

1 **Large language models to differentiate vasospastic angina using patient information**

2

3

4 Yuko Kiyohara¹, Satoshi Kodera¹, Masaya Sato², Kota Ninomiya³, Masataka Sato¹, Hiroki
5 Shinohara¹, Norifumi Takeda¹, Hiroshi Akazawa¹, Hiroyuki Morita¹, Issei Komuro¹

6

7

8 ¹ Department of Cardiovascular Medicine, Graduate School of Medicine, The University of
9 Tokyo, Tokyo, Japan

10 ² Department of Clinical Laboratory Medicine, Graduate School of Medicine, The University
11 of Tokyo, Tokyo, Japan.

12 ³ Department of Genome Analysis for Rare and Intractable Diseases, Tohoku University
13 Graduate School of Medicine, Miyagi, Japan.

14

15

16 * Corresponding author

17 E-mail: koderasatoshi@gmail.com (SK)

18 **Abstract**

19 **Background**

20 Vasospastic angina is sometimes suspected from patients' medical history. It is essential to
21 appropriately distinguish vasospastic angina from acute coronary syndrome because its
22 standard treatment is pharmacotherapy, not catheter intervention. Large language models
23 have recently been developed and are currently widely accessible. In this study, we aimed to
24 use large language models to distinguish between vasospastic angina and acute coronary
25 syndrome from patient information and compare the accuracies of these models.

26

27 **Method**

28 We searched for cases of vasospastic angina and acute coronary syndrome which were
29 written in Japanese and published in online-accessible abstracts and journals, and randomly
30 selected 66 cases as a test dataset. In addition, we selected another ten cases as data for few-
31 shot learning. We used generative pre-trained transformer-3.5 and 4, and Bard, with zero- and
32 few-shot learning. We evaluated the accuracies of the models using the test dataset.

33

34 **Results**

35 Generative pre-trained transformer-3.5 with zero-shot learning achieved an accuracy of 52%,
36 sensitivity of 68%, and specificity of 29%; with few-shot learning, it achieved an accuracy of
37 52%, sensitivity of 26%, and specificity of 86%. Generative pre-trained transformer-4 with
38 zero-shot learning achieved an accuracy of 58%, sensitivity of 29%, and specificity of 96%;
39 with few-shot learning, it achieved an accuracy of 61%, sensitivity of 63%, and specificity of
40 57%. Bard with zero-shot learning achieved an accuracy of 47%, sensitivity of 16%, and

41 specificity of 89%; with few-shot learning, this model could not be assessed because it failed
42 to produce output.

43

44 **Conclusion**

45 Generative pre-trained transformer-4 with few-shot learning was the best of all the models.

46 The accuracies of models with zero- and few-shot learning were almost the same. In the

47 future, models could be made more accurate by combining text data with other modalities.

48 **Introduction**

49 Vasospastic angina (VSA) is characterized by transient constriction of the coronary artery,
50 causing angina and other nonspecific symptoms [1–3]. VSA is known to be caused by
51 various factors, including emotional stress, exposure to low temperatures, and resting during
52 the period from midnight to early morning [3–7]. Conversely, acute coronary syndrome
53 (ACS), which is mostly caused by diseases that differ from VSA, frequently presents with
54 chest pain during physical exertion following a prolonged history of smoking, dyslipidemia,
55 hypertension, or diabetes mellitus [8–10]. Because VSA shares some symptoms with ACS,
56 which requires urgent catheter intervention treatment, distinguishing between VSA and ACS
57 is often challenging at the screening stage [3,4,8–11]. In cases where it is difficult to
58 distinguish between the two, VSA is confirmed by coronary spasm provocation testing, as
59 described in the guideline and international consensus [12,13]. However, provocation testing
60 is an invasive procedure that carries the risk of serious complications [14]. Therefore, there is
61 a need for a noninvasive method of screening for VSA in patients with angina.

62 In recent years, large language models (LLMs) have demonstrated great potential in a broad
63 range of fields including medicine [15–17]. In the medical field, generative pre-trained
64 transformer 4 (GPT-4) has achieved high scores in medical licensing examinations in the
65 United States and Japan [18–20]. In real-world clinical settings, physicians rely on patient
66 information gathered through interviews and examinations. However, to the best of our
67 knowledge, there has been no research on whether LLMs can effectively use patient
68 information to distinguish between specific diseases. Furthermore, although GPT-4 has
69 multilingual capabilities, much of the research conducted using this model has focused on
70 English documents [19,21]. In this study, we aimed to use LLMs to distinguish between VSA
71 and ACS from patient information written in Japanese and compare the accuracies of the
72 models.

73 **Materials and methods**

74 An overview of this study is shown in Fig 1.

75

76 **Fig 1. The flow of this study.**

77 The overview of this study is shown.

78

79 **Case selection**

80 In a previous study, we continuously collected cases of VSA and ACS published in abstracts
81 of the Japanese Society of Internal Medicine regional conferences and open-access journals.

82 In the present study, we randomly selected 66 of these cases (38 VSA and 28 ACS cases) as a
83 test dataset [22]. We defined ACS to include both unstable angina and acute myocardial

84 infarction. In addition, we selected another ten cases (five VSA and five ACS cases) from the
85 previously collected cases as data for few-shot learning. Because the data used in this study

86 were publicly accessible, there were no ethical concerns. Nonetheless, we handled cases

87 following the principles outlined in the Declaration of Helsinki.

88

89 **Data extraction**

90 We extracted data comprising age, sex, medical history, past medical history, and medication
91 from the selected abstracts and case reports, and organized them accordingly (S1 Fig).

92 Medication doses were not extracted, and the English notation of medication was translated
93 to Japanese and standardized.

94

95 **GPT-3.5 and GPT-4**

96 We used ChatGPT Plus (OpenAI), which is based on generative pre-trained transformer-3.5
97 (GPT-3.5) and GPT-4 [20]. In addition, we adopted the zero- and few-shot learning
98 approaches. To select appropriate cases to include in the data for few-shot learning, YK and
99 SK discussed and identified typical cases of VSA and ACS. We selected ten representative
100 cases from the data that were left behind after selecting the test dataset, and these cases were
101 used as data for few-shot learning.

102 At the beginning of the prompts, we asked GPT-3.5 and GPT-4 to answer whether each case
103 was predicted to be VSA or any other coronary artery disease. Technically, ACS can be
104 caused also by VSA, although rare, and therefore, we used the phrase of “any other coronary
105 artery disease” rather than ACS. In the zero-shot learning tests, we input the cases from the
106 test dataset just after the above question. In the few-shot learning tests, we inserted the set
107 comprising each case and its correct answer, for the ten cases of the learning data (S2 Fig).
108 We performed the experiment from May 10 to May 20, 2023. We used a new session for
109 each case by clearing all conversations before inputting any prompts.

110

111 **Bard**

112 We used Bard (Google) because it was announced on May 11, 2023 that Bard could respond
113 to Japanese text. We input the same request as we input to GPT-3.5 and GPT-4. In the zero-
114 shot learning tests, we input the cases in the test dataset following the request sentence. In the
115 few-shot learning tests, we input the same set of ten cases (with their correct answers) as we
116 input to GPT-3.5 and GPT-4. We performed the experiment from May 22 to May 24, 2023.
117 We reset the conversations every time we input a prompt.

118

119 **Evaluation of model accuracy**

120 In each of the three groups of experiments (GPT-3.5, GPT-4, and Bard), we compared their
121 answers with the correct answers, and calculated accuracy, sensitivity, specificity, precision,
122 and F-score. We compared the accuracies achieved by the three LLMs. The threshold for
123 statistical significance was set at 0.05. The accuracy of each model was calculated with a
124 95% confidence interval (CI) using the binomial test, implemented in R (R Foundation for
125 Statistical Computing, Vienna, Austria).

126

127 **Comparison with cardiologists' accuracy and accuracy of each** 128 **model**

129 In the previous study, we reported the accuracy with which cardiologists answered the cases
130 in the test dataset [22]. In this study, for each case, we evaluated the answer given by each
131 model with reference to those given by the cardiologists.

132

133 **Sensitivity analysis**

134 As a sensitivity analysis, we selected the subset of the test dataset comprising the cases for
135 which more than half of the cardiologists answered correctly in the previous study [22]. In 45
136 out of 66 cases, more than 50% of the cardiologists answered correctly. We hypothesized that
137 these cases contain enough information to distinguish between VSA and ACS. Therefore, in
138 the sensitivity analysis, we used these 45 cases only.

139

140 Results

141 Accuracy evaluation of GPT-3.5, GPT-4, and Bard

142 Table 1 shows the accuracy achieved in the three groups of experiments. For GPT-3.5 with
143 zero-shot learning, the accuracy was 52% (95% CI: 39–64%), sensitivity was 68%,
144 specificity was 29%, precision was 57%, and F-score was 62%; with few-shot learning, the
145 accuracy was 52% (95% CI: 39–64%), sensitivity was 26%, specificity was 86%, precision
146 was 71%, and F-score was 39%. For GPT-4 with zero-shot learning, the accuracy was 58%
147 (95% CI: 45–70%), sensitivity was 29%, specificity was 96%, precision was 92%, and F-
148 score was 44%; with few-shot learning, the accuracy was 61% (95% CI: 48–72%), sensitivity
149 was 63%, specificity was 57%, precision was 67%, and F-score was 65%. For Bard with
150 zero-shot learning, the accuracy was 47% (95% CI: 35–60%), sensitivity was 16%,
151 specificity was 89%, precision was 67%, and F-score was 26%; with few-shot learning, Bard
152 failed to respond to the input data.

153

154 **Table 1. The main results of accuracies of the models.**

	GPT-3.5/zero	GPT-3.5/few	GPT-4/zero	GPT-4/few	Bard/zero	Bard/few
Accuracy	51.5 (95%CI: 38.9 - 64.0)	51.5 (95%CI: 38.9 - 64.0)	57.6 (95%CI: 44.8 - 69.7)	60.6 (95%CI: 47.8 - 72.4)	47.0 (95%CI: 34.6 - 59.7)	NA
Sensitivity	68.4	26.3	28.9	63.2	15.8	NA
Spesificity	28.6	85.7	96.4	57.1	89.3	NA
Precision	56.5	71.4	91.7	66.7	66.7	NA
F-score	61.9	38.5	44.0	64.9	25.5	NA

155

156 The accuracy achieved in each of the three AI models (GPT-3.5, GPT-4, and Bard) with zero-
157 and few-shot learning is shown.

158

159 **Comparison with cardiologists' accuracy and result of each model**

160 Fig 2 shows the results of GPT-4 with few-shot learning with reference to the cardiologists'
161 accuracy (the proportion of cardiologists who answered correctly). GPT-4 tended to correctly
162 answer cases for which the cardiologists' accuracy was high. This tendency was observed for
163 the other models except for GPT-3.5 with zero-shot learning (S3–6 Figs).

164

165 **Fig 2. Comparison with cardiologists' accuracy and GPT-4 with few-shot learning.**

166 The result of GPT-4 with few-shot learning is shown with reference to the proportion of
167 cardiologists who answered correctly.

168

169 **Sensitivity analysis**

170 In the sensitivity analysis, the results were almost the same as the main results (Table 2).
171 GPT-4 yielded the highest accuracy of the three models. In GPT-4 with zero-shot learning,
172 the accuracy was 76%, sensitivity was 50%, specificity was 100%, precision was 100%, and
173 F-score was 67%; with few-shot learning, the accuracy was 71%, sensitivity was 77%,
174 specificity was 65%, precision was 68%, and F-score was 72%.

175

176 **Table 2. The results of the sensitivity analysis.**

	GPT-3.5/zero	GPT-3.5/few	GPT-4/zero	GPT-4/few	Bard/zero	Bard/few
Accuracy	48.9 (95%CI: 33.7 - 64.2)	60.0 (95%CI: 44.3 - 74.3)	75.6 (95%CI: 60.5 - 87.1)	71.1 (95%CI: 55.7 - 83.6)	57.8 (95%CI: 42.2 - 72.3)	NA
Sensitivity	68.2	31.8	50.0	77.3	27.3	NA
Specificity	30.4	87.0	100.0	65.2	87.0	NA
Precision	48.4	70.0	100.0	68.0	66.7	NA
F-score	56.6	43.8	66.7	72.3	38.7	NA

177

178 The results of experiments for the subset of the test dataset for which more than half of the
179 cardiologists answered correctly are shown.

180

181 **Discussion**

182 In this study, we used GPT-3.5, GPT-4, and Bard with zero- and few-shot learning for the
183 purpose of distinguishing between VSA and ACS from patient information. The results of the
184 study can be summarized as follows: 1) GPT-4 with few-shot learning yielded the most
185 accurate results of the three LLMs, 2) there were no significant differences in accuracy
186 between zero- and few-shot learning, and 3) our study was unique in that it processed data in
187 the form of Japanese text instead of English text.

188 GPT-4 was able to distinguish VSA and ACS from patient information and its accuracy was
189 almost the same as that of medical students. In a previous study, we compared a variant
190 model of BERT (Google) with cardiologists and medical students [22,23]. The test data were
191 the same as those used in this study. For cardiologists, the accuracy was 68%, sensitivity was
192 58%, and specificity was 82%, whereas for medical students, the accuracy was 61%,
193 sensitivity was 40%, and specificity was 89% [22]. In the current study, GPT-4 with few-shot
194 learning achieved almost the same accuracy as medical students. Because we collected cases
195 from conference abstracts of oral presentations and peer-reviewed articles, which ensured
196 linguistic consistency and no unique abbreviations or variations in sentences, we certainly
197 used high-quality data. However, the performance of GPT-4 in distinguishing VSA from
198 ACS did not match that of cardiologists. This implies that the advantages of enhancing the
199 quality of data are limited. Improvements to the methodology of models are required to
200 obtain further improvements in performance.

201 Surprisingly, we found no significant differences in accuracy between zero- and few-shot
202 learning. Previously, few-shot learning was reported to increase the accuracy of LLMs [24].
203 However, although we used GPT-3.5 and GPT-4 with both zero- and few-shot learning, there
204 were almost no differences in accuracy between the two methods. This suggests that few-shot
205 learning may not necessarily improve accuracy. Fine-tuning is another promising method for
206 enhancing the accuracy of models. Fine-tuning involves adjusting the parameters and
207 architecture of models, using training data to adapt them to specific tasks [25,26]. Initially,
208 we attempted to conduct fine-tuning using the remaining data for the purpose of tailoring the
209 models to our specific tasks. However, applying fine-tuning to GPT-4 was impossible as of
210 May 20, 2023. In the future, fine-tuning could potentially improve the accuracy of GPT-4
211 even further.

212 Our study has the potential to contribute to the advancement of natural language processing
213 of Japanese medical data. A comprehensive review showed that almost 90% of the
214 publications about clinical natural language processing addressed the processing of English
215 text data, whereas only 0.62% were on Japanese text data [27]. The multilingual capability of
216 GPT-4 will facilitate research on Japanese text. Much more research specific to the Japanese
217 population is required because there are some diseases, such as VSA, which are more
218 common in Japan than elsewhere.

219 There is scope for improvement in the accuracy of artificial intelligence (AI) models and their
220 applications to other diseases. It is expected that the accuracy of AI models can be improved
221 by combining other types of data with text data. In a previous study, an AI model was
222 developed that used numerical and image data, including laboratory data and resting 12-lead
223 electrocardiograms, to determine the presence or absence of coronary abnormalities in
224 patients with chest pain [28]. “Generalist” medical AI, capable of flexibly incorporating data
225 modalities including text data, numerical data (such as examination findings), and image data

226 (such as electrocardiograms), could potentially be useful in clinical settings [29].

227 Furthermore, the methodology of this study can be applied to diseases other than VSA,

228 thereby facilitating the development of AI models to distinguish between various diseases.

229 There are some limitations to this study. First, the number of conference abstracts and

230 published papers collected in the study was limited, and it is difficult to generalize its results.

231 One possible solution to the scarcity of data is the extraction of data from electronic health

232 records [30]. Although electronic health records contain patient information, the quality of

233 this information may be inadequate. Electronic health records are often written by busy

234 physicians and nurses; they therefore contain many abbreviations and incorrect or vague

235 expressions, and often ignore grammar [31,32]. Therefore, the quality of electronic health

236 record data is considered to be lower than that of the abstracts and reports used in this study.

237 Second, regarding the content of the data, the fact that the cardiologists' accuracy was only

238 around 70% suggests that the data used in this study may not have contained sufficient

239 information for distinguishing VSA. That said, in actual clinical practice, it is often difficult

240 to determine VSA solely from patients' medical history; therefore, the data used in this study

241 could not be necessarily reflective of real-world clinical practice. Third, we asked the models

242 to answer whether each case was VSA or any other coronary artery disease. We believe that

243 this prompt was very effective in distinguishing between VSA and ACS. However, we would

244 like to further improve models by directly comparing VSA and ACS. Fourth, in the data

245 collection process, we defined the ACS group as cases described as "unstable angina" or

246 "acute myocardial infarction." Cases described using synonyms or different expressions, such

247 as "acute coronary syndrome" or "ACS," may have been omitted, possibly limiting the

248 quantity of data available for training the AI model and evaluating its accuracy. Future

249 research should aim to collect cases comprehensively, including those described using

250 synonyms and different expressions.

251 In conclusion, we used GPT-3.5, GPT-4, and Bard to distinguish between VSA and ACS
252 from patient information. The predictive accuracy of GPT-4 with few-shot learning was the
253 best of the three models. In the future, by creating multimodal AI models that combine text
254 data with numerical data, image data, and other types of data, it is expected that the accuracy
255 will be further improved.

256

257 **Acknowledgments**

258 We employed GPT-4 to generate initial drafts, which were then critically reviewed. We take
259 full responsibility for the content presented in this article. We thank Edanz
260 (<https://jp.edanz.com/ac>) for editing a draft of this manuscript.

261

262 **References**

- 263 1. Stern S, DeLuna AB. Coronary artery spasm: a 2009 update. *Circulation*. 2009;119:
264 2531–2534. <https://doi.org/10.1161/CIRCULATIONAHA.108.843474> PMID:
265 19433770
- 266 2. Lanza GA, Careri G, Crea F. Mechanisms of coronary artery spasm. *Circulation*.
267 2011;124: 1774–1782. <https://doi.org/10.1161/CIRCULATIONAHA.111.037283>
268 PMID: 22007100
- 269 3. JCS Joint Working Group. Guidelines for diagnosis and treatment of patients with
270 vasospastic angina (Coronary Spastic Angina) (JCS 2013). *Circ J*. 2014;78: 2779–
271 2801. <https://doi.org/10.1253/CIRCJ.CJ-66-0098> PMID: 25273915

- 272 4. Yasue H, Nakagawa H, Itoh T, Harada E, Mizuno Y. Coronary artery spasm--clinical
273 features, diagnosis, pathogenesis, and treatment. *J Cardiol*. 2008;51: 2–17.
274 <https://doi.org/10.1016/J.JJCC.2008.01.001> PMID: 18522770
- 275 5. Yeung AC, Vekshtein VI, Krantz DS, Vita JA, Ryan TJ, Ganz P, et al. The effect of
276 atherosclerosis on the vasomotor response of coronary arteries to mental stress. *N Engl*
277 *J Med*. 1991;325: 1551–1556. <https://doi.org/10.1056/NEJM199111283252205> PMID:
278 1944439
- 279 6. Raizner AE, Chahine RA, Ishimori T, Verani MS, Zacca N, Jamal N, et al.
280 Provocation of coronary artery spasm by the cold pressor test. Hemodynamic,
281 arteriographic and quantitative angiographic observations. *Circulation*. 1980;62: 925–
282 932. <https://doi.org/10.1161/01.CIR.62.5.925> PMID: 7418176
- 283 7. Kugiyama K, Yasue H. Coronary-artery spasm. *N Engl J Med*. 1978;299: 617–620.
284 <https://doi.org/10.1056/NEJM197809282991305>
- 285 8. Writing Committee Members, Gulati M, Levy PD, Mukherjee D, Amsterdam E, Bhatt
286 DL, et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR Guideline for the
287 Evaluation and Diagnosis of Chest Pain: A Report of the American College of
288 Cardiology/American Heart Association Joint Committee on Clinical Practice
289 Guidelines. *J Am Coll Cardiol*. 2021;78: e187–e285.
290 <https://doi.org/10.1016/j.jacc.2021.07.053> PMID: 34756653
- 291 9. Collet JP, Thiele H, Barbato E, Bauersachs J, Dendale P, Edvardsen T, et al. 2020 ESC
292 Guidelines for the management of acute coronary syndromes in patients presenting
293 without persistent ST-segment elevation. *Eur Heart J*. 2021;42: 1289–1367.
294 <https://doi.org/10.1093/EURHEARTJ/EHAA575> PMID: 32860058

- 295 10. Kimura K, Kimura T, Ishihara M, Nakagawa Y, Nakao K, Miyauchi K, et al. JCS 2018
296 Guideline on Diagnosis and Treatment of Acute Coronary Syndrome. *Circ J.* 2019;83:
297 1085–1196. <https://doi.org/10.1253/CIRCJ.CJ-19-0133> PMID: 30930428
- 298 11. Bhatt DL, Lopes RD, Harrington RA. Diagnosis and Treatment of Acute Coronary
299 Syndromes: A Review. *JAMA.* 2022;327: 662–675.
300 <https://doi.org/10.1001/JAMA.2022.0358> PMID: 35166796
- 301 12. Yamagishi M, Tamaki N, Akasaka T, Ikeda T, Ueshima K, Uemura S, et al. JCS 2018
302 Guideline on Diagnosis of Chronic Coronary Heart Diseases. *Circ J.* 2021;85: 402–
303 572. <https://doi.org/10.1253/CIRCJ.CJ-19-1131> PMID: 33597320
- 304 13. Beltrame JF, Crea F, Kaski JC, Ogawa H, Ong P, Sechtem U, et al. International
305 standardization of diagnostic criteria for vasospastic angina. *Eur Heart J.* 2017;38:
306 2565–2568. <https://doi.org/10.1093/EURHEARTJ/EHV351> PMID: 26245334
- 307 14. Hamilton KK, Pepine CJ. A renaissance of provocative testing for coronary spasm? *J*
308 *Am Coll Cardiol.* 2000;35: 1857–1859. [https://doi.org/10.1016/S0735-1097\(00\)00653-](https://doi.org/10.1016/S0735-1097(00)00653-7)
309 [7](https://doi.org/10.1016/S0735-1097(00)00653-7) PMID: 10841235
- 310 15. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for
311 Medicine. *N Engl J Med.* 2023;388: 1233–1239.
312 <https://doi.org/10.1056/NEJMSR2214184> PMID: 36988602
- 313 16. Sanderson K. GPT-4 is here: what scientists think. *Nature.* 2023;615.
314 <https://doi.org/10.1038/D41586-023-00816-5> PMID: 36928404
- 315 17. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large
316 language models in medicine and medical research. *Lancet Digit Health.* 2023;0.
317 [https://doi.org/10.1016/S2589-7500\(23\)00083-3](https://doi.org/10.1016/S2589-7500(23)00083-3) PMID: 37120418

- 318 18. Nori H, King N, Mckinney SM, Carignan D, Horvitz E, Openai M 2. Capabilities of
319 GPT-4 on Medical Challenge Problems. 2023 [Available from:
320 <https://arxiv.org/abs/2303.13375v2>]
- 321 19. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D, Allen PG. Evaluating GPT-4 and
322 ChatGPT on Japanese Medical Licensing Examinations. 2023 [Available from:
323 <https://arxiv.org/abs/2303.18027v2>]
- 324 20. OpenAI. GPT-4 Technical Report. 2023 [Available from:
325 <https://arxiv.org/abs/2303.08774v3>]
- 326 21. Jiao W, Wang W, Huang J, Wang X, Tu Z. Is ChatGPT A Good Translator? Yes With
327 GPT-4 As The Engine. 2023 [Available from: <https://arxiv.org/abs/2301.08745v3>]
- 328 22. Kiyohara Y, Kodera S, Ninomiya K, Sawano S, Katsushika S, Shinohara H, et al.
329 [Natural Language Processing Artificial Intelligence Model to Distinguish Coronary
330 Spastic Angina by Medical History]. *Shinzo*. 2022;54: 1235–1242. [Available from:
331 https://jglobal.jst.go.jp/detail?JGLOBAL_ID=202202264102722109]. Japanese.
- 332 23. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional
333 Transformers for Language Understanding. NAACL HLT 2019 - 2019 Conference of
334 the North American Chapter of the Association for Computational Linguistics: Human
335 Language Technologies - Proceedings of the Conference. 2018;1: 4171–4186.
336 [Available from: <https://arxiv.org/abs/1810.04805v2>]
- 337 24. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language
338 Models are Few-Shot Learners. [Available from: <https://arxiv.org/abs/2005.14165>]
- 339 25. Min S, Seo M, Hajishirzi H. Question Answering through Transfer Learning from
340 Large Fine-grained Supervision Data. [Available from:
341 <https://arxiv.org/abs/1702.02171v5>]

- 342 26. Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. :
343 328–339. [Available from: <https://arxiv.org/abs/1801.06146>]
- 344 27. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural
345 Language Processing in languages other than English: opportunities and challenges. J
346 Biomed Semantics. 2018;9: 12. <https://doi.org/10.1186/S13326-018-0179-8> PMID:
347 29602312
- 348 28. Ahmad A, Shelly-Cohen M, Corban MT, Murphree Jr DH, Toya T, Sara JD, et al.
349 Machine learning aids clinical decision-making in patients presenting with angina and
350 non-obstructive coronary artery disease. European heart journal Digital health. 2021;2:
351 597–605. <https://doi.org/10.1093/EHJDH/ZTAB084> PMID: 36713103
- 352 29. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al.
353 Foundation models for generalist medical artificial intelligence. Nature. 2023;616:
354 259–265. <https://doi.org/10.1038/S41586-023-05881-4> PMID: 37045921
- 355 30. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, Rahimian F, Canoy D, Tran J, et al.
356 Deep learning for electronic health records: A comparative review of multiple deep
357 neural architectures. J Biomed Inform. 2020;101.
358 <https://doi.org/10.1016/J.JBI.2019.103337> PMID: 31916973
- 359 31. Roberts A. Language, Structure, and Reuse in the Electronic Health Record. AMA J
360 Ethics. 2017;19: 281–288.
361 <https://doi.org/10.1001/JOURNALOFETHICS.2017.19.3.STAS1-1703> PMID:
362 28323609
- 363 32. Greenhalgh T, Potts HWW, Wong G, Bark P, Swinglehurst D. Tensions and paradoxes
364 in electronic patient record research: a systematic literature review using the meta-
365 narrative method. Milbank Q. 2009;87: 729–788. [https://doi.org/10.1111/J.1468-
366 0009.2009.00578.X](https://doi.org/10.1111/J.1468-0009.2009.00578.X) PMID: 20021585

367

368 **Supporting information**

369 **S1 Fig. An example of the data.**

370 This example was extracted from the following case report published in an open-access
371 journal. We used only the original sentences written in Japanese above in the box. In the
372 figure, we translated the whole text into English to make it easy to understand.

373 **S2 Fig. The detailed data for few-shot learning.**

374 This figure shows the ten cases of the learning data used for few-shot learning. In the
375 prompts, we put only the original sentences written in Japanese in the above box. In the
376 figure, we translated the whole text into English to make it easy to understand.

377 **S3 Fig. Comparison with cardiologists' accuracy and GPT-3.5 with zero-shot learning.**

378 The result of GPT-3.5 with zero-shot learning is shown with reference to the proportion of
379 cardiologists who answered correctly.

380 **S4 Fig. Comparison with cardiologists' accuracy and GPT-3.5 with few-shot learning.**

381 The result of GPT-3.5 with few-shot learning is shown with reference to the proportion of
382 cardiologists who answered correctly.

383 **S5 Fig. Comparison with cardiologists' accuracy and GPT-4 with zero-shot learning.**

384 The result of GPT-4 with zero-shot learning is shown with reference to the proportion of
385 cardiologists who answered correctly.

386 **S6 Fig. Comparison with cardiologists' accuracy and Bard with zero-shot learning.**

387 The result of Bard with zero-shot learning is shown with reference to the proportion of
388 cardiologists who answered correctly.

Patient information



Zero-shot learning



Few-shot learning

GPT-3.5



GPT-4



Bard



distinguish

Vasospastic angina

Acute coronary syndrome

Fig 1

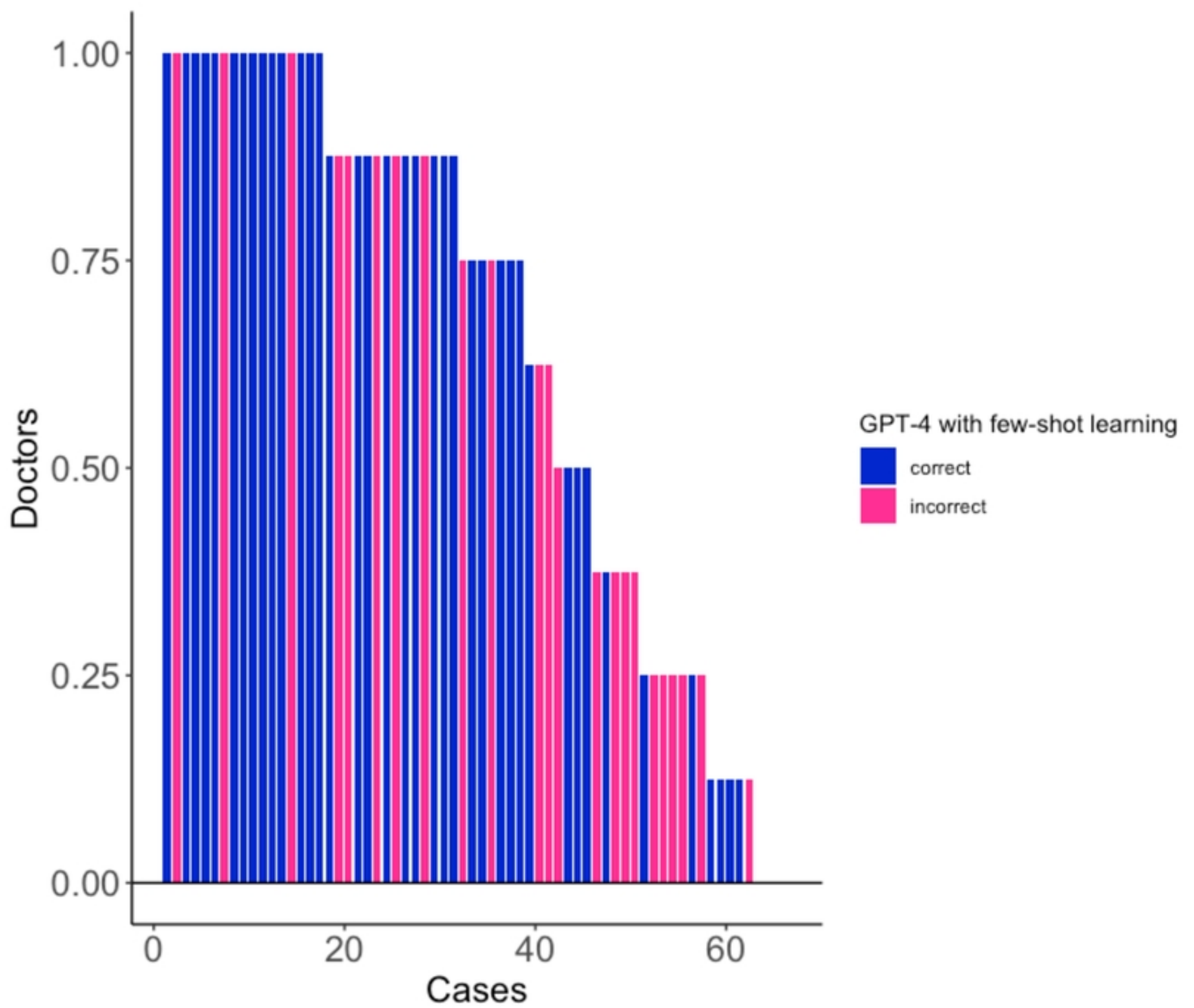


Fig 2