

1 Microsoft Bing outperforms five other generative artificial
2 intelligence chatbots in the Antwerp University multiple
3 choice medical license exam

4 Stefan Morreel^{1*}

5 Veronique Verhoeven¹

6 Danny Mathysen^{1,2}

7 ¹Department of Family Medicine and Population Health, University of Antwerp, Antwerp, Belgium

8 ²Dean's Department, University of Antwerp, Antwerp, Belgium

9 * Corresponding author:

10 E-mail: stefan.morreel@uantwerpen.be

11 **Short title:** Performance of AI bots in a medical license exam

12

13 Abstract

14 Recently developed chatbots based on large language models (further called bots) have promising
15 features which could facilitate medical education. Several bots are freely available, but their
16 proficiency has been insufficiently evaluated. In this study the authors have tested the current
17 performance on the multiple-choice medical licensing exam of University of Antwerp (Belgium) of six
18 widely used bots: ChatGPT (OpenAI), Bard (Google), New Bing (Microsoft), Claude instant (Anthropic),
19 Claude+ (Anthropic) and GPT-4 (OpenAI). The primary outcome was the performance on the exam
20 expressed as a proportion of correct answers. Secondary analyses were done for a variety of features
21 in the exam questions: easy versus difficult questions, grammatically positive versus negative
22 questions, and clinical vignettes versus theoretical questions. Reasoning errors and untruthful
23 statements (hallucinations) in the bots' answers were examined. All bots passed the exam; Bing and
24 GPT-4 (both 76% correct answers) outperformed the other bots (62-67%, $p = 0.03$) and students
25 (61%). Bots performed worse on difficult questions (62%, $p = 0.06$), but outperformed students (32%)
26 on those questions even more ($p < 0.01$). Hallucinations were found in 7% of Bing's and GPT4's
27 answers, significantly lower than Bard (22%, $p < 0.01$) and Claude Instant (19%, $p = 0.02$). Although the
28 creators of all bots try to some extent to avoid their bots being used as a medical doctor, none of the
29 tested bots succeeded as none refused to answer all clinical case questions.

30 Bing was able to detect weak or ambiguous exam questions. Bots could be used as a time efficient
31 tool to improve the quality of a multiple-choice exam.

32

33 Author Summary

34 Artificial chatbots such as ChatGPT have recently gained a lot of attention. They can pass exams for
35 medical doctors, sometimes they even perform better than regular students. In this study, we have
36 tested ChatGPT and five other (newer) chatbots in the multiple-choice exam that students in
37 Antwerp (Belgium) must pass to obtain the degree of medical doctor. All bots passed the exam with
38 results similar or better than the students. Microsoft Bing scored the best of all tested bots but still
39 produces hallucinations (untruthful statements or reasoning errors) in seven percent of the answers.
40 Bots performed worse on difficult questions but they outperformed students on those questions
41 even more. Maybe they are most useful when humans don't know the answer themselves? The
42 creators of the bots try to some extent to avoid their bots being used as a medical doctor, none of
43 the tested bots succeeded as none refused to answer all clinical case questions. Microsoft Bing also
44 turns out to be useful to find weak questions and as such improve the exam.

45 Introduction

46 The development of AI applications announces a new era in many fields of society including medicine
47 and medical education. Especially artificial intelligence (AI) chatbots based on large language models
48 (further called bots) have promising features which could facilitate education by offering simulation
49 training, by personalizing learning experiences with individualised feedback, or by acting as a decision
50 support in clinical training situations. However, before adopting this technology in the medical
51 curriculum, its capabilities have yet to be thoroughly tested.[1, 2]

52 Soon after the first bots became publicly available, higher medical education institutes started to
53 report on their performance in medical exam simulations.[3]

54 Whereas bots seem to be informative and logical in many of their responses, in others they answer
55 with obvious, sometimes dangerous, hallucinations (confident responses which however contain
56 reasoning errors or are unjustified by the current state of the art).[4] They will reproduce flaws in the
57 datasets they are trained by; they may reflect or even amplify societal inequality or biases or generate
58 inaccurate or fake information.[5]

59 Mostly, bots perform near the passing mark,[5-8] although they outperform students in some
60 reports.[9, 10] Performance is in general better on more easy questions and when the exam is written
61 in English.[11, 12] Notably their score is generally worse as exams at more advanced stages in the
62 medical curriculum are offered. However, bots seem to learn rapidly, and new versions do
63 considerably better than their prototypes [13-15] . As bots evolve, their proficiency needs continuous
64 monitoring and updating.

65 Whereas media articles state that higher education institutes already anticipate the dangers of bots in
66 terms of possible exam fraud, they also offer opportunities to assist in developing exams, for example
67 by identifying ambiguous or badly formulated exam questions.

68

69 Very few comparisons between different bots have been made, and those that do exist only compare
70 two or three bots and do not report hallucination rates.[16, 17]

71 In this study, we use the final theory exam that all medical students need to pass to obtain the degree
72 of Medical Doctor. It is followed by an oral exam which is not part of this study. The current exam was
73 used in 2021 at the University of Antwerp, Belgium. It is similar to countrywide exams used in other
74 countries, such as the United States Medical Licensing exam step 1 and step 2CK.[18]

75 In this study we have tested the current performance of six publicly available bots on the University of
76 Antwerp medical licensing exam. The primary outcomes concern the performance of each bot on the
77 exam. Secondary outcomes include performance on subsets of questions, interrater variability,
78 proportion of hallucinations and the detection of possible weak exam questions.

79 **Material and Methods**

80 **Ethics**

81 This experiment has been approved by the Ethics Committee of the University of Antwerp and the
82 Antwerp University Hospital (reference number MDF 21/03/037, amendment number 5462).

83 **Materials**

84 At the end of the undergraduate medical training at the University of Antwerp, medical students must
85 pass a general medical knowledge examination before being licensed as medical doctor. Besides an
86 oral viva examination, this general medical knowledge examination contains 102 multiple choice
87 questions covering the entire range of curricular courses. In this study, the exam as it was presented
88 to the students in their second master year (before their final year of clinical training) was used. The
89 scoring system was adapted afterwards, so the student's scores in this paper do not reflect the actual
90 grades given to the students. The questions were not available online, so they were not used for the
91 training of the studied bots.

92 Bot selection

93 Six bots that are publicly available and can currently be used by teachers and students were tested.
94 The most widely used free bots were selected: ChatGPT (OpenAI), Bard (Google), and New Bing
95 (Microsoft). Claude instant (Anthropic), Claude+(Anthropic) and GPT-4(OpenAI) were added to the list
96 because they allow for an evaluation of the difference between a free and a paying version. Even
97 though Bing is based on the GPT-4 large language model, it also uses other sources such as Bing
98 Search so it is a customized version of the pure GPT-4 bot.[19]

99 Data extraction

100 The exam was translated using DeepL (DeepL SE), a neural machine translation service. Clear
101 translation errors were corrected by author SM, but the writing style and grammar were not
102 improved in order to mimic an everyday testing situation. Questions containing images/tables (N=2)
103 and local questions were excluded (N=5). Local questions were excluded because they concern
104 theories, frameworks or models that have only been described in Dutch and are only applicable to
105 Belgium and the Netherlands. Literal translation of these questions leads to nonsense questions in
106 English.

107 Details on how and when the bots were used can be found in table 1. By coincidence, the authors
108 found out that when Bard refuses to answer a medical question, prompting it with “please regenerate
109 draft” may force it to answer the question anyhow. This was not the case for the other bots. In all
110 cases where Bard refused to answer, this additional prompt was used.

111 *Table 1: overview of the tested generative chat bots.*

Bot	Large Language Model	Properties	Avoiding memory retention	Log in?	Access dates	Price
Bing	GPT-4	Conversation style = More precise	“New topic” function is used after	Microsoft account	7-9/6/2023	Free

			each question			
Bard	PALM 2	Accessed using a virtual private network to emulate US location	“Reset Chat” function is used after each question	Google account	12-14/06/2023	Free
ChatGPT	GPT-3.5	Accessed through Poe*	A new chat is started using the broom button	Poe* log in	12-26/06/2023	Free, A paying version exists based on GPT-4.
Claude+	Claude version 1	Accessed through Poe*	Broom button	Poe* log in	12-26/06/2023	Free trial on Poe paying afterwards
Claude Instant	Lighter version of Claude version 1	Accessed through Poe*	Broom button	Poe* log in	12-26/06/2023	Free trial on Poe paying afterwards
GPT-4	GPT-4	Accessed through Poe*	Broom button	Poe* log in	12-26/06/2023	Free trial on Poe paying afterwards

112 GPT: generative pre-trained transformer

113 PaLM: Pathways Language Model

114 *: Poe (Platform for Open Exploration, Quora) was used because it allows fluent testing of
115 multiple bots at the same time. A trial subscription of one week was used.

116 Outcomes

117 The primary outcome was the performance on the exam expressed as a proportion of correct
118 answers (score). This outcome was also measured in the same way as the students were rated on this
119 exam (adapted score): eleven questions contained a second best answer (an acceptable alternative to
120 the best answer), a score of 0.33 was awarded when this option was chosen; twenty questions
121 contained a fatal answer (this option is dangerous for the patient) leading to a score of -1. For
122 calculation of the student’s scores, the image, table, and local questions were excluded as well.

123 The primary outcomes were assessed in four subsets of answers. Firstly, the difficulty of the
124 questions: thirteen questions were difficult (recorded P-value in question bank below 0.30 meaning
125 that less than 30% of the students answered the question correct[20]), 36 easy (recorded P-value in
126 question bank above 0.80) and 46 moderate (recorded P-value in question bank between 0.30 and
127 0.80). Secondly, the grammar of the questions: negative formulated questions (e.g., “which statement
128 is not correct?”) vs positive statements. Five questions were negatively formulated. Thirdly, the type
129 of question: theory (50 questions) or describing a patient (clinical vignette, 45 questions). Finally,
130 questions with vs without fatal answers.

131 In those cases where a bot answered a question incorrectly with a fatal answer, the proportion of
132 selected fatal answers among all wrong answers was calculated.

133 The primary outcome was also assessed for a virtual bot (called Ensemble Bot), the answer of this bot
134 was the mode (most common value) of the answers of all six bots.[21]

135 Three additional outcomes were assessed. Firstly, the proportion of hallucinations as rated by the
136 authors among the incorrect answers of the best scoring bot. Authors VV and DM read all incorrect
137 answers and judged them as containing a hallucination or not. In case of discordance, author SM
138 made a final decision. A hallucination was previously defined as content that is nonsensical or
139 untruthful in relation to certain sources.[22] This definition is not usable for the current research so
140 the authors defined a hallucination as content that either contains clear reasoning or is untruthful in
141 relation to current evidence based medical literature. To detect reasoning errors, no medical
142 knowledge is required. For example: “the risk is about 1 in 100 (3%)”. To detect untruthful answers,
143 the authors had to use their own background knowledge combined with common online resources to
144 verify the AI answers. One clear example of an untruthful answer given by several bots: “This is a
145 commonly used mnemonic to remember the order: "NAVEL" - Nerve, Artery, Vein, Empty space (from
146 medial to lateral).” This mnemonic does exist, but it should be used from lateral to medial. Because a
147 multiple-choice exam was studied, the hallucination could not be found in the answer itself but in the

148 arguments supporting the selected answer. Bots never answer with a simple letter, they all produce
149 written out answer of varying length. The authors wanted to report reasoning errors and untruthful
150 answers separately but found out that often, these two were both present in a bot's answer so this
151 outcome was suspended.

152 Secondly, the proportion of possible weak questions among the incorrect answers of the best scoring
153 bot. For this outcome, all authors discussed all incorrect answers of the best scoring bot and reached
154 unanimous consensus.

155 Thirdly, the interrater variability was examined. Originally, the authors planned to test whether user
156 interpretation of the answers would be different from strict interpretation of the bot's answer as this
157 difference was significant in a previous study.[8] This outcome was suspended because such cases
158 occurred only in ChatGPT and Bard.

159 **Analysis**

160 The differences in performance among the bots/students, differences in performance among
161 categories of questions, and differences in the proportion of hallucinations were tested with a one-
162 way ANOVA test and pairwise unpaired two-sample T-tests. P-values were 2-tailed where applicable,
163 and a p-value of less than 0.05 was considered statistically significant. A p-value between 0.05 and
164 0.10 was considered a trend. For the wrong answers on questions with a fatal answer, a χ^2 test was
165 used to assess the difference between the bot's proportion of fatal answers and the random
166 proportion of fatal answers (which equals 0.33). Fleiss' Kappa was used to assess the overall
167 agreement among the bots. Cohen's kappa was used to assess pairwise interrater agreement
168 between the different bots. Raw data was collected using Excel 2023 (Microsoft). JMP Pro version 17
169 (JMP Statistical Discovery LLC) was used for all analyses except Fleiss' kappa which was calculated in R
170 version 4.31 (DescTools package).

171 Results

172 Exam performance

173 See table 2 for an overview of the scores of the tested bots. Bing and GPT-4 scored the best with 76%
 174 correct answers and an adapted score (the way students were rated) of 76% as well. The mean score
 175 of all bots was 68%, the scores of the individual bots were not significantly different from this mean (p
 176 = 0.12). However, Bing and GPT-4 scored significantly better than Bard ($p=0.03$) and Claude Instant
 177 ($P=0.03$). GPT-4 had the same score as Bing but had more wrong answers (25 versus 13). Claude+ did
 178 not significantly score better than Claude Instant. All Bots gave one fatal answer (on different
 179 questions) except Bard which did not give any fatal answers. Bing gave four second best answers,
 180 ChatGPT/Bard/GPT three, Claud two and Claud Instant only one. For thirteen questions, Bard refused
 181 to answer. After prompting Bard up to five times with “regenerate draft”, it still refused to answer
 182 four questions, seven were answered correctly and two were wrongly. The performance of the bots
 183 using the adapted score was very similar because the added points of second-best answers were
 184 smoothed out by the lost points due to fatal answers. The mean score of the 95 students was 61%
 185 (standard deviation 9), the mean adapted score for students was 60% (standard deviation 21). The
 186 Ensemble Bot (answers with the most common answer among the six bots) scored the same as Bing
 187 (72 correct answers, 76%).

188 *Table 2. Performance of generative chat bots on the University of Antwerp Medical License Exam (95*
 189 *questions)*

	Correct Answers (N)	Score (%)	95% Lower CI	95% Upper CI	No answer (N)	Refusal to answer (N)	Several answers without clear choice (N)	Unclear answer (N)	Wrong answer (N)	Adapted score* (%)
Bing	72	76	66	83	3	1	1	5	13	76
ChatGPT	64	67	57	76	1	0	3	2	25	67
Bard	58	61	51	70	0	4	2	0	31	62
GPT-4	72	76	66	83	1	0	3	1	18	76
Claude+	64	67	57	76	1	2	5	0	23	67
Claude Instant	60	63	53	72	2	2	3	0	28	62

190
191 **This is the score that was used to assess students. A second-best answer was rated as +0.33*
192 *and a fatal answer as -1.*

193 *CI: confidence interval for the score (%)*

194 To illustrate this performance S1 Table contains a question and the responses from all selected bots.

195 Performance for subsets of questions

196 The bots scored on average 73% for easy questions and 62% for difficult questions ($P=0.06\%$). The
197 students scored on average 75% for easy questions and 32% for difficult questions ($p<0.01$). Assessing
198 difficult questions only, ChatGPT performed best with a score of 77%, Bing/GPT4 scored 69%. The
199 students scored 32% on difficult questions which is significantly lower as compared to ChatGPT, Bing,
200 and GPT-4 ($p<0.01$). A similar but smaller effect was found for moderate questions (Bing versus
201 students, 72% versus 59%, $p = 0.07$) but not for easy questions (69 vs 74%, $p=0.30$)

202 No significant difference in performance on negative versus positive questions ($p=0.16$) and on clinical
203 vignettes versus theory questions ($p=0.16$) was found. Such a difference was not found for the
204 students either ($p = 0.54$ and 0.38 respectively). When examining individual questions, errors on
205 clinical vignette questions were often caused because Bing missed an important clue in the context or
206 the history of the patient. For example, in a question concerning the timing of a flu vaccine for a
207 pregnant patient consulting in august, Bing answers that the flu vaccine was necessary now. Bing
208 missed the clue about august: flu vaccines should be given later and are generally not available yet in
209 august.[23]

210 The bots scored on average 72% on questions with a fatal answer which is not significantly different
211 from questions without a fatal answer (68%, $p=0.39$). Among the 34 wrong answers, the fatal answer
212 was chosen five times (15%) which is lower than can be expected by chance only (11 wrong answers
213 or 33%, $p=0.09$). The students did perform worse on these questions (mean 64% versus 52%, $p=0.03$).
214 Among the 843 wrong student answers, the fatal option was chosen in 111 answers (13%).

215 **Hallucinations**

216 Hallucinations were found in 7% of Bing’s and GPT4’s answers. This was significantly lower than Bard
 217 (22%, $p < 0.01$) and Claude Instant (19%, $p = 0.02$). ChatGPT had 15% hallucinations and Claude+ 12%,
 218 this was not significantly different from Bing and GPT4 ($P > 0.10$ for all these comparisons).
 219 See Table 3 for a question on which five bots hallucinated (reasoning errors).

220 *Table 3. Example of all generative chatbot hallucinations on one question. Reasoning errors are*
 221 *indicated in bold.*

Question	ChatGPT	Bard	Claude Instant	Claude+
What is the approximate risk of having a child with Down syndrome for a 40-year-old woman? a) 1 % (correct answer) b) 3 % c) 5 % d) 10 %	The approximate risk of having a child with Down syndrome for a 40-year-old woman is d) 10%. As women age, the risk of having a child with Down syndrome increases. At age 40, the risk is approximately 1 in 100 or 10%. This is compared to a risk of approximately 1 in 1,250 or 0.08% at age 25. ...	The correct answer is b) 3%. The risk of having a child with Down syndrome increases with the mother's age. For a 40-year-old woman, the risk is about 1 in 100 (3%). This means that about 3 out of every 100 babies born to a 40-year-old woman will have Down syndrome. ...	For a 40-year-old woman, the approximate risk of having a child with Down syndrome is about 1 in 50 or 2%. So the closest answer is b) 3%. The risk of Down syndrome increases with the mother's age. At age 25, the risk is about 1 in 1,250. At age 35, the risk is about 1 in 400. By age 40, the risk increases to about 1 in 100. And by age 45, the risk is about 1 in 30. ...	The risk of having a child with Down syndrome increases with maternal age. For a 40-year-old woman, the risk is approximately: c) 5% The risk of Down syndrome roughly doubles every 5 to 6 years after age 30: Under 30 years old: 0.1% At age 35: 1% At age 40: About 2% to 5% At age 45: About 5% to 10% So for a 40-year old woman, the approximate risk is in the ballpark of 5%, making C the closest option.

222 Detection of weak questions

223 Among the 23 incorrect answers of Bing, three questions were unclearly written and two were not in
224 line with current literature. An example of a detected weak questions is one concerning renal
225 replacement therapy: *“Complete. Renal function replacement therapy is indicated ... a) in any
226 symptomatic patient with an eGFR <15 ml/min/1.73m². b) only in patients under 65 years of age. c) in
227 anyone with an eGFR < 6 l/min/1.73m² d) only when urea is elevated”*. Bing answered “a)”. After
228 review of current literature, the authors judge that an eGFR below 15 is indeed a commonly used cut
229 of value for starting renal replacement therapy but it is not the only reason so start dialysis. Because
230 statement a contains “any”, Bing’s answer is wrong, but the authors do understand why Bing
231 answered this question and why a student might give this answer as well. The same argument applies
232 to answer c which is supposed to be the correct answer. Even more, the eGFR cut-off of six is odd.
233 This question needs improvement.

234 Interrater variability

235 For 34 questions (36%), all bots agreed. Fleiss’ Kappa for all raters was 0.54 (moderate agreement).
236 The agreement between ChatGPT and GPT-4 was the highest (Cohen’s Kappa=0.66, substantial
237 agreement). The agreement between Bing and Bard was the lowest (Cohen’s Kappa= 0.48, moderate
238 agreement).

239 Discussion

240 In this study, significant differences in the performance of publicly available AI chatbots on the
241 Antwerp Medical License Exam were found. Both GPT-4 and Bing scored the best, but Bing turns out
242 more reliable as it produces fewer wrong answers. This performance is in line with previous
243 research.[13-15] An ensemble bot which combines all tested bots scored equally. The proportion of
244 hallucinations was much lower for Bing than for Bard and Claude+/Claude Instant.

245 The improvement of these new bots both in scores as in proportion of hallucinations sounds
246 impressive, it might however increase the risk as users will have more confidence in wrong or even
247 dangerous answers as the bots (in general) answer more correctly. The risk of replicating biases in the
248 data on which these models are trained remains. Other authors already pointed out the meaning of
249 these results: bots can pass exams, but this does not make them medical doctors as this requires far
250 more capacities than reproduction of knowledge alone. The current study raises the questions
251 whether a multiple choice exam is a useful way to assess the competencies modern doctors need
252 (mostly concerning human interactions).[24] Bing performed equally as GPT-4 but with less wrong
253 answers, so currently it is not worth paying for a bot in order to test a medical exam, neither is it
254 useful to create an ensemble bot based on the mode of all bot's answers. Ensemble bots based on
255 more complex rules than just the mode of all answers should be studied further.

256 We can recommend the use of Bing to detect weak questions among the wrong answers. This is a
257 time-efficient way to improve the quality of a multiple-choice exam.

258 The trend we found towards better bot performance on easy questions is in line with previous
259 research.[11] However, the difference in performance between students and bots was large for
260 difficult questions and absent for easy questions. This compelling new finding demands further
261 research. Maybe bots are most useful in those situations that are difficult for humans?

262 The lack of a significant difference in performance between positive and negative questions, and
263 between clinical vignettes and theory questions needs confirmation on larger datasets and on other
264 exams.

265 Although the creators of all bots try, to a certain extent, to avoid their bots being used as a medical
266 doctor, none of the tested bots succeeded as none refused to answer all clinical case questions. Only
267 Claude+ and Claude instant refused (at times) to answer the question and closed the conversation.
268 For all other bots users can try to pursue them to answer the question anyhow. This finding was most

269 compelling for Bard where after entering the same questions repeatedly, Bard did answer it in nine
270 out of thirteen cases.

271 The rise of generative AI also raises many ethical and legal issues: their enormous energy
272 consumption, use of data sources without permission, use of sources protected by copyright, lack of
273 reporting guidelines and many more. Before widely implementing AI in medical exams, more
274 legislation and knowledge is necessary on these topics.[25, 26]

275 The strengths of this study mainly concern its novelty: the comparison of six different bots had not
276 been published yet. The bots tested are available to the public so our methodology can easily be re-
277 used. This study, however, has got several limitations as well. It only concerned one exam with a
278 moderate size set of questions. There was no usable definition of hallucinations, neither a validated
279 approach to detect them available at the time of writing. The definition we have used (chatbot
280 generated content that either contains clear reasoning or is untruthful in relation to current evidence
281 based medical literature) might inspire other authors although we found out that a distinction
282 between reasoning errors and untruthful statements was not feasible. The exclusion of tables, local
283 questions and images reduces the use of the comparison to real students. Future bots will most likely
284 be able to process such questions as well. Finally, the exam was translated in English to make the
285 current paper understandable for a broad audience. Further research on other languages is
286 necessary.

287 **Conclusion**

288 Six generative AI chatbots passed the Antwerp multiple choice exam necessary for obtaining a license
289 as an MD. Bing (and to a lesser extent GPT-4) outperformed all other bots and students. Bots
290 performed worse on difficult questions but outperformed students on those questions even more.
291 Bing can be used to detect weak multiple-choice questions. Bots should improve their algorithm if
292 they do not want to be used as a medical.

293 Acknowledgements

294 The authors would like to thank Professor David Martens for proofreading this manuscript.

295 References

- 296 1. Rudolph J, Tan S, Tan S. ChatGPT: Bullshit spewer or the end of traditional assessments in
297 higher education? *Journal of Applied Learning and Teaching*. 2023;6(1).
- 298 2. Chatterjee J, Dethlefs N. This new conversational AI model can be your friend, philosopher,
299 and guide... and even your worst enemy. *Patterns*. 2023;4(1).
- 300 3. Kung TH, Cheatham M, Medinilla A, ChatGPT, Sillos C, De Leon L, et al. Performance of
301 ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models.
302 medRxiv. 2022:2022.12.19.22283643.
- 303 4. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language
304 generation. *ACM Computing Surveys*. 2023;55(12):1-38.
- 305 5. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery
306 examination? Orthopaedic residents versus ChatGPT. *Clinical Orthopaedics and Related Research*®.
307 2022:10.1097.
- 308 6. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical
309 students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof*.
310 2023;20(1).
- 311 7. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style
312 examination: Insights into current strengths and limitations. *Radiology*. 2023:230582.
- 313 8. Morreel S, Mathysen D, Verhoeven V. Aye, AI! ChatGPT passes multiple-choice family
314 medicine exam. *Med Teach*. 2023;45(6):665-6. Epub 20230311. doi:
315 10.1080/0142159x.2023.2187684. PubMed PMID: 36905610.
- 316 9. Li SW, Kemp MW, Logan SJ, Dimri PS, Singh N, Mattar CN, et al. ChatGPT outscored human
317 candidates in a virtual objective structured clinical examination in obstetrics and gynecology.
318 *American Journal of Obstetrics and Gynecology*. 2023.
- 319 10. Subramani M, Jaleel I, Krishna Mohan S. Evaluating the performance of ChatGPT in medical
320 physiology university examination of phase I MBBS. *Advances in Physiology Education*.
321 2023;47(2):270-1.
- 322 11. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing
323 examination in Taiwan. *J Chin Med Assoc*. 2023;86(7):653-8. Epub 20230705. doi:
324 10.1097/jcma.0000000000000942. PubMed PMID: 37227901.
- 325 12. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style
326 Examination: Insights into Current Strengths and Limitations. *Radiology*. 2023;307(5):e230582. doi:
327 10.1148/radiol.230582.
- 328 13. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial Intelligence in
329 Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering
330 StatPearls Questions. *Cureus*. 2023;15(6):e40822. Epub 20230622. doi: 10.7759/cureus.40822.
331 PubMed PMID: 37485215; PubMed Central PMCID: PMC10362981.
- 332 14. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Sullivan PLZ, et al. Performance of ChatGPT,
333 GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*.
334 2022:10.1227.
- 335 15. Oh N, Choi G-S, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance
336 and its potential in surgical education and training in the era of large language models. *Annals of
337 Surgical Treatment and Research*. 2023;104(5):269.

- 338 16. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance
339 and its potential in surgical education and training in the era of large language models. *Ann Surg*
340 *Treat Res.* 2023;104(5):269-73. Epub 20230428. doi: 10.4174/astr.2023.104.5.269. PubMed PMID:
341 37179699; PubMed Central PMCID: PMCPMC10172028.
- 342 17. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT
343 Perform on the United States Medical Licensing Examination? The Implications of Large Language
344 Models for Medical Education and Knowledge Assessment. *JMIR Med Educ.* 2023;9:e45312. doi:
345 10.2196/45312.
- 346 18. Rashid H, Coppola KM, Lebeau R. Three Decades Later: A Scoping Review of the Literature
347 Related to the United States Medical Licensing Examination. *Acad Med.* 2020;95(11S Association of
348 American Medical Colleges Learn Serve Lead: Proceedings of the 59th Annual Research in Medical
349 Education Presentations):S114-s21. doi: 10.1097/acm.0000000000003639. PubMed PMID:
350 33105189.
- 351 19. Mehdi Y. Confirmed: the new Bing runs on OpenAI's GPT-4 2023 [09/08/2023]. Available
352 from: [https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-](https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4)
353 [OpenAI%E2%80%99s-GPT-4.](https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4)
- 354 20. Miller MD, Linn RL. Measurement and assessment in teaching. 11th ed. Boston: Pearson;
355 2013. xviii, 538 p. p.
- 356 21. Dietterich TG, editor Ensemble Methods in Machine Learning2000; Berlin, Heidelberg:
357 Springer Berlin Heidelberg.
- 358 22. OpenAI R. GPT-4 technical report. arXiv. 2023:2303.08774.
- 359 23. Prevention CfDca. Key Facts About Seasonal Flu Vaccine 2022 [11/08/2023]. Available from:
360 <https://www.cdc.gov/flu/prevent/keyfacts.htm>.
- 361 24. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a
362 spotlight on the flaws of medical education. *PLOS Digit Health.* 2023;2(2):e0000205. Epub 20230209.
363 doi: 10.1371/journal.pdig.0000205. PubMed PMID: 36812618; PubMed Central PMCID:
364 PMCPMC9931307.
- 365 25. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for
366 research. *Nature.* 2023;614(7947):224-6. doi: 10.1038/d41586-023-00288-7. PubMed PMID:
367 36737653.
- 368 26. Cacciamani GE, Collins GS, Gill IS. ChatGPT: standard reporting guidelines for responsible
369 use. *Nature.* 2023;618(7964):238. doi: 10.1038/d41586-023-01853-w. PubMed PMID: 37280286.

370

371

372 **Supplementary material captions**

373 S1 Table. Responses from all selected bots on an example question

374 S2 Selected Study Data. Study data excluding selected columns. See Data Availability Statement for
375 more information.

376 S3 Study Data Variables Overview. An overview of the properties of all variables used in file S2

377 Selected Study Data.