

# Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models

Tiffany H. Kung<sup>1,2</sup>; Morgan Cheatham<sup>3</sup>; ChatGPT<sup>4</sup>; Arielle Medenilla<sup>1</sup>; Czarina Sillos<sup>1</sup>; Lorie De Leon<sup>1</sup>; Camille Elepaño<sup>1</sup>; Maria Madriaga<sup>1</sup>; Rimel Aggabao<sup>1</sup>; Giezel Diaz-Candido<sup>1</sup>; James Maningo<sup>1</sup>; Victor Tseng<sup>\*1,5</sup>

## Author Affiliations:

<sup>1</sup>AnsibleHealth, Inc (Mountain View, CA)

<sup>2</sup>Department of Anesthesiology, Massachusetts General Hospital, Harvard School of Medicine (Boston, MA)

<sup>3</sup>Warren Alpert Medical School; Brown University (Providence, RI)

<sup>4</sup>OpenAI, Inc; (San Francisco, CA)

<sup>5</sup>Department of Medical Education, UWorld, LLC (Dallas, TX)

*\*Indicates corresponding author*

## Corresponding Author Information:

Victor Tseng, MD

Medical Director, Pulmonology

Ansible Health, Inc

229 Polaris Avenue, Ste 10

Mountain View, CA, 94043

Office Phone: (404) 595-7948

Email: [victor@ansiblehealth.com](mailto:victor@ansiblehealth.com)

**Running Title:** *ChatGPT and Medical Education*

**Subject Codes:** artificial intelligence; clinical decision support; medical education; standardized testing; ChatGPT; large language model; machine learning

**Word Count:** 3786 (Main Text: 3342)

**Main Figures:** 3    **Tables:** 0

**Supplementary Figures and Tables:** 2

## GLOSSARY OF NONSTANDARD ABBREVIATIONS

43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80

<b>ACI</b>	Accuracy-Concordance-Insight scoring system
<b>DOI</b>	Density of insight
<b>GPT</b>	Generative pretrained transformer
<b>LLM</b>	Large language model
<b>MCSA</b>	Multiple choice single answer
<b>MC-J</b>	Multiple choice single answer with forced justification
<b>MC-NJ</b>	Multiple choice single answer without forced justification
<b>NLP</b>	Natural language processing
<b>OE</b>	Open-ended question formulation
<b>Qn.m</b>	Question <i>n</i> , input run <i>m</i>
<b>USMLE</b>	United States Medical Licensing Exam

## ABSTRACT

81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117

We evaluated the performance of a large language model called ChatGPT on the United States Medical Licensing Exam (USMLE), which consists of three exams: Step 1, Step 2CK, and Step 3. ChatGPT performed at or near the passing threshold for all three exams without any specialized training or reinforcement. Additionally, ChatGPT demonstrated a high level of concordance and insight in its explanations. These results suggest that large language models may have the potential to assist with medical education, and potentially, clinical decision-making.

## INTRODUCTION

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

Over the past decade, advances in neural networks, deep learning, and artificial intelligence (AI) have transformed the way we approach a wide range of tasks and industries ranging from manufacturing and finance to consumer products. The ability to build highly accurate classification models rapidly and regardless of input data type (e.g. images, text, audio) has enabled widespread adoption of applications such as automated tagging of objects and users in photographs<sup>1</sup>, near-human level text translation<sup>2</sup>, automated scanning in bank ATMs, and even the generation of image captions<sup>3</sup>.

While these technologies have made significant impacts across many industries, applications in clinical care remain limited. The proliferation of clinical free-text fields combined with a lack of general interoperability between health IT systems contribute to a paucity of structured, machine-readable data required for the development of deep learning algorithms. Even when algorithms applicable to clinical care are developed, their quality tends to be highly variable, with many failing to generalize across settings due to limited technical, statistical, and conceptual reproducibility<sup>4</sup>. As a result, the overwhelming majority of successful healthcare applications currently support back-office functions ranging from payor operations, automated prior authorization processing, and management of supply chains and cybersecurity threats. With rare exceptions – even in medical imaging – there are relatively few applications of AI directly used in widespread clinical care today.

The proper development of clinical AI models<sup>5</sup> requires significant time, resources, and more importantly, highly domain and problem-specific training data, all of which are in short supply in the world of healthcare. One of the key developments that enabled image-based AI in clinical imaging has been the ability of large general domain models to perform as well as, or even outperform, domain-specific models. This development has catalyzed significant AI activity in medical imaging, where otherwise it would be challenging to obtain sufficient annotated clinical images. Indeed today, Inception-V3 serves as

144 the basic foundation of many of the top medical imaging models currently published, ranging from  
145 ophthalmology<sup>5,6</sup>, pathology<sup>7</sup>, to dermatology<sup>8</sup>.

146

147 In the past three weeks, a new AI model called ChatGPT captured significant attention due to its ability to  
148 perform a diverse array of natural language tasks<sup>9</sup>. ChatGPT is a general Large Language Model (LLM)  
149 developed recently by OpenAI. While the previous class of AI models have primarily been Deep Learning  
150 (DL) models, which are designed to learn and recognize patterns in data, LLMs are a new type of AI  
151 algorithm trained to predict the likelihood of a given sequence of words based on the context of the  
152 words that come before it. Thus, if LLMs are trained on sufficiently large amounts of text data, they are  
153 capable of generating novel sequences of words never observed previously by the model, but that  
154 represent plausible sequences based on natural human language. ChatGPT is powered by GPT3.5, an  
155 LLM trained on the OpenAI 175B parameter foundation model and a large corpus of text data from the  
156 Internet via reinforcement and supervised learning methods. Anecdotal usage indicates that ChatGPT  
157 exhibits evidence of deductive reasoning and chain of thought, as well as long-term dependency skills.

158

159 In this study, we evaluate the performance of ChatGPT, a non-domain specific LLM, on its ability to  
160 perform clinical reasoning by testing its performance on questions from the United States Medical  
161 Licensing Examination (USMLE). The USMLE is a high-stakes, comprehensive three-step standardized  
162 testing program covering all topics in physicians' fund of knowledge, spanning basic science, clinical  
163 reasoning, medical management, and bioethics. The difficulty and complexity of questions is highly  
164 standardized and regulated, making it an ideal input substrate for AI testing. The examination is well-  
165 established, showing remarkably stable raw scores and psychometric properties over the previous ten  
166 years<sup>10</sup>. The Step 1 exam is typically taken by medical students who have completed two years of  
167 didactic and problem-based learning and focuses on basic science, pharmacology, and pathophysiology;  
168 medical students often spend approximately 300-400 hours of dedicated study time in preparation for this  
169 exam<sup>11</sup>. The Step 2CK exam is usually taken by fourth-year medical students who have additionally

170 completed 1.5 to 2 years of clinical rotations; it emphasizes clinical reasoning, medical management, and  
171 bioethics. The Step 3 exam is taken by physicians who generally have completed at least a 0.5 to 1 year  
172 of postgraduate medical education.

173

174 USMLE questions are textually and conceptually dense; text vignettes contain multimodal clinical data  
175 (i.e., history, physical examination, laboratory values, and study results) often used to generate  
176 ambiguous scenarios with closely-related differential diagnoses. Due to its linguistic and conceptual  
177 richness, we reasoned that the USMLE would serve as an excellent challenge for ChatGPT.

178

179 Our work aims to provide both qualitative and quantitative feedback on the performance of ChatGPT and  
180 assess its potential for use in healthcare.

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

## METHODS

203

204

205

*Artificial Intelligence:* ChatGPT (OpenAI; San Francisco, CA), is a large language model that uses self-

206

attention mechanisms and a large amount of training data to generate natural language responses to

207

text input in a conversational context. It is particularly effective at handling long-range dependencies and

208

generating coherent and contextually appropriate responses. ChatGPT is a server-contained language

209

model that is unable to browse or perform internet searches. Therefore, all responses are generated *in*

210

*situ*, based on the abstract relationship between words (“tokens”) in the neural network. This contrasts to

211

other chatbots or conversational systems that are permitted to access external sources of information

212

(e.g. performing online searches or accessing databases) in order to provide directed responses to user

213

queries.

214

215

*Input Source:* 376 publicly-available test questions were obtained from the June 2022 sample exam

216

release on the official USMLE website. Random spot checking was performed to ensure that none of the

217

answers, explanations, or related content were indexed on Google prior to January 1, 2022, representing

218

the last date accessible to the ChatGPT training dataset. All sample test questions were screened, and

219

questions containing visual assets such as clinical images, medical photography, and graphs were

220

removed. After filtering, 305 USMLE items (Step 1: 93, Step 2CK: 99, Step 3: 113) were advanced to

221

encoding.

222

223

*Encoding:* Questions were formatted into three variants and input into ChatGPT in the following

224

sequence:

225

1. Open-ended (OE) format: Created by removing all answer choices, adding a variable lead-in

226

interrogative phrase. This format simulates free input and a natural user query pattern.

227

2. Multiple choice single answer without forced justification (MC-NJ): Created by reproducing the

228

original USMLE question verbatim.

- 229 3. Multiple choice single answer with forced justification (MC-J): Created by adding a variable lead-  
230 in imperative or interrogative phrase mandating ChatGPT to provide a rationale for each answer  
231 choice.

232  
233 Encoders employed deliberate variation in the lead-in prompts to avoid systematic errors that could be  
234 caused by stereotyped wording. To reduce memory retention bias, a new chat session was started in  
235 ChatGPT for each entry. Post-hoc analyses were performed to exclude systematic variation by encoder  
236 (data not shown).

237  
238 *Adjudication:* AI outputs were independently scored for Accuracy, Concordance, and Insight by two  
239 physician adjudicators using the rubric provided in **Supplemental Table 1**. To minimize within-item  
240 anchoring bias, adjudicators scored Accuracy for all items, followed by Concordance for all items,  
241 followed by Insight for all items. To minimize interrater cross-contamination, Physician 1 adjudicated  
242 Accuracy while Physician 2 adjudicated Concordance, and so forth. If consensus was not achieved for all  
243 three domains, the item was referred to a final physician adjudicator. Only 11 items (3.6% of the dataset)  
244 required arbitration.

245  
246 A schematic overview of the experimental protocol is provided in **Figure 1**.

247

248

249

250

251

252



253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284

## DATA AVAILABILITY

The data analyzed in this study were obtained from USMLE sample questions sets which are publicly available. The question index, raw inputs, and raw AI outputs are available in the **Online Data Supplement**. Inquiries and requests for additional dataset items and adjudication results can be provided upon reasonable request by contacting Victor Tseng, MD ([victor@ansiblehealth.com](mailto:victor@ansiblehealth.com)).

## RESULTS

### ChatGPT yields moderate accuracy approaching passing performance on USMLE

Exam items were first encoded as open-ended questions with variable lead-in prompts. This input format simulates a free natural user query pattern. With indeterminate responses censored/included, ChatGPT accuracy for USMLE Steps 1, 2CK, and 3 was 68.0%/42.9%, 58.3%/51.4%, and 62.4%/55.7%, respectively (**Figure 2A**).

Next, exam items were encoded as multiple choice single answer questions with no forced justification (MC-NJ). This input is the verbatim question format presented to test-takers. With indeterminate responses censored/included, ChatGPT accuracy for USMLE Steps 1, 2CK, and 3 was 55.1%/36.1%, 59.1%/56.9%, and 60.9%/54.9%, respectively.

Finally, items were encoded as multiple choice single answer questions with forced justification of positive and negative selections (MC-J). This input format simulates insight-seeking user behavior. With indeterminate responses censored/included, ChatGPT accuracy was 62.3%/ 40.3%, 51.9%/48.6%, and 64.6%/59.8%, respectively (**Figure 2B**).

### ChatGPT demonstrates high internal concordance

Concordance was independently adjudicated by two physician reviewers by inspection of the explanation content. Overall, ChatGPT outputted answers and explanations with 94.6% concordance across all questions. High global concordance was sustained across all exam levels, and across OE, MC-NJ, and MC-J question input formats (**Figure 3A**).

Next, we analyzed the contingency between accuracy and concordance in MC-J responses. ChatGPT was forced to justify its answer choice preference, and to defend its rejection of alternative choices.

314 Concordance amongst accurate responses was nearly perfect, and significantly greater than amongst  
315 inaccurate responses (99.1% vs. 85.1%,  $p < 0.001$ ) (**Figure 3B**).

316  
317 These data indicate that ChatGPT exhibits very high answer-explanation concordance, likely reflecting  
318 high internal consistency in its probabilistic language model.

319  
320 **Generative insight offered by ChatGPT may assist the human learner**

321  
322 Having established the accuracy and concordance of ChatGPT, we next examined its potential to  
323 augment human learning in the domain of medical education. AI-generated explanations were  
324 independently adjudicated by 2 physician reviewers. Explanation content was examined for significant  
325 insights, defined as instances that met the criteria (see **Supplemental Table 1**) of *novelty*,  
326 *nonobviousness*, and *validity*. The perspective of the target test audience was adopted by the  
327 adjudicator, as a second-year medical student for Step 1, fourth-year medical student for Step 2CK, and  
328 post-graduate year 1 resident for Step 3.

329  
330 Overall, ChatGPT produced at least one significant insight in 88.9% of all responses. The prevalence of  
331 insight was generally consistent between exam type and question input format (**Figure 3C**). In Step 2CK  
332 however, insight decreased by 10.3% ( $n = 11$  items) between MC-NJ and MC-J formulations. Review of  
333 this subset of questions did not reveal a discernible pattern for the paradoxical decrease (see  
334 **Supplemental Table 2B**).

335  
336 To quantify the density of insight (DOI) contained within AI-generated explanations, the number of unique  
337 insights was normalized to the number of possible answer choices. This analysis was performed on MC-  
338 J entries only. High quality outputs were generally characterized by DOI  $> 0.6$  (i.e. unique, novel,  
339 nonobvious, and valid insights provided for  $> 3$  out of 5 choices); low quality outputs were generally

340 characterized by DOI  $\leq 0.2$ . The upper limit on DOI is only bounded by the maximum length of text output.  
341 Across all exam types, we observed that DOI was significantly higher in questions items answered  
342 accurately versus inaccurately (0.458 versus 0.199%,  $p < 0.0001$ ) (**Figure 3D**).

343

344 These data indicate that a target human learner (e.g., such as a second-year medical student preparing  
345 for Step 1), if answering incorrectly, is likely to gain new or remedial insight from the ChatGPT AI output.  
346 Conversely, a human learner, if answering correctly, is less likely, but still able to access additional  
347 insight.

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

## DISCUSSION

366

367

368

In this study, we provide new and surprising evidence that ChatGPT is able to perform several intricate

369

tasks relevant to handling complex medical and clinical information. To assess ChatGPT's capabilities

370

against biomedical and clinical questions of *standardized* complexity and difficulty, we tested its

371

performance characteristics on the United States Medical Licensing Examination (USMLE).

372

373

Our findings can be organized into two major themes: (1) the rising accuracy of ChatGPT, which

374

approaches or exceeds the passing threshold for USMLE; and (2) the potential for this AI to generate

375

novel insights that can assist human learners in a medical education setting.

376

377

*The rising accuracy of ChatGPT:* The most recent iteration of the GPT LLM (GPT3) achieved 46%

378

accuracy with zero prompting<sup>12</sup>, which marginally improved to 50% with further model training. Previous

379

models, merely months prior, performed at 36.7%<sup>13</sup>. In this present study, ChatGPT performed at >50%

380

accuracy across all examinations, exceeding 60% in most analyses. The USMLE pass threshold, while

381

varying by year, is approximately 60%. Therefore, ChatGPT is now comfortably within the passing range.

382

Being the first experiment to reach this benchmark, we believe this is a surprising and impressive result.

383

Moreover, we provided no prompting or training to the AI, minimized grounding bias by expunging the AI

384

session prior to inputting each question variant, and avoided chain-of-thought biasing by requesting

385

forced justification only as the final input. Further model interaction and prompting could often produce

386

more accurate results (data not shown). Given this trajectory, it is likely that AI performance will continue

387

to rise as LLM models continue to mature.

388

389

Paradoxically, ChatGPT outperformed PubMedGPT (accuracy 50.8%, unpublished data), a counterpart

390

LLM with similar neural structure, but trained exclusively on biomedical domain literature. We speculate

391

that domain-specific training may have created greater ambivalence in the PubMedGPT model, as it

392 absorbs real-world text from ongoing academic discourse that tends to be inconclusive, contradictory, or  
393 highly conservative or noncommittal in its language. A foundation LLM trained on general content, such  
394 as ChatGPT, may therefore have an advantage because it is also exposed to broader clinical content,  
395 such as patient-facing disease primers and provider-facing drug package inserts, that are more definitive  
396 and congruent.

397

398 Consistent with the mechanism of generative language models, we observed that the accuracy of  
399 ChatGPT was strongly mediated by concordance and insight. High accuracy outputs were characterized  
400 by high concordance and high density of insight. Poorer accuracy was characterized by lower  
401 concordance and a poverty of insight. Therefore, inaccurate responses were driven primarily by missing  
402 information, leading to diminished insight and indecision in the AI, rather than overcommitment to the  
403 incorrect answer choice. These findings indicate that model performance could be significantly improved  
404 by merging foundation models, such as ChatGPT, with a domain-specific LLM or other model trained on  
405 a voluminous and highly validated medical knowledge resources, such as UpToDate, or other ACGME-  
406 accredited content.

407

408 Interestingly, the accuracy of ChatGPT tended to be lowest for Step 1, followed by Step 2CK, followed by  
409 Step 3. This mirrors both the subjective difficulty and objective performance for real-world test takers on  
410 Step 1, which is collectively regarded as the most difficult exam of the series. The low accuracy on Step  
411 1 could be explained by an undertrained model on the input side (e.g. underrepresentation of basic  
412 science content on the general information space) and/or the human side (e.g. insufficient or invalid  
413 human judgment at initial reinforcement stages). This result exposes a key vulnerability in pre-trained  
414 LLMs, such as ChatGPT: AI ability becomes yoked to human ability. ChatGPT's performance on Step 1  
415 is poorer precisely because human users perceive its subject matter (e.g., pathophysiology) as more  
416 difficult or opaque.

417

418 *The potential for AI-assisted human learning in medical education:* We also examined the ability of  
419 ChatGPT to assist the human learning process of its target audience (e.g., a second year medical  
420 student preparing for USMLE Step 1). As a proxy for the metric of helpfulness, we assessed the  
421 concordance and insight offered by the AI explanation outputs. ChatGPT responses were highly  
422 concordant, such that a human learner could easily follow the internal language, logic, and directionality  
423 of relationships contained within the explanation text (e.g., adrenal *hyper*cortisolism  $\Rightarrow$  *increased* bone  
424 osteoclast activity  $\Rightarrow$  *increased* calcium resorption  $\Rightarrow$  *decreased* bone mineral density  $\Rightarrow$  *increased* fracture  
425 risk). High internal concordance and low self-contradiction is a proxy of sound clinical reasoning and an  
426 important metric of explanation quality. It is reassuring that the directionality of relationships is preserved  
427 by the language processing model, where each verbal object is individually lemmatized.  
428  
429 AI-generated responses also offered significant insight, role-modeling a deductive reasoning process  
430 valuable to human learners (see **Supplemental Table 2**). At least one significant insight was present in  
431 approximately 90% of outputs. ChatGPT therefore possesses the partial ability to teach medicine by  
432 surfacing novel and nonobvious concepts that may not be in learners' sphere of awareness. This  
433 qualitative gain provides a basis for future real-world studies on the efficacy of generative AI to augment  
434 the human medical education process. For example, longitudinal exam performance can be studied in a  
435 quasi-controlled in AI-assisted and unassisted learners. Unit economic analysis may clarify the cost-  
436 effectiveness of incremental student performance gain in comparison to existing tools such as virtual  
437 tutors and study aids.  
438  
439 Medical education, licensing examinations, and test preparation services form a large industrial complex  
440 eclipsing a nine-figure market size annually. While its relevance remains debated, standardized testing  
441 has emerged as an important end-target of medical learning. In parallel, of the didactic techniques, a  
442 socratic teaching style is favored by medical students<sup>14</sup>. The rate-limiting step for fresh content

443 generation is the human cognitive effort required to craft realistic clinical vignettes that probe “high-yield”  
444 concepts in a subtle way, engage critical thinking, and offer pearls of knowledge even if answered  
445 incorrectly. Demand for new examination content continues to increase. For a national medical examiner,  
446 a single item typically requires 0.1 FTE work effort to produce (NBME, personal communication). Future  
447 studies may investigate the ability of generative language AI to offload this human effort by assisting in  
448 the question-explanation writing process or, in some cases, writing entire items autonomously.

449

450 Finally, the advent of AI in medical education demands an open science research infrastructure to  
451 standardize experimental methods, readouts, and benchmarks to describe and quantify human-AI  
452 interactions. Multiple dimensions must be covered, including user experience, learning environment,  
453 hybridization with other teaching modes, and effect on cognitive bias. In this report, we provide an initial  
454 basic protocol for adjudicating AI-generated responses along axes of accuracy, concordance, and  
455 insight.

456

457 Our study has several important limitations. The relatively small input size restricted the depth and range  
458 of analyses. For example, stratifying the output of ChatGPT by subject taxonomy (e.g., pharmacology,  
459 bioethics) or competency type (e.g., differential diagnosis, management) may be of great interest to  
460 medical educators, and could reveal heterogeneities in performance across language processing for  
461 different clinical reasoning tasks. Similarly, a more robust AI failure mode analysis (e.g., language  
462 parsing error) may lend insight into the etiology of inaccuracy and discordance. In addition to being  
463 laborious, human adjudication is error-prone and subject to greater variability and bias. Future studies  
464 will undoubtedly apply unbiased approaches, using quantitative natural language processing and text  
465 mining tools such as word network analysis. In addition to increasing validity and accelerating throughput  
466 by several orders of magnitude, these methods are likely to better characterize the depth, coherence,  
467 and learning value of AI output. Finally, to truly assess the utility of generative language AI for medical



468 education, ChatGPT and related applications must be studied in both controlled and real-world learning  
469 scenarios with students across the engagement and knowledge spectrum.

470

471 As AI becomes increasingly proficient, it will soon become ubiquitous, transforming clinical medicine  
472 across all healthcare sectors. Investigation of AI has now entered into the era of randomized controlled  
473 trials<sup>15</sup>. Additionally, a profusion of pragmatic and observational studies supports a versatile role of AI in  
474 virtually all medical disciplines and specialities by improving risk assessment<sup>16,17</sup>, data reduction, clinical  
475 decision support<sup>18,19</sup>, operational efficiency, and patient communication<sup>20,21</sup>.

476

477 Inspired by the remarkable performance of ChatGPT on the USMLE, clinicians within AnsibleHealth, a  
478 virtual chronic pulmonary disease clinic, have begun to experiment with ChatGPT as part of their  
479 workflows. Inputting queries in a secure and de-identified manner, our clinicians request ChatGPT to  
480 assist with traditionally onerous writing tasks such as composing appeal letters to payors, simplifying  
481 radiology reports (and other jargon-dense records) to facilitate patient comprehension, and even to  
482 brainstorm freely in a bid to kindle insight when faced with nebulous and diagnostically challenging  
483 cases. Overall, our clinicians reported a 33% decrease (future publication) in the time required to  
484 complete documentation and indirect patient care tasks. We believe this is an early but important signal  
485 that LLMs such as ChatGPT are reaching a maturity level that will soon impact clinical care at large and  
486 its ability to deliver truly individualized, compassionate, and scalable healthcare.

487

488

489

490

491

## ACKNOWLEDGEMENTS

492

The authors thank Dr. Kristine Vanijchroenkarn, MD and Ms. Audra Doyle RRT, NP for fruitful

493

discussions and technical assistance. We also thank Mr. Vangjush Vellahu for technical assistance with

494

graphical design and preparation.

495

496

## FUNDING

497

The work received no external funding.

498

499

## AUTHOR CONTRIBUTIONS

500

THK, MC, and VT conceived and designed the study, developed the study protocol, supervised the

501

research team, analyzed the data, and wrote the manuscript. AM, CS, LDL, CE, MM, DJC, and JM

502

encoded and input the data into ChatGPT. THK, VT, AM, and CS independently adjudicated the raw

503

ChatGPT outputs. JM and VT performed data synthesis, quality control, and statistical analyses.

504

ChatGPT contributed to the writing of several sections of this manuscript.

505

506

**Conflicts of Interest:** None

507

508

509

510

511

## FIGURE LEGENDS

512

513

514

515

### **Figure 1. Schematic of workflow for sourcing, encoding, and adjudicating results**

516

Abbreviations: **QC** = quality control; **MCSA-NJ** = multiple choice single answer without forced

517

justification; **MCSA-J** = multiple choice single answer with forced justification; **OE** = open-ended question

518

format

519

520

### **Figure 2. Accuracy of ChatGPT on USMLE**

521

For USMLE Steps 1, 2CK, and 3, AI outputs were adjudicated to be accurate, inaccurate, or

522

indeterminate based on the ACI scoring system provided in Supplemental Table 1.

523

**A:** Accuracy distribution for inputs encoded as open-ended questions

524

**B:** Accuracy distribution for inputs encoded as multiple choice single answer without (MC-NJ) or with

525

forced justification (MC-J)

526

527

### **Figure 3. Concordance and insight of ChatGPT on USMLE**

528

For USMLE Steps 1, 2CK, and 3, AI outputs were adjudicated on concordance and density of insight

529

(DOI) based on the ACI scoring system provided in Supplemental Table 1.

530

**A:** Overall concordance across all exam types and question encoding formats

531

**B:** Concordance rates stratified between accurate vs inaccurate outputs, across all exam types and

532

question encoding formats.  $p < 0.001$  for accurate vs inaccurate outputs by Fisher exact test

533

**C:** Overall insight prevalence, defined as proportion of outputs with  $\geq 1$  insight, across all exams for

534

questions encoded in MC-J format

535

**D:** DOI stratified between accurate vs inaccurate outputs, across all exam types for questions encoded in

536

MC-J format. Horizontal line indicates the mean.  $p$ -value determined by parametric 2-way ANOVA

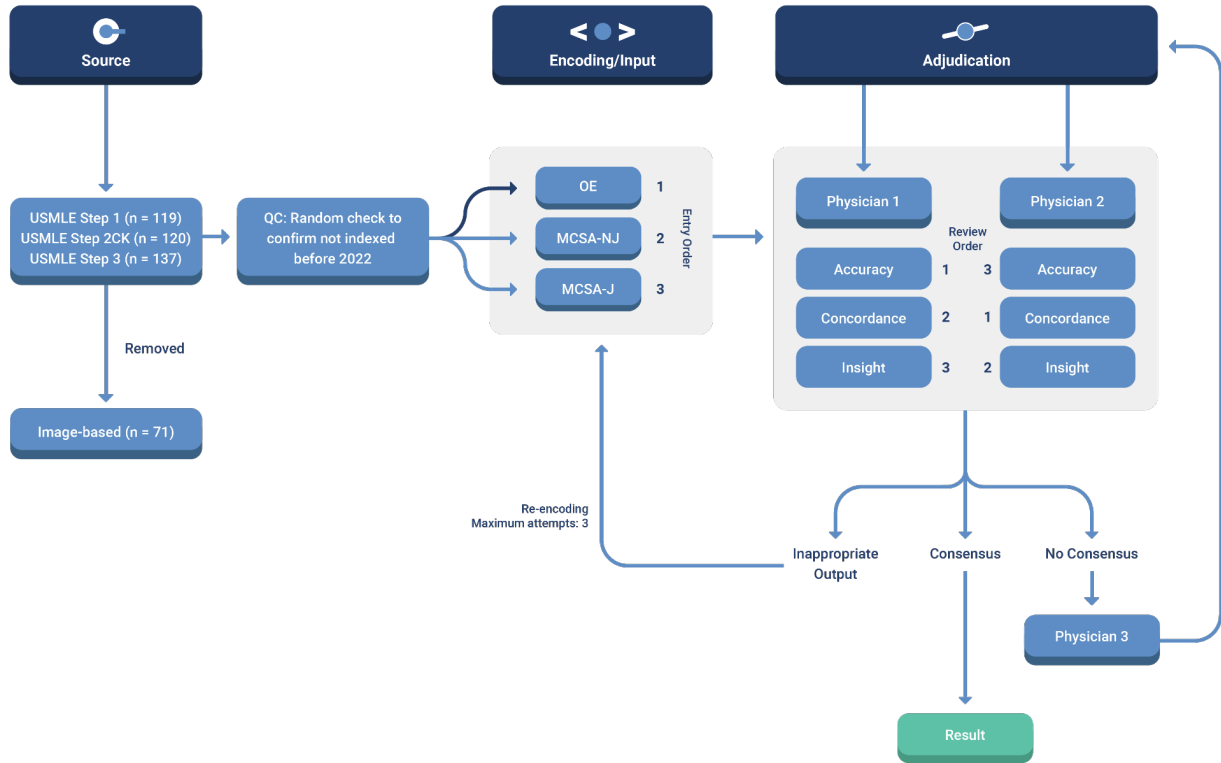
537

testing with Benjamini-Krieger-Yekutieli (BKY) *post hoc* to control for false discovery rate.

538

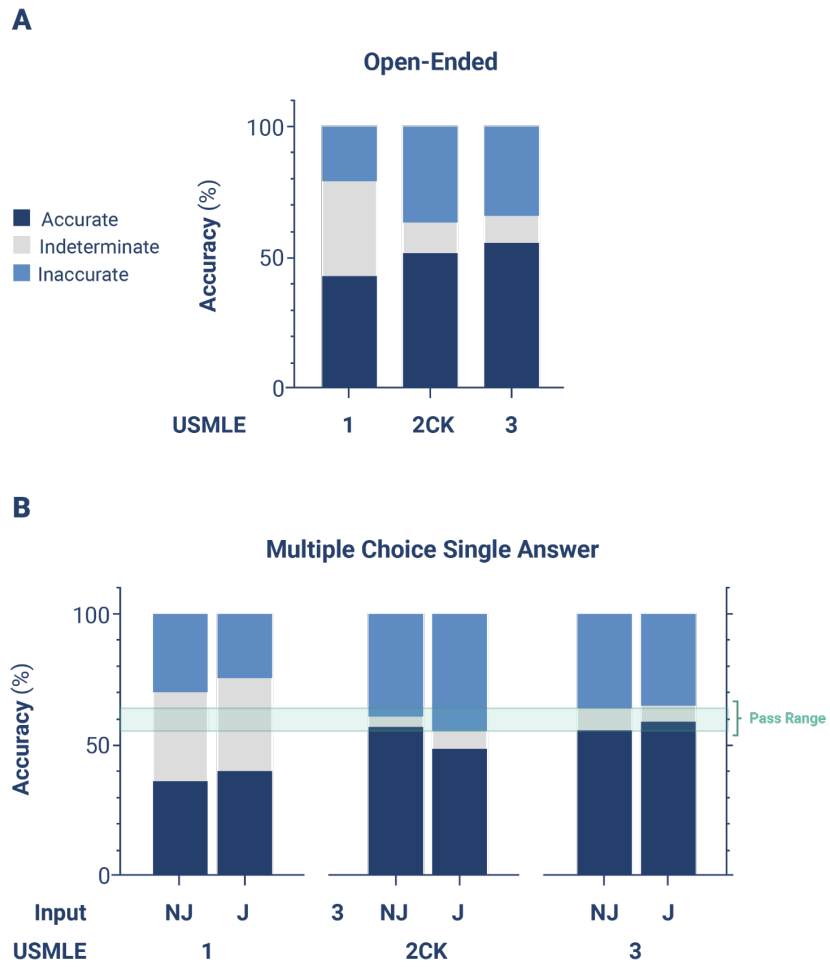
539  
540

## FIGURES AND TABLES



541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554

Figure 1



**Figure 2**

555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569

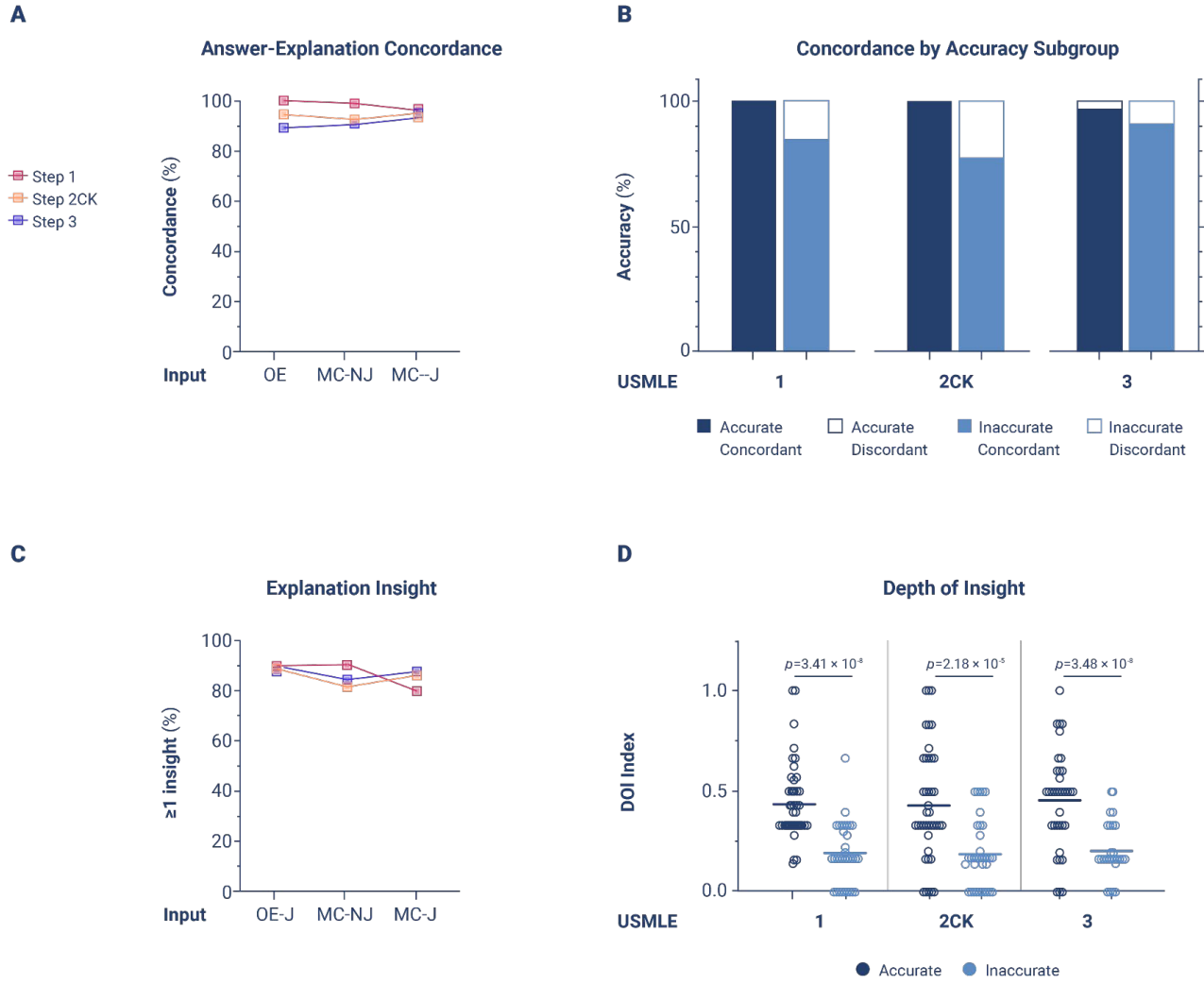


Figure 3

570  
571  
572  
573  
574  
575  
576  
577

578

579

<b>Adjudication Criteria: A-C-I Scoring System</b>	
<b>Accuracy</b>	<b>MC-NJ and MC-J</b>
	<ul style="list-style-type: none"> <li>● <b>Accurate:</b> Final answer matches the NBME key</li> <li>● <b>Inaccurate:</b> Incorrect answer choice is selected</li> <li>● <b>Indeterminate:</b> Response is not an answer choice, fails to select an answer, or claims that not enough information is available to commit to an answer</li> </ul>
	<b>OE</b>
	<ul style="list-style-type: none"> <li>● <b>Accurate:</b> Response identifies the correct concept, is specific, and is clinically sound</li> <li>● <b>Inaccurate:</b> Response targets an unrelated concept or is not clinically sound</li> <li>● <b>Indeterminate:</b> Any other response, including generic advice</li> </ul>
<b>Concordance</b>	<b>MC-J</b>
	<ul style="list-style-type: none"> <li>● <b>Concordant:</b> Explanation affirms the answer and negates <u>all</u> remaining choices</li> <li>● <b>Discordant:</b> <u>Any</u> part of explanation contradicts itself</li> </ul>
	<b>MC-NJ and OE</b>
	<ul style="list-style-type: none"> <li>● <b>Concordant:</b> Explanation affirms the answer</li> <li>● <b>Discordant:</b> <u>Any</u> part of explanation contradicts itself</li> </ul>
<b>Insight</b>	<p><b>Insight:</b> An instance of text in the explanation that is:</p> <ul style="list-style-type: none"> <li>● <i>Nondefinitional:</i> Does not simply define a term in the input question</li> <li>● <i>Unique:</i> A single insight may be used to eliminate several answer choices</li> <li>● <i>Nonobvious:</i> Requires deduction or knowledge external to the question input</li> <li>● <i>Valid:</i> In clinically or numerically accurate; preserves directionality</li> </ul> <p><b>Density of Insight (DOI):</b> <math>Number\ of\ insights / (number\ of\ answer\ choices + 1)</math></p>
	<ul style="list-style-type: none"> <li>● <b>Insightful:</b> DOI &gt;0 and offers a new concept or concept linkage</li> <li>● <b>Uninsightful:</b> DOI = 0</li> </ul>

580

581

582

583

584

**Supplemental Table 1**

585 **REFERENCES**

- 586 1. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture  
587 for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*  
588 Preprint at <https://doi.org/10.1109/cvpr.2016.308> (2016).
- 589 2. Zhang, W., Feng, Y., Meng, F., You, D. & Liu, Q. Bridging the Gap between Training and Inference  
590 for Neural Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for*  
591 *Computational Linguistics* Preprint at <https://doi.org/10.18653/v1/p19-1426> (2019).
- 592 3. Bhatia, Y., Bajpayee, A., Raghuvanshi, D. & Mittal, H. Image Captioning using Google’s Inception-  
593 resnet-v2 and Recurrent Neural Network. *2019 Twelfth International Conference on Contemporary*  
594 *Computing (IC3)* Preprint at <https://doi.org/10.1109/ic3.2019.8844921> (2019).
- 595 4. McDermott, M. B. A. *et al.* Reproducibility in machine learning for health research: Still a ways to go.  
596 *Sci. Transl. Med.* **13**, (2021).
- 597 5. Chen, P.-H. C., Liu, Y. & Peng, L. How to develop machine learning models for healthcare. *Nature*  
598 *Materials* vol. 18 410–414 Preprint at <https://doi.org/10.1038/s41563-019-0345-0> (2019).
- 599 6. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of  
600 Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).
- 601 7. Nagpal, K. *et al.* Development and validation of a deep learning algorithm for improving Gleason  
602 scoring of prostate cancer. *npj Digital Medicine* **2**, 1–10 (2019).
- 603 8. Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **26**, 900–  
604 908 (2020).
- 605 9. Website. <https://openai.com/blog/chatgpt/>.
- 606 10. Performance data. <https://www.usmle.org/performance-data>.
- 607 11. Burk-Rafel, J., Santen, S. A. & Purkiss, J. Study Behaviors and USMLE Step 1 Performance:  
608 Implications of a Student Self-Directed Parallel Curriculum. *Acad. Med.* **92**, S67–S74 (2017).
- 609 12. Liévin, V., Hother, C. E. & Winther, O. Can large language models reason about medical questions?



- 610 *arXiv [cs.CL]* (2022).
- 611 13. Jin, D. *et al.* What Disease does this Patient Have? A Large-scale Open Domain Question  
612 Answering Dataset from Medical Exams. *arXiv [cs.CL]* (2020).
- 613 14. Abou-Hanna, J. J., Owens, S. T., Kinnucan, J. A., Mian, S. I. & Kolars, J. C. Resuscitating the  
614 Socratic Method: Student and Faculty Perspectives on Posing Probing Questions During Clinical  
615 Teaching. *Acad. Med.* **96**, 113–117 (2021).
- 616 15. Plana, D. *et al.* Randomized Clinical Trials of Machine Learning Interventions in Health Care: A  
617 Systematic Review. *JAMA Netw Open* **5**, e2233946 (2022).
- 618 16. Kan, H. J. *et al.* Exploring the use of machine learning for risk adjustment: A comparison of standard  
619 and penalized linear regression models in predicting health care costs in older adults. *PLoS One* **14**,  
620 e0213258 (2019).
- 621 17. Delahanty, R. J., Kaufman, D. & Jones, S. S. Development and Evaluation of an Automated  
622 Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients.  
623 *Crit. Care Med.* **46**, e481–e488 (2018).
- 624 18. Vasey, B. *et al.* Reporting guideline for the early-stage clinical evaluation of decision support  
625 systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 (2022).
- 626 19. Garcia-Vidal, C., Sanjuan, G., Puerta-Alcalde, P., Moreno-García, E. & Soriano, A. Artificial  
627 intelligence to support clinical decision-making processes. *EBioMedicine* **46**, 27–29 (2019).
- 628 20. Bala, S., Keniston, A. & Burden, M. Patient Perception of Plain-Language Medical Notes Generated  
629 Using Artificial Intelligence Software: Pilot Mixed-Methods Study. *JMIR Form Res* **4**, e16670 (2020).
- 630 21. Milne-Ives, M. *et al.* The Effectiveness of Artificial Intelligence Conversational Agents in Health Care:  
631 Systematic Review. *J. Med. Internet Res.* **22**, e20346 (2020).