

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

Anonymous ACL submission

Abstract

Large Language models (LLMs) have demonstrated significant potential in transforming healthcare by automating tasks such as clinical documentation, information retrieval, and decision support. In this aspect, carefully engineered prompts have emerged as a powerful tool to use LLMs for medical scenarios, *e.g.*, patient clinical scenarios.

In this paper, we propose a modified version of the MedQA-USMLE dataset, which is subjective, to mimic real-life clinical scenarios. We explore the Chain of Thought (CoT) reasoning based on subjective response generation for the modified MedQA-USMLE dataset with appropriate LM driven forward-reasoning for correct responses to the medical questions. Keeping in mind the importance of response verification in the medical setting, we utilize a reward training mechanism, whereby for a particular response to a clinical question, the language model also provides an appropriate verified response. In this regard, we also include human-in-the-loop for different evaluation aspects. We develop better in-contrast learning strategies by modifying the 5-shot-codex-CoT-prompt from (Liévin et al., 2022) for subjective MedQA dataset and developing our incremental-reasoning prompt. Our evaluations show that the incremental-reasoning prompt performs better than the modified-codex prompt in certain scenarios. We also show that greedy decoding with the incremental-reasoning method performs better than other strategies, such as prompt chaining and eliminative reasoning.

1 Introduction

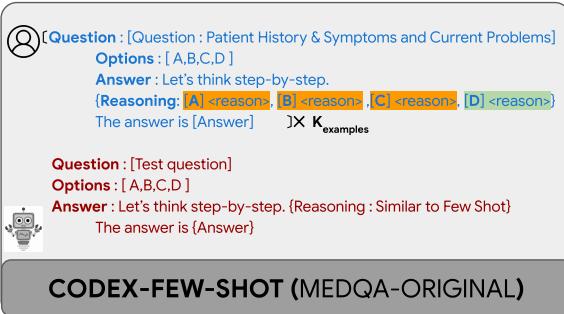
Large Language models (LLMs), are increasingly being utilized in the healthcare sector, particularly for patient query-related tasks. These LLM driven tools could potentially interpret and respond to patient inquiries, provide information on symptoms, diseases, treatments, and healthcare guidelines (Thirunavukarasu et al., 2023). By analyzing

vast amounts of medical literature and data, LLMs could also offer precise, up-to-date responses, improving patient education and engagement (Singhal et al., 2022). The ability of LLMs to understand and process natural language queries makes them accessible and user-friendly, thus enhancing patient experience and satisfaction (Clusmann et al., 2023). As technology evolves, LLMs are expected to play a pivotal role in delivering personalized healthcare information, contributing to informed decision-making and better health outcomes (Clusmann et al., 2023). However, with the recent advances in prompt engineering techniques in Large Language Models space, there is an underlying requirement for accuracy and verifiability. In recent years, several methods in the space of verifiable LLMs have been proposed such as Fact verification LLMs and COT based LLMs.

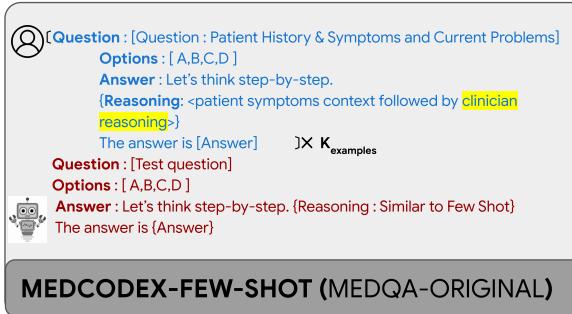
Fact verification LLMs: Fact verification in large language models (LLMs) is an evolving field that spans multiple disciplines, including computational linguistics, artificial intelligence, and digital media. (Guo et al., 2022) provides a comprehensive overview of automated fact-checking, defining it through stages such as claim detection, evidence retrieval, and claim verification. This framework emphasizes the multifaceted nature of fact-checking, which involves not only the assessment of a claim's veracity but also the identification of relevant evidence and the generation of explanations for the verdicts given.

Chain-of-Thought Prompting: Initially, scaling language models up appeared to benefit more knowledge-intensive tasks than reasoning-heavy ones (Rae et al., 2021). Nevertheless, (Wei et al., 2022) (2022) demonstrated that LLMs could be applied to System 2 problems by prompting the model to generate step-by-step solutions, coined “Chain-of-Thought” (CoT). CoT prompting led to substantial improvements in many reasoning-intensive tasks (Wei et al., 2022), (Zhou et al., 2022); (Droz-

Eliminative Codex Few Shot

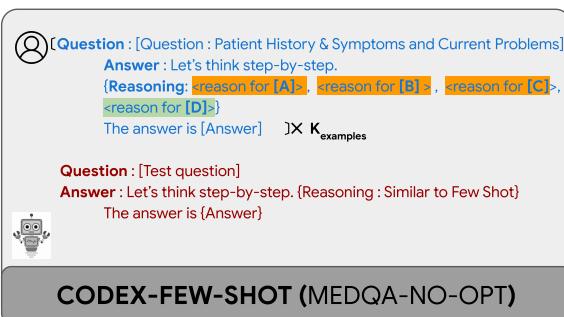


Forward Med-Codex Few Shot



Datasets

MEDCODEX-FEW-SHOT (MEDQA-NO-OPT)



Prompting Strategies

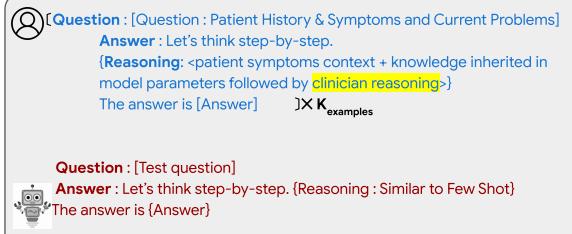


Figure 1: Dataset & Prompting Strategy wise Overview

dov et al., 2022); (Nye et al., 2021)), allowing to bridge the gap with human-level performances for most of the difficult BIG-bench tasks (Chung et al., 2022). As an alternative to writing reference step-by-step solutions, zero-shot CoT ((Kojima et al., 2022)) allows for generating CoTs using single and domain-agnostic cues: “Let’s think step by step” (see example 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

1.1 Key Contributions

We primarily focus on the USMLE-MedQA dataset (Jin et al., 2021), an objective medical exam dataset. The dataset consists of questions sourced from professional medical board exams in the USA and is available in three languages: English simplified Chinese, and traditional Chinese. The English subset utilized for this study consists of 12,723 questions, forming a comprehensive resource for investigating medical question answering in a standardized examination context. The current version of the dataset is not practically useful to be deployed in a patient query-based healthcare system since the responses to each such query are objective responses (either correct response or wrong response).

Our contributions here are manifold

- We propose a healthcare-specific language model response generation task that utilizes the best prompting methods.
- We demonstrate that prospective incremental reasoning driven prompting mimicking real life clinical scenarios performs significantly better at answering open ended medical questions.
- We present efforts towards building a novel medical corpus that includes human evaluated/verified subjective responses generated from a language model on the MedQA dataset.
- We also contribute a modified MedQA dataset conducive for testing medical question answering ability without options.

2 Background & Related Work

Recent advancements in the application of Large Language Models (LLMs) for medical question answering have highlighted the potential of these models to demonstrate applicability in medical diagnostics, education, and research. Venigalla et. al., introduced BioMedLM, demonstrating

135 its performance on several medical datasets, in-
136 cluding MedQA, underscoring the capabilities of
137 LLMs in handling complex medical queries (Veni-
138 galla et al., 2022; Jin et al., 2021). Furthermore,
139 Singhal et al.’s extensive experiments across a
140 broad spectrum of medical question tasks within
141 the MultiMedQA suite—encompassing datasets
142 such as MedQA, MedMCQA, PubMedQA, and
143 MMLU—have demonstrated the versatility and
144 depth of knowledge encoded in these models (Sing-
145 hal et al., 2022; Pal et al., 2022; Jin et al., 2019;
146 Hendrycks et al., 2020).

147 The exploration of LLMs in generating not only
148 accurate but also reasoning-based responses to medi-
149 cal questions marks a significant step forward.
150 Models like PubMedGPT (Bolton et al., 2022) and
151 Codex (Liévin et al., 2022) have established bench-
152 marks on datasets like MedQA through innovative
153 approaches including Classification head, Chain-of-
154 Thought, and Knowledge Grounding, highlighting
155 the importance of not just what is answered, but
156 how the answer is derived. Advanced LLMs, such
157 as Med-Palm2 (Singhal et al., 2022) and Flan-Palm
158 (Chung et al., 2022), have further raised the bar
159 for performance, although their limited availabil-
160 ity poses challenges for widespread research and
161 application in the medical field.

162 In addition to these contributions, the work by
163 authors such as Hendrycks *et. al.*, on the Measur-
164 ing Massive Multitask Language Understanding
165 (MMLU) dataset presents a comprehensive evalua-
166 tion of LLMs across a range of subjects, includ-
167 ing medicine, pointing to the broad applicability
168 and potential of LLMs beyond single-domain tasks
169 (Hendrycks et al., 2020). Another notable direc-
170 tion is the investigation into the interpretability and
171 explainability of model predictions, as highlighted
172 by Mesinovic *et. al.*, which is critical for trust and
173 reliability in medical applications (Mesinovic et al.,
174 2023).

175 3 Problem Setting and Preliminary 176 Notations

177 Let us consider a language model at inference time
178 $\mathcal{LM}_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by θ . Further, let us
179 assume we receive a particular question $q_i \in \mathcal{Q}$ i.e.,
180 from the universal set of questions \mathcal{Q} . Also, each
181 such question q_i is associated with a m -cardinal
182 set of options $\mathcal{O}_i = \{o_i^j\}_{j=1}^m$, out of which there
183 is one correct option $o_i^c \in \mathcal{O}_i$. Further we con-
184 sider for each option o_i^j , there exist a correspond-

185 ing reasoning distribution $\mathbb{R}^i(\bullet|q_i; \theta)$ associated
186 with the question q_i . Further let the reasoning as-
187 sociated with a particular option o_j^i be denoted as
188 $r_j \sim \mathbb{R}^i(o_j^i|q_i; \theta)$. Hence, we can rewrite the col-
189 lection of all such reasonings for all options as \mathcal{R}^i .

190 Let the prompt $\mathcal{P} = \langle q_i, \mathcal{O}_i, \mathcal{R}^i \rangle$ be passed as
191 an input to \mathcal{LM}_θ to generate corresponding re-
192 sponse $\mathcal{J}^i \in \mathcal{Y}$. Each such \mathcal{J}^i includes the corre-
193 sponding reasoning quality as well as correctness
194 of response based on ground-truth.

195 4 Experiments, Evaluation Methods and 196 Results

197 4.1 Datasets & Models

198 **Datasets** The MedQA (Zhang et al., 2018) dataset
199 consists of multiple choice questions based on
200 United States Medical License Exams (USMLE)
201 and collected from professional medical experts.
202 We modify the questions to seek descriptive re-
203 sponses and not remain objective multiple choice
204 questions. We consider two variations of the
205 MedQA dataset for our further experimentation

- **MEDQA-ORIGINAL** : We utilise the orig-
206 inal MCQ type options format of the
207 MedQA dataset, where we have the follow-
208 ing $\langle \text{Question}, \text{Option List} \rangle$ format $\{q_i, \mathcal{O}_i\}$
209 where the i -th option list \mathcal{O}_i consists of indi-
210 vidual 4 options: $\{o_j^i\}_{j=1}^4$. A sample is shown
211 in Table 1.
- **MEDQA-NO-OPT** : Here, we consider the
212 following format where the **option list for**
213 **a particular question q_i is not provided**
214 to the Language Model \mathcal{LM} i.e. we just
215 pass $\langle q_i, \emptyset, \mathcal{R}_i \rangle$, with a slight modification
216 (Table 1) to the question as showcased in
217 Section 4.1.1. We discuss in the following
218 sections regarding conversion strategies from
219 MEDQA-ORIGINAL \rightarrow MEDQA-No-OPT.

220 4.1.1 Conversion of MCQ type questions to 221 descriptive type

222 The conversion of MCQ-type questions
223 (MEDQA-ORIGINAL) to descriptive ques-
224 tions (MEDQA-No-OPT) is aimed to emulate
225 real-world medical scenarios where open-ended
226 inquiries are prevalent. This modification required
227 our model to respond without predefined choices,
228 fostering holistic reasoning and integration of
229 diverse knowledge sources. By eliminating answer
230 options, we assessed the model’s depth and quality

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

Table 1: Sample Questions from the USMLE-MedQA dataset along with modified dataset without any options. For more details check Appendix A

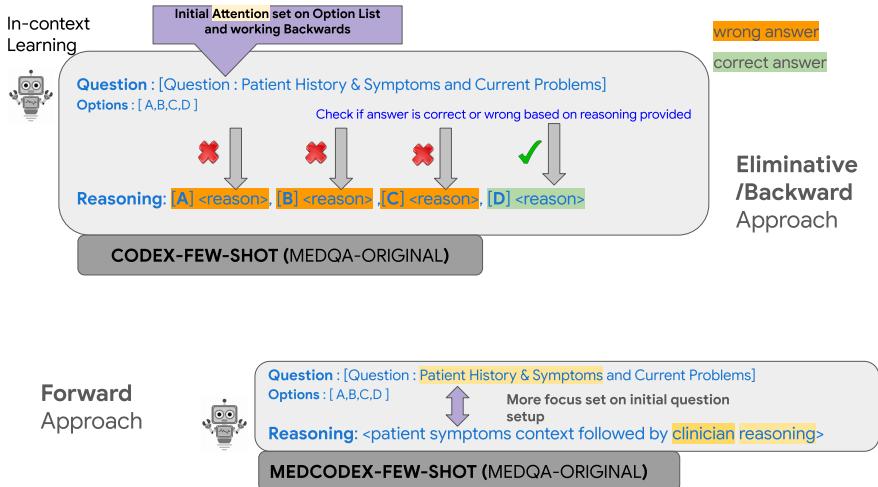


Figure 2: CODEX COT looks at generating options in a more eliminative approach often not catering to the context of clinical investigation unlike differential diagnostics as per MEDCODEX

of reasoning skills, ensuring a more realistic evaluation of its performance in complex medical scenarios. To cater to a descriptive scenario, we slightly modify the question q_i under the corresponding prompt \mathcal{P} with more descriptive question clauses \mathcal{C}_q^i like **What is the most likely ... ,** **The addition of what ... ,** **What is the most appropriate ...** with more details in Appendix. Again the modification of q_i under the objective MCQ regime to the Descriptive regime $q_i + \mathcal{C}_q^i$ is done under clinical supervision with the help of clinical experts.

Models We use Llama2-7B chat and Llama2-70B chat model (Touvron et al., 2023) for our entire experimental evaluations.

4.2 Prompting strategies

It is evident to see how we can utilise different prompting strategies for \mathcal{P} which can capture variations in the reasoning \mathcal{R}^i i.e., how we would like

the model to reason about reaching to a particular correct response.

CODEX FEWSHOT PROMPTS: Such prompting strategies first introduced in (Chen et al., 2021) and subsequently in medical literature (Liévin et al., 2022) has shown to perform well on MCQ type questionnaire. For a MCQ type question, the reasoning structure seem to follow the pattern:

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

However, such kind of prompting strategies tend to be eliminative in nature as shown in Fig. 2 and often contradicts with real life clinical scenarios where prospective additive reasoning is needed for reaching a differential diagnosis. Further, Codex prompt builds the entire context in one shot to reach the final answer.

MEDCODEX FEWSHOT PROMPTS: Here, we utilise the intermittent reasoning structure given a unique medical context which builds upon with every additive context in answering clinical questions

233
234
235
236
237
238
239
240
241
242
243
244
245
246
247

252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271

Forward/MedCodex Approach



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



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

The answer is (B) gallbladder cancer

Answer is correct

Eliminative/Codex Approach

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

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

The answer is (B) gallbladder cancer

Answer is correct

(a) MEDQA-ORIGINAL Codex vs MedCodex

Forward/MedCodex Approach



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



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

Answer: Intestinal atresia.

Answer is correct

Eliminative/Codex Approach

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

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

Answer: Congenital bowel obstruction (pyloric stenosis).

Answer is correct

(b) MEDQA-No-OPT Codex vs MedCodex

Figure 3: Illustrative Example: Showcasing different Prompting Strategies across two dataset variants. context of the answer and corresponding reasoning and Future actions are highlighted.

that mimics the usual way of reaching a final diagnosis. In real clinical scenarios, there are no four options to chose from; the clinician takes medical history, forms the mental structure for differential diagnosis, performs examination, adds or deletes the potential diagnosis based on contextual information and then ultimately takes into consideration the laboratory investigations to finally reach the diagnosis. The MEDCODEX FEWSHOT PROMPTS mimics this strategy to reach a final answer.

Figure 1 refers to different classification of

prompting strategies based on their reasoning structure as well as its basis on the corresponding dataset (MCQ/ Descriptive Questions). In both prompt strategies, we use few shots $K = 5$.

4.3 CODEX FEWSHOT PROMPTS vs MEDCODEX FEWSHOT PROMPTS on Original MedQA

Experimental Setup: We first utilise MEDCODEX FEWSHOT PROMPTS due to its underlying prospective incremental reasoning

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

structure which typically follows the line of reasoning employed by medical professionals. Along with this we compare responses collected using CODEX FEWSHOT PROMPTS where the reasoning is more eliminative in nature based on seeing the options and less attention towards the clinical flow of argument (which also involves forming the entire clinical context in one go). The Llama2 70B Base and 70B Chat model were prompted using both strategies.

Evaluation Strategy: The results of this experiment on the MEDQA-ORIGINAL’s 1273 questions from the Test set are described in the Table 2

Observations:

Methods	Accuracy Scores
MEDCODEX-70B Base (Codex)	54%
MEDCODEX-70B Base (MedCodex)	23%
MEDCODEX-70B Chat (Codex)	52%
MEDCODEX-70B Chat (MedCodex)	50%

Table 2: CODEX FEWSHOT PROMPTS vs MEDCODEX FEWSHOT PROMPTS on Llama7B and Llama70B base and chat models

Interpretation: On the MEDQA-ORIGINAL, dataset, the MEDCODEX FEWSHOT PROMPTS perform poorly as compared to CODEX FEWSHOT PROMPTS both on Llama70B base and 70B Chat model. Due to its inherent eliminative approach, the CODEX FEWSHOT PROMPTS has a smaller search space and results in higher accuracy.

4.4 CODEX FEWSHOT PROMPTS vs. MEDCODEX FEWSHOT PROMPTS on Descriptive MedQA questions

Experimental Setup: We perform the experiment of comparing a modified version of the CODEX FEWSHOT PROMPTS and MEDCODEX FEWSHOT PROMPTS on MEDQA-No-OPT dataset. The Codex prompt is modified to remove options and their labels, to make it suited for a no options setting (Section: 4.2).

Evaluation Strategy: We select 100 questions from MEDQA-No-OPT and evaluate via medical expert the final reasoning quality and final answer on a 3 point Likert scale (Batterton and Hale, 2017) Agree, Neutral, Disagree .

Observations: As indicated in Figures 6 and 7, for 82% and 77.85% of the questions, the expert medical personnel agreed with the reasoning and the evaluation provided by the MEDCODEX FEWSHOT PROMPTS after prompting Llama2-7B-chat and Llama2-70B-chat parameter models respectively. However, this was 72% and 82.5% subsequent to prompting Llama7B and Llama70B respectively in the case of CODEX FEWSHOT PROMPTS.

Interpretations: Given the open ended nature of MEDQA-NO-OPT dataset, the incremental generative approach of reasoning to reach the final answer by the MEDCODEX FEWSHOT PROMPTS is qualitatively superior compared to CODEX FEWSHOT PROMPTS for Llama2-7B-chat model. This could be because the CODEX FEWSHOT PROMPTS is designed to work best on a restricted universe of options for MCQ based questions. This also indicates that mimicking the human process of reaching the final answer in diverse medical scenarios (as performed by MEDCODEX FEWSHOT PROMPTS) is better in scenarios of open ended questions while prompting smaller parameter models. However, CODEX FEWSHOT PROMPTS better able to leverage a larger model (Llama2-70B-chat).

4.5 Differential diagnosis generation with Codex for Selection on MedQA-no-opt Dataset

Experimental Setup: Each question q_i in the MEDQA-NO-OPT dataset lacks options, therefore pertaining to a real world problem setting. We particularly utilize the MEDCODEX FEWSHOT PROMPTS to sample a number of options \mathcal{O}_i^{gen} for each question, thereby mimicking the MEDQA-ORIGINAL dataset format. The question q_i followed by generated options \mathcal{O}_i^{gen} are fed into the model \mathcal{LM} with the CODEX FEWSHOT PROMPTS to select the most appropriate option. To create a set of 4 options through sampling, 10 unique options are sampled by filtering on a word-level match basis (to avoid selecting options that are repetitive). Out of these, the top-4 are selected through the perplexity of model outputs.

Evaluation Methods: We evaluate via medical expert the final reasoning quality and final answer on a 3 point Likert scale (Batterton and Hale, 2017) Agree, Neutral, Disagree .

Observations: As indicated in Figures 6

and 7, for 80% and 89.5% of the questions, the expert medical personnel agreed with the reasoning and the evaluation provided by the Differential diagnosis generation (using MEDCODEX FEWSHOT PROMPTS) followed by CODEX FEWSHOT PROMPTS for Selection on MEDQA-No-OPT Dataset after prompting Llama2-7B-chat and Llama2-70B-chat parameter models respectively.

Interpretation: Since the MEDQA-No-OPT dataset questions are open ended, we tried to restrict the pool from which the correct answers could be selected driven by generating four top answers. This was inspired by Self-consistency approach of Codex Few Shot paper (Wei et al., 2022). As expected the larger Llama 70B model prompting worked better than Llama 7B model. Importantly, this process was significantly better than all prompting strategies indicating the utility of leveraging the option creation process using MEDCODEX FEWSHOT PROMPTS and subsequent selection by CODEX FEWSHOT PROMPTS.

4.6 Experiments with Verifier

Motivation for Verifier and its training

So far we have approached the problem with an in-context-learning perspective without changing the model parameters. We believe building on top of powerful models like Llama-2 chat series, could give us a significant advantage in this scenario. We endeavored to improve performance by substituting the CODEX FEWSHOT PROMPTS prompt with a verifier Reward learning model.

Verifier Dataset Contribution

The training dataset for the verifier \mathcal{D}_{ver} was constructed by gathering medical expert evaluations towards the correctness of model-generated responses. A representative sample of 100 questions was selected from the MEDQA-No-OPT and divided into two subsets for inter-annotator agreement evaluation. Multiple medical experts independently assessed the correctness of responses generated by the Llama2 70B chat model, utilizing a newly designed verification prompt on the 5-point scale. The response sets encompassed both positive (correct answer) and negative (incorrect answer) pairs to evaluate the verifier’s ability to distinguish between them. Anonymity was maintained by not disclosing the actual correctness of the responses to the medical experts. The medical experts exhibited a commendable ability to discern the reasoning behind incorrect answers, emphasizing the value of their annotations for training the verifier. This annotated data was employed to train our reward model, with positive pairs acting as chosen pairs and negative pairs as rejected pairs.

ing the value of their annotations for training the verifier. This annotated data was employed to train our reward model, with positive pairs acting as chosen pairs and negative pairs as rejected pairs.

We use pairwise “chosen-rejected” pairs of strings consisting of $\langle q_i; \mathcal{R}_i; \mathcal{A}_i \rangle$ <Question-Reasoning-Answer> triplets to train the reward model \mathcal{LM}_{ϕ}^{rw} . We use the MEDQA-ORIGINAL dataset with 4-options to create these pairs. Here, we take the known correct option for the ‘chosen’ part and switch the incorrect options for the ‘rejected’ part to create three pairs with one question sample. We borrow the Question and Answer parts of the triplet from the MedQA dataset and generate the Reasoning artificially by prompting the Llama-2-7b-chat model. These results have been verified by medical experts. The prompt and input format for this are given below in Fig 4.

Below is a question from the field of professional medicine, where the correct answer is given. Explain with concordant medical reasoning how the answer is reached. The reasoning should be in a single paragraph and strictly within 200 words, and should smoothly end with the phrase “Thus, the answer is <Given Answer>.”
 Question: A 3-month-old baby died suddenly at night while asleep. His mother noticed that he had died only after she awoke in the morning. No cause of death was determined based on the autopsy. What precautionary measures could have prevented the death of the baby?
 Answer: Placing the infant in a supine position on a firm mattress while sleeping
 Reasoning:

Figure 4: Verifier Reasoning Generation

Training

We have fine-tuned the Llama2-7b-chat model for the reward model training followed by a linear head. We are applying LoRA for fine-tuning using the trl library from huggingface for this purpose. The model is trained on a reward modelling loss from (Wang et al., 2024).

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

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

4.6.1 Differential diagnosis generation with Trained Verifier on MedQA-on-opt Dataset

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

471 experiment was performed only for Llama7B parameter model due to resource constraints.
 472

473 **Evaluation Methods:** We select 100 questions
 474 from MEDQA-No-OPT and evaluate via medical
 475 expert the final reasoning quality and final answer
 476 on a 3 point Likert scale (Batterton and Hale, 2017)
 477 **Agree**, **Neutral**, **Disagree**.

478 **Observations:** As indicated in Figures
 479 6 and 7, for 86% of the questions, the
 480 expert medical personnel agreed with the
 481 reasoning and the evaluation provided by
 482 the Differential diagnosis generation (using
 483 MEDCODEX FEWSHOT PROMPTS) followed
 484 by Verifier for Selection on MEDQA-No-OPT
 485 Dataset after prompting Llama2-7B-chat model.

486 **Interpretation:** The development of the Verifier
 487 substantially improves the ability of Llama7B
 488 parameter model to select the response with high
 489 quality of expalanability as well as high accuracy
 490 indicating the viability of such an approach to beat
 491 larger parameter models by alternative prompting
 492 strategies. We hypothesize that the Verifier being
 493 trained to identify good quality reasoning can re-
 494 sult in better abilities to select the most probable
 495 answer subsequent to 4 option generation driven by
 496 MEDCODEX FEWSHOT PROMPTS. We suspect
 497 that testing this strategy on the Llama70B model
 498 would provide significantly better results.

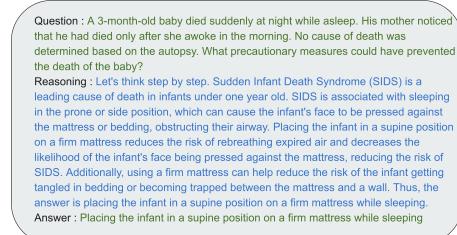


Figure 5: Verifier Input

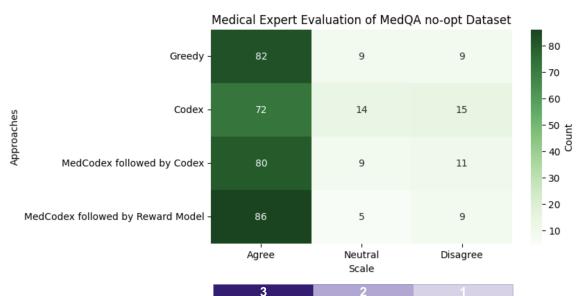


Figure 6: Results for experiments on MedQA-no-opt dataset with Llama-2-7B-chat

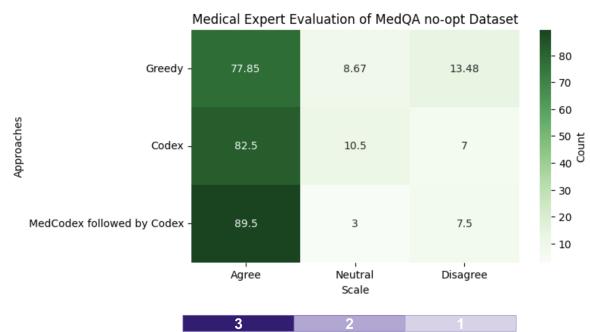


Figure 7: Results for experiments on MedQA-no-opt dataset with Llama-2-7B-chat

Conclusion and Future work

The incremental reasoning chain of thought prompting is a novel prompting methodology developed by us which follows the usual clinical approach of reaching a decision in real life clinical settings. We demonstrate that this strategy performs significantly good results over the CODEX prompting strategy which is designed for MCQ type questions. Further, we demonstrate that the verifier developed using reasoning performs much better at selecting agreeable responses from the Llama models. Further research will focus on testing the Verifier driven asnwer selection using the Llama 70B model and we would explore the generalizability of this approach by testing on other open source LLMs.

Acknowledgements

References

Katherine A Batterton and Kimberly N Hale. 2017. The likert scale what it is and how to use it. *Phalanx*, 50(2):32–39.

E Bolton et al. 2022. Pubmedgpt 2.7 b. Technical report, Technical report. Stanford University Center for Research on Foundation

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt,

535	Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. <i>Communications Medicine</i> , 3(1):141.	591
536		592
537		593
538		594
539		595
540	Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models. <i>arXiv preprint arXiv:2209.15003</i> .	596
541		597
542		598
543		599
544		600
545	Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. <i>Transactions of the Association for Computational Linguistics</i> , 10:178–206.	601
546		
547		
548		
549	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. <i>arXiv preprint arXiv:2009.03300</i> .	602
550		603
551		604
552		
553	Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. <i>Applied Sciences</i> , 11(14):6421.	605
554		606
555		607
556		608
557		
558	Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. <i>arXiv preprint arXiv:1909.06146</i> .	609
559		610
560		611
561		612
562	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. <i>Advances in neural information processing systems</i> , 35:22199–22213.	613
563		
564		
565		
566		
567	Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? <i>arXiv preprint arXiv:2207.08143</i> .	614
568		615
569		616
570		617
571	Munib Mesinovic, Peter Watkinson, and Tingting Zhu. 2023. Explainable ai for clinical risk prediction: a survey of concepts, methods, and modalities. <i>arXiv preprint arXiv:2308.08407</i> .	618
572		
573		
574		
575	Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. <i>arXiv preprint arXiv:2112.00114</i> .	619
576		620
577		621
578		622
579		
580		
581	Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In <i>Conference on Health, Inference, and Learning</i> , pages 248–260. PMLR.	623
582		624
583		625
584		626
585		
586	Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mhdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. <i>arXiv preprint arXiv:2212.13138</i> .	627
587		
588		
589		
590		
591	Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. <i>Nature medicine</i> , 29(8):1930–1940.	628
592		629
593		630
594		631
595		632
596	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. <i>arXiv preprint arXiv:2307.09288</i> .	633
597		634
598		635
599		636
600		637
601		638
602	A Venigalla, J Frankle, and M Carbin. 2022. Biomedlm: a domain-specific large language model for biomedical text. <i>MosaicML</i> . Accessed: Dec, 23(3):2.	639
603		640
604		641
605	Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. 2024. <i>Secrets of rlhf in large language models part ii: Reward modeling</i> .	642
606		643
607		644
608		645
609		646
610		647
611		648
612		649
613		650
614	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in Neural Information Processing Systems</i> , 35:24824–24837.	651
615		652
616		653
617		654
618		655
619	Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In <i>Proceedings of the AAAI conference on artificial intelligence</i> , volume 32.	656
620		657
621		658
622		659
623	Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. <i>International Journal of Computer Vision</i> , 130(9):2337–2348.	660
624		661
625		662
626		663

Appendix:

5 Results of Human Evaluations of Llama 7B chat responses

	Agree	Neutral	Disagree
Medical Expert 1	88	10	2
Medical Expert 2	86	12	2

Table 3: Human evaluated results of reasoning of Greedy based approach for Set 1 (50 questions)

	Agree	Neutral	Disagree
Medical Expert 1	82	2	16
Medical Expert 2	72	14	14

Table 4: Human evaluated results of reasoning of Greedy based approach for Set 2 (50 questions)

	Agree	Neutral	Disagree
Medical Expert 1	76	18	6
Medical Expert 2	28	24	48

Table 5: Human evaluated results of reasoning of Sampling Forward followed by Codex-based approach for Set 1 (50 questions)

	Agree	Neutral	Disagree
Medical Expert 1	72	14	14
Medical Expert 2	68	20	12

Table 6: Human evaluated results of reasoning of Sampling Forward followed by Codex-based approach for Set 2 (50 questions)

	Agree	Neutral	Disagree
Medical Expert 1	84	0	16
Medical Expert 2	86	4	8

Table 7: Human evaluated results of reasoning of Sampling Forward followed by Reward Model based approach for Set 1 (50 questions)

	Agree	Neutral	Disagree
Medical Expert 1	84	10	6
Medical Expert 2	88	6	6

Table 8: Human evaluated results of reasoning of Sampling Forward followed by Reward Model based approach for Set 2 (50 questions)

6 Results of Human Evaluations of Llama 70B chat responses

	Agree	Neutral	Disagree
Medical Expert 1	89	0	11
Medical Expert 2	92.3	7.7	0
Medical Expert 3	87	2	11

Table 9: Human evaluated results of reasoning of MedCodex approach

	Agree	Neutral	Disagree
Medical Expert 1	92	0	8
Medical Expert 2	72	18	10
Medical Expert 3	76	14	10

Table 10: Human evaluated results of reasoning of Codex approach

	Agree	Neutral	Disagree
Medical Expert 1	87	6	7
Medical Expert 2	92	0	8
Medical Expert 3	89	1	7

Table 11: Human evaluated results of reasoning of MedCodex followed by Codex approach