Science Bulletin 70 (2025) 934–942

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Article

Large language models for diabetes training: a prospective study

Haoxuan Li [a,1], Zehua Jiang [b,c,1], Zhouyu Guan [d,1], Yuqian Bao [d,1], Yuexing Liu [d], Tingting Hu [d], Jiajia Li [d,e], Ruhan Liu [d,e], Liang Wu [d], Di Cheng [d], Hongwei Ji [f], Yong Wang [g,h,i], Ya-Xing Wang [j], Carol Y. Cheung [k], Yingfeng Zheng [l], Jihong Wang [a], Zhen Li [a], Weibing Wu [a], Cynthia Ciwei Lim [m], Yong Mong Bee [n], Hong Chang Tan [n], Elif I. Ekinci [o,p,q], David C. Klonoff [r], Justin B. Echouffo-Tcheugui [s], Nestoras Mathioudakis [t], Leonor Corsino [u], Rafael Simó [v,w], Charumathi Sabanayagam [x,y], Gavin Siew Wei Tan [x], Ching-Yu Cheng [x,y,z], Tien Yin Wong [c,x], Huating Li [d,*], Chun Cai [d,*], Lijuan Mao [a,*], Lee-Ling Lim [aa,ab,ac,*], Yih-Chung Tham [x,y,ad,ae,*], Bin Sheng [a,d,e,*], Weiping Jia [d,*]

[a] *Shanghai University of Sport, Shanghai 200438, China*
[b] *School of Clinical Medicine, Beijing Tsinghua Changgung Hospital, Tsinghua Medicine, Tsinghua University, Beijing 100084, China*
[c] *Beijing Visual Science and Translational Eye Research Institute (BERI), Tsinghua Medicine, Tsinghua University, Beijing 100084, China*
[d] *Shanghai Belt and Road International Joint Laboratory of Intelligent Prevention and Treatment for Metabolic Diseases, Department of Computer Science and Engineering, School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Department of Endocrinology and Metabolism, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai Diabetes Institute, Shanghai Clinical Center for Diabetes, Shanghai 200240, China*
[e] *MoE Key Lab of Artificial Intelligence, Artificial Intelligence Institute, Shanghai Jiao Tong University, Shanghai 200240, China*
[f] *Department of Cardiology, the Affiliated Hospital of Qingdao University, Qingdao 266011, China*
[g] *CEMS, NCMIS, HCMS, MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*
[h] *School of Mathematics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100049, China*
[i] *Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China*
[j] *Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, Beijing Ophthalmology and Visual Sciences Key Laboratory, Beijing 100730, China*
[k] *Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong 999077, China*
[l] *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*
[m] *Department of Renal Medicine, Singapore General Hospital, SingHealth-Duke Academic Medical Centre, Singapore, 169608, Singapore*
[n] *Department of Endocrinology, Singapore General Hospital, Singapore, 169608, Singapore*
[o] *Department of Endocrinology, Austin Health, Melbourne, Heidelberg Victoria 3084, Australia*
[p] *Department of Medicine, The University of Melbourne (Austin Health), Melbourne, Parkville Victoria 3010, Australia*
[q] *Australian Centre for Accelerating Diabetes Innovations (ACADI), The University of Melbourne, Melbourne, Parkville Victoria, 3010, Australia*
[r] *Diabetes Research Institute, Mills-Peninsula Medical Center, San Mateo, CA 94010, USA*
[s] *Department of Medicine, Division of Endocrinology, Diabetes and Metabolism, Johns Hopkins School of Medicine, Baltimore, MD 21211, USA*
[t] *Division of Endocrinology, Diabetes, & Metabolism, Johns Hopkins University School of Medicine, Baltimore, MD 21211, USA*
[u] *Department of Medicine, Division of Endocrinology, Metabolism, and Nutrition, Duke University School of Medicine, Durham, NC 27710, USA*
[v] *Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Instituto de Salud Carlos III, Madrid, 08035, Spain*
[w] *Diabetes and Metabolism Research Unit, Vall d'Hebron Research Institut, Autonomous Univeristy of Barcelona, Barcelona, 08035, Spain*
[x] *Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, 168751, Singapore*
[y] *Ophthalmology and Visual Science Academic Clinical Program, Duke-NUS Medical School, Singapore, 169857, Singapore*
[z] *Centre for Innovation and Precision Eye Health; and Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, 119228, Singapore*
[aa] *Department of Medicine, Faculty of Medicine, Universiti Malaya, Kuala Lumpur, 50603, Malaysia*
[ab] *Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong 999077, China*
[ac] *Asia Diabetes Foundation, Hong Kong, China*
[ad] *Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, 119077, Singapore*
[ae] *Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, 119077, Singapore*

ARTICLE INFO

Article history:
Received 1 February 2024
Received in revised form 8 May 2024

ABSTRACT

Diabetes poses a considerable global health challenge, with varying levels of diabetes knowledge among healthcare professionals, highlighting the importance of diabetes training. Large Language Models (LLMs) provide new insights into diabetes training, but their performance in diabetes-related queries remains

* Corresponding authors.
E-mail addresses: huarting99@sjtu.edu.cn (H. Li), 1540422623@qq.com (C. Cai), maolijuan@sus.edu.cn (L. Mao), leeling.lim@ummc.edu.my (L.-L. Lim), thamyc@nus.edu.sg (Y.-C. Tham), shengbin@sjtu.edu.cn (B. Sheng), wpjia@sjtu.edu.cn (W. Jia).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2025.01.034
2095-9273/

Accepted 8 January 2025
Available online 27 January 2025

Keywords:
Diabetes
Diabetes training
Large language models
Primary diabetes care
Prospective study

uncertain, especially outside the English language like Chinese. We first evaluated the performance of ten LLMs: ChatGPT-3.5, ChatGPT-4.0, Google Bard, LlaMA-7B, LlaMA2-7B, Baidu ERNIE Bot, Ali Tongyi Qianwen, MedGPT, HuatuoGPT, and Chinese LlaMA2-7B on diabetes-related queries, based on the Chinese National Certificate Examination for Primary Diabetes Care in China (NCE-CPDC) and the English Specialty Certificate Examination in Endocrinology and Diabetes of Membership of the Royal College of Physicians of the United Kingdom. Second, we assessed the training of primary care physicians (PCPs) without and with the assistance of ChatGPT-4.0 in the NCE-CPDC examination to ascertain the reliability of LLMs as medical assistants. We found that ChatGPT-4.0 outperformed other LLMs in the English examination, achieving a passing accuracy of 62.50%, which was significantly higher than that of Google Bard, LlaMA-7B, and LlaMA2-7B. For the NCE-CPFC examination, ChatGPT-4.0, Ali Tongyi Qianwen, Baidu ERNIE Bot, Google Bard, MedGPT, and ChatGPT-3.5 successfully passed, whereas LlaMA2-7B, HuatuoGPT, Chinese LLaMA2-7B, and LlaMA-7B failed. ChatGPT-4.0 (84.82%) surpassed all PCPs and assisted most PCPs in the NCE-CPDC examination (improving by 1%–6.13%). In summary, LLMs demonstrated outstanding competence for diabetes-related questions in both the Chinese and English language, and hold great potential to assist future diabetes training for physicians globally.

1. Introduction

Diabetes constitutes a significant global health concern. The global age-standardized prevalence of diabetes is projected to rise by 59.7% between 2021 and 2050. It is estimated that by 2050, approximately 1.31 billion individuals will be affected by diabetes [1]. Long-term follow-up by endocrinology specialists is essential for monitoring diabetes and managing complications. Recent analyses of the medical workforce indicate a severe shortage of endocrinologists, a situation projected to deteriorate in the future [2]. Disparities exist in the levels of self-management knowledge among primary care physicians (PCPs) and patients [3–5]. Both groups, however, necessitate extensive professional education and training, which suggests that specialized training in diabetes care for PCPs is particularly crucial.

The advent of large language models (LLMs) appears to present a potential solution to this challenge. LLMs might offer professional guidance aid in primary care and daily management of diabetes in three ways: (1) by generating high-quality, human-like interactive text [6], (2) accessing clinical knowledge in the medical field [7–9], and (3) assisting in patient care [10–13]. Recently, LLMs specially developed in China, such as Baidu ERNIE Bot [14] and Alibaba Tongyi Qianwen[15], have exhibited strong performance in effectively addressing a wide range of queries. Moreover, significant medical LLMs like HuatuoGPT [16] and MedGPT [17] have also surfaced in China. The use of various LLMs has great potential for improving diabetes care and diabetes training by enhancing efficiency in medical consultations in both the English and Chinese languages [18].

One crucial measure of a physician's qualification is the successful completion of the medical licensing examination. Similarly, initial evaluation of the suitability of LLMs for medical practice involves assessing their performance in medical examinations. ChatGPT, a prominent open AI tool, has shown considerable promise in passing general medical licensing examinations, such as those in the United States[19–21], Australia [22], Peru [23], and Iran [24]. However, previous studies have suggested that ChatGPT's performance in subspecialty examinations has been controversial. While this LLM has exceeded passing scores in dermatology [25] and radiology [26], it has not met the subspecialty standards in orthopedics [27], plastic surgery[28], ophthalmology [29,30], and endocrinology[31]. There is currently no conclusive evidence of whether LLMs can provide accurate responses to diabetes-specific examination questions. Furthermore, most comparative studies of LLMs are based on English, while their performance in answering Chinese diabetes-specific questions remains unknown.

In this study, we first aimed to investigate the potential of LLMs for diabetes care by comparing the performance of ten popular LLMs to address diabetes-related queries in both the English and Chinese languages. Second, we aimed to determine whether LLMs would aid in preparing Chinese PCPs for medical examination questions related to diabetes care and to determine the LLM's ability to assist and correctness from feedback from PCPs.

2. Materials and methods

Ethical approval for this study was waived by the Ethics Committee of Shanghai Sixth People's Hospital because this study did not involve the collection of patient data or interventions for patients. All procedures were in accordance with the Declaration of Helsinki. Fig. 1 shows the whole study design flowchart.

2.1. Large language models inclusions

Ten popular LLMs were tested in this study, including: 1) ChatGPT-3.5 (Open AI), 2) ChatGPT-4.0 (Open AI), 3) Google Bard (Google), 4) LlaMA-7B (Meta AI), 5) LlaMA2-7B (Meta AI), 6) ERNIE Bot (Baidu), 7) Tongyi Qianwen (Alibaba), 8) MedGPT (Medlinker), 9) HuatuoGPT (Shenzhen Research Institute of Big Data and the Chinese University of Hong Kong, Shenzhen), and 10) Chinese LlaMA2-7B (GitHub). Detailed information on all the included LLMs is provided in Table 1.

ChatGPT-4.0 is available as a paid chatbot. It represents the fourth generation in the GPT foundation model series and is an improvement over the previous iteration, ChatGPT-3.5 [32,33]. Similarly, LlaMA2-7B [34] is an advanced version of LlaMA-7B [35], and Chinese LlaMA2-7B [36] is its Chinese counterpart with pre-trained Chinese language data. Google Bard was initially launched with the LaMDa LLM [37], and now utilizes the new PaLM2 model. It can access real-time information from the internet and deliver current data [38]. ERNIE Bot, Tongyi Qianwen, MedGPT, and HuatuoGPT are four LLM chatbots developed in China. Among them, MedGPT and HuatuoGPT are specifically medical LLMs while MedGPT was trained on real-world medical knowledge to diagnose nearly 3000 diseases [39] and a substantial Chinese medical corpus, in which MedGPT and HuatuoGPT are trained on a vast Chinese medical corpus [40].

2.2. Diabetes-related queries inclusions

Two sources of authoritative diabetes-related queries were included for LLM evaluation in Chinese and English. They were: the National Certificate Examination for Primary Diabetes Care in China (NCE-CPDC) [41] and sample questions from the Specialty Certificate Examination (SCE) in Endocrinology and Diabetes of

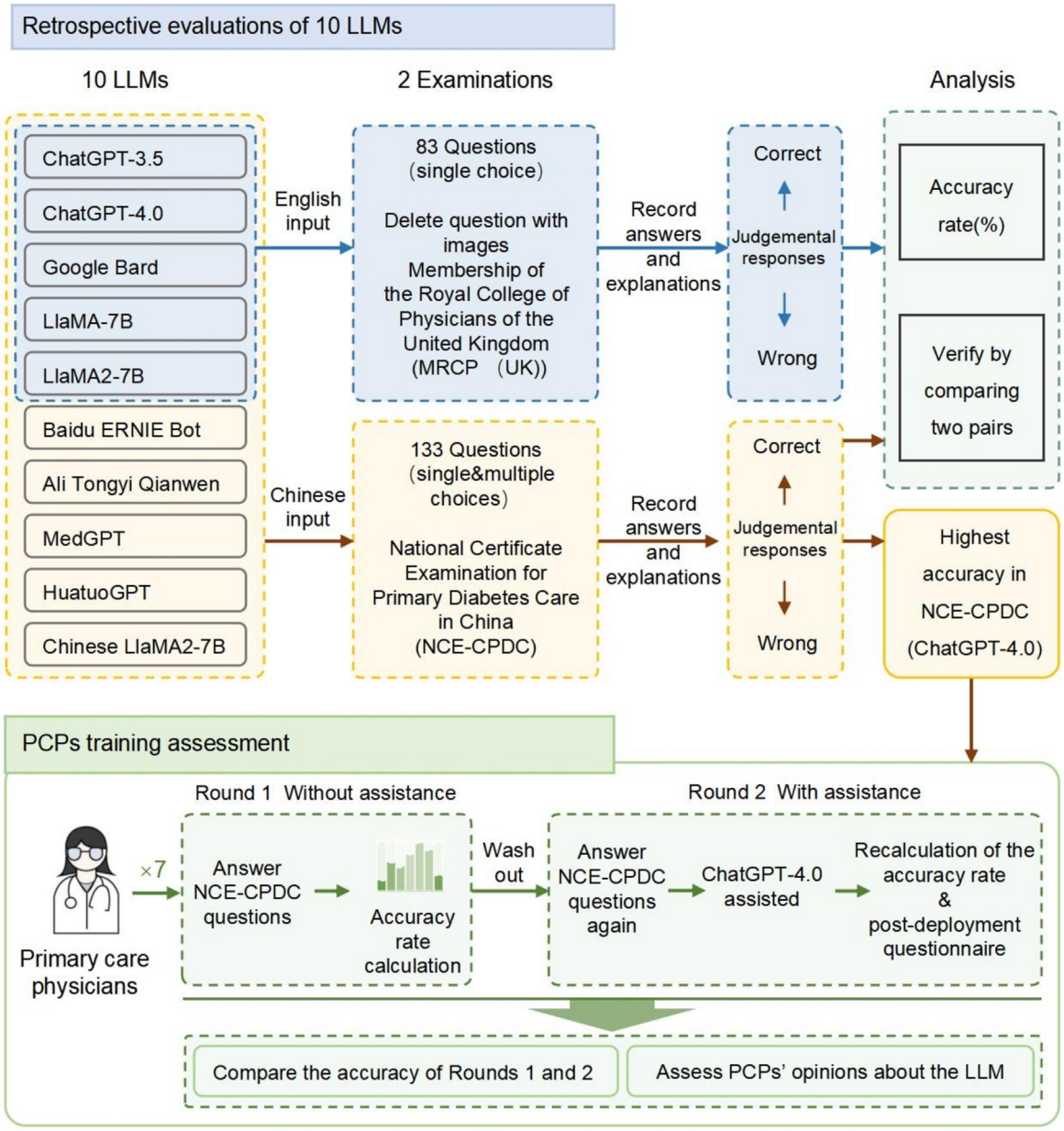

Fig. 1. Study Design Flowchart. A total of 5 LLMs were selected to take the MRCP (UK) and they answered 80 single-choice questions. 10 LLMs were selected to take the NCE-CPDC, and they answered 133 exam questions, including both single-choice and multiple-choice questions. The examination questions were sampled from official sources and excluded those with images. Subsequently, the LLMs' responses were differentiated based on the official answers and subjected to statistical analysis for comparison. Then, the LLMs with the highest correct rates were selected to participate in the subsequent stage of the PCPs' examinations. The PCPs were then given the Chinese examinations, with and without the assistance of the LLMs, and the results were analysed to determine the extent to which the LLMs could play a supporting role. Additionally, feedback from the PCPs was taken into account.

Membership of the Royal College of Physicians of the United Kingdom (MRCP [UK]) [42].

The NCE-CPDC is a professional examination designed for PCPs, who have completed formal training overseen by the Chinese Medical Doctor Association and organized by the National Office for Primary Diabetes Care of China. The training and examination are based on the National Guidelines for the Prevention and Control of Diabetes in Primary Care [41] and the certification of NCE-CPDC is widely recognized in China, particularly by healthcare professionals and individuals involved in diabetes care. Results with accuracy rates exceeding 60% were considered as passing and lower scores were deemed as failing. Sample questions (n = 133) from the NCE-CPDC question bank were used to evaluate the performance of ten selected LLMs, encompassing both single- and multiple-choice questions. These questions cover the area of definition, screening, diagnosis, referral, lifestyle interventions, drug therapies, and management of acute and chronic diabetes complications in primary diabetes care. These questions are non-public and come from the National Office for Primary Diabetes Care of China.

The MRCP (UK) conducts a series of postgraduate medical examinations that are widely recognized as benchmarks of excellence in the medical profession. SCEs are tailored to evaluate the knowledge, skills, and clinical acumen of specialists in various fields [42]. The passing mark for this examination was established by using the Hofstee compromise method, which combines the mean scores of the standard-setting group with UK trainees' performance. For the 2019 Endocrinology and Diabetes examination

Table 1
Detailed information on the included LLMs.

Included LLMs	Institutes	Date	Size (billion)[a]	How we accessed the LLM
ChatGPT-3.5	Open AI	December 2022	175	Accessed from https://chat.openai.com/
ChatGPT-4.0	Open AI	March 2023	Not available	Accessed from https://chat.openai.com/
Google Bard	Google	February 2023	137	Accessed from bard.google.com/
LlaMA-7B	Meta /Facebook	February 2023	7	Downloaded from https://llama.meta.com/
LlaMA2-7B	Meta /Facebook	July 2023	7	Accessed from https://labs.perplexity.ai/
ERNIE Bot	Baidu	July 2023	100	Accessed from https://yiyan.baidu.com/
Ali Tongyi Qianwen	Alibaba	April 2023	72	Accessed from https://tongyi.aliyun.com/
MedGPT	Medlinker	April 2023	100	Accessed from https://medgpt.co/
HuatuoGPT	Shenzhen Research Institute of Big Data and the Chinese University of Hong Kong, Shenzhen	June 2023	13	Accessed from https://www.huatuogpt.cn/
Chinese LlaMA2-7B	GitHub	July 2023	7	Accessed from https://chinese.llama.family/

LLM: large language model.
[a] The number of parameters that the model was trained on.

the pass mark was set at 60.5% (121/200) [43]. The sample questions from the Endocrinology and Diabetes SCE in MRCP (UK) can provide some insight into the performance of LLMs in addressing professional questions related to endocrinology and diabetes, but the passing accuracy rate can not be defined in this study. After excluding three image-based questions, 80 single-choice sample questions in English from the Endocrinology and Diabetes SCE in MRCP (UK) were included to assess the performance of five LLMs, namely ChatGPT-3.5, ChatGPT-4.0, Google Bard, LlaMA-7B, and LlaMA2-7B (Supplementary Material). These 80 questions included diabetes-related questions, which cover the area of pathophysiology, diagnosis, drug therapies, and management of acute and chronic diabetes complications in specialty diabetes care.

2.3. LLM performance comparisons

For each chatbot instance, a new account with no prior conversation history was utilized to ensure that the study remained unaffected by existing dialogue data. The questions and options of the diabetes care question banks from the NCE-CPDC were uploaded in Chinese, into all ten LLMs. The 80 questions for the SCE in endocrinology and diabetes of MRCP (UK) were uploaded in English into five LLM chatbots, including ChatGPT-3.5, ChatGPT-4.0, Google Bard, LlaMA-7B, and LlaMA2-7B. Both question banks required each LLM to provide the correct choice and corresponding explanations. Answers and explanations for each question from every chatbot were cross-referenced with official answers to determine the accuracy rate for assessing the performance of each chatbot.

2.4. Primary care physician training

ChatGPT-4.0 was further utilized in the training of Chinese PCPs by evaluating seven PCPs' performance in the NCE-CPDC sample questions (n = 133) without and with the assistance of ChatGPT-4.0. All PCPs had clinical practical experience ranging from 2 to 11 years (PCP#1-7 years, PCP#2-3 years, PCP#3-4 years, PCP#4-2 years, PCP#5-11 years, PCP#6-2 years, and PCP#7-5 years). They were each required to undergo two rounds of NCE-CPDC testing, one without any help and one with the assistance of ChatGPT-4.0. The first round necessitated independent completion of the NCE-CPDC examination by all PCPs, and their accuracy rates were recorded. To ensure consistency in the timing of the experiment, ChatGPT-4.0 also participated in the first-round testing simultaneously by answering all the NCE-CPDC questions. After a 3-week washout period, PCPs were notified of the second round of testing, which was conducted with the assistance of ChatGPT-4.0. Before the second round of testing began, all participants were informed that they could use ChatGPT-4.0 to inquire about exam questions but were also informed that the answers in the ChatGPT-4.0 were not completely correct and that they would have to rely on themselves to recognize the correctness of its answers and explanations. The accuracy rate of all participants in the second round of the NCE-CPDC examination was then evaluated and compared with their performance in the first round. Furthermore, PCPs who received assistance from ChatGPT-4.0 in the prospective study were asked to fill out a post-deployment questionnaire upon completion of the study. The questionnaire consisted of three items aimed at gauging the PCPs' opinions about the LLM. Each query was rated on a Likert scale ranging from 1 to 5, with 1 representing strong disagreement and 5 representing strong agreement.

2.5. Statistical analysis

A comprehensive statistical analysis of the performance of all LLMs and PCPs in providing correct responses was conducted with SPSS (IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.). The accuracy rate (%) of each chatbot and each PCP was calculated as the percentage of correct answers out of the total number of questions. Chi-square tests were employed to compare chatbot performance between different LLMs, with the adjusted p-values (Bonferroni correction) considered statistically significant when lower than 0.05. Paired t-tests were used to compare the PCPs' performance in the NCE-CPDC examination without and with ChatGPT-4.0's assistance. ($P < 0.05$ was considered to be statistically significant).

3. Results

The performance of five LLM-chatbots with English input in the SCE in endocrinology and diabetes of MRCP (UK) examination is depicted in Fig. 2. The accuracy rate of ChatGPT-4.0 was higher than the pass mark, achieving an accuracy of 62.50%, which was significantly higher than that of Google Bard (32.5%), LlaMA-7B (15.0%) and LlaMA2-7B (20.0%) (Chi-square test, adjusted $P < 0.05$). LlaMA-7B performed the worst, with a rate of correct responses of only 15.0%, which was significantly lower than those of ChatGPT-4.0 (62.5%), ChatGPT-3.5 (45.0%), and Google Bard (32.5%) (Chi-square test, adjusted $P < 0.05$).

The performance of ten LLM-chatbots with Chinese input in the NCE-CPDC exam is depicted in Fig. 3. ChatGPT-4.0 (accuracy rate 90.98%), Ali Tongyi Qianwen (81.20%), Baidu ERNIE Bot (71.43%), Google Bard (68.42%), MedGPT (67.67%), and ChatGPT-3.5 (63.16%) successfully passed the examination, whereas LlaMA2-7B (34.59%), HuatuoGPT (30.83%), Chinese LLaMA2-7B (24.81%), and LlaMA-7B (24.06%) failed (Fig. 3a). ChatGPT-4.0 significantly

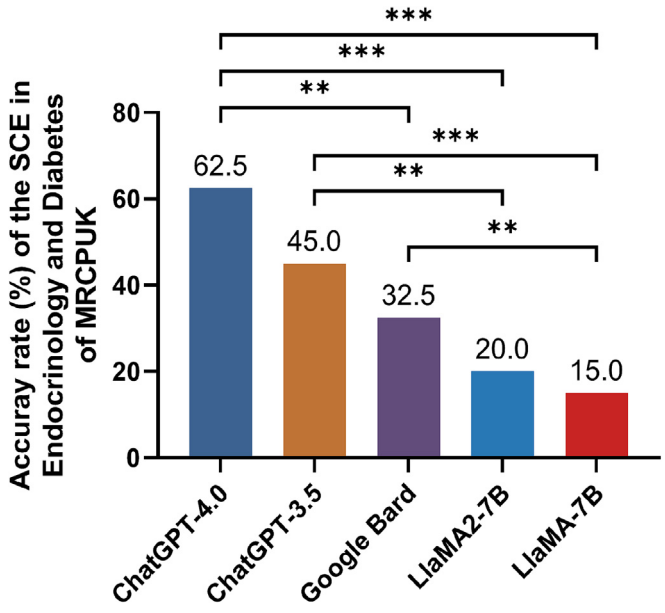

Fig. 2. The accuracy rate (%) in the Endocrinology and Diabetes SCE of MRCP (UK) with English Language Input of Five LLM-Chatbots. (Chi-square tests with Bonferroni corrected *P* values, ** adjusted $P < 0.01$, *** adjusted $P < 0.001$).

outperformed the other nine LLMs except Ali Tongyi Qianwen (Chi-square test, adjusted $P < 0.05$). The accuracy rate of the Ali Tongyi Qianwen was significantly superior to the ChatGPT-3.5 (Chi-square test, adjusted $P < 0.05$) in the NCE-CPDC exam. Additionally, Ali Tongyi Qianwen, Baidu ERNIE Bot, Google Bard, and MedGPT all performed significantly better than LlaMA-7B, LlaMA2-7B, HuatuoGPT, and Chinese LLaMA2-7B (Chi-square test, adjusted $P < 0.05$) (Fig. 3b).

All seven PCPs successfully passed the NCE-CPDC examination in the first-round assessment with a mean accuracy rate of 74.72% (±5.10), ranging from 68.57% to 81.16%. Notably, ChatGPT-4.0 demonstrated excellent performance in the NCE-CPDC, achieving an accuracy rate of 84.82%, surpassing all the PCPs in the first-round examination. In the second-round assessment, with the assistance of ChatGPT-4.0, the mean accuracy rate was 75.8% (±5.62). Most PCPs improved their accuracy rates from 1 to 6.13% in correctly answering the NCE-CPDC examination except PCP#3, (whose score decreased by 8.54%) and PCP#6 (whose score decreased by 0.92%) (Paired *t*-test, $P < 0.05$). Moreover, after training with ChatGPT-4.0, one participant's performance on the NCE-CPDC examination even exceeded that of everyone's ChatGPT-4.0 score with an accuracy rate of 85.57% (P5). Detailed performances of ChatGPT-4.0 and all seven PCPs in the first and second rounds are shown in Fig. 4.

Additionally, to capture the PCPs' perceptions and satisfaction with the LLM after using its insights, the seven PCPs who participated in the prospective study were asked to complete a post-deployment questionnaire. Across these PCPs, the LLM obtained an average score of 4.57 for internal concordance (out of 5.00), 4.33 for insights, and 4.71 for willingness to use in future diabetes training. (Table 2).

4. Discussion and conclusion

Diabetes management poses a significant global challenge because of its complex nature [1]. Multifaceted issues such as patient education, accessibility to healthcare, cost-effective treatments, and diabetes training demand innovative strategies[44]. Our study provided a thorough understanding of the LLMs' ability to comprehend diabetes-related knowledge with distinct topics, question styles, and languages based on the NCE-CPDC and the SCE in Endocrinology and Diabetes of MRCP (UK) examinations. ChatGPT-4.0 demonstrated superior performance in both the NCE-CPDC and the SCE in Endocrinology and Diabetes of the MRCP (UK), conducted in Chinese and English, respectively. Furthermore, ChatGPT-4.0 also outperformed PCPs and exhibited the potential to improve PCPs' performance in the NCE-CPDC examinations, especially for PCPs with more years of clinical practice.

Recently, some studies have demonstrated the efficacy of LLMs in examinations of medical topics. Rosoł et al. [45] assessed the performance of GPT-3.5 and GPT-4 in Polish and English for the Polish Medical Final Examination (Lekarski Egzamin Końcowy, or LEK). ChatGPT-4 consistently outperformed GPT-3.5 and outperformed students who graduated over 2 years ago. Novak et al. [46] conducted a study to assess the accuracy of Google Bard, GPT-3.5 Turbo, and GPT-4.0 in answering cardiology-specific questions of varying difficulty levels. In that study, LLMs showed promising results in possessing an ability to interpret and apply complex clinical guidelines, with a potential for enhancing patient outcomes through personalized advice. Skalidis et al. [47] challenged ChatGPT to answer questions from a demanding exam, post-graduate exam—the European Exam in Core Cardiology (EECC), which is the final exam for the completion of specialty training in Cardiology in many countries. Fijačko et al. [48] employed ChatGPT' to answer the Heart Association (AHA) Basic Life Support (BLS) and Advanced Cardiovascular Life Support (ACLS) exams. Although ChatGPT did not qualify for the exam, the authors recognized its capabilities and believe that ChatGPT has shown promising results in becoming a powerful reference and self-learning tool for preparing for the life support exams. Although various studies have indicated that ChatGPT performs well on various exams, there are very few studies comparing it with multiple other LLMs, although it is worthwhile to do a comparative analysis of the abilities of multiple LLMs because there are now many LLMs claiming that they have the same abilities as or can exceed the ability of ChatGPT.

The SCE in Endocrinology and Diabetes of MRCP (UK) is a highly specialized professional examination tailored for endocrinologists and diabetes care specialists. The exam's difficulty can be reflected in the 2023 pass rate, where only 28.6% of all candidates successfully passed, of which 48.5% were UK trainees [49]. The passing score for the 2019 Endocrinology and Diabetes examination was established at 60.5%. Among the five included LLMs, ChatGPT-4.0 achieved the highest accuracy rate of 62.5%, vividly demonstrating its proficiency in addressing complex endocrinology and diabetes care queries. A previous study also investigated the performance of Bard and ChatGPT using an investigator-designed set of 100 multiple-choice questions (MCQs) covering endocrinology (50 MCQs), diabetes, and diabetes technology (50 MAQs) [31]. The results showed that both Bard (49%) and ChatGPT (52%) failed in this investigator-designed examination, while in the diabetes and diabetes technology segment, ChatGPT outperformed Bard, attaining 46% accuracy compared to Bard's 40%. Further training of these LLMs using diabetes-related data and clinical-related knowledge will be required to improve their performance.

The NCE-CPDC covers questions related to primary diabetes care in the Chinese language, which can reflect LLMs' abilities to grasp diabetes primary care knowledge. ChatGPT-4.0 performed extremely well in NCE-CPDC with over 90% accuracy and outperformed all nine other LLMs except Ali Tongyi Qianwen, which means that it has a solid foundation of diabetes-related knowledge and may potentially be used to assist the training of Chinese grassroots diabetes care doctors. The test performance of ChatGPT-4.0 attests to the linguistic capabilities of the LLM, which enable it to meet certain standards required for medical professionals. LLMs

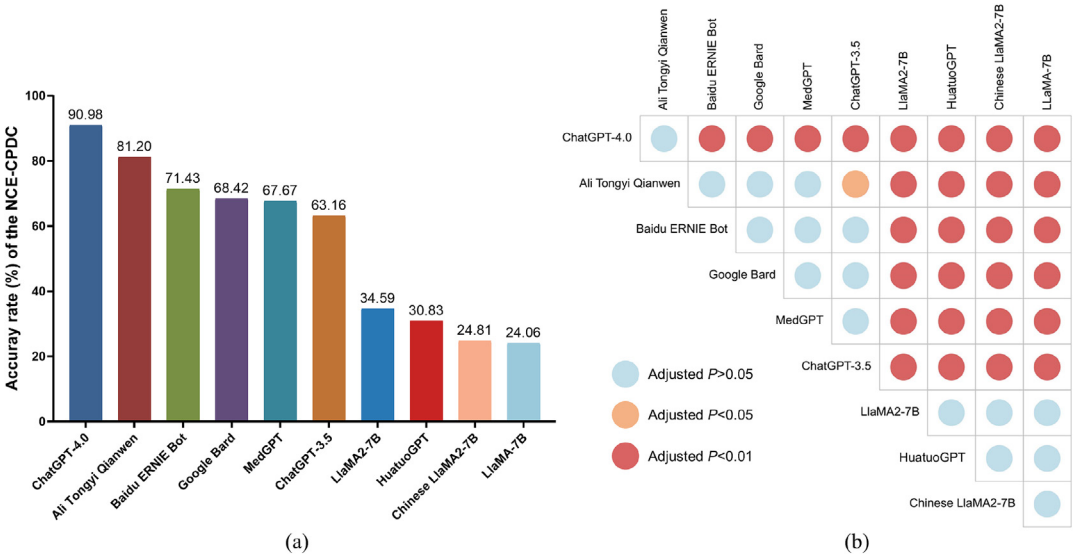

Fig. 3. (a) The accuracy rates (%) in NCE-CPDC with the Chinese Language between Ten LLM-Chatbots. (b) Presentation of Statistical Significance between Different Chatbots. (Chi-square tests with Bonferroni corrected *P* values).

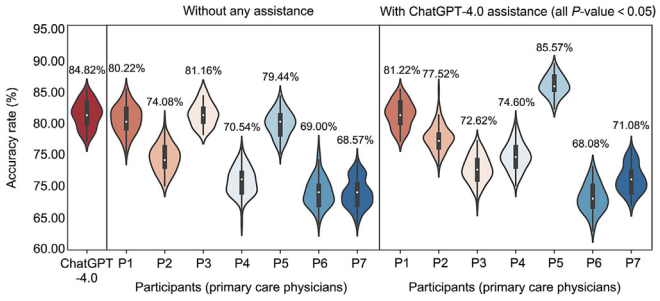

Fig. 4. A comparison of PCPs' Performance in NCE-CPDC Examination without and with the assistance of ChatGPT-4.0 (paired *t*-test, $P < 0.05$).

Table 2
Post-deployment assessment by PCPs using the ChatGPT-4.0.

Evaluation items	Mean score [a]	Standard error
Did the explanation contents of the large language model demonstrate internal concordance?	4.57	0.53
Did the explanation contents of the large language model offer you insights to answer the questions?	4.43	0.79
Do you want to use the large language model in the future diabetes training?	4.71	0.49

PCP: primary care physician.
[a] Every question was scored from Likert scales ranging from 1 to 5, with 1 representing strong disagreement and 5 representing strong agreement.

exhibited a higher error rate when dealing with multiple-choice questions, compared to single-choice questions, related to case analyses in NCE-CPDC. Even the top-performing LLM (ChatGPT-4.0) demonstrated a significant proportion of errors, predominantly associated with multiple-choice questions and case analysis items, which may be attributed to not having undergone specific training in medical knowledge integration and clinical thinking. This difficulty with multiple-choice questions highlights current limitations in the effective application of LLMs in clinical practice and also underscores the necessity for regular and systematic evaluation, supervision, and fine-tuning of chatbots to ensure consistent, trustworthy, and clinically safe medical recommendations.

LLMs developed in China have experienced rapid growth and advancement, yet few studies have explored their performance in the medical field. Notably, we discovered that the absolute accuracy rate of Ali Tongyi Qianwen was only slightly lower than ChatGPT-4.0 and significantly superior to ChatGPT-3.5. Specifically, the absolute accuracy rate of Ali Tongyi Qianwen and Baidu ERNIE Bot surpassed that of the Google Bard, and although MedGPT was slightly inferior to the Google Bard, it was better than ChatGPT-3.5. All these findings indicate the significant potential of LLMs developed in China within the medical realm and demonstrate their competitiveness in the emerging era of widespread use of LLMs to solve problems. Moreover, although both MedGPT and HuatuoGPT are prominent medical language models with enriched medical knowledge, MedGPT significantly outperformed HuatuoGPT. Impressively, although Ali Tongyi Qianwen and Baidu ERNIE Bot were trained without specialized medical knowledge, they demonstrated remarkable performance in addressing diabetes-related issues at the primary care level and achieved higher absolute accuracy rates than MedGPT, which was trained with specialized medical knowledge.

Interestingly, as the most widely used open-source LLM, LlaMA has spawned numerous fine-tuning models. However, the performance of the LlaMA-related LLMs, such as LlaMA-7B, LlaMA2-7B, HuatuoGPT, and Chinese LlaMA2-7B, proved to be unsatisfactory in both the Chinese and English language versions. Even though HuatuoGPT is an LLM trained with clinical medical knowledge, its effectiveness in diabetes care remains subpar. A previous study developed a medical fine-tuning LLM based on the LLaMA framework [50] called ChatDoctor with higher accurate scores than ChatGPT in answering English questions from the iCliniq database. However, its performance on Chinese questions has not yet been investigated. Hence, the development of a knowledge training database specific to diabetes care is essential to enhance the capabilities of fine-tuning LlaMA models in addressing diabetes-related issues, especially in different languages.

Most LLM-chatbots can provide relatively accurate answers and instructive explanations. The best-performing ChatGPT-4.0 in our study can achieve excellent accuracy in English and Chinese language examinations. This is why we assessed its potential to assist in the training of Chinese PCPs by comparing the performance of PCPs in the NCE-CPDC examination based

on traditional training and the training with the performance of the same PCPs based on receiving training with ChatGPT-4.0. Traditionally training of PCPs can help them pass the NCE-CPDC examination, but the accuracy rate varies from 68.57% to 81.16%, all are inferior to ChatGPT-4.0's performance (84.82%). In the second round of testing, having the assistance of ChatGPT 4.0, most of the participants showed improvement in the final passing accuracy rate. One PCP with 11 years of clinical practical experience even surpassed the performance of ChatGPT-4.0. However, two of the PCPs' performance after using ChatGPT-4.0 as the assistance conversely decreased, which may be attributed to their incapability of identifying the misleading explanations provided by ChatGPT-4.0. Notably, almost all PCPs showed high internal concordance, insights, and willingness to use LLMs as an assistive tool for diabetes training. Therefore, on the one hand, ChatGPT-4.0 could assist PCPs training for better comprehension of diabetes care knowledge and guidelines, but on the other hand, delivering misleading responses and hallucinations could happen and cannot be ignored. This is the reason why we chose PCPs for the LLMs adjunctive test rather than the general population.

This study evaluated the performance of prominent LLMs in the English SCE in Endocrinology and Diabetes of MRCPUK and the Chinese NCE-CPDC examinations, assessing their medical knowledge

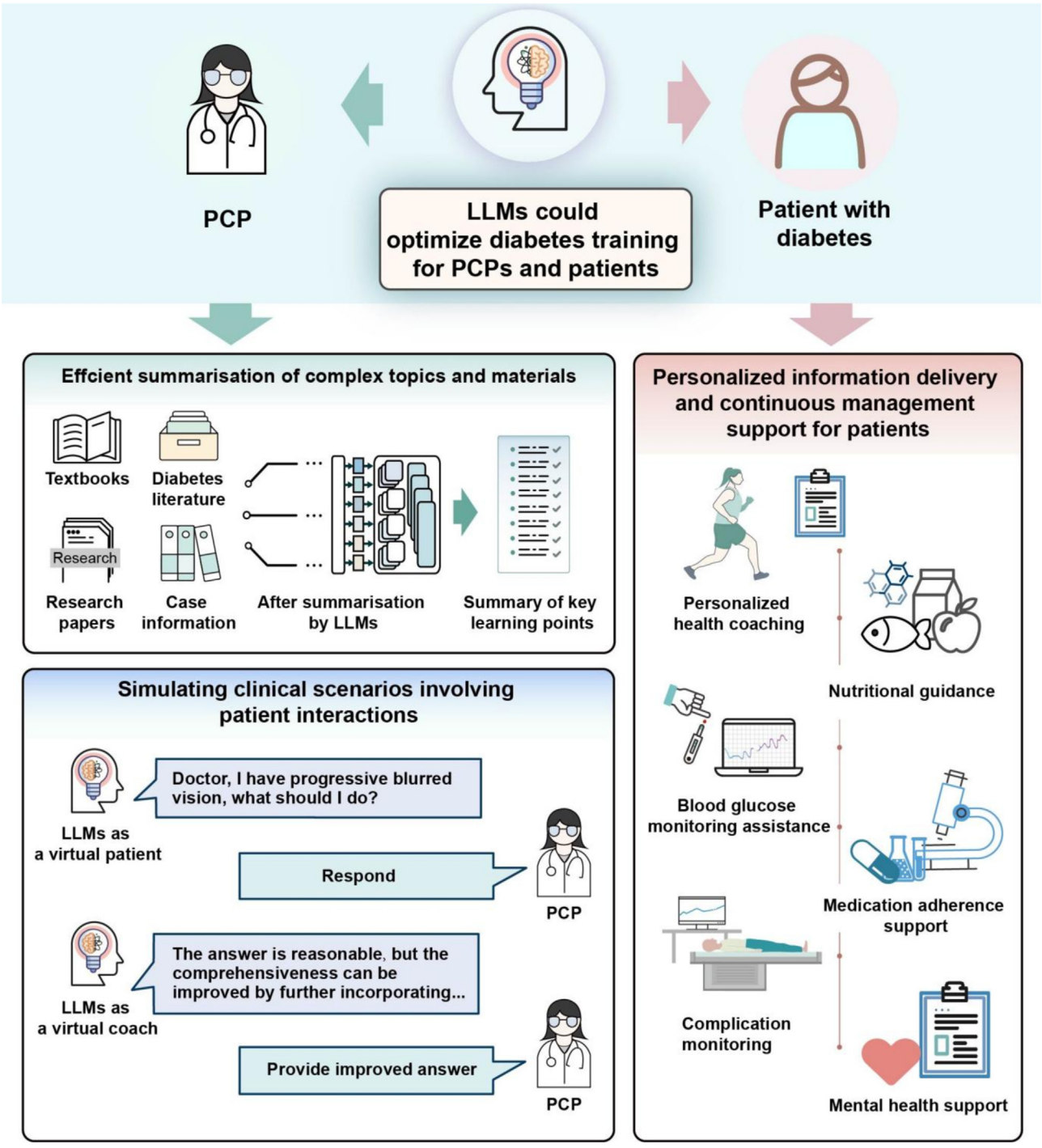

Fig. 5. Future prospects of large language models (LLMs) in diabetes training for primary care physicians (PCPs) and patients with diabetes.

and reasoning capabilities in the context of diabetes and diabetes training. Overall, most LLMs demonstrated a broad knowledge base and reasoning ability when answering diabetes questions, exemplified by models such as ChatGPT-4.0, Ali Tongyi Qianwen, and Baidu ERNIE Bot. Therefore, the LLMs' capability for equipping doctors with fundamental diabetes knowledge and the potential in diabetes training are significant and hold promising prospects for the future. Concurrently, there is mounting evidence that ChatGPT has been effective in medical licensing examinations, which could potentially result in a transformation of medical training. Although at present LLMs can efficiently process medical information and provide appropriate answers to questions, it is not a substitute for critical thinking, innovation, and creativity, which are essential traits for physicians. [47]

Our study had limitations. First, the selection of the 10 LLMs for evaluation was based on their popularity and availability at the time of the research; smaller or quantized models, which may have different performance characteristics, were not considered in this analysis. Second, the present study is limited by the current technical difficulties and complexity of medical diagnostics, which impede the ability of LLM to assist physicians in the diagnosis of diabetes. While the results indicate that LLMs are currently performing better in the area of direct medical reasoning about diabetes-related problems, no LLM received a perfect score on the exam. This suggests that there are still no LLMs that can be fully trusted. Similarly, their ability to assist in the diagnosis and treatment of more complex clinical problems and to help people with diabetes make decisions about their health management still requires the careful judgment of a medical professional. This indicates that while some current LLMs (e.g., ChatGPT-4.0) can be queried by healthcare professionals and patients for responses to specific medical inquiries, their responses are for informational purposes only and cannot currently be considered a reliable source of information for healthcare practitioners and patients. Third, the phenomenon of AI hallucination remains an unavoidable issue. Therefore, we should not neglect the possibility of hallucinations in the responses that LLMs provide when we use them as medical assistants. If AI tends to fabricate facts when answering questions, then it will mislead individuals lacking professional knowledge, necessitating the continued necessity for careful judgment in LLM answers and decisions. For them to be applied to actual medical diagnosis, they still require further professional training. Fourth, only three questions were included in the questionnaire to evaluate PCPs' perceptions and satisfaction with the LLM after using its insights. Further large-scale studies are warranted to comprehensively assess PCPs' attitudes towards the use of LLMs in diabetes training.

Looking ahead, LLMs fine-tuned with domain-specific knowledge have the potential to optimize diabetes training for both PCPs and patients with diabetes (Fig. 5). LLMs could swiftly summarise extensive texts, academic papers, or comprehensive diabetes literature, to distill key points and offer concise overviews on specific subjects. For instance, LLM combined with a deep learning model could bring surprising effects for physicians to give better clinical recommendations[51]. LLMs also could serve as virtual patients or coaches, by generating case studies and simulating clinical scenarios to help PCPs improve their communication and decision-making skills for primary diabetes care. For patients with diabetes, LLMs could potentially offer valuable support to patients by enhancing communication, providing personalized information, and assisting in continuous management. Therefore, LLMs could serve as a bridge and platform for integrating multiple technologies. Customized AI-LLM doctors offer patients the flexibility to seek medical advice and support anytime, anywhere. The personalized patient education, diagnoses, and treatment recommendations provided by LLMs allow doctors to stay updated on patients' conditions daily through the integration of diverse technological tools like voice, video, and virtual reality. This advancement shows potential for improving doctor-patient communication, cultivating positive relationships between healthcare providers and patients, and ushering in a new era in diabetes management.

In conclusion, our study assessed the performance of five prominent LLMs on English language-based diabetes care examinations and ten prominent LLMs on Chinese language-based diabetes care examinations. The performance of these LLMs demonstrated a broad knowledge base, strong reasoning ability, and the potential to assist in PCPs' diabetes training in many ways, which is significant for making decisions for individual patients and populations. The innovative application of LLMs in diabetes training signifies a transformative shift toward personalized, comprehensive, and accessible diabetes care and management. However, continued improvement and validation are essential to ensure their applications offer robust protection for public health.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0509202 and 2023ZD0509201), National Natural Science Foundation of China (62077037,8238810007, 82022012, 81870598, 62272298 and 82388101), the National Key Research and Development Program of China (2022YFC2502800 and 2022YFC2407000), the Shanghai Municipal Key Clinical Specialty, Shanghai Research Center for Endocrine and Metabolic Diseases (2022ZZ01002), the Chinese Academy of Engineering (2022-XY-08), the Innovative Research Team of High-level Local Universities in Shanghai (SHSMU-ZDCX20212700) and Beijing Natural Science Foundation (IS23096).

Author contributions

Haoxuan Li, Zehua Jiang, and Zhouyu Guan conceptualized and designed the research, analyzed the data, and wrote the initial draft of the manuscript. Haoxuan Li, Zehua Jiang, Zhouyu Guan and Jiajia Li performed verification of the experimental design. Ruhan Liu, Yuqian Bao, Yuexing Liu, Tingting Hu, Liang Wu, Di Cheng, Hongwei Ji, Yong Wang, Ya-Xing Wang, Carol Y. Cheung, Yingfeng Zheng, Jihong Wang, Zhen Li, Weibing Wu, Cynthia Ciwei Lim, Yong Mong Bee, Hong Chang Tan, Elif I. Ekinci, David C. Klonoff, Justin B. Echouffo-Tcheugui, Nestoras Mathioudakis, Leonor Corsino, Rafael Simó, Charumathi Sabanayagam Sabanayagam, Gavin Siew Wei Tan, Ching-Yu Cheng, and Tien Yin Wong reviewed and edited the first manuscript and suggested revisions. Huating Li, Chun Cai, Lijuan Mao, Lee-Ling Lim, Yih-Chung Tham, Bin Sheng, and Weiping Jia conducted the experimental idea construction and supervised the whole experiment.'

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scib.2025.01.034.

References

[1] Collaborators GBDD. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023;402:203–34.
[2] Ashrafzadeh S, Hamdy O. Patient-driven diabetes care of the future in the technology era. Cell Metab 2019;29:564–75.

[3] Martens TW, Simonson GD, Carlson AL, et al. Primary care and diabetes technologies and treatments. Diabetes Technol Ther 2021;23:S143–58.
[4] Rushforth B, McCrorie C, Glidewell L, et al. Barriers to effective management of type 2 diabetes in primary care: qualitative systematic review. Br J Gen Pract 2016;66:e114–27.
[5] Cho MK, Kim MY. Self-management nursing intervention for controlling glucose among diabetes: a systematic review and meta-analysis. Int J Environ Res Public Health 2021;18:12750.
[6] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620:172–80.
[7] Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616:259–65.
[8] Haupt CE, Marks M. AI-generated medical advice—GPT and beyond. JAMA 2023;329:1349–50.
[9] De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120.
[10] Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: evaluation study. JMIR Nurs 2023;6:e47305.
[11] Sng GGR, Tung JYM, Lim DYZ, et al. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. Diabetes Care 2023;46:e103–5.
[12] Nakhleh A, Spitzer S, Shehadeh N. ChatGPT's response to the diabetes knowledge questionnaire: implications for diabetes education. Diabetes Technol Ther 2023;25:571–3.
[13] Guan Z, Li H, Liu R, et al. Artificial intelligence in diabetes management: advancements, opportunities, and challenges. Cell Rep Med 2023;4:101213.
[14] Exclusive: Baidu Wenxin Large Model 3.5 enters internal testing, real-world performance exceeds ChatGPT. https://finance.sina.com.cn/wm/2023-06-20/doc-imyxyaxf8235213.shtml. Accessed 16 December 2024.
[15] Carter R. What is Tongyi Qianwen? Alibaba's ChatGPT Rival. UC Today. https://www.uctoday.com/unified-communications/what-is-tongyi-qianwen-alibabas-chatgpt-rival/. Accessed 16 December 2024.
[16] Zhong H, Chen J, Jiang F, et al. HuatuoGPT, towards taming language model to be a doctor. arXiv 2023; 2305.15075v1.
[17] Kraljevic Z, Shek A, Bean D, et al. MedGPT: medical concept prediction from clinical narratives. arXiv 2021; 2107.03134.
[18] Sheng B, Guan Z, Lim LL, et al. Large language models for diabetes care: potentials and prospects. Sci Bull 2024;69:583–8.
[19] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198.
[20] Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? the implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312.
[21] Yaneva V, Baldwin P, Jurich DP, et al. Examining ChatGPT performance on USMLE sample items and implications for assessment. Acad Med 2024;99:192–7.
[22] Kleinig O, Gao C, Bacchi S. This too shall pass: the performance of ChatGPT-3.5, ChatGPT-4 and new bing in an Australian medical licensing examination. Med J Aust 2023;219:237.
[23] Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the peruvian national licensing medical examination: cross-sectional study. JMIR Med Educ 2023;9:e48039.
[24] Ebrahimian M, Behnam B, Ghayebi N, et al. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health Care Inform 2023;30:e100815.
[25] Lewandowski M, Łukowicz P, Świetlik D, et al. An original study of ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the dermatology specialty certificate examinations. Clin Exp Dermatol 2023:llad255.
[26] Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023;307:e230582.
[27] Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg 2023;31:1173–9.
[28] Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J 2023;43:Np1078-np82.
[29] Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023;3:100324.
[30] Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023;141:589–97.
[31] Meo SA, Al-Khlaiwi T, AbuKhalaf AA, et al. The scientific knowledge of Bard and ChatGPT in endocrinology, diabetes, and diabetes technology: multiple-choice questions examination-based performance. *J Diabetes Sci Technol.* 2023: 19322968231203987.
[32] Wikipedia contributors. GPT-3. Wikipedia. https://en.wikipedia.org/wiki/GPT-3#GPT-3.5. Accessed 16 December 2024.
[33] OpenAI. GPT-4 research. OpenAI. https://openai.com/index/gpt-4-research/. Accessed 16 December 2024.
[34] Hugo T, Thibaut L, Gautier I, et al. LLaMA: open and efficient foundation language models. arXiv 2023; 2302.13971.
[35] Meta. Introducing LLaMA: a foundational, 65-billion-parameter large language model. https://ai.meta.com/blog/large-language-model-llama-meta-ai/. Accessed 16 December 2024.
[36] Hugo T, Louis M, Kevin S, et al. Llama 2: open foundation and fine-tuned chat models. arXiv 2023; 2307.09288.
[37] Krawczyk J. Bard is getting better at logic and reasoning. Google. https://blog.google/technology/ai/bard-improved-reasoning-google-sheets-export/. Accessed 16 December 2024.
[38] Pichai, S. An important next step on our AI journey. https://blog.google/technology/ai/bard-google-ai-search-updates/. Accessed 16 December 2024.
[39] MedLinker. China's first AI doctor MedGPT was launched. https://www.medlinker.com/index/xinwenzhongxin/1198.html. Accessed 16 December 2024.
[40] HuatuoGPT. https://www.huatuogpt.cn/#/. Accessed 16 December 2024.
[41] Chinese Diabetes Society, National Office for Primary Diabetes Care. National guidelines for the prevention and control of diabetes in primary care (2018) (in Chinese). Chinese Journal of Internal Medicine 2018;57:885–93.
[42] MRCP (UK) Specialty Examinations. https://rcpsg.ac.uk/physicians/exams/mrcp-uk-specialty-examinations. Accessed 16 December 2024.
[43] The Federation. Specialty certificate examinations in endocrinology and diabetes. https://www.thefederation.uk/examinations/specialty-certificate-examinations/specialties/endocrinology-and-diabete. Accessed 16 December 2024.
[44] Sheng B, Pushpanathan K, Guan Z, et al. Artificial intelligence for diabetes care: current and future prospects. Lancet Diabetes Endocrinol 2024;12:569–95.
[45] Rosoł M, Gąsior JS, Łaba J, et al. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023;13:20512.
[46] Novak A, Rode F, Lisičić A, et al. The pulse of artificial intelligence in cardiology: a comprehensive evaluation of state-of-the-art large language models for potential use in clinical cardiology. medRxiv 2023: 2023.08.08.23293689.
[47] Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European exam in core cardiology: an artificial intelligence success story? Eur Heart J Digit Health 2023;4:279–81.
[48] Fijačko N, Gosak L, Štiglic G, et al. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation 2023;185:109732.
[49] MRCP UK. MRCP(UK) examination results and pass rates. https://www.mrcpuk.org/mrcpuk-examinations/results/exam-pass-rates. Accessed 16 December 2024.
[50] Li Y, Li Z, Zhang K, et al. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. Cureus 2023;15:e40895.
[51] Li J, Guan Z, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care. Nat Med 2024;30:2886–96.