

# **Clinical NLP in the context of Large Language Models**

**a brief overview, sharing some clinical perspectives, related work, and areas of possible research**

**Sam Rawal**  
[scrawal2@illinois.edu](mailto:scrawal2@illinois.edu)

# Outline

- brief introduction to LLMs, including GPT-3/4 and ChatGPT
- related work
- some personal experiments with GPT applied to clinical domain + comments from physicians

# Motivation



Anyone else witnessing a panic x +

news.ycombinator.com/item?id=35180392&p=2

Hacker News

new | threads | past | comments | ask | show | jobs | submit | page 2

zora\_goron (259) | logout

▲ Anyone else witnessing a panic inside NLP orgs of big tech companies? (reddit.com)

398 points by georgehill 22 days ago | hide | past | favorite | 509 comments

▲ dserban 22 days ago | next [-]

The PR folks at my current company are in full panic mode on LinkedIn, judging from the passive-aggressive tone of their posts (sometimes very nearly begging customers not to use ChatGPT and friends).

They fully understand that LLMs are stealing lunch money from established information retrieval industry players selling overpriced search algorithms. For example, they are being protected by insurance companies and folks going through the publically on social media

AndriyMulyar on Twitter: "Serious question: What does an NLP Ph.D student work on nowadays with the presence of closed source GPT models that beat anything you can do in standard academic lab?"

@sleepinyourhat @srush\_nlp @chrmanning @mdredze @ChrisGPotts

5:56 PM · Mar 15, 2023 · 929.6K Views

209 Retweets 48 Quotes 1,468 Likes 600 Bookmarks

# Brief intro to LLMs (GPT, ChatGPT)

# General Purpose Transformer (GPT): Architecture

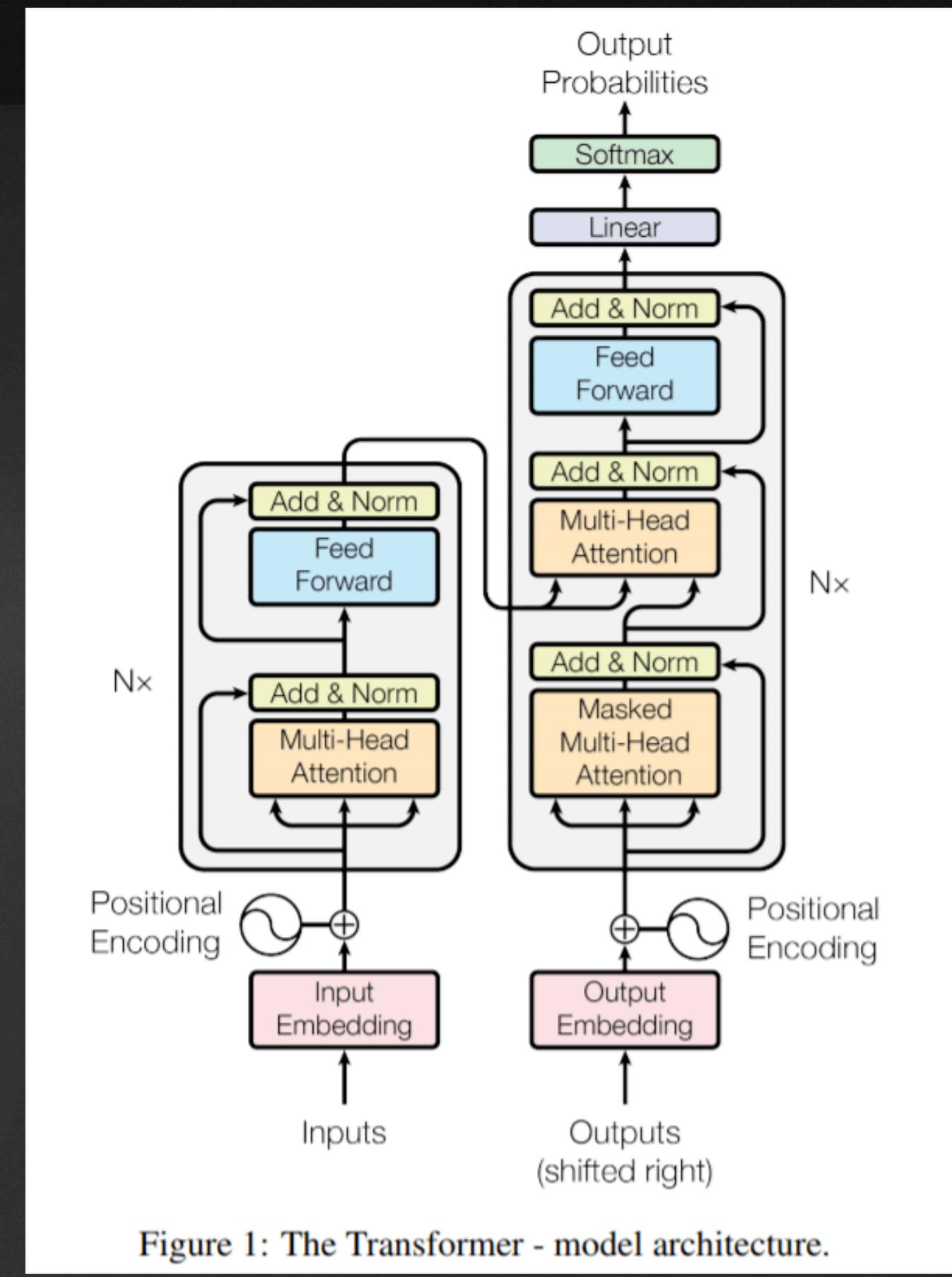
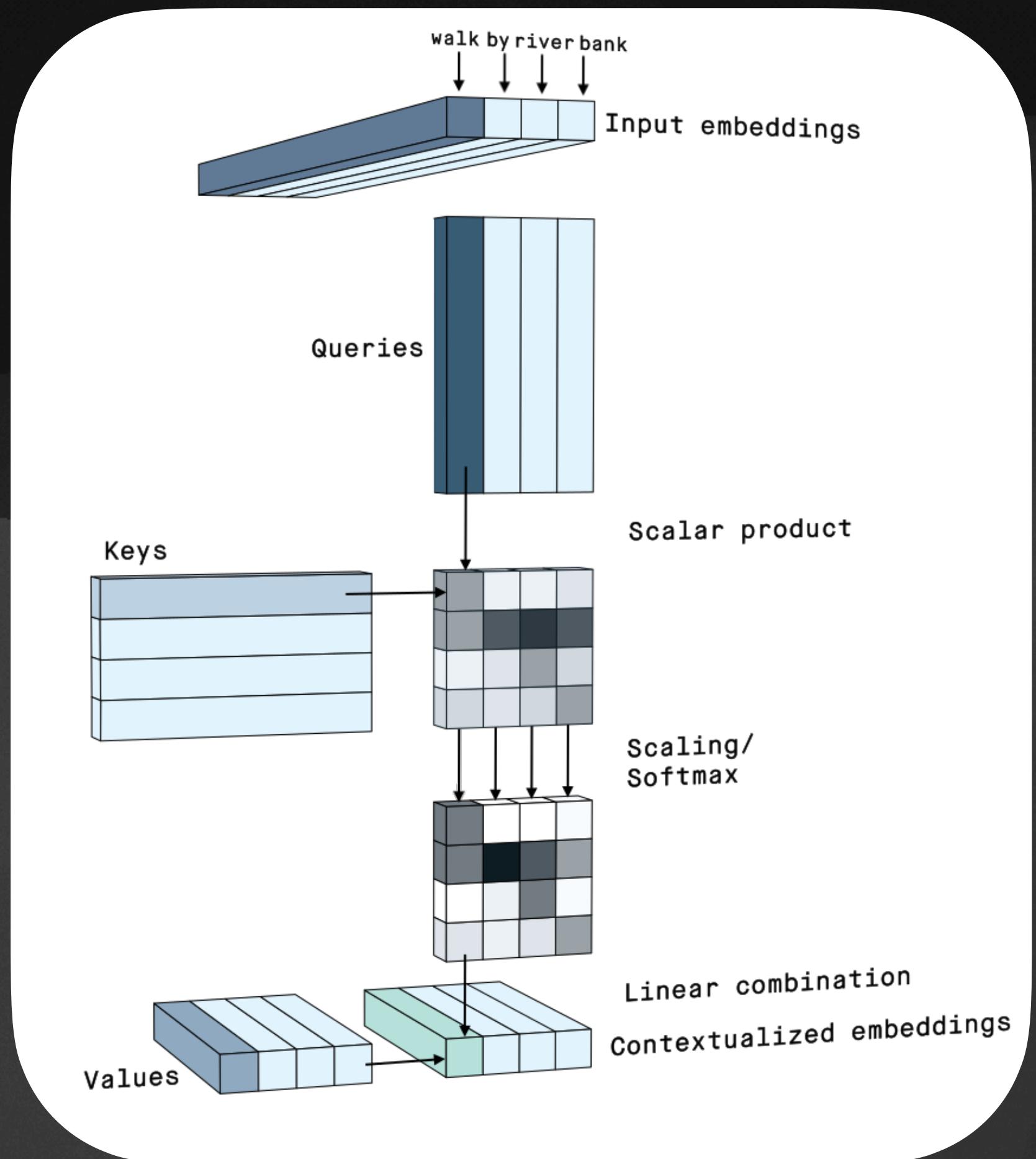
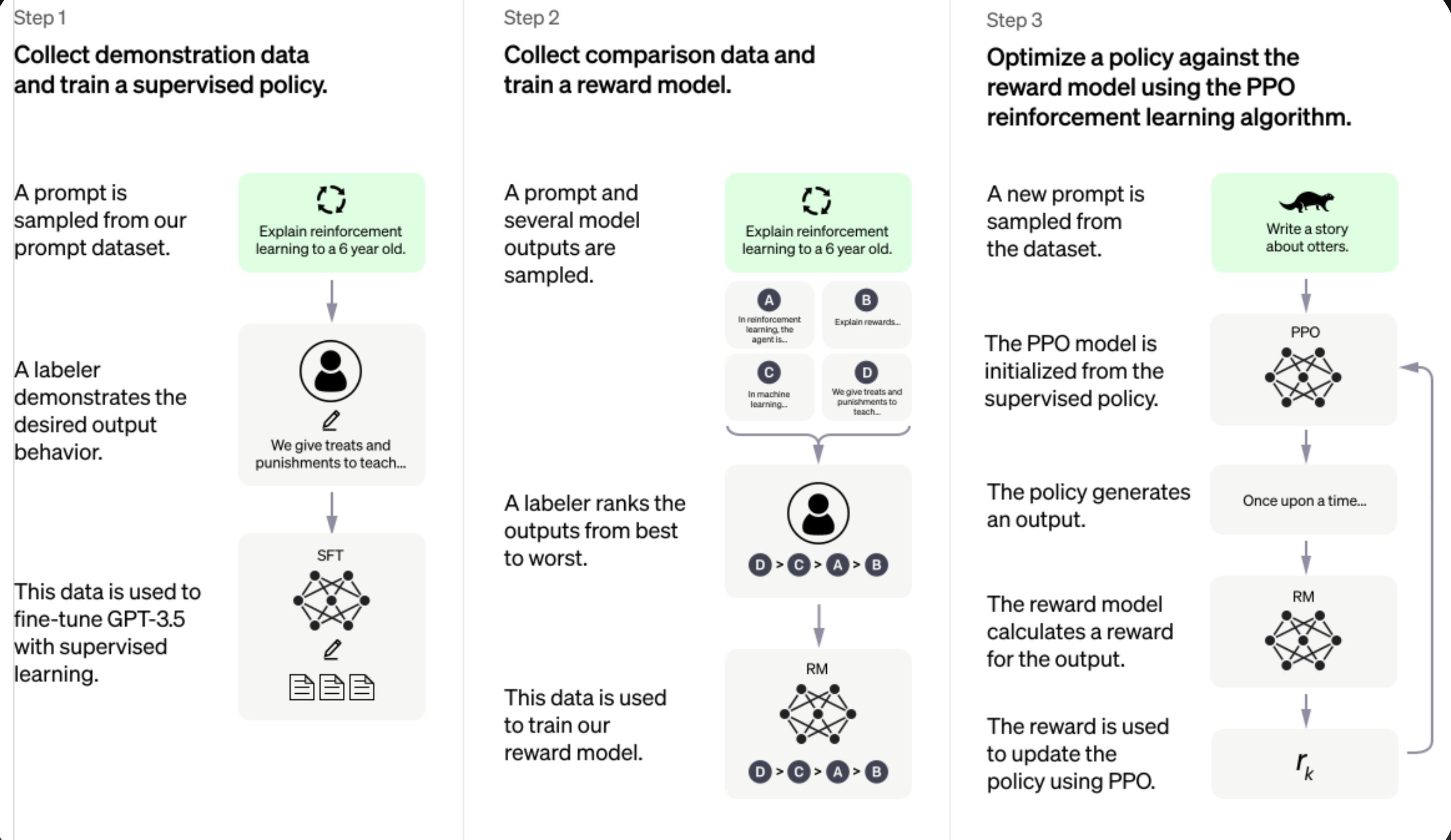


Figure 1: The Transformer - model architecture.

- GPT-2 and 3 have a similar architecture to the BERT model family, just with greater scale
  - GPT-3 ranges from 125M to 175B parameters
  - 125M  $\approx$  ‘bert-base’ model
- Training task: Next Word Prediction (self-supervised)
- At inference time: only use decoder (vs encoder for most BERT-based tasks)
- GPT-4 architecture is unknown but features longer context window + multimodal input (text + image)
- Possibly using OpenAI CLIP architecture

# GPT → ChatGPT

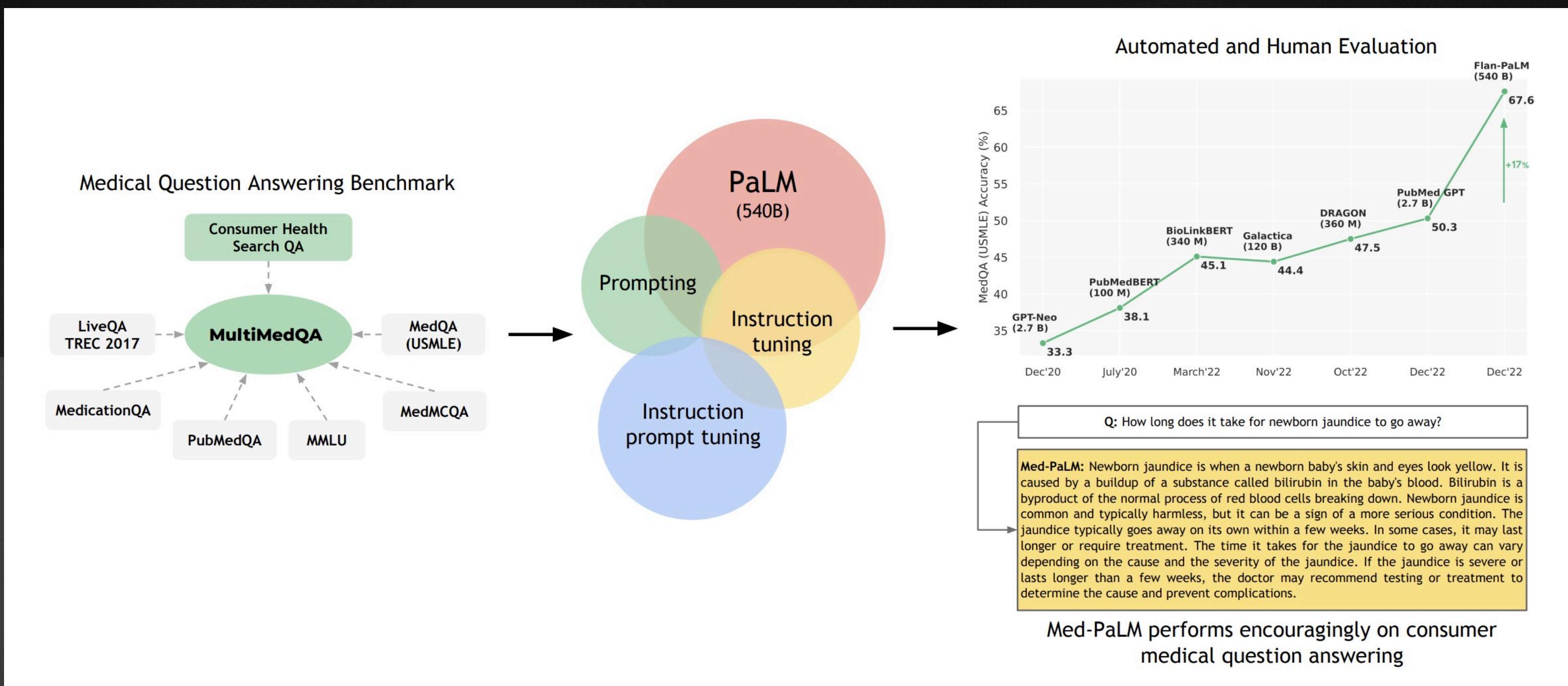


- ChatGPT is fine-tuned from the base GPT model using a technique called **Reinforcement Learning from Human Feedback (RLHF)**

# Related Work

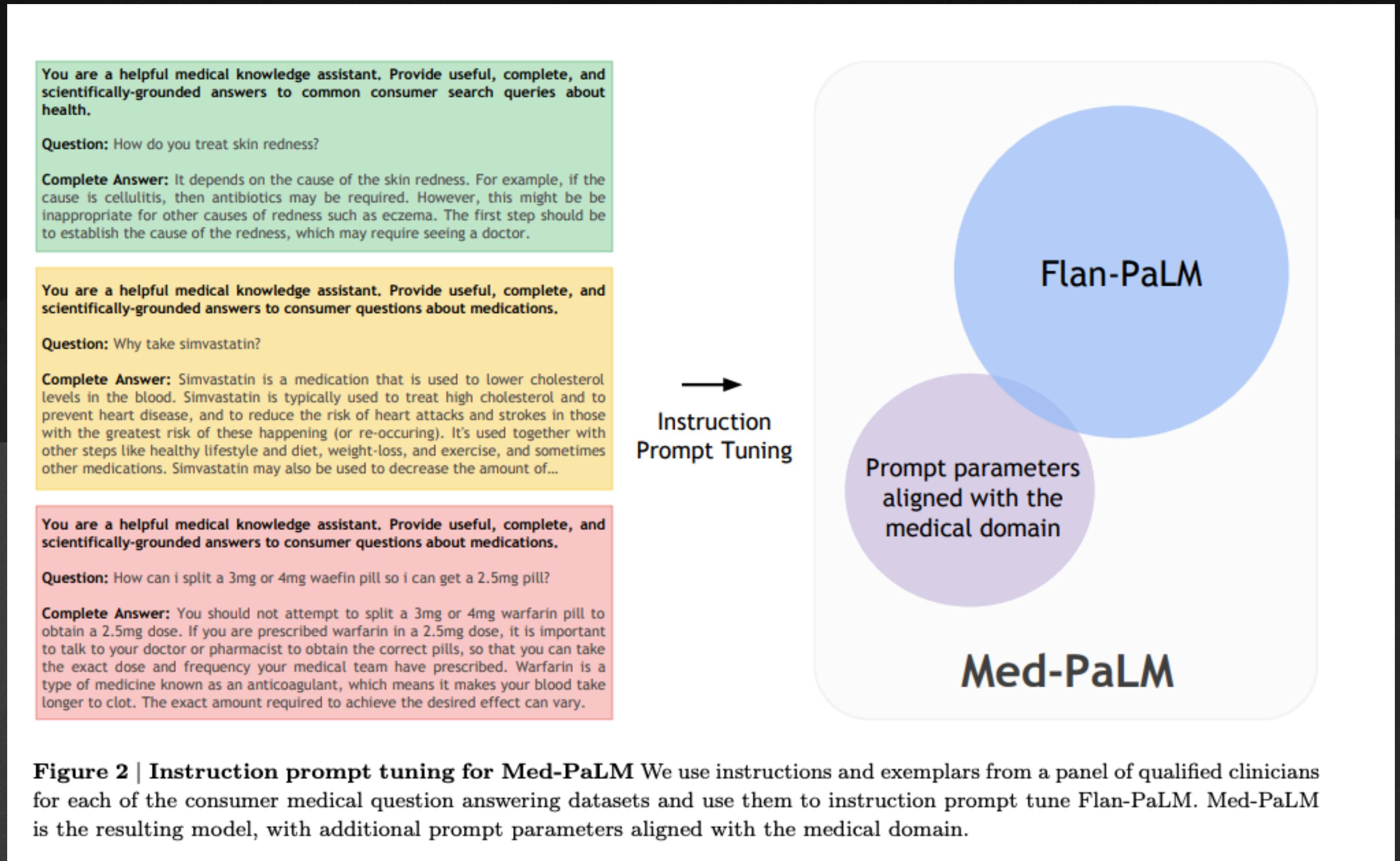
Google PaLM and MedPaLM, Meta AI LLaMA

# Google PaLM and MedPaLM



- PaLM is a Transformer-based LLM
- Flan-PaLM – instruction tuned (form of RLHF)
- MedPaLM
  - “Instruction Prompt Tuning” – 40 human-annotated responses used to further fine-tune the model
  - Evaluation methodology for clinical text

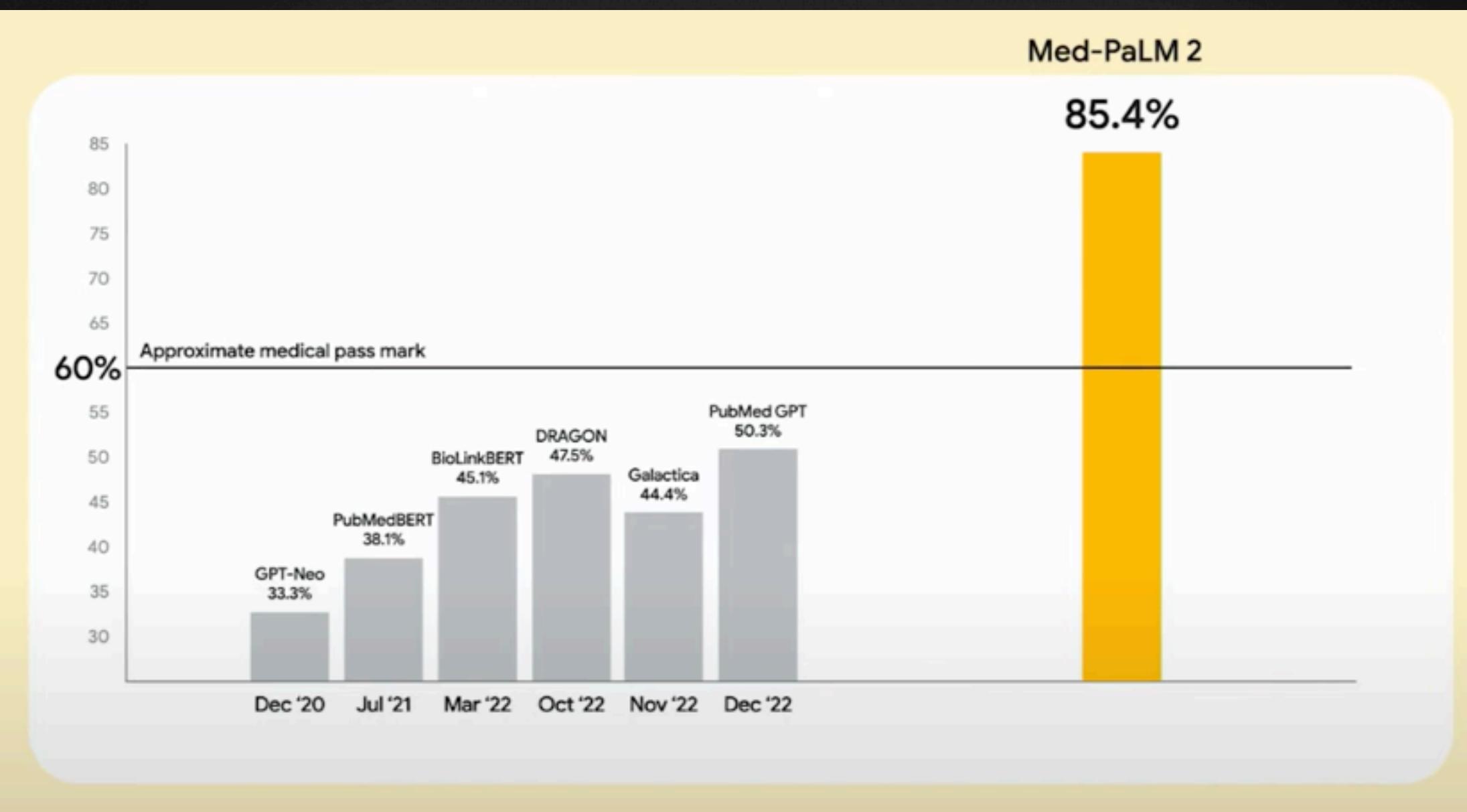
# Google PaLM and MedPaLM



**Figure 2 | Instruction prompt tuning for Med-PaLM** We use instructions and exemplars from a panel of qualified clinicians for each of the consumer medical question answering datasets and use them to instruction prompt tune Flan-PaLM. Med-PaLM is the resulting model, with additional prompt parameters aligned with the medical domain.

- PaLM is a Transformer-based LLM
- Flan-PaLM – instruction tuned (form of RLHF)
- MedPaLM
- “Instruction Prompt Tuning” – 40 human-annotated responses used to further fine-tune the model
- Evaluation methodology for clinical text

# Google PaLM and MedPaLM



- PaLM is a Transformer-based LLM
- Flan-PaLM – instruction tuned (form of RLHF)
- MedPaLM
  - “Instruction Prompt Tuning” – 40 human-annotated responses used to further fine-tune the model
  - Evaluation methodology for clinical text

# Meta AI LLaMA

The screenshot shows the arXiv preprint page for the paper "LLaMA: Open and Efficient Foundation Language Models" by Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample. The page is hosted by Cornell University and acknowledges support from the Simons Foundation and member institutions.

**Download:**

- PDF
- Other formats

(cc) BY

Current browse context:  
cs.CL  
< prev | next >  
new | recent | 2302

Change to browse by:  
cs

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

1 blog link (what is this?)

Export BibTeX Citation

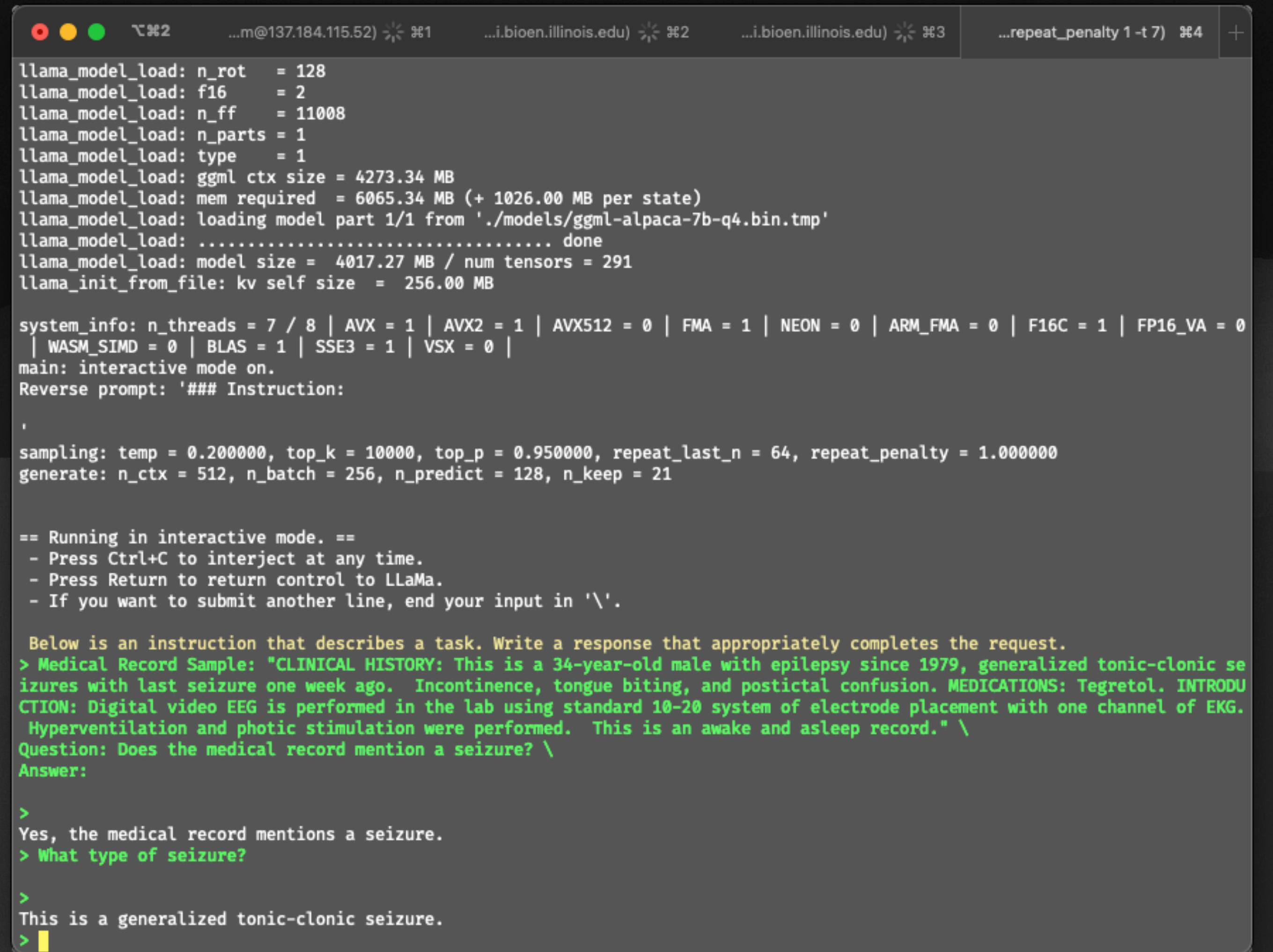
Bookmark

Subjects: Computation and Language (cs.CL)  
Cite as: arXiv:2302.13971 [cs.CL]  
(or arXiv:2302.13971v1 [cs.CL] for this version)  
<https://doi.org/10.48550/arXiv.2302.13971>

**Submission history**  
From: Gautier Izacard [view email]  
[v1] Mon, 27 Feb 2023 17:11:15 UTC (364 KB)

- Similar Transformer-based architecture on public datasets
- Several research groups/independent teams have done further instruction tuning, resulting in (relatively) small models that can be run locally

# Meta AI LLaMA



The screenshot shows a terminal window with four tabs open. The active tab displays the LLaMA model loading process and an interactive session. The output includes:

```
llama_model_load: n_rot    = 128
llama_model_load: f16      = 2
llama_model_load: n_ff     = 11008
llama_model_load: n_parts  = 1
llama_model_load: type     = 1
llama_model_load: ggml ctx size = 4273.34 MB
llama_model_load: mem required = 6065.34 MB (+ 1026.00 MB per state)
llama_model_load: loading model part 1/1 from './models/ggml-alpaca-7b-q4.bin.tmp'
llama_model_load: ..... done
llama_model_load: model size = 4017.27 MB / num tensors = 291
llama_init_from_file: kv self size = 256.00 MB

system_info: n_threads = 7 / 8 | AVX = 1 | AVX2 = 1 | AVX512 = 0 | FMA = 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0
| WASM SIMD = 0 | BLAS = 1 | SSE3 = 1 | VSX = 0 |
main: interactive mode on.
Reverse prompt: '### Instruction:

'

sampling: temp = 0.200000, top_k = 1000, top_p = 0.950000, repeat_last_n = 64, repeat_penalty = 1.000000
generate: n_ctx = 512, n_batch = 256, n_predict = 128, n_keep = 21

== Running in interactive mode. ==
- Press Ctrl+C to interject at any time.
- Press Return to return control to LLaMa.
- If you want to submit another line, end your input in '\'.

Below is an instruction that describes a task. Write a response that appropriately completes the request.
> Medical Record Sample: "CLINICAL HISTORY: This is a 34-year-old male with epilepsy since 1979, generalized tonic-clonic seizures with last seizure one week ago. Incontinence, tongue biting, and postictal confusion. MEDICATIONS: Tegretol. INTRODUCTION: Digital video EEG is performed in the lab using standard 10-20 system of electrode placement with one channel of EKG. Hyperventilation and photic stimulation were performed. This is an awake and asleep record." \
Question: Does the medical record mention a seizure? \
Answer:
>
Yes, the medical record mentions a seizure.
> What type of seizure?
>
This is a generalized tonic-clonic seizure.
>
```

- Similar Transformer-based architecture on public datasets
- Several research groups/independent teams have done further instruction tuning, resulting in (relatively) small models that can be run locally

# “Prompt Engineering”

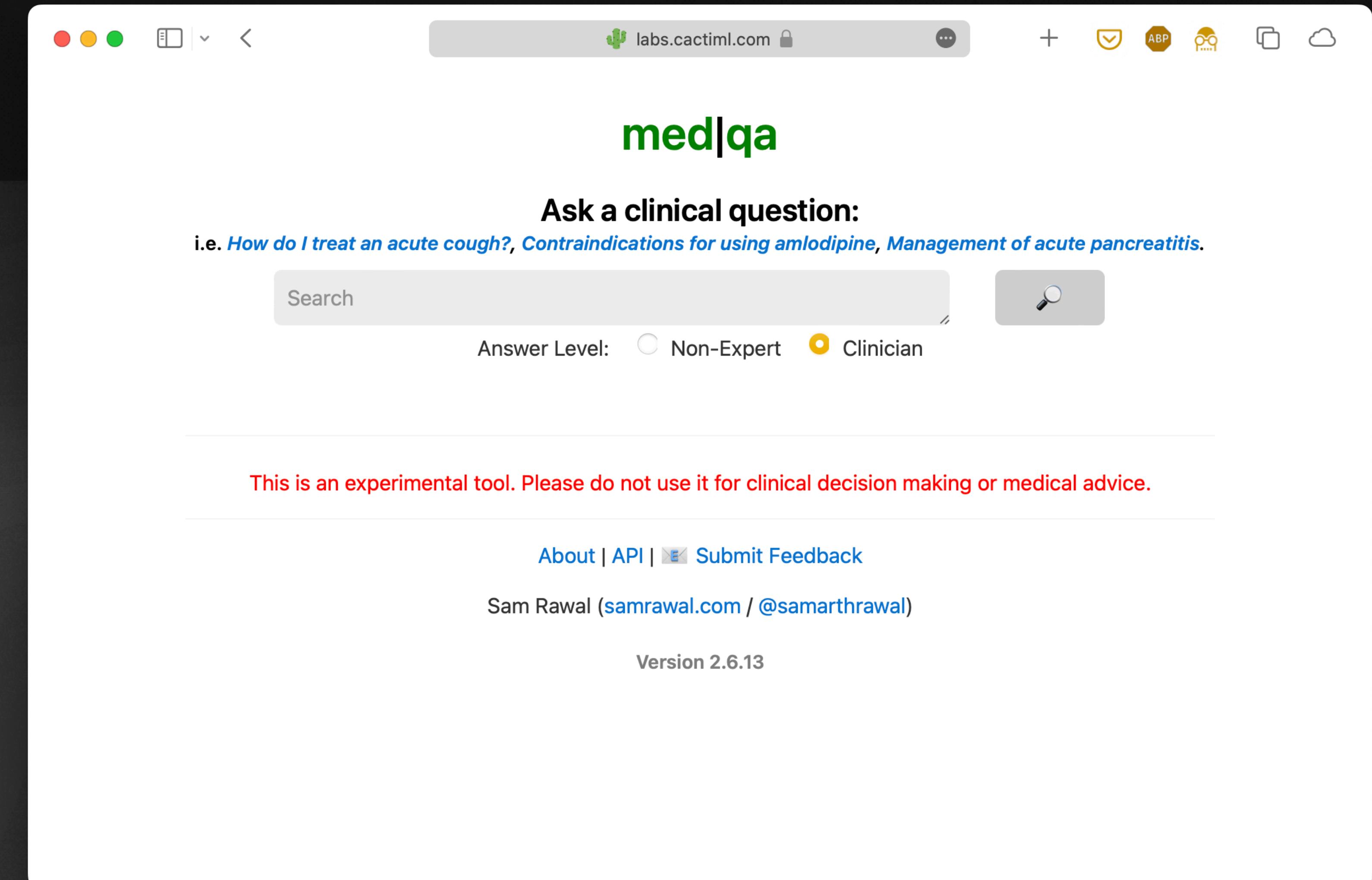
- aka “in-context learning”
- Designing methods to steer the output of LLMs for desired functionality
- “subtractive sculpting”: these models are trained to predict next sequences; prompting **narrows the probability distribution of likely words** depending on task
- Active area of study: both experimental work + more rigorous research
  - chain-of-thought prompting
  - gradient descent for optimizing prompts
  - domain-specific (ie math) work

# Clinical GPT

**Small personal experiments + feedback from physicians**

# MedQA: Clinical Information Retrieval + Question Answering System

- Experimenting with using GPT-3 for answering clinical questions, grounded in reference material (NCBI StatPearls)
- Goal: utilize the apparent reasoning + generative capabilities of GPT, while limiting the tendencies of "hallucination"\*
- Some abilities of LLMs that can be useful in clinical context:



<https://labs.cactiml.com/medqa>

# MedQA

- Grounding responses with proper context (reference material)

The screenshot shows a web browser window for 'med|qa' at 'labs.cactiml.com'. The page has a light gray header with a magnifying glass icon, a plus sign, a square icon, a cloud icon, and a double arrow icon. The main content area has a white background.

**med|qa**

**Ask a clinical question:**  
i.e. [How do I treat an acute cough?](#), [Contraindications for using amlodipine](#), [Management of acute pancreatitis](#).

Management of acute pancreatitis

Answer Level:  Non-Expert  Clinician

*Management of acute pancreatitis*

**Answer:**

Acute pancreatitis is an acute response to injury of the pancreas. Common etiologies of acute pancreatitis are gallstones and alcohol use, and it accounts for about 275,000 hospital admissions annually in the United States. The management of acute pancreatitis should involve specific counseling geared toward the etiology of pancreatitis and pain management is a critical priority. Intragastric balloon therapy can lead to acute pancreatitis, either due to direct injury from the balloon or the distended stomach, or via an indirect mechanism where a pancreatic injury is secondary to a duodenal obstruction.

[Try Again](#) | [More specific](#) | [More general](#) | Estimated Relevance Score: 1.0

Ask Follow Up Question

Example Follow Ups: [What are the symptoms of acute pancreatitis?](#), [What is the treatment for acute pancreatitis?](#), [What is the mechanism for acute pancreatitis?](#)

---

**Additional Information:**

- The two most common causes of acute pancreatitis in the United States are gallstones (35% to 40% of cases) and alcohol use (30% of cases)
- The essential stage of management of acute pancreatitis is the first 12 to 24 hours of admission
- An abdominal CT or ultrasound can be used for radiological diagnosis of acute pancreatitis

---

**References + Additional Reading:**

- [1] [Acute Pancreatitis. National Institute of Medicine StatPearls](#).
- [2] [Pancreatitis. National Institute of Medicine StatPearls](#).
- [3] [Intragastric Balloon. National Institute of Medicine StatPearls](#).

This is an experimental tool. Please do not use it for clinical decision making or medical advice.

# MedQA

## ● Follow-up question-answering

The image displays two side-by-side screenshots of the MedQA application interface, illustrating the process of follow-up question-answering.

**Screenshot 1 (Left):**

- Header:** med|qa
- Text Input:** Ask a clinical question:  
i.e. [How do I treat an acute cough?](#), [Contraindications for using amlodipine](#), [Management of acute pancreatitis](#).
- Search Bar:** How do I treat an acute cough?
- Text Area:** How do I treat an acute cough?
- Section:** Answer:  
Treatment for an acute cough typically includes supportive measures such as over-the-counter cough and cold medicines. It is important to remember, however, that a cough is a natural defense mechanism and should not be suppressed unless the underlying etiology is identified. For chronic or persistent coughing, a multidisciplinary approach is recommended to diagnose and treat the underlying cause.  
[Try Again](#)
- Search Bar:** When would imaging be indicated?
- Text Area:** Answering... (depending on demand, this could take a couple seconds!)
- Section:** Additional Information:
  - Treatment for an acute cough typically involves supportive measures such as over-the-counter medications.
  - Coughing is a natural defense mechanism and should not be suppressed without identifying the underlying etiology.
  - For chronic or persistent coughing, a multidisciplinary approach is recommended to diagnose and treat the underlying cause.
  - A thorough history is important when trying to establish the cause of a cough.
  - If a cough is prolonged and associated with other symptoms, a referral to a pulmonologist or an ENT surgeon is recommended.
- Section:** References + Additional Reading:
  - [\[1\] Cough \(Nursing\). National Institute of Medicine StatPearls](#)

**Screenshot 2 (Right):**

- Header:** med|qa
- Text Input:** Ask a clinical question:  
i.e. [How do I treat an acute cough?](#), [Contraindications for using amlodipine](#), [Management of acute pancreatitis](#).
- Search Bar:** When would imaging be indicated?
- Text Area:** When would imaging be indicated?
- Section:** Answer:  
Treatment for an acute cough typically involves managing the underlying cause, if it is known. If the cause is not known, then the cough is managed as a chronic cough. A focused history and physical examination as well as a chest x-ray should be conducted in patients with a cough lasting more than 8 weeks. If the cause is still not identified, then further testing such as high-resolution computed tomography, bronchoscopy, and echocardiography may be indicated.  
[Try Again](#)
- Search Bar:** Ask Follow Up Question
- Text Area:** Ask Follow Up Question
- Section:** Additional Information:
  - A diagnosis of a cough is an obvious clinical observation.
  - Most cases of acute cough should be treated empirically and will focus on symptomatic relief.
  - Initial evaluation of a cough of more than eight weeks duration should include a focused history and physical examination, and chest x-ray.
  - Patients should avoid exposure to cough-evoking irritants.
  - Patients who are taking an angiotensin-converting enzyme inhibitor should switch to another drug class.
- Section:** References + Additional Reading:
  - [1] Chronic Cough. National Institute of Medicine StatPearls.
  - [2] Cough (Nursing). National Institute of Medicine StatPearls.
  - [3] Cough. National Institute of Medicine StatPearls.

# MedQA: Clinical Information Retrieval + Question Answering System

- Identification of query intent
  - Change nature of answer (ie more procedure-focused for surgery questions, more details for epidemiological questions)
  - Determine if paragraph- or table-based response is more appropriate

The screenshot shows a web browser window for 'med|qa' at 'labs.cactiml.com'. The search bar contains the query 'bacterial versus viral conjunctivitis'. Below the search bar, the text 'Ask a clinical question:' is followed by an example: 'i.e. How do I treat an acute cough?, Contraindications for using amlodipine, Management of acute pancreatitis.' The main content area displays the search results in a table format.

Topic	Bacterial Conjunctivitis	Viral Conjunctivitis
Etiology	Bacterial	Viral, most commonly adenovirus
Pathophysiology	Colonization of the surface tissues by normal flora	Inflammation of the conjunctiva
Treatment/Management	Antibiotics to reduce duration of symptoms	Patient education to reduce transmission rate
Differential Diagnosis	Viral and allergic conjunctivitis	Bacterial conjunctivitis
Monitoring/Coordination of Care	High transmission rate	High transmission rate
Health Teaching/Health Promotion	High transmission rate	High transmission rate

A detailed description of bacterial and viral conjunctivitis follows:

Bacterial conjunctivitis is caused by bacterial infection, whereas viral conjunctivitis is caused by a virus, such as adenovirus. Bacterial conjunctivitis is typically characterized by conjunctival erythema and purulent discharge, whereas viral conjunctivitis is associated with watery discharge and less erythema. Treatment for bacterial conjunctivitis may include antibiotics, while viral conjunctivitis is generally self-limiting and does not require antibiotics.

# MedQA: Clinical Information Retrieval + Question Answering System

- Change the granularity of answers

The image displays two side-by-side screenshots of a web-based clinical information retrieval and question answering system, labeled "med|qa". Both screenshots show the same user interface for asking a clinical question and viewing the resulting answer, but with different user roles selected.

**Screenshot 1 (Left): Non-Expert User**

- Header:** localhost
- Section:** Ask a clinical question:  
i.e. [How do I treat an acute cough?](#), [Contraindications for using amlodipine](#), [Management of acute pancreatitis](#).
- Input Field:** What is pleural effusion? (with a magnifying glass icon)
- User Role:**  Non-Expert  Clinician
- Output:** What is pleural effusion?  
**Answer:**  
Pleural effusion is a condition where excess fluid accumulates in the pleural cavity, which is the space between the lungs and chest wall. This fluid can cause respiratory distress and reduce the lung's ability to expand, leading to difficulty breathing. It is often caused by infections, malignancies, or other underlying diseases. Diagnosis is usually done by chest X-ray or ultrasound and management of the condition will vary based on the cause.  
[Try Again](#) | [More specific](#) | [More general](#) | Estimated Relevance Score: 0.9  
**Follow-up:** Ask Follow Up Question (with a magnifying glass icon)  
**Example Follow Ups:** [What are the signs and symptoms of pleural effusion?](#), [What is the difference between exudative and transudative pleural effusions?](#), [What are the risk factors for malignant pleural effusion in the pediatric population?](#)

**Screenshot 2 (Right): Clinician User**

- Header:** localhost
- Section:** Ask a clinical question:  
i.e. [How do I treat an acute cough?](#), [Contraindications for using amlodipine](#), [Management of acute pancreatitis](#).
- Input Field:** What is pleural effusion? (with a magnifying glass icon)
- User Role:**  Non-Expert  Clinician
- Output:** What is pleural effusion?  
**Answer:**  
Pleural effusion is the accumulation of excess fluid in the pleural cavity. This fluid can be either exudative, resulting from an inflammatory process or a neoplasm, or transudative, resulting from an imbalance of hydrostatic and oncotic pressures. Pleural effusions can present with respiratory distress, lung hypoplasia, and circulatory diseases. Thoracentesis, paracentesis, and cardiocentesis are used for diagnosis.  
[Try Again](#) | [More specific](#) | [More general](#) | Estimated Relevance Score: 1.0  
**Follow-up:** Ask Follow Up Question (with a magnifying glass icon)  
**Example Follow Ups:** [What are the common causes of pleural effusion?](#), [What is the prognosis for patients with pleural effusion?](#), [What are the diagnostic tests for pleural effusion?](#)

# MedQA: Some Learnings + physician feedback

- Prompting:
  - Gradually built up extensive prompt (300+ word), covering clinical details, better grounding response to citations, type of information to extract/generate, etc.
- Use-case:
  - Not useful for looking up straightforward clinical algorithms/values – can already be easily done via database
  - Reasoning, summarization, comparison/contrast | given input – is a valuable use-case
  - Repurpose generative API for classification purposes

# Some learnings from MedQA: [link]

The screenshot shows a web browser window displaying a Substack article. The header of the browser reads "samrawal.substack.com". The main content area has a light gray background. At the top left, the title "m3 | music, medicine, machine learning" is displayed, followed by a "Dashboard" button with a yellow border. To the right of the dashboard are several small icons: a magnifying glass, a bell, a three-dot menu, and a user profile picture. Below the title, the article has a dark gray header with the title "Large Language Models for Clinical Applications: Part 1" in white text. A subtitle below it reads "Insights and considerations about the clinical applicability of LLMs, and the importance of domain knowledge in guiding these systems". Underneath the title, there is a small circular profile picture of the author, Sam Rawal, and the text "SAM RAWAL" and "APR 11, 2023". Below the author information are three small circular icons: a heart with the number "1", a speech bubble, and a circular arrow. To the right of these icons are two more buttons: "Share" and "...". A short note at the bottom of the article says "Here's a (very complementary 😊) TL;DR of this post, generated by GPT-4!" followed by a block of text about the author's background and work on MedQA.

**Large Language Models for Clinical Applications:  
Part 1**

Insights and considerations about the clinical applicability of LLMs, and the importance of domain knowledge in guiding these systems

SAM RAWAL  
APR 11, 2023

1    Share    ...

Here's a (very complementary 😊) TL;DR of this post, generated by GPT-4!

I've been involved in clinical Natural Language Processing research for the past eight years -- first as a Computer Science student and, more recently, as a medical student. A lot of my previous research has focused on Named Entity Recognition, Information Retrieval, and automatic document generation from patient medical records. More recently, I've been exploring the applicability of Large Language Models (LLMs) like GPT-3 in the clinical context, including on the tasks I mentioned above, but also new ways of using them to interact with clinical guidelines and records. I've been building MedQA in part to explore how to harness the strengths and limit the drawbacks of LLMs in the clinical domain.

There have been many public comments and demos on the roles of LLMs in the clinical domain from a generative sense (such as writing or autocompleting medical notes). I'd like to share some insights and considerations about the clinical applicability of these models from a *classification* and *decision-making perspective* that I've discovered over the course of building MedQA, using it at the hospital, and

