SELF-EXPLAIN: Teaching Large Language Models to Reason Complex Questions by Themselves

Jiachen Zhao, Zonghai Yao, Zhichao Yang, Hong Yu
University of Massachusetts Amherst
`jiachenzhao@umass.edu`

Abstract

Large language models (LLMs) can generate intermediate reasoning steps. To elicit the reliable reasoning, the common practice is to employ few-shot chain-of-thought prompting, where several in-context demonstrations for reasoning are prepended to the question. However, such chain-of-thought examples are expensive to craft, especially for professional domains, and can have high variance depending on human annotators. Therefore, this work investigates whether LLMs can teach themselves to reason without human-crafted demonstrations. We propose SELF-EXPLAIN to generate CoT examples by LLMs inspired by "encoding specificity" in human memory retrieval. We find using self-explanations makes LLMs more confident, more calibrated and less biased when answering complex questions. Moreover, we find prompting with self-explanations can even significantly outperform using human-crafted CoTs on several complex question-answering datasets.

1 Introduction

Recently, with the appearance of large langauge models (LLMs), some sparks of human-like intelligence are shown. LLMs can learn the underlying patterns from several in-context exemplars, which is known as in-context learning (ICL) (1). In ICL, no weight updates are implemented and LLMs construct answers based on their encoded knowledge following the given in-context demonstrations. On the other hand, Chain-of-Thought (CoT) prompting (32) shows that LLMs can generate rationale while answering the questions. Moreover, not only can LLMs learn to reason from in-context CoT examples, but also generate explanations in zero-shot prompting (13). However, despite the evidence of zero-shot reasoning ability, LLMs generally perform badly without any in-context demonstrations, especially when the tasks are complex. High-quality human-crafted CoT demonstrations, which are usually considered as the gold standard (17), are needed to guide LLMs to reason correctly through ICL.

However, designing and collecting CoT demonstrations for prompting LLMs can be expensive and difficult. On the one hand, for some professional domains such as medicine, physicians or medical students may be needed to produce fine-grained explanations of the answers (20; 33). On the other hand, collecting CoT examples (usually in the form of long sentences) through crowdsourcing can have great uncertainty and noise (30; 7). The style or the form of CoTs may mainly depend on human annotators' own experience. However, Coda-Forno et al. (3) show that different kinds of expressions of context can actually lead to distinct behaviors of LLMs. Additionally, it is difficult to measure whether those human-crafted CoT examples inform the correct distribution and feature space of the task-specific corpus for effective prompting (18).

Therefore, we are motivated to investigate whether LLMs can teach themselves to reason complex tasks without relying on human-designed CoTs. Our contributions are mainly in three folds: (1)

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

We find LLMs themselves can produce explanations of complex questions given the answer, which we refer to SELF-EXPLAIN. (2) Prompting with self-explanation [1] through ICL also makes LLMs more calibrated and less biased. For example, our experiments show self-explanation makes LLM 1.2% less biased in multiple-choice questions. (3) Prompting with self-explanation can outperform prompting with human-crafted CoT on challenging reasoning tasks even for clinical QA which requires professional medical knowledge. Our findings may challenge the common wisdom that human-crafted CoTs are the gold standard for prompting LLMs.

2 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 (16; 24; 4). On the other hand, in the cases of no access to task labels, Lyu et al. (18) proposed zero-shot ICL that employs pseudo exemplars with random labels for classification tasks. Such zero-shot method can be supported by the finding of Min et al. (19) on classification tasks that the label space carries more weight than correctness of labels in exemplars for ICL.

Prompting with Reasoning Prompting with reasoning triggers LLMs to generate intermediate steps of thinking, known as Chain-of-Thought (CoT) (32). Some works (22; 12; 21; 35) on multi-step reasoning prompt LLM to decompose a complex question into sub-tasks iteratively, showcasing the LLMs' capability of compositional reasoning under some human guidance. Additionally, Kojima et al. (13) propose zero-shot prompting to elicit LLMs' chain-of-thought reasoning without any human-crafted exemplars. This prompting method is leveraged to generate CoT exemplars for ICL (29; 34; 2). Wan et al. (29); Zhang et al. (34) mainly focus on selecting CoTs generated through Kojima et al. (13). Chen et al. (2) further incorporate pseudo task generation apart from pseudo label/ CoTs. Different from those works, we explore a new zero-shot prompting fashion inspired by encoding specificity in human's memory retrieval. Additionally, we further rigorously investigate the difference between human-crafted CoTs and self-generated CoTs and the according influence on model confidence.

3 Revisiting ICL

We first detail some annotations and give a formal setup of ICL. We denote the model parameters as θ, CoT 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^*.

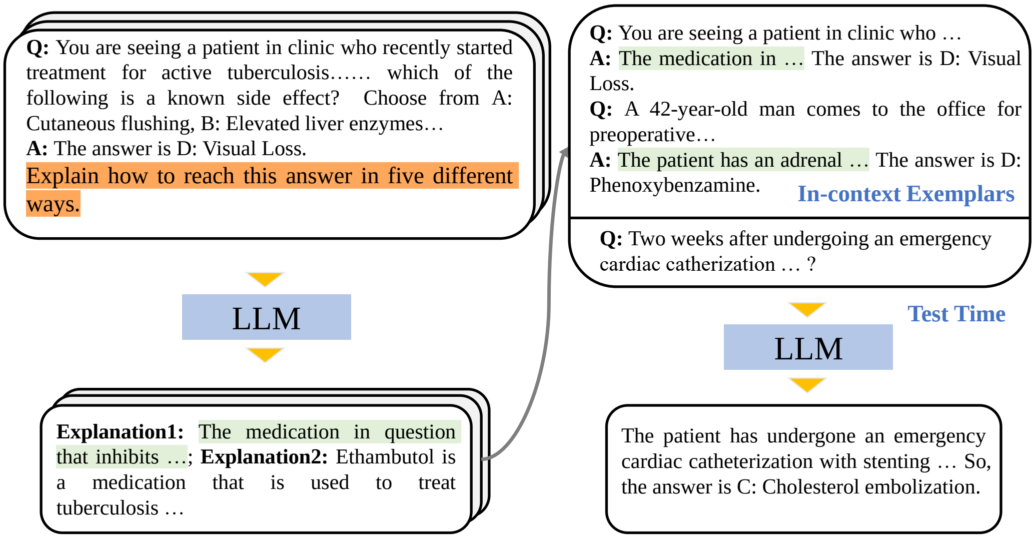
3.1 Encoding Specificity Hypothesis

A key question for CoT prompting is how to compose in-context CoTs for some specific dataset $\mathcal{D}_{(x,y)}$? CoTs 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 (6; 11; 26; 10; 23; 14). The feed-forwarding through hidden layers of LLM is to retrieve and generalize learnt 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 LLM.

From a memory view, CoTs should be composed in a way to match the context in the training corpus seen during encoding similar information to the test data. Such requirement on the match of context between testing and training is known as *encoding specificity* (28) in humans' memory retrieval. To see this hypothesis, a simple thought experiment can be considered: supposing that

---

[1] We use CoT and explanation interchangeably to refer to sentences showing reasoning in the prompt.



Q: You are seeing a patient in clinic who recently started treatment for active tuberculosis…… which of the following is a known side effect? Choose from A: Cutaneous flushing, B: Elevated liver enzymes…
A: The answer is D: Visual Loss.
Explain how to reach this answer in five different ways.

Explanation1: The medication in question that inhibits …; **Explanation2:** Ethambutol is a medication that is used to treat tuberculosis …

Q: You are seeing a patient in clinic who …
A: The medication in … The answer is D: Visual Loss.
Q: A 42-year-old man comes to the office for preoperative…
A: The patient has an adrenal … The answer is D: Phenoxybenzamine.

Q: Two weeks after undergoing an emergency cardiac catherization … ?

The patient has undergone an emergency cardiac catheterization with stenting … So, the answer is C: Cholesterol embolization.

Figure 1: The overall framework of SELF-EXPLAIN. We first generate self-explanations on some sampled training data. At test time, those self-explanations will be used as in-context exemplars for prompting.

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, prompting the LLM with (C, x) will trigger the LLM to generate y with great probability. All in all, the encoding specificity hypothesis implies in-context exemplars should match the distribution of context seen during pretraining, i.e., the training corpus encoded by LLMs containing similar information to (x, y). In that way, LLMs are expected to be motivated to search from related encoded knowledge to form the output generation.

4 SELF-EXPLAIN

Motivated by the encoding specificity, we would like our CoTs to match the training corpus similar to the test data distribution. In order to achieve this, we propose SELF-EXPLAIN where we directly prompt the LLM to elicit the explanation. The overall pipeline of our approach is shown in Fig. 1. The LLM will leverage the encoded knowledge it deems related to the provided question so as to finish the according generation. Those self-explanations are then used as in-context CoT exemplars for ICL. Formally, we consider a setting where we have some data (x, y) sampled from the distribution p^* and $p^{\text{test}} \approx p^*$. We show y does not have to be ground-truth in Sec. 5.2. We assume no human-crafted CoTs are available and define the best CoT as $\pi^* := \text{argmax}_\pi P(y|x, \pi, \theta)$. Self-explanation is represented as,

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

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. Moreover, we find $P(y|x, \pi^{\text{self}}, \theta) >> P(y|x, \pi^{\text{human}}, \theta)$ and $P(y|x, \pi^{\text{self}}, \theta) \approx 1$. Further discussion with supporting experimental results is in Sec. 6. We may arguably state that π^{self} is a very close estimation to π^* and have the following approximation,

$$\text{argmax}_\pi P(\pi|x, y, \gamma, \theta) \approx \text{argmax}_\pi P(y|x, \pi, \theta). \tag{3}$$

This shows the self-explanation to a specific x, is actually more effective and acceptable for LLM to generate the according y, which may evidence the encoding specificity hypothesis. Because self-explanation may be generalized from some related training data in a similar distribution to x.

Generalization through generation diversity. Generated self-explanation will then be used as in-context demonstrations together with corresponding question-answer pairs. However, the underlying logic of π^{self} may be very specific to its corresponding (x, y) and thus lacks generalizability to other different cases. 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 design γ so as to prompt the model to generate solutions employing distinct logics. Formally, we have,

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

where $n \in (1, N)$ and N is the number of different explanations to generate. Then π_i^{self} for some (x_i, y_i) as an in-context exemplar will be randomly sampled from the according $\{\pi_n^{\text{self}} | n \in (1, N)\}$ of (x_i, y_i) for unseen test cases.

5 Experiments

5.1 Experimental Setup

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 LMs (9; 27). We use challenging MedMCQA (20) and MedQA (8) with standard splits that involve multiple-choice questions to diagnose clinical cases, which are used for physician qualification exams. We also evaluate SELF-EXPLAIN on StrategyQA (5) for reasoning in general domains. In terms of implementations, we use GPT-3.5-turbo as our base LLM. Except for Auto-CoT (34), we randomly sample five exemplars from training data and keep exemplars the same for different baselines using ICL. For MedQA which does not provide human-crafted CoTs, in all cases, we use all the five exemplars provided and annotated in Singhal et al. (25).

Method\Dataset	**MedCQA**	**MedQA**	**StrategyQA**
No CoT	51.7	55.1	46.8
Zero-shot CoT (13)	51.1	60.6	45.6
Auto-CoT (34)	52.5	61.9	52.7
Human CoT	53.1	61.7	56.1
Self-Exp	**56.6**	**63.7**	**59.7**

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

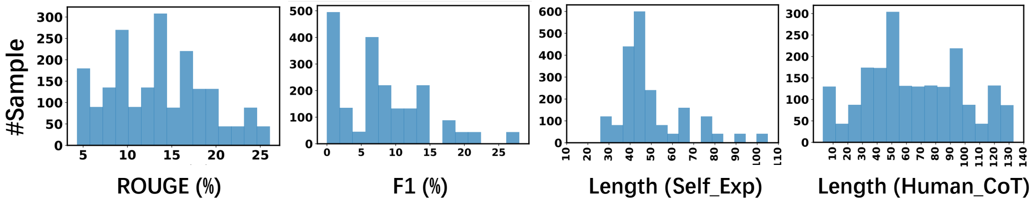
5.2 Results

As is shown in Table 1, using CoT prompting gives better performance than using input-output pairs solely. This demonstrates the importance of CoT in complex reasoning tasks to LLMs. More importantly, prompting with our proposed self-explanation can impressively outperform using CoTs crafted by human professionals by 2% in challenging Medical datasets. Considering the difficulty and expense of crafting CoTs by humans, employing SELF-EXPLAIN can thus be very useful in some domains. Our approach also outperforms Auto-CoT (34) and vanilla zero-shot CoT (13). Additionally, our results may challenge the common belief that human-crafted CoTs are gold standards which are expected to give better performance than using machine-generated CoTs.

	MedMCQA	**MedQA**	**StrategyQA**
Right	56.6	63.7	59.7
Wrong	56.0	63.5	59.1

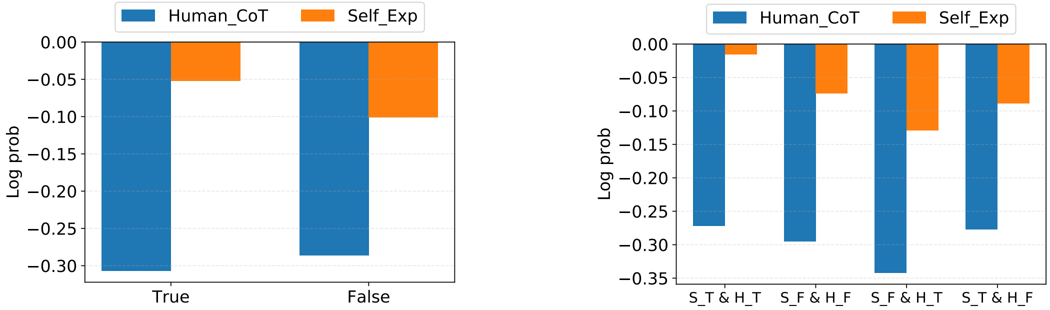
Table 2: Test accuracy of prompting with self-explanations that are generated provided by right answers and wrong answers.

Does the correctness of self-explanation matter? A natural question raised in SELF-EXPLAIN is what if the self-generated CoTs are wrong since the generation process is not supervised by humans. We look into this question by guiding LLMs to generate misleading explanation. More



(a) **(b)** **(c)** **(d)**

Figure 2: Similarity between human-crafted CoTs and self-explanations in terms of ROUGE-L score, terminology covered (terms are extracted by SciSpacy) and length on MedCQA.



(a) (b)

Figure 3: 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.

specifically, we provide the LLM with wrong answers to generate corresponding self-explanations. Those self-explanation for wrong answers 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 is insensitive to its correctness. We speculate that what carries more weight is the underlying distribution of self-explanations. This may evidence encoding specificity hypothesis that the matched distribution of exemplars at test with that of the context seen during encoding relevant information at pre-training is more important for effective prompting through ICL. This result suggests that a labeled dataset is not necessary to elicit CoTs in SELF-EXPLAIN.

How similar are self-explanations and human-crafted CoTs? The common standard to measure the quality of machine-generated samples is how similar they are to human-crafted ones (17; 31). The more similar, the better the quality is. However, despite the superior performance of self-explanation in prompting, we find that they are very different from human-crafted CoTs in terms of ROUGE-L score, terminology used and length. We use MedCQA as testbed which provides high-quality human-crafted explanations. For terminology comparison, we extract terms in both kinds of CoTs through scispaCy [2] and calculate F1 score between the two terms lists. Results are shown in Fig. 2. 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. We attach some example self-explanations in Appendix D.

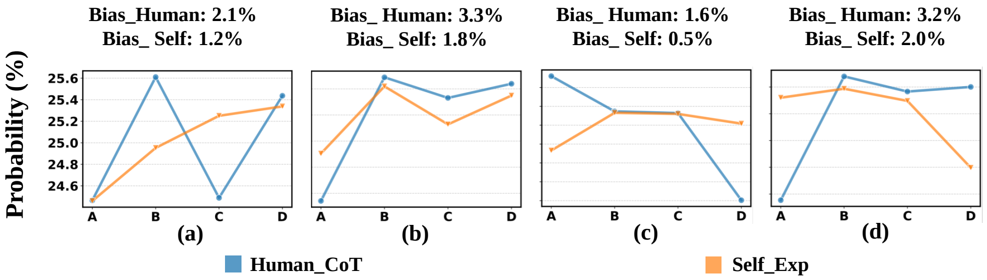


Figure 4: Average confidence of model on each option for a balanced dataset where the number of questions that each option is the correct answer is equal. The correct options in exemplars are fixed to the same choice in each sub-figure for more rigorous comparison. The selection bias is calculated as $\frac{1}{4}\sum_{i=1}^{4}|q_i - 0.25|$ where q_i is model's confidence for each option.

6 Influence on Model Confidence

Models are more confident when prompted by self-explanation. 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%, which may support our estimation of Eq. 3 in Sec. 4. 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 may be the most related context to elicit y. 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 Fig. 3. 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.

We next consider two metrics measuring the influence of prompting with self-explanation and human-crafted CoT on LLM. (1) Calibration: 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. (2) Selection bias: the model should give equal chance of each option for a balanced multi-choice QA dataset. The experiment implementations are detailed in Appendix C. We find that **(1) Self-explanation improves calibration.** Fig. 3a showcases the model is generally more calibrated when using self-explanation. Further results in Fig. 3b 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). Such results are consistent among different LLMs (see Fig. 5 in Appendix). **(2) Self-explanation reduces bias.** We control the correct option of in-context exemplars to be the same (e.g., all the correct options are A) for rigorous comparison. Results in Fig. 4 show that in all cases, using human-crafted CoTs leads to higher selection bias than using self-explanations.

7 Conclusion

In this work, we delve into whether LLMs can teach themselves to reason complex questions. Chain-of-Thought (CoT) can be costly to craft by humans and may have great uncertainty and noise. Therefore, we propose SELF-EXPLAIN inspired by encoding specificity hypothesis in memory retrieval to prompt LLMs to elicit reasonings given a question-answer pair. We find that when doing few-shot prompting with self-explanations as in-context CoT demonstrations, the LLM can surprisingly reach better performance than prompting with human-crafted CoTs on complex questions that require reasoning. We also find that LLM is more calibrated and less biased when prompting with self-explanations. However, comparison on four metrics shows that self-explanation is very different from human-crafted CoTs. Our work may thus challenge the common wisdom that human-crafted CoTs are the gold standard for prompting and possibly shed new light on understanding the effects of exemplars on ICL.

---

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

References

[1] 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.

[2] 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.

[3] Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*, 2023.

[4] 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*, 2023.

[5] 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.

[6] 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.

[7] Emily Jamison and Iryna Gurevych. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, 2015.

[8] 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, 2021.

[9] 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, 2021. URL `https://api.semanticscholar.org/CorpusID:237450545`.

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

[11] Pentti Kanerva. *Sparse distributed memory*. MIT press, 1988.

[12] 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.

[13] 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.

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

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

[16] Jinghui Lu, Dongsheng Zhu, Weidong Han, Rui Zhao, Brian Mac Namee, and Fei Tan. What makes pre-trained 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*, 2023.

[22] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

[23] 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.

[24] 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.

[25] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[26] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. *Advances in neural information processing systems*, 28, 2015.

[27] 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`.

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

[29] 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.

[30] Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47:9–31, 2013.

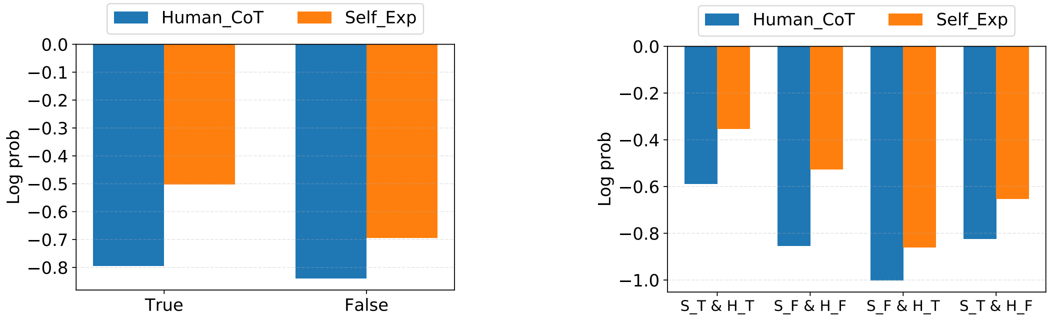
[31] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[32] 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.

[33] 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`.

[34] 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.

[35] 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.



(a) (b)

Figure 5: Comparison of model confidence in different cases by using human-crafted CoT and self-explanation. We use gpt-3.5-turbo-instruct as the tested LLM. **(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. Prompting with self-explanations makes the model more confident and more calibrated especially for cases "S_F & H_T" and "S_T & H_F".

	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 3: Different cues to elicit self-explanations.

A Effects of Cues in SELF-EXPLAIN

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. (15) as shown in Table 3. Since Liévin et al. (15) 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 following our SELF-EXPLAIN framework. The final test results are shown in Table 4. 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 SELF-EXPLAIN framework on the choice of cues.

B Effects of Generation Diversity

In this section, we examine the effects of generation diversity in SELF-EXPLAIN framework. By default, for each input of the exemplar, the LLM is prompted to generate five different ways of explanations to improve the generalization to test cases. At test time, one of five self-explanations is randomly chosen for an exemplar. To better understand the effects of this component in SELF-EXPLAIN, we therefore experiment with prompting different numbers of self-explanations for one exemplar input. Results are shown in Fig. 6. We find more diverse self-explanations to choose for an in-context exemplar can generally improve the test performance, while such improvement experiences diminishing return with further increased number of generations.

C Experiment Implementation Details for Sec. 6

We use MedCQA as testbed which offers human-crafted explanations for all the data. We use text-davinci-003 since it provides access to output probabilities. For the experiments on selection bias,

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

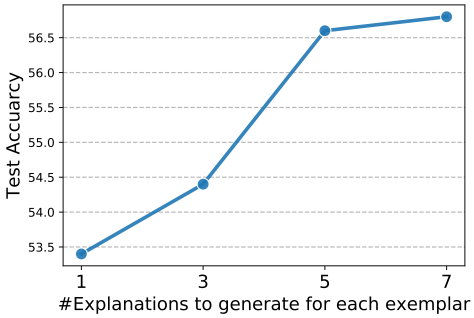


Figure 6: The test performance with respect to the number of available different styles of self-explanations to choose as exemplar CoT for ICL. See further details in Appendix B.

we modify the correct options in in-context exemplars to make them the same in each case, e.g., all correct options are option A so that to give better control of influence of ICL. We append each option to the question to get the model' confidence of the option token.

D Example self-explanation

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 patient2̆019s 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 it2̆019s 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.