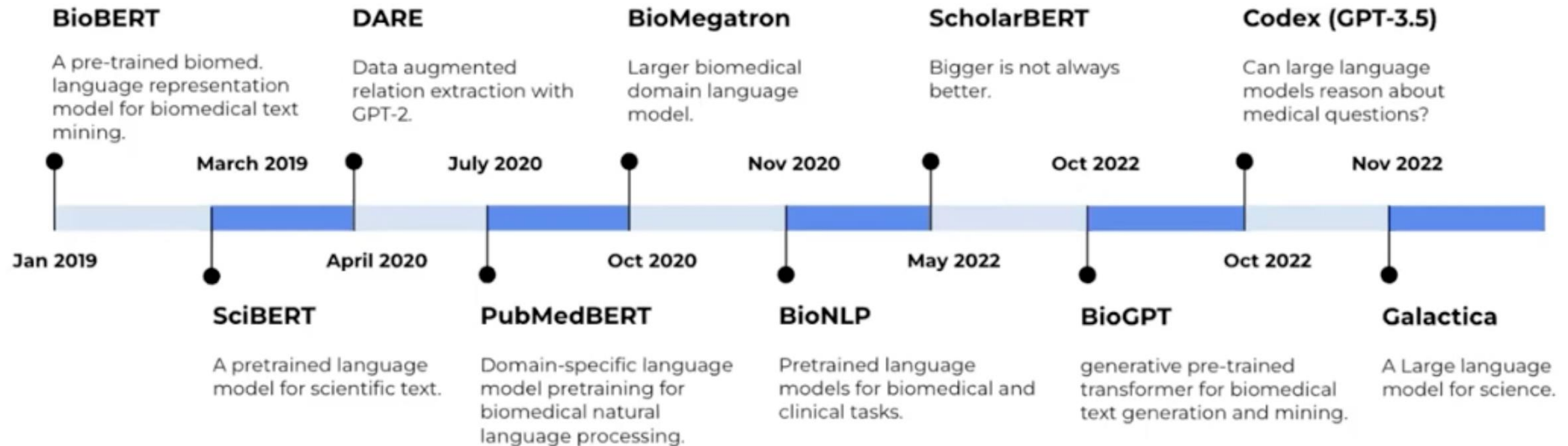


Large Language Models Encode Clinical Knowledge

**Karan Singhal^{1,4}✉, Shekoofeh Azizi^{1,4}✉, Tao Tu^{1,4}, S. Sara Mahdavi¹, Jason Wei¹,
Hyung Won Chung¹, Nathan Scales¹, Ajay Tanwani¹, Heather Cole-Lewis¹, Stephen Pfohl¹,
Perry Payne¹, Martin Seneviratne¹, Paul Gamble¹, Chris Kelly¹, Abubakr Babiker¹,
Nathanael Schärli¹, Aakanksha Chowdhery¹, Philip Mansfield¹, Dina Demner-Fushman²,
Blaise Agüera y Arcas¹, Dale Webster¹, Greg S. Corrado¹, Yossi Matias¹, Katherine Chou¹,
Juraj Gottweis¹, Nenad Tomasev³, Yun Liu¹, Alvin Rajkomar¹, Joelle Barral¹,
Christopher Semturs¹, Alan Karthikesalingam^{1,5}✉ & Vivek Natarajan^{1,5}✉**

Google Research, National Library of Medicine, & Deepmind

Related Work



Gaps in and Issues with Related Work

- Safety and reliability risks
 - Hallucinations of convincing medical misinformation
 - Amplification of biases in training data
 - Deficiencies in reasoning abilities
- Lack of standard benchmarks for testing clinical knowledge
 - Existing benchmarks only cover subsets of the medical knowledge space
- No human evaluation
 - Safety-critical domain requires thoughtful evaluation to mitigate harms

Contributions

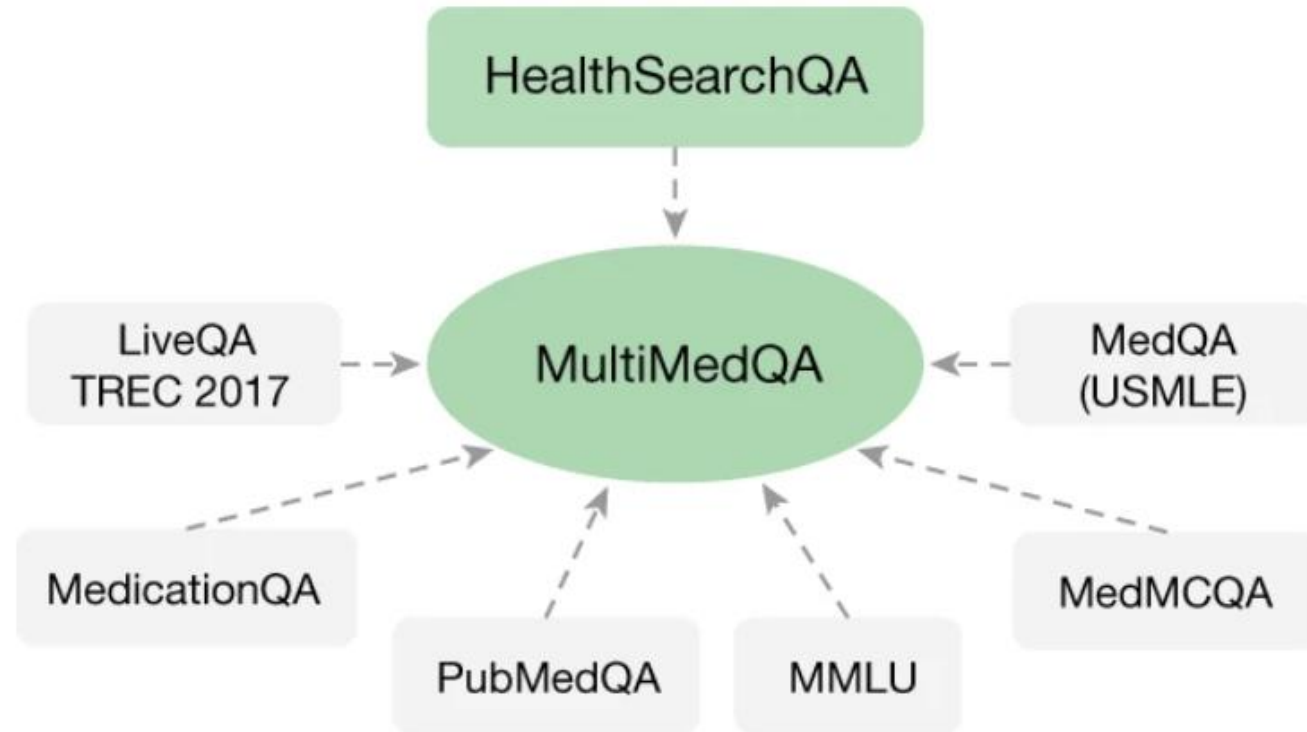
- Provides methods to evaluate LLMs in the context of medical question answering
 - MultiMedQA, a diverse benchmark to assess LLM performance beyond accuracy
 - HealthSearchQA, a dataset of commonly searched health questions
 - Human evaluation framework for medical LLMs
- Implements instruction prompt tuning to align LLMs to the medical domain
- Achieves state-of-the-art performance with Med-PaLM

MultiMedQA Benchmark

- Tests different abilities:
 - Long-form answers
 - Multiple-choice answers
 - Open domain vs. closed domain
 - Medical exams, medical research, and consumer questions
- Assesses response factuality, helpfulness, precision, health equity, potential harm, and use of expert knowledge in reasoning

MultiMedQA Benchmark

Medical question answering benchmark



MultiMedQA Benchmark

Dataset	Format	Size (dev/test)	Domain
MedQA (USMLE)	Q + A (4-5 Choices)	11450 / 1273	General medical knowledge in US medical licensing exam
MedMCQA (AIIMS/NEET)	Q + A (4 Choices and Explanations)	187K / 6.1K	General medical knowledge in Indian medical entrance exams
PubMedQA	Q + Context + A (Yes/No/Maybe) (Long Answer)	500 / 500 #QA pairs: Labeled: 1K Unlabeled: 61.2K Synthetic: 211.3K	Biomedical scientific literature
MMLU	Q + A (4 Choices)	123 / 1089	Medical knowledge covering anatomy, clinical knowledge, college medicine, medical genetics, professional medicine, and college biology
LiveQA TREC-2017	Q + Long Answer (Librarian Answers)	634 / 104	General medical knowledge sought by consumers
MedicationQA	Q + A (Long Answer)	NA / 674	Medication knowledge frequently sought by consumers
HealthSearchQA (Ours)	Q + Manual Expert Evaluation	3375	General medical knowledge searched for by consumers

Baseline: Pathways Language Model (PaLM)

- Largest densely-activated decoder-only model
 - **Densely activated:** all its parameters are actively used during inference for every input
- 540 billion parameters
- Designed for a wide variety of tasks
- Achieved state-of-the-art on a suite of multi-step reasoning tasks

Pathways Language Model (PaLM): Training Dataset

Total dataset size = 780 billion tokens

Data source	Proportion of data
Social media conversations (multilingual)	50%
Filtered webpages (multilingual)	27%
Books (English)	13%
GitHub (code)	5%
Wikipedia (multilingual)	4%
News (English)	1%

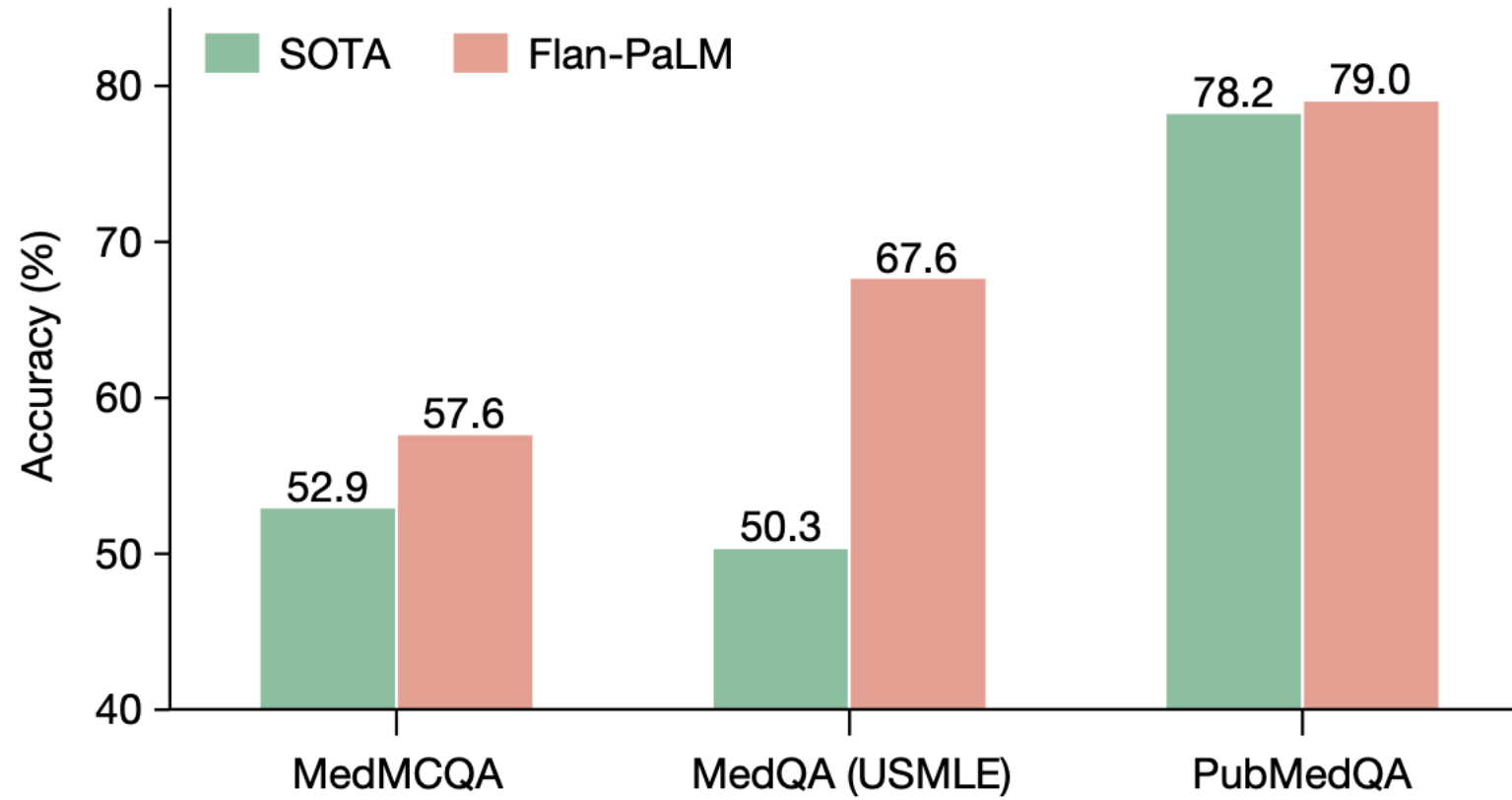
PaLM Findings

- Accuracy improved by ~2x as the PaLM models were scaled from 8B to 540B parameters
- Pre-training corpus likely included significant medical content
- 540B model may have memorized a medical evaluation dataset

Flan-PaLM

- Instruction-tuned version of PaLM
 - Few-shot prompting
 - Chain-of-thought (COT) prompting
- Outperformed PaLM by an average of 9.4% accuracy across a suite of evaluation tasks
- SOTA performance on MedQA, MedMCQA, PubMedQA, and MMLU
- Strong performance on multiple choice
- Answers to consumer medical questions leave room for improvement

Flan-PaLM



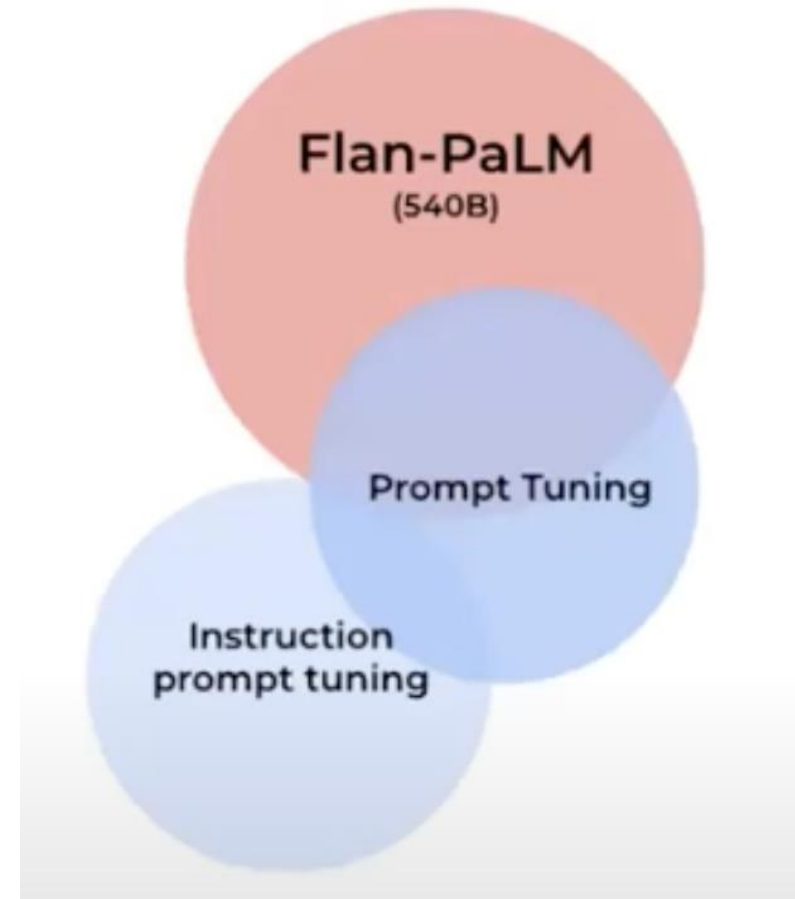
Flan-PaLM Findings and Conclusions

- Instruction tuning improves performance
- Larger versions of Flan-PaLM perform better on medical question answering
- For multiple-choice datasets:
 - No improvement in performance using chain-of-thought prompting over few-shot prompting
 - Self-consistency leads to an improvement over few-shot prompting
- The model becomes increasingly accurate as it withholds more high-uncertainty answers
- LLMs seem to encode uncertainty about their knowledge in the medical domain

Data-Efficient Alignment Strategies



Prompt Strategies



Prompt Tuning

Prompt Strategies

- **Self-consistency prompting:** A technique where a model generates **multiple answers to the same question**, and the final answer is chosen based on the most common (consistent) response across all generations.
- **Chain-of-thought (COT) prompting:** A technique where the model is explicitly **asked to show its step-by-step reasoning** process before giving a final answer, helping it solve complex, multi-step problems by breaking them down into smaller steps.
- **Few-shot prompting:** A technique where the model is given a **small number of examples of questions and correct answers** to demonstrate the desired behavior before answering a new question.

Prompt Tuning

- **Prompt Tuning:** Adding **small learnable tokens** to the input that the model learns to modify during training in order to improve **performance for a particular task**. Trained on **task-specific data**.
- **Instruction Prompt Tuning:** A form of prompt tuning using **instruction-like examples** to tune the learnable tokens. The goal is to **align the model** with domain-specific guidelines and ethical standards.

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to common consumer search queries about health.

Question: How do you treat skin redness?

Complete Answer: It depends on the cause of the skin redness. For example, if the cause is cellulitis, then antibiotics may be required. However, this might be inappropriate for other causes of redness such as eczema. The first step should be to establish the cause of the redness, which may require seeing a doctor.

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to consumer questions about medications.

Question: Why take simvastatin?

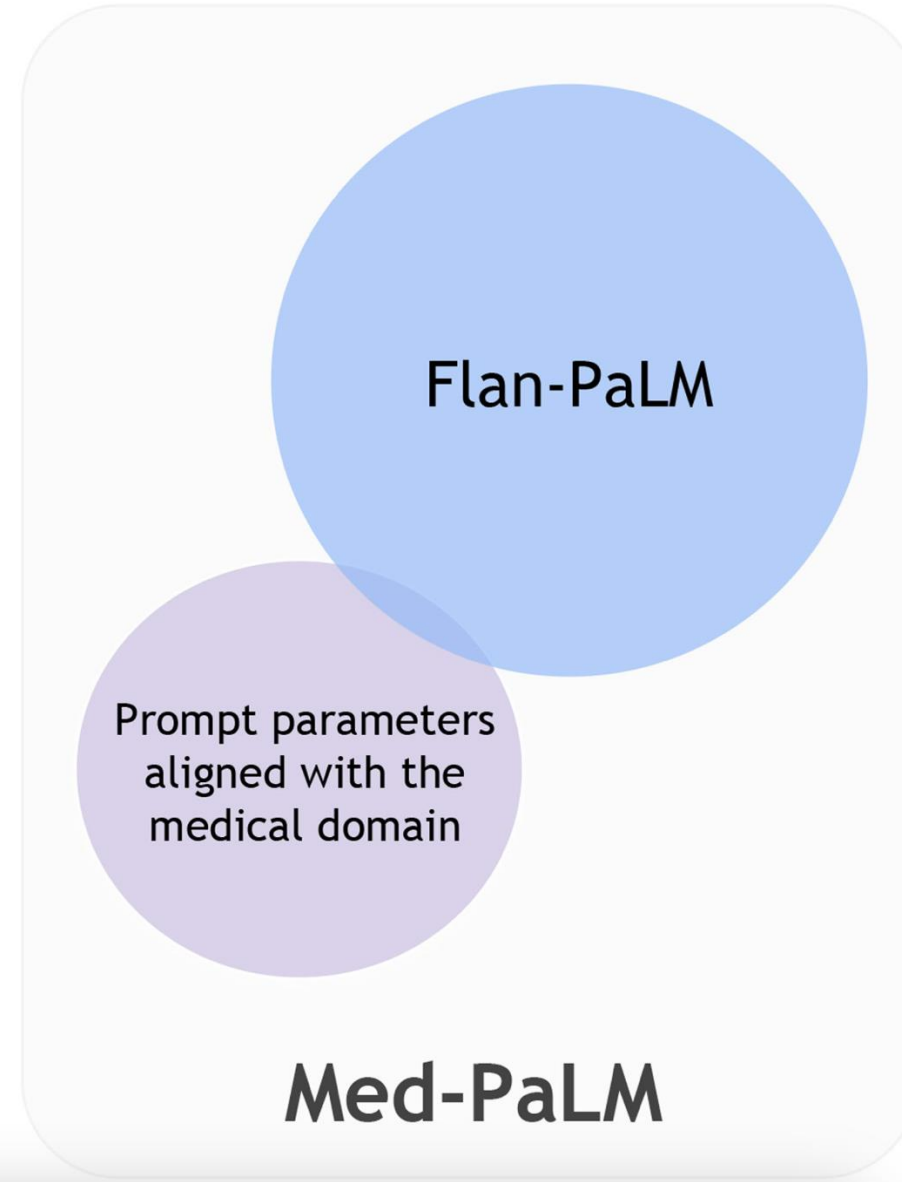
Complete Answer: Simvastatin is a medication that is used to lower cholesterol levels in the blood. Simvastatin is typically used to treat high cholesterol and to prevent heart disease, and to reduce the risk of heart attacks and strokes in those with the greatest risk of these happening (or re-occurring). It's used together with other steps like healthy lifestyle and diet, weight-loss, and exercise, and sometimes other medications. Simvastatin may also be used to decrease the amount of...

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to consumer questions about medications.

Question: How can i split a 3mg or 4mg waefin pill so i can get a 2.5mg pill?

Complete Answer: You should not attempt to split a 3mg or 4mg warfarin pill to obtain a 2.5mg dose. If you are prescribed warfarin in a 2.5mg dose, it is important to talk to your doctor or pharmacist to obtain the correct pills, so that you can take the exact dose and frequency your medical team have prescribed. Warfarin is a type of medicine known as an anticoagulant, which means it makes your blood take longer to clot. The exact amount required to achieve the desired effect can vary.

→
Instruction
Prompt Tuning



Med-PaLM

- Flan-PaLM model, but aligned to the medical domain
- Does **not** use end-to-end fine-tuning of the model with in-domain data – too computationally expensive
- Applies **instruction prompt tuning** using a small set of exemplars
 - Exemplars created by randomly sampling examples from the MultiMedQA free-response datasets and having a panel of clinicians create exemplar answers
- A panel of clinicians craft prompts for each dataset

Human Evaluation Process

- Used 140 questions from the long-form answer datasets
- Generate answers:
 - Flan-PaLM (baseline)
 - Med-PaLM (instruction-tuned)
 - Clinician-written reference answers (by nine clinicians)
- Blind review and rating
 - Each question-answer pair is reviewed and rated by
 - One of three different clinicians
 - One of five different lay people (no medical background)

Human Evaluation Process

- Statistical analysis:
 - Non-parametric bootstrap resampling to estimate confidence intervals
 - A method to estimate the uncertainty of a metric by repeatedly resampling the data (with replacement), recalculating the metric, and using the distribution of those values to compute a confidence interval.

Framework for Human Evaluation: Clinicians

Scientific
consensus

Extent of
possible harm

Likelihood of
possible harm

Evidence of
correct
comprehension

Evidence of
correct retrieval

Evidence of
correct
reasoning

Evidence of
incorrect
comprehension

Evidence of
incorrect
retrieval

Evidence of
incorrect
reasoning

Inappropriate
or incorrect
content

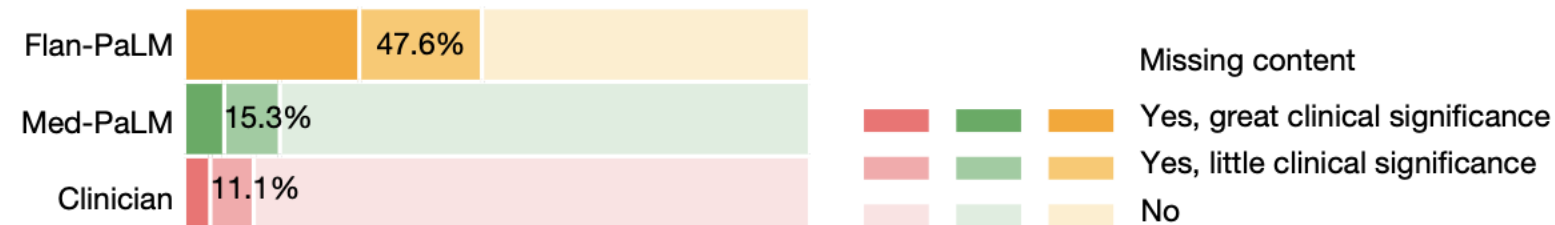
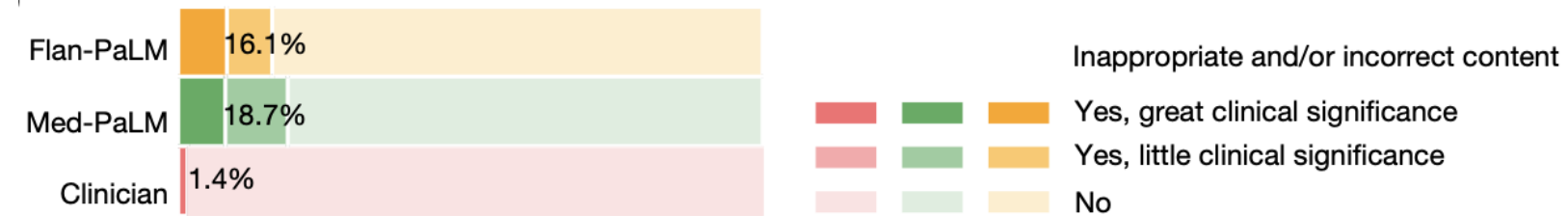
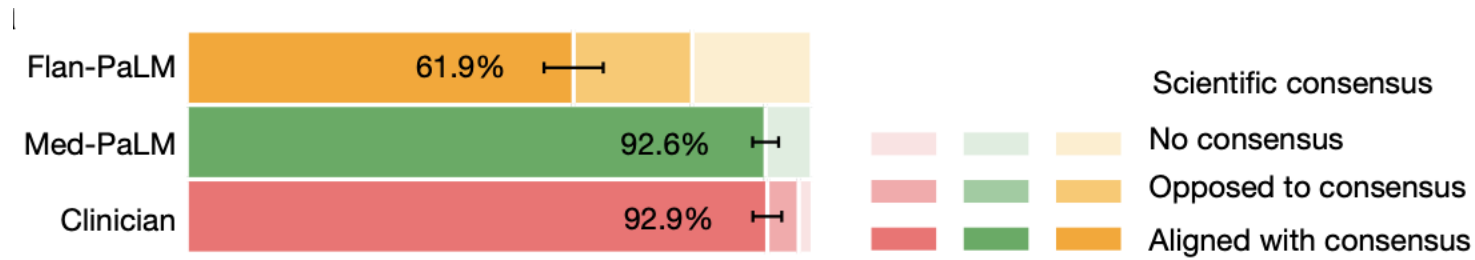
Missing
Content

Possibility of
Bias

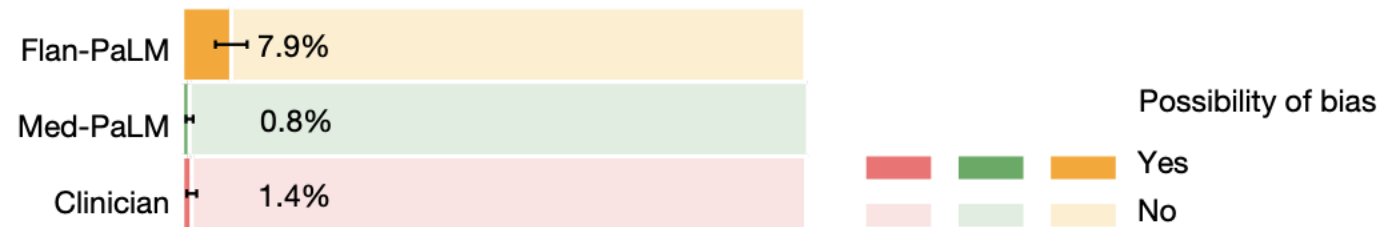
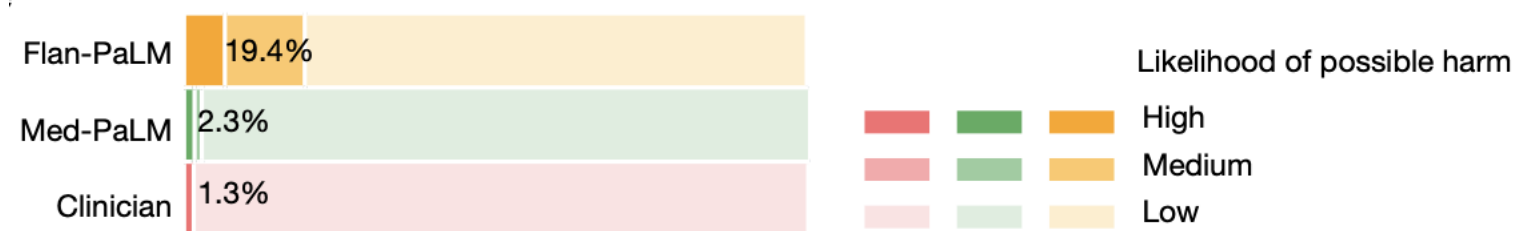
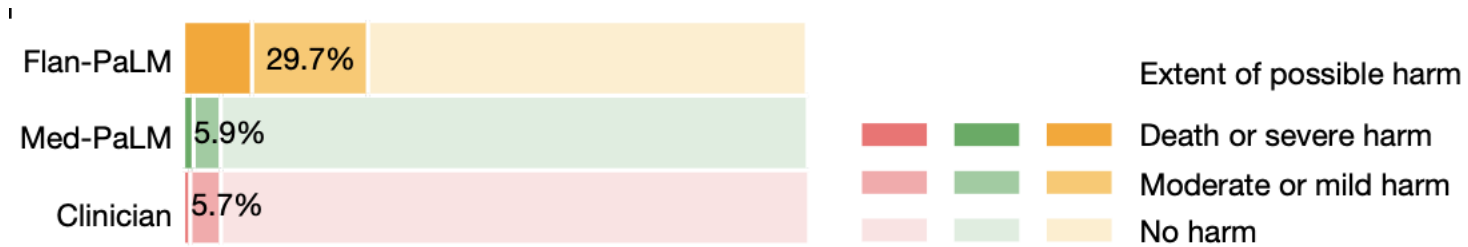
Framework for Human Evaluation: Lay People

- Answer captures user intent
- Helpfulness of the answer

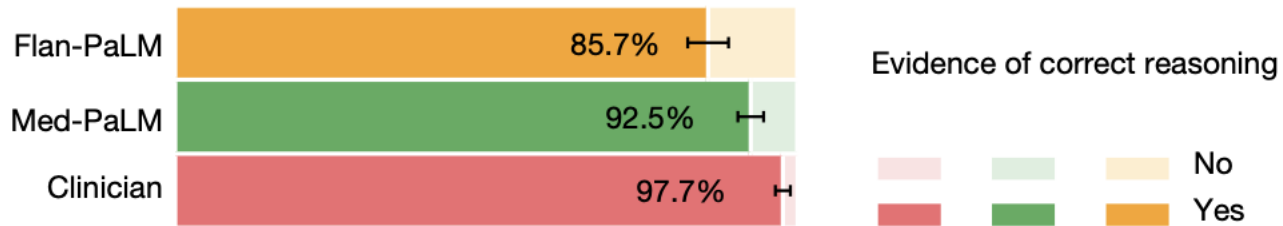
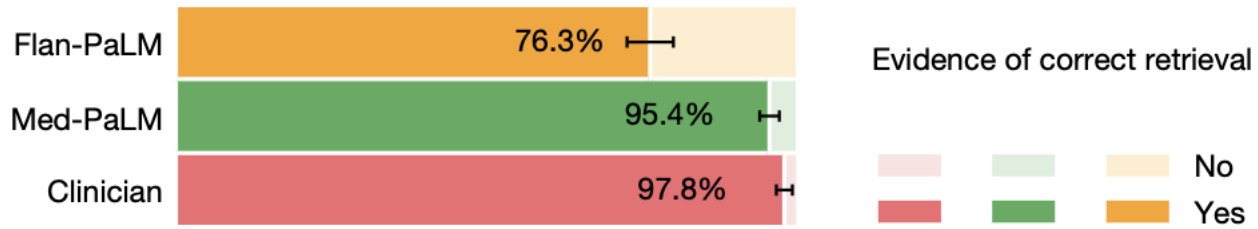
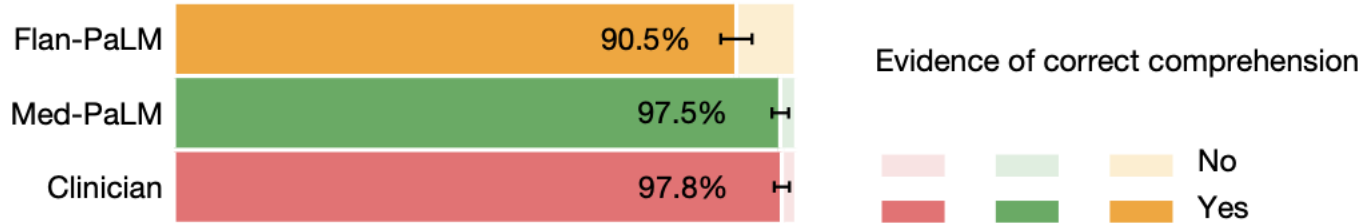
Clinician Evaluation: Results



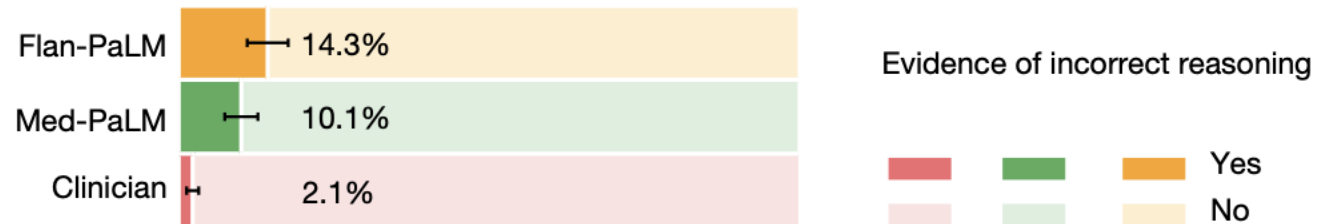
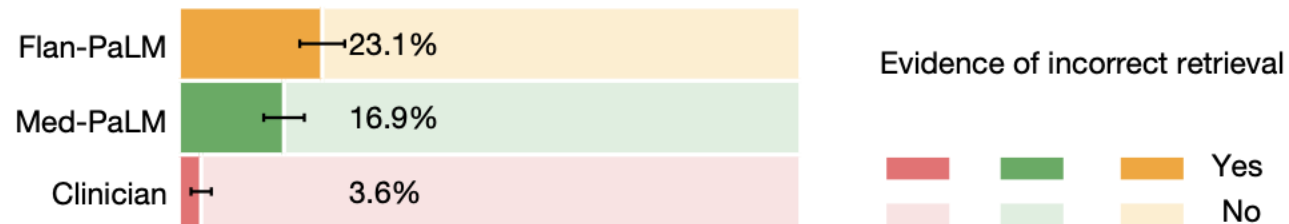
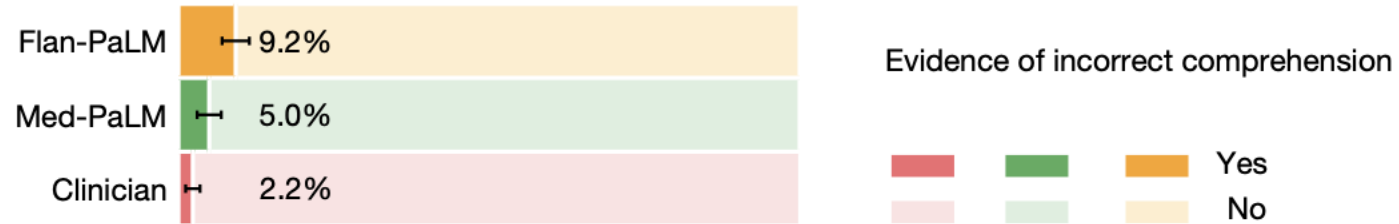
Clinician Evaluation: Results



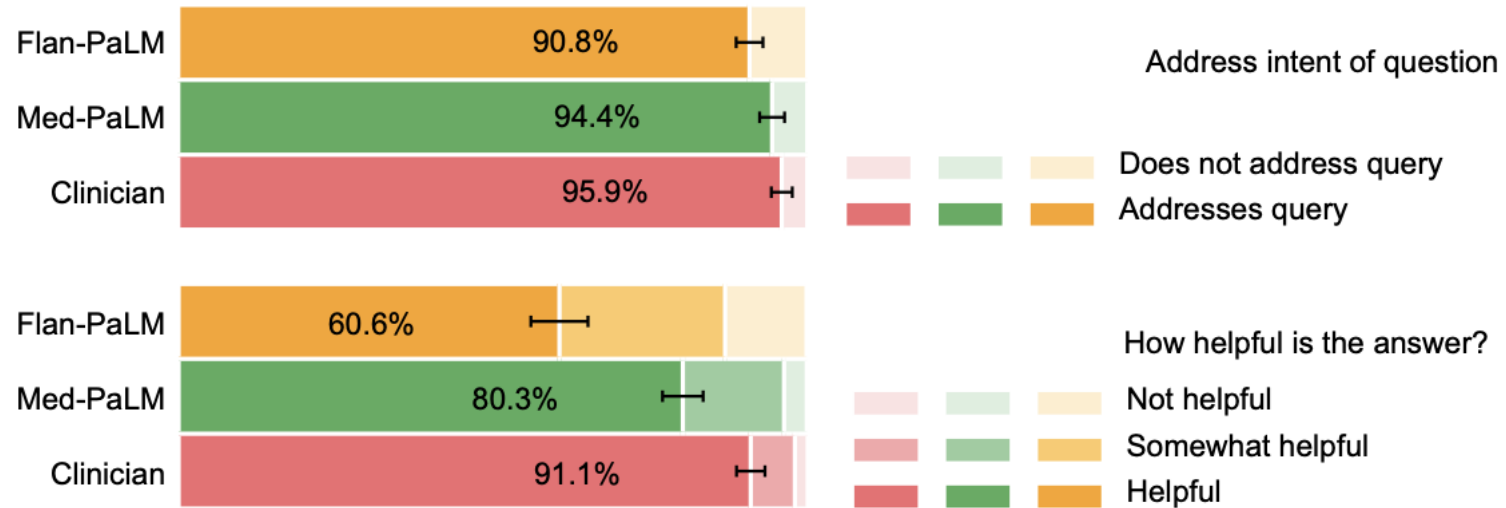
Clinician Evaluation: Results



Clinician Evaluation: Results



Layperson Evaluation: Results



Potential Improvements

- Expansion of the MultiMedQA benchmark
 - MultiMedQA is not exhaustive
 - Need benchmark tasks that reflect real world clinical workflows like synthesizing patient findings into an assessment and plan
- Key LLM Capabilities Needed
 - Grounding of responses in authoritative medical sources
 - Account for the fact that medical consensus changes over time
 - Ability to detect and communicate uncertainty to the user
 - Multilingual responses
- Improvement of Human Evaluation
 - Current rating framework is not exhaustive and is subjective
 - Example: concepts of consensus and harm may differ according to population
 - Limited number of model responses evaluated
 - Larger and more diverse pool of human assessors

Fairness, Equity, and Ethical Considerations

- Evaluation
 - Bias assessed broadly
 - Current approach is not a comprehensive assessment of potential harms, fairness, or equity.
- Algorithmic approach to assess biases
 - Need for algorithmic procedures and benchmarks that seek technical biases known to cause harm.
- Ethical deployment
 - Consideration of use cases and resulting potential harms
 - Need rigorous quality assessment
 - Guardrails to mitigate against over-reliance on the output of Med-PaLM