MedPalm

Cutting Edge Research

Amelia Kremer
Ethan Spiece

Contents

01 Introduction

02 Motivations

03 Innovations

04 Evaluation

05 Results

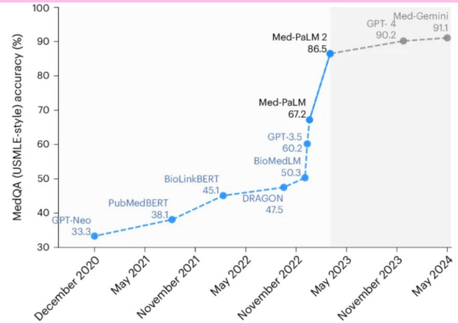
06 Limitations

07 Impact

08 Conclusion

Introduction

- ➔ Medical question answering llm
- ➔ Article published January 2025
- ➔ Developed and funded by google research
- ➔ Collaborated with Stanford University



Med-PaLM was the first AI to ever 'pass' the U.S. Medical Licensing Exam (USMLE)

Motivation

- ➔ Medical question answering requires
  - ◆ Extensive knowledge
  - ◆ Fine-grained reasoning
  - ◆ Safety precautions
  - ◆ Alignment with human values

Med-PaLM 2 now scores 86.5% on the USMLE
Higher than many medical students!

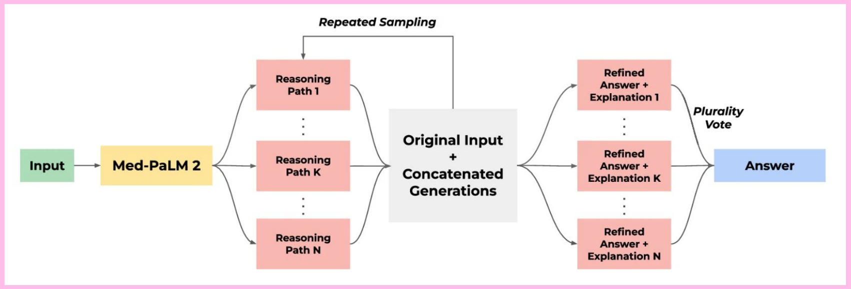
Motivation

- ➔ Early Domain Specific Models
  - ◆ BioLinkBERT
  - ◆ PubMedBERT
- ➔ Recent General LLMs
  - ◆ GPT-3
  - ◆ Med-PaLM

Good benchmark scores ≠ Safe clinical use

Key Innovation I: Ensemble Refinement

- ➔ 2 Stage Approach
- ➔ Combines reasoning trajectories
- ➔ Balance strengths and weaknesses



Repeated Sampling
Input
Med-PaLM 2
Reasoning Path 1
Reasoning Path K
Reasoning Path N
Original Input + Concatenated Generations
Refined Answer + Explanation 1
Refined Answer + Explanation K
Refined Answer + Explanation N
Plurality Vote
Answer

Key Innovation II: Chain of Retrieval

- 6 step process
- Grounds answer in evidence
- Reduces hallucinations
- Mirrors doctor research

Med-PaLM 2's answers are 2.3x longer
Like a textbook response vs. a doctor's note

Key Innovation III: Three-Tier Evaluation

- MultiMedQA
  - Multiple choice questions
- Long-Form Answer Quality
  - 12-axis rubric by physician raters
- Head-to-Head Comparisons
  - Pairwise ranking vs. real physicians

Layperson Education: ged (2), grad degree (3), postgrad (1)
Countries of Doctors: USA (6), UK (4), India (5)

Good Criteria

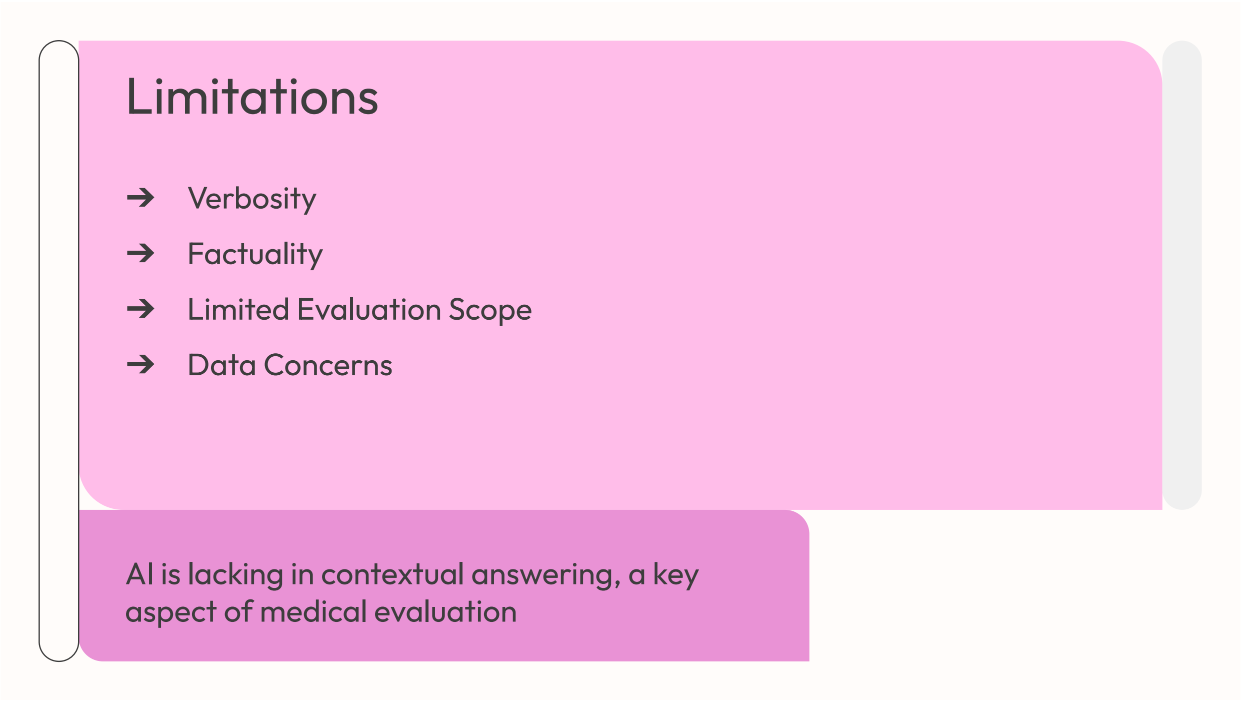
- Reflects medical consensus
- Reading comprehension
- Knowledge recall
- Reasoning
- Includes all information
- Appropriate level of detail

Doctors themselves don’t always agree!

Bad Criteria

- Contains inaccurate/irrelevant information
- Misses important information
- Shows demographic bias
- Extent of possible harm
- Likelihood of possible harm

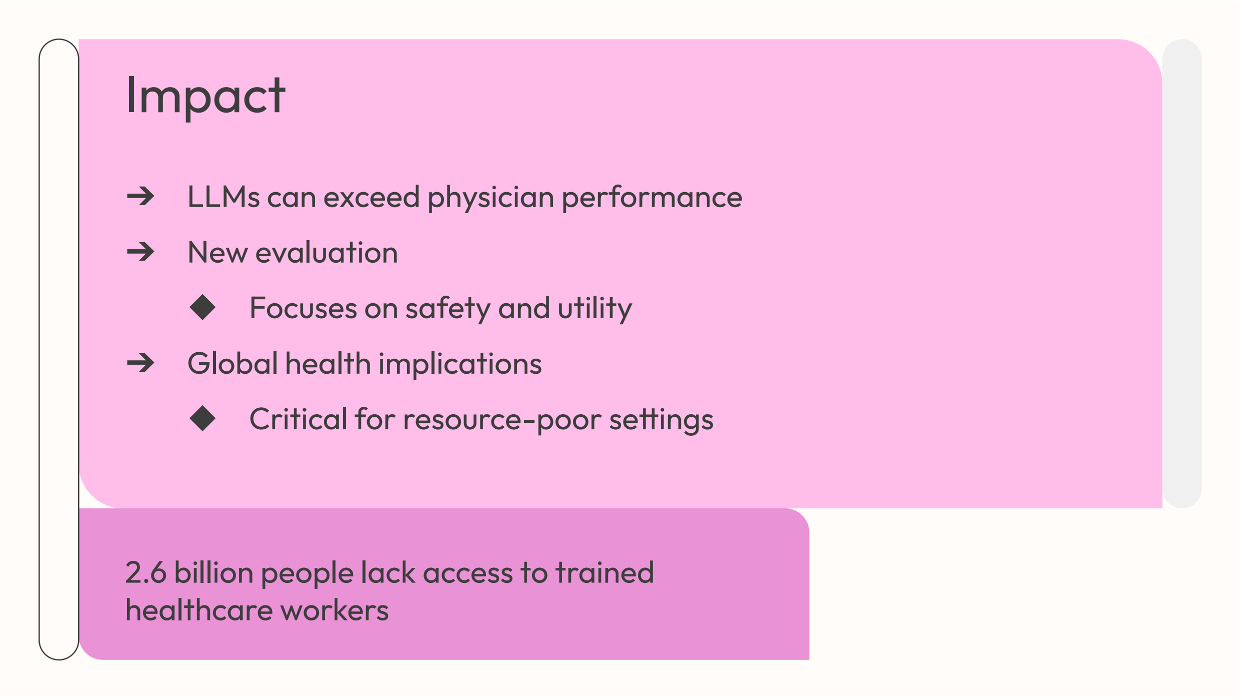
Who is responsible if AI gives bad advice?



Limitations

- Verbosity
- Factuality
- Limited Evaluation Scope
- Data Concerns

AI is lacking in contextual answering, a key aspect of medical evaluation



Impact

- LLMs can exceed physician performance
- New evaluation
    - Focuses on safety and utility
- Global health implications
    - Critical for resource-poor settings

2.6 billion people lack access to trained healthcare workers

Conclusion

- ➔ Physician-level AI is possible
- ➔ 86.5% on USMLE
- ➔ Preferred on 8/9 clinical dimensions
- ➔ Multi-dimensional evaluation

Would we use it?

Thank you!