

Original Paper

Slit Lamp Report Generation and Question Answering: Development and Validation of a Multimodal Transformer Model with Large Language Model Integration

Ziwei Zhao^{1*}, MD; Weiyi Zhang^{1*}, MS; Xiaolan Chen¹, MD; Fan Song¹, MD; James Gunasegaram², MD; Wenyong Huang³, MD, PhD; Danli Shi^{1,4*}, MD, PhD; Mingguang He^{1,4,5*}, MD, PhD; Na Liu^{6*}, MD

¹School of Optometry, The Hong Kong Polytechnic University, Hong Kong, China

²Monash University, Victoria, Australia

³Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

⁴Research Centre for SHARP Vision, The Hong Kong Polytechnic University, Hong Kong, China

⁵Centre for Eye and Vision Research (CEVR), Hong Kong, China

⁶Guangzhou Cadre and Talent Health Management Center, Guangzhou, China

*these authors contributed equally

Corresponding Author:

Na Liu, MD

Guangzhou Cadre and Talent Health Management Center

No. 109 Changling Road

Huangpu District

Guangzhou, 510700

China

Phone: 86 18701985445

Email: 1256695904@qq.com

Abstract

Background: Large language models have shown remarkable efficacy in various medical research and clinical applications. However, their skills in medical image recognition and subsequent report generation or question answering (QA) remain limited.

Objective: We aim to finetune a multimodal, transformer-based model for generating medical reports from slit lamp images and develop a QA system using Llama2. We term this entire process slit lamp–GPT.

Methods: Our research used a dataset of 25,051 slit lamp images from 3409 participants, paired with their corresponding physician-created medical reports. We used these data, split into training, validation, and test sets, to finetune the Bootstrapping Language-Image Pre-training framework toward report generation. The generated text reports and human-posed questions were then input into Llama2 for subsequent QA. We evaluated performance using qualitative metrics (including BLEU [bilingual evaluation understudy], CIDEr [consensus-based image description evaluation], ROUGE-L [Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence], SPICE [Semantic Propositional Image Caption Evaluation], accuracy, sensitivity, specificity, precision, and F_1 -score) and the subjective assessments of two experienced ophthalmologists on a 1-3 scale (1 referring to high quality).

Results: We identified 50 conditions related to diseases or postoperative complications through keyword matching in initial reports. The refined slit lamp–GPT model demonstrated BLEU scores (1-4) of 0.67, 0.66, 0.65, and 0.65, respectively, with a CIDEr score of 3.24, a ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score of 0.61, and a Semantic Propositional Image Caption Evaluation score of 0.37. The most frequently identified conditions were cataracts (22.95%), age-related cataracts (22.03%), and conjunctival concretion (13.13%). Disease classification metrics demonstrated an overall accuracy of 0.82 and an F_1 -score of 0.64, with high accuracies (≥ 0.9) observed for intraocular lens, conjunctivitis, and chronic conjunctivitis, and high F_1 -scores (≥ 0.9) observed for cataract and age-related cataract. For both report generation and QA components, the two evaluating ophthalmologists reached substantial agreement, with κ scores between 0.71 and 0.84. In assessing 100 generated reports, they awarded scores of 1.36 for both completeness and correctness; 64% (64/100) were considered “entirely good,” and 93% (93/100) were “acceptable.” In the evaluation of 300 generated answers to questions, the scores were 1.33 for completeness, 1.14 for correctness, and 1.15 for possible harm, with 66.3% (199/300) rated as “entirely good” and 91.3% (274/300) as “acceptable.”

Conclusions: This study introduces the slit lamp–GPT model for report generation and subsequent QA, highlighting the potential of large language models to assist ophthalmologists and patients.

(*J Med Internet Res* 2024;26:e54047) doi: [10.2196/54047](https://doi.org/10.2196/54047)

KEYWORDS

large language model; slit lamp; medical report generation; question answering

Introduction

The slit lamp, a cornerstone in ophthalmology, allows for detailed examination of the eye's anterior segment [1]. Using an illuminated, narrow beam, this noninvasive method facilitates the evaluation of abnormalities by depth and size. While instrumental in diagnosing common eye diseases such as keratitis, conjunctivitis, conjunctival concretions, and cataracts, interpreting slit lamp results can be challenging for primary care physicians due to the need for specialized training. This can result in overlooked abnormalities or misdiagnosis. Furthermore, ophthalmologists are tasked with interpreting, documenting, and effectively communicating these results to patients, a time and effort-intensive process. The scarcity of experienced ophthalmologists, particularly in rural areas, further exacerbates the situation [2].

Artificial intelligence (AI) and large language models (LLMs) have made significant strides in the medical field, enhancing the capabilities of health care professionals in interpreting, and analyzing medical imagery. For instance, AI has been instrumental in advancing the analysis of x-rays [3], magnetic resonance images [4], ultrasounds [5], and dermatological images [6]. Generative pretrained transformers (GPT) models such as ChatGPT [7] and Llama2 [8], have showcased remarkable capabilities in problem-solving scenarios across a spectrum of medical applications. These AI models are instrumental in streamlining clinical documentation [9], refining patient communication [10], aiding administrative tasks [11], enriching textual data [12], and bolstering evidence-based decision-making [13]. Their versatility extends to comprehensive patient assessments [14], precise disease diagnostics [15], informed treatment proposals [16], meticulous medical writing [17], innovative teaching methodologies [18], and robust question answering (QA) systems [19], embodying a multifaceted impact on the health care industry.

Deep learning strategies currently used to transform images into high-quality features include convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformer networks, and their variants such as long short-term memory (LSTM) and gated recurrent units (GRUs). CNNs are often combined with other networks such as RNNs to generate text [20]. RNNs and their variants, recognized for their prowess in

handling sequential data, account for element dependencies within sequences. Despite their effectiveness, RNNs face challenges with extended sequences and potential gradient issues which are mitigated by long short-term memories and GRUs through a gate mechanism. Transformer networks, proposed in 2017, use self-attention mechanisms to manage long sequences and parallel computations, thus boasting swift training speed at the cost of substantial computational resources [21]. Bootstrapping Language-Image Pre-training (BLIP), a hybrid approach leveraging transformer networks' architecture and amalgamating natural language processing and computer vision, enhances model performance via pretraining. BLIP's principal strength lies in its multimodal capacity to concurrently handle image and text data, allowing it to excel in specific tasks such as image description generation.

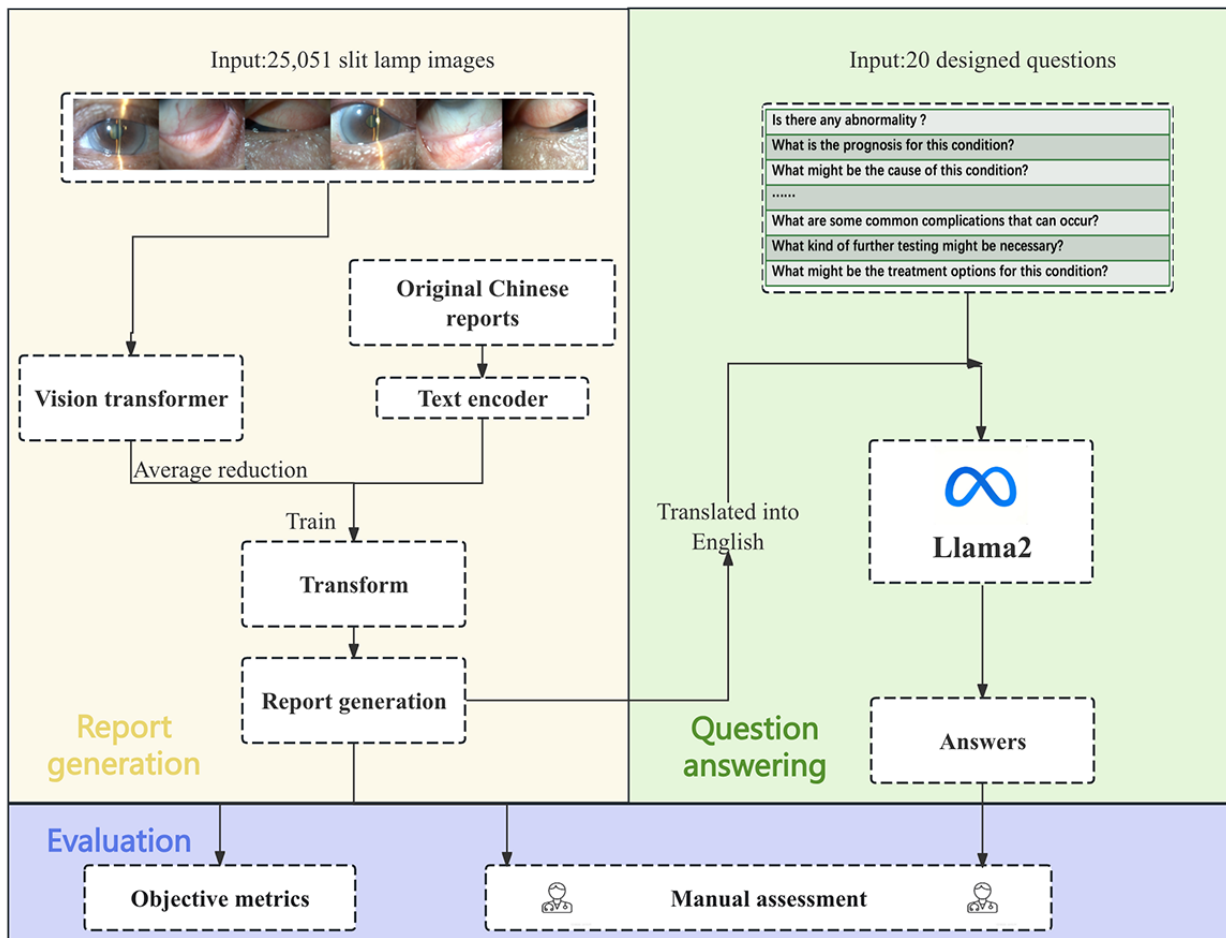
In the specific context of slit lamp imaging augmented with AI, research has primarily concentrated on individual disease detection and grading, such as in the case of cataracts [22,23], pterygium [24], and infectious keratitis [25]. However, there is a noticeable lack of a unified system that uses slit lamp images for the generation of systematic anterior segment reports and QA. While the advent of OpenAI's GPT-4V offered the possibility of image-based AI medical dialogue, its direct clinical application has been limited by inaccuracies and the generation of unreliable information, which was termed "hallucinations" [26,27]. Additionally, due to its closed-source nature, there is a constraint on the fine-tuning ability, which is paramount for medical applications. In response to this, our study has used Llama2, an open-source model, to harness the anticipated benefits of a specialized LLM tool that ensures enhanced control and reliability in the subsequent QA scenarios. Based on our experience in ophthalmic QA tasks and LLMs, including fundus fluorescein angiography and indocyanine green angiography QA [12,28,29] we aim to extend these methodologies to slit lamp imaging by developing a novel slit lamp–GPT system, using BLIP and LLMs specifically tailored for ophthalmology, with dual objectives: to generate reports and to facilitate QA.

Methods

Overview

The flow of our study is outlined in [Figure 1](#).

Figure 1. Flow diagram of this study.



Dataset

We collected data from a Chinese physical examination center for this retrospective study, which included both essential clinical information and annual slit lamp images. We included slit lamp photographs with corresponding medical reports, excluding any of inadequate quality. This study used data collected from a previous study [30], all participants’ information was deidentified per the Declaration of Helsinki’s guidelines. All slit lamp images, captured via a Haag-Streit BQ-900 at a 2048×1536-pixel resolution, included at least 4 images per participant showcasing the pupil, upper eyelid, and lower eyelid. Initial reports, written by ophthalmologists in Chinese, contained disease diagnoses, recommendations, or detailed descriptions of ocular signs. A subset of representative reports from the dataset was selected for translation into English to form a bilingual dataset.

Model Construction

Similar to other studies [28,29], we initially trained and tested the BLIP [28] network for report generation. Subsequently, the generated reports from the test set were input into Llama2 for QA validation, further evaluating the quality and practicality of the reports.

During the report generation phase, we used the BLIP framework, a multimodal transformer model skilled at aligning

visual interpretation with text generation. The model filtered out noisy data during training and generated slit lamp reports from paired images and text inputs. Our design incorporated a vision transformer [31] and BERT [32] as the image and language encoder and decoder, respectively. The vision transformer converts an image into encoded patch sequences, while BERT, trained on extensive unlabeled text data, enables deep contextualized representation learning. The pretrained BLIP model was fine-tuned using slit lamp images and associated reports, with each case providing at least four images during training, resized to 224×224 pixels. We applied the AdamW optimizer (the University of Freiburg), using an initial learning rate of 0.00002, a weight decay of 0.05, and a cosine learning rate schedule, across 50 epochs on one NVIDIA Tesla V100 GPU (NVIDIA Corp). The model with the highest BLEU1 (bilingual evaluation understudy) score (detailed in the performance evaluation part) on the validation set was selected for testing.

For the question and answering phase, we created a question set related to slit lamp examination and reporting based on prior studies [33] and our clinical expertise. These questions, along with the corresponding reports, were seamlessly input into the Llama2 model. This integration allowed for QA without the need for fine-tuning, while enhancing the interpretation of the generated reports. The process involved instructing the model

using a specific prompt: “Answer based on: [slit lamp report content here].”

Performance Evaluation

We used both language-based and disease classification metrics for quantitative evaluations of report quality, supplemented by manual assessments for report generation and QA.

For language-based metrics, we used BLEU [34], CIDEr [35], ROUGE-L [36]), and Semantic Propositional Image Caption Evaluation [37], each with its strengths. However, traditional language metrics may be less dependable for medical conditions due to the infrequent occurrence of disease-related keywords in reports. To address this, we introduced a classification evaluation procedure that used a manually curated dictionary to identify disease-related conditions or postoperative statuses from both original and generated reports. Disease classification metrics, such as specificity, accuracy, precision, sensitivity, and the F_1 -score, provided a comprehensive performance review of the model.

Considering the complexity of medical terminology and the potential harm of inaccurate reporting, manual assessment remains crucial. For report generation, 100 test set cases were randomly selected and independently evaluated by 2 ophthalmologists (ZZ and FS) using a 3-point scale, focusing on “completeness” (how well the generated reports matched the ground truth conditions) and “correctness” (the accuracy of diagnosis and condition descriptions). Scores ranged from 1 (excellent) to 3 (poor), with 2 representing an acceptable rating. The final score was the average of the scores from the 2 evaluators. For QA, 20 prepared human-posed questions and the translated report were put into Llama2 to generate answers, which were evaluated based on “completeness,” “correctness,” and “possible harm.” Scores ranged from 1 (recommendable to

patients) to 3 (not recommendable for patients), with 2 indicating that minor adjustments could make the answer suitable for recommendation. The average score was also used as the final score. For detailed scoring criteria in these 2 sections, refer to Table S1 in [Multimedia Appendix 1](#).

Ethical Considerations

This study used data collected from a previous study [30]. All patient data were anonymized and de-identified following the Declaration of Helsinki. Individual consent was waived due to the retrospective nature and the thorough anonymization process of the study. The Institutional Review Board of the Hong Kong Polytechnic University approved the study (HSEARS20240301004).

Results

Data

Our final dataset includes 25,051 slit-lamp images and 3409 reports. Most images (12,496, 49.89%) focus on the cornea, with 32.74% (n=8202) on the upper eyelid and 17.38% (n=4353) on the lower eyelid. The median age of participants is 65, with an IQR of 60 to 72 years, and the majority (2009/3409, 58.93%) are male. The demographics and image types are similar across all sets.

The distribution of images across years is as follows: 1257 (5.02%) from 2013, 12,206 (48.72%) from 2015, and 11,588 (46.26%) from 2016. The 2013 and 2015 images form the training set, while the 2016 images are partitioned evenly into validation and testing sets. There were no significant differences in demographic characteristics and positioning type between these datasets. [Table 1](#) provides a comprehensive overview of the dataset characteristics.

Table 1. Slit lamp images: dataset characteristics.

| | Total | Train | Validation | Test | P value |
|--|----------------------|----------------------|---------------------|----------------------|---------|
| Participants | | | | | |
| Number | 3409 | 1846 | 781 | 782 | |
| Age, median (Q1 ^a , Q3 ^b) | 65.46 (60.52, 72.47) | 65.31 (60.12, 72.13) | 62.04 (59.58, 65.9) | 71.04 (65.47, 77.03) | <.001 |
| Sex, n (%) | | | | | .002 |
| Male | 2009 (58.93) | 1101 (59.64) | 420 (53.78) | 488 (62.4) | |
| Female | 1400 (41.07) | 745 (40.36) | 361 (46.22) | 294 (37.6) | |
| Slit lamp images | | | | | |
| Number | 25,051 | 13,463 | 5987 | 5601 | |
| Position, n (%) | | | | | .002 |
| Upper eyelid | 8202 (32.74) | 4315 (32.05) | 2046 (34.2) | 1841 (32.9) | |
| Lower eyelid | 4353 (17.38) | 2423 (18) | 951 (15.9) | 979 (17.5) | |
| Cornea | 12,496 (49.89) | 6725 (49.95) | 2990 (49.9) | 2781 (49.7) | |

^aQ1: first quartile.

^bQ3: third quartile.

We used a custom dictionary to extract diagnoses and physical signs by keyword matching from the Chinese reports. We

identified 50 conditions, including age-related cataracts (478/2170, 22.03%), cataracts (498/2170, 22.95%), conjunctival

concretion (285/2170, 13.13%), after intraocular lens implantation (151/2170, 6.96%), pterygium (144/2170, 6.64%), conjunctivitis (97/2170, 4.47%), chronic conjunctivitis (93/2170, 4.29%), and other eye conditions with lower proportions. This led to 1377 Chinese reports primarily featuring diagnostic terms or descriptions of ocular signs.

Quantitative Model Performance

Language-based metrics are provided in Table 2, with BLEU (1-4) scores (0.67, 0.66, 0.65, and 0.65) indicating good lexical accuracy and a ROUGE-L score of 0.61 highlighting effective content retention. The CIDEr score of 3.24 reflects its ability to align closely with human judgment on sentence quality, while a SPICE score of 0.37 demonstrates moderate success in

capturing complex semantic relationships. For disease classification metrics (see Table 3), our model achieved a weighted accuracy of 0.82 and a weighted F_1 -score of 0.64. However, performance varied across diseases. It was highly accurate (≥ 0.9) for conditions of intraocular lens, conjunctivitis, and chronic conjunctivitis, and had high F_1 -scores (≥ 0.9) for cataracts and age-related cataracts. The model demonstrated excellent accuracy for positive cases of cataracts and age-related cataracts. Despite high accuracy, specificity, and precision for postoperative intraocular lens implantation, sensitivity was relatively low: a clinically acceptable trade-off. However, for conjunctival concretions, conjunctivitis, and chronic conjunctivitis, the model's overall predictive capacity fell short.

Table 2. Language-based metrics of report generation in the test set (5601 images from 782 participants).

| BLEU_1 ^a | BLEU_2 ^a | BLEU_3 ^a | BLEU_4 ^a | CIDEr ^b | ROUGE ^c | SPICE ^d |
|---------------------|---------------------|---------------------|---------------------|--------------------|--------------------|--------------------|
| 0.67 | 0.66 | 0.65 | 0.65 | 3.24 | 0.61 | 0.37 |

^aBLEU: bilingual evaluation understudy.

^bCIDEr: consensus-based image description evaluation.

^cROUGE: Recall-Oriented Understudy for Gisting Evaluation.

^dSPICE: Semantic Propositional Image Caption Evaluation.

Table 3. Disease classification metrics of report generation in the test set.

| Condition | Specificity | Accuracy | Precision | Sensitivity | F_1 -score |
|-------------------------------------|-------------|----------|-----------|-------------|--------------|
| Age-related cataract | 0.6 | 0.79 | 0.9 | 0.8 | 0.84 |
| Cataract | 0.58 | 0.78 | 0.9 | 0.79 | 0.84 |
| After intraocular lens implantation | 0.93 | 0.94 | 0.96 | 0.48 | 0.64 |
| Conjunctival concretion | 0.83 | 0.7 | 0.37 | 0.44 | 0.4 |
| Chronic conjunctivitis | 0.92 | 0.9 | 0.34 | 0.15 | 0.2 |
| Conjunctivitis | 0.92 | 0.9 | 0.34 | 0.14 | 0.2 |

Qualitative Model Performance

Overview

The score distribution is depicted in Figure S1 in Multimedia Appendix 2.

Report Generation

Two ophthalmologists scored the model highly for completeness (mean 1.36, SD 0.61, $\kappa=0.84$) and correctness (mean 1.36, SD 0.59, $\kappa=0.72$). Reports that received a score of 1 for both completeness and correctness were defined as entirely good and constituted 64% (64/100) of the evaluated reports. Reports that scored either 1 or 2 by both reviewers for both completeness and correctness were deemed acceptable, representing 93% (93/100) of the reports. These scores primarily corresponded to reports detailing specific conditions such as cataracts, age-related cataracts, and negative findings. However, 7% (7/100) of reports scored a 3, indicating deficiencies.

We discovered that lower scores were linked to issues such as limited sample sizes for specific diseases, difficulties in clearly identifying lesions, and challenges in interpreting diseases or signs from images due to the unique aspects of slit lamp

photography. These complications were common in conditions such as xanthomas, trichiasis, after-glaucoma surgery, lagophthalmos, and some small conjunctival concretions. Additionally, images not focused on the cornea made it difficult to detect corneal lesions.

Through our hands-on evaluation, we noticed that the model sometimes added diagnoses that were not in the original reports but were still acceptable based on the images. For example, it sometimes diagnosed mild cataracts even when the images did not show apparent lens abnormalities. We considered these decisions acceptable when considering the challenge faced by an ophthalmologist in making a precise distinction based solely on images.

About QA

Our constructed questionnaires included 20 items, addressing a breadth of topics such as diagnosis, pathologic localization, severity grading, visual impairment, prognosis, associated complications, therapeutic recommendations, suggested further examinations, preventive advice, and scientific education pertinent to slit lamp examination (Table S2 in Multimedia Appendix 1).

We selectively curated 15 representative English reports on conditions including cataracts, conjunctival concretions, conjunctivitis, postintraocular lens implantation, and pterygium, as well as their mixed states. Each report contributed 20 questions, culminating in a total of 300 questions.

Our model scored well on completeness (1.33, $\kappa=0.84$), correctness (1.14, $\kappa=0.71$), and possible harm (1.15, $\kappa=0.82$). Similarly, QA responses that scored a 1 in completeness, correctness, and possible harm were defined as entirely good, representing 66.3% (199/300) of the QA responses. Responses scoring either 1 or 2 across these categories were considered acceptable, comprising 91.3% (274/300). Less than 9% (26/300) of the 300 questions scored a 3 in any category. These were

typically related to reports focusing more on physical signs than diagnoses or conditions and statements about binocular intraocular lenses. Figure S2 in [Multimedia Appendix 3](#) provides examples of generated answers with different scores.

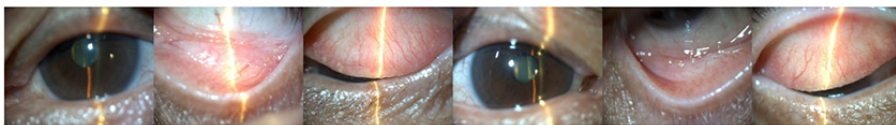
Discussion

Principal Findings

Our study introduces a novel method for analyzing slit lamp images through the integration of a multimodal transformer with an LLM. This approach has enabled the accurate identification of common anterior segment eye diseases and supports a QA system that directly addresses symptoms, diagnosis, and treatment options, as illustrated in [Figure 2](#).

Figure 2. Demonstration of the question-answering system. (A) Input image, ground truth, and model prediction. (B) Question answering. Blue highlight: corresponds to accurate diagnosis matches. Yellow highlight: supplementary predicted information (not in the manual report but correct).

(A)

| Input | |
|--|--|
|  | |
| Ground Truth | |
| 1. 双眼慢性结膜炎 2. 双眼年龄相关性白内障。建议:1.注意休息,注意用眼卫生,不适时可点抗生素眼水。2.多吃深绿色新鲜蔬菜、多喝水少食盐、补充维生素,注意养生。每年复查白内障情况。 | |
| Prediction | |
| 1. 双眼慢性结膜炎 2. 双眼结膜结石 3. 双眼年龄相关性白内障。建议:1.注意休息,注意用眼卫生,不适时可点抗生素眼水。2.结膜结石一般无须特殊处理,如异物感则剔除结膜结石后点抗生素眼水及眼膏。3.多吃深绿色新鲜蔬菜、多喝水少食盐、补充维生素,注意养生。每年复查白内障情况。 | |
| Translation | |
| 1. Chronic conjunctivitis in both eyes. 2. Conjunctival calculi (conjunctival stones) in both eyes. 3. Age-related cataracts in both eyes. | |
| Recommendations: | |
| 1. Pay attention to rest, maintain eye hygiene, and use antibiotic eye drops if discomfort occurs. 2. Conjunctival calculi generally do not require special treatment. If there is a foreign body sensation, the calculi can be removed, followed by the application of antibiotic eye drops and ointment. 3. Eat more dark green fresh vegetables, drink plenty of water, reduce salt intake, supplement vitamins, and pay attention to health. Have regular check-ups each year to monitor the condition of the cataracts. | |

(B)

Intelligent assistant

Chatbot

What might be the treatment options for this condition?

Based on the imaging information provided in the slit-lamp images, the following are some possible treatment options for the patient's condition:

- Chronic conjunctivitis in both eyes:
 - Antibiotic eye drops: The patient can be prescribed antibiotic eye drops to help manage the chronic conjunctivitis.
 - Rest and eye hygiene: The patient should be advised to take regular breaks to rest their eyes and practice good eye hygiene, such as avoiding rubbing their eyes and not sharing makeup or towels.
- Conjunctival concretions in both eyes:
 - Observation: Since conjunctival concretions generally do not require special treatment, the patient can be monitored regularly to see if the concretions change in size or number.
 - Removal of concretions (if necessary): If the patient experiences a foreign body sensation, the conjunctival concretions can be removed.
 - Antibiotic eye drops and ointment: If the concretions are removed, the patient can be prescribed antibiotic eye drops and ointment to help prevent infection.
- Age-related cataracts in both eyes:
 - Regular follow-up: The patient should be advised to regularly follow up with their eye doctor to monitor the progression of the cataracts and determine the best course of treatment.

Type a message... Submit

LLMs represent a breakthrough in AI with large knowledge bases and strong logical reasoning abilities. They have exhibited efficacy across various natural language processing tasks,

including text generation, summarization, translation, and QA. However, in the medical realm, the quality of these answers warrants further scrutiny. Previous research has shown mixed

results for the ability of LLMs to pass ophthalmology examinations. The study of Kung et al [38] indicates that ChatGPT can pass the United States Medical Licensing Examination without any specialized training or reinforcement. However, Thirunavukarasu's [39] attempt to assess ChatGPT's proficiency in the FRCOphth (Fellowship of the Royal College of Ophthalmologists) examination showed subpar performance, thereby underscoring the inability of LLMs to replace physicians in highly specialized fields. Conversely, in advising patients about symptoms or ongoing conditions—tasks less demanding of expertise—ChatGPT seems to demonstrate competence. Many patients turn to the internet for self-diagnosis before consulting a health care professional [40]. The use of LLMs for medical consultations can increase patient independence and potentially aid in accurate diagnosis. The release of GPT4V represents an innovative leap in the realm of LLM integration with computer vision, with promising prospects for extensive application in the medical field. Wu et al [26] assessed images from eight modalities across 17 human body systems and concluded that while GPT4V excels at identifying image modalities and anatomical structures, it encounters significant challenges in disease diagnosis and comprehensive report generation. In another study, we used a similar 1-3 evaluation scale to assess GPT4V's performance on ophthalmology-related tasks, including image interpretation and QA [27]. The model performed best in analyzing slit lamp images; however, it only reached 42% (42/100) in accuracy, 38.5% (34.7/90) in usability, and 68.5% (61.7/90) in safety of the responses. These results are significantly lower than the “entirely good” rates we reported previously—64% (64/100) for report generation and 66.3% (199/300) for QA. This discrepancy underscores the need for models tailored to ophthalmology to ensure high-quality outcomes. To address this gap, we implemented an experimental model, slit lamp-GPT, harnessing the BLIP and Llama2 frameworks. This initiative represents merely the first step in a broader journey toward refining AI applications in ophthalmology.

The model demonstrated proficiency in identifying and reporting common anterior segment eye diseases within our dataset. However, its performance on rare conditions highlighted a critical area for improvement, suggesting that its effectiveness is closely tied to the diversity and representation of conditions in the training data. Per report generation, suboptimal performance was linked to specific diseases such as trichiasis, postglaucoma surgery complications, and corneal pathologies. Given our dataset's origin in routine health examination data, these conditions were underrepresented, likely contributing to the poor performance. Another hypothesis considers the dynamic nature of slit lamp examinations in clinical settings, where ophthalmologists manually focus to obtain the best diagnostic view, a process not fully captured by static images. Instances of misdiagnosed keratitis, where images did not focus precisely on the cornea, support this assumption. Integrating our model with a broader spectrum of ophthalmic imaging techniques—such as indocyanine green angiography, fundus fluorescein angiography, ocular ultrasound, optical coherence tomography, and fundus photography—may enhance diagnostic alignment with actual clinical observations and further improve overall performance.

The current results suggest potential applicability in cataract screening, particularly in regions with a shortage of ophthalmologists. Previous studies have primarily focused on applying deep learning to the diagnosis and grading of cataracts, fundamentally using classification models. In contrast, our model is a natural language processing system capable of generating free-text reports. It not only provides descriptive insights but also achieves cataract classification accuracy similar to existing models [41,42]. Beyond this, our model could function as an educational tool for patients. In bustling eye clinics, patients may lack sufficient time to fully comprehend their examination reports and medical conditions. As demonstrated in this study, the slit lamp-GPT can provide patients with basic clinical explanations and recommendations concerning causes, abnormalities, treatment, and follow-up, indicating its potential to reduce medical consultation expenditure and bolster the use of remote health care services.

The manual evaluation suggests that slit lamp-GPT exhibits a promising capacity to assist participants with minimal risk. During the QA stage, 89.3% (268/300) of the responses were deemed completely harmless, surpassing the performance of GPT4V. However, the potential risks of using LLMs are yet to be thoroughly understood. A common problem with LLMs is that they sometimes generate inaccuracies and false statements, which are often referred to as “hallucinations” in the field [43]. These incorrect assertions can appear to be true, which could harm patients. This was reflected in our study, where the model sometimes created content. For example, the Llama2 model wrongly identified a binocular intraocular lens as a disease instead of a postoperative condition, creating the nonexistent “binocular intraocular lens syndrome.” This led to poor scores on the related 20 questions, highlighting the need for specialized fine-tuned LLM and knowledge-based generation [44]. Nonetheless, it is important to recognize that LLMs should serve as adjuncts or supplements in the clinical diagnosis and treatment process, not as fully trusted entities devoid of physician oversight. As LLM technology evolves, it is incumbent on stakeholders to collaboratively establish best practice standards to ensure patient safety.

Limitations

This study has a few limitations. First, the dataset used is skewed, coming mainly from routine health checks of healthy people. The small sample size for certain diseases might affect the effectiveness of classification. Using datasets from high-quality outpatient clinics could lead to better results. Second, as with other language models, our model sometimes produces repetitive text, and the accuracy of the responses it generates can be inconsistent. At times, the model's answers show logical errors. For instance, it diagnosed both a postintraocular lens implantation status and a senile cataract in the same eye. These issues might be addressed by incorporating expert knowledge and fine-tuning LLMs. There are also notable concerns about bias, as a single mistake in report generation can lead to multiple errors during the question-and-answer process. This highlights the need for further improvements to increase the accuracy and completeness of report generation. Lastly, creating a standardized manual evaluation process for these types of models is challenging [45,46]. This study was

limited to slit lamp anterior segment images, indicating a need for future research to include diverse datasets. This will help evaluate the model's applicability across various types of imaging.

Conclusion

This research underscores the effectiveness and potential of using LLMs for slit lamp image report generation and QA tasks, showcasing their viability in ophthalmic medical image analysis.

Acknowledgments

This study was supported by the Global STEM Professorship Scheme (P0046113) and Henry G. Leong Endowed Professorship in Elderly Vision Health. The sponsor or funding organization had no role in the design or conduct of this research. We thank the InnoHK Hong Kong special administrative region government for providing valuable supports.

Authors' Contributions

NL is the primary corresponding author of this article, while MH and DS are co-corresponding author. DS and MH conceived this study. NL provided data. DS and WZ built the deep learning model. DS and ZZ did the literature search and analyzed the data. DS, ZZ, XC, FS, and MH contributed to key data interpretation. ZZ and FS did the manual evaluation. ZZ wrote this paper. JG edited and reviewed this paper. All authors commented and critically revised this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary tables with data.

[\[DOCX File , 23 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Distribution of scores in the manual assessment of report generation and question-answering. RG: report generation; QA: question answering.

[\[PNG File , 78 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Examples of generated answers with different scores. Red bold text indicates incorrect information.

[\[PNG File , 918 KB-Multimedia Appendix 3\]](#)

References

1. Doggart JH. SLIT-LAMP Microscopy. *Br J Ophthalmol*. 1948;32(4):232-247. [[FREE Full text](#)] [doi: [10.1136/bjo.32.4.232](https://doi.org/10.1136/bjo.32.4.232)] [Medline: [18170442](https://pubmed.ncbi.nlm.nih.gov/18170442/)]
2. Magyezi J, Arunga S. Eye care where there are no ophthalmologists: the Uganda experience. *Community Eye Health*. 2020;33(110):48-50. [[FREE Full text](#)] [Medline: [34007108](https://pubmed.ncbi.nlm.nih.gov/34007108/)]
3. Niehoff JH, Kalaizidis J, Kroeger JR, Schoenbeck D, Borggreffe J, Michael AE. Evaluation of the clinical performance of an AI-based application for the automated analysis of chest X-rays. *Sci Rep*. 2023;13(1):3680. [[FREE Full text](#)] [doi: [10.1038/s41598-023-30521-2](https://doi.org/10.1038/s41598-023-30521-2)] [Medline: [36872333](https://pubmed.ncbi.nlm.nih.gov/36872333/)]
4. Sheth D, Giger ML. Artificial intelligence in the interpretation of breast cancer on MRI. *J Magn Reson Imaging*. 2020;51(5):1310-1324. [doi: [10.1002/jmri.26878](https://doi.org/10.1002/jmri.26878)] [Medline: [31343790](https://pubmed.ncbi.nlm.nih.gov/31343790/)]
5. Shen Y, Chen L, Yue W, Xu H. Artificial intelligence in ultrasound. *Eur J Radiol*. 2021;139:109717. [doi: [10.1016/j.ejrad.2021.109717](https://doi.org/10.1016/j.ejrad.2021.109717)] [Medline: [33962110](https://pubmed.ncbi.nlm.nih.gov/33962110/)]
6. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26(6):900-908. [doi: [10.1038/s41591-020-0842-3](https://doi.org/10.1038/s41591-020-0842-3)] [Medline: [32424212](https://pubmed.ncbi.nlm.nih.gov/32424212/)]
7. Introducing ChatGPT. URL: <https://openai.com/blog/chatgpt> [accessed 2024-10-19]
8. Touvron H, Martin L, Stone K. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. :1-77. Preprint published on July 18, 2023
9. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5(3):e107-e108. [[FREE Full text](#)] [doi: [10.1016/S2589-7500\(23\)00021-3](https://doi.org/10.1016/S2589-7500(23)00021-3)] [Medline: [36754724](https://pubmed.ncbi.nlm.nih.gov/36754724/)]
10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
11. Kahambing J. ChatGPT, public health communication and 'intelligent patient companionship'. *J Public Health (Oxf)*. 2023;45(3):e590. [doi: [10.1093/pubmed/fdad028](https://doi.org/10.1093/pubmed/fdad028)] [Medline: [37036209](https://pubmed.ncbi.nlm.nih.gov/37036209/)]

12. Chen X, Xu P, Li Y, Zhang W, Song F, He M, et al. ChatFFA: an ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography. *iScience*. 2024;27(7):110021. [FREE Full text] [doi: [10.1016/j.isci.2024.110021](https://doi.org/10.1016/j.isci.2024.110021)] [Medline: [39055931](https://pubmed.ncbi.nlm.nih.gov/39055931/)]
13. Dai H, Liu Z, Liao W, Huang X, Cao Y, Wu Z, et al. AugGPT: leveraging ChatGPT for text data augmentation. arXiv:2302.13007. :1-12. Preprint published on February 25, 2023
14. Ebrahimian S, Homayounieh F, Rockenbach MABC, Putha P, Raj T, Dayan I, et al. Artificial intelligence matches subjective severity assessment of pneumonia for prediction of patient outcome and need for mechanical ventilation: a cohort study. *Sci Rep*. 2021;11(1):858. [FREE Full text] [doi: [10.1038/s41598-020-79470-0](https://doi.org/10.1038/s41598-020-79470-0)] [Medline: [33441578](https://pubmed.ncbi.nlm.nih.gov/33441578/)]
15. Liang H, Guo Y, Chen X, Ang K, He Y, Jiang N, et al. Artificial intelligence for stepwise diagnosis and monitoring of COVID-19. *Eur Radiol*. 2022;32(4):2235-2245. [FREE Full text] [doi: [10.1007/s00330-021-08334-6](https://doi.org/10.1007/s00330-021-08334-6)] [Medline: [34988656](https://pubmed.ncbi.nlm.nih.gov/34988656/)]
16. Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Badia SBBI. A human-AI collaborative approach for clinical decision making on rehabilitation assessment. 2021. Presented at: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021 May 07:1-14; Yokohama, Japan.
17. Biswas S. ChatGPT and the future of medical writing. *Radiology*. 2023;307(2):e223312. [doi: [10.1148/radiol.223312](https://doi.org/10.1148/radiol.223312)] [Medline: [36728748](https://pubmed.ncbi.nlm.nih.gov/36728748/)]
18. Masters K. Artificial intelligence in medical education. *Med Teach*. 2019;41(9):976-980. [doi: [10.1080/0142159X.2019.1595557](https://doi.org/10.1080/0142159X.2019.1595557)] [Medline: [31007106](https://pubmed.ncbi.nlm.nih.gov/31007106/)]
19. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv:2305.09617. :1-30. Preprint published on May 16, 2023
20. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *NeurIPS Proceedings*. 2012. Presented at: 26th Annual Conference on Neural Information Processing Systems 2012; 2012-12-3:1-9; Lake Tahoe, Nevada, United States. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
21. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *NeurIPS Proceedings*. 2017. Presented at: the 31st International Conference on Neural Information Processing Systems; 2017 December 04:6000-6010; Red Hook, NY, United States. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
22. Son KY, Ko J, Kim E, Lee SY, Kim M, Han J, et al. Deep learning-based cataract detection and grading from slit-lamp and retro-illumination photographs: model development and validation study. *Ophthalmol Sci*. 2022;2(2):100147. [FREE Full text] [doi: [10.1016/j.xops.2022.100147](https://doi.org/10.1016/j.xops.2022.100147)] [Medline: [36249697](https://pubmed.ncbi.nlm.nih.gov/36249697/)]
23. Lu Q, Wei L, He W, Zhang K, Wang J, Zhang Y, et al. Lens opacities classification system III-based artificial intelligence program for automatic cataract grading. *J Cataract Refract Surg*. 2022;48(5):528-534. [doi: [10.1097/j.jcrs.0000000000000790](https://doi.org/10.1097/j.jcrs.0000000000000790)] [Medline: [34433780](https://pubmed.ncbi.nlm.nih.gov/34433780/)]
24. Fang X, Deshmukh M, Chee ML, Soh Z, Teo ZL, Thakur S, et al. Deep learning algorithms for automatic detection of pterygium using anterior segment photographs from slit-lamp and hand-held cameras. *Br J Ophthalmol*. 2022;106(12):1642-1647. [FREE Full text] [doi: [10.1136/bjophthalmol-2021-318866](https://doi.org/10.1136/bjophthalmol-2021-318866)] [Medline: [34244208](https://pubmed.ncbi.nlm.nih.gov/34244208/)]
25. Soleimani M, Esmaili K, Rahdar A, Aminizadeh M, Cheraqpour K, Tabatabaei SA, et al. From the diagnosis of infectious keratitis to discriminating fungal subtypes; a deep learning-based study. *Sci Rep*. 2023;13(1):22200. [FREE Full text] [doi: [10.1038/s41598-023-49635-8](https://doi.org/10.1038/s41598-023-49635-8)] [Medline: [38097753](https://pubmed.ncbi.nlm.nih.gov/38097753/)]
26. Wu C, Lei J, Zheng Q, Zhao W, Lin W, Zhang X, et al. Can GPT-4V (ision) serve medical applications? case studies on GPT-4V for multimodal medical diagnosis. arXiv:2310.09909. :1-178. Preprint published October 15 2023
27. Xu P, Chen X, Zhao Z, Shi D. Unveiling the clinical incapacities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. *Br J Ophthalmol*. 2024;108(10):1384-1389. [doi: [10.1136/bjo-2023-325054](https://doi.org/10.1136/bjo-2023-325054)] [Medline: [38789133](https://pubmed.ncbi.nlm.nih.gov/38789133/)]
28. Chen X, Zhang W, Xu P, Zhao Z, Zheng Y, Shi D, et al. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. *NPJ Digit Med*. 2024;7(1):111. [FREE Full text] [doi: [10.1038/s41746-024-01101-z](https://doi.org/10.1038/s41746-024-01101-z)] [Medline: [38702471](https://pubmed.ncbi.nlm.nih.gov/38702471/)]
29. Chen X, Zhang W, Zhao Z, Xu P, Zheng Y, Shi D, et al. ICGA-GPT: report generation and question answering for indocyanine green angiography images. *Br J Ophthalmol*. 2024;108(10):1450-1456. [doi: [10.1136/bjo-2023-324446](https://doi.org/10.1136/bjo-2023-324446)] [Medline: [38508675](https://pubmed.ncbi.nlm.nih.gov/38508675/)]
30. Hu Y, Niu Y, Wang D, Wang Y, Holden BA, He M. The association of longitudinal trend of fasting plasma glucose with retinal microvasculature in people without established diabetes. *Invest Ophthalmol Vis Sci*. 2015;56(2):842-848. [doi: [10.1167/iovs.14-15943](https://doi.org/10.1167/iovs.14-15943)] [Medline: [25613941](https://pubmed.ncbi.nlm.nih.gov/25613941/)]
31. Li J, Li D, Xiong C, Hoi S. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. 2022. Presented at: Proceedings of the 39th International Conference on Machine Learning; 2022 January 28; Baltimore, Maryland, USA.
32. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Presented October 11, 2018. 2018:1-16.

33. Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina*. 2023;7(10):862-868. [doi: [10.1016/j.oret.2023.05.022](https://doi.org/10.1016/j.oret.2023.05.022)] [Medline: [37277096](https://pubmed.ncbi.nlm.nih.gov/37277096/)]
34. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. 2002. Presented at: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; 2002 July 06; United States.
35. Vedantam R, Lawrence ZC, Parikh D. CIDEr: Consensus-based image description evaluation. 2015. Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 June 07-12; Boston, MA, USA.
36. Lin CY. Rouge: A package for automatic evaluation summaries. 2004. Presented at: Proceedings of the Workshop on Text Summarization Branches Out; 2004 July 25-26:74-81; Barcelona, Spain.
37. Anderson P, Fernando B, Johnson M, Gould S. SPICE: semantic propositional image caption evaluation. 2016. Presented at: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V; 2016 October 11-14; Amsterdam, The Netherlands.
38. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. [FREE Full text] [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
39. Thirunavukarasu AJ. ChatGPT cannot pass FRCOphth examinations: implications for ophthalmology and large language model artificial intelligence. *Eye News*. 2023. URL: <https://www.eyenews.uk.com/media/31505/eye-am23-onex-arun-proof-2.pdf> [accessed 2024-10-24]
40. Kuehn BM. More than one-third of US individuals use the internet to self-diagnose. *JAMA*. 2013;309(8):756-757. [doi: [10.1001/jama.2013.629](https://doi.org/10.1001/jama.2013.629)] [Medline: [23443421](https://pubmed.ncbi.nlm.nih.gov/23443421/)]
41. Shimizu E, Yazu H, Aketa N, Tanji M, Sakasegawa A, Nakayama S, et al. Innovative artificial intelligence-based cataract diagnostic method uses a slit-lamp video recording device and multiple machine-learning. *Invest Ophthalmol Vis Sci*. 2021;62(8):1031.
42. Goh JHL, Lim ZW, Fang X, Anees A, Nusinovi S, Rim TH, et al. Artificial intelligence for cataract detection and management. *Asia Pac J Ophthalmol (Phila)*. 2020;9(2):88-95. [FREE Full text] [doi: [10.1097/01.APO.0000656988.16221.04](https://doi.org/10.1097/01.APO.0000656988.16221.04)] [Medline: [32349116](https://pubmed.ncbi.nlm.nih.gov/32349116/)]
43. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. *J R Soc Med*. 2023;116(5):181-182. [FREE Full text] [doi: [10.1177/01410768231173123](https://doi.org/10.1177/01410768231173123)] [Medline: [37199678](https://pubmed.ncbi.nlm.nih.gov/37199678/)]
44. Chen X, Zhao Z, Zhang W, Xu P, Wu Y, Xu M, et al. EyeGPT: ophthalmic assistant with large language models for patient inquiries and medical education. *J Med Internet Res*. Nov 1, 2024:-6063. (forthcoming)(forthcoming). [doi: [10.2196/60063](https://doi.org/10.2196/60063)]
45. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. *Assoc Comput Mach*. 2024;15(3):39. [doi: [10.1145/364128](https://doi.org/10.1145/364128)]
46. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models in medical applications: a survey. arXiv:2405.07468. :1-42. Preprint published on May 13, 2024

Abbreviations

AI: artificial intelligence

BLEU: bilingual evaluation understudy

BLIP: Bootstrapping Language-Image Pre-training

CIDEr: consensus-based image description evaluation

CNNs: convolutional neural networks

GRUs: gated recurrent units

LLM: large language model

QA: question answering

ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation—Longest Common Subsequence

RNN: recurrent neural network

SPICE: Semantic Propositional Image Caption Evaluation

Edited by G Eysenbach; submitted 27.10.23; peer-reviewed by TK Yoo, S McRoy; comments to author 03.01.24; revised version received 24.02.24; accepted 05.09.24; published 30.12.24

Please cite as:

Zhao Z, Zhang W, Chen X, Song F, Gunasegaram J, Huang W, Shi D, He M, Liu N

Slit Lamp Report Generation and Question Answering: Development and Validation of a Multimodal Transformer Model with Large Language Model Integration

J Med Internet Res 2024;26:e54047

URL: <https://www.jmir.org/2024/1/e54047>

doi: [10.2196/54047](https://doi.org/10.2196/54047)

PMID:

©Ziwei Zhao, Weiyi Zhang, Xiaolan Chen, Fan Song, James Gunasegaram, Wenyong Huang, Danli Shi, Mingguang He, Na Liu. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 30.12.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.