OPEN ACCESS

Clinical science

Development and evaluation of a large language model of ophthalmology in Chinese

Ce Zheng,[1,2] Hongfei Ye ,[1,2] Jinming Guo,[3] Junrui Yang,[4] Ping Fei ,[1] Yuanzhi Yuan,[5] Danqing Huang,[6] Yuqiang Huang,[3] Jie Peng,[7] Xiaoling Xie ,[3] Meng Xie ,[1] Peiquan Zhao ,[1] Li Chen,[1] Mingzhi Zhang[3]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bjo-2023-324526).

For numbered affiliations see end of article.

Correspondence to
Professor Mingzhi Zhang; zmz0754@126.com and Dr Li Chen; chenree@163.com

CZ and HY contributed equally. LC and MZ contributed equally.

Received 5 September 2023
Accepted 3 June 2024
Published Online First 17 July 2024

ABSTRACT

Background Large language models (LLMs), such as ChatGPT, have considerable implications for various medical applications. However, ChatGPT's training primarily draws from English-centric internet data and is not tailored explicitly to the medical domain. Thus, an ophthalmic LLM in Chinese is clinically essential for both healthcare providers and patients in mainland China.

Methods We developed an LLM of ophthalmology (MOPH) using Chinese corpora and evaluated its performance in three clinical scenarios: ophthalmic board exams in Chinese, answering evidence-based medicine-oriented ophthalmic questions and diagnostic accuracy for clinical vignettes. Additionally, we compared MOPH's performance to that of human doctors.

Results In the ophthalmic exam, MOPH's average score closely aligned with the mean score of trainees (64.7 (range 62–68) vs 66.2 (range 50–92), p=0.817), but achieving a score above 60 in all seven mock exams. In answering ophthalmic questions, MOPH demonstrated an adherence of 83.3% (25/30) of responses following Chinese guidelines (Likert scale 4–5). Only 6.7% (2/30, Likert scale 1–2) and 10% (3/30, Likert scale 3) of responses were rated as 'poor or very poor' or 'potentially misinterpretable inaccuracies' by reviewers. In diagnostic accuracy, although the rate of correct diagnosis by ophthalmologists was superior to that by MOPH (96.1% vs 81.1%, p>0.05), the difference was not statistically significant.

Conclusion This study demonstrated the promising performance of MOPH, a Chinese-specific ophthalmic LLM, in diverse clinical scenarios. MOPH has potential real-world applications in Chinese-language ophthalmology settings.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence-based large language model (LLM) has significant implications in medical applications, and it has shown great potential in the diagnosis of ophthalmic conditions and preparation of the board certification in the field of ophthalmology. While there is a lack of LLM training with non-English languages and explicit up-to-date medical domains.

WHAT THIS STUDY ADDS

⇒ This study developed a Chinese-specific LLM of ophthalmology (MOPH), and further demonstrated its accuracy and reliability in three different clinical scenarios.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Our exploration revolves around safeguarding user privacy and security while leveraging LLMs in the healthcare domain, and based on this, further researches can be continued to evaluate MOPH's real-world clinical performance.

INTRODUCTION

Artificial intelligence (AI) has been expanding its applications in various medical domains, such as image analysis, patients' risk stratification and clinical note processing.[1 2] However, recent AI advancements mostly focus on narrow and well-defined tasks and challenges, such as detecting diabetic retinopathy from fundus images.[3] With the rapid development of the latest generation of AI models, which are trained on massive and diverse datasets, the aspiration of moving from the 'narrow AI' to 'artificial general intelligence' (AGI) demonstrated broad capabilities of intelligence. A noteworthy recent development is ChatGPT (Open AI, San Francisco, California), an AI-based large language model (LLM) that has significant implications in diverse scientific and medical applications.[4 5] These neural network models are based on the Transformer architecture and trained on massive corpora of web-text data and can be applied to numerous downstream tasks. In the field of ophthalmology, researches have been conducted to evaluate the performance and potential of LLMs.[6 7] It was showed that ChatGPT answered approximately half of the questions correctly in the OphthoQuestions free trial for ophthalmic board certification preparation.[8] Another study has even found that ChatGPT has the potential in the diagnosis of ophthalmic conditions, particularly for primary care providers.[9]

However, LLMs have crucial limitations. For instance, ChatGPT's training predominantly relies on English-centric internet data, which may impact the quality and diversity of their outputs, especially for non-English languages and domains.[10] Additionally, LLMs, usually deployed remotely, may have access to a wide range of patient characteristics, posing serious privacy

To cite: Zheng C, Ye H, Guo J, *et al*. *Br J Ophthalmol* 2024;**108**:1390–1397.

risks.[11] Furthermore, ChatGPT's knowledge cut-off is 2021, making its medical field output potentially outdated.

In this study, we aimed to develop an LLM of ophthalmology (MOPH) using Chinese corpora. We further assessed MOPH's performance in three different clinical scenarios: ophthalmic board exams in Chinese, answering ophthalmic questions following evidence-based medicine (EBM) and diagnostic accuracy for clinical vignettes. Our exploration revolves around safeguarding user privacy and security while leveraging LLMs in the healthcare domain.

MATERIALS AND METHODS

Overview

This study aimed to develop a Chinese LLM that can be deployed locally and dedicated to ophthalmology. We also tested its early AGI ability in various clinical scenarios. This observational study was approved by the Xinhua Hospital Ethics Review Committee (Approval No. XHEC-D-2023–131), and the study protocol followed the tenets of the Declaration of Helsinki. The review committee indicated that patient consent was not required in this research as we only used publicly accessible or deidentified data.

Development of MOPH in Chinese

We developed MOPH by adopting the open-source LLM (ChatGLM2-6B). ChatGLM2-6B is an open bilingual language model based on General Language Model (GLM) framework.[12 13] In brief, ChatGLM2-6B was trained on about one trillion tokens—equally of Chinese and English corpora, enabling the model to perform well in both languages (see Section A in the online supplemental material 1).

To customise the ChatGLM2-6B for our application scenarios, we first adopted prompt engineering to preprocess the users' input. Prompt engineering involves creating prompts based on specific questions or statements within a specific domain. This approach allowed us to leverage MOPH's semantic understanding while also providing the model with the most relevant information. We used publicly available and self-built Chinese ophthalmic knowledge databases, mainly referring to ophthalmic textbooks, guidelines and selected review papers in Chinese and AAO Eyewiki (translated by Internet Explorer's built-in function (online supplemental material 2).[14 15] To further address the unreliable and deceptive output of LLM, we performed prompt tuning (p-tuning) on our fine-tuning dataset in Chinese to refine MOPH (see Sections B and C in online supplemental material 1). We only chose Chinese ophthalmic contents for p-tuning purpose. P-tuning is an efficient fine-tuning technique that optimises continuous prompts, significantly reducing storage and memory usage per task. It has been shown to performs comparably to full parameter fine tuning with only 0.1%–3% of the fine-tuning parameters.[16] Figure 1 illustrates the implementation of MOPH's framework.

We conducted the p-tuning using the source codes from ChatGLM2-6B's GitHub.[12] The hyperparameters employed in the training process were as follows: the batch size of 1, a learning rate of 2e-4 with gradient accumulation steps of 16, a maximum source length of 128 tokens, and a maximum target length of 512 tokens. For prompt engineering, we used GanymedeNil/text2vec-large-Chinese for embedding and Facebook AI Similarity Search (Faiss) for efficient similarity search and clustering of dense vectors.[17 18] Figure 2 illustrates the details the prompt engineering and prompting generation process in our study. The hardware for MOPH training included an Intel 8th generation central processing unit (i5–8400, 2.81 GHz, 32 GB main memory) and two NVIDIA A4000 GPUs for 35 hours.

Evaluation of MOPH in ophthalmology

Some researchers believe that LLMs, such as ChatGPT, could be viewed as an early (yet still incomplete) version of the AGI system. Inspired by that, we propose here three clinical scenarios to investigate the capabilities of our MOPH model.

We first test the performance of MOPH in the Board of Ophthalmic Exams in Shanghai, China. We used a dataset of single-choice questions (SCQ) and the Written Qualifying Exam (WQE) from OphthoQuestions of the National Medical E-Book Packages, a common resource for board certification examination preparation.[19] We only included text-based questions and excluded questions requiring the input of images. We also asked the trainees in the same department to take the mock exam and compare their scores with MOPH's scores. Three senior ophthalmologists (all with over 10 years of clinical experience) independently reviewed each answer of WQEs. The overall mean score was determined by averaging the scores given by each grader. To avoid confirmation bias, we did not tell the graders in advance that the language model and humans were taking the exam together.

In the second clinical scenario, we investigated whether MOPH can respond following EBM. Based on clinical guidelines of the Chinese Medical Association, we generated 30 questions on the following six subspecialties of ophthalmology: glaucoma, lens and cataract, paediatric ophthalmology and strabismus, retina and vitreous, external disease and cornea, and uveitis and ocular inflammation (Section A in the online supplemental material 3). Three graders (more than 10 years of clinical experience) assessed the MOPH's responses using a Likert scale from 1 to 5 (1: very poor/unacceptable inaccuracies, 2: poor/minor potentially harmful inaccuracies, 3: moderate/potentially misinterpretable inaccuracies, 4: good/only minor non-harmful inaccuracies, 5: very good).

We evaluate the diagnostic accuracy of outpatient clinic notes from the clinical setting. We deidentified the clinical vignettes that only included: the patient's chief complaints, present illness, past ocular history, ocular medications, general medical and surgical history and physical examination with vital signs following the electronic medical record of the Hospital Information System, Xinhua Hospital. The following subspecialties of ophthalmology were included (30 clinical vignettes for each sub-specialty): glaucoma, lens and cataract, paediatric ophthalmology and strabismus, retina and vitreous, external disease and cornea and uveitis and ocular inflammation (Section B in the online supplemental material 3). We also measured the accuracy rate of diagnoses made by the above three graders using a majority consensus-based approach.

Finally, we compare MOPH's performance to commercial LLMs (ChatGPT). Assessing an LLM's performance has always been challenging. To this end, we selected MedQA as a medical benchmark alongside SCQ in ophthalmology.[20] The MedQA dataset comprises questions (compiled as SCQs) in the style of the US Medical License Exam (USMLE). We used an online translation tool (https://www.deepl.com/en/translator) to translate the MedQA questions into Chinese. For inference, we employed the default settings from ChatGLM2-6B's GitHub (top_p=0.7, temperature=0.95).[21] By default, MOPH's model parameters are loaded with F16 precision, requiring approximately 13 GB of GPU memory. After quantisation, MOPH can be deployed locally on consumer-grade

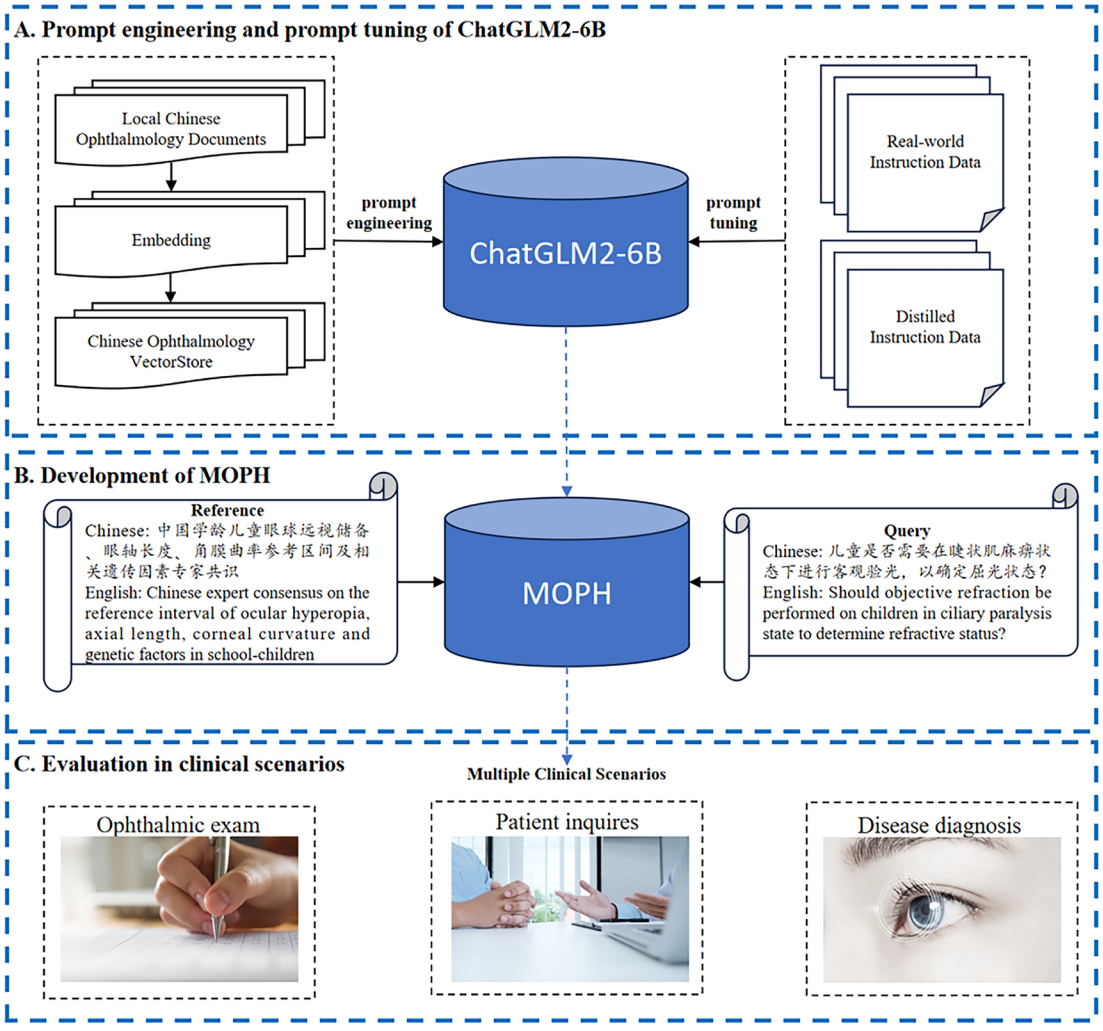

Figure 1 The Implementation of MOPH flowchart (drawn by CZ). A. Prompt engineering and prompt tuning of GhatGLM2-6B; B. Development of MOPH; C. Evaluation in clinical scenarios. MOPH, large language model of ophthalmology.

graphics cards (eg, only 6 GB of GPU memory is required at the INT8 quantisation level). We then evaluated the three LLMs (MOPH with Q8 and F16 precision, and ChatGPT) on questions from MedQA's general medical domain and SCQ's ophthalmology specialty.

Statistical analysis

We used the intraclass correlation coefficient (ICC) to measure inter-rater reliability. A t-test was used to compute the difference between observed means in two independent samples. Diagnosis accuracy was presented as numbers (percentages) and were compared using the χ^2 test. P values were two tailed, and a p value <0.05 was considered statistically significant.

RESULTS

We assessed the capability of MOPH in ophthalmic knowledge by preparing seven sets of mock exams, each consisting of 25 SCQs (50 scores in total) and 5 WQEs (50 scores in total). MOPH was required to complete all seven exams, while seven trainees were randomly assigned a set of exams each. On average, MOPH correctly answered 56% (range 52% (13/25) to 60% (15/25)) of SCQs, which was lower than the averages of trainees ($p<0.05$) (table 1 and figure 3).

However, even though we did not inform the scoring graders in advance that the language model and humans were taking the exam together in WQEs, no statistically significant difference was found between MOPH and trainees (73.4% (range 70%–82%) vs 59.5% (range 40%–88%), $p=0.07$). The final results showed that MOPH's average score was close to that of trainees (64.7 (range 62–68) vs 66.2 (range 50–92), $p=0.817$). Notably, MOPH achieved a score of over 60 in all seven mock exams, while three out of seven trainees failed to reach the passing requirement of 60 points.

Table 2 demonstrate some examples of responses from MOPH. We found that MOPH generated high-quality general information and provided good responses following EBM. MOPH had 83.3% (25/30) of responses following Chinese guidelines (Likert scale 4–5) (table 2a). Table 2b demonstrated the 'moderate/potentially misinterpretable inaccuracies' responses mainly due to the hallucinations produced by MOPH (10% (3/30), Likert scale 3). For instance, when asked 'For patients diagnosed with type 1 diabetes before puberty, should they start screening for diabetic retinopathy after puberty?', MOPH gave the correct response such as 'these guidelines recommend screening for diabetic retinopathy after puberty', however, MOPH returned the potentially misinterpretable reasons: 'This is because before

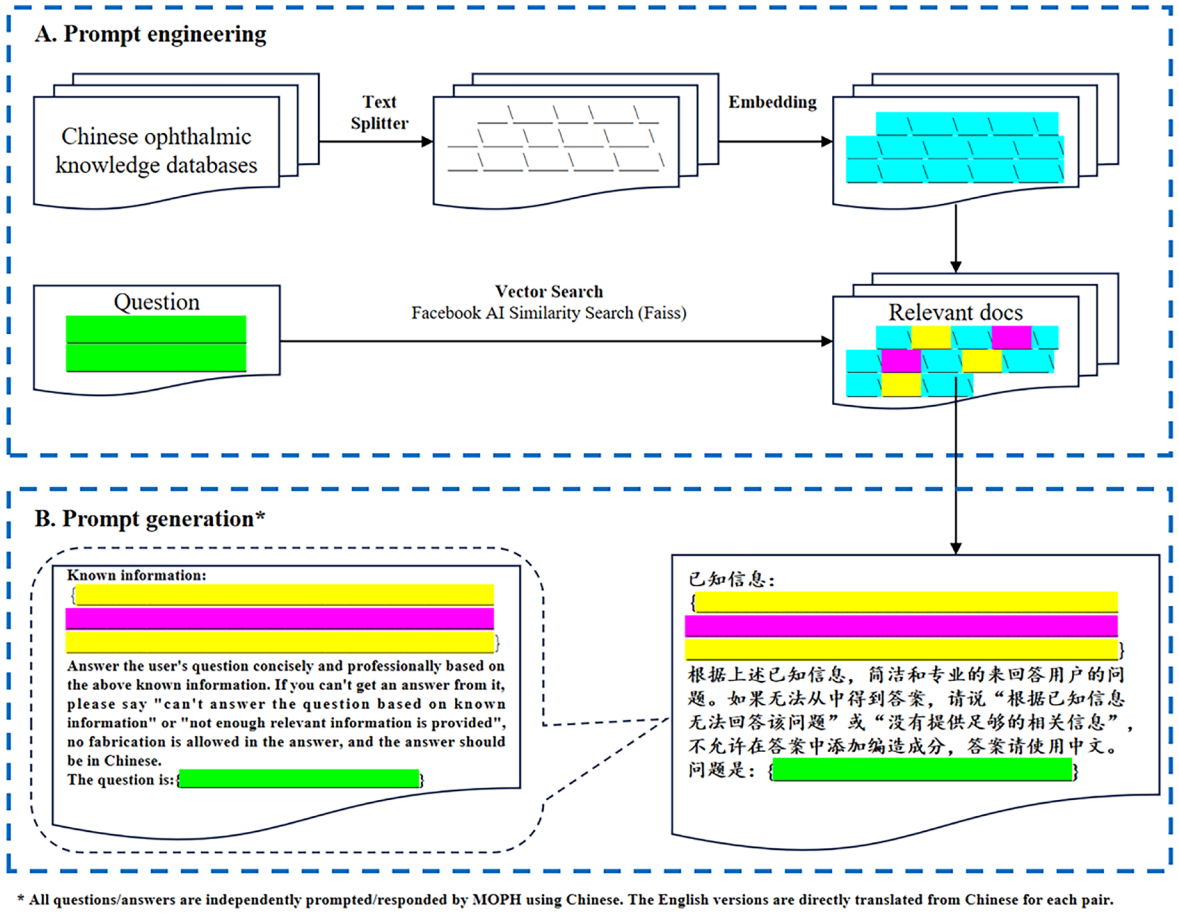

Figure 2 The illustration of prompting engineering and prompt generation (drawn by CZ). A. Prompt engineering; B. Prompt generation. MOPH, large language model of ophthalmology.

puberty, the patient's physical development is not fully mature, and the examination of fundus lesions may cause discomfort to the patient, with a high risk'. Only 6.7% (2/30, Likert scale 1–2) responses were graded as 'poor or very poor' by reviewers (table 2c). As examples in table 2c, MOPH inaccurately defined that 'The lower limit of normal vision reference value for children aged 3 to 5 years is 0.7, not 0.5'.

Table 3 illustrates the diagnostic accuracy of MOPH in the deidentified clinical vignettes. Overall, human doctors exhibited better than MOPH (accuracy 96.1% vs 81.1%, p>0.05). The performance of MOPH was still considered to be good without statistical difference. Especially in certain ophthalmic subspecialties, such as lens and cataract, MOPH attained near-human diagnostic accuracy (96.7% vs 100%, p=1.00). Conversely, in other subspecialties like retinal diseases, the performance of MOPH lagged considerably behind human doctors (63.3% vs 90%, p<0.03). The inter-rater reliability among senior graders was excellent, with ICC values of 0.95 and 0.91 for the first and third clinical scenarios, respectively.

Table 1 Comparisons of results between MOPH and trainees

	MOPH	Trainees	
	Mean±SD	Mean±SD	P value
SCQ	56.0±3.1	72.9±14.9	0.013
WQE	73.4±5.4	59.5±17.7	0.070
Total Scores	64.7±2.2	66.2±16.2	0.817

MOPH, large language model of ophthalmology; SCQ, single-choice questions; WQE, Written Qualifying Exam.

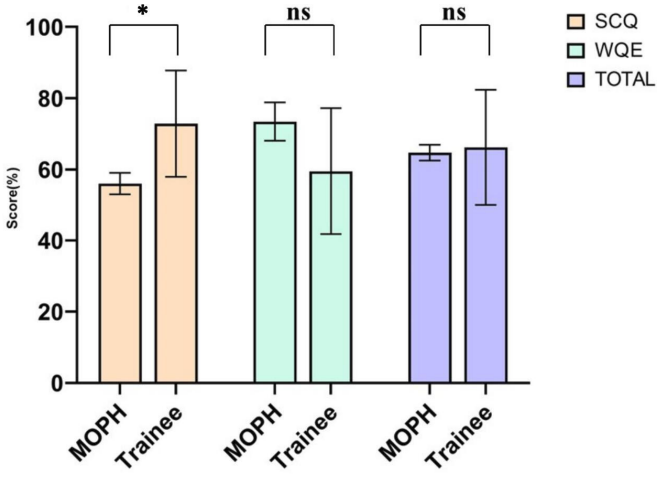

Figure 3 Comparisons of main results between MOPH and ophthalmology trainees. *Denotes p<0.05, ns denotes no statistical significance. MOPH, large language model of ophthalmology; SCQ, single-choice questions; WQE, Written Qualifying Exam.

Table 2 The performance of MOPH in answering ophthalmic questions following evidence-based medicine (EBM)*. (a) Examples of the good responses from MOPH; (b) Examples of 'moderate/potentially misinterpretable inaccuracies' responses from MOPH; (c) Examples of 'poor or very poor' responses from MOPH

Table 2a. Examples of the 'good' responses from MOPH

	Questions	Answers	Scores
1	Should objective refraction be performed on children in ciliary paralysis state to determine refractive status?	It is necessary for children to undergo objective refraction under cycloplegia to determine their refractive status. This can ensure that the objective refraction results accurately reflect the true refractive status of the child, thereby guiding clinical treatment.	5
2	Can emotions such as anxiety and depression also cause dry eye?	Emotions such as anxiety and depression can also cause dry eye. In addition to systemic and topical medication, emotional factors are also considered to be a factor that causes dry eye. Therefore, when treating dry eye, it is necessary to fully inquire about the patient's medical history and pay attention to their emotional state, to identify potential emotional factors such as anxiety or depression, and accordingly provide psychological guidance and treatment.	5

Table 2b. Examples of 'moderate/potentially misinterpretable inaccuracies' responses from MOPH

	Questions	Answers	Scores
1	For patients diagnosed with type 1 diabetes before puberty, should they start screening for diabetic retinopathy after puberty?	According to the relevant guidelines from countries such as Australia, Canada, UK and USA, patients diagnosed with type 1 diabetes before puberty should start screening for diabetic retinopathy by examining the fundus after puberty. This is because before puberty, the patient's body development is not fully mature, and the examination of fundus lesions may cause discomfort to the patient, with a higher risk. After puberty, the patient's body development is mature, and they can cooperate better with the examination, and also better detect diabetic retinopathy. Therefore, these guidelines suggest screening for diabetic retinopathy by examining the fundus after puberty.	3
2	For patients with uveal melanoma, who have orbital invasion, is it recommended to perform enucleation?	According to the known information, patients with uveal melanoma who have invaded the orbit are advised to undergo orbital content removal surgery. This is because in uveal melanoma, there are often symptoms of enophthalmos, and orbital content removal surgery can better protect the eyeball, which is beneficial for the patient's rapid healing. It should be noted that during the surgery, the eyelids should be preserved as much as possible to facilitate rapid healing.	3

Table 2c. Examples of 'poor or very poor' responses from MOPH

	Questions	Answers	Scores
1	Is the lower limit of normal visual acuity for children aged 3–5 years 0.5?	The lower limit of normal vision reference value for children aged 3–5 years is 0.7, not 0.5.	1
2	Can patients with keratoconus or other types of corneal ectasia undergo transepithelial photorefractive keratectomy (TransPRK)?	Patients with keratoconus or other types of corneal ectasia can undergo transepithelial photorefractive keratectomy (TransPRK). TransPRK is a type of surface ablation laser surgery that aims to correct refractive errors. This surgery is suitable for general patients who have a desire to remove their glasses, have reasonable expectations of the surgical outcome, and are over 18 years old. However, patients need to understand the risks and benefits of the surgery before the surgery, and fully understand the surgical process. Absolute contraindications include severe ocular surface diseases that may affect corneal epithelial healing, ocular active inflammatory reactions, keratoconus or other types of corneal ectasia, corneal thinning, severe dry eye, severe ocular adnexal lesions, glaucoma, cataracts that affect vision, and uncontrolled systemic connective tissue diseases and autoimmune diseases.	1

*All questions/answers are independently prompted/responded by MOPH using Chinese. The English versions are directly translated from Chinese for each pair.
MOPH, large language model of ophthalmology.

Finally, we compare the performance of MOPH with commercial LLMs. The datasets involved seven distinct sets of 100 questions randomly selected from MedQA's testing dataset (1273 questions in total) and above-mentioned SCQs. Figure 4 illustrates that MOPH outperforms ChatGPT on SCQ in the domain of ophthalmology, with accuracies of 57.4%, 56.3% and 49.1% for MOPH(F16), MOPH(Q8) and ChatGPT, respectively. However, on MedQA in the general medical domain, ChatGPT achieved a higher score than MOPH (44.6%, 43.4% and 53.7% for MOPH(F16), MOPH(Q8) and ChatGPT, respectively). After quantisation, the performance of MOPH slightly decreased but this difference can be considered negligible (all with $p>0.05$). During the above study, we observed that MOPH did not produce outputs in English when prompted in Chinese, and vice versa.

Table 3 Comparisons of the diagnostic accuracy of deidentified clinical vignettes between MOPH and ophthalmologist

Subspeicalities	MOPH	Ophthalmologist	P value
External disease and cornea	86.7% (26/30)	100% (30/30)	0.112
Lens and cataract	96.7% (29/30)	100% (30/30)	1.000
Glaucoma	80% (24/30)	93.3% (28/30)	0.254
Paediatric ophthalmology and strabismus	83.3% (25/30)	100% (30/30)	0.052
Uveitis and ocular inflammation	76.6% (23/30)	93.3% (28/30)	0.145
Retina and vitreous	63.3% (19/30)	90% (27/30)	0.030

MOPH, large language model of ophthalmology.

DISCUSSION

AGI refers to systems that demonstrate broad capabilities of intelligence, including reasoning, planning and the ability to learn from experience, and with these capabilities at or above human level.[22] In this paper, we developed an offline and local placement Chinese MOPH, suggesting early AGI characteristics. In the ophthalmic knowledge assessment, MOPH achieved a 65% mark, which is comparable to that of the ophthalmology trainees in a University Teaching Hospital. In answering medical questions, MOPH had 83.3% of responses following the guidelines

in Chinese. In diagnostic accuracy, although the rate of correct diagnosis by ophthalmologists was superior to that by MOPH, no statistical difference was found.

LLMs have demonstrated their effectiveness in various general domain tasks. Nevertheless, LLMs have not yet performed optimally in biomedical domain tasks due to the need for medical expertise in the responses. Additionally, since LLMs are mainly trained in English, their ability to understand and respond in languages quite distinct from English, like Chinese, hinders their effective use in Chinese contexts. Consequently, ChatGPT might face challenges in grammar, accuracy and fluency when dealing with Chinese queries, particularly in specialised fields like ophthalmology.[23] China faces a wide spectrum of eye diseases that impact a considerable number of patients.[24] The prevalence of eye diseases continues to rise, presenting a significant challenge to global eye health.[25 26] Therefore, the demand for an ophthalmic LLM in Chinese cannot be ignored. Several studies demonstrate LLMs' impressive multilingual capability, but their performance varies substantially across different languages.[27] According to Zeng's report, ChatGLM is a bilingual LLM that has been pretrained on over 1 trillion English and Chinese tokens.[13] While it is not explicitly mentioned in the study whether there is internal knowledge transfer/translation between the two languages, it is safe to assume that ChatGLM has been trained on a diverse range of data from both languages. This means that the model has learnt to recognise and understand the nuances of both languages and can generate outputs in either language based on the input prompt.

LLMs may generate text that is semantically or syntactically plausible but is incorrect or non-sensical (known as hallucination).[28] In Potapenko's study, ChatGPT provided inadequate responses when addressing questions related to the management of retinal diseases.[6] A study on ChatGPT's response to vernal keratoconjunctivitis queries found that it provided inaccurate and potentially harmful information, especially regarding treatment and medication side effects.[7] To overcome model hallucinations in medical data screening during reference data retrieval, we adopted a method combining vector database retrieval with keyword retrieval, which allows MOPH to integrate external knowledge bases and effectively reduce the LLMs' hallucinations.[29] Rather than training on public Chinese medical databases, such as Chinese Medical Knowledge Graph,[14] as many previous language models did, we carefully designed ophthalmic domain fine-tuning datasets. Our preliminary results demonstrated that MOPH not only offers highly accurate general information about ophthalmology but also provides evidence-based responses regarding the treatment of diseases.

Reasoning is the ability to draw logical conclusions from given information. Some studies have shown that LLMs, such as GPT-3.5, can achieve impressive performance on some reasoning tasks, such as mathematical reasoning or logical reasoning.[30] Lin *et al* compared GPT-3.5 with human performance in American board certification in ophthalmology. The results showed that GPT-3.5 scored 63.1%, while humans scored higher at 72.6%.[31] In the study led by Mihalache *et al*, ChatGPT achieved an accuracy of 46%, but its performance did not meet the threshold for providing substantial assistance in board certification preparation.[8] Interestingly, in both our study and previous studies, the average scores of humans are higher than those of LLMs. There are several possible explanations. First, clinical reasoning (CR) is essential for clinicians, as it is the process they use to reach a diagnosis, treatment and/or management plan. However, in the clinical domain, most LLMs mainly focus on clinical classification or reading comprehension and underexplore CR for disease diagnosis due to the expensive rationale annotation with clinicians.[32] Our results indicate that MOPH may have preliminary CR abilities, although they still lag behind human doctors. Technique limitation, such as lacking multimodal capability, may be another possible reason.[33] For example, MOPH has poor diagnostic abilities for retinal diseases. This may result from MOPH's inability to process ophthalmic images in its current version.

The use of digital health data raises concerns regarding security and privacy.[34] As an LLM that can operate offline and deploy locally, MOPH ensures that patients' privacy and information are not stored or disclosed on the network, thus helping healthcare institutions strengthen the defenses for protecting patient privacy and information. Responsible usage of these AI systems in clinical practice is of utmost importance.[35] Currently, MOPH serves as an assistive tool, highlighting the necessity of human supervision. MOPH delivers health information conversationally, making it easier to comprehend compared with professional guidelines.

This study has several limitations. First, one way to enhance the in-context learning ability of a model is to use few-shot

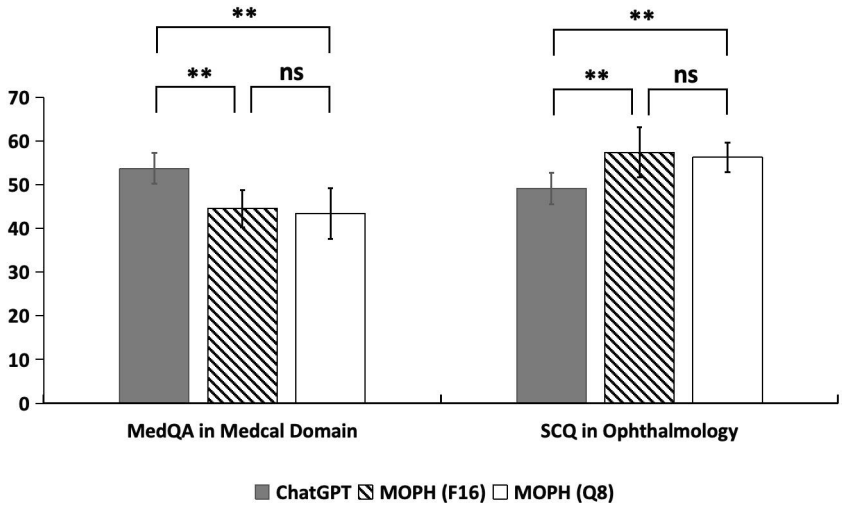

Figure 4 Comparing the performance of MOPH (F16 and Q8) with commercial LLMs in medical benchmarks. **Denotes p<0.001, ns denotes no statistical significance. LLM, large language model; MOPH, LLM of ophthalmology; SCQ, single-choice questions.

prompting, which involves providing several examples in the prompt to guide the model towards better performance. We did not assess few-shot prompting's effect on MOPH in this study. Previous research has highlighted its instability due to training example variations, order and prompt formats inconsistencies. Second, although we demonstrated the preliminary AGI capabilities of MOPH in ophthalmology through three different tasks, these cannot cover the entire clinical diagnosis and treatment spectrum, such as bedside consultation training for resident physicians, actual clinical visits and giving relevant medication and even surgical suggestions, etc. Future work would explore MOPH's performance in various other clinical scenarios. Third, MOPH currently only focuses on textual information and cannot analyse images or videos, while ophthalmic examination results primarily rely on images. Finally, given the rapid development of new LLM models and versions, the reported results should be interpreted cautiously. For instance, recent studies have shown GPT-4's improved performance over GPT-3.5 on medical assessments.[31] Further comparisons that encompass more advanced LLMs, such as GPT-4, Gemini (Google) or Ernie-4 (Baidu), would provide stronger insights into how LLMs might facilitate clinical workflows across different countries and languages.

We have developed a Chinese-specific ophthalmic LLM, MOPH, and demonstrated its accuracy and reliability in different clinical scenarios. As an AGI committed to patient privacy and data security, we will continually evaluate MOPH's real-world clinical performance.

Author affiliations
[1]Ophthalmology, Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China
[2]Institute of Hospital Development Strategy, China Hospital Development Institute, Shanghai Jiao Tong University, Shanghai, China
[3]Joint Shantou International Eye Center of Shantou University and The Chinese University of Hong Kong, Shantou, Guangdong, China
[4]Ophthalmology, The 74th Army Group Hospital, Guangzhou, Guangdong, China
[5]Ophthalmology, Zhongshan Hospital Fudan University, Shanghai, China
[6]Discipline Inspection & Supervision Office, Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China
[7]Opthalmology, Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China

Contributors Planning and design: CZ, HY, JG, JY, LC. Conduct and reporting: CZ, HY, JG, JY, JP, XX, MX and LC. Analysis and interpretation of data: CZ, HY, JG, JY, PF, YY, DH, YH, JP, XX, MX and LC. Critical revision: CZ, HY, MX and LC. Supervision and guarantor: CZ, LC, PZ and MZ.

Funding This study was supported,the National Natural Science Foundation of China (82171044), the Special Strategic Project on Innovative Science and Technology of Guangdong Province(STKJ202209073), the Hospital Funded Clinical Research, Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (21XJMR02) and Hospital Management Research Program of Institute of Hospital Development Strategy, China Hospital Development Institute, Shanghai Jiao Tong University (HDSI-2022-A-001). Interdisciplinary Program of Shanghai Jiao Tong University (Number YG2021QN52).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This observational study was approved by the Xinhua Hospital Ethics Review Committee (Approval No. XHEC-D-2023-131), and the study protocol followed the tenets of the Declaration of Helsinki. The review committee indicated that patient consent was not required in this research as we only used publicly accessible or de-identified data.

Provenance and peer review Not commissioned; externally peer-reviewed.

Data availability statement Data are available upon reasonable request. Data are available on reasonable request.

ORCID iDs
Hongfei Ye http://orcid.org/0000-0002-7966-2466
Ping Fei http://orcid.org/0000-0001-7276-4630
Xiaoling Xie http://orcid.org/0009-0002-3255-0341
Meng Xie http://orcid.org/0009-0009-3071-3101
Peiquan Zhao http://orcid.org/0000-0002-5092-9550

REFERENCES

1 Placido D, Yuan B, Hjaltelin JX, *et al*. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat Med* 2023;29:1113–22.
2 Rajkomar A, Oren E, Chen K, *et al*. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018;1:18.
3 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.
4 Kung TH, Cheatham M, Medenilla A, *et al*. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Dig Health* 2023;2:e0000198.
5 Grünebaum A, Chervenak J, Pollet SL, *et al*. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol* 2023;228:696–705.
6 Potapenko I, Boberg-Ans LC, Stormly Hansen M, *et al*. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol* 2023;101:829–31.
7 Rasmussen MLR, Larsen A-C, Subhi Y, *et al*. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. *Graefes Arch Clin Exp Ophthalmol* 2023;261:3041–3.
8 Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol* 2023;141:589–97.
9 Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel pro differential diagnosis generator. *JFO Open Ophthalmol* 2023;1:100005.
10 Radford A, Wu J, Child R, *et al*. Language models are Unsupervised Multitask learners. 2019.
11 Li H, Guo D, Fan W, *et al*. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv* 2023.
12 THUDM. ChatGLM2-6B. Github Repository. Available: https://github.com/THUDM/ChatGLM2-6B [Accessed 13 Jul 2023].
13 Zeng A, Liu X, Du Z, *et al*. Glm-130b: an open bilingual pre-trained model. *arXiv* 2022.
14 King-yyf. CMeKG_Tools. Github Repository. Available: https://github.com/king-yyf/CMeKG_tools [Accessed 29 May 2023].
15 American Academy of Ophthalmology. Eyewiki. Available: https://eyewiki.aao.org/Main_Page [Accessed 29 May 2023].
16 Latapie H, Kilic O, Liu G, *et al*. A metamodel and framework for artificial general intelligence from theory to practice. *J AI Consci* 2021;08:205–27.
17 GanymedeNil. Hugging Face Model Repository. Available: https://huggingface.co/GanymedeNil/text2vec-large-chinese/tree/main [Accessed 29 May 2023].
18 Facebook AI research.. Github Repository. Available: https://github.com/facebookresearch/faiss [Accessed 30 May 2023].
19 Beijing Yishi Times Technology Development Co., Ltd. National medical e-book packages (in Chinese). Version 3.5.0. 2019. Available: http://www.imed.org.cn/ [Accessed 30 May 2023].
20 Jin D, Pan E, Oufattole N, *et al*. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *arXiv* 2020.
21 THUDM. ChatGLM2-6B. Hugging Face Repository. Available: https://huggingface.co/THUDM/chatglm-6b/blob/main/modeling_chatglm.py [Accessed 26 Jan 2023].
22 Bubeck S, Chandrasekaran V, Eldan R, *et al*. Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv* 2023.
23 Liu X, Fang C, Yan Z. Performance of ChatGPT on clinical medicine entrance examination for Chinese postgraduate in Chinese. *medRxiv* 2023.
24 Wei X. National Health Commission launches the white paper on eye health in China (in Chinese). *Chin Health Pict* 2020;52–3.

25 Song P, Wang J, Bucan K, *et al*. National and subnational prevalence and burden of glaucoma in China: a systematic analysis. *J Glob Health* 2017;7:020705.
26 Xu T, Wang B, Liu H, *et al*. Prevalence and causes of vision loss in China from 1990 to 2019: findings from the global burden of disease study 2019. *Lancet Public Health* 2020;5:e682–91.
27 Huang H, Tang T, Zhang D, *et al*. Not all languages are created equal in LLMs: improving multilingual capability by cross-lingual-thought prompting. *arXiv* 2023.
28 Li J, Cheng X, Zhao WX, *et al*. HaluEval: a large-scale hallucination evaluation benchmark for large language models. *arXiv* 2023.
29 Cui J, Li Z, Yan Y, *et al*. ChatLaw: open-source legal large language model with integrated external knowledge bases. *arXiv* 2023.
30 Liu H, Ning R, Teng Z, *et al*. Evaluating the logical reasoning ability of chatgpt and GPT-4. *arXiv* 2023.
31 Lin JC, Younessi DN, Kurapati SS, *et al*. Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination. *Eye (Lond)* 2023;37:3694–5.
32 Kwon T, Ong KTI, Kang D, *et al*. Large language models are clinical reasoners: reasoning-aware diagnosis framework with prompt-generated rationales. *arXiv* 2023.
33 Delsoz M, Raja H, Madadi Y, *et al*. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. *Ophthalmol Ther* 2023;12:3121–32.
34 Mijwil M, Aljanabi M, Ali AH. ChatGPT: exploring the role of cybersecurity in the protection of medical information. *MJCS* 2023;18–21.
35 Li H, Moon JT, Purkayastha S, *et al*. Ethics of large language models in medicine and medical research. *Lancet Digit Health* 2023;5:e333–5.