

## Research and Applications

# Reasoning with large language models for medical question answering

Mary M. Lucas , MS, RN<sup>1</sup>, Justin Yang<sup>2</sup>, Jon K. Pomeroy, DO, FAMIA<sup>1,3</sup>,  
Christopher C. Yang, PhD<sup>1,\*</sup>

<sup>1</sup>College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, United States, <sup>2</sup>Department of Computer Science, University of Maryland, College Park, MD 20742, United States, <sup>3</sup>Penn Medicine, Philadelphia, PA 19104, United States

\*Corresponding author: Christopher C. Yang, PhD, College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, United States (chris.yang@drexel.edu)

### Abstract

**Objectives:** To investigate approaches of reasoning with large language models (LLMs) and to propose a new prompting approach, ensemble reasoning, to improve medical question answering performance with refined reasoning and reduced inconsistency.

**Materials and Methods:** We used multiple choice questions from the USMLE Sample Exam question files on 2 closed-source commercial and 1 open-source clinical LLM to evaluate our proposed approach *ensemble reasoning*.

**Results:** On GPT-3.5 turbo and Med42-70B, our proposed ensemble reasoning approach outperformed zero-shot chain-of-thought with self-consistency on Steps 1, 2, and 3 questions (+3.44%, +4.00%, and +2.54%) and (2.3%, 5.00%, and 4.15%), respectively. With GPT-4 turbo, there were mixed results with ensemble reasoning again outperforming zero-shot chain-of-thought with self-consistency on Step 1 questions (+1.15%). In all cases, the results demonstrated improved consistency of responses with our approach. A qualitative analysis of the reasoning from the model demonstrated that the ensemble reasoning approach produces correct and helpful reasoning.

**Conclusion:** The proposed iterative ensemble reasoning has the potential to improve the performance of LLMs in medical question answering tasks, particularly with the less powerful LLMs like GPT-3.5 turbo and Med42-70B, which may suggest that this is a promising approach for LLMs with lower capabilities. Additionally, the findings show that our approach helps to refine the reasoning generated by the LLM and thereby improve consistency even with the more powerful GPT-4 turbo. We also identify the potential and need for human-artificial intelligence teaming to improve the reasoning beyond the limits of the model.

**Key words:** large language model; clinical reasoning; machine reasoning; artificial intelligence.

### Background and significance

The use of artificial intelligence (AI) in healthcare has evolved rapidly, fueled by availability of large datasets, paving the way for developments in predictive modeling using data from the electronic health record (EHR). Natural language processing (NLP) techniques have enabled information extraction from narrative notes, eg, physician and nursing notes, and radiology reports.<sup>1</sup> The advent of generative AI in the form of consumer-facing large language models (LLMs) that are pre-trained on vast amounts of data and able to mimic human language, such as ChatGPT from OpenAI and Gemini from Google, has opened up new possibilities for automating parts of the clinician workflow and enhancing care delivery.

However, because of the sheer potential that generative LLMs appear to offer clinicians and health consumers, it is imperative to be aware of the issues that make using AI in healthcare different from other domains. Prominent among these is the critical role of clinical reasoning in healthcare, and the importance of explainability, particularly considering the

“black-box” nature of most AI systems.<sup>2–4</sup> AI tools are being subjected to close scrutiny, with the US President releasing an executive order addressing safe, secure, and trustworthy use of AI in all domains (<https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/>). Therefore, those who train and develop AI systems for use in healthcare as well as the clinicians who implement them need to ensure that these systems are reliable and trustworthy when integrated into the clinical workflow.

The aim of this study is to evaluate the ability of LLMs to improve their reasoning about and answers to medical questions using an ensemble reasoning approach. We hope to identify in which situations or for what types of questions the LLM may or may not be able to refine its reasoning and performance when provided the opportunity to review diverse reasonings for an answer. We anticipate that articulating an understanding of where LLMs perform well on medical questions can allow more trustworthy and reliable application of these models, and that our approach can be extended to interrogating unstructured

notes to extract information on patient diagnoses, disease stage, prognosis, and other clinical factors, at scale.

## Literature review

In this section, we review the relevant studies on the use and applications of generative LLMs in healthcare and on the ability of LLMs to reason.

### LLMs in healthcare

Multiple studies have evaluated the use of LLMs on healthcare related tasks, including parsing and extracting information from clinical notes,<sup>5</sup> clinical decision-making and support,<sup>6–8</sup> generating clinical documentation,<sup>9–11</sup> and improving and facilitating clinical trials.<sup>12,13</sup> Others have examined the opportunities, applications, and challenges of using LLMs to aid/enhance medical education.<sup>14–16</sup> In medical question-answering, LLMs have been tested on multiple choice questions from a textbook on the Clinical Informatics Board Examination (CIBE),<sup>17</sup> the Membership of the Royal College of General Practitioners Applied Knowledge Test (AKT),<sup>18</sup> AMBOSS and the National Board of Medical Examiners (NBME),<sup>19</sup> trivia questions from the Lancet's Picture Quiz Gallery, specifically on items related to rheumatic diseases,<sup>20</sup> and real-world patient questions related to myopia care<sup>21</sup> and cataract surgery.<sup>22</sup> The studies highlighted report overall results of LLM performance, with some identifying specific areas in which the LLM performs better or worse. For example, Kumah-Crystal et al.<sup>17</sup> observed that ChatGPT performed best in questions about fundamental knowledge and skills and worst in those covering improving care delivery and outcomes. Thirunavukarasu et al.<sup>18</sup> compared performance of ChatGPT 1 and ChatGPT 2 based on subject areas on the AKT exam. In the studies using datasets from the NBME United States Medical Licensing Examination (USMLE), Gilson et al.<sup>19</sup> evaluated performance on Steps 1 and 2 of the Clinical Knowledge exam, concluding that ChatGPT performed at the level of a third-year medical student in questions assessing competency in medical knowledge, but not reporting a deeper analysis of why this may be the case.

### LLM reasoning

Several studies have proposed approaches to evaluate and improve the LLM reasoning process. The Chain-of-Thought (CoT) approach, first proposed by Wei et al.,<sup>23</sup> involves a series of steps, where the model is provided with examples (few-shot) and prompted to reason about the question before providing an answer. This approach was adopted by Kojima et al.<sup>24</sup> who proposed zero-shot CoT, where no examples are provided but rather the LLM is instructed "Let's think step by step." Both approaches showed improvement in performance when compared with standard zero-shot prompting. Furthermore, Wang et al.<sup>25</sup> proposed the self-consistency approach, which involves generating a diverse set of reasoning paths regarding a question by sampling the language model decoder and then determining the best answer. Self-consistency involves 3 steps: (1) prompt the LLM using CoT, (2) generate multiple reasoning paths, and (3) marginalize out the generated paths to choose the most consistent answer.

Building on CoT, Yao et al.<sup>26</sup> proposed the Tree-of-Thoughts (ToT) framework, which explicitly decomposes the intermediate thought steps into smaller units so that the LLM can generate more diverse samples and evaluate them. While ToT performed better than CoT, the authors cautioned that it is resource-heavy and may not be necessary for all tasks.

## Materials and methods

### Data

We use the NBME sample exam questions for the USMLE, which is divided into 3 "Steps." Step 1 questions are "designed to measure basic science knowledge" (<https://www.usmle.org/prepare-your-exam/step-1-materials/step-1-content-outline-and-specifications>), Step 2 questions assess the "...examinee's ability to apply medical knowledge, skills, and understanding of clinical science essential for the provision of patient care under supervision and includes emphasis on health promotion and disease prevention" (<https://www.usmle.org/prepare-your-exam/step-2-ck-materials/step-2-ck-content-outline-specifications>), and Step 3 questions expect the examinee "to assume the role of a general, as-yet undifferentiated physician" (<https://www.usmle.org/prepare-your-exam/step-3-materials/step-3-content-outline-and-specifications>).

The questions are in multiple choice format, the majority being plain text while others contain media elements such as graphs and images. For the plain text format questions, some have complicated formatting, eg, question 25 in Figure 1, that may be difficult for a machine to parse.

For our experiments, we focused on text questions with simple formatting only (eg, question 26 in Figure 1) to ensure that we captured the LLMs true capabilities with textual data when all the required information to answer the question is available in a logical format. Our final dataset comprised 87 Step 1 questions, 100 Step 2 questions, and 118 Step 3 questions.

### Methods

This section provides an overview of the approaches utilized in our study. We also propose a new approach, ensemble reasoning (EnsReas).

#### Zero-shot prompting

Given a problem of interest  $X$ , presented in textual format, prompting may be defined by a function  $f(X)$  that transforms the input  $X$  into a prompt  $X'$  as

$$X' : f(X) \rightarrow X'. \quad (1)$$

The answer generated by the LLM may be denoted as

$$A = llm(X'). \quad (2)$$

#### Zero-shot chain-of-thought (CoT) prompting

Zero-shot CoT<sup>24,27</sup> uses 2 sequential prompts to elicit the reasoning and answer from the LLM.

In the first step,  $X$  is transformed into a prompt  $X'$  using

25. A 19-year-old man who is in the US Army is brought to the emergency department 45 minutes after he sustained a knife wound to the right side of his chest during an altercation. He has no history of major medical illness and takes no medications. His temperature is 36.9°C (98.4°F), pulse is 110/min, respirations are 24/min, and blood pressure is 114/76 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 94%. On physical examination, the trachea appears to be shifted to the left. Pulmonary examination of the right chest is most likely to show which of the following findings?

	<b>Fremitus</b>	<b>Percussion</b>	<b>Breath Sounds</b>
(A)	Decreased	dull	decreased
(B)	Decreased	hyperresonant	decreased
(C)	Decreased	hyperresonant	dull
(D)	Increased	dull	bronchial
(E)	Increased	dull	decreased

26. A 34-year-old man comes to the office because of a 1-month history of diarrhea. He has a history of pheochromocytoma treated 2 years ago. His mother is being treated for a tumor of her parathyroid gland. He has no other history of major medical illness and takes no medications. His temperature is 37.0°C (98.6°F), pulse is 84/min, respirations are 10/min, and blood pressure is 120/75 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 97%. Vital signs are within normal limits. Physical examination shows a 3-cm, palpable mass on the right side of the neck. A biopsy specimen of the mass shows a neuroendocrine neoplasm of parafollicular cell origin. The most likely cause of the findings in this patient is a mutation in which of the following types of genes?

- (A) Cell cycle regulation gene  
(B) DNA mismatch repair gene  
(C) Metastasis suppressor gene  
(D) Proto-oncogene  
(E) Tumor suppressor gene

**Figure 1.** Example of complicated vs simple formatting.

$$f_{\text{reason}}(X, T_{\text{reason}}) = X', \quad (3)$$

where  $T_{\text{reason}}$  is a sentence that triggers the LLM to use a reasoning process, eg, “Let’s think step by step.” with the output

$$Z = llm(X'). \quad (4)$$

In the second step, the LLM is queried for the final answer using

$$A = llm(f_{\text{answer}}(X', Z, T_{\text{answer}})), \quad (5)$$

where  $T_{\text{answer}}$  is the answer triggering sentence, eg, “Therefore the answer is...”

#### Self-consistency

To implement self-consistency, we execute eqns (2) and (5)  $n$  times as demonstrated in *algorithm 1* → Self-consistency (SC). In our experiments, we set  $n = 10$  to achieve a balance between resource

#### Algorithm 1 Self-consistency “expert panel” approach for multiple-choice question-answering with LLM

**Data:** A set of questions  $Q$  with their multiple-choice options  $C$   
**Result:** A dataset  $D$  containing LLM answers and reasonings  $r$

```

1 foreach question  $q \in Q$  do
2   for  $i \leftarrow 1$  to  $n$  do
3     Construct prompt  $X'$  for  $q$  with its options from  $C$ ;
4     Send  $X'$  to LLM and get the answer  $c$  with reasoning  $r$ ;
5     Record  $q, c, r$  in  $D$ ;
6   end
7 end

```

use and obtaining enough diverse reasoning pathways for a robust analysis.

### Ensemble reasoning

We propose the iterative ensemble reasoning (EnsReas) approach, outlined in *algorithm 2* → EnsReas, a prompting method that probes whether, given a question with a combination of correct and incorrect answers and corresponding diverse set of reasonings for each answer, the LLM is able to self-critique and improve its reasoning, and subsequently, its performance.

After executing ZS-CoT with self-consistency (ZS-CoT-SC), each question is categorized into 1 of 2 groups based on whether all answers from the “expert panel” are the same, ie, are consistent (group  $Q^{con}$ ) or not (group  $Q^{inc}$ ). EnsReas only takes as input questions from  $Q^{inc}$ .

Given a set of  $n$  reasonings for a question  $q$  in  $Q^{inc}$  with multiple choice options  $C = \{A, B, C, \dots\}$ , we implement ensemble reasoning using the following steps:

- 1) Group the  $n$  reasonings by chosen answer  $c$  and then concatenate them. This yields, for each question, a set of grouped reasonings  $g_c$ :

$$g_c = \{r | r \text{ is reasoning for choice } c, r \in R, c \in C\}. \quad (6)$$

We denote the set of grouped reasonings  $g_c$  as  $T_{ensemble}$ .

- 2) For all questions in  $Q^{inc}$ , transform  $X$  into a prompt  $X'$  using

$$f_{ensemble}(X, T_{ensemble}) = X'. \quad (7)$$

We then query the LLM:

$$A = llm(X'). \quad (8)$$

We iteratively run *algorithm 2* → EnsReas with each round taking as input the results from previous round until we reach the stop criteria. We again set  $n=10$  as in *algorithm 1* → Self-consistency (SC).

The **stop criteria** for EnsReas is defined based on consistency of responses, ie, the proportion of questions with inconsistent answers after the iteration. It may not be possible for the LLM to reach perfect consistency of responses, and in some cases, continued iterations may cause a degradation in performance, in addition to incurring high resource cost. To mitigate these concerns, we set what we consider to be a “reasonable” early stop threshold, achieved when the number of questions with inconsistent responses reaches less than 10% of the total number of questions (or less than 5% for a high performing LLM like GPT-4 turbo).

### Performance and consistency evaluation

To evaluate performance and consistency, we define the LLM answer as the one chosen by majority vote, and categorize each question based on the distribution of responses into 5 different groups as shown in *algorithm 3* → CatSC. Note that category  $C$ : *Indeterminate* refers to questions where the majority vote is tied and the LLM answer cannot be determined.

## Experiments

### Experiment setup

We use GPT-3.5 turbo (gpt-3.5-turbo-0613) and GPT-4 turbo (gpt-4-1106-preview) via the OpenAI API with temperature set to 1 in both cases. We leave all other model parameters at their default values.

#### Algorithm 2 Prompting with EnsReas

**Data:** A set of questions  $Q^{inc}$  with their multiple-choice options  $C$ , and a dataset  $E$  containing grouped reasonings  $T_{ensemble}$  for each question

**Result:** A dataset  $D$  containing LLM answers and reasonings  $r$

```

1 foreach question  $q \in Q^{inc}$  do
2   for  $i \leftarrow 1$  to  $n$  do
3     Construct prompt  $X'$  for  $q$  with its options and ensembled reasonings  $T_{ensemble}$ ;
4     Send  $X'$  to LLM and get the new answer  $c$  with reasoning  $r$ ;
5     Record  $q, c, r$  in  $D$ ;
6   end
7 end
```

**Algorithm 3** Evaluation of LLM self-consistency responses

**Data:** A dataset  $D$  containing LLM responses and the correct answers for a set of questions

**Result:** Categorization of questions into groups  $\mathcal{A} - \mathcal{E}$

```

1  for each entry in dataset  $D$  do
2      Determine the distribution of answers from the LLM responses;
3      Identify the majority vote answer(s) and the correct answer;
4      if majority vote uniquely matches the correct answer then
5          if all experts chose the correct answer then
6              Categorize as Group  $\mathcal{A}$ : Completely correct;
7          else
8              Categorize as Group  $\mathcal{B}$ : Partially correct;
9          end
10     else
11         if majority vote is tied and includes the correct answer then
12             Categorize as Group  $\mathcal{C}$ : Indeterminate;
13         else
14             if majority vote is incorrect then
15                 if at least one response is correct then
16                     Categorize as Group  $\mathcal{D}$ : Partially incorrect;
17                 else
18                     Categorize as Group  $\mathcal{E}$ : Completely incorrect;
19                 end
20             end
21         end
22     end
23 end

```



**Zero-shot (ZS):** We design a simple prompt to send the question and multiple choice options to the LLM and elicit an answer.

```
system_instruction = "You are a helpful assistant that answers
multiple choice questions about medical knowledge."

prompt = ""
Question: {question}
Choices: {choice}

Please pick the correct answer.
""
```

**Zero-shot CoT (ZS-CoT):** We first present the question and multiple choice options to the LLM, instructing it to “think step by step.” We then prompt the LLM to use its logical step-by-step thinking output to choose the correct answer.

```
system_instruction = "You are a helpful assistant that answers
multiple choice questions about medical knowledge."

first_prompt = ""
Question: {question}
Choices: {choice}
Let us think step by step.
""

second_prompt = ""
Based on the reasoning process, please pick the correct
answer.
""
```

**Self-consistency (ZS-SC and ZS-CoT-SC):** We utilize the “expert panel” approach outlined in *algorithm 1* → Self-consistency (SC). Our approach differs from that described by Wang et al.<sup>25</sup> in that we do not vary the prompt. For ZS with self-consistency (ZS-SC), we capture the answer provided by each “expert,” while for ZS-CoT with self-consistency (ZS-CoT-SC), we capture both the final answer as well as the reasoning provided in response to the “think step by step” prompt.

**Ensemble reasoning (ZS-CoT-SC-EnsReas):** For our proposed EnsReas approach, we use the set of questions with inconsistent answers  $Q^{inc}$  as described in Ensemble reasoning section, with the reasonings grouped by chosen answer, and then integrate them into a new prompt as outlined in *eqns (6) and (7)*. *Figure 2* shows an example of grouped reasonings, where question 12 has consistent answers (all experts chose the answer “B”) while question 13 has inconsistent answers (8 experts chose the answer “C” and 2 experts chose the answer “E”). We only re-prompt the LLM with the questions with inconsistent answers.

```
system_instruction = "You are a helpful assistant who reviews
a variety of responses to multiple choice questions from a panel
of experts, analyzes them critically, and provides ONE final cor-
rect response."
```

```
prompt = ""
Question: {question}
Choices: {choice}
Panel Responses: {panel_responses}
```

You are provided with the question and the multiple choice options. You are also provided with the chosen answers from the panel of experts with the corresponding reasonings. The reasonings provided by the experts are aggregated by chosen answer.

Please review each question and its multiple choice options. Analyze the reasonings provided by the panel for the chosen answers and provide ONE final correct answer with your reasoning.

Keep in mind that the majority vote may not be the correct one, therefore you should review the question and available multiple choices carefully in addition to considering the panel reasonings.

We consider the possibility that based on the aggregated reasonings provided, the LLM may default to the majority vote as the updated answer and therefore instruct the LLM to “Keep in mind that the majority vote may not be the correct one.” In this way, we hope to induce the LLM to “think” more critically rather than resorting to a majority vote heuristic.

We iteratively run EnsReas prompting, ie, ZS-CoT-SC-EnsReas- $j$ , where  $j$  is the iteration number, until we meet the stop criterion defined in the Ensemble reasoning section. After each iteration, we separate the questions into  $Q^{con}$  and  $Q^{inc}$  to determine if the stop criterion has been reached. *Algorithm 3* → CatSC is used post-hoc for evaluation of performance and consistency.

**Qualitative evaluation of LLM reasoning:** To evaluate whether the LLM follows the instruction to not fall upon the majority vote heuristic, we manually review a selection of the reasonings generated by the EnsReas prompt.

To evaluate the quality of the reasonings provided by the LLM with the EnsReas approach, we provide a sample of 29 questions whose answers converged from partially correct or partially incorrect to fully correct to a medical expert (a board-certified physician) to evaluate the LLM reasonings for “correctness” and “helpfulness.” For “correctness,” we instruct the expert to rate, for each question, each of the 10 reasonings on a 4-point Likert-type scale 0-3, where 0 is

```
"12": {
  "B": [
    "First, let's review the patient's symptoms and findings. The patient has urinary hesitancy and dribbling a
    "The patient's symptoms (urinary hesitancy, dribbling after urination, nocturia) are suggestive of benign p
    "The patient's symptoms of urinary hesitancy, dribbling after urination, and nocturia are indicative of ben
    "First, let's analyze the patient's symptoms and physical examination findings. The patient is experiencing
    "First, let's understand the patient's clinical presentation. The 62-year-old man has urinary hesitancy, dr
    "Alright. Let's analyze the information given in the question stem.\n\nThe patient is a 62-year-old man who
    "First, let's review the patient's symptoms. A 62-year-old man presenting with urinary hesitancy, dribbling
    "First, let's go through the given information. The patient is a 62-year-old man with urinary hesitancy and
    "First, let's understand the patient's condition. The patient is experiencing urinary hesitancy (difficulty
    "The symptoms described, such as urinary hesitancy, dribbling after urination, and nocturia (frequent urina
  ]
},
"13": {
  "C": [
    "First, we are presented with an 18-year-old woman with sickle cell disease who presents with severe abdomi
    "Step 1: Assess the patient's symptoms and history.\n\nThe patient is an 18-year-old woman with sickle cell d
    "First, let's analyze the information given in the question. The patient has sickle cell disease, severe ab
    "First, let's consider the patient's clinical presentation. The patient is a young woman with sickle cell d
    "First, we know that the patient has sickle cell disease, which is important to consider because it can lea
    "Step 1: Analyze the clinical presentation\n\nThe patient presents with severe abdominal pain and tenderness (
    "First, we know that the patient has sickle cell disease, which is a genetic disorder affecting the shape o
    "First, let's review the information provided in the question.\n\nThe 18-year-old woman has sickle cell dise
  ],
  "E": [
    "First, we know that the patient has sickle cell disease, which is a genetic disorder that affects the shap
    "First, we need to identify the most likely underlying cause of the patient's condition.\n\nThe patient has
  ]
},
... ..
```

**Figure 2.** Example of reasonings grouped by chosen answer. For Question 12, all experts chose answer “B” while for Question 13, 8 experts chose answer “C” while 2 chose answer “E.”

**Table 1.** Comparison of ZS-CoT prompting with the proposed EnsReas approach using GPT-3.5 turbo and GPT-4 turbo.

	ZS (%)	ZS-SC (%)	ZS-CoT (%)	ZS-CoT-SC (%)	ZS-CoT-SC-Ens-Reas-1 (%)	ZS-CoT-SC-Ens-Reas-2 (%)
<b>GPT-3.5 turbo</b>						
Step 1	49.43	52.87	58.62	68.97	<b>72.41</b>	70.11
Step 2	58.00	59.00	64.00	67.00	<b>71.00</b>	67.00
Step 3	54.24	56.87	70.34	72.88	<b>75.42</b>	75.42
<b>GPT-4 turbo</b>						
Step 1	87.36	87.36	<b>93.10</b>	91.95	<b>93.10</b>	<b>93.10</b>
Step 2	84.00	85.00	88.00	<b>92.00</b>	91.00	<b>92.00</b>
Step 3	85.59	87.29	92.37	<b>93.22</b>	91.53	92.37

**Bold** indicates the best performance.

“completely incorrect,” 1 is “mostly incorrect,” 2 is “mostly correct,” and 3 is “completely correct.” We record the correctness score for each reasoning, and report the mean. For “helpfulness,” we ask the expert to evaluate the LLM reasonings for each step, focusing on whether the LLM reasonings provided any new information and whether they provided guidance in picking the correct answer.

**Evaluation on clinical LLM:** To further validate the effectiveness and performance of the proposed EnsReas approach on a different LLM that uses a different prompt mechanism/template, we run the ZS-CoT-SC and EnsReas experiments using Med42-70B, a smaller open-source clinical LLM (<https://huggingface.co/m42-health/med42-70b>), hosted on

a local server. We quantize Med42-70B to int8 for efficient memory usage and use 4 NVIDIA A40 GPUs.

**Results**

Table 1 presents the results using our proposed EnsReas approach compared to standard ZS, ZS-SC, ZS-CoT, and ZS-CoT-SC approaches. GPT-4 turbo outperforms GPT-3.5 turbo in all experiments, likely due to being a more powerful model and the recency of its training data.

Our proposed EnsReas approach yields the best performance at just 1 iteration on all questions with GPT-3.5 turbo. With GPT-4 turbo, we have mixed results, only noting

improvement with EnsReas in Step 1 questions. On Step 2 questions a second round of EnsReas performs as well as ZS-CoT-SC, and on Step 3 questions EnsReas performs marginally worse than ZS-CoT-SC.

In addition to the impact on performance, we analyze how EnsReas affects the consistency of responses from the LLMs by comparing the distribution of type *B*, *C*, and *D* questions at each iteration. These results are shown in Figure 3 (see corresponding data in Supplementary Table S1). There is a lower proportion of questions of type *B*, *C*, and *D* with GPT-4 turbo, indicating that it has overall better consistency in question answering than GPT-3.5. Because the stop criterion for EnsReas iterations is based on consistency, we note that there may be situations where the model performance peaks before the stop criterion is met. For example, for GPT-3.5 turbo, EnsReas-1 has better performance than EnsReas-2 for Steps 1 and 2, suggesting a trade-off between performance and consistency. However, GPT-4 turbo does not demonstrate this behavior, likely due to having a higher consistency, leading to more stable performance across ZS-CoT-SC and the EnsReas iterations (see Supplementary Table S1).

Notably, in all cases, our EnsReas approach improves the consistency of answers obtained from both LLMs, with a significant reduction in questions of type *B*, *C*, and *D* at each iteration.

On reviewing a random selection of the reasonings generated by the EnsReas prompt, we observe the LLM concedes to the majority vote in some cases (Figure 4), but not in others (Figure 5).

This suggests that our EnsReas prompt may be effective in steering the LLM to refine its reasoning and not default to the majority vote when reviewing the provided reasonings.

Analysis of the evaluation on correctness of the LLM reasonings shows that the reasonings from the EnsReas prompting strategy are scored as “mostly correct” or “completely correct” across all steps, with no reasonings for correct answers being scored as incorrect. The mean (SD) scores are 2.72 (0.33) for Step 1, 2.86 (0.30) for Step 2, and 2.99 (0.03) for Step 3, indicating that EnsReas demonstrated a very high level of correctness in its reasoning, particularly for Step 2 and Step 3 questions. On helpfulness, the qualitative analysis revealed that overall the LLM reasonings provided guidance in the right direction to answer the questions and demonstrated insights beyond basic medical knowledge.

To further validate the performance of EnsReas, Table 2 presents results using the smaller clinical Med42-70B LLM. Overall, the model performance is comparable to GPT-3.5 turbo, outperforming it on Step 1 which tests fundamental medical knowledge. Med42-70B performance patterns are similar to GPT-3.5 turbo with EnsReas outperforming ZS-CoT-SC on all steps at just the first iteration. We similarly note improvements in consistency with EnsReas (see Supplementary Figure S1).

## Discussion

The aim of this study was to investigate whether LLMs can improve their ability to reason about and answer medical questions, using our proposed ensemble reasoning approach EnsReas. By conditioning the LLM on its own diverse set of reasonings specifically for only those questions that were partially correct, indeterminate, or partially incorrect, we found that the LLMs evaluated showed improvement in the

consistency of their responses, leading to better performance, particularly with GPT-3.5 turbo and Med42-70B.

A qualitative evaluation of the LLM reasonings from the EnsReas approach by a board-certified physician shows that the responses directly address the question, reasoning for the answer, and relevant medical concepts. Responses provide information that is pertinent to the question’s context. A chain of logic is often presented around narrowing down the possible answers to the most likely option, comparing it to other options and providing adequate explanations. There are clear transitions between ideas with appropriate medical terminology utilized, making the LLM reasoning easy to follow. Incoherent or disjointed responses are not noticed in the LLM responses. Utilizing medical expertise and cross-referencing LLM responses with established USMLE board exam and medical references including MKSAP and UpToDate, we check the correctness of the LLM reasoning. MKSAP (Medical Knowledge Self-Assessment Program; <https://www.acponline.org/about-acp/who-we-are/history-of-acp/a-brief-introduction-to-the-origins-of-mksap>) is an educational resource provided by the American College of Physicians that offers continuing medical education to internal medicine physicians. UpToDate (<https://www.wolterskluwer.com/en/solutions/uptodate>) is an online resource that offers synthesized up to date medical information based on research, clinical evidence, and best practice, to assist clinicians in decision making. Clinical information relevant to the questions is queried from these resources by an experienced physician and the results compared with the LLMs generated reasoning. We do not find factual errors, misconceptions, or unsupported claims that could undermine the credibility of the response.

Furthermore, the LLM reasonings guide us in the right direction to answer the question. The LLM is able to provide insights that go beyond surface-level medical knowledge. Reasonings that demonstrate a deeper understanding of the underlying concepts and their implications are indicative of the LLM’s proficiency in medical question-answering in the EnsReas setting. When multiple LLM outputs from the EnsReas approach addressing the same board question are compared, reasoning is found to be similar, demonstrating consistency with LLM responses. The length of the reasoning statements does vary but does not impact the correctness of each statement. The LLM responses are ideal for the intended audience: medical board test takers. They align with the expectations and standards of medical professionals taking these exams and effectively address the requirements of the question in the given context. The reasoning provided is similar to what is seen in standard board resources.

These findings suggest that LLMs can improve their performance on medical question answering tasks when given the opportunity to review and refine their own reasonings on questions to which they initially respond with levels of uncertainty in the self-consistency setting. This approach mimics, to some extent, the concept of brainstorming, where exposure to diverse ideas, whether “common” or “moderately creative” has been shown to improve creativity and the ability to process information.<sup>28</sup> The variation in correctness scores across the Steps may suggest that the ensembling of reasoning or brainstorming approach may be more effective for questions that test ability to apply medical knowledge and skills rather than those that test basic foundational knowledge.



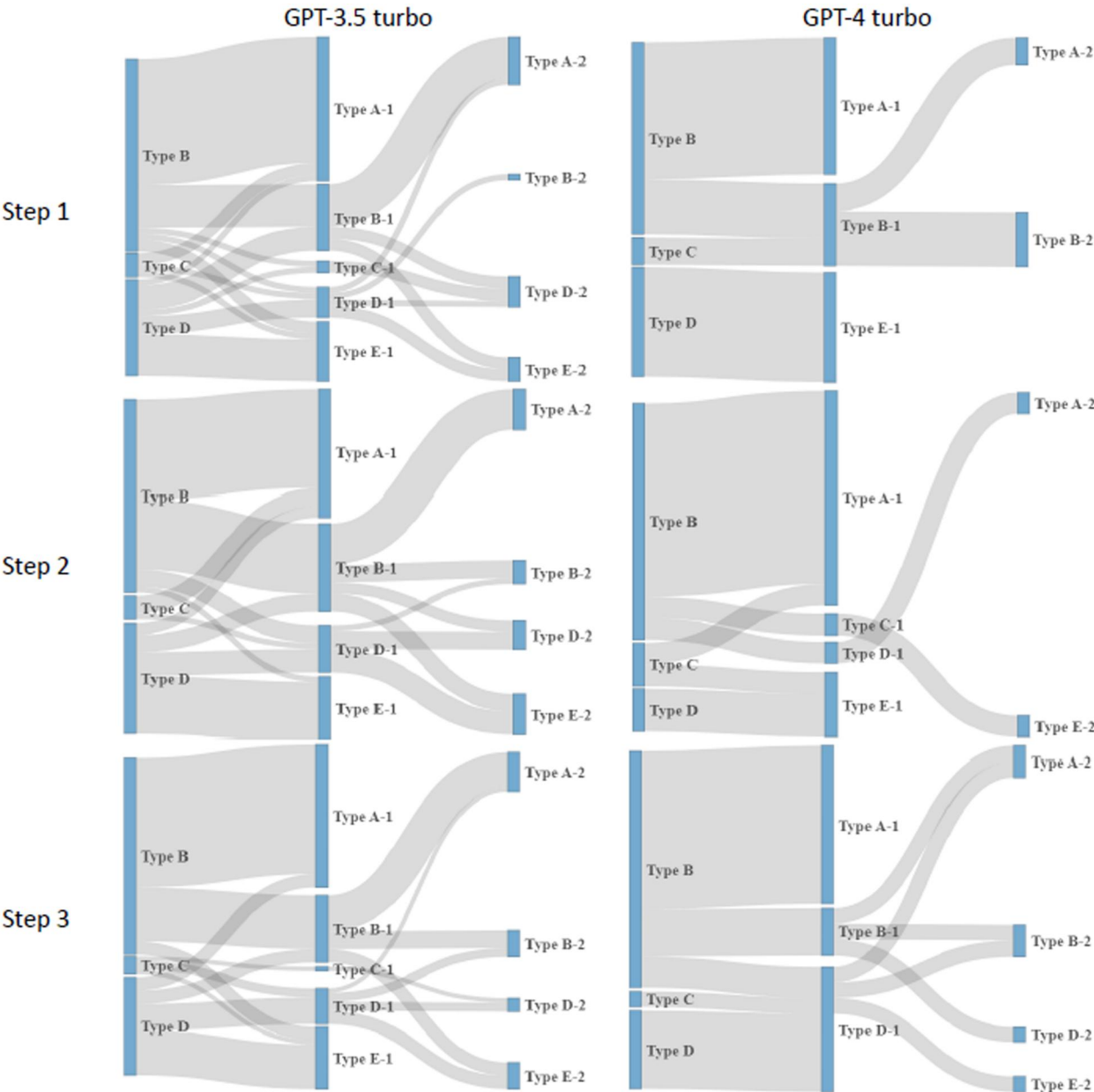


Figure 3. Distribution shift—ZS-CoT-SC to EnsReas-1 to EnsReas-2.

"The majority of the panel of experts agree that the most likely mechanism for increased bleeding risk in this patient with alcohol-induced cirrhosis is decreased production of clotting factors, and specifically, factor VII. The prolonged prothrombin time (PT) and elevated international normalized ratio (INR) suggest a defect in the liver's ability to synthesize vitamin K-dependent clotting factors. Among these, factor VII is the most sensitive marker of hepatic synthetic function since it has the shortest half-life. While inhibition of vitamin K epoxide reductase (Choice D) could contribute to this problem, it is not the primary issue in cirrhosis. The primary problem is the decreased hepatic synthetic capability. The reasonings support a coagulation factor production issue rather than primarily an enzymatic inhibition as would be seen with medications like warfarin. Therefore, the collective expert opinion as well as the clinical information supports choice A: Decreased production of factor VII as the most accurate answer."

Figure 4. EnsReas reasoning exemplar—majority vote agreement.

"The majority of the panel experts chose option E (Vancomycin and cefotaxime) as the most appropriate initial intravenous pharmacotherapy. Their reasoning is based on the assumption that the patient has osteomyelitis and bacterial endocarditis and that this combination covers both *Salmonella* species (common cause of osteomyelitis in sickle cell disease) and *Staphylococcus aureus* (including methicillin-resistant *S. aureus*, MRSA). However, upon reviewing the question and available multiple choices, the correct answer is option D (Nafcillin). The patient's symptoms and examination findings are consistent with acute hematogenous osteomyelitis (AHO), which is the most likely diagnosis. *Staphylococcus aureus* is the most common cause of AHO in children with sickle cell disease. Empiric antibiotic therapy should cover for *Staphylococcus aureus*, including methicillin-resistant *S. aureus* (MRSA), which is often the causative pathogen in this population. Nafcillin is a beta-lactam antibiotic that covers methicillin-susceptible *Staphylococcus aureus* (MSSA) and is the most appropriate initial intravenous pharmacotherapy for this patient. Therefore, the final correct answer is D: Nafcillin."

**Figure 5.** EnsReas reasoning exemplar—change in reasoning.

**Table 2.** Results of the proposed EnsReas approach using Med42-70B.

	ZS-CoT-SC (%)	ZS-CoT-SC-EnsReas-1 (%)	ZS-CoT-SC-EnsReas-2 (%)
<b>Med42-70B</b>			
Step 1	71.26	<b>73.56</b>	<b>73.56</b>
Step 2	66.00	<b>71.00</b>	70.00
Step 3	65.25	<b>66.10</b>	65.25

**Bold** indicates the best performance.

## Strengths

As described in Literature review section, previous studies have evaluated LLMs in medical question answering using various datasets, and their ability to reason about problems. However, there has not been, to our knowledge, a focus on evaluating and improving the certainty of the LLMs generations under self-consistency.

Because our approach is focused only on questions that the LLM was uncertain about, as evidenced by the distribution of answers, we were able to see an improvement in the certainty of the LLM with our EnsReas approach. Additionally, because we only re-prompt the LLM with questions where its responses are inconsistent in the previous iteration, and condition the LLM on concatenated reasonings that have been aggregated based on the selected answer, we increase the possibility that the LLM will focus on inconsistencies in reasoning and refine them. We hypothesize that this approach may reduce the possibility of confabulations as the LLM is not conditioned on wholly incorrect responses.

We include a qualitative analysis that demonstrates high levels of correctness and helpfulness of the reasonings generated from the EnsReas approach. This may translate to more trustworthiness of LLM generated responses in medical question-answering due to the provided reasonings being consistent with expert knowledge.

We further review performance by type of knowledge required of the examinee, as categorized by the NBME, which offers insights into where our approach could be best employed.

We observe that in the ZS setting, GPT-3.5 turbo performs best on Step 2 questions which is consistent with the results reported by Nori et al.<sup>29</sup> where GPT-3.5 zero-shot scored 51.26%, 60.83%, and 58.39% in Steps 1, 2, and 3, respectively. GPT-4 turbo on the other hand performs best in Step 1 questions unlike in Nori et al.<sup>29</sup> where GPT-4 zero-shot scores were 80.67%, 81.67%, and 89.78% on Steps 1, 2, and 3, respectively. We hypothesize that this is due to the recency of GPT-4 turbo training data (April 2023) compared

to GPT-4 (Sep 2021) (<https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>).

Overall, Step 2 questions, which require application of clinical knowledge in the different medical areas, appear more challenging for GPT-3.5 turbo and GPT-4 turbo, and for GPT-4 turbo, the EnsReas approach does not help the LLM to improve its reasoning or performance. For the clinical Med42-70B LLM, Step 3 questions appear to present a higher level of difficulty across the 3 prompting approaches.

Analyzing these questions and the reasonings generated by the LLM may help us identify where human-AI teaming would be most impactful, particularly as LLMs gain wider adoption for health applications. We may identify limits beyond which an LLM even as powerful as GPT-4 turbo is not able to further refine its reasoning or improve its performance, providing an optimum point for human expert intervention. We demonstrate, in Figure 6, a simple experiment executed in the OpenAI playground platform, using GPT-3.5.

The LLM is prompted with a question and asked to pick an answer. The human expert reviews the reasoning provided and then guides the LLM to focus on an important part of the question. With just this simple intervention, the LLM is able to self-correct and refine its reasoning, resulting in the correct answer. This simple example of human-AI teaming demonstrates that injecting human reasoning into the generated outputs where the LLM exhibits flawed reasoning may further improve the performance and reliability of LLMs for clinical tasks.

## Limitations and future work

Due to the resource cost of using the OpenAI LLMs, we restricted our experiments to the USMLE sample exam data because it has been widely tested before using different prompting approaches and LLMs.

Due to the black-box nature of neural networks and consequently of LLMs, we cannot fully understand how they generate their reasonings. This is further complicated by a lack

USER

An 18-year-old woman with sickle cell disease is brought to the emergency department by her parents because of a 2-hour history of severe abdominal pain and nausea. Her parents say that she had a cheeseburger, milk shake, and chocolate bar for lunch. Her temperature is 37.5°C (99.5°F). Physical examination shows tenderness over the right upper quadrant of the abdomen, radiating to the right shoulder. Ultrasonography of the right upper quadrant of the abdomen shows gallstones. Which of the following is the most likely underlying cause of this patient's current condition?  
  
A: Decreased hepatic secretion of lecithin  
B: Decreased reabsorption of bile salts  
C: High ratio of cholesterol to bile acids in bile  
D: Infestation with parasites secreting  $\beta$ -glucuronidase, E: Overload of unconjugated bilirubin

ASSISTANT

C: High ratio of cholesterol to bile acids in bile. The patient's presentation is consistent with acute cholecystitis, which is commonly caused by gallstones. The most common type of gallstones in the United States are cholesterol stones, which form due to an imbalance in the ratio of cholesterol to bile acids in bile. Option A is not a known cause of gallstones. Option B is a cause of intrahepatic cholestasis, but not of gallstones. Option D describes a condition called Gilbert's syndrome, which can cause mild elevations in unconjugated bilirubin levels but is not related to gallstones. Option E describes hemolytic disorders that can cause elevations in unconjugated bilirubin levels, but once again, this is not related to gallstones.

USER

Please reconsider the patients medical history. Does this change your answer?

ASSISTANT

Thank you for the clarification. The patient's sickle cell disease history may change the answer. Patients with sickle cell disease have a higher incidence of pigment gallstones due to the increased hemolysis of red blood cells and the subsequent release of bilirubin. Therefore, the most likely underlying cause of this patient's current condition is E: Overload of unconjugated bilirubin.

Add message

**Figure 6.** GPT-3.5 simple human-AI teaming example.

of knowledge of the exact data used to train the OpenAI GPT models.

In ongoing and future work, we intend to extend our experiments to different medical question-and-answer datasets and LLMs and to investigate more extensively the ability of different LLMs to improve their reasoning and reliability in medical question answering. We will then adapt these techniques to narrative clinical notes, evaluating the ability of LLMs to identify and extract or infer critical elements embedded in the narrative note such as disease status and important clinical findings.

Conclusion

Our study has demonstrated that the proposed iterative EnsReas approach has the potential to improve both the performance and consistency of LLMs in medical question answering tasks. Improvements in performance are most notable in the less powerful GPT-3.5 turbo and Med42-70B, which may suggest that this is a promising approach for smaller LLMs or those that do not do well using standard prompting approaches. The improvements in consistency, and the correctness and helpfulness of reasonings generated by the LLMs suggest that EnsReas has the potential to improve end-user trust of the LLM generations for medical question answering. While GPT-3.5 and GPT-4 are not trained specifically on medical data, Med42-70B is a Llama

model fine-tuned on medical data, demonstrating that EnsReas is able to improve performance on both general and clinical LLMs. In our ongoing and future work, we plan to identify relevant external source of medical data to integrate with the prompt, aiming to improve trustworthiness and reliability. We also plan to develop human-AI teaming approaches that allow human experts to interact with the generative AI models, with the goal of achieving explainable, reliable, and trustworthy outputs. Our current work with EnsReas builds a foundation on how LLMs can resolve inconsistencies and make further improvements.

Acknowledgments

This material is based upon work supported by (while serving at) the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Author contributions

Study conception and design: Mary M. Lucas and Christopher C. Yang; data preparation, experiments, and analyses: Mary M. Lucas, Justin Yang, and Jon K. Pomeroy; draft manuscript preparation: Mary M. Lucas with support from



Christopher C. Yang; manuscript review and revisions: Mary M. Lucas and Christopher C. Yang; overall project supervision: Christopher C. Yang.

## Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## Funding

This work was supported in part by the National Science Foundation under the Grants IIS-1741306 and IIS-2235548 and by the Department of Defense under the Grant DoD W91XWH-05-1-023.

## Conflicts of interest

The authors have no competing interests to declare.

## Data availability

The USMLE sample exam questions used for this study are available on the USMLE public website.

## References

1. Aramaki E, Wakamiya S, Yada S, et al. Natural language processing: from bedside to everywhere. *Yearbook Med Informat.* 2022;31(1):243-253. <https://doi.org/10.1055/s-0042-1742510>
2. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Mach Intell.* 2019;1(5):206-215. <https://doi.org/10.1038/s42256-019-0048-x>
3. Brożek B, Furman M, Jakubiec M, et al. The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artif Intell Law.* 2024;32:427-440. <https://doi.org/10.1007/s10506-023-09356-9>
4. Frank MC. Baby steps in evaluating the capacities of large language models. *Nat Rev Psychol.* 2023;2(8):451-452. <https://www.nature.com/articles/s44159-023-00211-x>
5. Vaid A, Landi I, Nadkarni G, et al. Using fine-tuned large language models to parse clinical notes in musculoskeletal pain disorders. *Lancet Digital Health.* 2023;5(12):e855-e858. [https://doi.org/10.1016/S2589-7500\(23\)00202-9](https://doi.org/10.1016/S2589-7500(23)00202-9)
6. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? *Lancet Infectious Dis.* 2023;23(4):405-406. [https://doi.org/10.1016/S1473-3099\(23\)00113-5](https://doi.org/10.1016/S1473-3099(23)00113-5)
7. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Informat Assoc.* 2023;30(7):1237-1245. <https://doi.org/10.1093/jamia/ocad072>
8. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res* 2023;25(1):e48659. <https://doi.org/10.2196/48659>
9. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digital Health.* 2023;5(3):e107-e108. [https://doi.org/10.1016/S2589-7500\(23\)00021-3](https://doi.org/10.1016/S2589-7500(23)00021-3)
10. Nguyen J, Pepping CA. The application of ChatGPT in healthcare progress notes: a commentary from a clinical and research perspective. *Clin Transl Med.* 2023;13(7):e1324. <https://doi.org/10.1002/ctm2.1324>
11. Preiksaitis C, Sinsky CA, Rose C. ChatGPT is not the solution to physicians' documentation burden. *Nat Med.* 2023;29(6):1296-1297. <https://doi.org/10.1038/s41591-023-02341-4>
12. Ghim J-L, Ahn S. Transforming clinical trials: the emerging roles of large language models. *Transl Clin Pharmacol.* 2023;31(3):131-138. <https://doi.org/10.12793/tcp.2023.31.e16>
13. den Hamer DM, Schoor P, Polak TB, Kapitan D. 2023. Improving patient pre-screening for clinical trials: assisting physicians with large language models. *arXiv*, arXiv:2304.07396 [cs]. <https://doi.org/10.48550/arXiv.2304.07396>
14. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ.* 2023;9(1):e48291. <https://doi.org/10.2196/48291>
15. Safranek CW, Sidamon-Eristoff AE, Gilson A, et al. The role of large language models in medical education: applications and implications. *JMIR Med Educ.* 2023;9(1):e50945. <https://doi.org/10.2196/50945>
16. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health.* 2023;2(2):e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
17. Kumah-Crystal Y, Mankowitz S, Embi P, et al. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? *J Am Med Informat Assoc.* 2023;30(9):1558-1560. <https://doi.org/10.1093/jamia/ocad104>
18. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Med Educ.* 2023;9(1):e46599. <https://doi.org/10.2196/46599>
19. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* 2023;9:e45312. <https://doi.org/10.2196/45312>
20. Venerito V, Puttaswamy D, Iannone F, et al. Large language models and rheumatology: a comparative evaluation. *Lancet Rheumatol.* 2023;5(10):e574-e578. [https://doi.org/10.1016/S2665-9913\(23\)00216-3](https://doi.org/10.1016/S2665-9913(23)00216-3)
21. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *eBioMedicine.* 2023;95:104770. <https://doi.org/10.1016/j.ebiom.2023.104770>
22. Chowdhury M, Lim E, Higham A, et al. Can large language models safely address patient questions following cataract surgery? In: Naumann T, ed. *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Toronto, Canada: Association for Computational Linguistics; 2023:131-137. <https://doi.org/10.18653/v1/2023.clinicalnlp-1.17>
23. Wei J, Wang X, Schuurmans D, et al. 2023. Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, arXiv:2201.11903 [cs]. <https://doi.org/10.48550/arXiv.2201.11903>
24. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. 2023. Large language models are zero-shot reasoners. *arXiv*, arXiv:2205.11916 [cs]. <https://doi.org/10.48550/arXiv.2205.11916>
25. Wang X, Wei J, Schuurmans D, et al. 2023. Self-consistency improves chain of thought reasoning in language models. *arXiv*, arXiv:2203.11171 [cs]. <https://doi.org/10.48550/arXiv.2203.11171>
26. Yao S, Yu D, Zhao J, et al. 2023. Tree of thoughts: deliberate problem solving with large language models. *arXiv*, arXiv:2305.10601 [cs]. <http://arxiv.org/abs/2305.10601>
27. Wang L, Xu W, Lan Y, et al. Plan-and-solve prompting: improving zero-shot chain-of-thought reasoning by large language models. In: Rogers A, Boyd-Graber J, and Okazaki N, eds. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics; 2023: 2609-2634. <https://doi.org/10.18653/v1/2023.acl-long.147>
28. Fink A, Koschutnig K, Benedek M, et al. Stimulating creativity via the exposure to other people's ideas. *Human Brain Mapping.* 2011;33(11):2603-2610. <https://doi.org/10.1002/hbm.21387>
29. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv*, arXiv:2303.13375 [cs]. <http://arxiv.org/abs/2303.13375>



