ACADEMIC MEDICINE
Journal of the Association of American Medical Colleges

Uncomposed, edited manuscript published online ahead of print.

This published ahead-of-print manuscript is not the final version of this article, but it may be cited and shared publicly.

Author: Matthias Carl Laupichler, MSc; Johanna Flora Rother, MSc; Ilona C. Grunwald Kadow, PhD; Seifollah Ahmadi, PhD; Tobias Raupach, MD, MME

Title: Large Language Models in Medical Education: Comparing ChatGPT- to Human-Generated Exam Questions

DOI: 10.1097/ACM.0000000000005626

Academic Medicine

DOI: 10.1097/ACM.0000000000005626

Large Language Models in Medical Education: Comparing ChatGPT- to Human-Generated Exam Questions

Matthias Carl Laupichler, MSc[1], Johanna Flora Rother, MSc[2], Ilona C. Grunwald Kadow, PhD[3], Seifollah Ahmadi, PhD[4], and Tobias Raupach, MD, MME[5]

[1]**M.C. Laupichler** is a research assistant, Institute of Medical Education, University Hospital Bonn, and doctoral student, Institute of Psychology, University of Bonn, Bonn, Germany; ORCID: https://orcid.org/0000-0003-3104-1123.

[2]**J.F. Rother** is a research assistant, Institute of Medical Education, University Hospital Bonn, Bonn, Germany; ORCID: https://orcid.org/0009-0004-3526-3211.

[3]**I.C. Grunwald Kadow** is full professor and head, Institute of Physiology II, Department of Medicine, University of Bonn, Bonn, Germany; ORCID: https://orcid.org/0000-0002-9085-4274.

[4]**S. Ahmadi** is a research associate, Institute of Physiology II, Department of Medicine, University of Bonn, Bonn, Germany.

[5]**T. Raupach** is consultant cardiologist and head, Institute of Medical Education, University Hospital Bonn, Bonn, Germany; ORCID: https://orcid.org/0000-0003-2555-8097.

Correspondence should be addressed to Matthias Carl Laupichler, Institute of Medical Education, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany; telephone: +49 228 287 52167; email: matthias.laupichler@ukbonn.de.

Acknowledgments: The authors thank Dr. Martin Breitbach for his support in setting up the study. Furthermore, the authors thank Ebru Bala and Gilda Masala for their assistance in analyzing the results.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: This study was approved by the ethics committee at the Medical Faculty of the Rheinische Friedrich-Wilhelms-University Bonn on July 3, 2023 (application number 175/23-EP).

Data: Data were collected pseudonymously, and all study participants gave informed consent for data processing. Outside sources of data were not used.

Abstract

Problem

Creating medical exam questions is time consuming, but well-written questions can be used for test-enhanced learning, which has been shown to have a positive effect on student learning. The automated generation of high-quality questions using large language models (LLMs), such as ChatGPT, would therefore be desirable. However, there are no current studies that compare students' performance on LLM-generated questions to questions developed by humans.

Approach

The authors compared student performance on questions generated by ChatGPT (LLM questions) with questions created by medical educators (human questions). Two sets of 25 multiple-choice questions (MCQs) were created, each with 5 answer options, 1 of which was correct. The first set of questions was written by an experienced medical educator, and the second set was created by ChatGPT after the authors identified learning objectives and extracted some specifications from the human questions. Students answered all questions in random order in a formative paper-and-pencil test that was offered leading up to the final summative neurophysiology exam (summer 2023). For each question, students also indicated whether they thought it had been written by a human or created by ChatGPT.

Outcomes

The final data set consisted of 161 participants and 46 MCQs (25 human and 21 LLM questions). There was no statistically significant difference in item difficulty between the 2 question sets, but discriminatory power was statistically significantly higher in human than LLM questions (mean = .36, standard deviation [SD] = .09 vs mean = .24, SD = .14; P = .001). On average, students identified 57% of question sources (human or LLM) correctly.

Next Steps

Future research should replicate the study procedure in other contexts (e.g., other medical subjects, semesters, countries, and languages). In addition, the question of whether LLMs are suitable for generating different question types, such as key feature questions, should be investigated.

Problem

Test-enhanced learning is resource intensive

Numerous studies have shown that repeated testing of knowledge leads to increased retention among learners.[1] This phenomenon is called the testing effect,[2] and test-enhanced learning[3] uses this effect by providing students with repeated, ungraded tests throughout a course. In test-enhanced learning, the traditional multiple-choice question (MCQ) format is often used, as MCQs allow a reliable and valid evaluation of knowledge[4] and are a mainstay of summative exams in many medical schools the world over. However, the development of MCQs by health care professionals and medical educators is costly and resource intensive. A common rule of thumb regarding the effort involved in creating these questions is that it takes about an hour of a health care professional’s or medical educator’s time to develop a single high-quality MCQ. Therefore, it would be of great benefit to the training of future physicians if this process could become (at least partially) automated.

Large language models (LLMs) in medical education

The concept of the automated creation of exam questions could benefit from the recent advent of LLMs, such as ChatGPT. LLMs are systems that use natural language processing methods to “recognize, interpret, and generate text.”[5(p.1930)] Following the recent hype around these artificial intelligence-based systems, which began with the release of OpenAI's ChatGPT in November 2022, a number of use cases have demonstrated how LLMs (and ChatGPT in particular) have been used to achieve results in various domains. In health care, for example, ChatGPT is has been used in efforts to improve doctor-patient communication and simplify clinical management processes.

In addition to these more general applications, the advantages and disadvantages of the use of LLMs in medical education and continuing medical education have been discussed in detail.

For example, Khan and colleagues describe 8 potential areas of ChatGPT application in medical education, including "teaching assistance," "personalized learning," and "creating content to facilitate learning."[6(p.606)] Other authors raise concerns that ChatGPT may pose a plagiarism threat or that ethical limitations play a role in the use of ChatGPT in medical education.[7] While these initial explorations of LLMs are important for identifying future research directions, they are not empirical in nature.

Using ChatGPT to automate exam question generation

Although the automated generation of exam questions by LLMs is a promising endeavor, question quality must be assessed before this approach can be advocated. Questions generated by ChatGPT would not only have to represent the learning objectives of the corresponding curriculum with sufficient validity, but quality criteria, such as item difficulty and discriminatory power, would also need to be acceptable.

Some researchers have already recognized the benefit of MCQs generated by LLMs and have conducted prospective studies. While attempts to automate the generation of medical exam questions using LLMs had been made prior to the introduction of ChatGPT,[8] ChatGPT provides a user-friendly, no-code alternative that enables the automated generation of items by individuals lacking significant programming or artificial intelligence expertise. For example, in one study, ChatGPT was used to create graduate medical education exam questions that were then evaluated by subject matter experts according to various criteria, such as appropriateness and clarity.[9] This and other prospective studies have found that automatically generated questions were equivalent to human-generated questions on most quality criteria. However, to our knowledge, no studies have compared students' performance on LLM-generated questions to their performance on questions developed by humans.

Approach

We compared ChatGPT-generated questions (LLM questions) with questions developed by medical educators (human questions) in a preparatory exam. We informed all students who were enrolled in the neurophysiology course at University of Bonn Medical School (in summer 2023) about the possibility of taking an ungraded preparatory exam before the final exam. We contacted students via email about the date and general conditions of participation in the study. Participation in the preparatory exam was voluntary, and we offered no financial or material incentives. We made it clear that consenting to data processing as part of the empirical study was voluntary and that participation in the formative exam would not affect the students' grade in the subsequent summative exam.

We created 2 sets of 25 MCQs, each with 5 answer options, 1 of which was correct. The first set of questions was written by an experienced medical educator on topics covered in the neurophysiology course lectures. We identified the specific topic of each human-generated question and the corresponding learning objective (e.g., specific topic: SNARE protein function, and learning objective: the function of SNARE proteins in the area of motor end plates) that were subsequently used in the ChatGPT (ChatGPT 3.5, May 24, 2023, version, OpenAI, San Francisco, California) prompts to create the LLM question set. The prompts were generated by the research team and consisted of a general section that was the same for each prompt. This general section specified the question type (e.g., an MCQ with 5 answer options, 1 of which is correct), defined the target audience (e.g., medical students), and provided certain supplementary details aimed at preventing cueing (e.g., answer options have to be mutually exclusive). The general section also indicated that the correct answer to the generated question was to be provided. For each question, the remainder of the prompt defined the specific topic and the individual learning objective of the question (see Table 1 for an example human question, ChatGPT prompt, and LLM question).

We presented all questions to undergraduate students near the end of their second year in a formative paper-and-pencil test in random order, which was the same across participants, that was offered leading up to the final summative neurophysiology exam (summer 2023). We used www.random.org to randomize the items. For each question, students were also asked to indicate whether they thought it had been written by a human or created by ChatGPT (i.e., dichotomous decision). Students had 60 minutes to answer the 50 MCQs and enter the 50 dichotomous decisions.

We used IBM SPSS Statistics (version 27, IBM, Armonk, New York) for data analyses. We conducted an independent t-test in which the percentage of correctly answered LLM questions was compared with the percentage of correctly answered human questions. In addition, we performed Levene tests to assess variance homogeneity and Kolmogorov-Smirnov tests to test the normal distribution assumption. Furthermore, we calculated the discriminatory power of each question using the point-biserial correlation coefficient. The mean discriminatory power of LLM questions and human questions were compared using independent t-tests. Finally, we performed a binomial test to evaluate whether students were better at identifying LLM questions than would be expected based on the guessing probability of 50%.

This study was approved by the ethics committee at the Medical Faculty of the Rheinische Friedrich-Wilhelms-University Bonn (application number 175/23-EP), and all participants provided written consent.

Outcomes

Of 179 medical students who participated in the exam, 175 (98%) consented to having their results processed as part of the empirical study. In addition, we excluded 14 (8%) students because they left more than 10 questions unanswered, resulting in a sample size of 161 (90%)

students. We had to exclude 4 of the 25 (16%) LLM questions from the analysis because ChatGPT had produced completely or partially incorrect answer options. The incorrect options were corrected by the medical educator and could therefore no longer be interpreted as pure LLM questions. Thus, the final data set consisted of 161 participants and 46 MCQs (25 human questions and 21 LLM questions).

Comparison of test difficulty of LLM and human questions

Students answered 62% (standard deviation [SD] = 19.0) of all human questions and 69% (SD = 22.5) of all LLM questions correctly. Thus, they answered 7% more of the LLM questions correctly than they did human questions (95% confidence interval [CI] [–18.43, 6.24]). The results of the Kolmogorov-Smirnov test and the Levene test were not significant, so it was legitimate to perform an independent t-test. According to this, the difference between human questions and LLM questions was not statistically significant. This is also illustrated by the Bland-Altman plot (Figure 1), which shows that the item difficulties of the respective item pairs (LLM questions and human questions on the same topic) were comparable. This result could be interpreted as preliminary evidence that ChatGPT and other LLMs could be successfully employed to create questions for formative exams in medical school.

Comparison of discriminatory power of LLM and human questions

The discriminatory power (as expressed by the point-biserial correlation coefficient) of the LLM questions averaged .24 (SD = .14), whereas the discriminatory power of the human questions averaged .36 (SD = .09). This difference of .12 (95% CI [0.05, 0.19]) was statistically significant ($t(44) = 3.44$; $P = .001$). The discriminatory power of both item sets (i.e., human and LLM questions) was acceptable at > .20. However, only the average discriminatory power of the human questions at > .30 was in the ideal range. Finally, when looking at individual item characteristics, we found that some LLM questions had particularly poor discriminatory power, which negatively impacted the average discriminatory power of the full LLM item set (Figure 2). Thus, before using LLM questions in summative exams, it would be advisable to pilot the questions to identify and eliminate such outliers.

The significant difference in discriminatory power could be an indication that questions created by medical educators are better at distinguishing high- from low-performing students. We postulate that there are 2 potential interrelated explanations for this difference. First, the medical educator who created the MCQs, owing to his extensive knowledge in the specific discipline (i.e., neurophysiology) and familiarity with the learning objectives, was more likely to be adept at generating items with higher construct validity. Second, the medical educator knew the lecture content, allowing him to align the questions more closely with the lecture material compared to what ChatGPT was able to do. Therefore, in future studies, a refinement of the ChatGPT prompts used would be necessary to raise the discriminatory power of the questions to an acceptable level.

Identification of LLM questions

On average, students were 7% (95% CI [5.44, 8.60]; $t(160) = 8.79$; $P < .001$) better at distinguishing LLM from human questions (and vice versa) than would be expected by chance. We were surprised to find that only 57% of these decisions were correct, which conversely means that the source of 43% of the questions was not identified correctly. This could be interpreted as preliminary evidence that students are not yet familiar with the possibilities of LLMs for generating MCQs, which makes their application in medical education a promising and novel endeavor.

Next Steps

This study had 2 major limitations. First, we only examined results from one semester of one course at one medical school. Therefore, our findings cannot be generalized to all medical schools or education systems. Second, we did not collect information about prior knowledge and/or exam performance of the participants. This and the fact that ours was a voluntary sample could have potentially influenced the results. For instance, it is conceivable that, at the time of the voluntary formative assessment, only students with higher competencies in the subject matter might have participated. In future studies, these potential moderating variables should thus be assessed and controlled for. Nevertheless, we hope that this report, as one of the first of its kind, will motivate further work in this area that will eventually allow a representative evaluation of the use of LLMs in formative medical exams.

One of the most important next steps is to replicate the study procedure in other contexts, that is, in other medical subjects, semesters, countries (i.e., educational systems), languages, etc. This includes the use of LLM-generated questions in summative exams to explore possible influences of exam consequences (e.g., grading, passing) on item difficulty and discriminatory power.

Furthermore, the extent to which LLMs are suitable for generating different question types should be investigated. Another question type in which LLMs could play a role are key feature questions, which are used for the assessment of clinical reasoning skills and have also been used in test-enhanced learning.[10]

The wording of the prompts could also be further refined. For example, different prompt types could be compared in the context of A/B testing. In addition, we employed the freely available ChatGPT 3.5. However, it should be examined how ChatGPT 3.5 compares to

ChatGPT 4.0 (and other potential newer versions) and to medicine-specific LLMs. Researchers could use the same prompts in different LLMs and compare the quality of the generated items.

Last, it would be interesting to investigate whether medical professionals are better than medical students at distinguishing ChatGPT- from human-generated questions. Since it can be assumed that medical educators in particular are well versed in both question development and the course content taught, it would seem to follow that they should be better at identifying the source of questions than medical students.

References

1. Karpicke JD, Roediger HL. The critical importance of retrieval for learning. Science. 2008;319:966-968. doi:10.1126/science.1152408
2. McDaniel MA, Anderson JL, Derbish MH, Morrisette N. Testing the testing effect in the classroom. Eur J Cogn Psychol. 2007;19:494-513. doi:10.1080/09541440701326154
3. Pan SC, Rickard TC. Transfer of test-enhanced learning: Meta-analytic review and synthesis. Psychol Bull. 2018;144:710-756. doi: 10.1037/bul0000151
4. Coughlin PA, Featherstone, CR. How to write a high quality multiple choice question (MCQ): a guide for clinicians. Eur J Vasc Endovasc Surg. 2017;54:654-658. doi: 10.1016/j.ejvs.2017.07.012
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29:1930-1940. doi:10.1038/s41591-023-02448-8
6. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci. 2023;39:605-607. doi:10.12669/pjms.39.2.7653
7. Lee H. The rise of ChatGPT: exploring its potential in medical education [published online ahead of print March 14, 2023]. Anat Sci Educ. doi:10.1002/ase.2270
8. von Davier M. Training Optimus Prime, M.D. a case study of automated item generation using artificial intelligence—from fine-tuned GPT2 to GPT3 and beyond. In: Yaneva V, von Davier M, eds. Advancing Natural Language Processing in Natural Assessment. New York, NY: Routledge; 2023.
9. Cheung BHH, Kui Kai Lau G, Tin Chun Wong G, et al. ChatGPT versus human in generating medical graduate exam questions—an international prospective study. medRxiv. 2023. doi:10.1101/2023.05.13.23289943

10. Berens M, Becker T, Anders S, Sam AH, Raupach T. Effects of elaboration and instructor feedback on retention of clinical reasoning competence among undergraduate medical students: a randomized crossover trial. JAMA Netw Open. 2022;5:e2245491. doi:10.1001/jamanetworkopen.2022.45491

Figure Legends

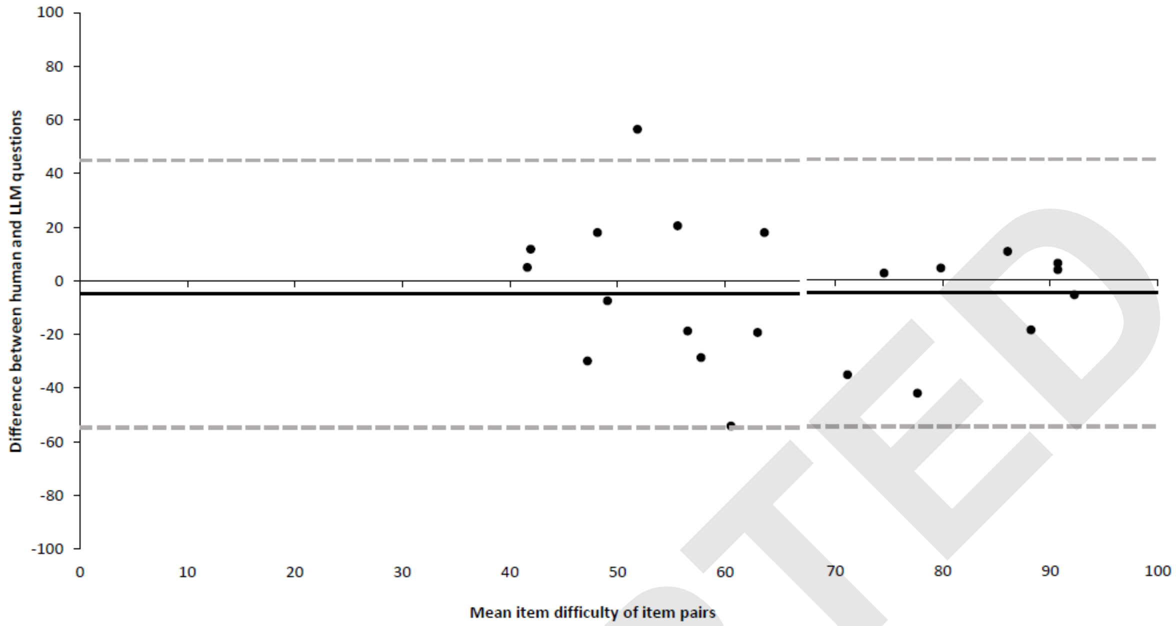
Figure 1

Bland-Altman plot depicting the differences in item difficulty between exam questions developed by a human and LLM-generated questions, University of Bonn Medical School, summer 2023. There were 21 item pairs of human and LLM questions. Each dot represents the difference in item difficulty between the human and LLM multiple-choice question on the same topic that were used in a preparatory exam. The solid line represents the mean difference between human and LLM questions, while the dashed lines represent the upper and lower limit of agreement (i.e., the upper and lower end of the 95% confidence interval). LLM questions were generated by ChatGPT 3.5 (May 24, 2023, version, OpenAI, San Francisco, California). Abbreviation: LLM, large language model.

Figure 2

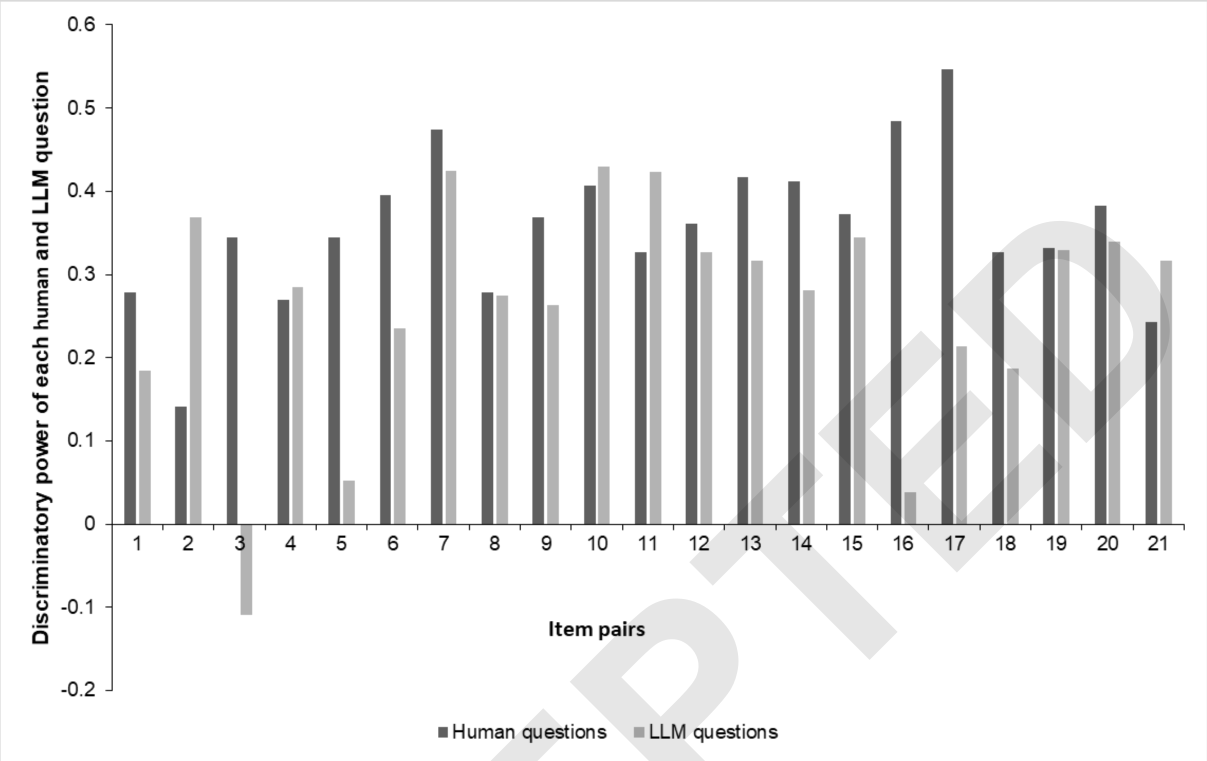
Bar chart showing the discriminatory power of each exam question developed by a human (dark gray) and LLM-generated question (light gray), University of Bonn Medical School, summer 2023. Items on the same topic are shown as pairs. LLM questions were generated by ChatGPT 3.5 (May 24, 2023, version, OpenAI, San Francisco, California).

Figure 1



Difference between human and LLM questions
100
80
60
40
20
0
-20
-40
-60
-80
-100
0
10
20
30
40
50
60
70
80
90
100
Mean item difficulty of item pairs

Figure 2



Discriminatory power of each human and LLM question
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Item pairs
Human questions
LLM questions

Table 1

Example of an Exam Question Developed by a Human and the Corresponding LLM-Generated Question, Including the Prompt Used to Create the LLM Question, University of Bonn Medical School, Summer 2023

Element	Example[a]
Human question	In some cases, botulinum toxin can be used to treat specific neuromuscular disorders (e.g., spasms). Botulinum toxin cleaves proteins of the so-called SNARE complex, the function of which is thereby disturbed in the area of the motor end plates. What function do these SNARE proteins normally have? (A) They function as nicotinic acetylcholine receptors in the subsynaptic cell membrane (B) They function as tension-gated Ca2+ channels in the neuronal cell membrane (C) They catalyze the biosynthesis of acetylcholine from acetyl-CoA and choline in the presynaptic terminal (D) They catalyze the cleavage of acetylcholine in the synaptic cleft (E) They are required for the fusion of synaptic acetylcholine-containing vesicles with the neuronal cell membrane Correct answer: E
Prompt used to generate LLM question[b]	Create an MCQ with 5 answer options and a single correct answer. The question is to be answered by medical students. The topic of the question is [specific topic]. The question should be about [learning objective]. However, the terms do not have to appear in each answer option. Make sure that the answer options do not exclude each other—that is, different answer options must not contain directly opposite statements. Provide the correct answer at the end.
Specific topic indicated in prompt	SNARE protein function
Learning objective specified in prompt	The function of SNARE proteins in the area of motor end plates
LLM question	What are the functions of SNARE proteins in motor end plates? (A) They enable the binding of neurotransmitters to postsynaptic receptors (B) They regulate the release of acetylcholine at the synaptic cleft (C) They mediate vesicle fusion with the presynaptic membrane (D) They enhance the formation of action potentials in motor neurons (E) They inhibit the breakdown of neurotransmitters by enzymes in the synaptic cleft Correct answer: C

Abbreviations: LLM, large language model; SNARE, soluble N-ethylmaleimide-sensitive factor activating protein receptor; MCQ, multiple-choice question.
[a]The human questions, LLM questions, and prompts were initially generated in German and the examples for this table were translated into English using DeepL (July 02, 2023, version, DeepL SE, Cologne, Germany). A domain expert subsequently reviewed the translations.
[b]Prompts were put into ChatGPT 3.5 (May 24, 2023, version, OpenAI, San Francisco, California).