ISMB 2024 • JULY 12-16
ISCB MONTREAL CANADA

Introduction to LLMs with a Focus on Biomedical Data Science

Presenter: Shubo Tian

NIH National Library of Medicine
National Center for Biotechnology Information

Learning Objective

- Understanding what is language model, type of language models, and existing major language models and their applications in general and in biomedicine.

What is Language Model

Given a sequence of words/tokens:

$$w_1, w_2, \dots, w_n$$

a model that computes either of the following probabilities:

$P(w_1, w_2, \dots, w_n)$ or
$P(w_n | w_1, w_2, \dots, w_{n-1})$

is called a language model.

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \dots, w_{i-1})$$

***P*(“P53 is a tumor suppressor gene.”)**

= *P*(“P53”)•

P(“is” |“P53”) •

P(“a”|“P53 is”)•

P(“tumor”|“P53 is a”)•

P(“suppressor”|“P53 is a tumor”)•

P(“gene”|“P53 is a tumor suppressor”)•

P(“.”|“P53 is a tumor suppressor gene”)

Jurafsky, D. and Martin, J.H. (2023) Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3

Language Modeling

Word count modeling (n-gram)

$$P(w_n|w_1, w_2, \dots, w_{n-1}) = \frac{\text{count}(w_1, w_2, \dots, w_n)}{\text{count}(w_1, w_2, \dots, w_{n-1})}$$

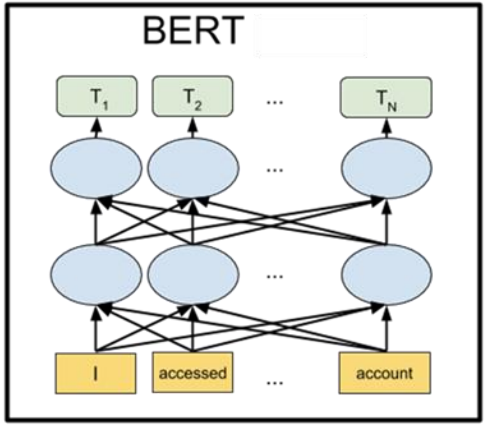
Neural network modeling

$$\begin{aligned} P(w_n|w_1, w_2, \dots, w_{n-1}) &= P(w_n|h) \\ &= \frac{\exp(h^T \text{emb}(w_n))}{\sum_{w' \in \text{Vocab}} \exp(h^T \text{emb}(w'))} \end{aligned}$$

$$h = \text{Encoder}(w_1, w_2, \dots, w_{n-1})$$

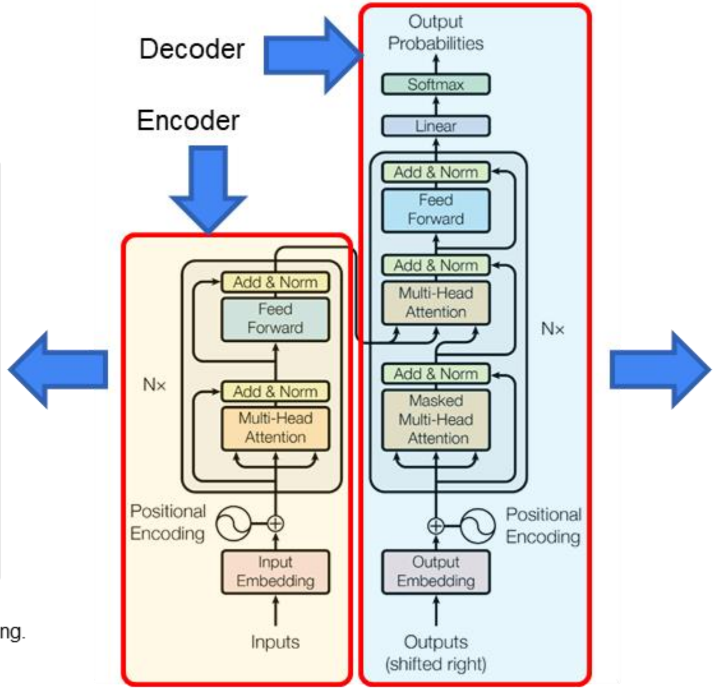
Jurafsky, D. and Martin, J.H. (2023) Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3

Transformer and LLMs

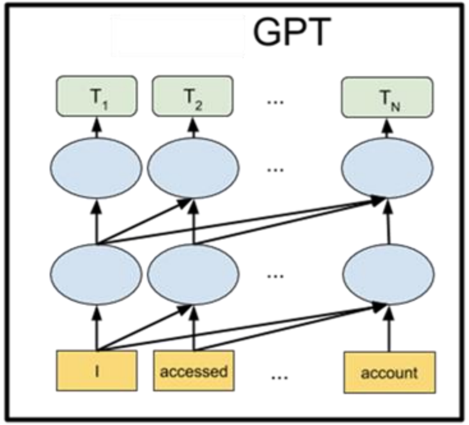


Devlin, J. *et al.* (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs].*

https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html



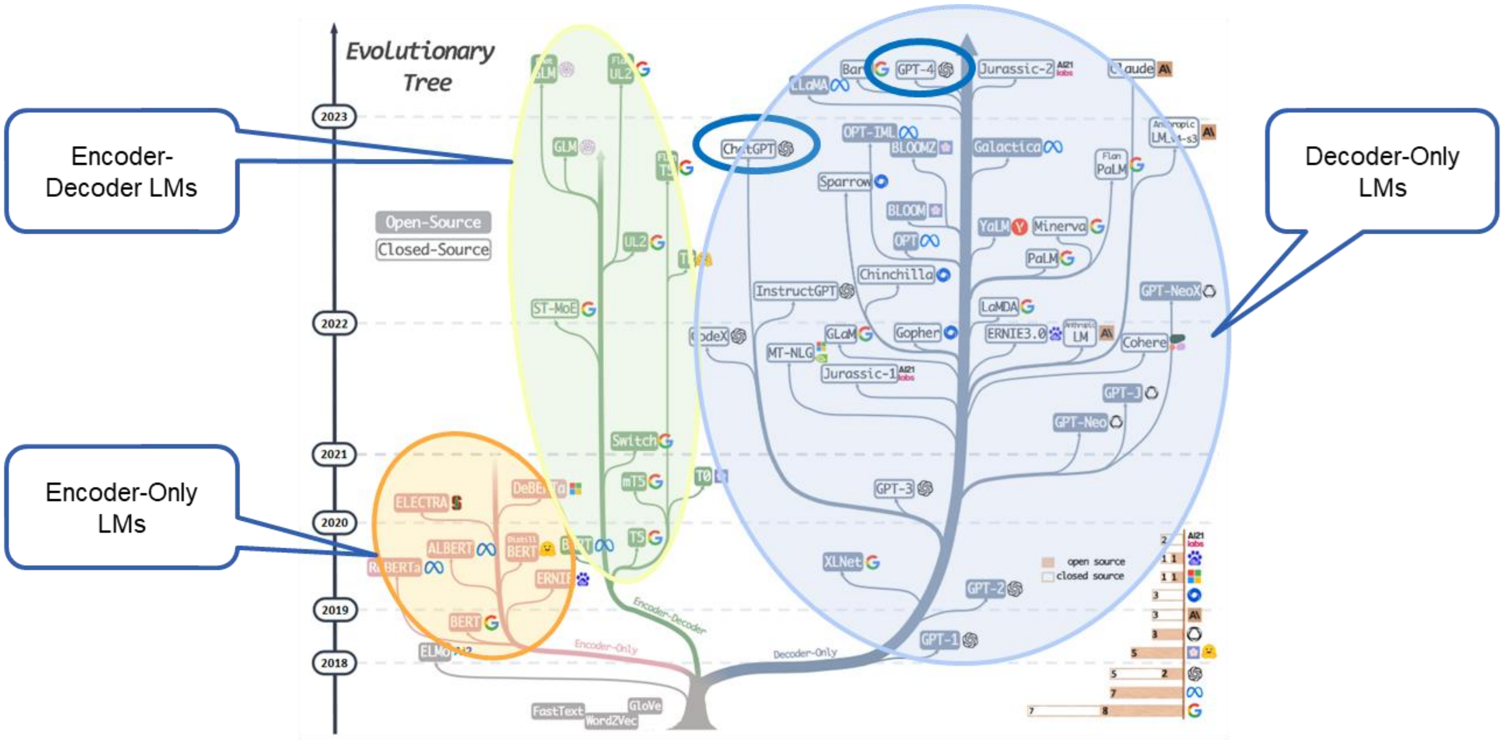
Vaswani, A. *et al.* (2017) Attention Is All You Need. *arXiv:1706.03762 [cs].*



Radford, A. *et al.* Improving Language Understanding by Generative Pre-Training.

https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Type of LLMs



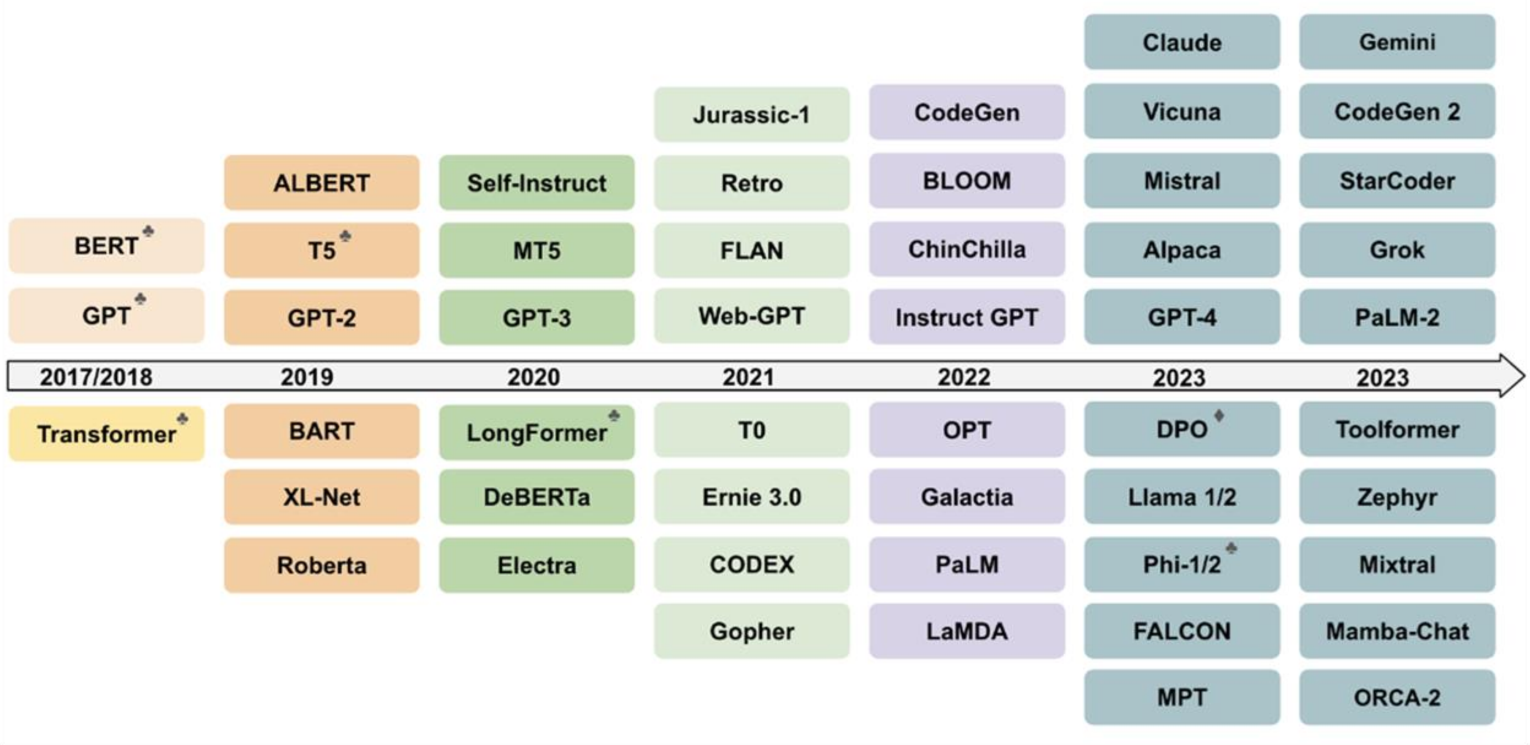
Encoder-Decoder LMs

Decoder-Only LMs

Encoder-Only LMs

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B. and Hu, X. (2023) Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. 10.48550/arXiv.2304.13712.

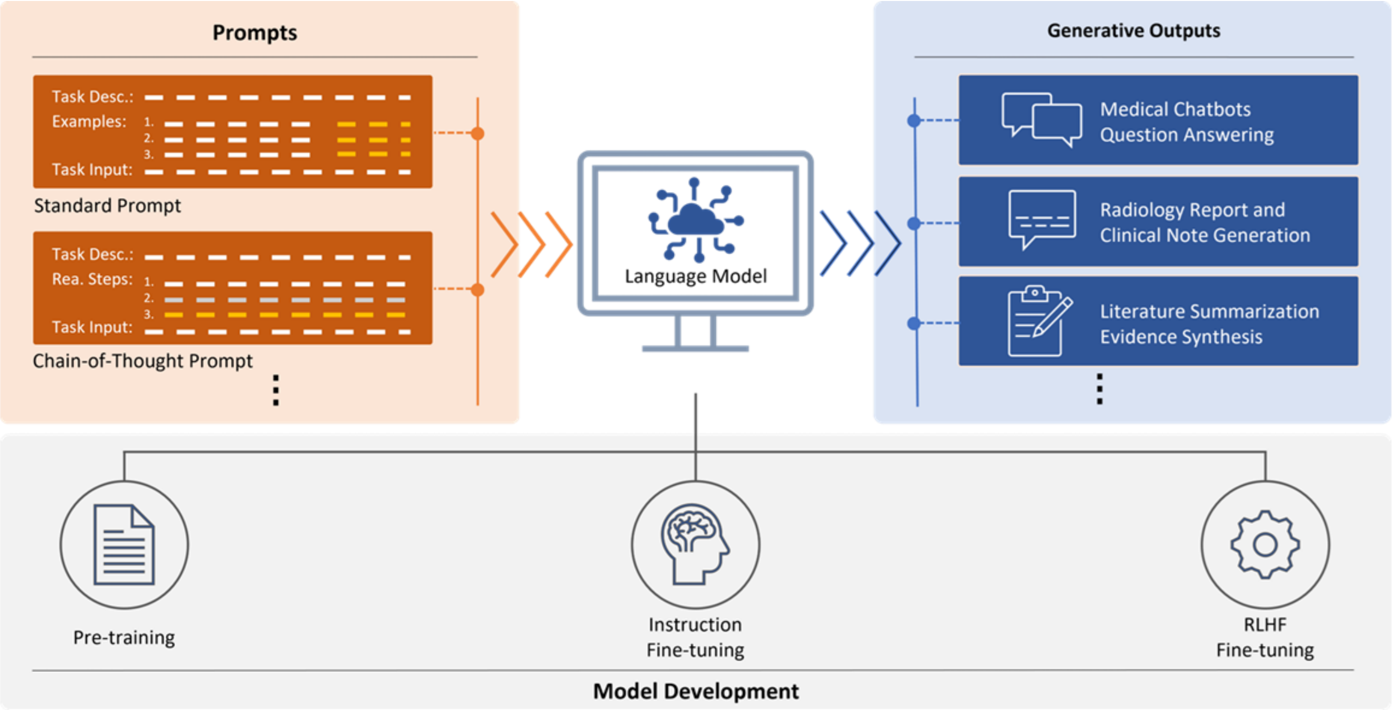
Timeline of Representative LLMs



					Claude	Gemini
			Jurassic-1	CodeGen	Vicuna	CodeGen 2
	ALBERT	Self-Instruct	Retro	BLOOM	Mistral	StarCoder
BERT ♣	T5 ♣	MT5	FLAN	ChinChilla	Alpaca	Grok
GPT ♣	GPT-2	GPT-3	Web-GPT	Instruct GPT	GPT-4	PaLM-2
2017/2018	2019	2020	2021	2022	2023	2023
Transformer ♣	BART	LongFormer ♣	T0	OPT	DPO ♦	Toolformer
	XL-Net	DeBERTa	Ernie 3.0	Galactia	Llama 1/2	Zephyr
	Roberta	Electra	CODEX	PaLM	Phi-1/2 ♣	Mixtral
			Gopher	LaMDA	FALCON	Mamba-Chat
					MPT	ORCA-2

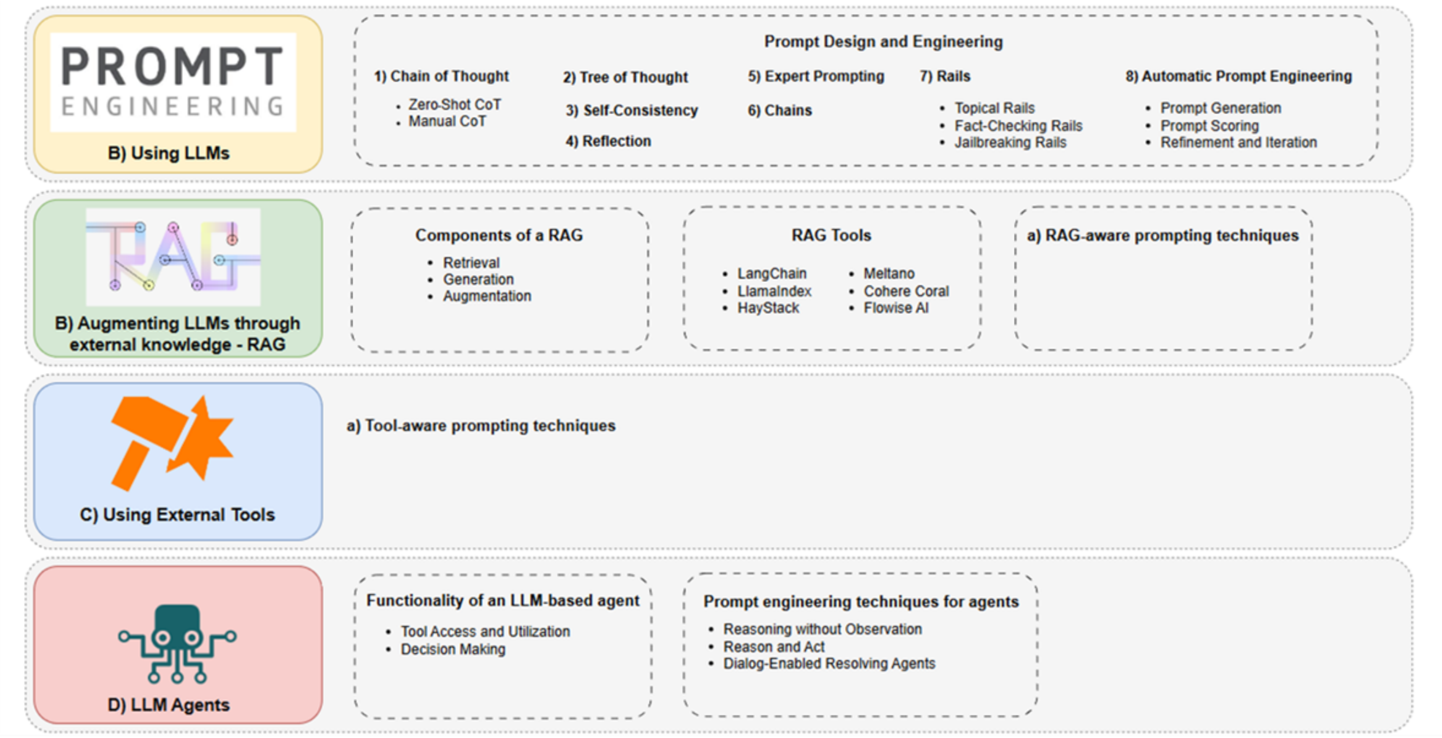
Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X. and Gao, J. (2024) Large Language Models: A Survey.

The Paradigm of LLMs



Tian, S., Jin, Q., Yeganova, L., Lai, P.-T., Zhu, Q., Chen, X., Yang, Y., Chen, Q., Kim, W., Comeau, D.C., et al. (2024) Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics, 25, bbad493.

How LLMs are Used



PROMPT ENGINEERING

B) Using LLMs

Prompt Design and Engineering

1) Chain of Thought
- Zero-Shot CoT
- Manual CoT

2) Tree of Thought

3) Self-Consistency

4) Reflection

5) Expert Prompting

6) Chains

7) Rails
- Topical Rails
- Fact-Checking Rails
- Jailbreaking Rails

8) Automatic Prompt Engineering
- Prompt Generation
- Prompt Scoring
- Refinement and Iteration

RAG

B) Augmenting LLMs through external knowledge - RAG

Components of a RAG
- Retrieval
- Generation
- Augmentation

RAG Tools
- LangChain
- LlamaIndex
- HayStack
- Meltano
- Cohere Coral
- Flowise AI

a) RAG-aware prompting techniques

C) Using External Tools

a) Tool-aware prompting techniques

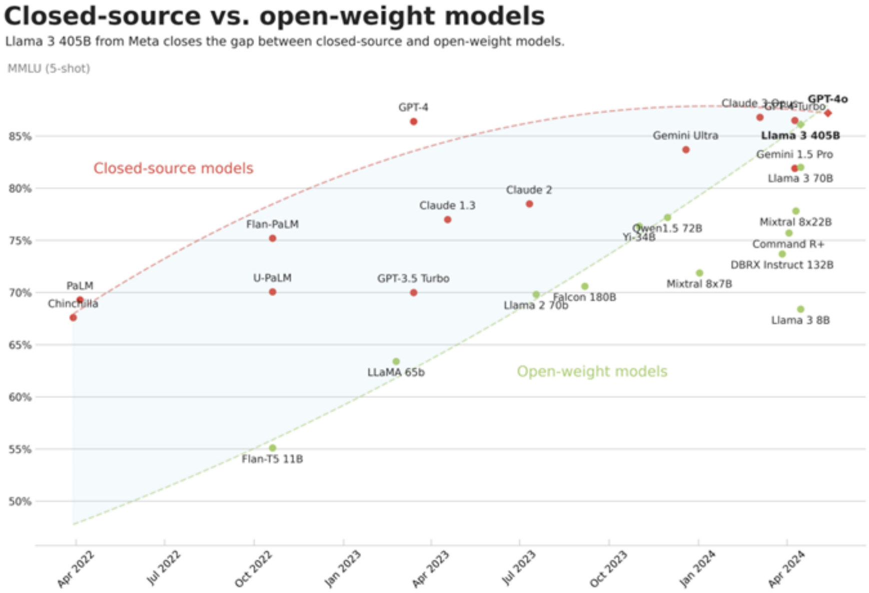
D) LLM Agents

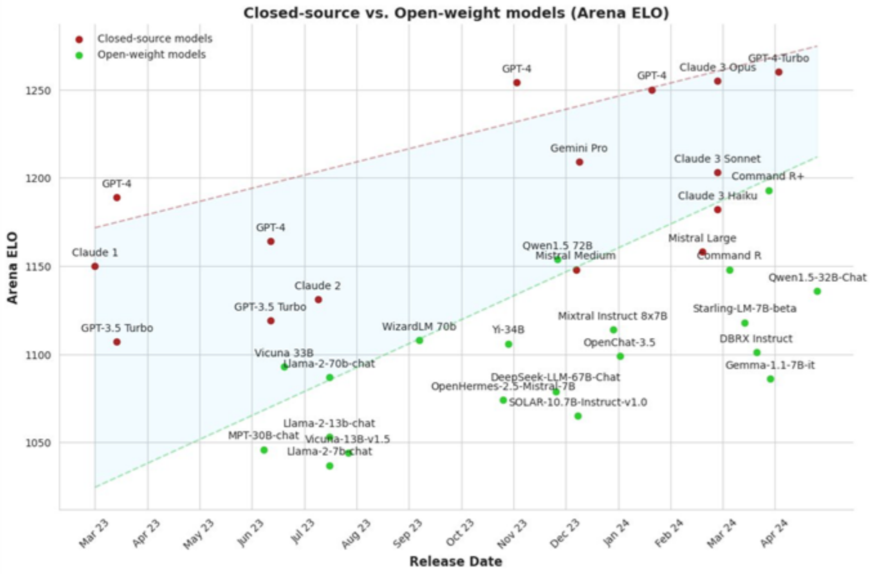
Functionality of an LLM-based agent
- Tool Access and Utilization
- Decision Making

Prompt engineering techniques for agents
- Reasoning without Observation
- Reason and Act
- Dialog-Enabled Resolving Agents

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X. and Gao, J. (2024) Large Language Models: A Survey.

Major LLMs and Their Performance

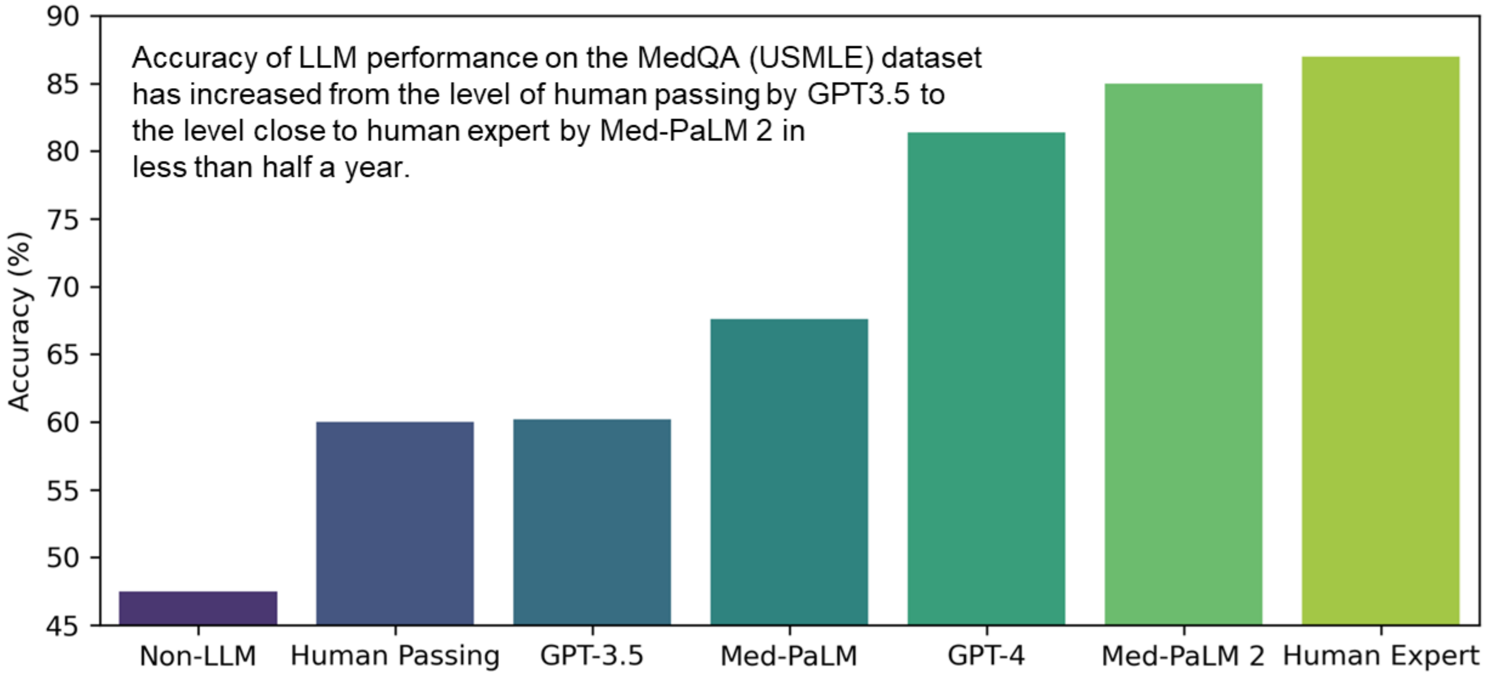




https://twitter.com/maximelabonne/status/1793947985430057288
https://crfm.stanford.edu/helm/mmlu/latest/#/leaderboard

https://twitter.com/maximelabonne/status/1779801605702836454
https://chat.lmsys.org/?leaderboard

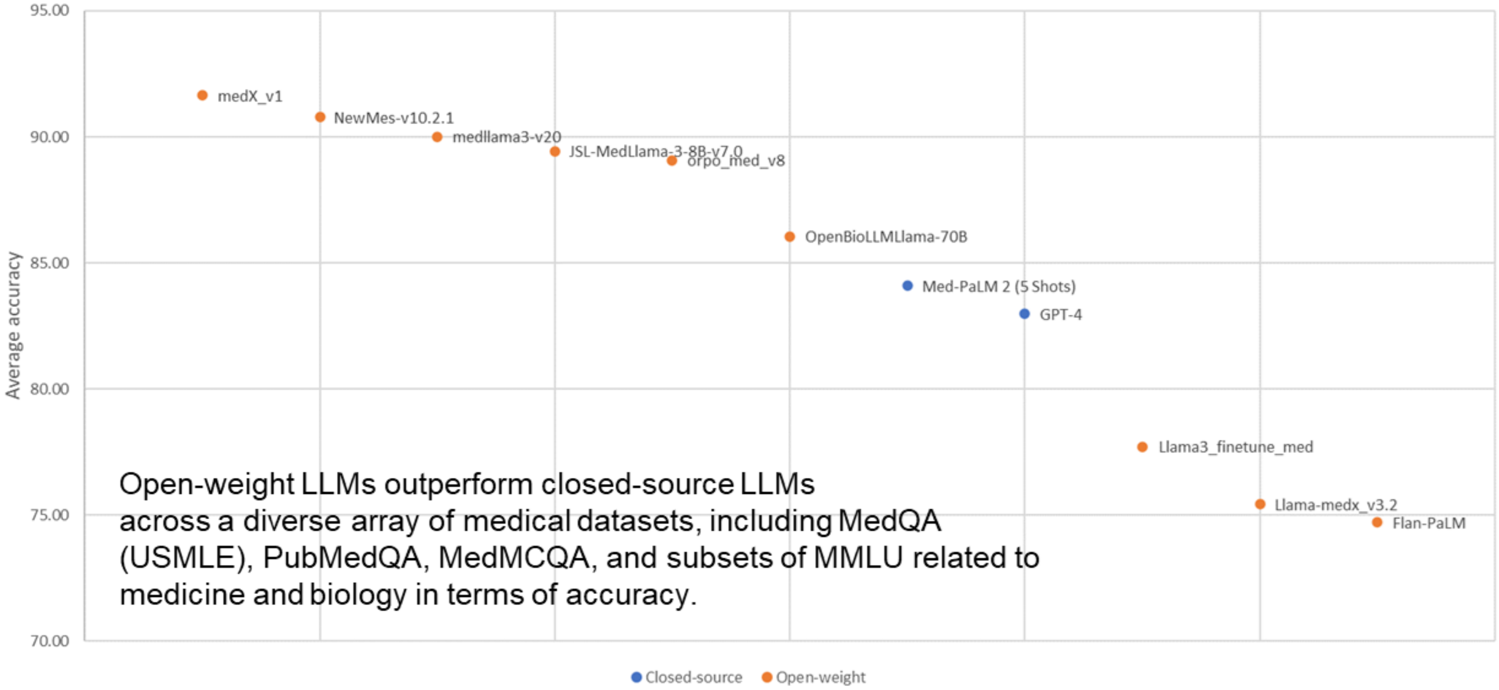
LLMs in Biomedicine



Accuracy of LLM performance on the MedQA (USMLE) dataset has increased from the level of human passing by GPT3.5 to the level close to human expert by Med-PaLM 2 in less than half a year.

Tian, S., Jin, Q., Yeganova, L., Lai, P.-T., Zhu, Q., Chen, X., Yang, Y., Chen, Q., Kim, W., Comeau, D.C., et al. (2024) Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics, 25, bbad493.

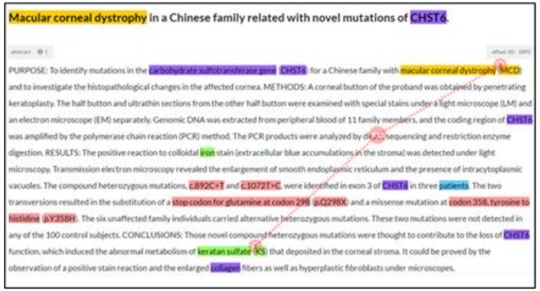
Major LLMs in Biomedicine and Their Performance



Open-weight LLMs outperform closed-source LLMs across a diverse array of medical datasets, including MedQA (USMLE), PubMedQA, MedMCQA, and subsets of MMLU related to medicine and biology in terms of accuracy.

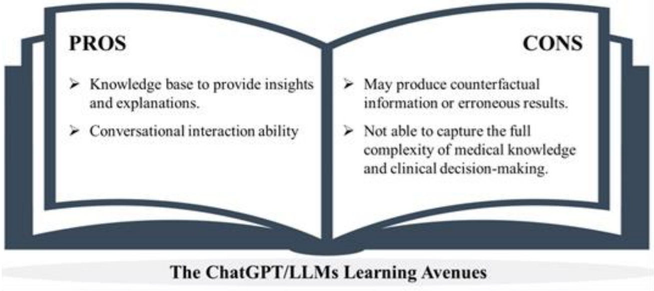
https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard

Applications of LLMs in Biomedicine



Macular corneal dystrophy in a Chinese family related with novel mutations of CHST6.

PURPOSE: To identify mutations in the carbohydrate sulfotransferase gene (CHST6) for a Chinese family with macular corneal dystrophy (MCD) and to investigate the histopathological changes in the affected cornea. METHODS: A corneal button of the proband was obtained by penetrating keratoplasty. The half button and ultrathin sections from the other half button were examined with special stains under a light microscope (LM) and an electron microscope (EM) separately. Genomic DNA was extracted from peripheral blood of 11 family members, and the coding region of CHST6 was amplified by the polymerase chain reaction (PCR) method. The PCR products were analyzed by direct sequencing and restriction enzyme digestion. RESULTS: The positive reaction to colloidal iron stain (extracellular blue accumulations in the stroma) was detected under light microscopy. Transmission electron microscopy revealed the enlargement of smooth endoplasmic reticulum and the presence of intracytoplasmic vacuoles. The compound heterozygous mutations, c.892C>T and c.1072T>C, were identified in exon 3 of CHST6 in three patients. The two transversions resulted in the substitution of a stop codon for glutamine at codon 298 (p.Q298X) and a missense mutation at codon 358, tyrosine to histidine (p.Y358H). The six unaffected family individuals carried alternative heterozygous mutations. These two mutations were not detected in any of the 100 control subjects. CONCLUSIONS: Those novel compound heterozygous mutations were thought to contribute to the loss of CHST6 function, which induced the abnormal metabolism of keratan sulfate (KS) that deposited in the corneal stroma. It could be proved by the observation of a positive stain reaction and the enlarged collagen fibers as well as hyperplastic fibroblasts under microscopes.



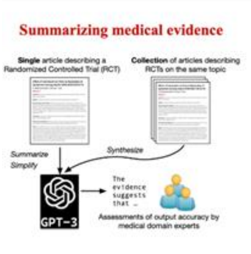
PROS

- Knowledge base to provide insights and explanations.
- Conversational interaction ability

CONS

- May produce counterfactual information or erroneous results.
- Not able to capture the full complexity of medical knowledge and clinical decision-making.

The ChatGPT/LLMs Learning Avenues



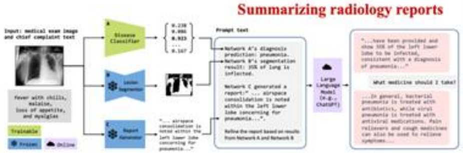
Summarizing medical evidence

Single article describing a Randomized Controlled Trial (RCT)
Collection of articles describing RCTs on the same topic
Summarize
Simplify
Synthesize
GPT-3
The evidence suggests that …
Assessments of output accuracy by medical domain experts

ChatGPT: the future of discharge summaries?

Sajan B Patel • Kyle Lam

Open Access • Published: February 06, 2023 • DOI: https://doi.org/10.1016/S2589-7500(23)00021-3 •



Summarizing radiology reports

Input: medical exam image and chief complaint text
Fever with chills, malaise, loss of appetite, and myalgias
Trainable
Frozen
Online
A
Disease Classifier
B
Lesion Segmentor
C
Report Generator
0.238
0.086
0.923
...
0.167
"... airspace consolidation is noted within the left lower lobe concerning for pneumonia..."
Prompt text
Large Language Model (e.g., ChatGPT)
What medicine should I take?

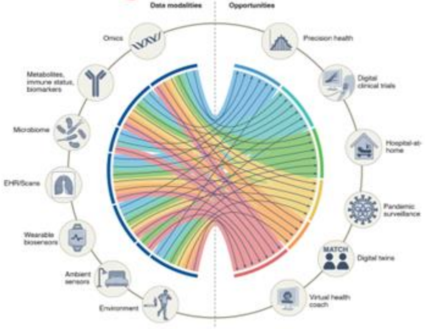
Network A's diagnosis prediction: pneumonia.
Network B's segmentation result: 35% of lung is infected.

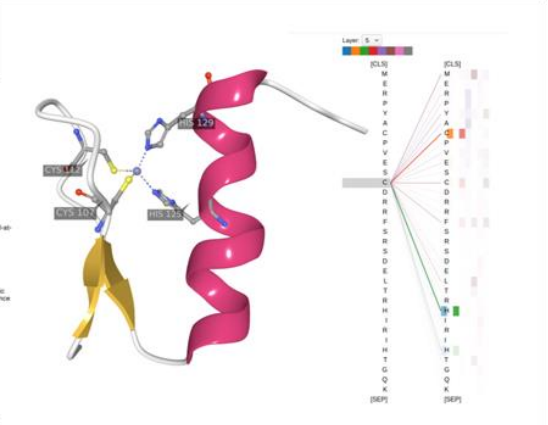
Network C generated a report:" ... airspace consolidation is noted within the left lower lobe concerning for pneumonia...".

Refine the report based on results from Network A and Network B

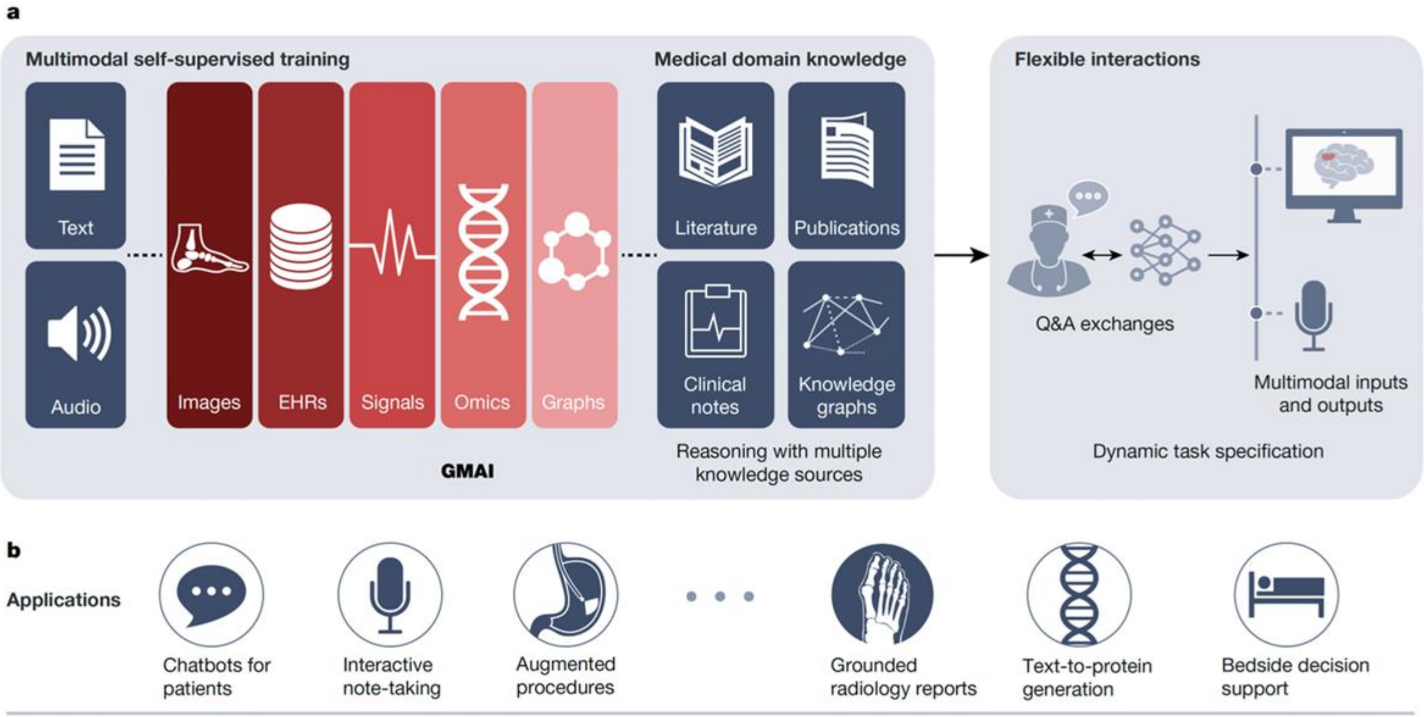
"...have been provided and show 35% of the left lower lobe to be infected, consistent with a diagnosis of pneumonia..."

...In general, bacterial pneumonia is treated with antibiotics, while viral pneumonia is treated with antiviral medications. Pain relievers and cough medicines can also be used to relieve symptoms...





LLMs in the Context of AI in Biomedicine



Moor,M. et al. (2023) Foundation models for generalist medical artificial intelligence. Nature, 616, 259–265.

Thank you!

Next talk in line: How to Use GPT-3.5 and GPT-4 with Python, Qiao Jin, M.D.