

1           **AI-Powered Test Question Generation in Medical Education: The DailyMed Approach**

2

3                                   J. van Uhm<sup>1#</sup>, M.M. van Haelst<sup>2,3,4</sup>, P.R. Jansen<sup>2,3,4,5</sup>

4

5   **Affiliations**

6           1. Erasmus University Medical Center, Department of Pediatrics, Rotterdam, the  
7           Netherlands

8           2. Amsterdam UMC, Department of Human Genetics, University of Amsterdam,  
9           Amsterdam, Netherlands

10          3. Amsterdam Reproduction and Development Research Institute, Amsterdam,  
11          Netherlands

12          4. Amsterdam UMC, Emma Center for Personalized Medicine, University of Amsterdam,  
13          Amsterdam, Netherlands

14          5. Department of Complex Trait Genetics, VU University, Amsterdam, the Netherlands

15

16   # Correspondence should be addressed to: J. van Uhm, [j.vanuhm@erasmusmc.nl](mailto:j.vanuhm@erasmusmc.nl)

17

18

19

20

21

22   Words: 3,155

23   Figures: 4

24   Tables: 2

25   References: 16

26

27 **ABSTRACT**

28 **Introduction:** Large language models (LLMs) presents opportunities to improve the efficiency and  
29 quality of tools in medical education, such as the generation of multiple-choice questions (MCQs).  
30 However, ensuring that these questions are clinically relevant, accurate, and easily accesible and  
31 reusable remains challenging. Here, we developed DailyMed, an online automated pipeline using  
32 LLMs to generate high-quality medical MCQs.

33 **Methods:** Our DailyMed pipeline involves several key steps: 1) topic generation, 2) question  
34 creation, 3) validation using Semantic Scholar, 4) difficulty grading, 5) iterative improvement of  
35 simpler questions, and 6) final human review. The Chain-of-Thought (CoT) prompting technique  
36 was applied to enhance LLM reasoning. Three state-of the art LLMs—OpenBioLLM-70B, GPT-4o,  
37 and Claude 3.5 Sonnet—were evaluated within the area of clinical genetics, and the generated  
38 questions were rated by clinical experts for validity, clarity, originality, relevance, and difficulty.

39 **Results:** GPT-4o produced the highest-rated questions, excelling in validity, originality, clarity,  
40 and relevance. Although OpenBioLLM was more cost-efficient, it consistently scored lower in all  
41 categories. GPT-4o also achieved the greatest topic diversity (89.8%), followed by Claude Sonnet  
42 (86.9%) and OpenBioLLM (80.0%). In terms of cost and performance, GPT-4o was the most  
43 efficient model, with an average cost of \$0.51 per quiz and a runtime of 16 seconds per question.

44 **Conclusions:** Our pipeline provides a scalable, effective and online-accessible solution for  
45 generating diverse, clinically relevant MCQs. GPT-4o demonstrated the highest overall  
46 performance, making it the preferred model for this task, while OpenBioLLM offers a cost-  
47 effective alternative.

48

49

50

51

## 52 INTRODUCTION

53 The development of novel educational tools is essential for advancing medical education and  
54 finding new ways to challenge the professional growth of clinicians and those in training.  
55 Generative artificial intelligence (AI), more recently in the form of large-language models (LLM)<sup>1</sup>,  
56 hold promises to positively impact education by its ability to explain complex information,  
57 interact with learners to improve learning experiences, and create personalized content that  
58 adapts to individual learning levels<sup>1,2</sup>. These models, trained on large input datasets and including  
59 over 100 billion parameters<sup>3</sup>, are able to provide trustworthy information and pass professional-  
60 level exams in the field of medicine<sup>4</sup>, law<sup>5</sup> and language<sup>6</sup>. The relative strength of these models  
61 can be even further improved by the sequential use of different LLM models, where output from  
62 one model can be further improved by a chain of subsequent LLM models<sup>6</sup>. Since LLMs have  
63 shown the capability to pass test-questions and provide underlying reasoning for their answers,  
64 LLMs may be equally suitable to develop high-quality test-questions themselves and assess the  
65 knowledge of others through multiple choice questions (MCQ). Indeed, in recent articles several  
66 LLMs models have proven to be useful in MCQ generation in various medical fields<sup>7</sup>, such as  
67 pathology<sup>8</sup> and radiology<sup>9</sup>. The capability of LLMs to generate these questions offers infinite  
68 potential for education and training purposes, as the variety of possible questions they can  
69 develop is theoretically limitless. Also, question development is flexible and can be responsive to  
70 the learner, can be quickly improved, and developed on a very large scale. Currently, many of  
71 these AI-driven educational tools lack the ability to generate dynamic, high-quality MCQs that  
72 are tailored to specific medical specialties. Also, there is a need for an online and easily accessible  
73 platform that leverages LLMs to generate medical test questions, ensuring that these powerful  
74 educational tools are easily accessible by a diverse range of users.

75 In this context, we introduce *DailyMed*, an innovative LLM-based educational tool that can  
76 generate test questions such as MCQ across a wide variety of domains within the medical field.  
77 We outline the development process of this method, which leverages the capabilities of a  
78 number of LLMs. *DailyMed* combines state-of-the-art artificial intelligence with expert clinical  
79 input to create challenging and educational MCQs tailored for medical professionals at various  
80 stages of their careers. Furthermore, we evaluate our *DailyMed* pipeline for questions in the field

81 of clinical genetics, as a benchmark to assess its accuracy and efficacy in producing relevant and  
82 reliable questions.

83

## 84 **METHODS**

### 85 **1. Quiz generation**

86 DailyMed automatically generates a set of MCQ, covering a wide range of topics within certain  
87 medical specialties. The steps in our pipeline are illustrated in **Fig. 1**. The OpenBioLLM model was  
88 obtained via the Ollama server (<https://ollama.com/taozhiyuai/openbiollm-llama-3/tags>), for  
89 the other two models we used the API, as these weights have not been publicly released. (**Table**  
90 **1**).

91

92 *< Insert figure 1 here >*

93

94 The included steps in question generation are as follows:

#### 95 *1.1 Topic generation*

96 The LLM model is prompted to generate a set of 10 topics (**Fig. 1**, Step 1). These topics are based  
97 on specific guidelines, such that topics: 1) are specific and focused; 2) cover a wide range of sub-  
98 areas within the given subject; 3) test both knowledge and (clinical) application of concepts; 4)  
99 are not overly broad or vague; 5) are related to recent developments or current issues in the  
100 field; 6) avoid subjective, country-specific or sensitive topics (law, ethics, psychosocial factors).

#### 101 *1.2 Question generation*

102 Results from *query 1.1* are forwarded to the LLM model that iterates over the set of topics (**Fig.**  
103 **1**, Step 2). For each topic, the model is instructed to provide a question and a set of answering  
104 options, a hint which users can help to identify the right answer, and an explanation on why a  
105 certain answer is correct, so that a learner can learn from mistakes.

### 106 1.3 Semantic Scholar

107 Questions are checked on validity based on papers retrieved from Semantic Scholar<sup>10</sup> (**Fig.1**, Step  
108 3). For each question from *1.2 Question Generation* a search query is generated by Claude 3.5  
109 Sonnet. This query is based on the question, correct answer, explanation, and is limited to two  
110 concepts of two keywords each (e.g., ‘genetic testing AND Huntington's disease’). From the top  
111 5 papers (closest vector distance to query), the title, abstract, journal, influential citation count,  
112 journal, URL and other identifiers (such as PubMed) are then retrieved and used alongside the  
113 question information to let the LLM verify this question based on sources (**Fig. 1**, Step 4). If no  
114 supportive literature is found, the pipeline returns to question generation (Step 2).

### 115 1.4. Grading

116 Validated questions are then automatically graded according to difficulty levels: easy, medium,  
117 hard, or expert (**Fig. 1**, Step 5). This seeks to evaluate its difficulty level, which are subclassified  
118 as ‘easy’ which can be regarded as high-school level knowledge; ‘medium’ which is equivalent to  
119 college level; ‘hard’ questions on university level, and ‘expert’ level for questions requiring  
120 knowledge typical of an expert or healthcare provider working in the field. If the grade is  
121 considered ‘hard’ or ‘expert’, the improvement round (**Fig. 1**, Step 6) is not necessary, and the  
122 pipeline will go to *1.5 validation*. We thus chose not to include questions of easy or medium level.

### 123 1.5. Improvement round

124 Questions that are considered easy, or medium are required to go through this improvement  
125 process (**Fig.1**, Step 6). Here, the language model evaluates the question and is instructed to  
126 provide 3 points for improvement of the question (critique). Improvement ideas include for  
127 example:

- 128 1. Changing the question format to a case study where a patient presents with symptoms  
129 consistent with Duchenne Muscular Dystrophy (DMD), and asking about the likely  
130 molecular mechanism underlying their condition.
- 131 2. Adding more answer choices that involve other muscular dystrophies or genetic disorders  
132 affecting muscle function, requiring the test-taker to differentiate between them based  
133 on molecular mechanisms.

134 3. Asking about the specific location within the gene where pathogenic variants typically  
135 occur in DMD, such as 'What region of the DMD gene is most commonly affected by  
136 variants leading to Duchenne Muscular Dystrophy?'

137 Following this, the critique is piped into the next LLM query, which chooses one of the three ideas  
138 given, and incorporates this into the question. The question is then passed back to Step 4. There  
139 is a maximum of three attempts to improve the question; if it's still easy or medium after these,  
140 it will be used as is, without further efforts to increase its difficulty.

#### 141 *1.6. Validation*

142 After one or multiple cycles across previous steps, a question is subjected to validation in the last  
143 step (**Fig. 1**, Step 7). This validation evaluates whether the question, answer options, hint and  
144 explanations are highly likely to be correct. Conflicting and low-confidence questions are being  
145 returned to the question generation (step 2). Valid questions are then saved to the DailyMed  
146 database (including sources, the un-edited question).

#### 147 *1.7 Human Review and Publishing*

148 Finally, completed quizzes are subjected to human review to ensure quality, accuracy and validity  
149 before being published. In this stage, the expert reviewer can change any part of the questions.  
150 All these changes are tracked and stored in the database. The end results of these pipeline steps  
151 are a high-quality question on a sufficient difficulty level and distinctive answers. The reviewed  
152 and published quizzes are available on the DailyMed website and rotated throughout different  
153 10-question quizzes.

## 154 **2. Prompting**

155 In our prompting strategy, we use the Chain-of-Thought (CoT) technique to increase the  
156 reasoning capabilities of the LLM during generation pipeline. In Step 3 (Semantic Scholar), the  
157 LLM is instructed to evaluate the generated question (1.2) based on the abstracts of the retrieved  
158 sources. Here, we let the LLM reason first, before providing an answer (example provided in **Table**  
159 **2**), allowing us to capture both the reasoning process and the resulting evaluation. This approach

160 enables the LLM to self-assess and refine questions based on its own reasoning, aligning with  
161 findings that LLMs can self-improve through self-generated reasoning without extensive external  
162 supervision<sup>11</sup>. We use the same CoT technique for 1.4 grading, 1.5 improvement, 1.6 validation,  
163 albeit with different reasoning fields and criteria.  
164 The applied prompts can be found in the supplementary methods and/or on github  
165 (<https://github.com/Uhm-J/DailyMed>).

166

### 167 3. Models

168 OpenBioLLM-70B (8b quantized) (<https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B>) is a  
169 finetuned Llama-3 model on a biomedical dataset. It consists of 70 billion parameters and  
170 requires approximately 75 gigabytes of VRAM. To run this, we use RunPod, an external hosting  
171 platform. On this platform, it is possible to rent a temporary container (or Pod) with powerful  
172 GPUs. Inference is run through Ollama (version 0.3.12). To reduce costs, we do not rent storage  
173 containers, but instead download the model upon starting the quiz generation. And generate  
174 multiple quizzes sequentially. Additionally, we ran the pipeline with OpenAI's GPT-4o, and  
175 Anthropic's Claude 3.5 Sonnet, which are two state-of-the-art foundation models (**Table 1**).

### 177 4. Question evaluation

178 To assess the validity of our approach to generate valid multiple-choice questions, human ratings  
179 by clinical experts are still necessary. We analyzed these questions along several quality criteria:

- 180 - 1. *Validity*: the question and its proposed correct answer are rated for being factually  
181 correct or incorrect as a binary outcome.
- 182 - 2. *Clarity*: Questions are evaluated whether the question (and answers) is clearly  
183 formulated and unambiguous (i.e. not being multi-interpretable) and can be answered  
184 from the provided information. *Clarity* is rated on a scale from 1 (=question and or  
185 answers are unclear) to 5 (=clearly formulated question and answers).
- 186 - 3. *Originality*: the question is assessed for originality in its topic, the type of question  
187 and the provided set of answers. *Originality* is rated on a scale from 1 (=low creativity)  
188 to 5 (=high creativity).
- 189 - 4. *Relevance*: questions are scored based on their relevance for clinicians and their  
190 usefulness in daily practice. Relevance is rated on a scale from 1 (=irrelevant) to 5  
191 (=highly relevant).
- 192 - 5. *Difficulty*: question difficulty is assessed by the requirement of prior knowledge and  
193 experience, understanding of the topic on a deeper level, and distinctiveness between



194 correct and alternative answers (i.e. the presence of 'distractors' that are similar to the  
195 correct answer). Difficulty is rated on a scale from 1 (=easy) to 5 (=difficult).

196

## 197 **RESULTS**

### 198 *DailyMed website*

199 The DailyMed website (v1.0) is freely available via <https://dailymed.ai/>. A user-account is  
200 necessary that is able to track progress across different sets of questions and compare results  
201 with score results from other users.

### 202 *Question analysis*

203 We rated generated questions among five dimensions, including question validity, difficulty,  
204 clarity, originality, and clinical relevance, generated by the three different LLM models. Questions  
205 were specifically aimed at cases in the field of clinical genetics. Overall, OpenAI's GPT4-o model  
206 showed the highest average scores on all variables (**Fig. 2**).

207 *< Insert figure 2 here >*

208

209 Regarding validity, we observed high validity ratings of questions, answers and rationale  
210 generated by the three LLM models, suggesting all three can be used in our pipeline to develop  
211 valid questions. The highest validity score was observed by GPT4-o generated quiz questions  
212 question (average: 4.6, range: 3-5) compared to Claude (average: 4.4, range: 2-5) and  
213 OpenBioLLM (average: 3.9, range: 1-5). Statistical testing showed that the Validity score of  
214 OpenBioLLM was significantly lower than that of OpenAI\_GPT-4o (mean difference = -0.66,  
215  $p=0.0082$ ), while no significant differences were found between Claude\_3.5 and the other  
216 models. On the originality axis, Claude and GPT-4o scored similar ( $p = 0.78$ ), whereas  
217 OpenBioLLM had significantly lower scores compared to both these models ( $p = 0.0002$  and  $p <$   
218  $0.001$ , respectively). OpenBioLLM created significantly easier questions compared to both Claude  
219 ( $p = 0.001$ ) and OpenAI ( $p = 0.019$ ), with again no differences between these last two models.

220 Also, for relevance, there was no significant difference scores between Claude\_3.5 and GPT-4o  
221 ( $p = 0.64$ ), whereas OpenBioLLM had significantly lower scores compared to the first two models  
222 (Claude\_3.5,  $p = 0.0013$ ; OpenAI\_GPT-4o  $p < 0.001$ ). Lastly, questions showed slightly lower  
223 clarity in OpenBioLLM-compared to GPT-4o ( $p = 0.040$ ), but not with the Claude model. As a  
224 summary, we found that the GPT-4o model achieves the most valid questions that are also clearly  
225 formulated, and rated as more creative, clinically relevant and more difficult than the other  
226 models.

### 227 *Cross-quiz analysis*

228 In addition to question quality, it is of importance for a multiple-choice quiz to provide sufficient  
229 variety in topics. Excessive repetition of the same topics across sessions may limit the learning  
230 experience by reducing engagement and hindering exposure to a broader range of concepts. We  
231 analysed how often the different models were able to base questions on unique clinical topics  
232 (**Fig. 2c**). Here, we found that the GPT-4o model showed the largest diversity (89.8% of questions  
233 based on unique topics) compared to Claude (86.9%) and OpenLLM (80.0%).

### 234 *Low/high score examples*

235 Examples of a question that was scored relatively high and low respectively are shown in **Fig. 3**.

236

237 *< Insert figure 3 here >*

238

239 The first question shows a high score on the five factors (model: GPT-4o). The clinical case  
240 description makes the question more engaging and clinically relevant. The text provides clues  
241 about the potential differential diagnosis of this clinical presentation. The question formulation  
242 is clear and provides several plausible distractors as alternative answers (e.g. the case description  
243 concerns a boy and several diseases with an x-linked inheritance are provided). It concerns a  
244 relatively rare disease that is not often encountered in the clinic, which contributes to its  
245 originality rating. Also, the difficulty of the question is here influenced by the rarity of the disease  
246 and the alternative answers. We note that, although the difficulty level is scored higher, this does

247 not mean that it is a better question per se. For a relatively new audience, a somewhat less  
248 difficult question may be preferred (e.g. genetic disorders that are more often seen in the clinic,  
249 such as Fragile X syndrome, hereditary breast/ovarian cancer, Lynch syndrome).

250 In contrast, from the lower scored question (model Claude 3.5), it is clear that the topic is  
251 (currently) less clinically relevant (CRISPR-CAS gene-editing), as it is not formulated in the context  
252 of a clinical problem. There are less provided answers, and we rate these as not being valid. The  
253 alternative answer of possibilities for cancer therapy is just as correct as the suggested  
254 answer<sup>12,13</sup>. The topic and formulation of the questions are low in their originality, and clarity is  
255 low ('potential clinical application' is vague and quite broad, anything could have a potential  
256 clinical application in some sense).

#### 257 *Performance and cost-efficiency*

258 The pipeline for generating quizzes using large language models (LLMs) was evaluated in terms  
259 of runtime and cost efficiency. **Fig. 4** shows the runtime of the pipeline per question for each  
260 model. The cost per quiz varied based on the model used.

261 *< Insert figure 4 here >*

262 For OpenAI's GPT-4o, the average cost per quiz (=10 questions) was \$0.51, with an average  
263 runtime of 16 seconds per question. For Anthropic's Claude 3.5 Sonnet, the cost was higher at  
264 \$0.79 per quiz, with an average runtime of 31 seconds per question. For the OpenBioLLM model,  
265 although the average runtime was significantly longer at 53 seconds per question, the overall  
266 cost efficiency depended on the number of quizzes generated. The model requires a download  
267 time of approximately 25 minutes at the start, translating to a fixed cost of \$0.80. After model  
268 initialization, the average runtime for a single quiz was 544 seconds, resulting in an average cost  
269 of \$0.11 per quiz, excluding the model download cost.

270

## 271 **DISCUSSION**

272 In this study, we found that our DailyMed pipeline is able to generate high-quality multiple-choice  
273 questions that can be used at scale. Of the state-of-the-art LLM models, we found that for this  
274 application in clinical genetics, OpenAI's GPT-4o model performed best across different quality  
275 measures. One intuitive reason for the differences in model performance could be the size of the  
276 models. GPT-4o and Claude are significantly larger (with unknown exact parameter counts but  
277 speculated to be 200 and 175 billion, respectively). This larger size enables these models to  
278 construct more complex sentences with better attention mechanisms, which in turn improves  
279 adherence to guidelines and overall quality. Another factor could be better reinforcement  
280 learning to improve output preferences. Models like GPT-4o and Claude 3.5 Sonnet likely benefit  
281 from more reinforcement learning techniques and data, allowing them to align generated  
282 outputs more closely with human preferences, resulting in higher quality multiple-choice  
283 questions. At the same time, we observed considerable differences in run time (best model: GPT-  
284 4o) and costs (best model: OpenBioLLM) between these models.

285 As this overview describes the first version of our DailyMed pipeline, we identify several possible  
286 improvements. One potential improvement could involve utilizing OpenBioLLM to generate the  
287 initial question, followed by GPT-4o mini to refine or rewrite it. This approach leverages the  
288 efficiency of OpenBioLLM for question generation while taking advantage of GPT-4o mini's ability  
289 to perform simple rewriting tasks without relying heavily on its extensive knowledge base. This  
290 could optimize both performance and resource use.

291 Furthermore, other subspecialties will be added to the website in addition to clinical genetics,  
292 spanning other areas within the medical field (e.g. pediatrics, internal medicine). With the further  
293 development of AI and LLM models specifically, regular tuning and updating to the latest LLM  
294 models is necessary. The current pipeline will also be explored for additional models that may  
295 provide cost-effective or efficient alternatives (such as the Gemini model<sup>14</sup>).

296 We envision several applications and target audiences for whom our online tool can provide a  
297 useful learning experience. First, medical student with an interest in certain medical specialties,  
298 such as clinical genetics, can get more familiar with different case presentations in the clinic and

299 the decision-making process in these cases (e.g. before starting a residency in this subspecialty).  
300 With an estimated 45 unique topics (89% for GPT-4o) that are found in 5 quizzes (50 questions),  
301 a student that is not familiar with genetics may thus encounter 180 topics in a month' time with  
302 only a couple of minutes of daily practice and reflection. Secondly, coordinators of residency  
303 programs for clinical genetics residencies may incorporate quiz questions in their curriculum,  
304 either as a formal test, or as a way of self-reflection on possible knowledge gaps. Given limited  
305 time that residents may have for self-study in between clinical duties, an automated approach of  
306 using LLM for topic and question generation may be a valuable solution that allows residents to  
307 focus on high-yield content while minimizing the time spent on seeking new information. Finally,  
308 experienced clinicians can use DailyMed to stay updated on evolving knowledge, particularly in  
309 rapidly advancing fields like clinical genetics. Additionally, this tool can support continuing  
310 medical education (CME) initiatives, enabling clinicians to efficiently assess their knowledge in  
311 areas where they may need review.

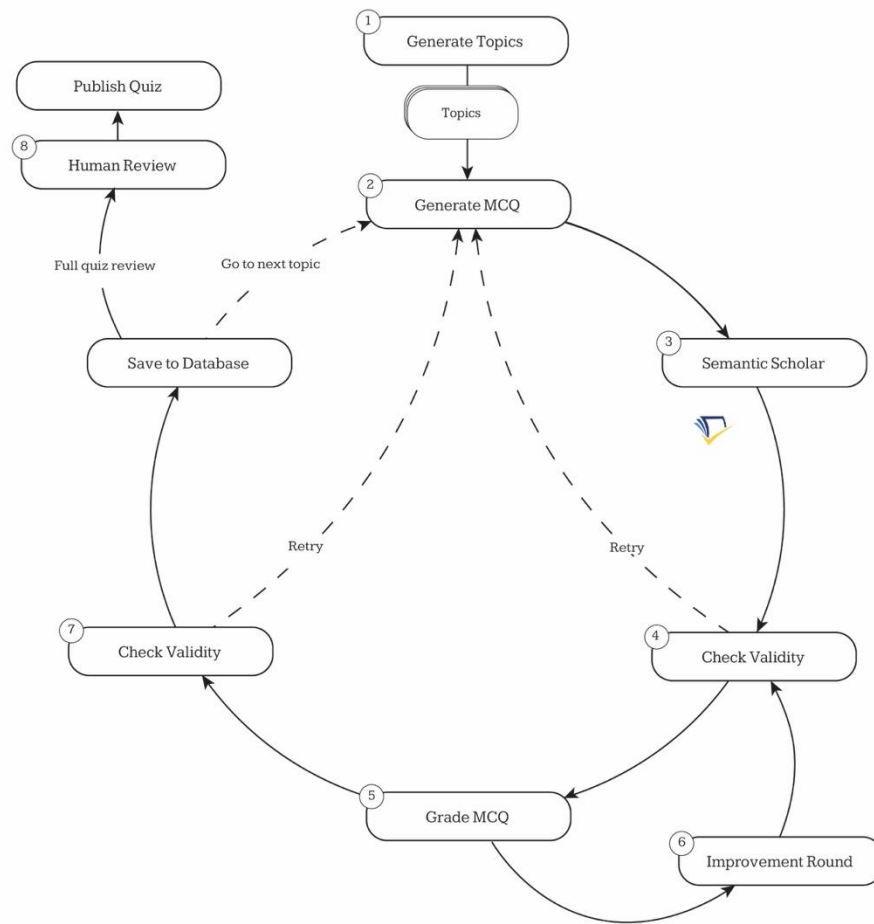
312 LLM models continue to develop rapidly, as illustrated by the number of parameters included in  
313 the latest models (>200 billion in GPT-4o). We envision that such models may become more  
314 responsive to the learner's progress and can be continuously updated to provide new material  
315 and/or new areas in which the learner can develop. Also, as generative AI and advanced LLMs  
316 allow the generation of images from text<sup>15</sup>, or even video<sup>16</sup>, multimedia material may be  
317 integrated in multiple choice questions, providing a more dynamic and interactive learning  
318 experience. Examples could be the presentation of certain clinical symptoms, dysmorphological  
319 features or congenital anomalies (which would not influence a patient's privacy) and visualization  
320 of imaging results, anatomical diagrams, or 3D reconstructions.

321 To conclude, our pipeline demonstrates the potential of large language models, particularly GPT-  
322 4o, to efficiently generate high-quality multiple-choice questions for medical education. Future  
323 versions of this system could integrate more specialized clinical input and explore additional  
324 models to further optimize question diversity, accuracy, and accessibility for a broader range of  
325 medical professionals.

326 **FIGURES**

327 **Figure 1. Overview of the Quiz Generation Pipeline from DailyMed.**

328 **(1)** Generate a set of 10 topics. **(2)** For each topic generate a question, answer options, a hint,  
329 and an explanation. **(3)** Retrieve papers from semantic scholar based on a query generated  
330 from the question. **(4)** Use the abstracts retrieved from step 3 to verify the quiz question. If the  
331 question could not be verified, the pipeline returns to step 2. **(5)** Grade question based on given  
332 criteria. **(6)** Critique the question and apply the critique given on the MCQ. **(7)** Check the validity  
333 of the question based on different criteria. If so, the question is saved to the database, and the  
334 next question is generated. **(8)** After quiz generation, the quizzes are reviewed before being  
335 published. MCQ: Multiple Choice Question.



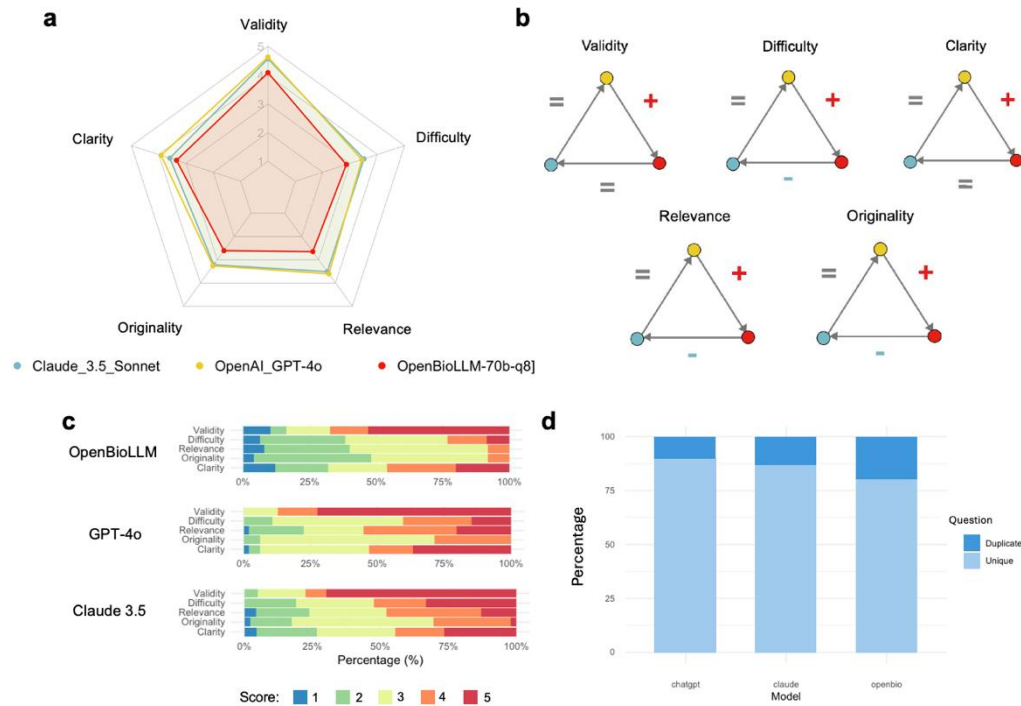
336

337

338

339

340 **Figure 2. Analyses of multiple choice question generation by the DailyMed pipeline.** A)  
341 spiderplot showing average scores on each question quality axis for the three tested LLM  
342 models; B) plot showing statistical comparisons between these models, where a plus sign  
343 represents a higher score, minus sign a lower score, and equal sign an equal score, colored dots  
344 represent the different models; C) distribution of quality scores as a percentage of rated  
345 questions (n=50 for each model); D) Proportions of unique and duplicate topics covered across  
346 quizzes



347

348

349 **Figure 3. Examples of questions that scored high and low respectively on question ratings. A)**  
350 **example of a highly rated question and rating scores; B) example of a lowly scored question and**  
351 **rating scores.**

352

**a** Rating: high

Model: OpenBioLLM

**Question 5** [1] [2] [3] [4] [5] Grade: Hard

A 2-year-old boy presents with recurrent severe bacterial infections, including pneumonia and sepsis. His medical history reveals delayed umbilical cord separation and poor wound healing. Laboratory tests show neutrophilia, but the neutrophils fail to migrate to sites of infection. Which of the following genetic defects is most likely responsible for this clinical presentation?

- A. CYBB mutation (X-linked chronic granulomatous disease)
- B. ITGB2 mutation (Leukocyte adhesion deficiency type I)**
- C. IL2RG mutation (X-linked severe combined immunodeficiency)
- D. BTK mutation (X-linked agammaglobulinemia)
- E. WAS mutation (Wiskott-Aldrich syndrome)

**Correct Answer:** B. ITGB2 mutation (Leukocyte adhesion deficiency type I)

**Explanation:** The clinical presentation of recurrent severe bacterial infections, delayed umbilical cord separation, poor wound healing, and neutrophilia with impaired neutrophil migration is characteristic of Leukocyte Adhesion Deficiency Type I (LAD-I). LAD-I is caused by mutations in the ITGB2 gene, which encodes the  $\beta 2$  integrin subunit CD18. This defect impairs the ability of neutrophils to adhere to blood vessel walls and migrate to sites of infection, resulting in the described symptoms.

Validity: 5  
Clarity: 5  
Originality: 4  
Relevance: 5  
Difficulty: 5

**b** Rating: low

Model: Claude 3.5

**Question 2** [1] [2] [3] [4] [5] Grade: Expert

Which of the following is a potential clinical application of CRISPR-Cas9 technology?

- A. Treatment of genetic disorders**
- B. Cancer therapy
- C. Infectious disease prevention

**Correct Answer:** A. Treatment of genetic disorders

**Explanation:** CRISPR-Cas9 has shown promise in treating genetic disorders by directly editing the DNA sequences responsible for these conditions. This can involve correcting or modifying specific mutations to restore normal gene function.

Validity: 2  
Clarity: 1  
Originality: 1  
Relevance: 2  
Difficulty: 1

353

354

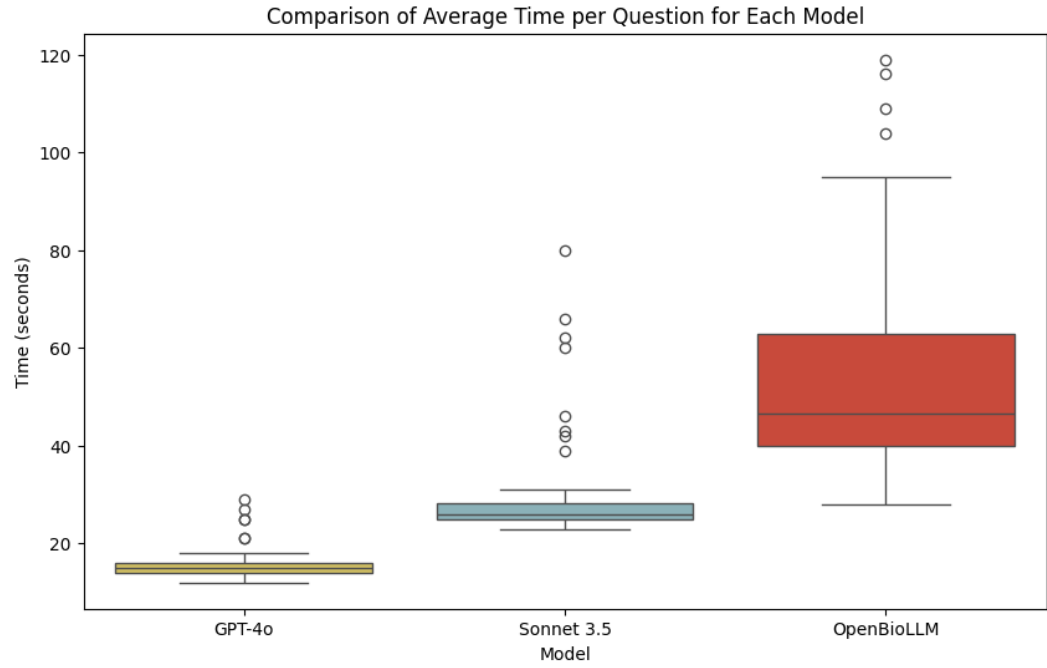
355

356

357



358 **Figure 4. Analyses of runtime per question for the three different models.** Boxplot showing  
359 the time for question generation, where colors represent three different LLM models.



360

361

362 **Table 1. Overview of three large-language models that were used in these analyses.**

363

<b>Model</b>	<b>Source</b>	<b>Inference method</b>	<b>Reference</b>	<b>Number of parameters</b>
GPT-4o	OpenAI	API	<a href="https://platform.openai.com/docs/models">https://platform.openai.com/docs/models</a>	>200 billion
OpenBioLLM-70B	Hugging Face (aaditya)	Ollama	<a href="https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B">https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B</a>	70 billion
Claude 3.5 Sonnet	Anthropic	API	<a href="https://www.anthropic.com/news/claude-3-5-sonnet">https://www.anthropic.com/news/claude-3-5-sonnet</a>	175 billion

364

365 **Table 2. Overview of LLM reasoning output.**

<b>Category</b>	<b>Details</b>
<b>Verification</b>	The quiz question is verified. DNA methylation is involved in the repression of gene expression. Sources 1 and 4 provide evidence that DNA methylation plays a role in repressing transcription by binding to specific regions of genes, leading to a decrease in their expression. Source 3 also supports this notion by showing that direct inhibition of transcription factor binding is the prevailing mode of gene repression by DNA methylation. Although source 2 does not directly mention DNA methylation's role in gene repression, it discusses MeCP2's function as a transcriptional repressor, which can be linked to DNA methylation. Source 5 further emphasizes the importance of DNA methylation in regulating gene expression and development.
<b>Verification Status</b>	PASS
<b>Difficulty Reasoning</b>	The question pertains to the role of DNA methylation in gene expression, which is an advanced genetic topic requiring a university level of understanding.
<b>Difficulty</b>	Hard

366

367 **REFERENCES**

- 368 1. Kasneci, E. *et al.* ChatGPT for good? On opportunities and challenges of large language  
369 models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).
- 370 2. Meyer, J. G. *et al.* ChatGPT and large language models in academia: opportunities and  
371 challenges. *BioData Min.* **16**, 20 (2023).
- 372 3. Wu, T. *et al.* A brief overview of ChatGPT: The history, status quo and potential future  
373 development. *IEEE/CAA J. Autom. Sin.* **10**, 1122–1136 (2023).
- 374 4. Gilson, A. *et al.* How does ChatGPT perform on the United States Medical Licensing  
375 Examination (USMLE)? The implications of large language models for medical education  
376 and knowledge assessment. *JMIR Med. Educ.* **9**, e45312 (2023).
- 377 5. Martínez, E. Re-evaluating GPT-4's bar exam performance. *Artif. Intell. Law* 1–24 (2024).
- 378 6. Zhang, Y. *et al.* Chain of Agents: Large Language Models Collaborating on Long-Context  
379 Tasks. *arXiv Prepr. arXiv2406.02818* (2024).
- 380 7. Biancini, G., Ferrato, A. & Limongelli, C. Multiple-choice question generation using large  
381 language models: Methodology and educator insights. in *Adjunct Proceedings of the*  
382 *32nd ACM Conference on User Modeling, Adaptation and Personalization* 584–590  
383 (2024).
- 384 8. Du, W. *et al.* Large Language Models in Pathology: A Comparative Study on Multiple  
385 Choice Question Performance with Pathology Trainees. *medRxiv* 2024.07.10.24310093  
386 (2024). doi:10.1101/2024.07.10.24310093
- 387 9. Mistry, N. P. *et al.* Large language models as tools to generate radiology board-style  
388 multiple-choice questions. *Acad. Radiol.* (2024).
- 389 10. Fricke, S. Semantic scholar. *J. Med. Libr. Assoc. JMLA* **106**, 145 (2018).
- 390 11. Huang, J. *et al.* Large language models can self-improve. *arXiv Prepr. arXiv2210.11610*  
391 (2022).

- 392 12. Song, X. *et al.* Delivery of CRISPR/Cas systems for cancer gene therapy and  
393 immunotherapy. *Adv. Drug Deliv. Rev.* **168**, 158–180 (2021).
- 394 13. Zhang, B. CRISPR/Cas gene therapy. *J. Cell. Physiol.* **236**, 2459–2481 (2021).
- 395 14. Team, G. *et al.* Gemini: a family of highly capable multimodal models. *arXiv Prepr.*  
396 *arXiv2312.11805* (2023).
- 397 15. Qin, J. *et al.* Diffusiongpt: LLM-driven text-to-image generation system. *arXiv Prepr.*  
398 *arXiv2401.10061* (2024).
- 399 16. Lin, H., Zala, A., Cho, J. & Bansal, M. Videodirectorgpt: Consistent multi-scene video  
400 generation via llm-guided planning. *arXiv Prepr. arXiv2309.15091* (2023).

401