**MiniMedMind: Small-Scale LLM**  
**for High-Performance Medical VQA**   
**on Chest X-ray**

Duong Quoc Khanh, Hoang Trung Kien, Duong Thanh Nam,   
Phan Duy Hung

FPT University, Hanoi, Vietnam  
khanhdqhe171671@fpt.edu.vn, kienhthe170770@fpt.edu.vn, [namdthe170744@fpt.edu.vn](mailto:namdthe170744@fpt.edu.vn), hungpd2@fe.edu.vn

**Abstract.** Visual Question Answering (VQA) in medical imaging has the potential to transform clinical practice by enabling automated, image-based responses to medical queries, particularly in radiology. However, most Medical VQA models are large-scale and computationally demanding, making them impractical for widespread use in resource-limited healthcare settings. In this work, we introduce MiniMedMind, a lightweight and efficient VQA model designed to perform medical question-answering tasks on chest X-Ray (CXR) images. MiniMedMind combines a vision encoder, a projector layer, and a fine-tuned small-scale language model to achieve robust performance with minimal computational requirements. The vision encoder leverages pretrained weights from CheXAgent for effective feature extraction, while the projector aligns these features with a fine-tuned Llama 3.2 model (3B parameters) enhanced with Low-Rank Adaptation (LoRA) to generate clinically accurate responses. The training dataset combines patient-doctor conversations, MIMIC-CXR data, and synthetic conversations generated by GPT-3.5, enabling MiniMedMind to effectively interpret and respond to medical queries. Evaluated on report generation and VQA tasks, MiniMedMind performs near top models like XrayGPT and Med-MoE, achieving competitive accuracy with a significantly lighter architecture. These results position MiniMedMind as an efficient, effective solution for resource-constrained medical AI applications. MiniMedMind offers a computationally efficient solution for Medical VQA, enabling AI-driven diagnostic support that can be applied in diverse clinical and educational settings.

**Keywords:** Medical VQA, AI-driven diagnostic support, Chest X-ray, Lightweight model, LoRA.

1. Introduction
   1. Problem & Motivation

Visual Question Answering (VQA) models have emerged as valuable tools across various fields by enabling automated responses from visual inputs, especially useful in medical settings. Medical VQA models, developed to interpret complex data from diagnostic images such as X-rays, CT scans, and MRIs, hold significant potential for aiding clinical decision-making. However, these models are often large and computationally demanding, which limits their accessibility and scalability in real-world healthcare environments, where high computational resources may not be readily available. As a result, their usage is often restricted to well-funded institutions, making them inaccessible to smaller clinics, research groups, and educational institutions.

A lightweight VQA model tailored for medical use could address these challenges, making medical AI more accessible and practical in diverse settings. Such a model would lower entry barriers for research teams, enabling innovation without the need for costly infrastructure. For radiologists and other clinicians, a compact VQA model could serve as an intelligent assistant, supporting routine tasks like early disease detection and assisting in complex case interpretation. Its accessibility would also make it viable for practitioners in resource-limited settings, enhancing its clinical impact.

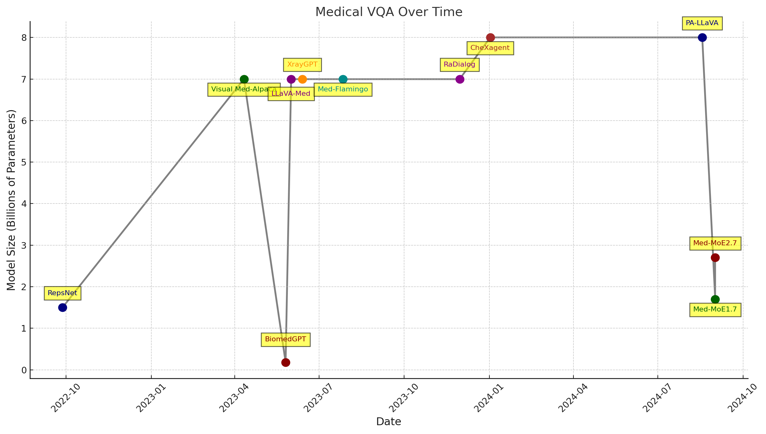
In education, a lightweight Medical VQA model could serve as a powerful teaching aid, providing real-time feedback to medical students and junior doctors on their image interpretations. This feedback would reinforce accurate diagnoses and offer guidance for improvement, enriching the learning experience. Such a model could support interactive training, bridging the gap between advanced AI capabilities and practical application in clinical, research, and educational settings.

* 1. Literature Review

Small Language Models have become essential in creating VQA systems by offering efficient alternatives to larger models. Models such as Gemma [1, 2] and StableLM [3] show that small-scale language models can achieve reasonable language understanding, while Qwen-1.8B [4] and Phi2/Phi3-Mini [5, 6] compress complex tasks into manageable architectures without substantial accuracy loss. TinyLlama [7] and Galactica [8] advance this trend by emphasizing efficient learning within smaller architectures, and models like Opt [9] and Pythia [10] focus on reducing computational demands for training and inference.

Small Vision-Language Models, including MobileVLM [11], MoE-LLaVA [12], and Vary-toy [13], exemplify the shift toward resource-efficient multimodal systems, ideal for real-time applications in medical environments. Models such as TinyGPT-V [14] and Mobile-Agent [15] effectively combine visual comprehension with natural language processing, offering results that balance accuracy with computational efficiency, laying a strong foundation for further developments in lightweight vision-language frameworks.

In the medical field (Fig. 1), Medical Vision-Language Models like LLaVA-Med [16] and Med-Flamingo [17] expand medical image comprehension using large-scale, domain-specific datasets, making them particularly suitable for visual question-answering tasks involving images like X-rays and MRIs. XrayGPT [18] and CheXagent [19] focus on diagnostics, providing context-aware insights from medical images, while models like Pa-LLaVA [20] and Visual Med-Alpaca [21] refine these capabilities for specific medical tasks that require integrated vision and language processing.



**Fig. 1.** The development of medical VQA model over time  
(The x-axis: date that model released, and the y-axis: the model size of its LLM)

Small Medical Vision-Language Models are emerging as compact solutions blending medical knowledge with efficiency. Models like Med-MoE [22] and RepsNet [23] achieve competitive performance with lightweight architectures, particularly in medical image analysis. BiomedGPT [24] further integrates multimodal medical data, providing a compact yet capable VQA tool for healthcare. These models aim to improve access to medical VQA systems, making them more viable for resource-limited healthcare environments. However, an optimized lightweight model specifically for chest X-ray imaging remains undeveloped, indicating an area for future advancements.

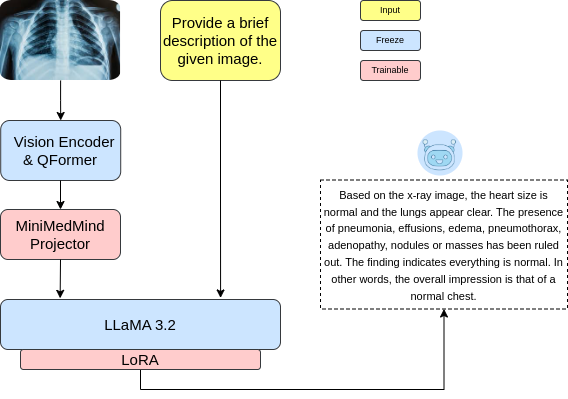
* 1. Contribution

Specifically, this paper makes the following contributions:

* *MiniMedMind*, a compact yet robust medical VQA model capable of generating accurate responses from CXR images.
* *MiniMedMind Instruction Dataset*, a custom VQA dataset generated by GPT-3.5, designed to enhance the model’s performance on multimodal medical question-answering tasks.
* *Small-scale language model,* optimized through patient-doctor conversations to emulate a doctor’s expertise.

1. Methodology
   1. Model Architecture

In this subsection, the architecture of MiniMedMind is outlined, consisting of a vision encoder, a projector layer, and a large language model, as shown in Fig. 2.



**Fig. 2.** The architecture of MiniMedMind

**Vision Encoder Backbone.** MiniMedMind’s vision encoder utilizes EVA-CLIP-g [25], a Vision Transformer (ViT)-based model [26] designed for feature extraction from chest X-ray (CXR) images. Enhanced with a QFormer [27] based on BERT, the encoder is tailored to capture complex visual patterns and contextual cues within medical imaging data. Unlike traditional models that require extensive retraining, this vision encoder is initialized with pretrained weights from CheXagent’s encoder, a model optimized specifically for CXR analysis. By leveraging these pretrained weights, MiniMedMind gains high-quality feature extraction capabilities specific to radiographic images, allowing it to focus on clinically relevant information with minimal additional training.

**Projector Layer.** The projector layer bridges the gap between the extracted visual features and the language model, facilitating seamless integration of visual and textual data. This layer consists of a linear projection that maps high-dimensional feature representations from the vision encoder to a space compatible with the language model’s embedding layer. Through this dimensional alignment, the projector ensures that visual information aligns effectively with the language model’s structure, enhancing efficiency while maintaining accuracy in outputs that correspond closely to the visual input. This streamlined projection approach helps MiniMedMind sustain a low parameter count, optimizing its efficiency in generating accurate responses.

**Large Language Model Backbone.** The generative backbone of MiniMedMind is the LLaMA 3.2 [28] model with 3 billion parameters, known for its robust language comprehension and generative capabilities. Fine-tuned with LoRA (Low-Rank Adaptation), a parameter-efficient tuning method, the model adapts to the medical context of MiniMedMind’s tasks without significantly increasing computational demands. LoRA fine-tuning enables MiniMedMind to excel in vision-language challenges like VQA and report generation, delivering coherent and contextually accurate responses tailored to medical imaging. This integrated architecture balances computational efficiency with high performance, making MiniMedMind effective in tasks that require both visual analysis and language understanding in the medical domain.

* 1. Training Stages

The training process for MiniMedMind in radiology follows a three-stage approach designed to optimize the model’s capacity for interpreting and responding to medical queries based on both text and visual data. Each stage plays a distinct role in enhancing the model’s performance with complex, multimodal inputs.

**Stage 0: Instruction Tuning LLM on Chest X-Ray Samples.** The initial stage involves instruction tuning of the large language model (Llama-3.2-3B) using a dataset of radiology-related conversational samples. This stage imparts specialized knowledge in medical concepts, focusing on radiology-specific question-answering scenarios. The aim is to train the model in medical terminology and reasoning patterns common in radiological practice, enabling it to address specialized medical inquiries with increased precision.

**Stage 1: Pretrain (Feature Alignment).** The second stage centers on vision-language alignment, where the model is pretrained on paired image-text data. This process aligns visual features from radiological images with the language model’s word embedding space, allowing it to associate visual content with relevant textual descriptions. By establishing this alignment, the model learns to integrate visual elements from medical images with language, ensuring it can produce contextually relevant responses that link visual and linguistic information effectively.

**Stage 2: Visual Instruction Tuning.** The final stage involves visual instruction tuning, enhancing the model's ability to respond to queries that include both text and image inputs. In this phase, the model is fine-tuned on a diverse set of visual instructions requiring the integration of textual and visual data. This step strengthens the model’s ability to interpret and execute complex instructions involving visual content, allowing it to handle diagnostic tasks that rely on the accurate interpretation of radiological images alongside textual guidance.

1. Experiments and Results
   1. Dataset

To address the limited availability of multimodal biomedical datasets for training an instruction-following assistant, we have curated a comprehensive dataset specifically tailored for medical imaging tasks, drawing from existing resources and generating synthetic data where necessary. This dataset comprises several parts: pretraining and fine-tuning datasets for the language model and the projector, as well as benchmark datasets for evaluation across multiple tasks, as described below.

**Language Model Fine-Tuning Data**. To optimize the language model for medical conversational tasks, we inherit conversational samples from XrayGPT, which were originally used to fine-tune their Vicuna model. This dataset contains 141,078 samples, capturing a broad range of medical conversations relevant to CXR (chest X-ray) interpretation. These conversations help align the language model with the types of dialogue and instructional responses commonly required in clinical settings, providing a strong foundation for handling complex, multimodal queries.

**Projector Pretraining Data**. The initial training for the projector layer leverages the MIMIC-CXR dataset [29], a large-scale, de-identified public dataset containing 377,110 chest radiographs in JPG format, along with structured captions derived from radiology reports. This dataset includes 214,539 paired images and captions, specifically adapted for supporting medical research in image understanding and natural language processing. The captions, inherited from XrayGPT, offer a standardized basis for associating textual summaries with visual data, enhancing the projector’s ability to map visual features effectively to the language model’s input space.

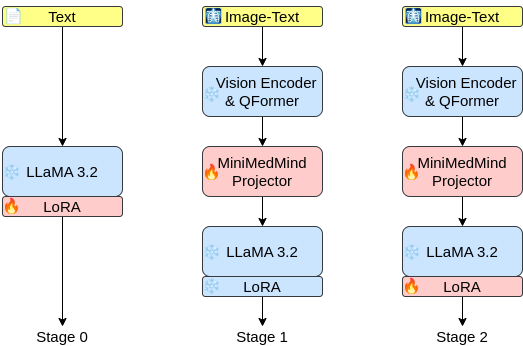
**Projector Fine-Tuning Data**. For fine-tuning the projector layer, we employ two additional datasets focused on enhancing conversational and contextual alignment with chest X-Ray images. The first, LLaVA-Med-Chest, consists of 6,542 images and 17,980 conversations filtered exclusively for chest-related content from the original dataset [16]. The second dataset, MiniMedMind-Instruct-80k, is custom-created to further align with the instructional needs of MiniMedMind. It includes 20,000 images paired with 81,851 conversations, which were synthetically generated using GPT-3.5 from the MIMIC-CXR dataset reports. This synthetic data generation allows for the creation of diverse conversational contexts, expanding the model’s ability to handle a wide array of questions and responses related to CXR images.

**Benchmark Datasets for Evaluation**. To assess the performance of MiniMedMind across both report generation and VQA tasks, we utilize several benchmark datasets. For report generation, we employ the MIMIC-CXR test set, containing 600 samples, and the OpenI [30] test set, comprising 640 samples. These datasets provide structured scenarios to evaluate the model’s ability to generate accurate, detailed radiology reports from CXR images. For the VQA task, we utilize the MIMIC-CXR-VQA [31] test set, which includes 600 samples evenly split between open and closed questions. This dataset, generated from MIMIC-CXR by ChatGPT-4, integrates clinical images with electronic health records to offer a realistic setting for medical VQA and multimodal AI model testing. Additionally, the SLAKE [32] test set, containing 361 samples with 68% open-ended questions, has been filtered to retain only chest-related and English-language content. Lastly, we include the VQA-RAD [33] test set, consisting of 131 samples with 35% open questions, also filtered to focus solely on chest-related queries. These datasets collectively enable a comprehensive evaluation of MiniMedMind’s capacity to interpret and respond to both structured and unstructured visual-linguistic tasks in the medical imaging domain.

* 1. Implementation Details

**Training Process.** MiniMedMind’s training consisted of three stages aimed at refining model components for optimal multimodal alignment.

**Stage 0: Instruction Tuning of the Language Model**. This stage focused on tuning LLaMA 3.2 with Low-Rank Adaptation (LoRA) using a dataset of 141,078 medical conversations related to chest X-rays. Training occurred over 1 epoch with a batch size of 10, lasting approximately 5 hours on an RTX 6000 Ada GPU. The prompt template guided the model to generate contextually accurate responses to radiological queries.



**Fig. 3.** Training stages  
Stage 0: Finetuning LoRA LLaMA3.2; Stage 1: Pretraining projector; Stage 2: Finetuning Projector and LoRA on LLaMA3.2 simultaneously

**Stage 1: Pretraining the Projector for Feature Alignment**. In this stage, the linear projector layer was pretrained to align visual features from chest X-ray images with the language model's embedding space, using 213,514 image-caption pairs from the MIMIC-CXR dataset. The model was trained for 160 epochs, each with 500 iterations and a batch size of 120, over 55 hours on four NVIDIA RTX A6000 GPUs.

**Stage 2: Visual Instruction Tuning**. The final stage fine-tuned both the projector and the LoRA-tuned LLaMA model simultaneously on a combined dataset of LLaVA-Med-Chest and MiniMedMind-Instruct-80k (99,363 samples). Training spanned 1 epoch with 1600 iterations per epoch, a batch size of 20, and a total training time of 21 hours on four NVIDIA RTX A6000 GPUs. This process enabled MiniMedMind to generate contextually appropriate responses to complex medical queries by integrating visual data with language understanding.

* 1. Metric Evaluation

Evaluating MiniMedMind’s performance in report generation involves two key metrics: BERTScore [34] and ROUGE [35], both crucial for capturing semantic fidelity and content completeness in clinical summaries from chest X-rays. BERTScore assesses semantic similarity using BERT embeddings, accommodating lexical and semantic variations critical in medical contexts where exact wording may vary without altering clinical meaning. ROUGE, focusing on n-gram and sequence overlap (ROUGE-1, ROUGE-2, ROUGE-L), ensures reports retain essential medical terms and findings, supporting factual accuracy and interpretability.

For VQA, performance is assessed separately for open-ended and close-ended questions. Recall is used for open-ended questions to evaluate the model's ability to capture pertinent clinical details, which is essential for comprehensive diagnostic responses. Accuracy, applied to close-ended yes/no queries, measures correctness in binary decisions, reflecting the model’s reliability in interpreting straightforward diagnostic prompts—an essential function in medical imaging analysis.

* 1. Results

The evaluation of MiniMedMind across VQA and Report Generation tasks reveals its competitive standing against larger models like CheXAgent, Med-MoE, LlavaMed, and XrayGPT across various datasets.

**Visual Question Answering.** MiniMedMind's performance is assessed on MIMIC-VQA, SLAKE, and VQA-RAD datasets, covering “Close” (closed-ended) and “Open” (open-ended) question types (Table 1). In MIMIC-VQA, MiniMedMind shows lower performance in “Close” questions, scoring 0.297 compared to CheXAgent (0.530) and Med-MoE (0.533), but surpasses LlavaMed (0.440). In “Open” questions, it scores 0.454, trailing CheXAgent (0.547) but outperforming LlavaMed (0.367), with Med-MoE leading in the closed category. In SLAKE, MiniMedMind demonstrates stronger performance, scoring 0.465 in “Close” questions, slightly ahead of CheXAgent (0.456) and LlavaMed (0.325), while Med-MoE leads at 0.816. For “Open” questions in SLAKE, MiniMedMind achieves 0.585, closely trailing Med-MoE (0.830). In VQA-RAD, MiniMedMind scores 0.635 in “Close” questions, ranking second after Med-MoE (0.929) and outperforming both CheXAgent (0.529) and LlavaMed (0.634). For “Open” questions in VQA-RAD, MiniMedMind scores 0.453, again securing second place after Med-MoE (0.870). These results underscore MiniMedMind’s competitive performance, especially in SLAKE and VQA-RAD where it frequently ranks second, though it faces challenges on MIMIC-VQA, particularly in “Close” questions, indicating potential dataset-specific limitations.

**Report Generation.** MiniMedMind's report generation capabilities are evaluated on MIMIC-VQA and OPENI datasets, using BERT, R1, R2, and RL metrics (Table 2). In MIMIC-VQA, MiniMedMind achieves a BERT score of 0.601, surpassing CheXAgent (0.479) and LlavaMed (0.586), though still behind XrayGPT (0.708), the leader. For the R1, R2, and RL metrics, it scores 0.279, 0.074, and 0.16, respectively, maintaining an edge over CheXAgent and remaining competitive with LlavaMed, particularly on RL. On the OPENI dataset, MiniMedMind scores 0.601 on BERT, demonstrating strong performance nearly matching XrayGPT (0.626). Its R1 score is 0.273, close to XrayGPT (0.316) and ahead of CheXAgent. For R2 and RL, it scores 0.064 and 0.168, retaining a competitive edge over CheXAgent but slightly behind XrayGPT and LlavaMed in certain submetrics. Overall, MiniMedMind performs well in report generation, especially on BERT, where it approaches the top model. Although it does not surpass XrayGPT across all metrics, MiniMedMind consistently outperforms CheXAgent and competes closely with LlavaMed, highlighting its suitability for producing coherent and contextually accurate medical reports.

**Table 1.** VQA Task

**Bold** denotes the best performance; underlined denotes the second-best.   
 Use Accuracy for Close-end Answer and Recall for Open-end Answer

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | LLM Params | MIMIC-VQA | | SLAKE | | VQA-RAD | |
| Close | Open | Close | Open | Close | Open |
| CheXAgent | 8B | 0.530 | **0.547** | 0.456 | 0.270 | 0.529 | 0.202 |
| Med-MoE | 1.8Bx4 | **0.533** | 0.133 | **0.816** | **0.830** | **0.929** | **0.870** |
| LlavaMed | 7B | 0.440 | 0.367 | 0.325 | 0.478 | 0.634 | 0.399 |
| MiniMedMind | 3B | 0.297 | 0.454 | 0.465 | 0.585 | 0.635 | 0.453 |

**Table 2.** Report Generation Task

**Bold** denotes the best performance; underlined denotes the second-best.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | LLM | MIMIC-VQA | | | | OPENI | | | |
| Bert | R1 | R2 | Rl | Bert | R1 | R2 | Rl |
| XrayGPT | 7B | **0.708** | **0.437** | **0.195** | **0.326** | **0.626** | **0.316** | **0.08** | 0.189 |
| CheXAgent | 8B | 0.479 | 0.154 | 0.049 | 0.112 | 0.479 | 0.164 | 0.047 | 0.119 |
| LlavaMed | 7B | 0.586 | 0.275 | 0.091 | 0.205 | 0.586 | 0.286 | 0.076 | **0.200** |
| MiniMedMind | 3B | 0.601 | 0.279 | 0.074 | 0.16 | 0.601 | 0.273 | 0.064 | 0.168 |

1. Conclusion and Future Works

MiniMedMind, with a modest parameter count of 3B, performs well compared to larger models like CheXAgent (8B) and LlavaMed (7B). It demonstrates particular strengths on VQA tasks in the VQA-RAD and SLAKE datasets, as well as competitive scores in report generation, especially in the OPENI dataset’s BERT metric.

Despite its smaller size, MiniMedMind frequently matches or surpasses larger models in certain metrics, reflecting its efficiency and focused specialization. However, it encounters challenges with more complex or information-rich datasets like MIMIC-VQA, suggesting areas for enhancement, such as fine-tuning on MIMIC data or further training to improve response diversity for open-ended questions.

Overall, MiniMedMind strikes a commendable balance of performance relative to its size, excelling on SLAKE and VQA-RAD datasets and performing competitively in report generation. Future enhancements could further strengthen its capabilities on MIMIC-VQA and refine its open-ended response accuracy.

While MiniMedMind shows considerable progress in lightweight medical VQA and report generation, certain limitations suggest areas for future enhancement. One issue is the occasional generation of overly lengthy responses, with sections that may be redundant or lack relevance—an aspect that can reduce clarity in clinical settings where conciseness is vital. Additionally, like many language models, MiniMedMind sometimes generates plausible-sounding but inaccurate information, risking clinical reliability.

Future improvements will focus on refining response clarity and accuracy, with plans to extend MiniMedMind’s capabilities to dynamic medical data, such as video and 3D imaging, supporting a broader range of diagnostic applications. Expanding beyond radiology, MiniMedMind could also adapt to fields like cardiology, pathology, and dermatology. To validate its practical utility, clinical evaluations will be conducted to assess MiniMedMind’s reliability, usability, and overall impact on clinical workflows. These advancements aim to establish MiniMedMind as a versatile, clinically reliable tool in healthcare AI.

References

1. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, et al.: GemMa: Open models based on Gemini research and technology, https://arxiv.org/abs/2403.08295.
2. Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., et al.: Gemma 2: Improving open language models at a practical size, https://arxiv.org/abs/2408.00118.
3. Stable LM 2 1.6B Technical Report, https://arxiv.org/html/2402.17834v1.
4. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen Technical Report, https://arxiv.org/abs/2309.16609.
5. Hughes, A.: Phi-2: The surprising power of small language models - Microsoft Research, https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.
6. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., et al.: PHI-3 Technical Report: A highly capable language model locally on your phone, https://arxiv.org/abs/2404.14219.
7. Zhang, P., Zeng, G., Wang, T., Lu, W.: TinyLlama: an Open-Source Small Language Model, https://arxiv.org/abs/2401.02385.
8. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., Stojnic, R.: Galactica: a large language model for science, https://arxiv.org/abs/2211.09085.
9. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.: OPT: Open Pre-trained Transformer Language Models, https://arxiv.org/abs/2205.01068.
10. Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O’Brien, K., Hallahan, E., Khan, M.A., Purohit, S., Prashanth, U.S., Raff, E., Skowron, A., Sutawika, L., Oskar, V.D.W.: Pythia: a suite for analyzing large language models across training and scaling, https://arxiv.org/abs/2304.01373.
11. Chu, X., Qiao, L., Lin, X., Xu, S., Yang, Y., Hu, Y., Wei, F., Zhang, X., Zhang, B., Wei, X., Shen, C.: MobileVLM : a fast, strong and open vision language assistant for mobile devices, https://arxiv.org/abs/2312.16886.
12. Lin, B., Tang, Z., Ye, Y., Huang, J., Zhang, J., Pang, Y., Jin, P., Ning, M., Luo, J., Yuan, L.: MOE-LLAVA: mixture of experts for large Vision-Language models, https://arxiv.org/abs/2401.15947.
13. Wei, H., Kong, L., Chen, J., Zhao, L., Ge, Z., Yu, E., Sun, J., Han, C., Zhang, X.: Small Language Model Meets with Reinforced Vision Vocabulary, https://arxiv.org/abs/2401.12503.
14. Yuan, Z., Li, Z., Huang, W., Ye, Y., Sun, L.: TinyGPT-V: efficient multimodal large language model via small backbones, https://arxiv.org/abs/2312.16862.
15. Wang, J., Xu, H., Ye, J., Yan, M., Shen, W., Zhang, J., Huang, F., Sang, J.: Mobile-Agent: Autonomous Multi-Modal Mobile Device Agent with Visual Perception, https://arxiv.org/abs/2401.16158.
16. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLAVA-MeD: Training a large Language-and-Vision assistant for biomedicine in one day, https://arxiv.org/abs/2306.00890.
17. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Zakka, C., Dalmia, Y., Reis, E.P., Rajpurkar, P., Leskovec, J.: Med-Flamingo: a Multimodal Medical Few-shot Learner, https://arxiv.org/abs/2307.15189.
18. Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S.: XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models, https://arxiv.org/abs/2306.07971.
19. Chen, Z., Varma, M., Xu, J., Paschali, M., Dave, V.V., Johnston, A., Youssef, A., Blankemeier, L., Bluethgen, C., Altmayer, S., Valanarasu, J.M.J., Muneer, M.S.E., Reis, E.P., Cohen, J.P., Olsen, C., Abraham, T.M., Tsai, E.B., Beaulieu, C.F., Jitsev, J., Gatidis, S., Delbrouck, J.-B., Chaudhari, A.S., Langlotz, C.P.: A Vision-Language foundation model to enhance efficiency of chest x-ray interpretation, https://arxiv.org/abs/2401.12208.
20. Dai, D., Zhang, Y., Xu, L., Yang, Q., Shen, X., Xia, S., Wang, G.: PA-LLAVA: a large Language-Vision assistant for human Pathology image understanding, https://arxiv.org/abs/2408.09530.
21. Visual Med-Alpaca, https://cambridgeltl.github.io/visual-med-alpaca/.
22. Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., Liu, Z.: Med-MOE: Mixture of Domain-Specific experts for lightweight medical Vision-Language models, https://arxiv.org/abs/2404.10237.
23. Tanwani, A.K., Barral, J., Freedman, D.: RepsNet: Combining Vision with Language for Automated Medical Reports. In: Lecture notes in computer science. pp. 714–724 (2022). https://doi.org/10.1007/978-3-031-16443-9\_68.
24. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., Huang, J., Chen, C., Zhou, Y., Fu, S., Liu, W., Liu, T., Li, X., Chen, Y., He, L., Zou, J., Li, Q., Liu, H., Sun, L.: A generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine. 30, 3129–3141 (2024). https://doi.org/10.1038/s41591-024-03185-2.
25. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: EVA-CLIP: Improved training techniques for CLIP at scale, https://arxiv.org/abs/2303.15389.
26. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, https://arxiv.org/abs/2010.11929.
27. Zhang, Q., Zhang, J., Xu, Y., Tao, D.: Vision Transformer with Quadrangle Attention, https://arxiv.org/abs/2303.15105.
28. meta-llama/Llama-3.2-3B-Instruct · Hugging Face, https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.
29. MIMIC-CXR Database v2.0.0, https://www.physionet.org/content/mimic-cxr/2.0.0/.
30. OpenINGWeb, <https://openi.nlm.nih.gov/faq>.
31. MIMIC-EXT-MIMIC-CXR-VQA: A Complex, Diverse, and Large-Scale Visual Question Answering Dataset for Chest X-ray Images V1.0.0, https://physionet.org/content/mimic-ext-mimic-cxr-vqa/1.0.0/.
32. Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., Wu, X.-M.: SLAKE: a Semantically-Labeled Knowledge-Enhanced dataset for medical visual question answering, https://arxiv.org/abs/2102.09542.
33. Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data. 5, (2018). https://doi.org/10.1038/sdata.2018.251.
34. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT, https://arxiv.org/abs/1904.09675.
35. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries, https://aclanthology.org/W04-1013/.