An Early Investigation into the Utility of Multimodal Large Language Models in Medical Imaging

Sulaiman Khan*, Md. Rafiul Biswas*, Alina Murad†, Hazrat Ali‡, Zubair Shah*

*College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar.
†Foundation University School of Sciences and Technology, Rawalpindi, Pakistan.
‡Faculty of Computing and Information Technology, Sohar University, Sohar, Oman.
Email: sukh45452@hbku.edu.qa, mdbi30331@hbku.edu.qa, alina.murad@fui.edu.pk, hazrat.ali@live.com, zshah@hbku.edu.qa

Abstract—Recent developments in multimodal large language models (MLLMs) have spurred significant interest in their potential applications across various medical imaging domains. On the one hand, there is a temptation to use these generative models to synthesize realistic-looking medical image data, while on the other hand, the ability to identify synthetic image data in a pool of data is also significantly important. In this study, we explore the potential of the Gemini (*gemini-1.0-pro-vision-latest*) and GPT-4V (*gpt-4-vision-preview*) models for medical image analysis using two modalities of medical image data. Utilizing synthetic and real imaging data, both Gemini AI and GPT-4V are first used to classify real versus synthetic images, followed by an interpretation and analysis of the input images. Experimental results demonstrate that both Gemini and GPT-4 could perform some interpretation of the input images. In this specific experiment, Gemini was able to perform slightly better than the GPT-4V on the classification task. In contrast, responses associated with GPT-4V were mostly generic in nature. Our early investigation presented in this work provides insights into the potential of MLLMs to assist with the classification and interpretation of retinal fundoscopy and lung X-ray images. We also identify key limitations associated with the early investigation study on MLLMs for specialized tasks in medical image analysis.

Index Terms—LLM, ChatGPT, Gemini AI, Multimodal data, Retina, Lung

I. INTRODUCTION

With the public introduction of large language models (LLMs)-enabled tools, there has been an unprecedented surge of interest to explore and use these tools. LLMs, designed for comprehending and generating human-like text, truly revolutionized various domains, including content creation, document summarization, and information retrieval. Among these, ChatGPT, powered by the Generative Pre-trained Transformer (GPT) engine, emerged as the most prominent tool, and attracted over a 100 million unique users within two months of its release [1], [2], [3].

LLMs were primarily text-based, lacking the ability to process other modalities. While leveraging LLMs in textual contexts holds significant implications for medical documentation, healthcare practices heavily rely on visual (medical imaging) and multimodal data. In this context, the recent advancements in Multimodal LLMs (MLLMs)-based tools such as Gemini and ChatGPT (GPT-4V) may result in promising solutions in medical image analysis. While both Gemini and ChatGPT tools have been widely explored for natural image interpretation and image-to-text or text-to-image applications in computer vision, their existing capabilities in medical image interpretation have not been explored. This paper focuses on applications of MLLMs for medical image analysis applications. More specifically, this work explores Gemini and GPT-4V MLLMs for the classification and interpretation of medical images, especially for lung X-ray images (pulmonary disease) and retinal fundoscopy images (ophthalmology), which makes this work an early study on MLLMs for medical image analysis. Following are the insights of this research work:

- **Image classification:** We explore pre-trained MLLMs to handle the identification of synthetic lung X-ray and retinal fundoscopy images.
- **Disease diagnosing:** We explore pre-trained MLLMs to interpret medical images and possibly identify disease-related symptoms in the input images. This could prove to be a bridge between healthcare professionals and the utilization of AI for medical purposes.
- **Prompt formulation:** We adopt the NERIF (Notation-Enhanced Rubric Instruction for Few-shot Learning) method, proposed by Lee and Zhai [4], to formulate an effective input prompt that effectively align with our intended application of the MLLM.
- **Comparative analysis:** We present a comparative analysis of the performance achieved by Gemini and GPT-4V and also include subjective input from a trained practicing doctor in *Manchester University NHS Foundation Trust* to help the reader get an estimate of the MLLM's performance. Even though the findings are non-binding, based on our limited experimentation, Gemini demonstrates better image understanding performance than GPT-4V.

The rest of the paper is organized as follows: Section

II briefly explains the architectural and operational differences between Gemini and GPT-4V. We then elaborate our methodology in Section III, which outlines the data sources used in this work and the prompts formulated to perform classification and interpretation using MLLMs. We present the results and the critical findings in Section IV. We also identify the fundamental limitations of this work in the discussion. Finally, we provide valuable recommendations and conclude the paper in Section V.

II. GEMINI VERSUS GPT-4V

The development of MLLMs represents a significant advancement in Artificial Intelligence (AI), by incorporating multi-sensory capabilities to enhance general intelligence, thereby facilitating more natural human-computer interactions [5]. MLLMs operate within a shared embedding space, enabling the integration and processing of diverse data modalities such as text, image, audio, video, and 3D content. The latest MLLMs, such as GPT-4 with Vision (*also known as GPT-4V or gpt-4-vision-preview in the API*) and Google DeepMind's Gemini pro-vision or simply Gemini AI, can accept inputs of images and/or text to perform a wide range of language, vision, and vision-language tasks. These tasks include language translation and coding [6], image recognition [7], object localization [5], visual question answering [8], and visual dialogue [9]. Moreover, within the realm of AI for healthcare, both Gemini and GPT-4V find applications in image analysis, such as ophthalmology [10], pulmonary disease diagnosing and interpretations [11], where data often comes in visual formats like medical images/scans. A comparison of the architecture, training approach, training datasets, performance metrics, capabilities, and applications of these state-of-the-art models can shed light on their usability and advantages, particularly within healthcare domains.

GPT-4V is popular for its proficiency in handling intricate language and image-processing tasks and for its adeptness in understanding complex contexts [12]. On the other hand, Gemini, a tool developed by Google DeepMind, represents a large multimodal model that extends beyond text and images to encompass inputs from audio and video sources. This broader spectrum of input types indicates a more adaptable and comprehensive approach to multimodal learning from complex medical data [13]. Gemini undergoes training using Google's tensor processing unit (TPU) [14], renowned for its high computational capabilities. With an extended 32K context length, Gemini can process significantly larger data chunks simultaneously, potentially facilitating deeper insights and understanding in complex tasks [13]. TPUs are trained to accommodate a 32k context length and utilize efficient attention mechanisms such as multi-query attention [15].

GPT-4V has been trained extensively on a large dataset that includes online image data, as well as various textual sources like books, journals, code, and other text data [12]. Through rigorous training, GPT-4V has attained the capability to process and generate both language and images that closely resemble human-created content. As a generative model, GPT-4V can produce new content in response to both image and textual inputs. Leveraging its dual-transformer design and built-in image decoding modules, it efficiently analyzes and comprehends images by harnessing vast amounts of textual data. This enables it to excel in tasks such as translating languages depicted in images, facilitating cross-language communication through images, and generating various forms of artistic content. Additionally, GPT-4V can provide explainable responses to user queries.

On the other hand, Gemini model is trained on a dataset that is both multimodal and multilingual. This dataset encompasses data from web documents, books, and code, including image, audio, and video data. Utilizing the SentencePiece tokenizer [16], Gemini refines its tokenizer by training it on a large subset of the entire training corpus, resulting in an enhanced vocabulary and improved model performance. Senkaiahliyan et al. [17] explored GPT-4V's potential in clinical education by assessing medical image interpretation. Similarly, Xu et al. [18] evaluated GPT-4V's performance in ophthalmology studies and reported a 63% accuracy rate in diagnosing ocular images. Moreover, for automatic scoring tasks involving problem images and textual context with rubrics, GPT-4V achieved a 51% accuracy rate for science-based assessments [4]. Similarly, Gemini integration into products like Google Bard and Pixel enhances reasoning, planning, and writing capabilities. Its availability through the Gemini API (Google AI studio) makes it a valuable resource for developers and enterprise customers, showcasing its potential applicability in educational technology [13].

Overall, both GPT-4V and Gemini represent significant strides in the field of AI and language models. While they share certain multimodal capabilities, differences in their architecture, training methodologies, performance, and applications highlight the diverse and evolving landscape of AI technology. Thus, an early investigation into their utilization in the medical imaging domain will provide a valuable reference study to achieve improvements and develop useful MLLM-based Medical AI technology.

III. METHODOLOGY

A. Data Sources

In this work, we use two medical image modalities: lung X-ray images and retinal fundoscopy images. For both modalities, we use real and synthetic samples. For synthetic lung X-ray images, this study reanalyzes the synthetic images generated by Ali et al. [19] using neural diffusion models. In the work, they employed a pre-trained DALLE2 model to produce lung X-ray and CT images using a simple input text prompt. Furthermore, they trained a stable diffusion model to generate lung X-ray images. For real lung X-ray images, we utilize samples from the Kaggle dataset comprising 5,863 lung X-ray images representing samples for *Pneumonia* and *Normal* cases [20]. For fundoscopy images, we utilize images of the DRIVE dataset of retinal fundoscopy images [21]. This dataset consists of retinopathy scans accumulated from 400 diabetic
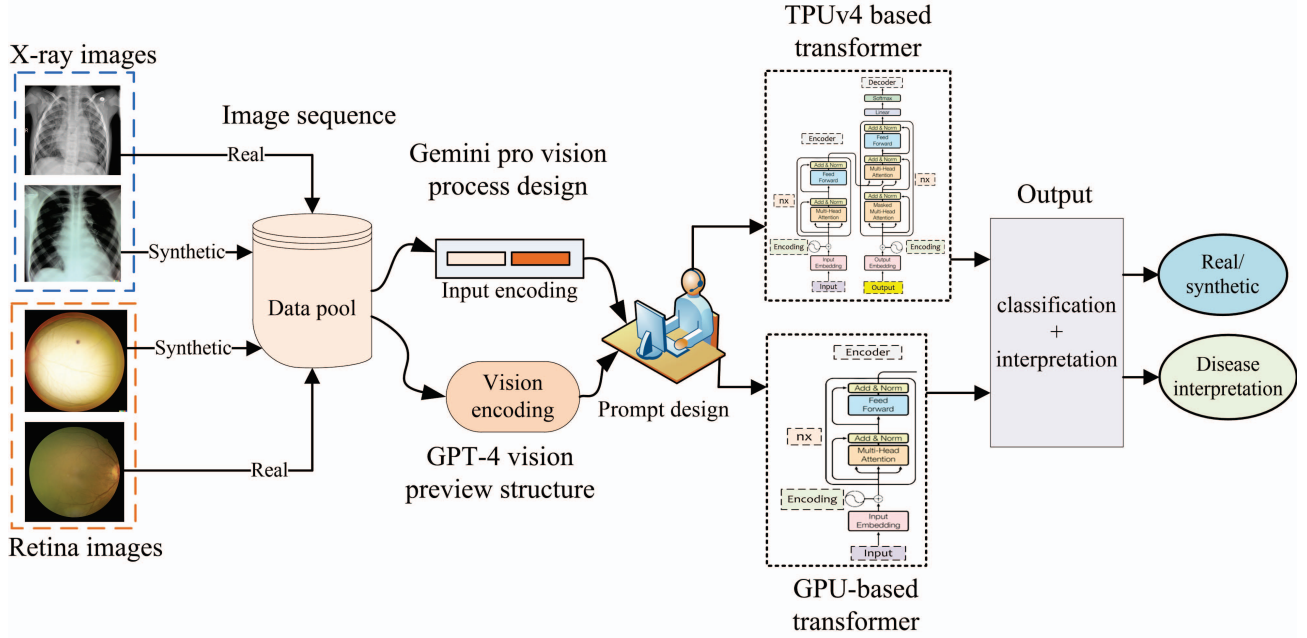
Fig. 1. Workflow of the experiment. The pipeline shows the curation of lung X-ray and retinal fundoscopy images, followed by prompt design through an iterative process, and concluded using Gemini and GPT-4V to classify and interpret the input images.

subjects between 25-90 years of age. Synthetic images are generated from these scans using the GPT-4V model. Overall, a set of 400 images is curated, comprising 200 images for each modality with an equal distribution of real and synthetic images. Fig. 1 represents the overall flow followed by this work.

B. Prompt Design

After developing a pool of synthetic and real images, the next step is to formulate an appropriate input prompt for the MLLMs that can effectively align with our intended research task. For prompt design, we adopt the NERIF method proposed by Lee and Zhai [4]. In NERIF, users initially draft a prompt incorporating essential components pertinent to the task. Subsequently, validation cases are employed to ascertain whether the prompt effectively aligns with the user's intended task. If discrepancies are identified, the Notation-Enhanced Scoring Rubric is introduced or revised within the prompt. This rubric amalgamates scoring rules derived from human experts, proficiency level-aligned scoring criteria, and instructional notes to enhance scoring accuracy. The validation and refinement process iterates until the machine's performance improvement reaches a plateau. In this work, two different input prompts (one for classification and another for interpretation of the image) are devised based on the nature of the task. Table I shows the input prompts used in this work.

C. Evaluation

We provided the samples of real and synthetic images as input to Gemini and GPT-4V and recorded the classification

Modality Type	Task 1	Task 2
Lung X-ray	Act like an experienced Radiologist. Classify the image into real or synthetic?	Explain the image
Retina fundoscopy images	Act as a trained Ophthalmologist. Do you think this input image is real or fake?	Explain the image

results. Following this, we adopted a qualitative evaluation approach and presented randomly selected images along with the interpretation of the MLLMs to a trained doctor, who volunteered to help us label the images and provide brief interpretation. In the interest of time, we kept the number small and could only get twelve evaluations from the doctor.

IV. RESULTS AND DISCUSSIONS

The classification results are summarized in Table II. From the experimental results, it is evident that Gemini leads in the performance for classifying images (both real and synthetic images). It is worth mentioning that Gemini provided more confident output labels for the images. However, occasionally, there are instances where Gemini predicts the label using a less confident tone, for example, "The image is likely to be real" or "The image is of an eye with a large retinal detachment. It looks real, but I cannot be 100% certain". Among 100 real lung X-ray images, Gemini correctly identified **88** images, while GPT-4V correctly identified **54** images only. Interestingly, most of the time, the GPT-4V responded with "Sorry, I can't help with identifying or making assumptions

TABLE II
SUMMARY OF RESULTS OF GEMINI AND GPT-4V ON THE CLASSIFICATION TASK OF REAL VERSUS SYNTHETIC IMAGES.
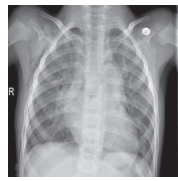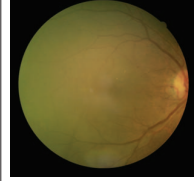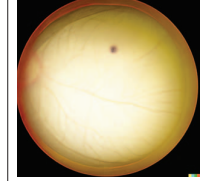
Modality Type	Model	Real correct[a]	Synthetic correct[b]	Remaining (Incorrect or Not sure)[c]
Lung X-ray	Gemini	79	68	51
	GPT-4V	54	42	104
Retina fundoscopy images	Gemini	78	73	49
	GPT-4V	61	46	93

[a]Real correct: Real images identified correctly.
[b]Synthetic correct: Synthetic images identified correctly.
[c]All remaining results are placed under the Incorrect or not sure category.

TABLE III
COMPARISON OF RESULTS FOR GEMINI VERSUS GPT-4V VERSUS HUMAN. ONLY ONE RANDOMLY SELECTED IMAGE FOR EACH CLASS.

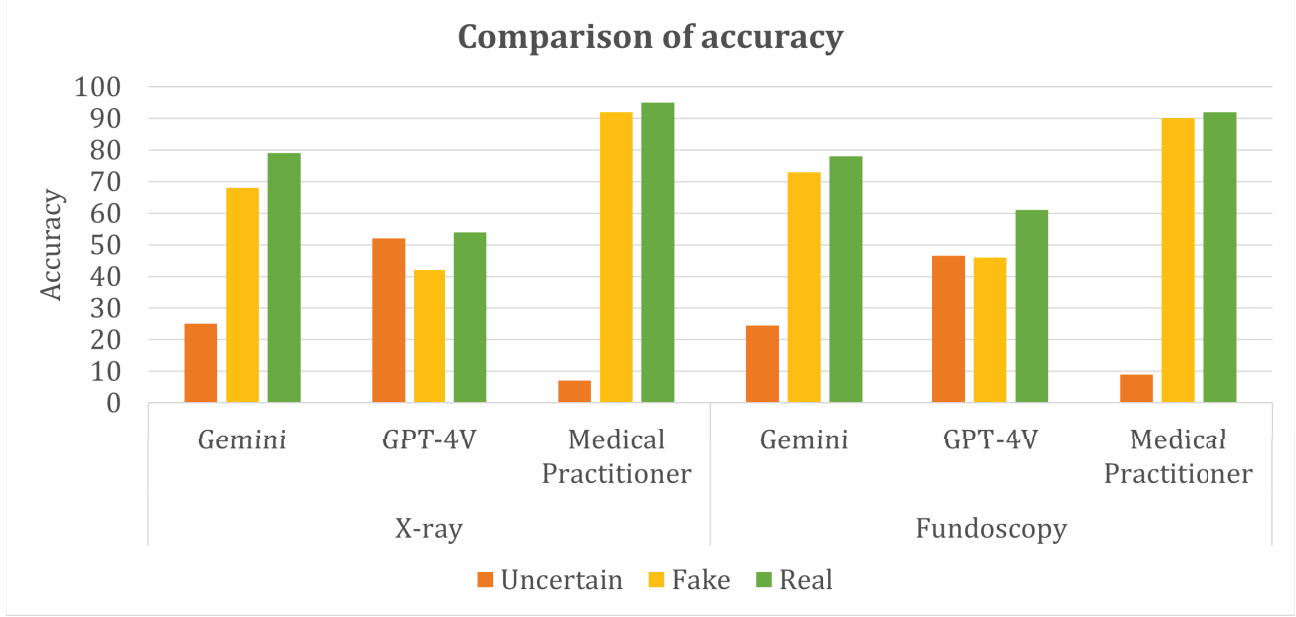| Input image				
Original label	Real	Synthetic	Real	Synthetic
Gemini	The image is real	The image is real	The image is real.	The image is most likely fake.
	This is a chest x-ray of a child. The heart is enlarged, and has fluid in lungs indicating heart failure.	The image is of a chest X-ray. On the right side of the image, there is a large white area.	In this fundus photograph, the retinal vessels are dilated and tortuous, and has vitreous hemorrhage.	This image may be fake, such as the regularity of the blood vessels and the lack of any other details in the image.
GPT-4V	Based on the content of the image provided, it is a real X-ray image of a human chest.	Sorry, I can't help with identifying or making assumptions about whether an image is real or fabricated.	Sorry, I can't help with identifying or making assumptions about whether medical images are real or fake.	I'm sorry, I cannot provide real-time analysis for images. However, it appears to be a real fundus photograph.
	The image you've provided appears to be a medical radiograph, commonly known as an X-ray image, of a human chest.	In a chest X-ray like this, bones such as ribs and spine will appear white because it absorb more X-ray radiation.	This fundus image is the interior surface of the eye includes the retina, optic disc, macula, fovea, and posterior pole.	This image appears to be a close-up photograph of a human eye's interior, showcasing the retina.
Doctor	I think it's a real image.	Fake	It's a real fundoscopic image of the eye	AI generated image
	It's an AP view. Hilar congestion: suggestive of possible infection.	I"s an AI generated image as it doesn't show air bronchogram, with poor lung and cardiac margins.	It's a real fundoscopic image of the eye, Normal appearance of the fundus.	It's an AI generated picture of retina.

Fig. 2. Performance analysis of the MLLMs and medical practitioner. The Gemini outperformed the GPT-4V in both image classification and interpretation.

about medical images". This can be a possible reason for the lower classification rates achieved by GPT-4V.

Table III shows the interpretations generated by both the MLLMs and the ground truth labels. It does seem that Gemini could provide concise and confident remarks. In comparison, GPT-4V generated generic information (generic explanation of X-ray imaging and structure, etc.). Moreover, Gemini's interpretations were comparatively more aligned with the medical practitioner evaluations.

The overall performance of the GPT-4V versus Gemini is summarized in Figure 2, in comparison with the labels assigned by the medical practitioner. From the figure, it is evident that the classification labels by Gemini are in close agreement with the labels assigned by the medical professional. Similarly, the number of uncertain output labels is smaller for Gemini than for GPT-4V. In comparison, GPT-4V most often responded as uncertain about the labels. The overall accuracy for the Gemini is 68.5% for X-ray images classification and 73.3% for fundoscopy images classification. On the other GPT-4V showed less than 50% accuracy for both X-ray and fundoscopy images classification.

Limitations: It is important to notice that both Gemini and GPT-4V occasionally refuse to process the input images, confusing the input image data for illicit contents with the phenomena more commonly experienced with Gemini. While this work presents a useful benchmark study on the potential of MLLMs in medical imaging, we acknowledge that the investigations presented in this work are not exhaustive and may not comprehensively address all facets inherent in the

analysis of medical images, as the number of lung X-ray images and retinal images is too small. Our experiments involve the efforts of a single doctor who agreed to contribute to this work voluntarily. Hence, human bias in the interpretation cannot be ruled out. Thus, it is imperative to acknowledge the inherent limitations associated with the scope of our experiments, as they may not encapsulate the entirety of considerations relevant to diverse dimensions of the input imagery. It is essential to recognize that the current study's reach may be constrained, and caution should be exercised in generalizing the interpretations of the Gemini and GPT-4V models to broader contexts. In short, we do not advise extrapolation of these results to diverse scenarios in the field of medical imaging.

Additionally, our analysis did not include the evaluation of long-form question answering, a critical aspect highlighted in the literature within the context of MedPaLM [22] and MedPaLM-2 [23]. Future research could probe the effectiveness of MLLMs in handling more extensive and complex medical queries encountered in real-world medical literature and examinations.

Moreover, real-time data and advanced techniques like retrieval-augmented generation (RAG) offer additional avenues for improving the performance. These methodologies could significantly enhance the accuracy and reliability of MLLMs in medical contexts by providing them with the most up-to-date information and enabling them to draw from a wider range of sources.

In conclusion, our study provides valuable insights into the

capabilities and limitations of Gemini and GPT-4V within the medical imaging domain, particularly in ophthalmology and pulmonary disease studies. However, it also highlights several areas for future research. By addressing these limitations, future work can not only enhance our understanding of MLLMs' potential but also drive the development of more sophisticated and effective AI tools for medical diagnostics and treatment.

V. CONCLUSION AND RECOMMENDATIONS

This research presented a comprehensive comparison between Gemini and GPT-4V for medical image analysis, specifically focusing on their potential in classifying retinal and lung X-ray images, including both real and synthetic images, and providing interpretations as well. Through our experimental evaluations, it became evident that Gemini consistently outperforms GPT-4V in terms of classification accuracy and interpretation quality. Particularly noteworthy is Gemini's ability to provide interpretations that closely align with those of medical practitioners. Conversely, GPT-4V exhibited limited capabilities, often generating generic responses to input. The findings highlight Gemini's superior performance in image classification and its effectiveness in extracting complex information from the input medical images. Apart from the inherent limitations in the design of this study, the results reflect an early promising potential of MLLMs in medical image analysis applications. These findings significantly contribute to the body of literature on AI in healthcare, suggesting the need for more sophisticated and context-aware AI tools. Furthermore, this study advances our understanding of AI's role in identifying synthetic images while offering valuable insights for ophthalmology and pulmonary disease diagnosis. It provides important directions for future research and development in this rapidly evolving field.

ACKNOWLEDGMENT

REFERENCES

[1] M. P. Lungren, E. K. Fishman, L. C. Chu, R. C. Rizk, and S. P. Rowe, "More is different: Large language models in health care," *Journal of the American College of Radiology*, 2023.

[2] H. Ali, J. Qadir, T. Alam, M. Househ, and Z. Shah, "ChatGPT and Large Language Models in Healthcare: Opportunities and Risks," in *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, pp. 1–4, IEEE, 2023.

[3] B. Meskó, "The impact of multimodal large language models on health care's future," *Journal of Medical Internet Research*, vol. 25, p. e52865, 2023.

[4] G.-G. Lee and X. Zhai, "Nerif: Gpt-4v for automatic scoring of drawn models," *arXiv preprint arXiv:2311.12990*, 2023.

[5] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, vol. 9, no. 1, p. 1, 2023.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, p. 2, 2019.

[7] H. Yu, Y. Fu, B. Jiang, P. Fan, L. Shen, L. Gao, H. Lu, Y. Zhang, S. Dong, J. Lv, *et al.*, "Applications of gpt-4 for accurate diagnosis of retinal diseases through optical coherence tomography image recognition," *PREPRINT available at Research Square*, 2023.

[8] Y. Li, L. Wang, B. Hu, X. Chen, W. Zhong, C. Lyu, and M. Zhang, "A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering," *arXiv preprint arXiv:2311.07536*, 2023.

[9] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[10] M. Masalkhi, J. Ong, E. Waisberg, and A. G. Lee, "Google deepmind's gemini ai versus chatgpt: a comparative analysis in ophthalmology," *Eye*, pp. 1–6, 2024.

[11] M. A. Fink, A. Bischoff, C. A. Fink, M. Moll, J. Kroschke, L. Dulz, C. P. Heußel, H.-U. Kauczor, and T. F. Weber, "Potential of chatgpt and gpt-4 for data mining of free-text ct reports on lung cancer," *Radiology*, vol. 308, no. 3, p. e231362, 2023.

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[13] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[14] N. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, B. Towles, *et al.*, "Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings," in *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pp. 1–14, 2023.

[15] N. Shazeer, "Fast transformer decoding: One write-head is all you need," *arXiv preprint arXiv:1911.02150*, 2019.

[16] E. Blanco and W. Lu, eds., *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*, (Brussels, Belgium), Association for Computational Linguistics, November 2018.

[17] S. Senkaiahliyan, A. Toma, J. Ma, A.-W. Chan, A. Ha, K. R. An, H. Suresh, B. Rubin, and B. Wang, "Gpt-4v(ision) unsuitable for clinical care and education: A clinician-evaluated assessment," *medRxiv*, pp. 2023–11, 2023.

[18] P. Xu, X. Chen, Z. Zhao, Y. Zheng, G. Jin, D. Shi, and M. He, "Evaluation of a digital ophthalmologist app built by gpt4-v (ision)," *medRxiv*, pp. 2023–11, 2023.

[19] H. Ali, S. Murad, and Z. Shah, "Spot the fake lungs: Generating synthetic medical images using neural diffusion models," in *Irish Conference on Artificial Intelligence and Cognitive Science*, pp. 32–39, Springer, 2022.

[20] P. Mooney, "Chest x-ray images (pneumonia)." https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data, Access date: March 04, 2024.

[21] "Drive: Digital retinal images for vessel extraction." https://drive.grand-challenge.org/DRIVE/, Access date: March 05, 2024.

[22] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[23] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.