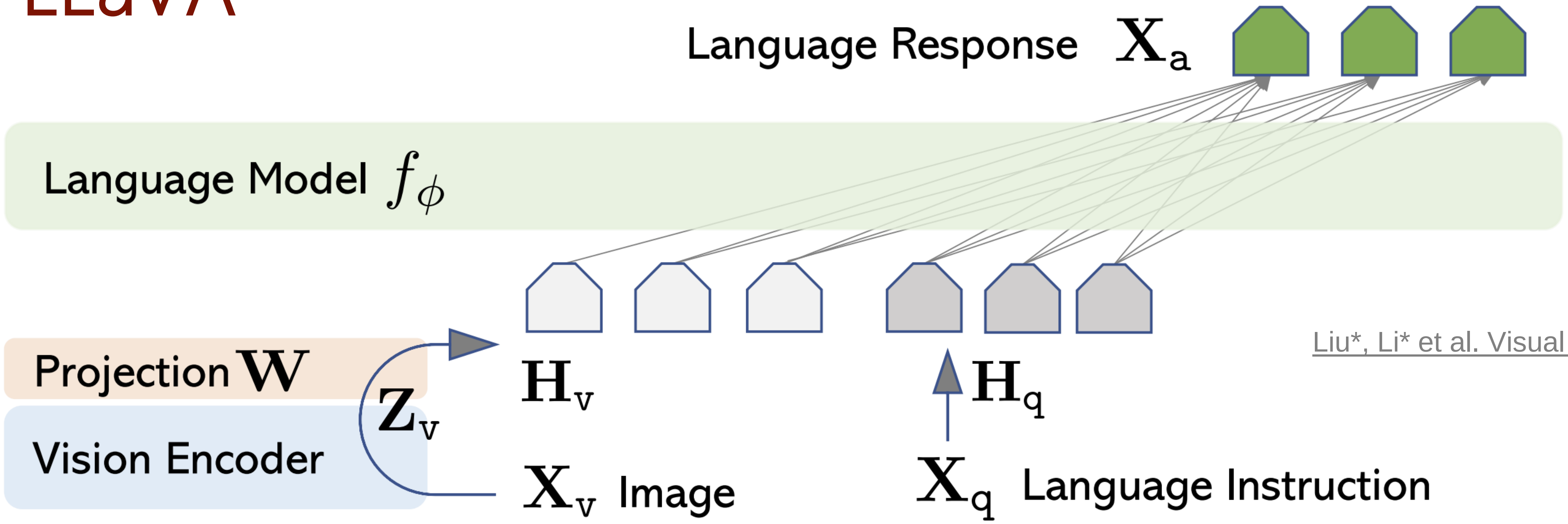
LLaVA



Liu*, Li* et al. Visual Instruction Tuning.

Stage 1
Frozen: language model, vision encoder
Train: projection
Instruction: describe image

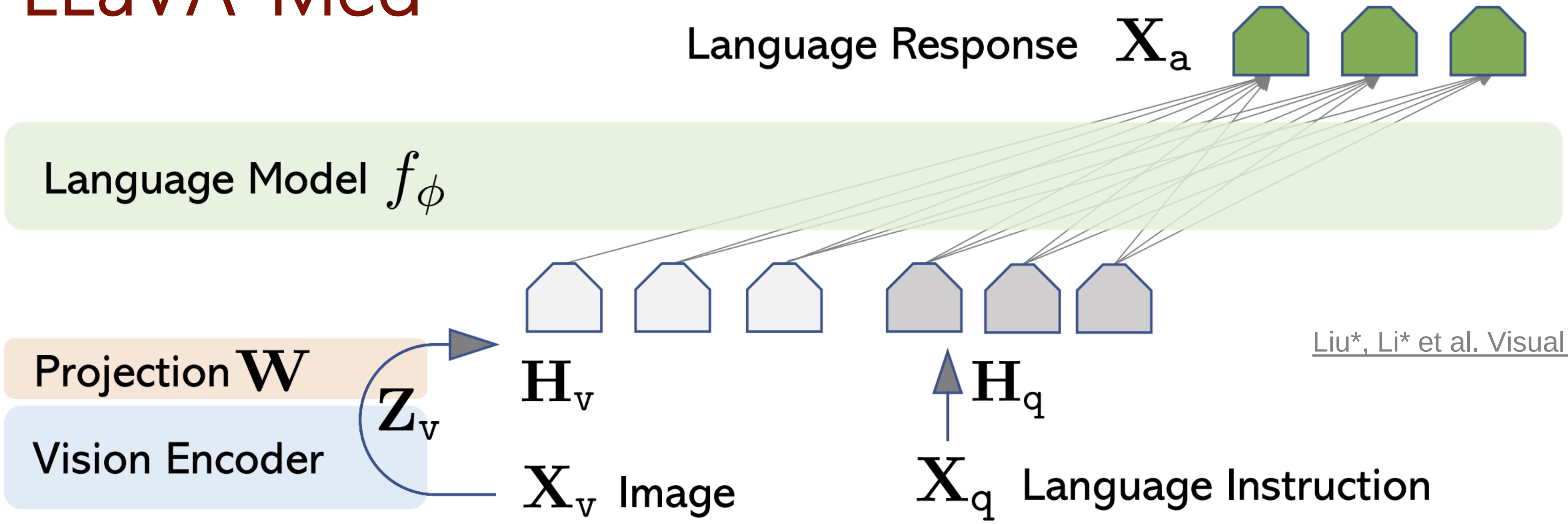
LLaVA



Liu*, Li* et al. Visual Instruction Tuning.

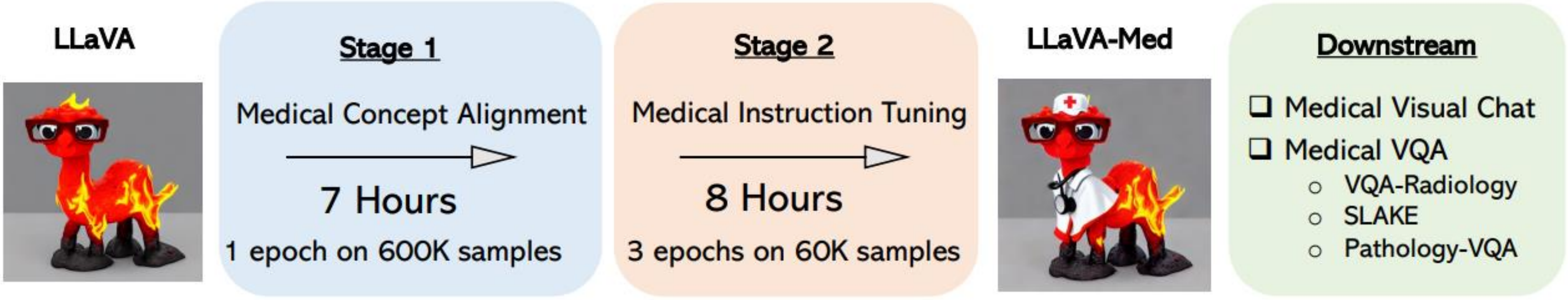
Stage 2
Frozen: vision encoder
Train: projection, language model
Instruction: GPT-4 generated conversations

LLaVA-Med



Liu*, Li* et al. Visual Instruction Tuning.

Language Model: LLaVA LM
Vision Encoder: BiomedCLIP
Projection: Linear



15 hours on eight A100s

MedPaLM-M

Tu*, Azizi* et al. Towards Generalist Biomedical AI.

LLaVA-Med [47] is perhaps most similar to our effort. The authors use PubMed and GPT-4 [48] to curate a multimodal instruction following dataset and finetune a LLaVA model with it.

Language Model: PaLM
Vision Encoder: ViT
Projection: Linear

Instruction-following: supervised
Prompt: task-specific

Task Type
Question Answering
Report Summarization
Visual Question Answering
Report Generation
Medical Image Classification

ELIXIR

Xu*, Yang*, Kelly*, et al. ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders.

Language Model: PaLM2
Vision Encoder: BLIP-2
Projection: Q-Former

Not general instruction-following
Radiology image/report pairs

Frozen LM/ViT (~ LLaVA stage 1)

Med-Flamingo

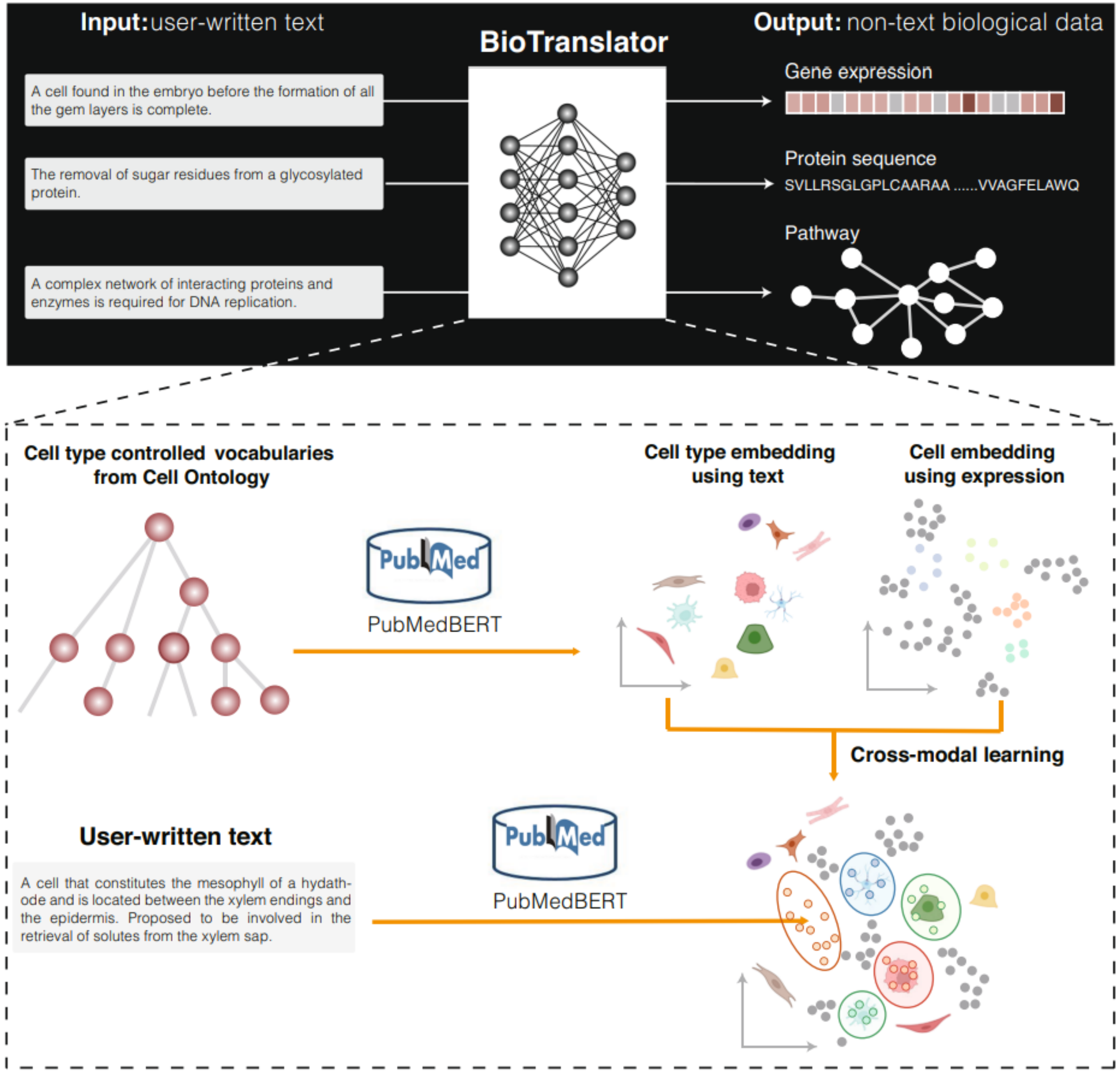
Moor*, Huang*, et al. Med-Flamingo: a Multimodal Medical Few-shot Learner.

Language Model: LLaMA
Vision Encoder: CLIP
Projection: Gated cross attention

Instruction-following:

- Publications / textbooks
- Supervised (VQA, Visual USMLE)

Multi-Modal: Universal Translator



Use text as common representation of knowledge & reasoning

nature communications

Article https://doi.org/10.1038/s41467-023-36476-2

Multilingual translation for zero-shot biomedical classification using BioTranslator

Received: 5 July 2022

Accepted: 1 February 2023

Hanwen Xu[1], Addie Woicik[1], Hoifung Poon[2], Russ B. Altman[3,4,5] & Sheng Wang[1]

Population-Level Health LLM

Patient → Serialized multimodal token sequence

Initialize: GPT-101 (consumed entire public web)

Continued pretraining: 8 billion "health documents"

What is the multimodal health scaling law?

Will there be emergent health capabilities?

Advancing Health at the Speed of AI