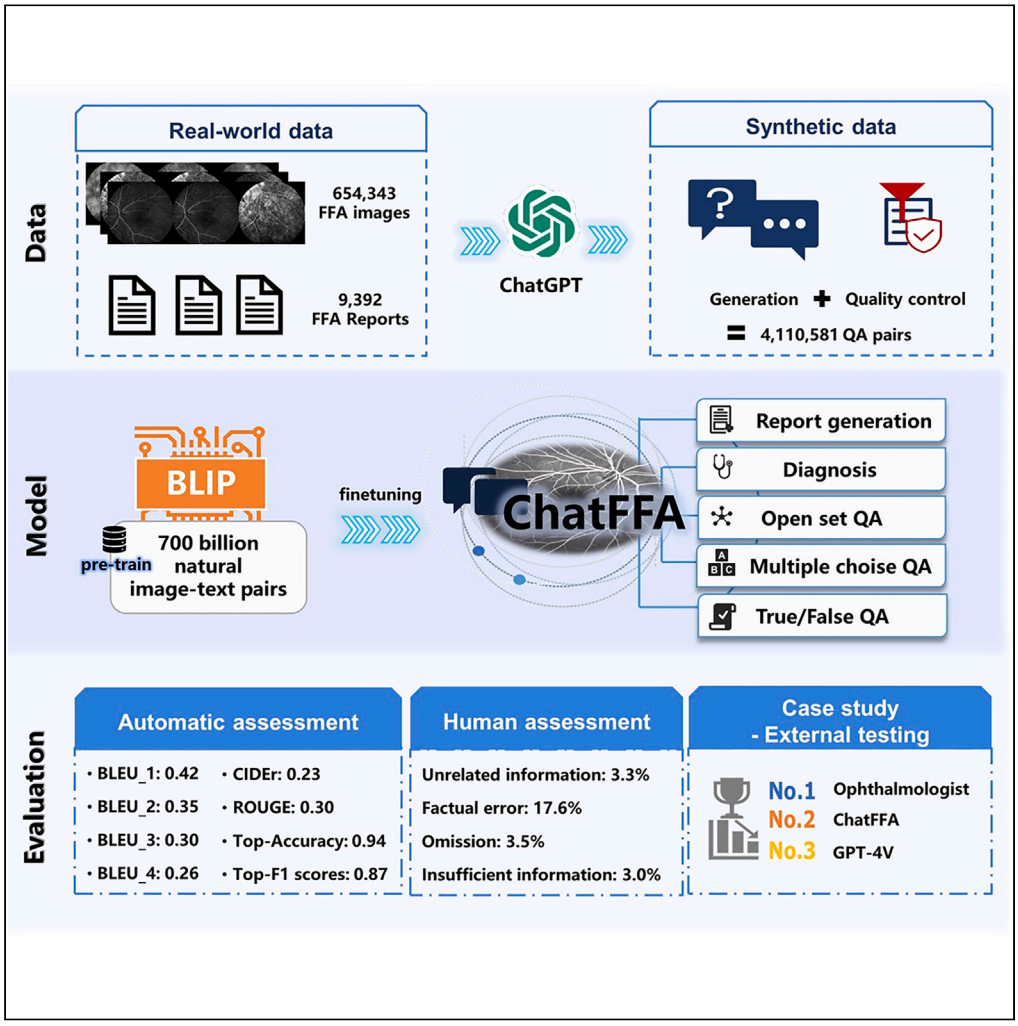
iScience

CellPress
OPEN ACCESS

Article

ChatFFA: An ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography



Xiaolan Chen, Pusheng Xu, Yao Li, Weiyi Zhang, Fan Song, Mingguang He, Danli Shi

danli.shi@polyu.edu.hk

Highlights

ChatFFA is a VQA system for fundus fluorescein angiography image interpretation

ChatFFA trained the BLIP framework using synthetic data for advanced VQA tasks

ChatFFA demonstrates satisfactory results in internal and external validation

ChatFFA shows potential for improving efficiency in medical imaging analysis

Chen et al., iScience 27, 110021
July 19, 2024
https://doi.org/10.1016/j.isci.2024.110021

iScience

CellPress
OPEN ACCESS

Article

ChatFFA: An ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography

Xiaolan Chen,[1,6] Pusheng Xu,[4,6] Yao Li,[5] Weiyi Zhang,[1] Fan Song,[1] Mingguang He,[1,2,3] and Danli Shi[1,2,7,*]

SUMMARY

Existing automatic analysis of fundus fluorescein angiography (FFA) images faces limitations, including a predetermined set of possible image classifications and being confined to text-based question-answering (QA) approaches. This study aims to address these limitations by developing an end-to-end unified model that utilizes synthetic data to train a visual question-answering model for FFA images. To achieve this, we employed ChatGPT to generate 4,110,581 QA pairs for a large FFA dataset, which encompassed a total of 654,343 FFA images from 9,392 participants. We then fine-tuned the Bootstrapping Language-Image Pre-training (BLIP) framework to enable simultaneous handling of vision and language. The performance of the fine-tuned model (ChatFFA) was thoroughly evaluated through automated and manual assessments, as well as case studies based on an external validation set, demonstrating satisfactory results. In conclusion, our ChatFFA system paves the way for improved efficiency and feasibility in medical imaging analysis by leveraging generative large language models.

INTRODUCTION

Ophthalmic images are vital for clinical decision-making as they offer essential diagnostic and prognostic insights regarding patient ocular health.[1] Among them, fundus fluorescein angiography (FFA) is a specialized modality for visualizing retinal vasculature, which can diagnose and monitor eye conditions including retinal vascular occlusion, diabetic retinopathy, and central serous chorioretinopathy with the help of fluorescent dyes.[2] The interpretation of FFA images frequently prompts queries from patients, medical students, and general practitioners, necessitating clarification from ophthalmic specialists for an enhanced understanding of these complex images. However, the scarcity of expert ophthalmologists, coupled with their frequent overwhelming academic and clinical workloads, makes it challenging for them to offer adequate support to patients or the educational needs of students.

Given the complexity of ophthalmic images, researchers have proposed various automated solutions to alleviate the burden on ophthalmic specialists. For instance, Gao et al.[3] proposed a deep learning model for FFA images that can help with prediagnosis assessment and lesion multilevel classification. However, existing models of this kind are either limited to a predetermined set of possible image classifications[4] or confined to text-based question-answering,[5] making it difficult to fully capture all the necessary semantic information about FFA images and limit the expressiveness and quality of the output results.

Generative large language models (LLMs) offer a unique prospect to reconsider medical image interpretation, owing to their extensive external knowledge and powerful cognitive reasoning ability. Among them, Chat Generative Pre-trained Transformers (ChatGPT) is especially compelling.[6] However, the existing state-of-the-art visual question-answering (VQA) model, GPT-4V, falls short in meeting the demands of professional ophthalmic interpretation tasks.[7] Therefore, this study aims to develop an end-to-end unified model that utilizes the power of ChatGPT to significantly enhance semantic comprehension in image analysis and address various VQA tasks associated with FFA images.

RESULTS

We extracted data from FFA reports created by physicians and employed ChatGPT-3.5 to generate question-answering (QA) pairs. These pairs were then utilized to perform fine-tuning and develop an interactive model for multi-tasks, including report generation, disease diagnosis, and VQA. The overview of the study is depicted in Figure 1.

[1]School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong
[2]Research Centre for SHARP Vision (RCSV), The Hong Kong Polytechnic University, Kowloon, Hong Kong
[3]Centre for Eye and Vision Research (CEVR), 17W Hong Kong Science Park, Shatin, Hong Kong
[4]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou 510060, China
[5]University of Waterloo, Computer Science, 200 University Avenue W 0, Waterloo, Canada
[6]These authors contributed equally
[7]Lead contact
*Correspondence: danli.shi@polyu.edu.hk
https://doi.org/10.1016/j.isci.2024.110021

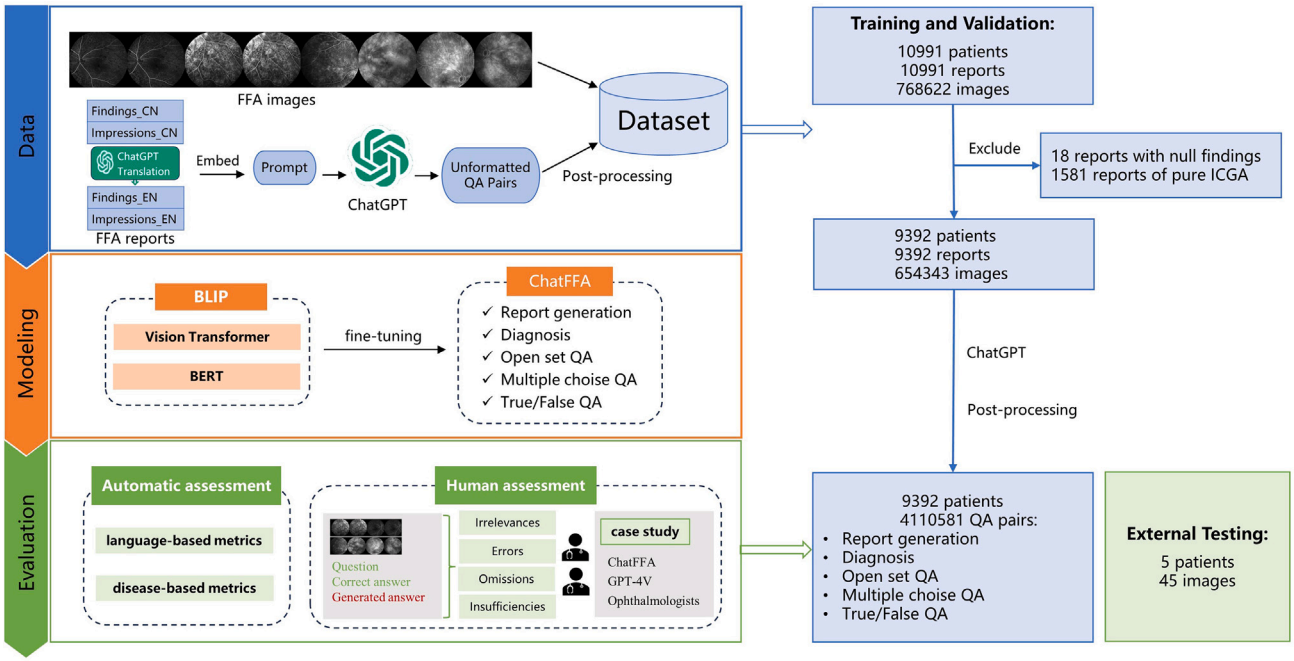


Figure 1. Overview of this study

FFA, fundus fluorescein angiography; CN, Chinese; EN, English; GPT, generative pre-trained transformer; QA, question answering; BLIP, Bootstrapping Language-Image Pre-training; BERT, Bidirectional Encoder Representations from Transformers; ICGA, indocyanine green angiography.

Data

Reports with null findings and images of pure indocyanine green angiography were excluded. The final dataset included 654,343 FFA images alongside 9,392 reports; 38,231 (5.8%) of them were in the arterial phase, 7,888 (1.2%) were in the arterial-venous phase, 366,450 (56.0%) were in the venous phase, and 241,774 (36.9%) were in the late phase. The median (interquartile range) age of the participants was 51 (36, 62) years, and 5,190 (55.3%) were male. Detailed information about the dataset can be found in Table 1.

Most of the participants were diagnosed with multiple retinal conditions, including hypopigmentation (6.0%), microaneurysm (5.0%), hemorrhage (3.3%), atrophy (2.9%), and laser spots (2.8%). There were a total of 172 conditions, and the number of images of these conditions is presented in Figure 2. We generated a total of 4,110,581 QA pairs, with 1,998,359 (48.62%) report generation, 1,108,812 (26.97%) diagnosis, 724,908 (17.64%) open-set QA, 139,278 (3.39%) multiple-choice QA, and 139,224 (3.38%) binary-choice QA.

Figure 3 illustrates the characteristics of the generated questions and answers. Based on the initial words of the questions in both Chinese and English, we conducted clustering and identified various question types. The Chinese question types mainly include inquiries such as "Is the retinal vasculature … " and "What is the fluorescein angiography diagnosis for the left eye … " and so on. The English question types mainly encompass inquiries such as "What is the condition of … " and "Is there any abnormality in … " and so forth. Further analysis of word frequency within the answers revealed that "retina" and "fluorescence" were among the most recurrent terms.

Model performance

Automatic assessment

The language-based and disease-based results of the model are shown in Table 2. Regarding language-based metrics, the Bilingual Evaluation Understudy (BLEU) scores (1–4) were 0.42, 0.35, 0.30, and 0.26, respectively. The model attained a Consensus-based Image Description Evaluation (CIDEr) score of 0.23 and a Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence (ROUGE-L) score of 0.30. In terms of disease-based metrics, the model achieved accuracy of 0.60 and 0.68 and F1 scores of 0.44 and 0.35 in binary and multiple-choice scenarios, respectively. For multi-class condition classification, keyword extraction of the eye conditions was conducted from English report generation and open-set QA types. The top-3 identified conditions were "microaneurysm," "diabetic retinopathy," and "arteriosclerosis." Remarkably, these conditions exhibited high accuracy values of 0.94, 0.94, and 0.87, respectively, indicating the excellent condition recognition performance of the model. Given the naturally imbalanced and long-tailed distribution of medical datasets, the F1 score, which combines precision and recall, is more suitable for comprehensive evaluation. Our model achieved satisfactory F1 scores, with the top three values being 0.87, 0.84, and 0.72.

The investigation into the impact of the number of input images on the model's performance revealed a remarkable improvement when multiple images were used. However, this improvement plateaued when at least four images were provided for the BLEU and ROUGE-L

Table 1. Fundus fluorescein angiography dataset characteristics

	Total	Train	Validation	Test	*p* value
Population					
No.	9,392	5,761	1,603	2,028	
Age, median (IQR)	51 (36, 62)	51 (37, 62)	50 (36, 61)	50 (33, 62)	0.009
Sex, n (%)					0.972
Female	4,202 (44.7)	2,583 (44.8)	714 (44.5)	905 (44.6)	
Male	5,190 (55.3)	3,178 (55.2)	889 (55.5)	1,123 (55.4)	
Year, n (%)					<0.001
2016	622 (6.6)	622 (10.8)	0 (0)	0 (0)	
2017	3,536 (37.6)	3,536 (61.4)	0 (0)	0 (0)	
2019	5,234 (55.7)	1,603 (27.8)	1,603 (100)	2,028 (100)	
FFA images					
No.	654,343	382,621	116,195	155,527	
Phase[a], n (%)					<0.001
Arterial	38,231 (5.8)	24,271 (6.3)	6,960 (6.0)	7,000 (4.5)	
Arterial-venous	7,888 (1.2)	4,866 (1.3)	1,368 (1.2)	1,654 (1.1)	
Venous	366,450 (56.0)	212,707 (55.6)	65,349 (56.2)	88,394 (56.8)	
Late	241,774 (36.9)	140,777 (36.8)	42,518 (36.6)	58,479 (37.6)	

IQR, interquartile range; FFA, fundus fluorescein angiography.
[a]Arterial: 20 to 30 s; Arterial-venous: 30 to 60 s; Venous: 60 s to 5 min; Late: 5 to 10 min.

metrics. As for the CIDEr metric, the performance of the model continued to improve when using more than four images and reached a stable level when 11 images were provided (Figure S1).

Human assessment

A total of 2,692 QA pairs generated by ChatFFA based on 100 images were evaluated by two ophthalmologists. The results of human assessment are shown in Table 2. Rater 1 identified 94 (3.5%) QA pairs as unrelated information, 474 (17.6%) QA pairs with apparent factual errors, 109 (4.0%) QA pairs with omissions, and 86 (3.2%) QA pairs as lacking sufficient information for a conclusive answer. Rater 2 found 81 (3.0%) QA pairs as unrelated information, 475 (17.6%) QA pairs with apparent factual errors, 80 (3.0%) QA pairs with omissions, and 77 (2.9%) QA pairs as lacking sufficient information for a conclusive answer. Examples of different types of errors are presented in Table S1. The inter-rater reliability as measured by kappa values was 0.746, 0.835, 0.743, and 0.741, respectively. After correcting the influence of the aforementioned errors, the specificity, accuracy, precision, and sensitivity of the binary-choice QA types were 0.85, 0.76, 0.84, and 0.67, respectively. And it reached 0.41, 0.43, 0.42, and 0.58, respectively, for multiple-choice QA types. In addition, there was an improvement in the F1 scores for binary-choice and multiple-choice QA pairs, which increased from 0.69 to 0.74 and 0.46 to 0.49, respectively. The model exhibits suboptimal performance in the multiple-choice QA task, even after errors have been corrected. Further analysis indicates that the primary reason for this may be the limited proficiency of the model in handling negation words, resulting in the occurrence of false negatives and false positives.

Case studies

We also performed a qualitative analysis of case studies using an external validation set, AngioReport.[8] The findings demonstrate that our model outperforms GPT-4V[9] in the specific VQA tasks (Table S2). For the direct report generation task, while GPT-4V demonstrated the capability to discern leakage lesions in FFA images, the generated descriptions are generally imprecise and lack essential details. In contrast, our model can mimic human physicians by generating standardized and detailed FFA reports, including accurate disease diagnoses in the impression section. Regarding the diagnosis task, our model successfully identifies pathologic myopia, whereas GPT-4V only provides pattern diagnoses based on general medical knowledge rather than candidate diagnoses specific to the input image. In the multiple-choice task, GPT-4V offers incorrect options, whereas our model accurately identifies the location of the leakage. In the cases of open-ended and true/false questions, both GPT-4V and our model generally provide consistent answers compared to those given by ophthalmologists.

DISCUSSION

In this study, we developed a Transformer-based system for multiple VQA tasks regarding FFA images. Our system demonstrates promising performance, as assessed through both automated and human evaluation, as well as case studies using an external dataset. This proof-of-concept model highlights the potential of ChatGPT in improving the interpretation of FFA images.

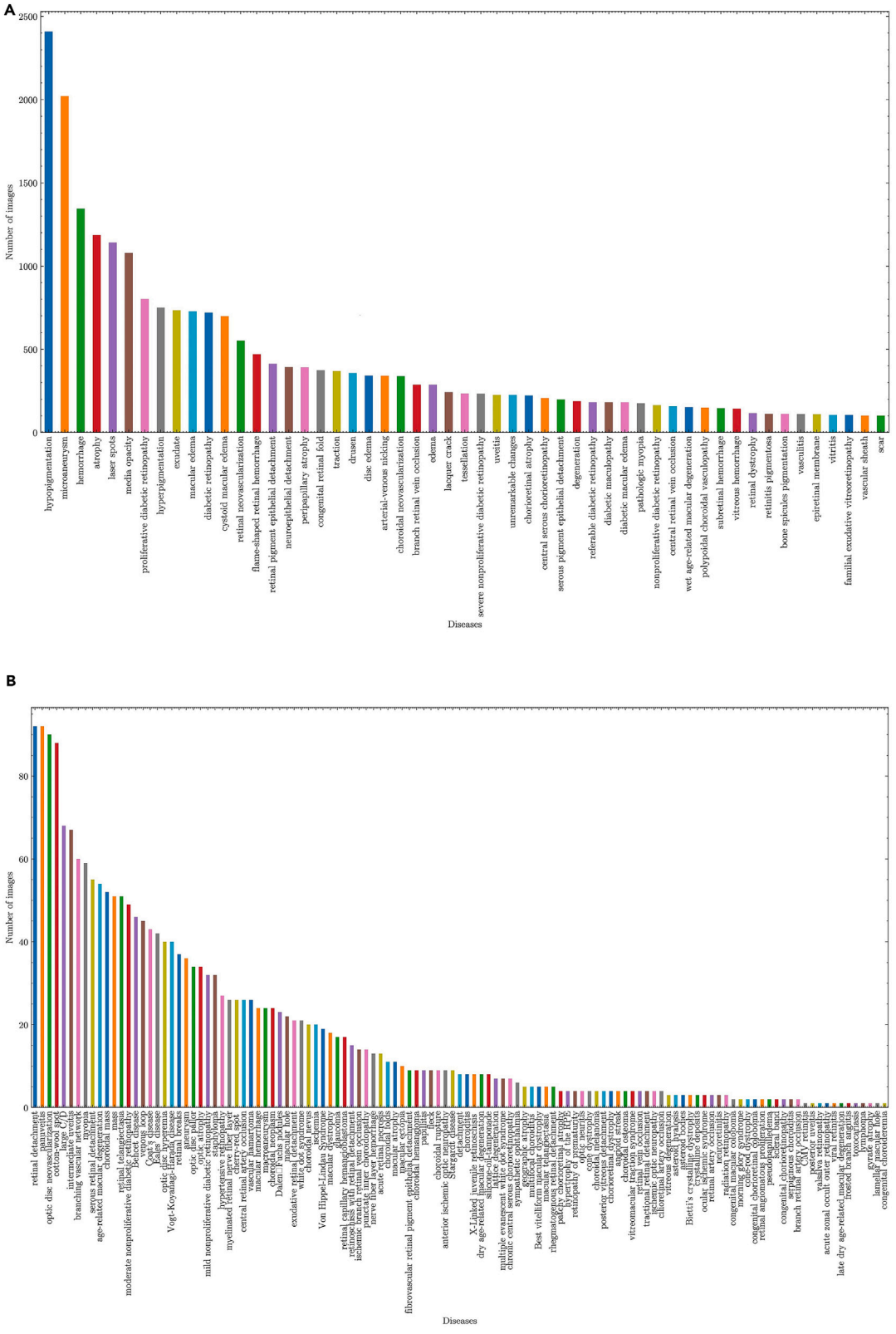


Figure 2. The number of images of various eye conditions

(A) Diseases with more than 100 patients; (B) Diseases with fewer than 100 patients.

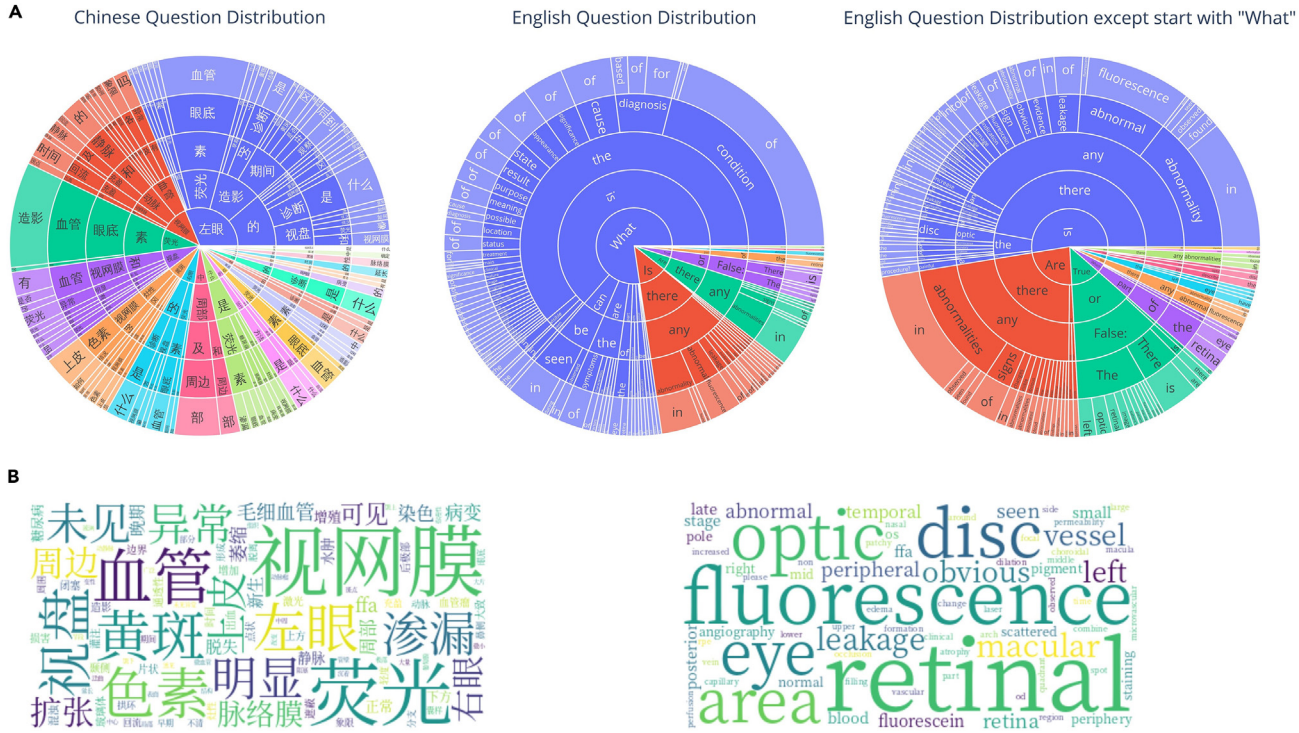


Figure 3. Characteristics of the generated questions and answers

(A) The distribution of English and Chinese-generated questions; (B) The English and Chinese answers' word cloud.

Medical VQA systems assist healthcare professionals in managing medical images by answering image-related questions. However, the construction of large-scale medical VQA datasets is challenging due to the significant involvement of professionals required, including image collection, quality control, and the generation of questions and answers.[10] Recent studies have demonstrated the feasibility of automating the generation of a substantial number of QA pairs through well-designed prompts, significantly enhancing the efficiency of database construction.[11,12] Our study extends this approach to ophthalmic imaging. Specifically, we have devised various prompt strategies for different kinds of questions based on FFA images. These strategies include generating direct report generation questions, closed-ended questions regarding the presence of abnormalities, and open-ended questions that facilitate further exploration of abnormal conditions. Furthermore, we have implemented a rigorous quality control process to ensure the reliability of the generated QA pairs. Our prompt-based approach yields a diverse and highly controllable dataset for training and evaluating VQA models specifically tailored to FFA images, offering time efficiency and wide-ranging real-world applicability.

The VQA model is a classic cross-modal transfer system that combines computer vision and natural language processing (NLP) techniques, providing an intuitive and natural way for human-computer interaction.[13] Features from different modalities could be transferred and utilized.[14,15] Previous studies on optimizing VQA models have mainly focused on cross-modality fusion of visual and language.[16] Among them, Li et al.[17] proposed the Bootstrapping Language-Image Pre-training (BLIP) framework for unifying visual language understanding and generation. The recent advancements in generative LLMs have inspired us to further enhance the performance of VQA models from the perspective of NLP.[18] These models can be directly integrated into the cross-modal model to overcome the limitations in model output and provide further interactive explanations.[19,20] Additionally, they can serve as external sources of knowledge and semantic refiners to optimize the fine-tuning process of cross-modal models. This enables the models to accurately capture details and semantic information in images, facilitating generalization to unseen objects or concepts and performing more complex downstream tasks. Our research is based on the latter approach, where we utilize ChatGPT as an external knowledge base. ChatGPT has been extensively tested and shown promising results in answering ophthalmology questions, including standard ophthalmic exam questions[21] and real-world eye care inquiries.[22] By fine-tuning the BLIP pre-training framework with the assistance of ChatGPT, we enable the model to interpret images and answer image-related questions, even outperforming GPT-4V in specific cases. This enhances the model's image analysis capability, optimizing VQA performance, and unlocking ChatGPT's potential in solving medical problems.

ChatFFA offers a distinct advantage over existing FFA analysis systems in clinical settings. Firstly, it excels in handling vision and language simultaneously and understanding free-form questions. Similar to dynamic and cross-modal brainstorming sessions focusing on FFA images, ChatFFA may serve as a knowledge assistant for medical students or junior clinicians. Given that limited experience can lead to misunderstandings, ChatFFA can support them in interpreting FFA images, thereby decreasing the risk of misdiagnosis.[23] Secondly, ChatFFA can

Table 2. Model performance in the automatic assessment (57,601 QA pairs of 2,028 participants in the test set) and manual assessment (2,692 QA pairs of 100 participants sampled from the test set)

A. Automatic assessment: language-based metrics

BLEU_1	BLEU_2	BLEU_3	BLEU_4	CIDEr	ROUGE-L
0.42	0.35	0.30	0.26	0.23	0.30

B. Automatic assessment: disease-based metrics

	Specificity	Accuracy	Precision	Sensitivity	F1 score
Answer classification					
Binary choice all[a]	0.56	0.60	0.75	0.31	0.44
Binary choice 100[b] before	0.81	0.70	0.80	0.61	0.69
Binary choice 100[b] after	0.85	0.76	0.84	0.67	0.74
Multiple choice all[a]	0.79	0.68	0.37	0.36	0.35
Multiple choice 100[b] before	0.39	0.39	0.39	0.55	0.46
Multiple choice 100[b] after	0.41	0.43	0.42	0.58	0.49
Condition classification					
Microaneurysm	0.96	0.94	0.89	0.86	0.87
Diabetic retinopathy	0.98	0.94	0.79	0.90	0.84
Arteriosclerosis	0.94	0.87	0.68	0.77	0.72

C. Manual assessment: error types of QA assessed by ophthalmologists

	Rater 1 N (%)	Rater 2 N (%)	Kappa
Unrelated information	94 (3.5%)	81 (3.0%)	0.746
Factual error	474 (17.6%)	475 (17.6%)	0.835
Omission	109 (4.0%)	80 (3.0%)	0.743
Insufficient information	86 (3.2%)	77 (2.9%)	0.741

BLEU, bilingual evaluation understudy; CIDEr, consensus-based image description evaluation; ROUGE-L, recall-oriented understudy for gisting evaluation-longest common subsequence.

[a]The metrics were calculated using all the data from the test set.

[b]The metrics were calculated using a randomly sampled subset of the test set, comprising 100 participants.

explain the professional content of FFA reports to patients in simple language and provide basic knowledge about diagnostic results and potential treatment options, empowering patients to actively engage in the clinical process and enhancing doctor-patient communication. Thirdly, ChatFFA can operate offline and be deployed locally in healthcare centers, reducing the risks associated with network data transmission and ensuring data security and privacy.[24] However, there are challenges to consider. Although our dataset includes a diverse range of conditions, there is a lack of representativeness for certain rare diseases, including choroidal hemangioma, Stargardt disease, and others, each with fewer than 10 cases. Additionally, imaging data sourced predominantly from specific manufacturers may limit variance in imaging conditions. To achieve seamless future integration, efforts should be made to expand specific disease datasets, incorporate a broader range of device-based images, and validate these images in diverse national and regional contexts, ensuring compatibility with various health information technology architectures. During the validation and integration process, we should prioritize data protection, patient privacy, and user training to highlight the user-friendliness of the system. Additionally, it is important to emphasize that while ChatFFA provides valuable support and explanations, the ultimate diagnosis and decision-making should still be conducted by trained and experienced healthcare professionals. ChatFFA should be regarded as an auxiliary tool rather than replacing the expertise and judgment of physicians.

In conclusion, we developed and validated a unified VQA system for FFA images by leveraging well-designed prompts and capabilities of ChatGPT. This initiative highlights that ChatGPT is a promising tool in enhancing the interpretation of medical images and ChatFFA demonstrates significant potential in reshaping medical education and clinical management.

Limitations of the study

There are several limitations in the study. Firstly, although our model incorporates information from ophthalmic image reports and the extensive general knowledge in ChatGPT, there are still limitations when it comes to broader medical expertise. Future efforts should focus on integrating prior knowledge from medical literature, clinical guidelines, and the expertise of ophthalmic specialists through reinforcement learning with human feedback to enhance the accuracy of medical VQA. Secondly, while our model generates QA data automatically based

on reports, real-world clinical scenarios often involve more personalized interactions. Therefore, gathering actual clinical dialogues is crucial to achieve a more comprehensive coverage in medical VQA. Finally, while our model underwent automatic and manual evaluations for VQA tasks and was validated using an external dataset, it is necessary to conduct more comprehensive comparisons with a broader range of FFA analysis AI tools and prospective trials involving collaborations with ophthalmologists at different levels of expertise. Furthermore, the model lacks interpretability. There is a need to explore the application of heatmaps, relevant scientific literature references, and authoritative medical websites in future work, subsequently providing clear insights into the decision-making process.[25]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data
  - Modelling
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Automatic assessment
  - Human assessment
  - Case studies
  - Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110021.

ACKNOWLEDGMENTS

D.S. and M.H. disclose support for the research and publication of this work from the Start-up Fund for RAPs under the Strategic Hiring Scheme (grant number: P0048623) and the Global STEM Professorship Scheme (grant number: P0046113) from HKSAR. The sponsor or funding organization had no role in the design or conduct of this research.

AUTHOR CONTRIBUTIONS

D.S. conceived the study. D.S., Y.L., and W.Z. built the deep learning model. X.C. and P.X. did the literature search and analyzed the data. D.S., X.C., P.X., and F.S. contributed to key data interpretation. X.C. wrote the manuscript. All authors have commented on the manuscript. X.C. and P.X. contributed equally in this study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 26, 2023
Revised: April 29, 2024
Accepted: May 15, 2024
Published: May 17, 2024

REFERENCES

1. Sengupta, S., Singh, A., Leopold, H.A., Gulati, T., and Lakshminarayanan, V. (2020). Ophthalmic diagnosis using deep learning with fundus images–A critical review. Artif. Intell. Med. *102*, 101758. https://doi.org/10.1016/j.artmed.2019.101758.
2. Schreur, V., Larsen, M.B., Sobrin, L., Bhavsar, A.R., den Hollander, A.I., Klevering, B.J., Hoyng, C.B., de Jong, E.K., Grauslund, J., and Peto, T. (2022). Imaging diabetic retinal disease: clinical imaging requirements. Acta Ophthalmol. *100*, 752–762. https://doi.org/10.1111/aos.15110.
3. Gao, Z., Pan, X., Shao, J., Jiang, X., Su, Z., Jin, K., and Ye, J. (2023). Automatic interpretation and clinical evaluation for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning. Br. J. Ophthalmol. *107*, 1852–1858. https://doi.org/10.1136/bjo-2022-321472.
4. Pan, X., Jin, K., Cao, J., Liu, Z., Wu, J., You, K., Lu, Y., Xu, Y., Su, Z., Jiang, J., et al. (2020). Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning. Graefe's Archive for Clinical and Experimental Ophthalmology *258*, 779–785. https://doi.org/10.1007/s00417-019-04575-w.
5. Chen, X., Zhao, Z., Zhang, W., Xu, P., Gao, L., Xu, M., Wu, Y., Li, Y., Shi, D., and He, M. (2024). EyeGPT: Ophthalmic Assistant with Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.00840.

6. Introducing ChatGPT. https://openai.com/blog/chatgpt.
7. Xu, P., Chen, X., Zhao, Z., and Shi, D. (2024). Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. Br. J. Ophthalmol. bjo-2023-325054. https://doi.org/10.1136/bjo-2023-325054.
8. Zhang, W., Chotcomwongse, P., Chen, X., Chung, F.H., Song, F., Zhang, X., He, M., Shi, D., and Ruamviboonsuk, P. (2023). Angiographic Report Generation for the 3rd APTOS's Competition: Dataset and Baseline Methods. Preprint at medRxiv. https://doi.org/10.1101/2023.11.26.23299021.
9. GPT-4V(ision) System Card. https://openai.com/research/gpt-4v-system-card.
10. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., and Ge, Z. (2023). Medical visual question answering: A survey. Artif. Intell. Med. *143*, 102611. https://doi.org/10.1016/j.artmed.2023.102611.
11. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., and Xie, W. (2023). PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.10415.
12. Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In The Twelfth International Conference on Learning Representations.
13. Li, Y., Long, S., Yang, Z., Weng, H., Zeng, K., Huang, Z., Lee Wang, F., and Hao, T. (2022). A Bi-level representation learning model for medical visual question answering. J. Biomed. Inform. *134*, 104183. https://doi.org/10.1016/j.jbi.2022.104183.
14. Shi, D., He, S., Yang, J., Zheng, Y., and He, M. (2024). One-shot retinal artery and vein segmentation via cross-modality pretraining. Ophthalmol. Sci. *4*, 100363. https://doi.org/10.1016/j.xops.2023.100363.
15. Shi, D., Zhang, W., He, S., Chen, Y., Song, F., Liu, S., Wang, R., Zheng, Y., and He, M. (2023). Translation of color fundus photography into fluorescein angiography using deep learning for enhanced diabetic retinopathy screening. Ophthalmol. Sci. *3*, 100401. https://doi.org/10.1016/j.xops.2023.100401.
16. Zhang, J., Huang, J., Jin, S., and Lu, S. (2024). Vision-Language Models for Vision Tasks: A Survey. IEEE Trans. Pattern Anal. Mach. Intell. 1–20. https://doi.org/10.1109/TPAMI.2024.3369699.
17. Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, pp. 12888–12900. https://proceedings.mlr.press/v162/li22n.html.
18. Betzler, B.K., Chen, H., Cheng, C.-Y., Lee, C.S., Ning, G., Song, S.J., Lee, A.Y., Kawasaki, R., van Wijngaarden, P., Grzybowski, A., et al. (2023). Large language models and their impact in ophthalmology. Lancet. Digit. Health *5*, e917–e924. https://doi.org/10.1016/S2589-7500(23)00201-7.
19. Chen, X., Zhang, W., Zhao, Z., Xu, P., Zheng, Y., Shi, D., and He, M. (2024). ICGA-GPT: report generation and question answering for indocyanine green angiography images. Br. J. Ophthalmol. https://doi.org/10.1136/bjo-2023-324446.
20. Chen, X., Zhang, W., Xu, P., Zhao, Z., Zheng, Y., Shi, D., and He, M. (2024). FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. NPJ Digit. Med. *7*, 111. https://doi.org/10.1038/s41746-024-01101-z.
21. Mihalache, A., Popovic, M.M., and Muni, R.H. (2023). Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmol. *141*, 589–597. https://doi.org/10.1001/jamaophthalmol.2023.1144.
22. Bernstein, I.A., Zhang, Y.V., Govil, D., Majid, I., Chang, R.T., Sun, Y., Shue, A., Chou, J.C., Schehlein, E., Christopher, K.L., et al. (2023). Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. JAMA Netw. Open *6*, e2330320. https://doi.org/10.1001/jamanetworkopen.2023.30320.
23. Tong, W.-J., Wu, S.-H., Cheng, M.-Q., Huang, H., Liang, J.-Y., Li, C.-Q., Guo, H.-L., He, D.-N., Liu, Y.-H., Xiao, H., et al. (2023). Integration of Artificial Intelligence Decision Aids to Reduce Workload and Enhance Efficiency in Thyroid Nodule Management. JAMA Netw. Open *6*, e2313674. https://doi.org/10.1001/jamanetworkopen.2023.13674.
24. Liu, J., Wang, C., and Liu, S. (2023). Utility of ChatGPT in clinical practice. J. Med. Internet Res. *25*, e48568. https://doi.org/10.2196/48568.
25. Cutillo, C.M., Sharma, K.R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K.D., MI in Healthcare Workshop Working Group, Collier, E., Colvis, C., Gersing, K., et al. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit. Med. *3*, 47. https://doi.org/10.1038/s41746-020-0254-2.
26. Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., De Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al. (2020). A deep learning system for differential diagnosis of skin diseases. Nat. Med. *26*, 900–908. https://doi.org/10.1038/s41591-020-0842-3.
27. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT 2019, *1*, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
28. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. https://openreview.net/forum?id=YicbFdNTTy.
29. Loshchilov, I., and Hutter, F. (2017). Decoupled Weight Decay Regularization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1711.05101.
30. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. *15*, 1–45. https://doi.org/10.1145/3641289.
31. Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. https://doi.org/10.3115/1073083.1073135.
32. Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087.
33. Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pp. 74–81. https://aclanthology.org/W04-1013.
34. Mandrekar, J.N. (2011). Measures of interrater agreement. J. Thorac. Oncol. *6*, 6–7. https://doi.org/10.1097/JTO.0b013e318200f983.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
ChatGPT	OpenAI	https://chat.openai.com/chat
PaddleOCR	Github	https://github.com/PaddlePaddle/PaddleOCR
Bootstrapping Language-Image Pre-training (BLIP)	Github	https://github.com/salesforce/BLIP
Caption evaluation	Github	https://github.com/salaniz/pycocoevalcap
Python (Version 3.1)	Python Software Foundation	https://www.python.org/
R (Version 4.3.1)	R software	https://www.r-project.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dr. Danli Shi (danli.shi@polyu.edu.hk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- De-identified patient standardized data used in this study are available via https://asiateleophth.org/cross-country-datasets/ upon request. Additional datasets related to this research can be provided by the lead contact upon request.
- Code is available at https://github.com/salesforce/BLIP.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

The study took place between July 26th, 2023 and September 26th, 2023 at the School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China.

Data

FFA images were retrospectively collected from a tertiary center between 2016 and 2019.[20] After data cleaning by removing duplicates, handling missing values, and correcting erroneous data, a total of 654,343 FFA images from 9,392 patients were used for model development. All patient data were anonymized and de-identified, as well as stored in a secure data center, managed and analyzed by authorized personnel. FFA images were obtained using Zeiss FF450 Plus and Heidelberg Spectralis (Heidelberg, Germany) cameras, with a resolution of 768*768. The image data normalization process involves using the ImageNet mean and standard deviation values to standardize the data, resizing the images to a fixed resolution of 320*320 and converting them into PyTorch tensors. This consistent normalization was applied during both training and testing for a uniform data distribution. Additionally, we excluded low-quality images, with a vascular ratio of less than 0.05. After filtering, we obtained 768,622 images from the original dataset of 820,853. These filtered images were then matched with the FFA reports, yielding 654,343 matched FFA images.

For training, we selected images from the period of 2016 to 2017. The images from January to May 2019 were chosen for validation, while the remaining images were allocated for testing. This temporal split strategy emulates a scenario where a model is developed using historical data and then assessed on future cases, thus serving as external validation.[26]

The study adheres to the tenets of the Declaration of Helsinki. The Institutional Review Board of the Hong Kong Polytechnic University approved the study. Informed consent was waived as the data were retrospectively collected and de-identified.

Data construction

All FFA reports typically comprise two main sections: findings and impressions. These reports contain crucial information, including the examination location, detailed descriptions of the angiography process and clinical indications. We used PaddleOCR to extract the data from the original Chinese reports, denoted as Findings_CN and Impression_CN. To generate bilingual (Chinese and English) versions of the

reports, we employed ChatGPT to translate Findings_CN and Impressions_CN from Chinese into English using prompt strategy, denoted as Findings_EN and Impressions_EN. The translation prompt can be found in Table S3A.

Subsequently, we used the extraction as initial inputs for ChatGPT and generated five types of questions:

(1) Direct report generation: Generating the complete angiography report directly from the image.
(2) Diagnosis questions: Generating diagnostic questions related to the images in the report, which usually provides one or more potential diagnoses based on the impressions of reports.
(3) Open-ended questions: Generating free-text questions that can guide a more detailed description or provide additional information about the report content.
(4) Multiple-choice questions: Generating multiple-choice questions that provide several options for selection, where the respondent can choose one correct answer.
(5) True or false questions: Generating binary questions that require the respondent to determine whether a given statement is true or false.

Each type of question was generated in both Chinese and English to obtain a comprehensive bilingual dataset. For the open-ended QA pairs, we generated ten sets following the prompt illustrated in Table S3B. Our prompt strategy for closed-ended QA pairs drew inspiration from the process established by Zhang et al.[11] Specifically, we modified the method and utilized ChatGPT-3.5 to generate ten sets of bilingual closed-ended QA pairs. These pairs consisted of five sets of multiple-choice questions, each containing four options (Question_EN, Options_EN, Answer_EN, Question_CN, Options_CN, Answer_CN), and five sets of true/false questions, following a format similar to the open-ended QA pairs. The modified prompt we used is shown in Table S3C.

Post-processing and quality control

We extracted the pertinent data from the raw outputs produced by ChatGPT and carefully filtered out QA pairs in incorrect formats. Since ChatGPT could fabricate responses that contained factual errors, we excluded QA pairs that involved causal inference. Furthermore, we removed pairs that contained information such as visual acuity and prognosis, as these aspects are hardly inferred from the provided report alone. To ensure a balanced representation of disease conditions and QA types, we sampled and balanced the QA categories.

Modelling

Our model incorporated the BLIP framework,[17] which consists of two main components: Bidirectional Encoder Representations from Transformers (BERT) and Vision Transformer (ViT). BERT is a pre-trained language model based on the Transformer architecture.[27] Unlike traditional unidirectional language models, BERT uses a bidirectional Transformer encoder to effectively capture contextual information during the pre-training phase. BERT is trained on large-scale text corpora to learn universal language representations, which can then be fine-tuned for various NLP tasks. In this study, we used it as the language encoder and decoder. ViT, introduced by Dosovitskiy et al.,[28] is an advanced architecture for CV tasks. It divides an image into patches and treats them as tokens, which are then inputted into a series of Transformer layers. After pre-training, ViT can be fine-tuned for specific downstream tasks such as image classification, object detection, or image segmentation. In our study, we used it as the image encoder.

We finetuned the pre-trained model on FFA images alongside the generated QA pairs. Input images of size 320*320 were fed to the encoder, and we employed the AdamW[29] optimizer during the fine-tuning process. The initial learning rate was set at 0.00002, accompanied by a weight decay of 0.05, and a cosine learning rate schedule. The fine-tuning was conducted with a maximum epoch of 20 using two NVIDIA Tesla V100 GPUs, the model with the highest BLEU1 score on the validation set was selected for testing.

QUANTIFICATION AND STATISTICAL ANALYSIS

Automatic assessment

Language-based metrics are widely used in previous NLP tasks.[30] We utilized metrics such as BLEU, CIDEr and ROUGE-L. BLEU[31] calculates the similarity between generated and reference sentences by measuring n-gram matching and we used 1-4 grams considering the domain-specific terminology of medicine. CIDEr[32] is a precision-and-recall metric that considers word frequency, repetition and phrase diversity to accurately evaluate the quality of generated descriptions. ROUGE-L[33] measures the longest common subsequence of word matches. These metric values range between 0 and 1, with higher values indicating better performance. Additionally, we conducted a study to investigate the impact of varying numbers of input images on the model's performance. We inputted one to twelve images into the model and observed the trend in the model's performance.

Disease-based metrics are commonly employed to overcome the limitations of language-based metrics in medical abnormality detection. In our evaluation, we utilized a classification pipeline that involved extracting diagnostic conditions using a manually constructed dictionary mapping for both the original and generated reports. The evaluation metrics include accuracy, sensitivity, specificity, precision, and F1 score.

Human assessment

To address evaluation aspects that are not covered by automatic assessment, we conducted an extensive human evaluation of the QA generated by ChatFFA. We randomly sampled 100 FFA images from the test set. Two ophthalmologists (P.X. and F.S.) assessed the quality of the

ChatFFA-generated QA pairs based on the ground truth and their expert judgment of the FFA images. The QA pairs were analyzed to identify occurrences of unrelated information (Irrelevances), factual errors (Errors), incomplete information (Omissions) and insufficiency of information to arrive at an answer (Insufficiencies).

Case studies

To further investigate the quality of answers generated by our model, we conducted a qualitative case study using the AngioReport[8] dataset. This dataset encompasses retrospectively collected cases from another center, thereby offering additional external validation. Specifically, we selected five cases covering five types of VQA tasks. For comparison, we selected the state-of-the-art VQA model, GPT-4V.[9] We used the answers provided by ophthalmologists as a benchmark and compared the answers generated by the two models. To address the issue of lengthy answers commonly provided by the models, we imposed a constraint on both to generate short, concise answers without lengthy explanations.

Statistical analysis

Statistical analyses were performed using R (Version 4.3.1, R Foundation, Vienna, Austria). To evaluate the differences in features of data from three datasets (training set, validation set and test set), we conducted the Kruskal-Wallis rank sum test and Dunn's multiple comparison post hoc test. To examine the reliability of the manual scoring results, we assessed the rating consistency between two ophthalmologists using Cohen's Kappa.[34] Its values represent different levels of agreement: 0.01 to 0.20 (slight agreement), 0.21 to 0.40 (fair agreement), 0.41 to 0.60 (moderate agreement), 0.61 to 0.80 (substantial agreement) and 0.81 to 0.99 (almost perfect agreement). $P < 0.05$ was considered statistically significant.