

1 Human versus Artificial Intelligence: ChatGPT-4
2 Outperforming Bing, Bard, ChatGPT-3.5, and Humans in
3 Clinical Chemistry Multiple-Choice Questions

4 Short title: AI-Based Models' Performance in Clinical Chemistry

5
6 Malik Sallam^{1,2,3,*}, Khaled Al-Salahat^{1,3}, Huda Eid³, Jan Egger⁴, Behrus Puladi⁵

7 ¹Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The
8 University of Jordan, Amman, Jordan

9 ²Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital,
10 Amman, Jordan

11 ³Scientific Approaches to Fight Epidemics of Infectious Diseases (SAFE-ID) Research
12 Group, The University of Jordan, Amman, Jordan

13 ⁴Institute for AI in Medicine (IKIM), University Medicine Essen (AöR), Essen, Germany

14 ⁵Institute of Medical Informatics, University Hospital RWTH Aachen, Aachen, Germany

15
16 ***Corresponding Author:**

17 E-mail: malik.sallam@ju.edu.jo (MS)

20 **Abstract**

21 The advances in large language models (LLMs) are evolving rapidly. Artificial intelligence
22 (AI) chatbots based on LLMs excel in language understanding and generation, with potential
23 utility to transform healthcare education and practice. However, it is important to assess the
24 performance of such AI models in various topics to highlight its strengths and possible
25 limitations. Therefore, this study aimed to evaluate the performance of ChatGPT (GPT-3.5 and
26 GPT-4), Bing, and Bard compared to human students at a postgraduate master's (MSc) level
27 in Medical Laboratory Sciences. The study design was based on the METRICS checklist for
28 the design and reporting of AI-based studies in healthcare. The study utilized a dataset of 60
29 Clinical Chemistry multiple-choice questions (MCQs) initially conceived for assessment of 20
30 MSc students. The revised Bloom's taxonomy was used as the framework for classifying the
31 MCQs into four cognitive categories: Remember, Understand, Analyze, and Apply. A
32 modified version of the CLEAR tool was used for assessment of the quality of AI-generated
33 content, with Cohen's κ for inter-rater agreement. Compared to the mean students' score which
34 was 40/60 (66.8%), GPT-4 scored 54/60 (90.0%), followed by Bing (46/60, 76.7%), GPT-3.5
35 (44/60, 73.3%), and Bard (40/60, 66.7%). Statistically significant better performance was noted
36 in lower cognitive domains (Remember and Understand) in GPT-3.5, GPT-4, and Bard. The
37 CLEAR scores indicated that ChatGPT-4 performance was "Excellent" compared to "Above
38 average" performance of ChatGPT-3.5, Bing, and Bard. The findings indicated that ChatGPT-4
39 excelled in the Clinical Chemistry exam, while ChatGPT-3.5, Bing, and Bard were above-
40 average. Given that the MCQs were directed to postgraduate students with a high degree of
41 specialization, the performance of these AI chatbots was remarkable. Due to the risks of
42 academic dishonesty and possible dependence on these AI models, the appropriateness of
43 MCQs as an assessment tool in higher education should be re-evaluated.

44 **Keywords:** AI in healthcare education; higher education; large language models; evaluation.

45 Abstract word count: **300** words, limit: 300 words.

46

47 **Introduction**

48 The domain of higher education is set for a new transformative era [1, 2]. This transformation
49 will be driven by the infiltration of artificial intelligence (AI) into various academic aspects [3-
50 7]. Specifically, the incorporation of AI into higher education can help in enhancing
51 personalized learning, supporting research, automating the grading, facilitating the human-
52 computer interaction, time-saving assistance, and enhancing the students' satisfaction [8-12].

53 Nevertheless, the AI utility in higher education does not only hold promising opportunities but
54 also valid concerns, both of which warrant critical and robust examination [13-15]. This
55 research endeavor is necessary to guide the ethical, responsible, and productive use of AI to
56 enhance higher education guided by a robust scientific evidence [14, 16, 17]. The relevance of
57 the quest to meticulously examine the benefits and challenges of AI in higher education is also
58 important in light of the current evidence showing that a substantial number of university
59 students are already using AI chatbots [18-22].

60 Despite the benefits of AI in higher education, it simultaneously raises valid concerns about
61 the academic integrity [23, 24]. The ease with which AI can perform complex tasks might
62 inadvertently encourage academic dishonesty, potentially undermining the educational ethics
63 [23, 25]. Furthermore, the reliance on AI for academic tasks could trigger a decline in critical
64 thinking and personal development skills among students, both of which are essential outcomes
65 to enable the graduates in achieving economic, technological, and social advancements [26,
66 27].

67 Ultimately, with notable capabilities of AI chatbots in understanding and engaging in helpful
68 conversations, would usher a paradigm shift in higher education [8, 14, 28]. This AI-driven
69 change could be a key moment in educational history, with impact surpassing the advent of the

70 internet and the transition to online teaching [16, 17, 29]. Therefore, the stakeholders in the
71 academia must strike the right balance between embracing technological innovations while
72 preserving the core values of education [30-32]. Thus, the integration of AI into higher
73 education is inevitable, and the academic organizations must adapt to this evolution [3]. This
74 adaptation involves the need to emphasize educational aspects such as self-reflection, critical
75 thinking, problem-solving, and independent learning [33]. Consequently, the educational
76 systems can benefit from AI as a tool to complement, rather than replace, human intellect and
77 creativity [34].

78 In the quest of transition to the AI era in education, guidance by robust scientific evidence is
79 crucial. One of the primary steps in this process is to scientifically evaluate the performance of
80 the commonly used and popular AI tools, such as ChatGPT (by OpenAI, San Francisco, CA),
81 Bing (by Microsoft Corporation, Redmond, WA), and Bard (by Google, Mountain View, CA).
82 Several recent studies explored the performance of AI-based models in multiple-choice
83 questions (MCQs), particularly within a broad spectrum of healthcare fields as recently
84 reviewed by Newton and Xiromeriti [35]. The observed variability in AI performance can be
85 ascribed to several factors, including the different AI models tested, varying approaches to
86 prompting, language variations, and the diversity of the topics tested, among others [35-37].
87 Thus, continued investigation into this research area is needed to elucidate the determinants of
88 AI model performance across various dimensions which can guide improvements in AI
89 algorithms. However, it is essential that such explorations are conducted utilizing a
90 standardized, refined methodology [36, 38].

91 The use of multiple-choice questions (MCQs) have traditionally been fundamental as an
92 objective approach in academic evaluation [39]. The versatility of MCQs is shown through the
93 use of the Bloom's taxonomy and its subsequent revised framework [40, 41]. The Bloom's

94 taxonomy can guide structuring MCQs to align with specific cognitive functions needed to
95 provide correct answers [42]. This alignment is key in assessing the students' achievement of
96 the intended learning outcomes [43]. The taxonomy stratifies these cognitive functions into
97 distinct categories. The lower cognitive levels encompass knowledge, which emphasizes
98 "Recall", and comprehension, centered on "Understanding". Conversely, the higher cognitive
99 functions include "Apply", key in problem-solving, and "Analyze", entailing the systematic
100 breakdown of information [40, 41].

101 In the context of assessing the performance of AI model performance in MCQs based on the
102 Bloom's taxonomy, a pioneering study by Herrmann-Werner et al. assessed ChatGPT-4 with
103 307 psychosomatic medicine MCQs [44]. The study demonstrated ChatGPT-4 ability to pass
104 the exam irrespective of the prompting method. Notably, cognitive errors were more prevalent
105 in "Remember" and "Understand" categories. Another recent study demonstrated that
106 ChatGPT-3.5 correctly answered 64 of 80 medical microbiology MCQs, albeit below student
107 averages, with better performance in the "Remember" and "Understand" categories and more
108 frequent errors in MCQ with longer choices in terms of word count [45].

109 In this study, the objective was to synthesize and expand upon recent research examining the
110 performance of AI chatbots in various examinations. This research was informed by seminal
111 studies, such as the Kung et al. evaluation of ChatGPT in the United States Medical Licensing
112 Examination (USMLE) [37], and also it aimed to extend the evidence of AI chatbot
113 performance in a topic rarely encountered in literature, namely the Clinical Chemistry at
114 postgraduate level. The novel contribution of the current study lies in employing a standardized
115 framework, termed "METRICS" for the design and reporting of AI assessment studies, coupled
116 with an in-depth analysis of AI models' rationale behind responses, using an evaluation tool
117 specifically tailored for AI content evaluation referred to as "CLEAR" [36, 38].

118 The study hypothesized that postgraduate students, particularly in the field of clinical
119 chemistry, will demonstrate superior performance compared to AI models. We anticipate that
120 this disparity will be especially evident in tasks requiring higher cognitive functions, such as
121 “Apply” and “Analyze”. This study aimed to critically assess the current capabilities of AI in
122 an academic setting and explore the differences of human versus artificial intelligence in
123 complex problem-solving scenarios.

124

125

126

127

128

129 **Materials and Methods**

130 **Study design**

131 The study utilized the METRICS checklist for the design and reporting of AI studies in
132 healthcare [36]. The basis of the study was a dataset of 60 MCQs, used in a Clinical Chemistry
133 examination. This examination was part of the Medical Laboratory Sciences Clinical
134 Chemistry course, tailored for Master of Science (MSc) students in Medical Laboratory
135 Sciences at the School of Medicine, University of Jordan.

136 The specific exam in focus was conducted in-person and 20 students undertook the examination
137 during the Autumn Semester of the 2019/2020 academic year. The students' performance in
138 each question was available for comparison with AI models.

139 The MCQs utilized in this exam were designed by the first author (M.S.), who is a Jordan
140 Medical Council (JMC) certified consultant in Clinical Pathology. Additionally, the first author
141 (M.S.) has been a dedicated instructor for this course since the Academic Year 2018/2019. The
142 MCQs were original, ensuring there were no copyright concerns.

143 **Ethical considerations**

144 In conducting this study, we paid careful attention to ethical implications. Ethical clearance for
145 this research was determined to be non-essential, given the nature of the data involved. The
146 data utilized were entirely anonymized, ensuring no breach of confidentiality or personal
147 privacy. Additionally, the university examination results, which formed part of our dataset, are
148 publicly accessible and open for academic scrutiny. Moreover, the MCQs employed in the
149 study were originally created by the first author. These questions are devoid of any copyright
150 concerns, further reinforcing the ethical integrity of our research approach.

151 **MCQ features and indices of human students' performance**

152 The indices of student performance included facility index defined as the proportion of students
153 who correctly answered the MCQ divided by the total number of students ($n=20$). The students
154 were then divided into the upper group comprising the top 5 performing students, and the
155 middle group comprising the middle 10 students and the lower group comprising the lower
156 scoring 5 students. The “Discrimination Index” (DI) was then calculated based on the
157 difference between the percent of correct responses in the upper group and the percent of
158 correct responses in the lower group. This was followed by the calculation of the “Maximum
159 Discrimination” based on the sum of the percent in the upper and lower groups marking the
160 item correctly. Then, the Discrimination Efficiency (DE) of the MCQ was calculated as the
161 ratio of DI to the Maximum Discrimination. The classification of the MCQs based on the
162 revised Bloom’s taxonomy four cognitive levels “Remember”, “Understand”, “Apply”, and
163 “Analyze” was based on a consensus between the first and second authors, both of which are
164 certified Clinical Pathologists.

165 **Models of AI tested, settings, testing time, and duration**

166 In this study, a detailed evaluation of four AI models was conducted, each selected for its
167 relevance, popularity, and advanced capabilities in language processing as follows: First,
168 ChatGPT-3.5 (OpenAI, San Francisco, CA) [46]: This model is grounded in the GPT-3.5
169 architecture deployed using its default settings and was assessed as of its latest update at time
170 of testing as of January 2022.

171 Second, ChatGPT-4 (OpenAI, San Francisco, CA) [46]: An advancement in the Generative
172 Pre-trained Transformer (GPT) series, with the most recent update from April 2023 at time of
173 testing. Third, Bing Chat (GPT-4 Turbo) [47]: This model uses the GPT-4 Turbo model. At the

174 time of testing, the version was updated until April 2023 and we selected the more balanced
175 conversation style. Fourth, Bard (Google, Mountain View, CA) [48]: This Google AI GPT
176 model was last updated on October 4, 2023, at time of testing.

177 The testing of these models was conducted over a concise period, spanning November 27 to
178 November 28, 2023. Our methodological approach involved initiating interactions with GPT-
179 3.5, GPT-4, and Bard using a single page. For Bing Chat, we used the “New Topic” option
180 considering the limit of responses posed by this model (50 at maximum). Additionally, we
181 opted not to use the “regenerate response” feature in ChatGPT and abstained from providing
182 feedback in all models to avoid feedback bias.

183 **Prompt and language Specificity**

184 In this study, we meticulously crafted the prompts used for interacting with the AI models to
185 ensure clarity and consistency in the testing process. For ChatGPT-3.5, ChatGPT-4, and Bard,
186 the following exact prompt was used: “For the following 60 Clinical Chemistry MCQs that
187 will be provided one by one, please select the most appropriate answer for each MCQ, with an
188 explanation for the rationale behind selecting this choice and excluding the other choices.
189 Please note that only one choice is correct while the other four choices are incorrect. Please
190 note that these questions were designed for masters students in medical laboratory sciences.”
191 This was followed by prompting each MCQ one by one. For Bing, the following prompt was
192 used for each MCQ: “For the following 60 Clinical Chemistry MCQs that will be provided one
193 by one, please select the most appropriate answer for each MCQ, with an explanation for the
194 rationale behind selecting this choice and excluding the other choices. Please note that only
195 one choice is correct while the other four choices are incorrect. Please note that these questions
196 were designed for masters students in medical laboratory sciences.”

197 All MCQs were presented in English. This choice was based on the fact that English is the
198 official language of instruction for the MSc program in Medical Laboratory Sciences at the
199 University of Jordan.

200 **AI content evaluation approach and individual involvement in** 201 **evaluation**

202 First, we objectively assessed the correctness of responses based on the key answers of the
203 MCQs. Then, subjective evaluation of the AI generated content was based on a modified
204 version of the CLEAR tool. This involved assessing the content on three dimensions as follows:
205 First, completeness of the generated response. Second, accuracy reflected by lack of false
206 knowledge and the content being evidence-based. Third, appropriateness and relevance of
207 content being easy to understand, well organized, and free from irrelevant content [38]. Each
208 dimension was scored on a 5-point Likert scale ranging from 1=poor, 2=satisfactory, 3=good,
209 4=very good, to 5=excellent. A list of the key points to be considered in the assessment was
210 set beforehand to increase objectivity.

211 The content generated by the four models was evaluated by two raters independently; the first
212 author (M.S.) a consultant in Clinical Pathology, and the second author (K.A.) a specialist in
213 Clinical Pathology, both certified in Clinical Pathology from the Jordan Medical Council
214 (JMC).

215 **Data source transparency and topic range**

216 The MCQs were totally conceived by the first author and sole instructor of the course. Sources
217 of the material taught during the course were the following three textbooks: Tietz Textbook of
218 Clinical Chemistry and Molecular Diagnostics; Clinical Chemistry: Principles, Techniques,

219 and Correlations; and Henry's Clinical Diagnosis and Management by Laboratory Methods
220 [49-51].

221 The scope of topics covered in the MCQs were as follows: Adrenal Function, Amino Acids
222 and Proteins, Body Fluid Analysis, Clinical Enzymology, Electrolytes, Gastrointestinal
223 Function, Gonadal Function, Liver Function, Nutrition Assessment, Pancreatic Function,
224 Pituitary Function, Thyroid Gland, and Trace Elements.

225 **Statistical and data analyses**

226 The statistical analysis was conducted using IBM SPSS Statistics Version 26.0 (Armonk, NY:
227 IBM Corp). The continuous variables were presented as means and standard deviations (SD),
228 while categorical data were summarized as frequencies and percentages [N (%)]. To explore
229 the associations between categorical variables, we employed the chi-squared test (χ^2), while to
230 explore the associations between scale variables and categorical variables, non-parametric tests
231 were utilized: the Mann–Whitney *U* test (M-W) and the Kruskal Wallis test (K-W). The
232 Kolmogorov-Smirnov test was employed to confirm the non-normality of the scale variables:
233 facility index (FI, $P=.042$), discriminative efficiency (DE, $P=.011$), word count for both stem
234 and choices ($P<.001$ for both), average completeness, accuracy/evidence,
235 appropriateness/relevance, and the mCLEAR scores ($P<.001$ for the four scores). *P* values
236 <0.050 were considered statistically significant. For multiple comparisons, post hoc analysis
237 was conducted using the M-W test. To account type I error due to multiple comparisons, we
238 adjusted the α level using the Bonferroni correction. Consequently, the adjusted α level for
239 conducting pairwise comparisons between the four AI models was set at $P=0.0083$.

240 The MCQs were categorized based on the FI as “difficult” for an FI of 0.40 or less, “average”
241 for an FI > 0.40 and ≤ 0.80 , and “easy” for an FI > 0.80 . Additionally, the DE was stratified

242 into “poor discrimination” if the DE was between -1 to zero, “satisfactory discrimination” for
243 $DE > zero$ to < 0.40 as satisfactory, and $DE \geq 0.4$ indicating “good discrimination”.

244 The inter-rater agreement was assessed using Cohen’s kappa (κ) values, which ranged from
245 very good to excellent. For ChatGPT-3.5, the agreement was $\kappa=0.874$ for Completeness,
246 $\kappa=0.921$ for Accuracy, and $\kappa=0.723$ for Relevance. ChatGPT-4 showed $\kappa=0.845$ for
247 Completeness, a perfect $\kappa=1$ for Accuracy, and $\kappa=0.731$ for Relevance. Bing displayed κ values
248 of 0.911 for Completeness, 0.871 for Accuracy, and 0.840 for Relevance. Lastly, Bard's
249 agreement was $\kappa=0.903$ for Completeness, $\kappa=1$ for Accuracy, and $\kappa=0.693$ for Relevance.
250 Finally, the overall modified CLEAR (mCLEAR) scores for AI content quality were averaged
251 based on the scores of the two raters and categorized as: “Poor” (1–1.79), “Below average”
252 (1.80–2.59), “Average” (2.60–3.39), “Above average” (3.40–4.19), and “Excellent” (4.20–
253 5.00) similar to the previous approach in [52].

254

255

256

257 **Results**

258 **Overall performance of the tested AI models compared to the** 259 **human students**

260 The overall performance of the MSc students in the exam was reflected in the average score of
261 40.05 ± 7.23 (66.75%), with the range of scores of the students of 24–54 (40.00%–90.00%). The
262 performance of the four AI models varied with the best performance for ChatGPT-4 scoring
263 54/60 (90.00%), followed by Bing scoring 46/60 (76.67%), ChatGPT-3.5 scoring 44/60
264 (73.33%), and finally Bard scoring 40/60 (66.67%).

265 **Human students’ performance based on the revised Bloom’s** 266 **taxonomy**

267 The MCQ metrics were derived from the performance of the 20 MSc students in the exam. The
268 best performance was in the “Remember” category, followed by the “Apply” category,
269 “Understand” category, while the worst performance was in the “Analyze” category; however,
270 these differences lacked statistical significance (**Table 1**).

271

272

273

274

275

276 **Table 1. Multiple-choice questions (MCQs) metrics stratified by the revised Bloom’s**
277 **taxonomy as derived from the performance of 20 MSc students.**

Revised Bloom's taxonomy	Remember	Understand	Apply	Analyze	P value³
MCQ metric	Mean±SD ²	Mean±SD	Mean±SD	Mean±SD	
Facility index	0.74±0.22	0.61±0.28	0.71±0.21	0.6±0.17	.180
Discriminative efficiency	0.24±0.25	0.24±0.27	0.17±0.43	0.43±0.41	.482
MCQ stem word count	15.04±6.5	24±16.95	73.4±57.06	25.31±27.55	.052
MCQ choices word count	13.5±9.29	24.83±17.76	22.2±8.07	29.54±28.64	.153
Revised Bloom's cognitive level¹	Lower		Higher		P value⁴
Facility index	0.68±0.25		0.63±0.18		.225
Discriminative efficiency	0.24±0.25		0.36±0.42		.205
MCQ stem word count	18.88±12.77		38.67±42.34		.265
MCQ choices word count	18.36±14.54		27.5±24.61		.268

278 ¹Lower cognitive level includes “Remember” and “Understand” categories, while the higher cognitive
279 level includes “Apply” and “Analyze” categories; ²SD: Standard deviation; ³Calculated using the
280 Kruskal Wallis test; ⁴Calculated using the Mann Whiteny *U* test.

281

282 **Performance of the AI models based on the MCQ metrics**

283 The performance of the four tested AI models was stratified based on the MCQ metrics.

284 Significantly lower number of correct answers was seen in difficult MCQs in both Bing and

285 Bard (**Table 2**), while the MCQ stem and choices word counts were not associated with AI

286 models’ performance.

287

288

289 **Table 2. Artificial intelligence (AI)-based model performance based on the multiple-choice question (MCQ) metrics.**

AI model	Answer	FI ¹ category			P value, χ^2	DE ³ category			P value, χ^2	MCQ stem word count		MCQ choices word count	
		Easy N ² (%)	Average N (%)	Difficult N (%)		Poor N (%)	Satisfactory N (%)	Good N (%)		Mean±SD ⁴	P value ⁵	Mean±SD	P value ⁵
GPT-3.5	Correct	15 (78.9)	22 (68.8)	7 (77.8)	.690, 0.741	10 (62.5)	15 (65.2)	19 (90.5)	.087, 4.891	23.8±28.49		18.2±16.07	
	Incorrect	4 (21.1)	10 (31.3)	2 (22.2)		6 (37.5)	8 (34.8)	2 (9.5)		27.63±21.62		29.06±22.41	
GPT-4	Correct	19 (100)	26 (81.3)	9 (100)	.054, 5.833	14 (87.5)	21 (91.3)	19 (90.5)	.923, .160	24.33±27.49		20.5±18.67	
	Incorrect	0	6 (18.8)	0		2 (12.5)	2 (8.7)	2 (9.5)		29.17±19.57		26.5±16.49	
Bing	Correct	16 (84.2)	26 (81.3)	4 (44.4)	.045, 6.204	11 (68.8)	18 (78.3)	17 (81.0)	.667, .809	25.89±29.62		19.57±15.17	
	Incorrect	3 (15.8)	6 (18.8)	5 (55.6)		5 (31.3)	5 (21.7)	4 (19.0)		21.29±13.53		26.14±26.6	
Bard	Correct	17 (89.5)	19 (59.4)	4 (44.4)	.027, 7.213	9 (56.3)	18 (78.3)	13 (61.9)	.303, 2.387	26.9±31.18		17.63±14.08	
	Incorrect	2 (10.5)	13 (40.6)	5 (55.6)		7 (43.8)	5 (21.7)	8 (38.1)		20.65±13.9		28.05±23.89	

290 ¹FI: Facility index of the MCQ; ²N: Number; ³DE: Discriminative efficiency of the MCQ; ⁴SD: Standard deviation; ⁵Calculated using the Mann Whiteny *U* test.

291

292

293

294

295 **Performance of the AI models based on the revised Bloom's**
296 **taxonomy**

297 Upon analyzing the AI models' performance in MCQs stratified per the four revised Bloom's
298 categories, only ChatGPT-4 showed statistically significant better performance in the
299 Remember and Understand categories compared to Apply and Analyze categories (Table 3).

300

301 **Table 3. The performance of the four artificial intelligence (AI)-based models in the**
302 **Clinical Chemistry multiple-choice question (MCQs) stratified per the four revised**
303 **Bloom's categories.**

Revised Bloom's taxonomy	Answer	Remember	Understand	Apply	Analyze	P value, χ^2
		N ¹ (%)	N (%)	N (%)	N (%)	
GPT-3.5	Correct	19 (79.2)	15 (83.3)	2 (40.0)	8 (61.5)	.164,
	Incorrect	5 (20.8)	3 (16.7)	3 (60.0)	5 (38.5)	5.104
GPT-4	Correct	24 (100)	17 (94.4)	3 (60.0)	10 (76.9)	.015,
	Incorrect	0	1 (5.6)	2 (40.0)	3 (23.1)	10.532
Bing	Correct	20 (83.3)	15 (83.3)	4 (80.0)	7 (53.8)	.182,
	Incorrect	4 (16.7)	3 (16.7)	1 (20.0)	6 (46.2)	4.859
Bard	Correct	18 (75.0)	14 (77.8)	3 (60.0)	5 (38.5)	.090,
	Incorrect	6 (25.0)	4 (22.2)	2 (40.0)	8 (61.5)	6.504

304 ¹N: Number.

305

306 On the other hand, ChatGPT-3.5, ChatGPT-4, and Bard showed statistically better performance
307 in the lower cognitive MCQs compared to the higher cognitive MCQs (**Figure 1**).

308 **Fig 1. The performance of the four artificial intelligence (AI)-based models in the MCQs**
309 **stratified per the revised Bloom cognitive levels.**

310

311 **Performance of the AI models based on the modified CLEAR tool**

312 In our assessment of completeness, accuracy/evidence, and appropriateness/relevance, based
313 on the modified CLEAR tool, ChatGPT-4 was the only model rated as "Excellent" across all
314 categories. Bing achieved an "Excellent" rating solely in appropriateness/relevance. The other
315 AI models were categorized as "Above average" in performance (**Table 4**). The statistical
316 analysis revealed significant superiority of ChatGPT-4 compared to the other models in all
317 CLEAR categories, with the exception of Bing where the difference was only significant in the
318 completeness and the overall mCLEAR score (**Table 4**).

319

320

321 **Table 4. Modified CLEAR average scores for the four AI models in explaining the rationale for selecting choices.**

Assessment category	Mean±SD ²	Rank	P value ³	Post hoc test (Mann Whiteny U test)						
				GPT-3.5 vs. GPT-4	GPT-3.5 vs. Bing	GPT-3.5 vs. Bard	GPT-4 vs. Bing	GPT-4 vs. Bard	Bing vs. Bard	
ChatGPT-3.5 completeness score	4.03±1.26	Above average								
ChatGPT-4 completeness score	4.73±0.77	Excellent	<.001	<.001	.343	.249	.001	.003	.745	
Bing completeness score	4.14±1.34	Above average								
Bard completeness score	4.19±1.18	Above average								
ChatGPT-3.5 accuracy/evidence score	3.87±1.80	Above average								
ChatGPT-4 accuracy/evidence score	4.6±1.21	Excellent	0.016	0.007	.633	.604	.023	.002	.324	
Bing accuracy/evidence score	4.07±1.66	Above average								
Bard accuracy/evidence score	3.67±1.90	Above average								
ChatGPT-3.5 appropriate/relevance score	4.18±1.33	Above average								
ChatGPT-4 appropriate/relevance score	4.76±0.76	Excellent	0.011	0.005	.645	.691	.023	.001	.355	
Bing appropriate/relevance score	4.27±1.35	Excellent								
Bard appropriate/relevance score	4.15±1.22	Above average								
ChatGPT-3.5 mCLEAR ¹ score	4.03±1.41	Above average								
ChatGPT-4 mCLEAR score	4.70±0.90	Excellent	<.001	<.001	.270	.213	.001	.002	.868	
Bing mCLEAR score	4.16±1.43	Above average								
Bard mCLEAR score	4.00±1.41	Above average								

322 ¹mCLEAR: Modified CLEAR score based on the study by Sallam et al. [38]; ²SD: Standard deviation; ³Calculated using the Kruskal Wallis test. Significant P
 323 values are highlighted in bold style.

324

325

326 **Discussion**

327 The whole landscape of education, including higher education is set for a new era that can be
328 described as a paradigm shift with the widespread and popularity of AI [13, 53, 54]. In this
329 study, a comparison between the human and AI abilities in a highly specialized field at a high
330 level was undertaken. Specifically, the performance of MSc students in a Clinical Chemistry
331 exam, with an average score of 40.05 ± 7.23 (66.75%), was used as a benchmark for comparison.
332 Remarkably, ChatGPT-4 surpassed this human benchmark, achieving a score of 54/60
333 (90.00%). Bing followed with 46/60 (76.67%), outperforming both ChatGPT-3.5 (44/60,
334 73.33%) and Bard (40/60, 66.67%). Overall, the level of AI models' performance underlines
335 the advancements in AI capabilities. Additionally, these results could pave the way for a
336 broader scientific inquiry into both the potential role of AI in educational settings as well as
337 the usefulness of the current assessment tools in higher education.

338 In this study, the initial central hypothesis assumed that the human students at a postgraduate
339 level who undertook a specialized course in a highly specialized field, namely Clinical
340 Chemistry, would show a superior performance compared to the tested AI models. The findings
341 of this study showed that the AI models tested not only passed the exam but showed a
342 noteworthy performance. For example, ChatGPT-4 score equaled the highest student score and
343 thus would be rated as an "A" student. On the other hand, the performance of the AI models in
344 this study was not entirely an unexpected finding. This comes in light of the recent evidence
345 showing AI models' abilities to pass reputable exams in multiple languages such as the
346 USMLE [37], the German State Examination in Medicine [55], the National Medical Licensing
347 Examination in Japan [56, 57], and the Brazilian National Examination for Medical Degree
348 Revalidation [58].

349 From a broader perspective, a recent systematic review highlighted the abilities of ChatGPT as
350 an example of LLMs in various exams [35]. The review by Newton and Xiromeriti highlighted
351 the capabilities of this popular AI model, with ChatGPT-3 outperforming human students in
352 11% of the included exams, with ChatGPT 4 achieving superior performance and outscoring
353 the human performance in 35% of the included exams [35]. The current study findings were in
354 line with the finding of better GPT-4 performance as opposed to the earlier and free GPT-3.5
355 version. Yet, the performance of ChatGPT-4 in comparison to the human students was
356 noteworthy highlighting the refinements of LLMs over a short period of time.

357 In this study, analyzing the human students' performance based on the revised Bloom's
358 taxonomy enabled elucidation of deeper insights into the assessment of cognitive aspects. The
359 human students excelled in the "Remember" domain which is indicative of strong recalling and
360 recognizing abilities. Additionally, the human students demonstrated a high performance in the
361 "Understand" and "Apply" categories, with the lowest performance shown in the "Analyze"
362 category. The lack of statistical significance in these differences suggest a balanced level of
363 cognitive skills acquired among the students during the course despite the potential for
364 improvement in higher-order cognitive skills entailing breakdown and organization of acquired
365 knowledge.

366 On the other hand, the study findings revealed an interesting observation manifested in worse
367 AI models' performance across the higher cognitive domains. This observation stands in
368 contrast to the findings of Herrmann-Werner et al., which pioneered the use of the Bloom's
369 taxonomy in AI model performance in MCQs [44]. Herrmann-Werner et al. demonstrated a
370 lower level of ChatGPT performance in the lower cognitive skills in contrast to the findings of
371 this study [44]. To the contrary, a recent study that assessed ChatGPT-3 performance in
372 medical microbiology MCQs showed a trend similar to our findings where the AI model

373 performed at a higher level in the lower cognitive domains [45]. This divergence of findings
374 suggests the need for more comprehensive studies to discern the abilities of AI models in
375 different cognitive domains, which would be helpful to guide improvements in these models
376 and to enhance their utility in higher education.

377 Upon examining the performance of AI models in this study based on the MCQ metrics (FI,
378 DE, stem and choices word count), a significant drop in performance was noted in Bing and
379 Bard for more difficult MCQs. This finding suggests that some AI models have yet to show
380 evolution into the level where it can handle complex queries. The absence of a correlation
381 between MCQ stem and choice word counts and AI performance indicates that the challenge
382 was not related to the length of the queries but rather in the inherent complexity of the prompts.

383 In this study, the use of the validated CLEAR tool for assessment of the quality of AI generated
384 content presented a robust approach [38]. The rating of ChatGPT-4 as “Excellent” across all
385 categories of completeness, accuracy/evidence, and appropriateness/relevance serves as a clear
386 demonstration of its superiority. The Bing’s — which uses similar GPT-4 architecture — rating
387 as “Excellent” in appropriateness/relevance was a noteworthy finding; nevertheless, the
388 performance of this Microsoft AI model did not match ChatGPT-4 in terms of completeness
389 and accuracy. The other AI models in this study were rated as “Above average” based on the
390 modified CLEAR tool. This result, albeit lower than ChatGPT-4, still showed the huge
391 potential of these freely available models, but with an evident room for improvement. The
392 significant superiority of ChatGPT-4 over the other AI models tested in this study highlights
393 the swift evolution of AI capabilities [59].

394 In the field of higher education, the implications of the study findings can be profound. The
395 noteworthy capabilities of AI models, especially those shown by ChatGPT-4, to outperform
396 humans at a postgraduate level could serve as a red flag necessitating the re-evaluation of

397 traditional assessment approaches currently utilized for evaluation of students' achievement of
398 learning outcomes [53, 60]. Additionally, the study findings highlighted the current possible
399 AI limitations in addressing higher-order cognitive tasks, which shows the unique value of
400 human critical thinking and analytical skills [61]. Nevertheless, more studies are needed to
401 confirm this finding based on a recent evidence showing the satisfactory performance of
402 ChatGPT in tasks requiring higher-order thinking specifically in the field of medical
403 biochemistry as shown by Ghosh and Bir [62].

404 Future research could focus on investigating the feasibility of integrating AI into higher
405 education frameworks in terms of utilizing an approach that could augment the human learning
406 (e.g., through enhancing personalized learning experience and providing instantaneous
407 feedback) without compromising the development of critical thinking and analytical skills [5,
408 9, 53, 63-65]. Additionally, the ethical considerations of academic integrity should be
409 considered in light of opportunities of academic dishonesty posed by AI models in educational
410 settings [66-68]. This issue also extends to warrant a thorough investigation into the
411 implications of possible decline in students' analytical and critical thinking skills and
412 prioritizing the human needs and value [27, 69, 70].

413 Finally, while the current study can provide valuable insights into the performance of AI
414 models compared to human students in the context of Clinical Chemistry topic, several
415 limitations should be considered when interpreting the results. Future research in this area
416 would benefit from addressing these limitations that included: First, this study employed a
417 limited dataset of 60 MCQs. This limited number of MCQs inherently restricts the scope of
418 performance evaluation. Second, the use of the CLEAR tool, albeit standardized, introduces a
419 subjective element in evaluating the content generated by AI models. This subjectivity could
420 lead to a potential bias in the assessment of AI responses if approached by different raters.

421 Thus, the AI content evaluation was not entirely devoid of subjective judgment despite the use
422 of key answers to reduce this subjectivity bias. Third, the exclusive concentration on Clinical
423 Chemistry as a subject is both a strength and a limitation. While it allowed for a deep insight
424 this specific health discipline, it limits the generalizability of the findings to other academic
425 fields, since different subjects may present unique challenges that were not addressed in this
426 study. Fourth, LLMs are evolving rapidly, and this study only provided a snapshot of AI
427 models' performance at a specific time point. Therefore, this study may not fully represent the
428 potential improvements or advancements in AI capabilities that have occurred or may occur
429 shortly after the study period. Fifth, the exam metrics, derived from the performance of a
430 limited number of students ($n=20$), might have been influenced by various external factors.
431 These include the format of the exam and its time limits and the specific cohort of students.

432 In conclusion, the current study provided a comparative analysis of the human versus AI
433 performance in a highly specialized academic context at the postgraduate level. The results
434 could motivate future research to address the possible role of AI in higher education reaping
435 its benefits while avoiding its limitations. The ideal approach would be to use the strengths of
436 AI as a complement to the unique capabilities of human intellect. This can ensure the evolution
437 of the educational process in an innovative way aiding in students' intellectual development.
438 Importantly, the study results call for a revision of the current assessment tools in higher
439 education with a focus on improving the assessment of higher cognitive skills.

440

441

442

443

444 **Acknowledgments**

445 NA.

446 **Funding**

447 We declare that we received no funding nor financial support/grants by any institutional,
448 private, or corporate entity.

449 **Conflicts of Interest**

450 We declare that we have no competing interest nor conflict of interest.

451 **Data availability statement**

452 The data that support the findings of this study are available on request from the
453 corresponding author (M.S.). The data are not publicly available due to the confidentiality of
454 the questions created for an exam purposes.

455 **Author contribution**

456 **Conceptualization:** Malik Sallam

457 **Data Curation:** Malik Sallam, Khaled Al-Salahat

458 **Formal Analysis:** Malik Sallam

459 **Investigation:** Malik Sallam, Khaled Al-Salahat, Huda Eid, Jan Egger, Behrus Puladi

460 **Methodology:** Malik Sallam, Khaled Al-Salahat, Huda Eid, Jan Egger, Behrus Puladi

461 **Project administration:** Malik Sallam

462 **Supervision:** Malik Sallam

463 **Visualization:** Malik Sallam

464 **Writing – Original Draft Preparation:** Malik Sallam

465 **Writing – Review & Editing:** Malik Sallam, Khaled Al-Salahat, Huda Eid, Jan Egger,

466 Behrus Puladi

467

468 **References**

- 469 1. Chiu TKF. Future research recommendations for transforming higher education with
470 generative AI. *Computers and Education: Artificial Intelligence*. 2023;In Press:100197. doi:
471 10.1016/j.caeai.2023.100197.
- 472 2. Rawas S. ChatGPT: Empowering lifelong learning in the digital age of higher education.
473 *Education and Information Technologies*. 2023. doi: 10.1007/s10639-023-12114-8.
- 474 3. Rahiman HU, Kodikal R. Revolutionizing education: Artificial intelligence empowered
475 learning in higher education. *Cogent Education*. 2024;11(1):2293431. doi:
476 10.1080/2331186X.2023.2293431.
- 477 4. Crompton H, Burke D. Artificial intelligence in higher education: the state of the field.
478 *International Journal of Educational Technology in Higher Education*. 2023;20(1):22. doi:
479 10.1186/s41239-023-00392-8.
- 480 5. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The Advent of Generative
481 Language Models in Medical Education. *JMIR Med Educ*. 2023;9:e48163. Epub 20230606. doi:
482 10.2196/48163.
- 483 6. Rodway P, Schepman A. The impact of adopting AI educational technologies on projected
484 course satisfaction in university students. *Computers and Education: Artificial Intelligence*.
485 2023;5:100150. doi: 10.1016/j.caeai.2023.100150.
- 486 7. Giansanti D. The Chatbots Are Invading Us: A Map Point on the Evolution, Applications,
487 Opportunities, and Emerging Problems in the Health Domain. *Life [Internet]*. 2023; 13(5):[1130 p.].
- 488 8. Dempere J, Modugu K, Hesham A, Ramasamy LK. The impact of ChatGPT on higher
489 education. *Frontiers in Education*. 2023;8:1206936. doi: 10.3389/feduc.2023.1206936.
- 490 9. Sallam M, Salim NA, Barakat M, Al-Tammemi AB. ChatGPT applications in medical,
491 dental, pharmacy, and public health education: A descriptive study highlighting the advantages and
492 limitations. *Narra J*. 2023;3(1):e103. doi: 10.52225/narra.v3i1.103.
- 493 10. Sáiz-Manzanares MC, Marticorena-Sánchez R, Martín-Antón LJ, González Díez I, Almeida
494 L. Perceived satisfaction of university students with the use of chatbots as a tool for self-regulated
495 learning. *Heliyon*. 2023;9(1):e12843. Epub 20230113. doi: 10.1016/j.heliyon.2023.e12843.
- 496 11. Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: systematic literature
497 review. *International Journal of Educational Technology in Higher Education*. 2023;20(1):56. doi:
498 10.1186/s41239-023-00426-1.
- 499 12. Imran M, Almusharraf N. Analyzing the role of ChatGPT as a writing assistant at higher
500 education level: A systematic review of the literature. *Contemporary Educational Technology*.
501 2023;15(4):ep464. doi: 10.30935/cedtech/13605.
- 502 13. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic
503 Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel)*. 2023;11(6):887. Epub
504 20230319. doi: 10.3390/healthcare11060887.
- 505 14. Kooli C. Chatbots in Education and Research: A Critical Examination of Ethical Implications
506 and Solutions. *Sustainability [Internet]*. 2023; 15(7):[5614 p.].

- 507 15. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges,
508 bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*. 2023;3:121-
509 54. doi: 10.1016/j.iotcps.2023.04.003.
- 510 16. Grassini S. Shaping the Future of Education: Exploring the Potential and Consequences of AI
511 and ChatGPT in Educational Settings. *Education Sciences [Internet]*. 2023; 13(7):[692 p.].
- 512 17. Kamalov F, Santandreu Calonge D, Gurrib I. New Era of Artificial Intelligence in Education:
513 Towards a Sustainable Multifaceted Revolution. *Sustainability [Internet]*. 2023; 15(16):[12451 p.].
- 514 18. von Garrel J, Mayer J. Artificial Intelligence in studies—use of ChatGPT and AI-based tools
515 among students in Germany. *Humanities and Social Sciences Communications*. 2023;10(1):799. doi:
516 10.1057/s41599-023-02304-7.
- 517 19. Sallam M, Salim NA, Barakat M, Al-Mahzoum K, Al-Tammemi AB, Malaeb D, et al.
518 Assessing Health Students' Attitudes and Usage of ChatGPT in Jordan: Validation Study. *JMIR Med*
519 *Educ*. 2023;9:e48254. Epub 20230905. doi: 10.2196/48254.
- 520 20. Abdaljaleel M, Barakat M, Alsanafi M, Salim NA, Abazid H, Malaeb D, et al. Factors
521 Influencing Attitudes of University Students towards ChatGPT and its Usage: A Multi-National Study
522 Validating the TAME-ChatGPT Survey Instrument. *Research Square*. 2023. doi: 10.21203/rs.3.rs-
523 3400248/v1.
- 524 21. Malik AR, Pratiwi Y, Andajani K, Numertayasa IW, Suharti S, Darwis A, et al. Exploring
525 Artificial Intelligence in Academic Essay: Higher Education Student's Perspective. *International*
526 *Journal of Educational Research Open*. 2023;5:100296. doi: 10.1016/j.ijedro.2023.100296.
- 527 22. Rodríguez JMR, Montoya MSR, Fernández MB, Lara FL. Use of ChatGPT at university as a
528 tool for complex thinking: Students' perceived usefulness. *NAER: Journal of New Approaches in*
529 *Educational Research*. 2023;12(2):323-39. doi: 10.7821/naer.2023.7.1458.
- 530 23. Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: Ensuring academic integrity in
531 the era of ChatGPT. *Innovations in Education and Teaching International*. 2023:1-12. doi:
532 10.1080/14703297.2023.2190148.
- 533 24. Bin-Nashwan SA, Sadallah M, Bouteraa M. Use of ChatGPT in academia: Academic
534 integrity hangs in the balance. *Technology in Society*. 2023;75:102370. doi:
535 10.1016/j.techsoc.2023.102370.
- 536 25. Birks D, Clare J. Linking artificial intelligence facilitated academic misconduct to existing
537 prevention frameworks. *International Journal for Educational Integrity*. 2023;19(1):20. doi:
538 10.1007/s40979-023-00142-3.
- 539 26. Hasanein AM, Sobaih AEE. Drivers and Consequences of ChatGPT Use in Higher
540 Education: Key Stakeholder Perspectives. *Eur J Invest Health Psychol Educ*. 2023;13(11):2599-614.
541 Epub 20231109. doi: 10.3390/ejihpe13110181.
- 542 27. Ahmad SF, Han H, Alam MM, Rehmat MK, Irshad M, Arraño-Muñoz M, et al. Impact of
543 artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities*
544 *and Social Sciences Communications*. 2023;10(1):311. doi: 10.1057/s41599-023-01787-8.
- 545 28. George B, Wooden O. Managing the Strategic Transformation of Higher Education through
546 Artificial Intelligence. *Administrative Sciences*. 2023;13(9):196. doi: 10.3390/admsci13090196.

- 547 29. Roll I, Wylie R. Evolution and Revolution in Artificial Intelligence in Education.
548 International Journal of Artificial Intelligence in Education. 2016;26(2):582-99. doi: 10.1007/s40593-
549 016-0110-3.
- 550 30. Chan CKY. A comprehensive AI policy education framework for university teaching and
551 learning. International Journal of Educational Technology in Higher Education. 2023;20(1):38. doi:
552 10.1186/s41239-023-00408-3.
- 553 31. Liu M, Ren Y, Nyagoga LM, Stonier F, Wu Z, Yu L. Future of education in the era of
554 generative artificial intelligence: Consensus among Chinese scholars on applications of ChatGPT in
555 schools. Future in Educational Research. 2023;1(1):72-101. doi: 10.1002/fer3.10.
- 556 32. McCarthy AM, Maor D, McConney A, Cavanaugh C. Digital transformation in education:
557 Critical components for leaders of system change. Social Sciences & Humanities Open.
558 2023;8(1):100479. doi: 10.1016/j.ssaho.2023.100479.
- 559 33. Spector JM, Ma S. Inquiry and critical thinking skills for the next generation: from artificial
560 intelligence back to human intelligence. Smart Learning Environments. 2019;6(1):8. doi:
561 10.1186/s40561-019-0088-z.
- 562 34. Essel HB, Vlachopoulos D, Essuman AB, Amankwa JO. ChatGPT effects on cognitive skills
563 of undergraduate students: Receiving instant responses from AI-based conversational large language
564 models (LLMs). Computers and Education: Artificial Intelligence. 2024;6:100198. doi:
565 10.1016/j.caeai.2023.100198.
- 566 35. Newton PM, Xiromeriti M. ChatGPT performance on MCQ exams in higher education. A
567 pragmatic scoping review. EdArXiv. 2023;Preprint. doi: 10.35542/osf.io/sytu3.
- 568 36. Sallam M, Barakat M, Sallam M. METRICS: Establishing a Preliminary Checklist to
569 Standardize Design and Reporting of Artificial Intelligence-Based Studies in Healthcare. JMIR
570 Preprints. 2023;Preprint. doi: 19/11/2023:54704.
- 571 37. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of
572 ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.
573 PLOS Digit Health. 2023;2(2):e0000198. Epub 20230209. doi: 10.1371/journal.pdig.0000198.
- 574 38. Sallam M, Barakat M, Sallam M. Pilot Testing of a Tool to Standardize the Assessment of the
575 Quality of Health Information Generated by Artificial Intelligence-Based Models. Cureus.
576 2023;15(11):e49373. Epub 20231124. doi: 10.7759/cureus.49373.
- 577 39. Douglas M, Wilson J, Ennis S. Multiple-choice question tests: a convenient, flexible and
578 effective learning tool? A case study. Innovations in Education and Teaching International.
579 2012;49(2):111-21. doi: 10.1080/14703297.2012.677596.
- 580 40. Bloom BS, Krathwohl DR. Taxonomy of Educational Objectives: The Classification of
581 Educational Goals: Longmans, Green; 1956. 403 p.
- 582 41. Seaman M. BLOOM'S TAXONOMY: Its Evolution, Revision, and Use in the Field of
583 Education. Curriculum and Teaching Dialogue. 2011;13(1/2):29-131A.
- 584 42. Liu Q, Wald N, Daskon C, Harland T. Multiple-choice questions (MCQs) for higher-order
585 cognition: Perspectives of university teachers. Innovations in Education and Teaching International.
586 2023;1-13. doi: 10.1080/14703297.2023.2222715.

- 587 43. Karanja E, Malone LC. Improving project management curriculum by aligning course
588 learning outcomes with Bloom's taxonomy framework. *Journal of International Education in*
589 *Business*. 2021;14(2):197-218. doi: 10.1108/JIEB-05-2020-0038.
- 590 44. Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, et
591 al. Assessing ChatGPT's Mastery of Bloom's Taxonomy using psychosomatic medicine exam
592 questions. medRxiv. 2023;Preprint:2023.08.18.23294159. doi: 10.1101/2023.08.18.23294159.
- 593 45. Sallam M, Al-Salahat K. Below average ChatGPT performance in medical microbiology
594 exam compared to university students. *Frontiers in Education*. 2023;8:1333415. doi:
595 10.3389/feduc.2023.1333415.
- 596 46. OpenAI. GPT-3.5 2023 [cited 2023 27 November 2023]. Available from: <https://openai.com/>.
- 597 47. Microsoft, OpenAI. Bing is your AI-powered copilot for the web 2023 [cited 2023 27
598 November 2023]. Available from:
599 <https://www.bing.com/search?q=Bing+AI&showconv=1&FORM=hpcodx>.
- 600 48. Google. Bard 2023 [cited 2023 27 November 2023]. Available from:
601 <https://bard.google.com/chat>.
- 602 49. Burtis CA, Ashwood ER, Bruns DE, Tietz NW. *Tietz textbook of clinical chemistry and*
603 *molecular diagnostics*. 5th ed. St. Louis, Mo.: Saunders; 2013. xviii, 2,238 p. p.
- 604 50. Bishop ML, Fody EP, Schoeff LE. *Clinical chemistry : principles, techniques, and*
605 *correlations*. Eighth edition. ed. Philadelphia: Wolters Kluwer; 2018. xxviii, 736 pages p.
- 606 51. McPherson RA, Pincus MR. *Henry's clinical diagnosis and management by laboratory*
607 *methods*. 24. ed. Philadelphia: Elsevier; 2021. pages cm p.
- 608 52. Sallam M, Al-Salahat K, Al-Ajlouni E. ChatGPT Performance in Diagnostic Clinical
609 Microbiology Laboratory-Oriented Case Scenarios. *Cureus*. 2023;15(12):e50629. Epub 20231216.
610 doi: 10.7759/cureus.50629.
- 611 53. Lo CK. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature.
612 *Education Sciences* [Internet]. 2023; 13(4):[410 p.].
- 613 54. Sallam M, Salim NA, Al-Tammemi AB, Barakat M, Fayyad D, Hallit S, et al. ChatGPT
614 Output Regarding Compulsory Vaccination and COVID-19 Vaccine Conspiracy: A Descriptive Study
615 at the Outset of a Paradigm Shift in Online Search for Information. *Cureus*. 2023;15(2):e35029. Epub
616 20230215. doi: 10.7759/cureus.35029.
- 617 55. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT
618 Passes German State Examination in Medicine With Picture Questions Omitted. *Dtsch Arztebl Int*.
619 2023;120(21):373-4. doi: 10.3238/arztebl.m2023.0113.
- 620 56. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on Medical
621 Questions in the National Medical Licensing Examination in Japan: Evaluation Study. *JMIR Form*
622 *Res*. 2023;7:e48023. Epub 20231013. doi: 10.2196/48023.
- 623 57. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the
624 Japanese Medical Licensing Examination: Comparison Study. *JMIR Med Educ*. 2023;9:e48002. Epub
625 20230629. doi: 10.2196/48002.

- 626 58. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, Belfort R, Jr. Performance
627 of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree
628 Revalidation. *Rev Assoc Med Bras* (1992). 2023;69(10):e20230848. Epub 20230925. doi:
629 10.1590/1806-9282.20230848.
- 630 59. Hofmann Hayden L, Guerra Gage A, Le Jonathan L, Wong Alexander M, Hofmann Grady H,
631 Mayfield Cory K, et al. The Rapid Development of Artificial Intelligence: GPT-4's Performance on
632 Orthopedic Surgery Board Questions. *Orthopedics*. 2023;0(0):1-5. doi: 10.3928/01477447-20230922-
633 05.
- 634 60. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE
635 shines a spotlight on the flaws of medical education. *PLOS Digital Health*. 2023;2(2):e0000205. doi:
636 10.1371/journal.pdig.0000205.
- 637 61. Zhai X, Nyaaba M, Ma W. Can AI Outperform Humans on Cognitive-demanding Tasks in
638 Science? *SSRN*. 2023;Preprint. doi: 10.2139/ssrn.4451722.
- 639 62. Ghosh A, Bir A. Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the
640 Competency-Based Medical Education Curriculum in Medical Biochemistry. *Cureus*.
641 2023;15(4):e37023. Epub 20230402. doi: 10.7759/cureus.37023.
- 642 63. Tlili A, Shehata B, Adarkwah MA, Bozkurt A, Hickey DT, Huang R, et al. What if the devil
643 is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning
644 Environments*. 2023;10(1):15. doi: 10.1186/s40561-023-00237-x.
- 645 64. Dai W, Lin J, Jin H, Li T, Tsai YS, Gašević D, et al., editors. Can Large Language Models
646 Provide Feedback to Students? A Case Study on ChatGPT. 2023 IEEE International Conference on
647 Advanced Learning Technologies (ICALT); 2023 10-13 July 2023.
- 648 65. Schleiss J, Laupichler MC, Raupach T, Stober S. AI Course Design Planning Framework:
649 Developing Domain-Specific AI Education Courses. *Education Sciences* [Internet]. 2023; 13(9):[954
650 p.].
- 651 66. Perkins M. Academic integrity considerations of AI Large Language Models in the post-
652 pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*. 2023;20.
653 doi: 10.53761/1.20.02.07.
- 654 67. Memarian B, Doleck T. ChatGPT in education: Methods, potentials, and limitations.
655 *Computers in Human Behavior: Artificial Humans*. 2023;1(2):100022. doi:
656 10.1016/j.chbah.2023.100022.
- 657 68. Saylam S, Duman N, Yildirim Y, Satsevich K. Empowering education with AI: Addressing
658 ethical concerns. *London Journal of Social Sciences*. 2023;(6):39-48. doi: 10.31039/ljss.2023.6.103.
- 659 69. Grájeda A, Burgos J, Córdova P, Sanjinés A. Assessing student-perceived impact of using
660 artificial intelligence tools: Construction of a synthetic index of application in higher education.
661 *Cogent Education*. 2024;11(1):2287917. doi: 10.1080/2331186X.2023.2287917.
- 662 70. Hadi Mogavi R, Deng C, Juho Kim J, Zhou P, D. Kwon Y, Hosny Saleh Metwally A, et al.
663 ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization
664 and perceptions. *Computers in Human Behavior: Artificial Humans*. 2024;2(1):100027. doi:
665 10.1016/j.chbah.2023.100027.

666

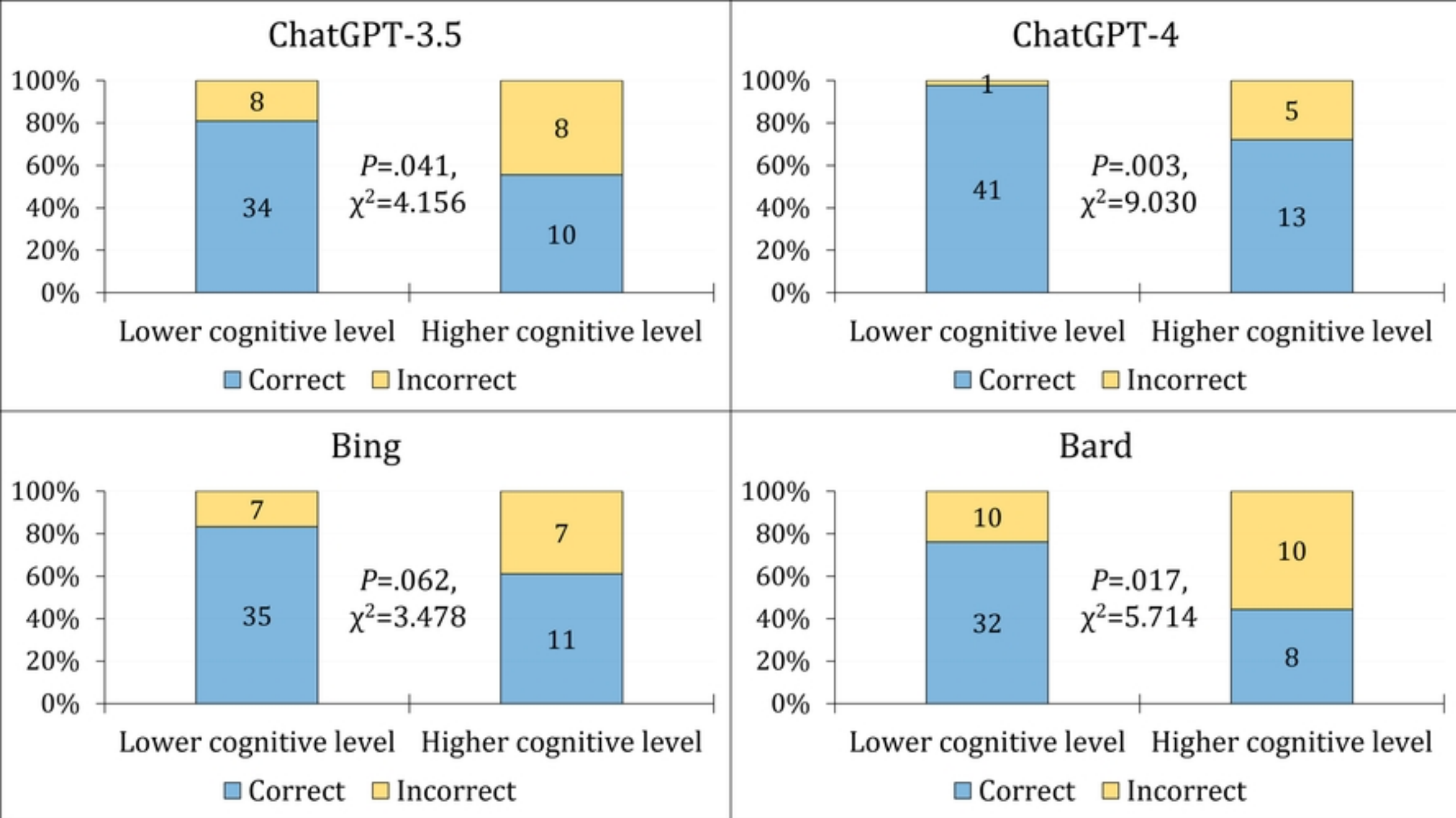


Figure 1