

1 Performance of Generative Pretrained Transformer on the National Medical 2 Licensing Examination in Japan

3
4
5 Yudai Tanaka^{1,2,3*}, Takuto Nakata^{1,2,3*}, Ko Aiga^{2*}, Takahide Etani^{1,4,5}, Ryota Muramatsu^{1,3}, Shun
6 Katagiri¹, Hiroyuki Kawai¹, Fumiya Higashino¹, Masahiro Enomoto¹, Masao Noda⁶, Mitsuhiro
7 Kometani², Masayuki Takamura⁷, Takashi Yoneda^{2,8}, Hiroaki Kakizaki⁹, Akihiro
8 Nomura^{2,7,8,10,11*}

9
10
11 ¹School of Medicine, Kanazawa University, Kanazawa, Japan
12 ²Department of Health Promotion and Medicine of the Future, Kanazawa University Graduate
13 School of Medicine, Kanazawa, Japan
14 ³Department of Molecular and Cellular Pathology, Kanazawa University Graduate School of
15 Medicine, Kanazawa, Japan
16 ⁴Graduate School of Media and Governance, Keio University, Fujisawa, Japan
17 ⁵Advanced Research Center for Human Sciences, Waseda University, Saitama, Japan
18 ⁶Department of Otolaryngology and Head and Neck Surgery, Jichi Medical University, Tochigi,
19 Japan
20 ⁷Department of Cardiovascular Medicine, Kanazawa University Graduate School of Medical
21 Sciences, Kanazawa, Japan
22 ⁸College of Transdisciplinary Sciences for Innovation, Kanazawa University, Kanazawa Japan
23 ⁹MICIN, Inc., Tokyo, Japan
24 ¹⁰Frontier Institute for Tourism Science, Kanazawa University, Kanazawa, Japan
25 ¹¹CureApp Institute, Karuizawa, Japan
26 *Contributed equally

27
28
29

30 Corresponding author:

31 Akihiro Nomura, MD, PhD
32 Associate Professor
33 College of Transdisciplinary Sciences for Innovation, Kanazawa University
34 Kakuma-machi, Kanazawa, Ishikawa, 9201192, Japan
35 E-mail: anomura@med.kanazawa-u.ac.jp
36 ORCID ID: 0000-0001-6647-8240

37 **Abstract**

38 The remarkable performance of ChatGPT, launched in November 2022, has significantly impacted
39 the field of natural language processing, inspiring the application of large language models as
40 supportive tools in clinical practice and research worldwide. Although ChatGPT recently scored
41 high on the United States Medical Licensing Examination, its performance on medical licensing
42 examinations of other nations, especially non-English speaking nations, has not been sufficiently
43 evaluated. This study assessed ChatGPT’s performance on the National Medical Licensing
44 Examination (NMLE) in Japan and compared it with the actual minimal passing rate for this exam.
45 In particular, the performances of both the GPT-3.5 and GPT-4 models were considered for the
46 comparative analysis. We initially used a model and prompt tuning set of 290 questions without
47 image data from the previous 116th NMLE (held in February 2022) to maximize the performance
48 for delivering correct answers and explanations of the questions. Thereafter, we tested the
49 performance of the best ChatGPT model (GPT-4) with tuned prompts on a dataset of 262 questions
50 without images from the latest 117th NMLE (held in February 2023). The best model with the
51 tuned prompts scored 82.7% for the essential questions and 77.2% for the basic and clinical
52 questions, both of which sufficed the minimum passing rates of 80.0% and 74.6%, respectively.
53 Simultaneously, we identified the three major factors contributing to the generation of the incorrect
54 answers—insufficient medical knowledge, information on Japan-specific medical system and
55 guidelines, and mathematical errors. In conclusion, GPT-4 powered ChatGPT with our optimally
56 tuned prompts achieved a minimum passing rate in the latest 117th NMLE in Japan. Although we
57 express strong concerns regarding the use of the current ChatGPT for medical purposes so far,
58 these artificial intelligence models may soon have the potential to serve as one of the best
59 “sidekicks” for solving medical and healthcare problems.

60

61

62 **Key words:** medical licensing examination, large language model, generative pretrained
63 transformer, artificial intelligence

64 **Author summary (150 words)**

65 ChatGPT's remarkable performance has inspired the use of large language models as supportive
66 tools in clinical practice and research. Although it scored well in the US Medical Licensing
67 Examination, its effectiveness in relevant examinations of non-English speaking countries remain
68 unexplored. This study assessed the performance of ChatGPT with GPT-3.5 and GPT-4 models in
69 Japan's National Medical Licensing Examination (NMLE). Initially, we used a tuning set of 290
70 questions from the 116th NMLE, and then the GPT-4 model with tuned prompts was tested on 262
71 questions from the 117th NMLE. The model scored 82.7% for essential and 77.2% for basic and
72 clinical questions, surpassing the minimum passing rates. Incorrect answers were attributed to
73 insufficient medical knowledge, Japan-specific medical system information, and mathematical
74 errors. In conclusion, GPT-4 powered ChatGPT achieved a minimum passing rate and might have
75 the potential for a valuable tool for fulfilling the needs of medical and healthcare fields.

76 **Introduction**

77 In recent decades, artificial intelligence (AI) algorithms have been widely applied in medical and
78 healthcare fields [1]. Currently, the AI algorithms available for clinical applications have been
79 developed using previous rule-based methods as well as recent machine learning (ML) methods
80 including its subfield of deep learning, promoted by the continually increasing availability of
81 computer resources and vast amount of medical data [2]. Consequently, these medical AI products
82 have been implemented to obtain targeted outputs such as the *prediction* of future disease risk,
83 *classification* as diagnostic support, or *generation* of various texts or images using natural
84 language processing (NLP) in medicine [1-3].

85 NLP is an area of AI that addresses the interaction between human languages and machines
86 [4]. The major roles of NLP in medicine and healthcare include serving as supportive tools in
87 clinical practice and research [3]. Beyond the prediction of certain risk factors or clinical decision-
88 making, NLP assists physicians and researchers to efficiently extract, translate, classify and
89 analyze patients' information and clinical-free text in electronic medical and health records, in
90 addition to dialogue generation and answering medical information [3, 4]. The performance of
91 NLP has dramatically improved following the emergence of transformer-based large language
92 models (LLMs). A transformer is a type of neural network model that employs self-attention
93 mechanism, relating multiple positions of a single sequence to compute a representation of the
94 sequence [5]. LLMs are created using advanced ML techniques, especially deep neural networks,
95 trained on enormous amounts of text data from the Internet and other sources [4]. A few notable
96 LLMs include pretrained Bidirectional Encoder Representations from Transformers (BERT) [6],
97 Language Models for Dialog Applications (LaMDA) [7], Pathway Language Model (PaLM) [8],
98 Large Language Model Meta (LLaMA) [9], and Generative Pretrained Transformer (GPT)-3 and
99 later models [10-12].

100 Recently, InstructGPT (GPT-3.5)—a GPT model employing 175 billion parameters with
101 supervised fine-tuning and reinforcement learning from human feedback [11]—and its dialogue-
102 optimized chatbot (ChatGPT) launched in November 2022 have significantly impacted NLP fields
103 [13]. By predicting the subsequent element of the texts, ChatGPT can comprehend user prompts
104 and generate human-like responses, expressed in ethical, sentimental, logical, and creative manner,
105 without any additional training (*e.g.*, foundation model) [14]. Although GPT is a non-domain-
106 specific LLM, not exclusively intended to be used for medical or healthcare fields, recent

107 publications have demonstrated that ChatGPT (GPT-3.5) possesses sufficient ability to pass the
108 United States Medical Licensing Examination [15, 16]. In contrast, another study reported
109 ChatGPT's inadequate performance on non-English-based Korean medical questions [17].
110 Although the performance variation can be attributed to differences in languages, domestic
111 healthcare systems, diagnostic criteria, and treatment strategies, the relationship between these
112 differences and ChatGPT's performance in answering medical questions remains unclear.
113 Furthermore, the performance of ChatGPT with the current GPT-4 model employing an estimated
114 10 trillion parameters [12] has not yet been evaluated on the latest Medical Licensing Examination,
115 which was originally written in non-English texts and held after the completion of GPT-4 model
116 training (August 2022) [18].

117 Therefore, this study tested the performance of GPT (both GPT-3.5 and GPT-4 models) on
118 the 117th National Medical Licensing Examination (NMLE) (held in February 2023 in Japan),
119 which was originally conducted in the Japanese language. In particular, questions from the
120 previous year (116th NMLE exam held in February 2022) were used as a model and prompt
121 performance tuning set before using the latest questions (117th exam held in February 2023) as a
122 performance testing set to verify whether GPT can qualify for the actual minimal passing rate of
123 this examination.

124
125

126 **Results**

127 *Improving performance through English translation and tuned prompts in 116th NMLE (2022)*

128 Initially, we used the non-image-based questions from 116th NMLE in Japan to develop the optimal
129 input prompts for ChatGPT to maximize the correct answer rate. We extracted the question data
130 from the 116th NMLE containing 394 questions (originally 400 questions, but six were officially
131 removed from scoring evaluation). Thereafter, we removed questions with image data (n = 104)
132 and analyzed the remaining 290 questions without image data (**Figure 1**).

133 Using the ChatGPT API powered by GPT3.5, we initially tested its performance for the
134 original questions in Japanese language. Initially, we obtained a correct answer rate of 52.8%
135 (153/290) with an output error rate of 5.5% (16/290). Accordingly, we used updated prompts to
136 translate the original Japanese NMLE questions into English using ChatGPT before inputting them

137 as questions. Although this marginally increased the correct answer rate to 56.2% (163/290), the
138 output errors increased to 14.8% (43/290; **Figure 2**).

139 To further improve the correct answer rate and reduce the errors, we tuned our prompts
140 for each question type (Basics of Medicine, Clinical Medicine, and Comprehension). In particular,
141 we provided sample outputs and directed the model to translate the questions into plain English
142 and create summaries before answering the questions (**Figure 3**). This tuned prompt improved the
143 correct answer rate to 63.1% (183/290) with a reduced output error rate of 7.6% (22/290).
144 Furthermore, we applied the above-tuned prompts to the GPT-4-based ChatGPT, which
145 demonstrated a correct answer rate of 82.8% (240/290) and a minimal error rate of 1.0% (3/290)
146 (**Figure 2**).

147

148 *GPT-4-based ChatGPT performance on 117th (2023) NMLE with tuned prompt*

149 Thereafter, we evaluated that the performance of the best model (GPT-4) with a tuned prompt for
150 the test set of 262 questions without image data from the 117th NMLE in Japan, held in February
151 4th and 5th, 2023, after the completion of GPT-4 model training in August 2022 (**Figure 1**). With a
152 tuned prompt, the best model achieved a correct answer rate of 78.6% (206/262) and an output
153 error rate of 0.8% (2/262) (**Table 1**).

154 The present results were compared with the actual minimal passing rate on the examination.
155 The current model with a tuned prompt scored 82.7% (129/156) for essential questions and 77.2%
156 (139/180) for basic and clinical questions, both of which qualified the minimum passing rates of
157 80.0% and 74.6%, respectively (**Figure 2**) [19]. Notably, we applied the GPT-4 model with tuned
158 prompts to the entire set of 395 questions (text-only) in the 117th NMLE, regardless of containing
159 image data (originally 400 questions, but five were officially removed from scoring evaluation).
160 This optimal model attained near-passing levels of 78.5% (157/200) for essential questions and
161 73.2% (216/295) for basic and clinical questions.

162

163 *Exploratory analysis of incorrect ChatGPT responses and their associated explanations*

164 To further enhance the performance of the model, we performed an exploratory analysis of 56
165 incorrect answers provided by the optimal GPT-4 model with tuned prompts for the 117th NMLE
166 questions. As listed in **Table 2**, the three primary factors contributing to the generation of incorrect
167 answers by the model included insufficient medical knowledge (33/56, 58.9%), Japan-specific

168 medical system information (17/56, 30.4%), and mathematical errors (4/56, 7.1%). Concerning the
169 insufficient medical knowledge, the areas of incorrect answers were not specific and spanned
170 across various medical fields. Notably, certain answers were outdated or critically incorrect in
171 current medical contexts (**Figure 4**). In terms of Japan-specific medical system, ChatGPT failed
172 to adequately answer questions related to Japanese medicolegal laws applicable in the medical and
173 healthcare field, guidance from the Ministry of Health, Labour, and Welfare (MHLW) in Japan,
174 and guidelines, especially those related to public health. Additionally, we noted several
175 mathematical errors such as in addition calculations (*e.g.*, the explanation and addition formula
176 were correct, but the answer was wrong) and handling decimal points (because of translation errors
177 from the phrase “rounding to first decimal point” from Japanese).

178
179

180 **Discussion**

181 This study evaluated the performance of GPT on the Japanese Medical Licensing Examination.
182 The results indicate that 1) GPT-4 with a tuned prompt cleared the minimal passing rate on the
183 116th (2022) NMLE in Japan; 2) GPT-4 with tuned prompt qualified the minimum passing rate on
184 the latest 117th NMLE (2023); and 3) Inadequate medical knowledge, Japan-specific medical
185 system information, and mathematical errors were the primary factors associated with the incorrect
186 answers generated by the optimal model. Despite the absence of image data in the questions, this
187 study demonstrated the first attempt to use the best available ChatGPT model with tuned prompts
188 to achieve a minimum passing rate for the latest 117th NMLE in Japan.

189 This study provides several conclusions. First, GPT-4 with a tuned prompt cleared the
190 minimal passing rate on the 116th NMLE in Japan held in February 2022. Although GPT-3.5-based
191 ChatGPT achieved a correct answer rate of 52.8% for Japanese questions, it increased to 56.2%
192 after translating the questions into English. As GPT-3, the original GPT-3.5, was primarily trained
193 in English, it delivers a higher performance when responding to prompts in English compared to
194 other languages [10]. Similarly, a recent multilingual performance evaluation of GPT-4, an
195 improved version of GPT-3, confirmed that the best performance is more generally obtained with
196 English prompts [12]. After tuning our prompts to include a translation procedure into plain
197 English and modifying the output format based on the question type, the correct response rate
198 increased to 63.1%. This finding is consistent with previous studies claiming that prompt

199 engineering can improve model task performance [12, 20]. These improved correct response rates
200 can be attributed to English being the majority of the language in the training data, *i.e.*, the Internet,
201 used by non-experts [21]. Although the error rate increased to 14.8% upon translating the Japanese
202 questions into English, it notably decreased to 7.6% after tuning the prompts by including the
203 format of the output. This result suggests that providing samples and standardizing the output
204 format can produce the desired output format and reduce the number of errors. Finally, upon
205 applying these optimized prompts to GPT-4-based ChatGPT, the correct response rate increased to
206 82.8% and the error rate plummeted to 1.0%. This significant improvement in performance can be
207 ascribed to the advanced architecture and training of GPT-4 [12].

208 Second, even in case of the latest 117th NMLE (2023), GPT-4 with tuned prompt qualified
209 the actual minimum passing rate. GPT-4 has passed various professional examinations in English,
210 including the practice bar exam with a score in the top 10% of examinees [12]. A previous study
211 reported that ChatGPT (GPT-3.5) failed to achieve the minimum passing rates [22]. However, this
212 study demonstrated that ChatGPT (GPT-4) can pass the 117th NMLE with the optimized prompt
213 tuning method proposed herein. The current results can be derived from the exquisite combination
214 of essential factors such as English translation and optimally tuned prompts for obtaining correct
215 answers through the best performance of the latest ChatGPT model.

216 Third, inadequate medical knowledge, information related to the medical and healthcare
217 system guidelines of Japan, and mathematical errors formed the three major factors of the incorrect
218 answers generated by the best available ChatGPT model with tuned prompts. Among the incorrect
219 answers associated under inadequate medical knowledge, no significant bias was observed for the
220 medical fields relevant to each question. Furthermore, even after providing incorrect answers, the
221 model output plausible but wrong medical explanations (so-called hallucinations in LLM outputs
222 [23]). Therefore, even if the model exhibits a performance level that surpasses the minimum score
223 for the NMLE, a broader range of specialized and up-to-date medical knowledge regarding
224 standard treatments should be inputted. In addition, output receivers should be equipped with
225 professional medical knowledge to assess the correctness of the output. For the Japan-specific
226 system, several incorrect answers were observed, especially in public health-related questions,
227 which are based on Japanese laws, guidelines, and unique systems. Although the GPT-4 powered
228 ChatGPT delivered improved performance in terms of output differences between the languages,
229 every country should perform their individual localization in terms of the applicable laws and

230 systems considering the language differences. Furthermore, in certain cases related to
231 mathematical errors, the calculation formula in the explanation was correct, but the result and the
232 final answer output were incorrect. Moreover, an instruction of "approximating the decimal place"
233 was not properly comprehended by ChatGPT during the Japanese-to-English translation. As such,
234 calculation problems are reported as one of the areas where LLMs still exhibit relatively low
235 accuracy [24], indicating that calculation problems may be a relatively unsuitable field for current
236 ChatGPT.

237 As discussed, we express strong concerns regarding the use of the current ChatGPT for
238 medical purposes, as OpenAI has already indicated that the models should not be used for
239 providing triage, diagnosis, or treatment options for life-threatening issues or severe medical
240 conditions [25]. Indeed, for use in medical settings, an approval must be obtained from regulatory
241 agencies, *e.g.*, software as a medical device. Moreover, utilizing such technology is already
242 difficult with its several black-box aspects [12]. Various countries have released statements
243 regarding the applications of LLMs in medical fields [26, 27]. Although the versatility of these
244 models hinders the verification of their validity and they require enormous computational
245 resources and costs, we believe that the advanced medical foundation AI model [28] can replace
246 task-specific approach AI models and will appear not far off, with scientifically proven clinical
247 efficacy and safety in medical and healthcare fields.

248 The novelty of this study is that it is the first research to achieve a minimum passing rate
249 using 262 non-image questions in the latest 117th NMLE in Japan with the ChatGPT GPT-4 version
250 with the optimally tuned prompts. The limitations of this study were as follows. First, we only
251 used questions without image data to evaluate the performance of the best available model with
252 tuned prompts, although it might be fair to assess the ability of the model to pass the examination
253 using all questions, regardless of image data. However, as revealed from the Results, we observed
254 a favorable model performance even upon using the entire question set in the 117th NMLE in Japan.
255 Second, the NMLE in Japan uniquely included strongly not-recommended "contraindication"
256 answer choices within the questions. The MHLW in Japan has set the minimum passing criteria
257 regarding selecting contraindication answer choices to be equal or less than three for the 116th
258 NMLE or two for the 117th NMLE. As the real number of contraindication answer choices were
259 not officially announced by the MHLW, we could not use them in the current performance
260 evaluation.

261 In conclusion, GPT-4 powered ChatGPT with optimally tuned prompts achieved a
262 minimum passing rate in the latest 117th NMLE in Japan. In addition, the model scored near-
263 passing levels for the entire test dataset of 395 questions, regardless of medical image data. The
264 upcoming GPT-4 version, which features enhanced image recognition capabilities, will easily
265 qualify the minimum passing rate and achieve top-tier scores, as reported in other English-based
266 examinations [12]. We again express strong concerns in terms of using of the current ChatGPT for
267 medical purposes so far. However, beyond its original design of answering examination questions
268 for humans, these AI models might have the potential be regarded as one of the best “sidekicks”
269 for solving problems and fulfilling the current needs in the medical and healthcare fields in the
270 near future.

271 **Materials and methods**

272 *Study overview*

273 This study evaluated the performance of GPT models on the NMLE in Japan. We utilized both the
274 GPT-3.5 and GPT-4 models of ChatGPT (Open AI, Inc., San Francisco, CA, USA). Initially, the
275 questions from the 116th NMLE in Japan (February 2022) were used as a model and prompt tuning
276 set to optimize the performance of obtaining the correct answers and explanations. Subsequently,
277 we assessed the performance of the best ChatGPT model (GPT-4) with the tuned prompts for
278 answering the questions from the 117th NMLE in Japan (February 2023).

279

280 *Input source*

281 The questions and answers for the 116th NMLE in Japan were obtained from the official website
282 of the MHLW, Japan [29]. For the latest 117th NMLE, we manually performed optical character
283 recognition on the original question papers to create input data and extracted the official answers
284 from the MHLW website [19]. The examination comprised six blocks (A–F), with 75 questions in
285 blocks A, C, D, and E, and 50 questions in blocks B and F. Note that six questions in the 116th
286 NMLE and five in the 117th NMLE were excluded. In addition, all image-containing questions
287 were removed from both the prompt-tuning and the performance-testing datasets, because up till
288 early April 2023, only text-based questions could be used as input to the ChatGPT interface,
289 including the API. The number of image-containing questions was 104 in the 116th NMLE and 133
290 in the 117th NMLE. Thereafter, according to the Japanese NMLE scoring method, the remaining
291 questions without image data were classified into the categories of "Essential" and "Basic and
292 Clinical." The 116th NMLE in Japan included 47 questions related to basics of medicine (essential),
293 24 questions of clinical medicine (essential), 14 questions on comprehension (essential), 65
294 questions regarding basics of medicine (general), 30 questions in basics of medicine (specifics),
295 31 questions of clinical medicine (general), 60 questions of clinical medicine (specifics), and 19
296 questions on comprehension. The 117th NMLE in Japan comprised 45 questions related to basics
297 of medicine (essential), 22 questions of clinical medicine (essential), 15 questions on
298 comprehension (essential), 61 questions from the basics of medicine (general), 27 questions on the
299 basics of medicine (specifics), 36 of clinical medicine (general), 46 questions related to clinical
300 medicine (specifics), and 10 questions regarding comprehension. Finally, we used 290 questions
301 (without image data) from the 116th NMLE and 262 questions (without image data) from the 117th

302 NMLE in Japan for analyses. The entire set of 395 text-based questions, irrespective of image data,
303 from the 117th NMLE in Japan was considered for the exploratory analysis.

304

305 *Generative Pretrained Transformer*

306 The GPT, developed by OpenAI [14], is a type of AI model used for NLP tasks. Following the
307 research path from the original GPT, GPT-2, and GPT-3, OpenAI's DL approach leverages
308 extensive amounts of data and intensive computation to create increasingly sophisticated and
309 capable language models [18]. ChatGPT has been fine-tuned from the initial GPT-3.5, and later,
310 GPT-4—a LLM trained in early 2022 to produce text [13, 30]. GPT-4 is OpenAI's latest and most
311 advanced AI model that can solve difficult problems with greater accuracy [18]. In this study, we
312 used ChatGPT powered by both the GPT-3.5 and GPT-4 versions.

313

314 *Prompt engineering to maximize the correct answer rate*

315 We used the 116th NMLE in Japan to generate the most suitable prompts for ChatGPT to answer
316 the 117th NMLE questions. Using the ChatGPT API, we first instructed ChatGPT to respond to the
317 original questions in Japanese language. We manually coded the Hyper Text Markup Language
318 (HTML) to represent the bold, italic, superscript, and subscript characters in the original text
319 (**Figure 3A**). Second, we instructed ChatGPT to translate the original Japanese NMLE questions
320 into English using its own capabilities before inputting them as questions (**Figure 3B**). In addition,
321 we compiled and analyzed the output errors. Thereafter, we provided prompts with restriction
322 sentences designed to prevent the reoccurrence of these errors, along with sample outputs
323 illustrating the desired output format. Finally, we inquired ChatGPT to improve the prompt itself.
324 We further refined the prompts using the 116th NMLE questions to achieve higher rates of correct
325 answers and output in the desired format, because prompt tuning can improve the task accuracy
326 compared to training the entire model [12, 20]. The final optimized two-step prompts for the
327 English translation process and the process of answering the medical questions are illustrated in
328 **Figure 3C**, wherein each process comprised "system," "sample output," and "question input"
329 sections. We organized the output examples according to each medical question category (basics
330 of medicine, clinical medicine, and comprehension). In brief, ChatGPT was initially instructed to
331 translate the HTML-based Japanese questions into plain, direct, and improved English, while
332 maintaining the original HTML codes without deleting or adding new text. In both processes, the

333 system of requirement and an exemplary output scenario were provided within the prompts. In the
334 question input section, the HTML-based Japanese questions were inputted for the English
335 translation process, and the English-translated questions were consequently inputted to the process
336 of answering the medical questions (**Figure 3**). To minimize output variability, all input prompts
337 were executed with the temperature parameter set to 0.

338 GPT-3.5-based analyses were performed using the ChatGPT API with custom Python code
339 on the Google Colaboratory interface. GPT-4-based analysis was conducted using ChatGPT
340 website console, with eight investigators (Y. T., T. N., K. A., T. E., R. M., S. K., H. K., and F. H.)
341 manually inputting prompts one by one and changing a thread each time. Specifically, they inputted
342 the questions, choices, and appropriate prompts into ChatGPT and summarized the output answers.
343 We used the GPT-3.5 version GPT3.5-turbo-0301 for the "Japanese," "English," and "English with
344 tuned prompt" analyses, and the GPT-4 model version released on March 14th 2023 for the
345 "English with tuned prompt" analysis.

346

347 *Outcomes*

348 The target outcome of this study is the correct answer rate. We manually compared ChatGPT's
349 output answers with the official answers to determine the correctness of the output answers.
350 Accordingly, the correct answer rate was calculated as the number of correct answers divided by
351 the number of questions. We defined the output errors as incorrect answers. To evaluate the
352 potential performance for passing the 117th NMLE in Japan, we applied the minimum passing rates,
353 not the minimum passing scores, to evaluate the model performance because the image-containing
354 questions were excluded from the analyses.

355

356 *Performance evaluation*

357 In the primary performance evaluation, we assessed the correct answer rate for questions without
358 images in the 117th NMLE in Japan using the best ChatGPT model (GPT-4) with tuned prompt,
359 which was compared to the actual minimally passing rate on the examination. In the secondary
360 performance evaluation, we examined the correct answer rate for all questions in the 117th exam
361 using the optimal model and prompts. In addition, the medical reasonableness of the generated
362 explanations for each answer was assessed by two independent clinical physicians (M.N. and
363 M.K.) and was double-checked by another independent clinical physician (A.N.). Furthermore, we

364 analyzed the content of the incorrect answers along with their explanations to identify the areas in
365 which the application of the current ChatGPT for medicine may be relatively weak.

366

367

368 **Acknowledgments**

369 We express our gratitude to Yasuhiro Onogi and Yuichi Miyamae at MICIN, Inc. for their insightful
370 online discussions regarding this project. We thank Dr. Hozumi for dedicating his time to discuss
371 this topic with us. We also thank ChatGPT (GPT-4) and Enago English proofreading service for
372 English proofreading.

373

374

375 **Data availability**

376 The ChatGPT APIs used in this study are accessible via GitHub
377 (https://github.com/yudaitanaka1026/ChatGPT_NMLE_Japan).

378

379

380 **Conflict of Interest**

381 The authors declare no conflicts of interest relevant to this article.

382

383

384 **Financial disclosure**

385 None.

386

387

388 **Author contributions**

389 **Conceptualization:** Yudai Tanaka, Takuto Nakata, Ko Aiga, Hiroaki Kakizaki, and Akihiro
390 Nomura.

391 **Data curation:** Yudai Tanaka, Takuto Nakata, Ko Aiga, Takahide Etani, Ryota Muramatsu, Shun
392 Katagiri, Hiroyuki Kawai, Fumiya Higashino, and Masahiro Enomoto.

393 **Formal analysis:** Yudai Tanaka.

394 **Methodology:** Yudai Tanaka, Takuto Nakata, Hiroaki Kakizaki, and Akihiro Nomura.

395 **Project administration:** Akihiro Nomura.

396 **Supervision:** Masayuki Takamura, Takashi Yoneda, and Hiroaki Kakizaki.

397 **Validation:** Masao Noda, Mitsuhiro Kometani, and Akihiro Nomura.

398 **Visualization:** Yudai Tanaka, Takuto Nakata, Ko Aiga, and Akihiro Nomura.

399 **Writing – original draft:** Yudai Tanaka, Takuto Nakata, Ko Aiga, and Akihiro Nomura.

400 **Writing – review and editing:** Yudai Tanaka, Takuto Nakata, Ko Aiga, Takahide Etani, Ryota

401 Muramatsu, Shun Katagiri, Hiroyuki Kawai, Fumiya Higashino, Masahiro Enomoto, Masao

402 Noda, Masayuki Takamura, Mitsuhiro Kometani, Takashi Yoneda, Hiroaki Kakizaki, and Akihiro

403 Nomura.

404 **References**

- 405 1. Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine,
406 2023. *N Engl J Med*. 2023;388(13):1201-8. doi: 10.1056/NEJMra2302038. PubMed PMID:
407 36988595.
- 408 2. Nomura A, Noguchi M, Kometani M, Furukawa K, Yoneda T. Artificial Intelligence in
409 Current Diabetes Management and Prediction. *Curr Diab Rep*. 2021;21(12):61. Epub 20211213.
410 doi: 10.1007/s11892-021-01423-2. PubMed PMID: 34902070; PubMed Central PMCID:
411 PMCPMC8668843.
- 412 3. Aramaki E, Wakamiya S, Yada S, Nakamura Y. Natural Language Processing: from
413 Bedside to Everywhere. *Yearb Med Inform*. 2022;31(1):243-53. Epub 20220602. doi: 10.1055/s-
414 0042-1742510. PubMed PMID: 35654422; PubMed Central PMCID: PMCPMC9719781.
- 415 4. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art,
416 current trends and challenges. *Multimed Tools Appl*. 2023;82(3):3713-44. Epub 20220714. doi:
417 10.1007/s11042-022-13428-4. PubMed PMID: 35855771; PubMed Central PMCID:
418 PMCPMC9281254.
- 419 5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all
420 you need. *Advances in neural information processing systems*. 2017;30.
- 421 6. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional
422 Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North
423 American Chapter of the Association for Computational Linguistics: Human Language
424 Technologies*. 2019;1:4171-86. doi: <https://doi.org/10.18653/v1/N19-1423>.
- 425 7. Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng H-T, et al.
426 LaMDA: Language Models for Dialog Applications2022 January 01, 2022:[arXiv:2201.08239
427 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2022arXiv220108239T>.
- 428 8. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling
429 Language Modeling with Pathways2022 April 01, 2022:[arXiv:2204.02311 p.]. Available from:
430 <https://ui.adsabs.harvard.edu/abs/2022arXiv220402311C>.
- 431 9. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. LLaMA:
432 Open and Efficient Foundation Language Models2023 February 01, 2023:[arXiv:2302.13971 p.].
433 Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv230213971T>.

- 434 10. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models
435 are few-shot learners. *Advances in neural information processing systems*. 2020;33:1877-901.
- 436 11. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language
437 models to follow instructions with human feedback. *Advances in Neural Information Processing*
438 *Systems*. 2022;35:27730-44.
- 439 12. OpenAI. GPT-4 Technical Report 2023 March 01, 2023:[arXiv:2303.08774 p.]. Available
440 from: <https://ui.adsabs.harvard.edu/abs/2023arXiv230308774O>.
- 441 13. OpenAI. Introducing ChatGPT 2022 [cited 2023 Apr 8]. Available from:
442 <https://openai.com/blog/chatgpt>.
- 443 14. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for
444 Medicine. *N Engl J Med*. 2023;388(13):1233-9. doi: 10.1056/NEJMsr2214184. PubMed PMID:
445 36988602.
- 446 15. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, et al. Performance
447 of ChatGPT on USMLE: Potential for AI-assisted medical education using large language
448 models. *PLOS Digit Health*. 2023;2(2):e0000198. Epub 20230209. doi:
449 10.1371/journal.pdig.0000198. PubMed PMID: 36812645; PubMed Central PMCID:
450 PMC9931230.
- 451 16. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does
452 ChatGPT Perform on the United States Medical Licensing Examination? The Implications of
453 Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*.
454 2023;9:e45312. Epub 20230208. doi: 10.2196/45312. PubMed PMID: 36753318; PubMed
455 Central PMCID: PMC9947764.
- 456 17. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of
457 medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ*
458 *Eval Health Prof*. 2023;20:1. Epub 20230111. doi: 10.3352/jeehp.2023.20.1. PubMed PMID:
459 36627845; PubMed Central PMCID: PMC9905868.
- 460 18. OpenAI. GPT-4 is OpenAI's most advanced system, producing safer and more useful
461 responses. 2023 [cited 2023 Apr 8]. Available from: <https://openai.com/product/gpt-4>.
- 462 19. Ministry of Health Labour and Welfare in Japan. Announcement of the results of the
463 117th National Medical Licensing Examination in Japan. 2023 [cited 2023 Apr 10]. Available
464 from: <https://www.mhlw.go.jp/general/sikaku/successlist/2023/siken01/about.html>.

- 465 20. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt
466 tuning. arXiv preprint arXiv:210408691. 2021.
- 467 21. Web Technology Surveys. Usage statistics of content languages for websites. Available
468 from: https://w3techs.com/technologies/overview/content_language.
- 469 22. Kaneda Y, Tanimoto T, Ozaki A, Sato T, Takahashi K. Can ChatGPT Pass the 2023
470 Japanese National Medical Licensing Examination? Preprintsorg. 2023;2023030191. doi:
471 <https://doi.org/10.20944/preprints202303.0191.v1>.
- 472 23. Lee K, Firat O, Agarwal A, Fannjiang C, Sussillo D. Hallucinations in neural machine
473 translation. 2018.
- 474 24. Shakarian P, Koyyalamudi A, Ngu N, Mareedu L. An Independent Evaluation of
475 ChatGPT on Mathematical Word Problems (MWP)2023 February 01, 2023:[arXiv:2302.13814
476 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv230213814S>.
- 477 25. OpenAI. Usage policies. 2023. Available from: [https://openai.com/policies/usage-](https://openai.com/policies/usage-policies)
478 [policies](https://openai.com/policies/usage-policies).
- 479 26. Ordish J. Large Language Models and software as a medical device.: Medicines and
480 Healthcare products Regulatory Agency (MHRA); 2023. Available from:
481 [https://medregs.blog.gov.uk/2023/03/03/large-language-models-and-software-as-a-medical-](https://medregs.blog.gov.uk/2023/03/03/large-language-models-and-software-as-a-medical-device/)
482 [device/](https://medregs.blog.gov.uk/2023/03/03/large-language-models-and-software-as-a-medical-device/).
- 483 27. Ministry of Health Labour and Welfare in Japan. Software as a Medical Device
484 (Japanese). 2023. Available from:
485 https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000179749_00004.html.
- 486 28. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation
487 models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65. Epub
488 20230412. doi: 10.1038/s41586-023-05881-4. PubMed PMID: 37045921.
- 489 29. Ministry of Health Labour and Welfare in Japan. Questions and answers of the 116th
490 National Medical Licensing Examination in Japan. 2022. Available from:
491 https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryuu/iryuu/topics/tp220421-01.html.
- 492 30. OpenAI. What is ChatGPT? 2023. Available from:
493 <https://help.openai.com/en/articles/6783457-what-is-chatgpt>.
- 494

495 **Figure legends**

496 **Figure 1. Study overview.**

497 Questions from the 116th NMLE in Japan were used as the prompt-tuning dataset and those from
498 117th NMLE were utilized as the performance-testing dataset after removing the image-based
499 questions. During the prompt tuning process, questions from the prompt-tuning dataset were
500 input into GPT-3.5-turbo and GPT-4, using simple prompts in both Japanese and English along
501 with tuned prompts in English. Subsequently, we evaluated the outputs from GPT-3.5-turbo and
502 GPT-4 with tuned prompts. After tuning the prompts, the ChatGPT (GPT-4) model optimized
503 with the tuned prompts was tested on the performance-testing dataset (117th NMLE).

504

505 **Figure 2. Variations in the rate of correct answers across languages, prompt tuning levels, 506 and GPT models.**

507 Translating the Japanese questions into English text improved the correct answer rate; however,
508 it increased the output error rate. Upon further tuning the prompts, the correct answer rate
509 improved and the output error decreased. Moreover, switching from the GPT-3.5 model to the
510 GPT-4 model enhanced the correct answer rate and almost eliminated errors.

511

512 **Figure 3. Examples of prompts for English translation and answering medical questions.**

513 **A:** A simple “Japanese prompt” used for answering Japanese questions.

514 **B:** Simple “English prompts” used for Japanese-to-English translation and answering translated
515 questions.

516 **C:** Our optimized “English with tuned prompts”.

517 The final optimized two-step prompts comprised a "system," "sample output," and "question
518 input" sections. ChatGPT was initially instructed to translate HTML-based Japanese questions
519 into simple, direct, and improved English. In both processes, the system of requirement and an
520 exemplary output scenario were provided within the prompts. In the question input section, the
521 Japanese questions were inputted to the English translation process, and sequentially, the
522 English-translated questions were used to obtain the answers of the 117th NMLE questions.

523

524 **Figure 4. Examples of potentially outdated or critically incorrect outputs from the model in 525 current medical contexts.**

526 **A:** A question on the primary treatment for hyperventilation syndrome in the emergency
527 department. The suggestion of paper bag method for raising the carbon-dioxide concentration in
528 the blood has been commonly used in the past, but it is not always the first choice, as it can
529 worsen symptoms in certain patients with secondary hyperventilation, *e.g.*, those with lung
530 diseases causing low blood oxygen levels. In such answers, it seems that outdated, traditional
531 information can prevail over the latest information, especially if it has been a standard practice
532 over a period and related information is widely available on the Internet.

533 **B:** A question on the initial outpatient treatment for a type-2 diabetes patient with poor control
534 and combined diabetic retinopathy and neuropathy. The long-term treatment goal for diabetes is
535 strict blood sugar control, but in this case, strict blood sugar control with sulfonylurea drugs
536 during the initial treatment may aggravate the risk of diabetic retinopathy, raising strong
537 concerns on ChatGPT's answer.

538

Tables

Table 1. Performance of optimal GPT-4 model with tuned prompt for the 117th NMLE in Japan.

	Essential			Basics and Clinical				
	Basics of medicine (essential)	Clinical medicine (essential)	Comprehension (essential)	Basics of medicine (general)	Basics of medicine (specifics)	Clinical medicine (general)	Clinical medicine (specifics)	Comprehension
No. of questions without image data	45	22	15	61	27	36	46	10
No. of correct answers	36	19	12	47	25	22	37	8
No. of output errors	1	0	0	0	1	0	0	0
No. of incorrect answers	8	3	3	14	1	14	9	2
Correct answer rate	80.0%	86.4%	80.0%	77.0%	92.6%	61.1%	80.4%	80.0%
Output error rate	2.2%	0.0%	0.0%	0.0%	3.7%	0.0%	0.0%	0.0%
Score weight	x1	x3		x1				
Total score (correct answer rate)	129/156 (82.7%)			139/180 (77.2%)				
Minimum passing rate	80.0%			74.6%				

Table 2. Summary of incorrect answers from the optimal model.

Total incorrect answer	N = 56
Insufficient medical knowledge	33 (58.9%)
Breast surgery	1
Dermatology	2
Emergency medicine	2
Endocrinology	6
Gastroenterology	2
Immunology	1
Medical interview	1
Medical procedure	1
Nephrology	2
Neurology	1
Obstetrics and gynecology	2
Ophthalmology	1
Pediatrics	2
Physical examination	1
Psychiatry	1
Public health	1
Rehabilitation	1
Respiratory medicine	3
Rheumatology	1
Urology	1
Japan-specific medical system	17 (30.4%)
Clinical research	1
Emergency	1
Psychiatry	1
Public health	14
Mathematical errors	4 (7.1%)
Respiratory	1
Pediatrics	1
Cardiology	1
Medical interview	1
Others	2 (3.6%)
Issue in English translation	1
Not providing an answer	1

Figures
Figure 1.

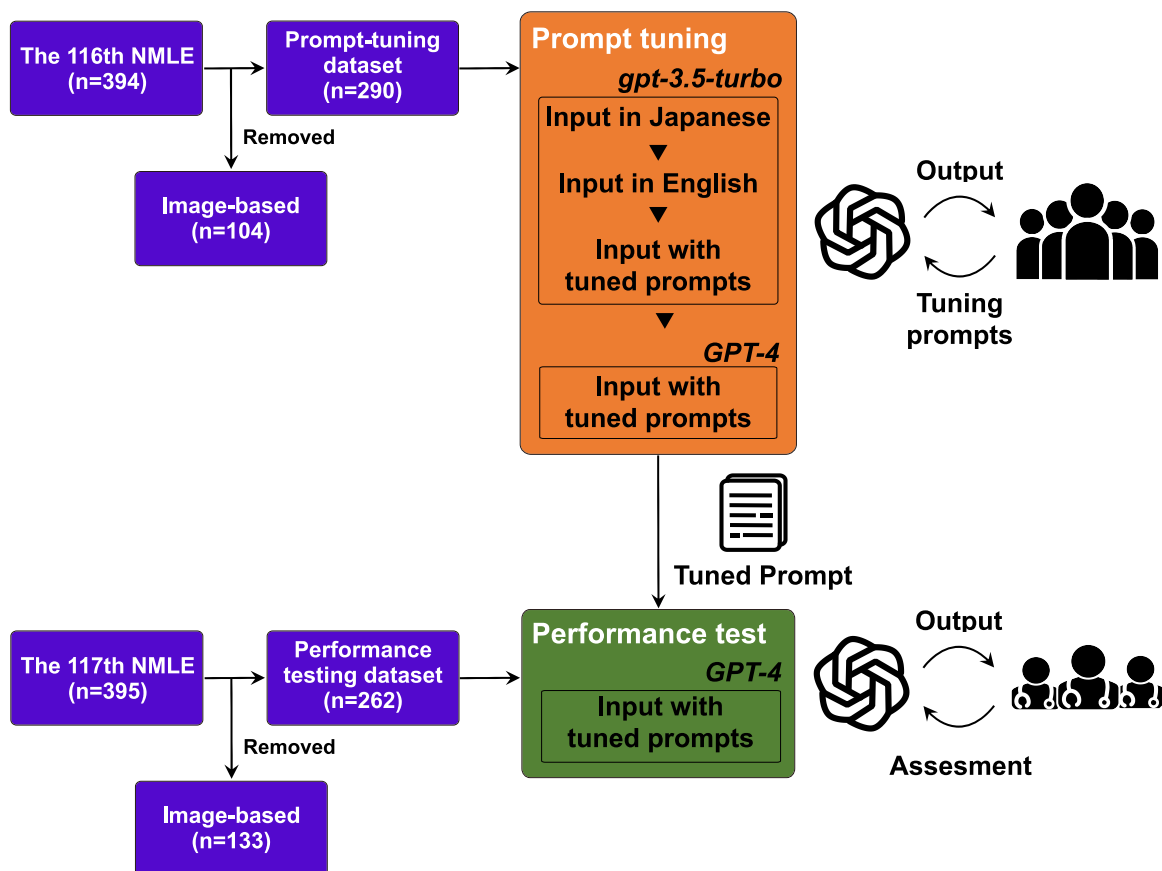


Figure 2.

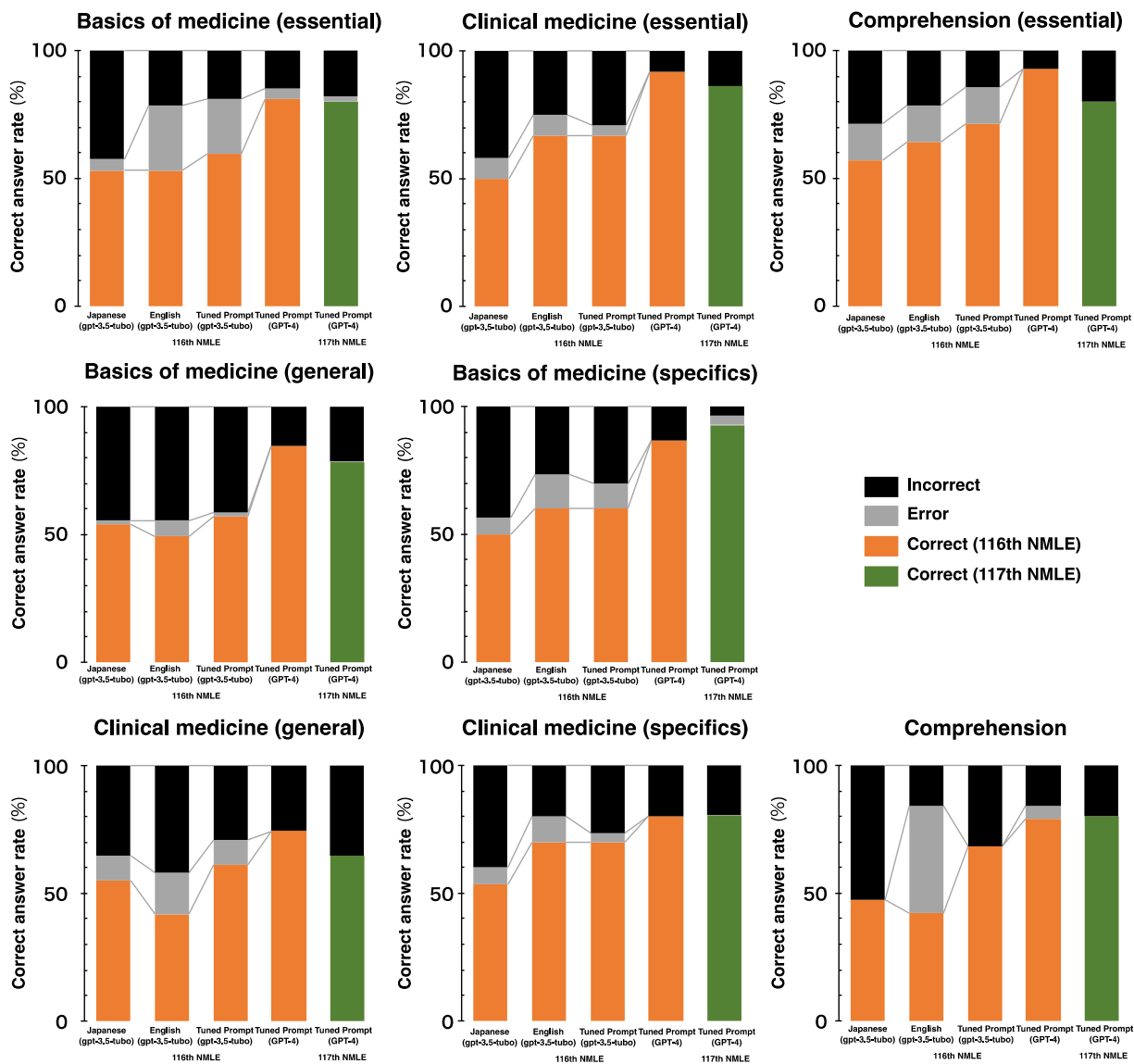


Figure 3.

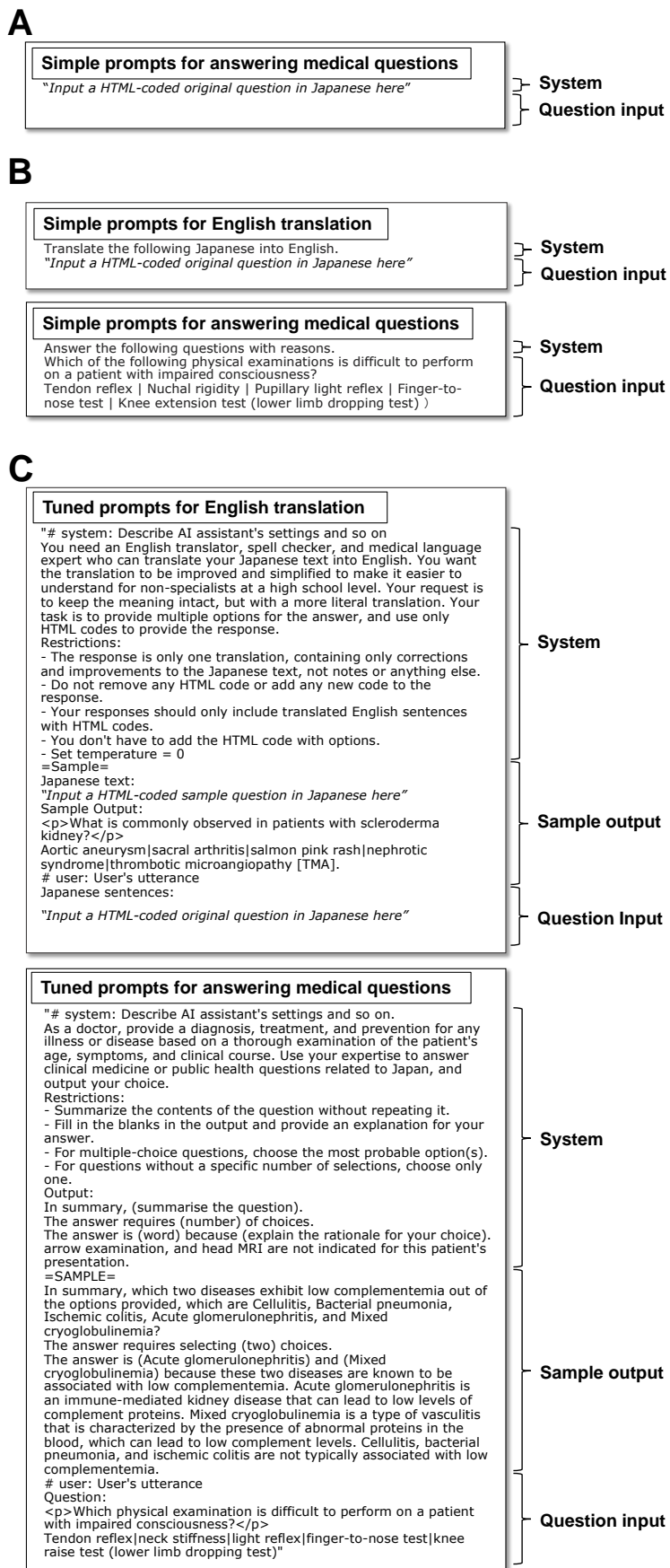


Figure 4.

A (117B37)

Model: GPT-4

Y # system: Describe AI assistant's settings and so on
As a doctor, provide a diagnosis, treatment, and prevention for any illness or disease based on a thorough examination of the patient's age, symptoms, and clinical course. Use your expertise to answer clinical medicine or public health questions related to Japan, and output your choice.
Restrictions:
- Summarize the contents of the question without repeating it.
- Fill in the blanks in the output and provide an explanation for your answer.
- For multiple-choice questions, choose the most probable option(s).
- For questions without a specific number of selections, choose only one.
- Set temperature = 0
Output:
In summary, (summarise the question).
The answer requires (number) of choices.
The answer is (word) because (explain the rationale for your choice).
=SAMPLE=
In summary, a 28-year-old woman with a history of spontaneous abortions presents with swelling and pain in her left calf and shortness of breath during exertion. Her vital signs and physical examination are provided, and her blood tests show low platelet count and high D-dimer levels. Which two tests should be performed from the options of blood culture, bone marrow examination, head MRI, chest CT scan, and lower extremity venous ultrasound?
The answer requires selecting (two) choices.
The answer is (chest CT scan) and (lower extremity venous ultrasound) because the patient's symptoms and history raise suspicion for deep vein thrombosis (DVT) or pulmonary embolism (PE), which can cause calf swelling and pain, shortness of breath, and an accentuated second heart sound. The low oxygen saturation level can also be a sign of PE. The elevated D-dimer level further supports the possibility of thrombosis, but it is not specific and can be elevated in other conditions. Chest CT scan is the most appropriate imaging test for diagnosing PE, while lower extremity venous ultrasound is the preferred test for diagnosing DVT. Therefore, both tests should be performed to rule out both conditions. Blood culture, bone marrow examination, and head MRI are not indicated for this patient's presentation.

user: User's utterance
Question:
<p>An 18-year-old female visited the emergency department with the main complaint of difficulty breathing. On her way to school, she suddenly felt short of breath and her breathing became labored while in a crowded bus. She is receiving treatment for panic disorder at a nearby clinic, but has no other underlying diseases. She is conscious and alert. Her body temperature is 36.8°C. Her pulse is 104 beats per minute and regular. Her blood pressure is 112/72mmHg. Her respiratory rate is 48 breaths per minute. Her SpO₂ is 100% (room air). There is no distention of her jugular veins. No abnormalities are noted in her heart and breath sounds. Her abdomen is flat and soft, with no tenderness on palpation. There is no cyanosis in her extremities. The patient says, "I can't breathe. My fingertips and the area around my mouth are numb."
</p><p>Which of the following actions is appropriate to perform in the emergency department for this patient?</p><p>Administer antipsychotic medication. | Let her go to school as is. | Allow her to inhale low-flow oxygen. | Have her breathe into a paper bag. | Instruct her to breathe slowly.*</p>

Y In summary, an 18-year-old female with a history of panic disorder presents with difficulty breathing, numbness around her mouth and fingertips, and no other significant findings on examination.
The answer requires selecting one choice.
The answer is (have her breathe into a paper bag) because her symptoms are consistent with hyperventilation, which can occur during a panic attack. Hyperventilation causes a decrease in the carbon dioxide (CO2) levels in the blood, leading to numbness and tingling around the mouth and fingertips. Breathing into a paper bag helps the patient re-breathe their exhaled CO2, which can help alleviate these symptoms. Other options, such as administering antipsychotic medication, are not appropriate for treating acute hyperventilation. Letting her go to school as is and allowing her to inhale low-flow oxygen would not address the underlying issue, and instructing her to breathe slowly may not be sufficient to correct her CO2 levels quickly.

B (117E44)

Model: GPT-4

Y # system: Describe AI assistant's settings and so on
As a doctor, provide a diagnosis, treatment, and prevention for any illness or disease based on a thorough examination of the patient's age, symptoms, and clinical course. Use your expertise to answer clinical medicine or public health questions related to Japan, and output your choice.
Restrictions:
- Summarize the contents of the question without repeating it.
- Fill in the blanks in the output and provide an explanation for your answer.
- For multiple-choice questions, choose the most probable option(s).
- For questions without a specific number of selections, choose only one.
- Set temperature = 0
Output:
In summary, (summarise the question).
The answer requires (number) of choices.
The answer is (word) because (explain the rationale for your choice).
=SAMPLE=
In summary, (the patient is a 43-year-old man with second to third degree burns covering 38% of his total body surface area on his face, both upper limbs, and chest and abdomen. He has airway thermal injury and underwent tracheal intubation. He is currently on a dietary therapy for hypertension and dyslipidemia, and has a history of smoking and drinking. On admission, he has a low blood pressure, high respiratory rate, and reduced SpO2 with burn injuries and soot in his mouth and throat.)
Q1:
(One) answer is required.
The answer is (Lactated Ringer's solution).
The reasons are that (Lactated Ringer's solution is the most appropriate initial fluid resuscitation for burn patients due to its ability to restore electrolyte balance and promote urine output. It also has a similar electrolyte composition to plasma, making it ideal for patients with burn injuries. Other options, such as saline solution or glucose solution, may exacerbate the electrolyte imbalances that can occur with burn injuries.)
Q2:
(One) answer is required.
The answer is (Breath sounds are heard over the epigastric area).
The reason is that (breath sounds should be heard over the lung fields, not the epigastric area, indicating incorrect placement of the endotracheal tube in the esophagus instead of the trachea. The presence of CO2 in exhalation, symmetrical chest movements, equal breath sounds over both lung fields, and fogging of the endotracheal tube wall are all signs of correct endotracheal tube placement.)

user: User's utterance
Question:
Q1:
<p>Please read the following text and answer questions 43 and 44.</p><p>A 68-year-old man visited the clinic with the main complaints of blurred vision and discomfort in his feet.</p><p>Present illness:</p><p>He was informed of glycosuria during a health check-up 20 years ago. He received lifestyle guidance at a medical institution, but stopped visiting due to moving. He has not had a health check-up since retiring 10 years ago. He has been aware of a tingling sensation in both feet for about 2 years. He noticed a decline in his vision six months ago, but attributed it to aging. Blurred vision in his right eye appeared three days ago.</p><p>Medical history:</p><p>Appendicitis at 18 years old. No history of blood transfusion.</p><p>Personal history:</p><p>Living alone since the age of 60. He spends most of the day at home. He eats convenience store boxed meals or side dishes once or twice a day. He has been smoking 15 cigarettes a day for 48 years. He drinks 150mL of beer per day or about 180mL of shochu per day, 5 or 6 times a week.</p><p>Family history:</p><p>His father died at 72 due to a stroke. His mother died at 88 due to senility.</p><p>Current conditions:</p><p>Height 170 cm, weight 72 kg, waist circumference 86 cm. Blood pressure 128/72 mmHg. No abnormalities found in the chest or abdomen. Soft white spots and neovascularization were found in both eyes, with vitreous hemorrhage in the right eye.</p><p>Examination results:</p><p>Urinalysis: Protein (trace), glucose 3+, ketones (-), occult blood (-), and no abnormalities in sediment. Blood test results: Red blood cell count of 4.4 million, Hb 12.8 g/dL, Ht 43%, white blood cell count of 6,000 (neutrophils 54%, eosinophils 2%, basophils 0%, monocytes 8%, lymphocytes 36%), platelet count of 190,000. Blood chemistry test results: Total protein 8.9 g/dL, albumin 2.6 g/dL, direct bilirubin 0.3 mg/dL, AST 22 U/L, ALT 19 U/L, LD 186 U/L (normal range: 120-245), γ-GT 17 U/L (normal range: 8-50), amylase 152 U/L (normal range: 37-180), CK 192 U/L (normal range: 30-140), blood urea nitrogen 20 mg/dL, creatinine 0.8 mg/dL, eGFR 72.8 mL/min/1.73m²/1.73m², uric acid 4.0 mg/dL, blood glucose 235 mg/dL, HbA1c 8.9% (normal range: 4.6-6.2), total cholesterol 247 mg/dL, triglycerides 64 mg/dL, HDL cholesterol 51 mg/dL, Na 140 mEq/L, K 4.4 mEq/L, Cl 106 mEq/L, Ca 91 mg/dL, P 3.0 mg/dL, TSH 3.0 μU/mL (normal range: 0.2-4.0), FT₄ 1.2 ng/dL (normal range: 0.8-2.2).</p><p>When examining this patient's feet, which of the following items has the lowest priority?</p><p>Achilles tendon reflex | Dorsalis pedis artery pulse | Presence of skin lesions | Vibration sense at the medial malleolus | Presence of flat feet</p>

Q2:
<p>Which of the following is the correct approach?</p><p>Protein-restricted diet | Intensive exercise therapy | 1,200 kcal per day dietary therapy | Strict blood sugar control with sulfonylurea drugs | Setting blood sugar control goals based on psychosomatic function evaluation.*</p>

Y In summary, a 68-year-old man presents with blurred vision and discomfort in his feet. He has a history of glycosuria and a sedentary lifestyle with poor diet choices. His examination results show high blood glucose and HbA1c levels, among other findings.
Q1:
One answer is required.
The answer is Presence of flat feet because the patient's symptoms and examination results suggest possible diabetic complications such as neuropathy and retinopathy, making it more important to assess the Achilles tendon reflex, dorsalis pedis artery pulse, presence of skin lesions, and vibration sense at the medial malleolus, which are more relevant to diabetic foot assessment.
Q2:
One answer is required.
The answer is Strict blood sugar control with sulfonylurea drugs because the patient has poorly controlled diabetes with an HbA1c of 8.9% and blood glucose of 235 mg/dL. The priority should be to control blood sugar levels to prevent further complications. While lifestyle modifications and exercise are important, they alone may not be sufficient to achieve the necessary blood sugar control in this case.