Real-World Evidence in the Age of LLMs

Tristan Naumann

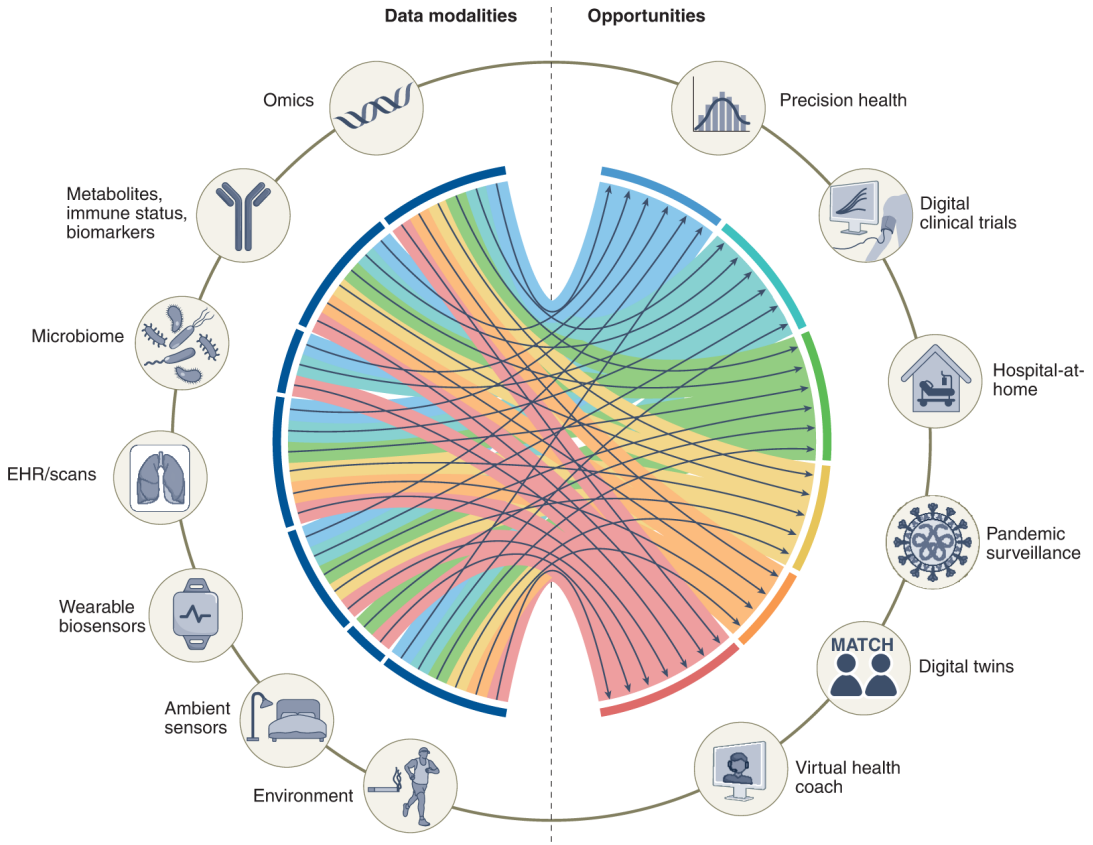
Principal Researcher
Microsoft Health Futures

tristan@microsoft.com
@TristanNaumann

NLP is key to unlocking the future of medicine from clinical text

Multi-modal is key to unlocking new opportunity

- Health is not limited to a single data modality
- Health is not limited to a single setting
- Holistic models build understanding of health



Acosta et al. Multimodal Biomedical AI. Nat Med. 2022

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

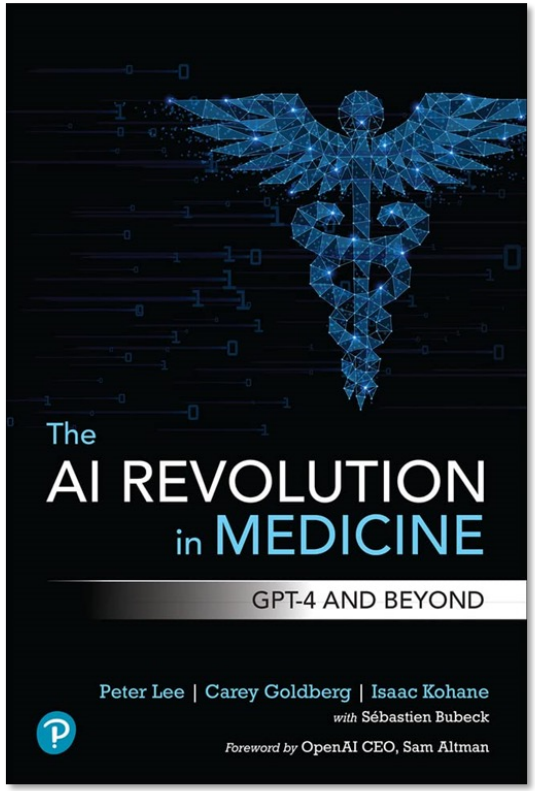
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

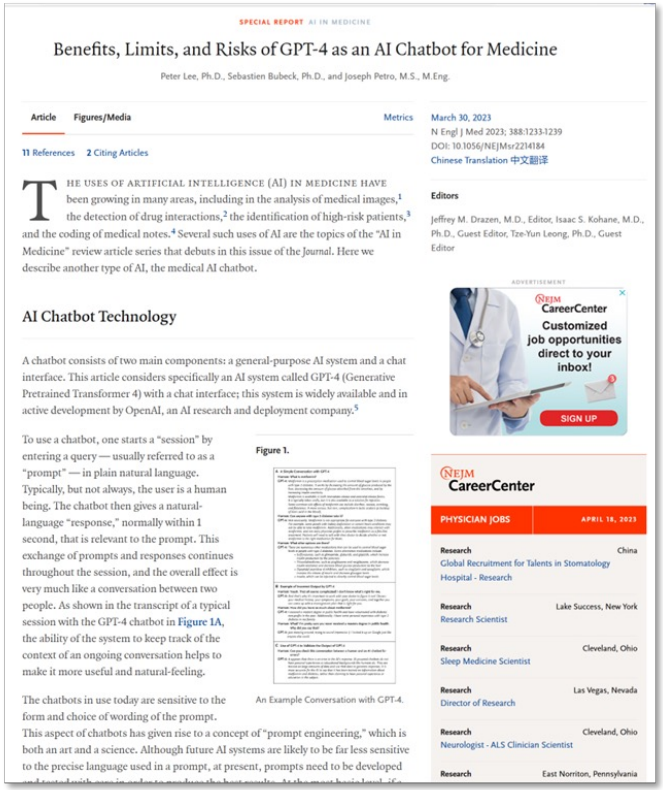
Microsoft Research

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





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

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

Editors

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

Use large language models to promote equity

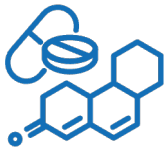
Emma Pierson[1, 2, 14], Divya Shanmugam[3, 14], Rajiv Movva[1, 14], Jon Kleinberg[4, 14], Monica Agrawal[5], Mark Dredze[6], Kadija Ferryman[6], Judy Wawira Gichoya[7], Dan Jurafsky[8], Pang Wei Koh[9], Karen Levy[4], Sendhil Mullainathan[10], Ziad Obermeyer[11], Harini Suresh[12], Keyon Vafa[13]

85% of LLM papers studying equity impacts focus on equity harms

- LLMs can improve detection of bias
- LLMs can create structured datasets of equity-relevant information
- LLMs can improve equity of access
- LLMs can improve equity in matching systems

Insight Consumer

Pharma, Payor, Regulator



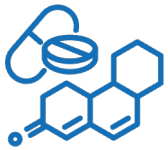
US: Less than 3% cancer patients enroll in trials
40% cancer trial failures due to insufficient patients
New drug costs $2-10 billion and takes 10+ years

Data Producer

Provider, EHR Vendor

Insight Consumer

Pharma, Payor, Regulator



Large language models → universal structuring
Instantly unlock value

Data Producer

Provider, EHR Vendor

Cabozantinib in High Grade Neuroendocrine Neoplasms

ClinicalTrials.gov Identifier: NCT04412629

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

See Contacts and Locations

Sponsor:
Washington University School of Medicine

Collaborator:
Exelixis

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

Study Description

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

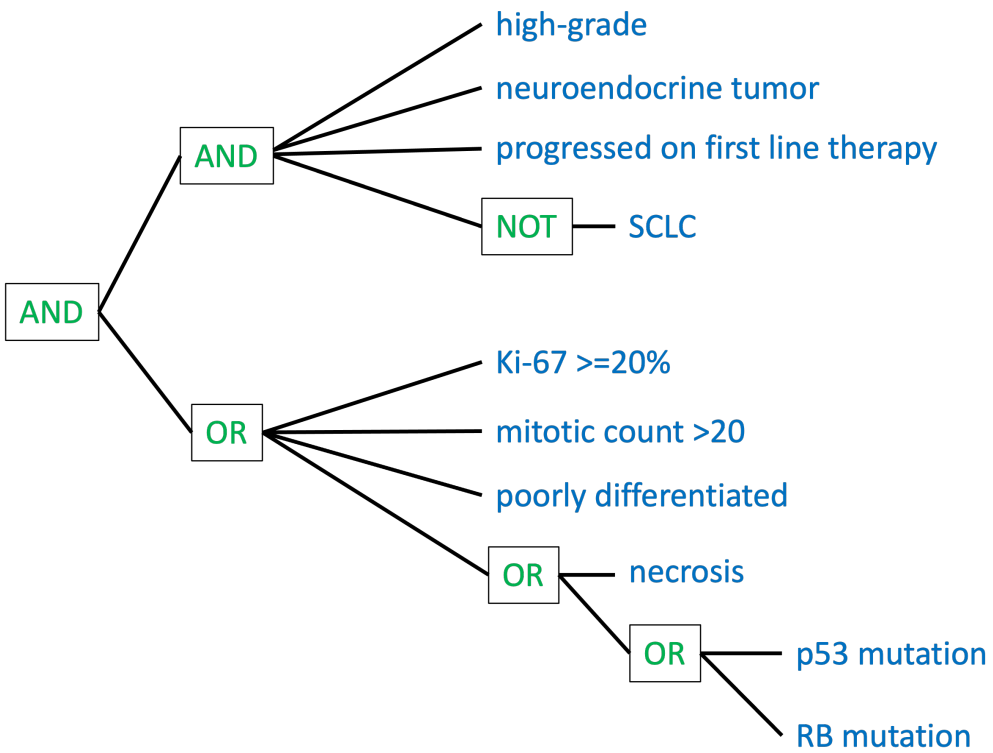
Input:
"
-Histologically or cytologically confirmed high-grade neuroendocrine tumor that has progressed on first line therapy, excluding small cell lung cancer (SCLC). High grade includes any neuroendocrine neoplasm with a Ki-67 of >=20% or with mitotic count of more than 20 mitoses per high power field 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)))
"

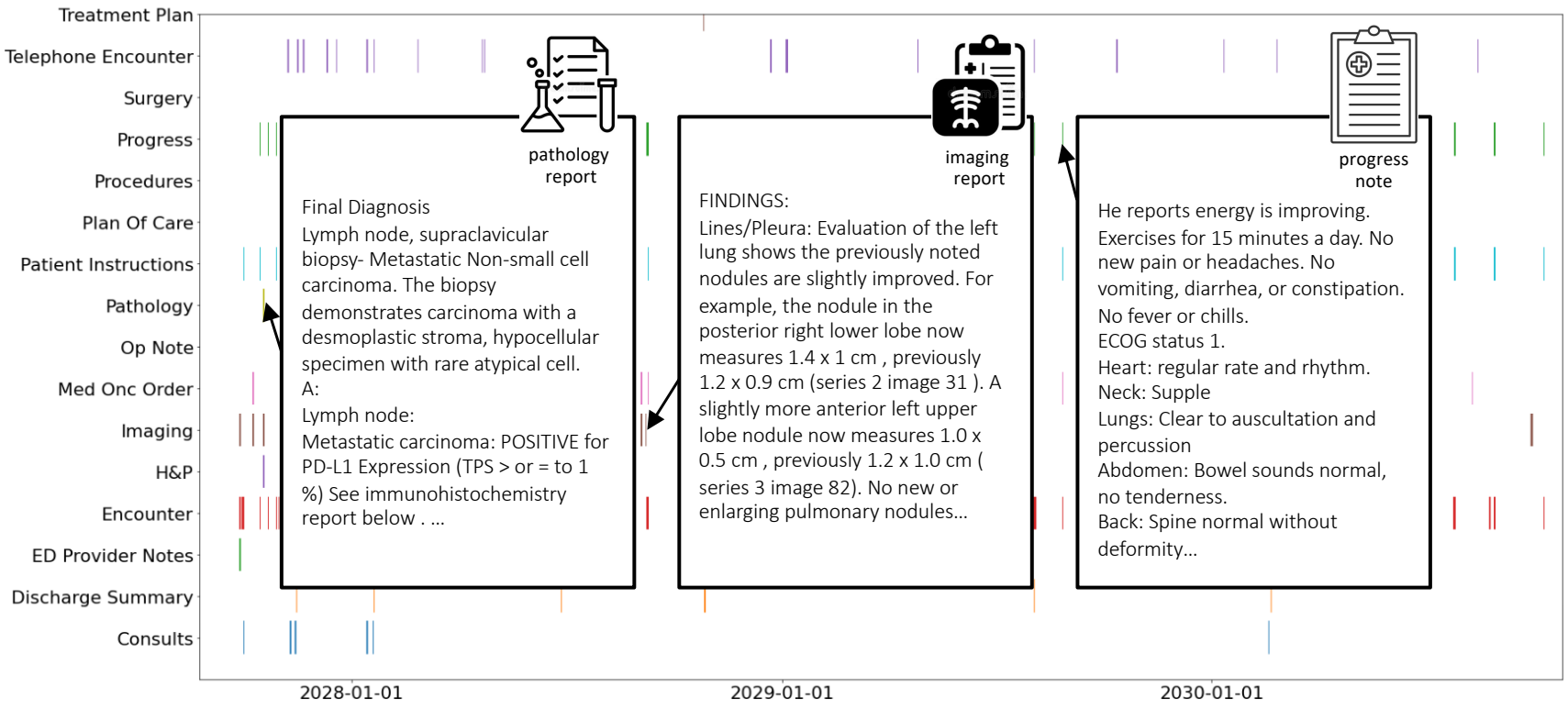


LLM: Universal Structuring

	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

Providence

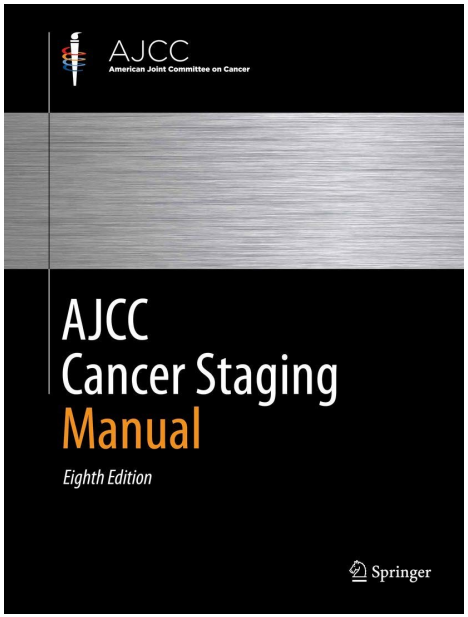
	Tumor Site	Histology	Clinical T	N	M	Pathological T	N	M
Ontology	19.4	19.2	-	-	-	-	-	-
BOW	62.8	76.6	70.4	96.6	98.4	72.1	90.7	**98.9**
OncoGloVe + CNN	72.0	84.4	74.2	96.5	98.6	83.9	93.1	98.5
OncoGloVe + HAN/GRU	74.0	85.9	76.2	97.1	98.7	86.4	94.2	98.5
BERT + HAN/GRU	75.1	86.2	77.0	96.6	98.4	86.4	94.4	98.2
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**

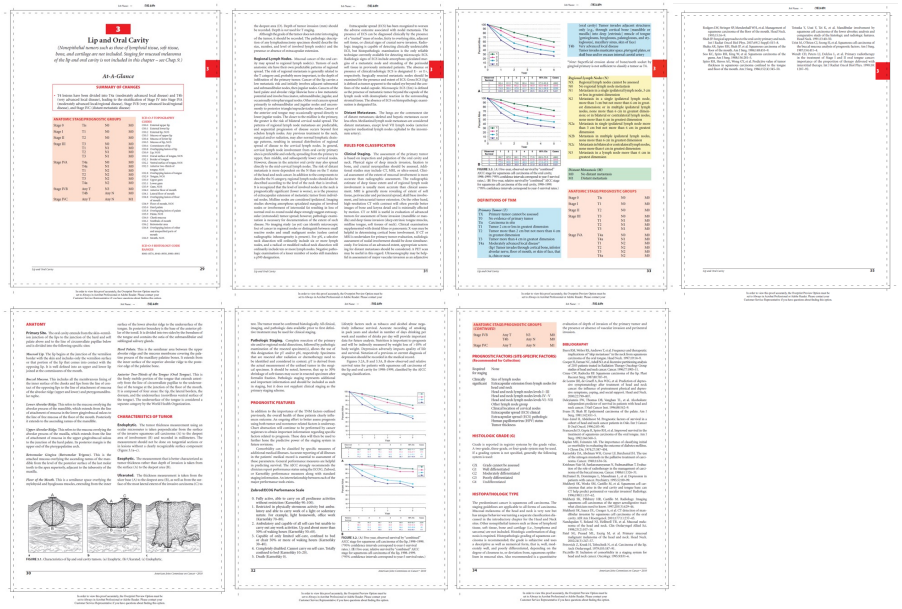
Preston, Wei, et al. "Towards Structuring Real-World Data at Scale: Deep Learning for Extracting Key Oncology Information from Clinical Text with Patient-Level Supervision", *Patterns 2023*.

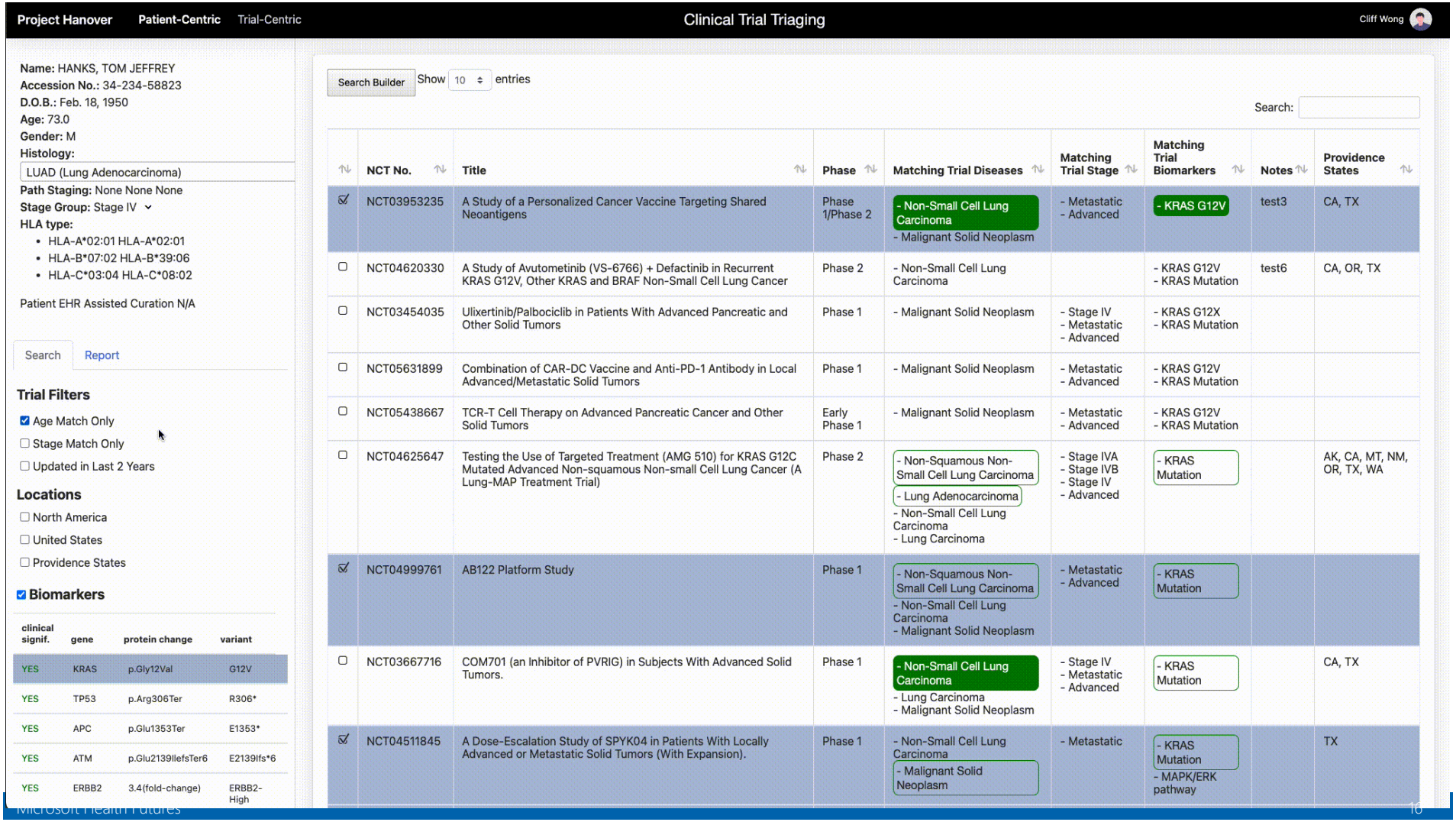
GPT-4: Structure Real-World Data

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
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

Biomedical LLMs

General vs Health Labeled Data

IMPRESSION

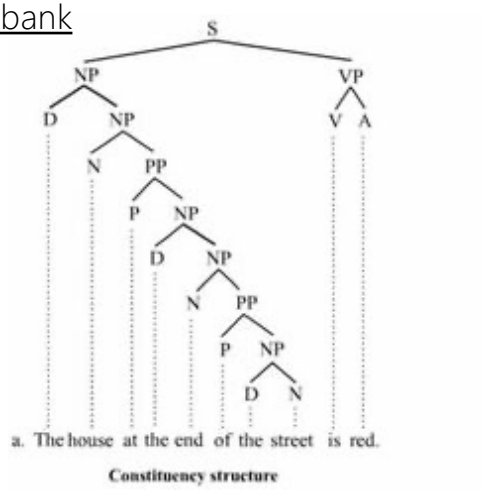
No significant change in right middle and low lobe pneumonia. Small increase in left pleural effusion. ……

Two cows are grazing in the field.

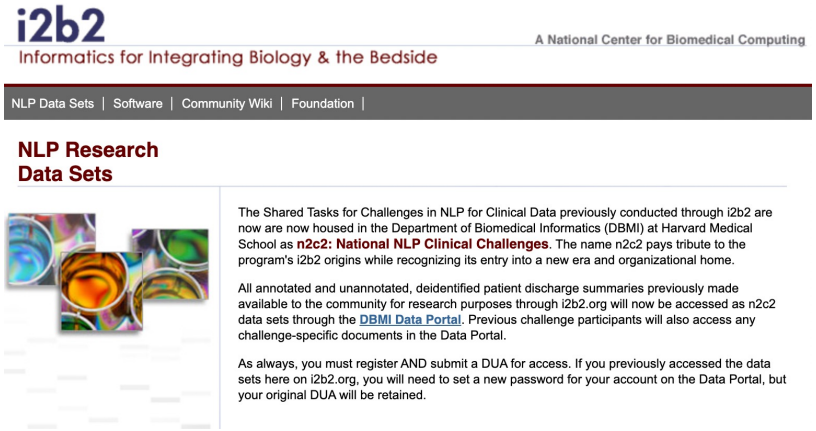
Biomedical and clinical domain label require expertise

General vs Health Data Availability

Penn Treebank



1992

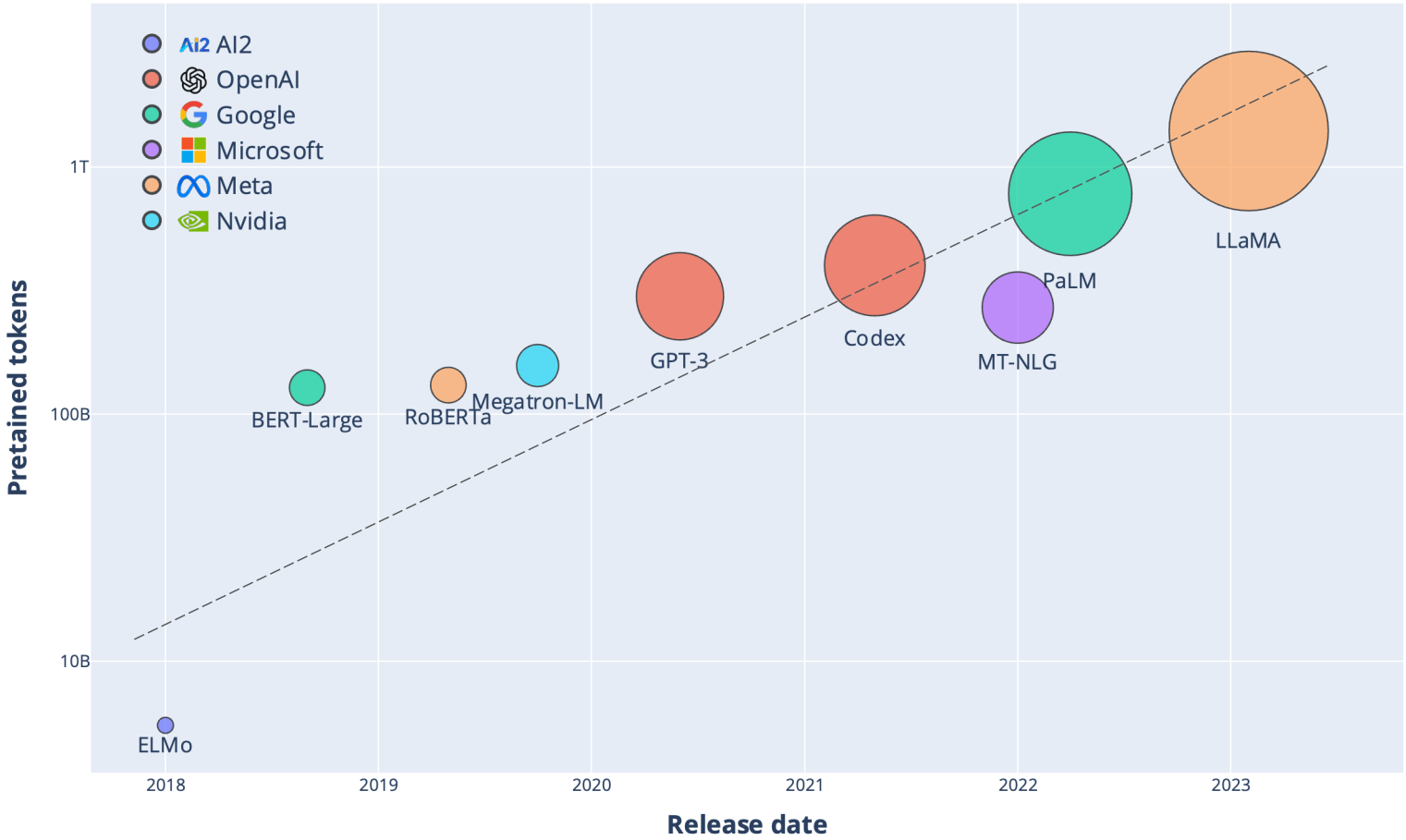


2006

Comparable datasets over a decade later

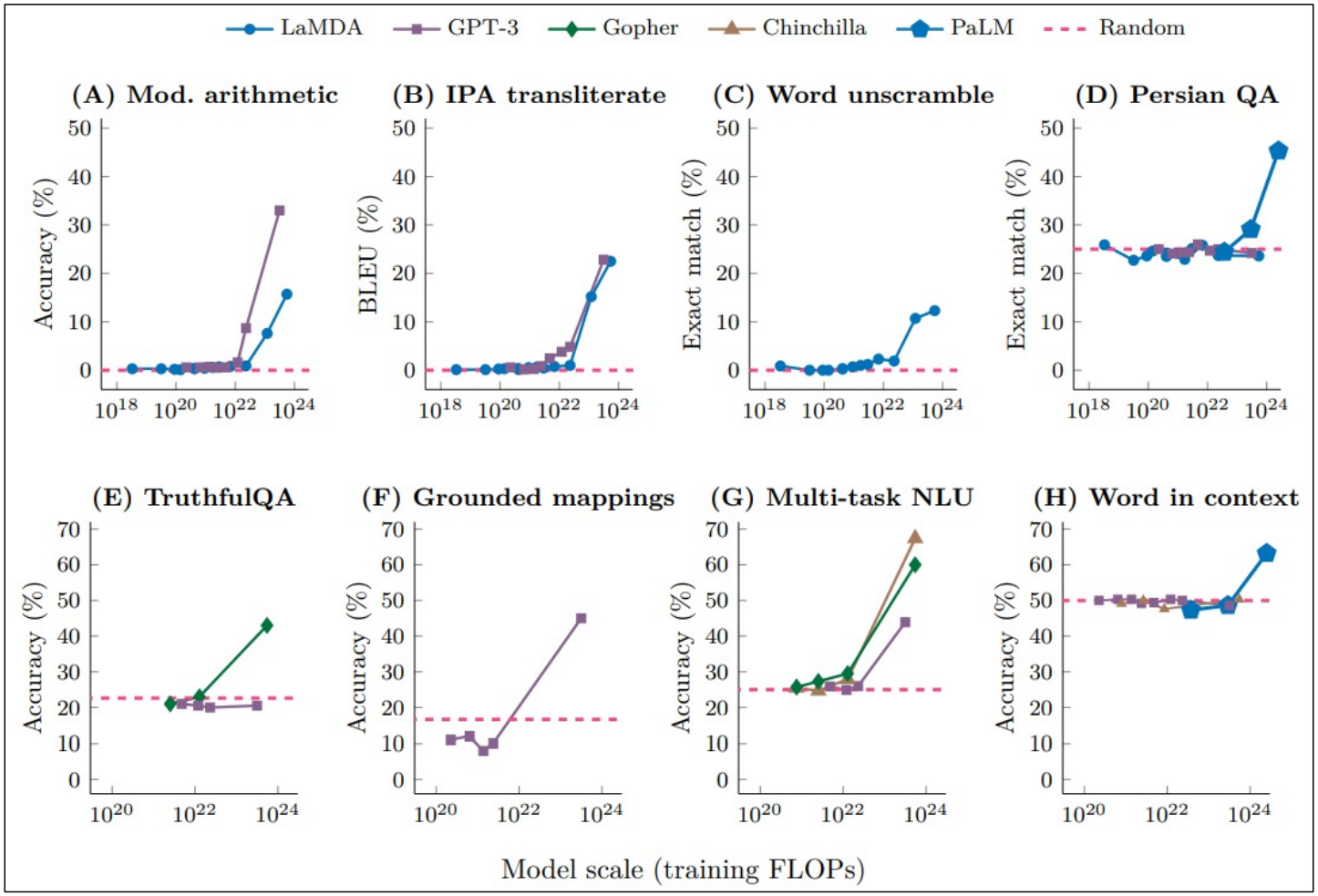
Growth of Data (5B → 1T)

GPT-4?



Growth of Model Size (100M → 1T+)

GPT-4?



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

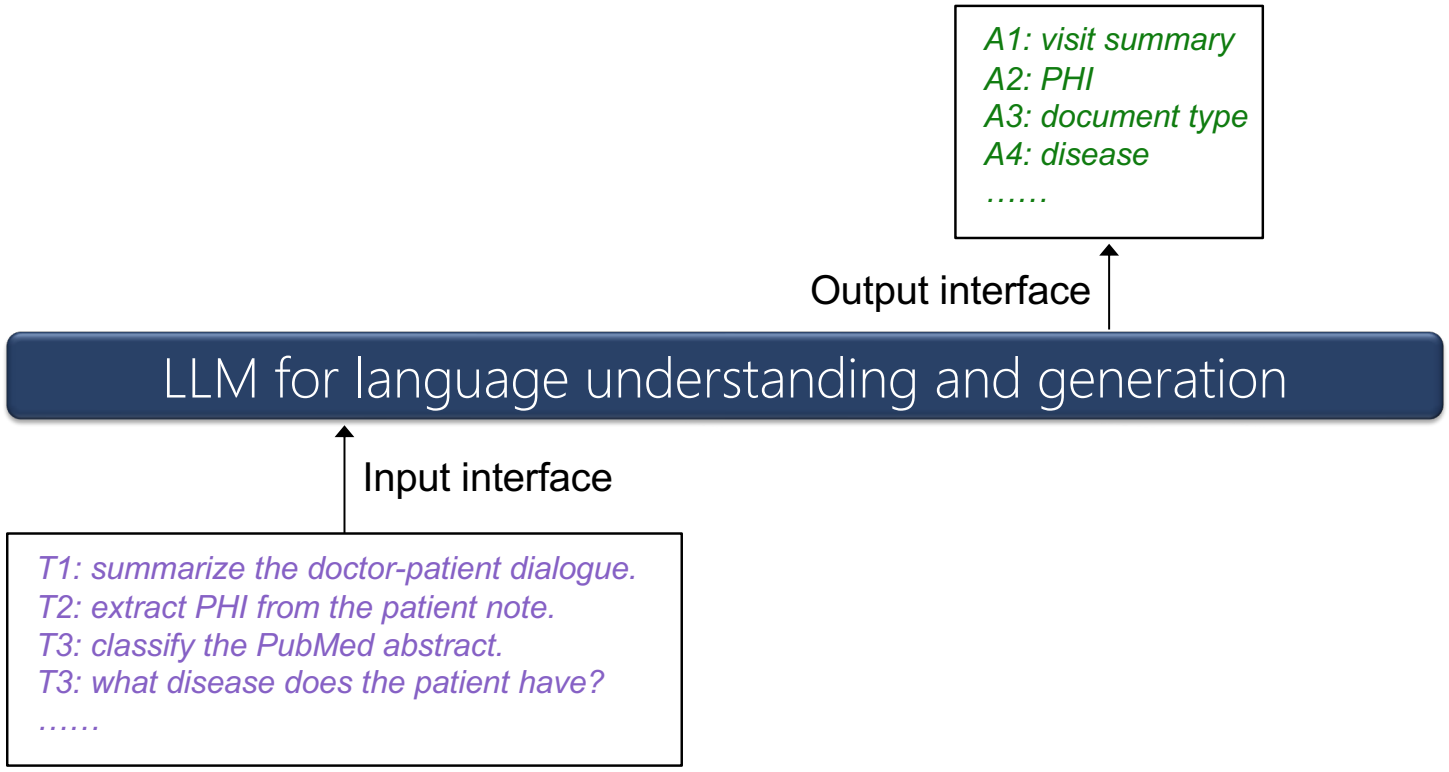
Effects of Scale

350M | 750M | 3B | 20B

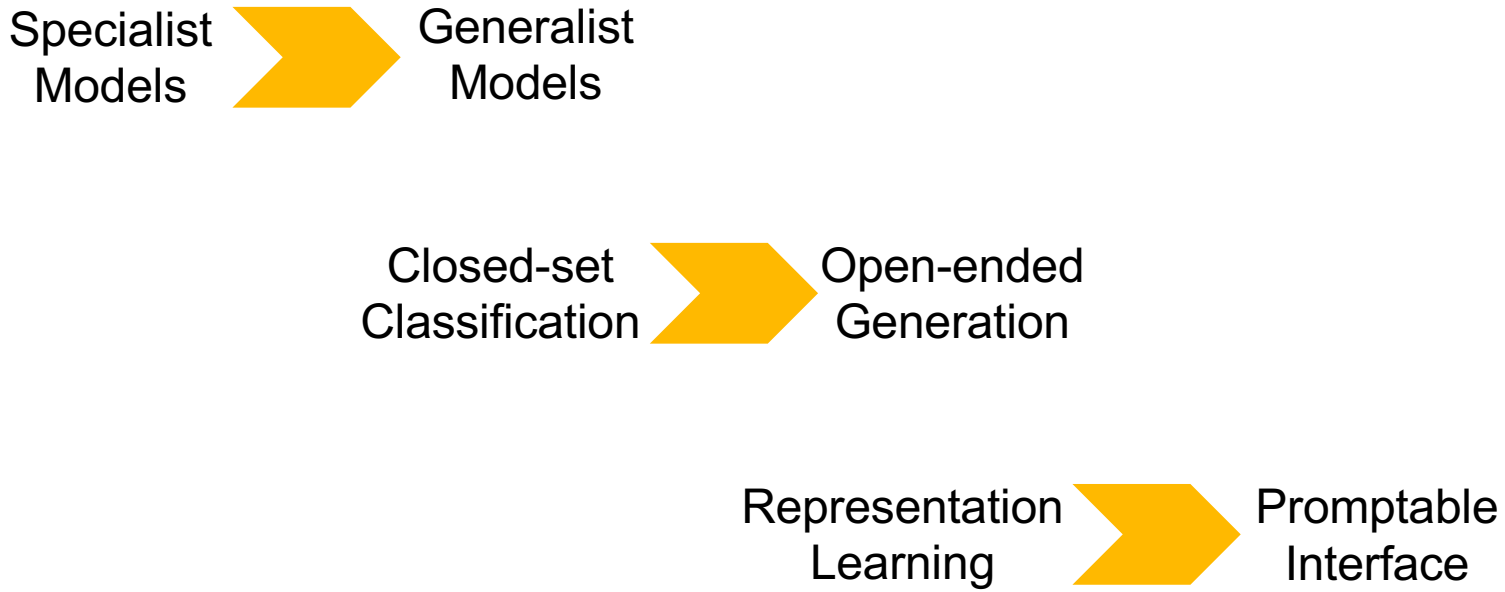
A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

https://parti.research.google/

General-purpose Interface



Paradigm Shifts with LLMs

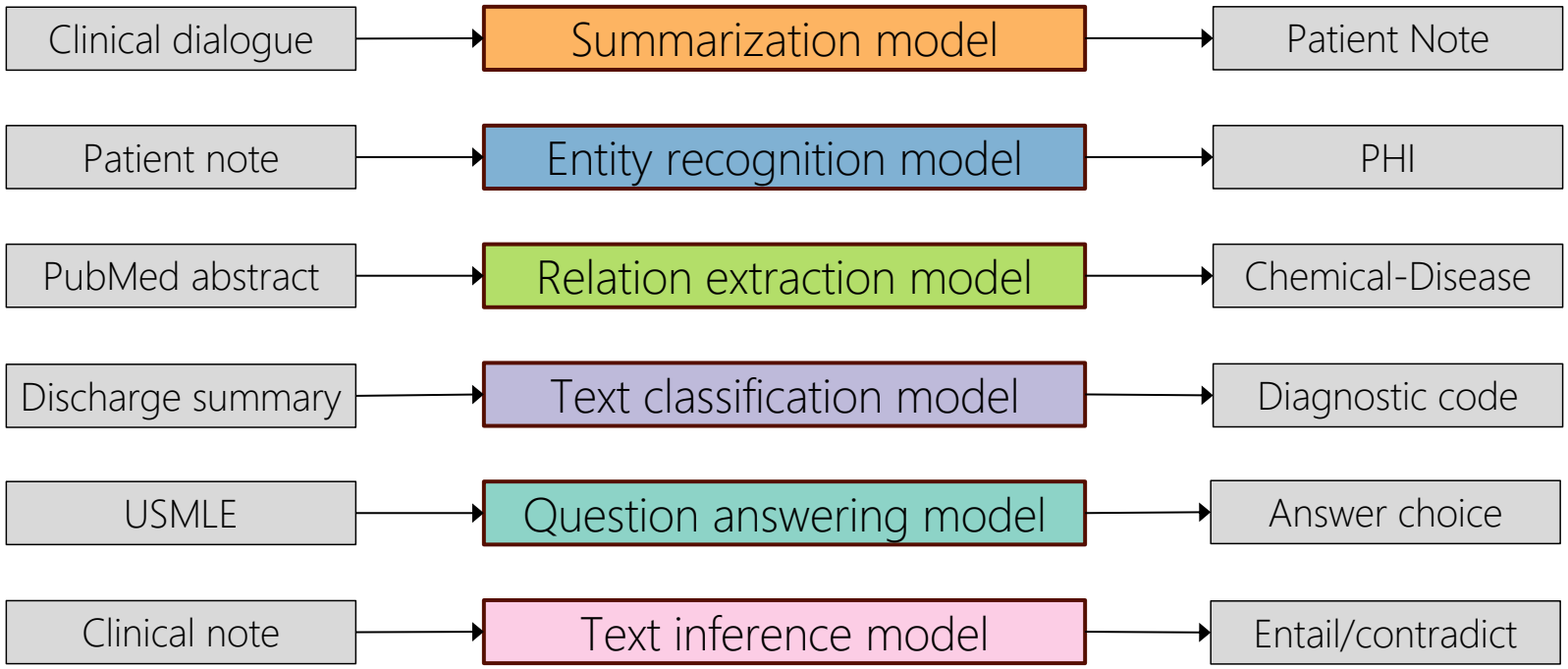


Paradigm Shifts with LLMs

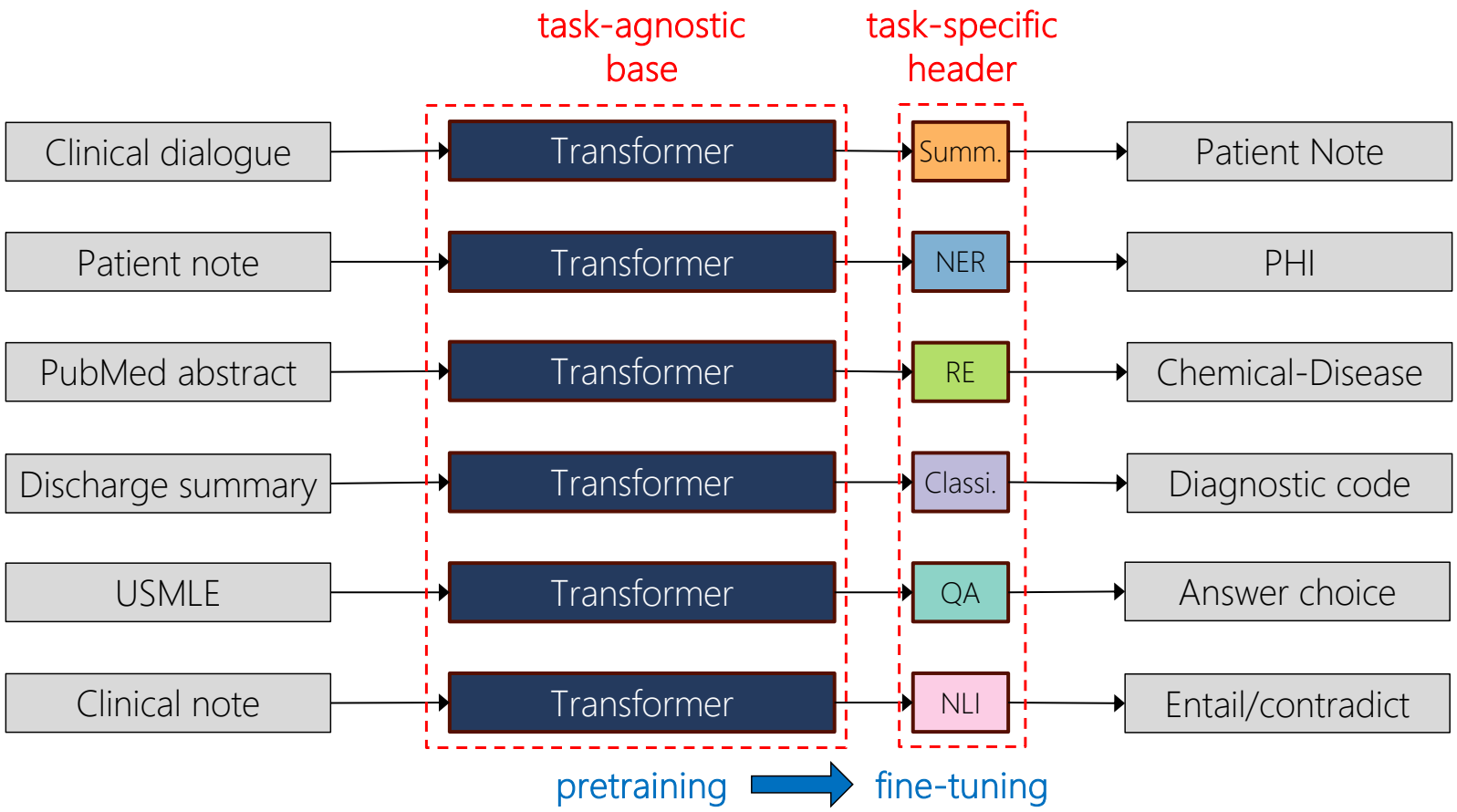
Closed-set
Classification
Open-ended
Generation

Representation
Learning
Promptable
Interface

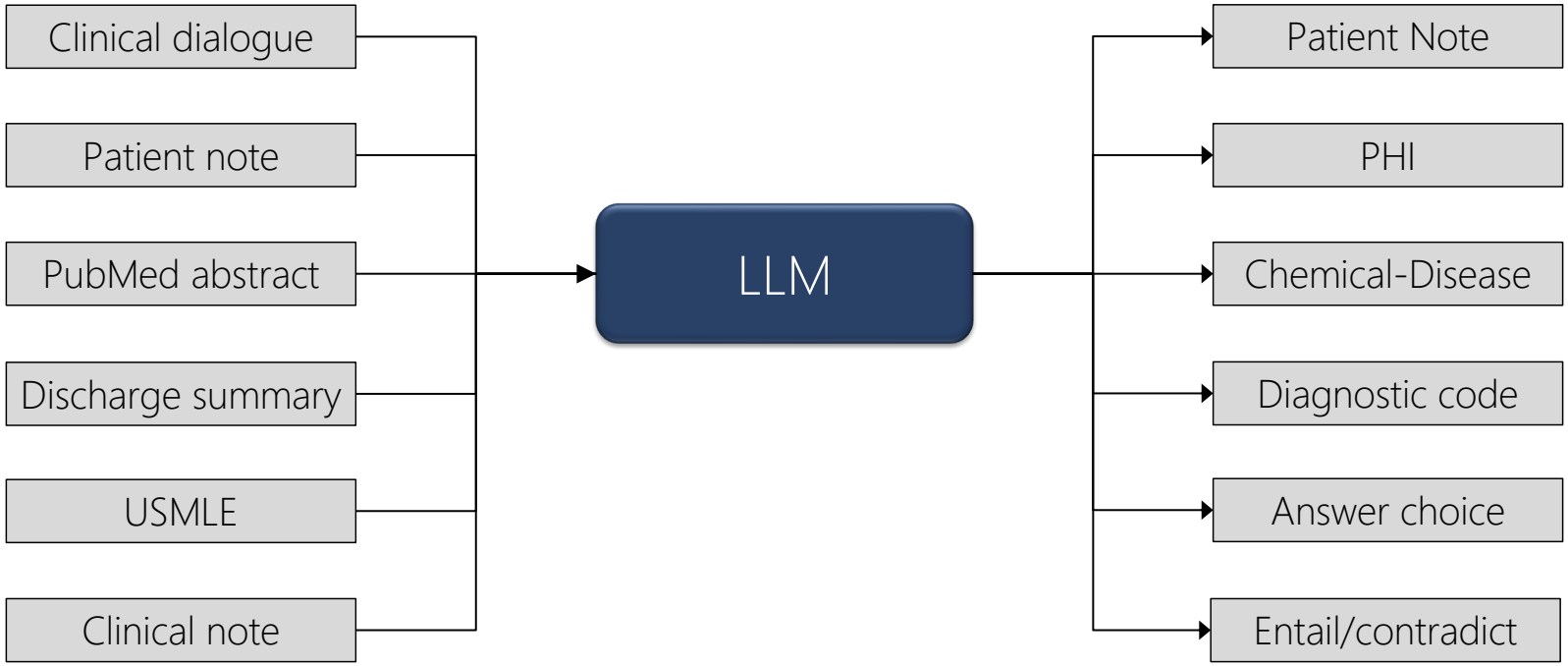
Specialist Models



Specialist Headers



Generalist Models

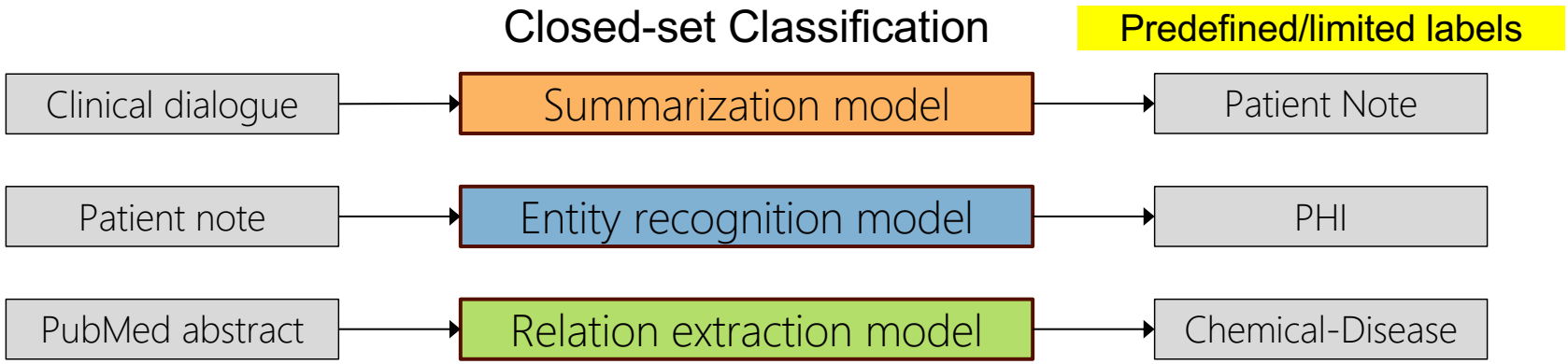


Specialist Models → Generalist Models

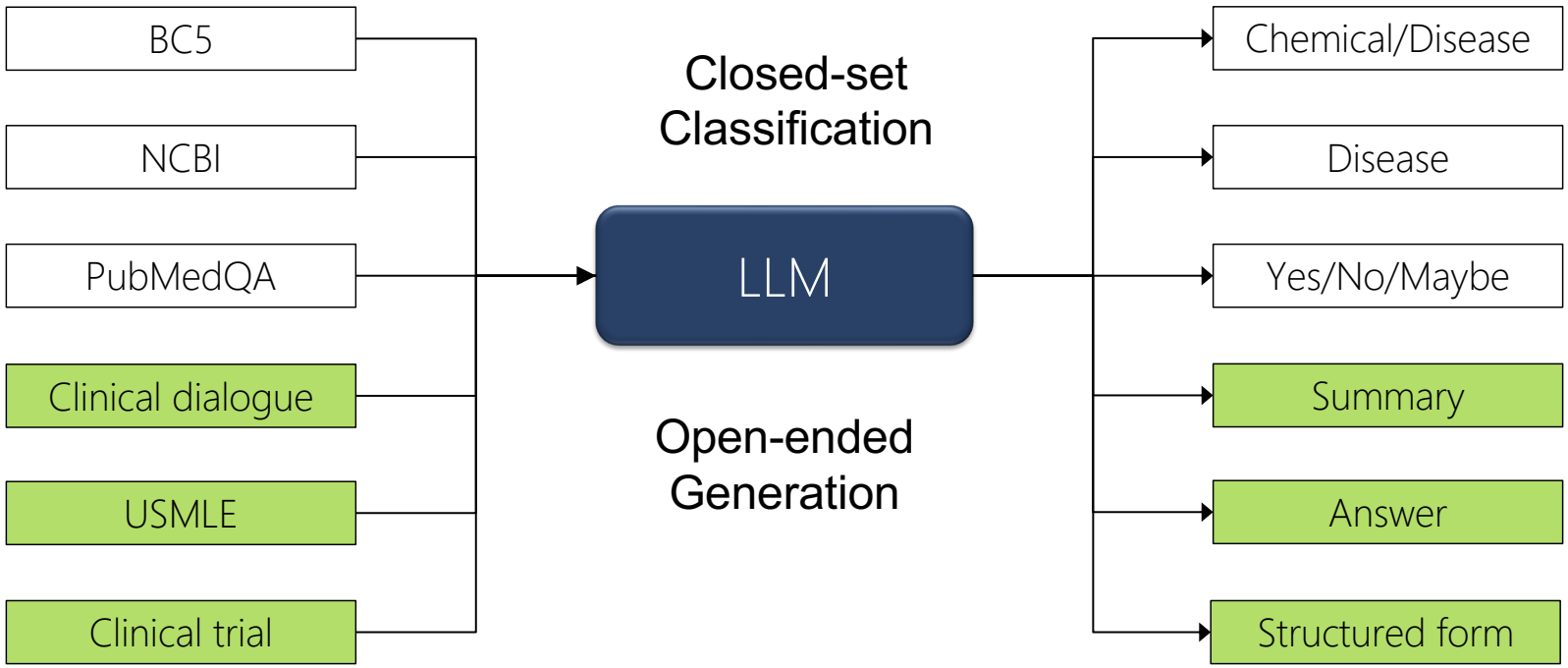
Closed-set Classification → Open-ended Generation

Representation Learning → Promptable Interface

Paradigm Shifts with LLMs



Paradigm Shifts with LLMs



Paradigm Shifts with LLMs

Specialist Models → Generalist Models

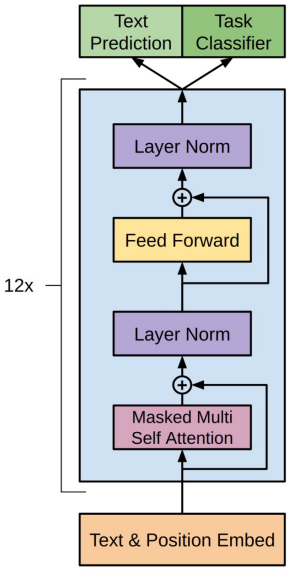
Closed-set Classification → Open-set Generation

Representation Learning → Promptable Interface

Paradigm Shifts with LLMs

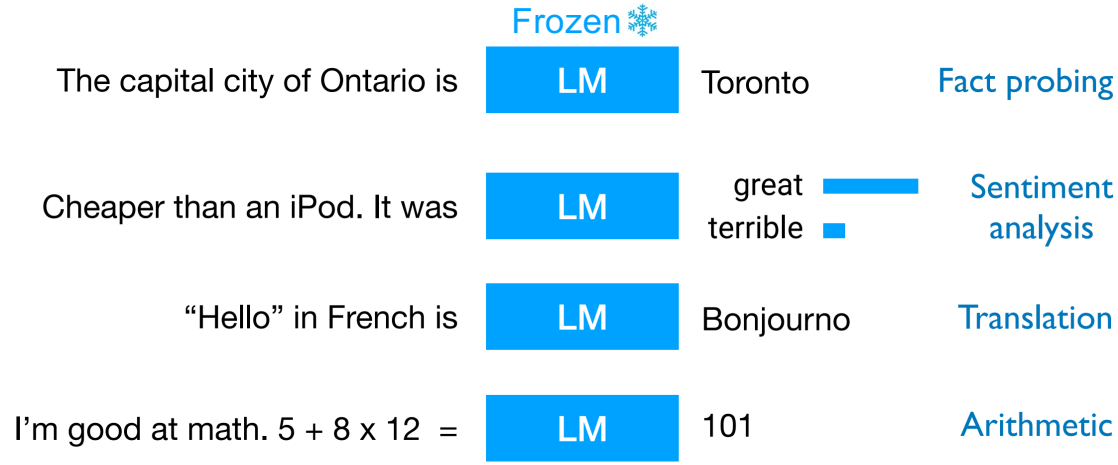
Representation learning

- Expensive
- Engineering heavy
- Task-specific



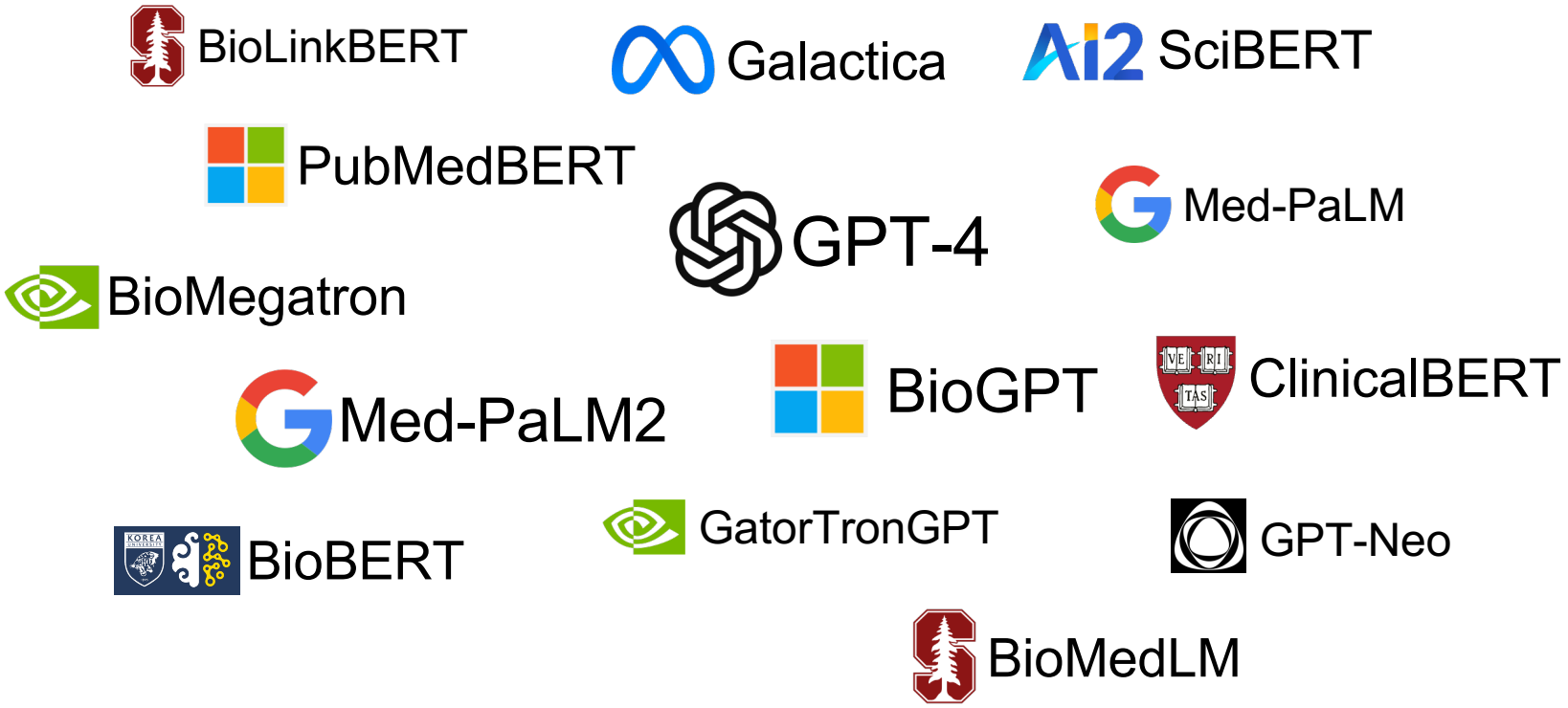
Promptable interface

- Training free
- Universal interface – natural language

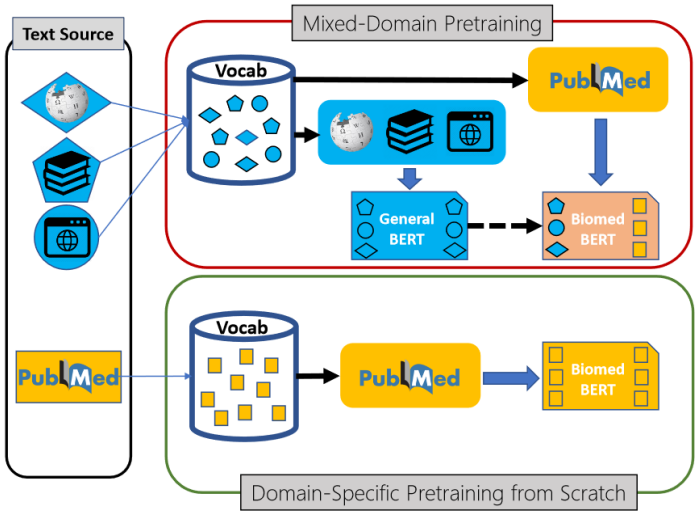


Improving Language Understanding by Generative Pre-Training
Retrieval-based Language Models and Applications

Biomedical LLMs



Domain-Specific Pretraining



Microsoft Research

PubMedBERT

Med-PaLM

DRAGON

Galactica

PubMedGPT

BioGPT

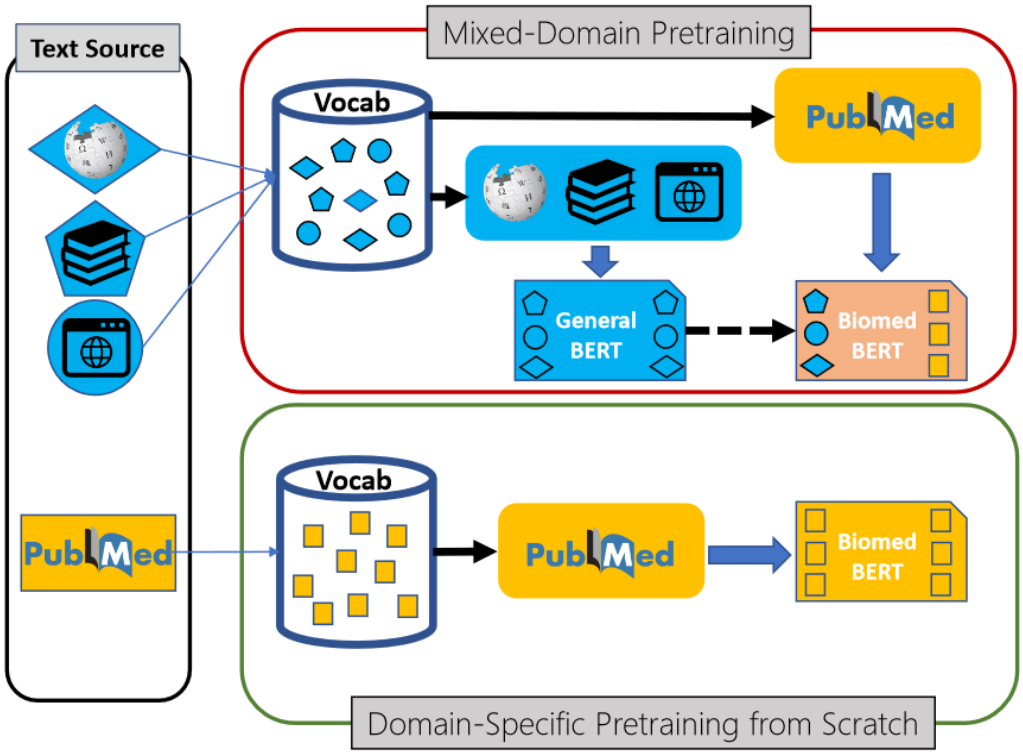
BioLinkBERT

.....

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.*

BioMedBERT (formerly 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	✓

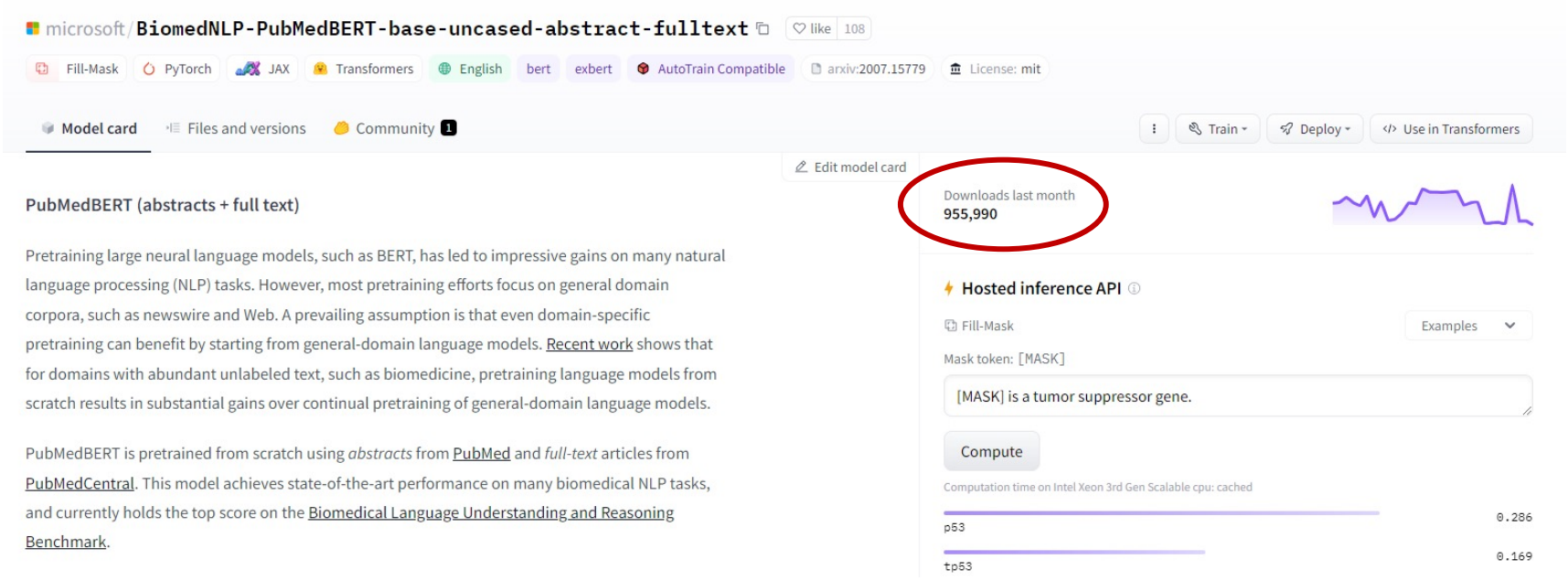
Yu, et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing", *Special Issue on Computational Methods for Biomedical Natural Language Processing, ACM Transactions on Computing for Health 2021.*

Domain-specific Vocab

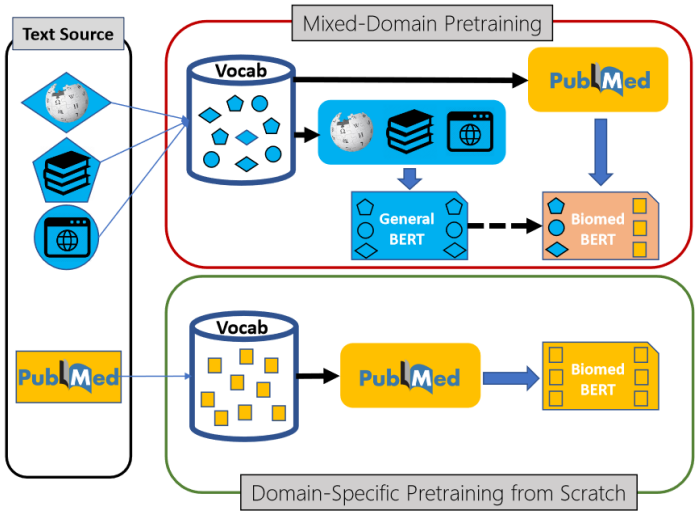
Preserves the integrity of

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

BioMedBERT (formerly PubMedBERT)



Domain-Specific Pretraining → Generalist Model



Microsoft Research

PubMedBERT

Med-PaLM

DRAGON

Galactica

PubMedGPT

BioGPT

BioLinkBERT

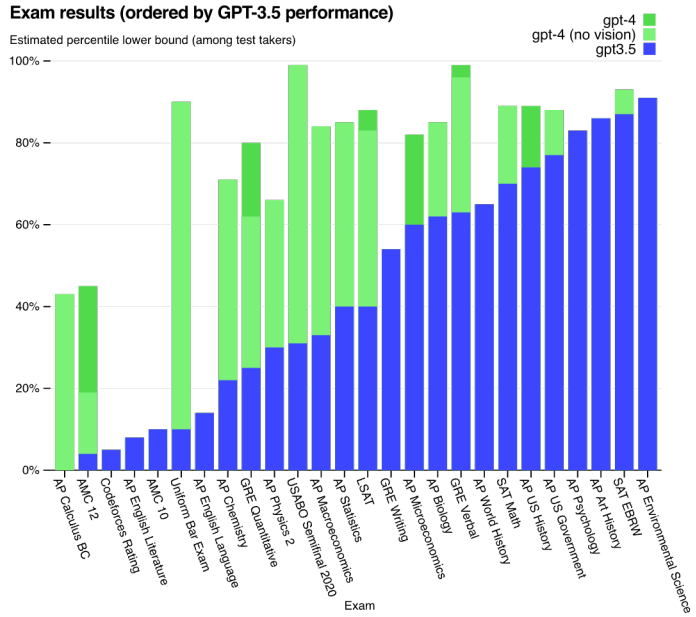
.....

GPT-4

2020 2022 2023

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

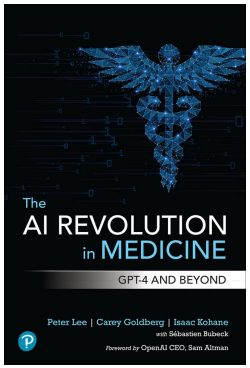
The most powerful general-purpose LLM
Human-level performance on many tasks



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



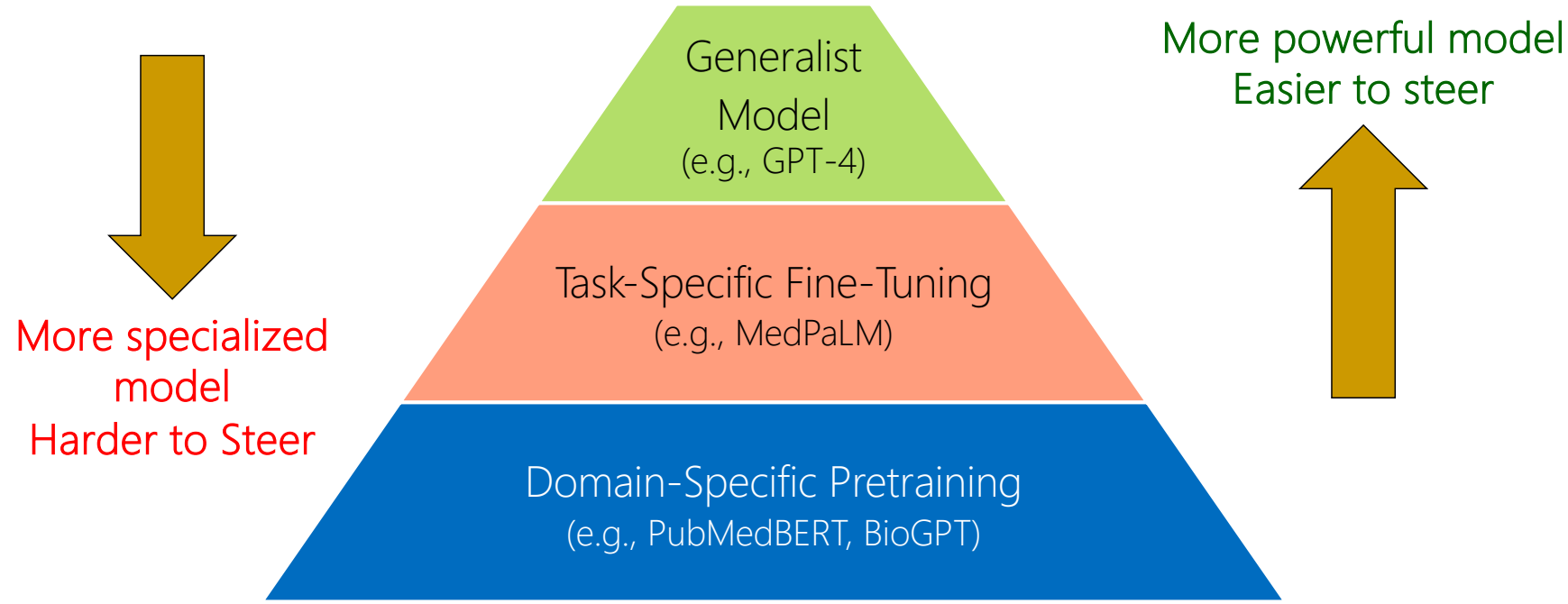
Capabilities of GPT-4 on Medical Challenge Problems
The AI Revolution in Medicine: GPT-4 and Beyond

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**

Generalist Models: Superior Steerability



Population-Level Health LLM

Patient → Serialized multimodal token sequence

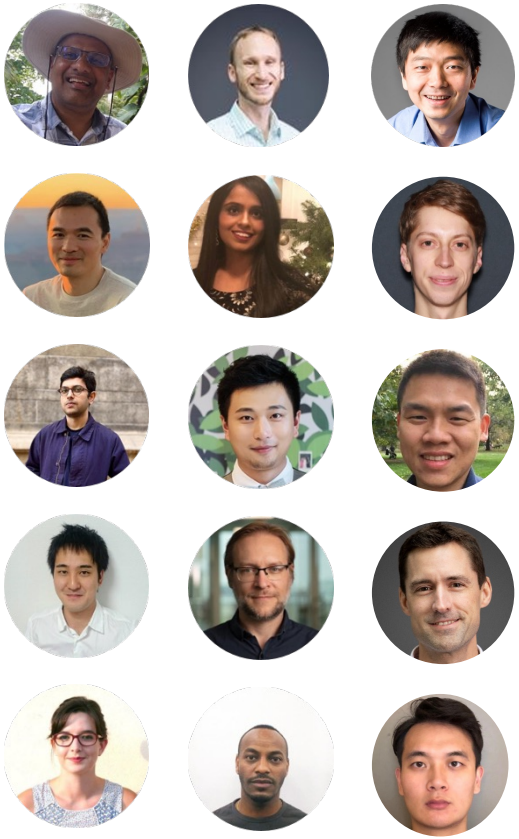
Initialize: GPT-101 (consumed entire public web)

Continued pretraining: 8 billion "health documents"

What is the multimodal health scaling law?

Will there be emergent health capabilities?

Real-World Evidence



- **JAX:** Susan Mockus, Sara Patterson
- **Fred Hutchinson**: Christopher Li, Kathi Malone
- **Providence:** Carlo Bifulco, Brian Piening
- **MSR:** Xiaodong Liu, Hao Cheng, Jianfeng Gao, Ozan Oktay, Javier Alvarez-Valle, Naveen Valluri
- **Interns:** Maxim Grechkin, Ankur Parikh, Victoria Lin, Sheng Wang, Stephen Mayhew, Daniel Fried, Violet Peng, Hai Wang, Robin Jia, Matthew McDermott, Alexis Ross, Zelalem Gero, Sarthak Jain, Jenny Chen, Hunter Lang, Benedikt Boecking, Varsha Kishore, Jinfeng Xiao, Wenxuan Zhou, Neha Hulkund, Michelle Li, Peniel Argaw, Hanwen Xu, Mars Huang, Juan Manuel Zambrano Chaves, Yiqing Xie, Isabel Chien, Alicia Curth