

# Few shot chain-of-thought driven reasoning to prompt LLMs for open-ended medical question answering

Anonymous

## Abstract

In this paper, we propose a modified version of the MedQA-USMLE dataset (MEDQANOOPT), which contains (i) open-ended medical questions without options, to mimic clinical scenarios along with (ii) clinician-approved reasoned answers. Further, we implement a prompt, driven by Chain of Thought (CoT) reasoning (MEDCODEX) to mirror the prospective process of (incremental reasoning) reaching a correct response to the medical questions. We empirically demonstrate how MEDCODEX outperforms the state-of-the-art 5-shot-codex-CoT-prompt (Liévin et al., 2022). We also present an approach that mirrors real-life clinical practice by first exploring multiple differential diagnoses (MEDCODEXFEWSHOTPROMPT) and subsequently narrowing down to a final diagnosis (CODEXFEWSHOTPROMPT). Finally, keeping in mind the importance of response verification in the medical setting, we utilize a reward model mechanism replacing the elimination performed by CODEXFEWSHOTPROMPT<sup>1</sup>.

## 1 Introduction

Large Language models (LLMs) are increasingly utilized in the healthcare sector, particularly for patient query-related tasks. These LLM-driven tools could potentially interpret and respond to patient inquiries, and provide information on symptoms, diseases, treatments, and healthcare guidelines (Thirunavukarasu et al., 2023). By analyzing vast amounts of medical literature and data, LLMs could also offer precise, up-to-date responses, improving patient education and engagement (Singhal et al., 2022). The ability of LLMs to process natural language queries makes them accessible and user-friendly, thus enhancing the patient experience and

<sup>1</sup>We plan to release the code and the developed prompts for our experiments and a Demo for our methods on Github upon acceptance of this submission. We will also be releasing any trained models on Huggingface

satisfaction (Clusmann et al., 2023). As technology evolves, LLMs are expected to play a pivotal role in delivering personalized healthcare information, contributing to informed decision-making and better health outcomes (Clusmann et al., 2023). However, with the recent advances in prompt engineering techniques in the Large Language Models space, there is an underlying requirement for accuracy and verifiability.

### 1.1 Key Contributions

We propose a healthcare-specific language model response generation task that utilizes the best prompting methods.

We build upon the MedQA-USMLE dataset (MEDQAORIGINAL) with the help of medical experts and construct a novel medical question corpus that contains open-ended medical questions, along with expert-verified appropriate reasoning as the answer (MEDQANOOPT).

We propose the MEDCODEXFEWSHOTPROMPT designed to reach a correct response to the medical questions by incremental reasoning. This is a 5-shot prompt (Appendix A) that uses Chain-of-Thought reasoning drawing successive inferences from a given patient history. This prompt integrates signs, as well as investigation reports to lead to a final reasoned response. In our comparisons, we show that the (MEDCODEXFEWSHOTPROMPT) outperforms the 5-shot-codex-CoT-prompt from Liévin et al. (2022).

We demonstrate that prospective, incremental and reasoning-driven prompting, that mimicks real-life clinical scenarios, performs significantly better at answering open-ended medical questions.

Further, we implemented a method of developing several differential diagnoses (MEDCODEXFEWSHOTPROMPT) followed by trimming down to a final diagnosis (CODEXFEWSHOTPROMPT) mirroring real-life clinical practice. This *forward-backward*

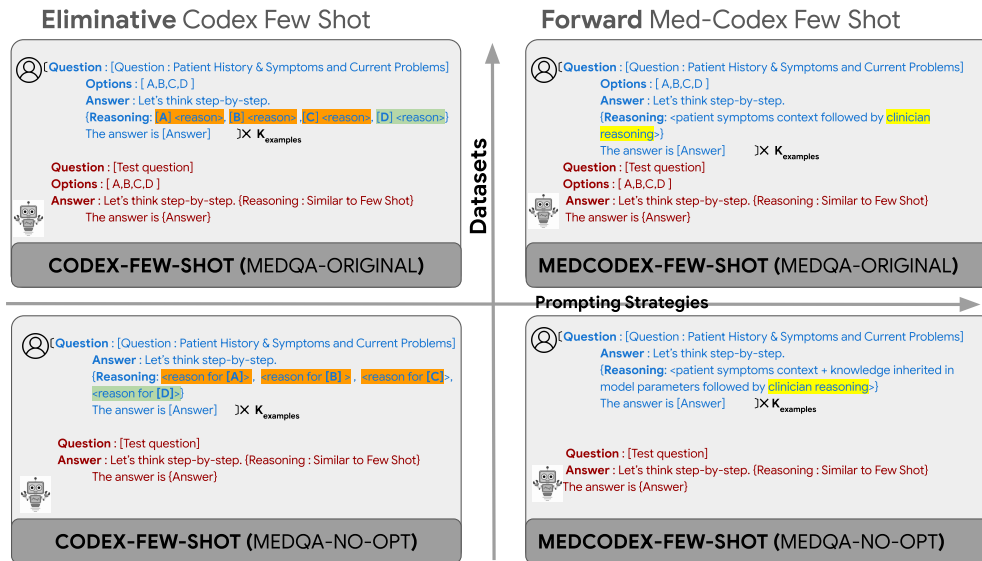


Figure 1: The overall structure of CODEXFEWSHOTPROMPT and MEDCODEXFEWSHOTPROMPT

method mimics real-life clinicians’ thought processes. We created a reward modeling dataset for building a verifier on medical question-response pairs. In this dataset, we include *Correct* and *Incorrect* Question-Reasoning-Answer triplets, a representative sample of which have been evaluated and verified by medical experts. We use the MEDQA-ORIGINAL dataset for the Question-Answer pair and artificially generate the Reasoning. We also release the reward model *verifier* trained with this dataset. Finally, the reward model replaced the (CODEXFEWSHOTPROMPT) method in the (*forward backward*) method and eliminated the inaccurate options to generate the final diagnosis.

## 2 Background & Related Work

The USMLE-MedQA dataset (Jin et al., 2021), a medical exam dataset, consists of questions sourced from professional medical board exams in the USA. The dataset is a comprehensive resource for investigating medical question answering in a standardized MCQ examination format. The dataset offers complexity and challenge, however, the current version of the dataset is not practically useful to be deployed in a patient query-based healthcare system since four options are associated with each question thereby not being similar to in-clinic scenarios. Recent advancements in the application of Large Language Models (LLMs) for medical question answering have highlighted the potential of these models to demonstrate applicability

in medical diagnostics, education, and research. Venigalla et al. (2022) introduced BioMedLM, demonstrating its performance on several medical datasets, including MedQA, underscoring the capabilities of LLMs in handling complex medical queries (Venigalla et al., 2022; Jin et al., 2021). Furthermore, Singhal et al. (2022) conducted extensive experiments across a broad spectrum of medical question tasks within the MultiMedQA suite—encompassing datasets such as MedQA, MedMCQA, PubMedQA, and MMLU—have, and demonstrated the versatility and depth of knowledge encoded in these models (Singhal et al., 2022; Pal et al., 2022; Jin et al., 2019; Hendrycks et al., 2020). The exploration of LLMs in generating not only accurate but also reasoning-based responses to medical questions, marks a significant step forward. Models such as PubMedGPT (Bolton et al., 2022) and CODEX (Liévin et al., 2022) have established benchmarks on datasets like MedQA through innovative approaches, including Classification head, Chain-of-Thought, and Knowledge Grounding, highlighting the importance of not just what is answered, but how the answer is derived. Advanced LLMs, such as Med-Palm2 (Singhal et al., 2022), and Flan-Palm (Chung et al., 2022), have claimed recent benchmarks for problem settings such as MedQA, although the limited availability of these models poses challenges for widespread research and application in medicine.

Answering medical questions involves the requirement of fact verification as well as the neces-

sity of reasoning skills. These have been attempted to be addressed through the following methods:

**Fact verification in LLMs:** Fact verification in LLMs involves computational linguistics, AI, and digital media. Guo et al. (2022) present detailed approaches to automated fact-checking, including claim detection, evidence retrieval, and verification. This process assesses truthfulness, identifies evidence, and generates explanations, enhancing robust verification systems (Guo et al., 2022).

**Chain-of-Thought Prompting:** Initially, scaling language models up appeared to benefit more knowledge-intensive tasks than reasoning-heavy ones (Rae et al., 2022). Nevertheless, Wei et al. (2022) demonstrated that LLMs could be applied to reasoning intensive problems by prompting the model to generate step-by-step solutions, coined “Chain-of-Thought” (CoT). CoT prompting led to substantial improvements in many reasoning-intensive tasks (Wei et al., 2022; Zhou et al., 2022; Drozdov et al., 2022; Nye et al., 2021), allowing to bridge the gap with human-level performances. The work of Hendrycks et al. (2020) evaluates LLMs using the MMLU dataset across diverse subjects like medicine, showcasing their broad applicability. Additionally, Mesinovic et al. (2023) emphasize the importance of model interpretability and explainability for trust in medical applications.

### 3 Problem Setting and Preliminary Notations

Let us consider a language model at inference time  $\mathcal{LM}_\theta : \mathcal{P} \rightarrow \mathcal{Y}$  parameterized by  $\theta$  with the prompt  $\mathcal{P} = \langle q_i, \mathcal{O}_i, \mathcal{R}_i \rangle$  being passed as an input to  $\mathcal{LM}_\theta$  to generate corresponding response  $y_i \in \mathcal{Y}$ . We next get into some detailed specification of the components of the prompt  $\mathcal{P} = \langle q_i, \mathcal{O}_i, \mathcal{R}_i \rangle$ .

Assume we receive a particular question  $q_i \in \mathcal{Q}$  i.e., from the universal set of questions  $\mathcal{Q}$ . Each question  $q_i$  is associated with a set of options  $\mathcal{O}_i = \{o_i^j\}_{j=1}^m$  of cardinality  $m$ , of which there is one correct option  $o_i^c \in \mathcal{O}_i$ . For each option  $o_i^j \in \mathcal{O}_i$ , consider  $\mathbb{R}^i$  to be the reasoning distribution associated with the corresponding question  $q_i$ . Let a specific reasoning associated with a particular option  $o_i^j$  be denoted as  $r_j \sim \mathbb{R}^i(o_i^j | q_i; \theta)$ .

We denote by  $\mathcal{R}_i$ , the collection of all such reasonings for all options  $\mathcal{O}_i$  of question  $q_i$ . The model produces a response  $y_i$ , each of which includes the corresponding reasoning quality as well as the cor-

rectness of the response based on ground truth for question  $q_i$ .

## 4 Experiments, Evaluation Methods and Results

### 4.1 Datasets & Models

The MedQA-USMLE dataset (Zhang et al., 2018) consists of multiple-choice questions based on United States Medical License Exams (USMLE). We refer to this as the MEDQAORIGINAL dataset here after. We modify the questions to seek descriptive responses without providing options (MEDQANOOPT)<sup>2</sup>, unlike the original setting. The MEDQANOOPT dataset helps us evaluate the capability of our methods on the tasks effectively, in a setting that is close to a real-life clinical scenario. We consider both of these two variations of the MedQA dataset for our further experimentation.

The MEDQAORIGINAL and MEDQANOOPT datasets are an open-source dataset and a modification of it respectively. However other than that, we also obtained 25 real-life Question-Answer pairs from practicing medical professionals that are based on cases encountered by them on a daily basis. We refer to this as the CLINICIANCASESTUDY dataset.

### Conversion of MCQ type questions to open-ended questions

The conversion of MCQ-type questions (MEDQAORIGINAL) to open-ended questions (MEDQANOOPT) is aimed to emulate real-world medical scenarios where open-ended inquiries are prevalent. This modification required our model to respond without predefined choices, fostering holistic reasoning by leveraging the entire parametric knowledge of  $\mathcal{LM}$ . By eliminating answer options, we assessed the model’s depth and quality of reasoning skills, ensuring a more realistic evaluation of its performance in complex medical scenarios.

We also make the prompt  $\mathcal{P}$  cater to an open-ended QA scenario by modifying the question  $q_i$  under the prompt  $\mathcal{P}$  with descriptive question clauses  $\mathcal{C}_q^i$  such as, “What is the most likely . . .”, “The addition of what . . .”, “What is the most appropriate . . .”. This modification of  $q_i$  under the objective MCQ regime to the

<sup>2</sup>Examples provided in ppendix B, we are also providing the complete MEDQANOOPTdataset

MEDQAORIGINAL (MCQ Type)	MEDQANOOPT (Descriptive Type)
<p><b>Question:</b> Four weeks after starting hydrochlorothiazide, a 49-year-old man with hypertension comes to the physician because of muscle cramps and weakness. His home medications also include amlodipine. His blood pressure today is 176/87 mm Hg. Physical examination shows no abnormalities. The precordial leads of a 12-lead ECG are shown. The addition of <b>which of the following</b> is most likely to have prevented this patient's condition?  <b>(A)</b> Torsemide <b>(B)</b> Nifedipine <b>(C)</b> Eplerenone <b>(D)</b> Hydralazine</p>	<p><b>Question:</b> Four weeks after starting hydrochlorothiazide, a 49-year-old man with hypertension comes to the physician because of muscle cramps and weakness. His home medications also include amlodipine. His blood pressure today is 176/87 mm Hg. Physical examination shows no abnormalities. The precordial leads of a 12-lead ECG are shown. The addition of <b>which medication is most likely</b> to have prevented this patient's condition?</p>

Table 1: Sample Question from the MEDQAORIGINAL dataset along with its modified counterpart, which is a part of MEDQANOOPT dataset. For more details, check Appendix B

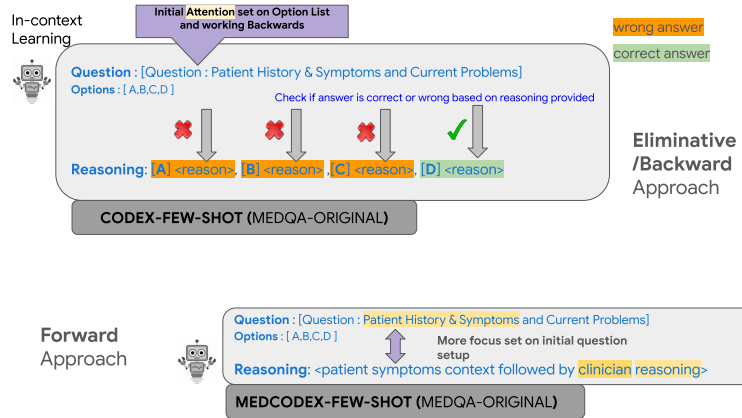


Figure 2: CODEXFEWSHOTPROMPT contains an eliminative form of reasoning, which iterates over options, accepting or discarding them as per their correctness. Often this does not cater to the context of real-life clinical investigation, unlike the incremental reasoning method as per MEDCODEXFEWSHOTPROMPT

open-ended regime  $q_i + \mathcal{C}_q^i$  is performed under clinical supervision with the help of medical experts.

## Models

We use Llama-2-7B-Chat and 70B-Chat model (Touvron et al., 2023) for experiments<sup>3</sup> & evaluations.

## 4.2 Prompting strategies<sup>4</sup>

We utilize different prompting strategies for  $\mathcal{P}$  which can capture variations in the reasoning  $\mathcal{R}_i$  i.e., how we would like the model to reason about reaching a particular correct response.

**CODXFEWSHOTPROMPT:** Codex is a GPT-based model that was first introduced in (Chen et al., 2021). The model was also evaluated on problems from professional medicine (Liévin et al., 2022) and is shown to perform well with prompt-

<sup>3</sup>For generating responses, we use *repetition\_penalty* = 1.1. We have used **greedy** decoding, unless specified otherwise. For when **sample** decoding is specified, we use *temperature* = 1, *top\_p* = 1, *top\_k* = 50, *typical\_p* = 1

<sup>4</sup>The prompts are presented in Appendix A

ing strategies on the MCQ-type questionnaire. The paper introduces *Codex-few-shot prompt* which we adapt for our problem as well. For an MCQ-type question, the reasoning structure in the prompt seems to follow this pattern: The prompt states an option  $o_j^i$ , followed by corresponding reasoning  $r_{ij}$ . The reasoning  $r_{ij}$  clarifies whether the option is considered appropriate or not, before moving onto the next option.

$$\mathcal{R}_i = \{o_j^i \in \mathcal{O}_i, r_{ij}\}_{j=1}^m$$

However, such prompting strategies tend to be eliminative in nature, as shown in Figure 2 and often contradict real-life clinical scenarios where prospective additive reasoning is needed for reaching a differential diagnosis.

**MEDCODEXFEWSHOTPROMPT:** Here, we utilize the intermittent reasoning structure. Given a unique medical context, it builds upon every additional context in answering clinical questions that mimics the usual way of reaching a final diagnosis. In real clinical scenarios, there are no four options

### Forward/MedCodex Approach



**Question:** A 59-year-old overweight woman presents to the urgent care clinic with the complaint of severe abdominal pain for the past 2 hours. She also complains of a dull pain in her back with nausea and vomiting several times. Her pain has no relation with food. Her past medical history is significant for recurrent abdominal pain due to cholelithiasis. Her father died at the age of 60 with some form of abdominal cancer. Her temperature is 37°C (98.6°F), respirations are 15/min, pulse is 67/min, and blood pressure is 122/98 mm Hg. Physical exam is unremarkable. However, a CT scan of the abdomen shows a calcified mass near her gallbladder. Which of the following diagnoses should be excluded first in this patient?  
(A) Acute cholecystitis (B) Gallbladder cancer (C) Cholelithiasis (D) Pancreatitis



**Answer:** Let's think step-by-step. 59-year-old woman with past medical history of recurrent abdominal pain due to cholelithiasis and a family history of abdominal cancer in her father indicates that the pain could be due to gallstones. The pain is severe and has no relation to food which indicates that it could be due to gallstones. The patient is overweight which is a risk factor for gallstones. The patient has nausea and vomiting which could be due to gallstones. The patient has a dull pain in her back which could be due to gallstones. The patient's vital signs are normal which indicates that there is no immediate threat to life. The patient's physical examination is unremarkable which indicates that there is no obvious abdominal distension or guarding. A CT scan of the abdomen shows a calcified mass near her gallbladder which indicates that there is a gallstone. Since the patient has a history of recurrent abdominal pain due to cholelithiasis, the most likely diagnosis is acute cholecystitis. Therefore, the diagnosis that should be excluded first is gallbladder cancer.  
The answer is (B) gallbladder cancer

Answer is correct

### Eliminative/Codex Approach

**Question:** A 59-year-old overweight woman presents to the urgent care clinic with the complaint of severe abdominal pain for the past 2 hours. She also complains of a dull pain in her back with nausea and vomiting several times. Her pain has no relation with food. Her past medical history is significant for recurrent abdominal pain due to cholelithiasis. Her father died at the age of 60 with some form of abdominal cancer. Her temperature is 37°C (98.6°F), respirations are 15/min, pulse is 67/min, and blood pressure is 122/98 mm Hg. Physical exam is unremarkable. However, a CT scan of the abdomen shows a calcified mass near her gallbladder. Which of the following diagnoses should be excluded first in this patient?  
(A) Acute cholecystitis (B) Gallbladder cancer (C) Cholelithiasis (D) Pancreatitis

**Answer:** A: Let's think step-by-step. We refer to Wikipedia articles on medicine for help. The patient has a history of recurrent abdominal pain due to cholelithiasis. Her father died of abdominal cancer at the age of 60. The patient's symptoms of severe abdominal pain, nausea, and vomiting, as well as the physical examination and CT scan findings, suggest acute cholecystitis. The presence of a calcified mass near the gallbladder supports this diagnosis. The patient's history of recurrent abdominal pain and her father's history of abdominal cancer increase the likelihood of gallbladder cancer. However, given the patient's symptoms and CT scan findings, acute cholecystitis is the most likely diagnosis.  
The answer is (B) gallbladder cancer

Answer is correct

### (a) Responses produces by CODEX vs MEDCODEX on the MEDQAORIGINAL dataset

### Forward/MedCodex Approach



**Question:** A one-day-old male is evaluated in the hospital nursery for bilious vomiting. The patient has urinated multiple times since birth but has not yet passed meconium. He was born at 36 weeks gestation to a 36-year-old gravida 3 via vaginal delivery. The pregnancy was complicated by gestational diabetes. The patient's mother received routine prenatal care and declined first trimester screening. She subsequently underwent a quadruple screen at 15 weeks gestation that demonstrated decreased maternal serum alpha-fetoprotein (AFP), increased beta-hCG, decreased unconjugated estriol, and increased inhibin A. Her last ultrasound prior to onset of labor demonstrated an amniotic fluid index (AFI) of 28 cm. The patient's two older siblings are both healthy. The patient's temperature is 97.8°F (36.6°C), blood pressure is 58/37 mmHg, pulse is 166/min, and respirations are 38/min. On physical exam, the patient is in moderate distress. His abdomen is firm and distended with hypocoactive bowel sounds. What is the most likely etiology of this fetus's presentation?



**Answer:** Let's think step by step. The patient is one day old and has a history of bilious vomiting. The patient has urinated multiple times since birth but has not yet passed meconium. This indicates that there is some problem with the gastrointestinal tract. The patient was born at 36 weeks gestation which is slightly premature. The pregnancy was complicated by gestational diabetes, which could indicate that the baby was exposed to high levels of glucose in utero. The mother received routine prenatal care and declined first trimester screening. She subsequently underwent a quadruple screen at 15 weeks gestation that demonstrated decreased maternal serum alpha-fetoprotein (AFP), increased beta-hCG, decreased unconjugated estriol, and increased inhibin A. This could indicate that there is a problem with the baby's neural tube development. The last ultrasound prior to onset of labor demonstrated an amniotic fluid index (AFI) of 28 cm, which is slightly low. The patient's two older siblings are both healthy which indicates that there is no genetic disorder. The patient's temperature is 97.8°F (36.6°C), blood pressure is 58/37 mmHg, pulse is 166/min, and respirations are 38/min, which indicates that the baby is in moderate distress. On physical exam, the patient is in moderate distress. His abdomen is firm and distended with hypocoactive bowel sounds. This could indicate that there is a problem with the baby's intestines. Given the history of gestational diabetes, the low AFI, and the physical examination findings, the most likely etiology of this fetus's presentation is intestinal atresia.  
Answer: Intestinal atresia.

Answer is correct

### Eliminative/Codex Approach

**Question:** A one-day-old male is evaluated in the hospital nursery for bilious vomiting. The patient has urinated multiple times since birth but has not yet passed meconium. He was born at 36 weeks gestation to a 36-year-old gravida 3 via vaginal delivery. The pregnancy was complicated by gestational diabetes. The patient's mother received routine prenatal care and declined first trimester screening. She subsequently underwent a quadruple screen at 15 weeks gestation that demonstrated decreased maternal serum alpha-fetoprotein (AFP), increased beta-hCG, decreased unconjugated estriol, and increased inhibin A. Her last ultrasound prior to onset of labor demonstrated an amniotic fluid index (AFI) of 28 cm. The patient's two older siblings are both healthy. The patient's temperature is 97.8°F (36.6°C), blood pressure is 58/37 mmHg, pulse is 166/min, and respirations are 38/min. On physical exam, the patient is in moderate distress. His abdomen is firm and distended with hypocoactive bowel sounds. What is the most likely etiology of this fetus's presentation?

**Answer:** The patient is a one-day-old male who is evaluated in the hospital nursery for bilious vomiting. The patient has urinated multiple times since birth but has not yet passed meconium. The patient's mother had gestational diabetes during pregnancy. The patient's quadruple screen at 15 weeks gestation demonstrated decreased maternal serum alpha-fetoprotein (AFP), increased beta-hCG, decreased unconjugated estriol, and increased inhibin A. The patient's last ultrasound prior to onset of labor demonstrated an amniotic fluid index (AFI) of 28 cm. The patient's two older siblings are both healthy. The patient's temperature is 97.8°F (36.6°C), blood pressure is 58/37 mmHg, pulse is 166/min, and respirations are 38/min. On physical exam, the patient is in moderate distress. His abdomen is firm and distended with hypocoactive bowel sounds. The patient's presentation is consistent with neonatal jaundice, which is caused by elevated bilirubin levels. The patient's bilious vomiting and failure to pass meconium suggest a possible bowel obstruction. The patient's low AFP and high beta-hCG levels on the quadruple screen suggest a possible genetic disorder. The patient's firm and distended abdomen with hypocoactive bowel sounds suggest a possible mechanical obstruction. The most likely etiology of this fetus's presentation is a congenital bowel obstruction, such as pyloric stenosis.  
Answer: Congenital bowel obstruction (pyloric stenosis).

Answer is correct

### (b) Responses produces by Codex vs MedCodex on MEDQANOOPT dataset

Figure 3: Illustrative Example: Showcasing two different Prompting Strategies: CODEXFEWSHOTPROMPT and MEDCODEXFEWSHOTPROMPT, with their responses across two dataset variants: (a) MEDQAORIGINAL and (b) MEDQANOOPT. context of the answer and corresponding reasoning are highlighted.

to choose from; the clinician takes a medical history, forms the mental structure for differential diagnosis, performs examinations, adds or deletes the potential diagnosis based on contextual information, and then ultimately takes into consideration laboratory investigations to finally reach the diagnosis. The MEDCODEXFEWSHOTPROMPT mimics this strategy to reach a final answer.

We have used the above-mentioned prompting strategies to experiment with both MEDQAORIGINAL and MEDQANOOPT datasets, where the former contains options and the latter does not. Originally CODEXFEWSHOTPROMPT was designed by

Liévin et al. (2022) for an MCQ setting that had options. To make the prompt suitable for a no-opt setting, we modified it by removing options, option labels, and references. Similarly MEDCODEXFEWSHOTPROMPT, which we originally designed for no-opt questions, can be altered by simply adding options after questions to use in an MCQ setting.

In Figure 1, we present different classifications of prompting strategies based on their reasoning structure as well as their basis on the corresponding dataset (MCQ/ Descriptive Questions). In both prompt strategies, we use few shots  $K = 5$ . The human medical expertise evaluations were performed

by eight final year Medical School students.

### 4.3 CODEXFEWSHOTPROMPTvs MEDCODEXFEWSHOTPROMPTon MEDQAORIGINAL dataset

**Experimental Setup:** We first utilize MEDCODEXFEWSHOTPROMPT due to its underlying prospective incremental reasoning structure, which typically follows the line of reasoning employed by medical professionals (Appendix A). Along with this, we compare responses collected using CODEXFEWSHOTPROMPT where the reasoning is more eliminative in nature, is based on seeing the options, and less attention is provided towards the clinical flow of argument (The responses tend to form the entire clinical context in one go, *c.f.*, Figure 3). The Llama-2-7B-Chat and 70B-Chat models were prompted using both strategies.

**Evaluation Strategy:** The results were produced by parsing the model-generated outputs, and compared with the gold truth answers. The evaluation results of this experiment on the MEDQAORIGINAL’s 1273 questions from the Test set are described in Table 2

**Observations:**

Methods	Accuracy Scores
Llama-2-7B-Chat (CODEXFEWSHOTPROMPT)	41%
Llama-2-7B-Chat (MEDCODEXFEWSHOTPROMPT)	40.6%
Llama-2-70B-Chat (CODEXFEWSHOTPROMPT)	52%
Llama-2-70B-Chat (MEDCODEXFEWSHOTPROMPT)	50%

Table 2: CODEXFEWSHOTPROMPTvs. MEDCODEXFEWSHOTPROMPTon Llama-2 7B-Chat and 70B-Chat models

### 4.4 CODEXFEWSHOTPROMPT vs. MEDCODEXFEWSHOTPROMPT on MEDQANOOPT and CLINICIANCASESTUDY

**Experimental Setup:** This experiment was performed to determine the ability of both the prompting strategies to handle open-ended medical questions which are similar to in-clinic scenarios. This is done by comparing the results of both prompting strategies by generating outputs with Llama-2-7B-chat and 70B-chat models.

**Evaluation Strategy:** We select 500 questions from MEDQANOOPT and 25 questions from

CLINICIANCASESTUDY for this. We evaluate, via medical experts, the final reasoning quality and final answer on a 3-point Likert scale (Batterton and Hale, 2017) Agree, Neutral, Disagree.

**Observations:** The MEDCODEXFEWSHOTPROMPT achieves a score of 83% and 87% respectively from Llama-2-7B-chat and 70B-chat models respectively with MEDQANOOPT against 56% and 84% from CODEXFEWSHOTPROMPT, (Figures 6 and 7). While MEDCODEXFEWSHOTPROMPT achieves 88% and 82% with Llama-2-7B-chat and 70B-chat with CLINICIANCASESTUDY dataset, CODEXFEWSHOTPROMPT achieves 90% and 86% respectively (Figures 8 and 9).

### 4.5 MedCodex diagnoses generation with Codex Pruning (MedCodex-Codex) on MEDQANOOPT and CLINICIANCASESTUDY

**Experimental Setup:** Each question  $q_i$  in the MEDQANOOPT dataset lacks options, therefore pertaining to a real-world problem setting. We particularly utilize the MEDCODEXFEWSHOTPROMPT to sample a number of options  $\mathcal{O}_i^{gen}$  for each question, thereby mimicking the MEDQAORIGINAL dataset format. The question  $q_i$  followed by generated options  $\mathcal{O}_i^{gen}$  are fed into the model  $\mathcal{LM}$  with the CODEXFEWSHOTPROMPT to select the most appropriate option. To create a set of 4 options through sampling, 10 unique options are sampled by filtering on a word-level match basis (to avoid selecting options that are repetitive). Out of these, the top 4 are selected through the perplexity of model outputs.

**Evaluation Methods:** We evaluate via medical expert the final reasoning quality and final answer on a 3-point Likert scale(Batterton and Hale, 2017) Agree, Neutral, Disagree. Both datasets MEDQANOOPT (500Qs) and CLINICIANCASESTUDY (25Qs) are evaluated).

**Observations:** The medical experts agreed with 81% and 87% of the questions from MEDQANOOPT evaluated with Llama-2-7B-chat and 70B-chat respectively (MEDCODEX-CODEX approach - *c.f.*, Figures 6 and 7), while one the CLINICIANCASESTUDY dataset, the corresponding numbers were 88% and 84% (Figures 8 and 9).

### 4.6 Development of Verifier

#### Motivation for Verifier and its training

So far, we have approached the problem with an

in-context learning perspective without changing the model parameters. We believe building on top of powerful models like the Llama-2 chat series could give us a significant advantage in this scenario. We endeavored to improve performance by substituting the CODEXFEWSHOTPROMPT prompt with a verifier Reward learning model.

### Verifier dataset creation

We borrow the Question and Answer parts of the triplet (question, reasoning, answer) denoted by  $\langle q_i; \mathcal{R}_i; \mathcal{A}_i \rangle$ , from the MEDQAORIGINAL dataset and generate the Reasoning for each option (irrespective of its being correct or incorrect) by prompting the Llama-2-70b-chat model (Figure 4). We iterate over the 4-options for a question provided by the MEDQAORIGINAL dataset and generate reasoning for the three incorrect and one correct option separately. We then pair the three "incorrect" reasoning-answer's with one correct generated reasoning answer to create three sets of "chosen-rejected" pairs of reasoning-answer units, each of which contributes to a data-point for training the verifier  $\mathcal{D}_{ver}$ . A representative sample was selected to evaluate the quality of reasoning generated for each option. Multiple medical experts independently assessed and authenticated the correctness of responses generated by the Llama-2-70B-Chat model.

Below is a question from the field of professional medicine, where the correct answer is given. Explain with concordant medical reasoning how the answer is reached. The reasoning should be in a single paragraph and strictly within 200 words, and should smoothly end with the phrase "Thus, the answer is <Given Answer>."

Question: A 3-month-old baby died suddenly at night while asleep. His mother noticed that he had died only after she awoke in the morning. No cause of death was determined based on the autopsy. What precautionary measures could have prevented the death of the baby?

Answer: Placing the infant in a supine position on a firm mattress while sleeping

Reasoning:

Figure 4: Prompt to generate the reasoning for Verifier dataset

### Training

We have used the Llama-2-7B-chat and 70B-chat models separately followed by a linear head as the architecture for training the reward model. We apply Low-Ranked Adaptation (LoRA) for fine-tuning using the *trl* library from *Hugging Face* for this purpose. The model is trained on a reward modeling loss from (Wang et al., 2024). The training parameters are as follows : For 7B-chat model, we used a learning rate of  $5e-5$ , with a *batch\_size* = 2 and *gradient\_accumulation\_steps* = 16. Optimizer used was *AdamW*. The LoRA parameters

were :  $r = 16$ , *lora\_alpha* = 16. We used 1 Nvidia A100 80GB GPU for training, which took around 72 hrs. For 70B-chat model, we had to use 4 Nvidia A100 80GB GPUs, and performed the training for 72 hrs.

$\mathcal{L}^{rw}(\bullet; \phi) = -\log(\sigma(r_\phi(q_i, o_i^c) - r_\phi(q_i, o_i^r) - m(r)))$

Where  $q_i$  is the input question string,  $o_i^c$  is the chosen output (in our case the correct output),  $o_i^r$  is the rejected output,  $m(r)$  denotes the reward margin. In our case, reward margin = 0.

### MedCodex diagnoses generation with Verification (MedCodex-Reward) on MEDQANOOPT and CLINICIANCASESTUDY

**Experimental Setup:** We sample forward options as in the previous experiment (4.5), then pass the generated options with their reasoning, as shown in Figure 5 figure, to the verifier. We choose the option that produces the highest scalar reward.

**Evaluation Methods:** We select 500 questions from MEDQANOOPT and evaluate via medical experts the final reasoning quality and final answer on a 3-point Likert scale (Batterton and Hale, 2017) Agree, Neutral, Disagree.

**Observations:** Both Llama-2-7B-chat and 70B-chat achieved scores of 87% with this method on MEDQANOOPT dataset (MedCodex-Reward Approach, Figures 6 and 7), while the CLINICIAN-CASESTUDY dataset evaluations reported scores of 90% and 82% respectively (Figures 8 and 9).

Question : A 3-month-old baby died suddenly at night while asleep. His mother noticed that he had died only after she awoke in the morning. No cause of death was determined based on the autopsy. What precautionary measures could have prevented the death of the baby?

Reasoning : Let's think step by step. Sudden Infant Death Syndrome (SIDS) is a leading cause of death in infants under one year old. SIDS is associated with sleeping in the prone or side position, which can cause the infant's face to be pressed against the mattress or bedding, obstructing their airway. Placing the infant in a supine position on a firm mattress reduces the risk of rebreathing expired air and decreases the likelihood of the infant's face being pressed against the mattress, reducing the risk of SIDS. Additionally, using a firm mattress can help reduce the risk of the infant getting tangled in bedding or becoming trapped between the mattress and a wall. Thus, the answer is placing the infant in a supine position on a firm mattress while sleeping.

Answer : Placing the infant in a supine position on a firm mattress while sleeping

Figure 5: Input prompt to the Verifier

### Interpreting the Experiment Results

From medical expert evaluations of Llama-2-7B-chat on MEDQANOOPT (Figure 6), we see that all methods except CODEXFEWSHOTPROMPT perform very well, with MedCodex-Reward Model setting giving the highest performance 87% and the lowest disagreement 6.9% in a lower resource setting such as with a Llama-2-7B-chat model. CODEXFEWSHOTPROMPT shows comparatively poorer results 56% with a high level of disagreements as well 30%, while MEDCODEXFEWSHOT-

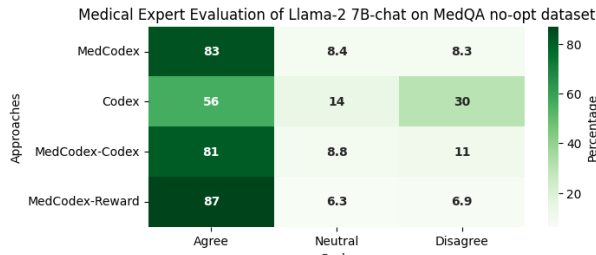


Figure 6: Results of experiments on MEDQANOOPT dataset with Llama-2-7B-chat model

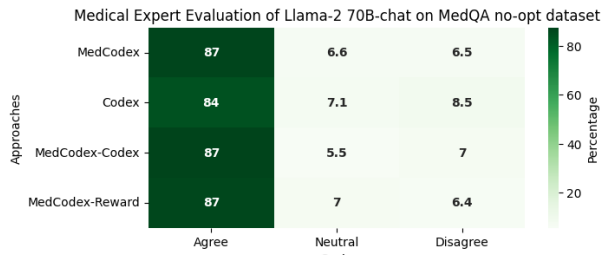


Figure 7: Results of experiments on MEDQANOOPT dataset with Llama-2-70B-chat model

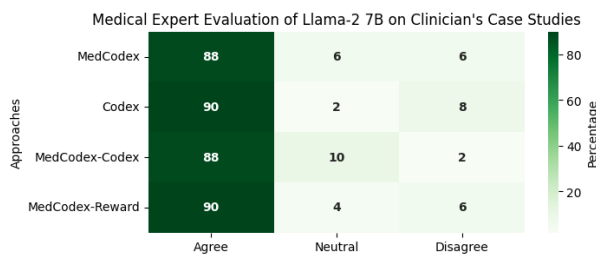


Figure 8: Results of experiments on CLINICIANCAS-ESTUDY with Llama-2-7B-chat model

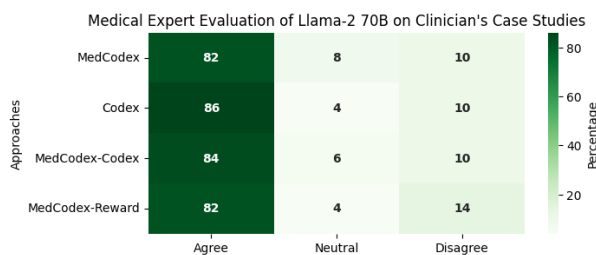


Figure 9: Results of experiments on CLINICIANCAS-ESTUDY with Llama-2-70B-chat model

PROMPT performs on par with the other methods 83%. We attribute this to the CODEXFEWSHOT-PROMPT lacking the ability to extract information from the model parameters as effectively as the others due to its eliminative nature and an absence of options in the open-ended-QA setting. MED-CODEXFEWSHOT-PROMPT, on the other hand, is

based on an incremental reasoning structure that is designed to draw out inferences from the context and reaches a differential diagnosis. We see this occurrence disappearing in the evaluations of Llama-2-70B-chat for the MEDQANOOPT dataset. Here all the methods including CODEXFEWSHOT-PROMPT, score really high on the experts' metrics, and are able to extract the information and reasoning quite effectively from the large knowledge pool of the 70B-chat model.

With the evaluations of the CLINICIANCAS-ESTUDY dataset on Llama-2-7B-chat model, we see all the methods performing very well with CODEXFEWSHOT-PROMPT showing slightly higher disagreements than others with 8%. The high performance of all methods is maintained across the 70B-chat model as well, where slightly higher disagreement is displayed by the MedCodex-Reward model method.

## 5 Conclusion and Future work

The incremental reasoning chain of thought prompting is a novel prompting methodology developed by us that follows the usual clinical approach of deciding on real-life clinical settings. We demonstrate that this strategy gives significantly better results than the CODEX prompting strategy, which is designed for MCQ-type questions. Further, we demonstrate that the verifier developed using reasoning performs much better at selecting agreeable responses from the Llama-2 models. Future research will explore the generalizability of this approach by testing on other open-source LLMs.

So far our research focuses on improving the quality of responses through in-context learning and use of the models' parametric knowledge of medicine. We could further aim to integrate these methods with non-parametric knowledge sources, to boost the range of knowledge available, and improve the confidence within the output. For this methods like Retrieval Augmented Generation (RAG), and Knowledge Graph integration could be explored.

## 6 Limitations:

This paper does not train an LLM from scratch and only leverages pre-trained models. The quality of the model response depends on the quality of the forward-looking prompt. The current process has been only demonstrated on Llama-2 models and needs to be tested on other models to demonstrate generalizability.



## 7 Ethics Statement

The proposed research has been designed with a strong emphasis on ethical considerations. This study utilizes a de-identified medical question and answer dataset (MEDQANOOPT). No human subjects were involved. All the methods were evaluated for their performance on open-source datasets, but did not interact with actual patients for diagnosis. The use of MEDCODEXFEWSHOTPROMPT, a CoT-driven prompt, is designed to mirror the incremental reasoning process in medical diagnosis, ensuring that the decision-making process is transparent and explainable. The utilization of a reward model mechanism for response verification ensures that the system is held accountable for its decisions, further emphasizing the importance of ethical considerations in AI Research, especially in medicine. Future work involving real-world clinical applications would require a separate ethics review and informed consent procedures.

## Acknowledgements

## References

- Katherine A Batterton and Kimberly N Hale. 2017. The likert scale what it is and how to use it. *Phalanx*, 50(2):32–39.
- E Bolton et al. 2022. Pubmedgpt 2.7 b. Technical report, Technical report. Stanford University Center for Research on Foundation . . .
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141.
- Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models. *arXiv preprint arXiv:2209.15003*.

- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.
- Munib Mesinovic, Peter Watkinson, and Tingting Zhu. 2023. Explainable ai for clinical risk prediction: a survey of concepts, methods, and modalities. *arXiv preprint arXiv:2308.08407*.
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsim-poukelli, Nikolai Grigorev, Doug Fritz, Thibault Sotiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy,

636 Chris Jones, James Bradbury, Matthew Johnson,  
637 Blake Hechtman, Laura Weidinger, Iason Gabriel,  
638 William Isaac, Ed Lockhart, Simon Osindero, Laura  
639 Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub,  
640 Jeff Stanway, Lorraine Bennett, Demis Hassabis, Ko-  
641 ray Kavukcuoglu, and Geoffrey Irving. 2022. [Scaling](#)  
642 [language models: Methods, analysis insights from](#)  
643 [training gopher](#).

644 Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mah-  
645 davi, Jason Wei, Hyung Won Chung, Nathan Scales,  
646 Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl,  
647 et al. 2022. Large language models encode clinical  
648 knowledge. *arXiv preprint arXiv:2212.13138*.

649 Arun James Thirunavukarasu, Darren Shu Jeng Ting,  
650 Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan,  
651 and Daniel Shu Wei Ting. 2023. Large language  
652 models in medicine. *Nature medicine*, 29(8):1930–  
653 1940.

654 Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-  
655 bert, Amjad Almahairi, Yasmine Babaei, Nikolay  
656 Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti  
657 Bhosale, et al. 2023. Llama 2: Open founda-  
658 tion and fine-tuned chat models. *arXiv preprint*  
659 *arXiv:2307.09288*.

660 A Venigalla, J Frankle, and M Carbin. 2022. Biomedlm:  
661 a domain-specific large language model for biomed-  
662 ical text. *MosaicML. Accessed: Dec, 23(3):2*.

663 Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan  
664 Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu  
665 Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao  
666 Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao  
667 Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen,  
668 Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang,  
669 Zuxuan Wu, and Yu-Gang Jiang. 2024. [Secrets of](#)  
670 [rlhf in large language models part ii: Reward model-](#)  
671 [ing](#).

672 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten  
673 Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,  
674 et al. 2022. Chain-of-thought prompting elicits rea-  
675 soning in large language models. *Advances in Neural*  
676 *Information Processing Systems*, 35:24824–24837.

677 Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su.  
678 2018. Medical exam question answering with large-  
679 scale reading comprehension. In *Proceedings of the*  
680 *AAAI conference on artificial intelligence*, volume 32.

681 Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and  
682 Ziwei Liu. 2022. Learning to prompt for vision-  
683 language models. *International Journal of Computer*  
684 *Vision*, 130(9):2337–2348.

## A Appendix 1: Prompts used in the research

686

This section provides the 5-shot prompts that we used for MedCodex and Codex strategies on original and no-opt datasets

687

688

### A.1 MedCodex prompt for dataset with options

689

The following are multiple choice questions (with answers) about professional medicine.

690

691

Q: A 4670-g (10-lb 5-oz) male newborn is delivered at term to a 26-year-old woman after prolonged labor. Apgar scores are 9 and 9 at 1 and 5 minutes. Examination in the delivery room shows swelling, tenderness, and crepitus over the left clavicle. There is decreased movement of the left upper extremity. Movement of the hands and wrists are normal. A grasping reflex is normal in both hands. An asymmetric Moro reflex is present. The remainder of the examination shows no abnormalities and an anteroposterior x-ray confirms the diagnosis. Which of the following is the most appropriate next step in management? (A) Physical therapy (B) Pin sleeve to the shirt (C) Splinting of the arm (D) MRI of the clavicle

692

693

694

695

696

697

698

699

A: Let's think step-by-step. Prolonged labour would mean it would be a difficult child birth. Apgar score is 9 and does not drop after repeated examination which means that there is no respiratory or neurological problem. Clavicular crepitus, tenderness and swelling indicates that there could be clavicular injury since this is common in difficult child birth. Decreased movement in the upper extremity supports some injury indication in the upper limb. Hand and wrist movement being normal as well as grasping movement being normal means that there is no neurological deficit and the problem lies with the arm or shoulder and not the forearm or wrist. Grasping reflex in both hands again indicates that there is no deficit between communication from the spinal cord. An asymmetric Moros reflex indicates some local injury which is due to either damage to a peripheral nerve, cervical cord or fracture of the clavicle. However, presence of clavicular crepitus and tenderness indicates towards clavicular fracture and not other conditions. So, the most probable diagnosis is clavicular fracture. Since Clavicular fractures heal very quickly in newborns, hence only restraining is necessary of the injured arm. Thus, the first line of treatment is pinning sleeve to the shirt. The answer is (B).

700

701

702

703

704

705

706

707

708

709

710

711

712

713

Q: A previously healthy 32-year-old man is brought to the emergency department by his girlfriend after having a seizure. Earlier that day, he also experienced a nosebleed that took 30 minutes to stop when applying pressure. He has had no sick contacts or history of epilepsy or other seizure disorder. He does not take any medications. His temperature is 39.1 °C (102.4 °F), pulse is 106/min, respirations are 26/min, and blood pressure is 128/70 mm Hg. He is confused and disoriented. Examination shows pallor and scattered petechiae over the trunk and arms. The neck is supple, and neurological examination is otherwise within normal limits. Laboratory studies show: Hemoglobin 9 g/dL Leukocyte count 8,200/mm<sup>3</sup> Platelet count 34,000/mm<sup>3</sup> Prothrombin time 13 seconds Partial thromboplastin time 30 seconds Fibrin split products negative Serum Creatinine 2.9 mg/dL Bilirubin Total 3.2 mg/dL Direct 0.4 mg/dL Lactate dehydrogenase 559 U/L A peripheral blood smear shows numerous schistocytes. Which of the following is the most appropriate next step in management?" (A) Transfusion of packed red blood cells (B) Plasma exchange therapy (C) Platelet transfusion (D) Intravenous tranexamic acid

714

715

716

717

718

719

720

721

722

723

724

725

726

A: Let's think step-by-step. Patient arrives with seizures but has no previous history of seizures, so previous neurological problems are less likely. Presence of nose bleed that did not stop for 30 mins indicates that there can be a clotting problem which could be either due to platelet insufficiency or clotting factor problems. There is fever along with seizures which could indicate meningitis. However, neck being supple, meaning not rigid goes against meningitis. Patient is not on any medication which rules out involuntary overdose. Pulse is slightly elevated which indicates that there is hemodynamic instability. Patient is confused and disoriented which could be post seizure effect. There is history of nose bleed as

727

728

729

730

731

732

733

734 well as petechiae all over the body indicates a problem of clotting which is supported by low platelet  
735 from the lab tests. There is fever but leucocyte count is normal which indicates against viral or bacterial  
736 infection. The creatinine level is elevated which indicates that the kidneys are being affected. Bilirubin  
737 level is elevated, but direct bilirubin is normal indicating that there is excessive bilirubin production  
738 potentially due to increased haemolysis which is supported by low Haemoglobin level. Presence of  
739 haemolysis is supported by schistocytes observed in peripheral blood smear. However, fragmented blood  
740 cells are seen and platelet count is low. All of this points towards a diagnosis of thrombocytopenic  
741 purpura. Hence, plasma exchange therapy is the correct answer. The answer is (B).  
742

743 Q: A 50-year-old man comes to the physician because of a 6-month history of difficulties having sexual  
744 intercourse due to erectile dysfunction. He has type 2 diabetes mellitus that is well controlled with  
745 metformin. He does not smoke. He drinks 5–6 beers daily. His vital signs are within normal limits.  
746 Physical examination shows bilateral pedal edema, decreased testicular volume, and increased breast  
747 tissue. The spleen is palpable 2 cm below the left costal margin. Abdominal ultrasound shows an atrophic,  
748 hyperechoic, nodular liver. An upper endoscopy is performed and shows dilated submucosal veins 2 mm  
749 in diameter with red spots on their surface in the distal esophagus. Therapy with a sildenafil is initiated  
750 for his erectile dysfunction. Which of the following is the most appropriate next step in management  
751 of this patient's esophageal findings? (A) Injection sclerotherapy (B) Nadolol therapy (C) Isosorbide  
752 mononitrate therapy (D) Transjugular intrahepatic portosystemic shunt  
753

754 A: Let's think step-by-step. The patient has a history of erectile dysfunction during sexual intercourse.  
755 The patient also has a history of Diabetes but it is well controlled with metformin therapy. However, the  
756 patient has a long history of alcohol abuse as indicated by 5 beers daily. There is decreased testicular  
757 volume and increased breast tissue which could indicate androgenic insufficiency. There is also pedal  
758 oedema which can be due to uncontrolled blood pressure or due to liver dysfunction. The spleen is  
759 enlarged as indicated by being palpable 2cm below left costal margin which indicates increased portal  
760 pressure. Abdominal ultrasonography indicates pressure of alcoholic liver cirrhosis as indicated by  
761 atrophic nodular liver. The enlarged spleen can then be potentially due to portal hypertension secondary  
762 to liver cirrhosis. Liver cirrhosis also causes decreased protein synthesis which causes pedal oedema due  
763 to decreased blood osmotic pressure. Liver cirrhosis also causes increased estrogen to testosterone ratio  
764 which results in the physical manifestation of decreased testicular volume and increased breast tissue as  
765 well as erectile dysfunction. Endoscopy shows dilated submucosal veins 2 mm in diameter with red spots  
766 on their surface in the distal esophagus which again indicates portal hypertension. The diagnosis of this  
767 condition is portal hypertension secondary to liver cirrhosis due to chronic alcoholism. Primary therapy to  
768 treat portal hypertension here is to decrease portal pressure which is done by administering beta blockers.  
769 So, Nadolol therapy should be initiated. The answer is (B).  
770

771 Q: A 24-year-old primigravid woman at 38 weeks' gestation comes to the physician for a prenatal visit.  
772 At the last two prenatal visits, transabdominal ultrasound showed the fetus in breech presentation. She has  
773 no medical conditions and only takes prenatal vitamins. Her pulse is 95/min, respirations are 16/min, and  
774 blood pressure is 130/76 mm Hg. The abdomen is soft and nontender; no contractions are felt. Pelvic exam-  
775 ination shows a closed cervical os and a uterus consistent with 38 weeks' gestation. The fetal rate tracing  
776 shows a baseline heart rate of 152/min and 2 accelerations over 10 minutes. Repeat ultrasound today shows  
777 a persistent breech presentation. The patient states that she would like to attempt a vaginal delivery. Which  
778 of the following is the most appropriate next step in management? (A) Recommend cesarean section  
779 (B) Observe until spontaneous labor (C) Repeat ultrasound in one week (D) Offer external cephalic version  
780

781 A: Let's think step-by-step. This is the first pregnancy of the woman and she is close to term as indicated  
782 by 38 weeks of gestation. Previous two abdominal USG's indicated a breech presentation of the foetus.  
783 The vitals are normal, i.e pulse and respiratory rate is normal and blood pressure is normal indicating not  
784 preeclampsia so no immediate threat to the mother's health and there are no contractions indicating that  
785 while the mother is close to term she is not in labour. Also, the abdomen is soft and tender and there are

no contractions which indicates that the woman is not in labour. Also, the pelvic examination shows that the cervical os is closed again indicating that the mother is not in labour. The foetal heart rate is 152/min and 2 accelerations over 10 mins again indicating that mother is not in labour and the foetus is healthy. An ultrasound done today confirms the breech presentation of the foetus. However, since the woman is close to term it is unlikely that the fetus will turn by itself and will continue to stay in breech position. While a C-section is an alternative in this situation, since the mother has indicated that she would want to attempt a normal vaginal delivery, an external cephalic version can be offered to the woman. The answer is (D).

Q: A 4-year old boy is brought to the emergency department with fever, painful swallowing, headache, and neck spasm that began shortly after waking up. He has had a sore throat over the last week that acutely worsened this morning. He has no history of serious illness and takes no medications. He lives at home with his mother. His older brother has asthma. His immunizations are up-to-date. He appears acutely ill. His temperature is 38.4°C (101.2°F), pulse is 95/min, respirations are 33/min, and blood pressure is 93/60 mm Hg. Examination shows drooling. The neck is stiff and extension is limited. Respirations appear labored with accessory muscle use. Inspiratory stridor is heard on auscultation of the chest. Cardiac examination shows no abnormalities. Oropharyngeal examination shows a bulge in the posterior pharyngeal wall. Intravenous access is obtained and laboratory studies are ordered. Which of the following is the most appropriate next step in the management of this patient? (A) Endotracheal intubation (B) IV antibiotics (C) Blood cultures (D) IV corticosteroids

A: Let's think step-by-step. The 4-year-old boy has fever 38.4 C, painful swallowing and headache which indicates upper respiratory tract infection. There is neck spasm and fever which could indicate meningitis. Sore throat presence again indicates upper respiratory tract infection which is worsening since this morning. No history of medication indicates that there is less likelihood of accidental overdose. Brother has a history of asthma which increases his own risk of autoimmune conditions. Neck is stiff, there is drooling which could indicate neurological affliction which points towards meningitis. This could be meningitis since meningococcal vaccine is given around 10-12 years. Drooling could also indicate extreme difficulty in swallowing. Respiration is laboured with use of accessory muscle indicates extreme respiratory distress. No cardiac abnormalities so, respiratory distress is primarily respiratory in origin. There is a huge bulge in the posterior pharyngeal wall which could be of acute origin since the sore throat is acutely worsened since morning. This could be either meningitis secondary to upper respiratory tract infection but the first step is to secure the airway since there is severe respiratory distress. So endotracheal intubation should be done immediately. The answer is (A).

## A.2 MedCodex prompt for no-opt dataset

Use just the given patient history to answer the question. Do not assume any further information about the patient. Strictly Limit your response to 200 words.

Q: A 4670-g (10-lb 5-oz) male newborn is delivered at term to a 26-year-old woman after prolonged labor. Apgar scores are 9 and 9 at 1 and 5 minutes. Examination in the delivery room shows swelling, tenderness, and crepitus over the left clavicle. There is decreased movement of the left upper extremity. Movement of the hands and wrists are normal. A grasping reflex is normal in both hands. An asymmetric Moro reflex is present. The remainder of the examination shows no abnormalities and an anteroposterior x-ray confirms the diagnosis. What is the most suitable course of action to manage this case?

A: Let's think step-by-step. Prolonged labour would mean it would be a difficult child birth. Apgar score is 9 and does not drop after repeated examination which means that there is no respiratory or neurological problem. Clavicular crepitus, tenderness and swelling indicates that there could be clavicular injury since this is common in difficult child birth. Decreased movement in the upper extremity supports some injury indication in the upper limb. Hand and wrist movement being normal as well as grasping movement being normal means that there is no neurological deficit and the problem lies

837 with the arm or shoulder and not the forearm or wrist. Grasping reflex in both hands again indicates  
838 that there is no deficit between communication from the spinal cord. An asymmetric Moro reflex  
839 indicates some local injury which is due to either damage to a peripheral nerve, cervical cord or  
840 fracture of the clavicle. However, presence of clavicular crepitus and tenderness indicates towards  
841 clavicular fracture and not other conditions. So, the most probable diagnosis is clavicular fracture.  
842 Since clavicular fractures heal very quickly in newborns, hence only restraining is necessary of the in-  
843 jured arm. Thus, the first line of treatment is pinning sleeve to the shirt. Answer: pinning sleeve to the shirt.  
844

845 Use just the given patient history to answer the question. Do not assume any further information about the  
846 patient. Strictly Limit your response to 200 words.  
847

848 Q: A previously healthy 32-year-old man is brought to the emergency department by his girlfriend after  
849 having a seizure. Earlier that day, he also experienced a nosebleed that took 30 minutes to stop when  
850 applying pressure. He has had no sick contacts or history of epilepsy or other seizure disorder. He  
851 does not take any medications. His temperature is 39.1 °C (102.4 °F), pulse is 106/min, respirations  
852 are 26/min, and blood pressure is 128/70 mm Hg. He is confused and disoriented. Examination shows  
853 pallor and scattered petechiae over the trunk and arms. The neck is supple, and neurological examination  
854 is otherwise within normal limits. Laboratory studies show: Hemoglobin 9 g/dL Leukocyte count  
855 8,200/mm<sup>3</sup> Platelet count 34,000/mm<sup>3</sup> Prothrombin time 13 seconds Partial thromboplastin time 30  
856 seconds Fibrin split products negative Serum Creatinine 2.9 mg/dL Bilirubin Total 3.2 mg/dL Direct 0.4  
857 mg/dL Lactate dehydrogenase 559 U/L A peripheral blood smear shows numerous schistocytes. What is  
858 the most appropriate next step in the management of this patient?  
859

860 A: Let's think step-by-step. Patient arrives with seizures but has no previous history of seizures, so  
861 previous neurological problems are less likely. Presence of nose bleed that did not stop for 30 mins  
862 indicates that there can be a clotting problem which could be either due to platelet insufficiency or clotting  
863 factor problems. There is fever along with seizures which could indicate meningitis. However, neck being  
864 supple, meaning not rigid goes against meningitis. Patient is not on any medication which rules out  
865 involuntary overdose. Pulse is slightly elevated which indicates that there is hemodynamic instability.  
866 Patient is confused and disoriented which could be post seizure effect. There is history of nose bleed as  
867 well as petechiae all over the body indicates a problem of clotting which is supported by low platelet  
868 from the lab tests. There is fever but leucocyte count is normal which indicates against viral or bacterial  
869 infection. The creatinine level is elevated which indicates that the kidneys are being affected. Bilirubin  
870 level is elevated, but direct bilirubin is normal indicating that there is excessive bilirubin production  
871 potentially due to increased haemolysis which is supported by low Haemoglobin level. Presence of  
872 haemolysis is supported by schistocytes observed in peripheral blood smear. However, fragmented blood  
873 cells are seen and platelet count is low. All of this points towards a diagnosis of thrombocytopenic  
874 purpura. Hence, plasma exchange therapy is the correct answer. Answer: plasma exchange therapy.  
875

876 Use just the given patient history to answer the question. Do not assume any further information about the  
877 patient. Strictly Limit your response to 200 words.  
878

879 Q: A 67-year-old man comes to the emergency department for the evaluation of two episodes of red urine  
880 since this morning. He has no pain with urination. He reports lower back pain and fever. Six months  
881 ago, he was diagnosed with osteoarthritis of the right knee that he manages with 1–2 tablets of ibuprofen  
882 per day. He has smoked one pack of cigarettes daily for the past 45 years. He does not drink alcohol.  
883 His temperature is 38.5°C (101.3°F), pulse is 95/min, and blood pressure is 130/80 mm Hg. Physical  
884 examination shows faint, diffuse maculopapular rash, and bilateral flank pain. The remainder of the  
885 examination shows no abnormalities. What is the most likely diagnosis for this patient?  
886

887 A: Let's think step-by-step. The patient has a history of erectile dysfunction during sexual intercourse.  
888 The patient also has a history of Diabetes but it is well controlled with metformin therapy. However, the

patient has a long history of alcohol abuse as indicated by 5 beers daily. There is decreased testicular volume and increased breast tissue which could indicate androgenic insufficiency. There is also pedal oedema which can be due to uncontrolled blood pressure or due to liver dysfunction. The spleen is enlarged as indicated by being palpable 2cm below left costal margin which indicates increased portal pressure. Abdominal ultrasonography indicates pressure of alcoholic liver cirrhosis as indicated by atrophic nodular liver. The enlarged spleen can then be potentially due to portal hypertension secondary to liver cirrhosis. Liver cirrhosis also causes decreased protein synthesis which causes pedal oedema due to decreased blood osmotic pressure. Liver cirrhosis also causes increased estrogen to testosterone ratio which results in the physical manifestation of decreased testicular volume and increased breast tissue as well as erectile dysfunction. Endoscopy shows dilated submucosal veins 2 mm in diameter with red spots on their surface in the distal esophagus which again indicates portal hypertension. The diagnosis of this condition is portal hypertension secondary to liver cirrhosis due to chronic alcoholism. Primary therapy to treat portal hypertension here is to decrease portal pressure which is done by administering beta blockers. So, Nadolol therapy should be initiated. Answer: Nadolol Therapy.

Use just the given patient history to answer the question. Do not assume any further information about the patient. Strictly Limit your response to 200 words.

Q: A 24-year-old primigravid woman at 38 weeks' gestation comes to the physician for a prenatal visit. At the last two prenatal visits, transabdominal ultrasound showed the fetus in breech presentation. She has no medical conditions and only takes prenatal vitamins. Her pulse is 95/min, respirations are 16/min, and blood pressure is 130/76 mm Hg. The abdomen is soft and nontender; no contractions are felt. Pelvic examination shows a closed cervical os and a uterus consistent with 38 weeks' gestation. The fetal rate tracing shows a baseline heart rate of 152/min and 2 accelerations over 10 minutes. Repeat ultrasound today shows a persistent breech presentation. The patient states that she would like to attempt a vaginal delivery. Describe the next best step for managing this patient.

A: Let's think step-by-step. This is the first pregnancy of the woman and she is close to term as indicated by 38 weeks of gestation. Previous two abdominal USG's indicated a breech presentation of the foetus. The vitals are normal, i.e pulse and respiratory rate is normal and blood pressure is normal indicating not preeclampsia so no immediate threat to the mother's health and there are no contractions indicating that while the mother is close to term she is not in labour. Also, the abdomen is soft and tender and there are no contractions which indicates that the woman is not in labour. Also, the pelvic examination shows that the cervical os is closed again indicating that the mother is not in labour. The foetal heart rate is 152/min and 2 accelerations over 10 mins again indicating that mother is not in labour and the foetus is healthy. An ultrasound done today confirms the breech presentation of the foetus. However, since the woman is close to term it is unlikely that the foetus will turn by itself and will continue to stay in breech position. While a C-section is an alternative in this situation, since the mother has indicated that she would want to attempt a normal vaginal delivery, an external cephalic version can be offered to the woman. Answer: offering external cephalic version.

Use just the given patient history to answer the question. Do not assume any further information about the patient. Strictly Limit your response to 200 words.

Q: A 4-year old boy is brought to the emergency department with fever, painful swallowing, headache, and neck spasm that began shortly after waking up. He has had a sore throat over the last week that acutely worsened this morning. He has no history of serious illness and takes no medications. He lives at home with his mother. His older brother has asthma. His immunizations are up-to-date. He appears acutely ill. His temperature is 38.4°C (101.2°F), pulse is 95/min, respirations are 33/min, and blood pressure is 93/60 mm Hg. Examination shows drooling. The neck is stiff and extension is limited. Respirations appear labored with accessory muscle use. Inspiratory stridor is heard on auscultation of the chest. Cardiac examination shows no abnormalities. Oropharyngeal examination shows a bulge in the

941 posterior pharyngeal wall. Intravenous access is obtained and laboratory studies are ordered. Explain the  
942 most suitable course of action for managing this patient's condition.

943  
944 A: Let's think step-by-step. The 4-year-old boy has fever 38.4 C, painful swallowing and headache  
945 which indicates upper respiratory tract infection. There is neck spasm and fever which could indicate  
946 meningitis. Sore throat presence again indicates upper respiratory tract infection which is worsening  
947 since this morning. No history of medication indicates that there is less likelihood of accidental overdose.  
948 Brother has a history of asthma which increases his own risk of autoimmune conditions. Neck is stiff,  
949 there is drooling which could indicate neurological affliction which points towards meningitis. This  
950 could be meningitis since meningococcal vaccine is given around 10-12 years. Drooling could also  
951 indicate extreme difficulty in swallowing. Respiration is laboured with use of accessory muscle indicates  
952 extreme respiratory distress. No cardiac abnormalities so, respiratory distress is primarily respiratory  
953 in origin. There is a huge bulge in the posterior pharyngeal wall which could be of acute origin since  
954 the sore throat is acutely worsened since morning. This could be either meningitis secondary to upper  
955 respiratory tract infection but the first step is to secure the airway since there is severe respiratory distress.  
956 So endotracheal intubation should be done immediately. Answer: endotracheal intubation.

### 958 **A.3 Codex prompt for dataset with options**

959 The following are multiple choice questions (with answers) about professional medicine.

960  
961 Q: Four weeks after starting hydrochlorothiazide, a 49-year-old man with hypertension comes to the  
962 physician because of muscle cramps and weakness. His home medications also include amlodipine. His  
963 blood pressure today is 176/87 mm Hg. Physical examination shows no abnormalities. The precordial  
964 leads of a 12-lead ECG are shown. The addition of which of the following is most likely to have prevented  
965 this patient's condition? (A) Torsemide (B) Nifedipine (C) Eplerenone (D) Hydralazine

966  
967 A: Let's think step by step. The patient has started hydrochlorothiazide. He now presents with muscle  
968 cramps and weakness and an ECG that supports the diagnosis of hypokalemia. The addition of (A)  
969 Torsemide is a loop diuretic and would likely aggravate the hypokalemia, (B) Nifedipine is a calcium  
970 antagonist and would not alleviate the hypocalcemia. (C) Eplerenone is a potassium-sparing diuretic and  
971 would likely decrease the chance of hypokalemia. (C) Eplerenone is a potent vasodilator and would not  
972 decrease the risk of hypokalemia. The answer is (C).

973  
974 Q: A 57-year-old woman comes to the emergency department because of severe pain around her right eye,  
975 blurred vision in the same eye, and a headache for the past 4 hours. She is nauseous but has not vomited.  
976 She can see colored bright circles when she looks at a light source. She is currently being treated for a  
977 urinary tract infection with trimethoprim-sulfamethoxazole. She appears uncomfortable. Vital signs are  
978 within normal limits. Examination shows visual acuity of 20/20 in the left eye and counting fingers at 5  
979 feet in the right eye. The right eye shows conjunctival injection and edema of the cornea. The right pupil  
980 is dilated and fixed. Intravenous analgesia and antiemetics are administered. Which of the following  
981 is the most appropriate next step in management? (A) Perform ultrasound biomicroscopy (B) Perform  
982 gonioscopy (C) Perform funduscopy (D) Administer topical steroids

983  
984 A: Let's think step by step. The patient has severe pain, blurred vision, and a headache. She is also  
985 nauseous but has not vomited. She can see bright colored circles when she looks at a light source. The  
986 right eye shows conjunctival injection and edema of the cornea. The right pupil is dilated and fixed. The  
987 tentative (most likely) diagnosis, given the patient's symptoms, is acute primary angle-closure glaucoma.  
988 Analgesics have been administered. The next step is to confirm the tentative diagnosis using the gold  
989 standard test. (A) ultrasound biomicroscopy could be used but are not widely available, (B) Gonioscopy is  
990 the gold-standard test to confirm the diagnosis of acute primary angle-closure glaucoma, (C) funduscopy  
991 is not as reliable as gonioscopy, (D) topical steroids is not relevant, the possible medical therapy would be



timolol, apraclonidine or pilocarpine. The most appropriate next step in management is to administer intravenous analgesia and antiemetics. The answer is (B).

Q: A 55-year-old man with a history of hypertension and benign prostate hyperplasia presents for follow-up 4 days into the treatment of a urinary tract infection with trimethoprim-sulfamethoxazole. His symptoms have resolved, and he reports no problems with urination, with the exception of a weak urine stream and hesitancy, which he has had for the past 2 years. At the time of this visit, the patient is afebrile; the blood pressure is 130/88 mm Hg and the heart rate is 80/min. There is no flank tenderness. A urinalysis reveals no leukocytes and is negative for esterase. The urinalysis reveals 2 red blood cells (RBCs)/ high power field (HPF), and there are no casts on urinary sediment analysis. The physician, however, notices the following abnormality: Prior treatment BUN 12 mg/dL Creatinine 1.2 mg/dL Today's visit BUN 13 mg/dL Creatinine 2.1 mg/dL (A) Reassure the patient, stop trimethoprim-sulfamethoxazole and repeat the measurement in 1–2 weeks (B) Schedule an intravenous pyelography for urinary obstruction (C) Schedule a cystoscopy for urethral obstruction (D) Admit the patient for further management of acute interstitial nephritis

A: Let's think step by step. The patient has a history of hypertension and benign prostate hyperplasia. These are both risk factors for urinary tract infections. The patient is being treated with trimethoprim-sulfamethoxazole for a urinary tract infection. The symptoms have resolved, but the patient has a weak urine stream and hesitancy. The urinalysis reveals 2 red blood cells (RBCs)/ high power field (HPF), and there are no casts on urinary sediment analysis. The physician, however, notices the following abnormality: BUN 12 mg/dL, Creatinine 1.2 mg/dL (prior treatment) vs BUN 13 mg/dL, Creatinine 2.1 mg/dL (today's visit). The BUN and creatinine levels have increased from the prior visit. This could be due to acute interstitial nephritis. A known but rare side effect of trimethoprim-sulfamethoxazole. Interstitial nephritis should be suspected if there is an elevated creatinine and a urinalysis that shows white cells, white cell casts, and, in some cases, eosinophiluria. Since none of the urinalysis results is present and the patient has no typical symptoms such as fever, rash, or acute interstitial nephritis is an unlikely diagnosis. The answer is (A).

Q: A 32-year-old woman comes to the physician because of weight gain, generalized weakness, and irregular menstrual cycles for the past 16 months. She began having symptoms of insomnia and depression 10 months ago. More recently, she has been having difficulties rising from a chair. She has a 2-year history of hypertension. Current medications include citalopram and hydrochlorothiazide. She is 168 cm (5 ft 6 in) tall and weighs 100 kg (220 lb). BMI is 36 kg/m<sup>2</sup>. Her pulse is 92/min, respirations are 18/min, and blood pressure is 134/76 mm Hg. She appears tired and has a full, plethoric face and central obesity. Examination of the skin shows violaceous linear striations on her lower abdomen. Two midnight serum cortisol studies show measurements of 288 g/L and 253 g/L (N < 90); a 24-hour urinary cortisol measurement was 395 g (N < 300). Upon follow-up laboratory examination, the patient's serum ACTH levels were also elevated at 136 pg/mL (N = 7–50). Which of the following is the most appropriate next step in evaluation? (A) MRI of the head with contrast (B) Measure ACTH levels in inferior petrosal sinuses (C) High-dose dexamethasone suppression test (D) CT scan of the abdomen with contrast

A: Let's think step by step. The patient has Cushing's syndrome, which is characterized by weight gain, generalized weakness, and irregular menstrual cycles. She also has symptoms of insomnia and depression, which could be caused by Cushing's syndrome. Additionally, she has difficulties rising from a chair, which could be caused by muscle weakness. The two elevated serum cortisol measurements and the elevated ACTH levels confirm the diagnosis of Cushing's syndrome. A high dose dexamethasone suppression test will help determine if the cause of Cushing's syndrome is Cushing's disease (CS caused by pituitary hypersecretion of ACTH) or ectopic ACTH syndrome (CS caused by nonpituitary ACTH-secreting tumors). Since the diagnosis of Cushing's syndrome is established, the most appropriate next step is to distinguish between Cushing's disease and ectopic ACTH syndrome using the high-dose dexamethasone suppression test. The answer is (C).

1044  
1045 Q: A 53-year-old man comes to the physician because of a 3-month history of a nonpruritic rash. He has  
1046 been feeling more tired than usual and occasionally experiences pain in his wrists and ankles. He does not  
1047 smoke or drink alcohol. His temperature is 37.6°C (99.7°F), pulse is 98/min, respirations are 18/min, and  
1048 blood pressure is 130/75 mm Hg. Physical examination shows multiple, erythematous, purpuric papules  
1049 on his trunk and extremities that do not blanch when pressed. The remainder of the examination shows no  
1050 abnormalities. The patient's hemoglobin is 14 g/dL, leukocyte count is 9,500/mm<sup>3</sup>, and platelet count  
1051 is 228,000/mm<sup>3</sup>. Urinalysis and liver function tests are within normal limits. The test for rheumatoid  
1052 factor is positive. Serum ANA is negative. Serum complement levels are decreased. Serum protein  
1053 electrophoresis and immunofixation shows increased gammaglobulins with pronounced polyclonal IgM  
1054 and IgG bands. Testing for cryoglobulins shows no precipitate after 24 hours. Chest x-ray and ECG show  
1055 no abnormalities. Which of the following is the most appropriate next step in management? (A) Rapid  
1056 plasma reagin test (B) Hepatitis C serology (C) pANCA assay (D) Bence Jones protein test  
1057

1058 A: Let's think step by step. The patient has a rash, fatigue, and pain in his wrists and ankles. He has  
1059 normal hemoglobin (normal range 13.8-17.2 g/dL), normal leukocyte count (normal range 4500-11000  
1060 wbc/microliter), and normal platelet count (normal range 150-450 platelets/microliter). His serum  
1061 complement levels are decreased. His serum protein electrophoresis and immunofixation show increased  
1062 gammaglobulins with pronounced polyclonal IgM and IgG bands. The cryoglobulin precipitation test  
1063 is normal, ruling out cryoglobulinemia. Serum ANA is negative, so Lupus is less likely. Physical  
1064 examination and symptoms are consistent with possible non-cryoglobulinemic vasculitis. The most likely  
1065 associated diagnosis is Hepatitis C. The answer is (B).  
1066

#### 1067 **A.4 Codex prompt for no-opt dataset**

1068 Q: Four weeks after starting hydrochlorothiazide, a 49-year-old man with hypertension comes to the physi-  
1069 cian because of muscle cramps and weakness. His home medications also include amlodipine. His blood  
1070 pressure today is 176/87 mm Hg. Physical examination shows no abnormalities. The precordial leads  
1071 of a 12-lead ECG are shown. The addition of what is most likely to have prevented this patient's condition?  
1072

1073 A: Let's think step by step. The patient has started hydrochlorothiazide. He now presents with muscle  
1074 cramps and weakness and an ECG that supports the diagnosis of hypokalemia. The addition of Torsemide  
1075 is a loop diuretic and would likely aggravate the hypokalemia, Nifedipine is a calcium antagonist and  
1076 would not alleviate the hypocalcemia. Eplerenone is a potassium-sparing diuretic and would likely  
1077 decrease the chance of hypokalemia. Eplerenone is a potent vasodilator and would not decrease the risk  
1078 of hypokalemia. Answer: Eplerenone.  
1079

1080 Q: A 57-year-old woman comes to the emergency department because of severe pain around her right eye,  
1081 blurred vision in the same eye, and a headache for the past 4 hours. She is nauseous but has not vomited.  
1082 She can see colored bright circles when she looks at a light source. She is currently being treated for a  
1083 urinary tract infection with trimethoprim-sulfamethoxazole. She appears uncomfortable. Vital signs are  
1084 within normal limits. Examination shows visual acuity of 20/20 in the left eye and counting fingers at 5  
1085 feet in the right eye. The right eye shows conjunctival injection and edema of the cornea. The right pupil  
1086 is dilated and fixed. Intravenous analgesia and antiemetics are administered. What is the most appropriate  
1087 next step in management?  
1088

1089 A: Let's think step by step. The patient has severe pain, blurred vision, and a headache. She is also  
1090 nauseous but has not vomited. She can see bright colored circles when she looks at a light source. The  
1091 right eye shows conjunctival injection and edema of the cornea. The right pupil is dilated and fixed. The  
1092 tentative (most likely) diagnosis, given the patient's symptoms, is acute primary angle-closure glaucoma.  
1093 Analgesics have been administered. The next step is to confirm the tentative diagnosis using the gold  
1094 standard test. Ultrasound biomicroscopy could be used but are not widely available, Gonioscopy is the

gold-standard test to confirm the diagnosis of acute primary angle-closure glaucoma, funduscopy is not as reliable as gonioscopy, topical steroids is not relevant, the possible medical therapy would be timolol, apraclonidine or pilocarpine. The most appropriate next step in management is to administer intravenous analgesia and antiemetics. Answer: Gonioscopy.

Q: A 55-year-old man with a history of hypertension and benign prostate hyperplasia presents for follow-up 4 days into the treatment of a urinary tract infection with trimethoprim-sulfamethoxazole. His symptoms have resolved, and he reports no problems with urination, with the exception of a weak urine stream and hesitancy, which he has had for the past 2 years. At the time of this visit, the patient is afebrile; the blood pressure is 130/88 mm Hg and the heart rate is 80/min. There is no flank tenderness. A urinalysis reveals no leukocytes and is negative for esterase. The urinalysis reveals 2 red blood cells (RBCs)/ high power field (HPF), and there are no casts on urinary sediment analysis. The physician, however, notices the following abnormality: Prior treatment BUN 12 mg/dL Creatinine 1.2 mg/dL Today's visit BUN 13 mg/dL Creatinine 2.1 mg/dL

A: Let's think step by step. The patient has a history of hypertension and benign prostate hyperplasia. These are both risk factors for urinary tract infections. The patient is being treated with trimethoprim-sulfamethoxazole for a urinary tract infection. The symptoms have resolved, but the patient has a weak urine stream and hesitancy. The urinalysis reveals 2 red blood cells (RBCs)/ high power field (HPF), and there are no casts on urinary sediment analysis. The physician, however, notices the following abnormality: BUN 12 mg/dL, Creatinine 1.2 mg/dL (prior treatment) vs BUN 13 mg/dL, Creatinine 2.1 mg/dL (today's visit). The BUN and creatinine levels have increased from the prior visit. This could be due to acute interstitial nephritis. A known but rare side effect of trimethoprim-sulfamethoxazole. Interstitial nephritis should be suspected if there is an elevated creatinine and a urinalysis that shows white cells, white cell casts, and, in some cases, eosinophiluria. Since none of the urinalysis results is present and the patient has no typical symptoms such as fever, rash, or acute interstitial nephritis is an unlikely diagnosis. Answer: Reassure the patient, stop trimethoprim-sulfamethoxazole and repeat the measurement in 1–2 weeks.

Q: A 32-year-old woman comes to the physician because of weight gain, generalized weakness, and irregular menstrual cycles for the past 16 months. She began having symptoms of insomnia and depression 10 months ago. More recently, she has been having difficulties rising from a chair. She has a 2-year history of hypertension. Current medications include citalopram and hydrochlorothiazide. She is 168 cm (5 ft 6 in) tall and weighs 100 kg (220 lb). BMI is 36 kg/m<sup>2</sup>. Her pulse is 92/min, respirations are 18/min, and blood pressure is 134/76 mm Hg. She appears tired and has a full, plethoric face and central obesity. Examination of the skin shows violaceous linear striae on her lower abdomen. Two midnight serum cortisol studies show measurements of 288 g/L and 253 g/L (N < 90); a 24-hour urinary cortisol measurement was 395 g (N < 300). Upon follow-up laboratory examination, the patient's serum ACTH levels were also elevated at 136 pg/mL (N = 7–50). What is the most appropriate next step in evaluation?

A: Let's think step by step. The patient has Cushing's syndrome, which is characterized by weight gain, generalized weakness, and irregular menstrual cycles. She also has symptoms of insomnia and depression, which could be caused by Cushing's syndrome. Additionally, she has difficulties rising from a chair, which could be caused by muscle weakness. The two elevated serum cortisol measurements and the elevated ACTH levels confirm the diagnosis of Cushing's syndrome. A high dose dexamethasone suppression test will help determine if the cause of Cushing's syndrome is Cushing's disease (CS caused by pituitary hypersecretion of ACTH) or ectopic ACTH syndrome (CS caused by nonpituitary ACTH-secreting tumors). Since the diagnosis of Cushing's syndrome is established, the most appropriate next step is to distinguish between Cushing's disease and ectopic ACTH syndrome using the high-dose dexamethasone suppression test. Answer: High-dose dexamethasone suppression test.

Q: A 53-year-old man comes to the physician because of a 3-month history of a nonpruritic rash. He has been feeling more tired than usual and occasionally experiences pain in his wrists and ankles. He does not

1147 smoke or drink alcohol. His temperature is 37.6°C (99.7°F), pulse is 98/min, respirations are 18/min, and  
1148 blood pressure is 130/75 mm Hg. Physical examination shows multiple, erythematous, purpuric papules  
1149 on his trunk and extremities that do not blanch when pressed. The remainder of the examination shows no  
1150 abnormalities. The patient's hemoglobin is 14 g/dL, leukocyte count is 9,500/mm<sup>3</sup>, and platelet count  
1151 is 228,000/mm<sup>3</sup>. Urinalysis and liver function tests are within normal limits. The test for rheumatoid  
1152 factor is positive. Serum ANA is negative. Serum complement levels are decreased. Serum protein  
1153 electrophoresis and immunofixation shows increased gammaglobulins with pronounced polyclonal IgM  
1154 and IgG bands. Testing for cryoglobulins shows no precipitate after 24 hours. Chest x-ray and ECG show  
1155 no abnormalities. What is the most appropriate next step in management?

1156  
1157 A: Let's think step by step. The patient has a rash, fatigue, and pain in his wrists and ankles. He has  
1158 normal hemoglobin (normal range 13.8-17.2 g/dL), normal leukocyte count (normal range 4500-11000  
1159 wbc/microliter), and normal platelet count (normal range 150-450 platelets/microliter). His serum  
1160 complement levels are decreased. His serum protein electrophoresis and immunofixation show increased  
1161 gammaglobulins with pronounced polyclonal IgM and IgG bands. The cryoglobulin precipitation test  
1162 is normal, ruling out cryoglobulinemia. Serum ANA is negative, so Lupus is less likely. Physical  
1163 examination and symptoms are consistent with possible non-cryoglobulinemic vasculitis. The most likely  
1164 associated diagnosis is Hepatitis C. Answer: Hepatitis C serology.  
1165

## 1166 B Appendix 2: Datasets released

1167 This appendix provides examples of questions from the datasets released, categorized into MedQA  
1168 (original and no-opt) and Clinician's Case Studies. We also demonstrate how the conversion of a question  
1169 from MedQA-original to MedQA-no-opt (open-ended) format.

### 1170 B.1 MedQA Dataset

#### 1171 1. MedQA Original Question No. 1

1172 *A 30-year-old G4P3 woman at 38 weeks gestation is admitted to the labor and delivery unit*  
1173 *complaining of contractions every 5 minutes for the past hour. Her previous births have been via*  
1174 *uncomplicated caesarean sections, but she wishes to attempt vaginal delivery this time. Her prenatal*  
1175 *care is notable for gestational diabetes controlled with diet and exercise. The delivery is prolonged,*  
1176 *but the patient's pain is controlled with epidural analgesia. She delivers a male infant with Apgar*  
1177 *scores of 8 and 9 at 1 and 5 minutes, respectively. Fundal massage is performed, but the placenta*  
1178 *does not pass. The obstetrician manually removes the placenta, but a red mass protrudes through the*  
1179 *vagina attached to the placenta. The patient loses 500 mL of blood over the next minute, during*  
1180 *which her blood pressure decreases from 120/80 mmHg to 90/65 mmHg. What is the best next step in*  
1181 *management?*

1182  
1183 **Options:** 'A': 'Hysterectomy', 'B': 'Intravenous oxytocin', 'C': 'Elevate posterior fornix', 'D':  
1184 'Red blood cell transfusion'  
1185

1186 **Correct Answer:** The correct answer is **C: Elevate posterior fornix**

- 1187 • **Converted Descriptive Version of the Question** A 30-year-old G4P3 woman at 38 weeks  
1188 gestation is admitted to the labor and delivery unit complaining of contractions every 5 minutes  
1189 for the past hour. Her previous births have been via uncomplicated caesarean sections, but  
1190 she wishes to attempt vaginal delivery this time. Her prenatal care is notable for gestational  
1191 diabetes controlled with diet and exercise. The delivery is prolonged, but the patient's pain  
1192 is controlled with epidural analgesia. She delivers a male infant with Apgar scores of 8 and  
1193 9 at 1 and 5 minutes, respectively. Fundal massage is performed, but the placenta does not  
1194 pass. The obstetrician manually removes the placenta, but a red mass protrudes through the

vagina attached to the placenta. The patient loses 500 mL of blood over the next minute, during which her blood pressure decreases from 120/80 mmHg to 90/65 mmHg. Elaborate the steps of management of this case ?	1195 1196 1197
<ul style="list-style-type: none"> <li>• <b>Reasoning for the no-opt question</b> In this scenario, the patient has a retained placenta with a red mass protruding through the vagina, which suggests that there may be a placental abnormality such as a placenta accreta or placenta percreta. This can cause excessive bleeding, which can lead to hypovolemic shock and potentially life-threatening complications. The first step in managing this case would be to elevate the posterior fornix, which will help to reduce the bleeding by increasing the intrauterine pressure and compressing the bleeding vessels. This will also help to prevent further descent of the placenta and allow for easier removal. Additionally, the patient's blood loss should be closely monitored and fluid replacement should be administered as needed to maintain blood pressure and prevent shock. Thus, the answer is Elevate posterior fornix.</li> </ul>	1198 1199 1200 1201 1202 1203 1204 1205 1206 1207
<b>2. MedQA Original Question No. 2</b>	1208
<i>A 16-year-old boy is brought to the emergency department by ambulance from a soccer game. During the game, he was about to kick the ball when another player collided with his leg from the front. He was unable to stand up after this collision and reported severe knee pain. On presentation, he was found to have a mild knee effusion. Physical exam showed that his knee could be pushed posteriorly at 90 degrees of flexion but it could not be pulled anteriorly in the same position. The anatomic structure that was most likely injured in this patient has which of the following characteristics?</i>	1209 1210 1211 1212 1213 1214
<b>Options:</b> 'A': 'Runs anteriorly from the medial femoral condyle', 'B': 'Runs medially from the lateral femoral condyle', 'C': 'Runs posteriorly from the lateral femoral condyle', 'D': 'Runs posteriorly from the medial femoral condyle'	1215 1216 1217 1218
<b>Correct Answer:</b> The correct answer is <b>D: Runs posteriorly from the medial femoral condyle</b>	1219
<ul style="list-style-type: none"> <li>• <b>Converted Descriptive Version of the Question</b> A 16-year-old boy is brought to the emergency department by ambulance from a soccer game. During the game, he was about to kick the ball when another player collided with his leg from the front. He was unable to stand up after this collision and reported severe knee pain. On presentation, he was found to have a mild knee effusion. Physical exam showed that his knee could be pushed posteriorly at 90 degrees of flexion but it could not be pulled anteriorly in the same position. Corresponding to the signs and symptoms, the patient presents with and the investigations performed already, what immediate measures of management are to be initiated, prioritising their urgency, and keeping in mind the probable diagnosis for the same?</li> <li>• <b>Reasoning for the no-opt question</b> The patient's history of sudden onset knee pain following a direct blow to the knee, along with the physical examination findings of a mild knee effusion and inability to pull the knee anteriorly at 90 degrees of flexion, suggests a possible diagnosis of a medial tibial stress syndrome (MTSS) or shin splints. This condition is caused by repetitive trauma or overuse of the tibia, leading to inflammation and pain in the surrounding soft tissues. The patient's age and participation in sports also increase the likelihood of this diagnosis. To confirm the diagnosis, imaging studies such as X-rays or an MRI may be ordered. In the meantime, conservative management options such as rest, ice, compression, and elevation (RICE) should be initiated to alleviate pain and reduce inflammation. Additionally, the patient should be advised to avoid activities that exacerbate the condition and consider wearing proper footwear and orthotics to prevent further injury. Thus, the answer is Runs posteriorly from the medial femoral condyle.</li> </ul>	1220 1221 1222 1223 1224 1225 1226 1227 1228 1229 1230 1231 1232 1233 1234 1235 1236 1237 1238 1239 1240
<b>3. MedQA Original Question No. 3</b>	1241
<i>A 64-year-old male with a past medical history of two myocardial infarctions presents to the emergency room with shortness of breath. He notes that he stopped taking his furosemide two weeks</i>	1242 1243

1244 *prior, because he ran out of pills. On exam, his oxygen saturation is 78%, his lungs have crackles*  
1245 *throughout, and jugular venous pulsation is located at the earlobe. EKG and troponin levels are*  
1246 *normal. Which of the following is consistent with this man's pulmonary physiology?*

1247 **Options:** 'A': 'Decreased Aa gradient, decreased surface area for diffusion, normal diffusion  
1248 distance', 'B': 'Decreased Aa gradient, increased surface area for diffusion, decreased diffusion  
1249 distance', 'C': 'Increased Aa gradient, normal surface area for diffusion, increased diffusion dis-  
1250 tance', 'D': 'Increased Aa gradient, decreased surface area for diffusion, increased diffusion distance'

1252 **Correct Answer:** The correct answer is **D: Increased Aa gradient, decreased surface area for**  
1253 **diffusion, increased diffusion distance**

- 1254 • **Converted Descriptive Version of the Question** A 64-year-old male with a past medical  
1255 history of two myocardial infarctions presents to the emergency room with shortness of breath.  
1256 He notes that he stopped taking his furosemide two weeks prior, because he ran out of pills. On  
1257 exam, his oxygen saturation is 78%, his lungs have crackles throughout, and jugular venous  
1258 pulsation is located at the earlobe. EKG and troponin levels are normal. What pulmonary  
1259 pathology is this man suffering from?
- 1260 • **Reasoning for the no-opt question** The patient's symptoms of shortness of breath, low oxygen  
1261 saturation, and lung crackles suggest respiratory distress. The absence of abnormal EKG  
1262 and troponin levels rules out acute coronary syndrome or cardiac failure as the cause of his  
1263 symptoms. The patient's history of stopping furosemide two weeks ago suggests that he may  
1264 have developed fluid overload, which can lead to pulmonary edema. This is supported by the  
1265 presence of jugular venous pulsation at the earlobe, indicating elevated central venous pressure.  
1266 The patient's age and history of myocardial infarctions also increase his risk for developing  
1267 pulmonary disease. Therefore, the most likely diagnosis is increased Aa gradient, decreased  
1268 surface area for diffusion, and increased diffusion distance, consistent with pulmonary edema.  
1269 Thus, the answer is Increased Aa gradient, decreased surface area for diffusion, increased  
1270 diffusion distance.

## 1271 **B.2 Clinician's Case Studies Dataset**

### 1272 **1. Clinician's Case Studies Question No. 1**

1273 *A 44 year old woman visited the opd c/o heavy menstrual bleeding for the past 10 days preceded by*  
1274 *prolonged spotting for 8 days. she has recently noticed some irregularity in her menses in the last*  
1275 *3-6 months. She has noticed weight gain of about 3 kg in the same time period. She has had two*  
1276 *normal deliveries and no other surgeries. She has diabetes mellitus and has been on tab. Metformin*  
1277 *500 mg three times a day for the last 2 years. She reports that her blood sugars are regularly checked*  
1278 *but recently have been outside the limit advised by her GP. She is not a known hypertensive. An*  
1279 *ultrasonography of the pelvis reveals a bulky uterus 12\*7\*8 cm in size with an endometrial thickness*  
1280 *of 18mm showing cystic spaces within. What should be the next investigation and treatment plan in*  
1281 *this patient?*

1283 **Correct Answer:** The correct answer is **Pipelle endometrial biopsy followed by cyclical pro-**  
1284 **gesterone therapy or Mirena insertion, dilatation and curettage for endometrial biopsy or**  
1285 **hysteroscopy with endometrial biopsy , Followed by cyclical progesterone or Mirena insertion**

- 1286 • **Reasoning for the no-opt question** The patient's symptoms of heavy menstrual bleeding and  
1287 irregular periods suggest endometrial pathology. Her recent weight gain and history of diabetes  
1288 increase the likelihood of endometrial cancer. The ultrasound findings of a bulky uterus with  
1289 cystic spaces within the endometrium further support this diagnosis. To confirm the diagnosis,  
1290 a pipelle endometrial biopsy should be performed. If the biopsy results indicate endometrial  
1291 cancer, treatment options include cyclical progesterone therapy or Mirena insertion, or dilatation

and curettage for endometrial biopsy or hysteroscopy with endometrial biopsy. These treatments aim to reduce the risk of recurrent bleeding and prevent further complications. Thus, the answer is Pipelle endometrial biopsy followed by cyclical progesterone therapy or Mirena insertion, or dilatation and curettage for endometrial biopsy or hysteroscopy with endometrial biopsy, followed by cyclical progesterone or Mirena insertion.	1292 1293 1294 1295 1296
<b>2. Clinician’s Case Studies Question No. 2</b>	1297
<i>A 27 year old man presented with red painful nodules over both shins for five days. He also gives a history of persistent dry cough and shortness of breath. On further enquiry, he experiences blurred vision and occasional pain in his eyes. On examination, there was a single palpable asymptomatic nodule in the right axilla. Which one blood test will confirm the diagnosis?</i>	1298 1299 1300 1301 1302
<b>Correct Answer:</b> The correct answer is <b>ACE levels</b>	1303
<ul style="list-style-type: none"> <li>• <b>Reasoning for the no-opt question</b> The patient’s symptoms of red painful nodules on both shins, persistent dry cough, shortness of breath, blurred vision, and occasional eye pain are consistent with a diagnosis of sarcoidosis, a systemic inflammatory disease caused by the growth of granulomas. Elevated levels of angiotensin-converting enzyme (ACE) in the blood are commonly seen in patients with sarcoidosis, and can be used to support the diagnosis. Therefore, measuring ACE levels in this patient’s blood would help confirm the diagnosis of sarcoidosis. Thus, the answer is ACE levels.</li> </ul>	1304 1305 1306 1307 1308 1309 1310
<b>3. Clinician’s Case Studies Question No. 3</b>	1311
<i>A 46 year old man visits an outpatient department of dermatology with sudden loss of scalp hair in patches. On a detailed examination, partial loss of eyelashes were noticed. Hair pull test was positive at the margins of the bald patches. On dermoscopic evaluation, black dots, yellow dots, broken and tapering hair were observed. What is the drug of choice to halt the progression of the disease?</i>	1312 1313 1314 1315 1316
<b>Correct Answer</b> The correct answer is <b>Corticosteroids</b>	1317
<ul style="list-style-type: none"> <li>• <b>Reasoning for the no-opt question</b> Alopecia areata is an autoimmune condition that causes hair loss in patches or completely. In this case, the patient has partial loss of eyelashes as well, which supports the diagnosis. The presence of black dots, yellow dots, broken and tapering hair on dermoscopy suggests that there is an active inflammation in the area. Corticosteroids are effective in halting the progression of alopecia areata by reducing inflammation and suppressing the immune system. Topical corticosteroids can be applied directly to the affected areas, while intralesional corticosteroid injections can be used for more resistant cases. Thus, the answer is corticosteroids.</li> </ul>	1318 1319 1320 1321 1322 1323 1324 1325
<b>C Appendix 3: Examples of Chosen/Rejected Reasoning Pairs for Reward Model training</b>	1326 1327
<b>C.1 Examples of Chosen/Rejected Reasoning Pairs for Reward Model training</b>	1328
In this section, we discuss the need to create chosen and rejected reasoning pairs for training the verifier for our model.	1329 1330
The chosen reasoning is the reasoning that led to the correct answer, while the rejected reasoning is the reasoning that led to incorrect answers. By training the verifier with both chosen and rejected reasoning pairs, the model can learn to distinguish between valid and invalid reasoning processes, improving its accuracy in answering questions.	1331 1332 1333 1334 1335

1336 **C.2 Example Question No 1**

1337 **Question:** A 67-year-old woman presents to her primary care physician for urinary incontinence. She has  
1338 been waking up every morning with her underwear soaked in urine. She notices that at work if she does  
1339 not take regular bathroom breaks her underwear will have some urine in it. She urinates 5 to 11 times per  
1340 day but she claims it is a small volume. Her current medications include lisinopril, metformin, insulin,  
1341 aspirin, atorvastatin, sodium docusate, and loratadine. Her temperature is 98.2°F (36.8°C), blood pressure  
1342 is 167/108 mmHg, pulse is 90/min, respirations are 15/min, and oxygen saturation is 99% on room air.  
1343 Physical exam is notable for decreased pinprick sensation in the lower extremities and a systolic murmur  
1344 along the right upper sternal border. Which of the following is the best treatment for this patient?  
1345

1346 **Options :** 'A': 'Bethanechol and intermittent straight catheterization', 'B': 'Bethanechol during the day  
1347 and oxybutynin at night', 'C': 'No recommendations needed', 'D': 'Reduce fluid intake and discontinue  
1348 diuretics'  
1349

1350 **Correct Answer:** The correct answer is **Bethanechol and intermittent straight catheterization.**

1351 **C.2.1 Chosen and Rejected Reasoning Pairs**

1352 **Pair 1:**

1353  
1354 *Chosen Reasoning:* The patient's symptoms of urinary incontinence, frequent urination, and decreased  
1355 pinprick sensation in the lower extremities suggest a neurological cause, such as a spinal cord injury or  
1356 multiple sclerosis. However, the patient's history of hypertension, hyperglycemia, and hyperlipidemia also  
1357 suggest a possible contribution from an overactive bladder. Therefore, the best treatment approach would  
1358 be to start the patient on bethanechol, a cholinergic agent that can help improve bladder function and reduce  
1359 urinary frequency. Additionally, intermittent straight catheterization can help manage urinary retention  
1360 and prevent urinary tract infections. This combination of medication and catheterization can effectively  
1361 address both the neurological and non-neurological contributors to the patient's urinary incontinence,  
1362 thus improving her quality of life. Thus, the answer is Bethanechol and intermittent straight catheterization.  
1363

1364 *Rejected Reasoning:* The patient's symptoms of urinary incontinence, frequent urination, and decreased  
1365 pinprick sensation in the lower extremities suggest a neurological cause, such as overactive bladder  
1366 or spinal cord injury. The patient's history of hypertension and diabetes mellitus also increase  
1367 the likelihood of a neurological cause. Bethanechol, a cholinergic agonist, is used to treat urinary  
1368 retention and can help improve bladder function by increasing the frequency and amplitude of  
1369 contractions. Oxybutynin, an anticholinergic agent, can help reduce urgency and frequency of urination,  
1370 especially at night when the patient may not be able to access a bathroom quickly. Thus, the combination  
1371 of bethanechol during the day and oxybutynin at night would be the best treatment approach for this patient.  
1372

1373 **Pair 2:**

1374  
1375 *Chosen Reasoning:* The patient's symptoms of urinary incontinence, frequent urination, and decreased  
1376 pinprick sensation in the lower extremities suggest a neurological cause, such as a spinal cord injury or  
1377 multiple sclerosis. However, the patient's history of hypertension, hyperglycemia, and hyperlipidemia also  
1378 suggest a possible contribution from an overactive bladder. Therefore, the best treatment approach would  
1379 be to start the patient on bethanechol, a cholinergic agent that can help improve bladder function and reduce  
1380 urinary frequency. Additionally, intermittent straight catheterization can help manage urinary retention  
1381 and prevent urinary tract infections. This combination of medication and catheterization can effectively  
1382 address both the neurological and non-neurological contributors to the patient's urinary incontinence,  
1383 thus improving her quality of life. Thus, the answer is Bethanechol and intermittent straight catheterization.  
1384

1385 *Rejected Reasoning:* This patient's symptoms of urinary incontinence, frequency, and nocturia are  
1386 consistent with overactive bladder (OAB). However, her age, hypertension, diabetes mellitus, and



neuropathy suggest that she may also have underlying detrusor instability or urgency urinary incontinence. Given these factors, the best course of action would be to refer her to a urologist for further evaluation and management. A urologist can perform tests such as a urinalysis, postvoid residual measurement, and cystoscopy to confirm the diagnosis and determine the appropriate treatment. Thus, the answer is "No recommendations needed" as the patient requires specialized care beyond the scope of primary care.

### Pair 3:

*Chosen Reasoning:* The patient's symptoms of urinary incontinence, frequent urination, and decreased pinprick sensation in the lower extremities suggest a neurological cause, such as a spinal cord injury or multiple sclerosis. However, the patient's history of hypertension, hyperglycemia, and hyperlipidemia also suggest a possible contribution from an overactive bladder. Therefore, the best treatment approach would be to start the patient on bethanechol, a cholinergic agent that can help improve bladder function and reduce urinary frequency. Additionally, intermittent straight catheterization can help manage urinary retention and prevent urinary tract infections. This combination of medication and catheterization can effectively address both the neurological and non-neurological contributors to the patient's urinary incontinence, thus improving her quality of life. Thus, the answer is Bethanechol and intermittent straight catheterization.

*Rejected Reasoning:* The patient's symptoms of urinary incontinence and frequent urination suggest an overactive bladder. This can be caused by excessive fluid intake or certain medications such as diuretics. In this case, the patient's history of taking lisinopril, a diuretic, likely contributes to her symptoms. Additionally, her high blood pressure and systolic murmur suggest that reducing fluid intake may help alleviate these issues. Discontinuing the diuretic medication and limiting fluid intake would help reduce the frequency of urination and minimize the risk of incontinence. Thus, the answer is Reduce fluid intake and discontinue diuretics.

### C.3 Example Question No 2

**Question:** A 55-year-old man presents into the emergency department with a severe cough and difficulty breathing. He says that he finds himself out of breath after taking a few steps, and has to sit down and rest, in order to continue. He also says that, at night, he has the greatest difficulty in breathing and usually uses at least 3 pillows to sleep comfortably. He mentions a cough that appears only at night, but which is persistent enough to wake him up from sleep. He mentions that he has had a 'heart attack' 5 years ago. He also says that he continues to consume alcohol on a regular basis even though his doctor has advised against it. He has brought his lab reports which he had recently got done on the suggestions of his family doctor. An electrocardiogram (ECG) and a chest X-ray are found. Which of the following is the next step in this patient's management?

**Options:** 'A': 'Computed Tomography (CT)', 'B': 'Echocardiogram', 'C': 'Stress test', 'D': 'Cardiac biopsy'

**Correct Answer:** The correct answer is **Echocardiogram**.

#### C.3.1 Chosen and Rejected Reasoning Pairs

##### Pair 1:

*Chosen Reasoning:* This patient's symptoms of shortness of breath, especially at night, and use of multiple pillows to sleep suggest pulmonary congestion. His history of heart attack increases the likelihood of underlying cardiac disease. The presence of a persistent nighttime cough further supports this diagnosis. Additionally, the patient's continued alcohol consumption may have contributed to the development or exacerbation of any cardiac condition. An echocardiogram would provide valuable information regarding the structure and function of the heart, allowing for the assessment of potential

cardiac causes of his symptoms, such as left ventricular dysfunction or valvular heart disease. Thus, the answer is Echocardiogram.

*Rejected Reasoning:* This patient's symptoms of shortness of breath, particularly at night, and use of multiple pillows to sleep suggest possible pulmonary disease. His history of heart attack increases the likelihood of coronary artery disease. Alcohol consumption further increases the risk of lung disease. A chest X-ray does not provide adequate information to diagnose pulmonary embolism or lung cancer, which are possibilities based on the patient's symptoms and history. An ECG may help diagnose any cardiac arrhythmias, but it will not provide information about lung pathology. A CT scan of the lungs is the most appropriate next step in this patient's management as it can help diagnose pulmonary embolism, lung cancer, and other lung diseases. Thus, the answer is Computed Tomography (CT).

**Pair 2:**

*Chosen Reasoning:* This patient's symptoms of shortness of breath, especially at night, and use of multiple pillows to sleep suggest pulmonary congestion. His history of heart attack increases the likelihood of underlying cardiac disease. The presence of a persistent nighttime cough further supports this diagnosis. Additionally, the patient's continued alcohol consumption may have contributed to the development or exacerbation of any cardiac condition. An echocardiogram would provide valuable information regarding the structure and function of the heart, allowing for the assessment of potential cardiac causes of his symptoms, such as left ventricular dysfunction or valvular heart disease. Thus, the answer is Echocardiogram.

*Rejected Reasoning:* This patient's symptoms of shortness of breath, especially at night, and a history of heart attack suggest that he may have coronary artery disease or other cardiac issues. The fact that he experiences difficulty breathing when lying flat and requires multiple pillows to sleep suggests that he may have pulmonary congestion or left ventricular failure. Additionally, his persistent cough at night could be indicative of acid reflux or aspiration. Given these concerns, a stress test would be an appropriate next step in the patient's management to assess his cardiac function and identify any potential issues. Thus, the answer is Stress test.

**Pair 3:**

*Chosen Reasoning:* This patient's symptoms of shortness of breath, especially at night, and use of multiple pillows to sleep suggest pulmonary congestion. His history of heart attack increases the likelihood of underlying cardiac disease. The presence of a persistent nighttime cough further supports this diagnosis. Additionally, the patient's continued alcohol consumption may have contributed to the development or exacerbation of any cardiac condition. An echocardiogram would provide valuable information regarding the structure and function of the heart, allowing for the assessment of potential cardiac causes of his symptoms, such as left ventricular dysfunction or valvular heart disease. Thus, the answer is Echocardiogram.

*Rejected Reasoning:* The patient's symptoms of shortness of breath, especially at night, and persistent cough suggest chronic obstructive pulmonary disease (COPD). His history of heart attack and continued alcohol consumption increase the likelihood of cardiac involvement. The ECG and chest X-ray do not provide sufficient information to diagnose COPD or assess cardiac function. A cardiac biopsy would help establish the diagnosis of COPD and determine the extent of cardiac damage, guiding appropriate management and treatment. Thus, the answer is Cardiac biopsy.

**D Appendix 4: Examples of Responses generated using MedCodex and Codex approaches** 1486  
1487

This section provides more examples of the Responses generated using MedCodex and Codex approaches highlighting their differences. 1488  
1489  
1490

**D.1 Question No. 1** 1491

**Question:** 1492

A male newborn is evaluated 24 hours after delivery for high-pitched crying, poor feeding, rhinorrhea, and low-grade fever. He was born at 40 weeks' gestation at 2514 g (5.54 lb) to a 28-year-old woman, gravida 3, para 2, by an uncomplicated cesarean section. Apgar scores were 8 and 9 at 1 and 5 minutes, respectively. The mother did not receive prenatal care. The infant's temperature is 38.0°C (100.4°F), pulse is 170/min, and blood pressure is 71/39 mm Hg. Examination shows hyperreflexia, tremors, and an excessive startle response. These symptoms are mostly like due to maternal use of what substance? 1493  
1494  
1495  
1496  
1497  
1498

**Response generated using MedCodex approach:** 1499  
1500

The baby is showing symptoms such as high pitched crying, poor feeding, rhinorrhea and low grade fever. The baby was born full term at 40 weeks and had Apgar scores of 8 and 9 at 1 and 5 minutes which indicates that there was no distress during delivery. Mother did not receive prenatal care which could indicate lack of awareness or lack of resources. Baby has hyperreflexia, tremors and excessive startle response which could indicate neurological problems. It could be due to maternal drug use during pregnancy. Maternal drug use during pregnancy can cause withdrawal symptoms in the baby after birth. The most common drugs that cause withdrawal symptoms in the baby are opioids, benzodiazepines, alcohol, nicotine, and cocaine. As the mother did not receive prenatal care, it is more likely that the mother used drugs during pregnancy. Therefore, the baby's symptoms are most likely due to maternal use of drugs during pregnancy. Answer: maternal use of drugs during pregnancy. 1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511

**Response generated using Codex approach:** 1512

The baby has a high-pitched cry, poor feeding, rhinorrhea, and low-grade fever. The baby was born at 40 weeks' gestation at 2514 g (5.54 lb) to a 28-year-old woman, gravida 3, para 2, by an uncomplicated cesarean section. Apgar scores were 8 and 9 at 1 and 5 minutes, respectively. The mother did not receive prenatal care. The baby's temperature is 38.0°C (100.4°F), pulse is 170/min, and blood pressure is 71/39 mm Hg. Examination shows hyperreflexia, tremors, and an excessive startle response. These symptoms are most likely due to withdrawal from maternal drug use during pregnancy. The most common drugs that cause withdrawal symptoms in neonates are opioids, benzodiazepines, and stimulants. The mother did not receive prenatal care, which suggests that she may have used drugs during pregnancy. The baby's symptoms are consistent with withdrawal from opioids or benzodiazepines. Answer: Opioids or Benzodiazepines. 1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523

**D.2 Question No. 2** 1524

**Question:** 1525

A 59-year-old man presents to general medical clinic for his yearly checkup. He has no complaints except for a dry cough. He has a past medical history of type II diabetes, hypertension, hyperlipidemia, asthma, and depression. His home medications are sitagliptin/metformin, lisinopril, atorvastatin, albuterol inhaler, and citalopram. His vitals signs are stable, with blood pressure 126/79 mmHg. Hemoglobin A1C is 6.3%, and creatinine is 1.3 g/dL. The remainder of his physical exam is unremarkable. If this patient's cough is due to one of the medications he is taking, what would be the next step in management? 1526  
1527  
1528  
1529  
1530  
1531  
1532

**Response generated using MedCodex approach:** 1533

1534 The patient has a history of multiple comorbidities such as diabetes, hypertension, hyperlipidemia, asthma,  
1535 and depression. He is on several medications including sitagliptin/metformin, lisinopril, atorvastatin,  
1536 albuterol inhaler, and citalopram. The patient presents with a dry cough, which could be related to one of  
1537 the medications he is taking. The cough could be due to the ACE inhibitor lisinopril, which is a known  
1538 side effect. The next step in management would be to hold the lisinopril and monitor the patient's blood  
1539 pressure to ensure it remains controlled. Additionally, an alternative medication for hypertension could be  
1540 considered if the patient's blood pressure remains stable. Answer: Hold lisinopril and monitor blood  
1541 pressure.

1542  
1543 **Response generated using Codex approach:**

1544 The patient has a dry cough. It could be related to his medications. The patient takes several medications,  
1545 including sitagliptin/metformin, lisinopril, atorvastatin, albuterol inhaler, and citalopram. The cough  
1546 could be caused by any of these medications. However, it is important to note that a dry cough is a  
1547 common side effect of ACE inhibitors, such as lisinopril. Therefore, the next step in management would  
1548 be to consider alternative medications for hypertension, such as beta-blockers, calcium channel blockers,  
1549 or thiazide diuretics. Answer: Consider alternative medications for hypertension.

1551 **E Appendix 5: Instructions provided to the interns for evaluation of the experiments'**  
1552 **results**

1553 This section gives the instructions that we provided to the interns before the result sheets were provided to  
1554 them for evaluation.

1555  
1556 **E.1 Instructions**

1557 Each row in the task sheet contains a medical question (Column A) and reasoning  
1558 generated by different approaches in a shuffled order (Columns B, D, F, H) in the  
1559 excel sheet provided. The task is to evaluate the reasoning based on its quality on  
1560 a 3-point scale described below:

- 1562 • Agree: the answer is perfect with no mistakes OR the answer is mostly correct,  
1563 with one or two minor mistakes
- 1564 • Neutral: the answer is not incorrect, but there is missing information OR it is  
1565 hard to decide given my knowledge
- 1566 • Disagree: the answer is incorrect but it has a few things right OR the answer is  
1567 completely hallucinated

1568 Each Response column will be followed by a drop-down list column containing the  
1569 options for the 3-point scale. Please select one from the dropdown list.

1570  
1571 Please feel free to reach out to me in case of any doubt OR if you have any queries.

## F Medical Expert Evaluation results of all the experiments<sup>5</sup>

1573

### F.1 Results of Human Evaluations of Llama-2-7B chat responses on Clinician’s Case studies

1574

	Agree	Neutral	Disagree
MedCodex	88	4	8
Codex	92	0	8
MedCodex-Codex	92	8	0
MedCodex-Reward	92	4	4

Table 3: Medical Expert-1 evaluated results of all four approaches on Clinician’s Case studies (25 questions)

	Agree	Neutral	Disagree
MedCodex	88	8	4
Codex	88	4	8
MedCodex-Codex	84	12	4
MedCodex-Reward	88	4	8

Table 4: Medical Expert-2 evaluated results of all four approaches on Clinician’s Case studies (25 questions)

The average inter-annotator agreement for this section: **0.860**

1575

1576

### F.2 Results of Human Evaluations of Llama-2-7B chat responses on MedQA no-opt dataset

1577

	Agree	Neutral	Disagree
MedCodex	94	4	2
Codex	49	13	38
MedCodex-Codex	84	11	5
MedCodex-Reward	95	3	2

Table 5: Medical Expert-1 evaluated results of all four approaches on MedQA no-opt dataset Set 1 (100 questions)

	Agree	Neutral	Disagree
MedCodex	94	4	2
Codex	54	16	30
MedCodex-Codex	85	11	4
MedCodex-Reward	96	2	2

Table 6: Medical Expert-2 evaluated results of all four approaches on MedQA no-opt dataset Set 1 (100 questions)

<sup>5</sup>We have used Python Scipy library for calculating Inter-Annotator Agreement

	Agree	Neutral	Disagree
MedCodex	88	0	12
Codex	54	7	39
MedCodex-Codex	77	1	22
MedCodex-Reward	85	1	14

Table 7: Medical Expert-1 evaluated results of all four approaches on MedQA no-opt dataset Set 2 (100 questions)

	Agree	Neutral	Disagree
MedCodex	89	0	11
Codex	55	6	39
MedCodex-Codex	79	0	21
MedCodex-Reward	86	0	14

Table 8: Medical Expert-2 evaluated results of all four approaches on MedQA no-opt dataset Set 2 (100 questions)

	Agree	Neutral	Disagree
MedCodex	65	31	4
Codex	68	27	5
MedCodex-Codex	69	25	6
MedCodex-Reward	76	21	3

Table 9: Medical Expert-1 evaluated results of all four approaches on MedQA no-opt dataset Set 3 (100 questions)

	Agree	Neutral	Disagree
MedCodex	92	2	6
Codex	62	10	28
MedCodex-Codex	89	1	10
MedCodex-Reward	95	1	4

Table 10: Medical Expert-2 evaluated results of all four approaches on MedQA no-opt dataset Set 3 (100 questions)

	Agree	Neutral	Disagree
MedCodex	70	19	11
Codex	13	32	55
MedCodex-Codex	73	17	10
MedCodex-Reward	76	15	9

Table 11: Medical Expert-1 evaluated results of all four approaches on MedQA no-opt dataset Set 4 (100 questions)

	Agree	Neutral	Disagree
MedCodex	70	16	14
Codex	46	16	38
MedCodex-Codex	76	13	11
MedCodex-Reward	81	13	6

Table 12: Medical Expert-2 evaluated results of all four approaches on MedQA no-opt dataset Set 4 (100 questions)

	Agree	Neutral	Disagree
MedCodex	84	1	15
Codex	78	1	21
MedCodex-Codex	89	0	11
MedCodex-Reward	91	0	9

Table 13: Medical Expert-1 evaluated results of all four approaches on MedQA no-opt dataset Set 5 (100 questions)

	Agree	Neutral	Disagree
MedCodex	87	7	6
Codex	82	10	8
MedCodex-Codex	86	9	5
MedCodex-Reward	87	7	6

Table 14: Medical Expert-2 evaluated results of all four approaches on MedQA no-opt dataset Set 5 (100 questions)

The inter-annotator agreement for this section: **0.659**

1578  
1579

The average inter-annotator agreement for this section: **0.32**

1580  
1581

### F.3 Results of Human Evaluations of Llama-2-70B chat responses on Clinician’s Case studies

	Agree	Neutral	Disagree
MedCodex	92	0	8
Codex	96	0	4
MedCodex-Codex	92	0	8
MedCodex-Reward	88	0	12

Table 15: Medical Expert-1 evaluated results of all four approaches on Clinician’s Case studies (25 questions)

	Agree	Neutral	Disagree
MedCodex	72	16	12
Codex	76	8	16
MedCodex-Codex	76	12	12
MedCodex-Reward	76	8	16s

Table 16: Medical Expert-2 evaluated results of all four approaches on Clinician’s Case studies (25 questions)

The average inter-annotator agreement for this section: **0.32**



	Agree	Neutral	Disagree
MedCodex	94	1	5
Codex	94	0	6
MedCodex-Codex	94	1	5
MedCodex-Reward	97	1	2

Table 17: Medical Expert-1 evaluated results of all four approaches on MedQA no-opt dataset Set 1 (125 questions)

	Agree	Neutral	Disagree
MedCodex	86	9	5
Codex	90	6	4
MedCodex-Codex	89	7	4
MedCodex-Reward	85	10	5

Table 18: Medical Expert-2 evaluated results of all four approaches on MedQA no-opt dataset Set 2 (125 questions)

	Agree	Neutral	Disagree
MedCodex	95	4	1
Codex	90	6	5
MedCodex-Codex	97	1	2
MedCodex-Reward	90	6	4

Table 19: Medical Expert-1 evaluated results of all four approaches on MedQA no-opt dataset Set 2 (125 questions)

	Agree	Neutral	Disagree
MedCodex	95	4	1
Codex	90	4	6
MedCodex-Codex	97	2	2
MedCodex-Reward	90	6	5

Table 20: Medical Expert-2 evaluated results of all four approaches on MedQA no-opt dataset Set 2 (125 questions)

The average inter-annotator agreement for this section: **0.372**

	Agree	Neutral	Disagree
MedCodex	87	2	11
Codex	90	0	10
MedCodex-Codex	83	5	12
MedCodex-Reward	86	3	10

Table 21: Medical Expert-1 evaluated results of all four approaches on MedQA no-opt dataset Set 3 (125 questions)

	Agree	Neutral	Disagree
MedCodex	59	23	18
Codex	61	28	11
MedCodex-Codex	73	19	8
MedCodex-Reward	95	1	4

Table 22: Medical Expert-2 evaluated results of all four approaches on MedQA no-opt dataset Set 3 (125 questions)

	Agree	Neutral	Disagree
MedCodex	94	0	6
Codex	82	0	18
MedCodex-Codex	87	0	13
MedCodex-Reward	94	0	6

Table 23: Medical Expert-1 evaluated results of all four approaches on MedQA no-opt dataset Set 4 (125 questions)

	Agree	Neutral	Disagree
MedCodex	84	10	6
Codex	76	13	11
MedCodex-Codex	79	10	11
MedCodex-Reward	80	13	7

Table 24: Medical Expert-2 evaluated results of all four approaches on MedQA no-opt dataset Set 4 (125 questions)