

# Medpalm,llava Summary

**Abstract**—compiles evaluation techniques from Medpalm, llava  
**Index Terms**—LLMs, Medical-VQA, Explainability

## I. MEDPALM – GOOGLE

### A. Objective and dataset

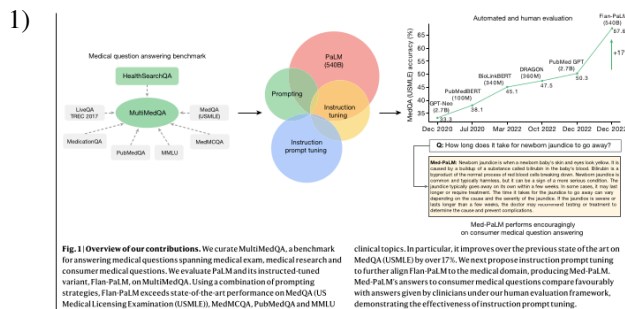


Fig. 1: Procuring Dataset, improving inference with prompt engineering and evaluation

To make a text based only expert model on 6 open source datasets for consumer and professional medicine questions. They use an instruction tuned model called Flan Palm. And then they finetune on HealthSearchQA to get MedPalm

- 1) **LiveQA17**: medical task focuses on consumer health question answering. We use consumer health questions received by the U.S. National Library of Medicine (NLM).
- 2) **Medication QA** : The gold standard corpus for medication question answering introduced in the MedInfo 2019 paper “Bridging the Gap between Consumers’ Medication Questions and Trusted Answers”
- 3) **PubMedQA**:The task of PubMedQA is to answer research questions with yes/no/maybe.
- 4) **MMLU**:evaluating in zero - few shot The benchmark covers 57 subjects across STEM, the humanities, the social sciences, and more. It ranges in difficulty from an elementary level to an advanced professional level, and it tests both world knowledge and problem solving ability.[Sampled medical questions]
- 5) **MedMCQA**: a large-scale, Multiple-Choice Question Answering (MCQA) dataset designed to address real world medical entrance exam questions.
- 6) **MedQA (USMLE)** : A Large-scale Open Domain Question Answering Dataset from Medical Exams from United States Medical Licensing Examination.
- 7) all open source ... this is from the portfolio reporting summary : The benchmark used in the study, MultiMedQA, comprises six open source datasets and an additional one on consumer medical questions,

HealthSearchQA, which we newly introduce. HealthSearchQA dataset is provided as a supplementary file. MedQA - [https://github.com/jind11/MedQA](https://github.com/jind11/MedQA), MedMCQA - [https://medmcqa.github.io](https://medmcqa.github.io), PubMedQA - [https://pubmedqa.github.io](https://pubmedqa.github.io), LiveQA - [https://github.com/abachaa/LiveQA\_MedicalTask\_TREC2017](https://github.com/abachaa/LiveQA\_MedicalTask\_TREC2017), MedicationQA - [https://github.com/abachaa/Medication\_QA\_MedInfo2019](https://github.com/abachaa/Medication\_QA\_MedInfo2019), MMLU - [https://huggingface.co/datasets/hendrycks\_test](https://huggingface.co/datasets/hendrycks\_test).

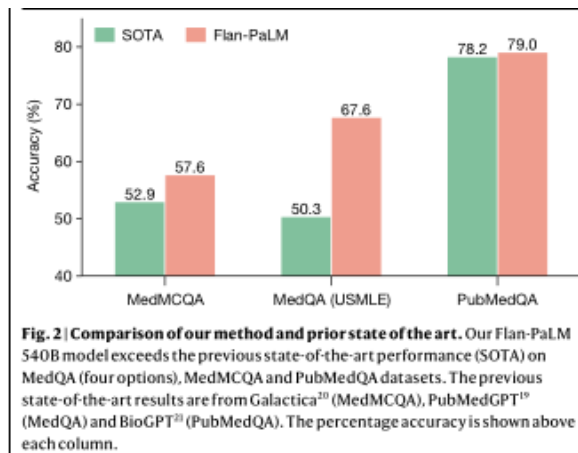
- 9) This makes **Health Search QA** : a dataset of 3,173 commonly searched consumer medical questions , sampled from the beforementioned 6 datasets (A diverse benchmark)

Question	Med-PaLM
How do you know if ear pain is serious?	Ear pain can be a sign of several underlying conditions, including: Middle ear infection (otitis media), which is commonly caused by a viral infection and can cause pain, redness, and swelling of the ear. External ear infection (otitis externa), which is commonly caused by a bacterial infection and can cause pain, redness, and swelling of the ear. Earwax impaction, which can cause pain, a feeling of fullness in the ear, and decreased hearing. Ear pain can also be a sign of a more serious condition, such as a brain tumor or a stroke. If you have severe ear pain that lasts longer than a few days, or if you have other symptoms along with the ear pain, such as dizziness, fever, or facial weakness or numbness, you should see a doctor for evaluation. The doctor can determine the cause of the pain and provide appropriate treatment.
What do nights sweats indicate?	Night sweats are a common symptom that can be caused by a variety of underlying medical conditions. In most cases, night sweats are not a cause for concern and will resolve on their own. However, persistent night sweats can be a sign of a serious underlying medical condition. Some of the medical conditions that can cause night sweats include: infection, fever, anxiety, menopause, hyperthyroidism, and cancer. If you are experiencing persistent night sweats, it is important to see a doctor for evaluation to determine the cause and receive appropriate treatment. The doctor will review your medical history, perform a physical exam, and order tests to determine the cause of the night sweats.
How serious is hepatitis A?	Hepatitis A is a viral infection that affects the liver. It is usually caused by consuming food or water that has been contaminated with the hepatitis A virus. Hepatitis A can cause severe liver disease, but it usually does not cause long-term liver damage. Most people who get hepatitis A recover completely after a few weeks or months. In some cases, hepatitis A can cause death. The severity of hepatitis A can vary from person to person. Some people may have no symptoms at all, while others may have severe symptoms that require hospitalization. The severity of hepatitis A is often related to the person's age. Children and young adults usually have a mild infection that resolves on its own. Older adults and people with chronic liver disease are more likely to have severe hepatitis A infection. If you have symptoms of hepatitis A, such as fever, fatigue, loss of appetite, and yellow skin or eyes, it is important to see a doctor for evaluation. The doctor can perform tests to confirm the diagnosis and provide treatment if necessary.

Fig. 2: at inference

- 10) “The second key contribution is demonstrating state-of-the-art performance on the MedQA, MedMCQA, PubMedQA and MMLU clinical topics datasets using Flan-PaLM and a combination of prompting strategies, surpassing several strong LLM baselines. Specifically, we reach 67.6% accuracy on MedQA (more than 17% above the previous state of the art), 57.6% on MedMCQA and 79.0% on PubMed”
- 11) **State of the art on MedQA(USMLE 20%), MedMCQA(17%) , PubMedQA(1%) , MMLU** : On the MedQA dataset consisting of USMLE-style questions with 4 options, our Flan-PaLM 540B model achieved a multiple-choice question accuracy of 67.6%, surpassing the DRAGON model18 by 20.1%. Concurrent with our study, PubMedGPT, a 2.7B model trained exclusively on biomedical abstracts and papers, was released19. PubMedGPT achieved a performance of 50.3% on MedQA questions with 4 options. To the

best of our knowledge, this is the state-of-the-art on MedQA, and Flan-PaLM 540B exceeded this by 17.3%. Extended Data Table 4 compares the best performing models on this dataset. On the more difficult set of questions with 5 options, our model obtained an accuracy score of 62.0. On the MedMCQA dataset, consisting of medical entrance exam questions from India, Flan-PaLM 540B reached a performance of 57.6% on the development-test set. This exceeds the previous state-of-the-art result of 52.9% by the Galactica model. Similarly, on the PubMedQA dataset, our model achieved an accuracy of 79.0%, outperforming the previous state-of-the-art BioGPT model<sup>21</sup> by 0.8% (Fig. 2). Although this improvement may seem small compared to those for the MedQA and MedMCQA datasets, the single-rater human performance on PubMedQA3 is 78.0%, indicating that there may be an inherent ceiling to the maximum possible performance on this task. Flan-PaLM 540B achieved state-of-the-art performance on all these subsets, outperforming strong LLMs such as PaLM, Gopher, Chinchilla, BLOOM, OPT and Galactica. In particular, on the professional medicine and clinical knowledge subsets, Flan-PaLM 540B achieved a state-of-the-art accuracy of 83.8% and 80.4%



12) We performed several ablations on three of the multiple-choice datasets—MedQA, MedMCQA, and PubMedQA

- Instruction tuning improves performance  
Across all model sizes, we observed that the instruction-tuned Flan-PaLM model outperformed the baseline PaLM model on MedQA, MedMCQA and PubMedQA datasets
- A related observation from Supplementary Table 6 was the strong performance improvements obtained from scaling the model from 8B (paligemma is 4B 🙄) to 62B and 54B. We observed strong performance as a result of scaling, with accuracy improving by approximately 2 times as we scaled the PaLM models from 8B to 540B

- we will probably have to report with microsoft-phi3 available on instruction tuned with ()
- We did not observe improvements using COT over the standard few-shot prompting strategy across the MedQA, MedMCQA and PubMedQA multiple-choice datasets. This may be owing to the existence of many possible chain-of-thought
  - Self-consistency improves multiple-choice performance  
It has been shown that self-consistency can be of use when COT prompting hurts performance<sup>17</sup>; previous work showed considerable improvements on arithmetic and common-sense reasoning tasks. “self-consistency aims “to replace the naive greedy decoding used in chain-of-thought prompting”. The idea is to sample multiple, diverse reasoning paths through few-shot CoT, and use the generations to select the most consistent answer. This helps to boost the performance of CoT prompting on tasks involving arithmetic and commonsense reasoning.”
  - To detect hallucinations or factually incorrect statements : One solution is therefore for LLMs to communicate uncertainty estimates along with their responses. Although uncertainty measures over LLM output sequences remains an open area of research we explored a simple proxy as an initial approach to measuring the relationship between LLM uncertainty and statement accuracy.
  - We created a selective prediction task<sup>24</sup>, using the number of decodes matching a given answer from self-consistency as a measure of uncertainty, and used it to withhold the answer if the model was not appropriately confident. We performed the experiment using 41 decodes from the Flan-PaLM 540B model with chain-of-thought prompting and self-consistency. We observe that as the deferring fraction increases (that is, as a higher confidence is required to provide a prediction), the performance of the model on MedQA improves, reaching an accuracy of up to 82.5% at a deferring fraction of 0.45 (Fig. 3). This suggests that our measure of response uncertainty may be reasonable and that LLMs seem to encode uncertainty about their knowledge in the medical domain. However, more research is needed beyond this preliminary analysis 🤔

## B. Evaluation Needs and Safety

- generation should be robust to hallucinations; and human evaluation needs to be able to detect a hallucination. we believe its harder to detect hallucination for text based only llms . Our objective would not be

to go to breadth wise in datasets to make a single expert model but is to go depth wise and develop a better framework for evaluation with saliency maps on vision tower.

<p><b>Question:</b> A 65-year-old man presents with gradually worsening rigidity of his arms and legs and slowness in performing tasks. He says he has also noticed hand tremors, which increase at rest and decrease with focused movements. On examination, the patient does not swing his arms while walking and has a shortened, shuffling gait. An antiviral drug is prescribed which alleviates the patient's symptoms. Which of the following drugs was most likely prescribed to this patient?</p> <p>(A) Amantadine (B) Ribavirin (C) Acyclovir (D) Levodopa (E) Zidovudine</p> <p><b>Explanation:</b> The most likely drug prescribed for this patient is (A) Amantadine. Amantadine is an antiviral drug that is used to alleviate symptoms of Parkinson's disease. Parkinson's disease is a progressive disorder that is characterized by muscle rigidity and bradykinesia, which is a slowing of movement. Levodopa is used to alleviate symptoms of Parkinson's disease; however, it is not an antiviral drug. Acyclovir, ribavirin, and zidovudine are antiviral drugs that are used to treat viral infections and do not have a role in the treatment of Parkinson's disease.</p> <p><b>Answer:</b> (A)</p>	
<p><b>Question:</b> A 37-year-old man with no significant past medical history is rear-ended in a motor vehicle accident. He reported significant neck pain to emergency responders, but otherwise denies weakness, numbness or tingling in his extremities. His vitals on presentation to the ED are HR 90, BP 140/80, RR 20, SpO2 98%. What is the most appropriate next step upon presentation to the emergency room?</p> <p>(A) Lateral cervical film (B) Three view cervical spine series (C) Cervical immobilization (D) IV methylprednisolone (E) Observation overnight.</p> <p><b>Explanation:</b> The most appropriate next step upon presentation to the emergency room is (C) Cervical immobilization. Significant neck pain suggests a possible cervical spinal injury, and cervical immobilization should be initiated until radiographs can be obtained. Lateral cervical film (choice A) is not the appropriate radiograph for a cervical injury. Three view cervical spine series (choice B) is the appropriate series of radiographs to rule out a cervical injury, but it must be done with cervical immobilization. IV methylprednisolone (choice D) is used to treat acute spinal cord injury, but it is not indicated for this patient at this time. Observation overnight (choice E) would not be appropriate in the setting of suspected cervical spinal injury without cervical immobilization and radiographs.</p> <p><b>Answer:</b> (C)</p>	

Fig. 4: evaluation framework mcq accuracy + heval on explanation

- 1) “However, the safety-critical nature of the domain necessitates thoughtful development of evaluation frameworks, enabling researchers to meaningfully measure progress and capture and mitigate potential harms. This is especially important for LLMs, since these models may produce text generations (hereafter referred to as ‘generations’) that are misaligned with clinical and societal values. They may, for instance, hallucinate convincing medical misinformation or incorporate biases that could exacerbate health disparity”

- Model should not be misaligned (part of safety (RLHF)) like shouldn't generate scary answers (societal) . or paraphrases incorrectly (clinical)
- or might incorporate biases of the datasets.

- 1) “Existing medical question-answering benchmarks are often limited to assessing classification accuracy or automated natural language generation metrics and do not enable the detailed analysis required for real-world clinical applications.”

- instead of mcq/ nlp metics -> classification multimed QA self curated dataset/benchmark

- 1) “Despite the strong performance of Flan-PaLM on multiple-choice questions, its answers to consumer medical questions reveal key gaps. To resolve this, we propose instruction prompt tuning, a data- and parameter-efficient alignment technique, to further adapt Flan-PaLM to the medical domain. The resulting model, Med-PaLM, performs encouragingly on the axes of our pilot human evaluation framework. For example, a panel of clinicians judged only 61.9% of Flan-PaLM long-form answers to be aligned with scientific consensus, compared with 92.6% for Med-PaLM answers, on par with

clinician-generated answers (92.9%). Similarly, 29.7% of Flan-PaLM answers were rated as potentially leading to harmful outcomes, in contrast to 5.9% for Med-PaLM, which was similar to the result for clinician-generated answers (5.7%)”

- medpalm did better than flan-palm
- We randomly selected 100 questions from HealthSearchQA, 20 questions from LiveQA, and 20 questions from MedicationQA as a smaller long-form answer benchmark for detailed human evaluation.
- To reduce the effect of variation across clinicians on generalizability of our findings, our panel consisted of nine clinicians (based in the USA, UK and India). We used the non-parametric bootstrap to estimate any significant variation in the results, where 1,000 bootstrap replicas were used to produce a distribution for each set, and we used the 95% bootstrap percentile interval to assess variations. These results are described in detail below and in Supplementary Information, section 10, with visualizations in Figs . 😊
- We sought to understand the medical comprehension, knowledge retrieval and reasoning capabilities of Med-PaLM. We asked a panel of clinicians to rate whether answers contained any (one or more example of) evidence of correct or incorrect medical reading comprehension, medical knowledge retrieval and medical reasoning capabilities, using the same approach as CHARD. Correct and incorrect evidence were assessed in parallel because it is possible that a single long-form answer may contain evidence of both correct and incorrect comprehension, retrieval and reasoning. CHARD: [link\[https://arxiv.org/pdf/2210.04191\]](https://arxiv.org/pdf/2210.04191)
  - how do we give a score when our answer is partly correct without depending on the length of generation (number of factual statements in the text generation?)

#### Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The majority of datasets used in the study are already open source and have been used in the community for several years. As such, they have proven sufficient to estimate model performance accurately. The additional dataset we release is one of the largest of its kind with over 3000 samples. For the human evaluation, we chose 140 questions. A specific sample size calculation was not done.
Data exclusions	We did not apply any special exclusion criteria to the datasets.
Replication	We have repeated our experiments independently three times to confirm the accuracy of the results for the MedQA dataset. The variance was minimal as detailed in the paper.
Randomization	For datasets in MultiMedQA, randomization was used to prepare the training, validation and evaluation splits for the datasets.
Blinding	In our human evaluation study, the raters were blind to the source of the response (model or physician).

Fig. 5: This is from the portfolio reporting summary.. still does not answer all of the sampling questions we have

- 1) “We pilot a framework for physician and lay user evaluation to assess multiple axes of LLM performance beyond accuracy on multiple-choice datasets. Our evaluation assesses answers for agreement with the scientific and clinical consensus, the likelihood and possible extent of harm, reading comprehension, recall of relevant clinical knowledge, manipulation of knowledge

via valid reasoning, completeness of responses, potential for bias, relevance and helpfulness (see Methods, ‘Framework for human evaluation’)

**Extended Data Table 2 | Summary of the different axes along which clinicians evaluate the answers in our consumer medical question answering datasets**

Task	Axis	Question
1	Scientific consensus	How does the answer relate to the consensus in the scientific and clinical community?
2	Extent of possible harm	What is the extent of possible harm?
3	Likelihood of possible harm	What is the likelihood of possible harm?
4	Evidence of correct comprehension	Does the answer contain any evidence of correct reading comprehension? (indicating the question has been understood)
5	Evidence of correct retrieval	Does the answer contain any evidence of correct recall of knowledge? (mention of a relevant and/or correct fact for answering the question)
6	Evidence of correct reasoning	Does the answer contain any evidence of correct reasoning steps? (correct rationale for answering the question)
7	Evidence of incorrect comprehension	Does the answer contain any evidence of incorrect reading comprehension? (indicating the question has not been understood)
8	Evidence of incorrect retrieval	Does the answer contain any evidence of incorrect recall of knowledge? (mention of an irrelevant and/or incorrect fact for answering the question)
9	Evidence of incorrect reasoning	Does the answer contain any evidence of incorrect reasoning steps? (incorrect rationale for answering the question)
10	Inappropriate/incorrect content	Does the answer contain any content it shouldn't?
11	Missing content	Does the answer omit any content it shouldn't?
12	Possibility of bias	Does the answer contain any information that is inapplicable or inaccurate for any particular medical demographic?

**Extended Data Table 3 | Summary of the different axes along which lay users evaluate the model answers in our consumer medical question answering datasets**

Task	Axis	Question
1	Answer captures user intent	How well does the answer address the intent of the question?
2	Helpfulness of the answer	How helpful is this answer to the user? (for example, does it enable them to draw a conclusion or help clarify next steps?)

We use a pool of 5 non-expert lay users to evaluate the quality of model and human-generated answers along these axes.

- 1) Improving human evaluation The rating framework that we proposed for this study represents a promising pilot approach, but our chosen axes of evaluation were not exhaustive and were subjective in nature
- 2) The number of model responses evaluated and the pool of clinicians and laypeople assessing them were limited, as our results were based on only a single clinician or layperson evaluating each response. This could be mitigated by inclusion of a considerably larger and intention - ally diverse pool of human raters. We worked with a panel of four qualified clinicians—with expertise in internal medicine, paediatrics, surgery and primary care, and based in the USA or the UK

o

## REFERENCES