Check for updates

| Research Article

Chatbot responses suggest that hypothetical biology questions are harder than realistic ones

Gregory J. Crowther,[1] Usha Sankar,[2] Leena S. Knight,[3] Deborah L. Myers,[1] Kevin T. Patton,[4] Lekelia D. Jenkins,[5] Thomas A. Knight[3]

AUTHOR AFFILIATIONS See affiliation list on p. 13.

ABSTRACT The biology education literature includes compelling assertions that unfamiliar problems are especially useful for revealing students' true understanding of biology. However, there is only limited evidence that such novel problems have different cognitive requirements than more familiar problems. Here, we sought additional evidence by using chatbots based on large language models as models of biology students. For human physiology and cell biology, we developed sets of realistic and hypothetical problems matched to the same lesson learning objectives (LLOs). Problems were considered hypothetical if (i) known biological entities (molecules and organs) were given atypical or counterfactual properties (redefinition) or (ii) fictitious biological entities were introduced (invention). Several chatbots scored significantly worse on hypothetical problems than on realistic problems, with scores declining by an average of 13%. Among hypothetical questions, redefinition questions appeared especially difficult, with many chatbots scoring as if guessing randomly. These results suggest that, for a given LLO, hypothetical problems may have different cognitive demands than realistic problems and may more accurately reveal students' ability to apply biology core concepts to diverse contexts. The Test Question Templates (TQT) framework, which explicitly connects LLOs with examples of assessment questions, can help educators generate problems that are challenging (due to their novelty), yet fair (due to their alignment with pre-specified LLOs). Finally, ChatGPT's rapid improvement toward expert-level answers suggests that future educators cannot reasonably expect to ignore or outwit chatbots but must do what we can to make assessments fair and equitable.

KEYWORDS artificial intelligence (AI), Google Bard, Bloom's taxonomy, cheating, exams, HOCS/LOCS, summative assessment, YouChat

Most biology instructors want their students to go beyond memorization to think critically, reason scientifically, and solve problems (1, 2). This range of cognitive tasks is often discussed via the lens of Bloom's taxonomy (3, 4), which includes both lower-order cognitive skills (LOCS: knowledge and comprehension) and higher-order cognitive skills (HOCS: application, analysis, synthesis, and evaluation) (5, 6).

Among its other advantages, Bloom's taxonomy provides a convenient language to discuss the cognitive requirements of biological tasks (7–9). Nevertheless, different instructors may perceive these cognitive requirements differently due to, for example, different awareness of students' prior activities (6). Likewise, students may perceive questions differently than their instructors; one study found that for a set of 24 multiple-choice questions, only 63% of medical students agreed with the instructor (i.e., not much better than a coin flip) on whether each question was lower order or higher order (10). Such discrepancies have been noted in student interviews when students, who were asked to solve problems designed by instructors to elicit HOCS, often avoided HOCS and instead took an LOCS approach of simply recalling related facts from memory (11).

To circumvent students' overreliance on fact recall and encourage the application of information contained within the problems, Semsar et al. (11) found a solution in "us[ing] primarily novel and/or invented scenarios" (p. 24). As an example, a scenario about the familiar roles of angiotensin and aldosterone in regulating blood pressure was replaced with a novel scenario about fluid balance in mosquitos. Student interviews indicated that the novel scenarios were indeed more successful in eliciting HOCS (11). This qualitative result provides some empirical support for others' stated preferences for unfamiliar scenarios in summative assessments (e.g., 12–15).

We know of one previous quantitative study that directly compared biology students' scores on familiar problems and novel problems. Deane-Coe et al. (16) reported that students scored 7.5% lower on high-novelty/low-complexity questions (their category B) compared to low-novelty/low-complexity questions (their category A). However, there was no such difference between high-novelty/high-complexity and low-novelty/high-complexity questions (their categories D and C, respectively). Moreover, while the study's methods generally appear sound, it is not clear whether their category-A questions and category-B questions assessed the same lesson learning objectives (LLOs). For example, among the 64 questions used, four category-A questions asked about the meaning of R-code output, but no category-B questions did. Meanwhile, four category-B questions asked students whether various hypotheses were consistent with pilot data (on fish jaw anatomy) provided with the questions, but no category-A questions asked this.

When we attempted to analyze our own archived exam data as Deane-Coe et al. did (16), our attempts were similarly hampered by an inability to control for LLO. That is, our archived exams did not ask the same students a familiar question and a novel question both matched to the same LLO. In principle, though, such matching is achievable via the Test Question Templates (TQT) framework, which explicitly and purposefully links LLOs to specific examples of assessment questions (17–19). Since TQT LLOs are written as "Given X, do Y," one can identify LLOs where "X" could reasonably be either a familiar starting point or a novel one and then generate both familiar and novel questions for those LLOs.

We therefore aimed to supplement previous evidence about familiar and novel questions (11, 16) with a new quantitative study using TQTs to control for the context of LLO. To collect data under well-controlled conditions that avoid the confounding impacts of student and instructor variability, we used chatbots based on large language models (LLMs) as models of undergraduate students in two subdisciplines of biology (human physiology and cell biology). We posed the following research question: Do chatbots perform better on realistic (familiar) questions than on hypothetical (novel) questions matched to the same LLO? We hypothesized that the answer would be affirmative, which would strengthen previous suggestions that the two types of questions have different cognitive requirements, which in turn would have clear implications for student training and assessment.

METHODS

Rationale for testing chatbots

As instructors who are not experts in artificial intelligence (AI) or educational technology, we are interested primarily in the learning and capabilities of (human) students rather than chatbots. Given this position, we studied LLM-driven chatbots as imperfect but useful models of undergraduate students. Although chatbots do not "think" like human students, a chatbot trained on enormous amounts of text might be considered analogous to a student taking an open-book/open-Internet test. In this context, a chatbot may approximate an inexperienced human student who organizes information according to "surface features" rather than critical underlying structures, principles, and core concepts, as experts do (20, 21).

For the aim of conducting a well-controlled research study, using chatbots has certain practical advantages over using actual students. First, a chatbot's ability to answer specific questions can be studied free of the influence of prior instructor-mediated practice, a major concern of Semsar and Casagrand (6), if no prior instructor-mediated

practice is provided. Second, a chatbot can be asked hundreds of questions without losing motivation or needing feedback or compensation. Third, a chatbot's output is not complicated by heterogeneity within and between students facing myriad concurrent conditions (e.g., mental or physical illness, socioeconomic disadvantage, discrimination, and limited familiarity with English). Thus, as an alternative to gathering limited noisy data from actual students, we gathered rich data from a few chatbots (ChatGPT, YouChat, and Google Bard).

A final practical reason for testing chatbots is that since chatbots such as ChatGPT are already widely used by students, it behooves us to understand the chatbots' capabilities and limitations. Within a couple of months of the release of the GPT 3.5-based version of ChatGPT, over a third of students were already using it for coursework, according to one survey (22), and this fraction has undoubtedly even climbed higher since then. Thus, apart from the possibility that chatbots may be reasonable models of students (as proposed above), our characterization of chatbots should also help clarify the extent to which chatbots represent a threat to assessment integrity and/or a reason to change exam formats (23, 24).

Realistic questions vs. hypothetical questions

To study the effect of question novelty on chatbots, rather than students, some adjustments to assessment questions were necessary; a scenario that is novel to a student might nonetheless be familiar to a chatbot trained on enormous repositories of information. To ensure that questions would be novel to chatbots, we made them "hypothetical," that is, not consistent with known biological situations and thus unlikely to be represented in the chatbots' training data, as opposed to "realistic," that is, consistent with well-known biological situations. (Thus, for the remainder of this paper, we use the words "realistic" and "hypothetical" when referring to our specific study and "familiar" and "novel" when referring to the general issues of teaching and assessing students.) We considered questions to be hypothetical if they provided either (i) a redefinition (i.e., a novel context or role) of an existing biological entity (molecule or organ) or (ii) an invention of an unfamiliar (usually fictitious) biological entity. Examples of both subtypes of hypothetical questions, along with a realistic question matched to the same LLO, are shown in Fig. 1.

For clarity, note that we are not using "realistic" and "hypothetical" as synonyms for or direct analogs to "LOCS" and "HOCS," respectively, since we did not formally classify our questions by Bloom level, to which the terms "LOCS" and "HOCS" are usually aligned. We compared the realistic and hypothetical questions to test the hypothesis that these questions may have different cognitive requirements; however, the exact nature of any such difference might not correspond precisely to a difference in Bloom level or the LOCS/HOCS dividing line *per se*.

Creation of question banks for human physiology and cell biology

To create large banks of questions suitable for testing chatbots, we used TQTs, a framework for transparently aligning LLOs and assessments (17–19). Since TQTs make explicit the links between LLOs and specific questions, we used this framework to identify certain LLOs as being compatible with both realistic questions and hypothetical questions (Fig. 1). For each of two biology subdisciplines—human physiology and cell biology—we selected 22 to 25 suitable LLOs, and, for each LLO, we expanded our set of example questions to include four realistic questions and four hypothetical questions, all text-based (i.e., no images) and all in multiple-choice format. The number of choices per question varied (between 2 and 8) but was the same for each LLO-matched set of realistic questions and hypothetical questions. All questions had one and only one correct answer. Due to the varying numbers of choices, random guessing would yield scores of 32% for the human physiology questions and 22% for the cell biology questions. The summaries of each question bank are given in Tables 1 and 2, and the

Lesson Learning Objective (LLO):
Given information on electrical and chemical gradients, predict an ion's direction of flow (i.e., into the cell or out of the cell).

Realistic question: An animal's neurons have an extracellular $[Na^+]$ of 140 mM and an intracellular $[Na^+]$ of 15 mM. Now imagine a moment when the membrane potential (V_m) is -40 mV. If some Na^+ channels open at this moment, which way will the Na^+ ions flow? (a) into the cell (b) out of the cell **Correct answer:** (a).	**Hypothetical question (redefinition subtype):** Imagine an alien species whose neurons are governed by the same biophysical principles as ours but differ in their details. This alien's neurons have an extracellular $[Na^+]$ of 15 mM and an intracellular $[Na^+]$ of 140 mM. Now imagine a moment when the membrane potential (V_m) is +20 mV. If some Na^+ channels open at this moment, which way will the Na^+ ions flow? (a) into the cell (b) out of the cell **Correct answer:** (b).
	Hypothetical question (invention subtype): Imagine an alien species whose neurons are governed by the same biophysical principles as ours but differ in their details. This alien's neurons have an extracellular $[Br^-]$ of 110 mM and an intracellular $[Br^-]$ of 4 mM. Now imagine a moment when the membrane potential (V_m) is +10 mV. If some bromide channels open at this moment, which way will the Br^- ions flow? (a) into the cell (b) out of the cell **Correct answer:** (a).

FIG 1 An illustration of realistic versus hypothetical questions. In the redefinition type of hypothetical question, a familiar biological entity (Na^+ ions) has now been assigned an atypical property unlikely to be represented in the chatbots' training data (i.e., a higher concentration inside the cell than outside the cell). The invention type of hypothetical question, in contrast, introduces an ion (Br^-) that normally has no role at all in electrical signaling.

complete lists of LLOs and associated questions are given in Supplemental File S1 and Supplemental File S2.

As shown in Tables 1 and 2, most of our hypothetical questions were of the invention subtype (74 out of 100 in human physiology and 76 out of 88 in cell biology). Questions of the redefinition subtype are generally difficult to write since if a question redefines an existing biological entity, it must clearly specify which relevant properties are being altered and which are not, which can be difficult to do concisely.

We stress that these question banks were created for the specific purpose of answering this study's research question and may or may not be suitable for regular classroom exams. The study's unvarying use of text-based multiple-choice questions contrasts with our general preference for short-answer questions, often involving the interpretation of figures. Moreover, to complete our study efficiently during a time of rapid chatbot evolution, we vetted our questions less broadly than is done in multi-year developments of assessment instruments meant for widespread use (11, 12). In our streamlined approach, the first author (G.J.C.) was the primary writer and compiler of LLOs and questions; the human physiology question bank was reviewed, edited, and approved by two coauthors who regularly teach physiology (K.T.P. and U.S.); and the cell biology question bank was similarly checked by two authors who regularly teach cell biology (L.S.K. and D.L.M.). Thus, each question was ultimately approved by three coauthors with relevant expertise.

Administration of questions to chatbots

The chatbots used in this study are profiled in Table 3. We consider each chatbot version—a single product that is fundamentally stable, despite some flexibility in

TABLE 1 A summary of the human physiology question bank[a]

| LLO ID | Topic | Realistic questions | Hypothetical questions | |
			Redefinition questions	Invention questions
1.1	Homeostasis	4	1	3
1.2	Homeostasis	4	2	2
10.1	Muscle actions	4	0	4
10.2	Muscle actions	4	0	4
11.1 a	Neurons	4	2	2
11.1 c	Neurons	4	2	2
11.2b	Neurons	4	1	3
11.3	Neurons	4	2	2
12.1	Central nervous system tracts	4	0	4
16.1	Hormones	4	0	4
16.2 c	Hormones	4	0	4
16.2 f	Hormones	4	0	4
16.3 c	Hormones	4	0	4
16.3d	Hormones	4	0	4
17.1b	Blood transfusions	4	0	4
18.1 a	Heart	4	0	4
18.1b	Heart	4	0	4
18.2 a	Heart	4	0	4
18.2b	Heart	4	0	4
22.1	Gas transport	4	1	3
23.1	Digestive enzymes	4	4	0
23.2	Digestive enzymes	4	4	0
24.1	Micronutrients	4	0	4
25.1b	Nephrons	4	3	1
25.2b	Nephrons	4	4	0
	Total	100	26	74

[a]A complete list of LLOs and associated questions is given in Supplemental File S1. Note that the total number of hypothetical questions (redefinition questions + invention questions) matches the number of realistic questions for each LLO.

behavior—to be more analogous to an individual student than to an entire class of students, making our study reminiscent of case studies that collect rich data sets on a limited number of individuals (25). To underline this point, which escaped readers of early versions of this paper, we have given each chatbot version a human nickname (Table 3); however, for maximum clarity, results below refer to the Table 3 identifiers rather than the nicknames, with the four versions of ChatGPT designated as ChatGPT-a (oldest) through ChatGPT-d (newest).

Overall, we tested six chatbot versions from January to March 2023, using one version (ChatGPT-a) only for exploratory pilot testing, five versions for human physiology testing, and four versions for cell biology testing (Table 3). For all testing, the order of questions was randomized within each LLO. In rare instances where a chatbot did not select a single multiple-choice answer, it was asked the same question again until it made a clear choice. In general, chatbots were not asked to explain their answers but usually did so anyway. At no point were chatbots given feedback on their answers. Most rounds of testing occurred over several days due to constraints on authors' schedules and chatbot access (e.g., ChatGPT-d was limited to 25 questions per 3 hours, while YouChat froze after every ~50 questions for an hour). Finally, any question identified as ambiguous (<10% of questions used in testing before March 2023) was corrected and re-administered in

TABLE 2 A summary of the cell biology question bank[a]

LLO ID	Topic	Realistic questions	Hypothetical questions	
			Redefinition questions	Invention questions
2.1	Atomic structure	4	0	4
3.1	Proteins	4	0	4
3.2	Biomolecules	4	0	4
3.3	Proteins	4	0	4
3.4	Proteins	4	0	4
4.1	Protein transport	4	0	4
4.2L	Cytoskeleton	4	0	4
5.1	Cell metabolism	4	0	4
5.2	Membrane transport	4	0	4
5.3	Membrane transport	4	0	4
6.1	Cell metabolism	4	0	4
6.2L	Cell metabolism	4	0	4
8.1	Meiosis/mitosis	4	0	4
9.1 a	Blood-type genetics	4	0	4
9.1b	Blood-type genetics	4	0	4
10.1b	Frameshift mutations	4	4	0
10.2	Central dogma	4	4	0
10.3	Central dogma	4	4	0
10.4	Central dogma	4	0	4
11.1	Gene expression	4	0	4
11.2	Gene expression	4	0	4
11.3L	Cell signaling	4	0	4
	Total	88	12	76

[a]A complete list of LLOs and associated questions is given in Supplemental File S2. Note that the total number of hypothetical questions (redefinition questions + invention questions) matches the number of realistic questions for each LLO.

a new round of testing, with previous responses to that question thrown out. (Notably, while some ambiguities were noticed via coauthor review, others were identified when a chatbot's "incorrect" response was noticed to actually be consistent with the phrasing of the question at the time. We therefore encourage instructors to consider using chatbots as a tool for checking their assessment questions for clarity as well as for difficulty.)

TABLE 3 Chatbots used in this study[a]

Identifier (nickname)	ChatGPT-a ("Chau")	ChatGPT-b ("Chet")	ChatGPT-c ("Chita")	ChatGPT-d ("Chuck")	YouChat ("Yusuf")	Bard ("Barb")
Chatbot version	ChatGPT, 9 January version	ChatGPT, 30 January version	ChatGPT, 13 February version	ChatGPT Plus, 4 March version	YouChat 2.0	Bard
LLM	GPT-3.5	GPT-3.5	GPT-3.5	GPT-4	Chat, apps, and links (CAL)	Language model for dialogue applications (LaMDA)
Maker of chatbot	OpenAI (major investor: Microsoft)	OpenAI (major investor: Microsoft)	OpenAI (major investor: Microsoft)	OpenAI (major investor: Microsoft)	You.com	Google
Cost at time of testing	Free	Free	Free	$20 per month	Free	Free
Dates tested	16–29 January 2023 (pilot testing only)	31 January–9 February 2023	15 February–25 March 2023	14–25 March 2023	1–28 March 2023	22–25 March 2023
Subject areas tested	Human physiology	Human physiology	Human physiology, cell biology	Human physiology, cell biology	Human physiology, cell biology	Human physiology, cell biology

[a]LLM, Large Language Model; GPT, Generative Pre-trained Transformer.

In our initial round of formal testing, we asked ChatGPT-b our human physiology questions. Upon observing a difference in ChatGPT-b's scores on realistic and hypothetical questions, we expanded the study in three ways. First, we added questions in another field: cell biology. Second, we added other chatbots available to the public at little to no cost (Table 1). Third, we started asking each chatbot each question 3–5 times, rather than only once or twice, in order to more accurately assess its "knowledge." (While this approach would not typically be taken with a human student, it is somewhat analogous to an oral examination that circles back to previous questions in order to check whether the answers given are internally consistent.)

Statistical analysis

This study's principal goal was to test for differences in chatbot performance on realistic versus hypothetical questions. For each LLO, we calculated separate scores for the four realistic questions and the four hypothetical questions, averaging all repeats. For example, on a given set of four questions, if a chatbot got two questions right on all five attempts, one question right on two of five attempts, and one question wrong on all five attempts, its score would be 2.4 out of 4 or 60%. For scoring purposes, questions were marked as right or wrong based on the letter chosen, irrespective of any explanations offered. We used paired two-tailed t-tests (Microsoft Excel) to compare scores on the realistic and hypothetical questions across all 25 or 22 LLOs, with alpha set at 0.05.

To determine whether a chatbot's performance changed (i) during 3–5 consecutive attempts at the same question and/or (ii) during eight questions about the same LLO, we fit linear equations (Microsoft Excel) to (i) chatbot score versus attempt number and (ii) chatbot score versus question number. We then used one-sample two-tailed t-tests to check whether each chatbot's set of 22 or 25 slopes was significantly different from 0.

IRB approval

The student quotation used in the Discussion section was gathered in a separate study (19) approved by the Institutional Review Board of Everett Community College.

RESULTS

Since our primary expertise and interests are in science education, we did not characterize the chatbots' functions as AI researchers might do. However, we did evaluate three basic assumptions that underpinned our goal of using chatbots as models of students: chatbots understand the multiple-choice format, we can ask a chatbot a given question multiple times and take each response as an independent readout of its "understanding," and chatbots can support their multiple-choice answers with good reasons.

Chatbots readily answer multiple-choice questions

When we asked chatbots our multiple-choice questions, they unambiguously chose a single answer with a frequency of 93% (YouChat), 96% (ChatGPT-c), 98% (ChatGPT-d), or 99% (Bard; this information was not recorded for ChatGPT-b). For the remaining 1%–7% of questions, the chatbot either reported that it did not have enough information to answer the question or picked 2–3 answers rather than a single option. All chatbots answered correctly at greater-than-chance frequencies, except for the redefinition subtype of hypothetical questions (see below; due to varying numbers of choices per question, random chance would have yielded scores of 32% and 22% for the human physiology and cell biology questions, respectively).

Without feedback, chatbots appear to answer each question independently

Since we asked the chatbots large numbers of related questions, we wondered whether such cumulative exposure might result in improvement or "learning" over time, despite a lack of feedback. We explored this issue at two levels of granularity. First, if a chatbot is

asked the same question 3–5 times in a row, does it score better on the later attempts? Second, as a chatbot answers an LLO's set of eight related questions, does it score better on the later questions? For all chatbots tested, the answer to both questions was no in both human physiology and cell biology. When scores were regressed against attempt number, there was no relationship (slopes were not different from 0, $P > 0.05$). Additionally, we found no relationship between score and question number (slopes not different from 0, $P > 0.05$) regardless of whether each LLO's realistic and hypothetical questions were analyzed together or separately.

Chatbots usually explain their correct answers well

Chatbots usually included an explanation of one to three paragraphs in their responses even when no explanation was requested. We therefore tested whether two chatbots' explanations provided accurate information relevant to and consistent with their letter choices, indicating associations that might appear as appropriate reasoning, as we might expect from human students.

First, we examined ChatGPT-c's full responses to human physiology questions between 15 February and 6 March 2023. Of the 200 multiple-choice questions attempted, 127 were answered correctly. Of the explanations accompanying those 127 correct answers, 118 (93%) were entirely correct (with no errors) or mostly correct (with only minor errors), and 9 were judged incorrect. (We did not systematically examine explanations accompanying incorrect multiple-choice answers.) Explanations were similarly successful for the realistic questions and the hypothetical questions. For the realistic questions, of the 76 correct multiple-choice answers, 70 came with correct or largely correct explanations (minor flaws or no flaws), and 6 had incorrect explanations. For the hypothetical questions, of the 51 correct multiple-choice answers, 48 came with correct or largely correct explanations, and 3 had incorrect explanations.

Second, we examined YouChat's full responses to cell biology questions on 1 March 2023, which yielded similar results. Here, of the 160 multiple-choice questions attempted, 79 were answered correctly. Of these, 7 lacked explanations; of the other 72, 62 (86%) had predominantly or entirely correct explanations, while 10 had incorrect explanations. For the realistic questions, of the 39 correct multiple-choice answers with explanations, 36 of the explanations were correct or predominantly correct, and 3 were incorrect. For the hypothetical questions, out of 33 correct multiple-choice answers with explanations, 26 explanations were largely or entirely correct, and 7 were incorrect.

Chatbots score better on realistic questions than on hypothetical questions

As noted above, this study's main research question was as follows: Do chatbots perform better on realistic questions than on hypothetical questions matched to the same LLO? The answer was an overall "yes" across multiple rounds of testing chatbots on both human physiology questions (Table 4) and cell biology questions (Table 5).

We tested the performance of five chatbot versions on 25 LLO-matched sets of human physiology questions. Four versions (ChatGPT-b, ChatGPT-c, YouChat, and Bard) scored significantly higher on the realistic questions than on the hypothetical questions (Table 4; examples of responses to realistic and hypothetical questions matched by LLO are shown in Fig. 2). The difference in scores of the most recent ChatGPT version, ChatGPT-d, did not reach statistical significance ($P = 0.079$).

We also tested the performance of four chatbot versions on 22 LLO-matched sets of cell biology questions. Three versions (ChatGPT-c, YouChat, and Bard) returned significantly more correct answers on the realistic questions than on the hypothetical questions (Table 5). As above, the difference in scores of ChatGPT-d was not statistically significant ($P = 0.15$).

Overall, these results show that an average drop in score due to questions being hypothetical rather than realistic was 15.5% for human physiology (i.e., chatbots averaged 69.7% on the realistic questions vs. 54.2% on the hypothetical questions) and 9.5% for cell biology (74.6% vs. 65.1%).

TABLE 4 Chatbots' performance on human physiology questions[a]

Chatbot identifier	ChatGPT-b	ChatGPT-c	ChatGPT-d	YouChat	Bard
Dates tested	31 January–9 February 2023	9–11 March 2023	14–18 March 2023	8–28 March 2023	22 March 2023
Attempts per question	2	5	5	5	3
Score on realistic questions	61.3% ± 21.3%	75.8% ± 20.4%	86.4% ± 16.7%	71.8% ± 22.5%	53.0% ± 25.1%
Score on hypothetical questions	46.4% ± 22.4%	52.0% ± 22.2%	79.0% ± 20.2%	51.6% ± 20.7%	41.8% ± 22.2%
Difference in scores	14.9%	23.8%	7.4%	20.2%	11.2%
Paired two-tailed t-test	$P = 0.008$	$P < 0.0001$	$P = 0.079$	$P < 0.0001$	$P = 0.025$

[a]Percentages listed for realistic questions and hypothetical questions are means ± standard deviations.

Among hypothetical questions, chatbots score better on invention questions than on redefinition questions

Eight human physiology LLOs had hypothetical questions of both subtypes (Table 1), permitting a small-scale comparison between these subtypes. Three of the five chatbots tested scored significantly lower on the redefinition questions than on the invention questions (Table 6). The exceptions were ChatGPT-d, which scored well on both subtypes, and Bard, which scored poorly on both (Table 6). Since the cell biology questions did not permit a similar comparison and since the overall number of redefinition questions was quite low, our finding that redefinition questions are harder should be considered preliminary and is in need of further testing.

ChatGPT improved rapidly during this study

During February and March 2023, we were able to study three different versions of ChatGPT (Table 1). ChatGPT's human physiology scores improved rapidly over this time period (Fig. 3), with ChatGPT-d showing especially large gains in hypothetical question scores. Similar trends were evident in the cell biology scores of ChatGTP-c and ChatGPT-d (Table 5).

DISCUSSION

This study employed LLM-driven chatbots as imperfect but useful models of biology students. While chatbots do not "think" like students, they readily fielded hundreds of questions apiece, yielding data that are unobtainable in typical classroom settings. These data showed that aside from the most advanced version of ChatGPT available to us, all

TABLE 5 Chatbots' performance on cell biology questions[a]

Chatbot identifier	ChatGPT-c	ChatGPT-d	YouChat	Bard
Dates tested	8–25 March 2023	18–25 March 2023	10–25 March 2023	23–25 March 2023
Attempts per question	5	5	5	3
Score on realistic questions	81.8% ± 17.7%	88.4% ± 20.3%	71.4% ± 23.2%	56.8% ± 27.1%
Score on hypothetical questions	69.1% ± 22.4%	84.1% ± 20.6%	61.8% ± 25.3%	45.5% ± 24.9%
Difference in scores	12.7%	4.3%	9.6%	11.3%
Paired two-tailed t-test	$P < 0.0001$	$P = 0.15$	$P = 0.034$	$P = 0.003$

[a]Values were calculated and reported as in Table 4.

LLO 16.3d: Given information on the chemical structures of (real or hypothetical) hormones, predict the primary location of the hormones' receptors (at the cell surface or intracellular).

Realistic question 16.3d-R4: Imagine that a newly discovered mammal has hormones that include progesterone, whose chemical formula is C21H30O2. In this mammal, progesterone's chemical structure would lead to the prediction that its target-cell receptors are _____.

(a) at the surface of the cell
(b) intracellular

Correct answer: (b).

ChatGPT-b answer (Feb. 9, 2023): (b). Progesterone is a steroid hormone, which means it has a hydrophobic structure that allows it to diffuse through the cell membrane and bind to intracellular receptors. The chemical formula of progesterone indicates that it contains multiple hydrophobic regions, including a long hydrocarbon chain and several ring structures, which are characteristic of steroid hormones. As a result, the target-cell receptors for progesterone are typically located inside the cell, rather than on the cell surface.

Hypothetical question 16.3d-H4: Imagine that a newly discovered mammal has hormones that include prolapsine, whose chemical formula is C21H29N2. In this mammal, prolapsine's chemical structure would lead to the prediction that its target-cell receptors are _____.

(a) at the surface of the cell
(b) intracellular

Correct answer: (b).

ChatGPT-b answer (Feb. 9, 2023): (a). The chemical structure of a hormone, such as prolapsine, can give some indication of its target cell receptors. Hormones with a hydrophobic (water-repelling) structure, like prolapsine, typically bind to receptors located at the surface of the cell. This allows the hormone to cross the cell membrane and interact directly with the target cell's interior. Intracellular receptors, on the other hand, are located within the cell and are typically targeted by hydrophilic (water-soluble) hormones that cannot cross the cell membrane.

FIG 2 An example of ChatGPT-b correctly answering a realistic question while incorrectly answering a very similar hypothetical question (invention subtype).

chatbots tested scored significantly lower on hypothetical questions than on realistic questions matched to the same LLOs. Our findings support the development and use of more novel questions to prompt students to transfer their knowledge to new scenarios.

The chatbots we tested generally impressed us with their apparent knowledge of biology. Their frequent success in answering difficult multiple-choice questions and their often-lucid supporting explanations suggest that these chatbots can model the kind of understanding that we want our students to acquire. Our results are broadly consistent with previous reports of chatbots' fluency with content from undergraduate biology (23, 24), the Medical College Admission Test (MCAT) (26), and medical school (27, 28). For instance, a ChatGPT version based on LLM GPT-3.5 (likely equivalent to our ChatGPT-a, ChatGPT-b, or ChatGPT-c) performed at or above median scores on the MCAT (26), and the scores we report here are roughly comparable to those (e.g., ~75% correct for questions from the MCAT section that are most analogous to our question banks, Biological and Biochemical Foundations of Living Systems). Our results are also roughly similar to those from a report (27) on neurosurgery board preparation exam questions, for which a GPT-3.5-based ChatGPT, a GPT-4-based ChatGPT (equivalent to our ChatGPT-d), and Bard had overall scores of 62.4%, 84.6%, and 44.2%, respectively.

Despite the chatbots' frequent virtuosity, switching from realistic questions to hypothetical ones in our study lowered their scores by an average of 13 percentage points—an effect that, if applied to student exams, would often correspond to a drop of 1.3 letter grades (e.g., from a B-plus to a C). Our finding that only ChatGPT-d did about equally well on hypothetical and realistic questions mirrored the neurosurgery board preparation exam question study, in which only the GPT4-based ChatGPT did about equally well on higher-order and lower-order questions (27).

Our study's distinction between "realistic" and "hypothetical" questions bears some (possibly misleading) similarity to a distinction between "abstract" and "applied" questions in a 20-question homeostasis concept inventory (29). In that study, McFarland et al. classified nine questions as abstract; six of these concerned the general meaning

TABLE 6 Chatbots' scores on the redefinition subtype and the invention subtype of hypothetical human physiology questions[a]

Chatbot identifier	ChatGPT-b	ChatGPT-c	ChatGPT-d	YouChat	Bard
Score on invention questions	88%	64%	88%	64%	44%
	± 35%	± 31%	± 21%	± 31%	± 36%
Score on redefinition questions	23%	19%	73%	23%	33%
	± 25%	± 21%	± 23%	± 34%	± 35%
Difference in scores	62%	45%	15%	41%	11%
Paired two-tailed t-test	$P = 0.009$	$P = 0.02$	$P = 0.17$	$P = 0.03$	$P = 0.5$

[a]This table shows a subset of the data reported in Table 4, covering only the human physiology LLOs that included both subtypes of hypothetical questions.

of terms like "control center" (question #16) and "effector" (question #13), and three concerned the hypothetical regulation of metabolite X in the blood of a new species of deer (questions #2–4). Therefore their "abstract" category is quite different from our "hypothetical" category, while their "applied" category corresponds closely to our "realistic" category, so their finding of similar scores on abstract questions and applied questions (29) is not directly comparable to our finding of different scores on hypothetical questions and realistic questions. Nonetheless, McFarland et al. offer relevant insight into the ways in which question formats may influence student responses (29). They cite a prior claim that concrete problems containing specific details may be especially hard for students because the details may appear to conflict with prior knowledge and/or "may trigger application of inaccurate mental models" (p. 3). In this light, it makes sense that our questions on which the chatbots scored worst were the "redefinition" questions, in which details were changed so as to directly clash with prior knowledge.

Taken together, our results and those of others (11, 16) have the practical implication that novel problems do indeed offer a unique window to student understanding. The general task of applying previously learned information to new contexts is known in the cognitive psychology literature by many terms, including HOCS (discussed above),
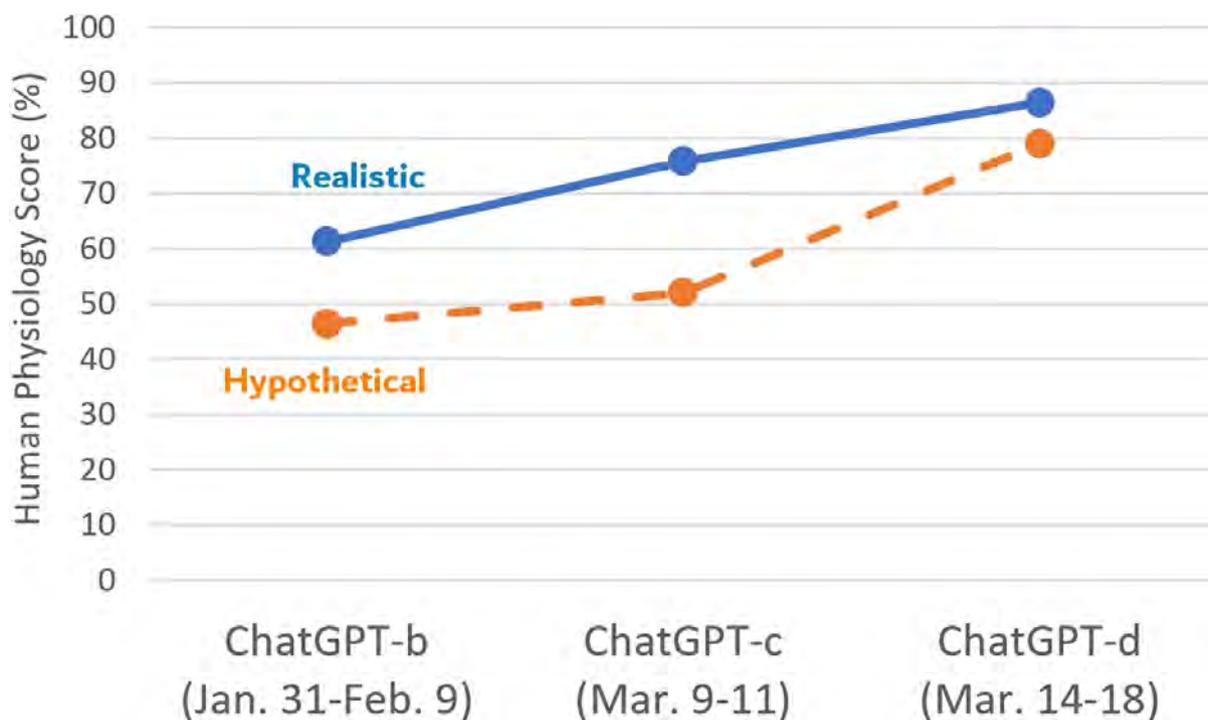
FIG 3 Improvements in ChatGPT human physiology scores over the course of this study. Data are replotted from Table 4 (error bars are omitted for clarity). Scores on both the realistic questions and the hypothetical questions improved significantly from the first round of testing to the third round, according to paired t-tests ($P < 0.001$ for both). Similar trends were seen for ChatGPT's cell biology scores (Table 5).

analogical thinking (30), case comparison (31), and transfer (20, 32). Regardless of terminology, this task is widely understood to be a central focus of education, yet students often fall short of faculty expectations (32, 33).

One likely reason for this difficulty is that students may not receive enough practice with scenarios that are truly novel (yet LLO-aligned). Momsen et al. have reported that introductory biology courses usually have few exam questions novel enough to demand HOCS (5). Similarly, we have observed that in popular human anatomy and physiology textbooks, only ~0.2% of questions concern non-human animals or aliens (Sankar et al., unpublished observations). When instructors create these kinds of questions, we may get feedback such as the following (given to G.J.C. by a human physiology student): "Give real patient examples and stop with the alien or monsters or other creatures; not everybody is aware or knowledgeable of these creatures unless you are a marine biologist of some sort. In real life, I would like to save and evaluate real people/person, not a Loch Ness monster."

This representative comment highlights the risk that novel problems, being unfamiliar to students, may be perceived as irrelevant and/or unfair. To avoid such misunderstandings, we urge instructors to be transparent with students on both the "what" and the "why" of these novel problems. That is, if students are likely to face exam problems about biological entities not previously covered, instructors should—well ahead of the exam—explicitly inform students of this, justify the inclusion of such novelty, and provide LLO-linked examples, perhaps via the TQT framework (17–19). Most broadly, instructors should help students appreciate that in both basic science and applied (e.g., clinical) science, we solve novel problems by applying what is known to what is not yet known. Novel or hypothetical problems thus serve as valuable practice for authentic challenges in research, medical care, etc. The invention subtype of hypothetical question may correspond to the discovery of a novel mechanism, the diagnosis of a patient with a novel disorder, or the treatment of a patient with a novel class of drug, while the redefinition subtype may correspond to situations where new test results overturn previous assumptions.

Finally, regarding cheating (34, 35), our results provide some reassurance about chatbots' current limitations in answering exam questions, as well as some warning, given the ongoing evolution of their abilities. As of this writing, many chatbots seem to struggle with hypothetical questions and, in the absence of feedback, do not improve their answers when repeatedly asked the same question or similar questions; the latter finding suggests that chatbots cannot necessarily improve in real time during the course of a single exam. In addition, while we made all of our questions text-based, we presumably could have stumped the chatbots with image-based questions. However, the high success rate of GPT-4-driven ChatGPT (ChatGPT-d) on our and others' hypothetical questions (36), as well as this ChatGPT version's ability to analyze images (37), suggests that with continuing advances in AI, even hypothetical and image-based questions may soon become straightforward for many chatbots. We advise against a strategy of trying to "outsmart" chatbots by writing ever more convoluted questions; instead, we favor approaches to assessment that simultaneously prioritize fairness (equity), stress reduction, and student learning (38). While our study did not directly investigate equity issues, we had to pay $20 per month to use the highest-scoring chatbot, implying that students with different resources might have access to chatbots with different capabilities.

In conclusion, we used the framework of TQTs to create well-matched sets of realistic and hypothetical questions relevant to undergraduate courses in human physiology and cell biology. The fact that LLM-based chatbots usually scored lower on the hypothetical questions constitutes new evidence to support previous suggestions that novel scenarios provide unique cognitive challenges. We hope that future work will further explore the issue of question novelty, perhaps via fuller comparisons of the redefinition and invention subtypes of novel questions, to further clarify how novel questions impact cognition and how they might be used optimally in instruction and assessment.

AUTHOR AFFILIATIONS

[1]Life Sciences Department, Everett Community College, Everett, Washington, USA
[2]Department of Biological Sciences, Fordham University, Bronx, New York, USA
[3]Biology Department, Whitman College, Walla Walla, Washington, USA
[4]Biology Department, St. Charles Community College, Cottleville, Missouri, USA
[5]School for the Future of Innovation in Society, Arizona State University, Tempe, Arizona, USA

AUTHOR ORCIDs

Gregory J. Crowther http://orcid.org/0000-0003-0530-9130
Usha Sankar http://orcid.org/0009-0005-9516-6939
Leena S. Knight http://orcid.org/0000-0003-1738-2976
Kevin T. Patton http://orcid.org/0000-0003-3219-667X
Lekelia D. Jenkins http://orcid.org/0000-0002-2375-2032

AUTHOR CONTRIBUTIONS

Gregory J. Crowther, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review and editing | Usha Sankar, Formal analysis, Investigation, Validation, Writing – review and editing | Leena S. Knight, Formal analysis, Investigation, Methodology, Validation, Writing – review and editing | Deborah L. Myers, Formal analysis, Investigation, Validation, Writing – review and editing | Kevin T. Patton, Formal analysis, Investigation, Validation, Writing – review and editing | Lekelia D. Jenkins, Conceptualization, Supervision, Validation, Writing – review and editing | Thomas A. Knight, Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, Writing – review and editing

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental file S1 (jmbe00153-23-s0001.xlsx). Human physiology questions.
Supplemental file S2 (jmbe00153-23-s0002.xlsx). Cell biology questions.

REFERENCES

1. Handelsman J, Ebert-May D, Beichner R, Bruns P, Chang A, DeHaan R, Gentile J, Lauffer S, Stewart J, Tilghman SM, Wood WB. 2004. Scientific teaching. Science 304:521–522. https://doi.org/10.1126/science.1096022
2. Clemmons AW, Timbrook J, Herron JC, Crowe AJ. 2020. BioSkills guide: development and national validation of a tool for interpreting the vision and change core competencies. CBE Life Sci Educ 19:ar53. https://doi.org/10.1187/cbe.19-11-0259
3. Krathwohl DR. 2002. A revision of Bloom's taxonomy: an overview. Theory Into Pract 41:212–218. https://doi.org/10.1207/-s15430421tip4104_2
4. Crowe A, Dirks C, Wenderoth MP. 2008. Biology in Bloom: implementing Bloom's taxonomy to enhance student learning in biology. CBE Life Sci Educ 7:368–381. https://doi.org/10.1187/cbe.08-05-0024
5. Momsen JL, Long TM, Wyse SA, Ebert-May D. 2010. Just the facts? Introductory undergraduate biology courses focus on low-level

cognitive skills. CBE Life Sci Educ 9:435–440. https://doi.org/10.1187/cbe. 10-01-0001

6. Semsar K, Casagrand J. 2017. Bloom's dichotomous key: a new tool for evaluating the cognitive difficulty of assessments. Adv Physiol Educ 41:170–177. https://doi.org/10.1152/advan.00101.2016

7. Thompson AR, O'Loughlin VD. 2015. The Blooming anatomy tool (BAT): a discipline‐specific rubric for utilizing Bloom's taxonomy in the design and evaluation of assessments in the anatomical sciences. Anat Sci Educ 8:493–501. https://doi.org/10.1002/ase.1507

8. Zaidi NB, Hwang C, Scott S, Stallard S, Purkiss J, Hortsch M. 2017. Climbing Bloom's taxonomy pyramid: lessons from a graduate histology course. Anat Sci Educ 10:456–464. https://doi.org/10.1002/ase.1685

9. Arneson JB, Offerdahl EG. 2018. Visual literacy in Bloom: using Bloom's taxonomy to support visual learning skills. CBE Life Sci Educ 17:ar7. https://doi.org/10.1187/cbe.17-08-0178

10. Monrad SU, Bibler Zaidi NL, Grob KL, Kurtz JB, Tai AW, Hortsch M, Gruppen LD, Santen SA. 2021. What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy. Med Teach 43:575–582. https://doi.org/10.1080/0142159X.2021. 1879376

11. Semsar K, Brownell S, Couch BA, Crowe AJ, Smith MK, Summers MM, Wright CD, Knight JK. 2019. Phys-MAPS: a programmatic physiology assessment for introductory and advanced undergraduates. Adv Physiol Educ 43:15–27. https://doi.org/10.1152/advan.00128.2018

12. Couch BA, Wright CD, Freeman S, Knight JK, Semsar K, Smith MK, Summers MM, Zheng Y, Crowe AJ, Brownell SE. 2019. GenBio-MAPS: a programmatic assessment to measure student understanding of vision and change core concepts across general biology programs. CBE Life Sci Educ 18:ar1. https://doi.org/10.1187/cbe.18-07-0117

13. Chirillo M, Silverthorn DU, Vujovic P. 2021. Core concepts in physiology: teaching homeostasis through pattern recognition. Adv Physiol Educ 45:812–828. https://doi.org/10.1152/advan.00106.2021

14. Silverthorn DU. 2022. Constructing the wiggers diagram using core concepts: a classroom activity. Adv Physiol Educ 46:714–723. https://doi. org/10.1152/advan.00046.2022

15. Stanfield E, Slown CD, Sedlacek Q, Worcester SE. 2022. A course-based undergraduate research experience (CURE) in biology: developing systems thinking through field experiences in restoration ecology. CBE Life Sci Educ 21:ar20. https://doi.org/10.1187/cbe.20-12-0300

16. Deane-Coe KK, Sarvary MA, Owens TG. 2017. Student performance along axes of scenario novelty and complexity in introductory biology: lessons from a unique factorial approach to assessment. CBE Life Sci Educ 16:ar3. https://doi.org/10.1187/cbe.16-06-0195

17. Crowther GJ, Wiggins BL, Jenkins LD. 2020. Testing in the age of active learning: test question templates help to align activities and assessments. HAPS Ed 24:592–599. https://doi.org/10.21692/haps.2020.006

18. Crowther GJ, Knight TA. 2023. Using test question templates to teach physiology core concepts. Adv Physiol Educ 47:202–214. https://doi.org/ 10.1152/advan.00024.2022

19. Evans DP, Jenkins LD, Crowther GJ. 2023. Student perceptions of a framework for facilitating transfer from lessons to exams, and the relevance of this framework to published lessons. J Microbiol Biol Educ 24:e00200-22. https://doi.org/10.1128/jmbe.00200-22

20. Kaminske AN, Kuepper-Tetzel CE, Nebel CL, Sumeracki MA, Ryan SP. 2020. Transfer: a review for biology and the life sciences. CBE Life Sci Educ 19:es9. https://doi.org/10.1187/cbe.19-11-0227

21. Doherty JH, Scott EE, Cerchiara JA, Jescovitch LN, McFarland JL, Haudek KC, Wenderoth MP. 2023. What a difference in pressure makes! A framework describing undergraduate students' reasoning about bulk flow down pressure gradients. CBE Life Sci Educ 22:ar23. https://doi.org/ 10.1187/cbe.20-01-0003

22. Sullivan M, Kelly A, McLaughlan P. 2023. ChatGPT in higher education: considerations for academic integrity and student learning. J Appl Learn Teach6:1–10. https://doi.org/10.37074/jalt.2023.6.1.17

23. Berezow A. 2022. We gave ChatGPT a college-level Microbiology quiz. It blew the quiz away. BigThink.com. Available from: https://bigthink.com/ the-future/chatgpt-microbiology-quiz-aced/

24. Schembri N. 2023. ThinkMagazine.mt. Is ChatGPT an Aid or a Cheat? Available from: https://thinkmagazine.mt/is-chatgpt-an-aid-or-a-cheat/

25. Devetak I, Glažar SA, Vogrinc J. 2010. The role of qualitative research in science education. Eurasia J Math Sci T 6:77–84. https://doi.org/10.12973/ejmste/75229

26. Bommineni VL, Bhagwagar S, Balcarcel D, Davatzikos C, Boyer D. 2023. Performance of ChatGPT on the MCAT: the road to personalized and equitable premedical learning. Med edu. https://doi.org/10.1101/2023. 03.05.23286533

27. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, Cielo D, Oyelese AA, Doberstein CE, Telfeian AE, Gokaslan ZL, Asaad WF. 2023. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery Publish Ahead of Print:2023–04. https://doi.org/10.1227/neu.0000000000002551

28. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. 2023. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. Med Educ. https://doi.org/10.1101/2023.02.13.23285879

29. McFarland JL, Price RM, Wenderoth MP, Martinková P, Cliff W, Michael J, Modell H, Wright A. 2017. Development and validation of the homeostasis concept inventory. CBE Life Sci Educ 16:ar35. https://doi.org/10.1187/ cbe.16-10-0305

30. Gentner D, Loewenstein J, Thompson L, Forbus KD. 2009. Reviving inert knowledge: analogical abstraction supports relational retrieval of past events. Cogn Sci 33:1343–1382. https://doi.org/10.1111/j.1551-6709. 2009.01070.x

31. Alfieri L, Nokes-Malach TJ, Schunn CD. 2013. Learning through case comparisons: a meta-analytic review. Educ Psych 48:87–113. https://doi. org/10.1080/00461520.2013.775712

32. Ambrose SA, Bridges MW, DiPietro M, Lovett MC, Norman MK. 2010. How learning works: seven research-based principles for smart teaching. John Wiley & Sons, Hoboken, NJ.

33. Michael J. 2022. Use of core concepts of physiology can facilitate student transfer of learning. Adv Physiol Educ 46:438–442. https://doi.org/10. 1152/advan.00005.2022

34. Cox C, Tzoc E. 2023. ChatGPT: implications for academic libraries. CRLN 84. https://doi.org/10.5860/crln.84.3.99

35. Murugesan S, Cherukuri AK. 2023. The rise of generative artificial intelligence and its impact on education: the promises and perils. Computer 56:116–121. https://doi.org/10.1109/MC.2023.3253292

36. Kosinski M. 2023. Theory of mind may have spontaneously emerged in large language models. arXiv. https://doi.org/10.48550/ arXiv.2302.02083.

37. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, Grigo J, Ben Tkhayat H, Frey B, Gaipl U, Distel L, Maier A, Fietkau R, Bert C, Putz F. 2023. Benchmarking ChatGPT-4 on ACR radiation oncology in-training (TXIT) Exam and Red Journal Gray Zone cases: potentials and challenges for AI-assisted medical education and decision making in radiation oncology. arXiv. https://doi.org/10.2139/ssrn.4457218

38. Hsu JL. 2021. Promoting academic integrity and student learning in online biology courses. J Microbiol Biol Educ 22:22.1.17. https://doi.org/ 10.1128/jmbe.v22i1.2291