

1 **Performance of ChatGPT on Clinical Medicine Entrance Examination for**

2 **Chinese Postgraduate in Chinese**

3 Xiao Liu^{*1,2}, M.D., Changchang Fang^{*3}, M.D., Ziwei Yan, M.P.T., Xiaoling Liu¹,

4 M.D., Yuan Jiang^{1,2}, M.D., Zhengyu Cao^{1,2}, M.D., Maoxiong Wu^{1,2}, M.D., Zhiteng

5 Chen^{1,2}, M.D., Jianyong Ma⁴, M.D., Peng Yu³, Wengen Zhu⁴, M.D., Ayiguli

6 Abudukeremu^{1,2}, M.D., Yue Wang^{1,2}, M.D. Yangxin Chen^{1,2}, M.D., Yuling Zhang^{1,2},

7 M.D.,Jingfeng Wang, M.D.^{1,2}

8 ¹Department of Cardiology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University,

9 Guangzhou, China

10 ²Guangdong Province Key Laboratory of Arrhythmia and Electrophysiology,

11 Guangzhou, China

12 ³ Department of Endocrine, the Second Affiliated Hospital of Nanchang University,

13 Jiangxi, China

14 ⁴ Department of Pharmacology and Systems Physiology, University of Cincinnati

15 College of Medicine, Cincinnati, United States

16 ⁵ Department of Cardiology, the first Affiliated Hospital of Sun Yat-sen University,

17 Guangzhou, China

18 *Co-first author

19 Word counts: 2359 (excluding abstract and references)

20 **Corresponding author:**

21 ✉ Yuling Zhang, Email: zzhangyuling@126.com

22 ✉ Jingfeng Wang, Email: wjingf@mail.sysu.edu.cn

24 **Abstract**

25 **Background:** The ChatGPT, a Large-scale language models-based Artificial
26 intelligence (AI), has fueled interest in medical care. However, the ability of AI to
27 understand and generate text is constrained by the quality and quantity of training data
28 available for that language. This study aims to provide qualitative feedback on
29 ChatGPT's problem-solving capabilities in medical education and clinical decision-
30 making in Chinese.

31 **Methods:** A dataset of Clinical Medicine Entrance Examination for Chinese
32 Postgraduate was used to assess the effectiveness of ChatGPT3.5 in medical
33 knowledge in Chinese language. The indicator of accuracy, concordance (explaining
34 affirms the answer) and frequency of insights was used to assess performance of
35 ChatGPT in original and encoding medical questions.

36 **Result:** According to our evaluation, ChatGPT received a score of 153.5/300 for
37 original questions in Chinese, which is slightly above the passing threshold of
38 129/300. Additionally, ChatGPT showed low accuracy in answering open-ended
39 medical questions, with total accuracy of 31.5%. While ChatGPT demonstrated a
40 commendable level of concordance (achieving 90% concordance across all questions)
41 and generated innovative insights for most problems (at least one significant insight
42 for 80% of all questions).

43 **Conclusion:** ChatGPT's performance was suboptimal for medical education and
44 clinical decision-making in Chinese compared with in English. However, ChatGPT

45 demonstrated high internal concordance and generated multiple insights in Chinese

46 language. Further research should investigate language-based differences in

47 ChatGPT's healthcare performance.

48 **Introduction**

49 Artificial intelligence (AI) is initially conceptualized in 1956¹, but it has only
50 gained significant momentum in recent years. AI aims to replicate human intelligence
51 and thinking processes through the use of brain-like computer systems to solve
52 complex problems. The most inspiring is that, AI systems can be trained on specific
53 data sets to improve prediction accuracy and address intricate problems²⁻⁴, which
54 means that one of the possible application of AI is the ability to helps doctors to
55 rapidly search through vast amounts of medical data, enhancing their creativity and
56 enabling them to make error-free decisions^{5,6}.

57 ChatGPT is an artificial intelligence model that has spurred great attention due to
58 the revolutionary innovations in its ability to perform a diverse array of natural
59 language tasks. By using a class of Large-scale language models (LLMs), GPT3.5 can
60 predict the likelihood of a sequence of words based on the context of the preceding
61 words. With sufficient training on vast amounts of text data, ChatGPT can generate
62 novel word sequences that closely resemble natural human language, but have never
63 been observed before by other AI⁷.

64 A study was conducted on the effectiveness of the version of GPT LLM (GPT3.5)
65 in passing the United States Medical Licensing Exam (USMLE). The results showed
66 that the AI model achieved an accuracy rate of over 50% in all the tests, and in some
67 analyses, it even surpassed 60% accuracy. It is imperative to highlight and emphasize
68 that the study was conducted mostly using English input and the AI model was also
69 trained in English.

70 However, like all language models, ChatGPT's ability to understand and generate
71 text in any given language is limited by the quality and quantity of training data
72 available in that language. Chinese is the second most widely spoken language in the
73 world, with more than 1.3 billion speakers globally, while the quality and quantity of
74 Chinese language data may be not compared with English due to some reasons, such
75 as complexity of the written characters. Thus, the performance of ChatGPT in
76 Chinese medical information warrants further investigation.

77 In this study, we evaluate ChatGPT's clinical reasoning ability by administering
78 questions from the Clinical Medicine Entrance Examination for Chinese Postgraduate
79 in Chinese. This standardized and regulated test assesses candidates' comprehensive
80 abilities, the questions textually and conceptually dense, and the difficulty and
81 complexity of questions is highly standardized and regulated. Additionally, this exam
82 has demonstrated remarkable stability in raw scores and psychometric properties over
83 the past years. Moreover, the exam comprises 43% basic science and medical
84 humanities, with 14% physiology, 10% biochemistry, 13% pathology, and 6%
85 medical humanities. Clinical medicine makes up the remaining 57%, with internal
86 medicine and surgery accounting for 37% and 20%, respectively. Due to the exam's
87 linguistic and conceptual complexity, we hypothesize that it will serve as an excellent
88 challenge for ChatGPT.

89 **Methods**

90 **Artificial Intelligence**

91 ChatGPT is a state-of-the-art language model that employs self-attention

92 mechanisms and vast training data to produce natural language responses in a
93 conversational context. Its key strengths include the ability to handle long-range
94 dependencies and generate coherent, contextually relevant replies. However, it is
95 worth noting that GPT3.5 is a server-based language model that lacks internet
96 browsing and search capabilities. As a result, all responses generated are based solely
97 on the abstract relationships between words, or "tokens," within its neural network⁷. It
98 should be noticed is that the OpenAI has developed a latest version GPT4 in March
99 2023, while the inputting date is Feb 2023 which the latest version is GPT3.5 at that
100 time.

101 **Input source**

102 The test questions for the Chinese Clinical Medicine Postgraduate Entrance
103 Examination in 2022 is not released by the official website. However, a complete set
104 of 165 questions with a total of 500 point was available online (Supplemental S1),
105 which is deemed as original questions. The point values for questions varied. Case
106 analysis questions were worth 2 points each, as were the Multi-Choice questions.
107 Additionally, there were 60 Common Questions worth 1.5 points each, and 30 other
108 Common Questions worth 2 points each.

109 All the inputs given for the GPT-3.5 model are valid samples that do not belong to
110 the training dataset. This is because the database has not been updated since
111 September 2021, which is prior to the release of these questions. In order to
112 streamline future research efforts, the 165 questions have been grouped into three
113 distinct categories, as listed below.

114 1.Common Questions (n=90): These questions are to evaluate the knowledge in basic
115 science in physiology, biochemistry, pathology, and medical humanities. There are
116 four choices for each question, and the respondent should collect the only correct
117 answers. For example: “The closing time of the aortic valve during the cardiac cycle
118 is? A. Atrial systolic end card B. Rapid ejection beginning C. Slow ejection beginning
119 D. Isovolumic diastole beginning”

120 2.Case Analysis Questions (n=45). It is a method used in clinical medicine to examine
121 and evaluate patient cases. It involves an in-depth review of a patient's medical
122 history, presenting symptoms, laboratory and imaging results, and diagnostic findings
123 to arrive at a diagnosis and treatment plan. There are four choices, and the respondent
124 should collect the only correct answers. The difference between Case Analysis
125 Questions and Common Questions is that Common Questions is focus on clinical
126 decision making. For example: “A 38-year-old male, suffering chest pain and fever
127 for 3 days, having a 5 years of diabetes history. Physical examination: T=37.6□, right
128 lower lung turbid knock, breathing sound is reduced. A chest X radiograph suggests a
129 right pleural effusion. Pleural aspiration liquefaction test showed WBC650×10⁶/L
130 with fine lymph Cell 90% in pleural fluid, with glucose of 3.2 mmol/L, the diagnosis
131 for this patient is? A. Tuberculous pleurisy B. malignant pleural effusion C. empyema
132 D. pneumonia-like pleural effusion”

133 3.Multi-Choices Questions (n=30): There are four choices, and the respondent should
134 collect the correct answers which is more than two. There are no points for choosing

135 more or less. For example: “The structures of auditory bone conduction include? A.
136 skull B. round window film C. ossicular chain D. cochlear bone wall”.

137 **Scoring**

138 Initially, we realized that modifying the question format was necessary to
139 accurately evaluate ChatGPT's performance in questions of the Chinese Clinical
140 Medicine Postgraduate Entrance Examination. Specifically, we found that reminding
141 the AI with "multi-choice" or "single-choice" was essential as ChatGPT produced
142 varying results without this specification. For the Multi-Choices Questions, we
143 modified it to read "Please choose one or more of the correct options," while the
144 Common Questions and Case Analysis Questions were modified to read "There is
145 only one correct answer." This only applies when evaluating the score of ChatGPT in
146 answering questions in the Chinese language.

147 We created a dataset consisting of questions from the Chinese Clinical Medicine
148 Postgraduate Entrance Examination and their corresponding answers. To ensure its
149 accuracy, we verified these answers by comparing them with those available on the
150 internet and consulting with senior doctors. We then used this dataset to evaluate
151 ChatGPT's performance on the exam by comparing its responses to the standard
152 answer. A high score on the exam would indicate that ChatGPT performed well on
153 this task.

154 **Encoding**

155 To better reflect the actual clinical situation, we modified these questions to be
156 open-ended. We presented the Case Analysis Questions to ChatGPT in different

157 variations without multiple-choice options and asked it to identify the disease the
158 patient had and explain its reasoning. For the Multi-Choices Questions, we removed
159 all the choices without reminding ChatGPT that there were multiple options. For the
160 Common Questions, we processed them in the same manner as the Multi-Choices
161 Questions. However, there was an exclusive group within the three subgroups that
162 could not be encoded in the same way as the others. These questions required
163 selecting one of the provided choices, and thus, we transformed them into a special
164 form (n=26), which is highlighted in yellow in Supplemental S1. For example, the
165 original question, "Which can inhibit insulin secretion? A. Increased free fatty acids
166 in blood B. Increased gastric inhibitory peptide secretion C. Sympathetic nerve
167 excitation D. Growth hormone secretion increases" was encoded as "Can an increase
168 in free fatty acids in the blood, an increase in gastric inhibitory peptide, an increase in
169 sympathetic nerve excitation, or an increase in growth hormone secretion can inhibit
170 insulin secretion?" The encoder was present in all three subgroups.

171 Furthermore, to minimize memory retention bias, a new chat session was initiated
172 for every enquiry.

173 **Adjudication**

174 AI outputs were independently scored for Accuracy, Concordance, and Insight by
175 two physician who were blinded to each other, adjudicators using the pre-defined
176 criteria Supplemental S2. A subset of 20 questions was used to train the physician
177 adjudicators who were not blinded to each other. The accuracy of ChatGPT's
178 responses was classified into three categories: accurate, inaccurate, and indeterminate.

179 Accurate responses mean that ChatGPT provided the correct answer. Inaccurate
180 responses included no answer, an incorrect answer, or multiple answers with incorrect
181 options. Indeterminate responses imply that the AI output does not provide a
182 definitive answer selection, or it believes that there is insufficient information to do so.
183 Concordance was defined as the ChatGPT's explanation affirms its provided answer,
184 while a discordant explanation contradicts it. Valuable insights were defined as
185 unique instances of text within the AI's explanations that met specific criteria: they
186 were nondefinitional, nonobvious, valid, and unique. Specifically, valuable insights
187 required additional knowledge or deduction beyond the input question, provided
188 accurate clinical or numerical information, and had the potential to eliminate multiple
189 answer choices with a single insight.

190 To reduce within-item anchoring bias, the adjudicators evaluated accuracy for all
191 items first, followed by concordance for all items. Two physicians were blinded to
192 each other. If there was discrepancy on the domains, a third physician adjudicator was
193 consulted. The number of third adjudicator for Common Questions and Multi-Choices
194 Questions was 7 and 3, respectively. The need for third adjudicator in Case Analysis
195 Questions was 1 for concordance. Ultimately, 11 items (6.8% of the dataset) required
196 the intervention of a third physician adjudicator. The interrater agreement between the
197 physicians was evaluated using the Cohen kappa (κ) statistic for the questions
198 (Supplemental S3).

199 A schematic overview of the study protocol is provided in Fig 1.

200 **Result**

201 **ChatGPT performance poor towards the original questions**

202 After inputting the original questions into ChatGPT and collecting its answers,
203 ChatGPT received a score of 153.5/300, which means that it only obtained 51.16% of
204 the total points on the test. This score is much lower than the expectation, but slightly
205 higher than the passing threshold (129/300) defined by official agencies.

206 Among three subgroups of questions, the evaluation revealed that out of a total of
207 90 Common Questions, ChatGPT only provided 50 (55.6%) correct answers.
208 Similarly, out of 45 Case Analysis Questions, ChatGPT provided 25 (55.6%) correct
209 answers. Furthermore, out of 30 Multi-Choices Questions, ChatGPT provided 10
210 (33.3%) completely accurate answers (Fig 2). These results suggest that ChatGPT's
211 ability to resolve medical problems in Chinese needs to be improvement.

212 **ChatGPT performs worse on encoded questions compared to the original** 213 **questions**

214 We encoded questions of the Chinese Clinical Medicine Postgraduate Entrance
215 Examination and inputted them into ChatGPT, which simulates scenarios where a
216 student asks a common medical question without answer choices, or a doctor tries to
217 diagnose a patient based on multimodal clinical data (i.e. symptoms, history, physical
218 examination, laboratory values). ChatGPT's accuracy was 31.5% for all questions.
219 Among the three subgroups, namely Common Questions, Multi-Choices Questions,
220 and Case Analysis Questions, the accuracy was 41.7%, 36.8%, and 16.7%,
221 respectively (Fig 2). Compared the original questions, the accuracy of their encoding
222 questions decreased by was 19%, 17%, and 14%, for Common Questions, Multi-

223 Choices Questions, and Case Analysis Questions, respectively, which demonstrates
224 the ability of ChatGPT answering the open-ended questions in Chinese is shortcoming.
225 During the adjudication stage, there was substantial agreement among physicians for
226 prompts in all three subgroups (κ ranged from 0.80 to 1.00).

227 *ChatGPT demonstrates high internal concordance*

228 Concordance, which is a measure of the level of agreement or similarity between
229 the option selected by AI and its subsequent explanation, was also taken into
230 consideration. The results showed that ChatGPT had 90% concordance across all
231 questions, and this high concordance was maintained across all three subgroups (Fig
232 3). Additionally, we analyzed the concordance difference between correct and
233 incorrect answers and found that concordance among accurate responses was perfect
234 and significantly greater than among inaccurate responses (100% vs. 50%, $p < 0.001$)
235 (Fig 3). These findings suggest that ChatGPT has a high level of answer-explanation
236 concordance in Chinese, likely due to its strong internal consistency in its
237 probabilistic language model.

238 *ChatGPT shows multiply insights towards the same questions*

239 Another evaluation index considered was the frequency of insights generated by the
240 AI model, which quantifies the quantity of insights produced. After evaluating the
241 score, accuracy, and concordance of ChatGPT, we investigated its potential to
242 enhance medical education by augmenting human learning. We examined the
243 frequency of insights provided by ChatGPT. Remarkably, ChatGPT generated at least
244 one significant insight in 80% of all questions (Fig 4). Moreover, the analysis

245 revealed that the accuracy response had the highest frequency of insights, with an
246 average of 2.95. The indeterminate response followed closely behind with an average
247 of 2.7, while the inaccurate response had a lower frequency of insights with an
248 average of 1.39 (Fig 4). The high frequency of insights in the accurate group suggests
249 that it may be feasible for a target learner to acquire new or remedial knowledge from
250 the ChatGPT AI output.

251 **Discussion**

252 **Major findings**

253 To evaluate ChatGPT's problem-solving capabilities and assess its potential for
254 integration into medical education in Chinese, we tested its performance on the
255 Chinese Clinical Medicine Postgraduate Entrance Examination. Our findings can be
256 organized into 2 major themes: (1) The score of ChatGPT, which needs to be
257 improved when facing questions asking in Chinese language; (2) There are still
258 potential for this AI to generate novel performance that can assist human due to the
259 high concordance and the frequency of insights. This is the first study to assess the
260 performance of ChatGPT on in medical care and clinical decision in Chinese.

261 **ChatGPT performance need improvement for medical questions in Chinese**

262 A recent study showed the ChatGPT3.5 performed with an accuracy rate of over
263 50% across all examinations and even exceeded 60% accuracy in some analyses when
264 facing the United States Medical Licensing Exam (USMLE)⁷. In our results, the study
265 found that ChatGPT exhibited moderate accuracy in answering open-ended medical

266 questions in Chinese, with accuracy was 31.5%. Given the differences between
267 English and Chinese inputs, we conclude that ChatGPT requires further improvement
268 in answering medical questions in the Chinese language.

269 We sought to understand why there is a significant discrepancy between the
270 performance of ChatGPT on Chinese and English language exams. To investigate this,
271 we asked the ChatGPT for the reason, it explains that the training data used to train AI
272 in different languages may be different, and the algorithms used to process and
273 analyze text may vary from language to language (data not shown). Therefore, even
274 for the same question, the output generated may vary slightly based on the language
275 and the available language-based data.

276 Upon analyzing the results of our research, we found that the accuracy of ChatGPT
277 was lowest for Multi-Choices Questions, followed by Common Questions, and Case
278 Analysis Questions. The lower accuracy on Multi-Choices Questions s may be due to
279 the model being undertrained on the input, as well as the Multi-Choices Questions
280 samples being significantly less than those of single-choice questions. On the other
281 hand, the Case Analysis Questions may have extensive training compared to Multi-
282 Choices Questions, is similar in type to the USMLE question.

283 Furthermore, we noticed that high accuracy outputs were associated with high
284 concordance and a high frequency of insight, whereas poorer accuracy was linked to
285 lower concordance and a lack of insight. Thus, we hypothesized that inaccurate
286 responses were primarily driven by missing information, which could result in
287 reduced insight and indecision in the AI, rather than an over-commitment to an

288 incorrect answer⁷. The results indicate that enhancing the database and providing
289 additional training with Chinese questions could lead to a substantial improvement in
290 the performance of the model.

291 **Challenges of AI in future applications**

292 Despite the promising potential of AI in medicine, it also poses some challenges.
293 Standards for use of AI in health care are still need to be developed^{8,9}, including
294 clinical care, quality, safety, malpractice, and communication guidelines. Furthermore,
295 the implementation of AI in healthcare requires a shift in medical culture, which poses
296 a challenge for both medical education and practice. Additionally, ethical
297 considerations must be taken into account, such as data privacy, informed consent,
298 and bias prevention, to ensure that AI is used ethically and for the benefit of patients.
299 Surprisingly, A recently launched AI system for autonomous detection of diabetic
300 retinopathy carries medical malpractice and liability insurance¹⁰.

301 **Prospective of AI**

302 AI is a rapidly growing technology. At this time of writing, the ChatGPT has
303 released version 4 with great improvement. Numerous practical and observational
304 studies have demonstrated the versatile role of AI in almost all medical disciplines
305 and specialties, particularly in improving risk assessment,^{11,12} data reduction, clinical
306 decision support^{13,14}, operational efficiency, and patient communication^{15,16}. We
307 anticipate that advanced language models such as ChatGPT are reaching a level of
308 maturity that will soon have a significant impact on clinical medicine, enhancing the
309 delivery of personalized, compassionate, and scalable healthcare.

310 **Limitations**

311 With the limitation of our research is the small sample size. We only access 165
312 samples to qualify its accuracy and 30 case analysis questions to qualify its
313 concordance and frequency of insight, Furthermore, the clinical situation is more
314 complicated than the test, larger and deep analyses were needed. Finally, bias and
315 error were inevitable for human-adjudication, although there was a good interrater
316 agreement between the physicians for the adjudication.

317 **Conclusion**

318 In conclusion, although the ChatGPT's got a score over the passing score in
319 Clinical Medicine Entrance Examination for Chinese Postgraduate in Chinese
320 language, the performance was limited when presented with open-ended questions.
321 On the other hand, ChatGPT demonstrated a high level of internal concordance,
322 which suggests that the explanations provided by ChatGPT support and affirm the
323 given answers. Moreover, ChatGPT generated multiple insights towards the same
324 questions, demonstrating its potential for generating a variety of useful information.
325 Further prospective studies are needed to explore whether there was a language-based
326 difference in performance of medical education setting and clinical decision-making,
327 such as Chinese and minority language.

328

329 **Data statement**

330 All data generated or analyzed during this study are included in this published article
331 [and its supplementary information files].

332 **Acknowledgments**

333 We acknowledge the ChatGPT for polishing our manuscript.

334 **Author contributions**

335 X-L and W.G.-Z was responsible for the entire project and revised the draft. J. W-C,
336 K.B-M and Q. W-H performed the study selection, data extraction, statistical analysis,
337 and interpretation of the data. J. W-C and X.L. drafted the first version of the
338 manuscript. All authors participated in the interpretation of the results and prepared
339 the final version of the manuscript.

340 **Funding**

341 None

342 **Declarations**

343 Ethics approval This is a systematic review and meta-analysis. No ethical approval is
344 required.

345 **Conflict of interest**

346 All authors declare no competing interests.

347

348

References

349 1. Haleem A, Javaid M, Khan IH. Current status and applications of Artificial
350 Intelligence (AI) in medical field: An overview. *Current Medicine Research and*
351 *Practice*. 2019;9(6):231-237.

352 2. Haleem A, Vaishya R, Javaid M, et al. Artificial Intelligence (AI) applications in
353 orthopaedics: An innovative technology to embrace. *Journal of Clinical Orthopaedics*

- 354 and Trauma. 2019(0976-5662 (Print)).
- 355 3. Jha S, Topol EJ. Information and artificial intelligence. Journal of the American
356 College of Radiology. 2018;15(3):509-511.
- 357 4. Lupton ML. Some ethical and legal consequences of the application of artificial
358 intelligence in the field of medicine. 2018.
- 359 5. Murdoch TB, Detsky AS. The inevitable application of big data to health care.
360 JAMA. 2013(1538-3598 (Electronic)).
- 361 6. Misawa M, Kudo S-e, Mori Y, et al. Artificial intelligence-assisted polyp
362 detection for colonoscopy: initial experience. Gastroenterology. 2018;154(8):2027-
363 2029.
- 364 7. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE:
365 Potential for AI-assisted medical education using large language models. PLOS
366 digital health. 2023;2(2):e0000198.
- 367 8. F. D-V. Considerations for the Practical Impact of AI in Healthcare Food and
368 Drug Administration.
- 369 9. Zweig M EBRH. How should the FDA approach the regulation of AI and
370 machine learning in healthcare?
371 Available: [https://rockhealth.com/how-should-the-fda-approach-the-regulation-of-ai-
372 and-machine-learning-in-healthcare/](https://rockhealth.com/how-should-the-fda-approach-the-regulation-of-ai-and-machine-learning-in-healthcare/).
- 373 10. Abramoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based
374 diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ
375 digital medicine. 2018;1(1):39.

- 376 11. Kan HJ, Kharrazi H, Chang H-Y, et al. Exploring the use of machine learning for
377 risk adjustment: A comparison of standard and penalized linear regression models in
378 predicting health care costs in older adults. *PloS one*. 2019;14(3):e0213258.
- 379 12. Delahanty RJ, Kaufman D, Jones SS. Development and evaluation of an
380 automated machine learning algorithm for in-hospital mortality risk adjustment
381 among critical care patients. *Critical care medicine*. 2018;46(6):e481-e488.
- 382 13. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage
383 clinical evaluation of decision support systems driven by artificial intelligence:
384 DECIDE-AI. *Nature medicine*
385 2022;28(5):924-933.
- 386 14. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, et al. Artificial intelligence to
387 support clinical decision-making processes. *EBioMedicine*
388 2019;46:27-29.
- 389 15. Bala S, Keniston A, Burden M. Patient perception of plain-language medical
390 notes generated using artificial intelligence software: pilot mixed-methods study.
391 *JMIR Formative Research*
392 2020;4(6):e16670.
- 393 16. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence
394 conversational agents in health care: systematic review. *Journal of medical Internet*
395 *research*
396 2020;22(10):e20346.
- 397

398

Figure legends

399 **Fig 1. Schematic of workflow for sourcing, encoding, and adjudicating results.**

400 Abbreviations: **CQ** =Common Questions; **CAQ** =Case Analysis Questions; **MCQ**
401 =Multi-Choices Questions. The 165 questions were categorized into three types: CQ,
402 CAQ, and MCQ, and each question was assessed for its score. The accuracy of the
403 CQ and MCQ questions were evaluated, while the MCQ questions were also assessed
404 for the accuracy, concordance, and frequency of insights. The adjudication process
405 was carried out by two physicians, and in case of any discrepancies in the domains, a
406 third physician was consulted for adjudication. Additionally, any inappropriate output
407 was identified and required re-encoding.

408 **Fig 2. Accuracy of ChatGPT on Chinese Clinical Medicine Postgraduate**

409 **Entrance Examination Test before and after encoding.** For the subgroup CQ, CAQ
410 and Multi-Choices Questions before encoding, AI output were comparing with the
411 standard answer key. For the subgroup CQ, CAQ and Multi-Choices Questions after
412 encoding, AI outputs were adjudicated to be accurate, inaccurate, or indeterminate
413 based on the scoring system provided in S2 Data. It demonstrates the different
414 accuracy distribution for inputs between the before and the after.

415 **Fig 3. Concordance of ChatGPT on Chinese Clinical Medicine Postgraduate**

416 **Entrance Examination after encoding.** For the subgroup “case analysis question”,
417 AI outputs were adjudicated to be concordant and discordant, based on the scoring
418 system provided in S2 Data. It demonstrates concordance rates stratified between

419 accurate, inaccurate and indeterminate outputs, across all of the CAQ

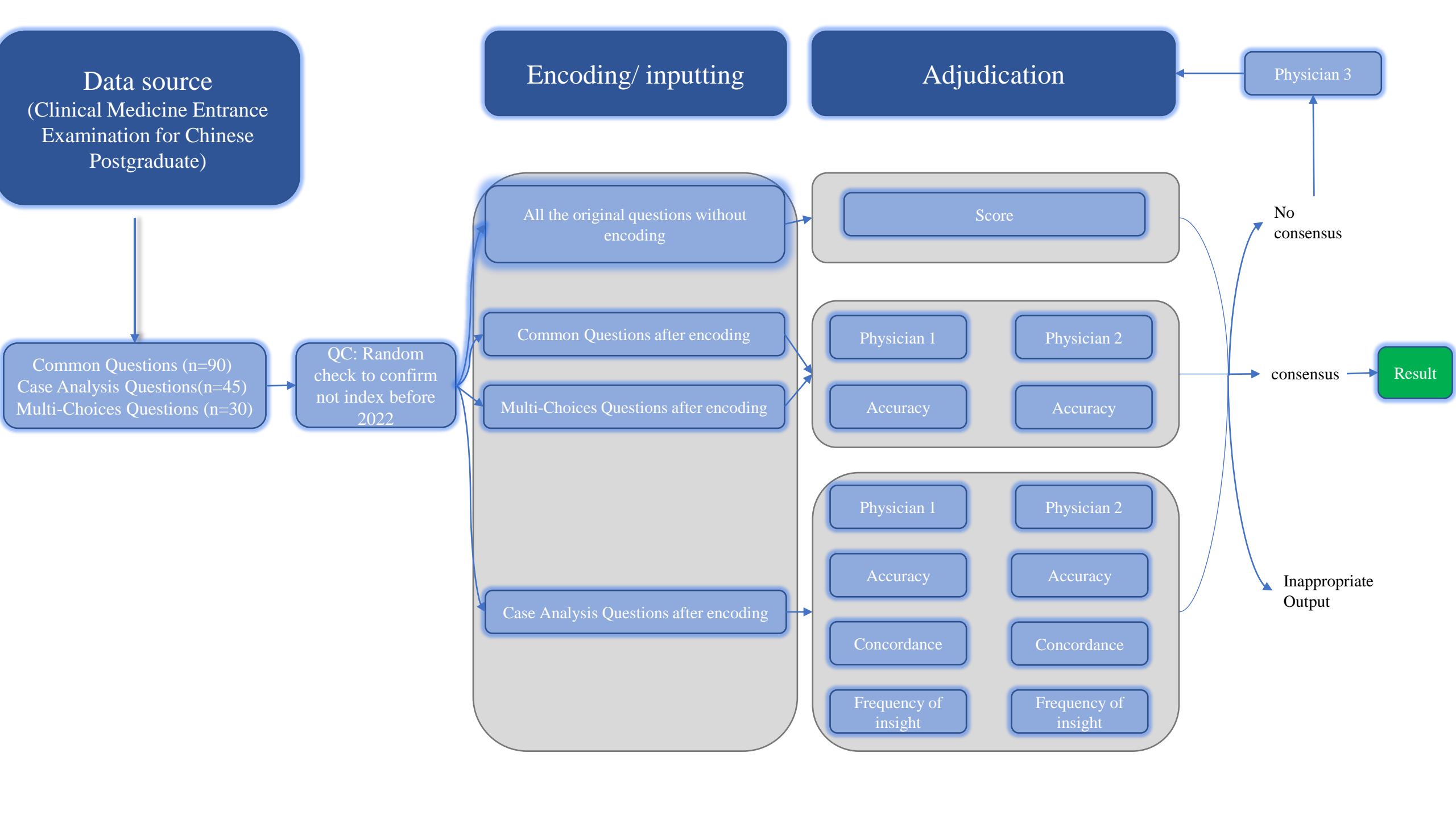
420 **Fig 4. The frequency of insights of ChatGPT on Chinese Clinical Medicine**

421 **Postgraduate Entrance Examination after encoding.** For the subgroup “case

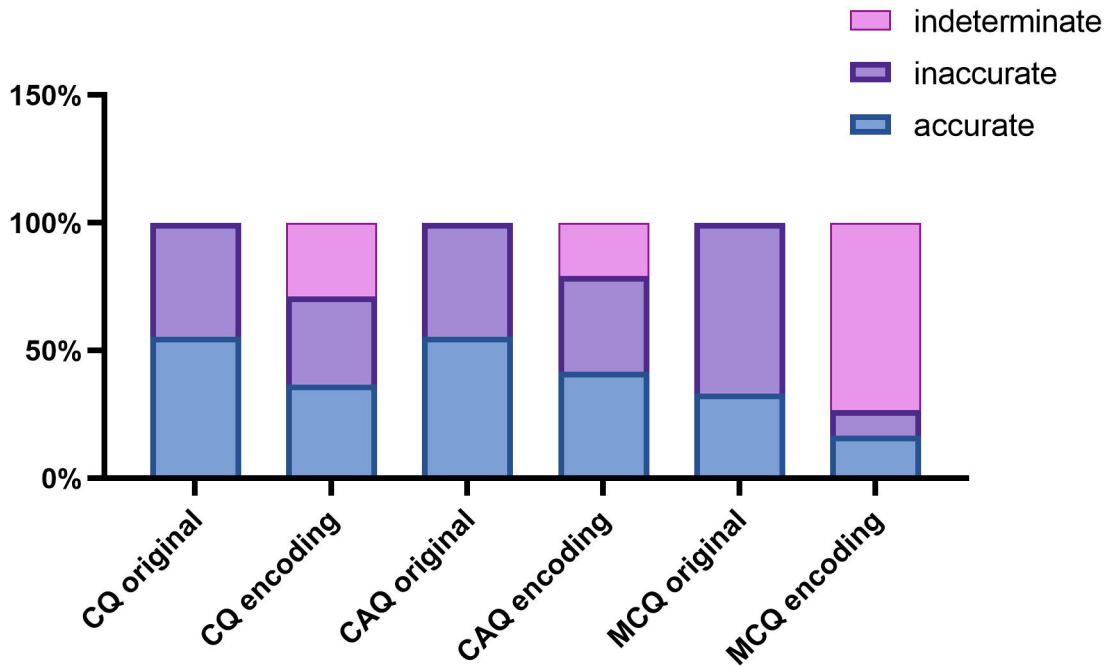
422 analysis question”, AI outputs were adjudicated to count the frequency of insights it

423 offered. It demonstrates frequency of insights stratified between accurate, inaccurate

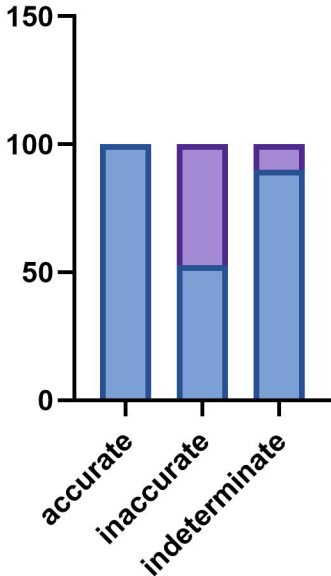
424 and indeterminate outputs, across all of the CAQ.



accuracy%



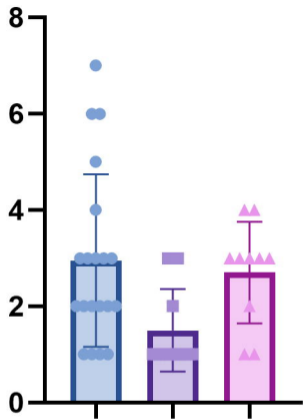
concordance %



concordance

discordance

frequency of insight



- Accurate
- Inaccurate
- ▲ Indeterminate

Accurate
Inaccurate
Indeterminate