

### Can large language models reason about medical questions?

Liévin, Valentin; Hother, Christoffer Egeberg; Motzfeldt, Andreas Geert; Winther, Ole

Published in: Patterns

Link to article, DOI: 10.1016/j.patter.2024.100943

Publication date: 2024

Document Version Publisher's PDF, also known as Version of record

### Link back to DTU Orbit

*Citation (APA):* Liévin, V., Hother, C. E., Motzfeldt, A. G., & Winther, O. (2024). Can large language models reason about medical questions? *Patterns*, *5*(3), Article 100943. https://doi.org/10.1016/j.patter.2024.100943

#### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.

- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# **Patterns**

# Can large language models reason about medical questions?

## **Graphical abstract**



## **Highlights**

- InstructGPT can mobilize expert medical knowledge and reasoning skills
- With adequate prompting, both closed- and open-source models pass the USMLE (MedQA)
- LLMs can quantify uncertainty when applied to medical exam multiple-choice questions
- LLMs are affected by the order of the answer options (positional bias)

## Authors

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, Ole Winther

## Correspondence

valentin.lievin@gmail.com

## In brief

Large language models (LLMs) are powerful but their limits are unknown. This paper probes the capabilities of LLMs using medical exam questions (e.g., USMLE). Based on detailed answers generated by LLMs, we show that LLMs can often answer challenging medical questions by mobilizing expert knowledge and advanced reasoning capabilities. Both closed- and opensource LLMs can pass the USMLE (>60% accuracy). Our results support that future LLMs might be applicable to critical realworld applications such as supporting healthcare professionals.





# **Patterns**



### Article

# Can large language models reason about medical questions?

Valentin Liévin,<sup>1,2,6,\*</sup> Christoffer Egeberg Hother,<sup>3</sup> Andreas Geert Motzfeldt,<sup>1</sup> and Ole Winther<sup>1,2,4,5</sup>

<sup>1</sup>Section for Cognitive Systems, Technical University of Denmark, Anker Engelunds Vej 101, 2800 Kongens Lyngby, Denmark <sup>2</sup>FindZebra, Rådvadsvej 36, 2400 Copenhagen, Denmark

<sup>3</sup>Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, Inge Lehmanns Vej 107, 2100 Copenhagen, Denmark <sup>4</sup>Center for Genomic Medicine, Copenhagen University Hospital, Rigshospitalet, Ørestads Boulevard 5, 2300 Copenhagen, Denmark <sup>5</sup>Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark <sup>6</sup>Lead contact

\*Correspondence: valentin.lievin@gmail.com

https://doi.org/10.1016/j.patter.2024.100943

**THE BIGGER PICTURE** Foundation models have changed the way machine learning is practiced. Foundation models applied to text, so-called large language models (LLMs), have proven to be a disruptive technology. They might radically change the way we interact with computers.

In early 2022, it was clear that generalist LLMs can outperform domain-specific approaches in many domains. Benchmarks that reflect real-world scenarios were still needed, and today, it remains unclear how to best use and evaluate these models. This paper probes the capabilities of LLMs using medical exam questions (e.g., USMLE). Based on detailed step-by-step answers generated by LLMs, we show that LLMs can often answer challenging medical questions by mobilizing expert knowledge and advanced reasoning capabilities. Our results support that future LLMs might be applicable to critical real-world applications such as supporting healthcare professionals.

#### SUMMARY

Although large language models often produce impressive outputs, it remains unclear how they perform in real-world scenarios requiring strong reasoning skills and expert domain knowledge. We set out to investigate whether closed- and open-source models (GPT-3.5, Llama 2, etc.) can be applied to answer and reason about difficult real-world-based questions. We focus on three popular medical benchmarks (MedQA-US Medical Licensing Examination [USMLE], MedMCQA, and PubMedQA) and multiple prompting scenarios: chain of thought (CoT; think step by step), few shot, and retrieval augmentation. Based on an expert annotation of the generated CoTs, we found that InstructGPT can often read, reason, and recall expert knowledge. Last, by leveraging advances in prompt engineering (few-shot and ensemble methods), we demonstrated that GPT-3.5 not only yields calibrated predictive distributions but also reaches the passing score on three datasets: MedQA-USMLE (60.2%), MedMCQA (62.7%), and PubMedQA (78.2%). Open-source models are closing the gap: Llama 2 70B also passed the MedQA-USMLE with 62.5% accuracy.

#### INTRODUCTION

Self-supervised pre-training promises to turn vast quantity of raw data (e.g., text, images, audio) into general-purpose models. Language representations have transformed the field of natural language processing from simple word vectors to deep contex-tualized representations,<sup>1–6</sup> and language models are now ubiquitous in natural language processing. Notably, this ubiquity is thanks to the Transformer architecture and its compatibility with massively parallel computation hardware.<sup>4</sup>

#### Large language models (LLMs)

In recent years, tremendous resources have been allocated to scale Transformer-based language models to using hundreds of billions of parameters and to training on gigabytes of text.<sup>7–16</sup> This has so far translated to sustained gains<sup>17</sup> and enabled new ways to interact with language models. This progress made many of the past benchmarks obsolete and sparked a general interest for designing difficult enough benchmarks (e.g., BIG-bench).<sup>18</sup> Pre-train, prompt, and predict<sup>19</sup> is an emerging paradigm for applying LLMs to new problems without fine-tuning





Figure 1. Answering a USMLE (US Medical Licensing Examination) question using zeroshot CoT prompting "Let's think step by step" and InstructGPT Selected example.

the weights on the task. Prompt-based learning consists of augmenting the problem with instructions such that the model's completion of the prompt will correspond to a solution. This allows for LLMs to learn from a few examples (coined shots), which are simply incorporated into the prompts.<sup>7</sup>

#### Chain-of-thought (CoT) prompting

Initially, scaling language models up appeared to benefit more knowledge-intensive tasks than the reasoning-heavy ones.<sup>8</sup> Nevertheless, it was demonstrated that LLMs could be applied to system 2 problems by prompting the model to generate step-by-step solutions, coined CoT.<sup>20</sup> CoT prompting led to substantial improvements on many reasoning-intensive tasks,<sup>20–23</sup> allowing us to bridge the gap with human-level performances for most of the hard BIG-bench tasks.<sup>24</sup> As an alternative to writing reference step-by-step solutions, zero-shot CoT allows for generating CoTs using a single and domain-agnostic cue: "Let's think step by step"<sup>25</sup> (see example generated by InstructGPT<sup>26</sup> in Figure 1). The CoTs that result from that prompt not only appear to expose valid reasoning but also translate into superior zero-shot performances (see example in Figure 1).

#### LLMs and medical applications

Applying LLMs to real-life scenarios will require implementing additional safeguards. Language models may amplify the social biases present in the training data, may hallucinate incorrect facts, and may lack robustness,<sup>27</sup> for instance to adversarial attacks.<sup>28</sup> Therefore, deploying LLMs into sensitive areas such as healthcare must be operated with great care.<sup>29,30</sup> Nonetheless,

LLMs are powerful tools and therefore have the potential to transform the field of machine intelligence. At the dawn of this research work, although LLMs had been tested on large benchmarks (MMLU,<sup>31</sup> BIG-bench<sup>18</sup>), studies applied to the medical domain were still needed. Specialized datasets such as the MedQA-US Medical Licensing Examination (USMLE) enable assessing the capabilities of LLMs in realistic clinical scenarios requiring specialized medical knowledge, advanced reasoning capabilities, and human-level reading comprehension skills.<sup>32</sup>

#### **Related work**

This article-written in three stages (v.1: July 2022, v.2: December 2022, and v.3: September 2023)-evolved along with the remaining of the field. December 2022 was a turning point in machine learning history; new records were achieved on medical benchmarks by the domain-specific Med-PaLM, 33,34 ChatGPT, and GPT-4.35 ChatGPT sparked the interest of the public and the research community, which hastened to benchmark it against USMLE questions,<sup>36,37</sup> turning to self-curated data instead of the peer-reviewed MedQA benchmark. Involving human experts to evaluate the generated explanations on USMLE guestions has also been explored in concurrent work.<sup>33,37</sup> Throughout the development of this research, significant progress happened in the open-source world (Llama 2<sup>38</sup>), and recently, there has been an investigation on both generalist and fine-tuned open-source LLMs applied to medical benchmarks.<sup>39</sup> CoT prompting and ensemble methods are now commonplace in the literature, whereas retrieval augmentation (grounding) remains less common.<sup>33–35,39–41</sup>

	Direct zero-shot	Zero-shot CoT
Question	Question: [Question]	Question: [Question]
CoT	Ø	Answer: Let's think step by step <cot></cot>
Answer	Answer: among A through D, the answer is <answer></answer>	Therefore, among A through D, the answer is <answer></answer>
	Direct zero-shot + grounding	One-shot CoT
Shot	Ø	Question: [Question]
	Ø	Answer: Let's think step by step [Explanation]
	Ø	Therefore, among A through D, the answer is [answer]
Context	Context: [Context]	ø
Question	Question: [Question]	Question: [Question]
CoT	Ø	Answer: Let's think step by step <cot></cot>
Answer	Answer: among A through D, the answer is <answer></answer>	Therefore, among A through D, the answer is* <answer></answer>

#### Figure 2. Prompt templates

In the table, we use typewriter style and brackets to represent [provided data] such as the question, additional context, or the answer and <completions> generated by GPT-3. The symbol  $\varnothing$  represents an empty string.

Table 1. Answering accuracy of leading models against human performance on USMLE (test), MedMCQA (validation), and PubMedQA (test) datasets

1 N N				
Model	Date	USMLE	MedMCQA	PubMedQA
Codex 5-shot CoT <sup>a</sup>	2022	60.2	59.7	78.2
Llama 2 5-shot CoT <sup>a</sup>	2023	62.5	53.6	-
Fine-tuned SOTA	2022	50.3	52.9	78.2
GPT-4	2023	86.1	73.7	<u>81.2</u>
MedPalm v.2	2023	86.5	72.3	77.4
Human (passing score)	-	60.0	50.0	-
Human (expert score)	-	87.0	90.0	78.0

Find an overview of our results in supplemental information section A. <sup>a</sup>Our best methods.

#### Contributions

This paper investigates the performances, interpretability, and limitations of CoT prompting for medical question answering. We utilized the GPT-3.5 series (InstructGPT and Codex). This research was conducted in three rounds; first, using InstructGPT, we investigated variations of zero-shot CoT prompting for medical reasoning (domain-specific CoT cues, retrieval augmentation), looking both at the answering performances and the limitations based on an expert evaluation. In the second round, thanks to the Codex beta program, we investigated how scaling inference-time compute could be applied to challenge both the human baseline and to quantify uncertainty. Last, we benchmarked a range of open-source models. Our contributions are as follows.

- We assess how GPT-3.5 performs on multiple-choice medical board exam question datasets (MedQA-USMLE and MedMCQA) and a medical reading comprehension dataset (PubMedQA) using prompt engineering. We explore zero-/few-shot, direct/CoT, domain-specific CoT cues and retrieval augmentation.
- We propose an evaluation protocol for evaluating generated CoTs (three main categories: reasoning, knowledge, and reading comprehension). A medical expert annotated a subset of CoTs generated by zero-shot InstructGPT and supports that InstructGPT, in many cases, can reason and exploit memorized expert knowledge.



- We demonstrate that scaling inference-time compute enables Codex 5-shot CoT to be well calibrated and to reach passing scores on the three medical datasets.
- We benchmark open-source LLMs on the MedQA-USMLE and MedMCQA.

#### Development

This article has evolved over three distinct versions, each exploring different facets of LLMs:

- v.1, July 2022: investigated InstructGPT (expert evaluation and benchmarking prompting strategies).
- v.2, December 2022: scaled experiments and passed the MedQA-USMLE using Codex.
- v.3, September 2023: evaluated open-source models Llama 2, Vicuna, Guanaco, Falcon, etc.

#### METHOD

This paper explores variations of prompt engineering for medical question answering. The prompt templates are summarized in Figure 2.

#### Zero shot

We studied two classes of prompts: the direct prompt and zeroshot CoT. The direct prompt triggers the model to generate the answer using a single completion step (i.e., "The answer is"), whereas when applying the zero-shot CoT framework, we use a two-step prompting scheme: first, an initial reasoning prompt with a CoT cue (e.g., "Let's think step by step"), the completion of which is the CoT, and second, an extractive prompt, the completion of which is the answer (e.g., "Therefore the answer is"). In the zero-shot CoT setting, this corresponds to the setup described in Kojima et al.,<sup>25</sup> and the direct setting corresponds to Brown et al.<sup>7</sup>

#### **Few-shot**

We experimented with inserting examplars (or shots) of question-answer pairs and question-explanation-answers triplets in the prompts. We built each shot using the zero-shot template, replacing the output with the reference explanations and answers. In the few-shot CoT setting, our setup matches the one from Wei et al.<sup>20</sup>

Table 2. Summary of the medical question answering datasets			
	MedQA-USMLE <sup>32</sup>	MedMCQA <sup>44</sup>	PubMedQA <sup>45</sup>
Answer options	A/B/C/D	A/B/C/D	yes/no/maybe
Questions (train/valid./test)	10,200/1,300/1,300	182,800/4,200/6,100	450/50/500
Words/question	116.6	12.7	253.3
Source (questions)	national medical board examination (US)	AIIMS and NEET PG entrance exams	expert-annotated PubMed abstracts
Words/explanation	41.6	66.2	43.2
Source (explanations)	5 human-written CoTs (sourced from MMLU <sup>46</sup> )	detailed explanations (original dataset)	long answer (original dataset)
valid validation.			

#### Patterns 5, 100943, March 8, 2024 3







Figure 3. Generative process and answer likelihood (ensemble model, i.e., self-consistency)

#### **Answer likelihood**

We denote **x** the answer string, **y** a prompt, and **z** a completion generated from an LLM denoted  $p_{\theta}$ . In the zero-shot setting, sampling  $\hat{\mathbf{z}} \sim p_{\theta}(\mathbf{z}|\mathbf{y})$  is a two-step process (first generate the CoT, then extract the answer) pictured in Table 2. Using a sampling temperature  $\tau$ , *k* completions  $\hat{\mathbf{z}}_1, ..., \hat{\mathbf{z}}_k$  can be sampled from the generative LLMs. We aggregate the completions and estimate the marginal answer likelihood as (Figure 3)<sup>42</sup>:

$$p_{\theta}(\mathbf{x}|\mathbf{y}) \approx \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}[\mathbf{x} \in \widehat{\mathbf{z}}_i], \ \widehat{\mathbf{z}}_1, \dots, \widehat{\mathbf{z}}_k \sim p_{\theta}(\mathbf{z}|\mathbf{y})$$
 (Equation 1)

where  $1[\mathbf{x} \in \hat{\mathbf{z}}_i]$  takes value one when the answer **x** can be matched in the completion  $\hat{\mathbf{z}}$  and otherwise takes zero. Sampling multiple completions may allow exploring multiple hypotheses. Combining multiple sampled CoTs (also known as self-consistency) has also been explored in past work, and showed improvements over single-sample CoT methods.<sup>42,43</sup>

#### **Retrieval augmentation**

LLMs memorize part of the knowledge embedded into the training data; nonetheless, models might fail to reuse this knowledge effectively during prediction. Conditioning the predictions on a knowledge base is an alternative research direction for improving language models.<sup>47–49</sup>

We investigated whether grounding the model with additional context could improve the answering accuracy. We experimented with a simple BM25 retriever and used Wikipedia as a knowledge base. Read more details in supplemental information section G.

#### **EXPERIMENTS**

This section is separated into three parts: (1) introducing the datasets and the GPT-3.5 models, (2) investigating zero-shot medical reasoning with InstructGPT, and (3) scaling inference-time compute with Codex (using longer few-shot prompts and sampling many completions per question).

#### **Resources availability**

#### Lead contact

Further information and requests for code and data should be directed to and will be fulfilled by the lead contact, Valentin Liévin (valentin.lievin@gmail.com).

#### Table 3. Selected domain-specific CoT cues

#### #1 – Let's think step by step

#2 – Let's think step by step like a medical expert

#3 – Let's use step by step inductive reasoning, given the medical nature of the question

#4 – Let's differentiate using step by step reasoning like a medical expert

#5 – Let's derive the differential diagnosis

#### Materials availability

## This study did not generate new unique materials or reagents. **Data and code availability**

Our source code is available on Github (https://github.com/vlievin/medicalreasoning).<sup>50</sup> A collection of generated CoTs, reusable for downstream tasks, are accessible through ToughtSource.<sup>51</sup> All our benchmark results are summarized in supplemental information section A and Table S2.

#### **Datasets and models**

#### Datasets

This study is centered around three medical multiple-choice question answering datasets: USMLE, which includes difficult real-world medical questions targeting medical professionals<sup>32</sup>; the MedMCQA, which gathers questions from medical school entrance exams<sup>44</sup>; and the PubMedQA, which includes reading comprehension questions about PubMed abstracts.<sup>45</sup> The three datasets are summarized in Table 2. For each dataset, we gathered questions with explanations (long answer), which we used as reference CoTs in few-shot learning scenarios. We present the three datasets in further details in supplemental information section C. Furthermore, we compare the MedQA-USMLE with the MMLU-USMLE dataset in supplemental information section D; we found the MedQA questions to be more challenging than the MMLU ones.<sup>31</sup>

#### Models

We study a collection of closed- and open-source models. The 175-billion parameter GPT-3.5 series: the human-aligned GPT-3 (InstructGPT, text-davinci-002<sup>26</sup>) and the code-fine-tuned GPT-3 (Codex, code-davinci-002).<sup>52</sup> A collection of open-source models ranging from 7 to 70 billion parameters: Llama 2,<sup>38</sup> Vicuna,<sup>53</sup> Guanaco,<sup>54</sup> Falcon,<sup>55</sup> MPT,<sup>56</sup> and GPT-NeoX.<sup>57</sup> We used greedy decoding (temperature  $\tau = 0$ ) with k = 1 sample unless specified otherwise (e.g., ensemble methods).

In supplemental information section E, we report the test USMLE accuracy for four GPT-3 versions: a small GPT-3, the largest GPT-3 trained without human alignment, InstructGPT, and Codex. The smaller model text-curie-002 delivered close to random performances, with a maximum accuracy of 27.9%. The non-aligned largest GPT-3 text-davinci-001 scored 40.2%, whereas the largest code pre-trained Codex scored 52.9%, and the code-pre-trained and human-aligned InstructGPT scored 47.1%.

#### Investigating zero-shot reasoning with InstructGPT

In this section, we investigate whether the good generative capabilities of LLMs can be applied to answer medical questions in a zero-shot setting. We investigate variations of the zero-shot CoT framework: using domain-specific CoT cues and augmenting the prompt with Wikipedia passages.

In addition to the original zero-shot CoT cue "Let's think step by step," we tested 29 other domain-specific variations such as "Let's think step by step like a medical expert." The study is available in supplemental information section B. We selected five CoT cues displayed in Table 3. In supplemental information section I, we display CoT samples for more exotic cues such as "Let's follow a Bayesian step by step approach," "Let's work by elimination," and "Let's reflect on each answer option."

#### Zero-shot benchmark

In Table 4, we report the performances of InstructGPT for the direct prompt and the aggregated performances for the five domain-specific CoT cues (Table 3). We explored augmenting the prompts with retrieved Wikipedia passages (grounding) and report the performances of an ensemble model with majority voting.<sup>42</sup>

#### Zero-shot direct

InstructGPT outperformed the domain-specific and fine-tuned BERT baselines on the three datasets. Without BM25 grounding, InstructGPT scored +1.4% on the USMLE questions, +1.0% on the MedMCQA exam questions, and +1.1% on PubMedQA over the best BERT methods.

#### Zero-shot CoT

Without BM25 grounding, the direct prompt remained, on average, a better alternative to the CoT prompts. Performances were lower for each of the considered CoT cues, except in the case of the USMLE dataset, for which half of the CoT prompts resulted in small improvements over the direct prompt (+1.1% using CoT prompt #1 vs. using the direct prompt). Nonetheless, the

Table 4. Zero-shot answering accuracy of InstructGPT (text-davinci-002) on the MedQA-USMLE (test), MedMCQA (valid), and PubMedQA (test) datasets

	<b>a</b>	-			
Model	Grounding	Prompt	USMLE	MedMCQA	PubMedQA
InstructGPT	Ø	direct	46.0	44.0	<u>73.2</u>
InstructGPT	Ø	CoT #1-#5	$46.1 \pm 0.7$	$40.4 \pm 2.2$	$59.9 \pm 3.5$
InstructGPT	BM25	direct	47.3	46.7	-
InstructGPT	BM25	CoT #1-#5	$46.4 \pm 0.7$	42.5 ± 1.7	-
InstructGPT	Ø	ensemble (n = $6$ )	50.0	42.4	70.4
InstructGPT	BM25	ensemble (n = $6$ )	49.3	<u>48.8</u>	-
InstructGPT	Ø + BM25	ensemble (n = 12)	<u>53.1</u>	47.6	-
Fine-tuned BERT	BM25, DPR, $\varnothing$	-	44.6	43.0	72.2
Human (passing score)	-	-	60.0	50.0	-
Human (expert score)	-	-	87.0	90.0	78.0

We report the best fine-tuned BERT-based methods. We tested 5 domain-specific CoT cues (#1–#5) and report the mean performances with standard deviations. Fine-tuned BERT, BioLinkBERT<sup>58</sup>; DPR, dense passage retrieval.<sup>59</sup> When multiple results are aggregated, we report the mean and standard deviation (±).

domain-specific CoT prompts #2-#5 did not significantly outperform the original CoT prompt #1.

#### Knowledge grounding

In an attempt to exploit the good reading comprehension skills of InstructGPT, we explored conditioning the completions on Wikipedia passages. When using the direct prompt, we recorded gains on the USMLE (+1.3%) and on the MedMCQA (+2.7%) datasets, suggesting that retrieval augmentation might be beneficial.

#### Ensemble

Combining the predictions of multiple prompts outperformed the singleprompt predictions, except in the case of the PubMedQA dataset, for which the direct prompt performed exceptionally well. The best performances on the USMLE and MedMCQA datasets were obtained by combining retrieval-augmented prompts and setting a maximum of 53.1% accuracy on the USMLE dataset and 48.8% validation accuracy on the MedMCQA dataset.

#### Expert evaluation of the generated CoTs Protocol

InstructGPT delivered strong performances using zero-shot CoT prompting. In this section, we investigate whether the CoTs are sound and seek to understand better how the model fails and succeeds. We considered three general skills that we expect are required to be mastered to answer medical questions: (1) the ability to perform non-trivial reasoning steps, (2) the ability to recall knowledge that is not provided in the context, and (3) the ability to comprehend the question and the context. Based on the three skills, we defined three success patterns (A, B, C) and three failure patterns (D, E, F).

A subset of 50 CoTs generated based on USMLE questions were annotated by a medical expert (C.E.H.) using the six categories. For each category and each CoT, we reported a match if the pattern could be observed at least once. This means that a CoT can be labeled with both a correct and an incorrect pattern for the same skill. We showcase thirty annotated CoTs (three in Figure 9 and 27 in supplemental information section I).

#### Analysis

We report the frequencies of occurrence for the six patterns in Table 5. We found that most of the questions answered incorrectly triggered generating CoTs that contained reasoning errors (pattern D, 86%) and that exhibited a lack of knowledge (pattern E, 74%). Misunderstanding of the questions or the context was less frequently observed (pattern F, 50%). We observed that CoTs leading to questions answered correctly could still show failure patterns but we also observed that the CoTs leading to incorrect naswers were not entirely incorrect, as 59% contained at least one correct reasoning step and 65% showed proper recall of knowledge. Furthermore, inspecting the CoTs leading to incorrect answers more closely, we found that 47% of those were inconclusive: the model could not narrow down the prediction to a single answer.

#### Answering bias

In Figure 4, we report the frequencies of the USMLE answers and the frequencies of predicted labels (zero-shot InstructGPT) for the direct and CoT prompts. Both prompting schemes led to biased predictive frequencies. Direct prompting led to overestimating labels C and D while underestimating label A. CoT prompting led to underestimating B and C while overestimating label D. We repeated the experiment using randomly permuted labels and observed similar patterns (see supplemental information section F).

#### Scaling inference-time compute with codex

In the second round of experiments, we investigated whether using more inference-time compute, thanks to the Codex beta program, could be utilized to obtain better performances and more interpretable outputs. Codex enables using longer prompts, so we used 5-shot prompts and experimented with sampling k = 100 completions with temperature  $\tau = 0.5$  for each question.

Table 5. Frequency of observed patterns (A, B, C, D, E, F) identified among 50 CoTs generated by InstructGPT with temperature $ au$ = 0				
	Pattern	Correct, % (16)	Incorrect, % (34)	Total, % (50)
A	correct reasoning step	94 (15)	59 (20)	<b>70</b> (35)
в	correct recall of knowledge	87 (14)	65 (22)	<b>72</b> (36)
С	correct reading comprehension	100 (16)	85 (29)	<b>90</b> (45)
D	incorrect reasoning step	12 (2)	86 (29)	<b>62</b> (31)
E	incorrect or insufficient knowledge	25 (4)	74 (25)	<b>58</b> (29)
F	Incorrect reading comprehension	6 (1)	50 (17)	<b>36</b> (18)

The CoTs are generated based on USMLE questions and using the CoT prompts #1-#5 (Table 3). We report the frequencies of CoTs leading to correct and incorrect predictions along with the total.







Figure 4. Frequencies of USMLE answers and InstructGPT (text-davinci-002) predictions for direct and CoT prompts (no grounding, zero-shot)

We report question-answering performances and results on uncertainty quantification.

#### Codex 5-shot CoT: Sampling and combining multiple CoTs

In Figure 5, we report the performances of Codex 5-shot CoT given subsets of k' < k CoTs. We report the best fine-tuned models and the human baseline. Increasing the budget of samples yields better results.<sup>42</sup> Using an ensemble of the *k* samples, Codex 5-shot CoT reaches the passing score on the three tasks (see Table 1): the USMLE dataset (60.2%  $\geq$  60%), the MedMCQA dataset (62.7%  $\geq$  50%), and the PubMedQA dataset (78.2%  $\geq$  78%). Additional results, including performances in zero-shot settings, are available in Table S2 (supplemental information section A). Although Codex performed exceptionally well with 5 shots, Codex yield feeble performances with zero-shot CoT; inspecting the generated CoTs revealed lesser-quality samples (supplemental information section I).

#### **Uncertainty quantification**

We investigate the answering likelihood Equation 1 given by Codex 5-shot CoT with k = 100 samples. In Figure 6, we report the maximum probability assigned by the model for correctly vs. incorrectly answered questions along with the calibration plots for the three datasets. Codex 5-shot CoT appears to be overall calibrated, although the calibration is worse for the PubMedQA dataset.

#### Benchmarking open-source models

In the rapidly evolving landscape of LLMs, a prevalent question is the performance gap between open-source and closed-source models. Our study focused on the capabilities of InstructGPT and Codex. Given a budget of 2.000 GPU hours (NVIDIA A100), we benchmarked a range of open-source LLMs, with parameter sizes ranging from 7 to 70 billion, against the 175-billion-parameter Codex. In Figure 7, we report the predictive performances, calibration plots, and biases for Llama 2, Vicuna 1.5, and Codex using up to k = 100 CoT samples. We provided additional results in Figure 8 in supplemental information section H (zero- and 5-shot, MedQA-USMLE, and MedMCQA).

#### DISCUSSION

60

55

50

Accuracy (%)

#### Zero-shot LLMs outperform fine-tuned BERT

Zero-shot InstructGPT and Codex outperformed fine-tuned BERT models on three challenging question-answering datasets (see zero-shot benchmark and supplemental information section A). In the case of the USMLE and the MedMCQA datasets, the

> USMLE MedMCQA PubMedQA 60.0 57.5 78 55.0 77 Human level 52.5  $Codex (\tau = 0.5)$ 76  $Codex(\tau = 0)$ 50.0 Finetuned baseline 75 6 6 10 20 61 80 23600 NA NA 236.0

retrieval-augmented BERT baselines were outperformed by several LLMs, regardless of augmenting the prompts with Wikipedia passages. This suggests that LLMs, without fine-tuning, can mobilize medical knowledge and problem-solving skills.

# Zero-shot CoT prompting often yields sound and interpretable step-by-step solutions

For both InstructGPT and Codex, single-sample CoT prompting was not found to be competitive with direct prompting (see zero-shot benchmark and supplemental information section A). Nevertheless, CoTs are human readable and therefore interpretable. Our expert evaluation (see expert evaluation of the generated CoTs) revealed that CoTs are often sound: even though InstructGPT still does make mistakes, it was often able to reason, recall medical knowledge, and comprehend the given problem. In the section investigating zero-shot reasoning with InstructGPT and supplemental information section B, we explored domain-specific CoTs cues such as "Let's think step by step like a medical expert." Although such prompts, taken separately, did not outperform the original zero-shot CoT prompt (see Table S2 in supplemental information section A), more specific prompts appeared to trigger alternative strategies such as working by elimination or manipulating equations (see supplemental information sections B and I). Investigating whether a task-specific prompt could help solve specific tasks will be left for future research. A collection of generated CoT samples are presented in supplemental information section I, and many more samples are available on our GitHub page.

#### LLMs memorize some expert knowledge

The expert evaluation of the generated CoTs (see expert evaluation of the generated CoTs) and the good results obtained on the medical exam questions (see Table S2; supplemental information section A) suggest that GPT-3.5 memorizes domain knowledge. Nevertheless, despite the simplicity of the BM25 retriever and the small number of retrieved documents prepended in each prompt, grounding InstructGPT resulted in slight improvements (see Table 4). This suggests that InstructGPT is not omniscient, and so (1) using stronger retrievers such as commercial search engines or dense retrievers,<sup>49</sup> (2) using a more complete knowledge base,<sup>48</sup> or (3) leveraging inference-time compute by retrieving, reranking, and processing more passages<sup>49</sup> might improve performances.

#### Bias

In the section answering bias, we exposed the biases induced by the use of direct and CoT prompts. In the case of the direct

## Figure 5. Sampling and combining multiple CoTs

Answering accuracy of Codex 5-shot CoT (codedavinci-002) on the USMLE (test), the MedMCQA (validatuin), and the PubMedQA (test) datasets for 100 CoTs sampled with temperature  $\tau \in \{0, 0.5\}$ . We report the average accuracy for ensemble models evaluated using random subsets of k' = 1...100 CoTs. We report the mean and standard deviation. We display the performances of the best fine-tuned methods along with the lower human baselines.







#### Figure 6. Uncertainty quantification

First row: distribution of the probability assigned to the correct label for correct predictions and incorrect predictions (see Equation 1). Second row: calibration plot. The probabilities are obtained using Codex 5-shot CoT and an ensemble of k = 100predictions sampled with temperature  $\tau = 0.5$ .

prompt, answer D was most often selected, which might be due to its proximity to the generated answer. In the case of the CoT prompts, labels A and D were selected more often, which might be a result of often beginning CoTs with content related to option A. Based on an inspection of the CoTs, we speculate that GPT-3 defaults to this behavior when it cannot answer but still attempts to complete the prompt with a default answer (D or A). Shuffling the answer options might be one way to overcome this limitation; however, other forms of biases might still be present.

# Generating and combining many CoTs bridges the gap with human-level performances

CoTs can be combined and/or filtered using human or automated feedback.<sup>42,60</sup> In the section scaling inference-time compute with Codex, we showed that sampling and combining up to k = 100 completions using Codex or Llama 2 with 5-shot CoT prompts was sufficient to pass both the MedMCQA and the challenging USMLE, although a large gap remains between our models and the human experts.

#### 5-Shot CoT-prompted LLMs are close to well calibrated

In the sections uncertainty quantification and benchmarking open-source models, we looked at the probability assigned to correct and incorrect predictions using the ensemble model from Equation 1. We found Codex and Llama 2 to be close to well calibrated, corroborating the results that "language models (mostly) know what they know."<sup>61</sup>

#### Scale, code pre-training, human-alignment, and fewshot learning

In supplemental information section E, we compared multiple GPT-3 models in the zero-shot setting. Best performances are obtained using Codex, outperforming the human-aligned

InstructGPT, which is a fine-tuned version of Codex. Human alignment might impair performances; Codex (without alignment) was not as robust as InstructGPT (with alignment) in zero-shot CoT setting (see performances in Table S2 in supplemental information section A and see CoT samples in supplemental information section I). Nevertheless, 5-shot prompting allowed us to bypass the zero-shot limitations of Codex. We observed a similar pattern when comparing the versions of LLama-2 70b: the base version outperformed the chat version (supplemental information section H). Instruction-fine-tuned models might lose in-context learning abilities.

# Open-source models narrow the gap with proprietary counterparts

Open-source models, despite having fewer parameters, are approaching the performance of proprietary ones (Figures 7 and 8). For instance, Llama 2 outperforms Codex with just half the parameters.

Instruction-fine-tuned LLMs like Guanaco and Vicuna performed exceptionally well (Figure 8). Surprisingly, Vicuna 1.5 13B's superior performance to both Llama 2 versions underscores the significance of high-quality datasets for instructionbased fine-tuning.<sup>62</sup>

#### Conclusion

We applied zero-shot, few-shot direct, and CoT prompting to medical question answering with and without retrieval augmentation. Zero-shot InstructGPT significantly outperformed the fine-tuned BERT baselines. CoT prompting proved to be a powerful tool leading to better performances and more interpretable predictions. Our expert evaluation suggests that LLMs can mostly comprehend complex medical questions, can often recall expert-domain knowledge, and can often perform non-trivial reasoning steps.



Figure 7. Comparing open-source LLMs against the closed-source Codex on the MedQA-USMLE benchmark ( $\tau = 0.9$ , up to k = 100 samples)

We report answering accuracy, model calibration, and answering bias.



#### Figure 8. MedQA-USMLE accuracy vs. model size

All experiments were performed using a 5-shot CoT prompting strategy and greedy decoding ( $\tau = 0$ ). Llama 2 70B outperforms Codex 175B (proprietary).

Although InstructGPT and Codex still make mistakes, we found that scaling inference-time compute by sampling many CoTs per question could overcome part of these limitations. With 100 samples, Codex 5-shot CoT delivered unprecedented performances on the three datasets, bridging the gap with human-level performances and virtually passing the USMLE by 0.2% points. Our exploration into open-source LLMs indicated their competitive stance in medical benchmarks. Llama 2 outperformed Codex by 2 points on the USMLE in spite of a much smaller parameter footprint.

However, deploying LLMs in real-life clinical scenarios will require the development of more robust techniques. We exposed one form of bias (ordering of the answer options affects the predictions), but many more might affect predictions, including those hidden in the training data (e.g., gender, race, ...). Nevertheless, a lack of knowledge might be more easily compensated; our experiment with BM25, albeit limited, suggests that augmenting the prompt with factual data improves performances.

Since the completion of v.2 of this work, both GPT-4 and MedPalm 2 have achieved performance on USMLE around 85%.<sup>35,63</sup> This is not unexpected given the evolution the LLM field has witnessed recently. Although benchmark contamination in training sets for both proprietary and open-source LLMs is a valid concern, these results indicate that both open- and closed-source LLMs hold great potential for assisting human decision-making in medicine and beyond.

#### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j. patter.2024.100943.

#### ACKNOWLEDGMENTS

We thank OpenAl for granting access to the Codex beta program. We acknowledge the EuroHPC Joint Undertaking for awarding us access to MeluXina at LuxProvide, Luxembourg. V.L.'s work was funded in part by Google DeepMind through a PhD grant. O.W.'s work was funded in part

by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). V.L., A.G.M., and O.W. acknowledge support from the Pioneer Center for AI, DNRF grant number P1.

Patterns

#### **AUTHOR CONTRIBUTIONS**

Conceptualization, V.L., C.E.H., and O.W.; methodology, V.L. and O.W.; software, V.L. and A.G.M.; investigation, V.L. and A.G.M.; writing – original draft, V.L.; writing – review & editing, all authors; data curation, C.E.H.; funding acquisition, O.W.; supervision, O.W. and V.L.

#### **DECLARATION OF INTERESTS**

The authors declare no competing interests.

#### DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors used GPT-3.5 and ChatGPT to help reformulate paragraphs in the writing process. After using generative technologies, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Received: December 29, 2022 Revised: March 18, 2023 Accepted: February 2, 2024 Published: March 1, 2024

#### REFERENCES

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 26.
- Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Association for Computational Linguistics), pp. 1532–1543. https://doi. org/10.3115/v1/D14-1162.
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.05365.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at arXiv. https://doi.org/10.04805/arXiv.1810.04805.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. https:// cdn.openai.com/research-covers/language-unsupervised/ language\_understanding\_paper.pdf.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al. (2020). Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33.
- Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, H.F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., et al. (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. Preprint at arXiv. https://doi.org/10.48550/arXiv. 2112.11446.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. J. Mach. Learn. Res. 24, 1–113.
- Thoppilan, R., Freitas, D.D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H., Jin, A., Bos, T., Baker, L., Du, Y., et al. (2022). LaMDA: Language



Figure 9. (Sample 1) Generated zero-shot

CoT from InstructGPT text-davinci-002 for

three CoT prompts on a sample for the

MedQA-USMLE test set

USMLE [test:731] Three days after induction chemotherapy, a 9-year-old boy is admitted to the hospital for fever and neutropenia. He was initially diagnosed with t(9;22)-positive acute T cell lymphoblastic leukemia following workup for fatigue and pancytopenia. He is fully conscious. His blood pressure is 110/65 mm Hg, temperature is 39.5°C (103.1°F), pulse is 98/min, and respiratory rate is 14/min. Physical examination including examination of the central venous catheter site shows no source of infection. Broad-spectrum antibiotics are initiated. Which of the following factors about this patient most strongly indicates a poor prognosis?

A) Age B) Cytogenetics C) Fever and neutropenia D) Initial presentation

✓ Let's use step by step in-	✗ Let's think step by step	✗ Let's derive the differen-
ductive reasoning, given the	like a medical expert. The	tial diagnosis step by step:
medical nature of the ques-	first step is to gather informa-	1) The patient has fever and
tion.	tion about the patient. In this	neutropenia.
The first thing that stands out	case, we are given the follow-	2) The patient is 9 years old.
is that the patient is 9 years	ing information: -The patient	3) The patient has t(9;22)-
old. In general, children have	is 9 years old -The patient has	positive acute T cell lym-
a better prognosis than adults	t(9;22)-positive acute T cell	phoblastic leukemia.
when it comes to leukemia.	lymphoblastic leukemia -The	4) The patient's initial presen-
The next thing to consider	patient is currently experienc-	tation was fatigue and pancy-
is the patient's cytogenetics.	ing fever and neutropenia	topenia.
The fact that the patient has	The next step is to analyze the	The most likely cause of the
t(9;22)-positive acute T cell	information and look for any	patient's fever and neutrope-
lymphoblastic leukemia indi-	patterns or clues that could	nia is infection. The patient's
cates that the patient has a	help us answer the question.	age, cytogenetics, and initial
specific chromosomal abnor-	In this case, the most im-	presentation are all risk fac-
mality that is associated with	portant factor that indicates a	tors for infection.
a poor prognosis.	poor prognosis is the patient's	Therefore, the factor that
The next thing to consider	age. Children under the age of	most strongly indicates a poor
is the patient's fever and neu-	10 have a much higher chance	prognosis is the patient's age.
tropenia. Fever and neutrope-	of relapsing after treatment	Therefore, among A
nia are both signs of infection,	and a lower chance of achiev-	through D, the answer
and infection is a serious com-	ing remission.	is A.
plication in leukemia patients,	Therefore, among A	
particularly in those who are	through D, the answer	
immunocompromised. Fi-	is A.	
nally, we need to consider		
the patient's initial presenta-		
tion. The fact that the patient		
was initially diagnosed with		
leukemia following workup		
for fatigue and pancytope-		
nia indicates that the patient's		
leukemia was advanced at the		
time of diagnosis.		
All of these factors indicate		
a poor prognosis for this pa-		
tient.		
Therefore, among A		
through D, the answer		
is B, cytogenetics.		

Pattern A	Correct reasoning step	Pattern D	Incorrect reasoning step
Pattern B	Correct recall of knowledge	Pattern E	Incorrect or insufficient knowledge
Pattern C	Correct reading comprehension	Pattern F	Incorrect reading comprehension

Models for Dialog Applications. Preprint at arXiv. https://doi.org/10. 48550/arXiv.2201.08239.

Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv. 2205.01068.

- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.D.L., Hendricks, L.A., Welbl, J., Clark, A., et al. (2022). Training Compute-optimal Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.15556.
- Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al. (2022). Using DeepSpeed and Megatron to Train Megatron-turing NLG 530B, A Largescale Generative Language Model. Preprint at arXiv. https://doi.org/10. 48550/arXiv.2201.11990.
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al. (2022). OPT: Open Pre-trained Transformer
- Lieber, O., Sharir, O., Lenz, B., and Shoham, Y. (2021). Jurassic-1: Technical details and evaluation. White Paper Al21 Labs *1*, 9.
  Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch transformers: Scaling
- to trillion parameter models with simple and efficient sparsity. J. Mach. Learn. Res. 23, 5232–5270.
- 16. Laurençon, H., Saulnier, L., Wang, T., Akiki, C., Villanova del Moral, A., Le Scao, T., Von Werra, L., Mou, C., González Ponferrada, E., Nguyen, H., et al. (2022). The BigScience ROOTS Corpus: A 1.6TB composite multilingual dataset. Adv. Neural Inf. Process. Syst. 35, 31809–31826.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., et al. (2020). Scaling Laws for Neural



Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv. 2001.08361.

- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. Preprint at arXiv. https://doi.org/10.48550/arXiv. 2206.04615.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. 55, 1–35.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. 35, 24824–24837.
- Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., and Chi, E. (2022). Least-to-most Prompting Enables Complex Reasoning in Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2205.10625.
- Drozdov, A., Schärli, N., Akyürek, E., Scales, N., Song, X., Chen, X., Bousquet, O., and Zhou, D. (2022). Compositional Semantic Parsing with Large Language Models. Preprint at arXiv. https://doi.org/10.48550/ arXiv.2209.15003.
- Nye, M.I., Andreassen, A.J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. (2021). Show Your Work: Scratchpads for Intermediate Computation with Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv. 2112.00114.
- Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H.W., Chowdhery, A., Le, Q.V., Chi, E.H., Zhou, D., and Wei, J. (2022). Challenging BIG-bench Tasks and Whether Chain-of-thought Can Solve Them. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.09261.
- Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Adv. Neural Inf. Process. Syst. 35, 22199–22213.
- 26. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., et al. (2022). Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 35, 27730–27744.
- 27. Bender, E.M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 610–623.
- Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., Awadallah, A.H., and Li, B. (2021). Adversarial GLUE: A Multi-task Benchmark for Robustness Evaluation of Language Models. Preprint at arXiv. https:// doi.org/10.48550/arXiv.2111.02840.
- Korngiebel, D.M., and Mooney, S.D. (2021). Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit. Med. 4, 93.
- 30. Sezgin, E., Sirrianni, J., and Linwood, S.L. (2022). Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. JMIR medical informatics, 10.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv. 2009.03300.
- 32. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., and Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Appl. Sci. *11*, 6421.
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. Nature 620, 172–180.
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. (2023). Towards Expert-level



Medical Question Answering with Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.09617.

- Nori, H., King, N., McKinney, S.M., Carignan, D., and Horvitz, E. (2023). Capabilities of GPT-4 on Medical Challenge Problems. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.13375.
- 36. Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A., and Chartash, D. (2023). 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. 9, e45312.
- Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., and Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for Al-assisted medical education using large language models. PLOS Digit. Health *2*, e0000198.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama
  Open Foundation and Fine-tuned Chat Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.09288.
- Chen, Z., Hernández-Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., et al. (2023). MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.16079.
- Wang, Y., Ma, X., and Chen, W. (2023). Augmenting Black-box LLMs with Medical Textbooks for Clinical Question Answering. Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.02233.
- Liévin, V., Motzfeldt, A.G., Jensen, I.R., and Winther, O. (2023). Variational open-domain question answering. In International Conference on Machine Learning (PMLR), pp. 20950–20977.
- Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., and Zhou, D. (2022). Self-consistency Improves Chain of Thought Reasoning in Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.11171.
- Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J., and Chen, W. (2022). On the Advance of Making Language Models Better Reasoners. Preprint at arXiv. https://doi.org/10.48550/arXiv.2206.02336.
- Pal, A., Umapathi, L.K., and Sankarasubbu, M. (2022). Medmcqa: A largescale multi-subject multi-choice dataset for medical domain question answering. PMLR *174*, 248–260.
- Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., and Lu, X. (2019). PubMedQA: A Dataset for Biomedical Research Question Answering. Preprint at arXiv. https://doi.org/10.48550/arXiv.1909.06146.
- Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., et al. (2022). Scaling Instruction-finetuned Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.11416.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T., et al. (2020). Retrievalaugmented generation for knowledge-intensive nlp tasks. Adv. Neural Inf. Process. Syst. 33, 9459–9474.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G.B., Lespiau, J.B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. PMLR *162*, 2206–2240.
- Lazaridou, A., Gribovskaya, E., Stokowiec, W., and Grigorev, N. (2022). Internet-augmented language models through few-shot prompting for open-domain question answering. Preprint at arXiv. https://doi.org/10. 48550/arXiv.2203.05115.
- Liévin, V. (2023). Vlievin/Medical-Reasoning: Official Release (Zenodo). https://doi.org/10.5281/zenodo.10301874.
- Ott, S., Hebenstreit, K., Li'evin, V., Hother, C.E., Moradi, M., Mayrhauser, M., Praas, R., Winther, O., and Samwald, M. (2023). ThoughtSource: A central hub for large language model reasoning data. Sci. Data, 10.
- 52. Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating





Large Language Models Trained on Code. Preprint at arXiv. https://doi.org/10.48550/arXiv.2107.03374.

- Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. Preprint at arXiv. https://doi.org/10.48550/arXiv. 2306.05685.
- Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.14314.
- Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., et al. (2023). Falcon-40B: an open large language model with state-of-the-art performance. https://falconllm.tii.ae.
- Team, M.N. (2023). Introducing Mpt-30b: Raising the Bar for Open-Source Foundation Models. www.mosaicml.com/blog/mpt-30b. (Accessed 22 June 2023). Accessed.
- Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al. (2022). GPT-NeoX-20B: An Open-source Autoregressive Language Model. Preprint at arXiv. https:// doi.org/10.48550/arXiv.2204.06745.

- Yasunaga, M., Leskovec, J., and Liang, P. (2022). LinkBERT: Pretraining Language Models with Document Links. Preprint at arXiv. https://doi. org/10.48550/arXiv.2203.15827.
- Karpukhin, V., Oguz, B., Min, S., Wu, L., Edunov, S., Chen, D., and Yih, W.T. (2020). Dense Passage Retrieval for Open-domain Question Answering. Preprint at arXiv. https://doi.org/10.48550/arXiv.2004.04906.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2021). Training Verifiers to Solve Math Word Problems. Preprint at arXiv. https://doi.org/ 10.48550/arXiv.2110.14168.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2022). Language Models (Mostly) Know What They Know. Preprint at arXiv. https://doi. org/10.48550/arXiv.2207.05221.
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al. (2023). LIMA: Less Is More for Alignment. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.11206.
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. (2023). Towards Expert-level Medical Question Answering with Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.09617.