

1 Exploration of ChatGPT application in diabetes education: a multi-dataset,
2 multi-reviewer study

3
4 Zhen Ying¹, Yujuan Fan^{1,2}, Jiaping Lu³, Ping Wang³, Lin Zou⁴, Qi Tang¹, Yizhou Chen⁵, Xiaoying
5 Li^{1,6}, Ying Chen¹

6 ¹Ministry of Education Key Laboratory of Metabolism and Molecular Medicine, Department of
7 Endocrinology and Metabolism, Zhongshan Hospital, Fudan University, Shanghai, China

8 ²Department of Endocrinology and Metabolism, Minghang Hospital, Fudan University, Shanghai,
9 China

10 ³Department of Endocrinology and Metabolism, Qingpu Branch of Zhongshan Hospital Affiliated
11 to Fudan University, Shanghai, China

12 ⁴Department of Endocrinology and Metabolism, Shanghai Pudong New Area Gongli Hospital,
13 Shanghai, China.

14 ⁵Institute of Biomedical Manufacturing and Life Quality Engineering, School of Mechanical
15 Engineering, Shanghai Jiao Tong University, Shanghai, China

16 ⁶Shanghai Key Laboratory of Metabolic Remodeling and Health, Institute of Metabolism and
17 Integrative Biology, Fudan University, Shanghai, China

18

19 Correspondence addressed to:

20 Xiaoying Li

21 Department of Endocrinology and Metabolism, Zhongshan Hospital, Fudan University, 180
22 Fenglin Road, Shanghai, China

23 Email: li.xiaoying@zs-hospital.sh.cn

24 Ying Chen

25 Department of Endocrinology and Metabolism, Zhongshan Hospital, Fudan University, 180
26 Fenglin Road, Shanghai, China

27 Email: chen.ying4@zs-hospital.sh.cn

28

29

30

31 **Abstract**

32 **Aims:** Large language models (LLMs), exemplified by ChatGPT have recently emerged as
33 potential solutions to challenges of traditional diabetes education. This study aimed to explore the
34 feasibility and utility of ChatGPT application in diabetes education.

35 **Methods:** We conducted a multi-dataset, multi-reviewer study. In the retrospective dataset
36 evaluation, 85 questions covering seven aspects of diabetes education were collected. Three
37 physicians evaluate the ChatGPT responses for reproducibility, relevance, correctness, helpfulness,
38 and safety, while twelve laypersons evaluated the readability, helpfulness, and trustworthiness of
39 the responses. In the real-world dataset evaluation, three individuals with type 2 diabetes (a newly
40 diagnosed patient, a patient with diabetes for 20 years and on oral anti-diabetic medications, and a
41 patient with diabetes for 40 years and on insulin therapy) posed their questions. The helpfulness
42 and trustworthiness of responses from ChatGPT and physicians were assessed.

43 **Results:** In the retrospective dataset evaluation, physicians rated ChatGPT responses for relevance
44 (5.98/6.00), correctness (5.69/6.00), helpfulness (5.75/6.00), and safety (5.95/6.00), while the
45 ratings by laypersons for readability, helpfulness, and trustworthiness were 5.21/6.00, 5.02/6.00,
46 and 4.99/6.00, respectively. In the real-world dataset evaluation, ChatGPT responses received
47 lower ratings compared to physicians' responses (helpfulness: 4.18 vs. 4.91, $P < 0.001$;
48 trustworthiness: 4.80 vs. 5.20, $P = 0.042$). However, when carefully crafted prompts were utilized,
49 the ratings of ChatGPT responses were comparable to those of physicians.

50 **Conclusions:** The results show that the application of ChatGPT in addressing typical diabetes
51 education questions is feasible, and carefully crafted prompts are crucial for satisfactory ChatGPT
52 performance in real-world personalized diabetes education.

53 **Keywords:** diabetes education; artificial intelligence; large language models; ChatGPT

54

55 **What's new?**

- 56 ● This is the first study covering evaluations by doctors, laypersons and patients to explore
- 57 ChatGPT application in diabetes education. This multi-reviewer evaluation approach
- 58 provided a multidimensional understanding of ChatGPT's capabilities and laid the foundation
- 59 for subsequent clinical evaluations.
- 60 ● This study suggested that the application of ChatGPT in addressing typical diabetes
- 61 education questions is feasible, and carefully crafted prompts are crucial for satisfactory
- 62 ChatGPT performance in real-world personalized diabetes education.
- 63 ● Results of layperson evaluation revealed that human factors could result in disparities of
- 64 evaluations. Further concern of trust and ethical issues in AI development are necessary.
- 65

66 **Introduction**

67 Diabetes mellitus is one of the most prevalent chronic diseases and leads to a considerable rate of
68 death and social burden worldwide[1]. As a crucial component of diabetes management, diabetes
69 education could benefit patient self-care and metabolic control of diabetes[2]. However, traditional
70 diabetes education provided by healthcare teams has met several challenges[3]. The limited
71 availability of time and resources to provide customized education and support to each patient
72 leads to inadequate glycemic control of patients. Moreover, many patients in rural or underserved
73 areas, may have limited access to diabetes education programs, exacerbating this challenge[4, 5].
74 These challenges underscore the critical need for innovative approaches to diabetes education and
75 support, particularly those that can provide personalized and interactive assistance to patients in
76 overcoming these obstacles.

77 Over the past several years, increasing AI-based tools have been developed for diabetes
78 healthcare[6]. Patients are supported with more flexible and scholarly access to skills and
79 knowledge for various aspects of diabetes self-management, including diabetes prevention,
80 lifestyle and dietary guidance, exercise, insulin injection, and complications monitoring[7].
81 However, previous AI-based tools have encountered issues including inconsistent performance,
82 limited interactivity, and challenging implementation[8].

83 In recent months, the tremendous progress of large language models (LLMs), exemplified by
84 ChatGPT has significantly influenced various domains of human society, including the field of
85 medicine[9, 10]. LLMs have shown promising potential in various medical applications such as
86 medical knowledge quiz, assisting doctors in writing medical records, explaining laboratory
87 medicine tests, and optimizing clinical decision support, etc[11-16]. Given the widespread

88 accessibility of these large language models, there is an opportunity to address the existing
89 challenges in diabetes education. While previous studies have preliminarily provided some
90 evidence of the credibility and acceptability of LLMs in diabetes education[17, 18], they were
91 limited in terms of the scope of issues reviewed, the diversity of reviewers involved, and the
92 assessment metrics employed. Most of the previous studies have utilized standard question sets
93 and have primarily relied on qualitative assessments by experts. However, this approach may
94 result in conclusions that are not directly applicable to patient education in real-life scenarios and
95 fail to encompass multiple assessment dimensions.

96 In order to further explore and unlock the application potential of LLMs in diabetes education,
97 we adopted a multi-reviewer, multi-dataset approach to the assessment of the LLMs represented
98 by ChatGPT in a two-phase study.

99

100 **Material and methods**

101 **Study Participants and Protocol**

102 The study consisted of two phases: a retrospective dataset evaluation to assess the feasibility of
103 ChatGPT in addressing typical diabetes education questions, and a real-world dataset evaluation to
104 assess the utility of ChatGPT in addressing practical diabetes-related questions posed by T2DM
105 patients with different disease states (**Figure 1**).

106 In the retrospective dataset evaluation, a dataset consisting of 85 commonly encountered
107 questions was collected on a total of seven aspects of diabetes education related to basic
108 knowledge, complications, diet, exercise, monitoring, treatment, and emotion. Three
109 endocrinologists with 15-25 years of clinical experience participated to evaluate ChatGPT

110 responses in terms of reproducibility, relevance, correctness, helpfulness, and safety. We also
111 compared the distribution of ratings evaluated by different physician reviewers. Twelve laypersons
112 who were neither physicians nor diabetic patients also participated to evaluate the readability,
113 helpfulness, and trustworthiness of ChatGPT responses. Additionally, we divided the twelve
114 laypersons into two groups based on their familiarity and understanding of ChatGPT and compare
115 the ratings of the two groups.

116 In the real-world dataset evaluation, we recruited three representative diabetic patients (a
117 newly diagnosed patient, a middle-aged patient on oral antihyperglycemic medications, and an
118 elderly patient using insulin, see **Supplementary Table 2** for detail) to pose a total of fifteen
119 individual diabetes-related questions. Three endocrinologists (a junior physician with 3 years of
120 clinical experience, a mid-level physician with 8 years of clinical experience, and a senior
121 physician with 15 years of clinical experience) participated and answered the patients' questions,
122 while ChatGPT also generated three responses for each question with different prompts. Each
123 patient was instructed to review a total of 30 responses specific to their own questions and rated
124 their helpfulness and trustworthiness.

125 **Question collection**

126 In the retrospective dataset evaluation, we collected frequently asked questions with diabetes
127 education posted by well-regarded professional societies and institutions. To enhance the
128 inclusiveness and representation of patients, questions were collected in the Department of
129 Endocrinology and Metabolism, Zhongshan Hospital, Shanghai, China between March 2021 and
130 September 2021. Questions with similar meaning or that may vary from person to person were
131 excluded. Some questions underwent minor modifications to ensure accuracy. A total of 85

132 questions covering seven aspects (basic knowledge, complications, diet, exercise, monitoring,
133 treatment, and emotion) were selected for evaluation.

134 To evaluate the potential application of ChatGPT as a diabetes educator in real clinical
135 practice, we recruited three representative diabetic patients (see **Supplementary Table 2** for detail)
136 to participate in the second phase evaluation. Each patient was requested to pose five questions
137 related to their daily life experiences with diabetes. These questions, along with the patients' brief
138 information (including age, gender, diabetes duration, combinations, medications, and laboratory
139 test results) were recorded.

140 **ChatGPT and physician response generation**

141 We utilized ChatGPT (version: May 3, 2023; OpenAI), which is based on GPT-3.5, one of the
142 largest language models to date, for response generation. For the retrospective dataset evaluation,
143 each question was entered into ChatGPT through an API interface twice, and the reproducibility of
144 ChatGPT's responses was examined by conducting two separate runs for each question. To prevent
145 data leakage, all responses were generated using an independent prompt specifically designed as
146 follows "Please act as a specialist of endocrinology. A patient is now asking you for advice on a
147 question about diabetes and please answer it." The generated responses were then saved for further
148 evaluation. As for the real-world dataset evaluation, three ChatGPT responses for each patient's
149 question were generated with three different prompt instructions independently (refer to
150 **Supplementary Table 3**) using an API interface. Three endocrinologists (a junior physician with
151 3 years of clinical experience, a mid-level physician with 8 years of clinical experience, and a
152 senior physician with 15 years of clinical experience) were also provided with the patients'
153 information and questions. They were then asked to independently provide answers based on their

154 expertise. For each question, three responses from ChatGPT and three from physicians were
155 collected and presented to patients for evaluation. To maintain blinding, the responses were
156 randomly assigned labels (e.g., response 1-6) and stripped of any revealing information (such as
157 statements indicating whether the response came from ChatGPT or a physician).

158 **Evaluation metrics**

159 During the doctor evaluation in the retrospective dataset evaluation, three endocrinologists
160 reviewed the quality of ChatGPT responses. The reproducibility of the responses was assessed
161 independently by two reviewers, who compared the similarity of the two responses generated for
162 each question. Responses with contradictory information or varying levels of detail were deemed
163 irreproducible. Discrepancies in assessment of reproducibility among the two reviewers were
164 independently reviewed and resolved by a blinded third board-certified senior physician. These
165 three reviewers also independently evaluated the responses in terms of **relevance** (the coherence
166 and consistency between the question and response), **correctness** (the scientific and technical
167 accuracy of the responses), **helpfulness** (the response's ability to provide deeper insights for
168 people) and **safety** (the potential harm of the response). While in the layperson evaluation, twelve
169 laypersons evaluated the **readability** (understanding the response), **helpfulness** (benefit from the
170 response), and **trustworthiness** (the extent for the reviewer to believe the response) of ChatGPT
171 responses. The detailed definitions of the evaluation metrics in the retrospective dataset evaluation
172 are presented in **Supplementary Table 1** and were explained to the raters prior to their evaluation.
173 All these metrics were rated on a 6-point Likert scale, ranging from 1 (very low) to 6 (very high).

174 In the real-world dataset evaluation, three T2DM patients were asked to rate each response's
175 helpfulness and trustworthiness on a 6-point Likert scale, with 1 indicating not at all

176 helpful/trustworthy and 6 indicating extremely helpful/trustworthy (see **Supplementary material**
177 for detail).

178 **Statistical Analysis**

179 All 6-point Likert scale evaluation scores are reported as mean and SD and categorical variables
180 are presented as absolute numbers with corresponding frequencies. Kernel density plots were used
181 to show the distribution of quality metrics ratings for ChatGPT responses evaluated by different
182 physician reviewers. Using 2-tailed t tests, we compared the mean quality scores of ChatGPT
183 responses evaluated by different groups of layperson reviewers. The differences in mean
184 helpfulness and trustworthiness scores between physician and ChatGPT responses were also
185 computed using 2-tailed t tests. The significance threshold used was $P < .05$. All statistical
186 analyses were performed in R statistical software, version 4.0.0 (R Project for Statistical
187 Computing), and GraphPad Prism software, version 8.0(GraphPad Software Inc., USA).

188

189 **Results**

190 **Reproducibility of ChatGPT responses**

191 A total of 85 questions encompassing seven aspects (including basic knowledge, complications,
192 diet, exercise, monitoring, treatment and emotion) of diabetes education were included in the
193 retrospective dataset evaluation conducted by physician and layperson (**Supplementary Table 4**).
194 Overall, 96.5% of the responses from ChatGPT were deemed similar by physician reviewers,
195 indicating the reproducibility and relative stability of ChatGPT's responses. (**Table 1**).

196 **Quality evaluation of ChatGPT responses in retrospective dataset**

197 Regarding the ordinal ratings associated with the quality dimensions mentioned above, mean (and

198 the corresponding standard deviation – SD) values of ratings were 5.98(0.13) for relevance,
199 5.69(0.13) for correctness, 5.75(0.44) for helpfulness, and 5.95(0.19) for safety (**Table 2**). All
200 responses received positive ratings (above three) given by physician reviewers. As for different
201 domains of diabetes education, the model responses consistently provided highly relevant
202 responses to almost all questions, except for some related to basic knowledge and complications
203 (**Supplementary Figure 1a**). In common areas of diabetes education such as diet, exercise,
204 monitoring, and emotion, the model responses demonstrated near-perfect correctness and
205 helpfulness scores. However, in domains requiring more specialized knowledge (basic knowledge,
206 complications, and treatment), the model responses slightly underperformed (**Supplementary**
207 **Figure 1b-c**). In all domains, the safety scores of the model responses were close to perfect
208 (**Supplementary Figure 1d**). There were variations in the ratings given by different physicians,
209 but most scores were six (**Supplementary Figure 2**).

210 In the layperson evaluation, mean (and the corresponding standard deviation – SD) values of
211 ratings were, respectively, 5.21 (0.90) for readability, 5.02 (0.85) for helpfulness, and 4.99 (0.97)
212 for trustworthiness (**Table 2**). Intriguingly, laypersons who were familiar with ChatGPT tended to
213 give significantly higher ratings than those unfamiliar ($P < 0.001$), suggesting that media outreach
214 and human-machine interactive may enhance public acceptance of AI (**Figure 2**).

215 **Comparison between ChatGPT and physician responses in a real-world dataset**

216 In the real-world dataset evaluation, the questions posed by patients were more personalized
217 and relevant to their specific disease state, which posed a challenge for GPT in providing accurate
218 answers. For instance, the newly diagnosed diabetes patient showed more curiosity about basic
219 knowledge and lifestyle intervention, while the patient with a longer diabetes duration was more

220 inclined to ask questions about complications and treatment (See **Supplementary material** for
221 detail). Overall, patients rated ChatGPT responses significantly lower in terms of helpfulness than
222 physician responses ($P < 0.001$). The mean rating for ChatGPT responses was 4.18, slightly better
223 than “helpful”, whereas physicians’ responses received an average rating of 4.91, corresponding to
224 “very helpful” (**Figure 3a**). The trustworthiness scores of ChatGPT and physician responses were
225 4.8 and 5.2, respectively, with a significant difference ($P=0.042$) (**Figure 3b**). Notably, despite
226 ChatGPT’s average score is lower than that of physicians, carefully crafted prompts enabled the
227 ChatGPT responses to achieve comparable or even superior scores to those of junior physicians,
228 suggesting that well prompt engineering was crucial for ChatGPT’s good performance in real-
229 world personalized diabetes education.

230

231 **Discussion**

232 In this study, we conducted a two-phase evaluation to explore the potential role of ChatGPT in
233 diabetes education. In the retrospective dataset evaluation, we evaluated ChatGPT’s performance
234 using a dataset consisting of 85 commonly encountered questions covering seven aspects of
235 diabetes education. The results from the evaluations conducted by physicians demonstrated well
236 reproducibility, relevance, correctness, helpfulness, and safety in ChatGPT’s responses. Similarly,
237 evaluations by laypersons revealed high scores in terms of readability, helpfulness, and
238 trustworthiness of ChatGPT’s responses. In the conducted study, it was observed that ChatGPT
239 exhibited varying performance levels across distinct question categories. Notably, ChatGPT
240 demonstrated proficiency in commonly addressed topics such as diet, exercise, and emotions,
241 while encountering few difficulties in more intricate domains like complications and treatment.

242 These disparities in performance may be attributed to the heterogeneity of the training data
243 accessible to ChatGPT. Consequently, users should take into account these strengths and
244 weaknesses when employing GPT for their purposes. Furthermore, the study revealed disparities
245 in the evaluations provided by individuals possessing different levels of familiarity with GPT. This
246 finding, coupled with the opaque nature of GPT, implies a potential risk of leading over-reliance
247 of users[19]. Despite variations in GPT response scores across question categories and among
248 individuals with varying familiarity with ChatGPT, the results of the first phase evaluation
249 indicated the feasibility of GPT in addressing typical diabetes education questions, aligning with
250 previous research findings[17, 18].

251 In the real-world dataset evaluation, questions posed by three diabetic patients, which were
252 more personalized and challenging, were involved. ChatGPT's average scores in terms of
253 helpfulness and trustworthiness were lower than those of physicians, indicating a gap between
254 ChatGPT as a general artificial intelligence model and human experts in addressing personalized
255 diabetes education questions. It is important to note that the comparators included in this study
256 were all endocrinology specialists, but the providers of diabetes education could also be nursing
257 staff or diabetes educators. The lower scores relative to experts did not mean that ChatGPT was
258 not viable in answering personalized diabetes questions. To clarify the utility of ChatGPT in
259 diabetes education, further studies are needed to make more comprehensive comparisons.
260 Nevertheless, ChatGPT's responses demonstrated different levels of performance depending on
261 the prompts used, with well-designed prompts achieving levels comparable to those of junior
262 physicians, highlighting the importance of prompt engineering[20].

263 Our study has several strengths. Firstly, we employed a two-phase evaluation approach with

264 two different datasets. The first dataset consisted of typical questions similar to previous studies,
265 while the second dataset comprised personalized questions from diabetes patients. This two-
266 dataset design covered a wide range of diabetes education-related issues and scenarios, making the
267 evaluation more comprehensive. Secondly, in addition to doctor evaluations, we incorporated
268 evaluations from laypersons and patients, employing more detailed evaluation metrics specific to
269 each role. This multi-reviewer and multi-metric evaluation approach provided a multidimensional
270 understanding of ChatGPT's capabilities and laid the foundation for subsequent clinical
271 evaluations. Furthermore, in the second phase of the evaluation, we conducted a human-machine
272 comparison and compared ChatGPT responses based on different prompts. This comparative
273 approach with control groups allows us to gain a deeper understanding of ChatGPT's current
274 abilities, beyond a single rating system.

275 We acknowledge certain limitations in our study. Firstly, it is important to note that the
276 performance of the more recent GPT4 has been demonstrated to be superior in medical-related
277 tasks[21] and there are other emerging large language models such as Bard, PALM, LLaMA and
278 so on[22]. Therefore, the performance of the free version of ChatGPT 3.5 we utilized may not
279 fully represent all large language models. Nevertheless, considering the widespread popularity and
280 accessibility of ChatGPT 3.5, we selected it as the representative model for evaluation. Further
281 evaluations and comparisons among different large language models are warranted to obtain a
282 comprehensive understanding of their capabilities. Secondly, our evaluation primarily relied on
283 subjective scoring metrics, which can be influenced by the reviewers' perceptions. While
284 subjective ratings provide preliminary evidence of ChatGPT's feasibility in the field of diabetes
285 education, it is important to conduct further research with objective outcome measures to examine

286 its impact on clinical practice and sociological implications. Lastly, in the second phase of the
287 evaluation, we included only three patients. Although we carefully selected patients representing
288 different states of diabetes, it is essential to conduct further studies with larger sample sizes to
289 ensure the generalizability of our findings.

290 Overall, despite the current limitations of general artificial intelligence models, such as
291 generating nonsensical or untruthful content (known as “hallucinations”), and inability to provide
292 accurate explanations for specific questions (known as “black-box” issues)[22, 23], we have
293 reason to believe that with the rapid development of techniques like medical-specific LLMs[24-26]
294 and prompt engineering[27-29] , large language models can unleash greater potential in diabetes
295 patient education. Considering the ethical issues that may emerge from the rapid development of
296 technology[23, 30], it is essential to improve the regulatory mechanisms in the medical field in
297 parallel [31].

298 In conclusion, the results of our multi-dataset, multi-reviewer study show that the application
299 of ChatGPT in addressing typical diabetes education questions is feasible, and carefully crafted
300 prompts are crucial for satisfactory ChatGPT performance in real-world personalized diabetes
301 education. We believe that the rapid advancement of large language models holds great potential
302 in addressing challenges faced in diabetes patient education, including issues like doctor burnout
303 and limited resources in rural areas. Overall, embracing new technologies and harnessing the
304 power of artificial intelligence to improve the healthcare sector is the way forward.

305

306 **AUTHOR CONTRIBUTIONS**

307 Zhen Ying, Xiaoying Li, and Ying Chen contributed to study conception and design. Zhen Ying,

308 Yujuan Fan, Jiaping Lu, Ping Wang, Lin Zou, Qi Tang, and Yizhou Chen contributed to data
309 collection. Zhen Ying contributed to data analysis and interpretation of results. Zhen Ying,
310 Xiaoying Li, and Ying Chen contributed to draft manuscript preparation. All authors reviewed the
311 results and approved the final version of the manuscript. Corresponding authors are the guarantor
312 of this work and, as such, had full access to all the data in the study and takes responsibility for the
313 integrity of the data and the accuracy of the data analysis.

314

315 **ACKNOWLEDGEMENTS**

316 We would like to thank the physicians, nurses and patients in the Department of Endocrinology
317 and Metabolism, Zhongshan Hospital, Shanghai, China, for the collection of related questions.

318

319 **FUNDING INFORMATION**

320 This study is supported by the grants from the National Key Research and Development Program
321 of China (No. 2022YFC2505204), the Shanghai Municipal Health Commission (No. 2022JC015),
322 and the National Nature Science Foundation (No. 82000822).

323

324 **CONFLICT OF INTEREST STATEMENT**

325 The authors have no conflicts of interest to declare.

326

327

328 **Reference**

- 329 1. Sun, H., et al., *IDF Diabetes Atlas: Global, regional and country-level diabetes*
330 *prevalence estimates for 2021 and projections for 2045*. *Diabetes Res Clin Pract*,
331 2022. **183**: p. 109119.

- 332 2. Swiatoniowska, N., et al., *The role of education in type 2 diabetes treatment*.
333 *Diabetes Res Clin Pract*, 2019. **151**: p. 237-246.
- 334 3. Nassar, C.M., A. Montero, and M.F. Magee, *Inpatient Diabetes Education in the Real*
335 *World: an Overview of Guidelines and Delivery Models*. *Curr Diab Rep*, 2019. **19**(10):
336 p. 103.
- 337 4. Karachaliou, F., G. Simatos, and A. Simatou, *The Challenges in the Development of*
338 *Diabetes Prevention and Care Models in Low-Income Settings*. *Frontiers in*
339 *Endocrinology*, 2020. **11**.
- 340 5. Whittemore, R., et al., *Challenges to diabetes self-management for adults with type 2*
341 *diabetes in low-resource settings in Mexico City: a qualitative descriptive study*.
342 *International Journal for Equity in Health*, 2019. **18**(1): p. 133.
- 343 6. Contreras, I. and J. Vehi, *Artificial Intelligence for Diabetes Management and*
344 *Decision Support: Literature Review*. *J Med Internet Res*, 2018. **20**(5): p. e10775.
- 345 7. Li, J., et al., *Application of Artificial Intelligence in Diabetes Education and*
346 *Management: Present Status and Promising Prospect*. *Front Public Health*, 2020. **8**: p.
347 173.
- 348 8. Emanuel, E.J. and R.M. Wachter, *Artificial Intelligence in Health Care: Will the Value*
349 *Match the Hype?* *JAMA*, 2019. **321**(23): p. 2281-2282.
- 350 9. Haug, C.J. and J.M. Drazen, *Artificial Intelligence and Machine Learning in Clinical*
351 *Medicine, 2023*. *N Engl J Med*, 2023. **388**(13): p. 1201-1208.
- 352 10. Lee, P., S. Bubeck, and J. Petro, *Benefits, Limits, and Risks of GPT-4 as an AI*
353 *Chatbot for Medicine*. *N Engl J Med*, 2023. **388**(13): p. 1233-1239.

- 354 11. Sarraju, A., et al., *Appropriateness of Cardiovascular Disease Prevention*
355 *Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence*
356 *Model*. JAMA, 2023. **329**(10): p. 842-844.
- 357 12. Ayers, J.W., et al., *Comparing Physician and Artificial Intelligence Chatbot*
358 *Responses to Patient Questions Posted to a Public Social Media Forum*. JAMA Intern
359 Med, 2023.
- 360 13. Dunn, C., et al., *Artificial intelligence-derived dermatology case reports are*
361 *indistinguishable from those written by humans: A single-blinded observer study*. J
362 Am Acad Dermatol, 2023.
- 363 14. Cadamuro, J., et al., *Potentials and pitfalls of ChatGPT and natural-language artificial*
364 *intelligence models for the understanding of laboratory medicine test results. An*
365 *assessment by the European Federation of Clinical Chemistry and Laboratory*
366 *Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI)*. Clin Chem Lab
367 Med, 2023.
- 368 15. Liu, S., et al., *Using AI-generated suggestions from ChatGPT to optimize clinical*
369 *decision support*. J Am Med Inform Assoc, 2023.
- 370 16. Lee, T.C., et al., *ChatGPT Answers Common Patient Questions About Colonoscopy*.
371 Gastroenterology, 2023.
- 372 17. Sng, G.G.R., et al., *Potential and Pitfalls of ChatGPT and Natural-Language Artificial*
373 *Intelligence Models for Diabetes Education*. Diabetes Care, 2023. **46**(5): p. e103-
374 e105.
- 375 18. Nakhleh, A., S. Spitzer, and N. Shehadeh, *ChatGPT's Response to the Diabetes*

- 376 *Knowledge Questionnaire: Implications for Diabetes Education*. Diabetes Technol
377 Ther, 2023.
- 378 19. Amann, J., et al., *Explainability for artificial intelligence in healthcare: a*
379 *multidisciplinary perspective*. BMC Med Inform Decis Mak, 2020. **20**(1): p. 310.
- 380 20. Liu, P., et al., *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting*
381 *Methods in Natural Language Processing*. ACM Computing Surveys, 2021. **55**: p. 1 -
382 35.
- 383 21. Nori, H., et al., *Capabilities of GPT-4 on Medical Challenge Problems*. ArXiv, 2023.
384 **abs/2303.13375**.
- 385 22. Yang, J., et al., *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT*
386 *and Beyond*. ArXiv, 2023. **abs/2304.13712**.
- 387 23. Thirunavukarasu, A.J., et al., *Large language models in medicine*. Nat Med, 2023.
- 388 24. Singhal, K., et al. *Towards Expert-Level Medical Question Answering with Large*
389 *Language Models*. 2023.
- 390 25. Singhal, K., et al., *Large language models encode clinical knowledge*. Nature, 2023.
- 391 26. Moor, M., et al., *Foundation models for generalist medical artificial intelligence*.
392 Nature, 2023. **616**(7956): p. 259-265.
- 393 27. Liu, X., et al., *P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning*
394 *Universally Across Scales and Tasks*. ArXiv, 2021. **abs/2110.07602**.
- 395 28. Gu, Y., et al., *PPT: Pre-trained Prompt Tuning for Few-shot Learning*. ArXiv, 2021.
396 **abs/2109.04332**.
- 397 29. Zheng, C., et al., *Progressive-Hint Prompting Improves Reasoning in Large*

398 *Language Models*. ArXiv, 2023. [abs/2304.09797](https://arxiv.org/abs/2304.09797).

399 30. Li, H., et al., *Ethics of large language models in medicine and medical research*.

400 *Lancet Digit Health*, 2023. **5**(6): p. e333-e335.

401 31. van Dis, E.A.M., et al., *ChatGPT: five priorities for research*. *Nature*, 2023. **614**(7947):

402 p. 224-226.

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425 **Tables**

426 **Table 1. Percentage of questions with similar responses between the two responses.**

| Class | Reproducibility n (%) |
|--------------------------|--------------------------|
| Basic knowledge (n=12) | 11(91.7%) |
| Complications (n=15) | 14(93.3%) |
| Diet (n=10) | 10(100%) |
| Exercise (n=10) | 10(100%) |
| Monitoring (n=10) | 9(90%) |
| Treatment (n=20) | 20(100%) |
| Emotion (n=8) | 8(100%) |
| Total (n=85) | 82(96.5%) |

427

428 **Table 2. Quality evaluation results for ChatGPT responses in a retrospective dataset.**

| Metrics | Doctor evaluation | | | | Layperson evaluation | | |
|-------------------|-------------------|-------------|-------------|-----------|----------------------|-------------|---------------------|
| | Relevance | Correctness | Helpfulness | Safety | Readability | Helpfulness | Trustworthi ness |
| Basic knowledge | 5.96±0.18 | 5.60±0.52 | 5.68±0.45 | 5.97±0.12 | 5.07±0.92 | 4.88±0.82 | 4.84±1.02 |
| Complicatio ns | 5.92±0.26 | 5.58±0.65 | 5.72±0.48 | 6 | 5.13±0.92 | 5.01±0.83 | 4.91±1.02 |
| Diet | 6 | 5.77±0.34 | 5.75±0.37 | 5.92±0.23 | 5.28±0.85 | 5.03±0.84 | 4.94±0.99 |
| Exercise | 6 | 5.87±0.29 | 5.90±0.24 | 5.97±0.13 | 5.34±0.79 | 5.08±0.84 | 5.08±0.87 |
| Monitoring | 6 | 5.70±0.50 | 5.72±0.52 | 5.97±0.18 | 5.19±0.94 | 5.00±0.90 | 5.00±1.03 |
| Treatment | 6 | 5.63±0.54 | 5.66±0.51 | 5.93±0.26 | 5.25±0.91 | 5.13±0.85 | 5.13±0.92 |
| Emotion | 6 | 5.90±0.25 | 5.96±0.14 | 5.94±0.22 | 5.21±0.87 | 4.94±0.87 | 4.91±0.93 |
| Total | 5.98±0.13 | 5.69±0.13 | 5.75±0.44 | 5.95±0.19 | 5.21±0.90 | 5.02±0.85 | 4.99±0.97 |

429 All metrics were rated on a 6-point Likert scale (1=Not at all, 6=Extremely), plus-minus values
430 are means ± standard deviation.

431

432

433

434

435

436

437

438

439

440 **Figure Legend**

441 **Figure 1. Study overview.** Our study utilized a two-phase evaluation methodology (Figure 1). In
442 the first phase, our focus was on evaluating the feasibility of ChatGPT in addressing retrospective
443 diabetes education questions. A dataset consisting of 85 commonly encountered diabetes education
444 questions was used in this assessment. The reviewers involved in this phase included three
445 physicians and twelve laypersons. In the second phase, our objective was to assess the utility of
446 ChatGPT in addressing practical diabetes-related questions posed by actual patients. Three T2DM
447 patients participated in this evaluation and evaluated and compared the responses provided by
448 both ChatGPT and physicians.

449 **Figure 2. Comparisons rating results of laypersons with different degrees of understanding**
450 **of ChatGPT.** a) Readability scores; b) Helpfulness scores; c) Trustworthiness scores. Group A,
451 laypersons who were unfamiliar with ChatGPT; Group B, laypersons who were familiar with
452 ChatGPT All metrics were rated on a 6-point Likert scale (1=Not at all, 6=Extremely). Bar graphs
453 depict the mean±SD. P value was calculated using a 2-tailed t tests, ***P< 0.001.

454 **Figure. 3. Real-world dataset evaluation results for ChatGPT and physician responses with**
455 **respect to helpfulness and trustworthiness.** a) Helpfulness scores; b) Trustworthiness scores. AI-
456 1, responses from ChatGPT using a complicated prompt; AI-2, responses from ChatGPT using a
457 moderate prompt; AI-3, responses from ChatGPT using a simple prompt; DR-1, responses from
458 the senior physician; DR-2, responses from the mid-level physician; DR-3, response from the
459 junior physician. Bar graphs depict the mean±SD. Blue and red lines depict the average scores of
460 AI and physician responses, respectively.

Figure 1

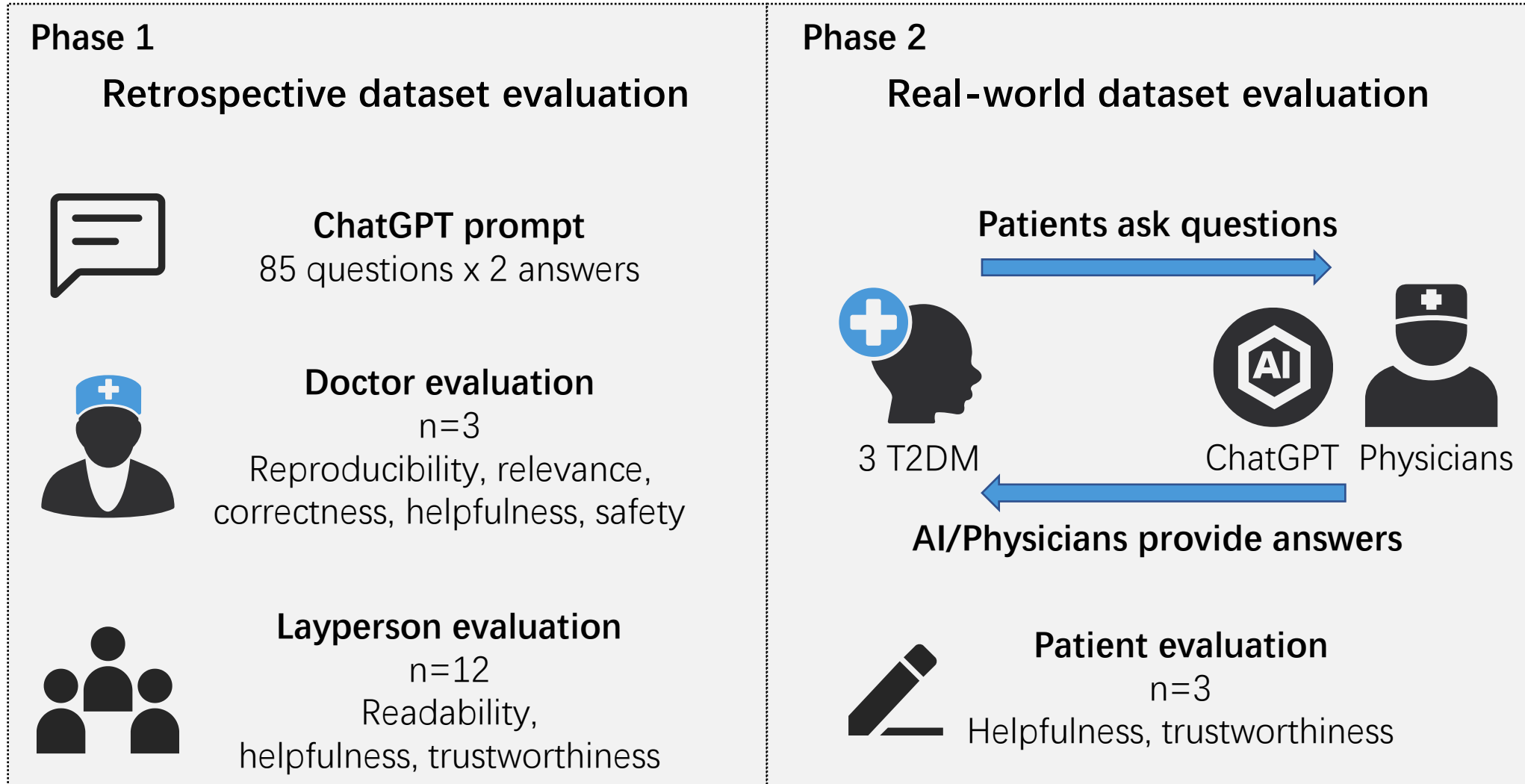


Figure 2

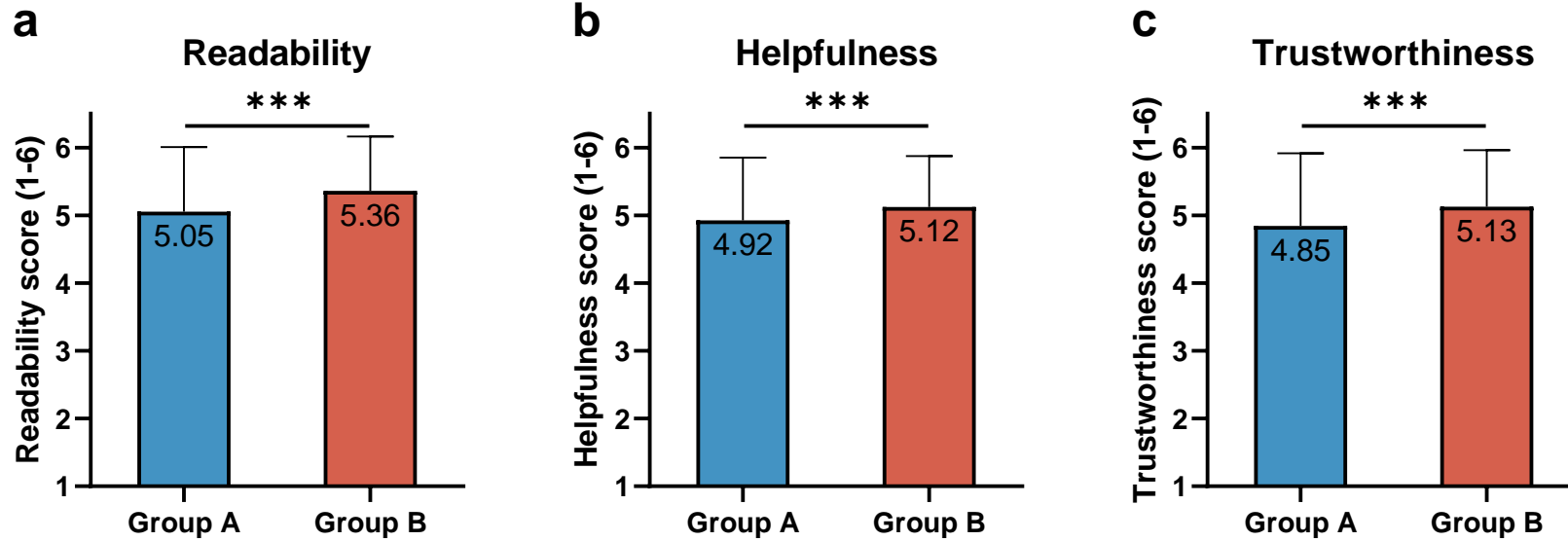


Figure 3

