MCQG-C++: Multiple Choice Question Generation and Evaluation With Critique, Correction, and Comparison Feedback

Anonymous submission

Abstract

Automatic question generation (QG) is essential for AI and NLP, particularly in intelligent tutoring, dialogue systems, and fact verification. Generating multiple-choice questions (MCQG) for professional exams, like the United States Medical Licensing Examination (USMLE), is particularly challenging, requiring domain expertise and complex multi-hop reasoning for high-quality questions. However, current large language models (LLMs) like GPT-4 struggle with professional MCQG due to outdated knowledge, hallucination issues, and prompt sensitivity, resulting in unsatisfactory quality and difficulty. To address these challenges, we propose MCQG-C++, an LLM self-refine-based (**C**ritique and **C**orrection) framework for converting medical cases into high-quality USMLE-style questions. By integrating expert-driven prompt engineering with iterative self-critique and self-correction feedback, MCQG-C++ significantly enhances human expert satisfaction regarding both the quality and difficulty of the questions. Furthermore, we introduce an `LLM-as-Judge`-based automatic metric to replace the complex and costly expert evaluation process, ensuring reliable and expert-aligned assessments. [1]

Introduction

In Artificial Intelligence (AI) and Natural Language Processing (NLP), automatic question generation (QG) from knowledge bases, texts, and images (Guo et al. 2024) plays a crucial role in enhancing question-answering (QA) models (Chen, Wu, and Zaki 2023; Guo et al. 2022), supporting intelligent tutoring systems (Zhao et al. 2022; Cai et al. 2023), improving dialogue systems, and aiding fact verification (Pan et al. 2021; Zhang and Gao 2023). Multiple-choice question generation (MCQG), a specialized type of QG, is extensively used in exams to assess students' knowledge efficiently (Zhang et al. 2021). However, creating MCQs is labor-intensive, requiring the design of effective stems, prediction of common errors as distractors, and provision of corrective feedback (Ch and Saha 2018). In professional fields, MCQs often require field experts because they need to reflect real-world scenarios and involve complex multi-hop reasoning. These are unique challenges not typically encountered in general QG tasks.

---

[1] We will release codes and data upon acceptance.

The United States Medical Licensing Examination (USMLE) exemplifies the need for high-quality MCQG (Scoles 2008). Preparing for the USMLE costs medical students over $5000 on average (Philip A Bucur 2019). For exam boards and instructors, creating MCQs is both time-consuming and expensive (Gierl, Lai, and Turner 2012). Any application that can automate this process is highly valuable to medical educators (Homolak 2023; Gilardi, Alizadeh, and Kubli 2023). Due to the high difficulty with the need for domain knowledge and complex reasoning, USMLE questions are becoming important large language models (LLMs) benchmarks (Jin et al. 2021). Top LLMs like GPT-4 have shown over 90% accuracy on sample USMLE questions (Achiam et al. 2023). Recent research explores leveraging GPT-4's potential in USMLE-MCQG to improve question generation efficiency for educators and assist students in exam preparation (Klang et al. 2023; Agarwal, Sharma, and Goswami 2023; Biswas 2023).

However, relying solely on LLMs like GPT-4 to generate USMLE questions presents several challenges. Firstly, their performance is constrained by their training data, leading to two major issues: outdated knowledge (Mousavi, Alghisi, and Riccardi 2024) and hallucination (Zhang et al. 2023) (**Limit1**). Outdated knowledge means that LLMs can only repeat or integrate old USMLE questions or medical documents in their training data during the generation. Consequently, they struggle to create new questions based on the latest medical cases or guidelines like medical experts. Hallucination refers to LLMs potentially providing misinformation in questions, which could harm students' learning outcomes. Secondly, generating questions on specific concepts requires precise prompts (**Limit2**), which students might not know how to formulate (Sahoo et al. 2024). For example, given one USMLE question about cardiovascular physiology, if GPT-4 is asked to "`Generate a more difficult USMLE question.`", it only provides a longer stem. It will focus on the specific topics and test points only when specifically asked to generate a question "`testing knowledge of the hemodynamic profile of aortic stenosis.`" Lastly, the quality and difficulty of the questions often do not meet expectations (Artsi et al. 2024; Benítez et al. 2024) (**Limit3**). As shown in Figure 1, the output from GPT-4 contains several errors and inconsistencies, such as extraneous information,
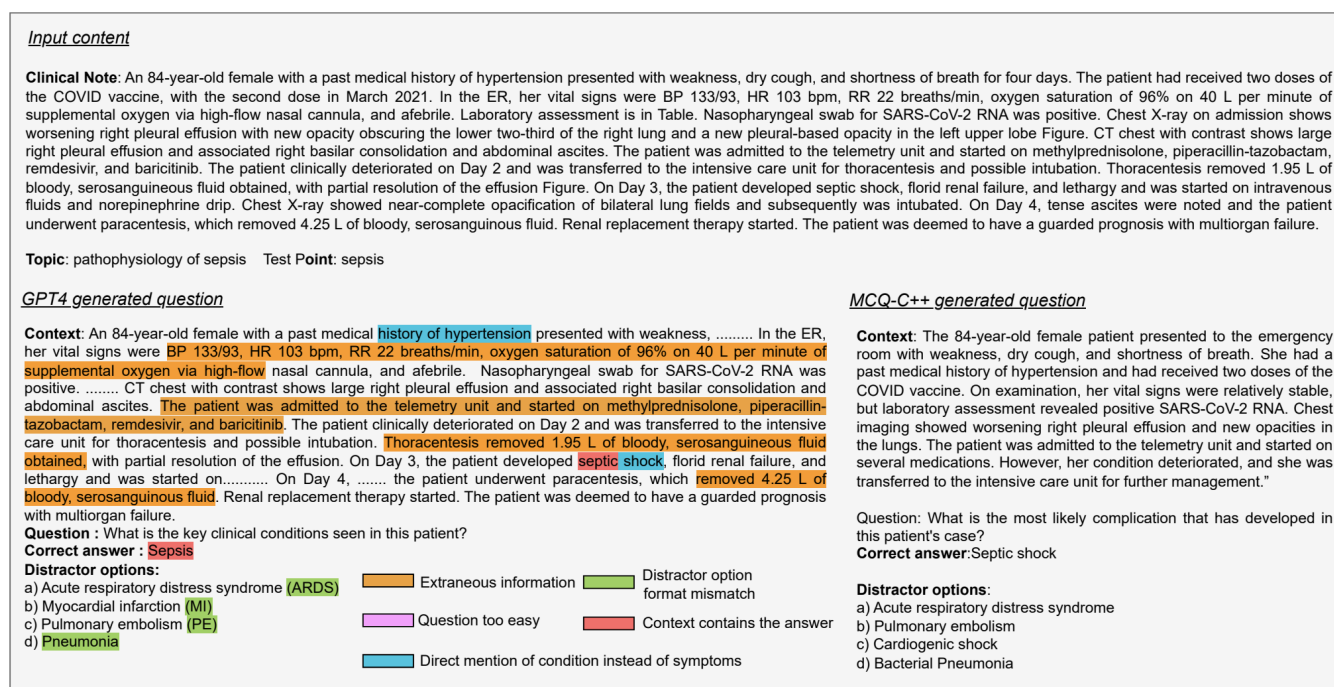
Figure 1: USMLE MCQ generated by GPT-4 and MCQG-C++. The GPT-4 question contains several errors and inconsistencies, such as extraneous information, a distractor option format mismatch, mentioning symptoms instead of conditions, and a context that contains the answer. The MCQG-C++ addresses these issues, resulting in a higher quality question that aligns the context, question, and answer options more coherently. Irrelevant details are removed, the question focuses on the key clinical condition of sepsis, distractor options are presented in a consistent format, and the context no longer gives away the answer.

mismatched distractor option formats, direct mention of the condition instead of symptoms, and context that includes the answer. The generated questions often lack the depth required for students' higher-order critical thinking.

To address these challenges, we propose a new task: converting medical cases that appear in case reports or clinical notes into USMLE-style MCQs. Our approach involves several key design elements to alleviate the above limitations:

For **Limit1**, previous research has attempted to prompt LLMs to generate USMLE-MCQs by following expert-crafted instructions in zero-shot setting or further adding existing questions as few-shot examples (Artsi et al. 2024). To our knowledge, we are the first to qualitatively and quantitatively study how to convert medical cases into USMLE-MCQs. These medical cases provide valuable real-world information on disease progression, accurate assessments, diagnoses, and potential treatment plans. Using the latest medical cases as input, LLMs can generate up-to-date questions, thereby minimizing the limitations of outdated knowledge. Additionally, grounding questions' key elements in the original medical cases can help LLMs reduce hallucinations, enhancing the reliability of the generated content.

For **Limit2**, we follow the National Board of Medical Examiners (NBME) guidelines [2] to establish a checklist of 41 target topics covering all potential exam areas. We then deployed a ColBERT retriever (Santhanam et al. 2021) using

USMLE Content Outline [3] as a collection for test points retrieval. Each input medical case in our experiments is evaluated by medical experts with exam experience to determine if it contains sufficient information to generate questions related to the specified target topics and test points. We also compared the topics and test points identified by experts with those generated by LLMs, assessing their impact on the quality and difficulty of the resulting questions.

For **Limit3**, as illustrated in Figure 1, we used the triplets (medical case, topic, and test point) as our question generation pipeline input. We work with experts for prompt engineering based on USMLE guidelines. We then created our MCQG-C++ (e.g., self-refine with iterative Critique and Correction feedback) following three steps. S1 - Initial MCQ Generation: Generate an initial USMLE-MCQ based on the triplets. S2 - Critique Feedback: Prompt the LLM itself to provide feedback on the S1 USMLE-MCQ. S3 - Correction Feedback: Correct the USMLE-MCQ based on the S2-generated critique feedback. Through iterative critique and correction, MCQG-C++ significantly enhances the quality and difficulty of generated USMLE-style MCQs. Human evaluations confirm its effectiveness, showing a strong preference for MCQG-C++ generated questions, with a preference ratio of 72.5% in win, 10% tie, and 17.5% loss when compared to GPT-4 generated questions. In terms of diffi-

INIT

u

QUESTION ANSWER

u,r,a_a

CRITIQUE

$u,r,a_{a'},f$

CORRECTION

STOP
< n,t,k >, u = <c, q, a, d>

no

yes

stopping criteria: Rubric (Context) Rubric (Question) Rubric (Correct Answer) Rubric (Distractors) Rubric (CoT/Reasoning)

1. INITIALIZATION

The initialization step generates the four USMLE question components in four substeps, each using the few-shot retriever and generator sequentially.

note topic keypoint → [context,question] fewshot exemplars → context

context → [context, question] fewshot exemplars → question

context question → [context, question, answer] fewshot exemplars → answer

context question answer → [context, question, answer, distractor options] fewshot exemplars → distractor options

2. QUESTION ANSWER

The generator is given the components in the form of a USMLE question and it generates an answer and it's reasoning.

context question options → attempted answer reasoning → context, question, answer, distractor options, attempted answer, reasoning

3. CRITIQUE

The critique step critiques the four USMLE question components and reasoning in five substeps using exemplars and component-specific rubrics; if feedback scores exceed a threshold, the generation process stops.

note topic keypoint

context, question, answer, distractor options, attempted answer, reasoning

context feedback fewshot exemplars → context feedback

question feedback fewshot exemplars → question feedback

answer feedback fewshot exemplars → answer feedback

distractor options feedback fewshot exemplars → distractor options feedback

reasoning feedback fewshot exemplars → reasoning feedback

4. CORRECTION

The correction step uses previous feedback to improve USMLE question components, followed by the answer step, iterating further as needed.

context, question, answer, distractor options, attempted answer, reasoning + component feedbacks, refinement fewshot exemplars → context, question, answer, distr. options
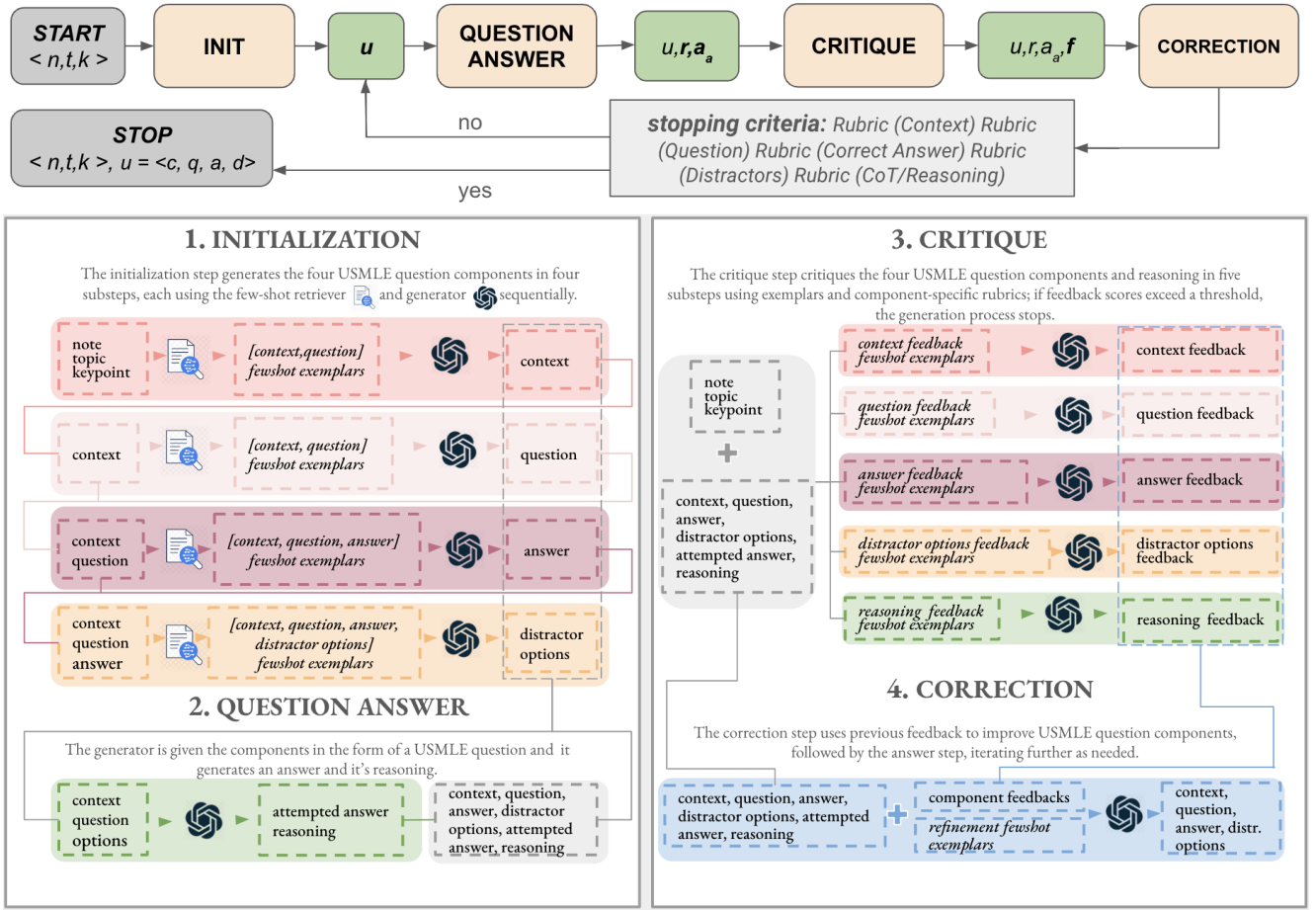
Figure 2: The framework for generating USMLE-style questions involves four main steps, as illustrated in the figure. First, the initialization generates the context, question, answer, and distractor options using retrieval and generation models. The generation model then answers the generated question along with a reasoning. Next, the feedback step evaluates the generated components on various rubrics and generates textual feedback and scores, stopping if feedback scores exceed a threshold. Finally, the refine step iterates by using the feedback to improve the components before cycling back to the answer step.

culty, MCQG-C++ generates significantly more challenging questions compared to GPT-4. Specifically, when provided with expert-identified topics and test points, there is an 80% reduction in easy questions, a 2.25-fold increase in medium questions, and a 4-fold increase in hard questions.

Finally, designing a reliable reference-free metric to automatically evaluate the quality of system-generated USMLE-MCQs is challenging. Recent research indicates that `LLM-as-Judge` correlates more closely with human evaluations than traditional metrics (Chen et al. 2023; Chiang and Lee 2023; Kocmi and Federmann 2023; Zheng et al. 2024; Zhang et al. 2024; Kim et al. 2023, 2024), though these methods remain underexplored in medical NLP tasks. Our goal is to replace the costly expert evaluation process in USMLE-MCQG with `LLM-as-Judge`. We used 30 criteria designed by experts in their human evaluation, covering different aspects of USMLE questions, to guide `LLM-as-Judge` in providing rating or comparison feed-back on different systems' questions. By further exploring the filtering methods for the 30 criteria, we finally screened out a combination of 10 key criteria. This improved the correlation between `LLM-as-Judge` and expert evaluations, as measured by Cohen's kappa, from 0.124 (slight reliability) to 0.539 (moderate reliability). Using the results from this automated evaluation system, it was shown that the preference rate for questions generated by MCQG-C++ over those generated by GPT-4 was 79.97% in favor and 20.03% against. Moreover, MCQG-C++ demonstrated overall improvements across most 10 criteria within the 5 components, not just in a specific area.

Method

Problem statement: Given a medical case n detailing a patient's history, diagnosis, treatment, and outcome, we aim to generate a USMLE question u. Here, $u = <c, q, a, d>$ consists of a context (c), which is a modified excerpt from n

tailored to align with the target style and obscure evidence information that can easily lead directly to the correct answer; a question (q) based on the generated context, which may be one or several sentences; the correct answer (a) to this question; and several distractor options (d).

Topic and test point identification

As discussed in **Limit2**, generating questions using LLMs without specific guidance, such as defined topics t and test points k, often results in questions that lack relevance, quality, and appropriate difficulty. These questions may fall outside the scope of the USMLE exam, being either too simple or overly complex. Therefore, the quality and difficulty of the generated USMLE questions are significantly influenced by the selection of topics t and test points k.

Topics t refer to a list of target topics selected from 41 potential topics outlined in the NBME official guidelines, categorized into 10 sections [4]. Both the LLM and human experts are provided with this list to generate a maximum of five topics (t) that are highly relevant to the medical case.

Test points k refer to the core concepts closely related to the correct answer. We employ the ColBERT retriever [5], denoted as π_{rtr}, to retrieve suitable test points from 18 sections [6] of the USMLE content outline. This test points retrieval process involves querying the model with the medical case (n) and the identified topics (t). The result is a list of highly specific USMLE concepts, such as 'keloid formation'. π_{llm} then generates specific test point concepts using these filtered concepts as a foundation, ensuring they are directly relevant to the context (c) and topics (t). These test points can either originate from the filtered list of USMLE concepts or be derived from the content of the medical case itself. Additionally, we ask human experts to identify specific test points they believe are related to their selected topics, referring to the key medical concepts mentioned in the USMLE content outline. More details can be found in Appendix Table 10. In Section , we will compare the topics and test points generated by humans and LLMs and evaluate their impact on the quality of the resulting questions.

Initialization

As illustrated in Figure 2, the MCQG-C++ pipeline begins with the INIT step, which comprises four generation steps, each targeting a component of the goal $u = <c, q, a, d>$. To assist the model in referencing similar examples for better generation of each component of u, we deploy a ColBERT retriever model R to retrieve a small set of USMLE examples from the MedQA (Jin et al. 2021) question bank. As shown in Figure 2, given the input $< n, t, k >$ from Topic and Test point identification ste, R first uses $< n, t, k >$ as a query to retrieve few-shot examples, and then LLM follows the INIT-c prompts in Appendix Table 14 to generate the context c. Subsequently, after obtaining c, we continue to use $< n, t, k, c >$ as a query to retrieve few-shot examples and follow the INIT-q prompts in Appendix Table 14 to

---

[4]Details can be found in Appendix: Topics and Test points

[5]https://github.com/stanford-futuredata/ColBERT

[6]Details can be found in Appendix: Topics and Test points

generate the question q. The exact process is applied for t and k. It is important to note that we trim the retrieved examples for each component; for instance, in INIT-c prompts, we only retain the context component of each example, and similarly, for the other three components q, a, d, only the relevant parts are kept. As demonstrated in Appendix Table 13, this INIT step already results in a USMLE MCQ for a given input $< n, t, k >$. However, as discussed in Figure 1, despite incorporating several advanced prompting engineering methods in the INIT step—including prompts designed by medical experts according to USMLE guidelines, tailored topics&test points for each input medical case, as well as step-by-step retrieval and generation—the USMLE MCQs generated by the LLM in the INIT step still fall short of the required quality and difficulty.

Question Answering Feedback Collection

Inspired by recent work that augments the standard QG model with an additional QA module to further constrain the generated questions (Su, Xu, and Fung 2022; Xie et al. 2020; Sun et al. 2019), we add a Question Answering Feedback Collection module. This module provides additional feedback from the question-answering perspective to further challenge the quality and difficulty of the questions. Our motivation stems from the fact that LLMs like GPT-4 have proven to perform exceptionally well on USMLE QA tasks, achieving human-expert levels in both accuracy and reasoning processes. By analyzing the rationale and correctness of the final answers produced by the language models during the QA process, we can gather valuable insights into the quality and difficulty of the questions. Specifically, given the context c, question q, and options composed of $a \cup d$ (with their order shuffled), we collect the LLM's generated attempt a_a along with the reasoning r that supports a_a in this step. An example output is shown in Appendix Table 13.

Critique

After generating all components in the INIT step and the Question Answering step, the LLM is asked to critique each component. The set to be critiqued is $u_{\text{critique}} = \langle c, q, a, d, r \rangle$. The LLM receives a scoring guide G, which includes all aspects that need to be evaluated for each component. The prompt includes this guide G as well as several manually written example critiques of scored components $E_{\text{critique}}^{\text{fs}} = \langle e_c^{\text{fs}}, e_q^{\text{fs}}, e_a^{\text{fs}}, e_r^{\text{fs}} \rangle$ and $\langle G, n, t, k, c, q, a, d, a_a \rangle$. The final output of this step is LLM critique feedback on all components, $f = \langle f_c, f_q, f_a, f_d, f_r \rangle$, which includes short text critiques and scores for each aspect. The aspects for scoring different components in G are as follows: **Context:** Relevant, Concise, Coherent, Consistent, Specific, Fluent, Clueing, Completeness, and Misdirection. **Question:** Relevant, Clear, Concluding, Difficulty, and Clarity. **Correct Answer:** Relevant, Occurrence, Justification, Depth of Understanding, and Prevention of Guesswork. **Distractors:** Format, Length, Relation, Variation, Plausibility, Differentiation, and Common Mistakes. **Chain of Thought/Reasoning:** Logical Flow, Evidence-Based Reasoning, and Consideration of Options. We provide detailed explanations for

each aspect of every component in Appendix Table 13, and LLM-Critique prompts in Table 14. The total score for each component is calculated by summing up all individual aspect scores. A sample output is provided in Appendix Table 13.

Correction

The Correction step aims to correct each of the generated components of u in the INIT step. LLM is prompted with $<E_{correction}^{fs}, n, t, k, f, c, q, a, d, a_a, r>$ and asked to generate $u_{correction}$, which can perform better on all the component's critique aspects. Here $E_{correction}^{fs}$ is a set of manually written few shot examples which are incrementally improving using the previous output's feedback. $u_{correction}$ is again given to the Critique step to check if the feedback scores are greater than a fixed threshold, which, if true, stops the iterative Critique and Correction and, if not, continues [7].

Experimental Design and Setup

Experimental Setup For all our experiments with MCQG-C++ we use the chat completions API from OpenAI and the GPT-4 release, which has a context window of 8192 tokens, and the values of the hyperparameters temperature and top-p are set to 1. Similarly, for all other models used for the `LLM-as-Judge` comparison feedback generation we used their default hyperparameter settings.

Experimental Design Our experimental design is motivated to answer the following two research questions [8]:

- RQ1: Evaluate whether MCQG-C++ improves both the quality and difficulty of the generated questions. RQ1-1: Question quality; RQ1-2: Question difficulty
- RQ2: Determine if `LLM-as-Judge` is an effective metric. RQ2-1: Rating or comparison; RQ2-2: Which LLM is the better judge?; RQ2-3: Aspect filtering to improve `LLM-as-Judge` evaluation reliability

RQ1: Efficacy of MCQG-C++ To address the first research question, we employed the MCQG-C++ pipeline to generate USMLE questions and compared them with baseline questions generated by GPT-4 under identical inputs and settings. Our inputs consisted of medical cases n from the PMC-Patients dataset, with topics t and test points k that were both human-annotated and generated by the LLM. We generated 373 questions from the human-annotated $\langle n, t, k \rangle$ set and 385 questions using the LLM-generated $\langle n, t, k \rangle$ set.

To assess the quality of the questions (**RQ1-1:** `Question quality`), we engaged two medical experts to express their preferences between the two sets of system-generated questions based on an annotation guideline (Appendix Table 15). System 1 consists of questions generated in the initial step (INIT-step) by GPT-4, referred to as GPT-4 questions[9], while System 2 comprises the MCQG-C++ questions, as illustrated in Figure 1. The evaluators

were blinded to the source of the questions, and the order of the questions was randomized for each data point. We calculated the Percentage Agreement (87.5%) and Cohen's kappa (0.66722) for the two evaluators' preferences, indicating substantial reliability of our human evaluation settings. Subsequently, a third human expert facilitated discussions with the initial evaluators to make the final decision for each data point, representing the final human expert preference (referred to as `Expert X`). These final preferences were used in the RQ1.1 results and subsequent discussions for RQ2 as the golden preference for `LLM-as-Judge`.

For evaluating the difficulty level of the questions (**RQ1-2:** `Question difficulty`), the human evaluators were also asked to classify the difficulty of both questions into one of three categories: Easy, Medium, and Difficult.

RQ2: Reliability of `LLM-as-Judge` To explore **RQ2-1**, we utilized ratings generated during the MCQG-C++ critique step as `LLM-as-Judge (rating)` scores. For `LLM-as-Judge (comparison)`, we employed the same guidelines and settings used in the human evaluation from RQ1-1, prompting the model to choose a preferred output between the two system-generated questions (e.g., GPT-4 and MCQG-C++). To ensure the results were not biased by the inherent positional preferences of the LLM (e.g., **position bias**) (Wang et al. 2023a; Zheng et al. 2023; Zeng et al. 2023), we collected preferences from two sequence settings: one where the GPT-4 generated output was shown first followed by the MCQG-C++ output, and the reverse order. We filtered out preferences that did not match across the two sequence settings, using the remaining data for further analysis. To ensure the results were not influenced by **length bias** (i.e., LLMs like GPT-4 prefer longer generations during their automatic evaluation) (Wang et al. 2023b; Zeng et al. 2023), we deliberately included those data points in the evaluation data which had the least context length ratio for the initial and the MCQG-C++ questions. These examples constituted half the human evaluation data and the rest were chosen randomly from the dataset. Notably, the guidelines used in human evaluation, `LLM-as-Judge (rating)`, and `LLM-as-Judge (comparison)` were consistent. Thus, by standardizing the evaluation data, we could fairly compare the judgments of Expert X and `LLM-as-Judge (rating/comparison)` under identical settings to assess the reliability of `LLM-as-Judge`.

To investigate **RQ2-2**, we prompted various LLMs [10] using the `LLM-as-Judge (comparison)` settings and compare the preference with Expert X.

In **RQ2-3**, we focused on exploring two heuristic algorithms to improve the effectiveness of `LLM-as-Judge (rating)` through aspect filtering [11]. Specifically, we first calculated the correlation of each aspect score from the collected rating feedback with human preferences, sorting them in descending order by percentage agreement or Cohen's kappa. In the **Greedy** algorithm [12], we sequentially

---

[7]Stopping criteria can be found in Appendix .

[8]We have included the Ablation Study and Output Analysis of MCQG-C++ in the Appendix .

[9]The INIT-step employs GPT-4 with advanced prompt engineering, hence the designation GPT-4 questions

---

[10]`GPT-3.5-turbo`, `GPT-4`, `GPT-4o`, `Claude-3-haiku/opus`, and `Claude-3.5-sonnet`

[11]5 components with 30 aspects shown in Appendix Table 13.

[12]The Appendix provides the pseudo-codes of Greedy Aspect

added new aspects to the final rating score calculation based on their correlation, from highest to lowest, then computed the new correlation between `LLM-as-Judge (rating)` and human preference. Similarly, in the **All-Combo** algorithm, we calculated the final rating score for all possible combinations of the top n aspects ($1 <= n <= 11$)[13], selecting the combination with the highest correlation to human preferences as the output of the All-Combo algorithm.

Results and Discussion

RQ1: Efficacy of MCQG-C++

Table 1 demonstrates the overwhelming advantage of MCQG-C++ over GPT-4 with a 70-80% win rate in human preference about question quality (**RQ1-1**). We also observe a decrease of about 80% in easy questions, a 2.25 times increase in medium questions, and a 4 times increase in hard questions with the input of these medical cases with expert-provided topics and key points (Table 1 Expert-x H.). For machine-provided topics and key point cases, the proportion of easy questions decreased by 33.3%, a 2 times increase in medium questions, but there was no increase in the proportion of hard questions (**RQ1-2**). This demonstrates the effectiveness of MCQG-C++ in increasing the difficulty of questions. This also indicates that the quality of topics and key points provided by experts is higher, so LLM can generate more difficult questions by thinking more deeply during the critique and correction steps. Further improving the quality of the machine-provided topics and key points can be a future direction for improvement.

One of the main issues with the questions generated by GPT-4 is that they often directly include the correct answer or too obvious relevant keywords within the context component (Appendix Table 17 Case Studies 1 and 3). So the questions directly generated by GPT-4 often make the answers obvious, but MCQG-C++ can modify this information into hints for the correct answer through its critical and corrective steps, which experts consider a better way to construct questions for candidates (Case Study 4).

Another finding is that the questions generated by MCQG-C++ are more concise compared to those by GPT-4 (Case Study 1). Experts pointed out that this conciseness makes them more similar to real USMLE questions. Our experiments show that GPT-4 adopted a very conservative strategy when generating context due to our emphasis on hallucination issues in the prompts. This strategy involves copying and pasting much information from raw medical case inputs to avoid generating potentially incorrect new facts. Although this does reduce the occurrence of hallucinations - our human assessment shows that 7.5% of GPT-4 problems contain factual errors - it inevitably sacrifices the typical simplicity and highly refined information presentation of USMLE problems, as well as the logical coherence of multi-hop reasoning between information in context, question, correct answer, and distractors. This is a significant stylistic difference between USMLE questions and

---

Selection and All-Combo algorithms.

[13]We observed diminishing returns in the top 9 10 11 aspects, and stopped at 11 for time and computational resources required.

| | t, k | Quality | | | Difficulty | |
		GPT4	Ours	Tie	GPT4	Ours
1	H.	15%	**80%**	5%	50:35:15	35:30:35
	M.	20%	**80%**	0%	70:20:10	45:35:20
2	H.	20%	**70%**	5%	60:35:5	30:65:5
	M.	20%	**75%**	0%	65:30:5	60:35:5
x	H.	15%	**75%**	10%	70:20:10	15:45:40
	M.	20%	**70%**	10%	75:20:5	50:40:5

Table 1: The Quality part shows expert preference counts for the GPT-4 generated question and the MCQG-C++ (Ours) question. The data is divided into Human (H.) and Machine (M.) based on how the topic and key points were generated. Expert x represents the preferences reached by the experts after a round of preliminary annotation. The percentage agreement between Expert 1 and Expert 2 is 87.5% (Human: 90%, Machine: 85%). The Cohen's kappa between Expert 1 and Expert 2 is 0.66722 (Human: 0.75, Machine: 0.571428), indicating substantial reliability. The Difficulty part shows the difficulty level distribution (e.g., Easy: Medium: Hard) for GPT-4 and Ours, annotated by human experts. Compared with GPT-4, the MCQG-C++ pipeline generates better quality USMLE multiple-choice questions while producing significantly more medium and hard questions.

the original medical cases. We also found that prompt engineering was ineffective in resolving this issue. We interpret this as a shortcut behavior learned by GPT-4 during aligning with human preferences stage training (e.g., RLHF (Ouyang et al. 2022)) to reduce output diversity to mitigate hallucinations (Kirk et al. 2023). In contrast, MCQG-C++ maintains a high level of factual accuracy (with 5% factual errors) and further improves the quality of USMLE questions by iteratively criticizing and correcting each component of the question, as well as shifting the perspective from QG to QA to make it closer to actual exam questions. This is another major reason why MCQG-C++ outperforms GPT-4 in human evaluation. However, when MCQG-C++ reduced the amount of useless information provided in the questions to increase quality and difficulty, omitting too much information sometimes made inferring the correct answer more challenging (Case Study 2), although this was a rare occurrence in our human evaluation (7.5%).

RQ2: Reliablity of LLM-as-Judge

For **RQ2-1**, we found that the **rating** method performs slightly better than the **comparison** method. Percentage Agreement between GPT-4-Comparison and Expert X: Total agreement: 53.8% (Human: 58.33%, Machine: 50%). Percentage Agreement between GPT-4-Rating and Expert X: Total agreement: 61.538% (Human: 58.33%, Machine: 64.285%). For **RQ2-2**, as shown in Appendix Table 9 and 16, we observed that GPT-4 as the LLM-as-judge has the least position bias (35% of the data showed inconsistent results under two different orders, while other models showed more than 50% inconsistency). Moreover, GPT-4 has a relatively higher correlation with human evaluations in the valid data (i.e., data without position bias) with a percent-

age agreement of 61.53% and a Cohen's kappa of 0.2121. On the other hand, Claude-3-opus showed 52.5% position bias but had the highest correlation with human evaluation in the valid data, with a percentage agreement of 68.42% and a Cohen's kappa of 0.3152. For **RQ2-3**, as shown in Appendix Table 22 23, the All-Combo method found our optimal aspect combination [14], including the following 10 aspects in Table 2. The LLM-as-judge model based on these 10 aspects achieved an 80% percentage agreement and a Cohen's kappa of 0.5389 with human evaluation. Ultimately, using this best `LLM-as-judge` metric, we evaluated all 341 human topics and key points, resulting in the win rate for MCQG-C++: GPT-4 = 79.8%: 20.2%. For all 359 machine topics and key points, the win rate was MCQG-C++: GPT-4 = 80.1%: 19.9%. We also observed that MCQG-C++ outperformed GPT-4 with improvements across all five components, rather than excelling in just one component.

Related Work

LLMs for Generating Medical MCQs Cheung et al. (2023) conducted the first study comparing LLMs with humans in medical exam MCQ generation. Using GPT-3.5, they generated MCQs from medical textbooks. GPT produced 50 MCQs in 21 minutes—10% of the time taken by humans—but human-written questions were of higher quality and effectiveness, outperforming AI in 60% of acceptable MCQs. Other studies, such as Klang et al. (2023), evaluated AI-generated MCQs using GPT-4 and found them mostly effective, though lacking in clinical context. Agarwal, Sharma, and Goswami (2023) compared GPT-3.5, Bard, and Bing in generating MCQs for a physiology course, noting that GPT produced more effective but less difficult questions. Overall, while LLMs can quickly generate MCQs, their quality and difficulty often do not meet expectations. We build on this work by qualitatively and quantitatively studying how to convert real-world medical cases into high-quality and challenging USMLE-style MCQs.

Self-Refine using LLM Feedback Learning from feedback, including human preferences (Ouyang et al. 2022) and AI-generated feedback (Lee et al. 2023), has shown promise in aligning LLM outputs with desired outcomes. Compared to scalar feedback, natural language feedback provides more detailed information (Scheurer et al. 2022; Ma et al. 2023) and has proven effective for self-correction (Pan et al. 2023). Recent research indicates that LLMs can self-correct their responses to meet diverse requirements, ranging from reducing harmful content and debugging code in general domains (Achiam et al. 2023; Madaan et al. 2024) to generating clinical dialogues and documentation in specialized medical contexts (Wang et al. 2024; Das, Albassam, and Sun 2024; Yao et al. 2024). Inspired by these findings, we explore iterative self-refinement for generating USMLE MCQs.

LLM-as-Judge LLMs have been increasingly used for automatic evaluation, with studies showing that models like

---

Corr. w/ Expert-x	P. Agreement	C. Kappa
w/o aspect filtering	67.5%	0.226190476
PA-based Greedy	72.5%	0.342301943
CK-based Greedy	72.5%	0.423328965
PA-based All-Comb	80.0%	0.505409582
CK-based All-Comb*	80.0%	0.538904899

LLM-as-Judge* preference		t k (H.)	t k (M.)
MCQG-C++ win		79.80%	80.10%
GPT-4 win		20.20%	19.90%

avg. score	MCQG-C++	GPT-4
Context	0.96 (0.96, 0.95, 0.99)	0.95 (0.91, 0.96, 0.99)
Question	0.91 (0.89, 0.93)	0.87 (0.84, 0.9)
CQT ANS	0.79 (0.64, 0.94)	0.64 (0.45, 0.83)
Distractor	0.94 (0.94)	0.89 (0.89)
Reasoning	0.985 (0.99, 0.98)	0.97 (0.98, 0.96)
Total	0.917	0.864

Table 2: The table (top) presents the correlation between the LLM-as-Judge and Expert-X across different aspect filtering methods. The table (middle)* refers to the win rate of questions generated by MCQG-C++ and GPT-4 using the best LLM-as-Judge as the metric from RQ2-3. The table (below)* also shows the rating results for the 10 different aspects: Context (concision, relevance, misdirection), Question (concluding, clarity), Correct Answer (occurrence, depth of understanding), Distractor (common mistakes), and Reasoning (logical flow, evidence-based reasoning).

GPT-4 produce results highly correlated with human evaluations in tasks like summarization and story generation (Chiang and Lee 2023; Kocmi and Federmann 2023). However, applying LLMs as judges in specialized fields, such as clinical NLP, remains challenging due to domain-specific knowledge requirements (Singhal et al. 2023). Our work is among the first to investigate the use of `LLM-as-Judge` in the USMLE MCQs generation.

Conclusion

In conclusion, the MCQG-C++ framework improves LLM ability to generate high-quality USMLE-style MCQs by integrating expert-driven prompts with iterative self-refinement. The framework significantly enhances question quality and difficulty while aligning closely with expert evaluations. Additionally, our `LLM-as-Judge` metric offers a scalable alternative to costly expert assessments.

Limitations and Ethical Considerations

Our research focuses on generating medical exam questions, with promising results, though its adaptability to other domains is yet to be explored. We evaluated the method with three medical experts, but future involvement of specialized USMLE exam experts could enhance credibility. Future research should address privacy, fairness, generalizability to other languages/domains, and potential biases. While our work marks an important step in medical exam question generation, broader societal impacts require careful considera-

tion. (Semi)-Automated USMLE-MCQG systems can boost the efficiency and accuracy of medical education, but over-reliance could reduce direct student-educator interactions, and LLMs' hallucinations might impact students' learning and real-world practice. Caution is necessary to ensure these technologies support, not replace, medical educators. Future research should focus on these challenges to maximize positive contributions to medical education and society.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agarwal, M.; Sharma, P.; and Goswami, A. 2023. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus*, 15(6).

Artsi, Y.; Sorin, V.; Konen, E.; Glicksberg, B. S.; Nadkarni, G.; and Klang, E. 2024. Large language models for generating medical examinations: systematic review. *BMC Medical Education*, 24(1): 354.

Benítez, T. M.; Xu, Y.; Boudreau, J. D.; Kow, A. W. C.; Bello, F.; Van Phuoc, L.; Wang, X.; Sun, X.; Leung, G. K.-K.; Lan, Y.; et al. 2024. Harnessing the potential of large language models in medical education: promise and pitfalls. *Journal of the American Medical Informatics Association*, 31(3): 776–783.

Biswas, S. 2023. Passing is great: Can ChatGPT conduct USMLE exams? *Annals of Biomedical Engineering*, 51(9): 1885–1886.

Cai, P.; Yao, Z.; Liu, F.; Wang, D.; Reilly, M.; Zhou, H.; Li, L.; Cao, Y.; Kapoor, A.; Bajracharya, A.; et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11: 1518–1536.

Ch, D. R.; and Saha, S. K. 2018. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1): 14–25.

Chen, H.; Vo, D. M.; Takamura, H.; Miyao, Y.; and Nakayama, H. 2023. StoryER: Automatic story evaluation via ranking, rating and reasoning. *Journal of Natural Language Processing*, 30(1): 243–249.

Chen, Y.; Wu, L.; and Zaki, M. J. 2023. Toward subgraph-guided knowledge graph question generation with graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Cheung, B. H. H.; Lau, G. K. K.; Wong, G. T. C.; Lee, E. Y. P.; Kulkarni, D.; Seow, C. S.; Wong, R.; and Co, M. T.-H. 2023. ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom). *PLoS One*, 18(8): e0290691.

Chiang, C.-H.; and Lee, H.-y. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Das, T.; Albassam, D.; and Sun, J. 2024. Synthetic Patient-Physician Dialogue Generation from Clinical Notes Using LLM. *arXiv preprint arXiv:2408.06285*.

Gierl, M. J.; Lai, H.; and Turner, S. R. 2012. Using automatic item generation to create multiple-choice test items. *Medical education*, 46(8): 757–765.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120.

Guo, S.; Liao, L.; Li, C.; and Chua, T.-S. 2024. A Survey on Neural Question Generation: Methods, Applications, and Prospects. *arXiv preprint arXiv:2402.18267*.

Guo, S.; Zhang, J.; Wang, Y.; Zhang, Q.; Li, C.; and Chen, H. 2022. Dsm: Question generation over knowledge base via modeling diverse subgraphs with meta-learner. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4194–4207.

Homolak, J. 2023. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. *Croatian Medical Journal*, 64(1): 1.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Kim, S.; Shin, J.; Cho, Y.; Jang, J.; Longpre, S.; Lee, H.; Yun, S.; Shin, S.; Kim, S.; Thorne, J.; et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Kim, S.; Suk, J.; Longpre, S.; Lin, B. Y.; Shin, J.; Welleck, S.; Neubig, G.; Lee, M.; Lee, K.; and Seo, M. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. *arXiv preprint arXiv:2405.01535*.

Kirk, R.; Mediratta, I.; Nalmpantis, C.; Luketina, J.; Hambro, E.; Grefenstette, E.; and Raileanu, R. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Klang, E.; Portugez, S.; Gross, R.; Brenner, A.; Gilboa, M.; Ortal, T.; Ron, S.; Robinzon, V.; Meiri, H.; Segal, G.; et al. 2023. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Medical Education*, 23.

Kocmi, T.; and Federmann, C. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Lee, H.; Phatale, S.; Mansoor, H.; Lu, K.; Mesnard, T.; Bishop, C.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Ma, Y. J.; Liang, W.; Wang, G.; Huang, D.-A.; Bastani, O.; Jayaraman, D.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang,

Y.; et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Mousavi, S. M.; Alghisi, S.; and Riccardi, G. 2024. Is Your LLM Outdated? Benchmarking LLMs & Alignment Algorithms for Time-Sensitive Knowledge. *arXiv preprint arXiv:2404.08700*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pan, L.; Chen, W.; Xiong, W.; Kan, M.-Y.; and Wang, W. Y. 2021. Zero-shot fact verification by claim generation. *arXiv preprint arXiv:2105.14682*.

Pan, L.; Saxon, M.; Xu, W.; Nathani, D.; Wang, X.; and Wang, W. Y. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Philip A Bucur, S. R. D., Vikrant Bhatnagar. 2019. A "U-shaped" Curve: Appreciating How Primary Care Residency Intention Relates to the Cost of Board Preparation and Examination. *cureus*.

Sahoo, P.; Singh, A. K.; Saha, S.; Jain, V.; Mondal, S.; and Chadha, A. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint arXiv:2402.07927*.

Santhanam, K.; Khattab, O.; Saad-Falcon, J.; Potts, C.; and Zaharia, M. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

Scheurer, J.; Campos, J. A.; Chan, J. S.; Chen, A.; Cho, K.; and Perez, E. 2022. Training language models with language feedback. *arXiv preprint arXiv:2204.14146*.

Scoles, P. V. 2008. Comprehensive review of the USMLE. *Advances in Physiology Education*, 32(2): 109–110.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Su, D.; Xu, P.; and Fung, P. 2022. QA4QG: using question answering to constrain multi-hop question generation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8232–8236. IEEE.

Sun, Y.; Tang, D.; Duan, N.; Qin, T.; Liu, S.; Yan, Z.; Zhou, M.; Lv, Y.; Yin, W.; Feng, X.; et al. 2019. Joint learning of question answering and question generation. *IEEE Transactions on Knowledge and Data Engineering*, 32(5): 971–982.

Wang, J.; Yao, Z.; Yang, Z.; Zhou, H.; Li, R.; Wang, X.; Xu, Y.; and Yu, H. 2024. NoteChat: A Dataset of Synthetic Patient-Physician Conversations Conditioned on Clinical Notes. In *Findings of the Association for Computational Linguistics ACL 2024*, 15183–15201.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Wang, Y.; Ivison, H.; Dasigi, P.; Hessel, J.; Khot, T.; Chandu, K.; Wadden, D.; MacMillan, K.; Smith, N. A.; Beltagy, I.; et al. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36: 74764–74786.

Xie, Y.; Pan, L.; Wang, D.; Kan, M.-Y.; and Feng, Y. 2020. Exploring question-specific rewards for generating deep questions. *arXiv preprint arXiv:2011.01102*.

Yao, Z.; Jaafar, A.; Wang, B.; Yang, Z.; and Yu, H. 2024. Do Clinicians Know How to Prompt? The Need for Automatic Prompt Optimization Help in Clinical Note Generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 182–201.

Zeng, Z.; Yu, J.; Gao, T.; Meng, Y.; Goyal, T.; and Chen, D. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Zhang, C.; D'Haro, L. F.; Chen, Y.; Zhang, M.; and Li, H. 2024. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19515–19524.

Zhang, R.; Guo, J.; Chen, L.; Fan, Y.; and Cheng, X. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1): 1–43.

Zhang, X.; and Gao, W. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhao, Z.; Hou, Y.; Wang, D.; Yu, M.; Liu, C.; and Ma, X. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. *arXiv preprint arXiv:2203.14187*.

Zhao, Z.; Jin, Q.; Chen, F.; Peng, T.; and Yu, S. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10(1).

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; and Li, D. 2023. Eric. *P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena*, 2(6): 7.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)

Does this paper make theoretical contributions? (no)
Does this paper rely on one or more datasets? (yes)

- A motivation is given for why the experiments are conducted on the selected datasets (yes)
- All novel datasets introduced in this paper are included in a data appendix. (yes)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (NA)

Does this paper include computational experiments? (yes)

- Any code required for pre-processing data is included in the appendix. (yes).
- All source code required for conducting and analyzing the experiments is included in a code appendix. (yes)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
- This paper states the number of algorithm runs used to compute each reported result. (yes)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (NA)

Appendix

Topics and Test points

Topics t refer to a list of target topics selected from 41 potential topics outlined in the NBME official guidelines, categorized into the following 10 sections:

1. Diagnosis - Causes and Mechanisms
2. Diagnosis - Obtaining and Predicting History and Physical Examination
3. Diagnosis - Selecting and Interpreting Laboratory and Diagnostic Studies
4. Diagnosis - Formulating the Diagnosis
5. Diagnosis - Determining Prognosis/Outcome
6. Management - Health Maintenance and Disease Prevention
7. Management - Selecting and Monitoring Pharmacotherapy
8. Management - Clinical Interventions/Treatments
9. Management - Selecting Clinical Interventions (Mixed Management)
10. Management - Monitoring/Surveillance for Disease Recurrence or Progression

Test points k refer to the core concepts closely related to the correct answer. We employ the ColBERT retriever [15], denoted as π_{rtr}, to retrieve suitable test points from 18 sections.

1. General Principles of Foundational Science
2. Immune System
3. Blood & Lymphoreticular System
4. Behavioral Health
5. Nervous System & Special Senses
6. Skin & Subcutaneous Tissue
7. Musculoskeletal System
8. Cardiovascular System
9. Respiratory System
10. Gastrointestinal System
11. Renal & Urinary System
12. Pregnancy, Childbirth, & the Puerperium
13. Female Reproductive System & Breast
14. Male Reproductive System
15. Endocrine System
16. Multisystem Processes & Disorders
17. Biostatistics, Epidemiology/Population Health, & Interpretation of the Medical Literature
18. Social Sciences

Stopping Criteria The stopping criteria for the MCQG-C++ included two checks. One was that the iteration would stop if the total critique score of all the components exceeded 90% of the maximum possible score, and the other was if the number of iterations exceeded 4.

---

[15]https://github.com/stanford-futuredata/ColBERT

Task Input and Response R_a, R_b
I: Generate a USMLE question based on the following **medical case**: A 84-year-old male presents with chest pain .... **Topic**: pathophysiology of sepsis **Test point**: sepsis R_a: system-a generated USMLE question R_b: system-b generated USMLE question
Critique $F_{crq}(I, R_b)$
The generated question fails to meet the quality and difficulty requirements for the USMLE because ...... The following suggestions could be considered: .....   **Score for Question** R_b: 1/10
Correction $F_{crr}(I, R_b, F_{crq})$ **with** F_{crq}
Here is the revision for question B: **Context**: The 84-year-old female patient presented to the emergency room with ......   **Question**: What is the most likely complication that has developed in this patient's case? **Correct answer**: Septic shock **Distractor options**: a) b) c) d)
Comparison $F_{cmp}(I, R_a, R_b)$
Question A has higher quality than B as it ...... On the contrary, question B ......   **Preference Response: A**

Table 3: Examples of different LLM feedback (Critique, Correction, Comparison) for USMLE MCQG.

Dataset

For the medical cases utilized in the generation of USMLE questions, we used patient summaries from the PMC-Patients dataset (Zhao et al. 2023). The average length of these patient summaries was ~419 words. The frequency of topics used is listed in Table 11 and 12.

Ablation study and output analysis of MCQG-C++

Round-wise metrics As shown in Table 4, the round-wise metrics for human data reveal nuanced trends in the model's performance through multiple feedback iterations. The total score exhibited a modest increase from Round 1 (0.9031) to Round 3 (0.9062), followed by a slight decline in Round 4 (0.9006), suggesting that while the self-feedback process contributes positively, its benefits may diminish with excessive iterations. The context score notably improved from 0.9683 in Round 1 to 0.9712 in Round 3, reflecting enhanced model comprehension of the context through feedback. However, performance in other areas fluctuated: the question score showed a slight decline across rounds, and while the correct answer score improved from Round 1 to Round 3, it decreased slightly in Round 4. Conversely, the distractor option score showed steady improvement, culminating in the highest score by Round 4. The reasoning score, however, demonstrated a gradual decline over the rounds. The standard deviation for most metrics either decreased or remained stable, indicating more consistent performance. Overall, while certain components of the model benefited

from the feedback process, others did not, highlighting the complexity of balancing improvements across different aspects of question generation.

As shown in Table 5, the round-wise metrics for machine data demonstrate a stable overall performance, with the total score showing minimal fluctuation across rounds, peaking slightly in Round 2 (0.907) and ending at 0.903 in Round 4. This suggests that while the feedback process maintains performance, it does not lead to significant improvements. The context score exhibits a notable increase from 0.9688 in Round 1 to 0.974 in Round 4, coupled with a decrease in standard deviation, indicating enhanced and more consistent context understanding. The C. Answer score also shows a gradual upward trend, improving from 0.795 in Round 1 to 0.810 in Round 4, reflecting slow but steady progress in answer generation. The distractor option score remains relatively high and stable across rounds, while the reasoning score experiences a decline from 0.967 in Round 1 to 0.948 in Round 3, with a slight recovery to 0.951 in Round 4. Variability in performance is evident, with some metrics, such as the context score, showing decreased standard deviation, indicating more consistency, while others, like the reasoning score, exhibit increased variability, highlighting differential effects on the consistency of various components.

Based on the analysis of the round-wise metrics from both human and machine data, the self-feedback process demonstrates the most significant improvements in context understanding and answer generation. These enhancements are evidenced by the consistent upward trends in context and correct answer scores across rounds. However, the data also suggests a potential trade-off between reasoning ability and other metrics, as seen in the decline of reasoning scores, particularly after the initial rounds. The plateau or slight decline in overall performance after 2-3 rounds indicates that the benefits of the feedback process diminish with excessive iterations, implying that a limited number of rounds may be optimal for maximizing improvements without compromising other aspects of question generation. These findings highlight the importance of balancing the feedback process to achieve comprehensive improvements across all key metrics.

In addition, the analysis of the round-wise metrics in Table 6 and 7 reveals several key trends in the performance of the question generation pipeline. As the rounds progress, the context length consistently decreases, indicating that the pipeline effectively refines the context by excluding extraneous information, leading to more precise and focused questions. The accuracy of the QA component improves over the rounds, suggesting that the iterative process enhances the overall quality of the questions, making them more answerable by the LLM. However, this improvement reaches a point of diminishing returns in the later rounds, implying that a limited number of iterations may be optimal. On the other hand, the equality between the correct answer and the keypoint deteriorates over the rounds, indicating that the pipeline makes the correct answer more subtly related to the keypoint rather than directly copying it. This shift suggests that while the pipeline reduces redundancy, it may also introduce complexity that could impact the clarity and directness
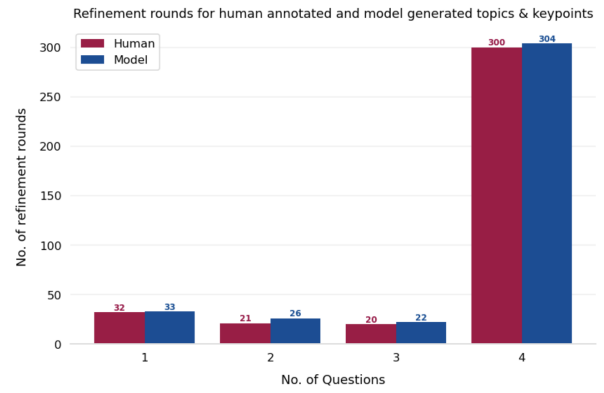

Figure 3: No. of refinement rounds. This histogram indicates that most of the data points(78.9% for human, 80.4% for machine) take the full 4 iterations in pursuit of the threshold score.
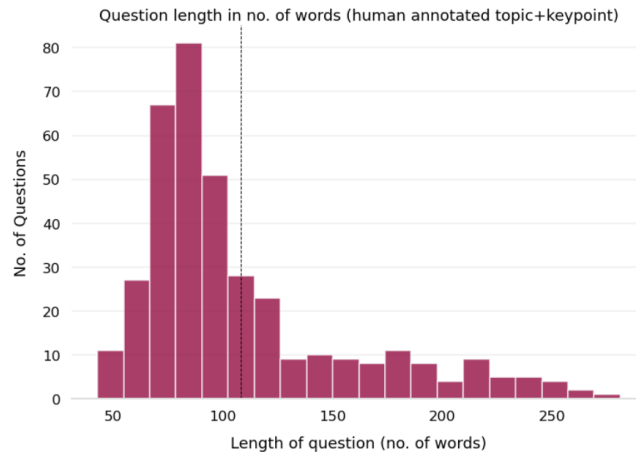

Figure 4: Length of the generated question human.

of the correct answer-keypoint relationship.

Other Basic statistics We also calculate Refinement rounds for human-annotated and model-generated topics & key points in Figure 3; Question length in no. of words (human annotated topic+keypoint) in Figure 4; Question length in no. of words (machine annotated topic+keypoint) in Figure 5; Best scoring rounds for human-generated topic+keypoint in Figure 6; Best scoring rounds for machine-generated topic+keypoint in Figure 7.

Correlation analysis Our analysis in Figure 8 9 and Table 8 reveals a strong inverse relationship between the length of clinical notes/questions and the number of rounds required to achieve the best score in both human-generated and machine-generated data. The perfect negative Spearman and Kendall correlations (-1.0) indicate that as the length of inputs increases, fewer rounds are consistently needed to reach optimal performance. Human-generated content shows slightly stronger linear correlations compared to machine-generated data. These findings suggest

Table 4: Round-wise metrics - Human data

Mean — Std. dev	Round 1	Round 2	Round 3	Round 4
Context score	0.9683 — 0.068	0.96705 — 0.060	**0.9712 — 0.055**	0.9679 — 0.062
Question score	**0.8394** — 0.098	0.8322 — 0.09	0.8277 — **0.089**	0.8195 — 0.091
C. Answer score	0.8051 — 0.168	0.8212 — 0.167	**0.8358 — 0.161**	0.8284 — 0.172
Distractor option score	0.9378 — 0.096	0.9344 — 0.089	0.9376 — 0.086	**0.9414 — 0.084**
Reasoning score	**0.9650 — 0.116**	0.96168 — 0.122	0.9589 — 0.132	0.9456 — 0.150
Total score	0.9031 — 0.061	0.9033 — **0.052**	**0.9062** — 0.0569	0.9006 — 0.058

Table 5: Round-wise metrics - Machine data

Mean — Std. dev	Round 1	Round 2	Round 3	Round 4
Context score	0.9688 — 0.0686	**0.975 — 0.053**	0.970 — 0.0647	0.974 — **0.0511**
Question score	0.838 — **0.090**	**0.848 — 0.092**	0.836 — 0.0941	0.837 — 0.0939
C. Answer score	0.795 — 0.150	0.809 — **0.158**	0.799 — 0.151	**0.810 — 0.154**
Distractor option score	0.943 — 0.080	0.940 — 0.088	0.941 — 0.081	**0.944 — 0.0825**
Reasoning score	**0.967 — 0.1098**	0.965 — 0.121	0.948 — 0.147	0.951 — 0.149
Total score	0.902 — **0.051**	**0.907 — 0.0564**	0.899 — 0.058	0.903 — 0.0561

Table 6: Round-wise metrics - Human data

Mean — Std. dev	Round 1	Round 2	Round 3	Round 4
Question length	203.008 — 46.15	111.09 — 24.14	91.47 — 19.68	**84.11 — 17.80**
Accuracy (QA)	0.898 — 0.302	0.89 — 0.311	**0.903 — 0.295**	0.86 — 0.339
C. Answer = Keypoint	0.193 — 0.394	0.170 — 0.375	0.153 — 0.360	**0.146 — 0.353**

Table 7: Round-wise metrics - Machine data

Mean — Std. dev	Round 1	Round 2	Round 3	Round 4
Question length	204.68 — 49.04	117.85 — 25.33	96.113 — 19.52	**87.5 — 17.21**
Accuracy (QA)	0.89 — 0.30	**0.908 — 0.288**	0.865 — 0.341	0.875 — 0.33
C. Answer = Keypoint	0.083 — 0.276	0.080 — 0.272	0.073 — 0.261	**0.065 — 0.247**

Data Type	Variable	Pearson	Spearman	Kendall
Human	Rounds with Clinical Note Length	-0.9103	-1.0	-1.0
Machine	Rounds with Clinical Note Length	-0.8869	-1.0	-1.0
Human	Rounds with Question Length	-0.9103	-1.0	-1.0
Machine	Rounds with Question Length	-0.8869	-1.0	-1.0

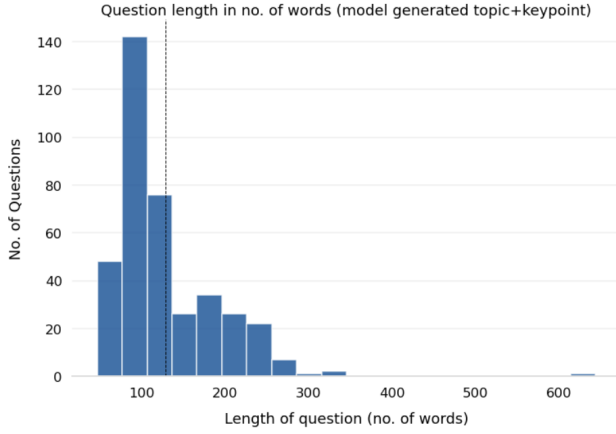Table 8: Correlation between different rounds with Clinical Note Length and Question Length for both Human and Machine data.

Figure 5: Length of the generated question machine.
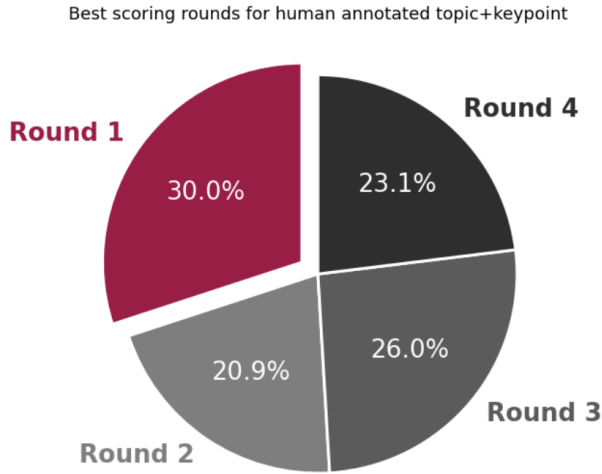
Figure 7: Best scoring rounds machine.

Figure 6: Best scoring rounds human.
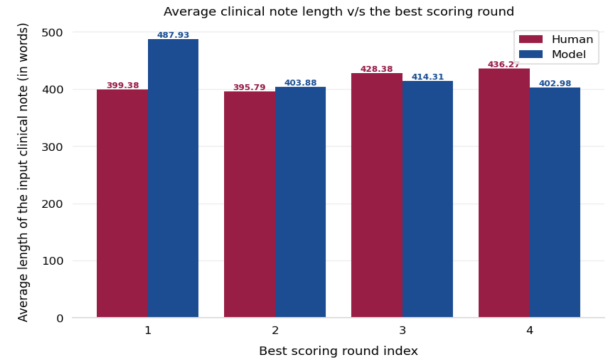
Figure 8: Clinical note length v/s best scoring round.

that longer, more detailed inputs provide richer information, allowing the model to converge on optimal performance more quickly. Practically, this implies that shorter inputs may benefit from more iterative rounds, while longer inputs may require fewer rounds to achieve the best results. The consistency of this pattern across both clinical note length and question length underscores the importance of input complexity in determining the efficiency of the iterative improvement process.

Greedy Algorithm

The Greedy algorithm iteratively constructs an optimal set of aspects by sequentially adding the most correlated aspects. Let $A = \{a_1, a_2, ..., a_n\}$ be the set of all aspects, and h be the human preference data. The algorithm proceeds as follows:
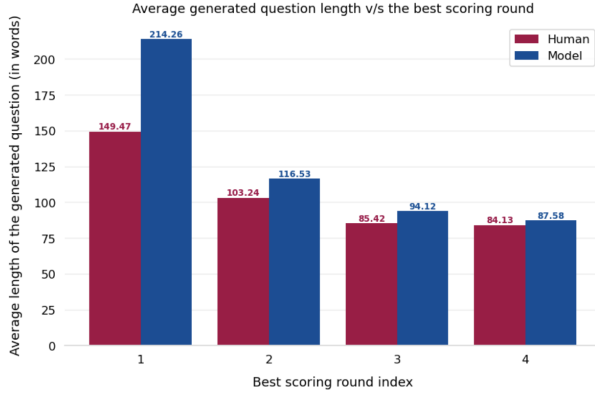
Figure 9: Question length v/s best scoring round.

Model	No POS Bias	PA.	τ
GPT-4	65%	61.53%	0.2121
GPT-4o	42.5%	58.82%	0.3340
GPT-3.5-turbo	35%	64.29%	0.0572
Claude-3-haiku	42.5%	41.18%	-0.0636
Claude-3.5-sonnet	37.5%	46.67%	0.1517
Claude-3-opus	47.5%	68.42%	0.3152

Table 9: Percentage Agreement (PA.) and Kendall's Tau (τ) for Different Models Compared to Expert X using valid data (e.g., order matching for both GPT-4 generated output first setting and MCQG-C++ generated output first setting).

Algorithm 1 Greedy Aspect Selection

1: Sort A in descending order of correlation with h
2: $S \leftarrow \emptyset$ ▷ Initialize empty set of selected aspects
3: $c_{best} \leftarrow 0$ ▷ Best correlation so far
4: **for** each $a_i \in A$ **do**
5: $\quad S' \leftarrow S \cup \{a_i\}$
6: $\quad r \leftarrow \text{CalculateRating}(S')$
7: $\quad c \leftarrow \text{Correlation}(r, h)$
8: $\quad$ **if** $c > c_{best}$ **then**
9: $\quad\quad S \leftarrow S'$
10: $\quad\quad c_{best} \leftarrow c$
11: $\quad$ **end if**
12: **end for**
13: **return** S, c_{best}

---

where CalculateRating(S) computes the final rating score using the aspects in set S, and Correlation(r, h) calculates either the percentage agreement or Cohen's kappa between the rating r and human preferences h.

All-Combo Algorithm

The All-Combo algorithm exhaustively evaluates all possible combinations of the top n aspects to find the optimal subset. Let k be the number of aspects in a combination, where $1 \le k \le n$.

Algorithm 2 All-Combo Aspect Selection

1: $C_{best}^{PA} \leftarrow \emptyset, c_{best}^{PA} \leftarrow 0$ ▷ Best for Percentage Agreement
2: $C_{best}^{\kappa} \leftarrow \emptyset, c_{best}^{\kappa} \leftarrow 0$ ▷ Best for Cohen's Kappa
3: **for** $k = 1$ to n **do**
4: $\quad$ **for** each combination $C \in \binom{A}{k}$ **do**
5: $\quad\quad r \leftarrow \text{CalculateRating}(C)$
6: $\quad\quad c_{PA} \leftarrow \text{PercentageAgreement}(r, h)$
7: $\quad\quad c_{\kappa} \leftarrow \text{CohenKappa}(r, h)$
8: $\quad\quad$ **if** $c_{PA} > c_{best}^{PA}$ **then**
9: $\quad\quad\quad C_{best}^{PA} \leftarrow C, c_{best}^{PA} \leftarrow c_{PA}$
10: $\quad\quad$ **end if**
11: $\quad\quad$ **if** $c_{\kappa} > c_{best}^{\kappa}$ **then**
12: $\quad\quad\quad C_{best}^{\kappa} \leftarrow C, c_{best}^{\kappa} \leftarrow c_{\kappa}$
13: $\quad\quad$ **end if**
14: $\quad$ **end for**
15: **end for**
16: **return** $C_{best}^{PA}, c_{best}^{PA}, C_{best}^{\kappa}, c_{best}^{\kappa}$

This algorithm returns the best aspect combinations and their corresponding correlation scores for both percentage agreement and Cohen's kappa, as these may differ depending on the evaluation metric used.

Patient History

An 84-year-old female with a past medical history of hypertension presented with weakness, dry cough, and shortness of breath for four days. The patient had received two doses of the COVID vaccine, with the second dose in March 2021. In the ER, her vital signs were BP 133/93, HR 103 bpm, RR 22 breaths/min, oxygen saturation of 96% on 40 L per minute of supplemental oxygen via high-flow nasal cannula, and afebrile. Laboratory assessment is in Table. Nasopharyngeal swab for SARS-CoV-2 RNA was positive. Chest X-ray on admission shows worsening right pleural effusion with new opacity obscuring the lower two-third of the right lung and a new pleural-based opacity in the left upper lobe Figure. CT chest with contrast shows large right pleural effusion and associated right basilar consolidation and abdominal ascites. The patient was admitted to the telemetry unit and started on methylprednisolone, piperacillin-tazobactam, remdesivir, and baricitinib. The patient clinically deteriorated on Day 2 and was transferred to the intensive care unit for thoracentesis and possible intubation. Thoracentesis removed 1.95 L of bloody, serosanguineous fluid obtained, with partial resolution of the effusion Figure. On Day 3, the patient developed septic shock, florid renal failure, and lethargy and was started on intravenous fluids and norepinephrine drip. Chest X-ray showed near-complete opacification of bilateral lung fields and subsequently was intubated. On Day 4, tense ascites were noted and the patient underwent paracentesis, which removed 4.25 L of bloody, serosanguineous fluid. Renal replacement therapy started. The patient was deemed to have a guarded prognosis with multiorgan failure.

Diagnosis

Causes and Mechanisms

- ☐ the cause/infectious agent or predisposing factor(s)
- ☒ underlying processes/pathways (pathophysiology of sepsis)
- ☐ underlying anatomic structure or physical location
- ☒ mechanisms, drugs (tazobactam, beta-lactamase inhibitors)

Obtaining and Predicting History and Physical Examination

- ☐ knows signs/symptoms of selected disorders
- ☐ knows individual's risk factors for development of condition
- ☐ knows what to ask to obtain pertinent additional history
- ☒ predicts the most likely additional physical finding (pleural effusion)

Selecting and Interpreting Laboratory and Diagnostic Studies

- ☐ select most appropriate laboratory or diagnostic study
- ☒ interprets laboratory or other study findings (pleural fluid analysis)
- ☒ predicts the most likely laboratory or diagnostic study result (pleural fluid protein/serum protein ratio > 0.5)
- ☐ most appropriate laboratory or diagnostic study after change in patient status

Formulating the Diagnosis

- ☒ select most likely diagnosis (correct answer: pleural effusion after pneumonia, sepsis, MODS)

Determining Prognosis/Outcome

- ☐ recognizes factors in the history, or physical or laboratory study findings
- ☐ interprets laboratory or other diagnostic study results and identifies current/future status of patient
- ☒ recognizes associated conditions of a disease (septic shock, MODS)
- ☐ recognizes characteristics of disease relating to natural history or course of disease

Management

Health Maintenance and Disease Prevention

- ☐ risk factors for conditions amenable to prevention or detection
- ☐ identifies patient groups at risk
- ☐ knows common screening tests
- ☐ selects appropriate preventive agent or technique
- ☐ knows appropriate counseling regarding current and future problems
- ☐ educates patients

Selecting and Monitoring Pharmacotherapy

- ☐ selects most appropriate pharmacotherapy
- ☐ assesses patient adherence, recognizes techniques to increase adherence
- ☐ recognizes factors that alter drug requirements
- ☐ Knows adverse effects of various drugs or recognizes signs and symptoms of drug (and drug-drug) interactions
- ☐ knows contraindications of various medications
- ☐ knows modifications of a therapeutic regimen within the context of continuing care
- ☐ appropriate monitoring to evaluate effectiveness of pharmacotherapy or adverse effects

Clinical Interventions/Treatments

- ☒ most appropriate management of selected conditions (sepsis)
- ☒ immediate management or priority in management (septic shock)
- ☐ follow-up or monitoring approach regarding the management plan
- ☒ current/short-term management (sepsis)
- ☐ severity of patient condition in terms of need for referral for surgical treatments/procedures
- ☐ appropriate surgical management
- ☐ preoperative/postoperative

Selecting Clinical Interventions (Mixed Management)

- ☐ Selecting Clinical Interventions (Mixed Management)

Monitoring/Surveillance for Disease Recurrence or Progression

- ☐ indications for surveillance for recurrence or progression of disease following treatment
- ☐ how to monitor a chronic disease in a stable patient where a change in patient status might indicate a need to change therapy
- ☐ most appropriate long-term treatment

Table 10: Human annotation example for topics and test points.

Topic Name	Frequency
Select most likely diagnosis	40
Selects most appropriate pharmacotherapy	37
Select most appropriate laboratory or diagnostic study	33
Most appropriate management of selected conditions	27
Mechanisms, drugs	21
Most appropriate long-term treatment	21
Predicts the most likely laboratory or diagnostic study result	20
Recognizes associated conditions of a disease	15
The cause/infectious agent or predisposing factor(s)	13
Underlying processes/pathways	11
Knows signs/symptoms of selected disorders	10
Current/short-term management	8
Immediate management or priority in management	7
The cause/infectious agent or predisposing factor	7
Knows adverse effects of various drugs or recognizes signs and symptoms of drug (and drug-drug) interactions	7
Appropriate surgical management	6
Interprets laboratory or other study findings	6
Underlying anatomic structure or physical location	5
Predicts the most likely additional physical finding	4
Interprets laboratory or other diagnostic study results and identifies current/future status of patient	4
Selecting Clinical Interventions (Mixed Management)	4
Follow-up or monitoring approach regarding the management plan	4
Identifies patient groups at risk	3
Selects appropriate preventive agent or technique	3
Knows common screening tests	3
Educates patients	2
Select most likely diagnosis yes	1
Selecting Clinical Interventions (Mixed Management)	1
Risk factors for conditions amenable to prevention or detection	1
Recognizes factors that alter drug requirements	1
Indications for surveillance for recurrence or progression of disease following treatment	1
Most appropriate long-term treatment	1
Severity of patient condition in terms of need for referral for surgical treatments/procedures	1
Knows what to ask to obtain pertinent additional history	1
Preoperative/postoperative	1
Knows contraindications of various medications	1
Recognizes characteristics of disease relating to natural history or course of disease	1
Appropriate monitoring to evaluate effectiveness of pharmacotherapy or adverse effects	1
Predicts the most likely laboratory or diagnostic study result	1

Table 11: Frequency of Human identified topics

Topic Name	Frequency
Select most appropriate laboratory or diagnostic study	66
The cause/infectious agent or predisposing factor(s)	64
Selects most appropriate pharmacotherapy	56
Appropriate surgical management	41
Select most likely diagnosis	37
Knows signs/symptoms of selected disorders	37
Most appropriate management of selected conditions	16
Most appropriate long-term treatment	10
Underlying processes/pathways	7
Mechanisms, drugs	4
Knows individual's risk factors for development of condition	4
Appropriate monitoring to evaluate effectiveness of pharmacotherapy or adverse effects	3
Educates patients	2
Interprets laboratory or other diagnostic study results and identifies current/future status of patient	2
Interpret laboratory or other study findings	1
Recognizes factors that alter drug requirements	1
Interprets laboratory or other study findings	1
Underlying anatomic structure or physical location	1
Knows modifications of a therapeutic regimen within the context of continuing care	1
Risk factors for conditions amenable to prevention or detection	1
Predicts the most likely additional physical finding	1
Immediate management or priority in management	1
Knows what to ask to obtain pertinent additional history	1
Knows adverse effects of various drugs and signs and symptoms of drug (and drug-drug) interactions	1

Table 12: Frequency of machine generated topics

Problem Statement

Given a medical case n detailing a patient's history, diagnosis, treatment, and outcome, we aim to generate a USMLE question u.

Here, $u :< c, q, a, d >$ consists of:

- **context** (c) which is a modified excerpt from n tailored to align with the target style and obscure evidence information that can easily lead directly to the correct answer, based on a topic t and a keypoint k;

- **question** (q) based on the generated context based on the generated context c, which may be one or several sentences;

- **correct answer** (a) to this question and,

- **distractor options** (d) to challenge the test-taker.

Our framework consists of a MCQG-C++ pipeline for generating the target u.

Input: Medical Cases

A 67-year-old Caucasian female presented to our hospital with a chief complaint of persistent bright red blood per rectum. Her medical history was significant for hypertension, hyperlipidemia, diabetes mellitus type 2, coronary artery disease with three prior myocardial infarctions, recurrent cerebrovascular accidents requiring anticoagulation with warfarin, gastroesophageal reflux disease, asthma, and endometrial cancer status post radiation therapy. Fifteen months prior to the current presentation, the patient was noted to have a grade 1 endometrial adenoma but was not considered a good surgical candidate due to multiple comorbidities. Vaginal hysterectomy was considered but due to her long and narrow vagina, this option was deferred initially. Her only treatment option was radiation therapy and brachytherapy. She eventually underwent total abdominal hysterectomy with bilateral salpingo-oopherectomy due to continued pelvic pain. The patient denied any prior history of gastrointestinal (GI) bleeding. Her bleeding was described as one large episode of bright red blood per rectum associated with blood clots. She denied any abdominal pain, nausea, vomiting, diarrhea, constipation, or melena. The most recent colonoscopy was performed four months ago and revealed three diminutive polyps in the transverse colon with pathology confirming tubular adenoma. Her physical examination was significant for mild left-sided abdominal tenderness but was otherwise unremarkable. Rectal examination was notable for nonbleeding hemorrhoids and no visible blood. Blood work revealed white blood cells of 14.3k/uL (normal range 4.3-10.0 k/uL) and hemoglobin of 9.6 g/dL (normal range 11.8-14.8 g/dL), which is similar to the patient's baseline. Creatinine was slightly elevated to 1.2 and blood urea nitrogen was elevated to 39. International normalized ratio was 2.0. Due the large volume of hematochezia and presence of anemia, the patient was admitted to the hospital and underwent a colonoscopy, which revealed a large, fungating, friable, and ulcerated nonobstructing mass in the sigmoid colon. The mass was noncircumferential, measured 4 cm in length, and was located 15-19 cm from the anal verge. Biopsies were obtained with cold forceps for histology and the proximal and distal margins of the mass were tattooed. Histology showed invasive and moderately differentiated carcinoma without visible goblet cells. Given the patient's history of endometrial cancer, immunohistochemistry was performed and was consistent with an endometrial (endometrioid subtype) primary. Pan-computed tomography was negative for any obvious distant disease. The tumor marker carcinoembryonic antigen level was normal as well. The patient subsequently underwent surgical resection of the affected colon followed by radiation therapy. At a follow-up visit one year later, the patient was doing well without active GI complaints.

Input: Topics

The topics come from an official list of 16 target topics from USMLE.

We use both human annotated and model identified topics for generating USMLE questions.

Some example topics:

- Causes and Mechanisms
    - the cause/infectious agent or predisposing factor(s)
    - underlying processes/pathways
    - underlying anatomic structure or physical location
    - mechanisms, drugs
- Selecting and Interpreting Laboratory and Diagnostic Studies
    - select most appropriate laboratory or diagnostic study
    - interprets laboratory or other study findings
- Monitoring/Surveillance for Disease Recurrence or Progression
    - most appropriate long-term treatment **(our topic for the current clinical note)**

Input: Key Points

The key points are the core concepts to which the target correct answer is highly related or is the same as the keypoint.

We use both human-annotated and model-identified key points to generate USMLE questions.

Example keypoints:
- covid-19 with copd
- Radiotherapy or chemotherapy **(our keypoint for the current clinical note)**
- ARDS
- osteonecrosis of the femoral head
- echocardiography for Ebstein's anomaly
- COVID-19 pneumonia

MCQG-C++ INIT Step

Initiation Step:
- Four sub-steps generate different components of the target output u.
- Examples for few-shot prompting are retrieved from the COLBERT retriever model R.
- These examples consist of USMLE questions from a question bank.

Context-Question Pair Retrieval:
- Use a query to retrieve the most relevant context-question pairs from R.
- Prompt model M with these pairs, along with other relevant information, to generate the context c.

Question Retrieval:
- Use the context to retrieve the next most relevant question from R.
- Prompt model M with this question, along with other information, to generate the question q.

Iteration:
- Repeat the process until generating the complete target output u, including context, question, correct answer, and distractor options.

MCQG-C++ INIT Output

Context: A 67-year-old Caucasian female with a significant medical history including hypertension, hyperlipidemia, diabetes mellitus type 2, coronary artery disease with three prior myocardial infarctions, recurrent cerebrovascular accidents, gastroesophageal reflux disease, asthma, and endometrial cancer presented with persistent bright red blood per rectum. She underwent radiation therapy and brachytherapy for her endometrial cancer but eventually required total abdominal hysterectomy with bilateral salpingo-oopherectomy due to persistent pelvic pain. She had no prior history of gastrointestinal bleeding. Her recent colonoscopy revealed three diminutive polyps in the transverse colon confirmed as tubular adenoma. Upon admission for an episode of hematochezia and anemia, she underwent a colonoscopy revealing a large, fungating, nonobstructing mass in the sigmoid colon. The biopsy was consistent with invasive and moderately differentiated carcinoma, and immunohistochemistry indicated an endometrial primary. The patient underwent surgical resection of the colon mass and subsequent radiation therapy. A follow-up visit a year later revealed no active GI complaints.

Question: What is the most appropriate long-term treatment for a 67-year-old female patient with a history of endometrial cancer who presented with a large, nonobstructing mass in the sigmoid colon due to invasive and moderately differentiated carcinoma?

Correct answer: Radiotherapy or chemotherapy

Distractor options:
- A: Oral corticosteroids
- B: Aspirin therapy
- C: Bisphosphonates
- D: Hormone replacement therapy
- E: Anticoagulation therapy

MCQG-C++ QUESTION ANSWER Step

- In this step u is presented as a USMLE question for M, i.e. u without the knowledge of the correct answer a, the options presented are $d \cup a$.
- M generates an attempted answer aa and a rationale which supports aa in the form of a reasoning r behind it.

MCQG-C++ QUESTION ANSWER Output

Attempted answer: Radiotherapy or chemotherapy

Reasoning: The patient's history of endometrial cancer and the finding of a large, nonobstructing mass in the sigmoid colon are significant. The biopsy results revealing invasive and moderately differentiated carcinoma with immunohistochemistry indicating an endometrial primary suggest that this mass is likely a metastasis from her endometrial cancer. In such a case, systemic therapy with chemotherapy or targeted radiation therapy is often the most appropriate treatment to control the disease and prevent further metastases.

- Option A (Oral corticosteroids) is not typically used as a primary treatment for cancer. It may be used in certain situations to manage side effects of cancer or cancer treatment, but it wouldn't be the mainstay of treatment in this case.
- Option B (Aspirin therapy) is sometimes used as a preventive measure against colorectal cancer, but it is not a treatment for existing cancer.
- Option C (Bisphosphonates) are used to prevent bone loss and fractures in patients with certain types of cancer, such as breast and prostate cancer, that have spread to the bones. They are not typically used as a primary treatment for colorectal or endometrial cancer.
- Option D (Hormone replacement therapy) is typically used to manage symptoms of menopause, but it's not typically used to treat endometrial cancer, particularly in a patient with a history of this disease.   - Option E (Anticoagulation therapy) is used to treat or prevent blood clots, but it does not treat cancer itself. Hence, based on the above discussion, the most appropriate long-term treatment for this patient would be radiotherapy or chemotherapy

MCQG-C++ CRITIQUE Step
- After generating all the components in the initialization and answer steps, the model M is tasked with critiquing each component in this feedback step.
- The set of components to be critiqued, $u_{critique}$, which includes the context (c), question (q), correct answer (a), distractor options (d), and rationale (r).
- Model M receives a scoring guide (G) containing all the aspects on which each component needs to be scored.
- The prompt also includes few-shot examples of already scored components and relevant information.
- The final output of this step is the feedback for all components, denoted as f. It includes a short textual critique and a score for each aspect of the respective components.

Rubric for the Feedback Output
Rubric for the Context:
  - `Relevant`: The context should be relevant to the topic provided.
  - `Concision`: The context should be concise and shouldn't include extraneous information or just be a copy of the clinical note.
  - `Coherent`: The context should be coherent enough and should organically be built for the question in the end.
  - `Consistent`: The context is consistent with the information in the clinical note and the topic.
  - `Specific`: The context should be specific and address the topic.
  - `Fluent`: The context is fluent in terms of grammar and flow of words and ideas.
  - `Clueing`: Instead of directly mentioning the diagnosis of a medical condition, it should be clued in through the mention of symptoms.
  - `Completeness`: The context should be complete and free from gaps or missing information that could lead to ambiguity in answering the question accurately.
  - `Misdirection`: Does the context avoid misleading the test taker intentionally or unintentionally?
Rubric for the Question:
  - `Relevant`: The question should be answerable from the information provided in the context and should not be abrupt.
  - `Clear`: The question should not be vague or ambiguous.
  - `Concluding`: The flow of ideas from the context should organically result into the question.
  - `Difficulty`: The question should not be too easy.
  - `Clarity`: Is the question unambiguous to avoid incorrect interpretations caused by ambiguity or poor wording?
Rubric for the Correct Answer:
  - `Relevant`: The correct answer should be either the keypoint or very related to it.
  - `Occurrence`: The correct answer or any of its semantic or syntactic forms and directly related medical concepts should not occur in the context.
  - `Justification`: Is the correct answer logically supported by the context and aligned with the provided information?
  - `Depth of Understanding`: Does the correct answer demand nuanced understanding of the context or concepts, ensuring the test taker genuinely grasps the material?
  - `Prevention of Guesswork`: Does the correct answer deter guessing and align with the context, avoiding common misconceptions?
Rubric for the Distractors:
  - `Format`: The distractors should be in similar formats as the correct answer, i.e. if it is an abbreviation or an explanation.

- `Length`: The distractors should have similar length as the correct answer.
   - `Relation`: The distractors should be related to the correct answer through some medical concepts or they should be the same kind of medical entities.
   - `Variation`: Distractors should be distinct from each other and from the correct answer.
   - `Plausibility`: Do the options align with the context and challenge critical thinking?
   - `Differentiation`: Are the options distinct and does the correct answer clearly outshine the distractors based on context and available information?
   - `Common Mistakes`: Distractors should align with common misconceptions to test genuine understanding?

Rubric for the Chain of Thought/Reasoning:
   - `Logical Flow`: Does the chain of thought exhibit a coherent sequence of steps or considerations that are easy to follow?
   - `Evidence-Based Reasoning`: Is the answer supported by evidence or information from the context, justifying the chosen response?
   - `Consideration of Options`: Does the chain of thought demonstrate critical evaluation of each option, employing a systematic process to eliminate distractors with supporting evidence?

Example of Distractors Feedback Output
 - **Format**: The format of the distractor options matches that of the correct answer. All options are treatment modalities, which is suitable for the question. **Score: 5/5**.
 - **Length**: The length of the distractor options is similar to the length of the correct answer. **Score: 5/5**.
 - **Relation**: The distractor options are related to the correct answer as they are all medical treatments, but not all are related to cancer treatment which is the specific context. **Score: 3/5**.
 - **Variation**: The distractor options are distinct from each other and from the correct answer, which helps to test the test-taker's understanding of the context. **Score: 5/5**.
 - **Plausibility**: The distractor options are plausible treatments for certain conditions, but not necessarily the condition described in the context, which is advanced cancer. **Score: 3/5**.
 - **Differentiation**: The correct answer (radiotherapy or chemotherapy) clearly outshines the distractors based on the context and available information. **Score: 5/5**.
 - **Common Mistakes**: The distractors do not align with common misconceptions about cancer treatment. They represent treatments for other conditions, not common mistakes in cancer management. **Score: 3/5**.
Distractor Score: Total Score: **29/35**

MCQG-C++ Correction Step
 - The feedback generated in the previous step is utilized to refine each of the generated components of the target output u from the initialization step.
 - M is prompted with inputs including examples for refinement, the feedback (f), and the current components (c, q, a, d, aa, r).
 - M generates $u_{refined}$, aiming to improve its performance on all aspects critiqued in the feedback.
 - $u_{refined}$ is then subjected to the answer step and then to the feedback step.
 - If the feedback scores surpass a threshold, the iterative refinement process halts. Otherwise, the process continues for up to four steps of feedback and refinement.

MCQG-C++ Correction Output
Context: A 67-year-old female with a history of endometrial cancer presented with symptoms of rectal bleeding. She had previously undergone radiation therapy and brachytherapy for her cancer. On admission, a colonoscopy was performed due to her history and the presenting symptoms, which revealed a large mass in the sigmoid colon. Biopsy of the mass showed invasive and moderately differentiated carcinoma, with immunohistochemistry suggesting an endometrial primary. The patient underwent surgical resection of the mass followed by radiation therapy.
Question: Following surgical resection of a colon mass in a patient with a history of endometrial cancer, what would be the most appropriate long-term management approach?
Correct answer: radiotherapy or chemotherapy
Distractor options:
   - a) oral corticosteroids
   - b) aspirin therapy
   - c) bisphosphonates
   - d) hormone replacement therapy
   - e) anticoagulation therapy

Table 13: Demonstration of MCQG-C++ Pipeline.

Type	Prompt
INIT Prompt - Context	```USMLE context based questions: Context and question: {context_question} Context and question: {context_question} Context and question: {context_question} Generate a context(not the question,in the format Context:) based on the given topic from the clinical ↪ note: Clinical Note: {clinical_note} Topic: {topic} Keypoint: {keypoint}"```
INIT Prompt - Question	```USMLE context based questions: Context and question: {context_question} Context and question: {context_question} Context and question: {context_question} Generate a one line question(in the format Question:) based on the given context: Context: {context} Topic: {topic} Keypoint: {keypoint}"```
INIT Prompt - Correct answer	```USMLE context based questions with their correct answers: Context and question: {context_question} Correct answer: {correct_answer} Context and question: {context_question} Correct answer: {correct_answer} Context and question: {context_question} Correct answer: {correct_answer} Generate the correct answer(in the format Correct answer:) to the question based on the given context, ↪ topic and keypoint(to which it should be highly related to) : Context: {context} Question: {question} Topic: {topic} Keypoint: {keypoint} "```
INIT Prompt - Correct answer	```USMLE context based questions with their correct answers: Context and question: {context_question} Correct answer: {correct_answer} Context and question: {context_question} Correct answer: {correct_answer} Context and question: {context_question} Correct answer: {correct_answer} Generate the correct answer(in the format Correct answer:) to the question based on the given context, ↪ topic and keypoint(to which it should be highly related to) : Context: {context} Question: {question} Topic: {topic} Keypoint: {keypoint} "```
INIT Prompt - Distractor options	```USMLE context based questions with their correct answers and distractor options: Context and Question: {question} Correct answer: {correct_answer} Distractor options: {distractor_options} Generate distractor options(in the format Distractor options:) for the context, question, and correct ↪ answer: Context: {context} Question: {question} Correct answer: {correct_answer} "```

| ANSWER Prompt | ```
# few shot exemplars (X2)
Context: {context}

Question: {question}

Options: {options}

Correct answer: {answer}

Reasoning: {reasoning}
Answer the USMLE question and provide a step by step reasoning for reaching that particular answer and
    ↪ rejecting other options
Context: {context}

Question: {question}

Options: {options}

Correct answer:
``` |
| CRITIQUE Prompt | ```
# Example USMLE component generated from the clinical note and its respective feedback:
Clinical note:{clinical_note}
Topic: {topic}
Keypoint: {keypoint}
Context:{context}
Question:{question}
Correct answer:{correct_answer}
Distractor options:{distractor_options}
{component_name} feedback: {feedback}
{component_name} score: {score}
In addition to the scoring rubrics in the examples above,give feedback and score the {component_name} using
    ↪ the attempted answer's(correct/incorrect) reasoning-based rubrics and their definitions below.
Please include both the previous scoring rubrics and the following reasoning-based rubrics before giving
    ↪ the feedback for a particular aspect and add up the scores for all the aspects for the total scores
    ↪ of the {component_name}.
Many of these feedback points for the {component_name} depend upon the reasoning and the attempted answer
    ↪ correctness so consider that while giving feedback for the {component_name}.
{component_name} reasoning-based rubrics: {reasoning_rubrics}
Give the output in just this format: {format_instructions}
Output just the JSON instance and nothing else.
Clinical note: {clinical_note}
Keypoint: {keypoint}
Topic: {topic}
Context: {context}
Question: {question}
Correct answer: {correct_answer}
Attempted answer: {attempted_answer}
Reasoning: {reasoning}
Distractor options: {distractor_options}
``` |
| CRITIQUE Prompt - Reasoning | ```
Give supporting textual feedback for each aspect and score(out of 5 for each aspect, in the format "2/5" if
    ↪ the score for that aspect is 2, also give supporting evidence for that score) the {component_name}
    ↪ using the attempted answer's(correct/incorrect) reasoning-based rubrics and their definitions
    ↪ below.
Please include the following reasoning-based rubrics before giving the feedback for a particular aspect and
    ↪ add up the scores for all the aspects for the total score of the {component_name}.
Many of these feedback points for the {component_name} depend upon the reasoning and the attempted answer
    ↪ correctness so consider that while giving feedback for the {component_name}.
{component_name} rubrics: {reasoning_rubrics}
Give the output in just this format: {format_instructions}
Output just the JSON instance and nothing else.
Clinical note: {clinical_note}
Keypoint: {keypoint}
Topic: {topic}
Context: {context}
Question: {question}
Correct answer: {correct_answer}
Attempted answer: {attempted_answer}
Reasoning: {reasoning}
Distractor options: {distractor_options}
``` |

| CORRECTION Prompt | ```
# few shot exemplars (x3), Eeach example is an improved version of the previous
Context: {context}

Question: {question}

Correct answer: {correct_answer}

Distractor Options: {distractor_options}

Feedback for the above components:

Context feedback: {context_feedback}

Context score: {context_score}

Question feedback: {question_feedback}

Question score: {question_score}

Correct answer feedback: {correct_answer_feedback}

Correct answer score: {correct_answer_score}

Distractor options feedback: {distractor_option_feedback}

Distractor options score: {distractor_option_score}

Improve the context,question, correct answer and distractor options using each previous components'
    ↪ feedback and the reasoning feedback.
Generate a context, question, correct answer and distractor options that can achieve high scores on all the
    ↪ above feedback rubrics, given the clinical note, keypoint and topic. Do not generate the feedback
    ↪ for any of the component.:

Clinical Note: {clinical_note}

Topic: {topic}

Keypoint: {keypoint}

Context: {context}

Question: {question}

Attempted answer: {attempted_answer}

Reasoning: {reasoning}

Correct answer: {correct_answer}

Distractor Options: {distractor_options}

Feedback on the generated content with respect to various rubrics.

Context feedback: {context_feedback}

Context score: {context_score}

Question feedback: {question_feedback}

Question score: {question_score}

Correct answer feedback: {correct_answer_feedback}

Correct answer score: {correct_answer_score}

Distractor options feedback: {distractor_option_feedback}

Distractor options score: {distractor_option_score}

Reasoning feedback: {reasoning_feedback}

Reasoning score: {reasoning_score}
``` |

Table 14: All Prompts.

Evaluation Guide

Your task is to evaluate two system outputs, each consisting of a machine-generated USMLE question. Each question has four components: context, a question based on the context, the correct answer, and distractor options.

The questions are generated using a clinical (EHR) note, a topic from the official USMLE topic list, and a keypoint. The keypoint is a specific aspect on which the student is tested, either related to or being the correct answer itself.

You have to annotate the following:
1. **Preference:** Please indicate your preference for the better question in the **'Preference'** column by writing **1** or **2**, signifying **Question 1** or **Question 2**, respectively. In the Remark column please give a short reasoning behind your choice.
2. **Difficulty level:** For each Question 1 and Question 2, you have to adjudge which level the question falls under: **Easy**, **Medium**, or **Hard**. Please indicate your choice in the respective Question difficulty level column, i.e., **Question 1 difficulty** and **Question 2 difficulty**.
3. **Errors:** For both Question 1 and Question 2, highlight any spans having errors, be it logical, syntactical, or medical. An example error could be if the question contradicts the clinical note. Write any such errors if they exist in the **Error** column.

Below are some rubrics you can use to evaluate the two outputs. You don't need to assess each question on every metric; simply use these as a guideline to understand what makes a good USMLE question before you start evaluating:

Metrics for the context:
   - `Relevant`: The context should be relevant to the topic provided.
   - `Concision`: The context should be concise and shouldn't include extraneous information or just be a copy of the clinical note.
   - `Coherent`: The context should be coherent enough and should organically be built for the question in the end.
   - `Consistent`: The context is consistent with the information in the clinical note and the topic.
   - `Specific`: The context should be specific and address the topic.
   - `Fluent`: The context is fluent in terms of grammar and flow of words and ideas.
   - `Clueing`: Instead of directly mentioning the diagnosis of a medical condition, it should be clued in through the mention of symptoms.
   - `Completeness`: The context should be complete and free from gaps or missing information that could lead to ambiguity in answering the question accurately.
   - `Misdirection`: Does the context avoid misleading the test taker intentionally or unintentionally?

Metrics for the question:
   - `Relevant`: The question should be answerable from the information provided in the context and should not be abrupt.
   - `Clear`: The question should not be vague or ambiguous.
   - `Concluding`: The flow of ideas from the context should organically result in the question.
   - `Difficulty`: The question should not be too easy.
   - `Clarity`: Is the question clear and unambiguous to avoid incorrect interpretations caused by ambiguity or poor wording?

Metrics for the correct answer:
   - `Relevant`: The correct answer should be either the keypoint or very related to it.
   - `Occurrence`: The correct answer or any of its semantic or syntactic forms and directly related medical concepts should not occur in the context.
   - `Justification`: Is the correct answer logically supported by the context and aligned with the provided information?
   - `Depth of Understanding`: Does the correct answer demand a nuanced understanding of the context or concepts, ensuring the test taker genuinely grasps the material?
   - `Prevention of Guesswork`: Does the correct answer deter guessing and align with the context, avoiding common misconceptions?

Metrics for the distractors:

    - `Format:` The distractors should be in similar formats as the correct answer, i.e., if it is an abbreviation or an explanation.

    - `Length:` The distractors should have a similar length as the correct answer.

    - `Relation:` The distractors should be related to the correct answer through some medical concepts, or they should be the same kind of medical entities.

    - `Variation:` Distractors should be distinct from each other and from the correct answer.

    - `Plausibility:` Do the options align with the context and challenge critical thinking?

    - `Differentiation:` Are the options distinct, and does the correct answer clearly outshine the distractors based on context and available information?

    - `Common Mistakes:` Distractors should align with common misconceptions to test genuine understanding.

Table 15: Human evaluation guide.

Cases	gpt-4o	gpt4	gpt3.5	opus	sonnet	haiku	Expert X
1	1	2	1	1	1	-1	1
2	1	1	-1	-1	2	2	2
3	-1	2	-1	-1	2	-1	2
4	1	2	-1	-1	-1	2	2
5	-1	-1	2	1	-1	1	2
6	1	-1	2	1	1	-1	2
7	-1	1	-1	2	2	2	0
8	1	1	-1	-1	-1	-1	1
9	1	1	-1	-1	-1	2	2
10	-1	1	-1	2	-1	2	2
11	-1	-1	-1	-1	-1	1	2
12	2	-1	2	1	-1	2	2
13	2	2	2	1	1	-1	1
14	2	1	-1	-1	1	2	2
15	1	-1	2	-1	-1	2	2
16	-1	-1	2	1	-1	2	0
17	1	2	2	1	2	-1	2
18	-1	2	2	1	-1	2	2
19	-1	-1	2	1	2	-1	2
20	1	-1	-1	-1	2	2	2
21	-1	1	-1	-1	-1	2	2
22	1	-1	1	1	-1	2	0
23	1	1	1	-1	1	2	2
24	-1	-1	1	1	-1	2	0
25	-1	-1	2	1	-1	1	2
26	1	2	2	1	2	-1	2
27	2	1	-1	-1	-1	-1	1
28	1	2	-1	-1	1	1	1
29	2	1	2	1	2	-1	1
30	-1	-1	1	1	-1	2	1
31	1	-1	1	1	-1	-1	2
32	-1	1	-1	1	2	-1	2
33	1	2	1	-1	-1	2	2
34	-1	1	1	-1	-1	-1	2
35	-1	1	1	-1	-1	-1	2
36	1	2	1	-1	-1	2	2
37	-1	1	-1	-1	2	-1	1
38	-1	1	-1	-1	-1	-1	2
39	1	-1	-1	1	-1	-1	2
40	1	2	1	1	2	2	2

Table 16: LLM-as-judge (comparison). 0 means tie, 1 means preferring GPT-4, 2 means preferring MCQG-C++, and -1 means position bias (LLM has different preferences in two different orders).

Case Study 1

Clinical Note: A 35-year-old man presented with the complaints of low back ache for 12 years, radiating to right leg for 4 months and numbness extending to lateral side of the sole of right foot. On examination, there was a 30% sensory loss in right S1 dermatome as compared with contralateral limb, with no bladder bowel involvement. Patient had no motor deficit. Magnetic resonance imaging (MRI) of the lumbosacral spine was suggestive of IM cystic lesion at L2-3 hypointense on T1-weighted images and hyperintense on T2-weighted images. MRI brain did not reveal any abnormality. Lumbar puncture and serologic studies were not performed. With the differential diagnosis of neoplastic lesion, the patient was taken up for posterior laminectomy. L2-3 laminectomy was done. A dural bulge was identified. On durotomy, the cord was found to be enlarged. Under microscopic guidance, posterior longitudinal myelotomy was done, the cysts were approached, and subtotal resection of cysts was done. Intraoperatively, three grayish white cysts were identified. Cysts were found to be adherent to the nerve roots causing their inflammation. As a result, one of the cysts could not be excised and was only decompressed. The remaining two cysts were completely excised. Histopathology revealed it to be NCC. The patient improved postoperatively. Back pain was relieved, and there was significant reduction in radiating pain. He was started on albendazole (15 mg/kg body weight) for 4 weeks and steroids for 2 weeks. The patient was discharged on the 4th post-operative day. He was followed-up biweekly for the first month. Thereafter, monthly follow-up was done for the next 2 months. MRI done at 6 months confirmed resolution of the cystic lesion. Thereafter, 6 monthly follow-up was done. Patient is symptom free and not on any medication at 2 years of follow-up.
  - **Topic:** select most appropriate laboratory or diagnostic study
  - **Keypoint:** MRI

GPT-4 generated USMLE question:
Context: A 35-year-old man has been experiencing low back ache for 12 years, with pain radiating to his right leg for the past 4 months. He also reports numbness extending to the lateral side of the sole of his right foot. Upon examination, a 30% sensory loss in right S1 dermatome was found compared to the contralateral limb, but there was no bladder or bowel involvement and no motor deficit. An MRI of his lumbosacral spine revealed an IM cystic lesion at L2-3, which was hypointense on T1-weighted images and hyperintense on T2-weighted images. No abnormalities were found in an MRI of the brain. The patient underwent a posterior laminectomy and L2-3 laminectomy, during which three grayish-white cysts were identified. These cysts were found to be adhering to the nerve roots, causing inflammation. One cyst could not be completely excised and was only decompressed, while the other two were completely removed. Histopathology confirmed the diagnosis of NCC. Postoperative improvement was noted in the patient, with relief from back pain and significant reduction in radiating pain. He was prescribed albendazole and steroids and was discharged on the 4th post-operative day. Follow-ups were conducted biweekly for the first month, monthly for the next two months, and then every six months. An MRI taken six months post-operatively confirmed the resolution of the cystic lesion. After two years, the patient is symptom-free and not on any medication.
Question: What is the most appropriate diagnostic study to confirm the resolution of a cystic lesion in the lumbosacral spine?
Correct answer: MRI
Distractor options:
  - A : CT scan
  - B : Ultrasound
  - C : X-ray
  - D : EEG

MCQG-C++ generated USMLE question
Context: A 35-year-old patient had been experiencing lower back pain for over a decade, which recently started radiating to his right leg and was accompanied by numbness in his right foot. Although the patient had no motor deficit, a 30% sensory loss was detected in the right S1 dermatome. A key diagnostic tool was utilized to identify the cause of his symptoms.

Question: Which imaging technique was pivotal in diagnosing the patient's condition and guiding subsequent treatment?
Correct answer: MRI
Distractor options:
- A : CT scan
- B : Ultrasound
- C : X-ray
- D : EEG

Human Evaluation:
Expert X:
Preference:
    MCQG-C++
Difficulty:
    GPT-4: Easy
    MCQG-C++: Hard
Reasons:
    1. Question (GPT-4) consists of the answer to the question in the context.
    2. Question (MCQG-C++) context more Concision.

Case Study 2

Clinical Note: This 33-year-old female patient had typical COVID-19 symptoms such as high fever, dry cough, headache, and dyspnea about 1 week before ICU admission. She was intubated and proned due to rapid respiratory deterioration. For the following 6 days, her situation was unstable, and physical therapy consisted of prone positioning and prevention of secondary damage. From day 7 onwards, she started to improve rapidly and could be mobilized passively into a side-edge position. After extubation, she presented postextubation dysphagia and severe ICU-acquired weakness (MRC-SS 36/60). She also suffered from pronounced delirium and anxiety and said repeatedly that she had been abducted and that she believed she had to die. She seemed to feel threatened by us and it was difficult to calm her down. Due to the pandemic measures of the Swiss government, hospital visits were not generally allowed, but because her anxiety was limiting her rehabilitation, her husband was granted an exceptional permission to visit her. This seemed to give the patient a short sense of security, and she started to participate in some basic functional activities (CPAx 21/50). Nevertheless, the delirium did not resolve upon her transfer to a peripheral acute hospital.
- **Topic:** the cause/infectious agent or predisposing factor(s)
- **Keypoint:** covid-19

GPT-4 generated USMLE question:
Context: A 33-year-old female patient presented with typical symptoms of COVID-19, including high fever, dry cough, headache, and dyspnea. She was admitted to the ICU due to rapid respiratory deterioration and was intubated and proned. Despite physical therapy and prone positioning, her condition remained unstable for 6 days. From the seventh day, she started showing signs of improvement and was passively mobilized into a side-edge position. However, post-extubation, she experienced dysphagia and severe ICU-acquired weakness along with pronounced delirium and anxiety. Her rehabilitation was further hampered due to her delirium, which led to her belief of being abducted and fear of death. Due to her heightened anxiety, her husband was granted special permission to visit her, which provided her with a temporary sense of security and encouraged her to participate in basic functional activities. Despite this, her delirium persisted even after her transfer to a peripheral acute hospital.
Question: What infectious agent is responsible for the symptoms and complications experienced by the 33-year-old female patient?
Correct answer: SARS-CoV-2 (COVID-19)
Distractor options:
- A : Influenza A virus
- B : Human immunodeficiency virus (HIV)
- C : Mycobacterium tuberculosis

- D : Streptococcus pneumoniae

MCQG-C++ generated USMLE question
Context: A 33-year-old female patient was admitted to the ICU following the onset of high fever, dry cough, headache, and dyspnea. Following intubation, her condition remained unstable for nearly a week. Upon extubation, she displayed severe weakness, dysphagia, and exhibited signs of delirium and anxiety.
Question: Given the patient's symptoms and the course of her illness, what is the most probable causative agent of her condition?
Correct answer: sars-cov-2 (covid-19)
Distractor options:
   - A : Influenza a virus
   - B : Pneumococcus bacteria
   - C : Mycobacterium tuberculosis
   - D : Epstein-Barr virus

Human Evaluation:
Expert X:
Preference:
   GPT-4
Difficulty:
   GPT-4: Easy
   MCQG-C++: Hard
Reasons:
   1. According to the context in Question (MCQG-C++), it is hard to get the right answer. Question (MCQG-C++) tries to make it harder, but at the same time, it becomes confusing.
   2. The answer to Question (MCQG-C++) can be all the following distractors. Because there is no evidence to suspect the infection of COVID-19. Question (GPT-4), however, starts the context by giving evidence that the patient was already diagnosed with Covid-19. But the answer also has COVID-19, so this is an error of the question.

Case Study 3

Clinical Note: A 35-year-old male patient presented with a six-month history of pain in the anterior chest and neck. His pain initially had been mild but had become severe in the last three weeks. He was afebrile and did not have any constitutional symptoms. On examination, marked tenderness of the right sternoclavicular joint was noted. Initial laboratory investigations, tuberculosis workup, and chest X-rays were unremarkable. However, further study with MRI revealed subchondral bone marrow edema and enhancement involving the medial end of the right clavicle. In addition, mild effusion of the right sternoclavicular joint and surrounding soft-tissue edema was also seen (Figure ,). Given the patient's radiological and clinical findings, suspicion of SAPHO syndrome was raised. He was given an initial trial of paracetamol and NSAIDs and he showed remarkable improvement on his monthly follow-up visits. Though a biopsy was offered, the patient refused to undergo one and has shown no relapse to date thanks to the dramatic response to NSAIDs.
   - **Topic:** select most likely diagnosis yes
   - **Keypoint:** SAPHO syndrome

GPT-4 generated USMLE question:

Context: A 35-year-old male patient has been experiencing anterior chest and neck pain for six months, which has become severe in the last three weeks. Despite being afebrile and showing no constitutional symptoms, the right sternoclavicular joint was notably tender on examination. Initial investigations, including laboratory tests, tuberculosis workup, and chest X-rays, yielded no remarkable findings. However, an MRI revealed subchondral bone marrow edema and enhancement at the medial end of the right clavicle, along with mild effusion of the right sternoclavicular joint and surrounding soft-tissue edema. Given these clinical and radiological findings, the suspicion of SAPHO syndrome was raised. The patient was started on paracetamol and NSAIDs, which led to a significant improvement in his condition during subsequent monthly follow-ups. Although a biopsy was considered, the patient declined. He has remained free of relapse due to the effective response to NSAIDs.

Question: Based on the clinical and radiological findings, what is the most likely diagnosis for a 35-year-old male patient experiencing severe anterior chest and neck pain, and showing signs of subchondral bone marrow edema and enhancement at the medial end of the right clavicle on an MRI?

Correct answer: SAPHO syndrome

Distractor options:
- A : Rheumatoid arthritis
- B : Osteomyelitis.
- C : Gout.
- D : Bone cancer.
- E : Paget's disease of bone.

MCQG-C++ generated USMLE question

Context: A 35-year-old male has been dealing with pain in his anterior chest and neck area for the past half a year, which has recently escalated in severity. Despite his discomfort, he had no other systemic symptoms and was afebrile. Upon physical examination, severe tenderness was noted in the right sternoclavicular joint. While initial investigations including laboratory tests and chest radiographs showed no abnormalities, an MRI scan revealed edema and enhancement in the bone marrow beneath the cartilage at the medial end of the right clavicle. There were also signs of mild effusion in the right sternoclavicular joint and surrounding soft tissue edema. After being administered paracetamol and NSAIDs, the patient showed significant improvement in his condition.

Question: What diagnosis does the clinical presentation and the results of the MRI scan most likely indicate?

Correct answer: SAPHO syndrome
- A : Rheumatoid arthritis
- B : Osteomyelitis
- C : Gout
- D : Bone cancer
- E : Paget's disease of bone

Human Evaluation:
Expert X:
Preference:
  MCQG-C++
Difficulty:
  GPT-4: Easy
  MCQG-C++: Hard
Reasons:
  1. In Question (GPT-4), the answer was directly listed in the context.
  2. In Question (MCQG-C++), only the signs/symptoms for SAPHO syndrome are given. "

Case Study 4

Clinical Note: A 35-year-old male patient presented with a six-month history of pain in the anterior chest and neck. His pain initially had been mild but had become severe in the last three weeks. He was afebrile and did not have any constitutional symptoms. On examination, marked tenderness of the right sternoclavicular joint was noted. Initial laboratory investigations, tuberculosis workup, and chest X-rays were unremarkable. However, further study with MRI revealed subchondral bone marrow edema and enhancement involving the medial end of the right clavicle. In addition, mild effusion of the right sternoclavicular joint and surrounding soft-tissue edema was also seen (Figure ,). Given the patient's radiological and clinical findings, suspicion of SAPHO syndrome was raised. He was given an initial trial of paracetamol and NSAIDs and he showed remarkable improvement on his monthly follow-up visits. Though a biopsy was offered, the patient refused to undergo one and has shown no relapse to date thanks to the dramatic response to NSAIDs.
- **Topic:** knows signs/symptoms of selected disorders
- **Keypoint:** SAPHO syndrome

GPT-4 generated USMLE question:
Context: A 35-year-old male presents with a six-month history of anterior chest and neck pain that has progressively worsened over the last three weeks. The patient is afebrile and does not display any constitutional symptoms. Upon examination, significant tenderness is noted in the right sternoclavicular joint. Initial laboratory tests, including a tuberculosis workup and chest X-rays, do not reveal any abnormalities. However, an MRI shows subchondral bone marrow edema and enhancement at the medial end of the right clavicle, along with mild effusion of the right sternoclavicular joint and surrounding soft-tissue edema. The patient's clinical and radiological findings lead to a suspicion of SAPHO syndrome. An initial treatment with paracetamol and NSAIDs results in significant improvement, and the patient continues to show no signs of relapse. Despite the offer of a biopsy, the patient chooses not to undergo the procedure.
Question: What are the common signs and symptoms of SAPHO syndrome?
Correct answer: Anterior chest and neck pain, tenderness in the sternoclavicular joint, subchondral bone marrow edema and enhancement, mild effusion of the sternoclavicular joint and surrounding soft-tissue edema.
Distractor options:
- A : Swelling and redness over the affected joint, accompanied by intense pain.
- B : Chronic diarrhea, abdominal pain, and weight loss.
- C : Presence of skin rashes, dry mouth and eyes.
- D : Persistent cough, chest pain, and difficulty in breathing.
- E : Severe headache, nausea, and light sensitivity.

MCQG-C++ generated USMLE question
Context: A 35-year-old male patient has been experiencing pain in his chest and neck for six months. This pain has been intensifying over the past three weeks. He does not have any fever or other systemic symptoms. During the physical exam, we noticed significant tenderness in his right sternoclavicular joint. While initial lab tests and chest X-rays didn't reveal anything unusual, an MRI showed signs of subchondral bone marrow edema in the medial end of the right clavicle, a mild effusion in the right sternoclavicular joint, and some soft-tissue edema surrounding the joint. These clinical and radiological findings are suggestive of a particular condition. The patient was treated with paracetamol and NSAIDs, which led to a significant improvement in his symptoms.
Question: Based on the described signs and symptoms, which syndrome is most likely to be suspected in this patient?
Correct answer: sapho syndrome
Distractor options:
- A : Ehlers-danlos syndrome
- B : Marfan syndrome
- C : Reiter's syndrome
- D : Guillain-barre syndrome

Human Evaluation:
Expert X:
Preference:
  MCQG-C++

Difficulty:
    GPT-4: Easy
    MCQG-C++: Hard
Reasons:
    1. For Question (GPT-4), the common signs and symptoms of SAPHO syndrome are not Anterior chest and neck pain, tenderness in the sternoclavicular joint, subchondral bone marrow edema and enhancement, mild effusion of the sternoclavicular joint and surrounding soft-tissue edema. This is just the symptom for the giving case.
    2. Question (MCQG-C++) consists the symptoms of SAPHO and asks to choose the right diagnosis for the patient which is a common USMLE type question. Question (GPT-4) also is a good way to test whether the test-taker knows the signs/symptoms of SAPHO. But the context is too long, and the answers are a little bit obvious.
    3. Question (GPT-4) is the missing context for diagnosis, but Question (MCQG-C++)'s Distractor options are not that relevant and need correction.
    4. In the Question (MCQG-C++) context, there is a hint about the signs and symptoms of the patient's disease.

Table 17: Human evaluation case study.

Component	Aspect	Percentage agreement	Cohen's kappa
Context	consistent	10.8108	-0.0008196
Context	misdirection	10.52631	-0.001550
Context	coherent	26.31578	0.065026
Context	clueing	28.947368	0.0393258
Context	concision	26.3157	0.04488
Context	relevant	18.42105	0.010084
Context	specific	15.7894	-2.22044
Context	completeness	13.15789	0.0079113
Context	fluent	10.5263	-0.00310
Question	difficulty	27.50000	-0.01665
Question	relevant	27.50000	-0.03386
Question	clarity	40.0	0.08309
Question	clear	42.5	0.05056
Question	concluding	52.5	0.197465
Correct Answer	depth of understanding	30.7692	-0.02631
Correct Answer	relevant	12.8205	0.008968
Correct Answer	justification	12.8205	0.008968
Correct Answer	occurrence	28.20512	0.07379
Correct Answer	prevention of guesswork	38.461	0.0370
Distractor	variation	10.2564	0.0
Distractor	plausibility	31.578	0.023715
Distractor	common mistakes	33.333	0.10582
Distractor	length	19.4444	0.035120
Distractor	differentiation	17.9487	-0.012165
Distractor	relation	28.205	-0.02535
Distractor	format	12.820	-0.06763
Reasoning	logical flow	7.5	-0.069364
Reasoning	correctness	7.5	-0.048158
Reasoning	consideration of options	10.0	-0.0020876
Reasoning	evidence˙based˙reasoning	5.0	-0.081081

Table 18: Aspect vs Expert-X Preference correlation (Human Eval 40 data).

Component	Percentage agreement	Cohen's kappa
Context	0.38235	0.0338
Question	0.5	0.05472
Correct Answer	0.47058	0.15522
Distractor	0.3235	-0.0358
Reasoning	0.08823	-0.07168

Table 19: Component vs Expert-X Preference correlation (Human Eval 40 data). The correct answer is the highest, maybe because it is easier to judge a correct answer since it is the shortest lengthwise, and in case it is right or wrong, it is very easy to reject a particular question, whereas other components are considerably more subjective.

	Component	Aspect	Percentage Agreement	Cohen's kappa
1	question	concluding	52.5	0.19746568109820484
2	question	clear	47.5	0.08296943231441045
3	question	clarity	47.5	0.05084745762711862
4	correct answer	prevention of guesswork	62.5	0.1724137931034483
5	distractor option	common mistakes of guesswork	62.5	0.16897506925202774
6	distractor option	plausibility	60.0	0.10987482614742705
7	correct answer	depth of understanding	60.0	0.11724137931034484
8	context	clueing	65.0	0.1384615384615384
9	correct answer	occurrence of guesswork	65.0	0.19308357348703176
10	distractor option	relation	67.5	0.21921921921921916
11	question	relevant	70.0	0.3083573487031701
12	question	difficulty	70.0	0.30535455861070904
13	context	coherent	72.5	0.34230194319880425
14	context	concision	72.5	0.34230194319880425
15	distractor option	length	70.0	0.30535455861070904
16	context	relevant	70.0	0.3083573487031701
17	distractor option	differentiation	70.0	0.3083573487031701
18	context	specific	70.0	0.3083573487031701
19	context	completeness	70.0	0.3083573487031701
20	correct answer	relevant	70.0	0.3083573487031701
21	correct answer	justification	70.0	0.3083573487031701
22	distractor option	format	67.5	0.22619047619047628
23	context	consistent	67.5	0.22619047619047628
24	context	misdirection	67.5	0.22619047619047628
25	context	fluent	67.5	0.22619047619047628
26	distractor option	variation	67.5	0.22619047619047628
27	reasoning	consideration of options	67.5	0.22619047619047628
28	reasoning	logical flow	67.5	0.22619047619047628
29	reasoning	correctness	67.5	0.22619047619047628
30	reasoning	evidence based reasoning	67.5	0.22619047619047628

Table 20: Percentage agreement based Greedy aspect filtering.

	Component	Aspect	Percentage Agreement	Cohen's kappa
1	question	concluding	52.5	0.19746568109820484
2	distractor option	common mistakes	60.0	0.2660550458715596
3	question	clarity	57.5	0.169719169719169718
4	correct answer	occurrence	67.5	0.3603936039360395
5	context	coherent	72.5	0.4233289646133682
6	question	clear	65.0	0.2660550458715596
7	context	concision	67.5	0.31847968545216254
8	context	clueing	65.0	0.24119241192411933
9	correct answer	prevention of guesswork	70.0	0.30535455861070904
10	distractor option	length	70.0	0.30535455861070904
11	distractor option	plausibility	70.0	0.3083573487031701
12	context	relevant	70.0	0.30535455861070904
13	correct answer	relevant	70.0	0.30535455861070904
14	correct answer	justification	70.0	0.30535455861070904
15	context	completeness	70.0	0.30535455861070904
16	distractor option	variation	70.0	0.3083573487031701
17	context	specific	67.5	0.2737430167597765
18	context	consistent	67.5	0.2737430167597765
19	context	misdirection	67.5	0.2737430167597765
20	reasoning	consideration of options	67.5	0.2737430167597765
21	context	fluent	67.5	0.2737430167597765
22	distractor option	differentiation	67.5	0.2737430167597765
23	question	difficulty	67.5	0.2737430167597765
24	distractor option	relation	67.5	0.22619047619047628
25	correct answer	depth of understanding	70.0	0.25811437403400295
26	question	relevant	70.0	0.3083573487031701
27	reasoning	correctness	70.0	0.3083573487031701
28	distractor option	format	67.5	0.22619047619047628
29	reasoning	logical flow	67.5	0.22619047619047628
30	reasoning	evidence based reasoning	67.5	0.22619047619047628

Table 21: Cohen's kappa-based Greedy aspect filtering.

	Combination	Percentage Agreement	Cohen's kappa
Top 1	question - concluding	52.5	0.1974656
Top 2	question - concluding correct answer - occurrence	67.5	0.39953
Top 3	question - concluding context - clueing correct answer - prevention of guesswork	72.5	0.445843
Top 4	question - concluding context - concision correct answer - prevention of guesswork correct answer - occurrence	75.0	0.479843
Top 5	question - concluding context - coherent correct answer - occurrence correct answer - depth of understanding distractor - common mistakes	77.5	0.51612
Top 6	question - concluding context - coherent context - consistent correct answer - occurrence correct answer - depth of understanding distractor - common mistakes	77.5	0.51612
Top 7	question - concluding context - coherent context - consistent context - misdirection correct answer - occurrence correct answer - depth of understanding distractor - common mistakes	77.5	0.51612
Top 8	question - concluding question - relevant context - clueing context - concision correct answer - prevention of guesswork correct answer - occurrence correct answer - depth of understanding distractor - plausibility	80.0	0.5054095826893353
Top 9	question - concluding question - relevant context - consistent context - clueing context - concision correct answer - occurrence correct answer - depth of understanding correct answer - prevention of guesswork distractor - plausibility	80.0	0.5054095826893353
Top 10	question - concluding question - relevant context - clueing context - completeness context - misdirection context - concision correct answer - occurrence correct answer - prevention of guesswork correct answer - depth of understanding distractor - plausibility	80.0	0.5054095826893353

Table 22: Percentage Agreement-based All-Combination aspect filtering.

	Combination	Percentage Agreement	Cohen's kappa
Top 1	question - concluding	52.5	0.1974656
Top 2	question - concluding correct answer - occurrence	67.5	0.39953
Top 3	question - concluding correct answer - depth of understanding correct answer - occurrence	72.5	0.445843
Top 4	question - concluding correct answer - depth of understanding correct answer - relevant correct answer - occurrence	75.0	0.479843
Top 5	question - concluding context - coherent correct answer - occurrence correct answer - depth of understanding distractor - common mistakes	77.5	0.51612
Top 6	question - concluding context - coherent context - consistent correct answer - occurrence correct answer - depth of understanding distractor - common mistakes	77.5	0.51612
Top 7	question - concluding context - coherent context - consistent context - misdirection correct answer - occurrence correct answer - depth of understanding distractor - common mistakes	77.5	0.51612
Top 8	question - concluding context - coherent context - consistent context - misdirection context - completeness correct answer - occurrence correct answer - depth of understanding distractor - common mistakes	77.5	0.51612
Top 9	question - concluding question - clarity context - concision context - relevant correct answer - occurrence correct answer - depth of understanding distractor - common mistakes reasoning - evidence based reasoning reasoning - logical flow	80.0	0.5389048991354467
Top 10	question - concluding question - clarity context - concision context - relevant context - misdirection correct answer - occurrence correct answer - depth of understanding distractor - common mistakes reasoning - evidence based reasoning reasoning - logical flow	80.0	0.5389048991354467

Table 23: Cohen's kappa-based All-Combination aspect filtering.