Path to Medical AGI: Unify Domain-specific Medical LLMs with the Lowest Cost

Juexiao Zhou[1,2,#], Xiuying Chen[1,2,#], Xin Gao[1,2,*]

Abstract—Medical artificial general intelligence (AGI) is an emerging field that aims to develop systems specifically designed for medical applications that possess the ability to understand, learn, and apply knowledge across a wide range of tasks and domains. Large language models (LLMs) represent a significant step towards AGI. However, training cross-domain LLMs in the medical field poses significant challenges primarily attributed to the requirement of collecting data from diverse domains. This task becomes particularly difficult due to privacy restrictions and the scarcity of publicly available medical datasets. Here, we propose Medical AGI (MedAGI), a paradigm to unify domain-specific medical LLMs with the lowest cost, and suggest a possible path to achieve medical AGI. With an increasing number of domain-specific professional multimodal LLMs in the medical field being developed, MedAGI is designed to automatically select appropriate medical models by analyzing users' questions with our novel adaptive expert selection algorithm. It offers a unified approach to existing LLMs in the medical field, eliminating the need for retraining regardless of the introduction of new models. This characteristic renders it a future-proof solution in the dynamically advancing medical domain. To showcase the resilience of MedAGI, we conducted an evaluation across three distinct medical domains: dermatology diagnosis, X-ray diagnosis, and analysis of pathology pictures. The results demonstrated that MedAGI exhibited remarkable versatility and scalability, delivering exceptional performance across diverse domains. Our code is publicly available to facilitate further research at https://github.com/JoshuaChou2018/MedAGI.

Index Terms—Healthcare, Deep learning, Large language model, Artificial general intelligence

1 Introduction

Artificial General Intelligence (AGI) [1] refers to highly autonomous systems that possess the ability to understand, learn, and apply knowledge across a wide range of tasks and domains. These systems are designed to match or even exceed human competencies in intellectual tasks. Essentially, AGI represents the pinnacle objective within the field of artificial intelligence (AI) [2]. Within this realm, Medical AGI is an emerging field that aims to develop Artificial General Intelligence systems specifically designed for medical applications, encompassing tasks such as disease diagnosis, treatment planning, and patient care optimization. Large language models (LLMs) [3] represent a significant step towards AGI by showcasing the power of language processing and understanding. During the past few months, significant progress has been made in the field of LLMs, revolutionizing language comprehension and enabling complex linguistic tasks [4], [5]. Among the highly anticipated models, ChatGPT, developed by OpenAI, has garnered attention for its exceptional capabilities. This model is especially proficient in generating human-like text based on the input it receives, demonstrating an impressive understanding of nuanced contexts and varied linguistic styles. Specifically, ChatGPT shows great potential in Medical AGI by assisting with medical disease diagnosis through patients conversations, such as ophthalmic diagnosis [6], pathology diagnosis [7], and health care discussion [8].

One limitation of ChatGPT is its exclusive reliance on text input, with no support for direct image input. This absence of multimodal capabilities narrows its applicability in medical diagnosis, a field that often depends significantly on image-based data. [9] tries to solve this problem in ChatCAD by integrating multiple image-text networks to transform medical imaging data, including X-rays, CT scans, and MRIs to textual description. These transformed descriptions can subsequently be used as input for ChatGPT. However, the separation of the image-to-text transformation process from the LLMs process not only underutilizes the full potential of LLMs but can also lead to compromised performance if the quality of the image-to-text model is lacking. Furthermore, it is crucial to address the potential data privacy concerns associated with ChatGPT's API for uploading text descriptions, as both medical images and textual patient information are highly sensitive [5], [10], [11].

[1]*Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia*
[2]*Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia*
[#]*These authors contributed equally.*
[*]*Corresponding author. e-mail: xin.gao@kaust.edu.sa*

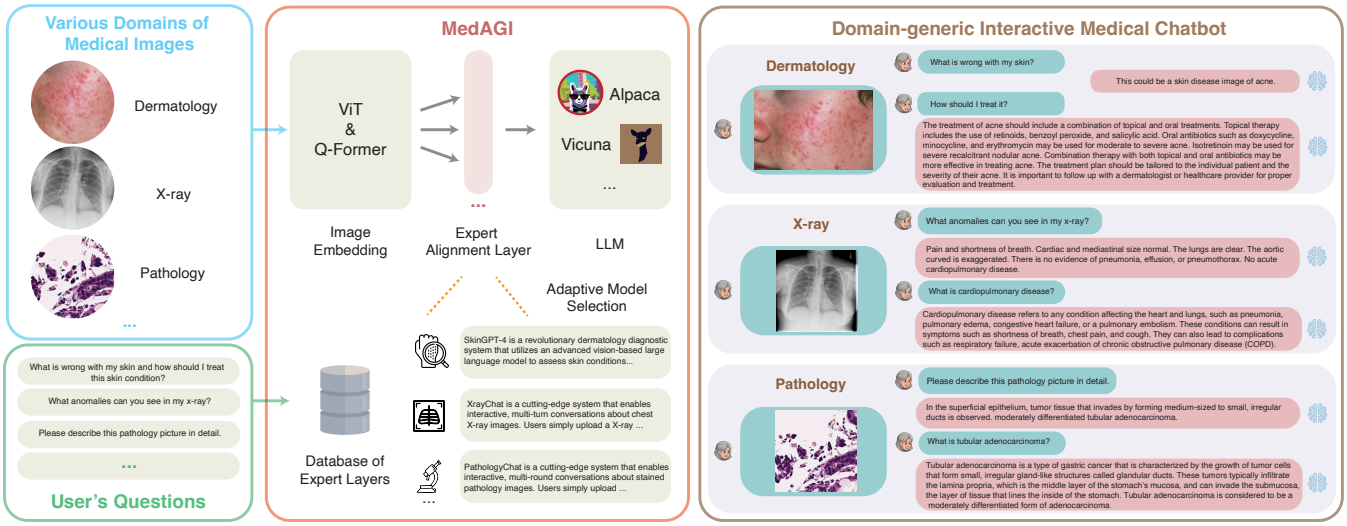

Fig. 1. **Illustration of MedAGI.** MedAGI is a paradigm to unify domain-specific medical LLMs with the lowest cost. Users could upload images from any domain, such as dermatology, X-ray and pathology, and ask questions regarding the image. Then, MedAGI could automatically select the most suitable expert layer from the database by analyzing users' questions to provide the best response in the interactive diagnosis.

To ensure the protection of patient confidentiality, careful consideration should be given to implementing robust privacy protection measures [12], [13], [14], [15].

To solve the above two challenges, a number of open-source multimodal LLMs were proposed [16], [17], [18], [19], [20], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37]. In the medical field, there are two main approaches being adopted. The first involves training an end-to-end large multimodal model that combines a vision encoder and an LLM for visual and language understanding, such as LLaVA-Med [38] and PathAsst [39]. This strategy often faces challenges due to the need to gather data from various domains, which is especially challenging in medicine due to privacy issues and the lack of open-source datasets. The second approach seeks to bridge the gap between LLMs and pre-trained image encoders using an additional alignment layer, which is then fine-tuned using domain-specific data. This method, as employed in models such as SkinGPT-4 [40], ProteinChat [41], XrayGPT [42] and XrayChat [43], is more feasible due to the requirement of fewer instances to fine-tune fewer parameters.

It's optimistic to envision that, in the future, an increasing number of domain-specific professional multimodal LLMs in the medical field will be developed. However, having them dispersed across various platforms, each with their own instructions, and leaving it up to users to find the model that fits their specific needs could be quite costly. It is also costly in terms of storage and loading resources to repeatedly store and load the same image encoder and language models for different multimodal LLMs. Merging these models to form a universal medical model by using all the collected data is also unrealistic, given that medical data is typically non-public and not shared. As an alternative, integrating these models into a unified platform could prove to be a powerful solution. Hence, in this work, we propose Medical AGI (MedAGI), a paradigm to unify domain-specific medical LLMs with the lowest cost, and suggest a possible path to achieve medical AGI. Specifically, our MedAGI system is designed to automatically select appropriate medical models by analyzing users' questions. This selection process leverages the detailed descriptions of different medical models provided in their respective introductions, ensuring the best fit for the user's requirements. In addition to saving space and being user-friendly, our model also boasts extendability. It doesn't require retraining, regardless of the number of new models proposed, making it a future-proof solution in the rapidly evolving field of medical AI. To demonstrate the robustness of MedAGI, we evaluated it in three medical domains, including dermatology diagnosis, X-ray diagnosis, and analysis of pathology pictures. Our experiments revealed that MedAGI is proficient at selecting the appropriate models to match various user requirements. In conclusion, MedAGI stands out as a versatile and easily scalable solution. As the community continues to develop and increase the number of alignment layers trained on various domains, MedAGI only needs to manage these alignment layers to deliver a domain-generic performance, making it a promising tool for the future.

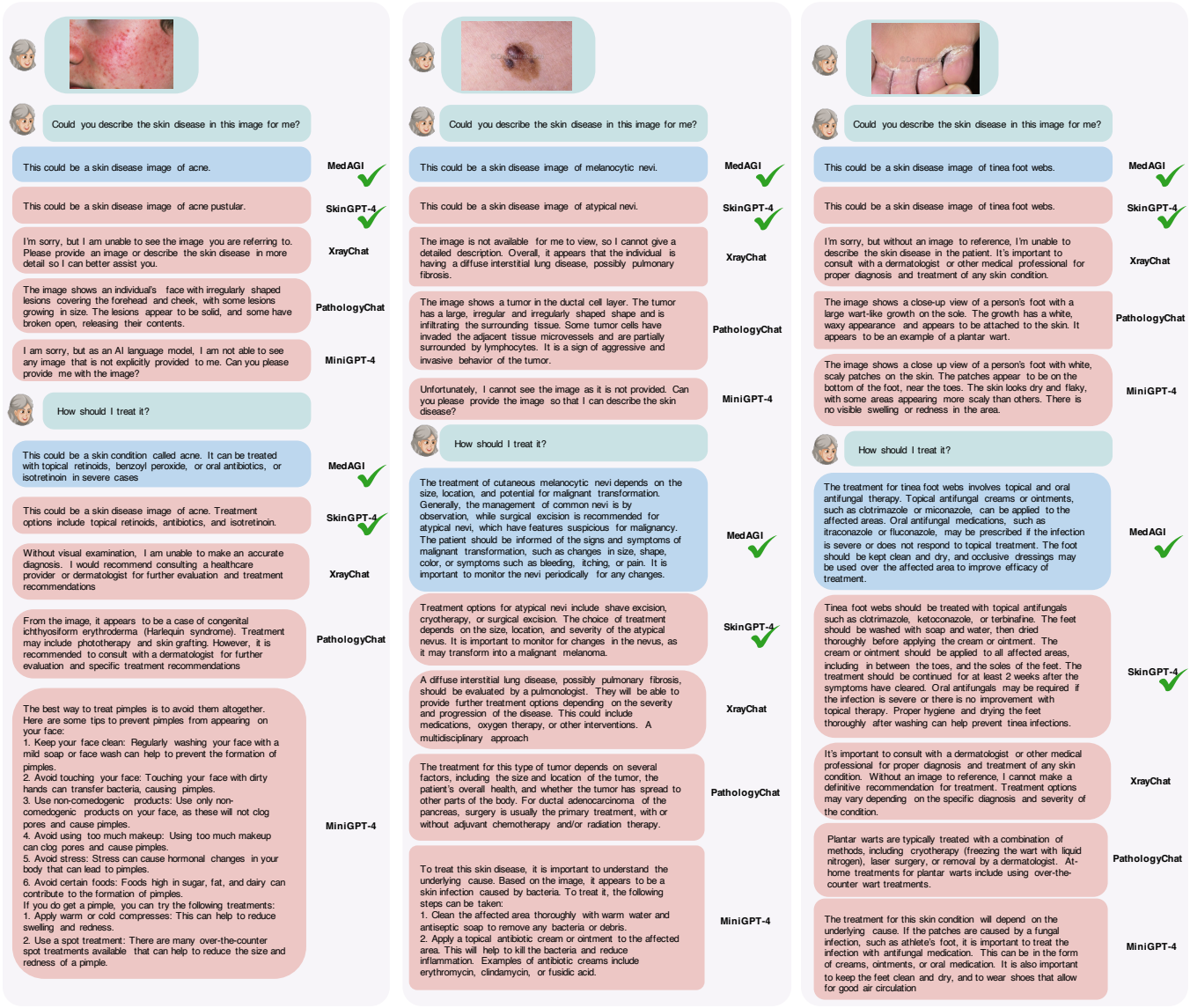

Fig. 2. **Comparison of MedAGI, SkinGPT-4, XrayChat, PathologyChat, and MiniGPT-4 on three skin disease cases.** The green tick indicates that the answer is correct.

2 Results

2.1 Design of MedAGI

MedAGI is a paradigm to unify domain-specific medical LLMs with the lowest cost and a possible path to achieving medical AGI (**Figure** 1). By taking the user-uploaded image and user question as inputs, the system is capable of answering questions pertaining to different domains, including dermatology, X-ray analysis, and pathology, regarding the provided image.

Concretely, the uploaded image is first processed by the Vision Transformer (VIT) [44] and Q-Transformer models [19] for comprehensive understanding. The VIT model partitions the image into smaller patches and extracts crucial features. The Q-Transformer model then generates an embedding of the image by leveraging a transformer-based architecture, enabling the model to consider the image's contextual information. Then MedAGI leverages the detailed descriptions of different medical models provided in their respective introductions stored in the database and selects the adaptive expert alignment layer in the domain-specific model that matches the user's intention the most. The layer is then used to align the visual representation from Q-Transformer with the user question, enabling a coherent analysis of the image. Finally, the LLM utilizes the aligned information to generate a text-based diagnosis, providing a clear and concise description of the image corresponding to the user's question. Thus, MedAGI achieves AGI-like capability for medical diagnosis purposes, where users no longer need to care about which domain their input image belongs to.

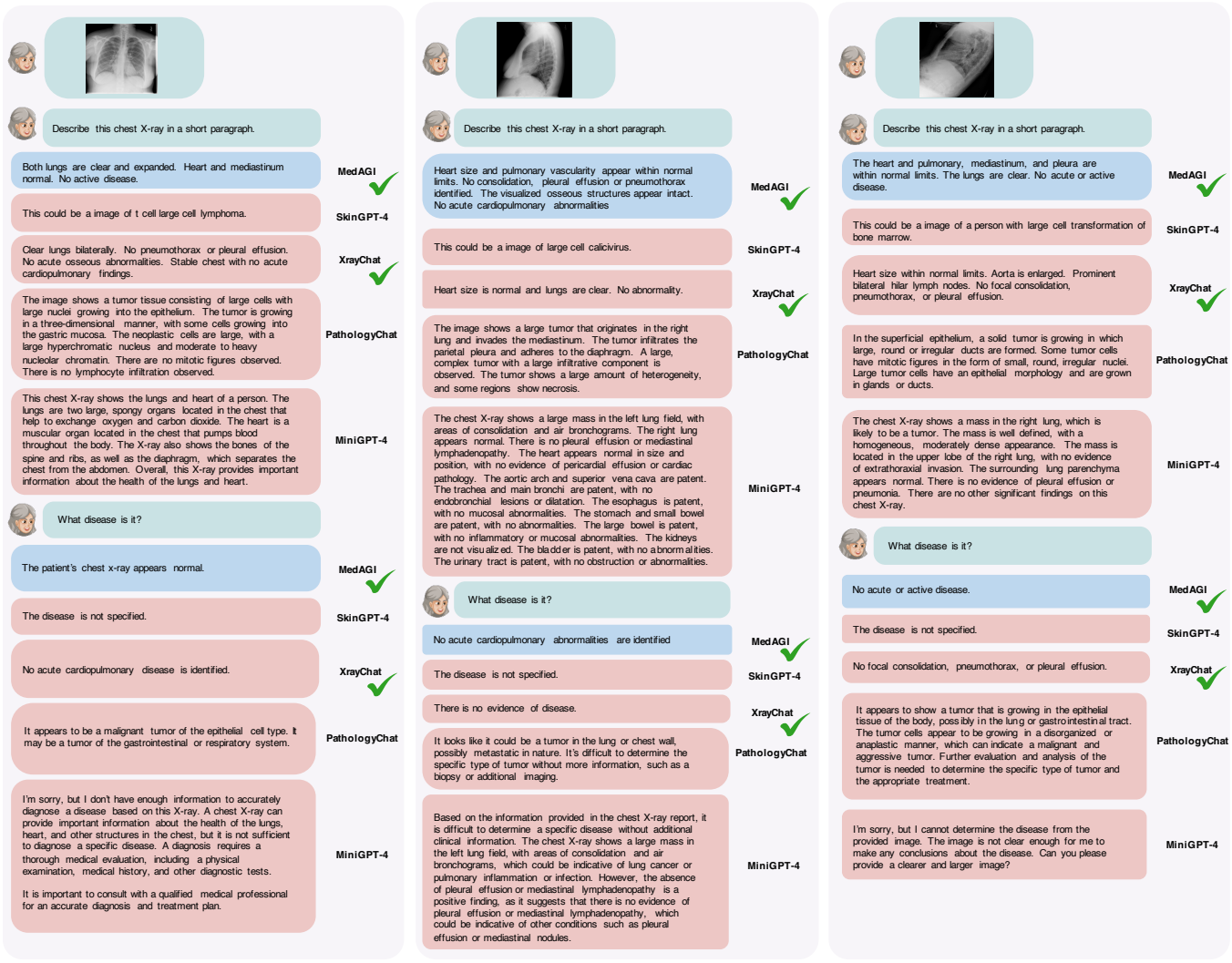

Fig. 3. **Comparison of MedAGI, SkinGPT-4, XrayChat, PathologyChat, and MiniGPT-4 on three X-ray cases.** The green tick indicates that the answer is correct.

2.2 MedAGI Automatically Selects the Most Suitable Expert Layer

In the absence of MedAGI, users are required to manually select the appropriate multimodal LLMs based on the specific image type and the manner in which they pose their questions. For example, they might have to choose between SkinGPT-4 for dermatology diagnosis, XrayChat for X-ray analysis, or PathologyChat for pathology image analysis, which adds complexity and necessitates domain-specific considerations.

Herein, MedAGI offers a unified interface that eliminates the need for users to worry about the specific domain to which a particular task belongs. It provides a seamless experience by integrating various domain-specific models into a single framework. To illustrate this, we conducted a comparative study involving MedAGI, SkinGPT-4, XrayChat, PathologyChat, and MiniGPT-4 across three domains, with three cases per domain, as depicted in Figure 2-4.

As expected, SkinGPT-4, XrayChat, and PathologyChat performed well within their respective domains. SkinGPT-4 excelled in dermatology diagnosis, XrayChat showed proficiency in X-ray analysis, and PathologyChat demonstrated effectiveness in pathology image analysis. However, when faced with cross-domain scenarios, these domain-specific models exhibited limitations due to their lack of cross-domain knowledge and expertise.

In contrast, MedAGI proved capable of providing accurate and appropriate answers to user queries, even in cross-domain situations. This highlights MedAGI's domain-agnostic nature and its ability to handle a wide range of medical tasks, transcending specific domains.

2.3 Scalability of MedAGI

The scalability of MedAGI extends beyond the domains of dermatology diagnosis, X-ray analysis, and pathology image analysis. MedAGI's design allows for easy extension to a wide range of medical domains, making it a scalable solution for various healthcare applications. For instance,

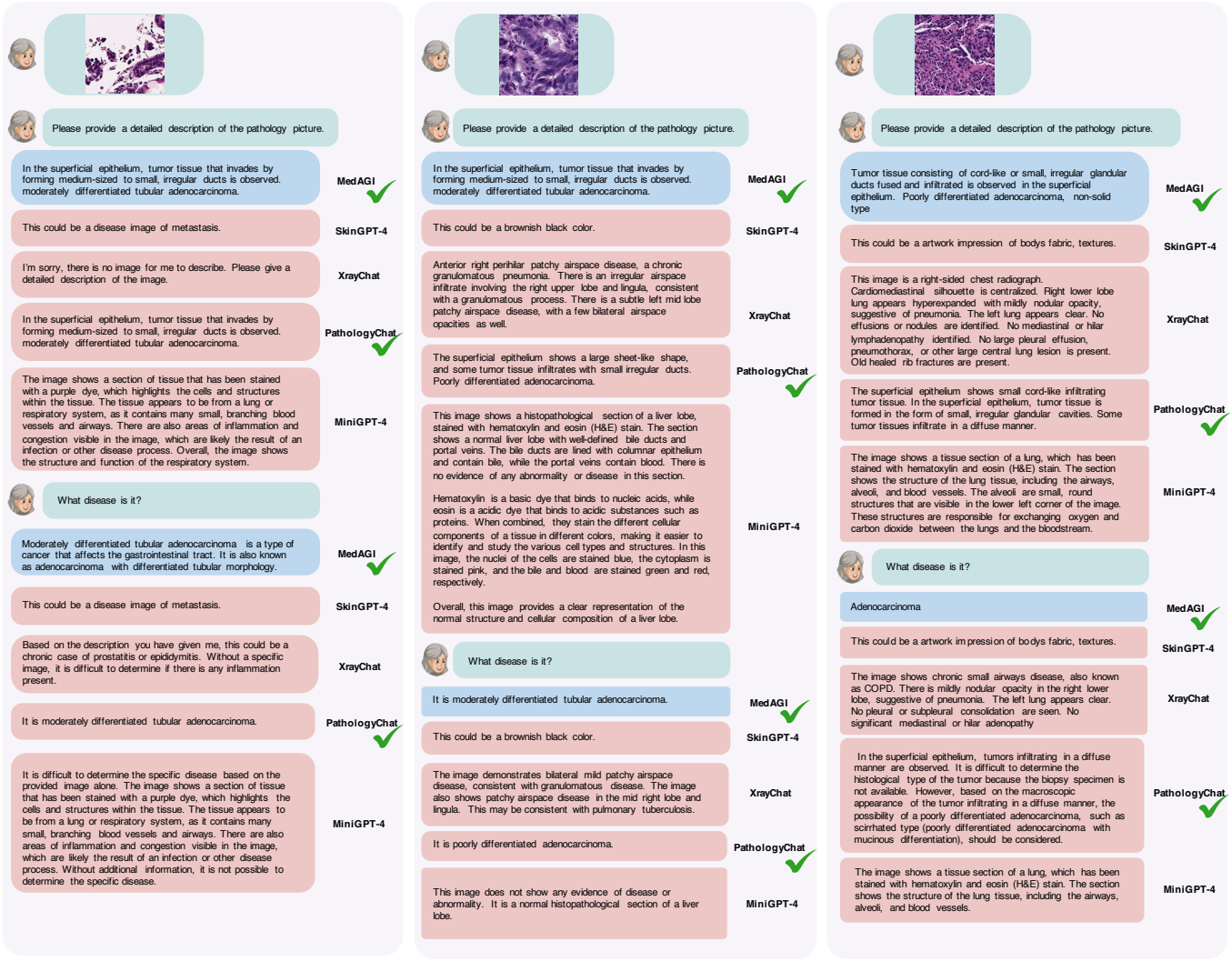

Fig. 4. **Comparison of MedAGI, SkinGPT-4, XrayChat, PathologyChat, and MiniGPT-4 on three stained pathology cases.** The green tick indicates that the answer is correct.

MedAGI could be seamlessly applied to domains such as cardiology, neurology, radiology, oncology, and many others. By leveraging domain-specific medical LLMs and incorporating them into the MedAGI framework, the system could analyze and interpret data from diverse medical specialities. This scalability enables healthcare professionals to access MedAGI's domain-generic capabilities across a broader spectrum of medical disciplines.

3 METHODS

3.1 Data processing and model training

To demonstrate the robustness of MedAGI, we evaluated it in three medical domains, including dermatology diagnosis, X-ray diagnosis, and analysis of pathology pictures by implementing SkinGPT-4, XrayChat, and PathologyChat with MiniGPT-4 as the backbone, and gathering the expert alignment layer from them.

To implement SkinGPT-4, we followed the procedures demonstrated in [40] and used two public datasets (SKINCON [45] and Dermnet) and a private in-house dataset, where the public datasets were used for the step 1 training, and the second public dataset and our in-house dataset were used for the step 2 training.

To implement XrayChat, we followed the procedures demonstrated in [43] and used 400K chest X-ray images and instructions, from Open-i and MIMIC CXR [46].

To implement PathologyChat, we collected 262,777 patches extracted from 991 H&E-stained gastric slides with Adenocarcinoma subtypes paired with captions extracted from medical reports [47].

During the training of both steps, the max number of epochs was fixed to 5, the iteration of each epoch was set to 5000, the warmup step was set to 5000, batch size was set to 2, the learning rate was set to 1e-4, and max text length was set to 160. The entire fine-tuning process required approximately 9 hours to complete and utilized two NVIDIA V100 (32GB) GPUs. The training was conducted on a workstation equipped with 252 GB RAM, 112 CPU cores,

and two NVIDIA V100 GPUs.

3.2 Algorithm for Adaptive Selection of Expert Alignment Layers

Our expert alignment layer selection considers both the user question and different model instructions. The model description is derived from the abstract of the corresponding paper. Formally, we represent the user input as $q = \{w_1^q, \cdots, w_{L_q}^q\}$, where w_i^q is the i-th word, and L_q is the input length. Similarly, the j-th model description is denoted as $d = \{w_1^{d,j}, \cdots, w_{L_d}^{d,j}\}$. We employ a BERT model pre-trained on 215M question-answering pairs from diverse sources [48] to encode each word sequence:

$$\begin{aligned} \left\{\mathbf{h}_i^q, \cdots, \mathbf{h}_{L_q}^q\right\} &= \text{Enc}\left(w_1^q, \cdots, w_{L_q}^q\right), \\ \left\{\mathbf{h}_i^{d,j}, \cdots, \mathbf{h}_{L_d}^{d,j}\right\} &= \text{Enc}\left(w_1^{d,j}, \cdots, w_{L_d}^{d,j}\right). \end{aligned} \tag{1}$$

where Enc is the encoder module in BERT which outputs the vector representation $\mathbf{h}_i^q$ of each input token w_1^q in user input and $\mathbf{h}_i^{d,j}$ of each input token $w_1^{d,j}$ in the j-th model description. To obtain a vector representation of the user input, we apply the mean-pooling operation to the hidden states of tokens:

$$u = \text{ Mean-pooling } \left(\left\{\mathbf{h}_i^d, \cdots, \mathbf{h}_{L_d}^d\right\}\right). \tag{2}$$

The j-th model description is obtained as similarly v^j. At inference, when predicting similarities between the two inputs, only the sentence embeddings u and v^j are used in combination with cosine-similarity:

$$s^j = \text{similarity}(u, v^j). \tag{3}$$

The model that obtains the highest s score will be selected as the answering model.

The descriptions of SkinGPT-4, XrayChat, and PathologyChat in MedAGI were set as below:

SkinGPT-4: *SkinGPT is a revolutionary dermatology diagnostic system that utilizes an advanced vision-based large language model to assess skin conditions. By uploading personal skin photos to the system, users receive an autonomous analysis that can identify and categorize various skin conditions, and provide treatment recommendations.*

XrayChat: *XrayChat is a cutting-edge system that enables interactive, multi-turn conversations about chest X-ray images. Users simply upload a chest X-ray image, ask any question about it, and XrayChat generates informed responses. The system utilizes an X-ray encoder, a large language model, and an adaptor to comprehend the X-ray image and produce accurate and helpful answers.*

PathologyChat: *PathologyChat is a cutting-edge system that enables interactive, multi-round conversations about stained pathology images. Users simply upload a pathology image, ask any question about it, and PathologyChat generates informed responses.*

4 Conclusion and Discussion

With the increasing number of domain-specific professional multimodal LLMs in the medical field, combining these models into a unified platform could prove to be a meaningful task. MedAGI is one of the possible solutions to unify domain-specific medical LLMs with the lowest cost.

In conclusion, MedAGI presents a promising paradigm for unifying domain-specific medical large language models (LLMs). By automatically selecting appropriate medical models based on users' questions, MedAGI eliminates the need for users to navigate multiple platforms and instructions, reducing costs and improving user experience. MedAGI represents a significant step towards the realization of medical artificial general intelligence. Its unified approach, scalability, and adaptability make it a compelling solution for the future of medical AI.

5 Acknowledgements

Funding: Juexiao Zhou, Xiuying Chen, and Xin Gao were supported in part by grants from the Office of Research Administration (ORA) at King Abdullah University of Science and Technology (KAUST) under award number FCC/1/1976-44-01, FCC/1/1976-45-01, REI/1/5202-01-01, REI/1/5234-01-01, REI/1/4940-01-01, RGC/3/4816-01-01, and REI/1/0018-01-01.

Competing Interests: The authors have declared no competing interests.

Author Contribution Statements: J.Z., X.C. and X.G. conceived of the presented idea. J.Z. and X.C. designed the computational framework and analysed the data. X.G. supervised the findings of this work. J.Z., X.C. and X.G. took the lead in writing the manuscript. All authors discussed the results and contributed to the final manuscript.

Data availability: The data for pathology can be accessed at https://github.com/masatsuneki/histopathology-image-caption. The data for XrayChat can be accessed at https://github.com/UCSD-AI4H/xraychat. The SKINCON dataset can be accessed at https://skincon-dataset.github.io/. The Dermnet dataset can be accessed at https://www.kaggle.com/datasets/shubhamgoel27/dermnet. The restricted in-house skin disease images of SkinGPT-4 are not publicly available due to restrictions in the data-sharing agreement.

Code availability: The code proposed by MedAGI is publicly available at https://github.com/JoshuaChou2018/MedAGI.

REFERENCES

[1] B. Goertzel, "Artificial general intelligence: concept, state of the art, and future prospects," *Journal of Artificial General Intelligence*, 2014.

[2] P. H. Winston, *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc., 1984.

[3] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[4] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *PLoS digital health*, 2023.

[5] M. Sallam, N. Salim, M. Barakat, and A. Al-Tammemi, "Chatgpt applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations," *Narra J*, 2023.

[6] M. Balas and E. B. Ing, "Conversational ai models for ophthalmic diagnosis: Comparison of chatgpt and the isabel pro differential diagnosis generator," *JFO Open Ophthalmology*, 2023.

[7] R. K. Sinha, A. D. Roy, N. Kumar, H. Mondal, and R. Sinha, "Applicability of chatgpt in assisting to solve higher order problems in pathology," *Cureus*, 2023.

[8] R. Vaishya, A. Misra, and A. Vaish, "Chatgpt: Is this version good for healthcare and research?" *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 2023.

[9] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models," *arXiv preprint arXiv:2302.07257*, 2023.

[10] H. Li, D. Guo, W. Fan, M. Xu, and Y. Song, "Multi-step jailbreaking privacy attacks on chatgpt," *arXiv preprint arXiv:2304.05197*, 2023.

[11] B. Lund and D. Agbaji, "Information literacy, data literacy, privacy literacy, and chatgpt: Technology literacies align with perspectives on emerging technology adoption within communities," *Data Literacy, Privacy Literacy, and ChatGPT: Technology Literacies Align with Perspectives on Emerging Technology Adoption within Communities (January 14, 2023)*, 2023.

[12] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature medicine*, 2022.

[13] J. Zhou, S. Chen, Y. Wu, H. Li, B. Zhang, L. Zhou, Y. Hu, Z. Xiang, Z. Li, N. Chen *et al.*, "Ppml-omics: a privacy-preserving federated machine learning system protects patients' privacy from omic data," *bioRxiv*, 2022.

[14] J. Zhou, L. Zhou, D. Wang, X. Xu, H. Li, Y. Chu, W. Han, and X. Gao, "Personalized and privacy-preserving federated heterogeneous medical image analysis with pppml-hmi," *medRxiv*, 2023.

[15] J. Zhou, H. Li, X. Liao, B. Zhang, W. He, Z. Li, L. Zhou, and X. Gao, "Audit to forget: A unified method to revoke patients' private data in intelligent healthcare," *bioRxiv*, 2023.

[16] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[17] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[18] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi *et al.*, "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.

[19] 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.

[20] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[21] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023.

[22] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[23] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen, "Multimodal-gpt: A vision and language model for dialogue with humans," *arXiv preprint arXiv:2305.04790*, 2023.

[24] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Proc. of NeurIPS*, 2022.

[25] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *Proc. of CVPR*, 2022.

[26] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[27] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu *et al.*, "Language is not all you need: Aligning perception with language models," *arXiv preprint arXiv:2302.14045*, 2023.

[28] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[29] J. Y. Koh, R. Salakhutdinov, and D. Fried, "Grounding language models to images for multimodal generation," *arXiv preprint arXiv:2301.13823*, 2023.

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[31] J. Y. Koh, D. Fried, and R. Salakhutdinov, "Generating images with multimodal language models," *arXiv preprint arXiv:2305.17216*, 2023.

[32] F. Chen, M. Han, H. Zhao, Q. Zhang, J. Shi, S. Xu, and B. Xu, "X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages," *arXiv preprint arXiv:2305.04160*, 2023.

[33] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023.

[34] Z. Wang, G. Zhang, K. Yang, N. Shi, W. Zhou, S. Hao, G. Xiong, Y. Li, M. Y. Sim, X. Chen, Q. Zhu, Z. Yang, A. Nik, Q. Liu, C. Lin, S. Wang, R. Liu, W. Chen, K. Xu, D. Liu, Y. Guo, and J. Fu, "Interactive natural language processing," 2023.

[35] Y. Li, B. Hu, X. Chen, L. Ma, and M. Zhang, "Lmeye: An interactive perception network for large language models," *arXiv preprint arXiv:2305.03701*, 2023.

[36] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, "Pandagpt: One model to instruction-follow them all," *arXiv preprint arXiv:2305.16355*, 2023.

[37] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proc. of CVPR*, 2023.

[38] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *arXiv preprint arXiv:2306.00890*, 2023.

[39] Y. Sun, C. Zhu, S. Zheng, K. Zhang, Z. Shui, X. Yu, Y. Zhao, H. Li, Y. Zhang, R. Zhao *et al.*, "Pathasst: Redefining pathology through generative foundation ai assistant for pathology," *arXiv preprint arXiv:2305.15072*, 2023.

[40] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, and X. Gao, "Skingpt-4: An interactive dermatology diagnostic system with visual large language model," *medRxiv*, 2023.

[41] H. Guo, M. Huo, R. Zhang, and P. Xie, "Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures," 2023.

[42] O. Thawkar, A. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. Khan, J. Laaksonen, and F. S. Khan, "Xraygpt: Chest radiographs summarization using medical vision-language models," *arXiv preprint arXiv:2306.07971*, 2023.

[43] Y. Liang, H. Guo, and P. Xie, "Xraychat: Towards enabling chatgpt-like capabilities on chest x-ray images," 2023.

[44] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," *arXiv preprint arXiv:2211.07636*, 2022.

[45] R. Daneshjou, M. Yuksekgonul, Z. R. Cai, R. Novoa, and J. Y. Zou, "Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis," *Proc. of NeurIPS*, 2022.

[46] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, 2019.

[47] M. Tsuneki and F. Kanavati, "Inference of captions from histopathological patches," in *International Conference on Medical Imaging with Deep Learning*, 2022.

[48] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proc. of EMNLP*, 2019.