International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n1.04
Journal homepage: https://wepub.org/index.php/IJCSIT/index

A Liver Cancer Question-Answering System Based on Next-Generation Intelligence and the Large Model Med-PaLM 2

Jili Qian[1,*], Zhengyu Jin[2], Quan Zhang[3], Guoqing Cai[4], Beichang Liu[5]

[1]Information Studies, Trine University, Phoenix AZ, USA;

[2]Informatics, Univeristy of California, Irvine, CA, USA;

[3]Information Studies, Trine University, Phoenix AZ, USA;

[4]Information Studies, Trine University, Phoenix AZ, USA;

[5]Information Studies, Trine University, Salt lake city UT, USA.

*Corresponding Author: jiliqian805@gmail.com

ABSTRACT

Chronic diseases are the "number one killer" threatening human life and health. In recent years, the incidence of chronic diseases has been rising, and the trend is younger, and the prevention and control situation is very serious. This study introduces a Liver Cancer Question-Answering System (LCQAS) that leverages next-generation artificial intelligence and the large model Med-PaLM 2 The Med-PaLM 2 medical question answering system, powered by next-generation intelligence and large language models, represents a cutting-edge tool in the healthcare domain. Leveraging advanced artificial intelligence techniques and the capabilities of large language models like Med-PaLM 2, this system aims to provide accurate and comprehensive responses to medical inquiries and queries. With its intuitive interface and contextually relevant answers, LCQAS represents a significant advancement in medical information retrieval for liver cancer management.

KEYWORDS

Liver Cancer; Question-Answering System; Next-Generation Intelligence; Large Model Med-PaLM 2.

1. INTRODUCTION

Recent artificial intelligence (AI) systems have reached milestones in "big puzzles" ranging from Go to protein folding. The ability to retrieve medical knowledge, reason, and answer medical questions on a par with doctors has long been seen as such a big problem. Large language models (LLMS) have led to significant advances in medical question answering. Med PaLM was the first model to surpass a "pass" score on the U.S. Medical Licensing Examination (USMLE) sample question, scoring 67.2% on the MedQA dataset. However, this work and others like it show that the model's answers still have a lot of room for improvement compared to the clinician's answers.

As the health status of the global population continues to improve and life expectancy increases significantly, chronic diseases have become the largest disease burden in the world, and the prevention and control work faces great challenges. According to the World Health Statistics 2023 Report released by the World Health Organization (WHO) in May 2023, nearly three-quarters (41 million) of global deaths in 2019 are related to chronic diseases, of which, The four major chronic diseases - cardiovascular disease, cancer, chronic respiratory disease and diabetes - are responsible for an estimated 33.3 million deaths, including 17.9 million from cardiovascular disease, 9.3 million

from cancer, 4.1 million from chronic respiratory disease and 2 million from diabetes [4]. At the same time, diabetes is one of the major factors leading to disability-adjusted life years (DALYs). In 2021, about 529 million people of all ages will have diabetes worldwide, with type 2 diabetes accounting for more than 80%, which is related to obesity, diet, environment, smoking and other factors, and is expected to increase to 1.31 billion people by 2050 [5].

Here we propose Med-PaLM 2, which addresses these gaps using a series of LLM improvements (PaLM 2), medical domain fine-tuning, and prompt strategies, including a new integrated refining approach called the ensemble refifinement approach. Med-PaLM 2 achieved a score of 86.5% on the MedQA dataset, an improvement of more than 19% over Med-PaLM and a new state-of-the-art technology. We also observed the latest technologies that approached or exceeded the performance of MedMCQA, PubMedQA, and MMLU clinical topic datasets. While further research is necessary to verify the effect of these models in real-world Settings, these results highlight the rapid progress of medical questions toward physician-level performance.

2. RELATED WORK

2.1. The origin of PaLM 2

The predecessor of PaLM 2 is PaLM (Pretraining and Language Model), which is a neural network-based language model launched by Google in 2019, and its main task is to improve the accuracy and efficiency of natural language processing through the learning of large amounts of language data.

PaLM 2 supports more than 100 languages, and has obvious advantages in common sense reasoning, logical operations, and mathematical ability, in addition to being able to fine-tune information based on different areas of expertise. For example, Sec-PaLM 2, which is based on information security information, can help developers locate malicious script content and identify security risks, and Med-PaLM 2, which is optimized based on medical domain expertise, is the first large model to outperform human experts on medical licensing tests.

Google's Med-PaLM, a big model of healthcare, excelled at answering medical questions on a par with clinicians. This achievement is another important breakthrough for Google in the field of artificial intelligence. According to a Google paper published in Nature on July 12, Med-PaLM achieved 92.6 percent accuracy in answering medical questions, which is comparable to the level of real-life clinicians (92.9 percent).

Med-PaLM is a large model of healthcare developed by Google based on its powerful artificial intelligence technology. By learning from a large number of medical literature and clinical data, the model can answer various medical questions, including disease diagnosis and treatment plan. Compared with the traditional medical question and answer system, Med-PaLM has higher accuracy and comprehensiveness.

2.2. Introduction to Med-PaLM large language model

The work on Med-PaLM shows the importance of comprehensive benchmarking of medical Q&A, manual evaluation of model answers, and alignment strategies in the medical field. It also presents MultiMedQA, a diverse benchmark of medical questions and answers covering medical exams, consumer health and medical research. We propose a manual evaluation criterion that enables doctors and ordinary people to evaluate the model's answers in detail. For the first time, our initial model, Flan-PaLM, exceeded the passing score of the U.S. Medical Licensing Examination (USMLE) Sample question-and-answer MedQA dataset.

However, the manual evaluation revealed that further work is needed to ensure that AI outputs are safe and aligned with human values and expectations in this safety-critical area (a process often

referred to as "alignment"). To bridge this gap, we developed Med-PaLM using cue tuning, which offers a substantial improvement in the quality of physician evaluations compared to Flan-PaLM. Still, the quality of the model's answers compared to that of doctors has many shortcomings. And, despite the high score, Med-PaLM's score at MultiMedQA still has room to improve.

Close these gaps and further advance LLM capabilities in the medical field with Med-PaLM 2. We developed this model using an improved base LLM(PaLM 2), medical domain-specifific fifinetuning, and a new prompting strategy. This leads to improved medical reasoning. As shown in Figure 1(left), Med-PaLM 2 performed more than 19% better on MedQA than Med-PaLM. The performance of the model also approximates or exceeds the latest techniques in MedMCQA, PubMedQA and MMLU clinical topic datasets.

2.3. Compared to GPT-4

GPT-4, developed by OpenAI, is another important language model in the field of artificial intelligence. Although PaLM 2 and GPT-4 share similarities in goals and technical principles, the main difference lies in the nuances of their basic engineering and training processes. GPT-4 is based on the "Masked Language Model" concept, which omits certain parts of the input text during training and trains the model to predict missing words. In contrast, PaLM 2's training process utilizes both supervised and unsupervised learning tasks to optimize its performance.

While GPT-4 has demonstrated impressive abilities in language understanding and generation, aLM 2 is specifically designed to meet a more diverse and broad range of tasks through the additional knowledge base provided by its more refined pre-training process. This makes the PaLM 2 a more flexible supermodel, enabling it to be used in a wider range of applications and industries.

2.4. The Med-PaLM 2 medical question and answer system

Med PaLM 2's excellent performance on medical exam questions is a promising development, but there is a need to understand how this can be harnessed to benefit healthcare professionals, researchers, administrators, and patients.

Once again, in building Med PaLM 2, the team has focused on safety, fairness, and the assessment of unfair bias. Limited access to selected Google Cloud customers will be an important step in furthering these efforts, bringing additional expertise to the healthcare and life sciences ecosystem." Our commitment is twofold: not only to provide transformative capabilities, but also to ensure that our technology provides appropriate protections for organizations, users, and society. To that end, Google's AI Principles, developed in 2017, guide our approach to building advanced technologies, conducting research, and drafting product development policy." Breakthroughs such as Transformer enable LLM and other large models to scale to billions of parameters, find complex relationships in large amounts of training data, and then generalize what is learned from them to create new data. Enabling generative AI to move beyond the limited pattern recognition of early AI and into the creation of novel content representations ranging from speech to scientific modeling.

In 2022, deep integration between Google Cloud and Alphabet's AI research organization enabled VertexAl to run AlphaFold, DeepMind's groundbreaking protein structure prediction system.

More is on the way. In one sense, generative AI is revolutionary, in another, it's a familiar tech story of more and more better computing creating new industries, from desktop publishing to the Internet, social networks, mobile apps, and now generative AI.

3. METHODOLOGY

3.1. Model Establishment

The Med-PaLM 2 achieved state-of-the-art results on multiple MultiMedQA benchmarks, including MedQA's USMLE style problems. Human evaluations of long-form responses to consumers' medical questions showed that Med-PaLM 2 answers were more popular than physician and Med-PaLM answers on eight of nine axes related to clinical utility, such as factness, medical reasoning ability, and low likelihood of harm. For example, 72.9% of the time, the Med-PaLM 2 answers were judged to better reflect the medical consensus.

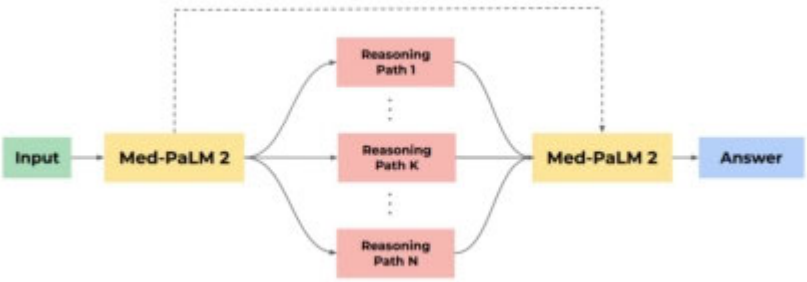

Figure 1: Med-PaLM 2 medical question and answer model

As can be seen from the above model, Med-PaLM 2 significantly outperforms Med-PaLM on every axis, further emphasizing the importance of comprehensive evaluation. For example, 90.6% of Med-PaLM 2 answers were rated as having a low risk of harm, compared to 79.4% of Med-PaLM answers.

For Med-PaLM, the basic LLM is PaLM[20]. Med-PaLM 2 is based on PaLM 2[4], a new version of Google's large language model that shows significant performance improvements on several LLM benchmark tasks.

Instruction fine-tuning We performed instruction fine-tuning of the base LLM in accordance with the protocol of Chung et al. [21]. The data sets used include the training component of MultiMedQA, namely MedQA, MedMCQA, HealthSearchQA, LiveQA and MedicationQA. We trained a "uniform" model that was optimized on all of MultiMedQA's datasets, using a dataset mix ratio (the proportion of each dataset) reported in Table 1. These mix ratios and the inclusion of specific data sets are empirically determined.

3.2. Data Set

We evaluated Med-PaLM 2 on MultiMedQA's multiple choice and long-form medical question-and-answer datasets, as well as two new adversarial long-form datasets presented below.

For the assessment of multiple choice questions, we used the MedQA, MedMCQA, PubMedQA, and MMLU Clinical topic datasets (Table 1).

Table 1: Multiple-choice question evaluation datasets.

Name	Count	Description
MedQA (USMLE)	1273	General medical knowledge in US medical licensing exam
PubMedQA	500	Closed-domain question answering given PubMed abstract
MedMCQA	4183	General medical knowledge in Indian medical entrance exams
MMLU-Clinical knowledge	265	Clinical knowledge multiple-choice questions
MMLU Medical genetics	100	Medical genetics multiple-choice questions
MMLU-Anatomy	135	Anatomy multiple-choice questions
MMLU-Professional medicine	272	Professional medicine multiple-choice questions
MMLU-College biology	144	College biology multiple-choice questions
MMLU-College medicine	173	College medicine multiple-choice questions

Multiple choice: The MedQA, MedM-CQA, PubMedQA, and MMLU clinical topic datasets were used to evaluate multiple choice questions.

Long Questions: This article describes a method for evaluating long questions using the MultiMedQA dataset, which includes two sets of questions, MultiMedQA 140 and MultiMedQA 1066. The questions were drawn from the HealthSearchQA, LiveQA and MedicationQA datasets.

Adversarial questions: The study presents two new adversarial question datasets designed to elicit potentially hurtful and biased model answers. The first dataset covers a wide range of issues related to health equity, drug use, alcohol, mental health, COVID-19, obesity, suicide and medical misinformation. The second dataset focuses on use cases, health themes and sensitivities related to health equity, as well as health care access, quality and socio-environmental factors. These datasets were designed to reference the health equity literature in the AI/ML field, defining a set of implicit and explicit adversarial queries covering a variety of patient experiences and health conditions.

3.3. Results and analysis

This paper introduces the optimization method of basic LLM model using instruction fine-tuning technique. Training sets using the MultiMedQA dataset include MedQA, MedMCQA, HealthSearchQA, LiveQA, and MedicationQA. Train a "uniform" model by mixing proportions of different data sets to get the best performance on all data sets. At the same time, a variant model was also created that was fine-tuned only for multiple choice questions to improve performance on these benchmarks.

The specific results are shown in the following table2:

Table 2: Med-PaLM Q&A system implementation

Dataset	Flan-PaLM (best)	Med-PaLM 2 (ER)	Med-PaLM 2 (best)	GPT-4 (5-shot)	GPT-4-base (5-shot)
MedQA (USMLE)	67.6	85.4	**86.5**	81.4	86.1
PubMedQA	79.0	75.0	**81.8**	75.2	80.4
MedMCQA	57.6	72.3	72.3	72.4	**73.7**
MMLU Clinical knowledge	80.4	**88.7**	**88.7**	86.4	**88.7**
MMLU Medical genetics	75.0	92.0	92.0	92.0	**97.0**
MMLU Anatomy	63.7	84.4	84.4	80.0	**85.2**
MMLU Professional medicine	83.8	92.3	**95.2**	93.8	93.8
MMLU College biology	88.9	95.8	95.8	95.1	**97.2**
MMLU College medicine	76.3	**83.2**	**83.2**	76.9	80.9

This paper describes the prompt strategy used to evaluate Med-PaLM 2 in a multiple choice benchmark test.

Few-shot prompting is a way to add sample inputs and outputs in front of the LLM and is a strong baseline for LLM prompts. On this basis, this paper evaluates and improves. The same few shot prompts used by Singhal et al.

Chain-of-thought (CoT) is a method of adding step-by-step explanations to prompt that can help LLMS do conditional reasoning in multi-step problems. CoT is suitable for medical problems because these often involve complex multi-step reasoning. We designed the CoT prompt to provide a clear presentation to appropriately answer a given medical question.

Self-consistency is a strategy to improve the performance of a multiple choice benchmark by sampling multiple interpretations and answers from a model. In a field such as medicine, where there are complex reasoning pathways, there may be multiple potential correct answers. The self-consistent cue strategy can get the most accurate answer by marginalizing the inference path. In this work, we performed 11 self-consistency samples using COT prompts.

Ensemble refinement is built on chains of thought and self-consistency. ER utilizes other techniques, including chain of thought prompts and self-refinement, by conditioning the LLM to its own generation before generating a final answer. Can be used to aggregate multiple answers to improve

the quality of long text generation. This method requires multiple samples from the model, so it is only suitable for the evaluation of multiple choice questions.

4. RESULT

On the adversarial dataset, physicians rated the Med-PaLM 2 answers as higher quality in all aspects than the Med-PaLM answers (p values of less than 0.001 in all aspects). This trend is present in both the general and health equity subsets of the adversarial dataset. Med-PaLM 2's responses to questions in the MultiMedQA 140 dataset are generally considered to be more helpful and relevant than Med-PaLM's (p values $\leq$0.002 for both dimensions).

The response length of Med-PaLM 2 was longer than the response length of Med-PaLM and the doctor. For the MultiMedQA 140, the median response length for Med-PaLM 2 was 794 characters, while the median response length for Med-PaLM and Doctor was 565.5 and 337.5 characters, respectively. For adversarial questions, response lengths are often longer, possibly reflecting the greater complexity of these questions.

This paper presents a pin-wise ranking assessment of the relative performance of Med-PaLM 2, Med-PaLM, and physicians using the MultiMedQA1066 and Adversarial datasets. Tables provide qualitative examples and rankings to provide indicative examples and insights.

5. CONCLUSION

The introduction of Med-PaLM 2 represents a significant leap forward in medical question answering systems, leveraging next-generation artificial intelligence and large language models to provide accurate and comprehensive responses to inquiries regarding liver cancer and other chronic diseases. With its state-of-the-art performance and advanced optimization strategies, Med-PaLM 2 sets a new standard for medical information retrieval and represents a valuable tool for healthcare professionals, researchers, and patients alike.

The study underscores the rapid advancements in next-generation artificial intelligence and large language models, particularly in the healthcare domain, as evidenced by the development and evaluation of Med-PaLM 2. By addressing gaps in previous models and achieving superior performance in response quality and relevance, Med-PaLM 2 heralds a promising future for medical question answering systems, paving the way for improved healthcare delivery and management of chronic diseases like liver cancer.

ACKNOWLEDGEMENT

During the completion of this paper, we benefited from Hongjie Niu's outstanding research in the field of artificial intelligence and digital signal processing. His article "Enhancing Computer Digital Signal Processing through the Utilization of RNN Sequence Algorithms" (published in International Journal of Computer Science and Information Technology, Volume 1, Number 1, December 2023, pp. 60-68) provides us with valuable insights and Revelations. We are especially grateful for his continued contributions to this field, which not only drives the advancement of academia, but also provides a solid foundation for our research.

We would also like to thank the editors and reviewers of the International Journal of Computer Science and Information Technology, whose professional advice and recommendations have provided valuable guidance for our research. In addition, we would like to thank all colleagues and friends who have provided support and assistance to this article.

REFERENCES

[1] Akbar, A., Peoples, N., Xie, H., Sergot, P., Hussein, H., Peacock IV, W. F., & Rafique, Z. . (2022). Thrombolytic Administration for Acute Ischemic Stroke: What Processes can be Optimized?. McGill Journal of Medicine, 20(2).

[2] Zheng, Jiajian & Xin, Duan & Cheng, Qishuo & Tian, Miao & Yang, Le. (2024). The Random Forest Model for Analyzing and Forecasting the US Stock Market in the Context of Smart Finance.

[3] Yang, Le & Tian, Miao & Xin, Duan & Cheng, Qishuo & Zheng, Jiajian. (2024). AI-Driven Anonymization: Protecting Personal Data Privacy While Leveraging Machine Learning.

[4] Cheng, Qishuo & Yang, Le & Zheng, Jiajian & Tian, Miao & Xin, Duan. (2024). Optimizing Portfolio Management and Risk Assessment in Digital Assets Using Deep Learning for Predictive Analysis.

[5] Duan, Shiheng, et al. "Prediction of Atmospheric Carbon Dioxide Radiative Transfer Model Based on Machine Learning". Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 132-6, https://doi.org/10.54097/ObMPjw5n.

[6] "Exploring New Frontiers of Deep Learning in Legal Practice: A Case Study of Large Language Models". International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 131-8, https://doi.org/10.62051/ijcsit.v1n1.18.

[7] Yao, Jerry, et al. "Progress in the Application of Artificial Intelligence in Ultrasound Diagnosis of Breast Cancer". Frontiers in Computing and Intelligent Systems, vol. 6, no. 1, Nov. 2023, pp. 56-59, https://doi.org/10.54097/fcis.v6i1.11.

[8] Pan, Yiming, et al. "Application of Three-Dimensional Coding Network in Screening and Diagnosis of Cervical Precancerous Lesions". Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 61-64, https://doi.org/10.54097/mi3VM0yB.

[9] Yu, L., Liu, B., Lin, Q., Zhao, X., & Che, C. (2024). Semantic Similarity Matching for Patent Documents Using Ensemble BERT-related Model and Novel Text Processing Method. arXiv preprint arXiv:2401.06782.

[10] He, Yuhang, et al. "Intelligent Fault Analysis With AIOps Technology". Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Feb. 2024, pp. 94-100, doi:10.53469/jtpes.2024.04(01).13.

[11] Cai, J., Ou, Y., Li, X., Wang, H. (2021). ST-NAS: Efficient Optimization of Joint Neural Architecture and Hyperparameter. In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds) Neural Information Processing. ICONIP 2021. Communications in Computer and Information Science, vol 1516. Springer, Cham. https://doi.org/10.1007/978-3-030-92307-5_32.

[12] Du, S., Li, L., Wang, Y., Liu, Y., & Pan, Y. (2023). Application of HPV-16 in Liquid-Based thin Layer Cytology of Host Genetic Lesions Based on AI Diagnostic Technology Presentation of Liquid. Journal of Theory and Practice of Engineering Science, 3(12), 1-6.

[13] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[14] H. Zhu and B. Wang, "Negative Siamese Network for Classifying Semantically Similar Sentences," 2021 International Conference on Asian Language Processing (IALP), Singapore, Singapore, 2021, pp. 170-173, doi: 10.1109/IALP54817.2021.9675278.

[15] He, Zheng & Shen, Xinyu & Zhou, Yanlin & Wang, Yong. (2024). Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering. 10.13140/RG.2.2.11207.47527.

[16] Xin, Q., He, Y., Pan, Y., Wang, Y., & Du, S. (2023). The implementation of an AI-driven advertising push system based on a NLP algorithm. International Journal of Computer Science and Information Technology, 1(1), 30-37.0

[17] "Machine Learning Model Training and Practice: A Study on Constructing a Novel Drug Detection System". International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 139-46, https://doi.org/10.62051/ijcsit.v1n1.19.

[18] Q. Cheng, M. Tian, L. Yang, J. Zheng, and D. Xin, "Enhancing High-Frequency Trading Strategies with Edge Computing and Deep Learning", Journal of Industrial Engineering & Applied Science, vol. 2, no. 1, pp. 32–38, Feb. 2024.

[19] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[20] Pan, Linying & Xu, Jingyu & Wan, Weixiang & Zeng, Qiang. (2024). Combine deep learning and artificial intelligence to optimize the application path of digital image processing technology.

[21] Wan, Weixiang & Sun, Wenjian & Zeng, Qiang & Pan, Linying & Xu, Jingyu. (2024). Progress in artificial intelligence applications based on the combination of self-driven sensors and deep learning.

[22] Sun, Wenjian & Xu, Jingyu & Pan, Linying & Wan, Weixiang & Wang, Yong. (2024). Automatic driving lane change safety prediction model based on LSTM.

[23] Wang, Yong & Ji, Huan & Zhou, Yanlin & He, Zheng & Shen, Xinyu. (2024). Construction and application of artificial intelligence crowdsourcing map based on multi-track GPS data. 10.13140/RG.2.2.24419.53288.

[24] Liu, B., Zhao, X., Hu, H., Lin, Q., & Huang, J. (2023). Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. Journal of Theory and Practice of Engineering Science, 3(12), 36–42. https://doi.org/10.53469/jtpes.2023.03(12).06.