Large Language Models are In-context Teachers for Knowledge Reasoning

Anonymous EMNLP submission

Abstract

Chain-of-thought (CoT) prompting teaches large language models (LLMs) in context to reason over queries that require more than mere information retrieval. However, human experts are usually required to craft demonstrations for in-context learning (ICL), which is expensive and has high variance. More importantly, how to craft helpful reasoning exemplars for ICL remains unclear. In this work, we investigate whether LLMs can be better in-context teachers for knowledge reasoning. We follow the "encoding specificity" hypothesis in human's memory retrieval to assume in-context exemplars at inference should match the encoding context in training data. We are thus motivated to propose Self-Explain to use one LLM's self-elicited explanations as in-context demonstrations for prompting it as they are generalized from the model's training examples. Self-Explain is shown to significantly outperform using human-crafted exemplars and other baselines. We further reveal that for in-context teaching, rationales by distinct teacher LLMs or human experts that more resemble the student LLM's self-explanations are better demonstrations, which supports our encoding specificity hypothesis. We then propose Teach-Back that aligns the teacher LLM with the student to enhance the in-context teaching performance. For example, Teach-Back enables a 7B model to teach the much larger GPT-3.5 in context, surpassing human teachers by around 5% in test accuracy on medical question answering.

1 Introduction

Knowledge reasoning, different from numerical reasoning, requires large language models (LLMs) to deduce the association between questions and answers that do not usually explicitly appear in the training corpus, although LLMs may have memorized all the facts involved in the question. Such a *compositionality gap* [Press et al., 2023] between testing and pretraining makes knowledge reasoning difficult and beyond mere fact retrieval. However, LLMs have demonstrated impressive knowledge reasoning performance on diverse tasks [Wei et al., 2022] with few-shot prompting. Exemplars of reasoning are provided in the prompt as context to teach LLMs to reason through in-context learning (ICL) [Brown et al., 2020] at inference. LLMs will generate intermediate reasoning steps (known as Chain-of-Thought (CoT)[1] [Wei et al., 2022]) for deducing the test cases.

Standard few-shot CoT prompting requires humans to first craft high-quality demonstrations of reasoning for LLMs, as depicted in the upper part of Figure 1. However, this may bring some issues. On the one hand, in professional domains such as medicine, experts like physicians are needed to produce fine-grained rationales with correct jargon, which is time-consuming and expensive [Pal et al., 2022, Yang et al., 2023]. On the other hand, different from labels, rationales can be phrased in varied ways, while all being correct [Yao et al., 2023]. Collecting reasoning examples through crowd-sourcing can thus have great uncertainty [Gebreegziabher et al., 2023]. The constructed rationales heavily depend on human annotators' own experience and thus, may be very subjective [Lee et al., 2022].

More fundamentally, there is a limited understanding of the principles behind constructing effective rationale exemplars for in-context learning. Currently, the majority of works depend on human-crafted demonstrations (usually by professionals) that are based on some heuristic rules [Fu et al., 2023b, Zhou et al., 2022, Khot et al., 2023]. However, it is unclear whether those sophisticated rationales crafted by humans are equally the most sensible to LLMs. Demonstrations of rationales from humans may not always be helpful [Yao et al., 2023], although they are often assumed to be gold standards [Muller et al., 2021].

Therefore, we are motivated to ask, **can an LLM teach itself or other models through in-context learning for knowledge reasoning, preferably better than humans**? We consider a generic framework of *in-context teaching*, where a teacher LLM provides example rationales that are then used as in-context demonstrations to prompt a student LLM. For constructing effective in-context exemplars, we assume the **Encoding**

[1] We use "CoT" and "rationale" interchangeably to refer to reasoning paths.

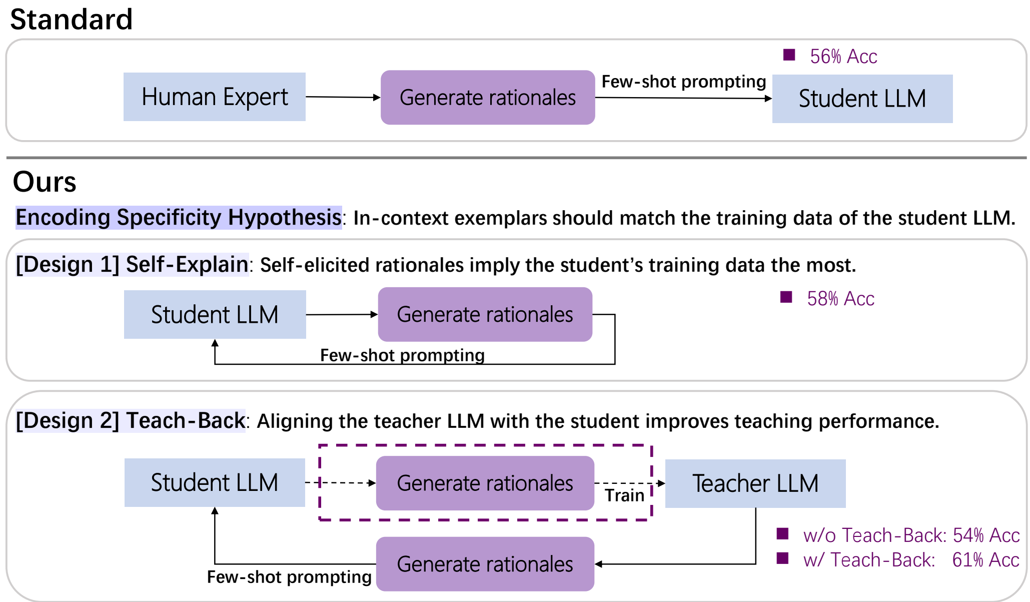


Figure 1: Overview of our approaches. Existing few-shot CoT prompting methods rely on human experts to craft rationales as in-context demonstrations. We propose Encoding Specificity Hypothesis to make large language models better in-context teachers than humans. We accordingly design Self-Explain for an LLM itself to be the teacher and Teach-Back to improve the LLM's capability of teaching another model in context.

Specificity Hypothesis [Tulving, 1972] from human's episodic memory as inspired by the convergence between attention and memory [Ramsauer et al., 2020, Bricken and Pehlevan, 2021, Zhao, 2023]. The hypothesis postulates the context during recalling information from the episodic memory should match the context during encoding it. Similarly, for few-shot prompting, ideal in-context exemplars of reasoning at test time are hypothesized to match the rationales from the training corpora related to the test domain. For example, when reasoning over medical questions at inference, in-context rationales are expected to be phrased similarly to examples in the medical corpus during pre-training.

The encoding specificity hypothesis can be easily satisfied when the teacher model is the same as the student model. We directly prompt the student model to *explain* the given answer to a question sampled from the same dataset as the test data as inspired by learning theory in cognitive science [Chi et al., 1989]. Those elicited self-explanations can represent the model's encoded knowledge for the test task and are then used as in-context demonstrations for ICL at inference. We refer to this approach as Self-Explain. On the other hand, when the teacher model is different from the student model (e.g., using a weak and small model to teach a much larger model [Burns et al., 2024]), we first let the teacher model learn from the student's explanations before eliciting the teacher's explanations (see the lower part of Figure 1). We refer to this method as Teach-Back, which is how healthcare providers (i.e., teacher) reduce the communication gap with patients (i.e., student) for effective health education [Talevski et al., 2020].

Our experimental results provide sources of evidence for the encoding specificity hypothesis. We find that the student model itself tends to be the best in-context teacher for it, surpassing human teachers or other LLM teachers (w/o Teach-Back). Our experiments across models of different sizes and reasoning abilities suggest that for in-context teaching, larger and stronger models are not necessarily better in-context teachers, though they may produce more reliable rationales. Instead, our findings demonstrate a strong linear correlation that in-context exemplars of reasoning that more resemble the student's self-explanations may yield better student's performance, which supports our encoding specificity hypothesis. Furthermore, applying Teach-Back can significantly improve the in-context teaching capability of a teacher model for the student model and even outperform Self-Explain depending on the teacher model. For example, Teach-Back enables a small deployable 7B model to teach the much larger GPT-3.5 in context, surpassing human teachers by around 5% in test accuracy on medical question answering.

In summary, our contributions are mainly in three folds:

- We investigate in-context teaching of LLMs for knowledge reasoning, where we use a teacher LLM to provide in-context demonstrations to teach a student model to reason over the question. We propose the Encoding Specificity Hypothesis as our guideline for composing in-context exemplars.
- We propose a new way of eliciting rationales from an LLM by prompting it to explain question-answer pairs. We then propose Self-Explain

prompting to use an LLM's self-explanations as its in-context exemplars, which outperforms using human-crafted CoTs.

- Our experiments suggest that in-context exemplars of rationales from LLM teachers or human teachers that more resemble the student's self-explanations may produce better reasoning performance. We then propose Teach-Back, demonstrating that the in-context teaching ability of LLMs can be improved by aligning the teacher with the student's self-explanations.

2 Revisiting ICL

We first detail some annotations and give a formal setup of ICL. We denote the model parameters as θ, rationale as π and assume a labeled dataset $\mathcal{D}$ with distribution p^*. Given a test query $\mathbf{x}$, the model will predict $\tilde{\mathbf{y}}$ by conditioning on the query and in-context exemplars. We can then have,

$$\tilde{\mathbf{y}} = \text{argmax}_y P(y|\mathbf{e}, \mathbf{x}, \theta), \tag{1}$$

where $\mathbf{e}$, is the sequence of all K in-context exemplars i.e., $\mathbf{e} = e_1, ..., e_K$) and $e_i = (x_i, \pi_i, y_i)$ where (x_i, y_i) is sampled from p^*.

2.1 Encoding Specificity Hypothesis

A key question for ICL is how to compose in-context rationales for some specific dataset $\mathcal{D}_{(x,y)}$? Rationales can be rephrased differently while delivering the same logic. To understand this question, we take a memory view of ICL by conceptualizing LLMs as memory networks [Hopfield, 1982, Kanerva, 1988, Kaiser and Bengio, 2016, Ramsauer et al., 2020, Krotov and Hopfield, 2016]. The feed-forwarding through hidden layers of LLM is to retrieve and generalize learned information in memory to construct the output $\mathbf{y}$ to complete the query $\mathbf{x}$ under the guidance of context $\mathbf{C}$ (i.e., in-context exemplars). The pretraining stage can be viewed as encoding information into the weights, i.e., memories of LLMs.

From a memory view, we are inspired to assume the encoding specificity [Tulving and Thomson, 1973] in humans' memory retrieval that requires the match of context between testing and training. To see this hypothesis, a simplified thought experiment can be considered: supposing that a specific datapoint (x, y) has been seen during language modeling in pretraining and C is the corresponding context prepending (x, y), (i.e., a consecutive string (C, x, y) is seen by LLM during training), at test time, to let the model generate y with great probability, we can prompt it with (C, x).

More generally, the encoding specificity hypothesis implies that **in-context exemplars of reasoning should match the distribution of reasoning examples seen during training**, especially the training corpus containing information similar to task data $\mathcal{D}_{(x,y)}$. LLMs may have seen many sentences involving reasoning during pre-training and also explicit examples of reasoning (e.g., responses to users' questions) during further instruction fine-tuning. Thus, for ICL, it may be easier for LLMs to generalize from in-context exemplars similar to those rationales from training data (e.g., having similar reasoning logic or using similar expressions/ jargon) to answer new questions at inference.

3 Methodology

The general framework of our proposed methods and our prompting format is shown in Figure 2. We first introduce Self-Explain, where the student and teacher are the same (Section 3.1), as a straightforward implementation of the encoding specificity hypothesis. We then extend this approach to employing a different teacher model (Section 3.2).

3.1 Self-Explain

Motivated by the encoding specificity, we would like our in-context exemplars of reasoning to match the LLMs' training corpus containing information similar to the task data. To achieve this, we directly prompt an LLM to elicit its explanation for some question-answer pairs of task data. Such self-explaining is actually how humans integrate new information with their existing knowledge [Chi et al., 1989]. Similarly, the LLM is expected to utilize its existing encoded knowledge relevant to the unseen question provided, in order to generate its explanations. These self-explanations are then used as in-context exemplars of reasoning to prompt the model itself.

Eliciting LLMs' Self-explanations. Formally, we assume access to labeled training data where we have some data (x, y) sampled from the distribution p^{train} and assume $p^{\text{test}} \approx p^{\text{train}}$. We consider a realistic setting where human-crafted CoTs are not available. We define an oracle CoT as

$$\pi^* := \text{argmax}_\pi P(y|x, \pi, \theta). \tag{2}$$

Self-explanation is then obtained as,

$$\pi^{\text{self}} = \text{argmax}_\pi P(\pi|x, y, \gamma, \theta), \tag{3}$$

where γ is an instruction. We hope LLMs to generate reasoning path based on given (x, y) by recalling from its according encoded knowledge so as to satisfy encoding specificity. we further find $P(y|x, \pi^{\text{self}}, \theta) >> P(y|x, \pi^{\text{human}}, \theta)$ (see Appendix C). We may arguably state that π^{self} is a more reasonable estimation to π^* than π^{human}.

Filtering Self-explanations. We filter out the elicited self-explanations based on the explanation *faithfulness* [Jacovi and Goldberg, 2020]. Explanations that fail to guide the model to produce the given answer, i.e., $y \neq \text{argmax}_{\tilde{y}} P(\tilde{y}|x, \pi^{\text{self}}, \theta)$, are screened. We empirically verify the self-explanation ability of different

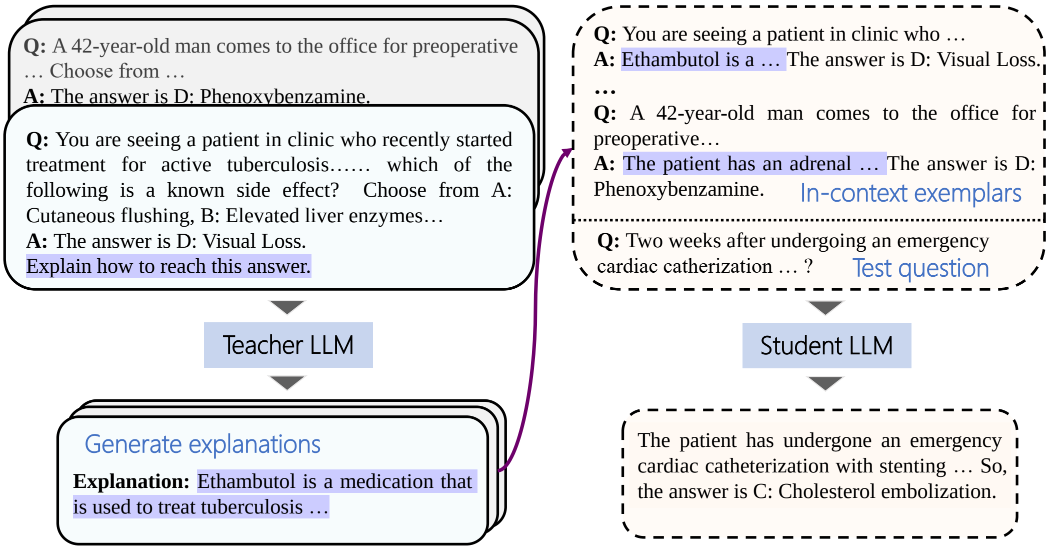


Figure 2: The overall framework and prompting format of our approach. The teacher LLM is prompted to generate explanations on sampled training data. Those teacher's explanations are used as in-context demonstrations for the student model at test time. The student model and the teacher model can be the same.

LMs that models actually succeed to justify the given (x, y) most of the time (see Section 5.1).

ICL with Self-explanations. The self-explanations π^{self} elicited by the model are then used as in-context exemplars for prompting it following Equation 1. This can be viewed as the model teaching itself via ICL to do reasoning. Additionally, the ICL performance is very close when using respective π^{self} elicited with either wrong or ground-truth y for the input question x in Equation 3 (see Section 4.2).

Generalization through generation diversity. The underlying logic of π^{self} might be very specific to its corresponding (x, y) and thus lacks generalizability to other different cases from test data. Then, the output explanation $\hat{\pi}_{\text{te}}$ at test time may fail to apply to the input cases, leading to wrong answers. To mitigate this issue, we also design a new instruction γ' so as to prompt the model to generate solutions employing distinct logics. Formally, we have,

$$(\pi_1^{\text{self}}, ..., \pi_n^{\text{self}}) = \text{argmax}_{\pi} P(\pi | x, y, \gamma', \theta), \quad (4)$$

where $n \in (1, N)$ and N is the number of different explanations to generate. For example, if $N = 5$, γ' will be "Explain how to reach this answer in five different ways". Then at test time, π_i^{self} for an in-context exemplar (x_i, y_i) will be randomly sampled from the according $\{\pi_n^{\text{self}} | n \in (1, N)\}$ of (x_i, y_i).

3.2 Teach-Back

Instead of ICL with self-explanations, those explanations as in-context exemplars can be provided by a separate model as the teacher, i.e.,

$$\pi^{\text{self}} = \text{argmax}_{\pi} P(\pi | x, y, \gamma, \theta_{\text{teacher}}). \quad (5)$$

However, explanations of one model may not serve as the most helpful reasoning demonstrations for another model, especially when the teacher's explanations are very distinct from the student's self-explanations (see results in Section 4.3). Based on our encoding specificity hypothesis, we propose to let the teacher model learn from the student's self-explanations (through supervised fine-tuning) before eliciting the teacher's explanations. This method is called Teach-Back, which is similar to how healthcare providers reduce the communication gap with patients for effective health education [Talevski et al., 2020]. Doctors will rephrase and clarify their explanations based on patients' explanations for better communication. In Section 4.4, we empirically show the effectiveness of Teach-Back in improving teaching efficacy and enhancing student's performance.

4 Experiments

4.1 Experimental Setup

Datasets. We are focused on knowledge-intensive question-answering tasks that require logical thinking on information and associating encoded knowledge but mere facts retrieval. Such knowledge-intensive QA is common and important for the applications of LLMs [Jin et al., 2021b, Tran et al., 2023]. We evaluate our method in both general domains and expert domains. We employ widely-used StrategyQA [Geva et al., 2021] for commonsense reasoning. For expert domains, we use challenging MedMCQA [Pal et al., 2022] and MedQA [Jin et al., 2021a] with standard splits. These datasets consist of multiple-choice questions to diagnose clinical cases, which are used for physician qualification exams.

Method\Dataset	MedCQA	MedQA	StrategyQA
No CoT	51.7	52.1	46.8
Zero-shot CoT [Kojima et al., 2022]	51.1	54.4	45.6
Auto-CoT [Zhang et al., 2023]	52.5	55.2	52.7
Human CoT	53.1	55.6	56.1
Self-Explain	53.2	57.5	58.5
w/ Multi-Exp	**56.6**	**59.6**	**59.7**

Table 1: Test accuracy of different prompting methods on three datasets for knowledge reasoning.

	MedMCQA	MedQA	StrategyQA
Right	56.6	59.6	59.7
Wrong	56.0	59.4	59.1

Table 2: Test accuracy of prompting with self-explanations that are generated provided by right answers and wrong answers.

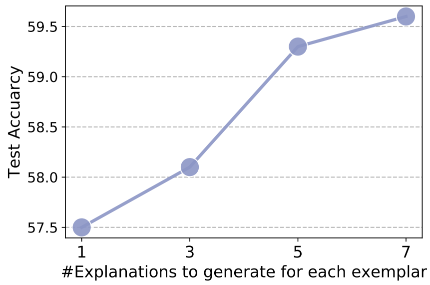


Figure 3: The test performance with respect to the number of self-explanations to generate for each exemplar.

Models. We use a variety of models. We employ the chat version of 7B model and 13B model of Llama2 [Touvron et al., 2023], the 7B model of Mistral [Jiang et al., 2023], the Phi3-128k-mini that has 3.8B parameters [Abdin et al., 2024] and the frozen version (0613) of GPT-3.5 [2].

Prompting. For the instruction used for eliciting models' self-explanations, an ablation study is conducted in Appendix A. For few-shot prompting at test time, we use five in-context exemplars sampled from the training data.

Baselines. Apart from comparing our approach with standard few-shot prompting with human CoTs, we include three more baselines. (1) "No CoT": We remove rationales and use input-output pairs only for in-context exemplars; (2) "Zero-shot CoT" [Kojima et al., 2022]: This method does not require human-crafted demonstrations as it is not few-shot prompting. It directly elicits reasoning from LLMs for the test question by using the prompt "Let's think step by step". (3) "Auto-CoT" [Zhang et al., 2023]: This work uses the same method as Kojima et al. [2022] to elicit rationales from LLMs. But it further proposes a way of exemplar selection to choose elicited rationales as in-context exemplars. For fair comparison, in each trial, we use the same question-answer pairs for few-shot demonstrations for all baselines.

4.2 Few-shot Prompting with Self-explanations

In this section, we evaluate the test performance of Self-Explain, i.e., when the model's self-explanations are used as in-context exemplars of reasoning for few-shot prompting. We conduct experiments with GPT-3.5-turbo on reasoning tasks in both general domains and expert domains.

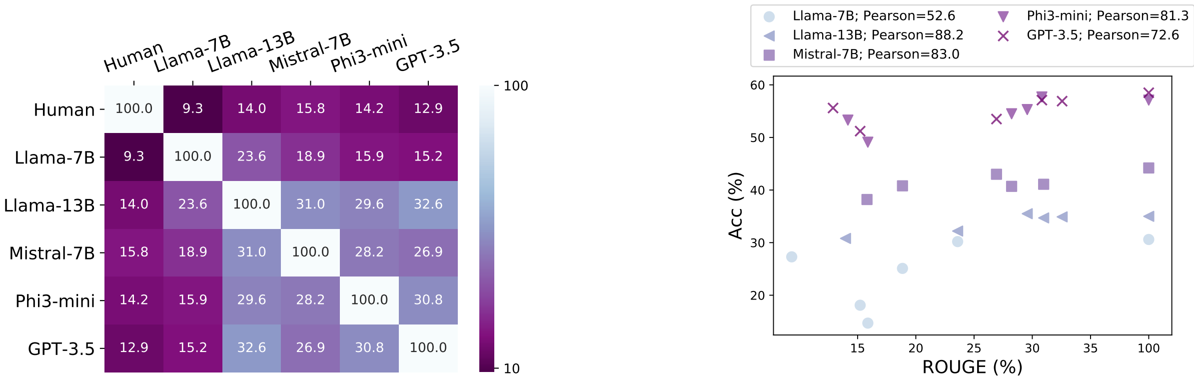
[2]https://platform.openai.com/docs/models/gpt-3-5-turbo

Prompting with self-explanations is better than using human-crafted CoT. Our results are shown in Table 1. Self-Explain can impressively outperform using CoTs crafted by human professionals by around 2% in both challenging MedQA and general domain, while reaching similar performance to Human CoT for MedCQA. Our approach also outperforms Auto-CoT [Zhang et al., 2023] and vanilla zero-shot CoT [Kojima et al., 2022], both of which cannot effectively surpass Human CoT for knowledge reasoning. The superior performance of Self-Explain may support our encoding specificity hypothesis. Overall, considering the difficulty and expense of crafting CoTs by humans, Self-Explain can thus be very useful in expert domains. Example self-explanations and human CoTs can be found in Appendix D.

Generation diversity is helpful. Apart from naive Self-Explain, we generate five different explanations for each in-context exemplar and randomly select one for ICL at test time (see explanation in Section 3.1). As shown in results of "w/ Multi-Exp" in Table 1, this approach further boosts the performance of Self-Explain to significantly surpass Human CoT by around 4% in all datasets. To better understand the effects of this component, we experiment with generating different numbers of self-explanations for one exemplar input. Results are shown in Figure 3. We find generating different self-explanations for an in-context exemplar can generally improve the test performance, while such improvement experiences diminishing returns with further increased numbers of generations.

Teacher\Student	Llama2-7B	Llama2-13B	Mistral-7B	Phi3-mini	GPT-3.5
No CoT	28.4	31.1	31.8	49.5	41.1
Human	27.3	31.4	38.2	53.3	55.6
Llama2-7B	**30.6**	32.2	40.8	49.1	51.2
Llama2-13B	30.2	**35.5**	41.1	55.3	56.9
Mistral-7B	25.1	34.7	**44.2**	54.4	53.5
Phi3-mini	18.7	35.1	40.7	57.1	57.1
GPT-3.5	18.1	34.4	43.1	**57.7**	**57.5**

Table 3: Results of teaching student LLMs with teachers' self-explanations through in-context learning. The best test accuracy is highlighted in bold.



(a) (b)

Figure 4: (a) ROUGE scores between self-explanations of teacher and student. For "Human" teachers, human-crafted CoTs are used for computation. (b) Strong linear correlation is observed between ROUGE scores of self-explanations of teacher and student and the student's test accuracy.

Does the correctness of self-explanations matter? A natural question raised in Self-Explain is what if the self-generated explanations are wrong since the generation process is not supervised by humans. We look into this question by providing the LLM with random wrong answers to generate misleading explanations of the input question (i.e., we use (x, y_{wrong}) in Equation 3). Those self-explanations with wrong answers i.e., $(x, \pi^{\text{self}}_{\text{wrong}}, y_{\text{wrong}})$ are then used for prompting as in-context exemplars. The results are shown in Table 2. We find that the performance of prompting with self-explanation seems insensitive to its correctness. This result suggests that a correctly labeled dataset may not be necessary for Self-Explain prompting. Similar results on text classification are observed that label space is more important for ICL than label correctness [Min et al., 2022]. We similarly speculate that what carries more weight is how self-explanations are phrased, as they should match the context seen during encoding relevant information according to our encoding specificity hypothesis. We look deeper into this hypothesis in Section 4.3.

4.3 In-context Teaching via Explantions

We have demonstrated LLMs can teach themselves with Self-Explain for better knowledge reasoning. We further extend this to study whether self-explanations of one model can be used as in-context exemplars to teach another model through ICL. Teaching through supervised learning on teacher's generated data has been widely investigated [Zhao et al., 2024, Ho et al., 2023, Hsieh et al., 2023], which can be framed as knowledge distillation. However, machines' supervision through ICL has not yet been well studied. In this section, we have a teacher LLM generate self-explanations that are then used as in-context exemplars to teach a student LLM for reasoning unseen test cases. Saha et al. [2023] have explored a similar research question, while they insert teacher's explanations into student's generation for test examples during inference. This may not be fully considered as *teaching* as the taught model receives assistance with test examples, and its generalization ability is thus not evaluated.

The student is often its own best teacher. Results are shown in Table 3. When doing few-shot prompting with the students' own self-explanations as in-context exemplars, the students can generally reach the best performance, which is aligned with results in Section 4.2. This may also support that the encoding specificity hypothesis for ICL is correct. Noticeably, larger or stronger models may not necessarily be better in-context teachers. For example, for Llama2 and Mistral, prompting them with GPT-3.5's explanations gives worse results than using those models' own self-explanations. The performance for Llama2-7B with GPT-3.5 as the in-context teacher is even worse than not using any demonstrations of reasoning (i.e., "No CoT").

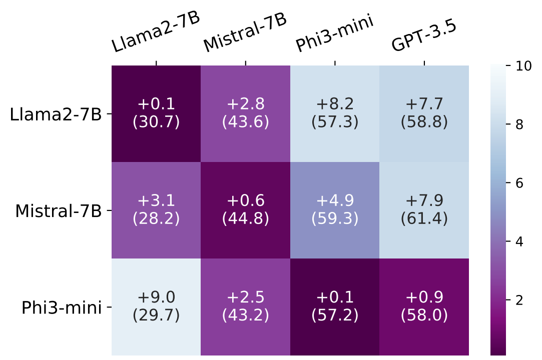


Figure 5: Students' performance improvement after applying Teach-Back. Values in brackets stand for students' respective test accuracy w/ Teach-Back. The x-axis represents students who do reasoning. The y-axis is teacher models that provide in-context demonstrations for students.

Better teachers tend to produce rationales that more resemble student's self-explanations. The performance or scale of the teacher model is shown not indicative of its in-context teaching ability via explanation. Instead, we observe a strong correlation between students' performance and the similarity between teachers' demonstrations and students' self-explanations. Specifically, ROUGE-L [Lin, 2004] is used to measure the closeness between rationales of teachers and students. We observe that explanations generated by different LLMs tend to be more similar to each other than to human-crafted rationales, as shown in Figure 4a. Meanwhile, LLM teachers tend to yield better student's performance than human teachers shown in Table 3. We further compute Pearson correlation coefficient between the student LLM's test accuracy and the ROUGE score between its self-explanations and the teacher LLMs' explanations/ human-crafted rationales. As shown in Figure 4b, evident linear correlations between accuracy and ROUGE score are observed, especially for the models stronger than Llama2-7B. This may further support our encoding specificity hypothesis: in-context exemplars for the student should match its training data because rationales similar to the student's self-explanations are more likely to align with the student's training data.

4.4 Learning from Students for Better Teaching

Given the strong correlation between students' test accuracy and closeness between teacher and student explanations, we are motivated to further look into the underlying causality, i.e., **whether in-context teaching can be improved by letting the teacher learn the student's self-explanations**. In this section, the teacher model will first be fine-tuned on the student's self-explanations to generate its new self-explanations that are then used as in-context exemplars for the student (i.e., Teach-Back introduced in Section 3.2).

We generate each student model's self-explanations for 500 held-out training examples (will not be used for in-context demonstrations) for fine-tuning teacher models. To accommodate our available computing resources, we only fine-tune teacher models whose sizes are smaller or equal to 7B with LoRA [Hu et al., 2021]. Detailed implementations for fine-tuning are shown in Appendix B. Example generations before and after Teach-Back are shown in Appendix E.

Teach-Back improves in-context teaching. As shown in Figure 5, when the teacher model is different from the student model, Teach-Back can greatly enhance the teaching performance of the teacher LLM, as evidenced by higher test accuracy among students. Noticeably, Some fine-tuned teachers using Teach-Back can enable students to achieve significantly higher accuracy than the former best teachers in Table 3. For example, a fine-tuned Mistral-7B can guide Phi3-mini to achieve 59.3% accuracy. This is 4.9% higher than the accuracy achieved with an unfine-tuned Mistral-7B teacher and 1.6% higher than the best unfine-tuned teacher (i.e., GPT-3.5, see the column for "Phi3-mini" in Table 3). Interestingly, Teach-Back enables the smaller Mistral-7B to teach the much larger GPT-3.5 in context, surpassing human teachers by around 5% and Self-Explain by around 4% as visualized in Figure 7 of Appendix. Our results showcase the promising use of Teach-Back in leveraging a small tunable model to improve the few-shot prompting performance of a much larger LM without human's supervision (i.e., human-crafted demonstrations).

5 Further Analysis of LLMs' Explanations

5.1 Faithfulness of Self-explanations

To elicit self-explanations π^{self} from LLMs, we prompt the model to explain a given pair of question and answer (x, y) as shown in Equation 3, and then $(x, \pi^{\text{self}}, y)$ will be used as one in-context demonstration. In this section, we evaluate how many of those raw self-explanations (before filtering) actually support the model to predict the given answer [Hase et al., 2020]. We append the elicited explanations π^{self} back to the given question x as the prompt fed to the model. We then examine whether the model will correctly output the given answer y. We empirically find that diverse models can produce faithful explanations most of the time. For example, Mistral-7B reaches 94.2 % rate of faithful explanations, GPT 3.5 reaches 98.3% and Llama2-13B reaches 93.9 %. Our results may confirm the ability of LLMs' explaining some given questions and answers.

5.2 How Similar are Self-explanations to Human-crafted CoTs?

The common standard to measure the quality of machine-generated samples is how similar they are to human-crafted ones [Lu et al., 2022, Wang et al., 2022]. The more similar, the better the quality is

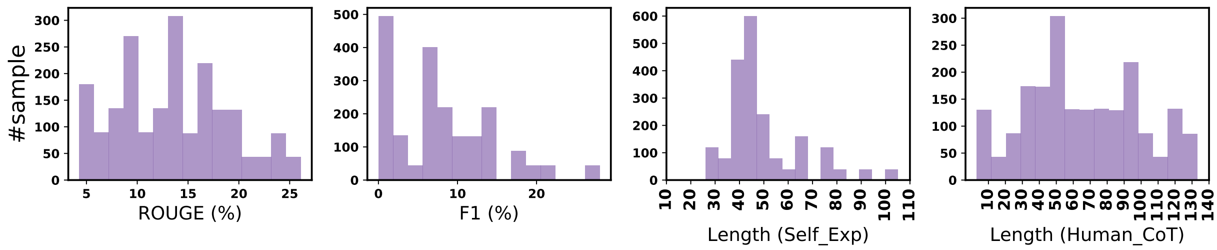


Figure 6: Similarity between human-crafted CoTs and self-explanations in terms of ROUGE score, terminology covered, and length.

assumed to be. However, Hase et al. [2020] have pointed out that evaluation based on plausibility by matching human explanations is not sufficient. Our results also challenge this evaluation criterion. We show that LLMs' self-explanations are very different from human-crafted CoTs in terms of ROUGE-L score, terminology used, and length. However, few-shot prompting with LLMs' self-explanations demonstrates superior performance to using human-crafted CoTs.

We use MedCQA as our testbed, which provides high-quality human-crafted explanations. For terminology comparison, we extract terms in both kinds of CoTs through scispaCy [3] and calculate F1 score between the two terms lists. Results are shown in Fig. 6. We find in terms of content (measured by ROUGE-L and term coverage), self-explanation differs from human-crafted CoTs greatly, given the average similarity is around 15%. The length distribution of self-explanation is more centric, while the human-crafted CoTs have more varied lengths.

6 Related Work

In-Context Learning. In-Contex Learning (ICL) is the ability of language models to induce answers from given demonstrations without weights updating in supervised tuning. In-context exemplars are the key to ICL which have dominating influence on the generation. Quite a few works have been proposed to optimize the selection of exemplars [Lu et al., 2023, Rubin et al., 2022, Fu et al., 2023b]. On the other hand, in the cases of no access to task labels, Lyu et al. [2023] proposed a zero-shot ICL that employs pseudo exemplars with random labels for classification tasks.

Chain-of-Thought prompting without human-crafted exemplars. Prompting with reasoning exemplars triggers LLMs to generate similar intermediate steps of thinking through ICL, known as Chain-of-Thought (CoT) [Wei et al., 2022]. Kojima et al. [2022] propose zero-shot CoT prompting to elicit LLMs' reasoning without human-crafted exemplars. This method is then leveraged to prompt LLMs to generate CoT exemplars by themselves for ICL [Zhang et al., 2023, Wan et al., 2023, Chen et al., 2023]. Different from our work, which focuses on eliciting rationales from LLMs, Zhang et al. [2023], Wan et al. [2023] concentrate on selecting rationales generated according to Kojima et al. [2022]. And Chen et al. [2023] further incorporate pseudo task generation alongside self-generated CoTs. Additionally, Yasunaga et al. [2024] propose analogical prompting to solve emerging new tasks without human-crafted demonstrations. Importantly, these works mainly focus on prompting engineering for very large, closed-source LMs (e.g., GPT-4). None of them formally investigate the fundamental in-context teaching among different LMs. Instead, our work proposes encoding specificity hypothesis to understand in-context teaching for LLMs, which is evidenced by experiments across different models.

Teaching via explanations. Many past works have explored teaching student LLMs with teacher model's explanations through supervised fine-tuning [Ho et al., 2023, Hsieh et al., 2023, Fu et al., 2023a]. Few have investigated in-context teaching. Lampinen et al. [2022] demonstrate that LLMs can learn from human-crafted explanations in context. Saha et al. [2023] feed the teacher's explanations of test examples to the student model during inference. However, this may not be fully considered as *teaching* as the taught model receives assistance with test examples, and thus, the student model's generalization ability from the teacher is not evaluated.

7 Conclusion

In this work, we investigate whether LLMs can teach themselves or other models in the context of knowledge reasoning. We introduce the encoding specificity hypothesis that in-context exemplars at inference should match the encoding context of the model's training data. Motivated by our hypothesis, we propose Self-Explain to let an LLM teach itself with its self-explanations through in-context learning, which outperforms human-crafted chain-of-thoughts and other baselines in different reasoning tasks. We reveal that for in-context teaching, rationales by distinct teacher LLMs or human teachers that more resemble the student LLM's self-explanations are better demonstrations, which further supports the encoding specificity hypothesis. We then propose Teach-Back that aligns the teacher LLM with the student that can enhance the in-context teaching performance.

[3] https://allenai.github.io/scispacy/

8 Limitations

We propose Self-Explain and Teach-Back that verify our encoding specificity hypothesis for few-shot prompting. They also demonstrate impressive performance on diverse models for knowledge reasoning without human guidance. The student model's performance with Self-Explain is consistently among the best. However, the student performance in Teach-Back does not necessarily surpass standard prompting with human CoTs, depending on the teacher model. For example, Mistral-7B with Teach-Back enables different student models to reach optimal test performance, while teachers like Llama2-7B are less effective. Therefore, we suggest using Self-Explain as a starting point in real applications. In the future, we will further investigate the influence of the teacher model in Teach-Back on student performance and how fine-tuning affects the teacher model's self-explanations. Overall, in this work, the main contribution of our proposed Teach-Back is that it can greatly improve the ability of one LLM to teach a different student model.

In addition, our work is limited to only one teacher. Future work could explore many teachers, including mixture of experts. Moreover, there are various emerging advanced prompting methods for different kinds of reasoning tasks, e.g., tree-of-thoughts [Yao et al., 2024] or multi-round prompting [Khot et al., 2023, Zhou et al., 2022]. In this work, we do not consider these more advanced designs of prompting, but focus on commonly used CoT prompting to eliminate the need of human-crafted CoTs. However, our approaches can be adapted to these methods e.g., by modifying the instructions used to elicit LLMs' rationales. The majority of these methods still require human-crafted demonstrations. We will further investigate whether LLMs can implement these advanced prompting methods without human-crafted exemplars under our framework in the future.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Trenton Bricken and Cengiz Pehlevan. Attention approximates sparse distributed memory. *Advances in Neural Information Processing Systems*, 34:15301–15315, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *The Forty-first International Conference on Machine Learning*, 2024.

Wei-Lin Chen, Cheng-Kuang Wu, and Hsin-Hsi Chen. Self-icl: Zero-shot in-context learning with self-generated demonstrations. *arXiv preprint arXiv:2305.15035*, 2023.

Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13 (2):145–182, 1989.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*, pages 10421–10430. PMLR, 2023a.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023b.

Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L Glassman, and Toby Jia-Jun Li. Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.390. URL https://aclanthology.org/2020.findings-emnlp.390.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, 2023.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021a.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qian Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55:1 – 36, 2021b. URL https://api.semanticscholar.org/CorpusID:237450545.

Łukasz Kaiser and Samy Bengio. Can active memory replace attention? *Advances in Neural Information Processing Systems*, 29, 2016.

Pentti Kanerva. *Sparse distributed memory*. MIT press, 1988.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, December 2022.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*, 2022.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*, 2022.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Jinghui Lu, Dongsheng Zhu, Weidong Han, Rui Zhao, Brian Mac Namee, and Fei Tan. What makes pretrained language models better zero-shot learners? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2303, 2023.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064. Association for Computational Linguistics, December 2022.

Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711. Association for Computational Linguistics, December 2023.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, July 2022.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. Can language models teach? teacher explanations improve student performance via personalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=IacxcFpvWQ.

Jason Talevski, Anna Wong Shee, Bodil Rasmussen, Georgie Kemp, and Alison Beauchamp. Teach-back: A systematic review of implementation and impacts. *PloS one*, 15(4):e0231350, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. 2023. URL https://api.semanticscholar.org/CorpusID:264744285.

Endel Tulving. Episodic and semantic memory. 1972.

Endel Tulving and Donald M Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352, 1973.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan O Arik, and Tomas Pfister. Better zero-shot reasoning with self-adaptive prompting. *arXiv preprint arXiv:2305.14106*, 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Beining Wang, Dan Berlowitz, and Hong Yu. Performance of multimodal gpt-4v on usmle with image: Potential for imaging diagnostic support with explanations. *medRxiv*, 2023. doi: 10.1101/2023.10.26.23297629. URL https://www.medrxiv.org/content/early/2023/10/26/2023.10.26.23297629.

Bingsheng Yao, Prithviraj Sen, Lucian Popa, James Hendler, and Dakuo Wang. Are human explanations always helpful? towards objective evaluation of human natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14698–14713, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.821. URL https://aclanthology.org/2023.acl-long.821.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. Large language models as analogical reasoners. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AgDICX1h50.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

Jiachen Zhao. In-context exemplars as clues to retrieving from large associative memory. In *Associative Memory & Hopfield Networks in 2023*, 2023. URL https://openreview.net/forum?id=pgPAsSv5ga.

Jiachen Zhao, Wenlong Zhao, Andrew Drozdov, Benjamin Rozonoyer, Md Arafat Sultan, Jay-Yoon Lee, Mohit Iyyer, and Andrew McCallum. Multistage collaborative knowledge distillation from a large language model for semi-supervised sequence generation. In *Association for Computational Linguistics*, 2024.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

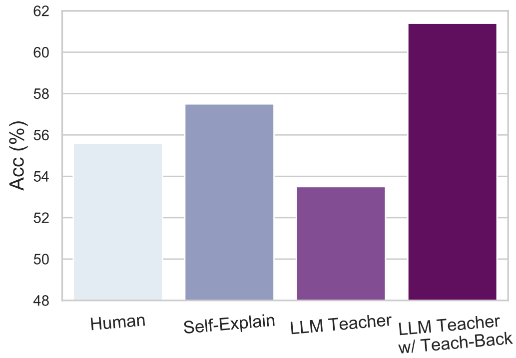


Figure 7: Few-shot prompting performance of GPT-3.5 on knowledge reasoning with different kinds of in-context exemplars of reasoning. "Human" is using chain-of-thought examples crafted by humans. "Self-Explain" is prompting GPT-3.5 with its self-explanations elicited. "LLM Teacher" is using rationales generated by a separate model (Mistral-7B) as teacher, while for "w/ Teach-Back", the teacher model has been first fine-tuned with GPT-3.5's self-explanations.

A Effects of Instructions in Eliciting Self-explanations

In this section, we examine the performance of ICL with self-explanations prompted by different cues in our framework. We mainly follow cues in Liévin et al. [2022] as shown in Table 4. The first one is by default used in our framework. Since Liévin et al. [2022] focuses on medical domains, for general domains, we modify its cues by removing information specific to medical domains. We then generate self-explanations and perform ICL with them. The final test results are shown in Table 5. We find no matter what cues are employed, ICL with self-generations elicited can all outperform using human-crafted CoTs, which demonstrates the robustness of our proposed Self-Explain on the choice of cues.

B Implementation for Teach-Back with Fine-tuning

We reformat the training data with students' self-explanations following the template in Table 6. We set the learning rate as 1×10^{-5} and fine-tune the teacher model with five epochs. We use the default setting for LoRA.

C Analysis on Model Confidence

We find models are more confident with their predictions when using Self-Explain. For exemplars selected for generating self-explanation, we use Text-Davinci-003 to compute the average $P(y|x, \pi^{\text{self}}, \theta)$ which reaches 99.96%. In comparison, for human-crafted explanation of the same exemplars, average $P(y|x, \pi^{\text{human}}, \theta)$ is lower, reaching 89.05%. This implies that given (x, y), self-explanation π^{self} may be

	Medical Domain	General Domain
1	Explain how to reach this answer.	
2	Let's think step by step.	
3	Let's think step by step like a medical expert.	Let's think step by step like an expert.
4	Let's use step by step inductive reasoning, given the medical nature of the question.	Let's use step by step inductive reasoning.

Table 4: Different cues to elicit self-explanations.

Dataset	Cue #1	Cue #2	Cue #3	Cue #4	Human
MedMCQA	56.6	54.6	54.3	54.2	53.1
MedQA	63.7	63.3	62.6	62.3	61.7
StrategyQA	59.7	57.7	57.2	57.3	56.1

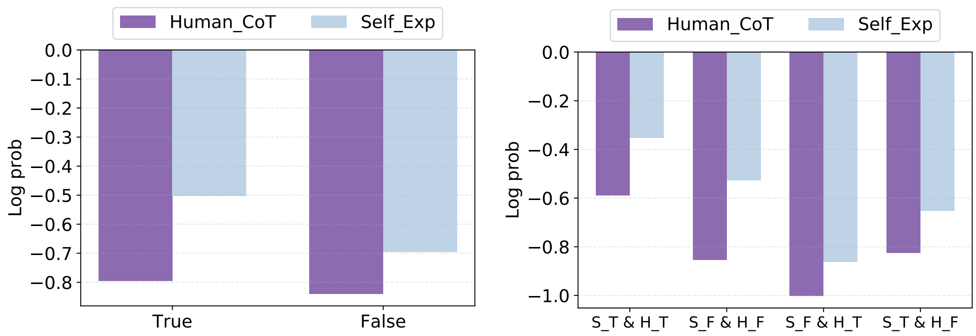
Table 5: Test results of ICL with self-explanations elicited by different cues.

Input:$\{$input string of training example $i\}$
Output:$\{$output result of training example $i\}$
Explain how to reach this answer.
$\{$explanation for training example $i\}$

Table 6: The format of training data for fine-tuning teacher LLMs on students' self-explanations.

much more related context to elicit y than π^{human}.

In addition, for inference with self-explanation as demonstrations in ICL, log probabilities are computed for correct and wrong model outputs prompted with self-explanation and human-crafted one. Results are shown in Figure 8. We can observe that models' output log probabilities with self-explanation are much higher than with human-crafted explanation, indicating greater model's confidence in its output. This suggests self-explanation can be more acceptable and effective for LLMs to elicit reasoning. Self-Explain may also improve model's calibration. Calibration requires the model's output confidence should indicate the correctness of answers (e.g., wrong answers have lower confidence). Well-calibrated output confidence can assist human users to determine whether to trust model's outputs or to look for alternatives in high-stakes applications. Figure 8a showcases the model is generally more calibrated when using self-explanation. Further results in Fig. 8b show that output confidence by using human CoT is not indicative especially when the question is debatable (i.e., using self-explanation and human CoT generate different answers).



(a)

(b)

Figure 8: Comparison of model confidence in different cases by using human-crafted CoT and self-explanation. Prompting with self-explanations makes the model more calibrated and more confident as well. **(a)** Comparison between confidence of true answers and false ones. **(b)** Further comparison of confidence in more specific cases. "S" denotes "Self_exp", "H" denotes "Human_cot", "T" denotes "True" and "F" denotes "False", e.g., "S_T & H_F" means self explanation gives true answer, while human-crafted CoT leads to false answer.

D Example self-explanations

D.1 MedQA

Question: A 22-year-old male marathon runner presents to the office with the complaint of right-sided rib pain when he runs long distances. Physical examination reveals normal heart and lung findings and an exhalation dysfunction at ribs 4-5 on the right. Which of the following muscles or muscle groups will be most useful in correcting this dysfunction utilizing a direct method? (A) anterior scalene (B) latissimus dorsi (C) pectoralis minor (D) quadratus lumborum

Self-explanation: The patient is a pregnant woman, and nitrofurantoin is considered safe for use during pregnancy. Other antibiotics, such as ciprofloxacin and doxycycline, are contraindicated in pregnancy due to potential harm to the fetus.

Human-crafted explanation: We refer to Wikipedia articles on medicine for help. Among the options, only pectoralis minor muscle origins from the outer surfaces of the 3rd to 5th ribs.

Question: A 44-year-old man comes to the office because of a 3-day history of sore throat, nonproductive cough, runny nose, and frontal headache. He says the headache is worse in the morning and ibuprofen does provide some relief. He has not had shortness of breath. Medical history is unremarkable. He takes no medications other than the ibuprofen for pain. Vital signs are temperature 37.4 C (99.4 F), pulse 88/min, respirations 18/min, and blood pressure 120/84 mm Hg. Examination of the nares shows erythematous mucous membranes. Examination of the throat shows erythema and follicular lymphoid hyperplasia on the posterior oropharynx. There is no palpable cervical adenopathy. Lungs are clear to auscultation. Which of the following is the most likely cause of this patient2̌019s symptoms? (A) Allergic rhinitis (B) Epstein-Barr virus (C) Mycoplasma pneumonia (D) Rhinovirus

Self-explanation: The presentation of a fussy infant who is regurgitating all of his feeds and has yellow vomit is consistent with pyloric stenosis, which is characterized by abnormal hypertrophy of the pylorus. This condition causes narrowing of the passage between the stomach and the small intestine, leading to difficulty in food passage and subsequent vomiting.

Human-crafted explanation: We refer to Wikipedia articles on medicine for help. The symptoms, especially the headache, suggest that the most likely cause is Rhinovirus. Epstein-Barr virus will cause swollen lymph nodes but there is no palpable cervical adenopathy. Lungs are clear to auscultation suggests it2̌019s not Mycoplasma pneumonia.

D.2 MedCQA

Question: Which of the following statements about Total Infra Venous Anaesthesia (TIVA) is true? (A) Causes More Renal Toxicity (B) Reduces Cerebral Metabolic Rate (C) Risk of Malignant Hyperthermia is high (D) Inhibits Hypoxic Pulmonary Vasconstriction

Self-explanation: By understanding the mechanism of action: Total Intravenous Anaesthesia (TIVA) involves the administration of intravenous drugs to induce and maintain anesthesia. One of the main drugs used in TIVA is propofol, which is known to reduce cerebral metabolic rate.

Human-crafted explanation: Total intravenous anaesthesia refers to a technique in general anaesthesia using a compination of agents given solely by IV route and in the absence of all inhalation agents. TIVA (Total intravenous anaesthesia) is produced by IV propoful. Propofol decreases the cerebral metabolic rate. Propofol does not impair renal function, does not trigger malignant hyperthermia and does not inhibit hypoxic pulmonary vasoconstriction.

Question: Shock causes: Choose the answer from A to D. A: Stagnant hypoxia. B: Anemic hypoxia. C: Hypoxic hypoxia. D: Histotoxic hypoxia.

Self-explanation: Shock is characterized by a decrease in blood pressure and impaired circulation. This can result in reduced blood flow to the tissues, leading to stagnant hypoxia.

Human-crafted explanation: Ans: A (Stagnant hypoxia)Ref: Pat GK, Hypoxia and Oxygen Therapy, In: Textbook of Medical Physiology, 2nd ed, Abuja Publishing House 2011: 94: 754-55

D.3 StrategyQA

Question: Would ramen be bad for someone with heart failure?

Self-explanation: Ramen can be bad for someone with heart failure due to its high sodium content. Sodium can cause fluid retention and increase blood pressure, which can put additional strain on the heart.

Human-crafted explanation: People with heart failure have to limit their sodium intake. Ramen is notorious for having incredibly high sodium levels.

Questions: Is Cape Town south of the Equator?

Self-explanation: By looking at a map: You can visually locate Cape Town on a map and observe that it is situated below the Equator, in the southern hemisphere.

Human-crafted explanation: Cape Town is an important city in South Africa. South Africa is located entirely south of the Equator.

E Examples of Different LLMs' Self-explanations

Llama2-7B: The patient's symptoms of decreased appetite, difficulty falling asleep, sadness, and compulsive behaviors (rechecking the door lock and counting toilet paper) are consistent with a diagnosis of obsessive-compulsive disorder (OCD). OCD is characterized by recurrent, intrusive thoughts (obsessions) and repetitive behaviors or mental acts (compulsions) that are time-consuming and cause significant distress or impairment.
The neurotransmitter serotonin is involved in the pathophysiology of OCD. Selective serotonin reuptake inhibitors (SSRIs), which increase the levels of serotonin in the synaptic cleft, are the most commonly used medications to treat OCD.

Mistral-7B: The patient's symptoms are consistent with major depressive disorder and obsessive-compulsive disorder (OCD). OCD is characterized by recurrent and intrusive thoughts (obsessions) and repetitive behaviors or mental acts (compulsions) aimed at reducing anxiety or distress caused by the obsessions. Serotonin is a neurotransmitter that plays a crucial role in regulating mood and anxiety, as well as in the modulation of obsessive-compulsive symptoms. Therefore, pharmacotherapy for this patient should be targeted towards increasing serotonin levels in the brain. Dopamine is associated with motivation and reward, glutamate with learning and memory, and norepinephrine with the fight-or-flight response.

E.1 Examples of Self-explanations after Teach-Back

Mistral-7B fine-tuned on Llama2-7B: The patient's symptoms are consistent with major depressive disorder and obsessive-compulsive disorder (OCD). Both conditions are associated with alterations in neurotransmission, particularly serotonin. Serotonin plays a role in mood regulation and impulse control, which are affected in major depressive disorder and OCD, respectively. Additionally, selective serotonin reuptake inhibitors (SSRIs) are commonly used to treat both conditions.

Llama2-7B fine-tuned on Mistral-7B: The patient's symptoms are consistent with a diagnosis of obsessive-compulsive disorder (OCD). OCD is a neuropsychiatric disorder characterized by recurrent, intrusive thoughts (obsessions) and repetitive behaviors or mental acts (compulsions). The most commonly affected neurotransmitter in OCD is serotonin, which is involved in regulating mood, impulse control, and cognitive function. Therefore, pharmacotherapy targeting serotonin should be considered as a first-line treatment for the patient's symptoms.
The other options are not as relevant to the patient's symptoms:
A) Dopamine: While dopamine is involved in reward processing and motivation, it is not typically implicated in OCD.

B) Glutamate: Glutamate is involved in excitatory neurotransmission and is not typically associated with OCD.
C) Norepinephrine: While norepinephrine is involved in arousal and attention, it is not typically implicated in OCD.
In summary, the patient's symptoms are best explained by a diagnosis of obsessive-compulsive disorder, and pharmacotherapy targeting serotonin is the most appropriate treatment option.