EchoPrompt: Instructing the Model to Rephrase Queries for Improved In-context Learning

Anonymous ACL submission

Abstract

Language models are achieving impressive performance on various tasks by aggressively adopting inference-time prompting techniques, such as zero-shot and few-shot prompting. In this work, we introduce *EchoPrompt*, a simple yet effective approach that prompts the model to rephrase its queries before answering them. EchoPrompt is tailored for four scenarios, including standard and chain-of-thought prompting, in both zero-shot and few-shot settings. Experimental results show that EchoPrompt yields substantial improvements across all these settings for four families of causal language models. These improvements are observed across various numerical reasoning (e.g., GSM8K, SVAMP), reading comprehension (e.g., DROP), and logical reasoning (e.g., Coin flipping) tasks. On average, EchoPrompt improves the Zero-shot-CoT performance of code-davinci-002 by 5% in numerical tasks and 13% in reading comprehension tasks. Our empirical results indicate that EchoPrompt is an effective technique that enhances in-context learning performance.

1 Introduction

Large language models have revolutionized natural language task-solving through prompting (Brown et al., 2020). This technique involves conditioning the language model with an instruction (zero-shot) or augmenting the prompt with a small set of task-specific examples (few-shot), resulting in the model generalizing and responding effectively to tasks.

A rapidly advancing body of research has introduced techniques to enhance these prompting methodologies. Notably, chain-of-thought prompting (Wei et al., 2023; Kojima et al., 2022) has emerged as a powerful method for enhancing language model performance in reasoning tasks. Least-to-most prompting (Zhou et al., 2022a) and Tree of Thoughts (Yao et al., 2023) further support chain-of-thought by breaking down complex problems into simpler subproblems.

Q: Kelly has 5 quarters and 2 dimes. If she buys a can of pop for 55 cents, how many cents will she have left?

Zero-Shot-CoT
A: Let's think step by step. *Kelly has 5 quarters and 2 dimes. So, she has 5 quarters and 2 dimes left. 5 quarters = 5 x 25 cents = 125 cents and 2 dimes = 2 x 10 cents = 20 cents. So, she has 125 + 20 = 145 cents left.*
Therefore, the answer is *145 cents.* ✗

Zero-Shot-CoT + EchoPrompt
A: Let's repeat the question and also think step by step. *Kelly has 5 quarters and 2 dimes. She buys a can of pop for 55 cents. How many cents will she have left? Let's start with the first step. Kelly has 5 quarters and 2 dimes. We can represent this as: 5 quarters = 5 x 25 cents = 125 cents. 2 dimes = 2 x 10 cents = 20 cents. So, Kelly has 125 + 20 = 145 cents. She buys a can of pop for 55 cents. We can represent this as: 55 cents. So, Kelly has 145 - 55 = 90 cents left.*
Therefore, the answer is *90 cents.* ✔

Figure 1: Comparison of prompts in Zero-shot-CoT with and without EchoPrompt, highlighting the modification in prompts. Zero-shot-CoT with EchoPrompt uses the prompt "Let's repeat the question and also think step by step" to aid the model in recalling the query before solving it.

In this paper, we propose EchoPrompt, a prompting strategy that builds upon existing prompting approaches by incorporating *query-rephrasing* as a preliminary task in the in-context learning process. EchoPrompt draws inspiration from the innate cognitive strategies employed by humans, precisely the act of self-questioning, when answering queries. By verbalizing queries before answering them, humans establish a cognitive checkpoint to refine their thoughts (Joseph and Ross, 2018; Joseph et al., 2019). Figure 1 provides an illustrative example of EchoPrompt in Zero-shot-CoT settings. While the approach proposed by (Kojima et al., 2022) uses the prompt *"Let's think step by step."* to elicit chain-of-thought reasoning and extracts the answer

Standard Prompting	**Standard Prompting + Echo Prompt**
Q: If Pam is currently twice as young as Rena is, and in 10 years Rena will be 5 years older than her, how old is Pam now? A: The answer is 5. [more exemplars] Q: Kelly has 5 quarters and 2 dimes. If she buys a can of pop for 55 cents, how many cents will she have left? A: *The answer is 145.* ✗	Q: If Pam is currently twice as young as Rena is, and in 10 years Rena will be 5 years older than her, how old is Pam now? A: Rewriting in simple words, the question is: "Given that Pam is currently twice as young as Rena and that in 10 years Rena will be 5 years older than Pam, how old is Pam now?" Now, to answer the rewritten question, the answer is 5. [more exemplars] Q: Kelly has 5 quarters and 2 dimes. If she buys a can of pop for 55 cents, how many cents will she have left? A: *Rewriting in simple words, the question is: "Given that Kelly has 5 quarters and 2 dimes, and she buys a can of pop for 55 cents, how many cents will she have left?"* *Now, to answer the rewritten question, the answer is 90.* ✔

Figure 2: Example of EchoPrompt with compound sentence rephrasing. EchoPrompt incorporates a *query-rephrase* step, where the exemplars showcase a rephrased version of the query along with the answering format.

using the prompt *"Therefore, the answer is"*, we modify the first prompt to *"Let's repeat the question and also think step by step."* or similar texts. This modification guides the model to generate a version of the original query before solving it.

We empirically evaluate our approach against various prompting baselines using a wide variety of model families with different sizes, including code-davinci-002, GPT-3.5-Turbo. Our results show that EchoPrompt significantly improves the performance of language models on arithmetic, reading comprehension, and logical reasoning tasks. We observe substantial performance gains with both standard and chain-of-thought prompting, particularly in zero-shot scenarios for large language models (code-davinci-002, GPT-3.5-turbo, CodeLlama-Instruct-34B) and with standard prompting on smaller models (Starcoder-15B, and GPT-J-6B). For example, EchoPrompt improves the chain-of-thought performance from 59.7% to 65.2% on DROP (Break) and from 76.1% to 82.6% on GSM8K on GPT-3.5-Turbo. Overall, our study indicates that EchoPrompt fundamentally improves in-context learning performance and finds broad applicability as a building block in emerging complex techniques that leverage prompting in multiple stages.

2 EchoPrompt

EchoPrompt teaches language models to generate a version of the query before solving it. The fine-grained details of this technique are explained in the following subsections, with examples.

2.1 Zero-shot EchoPrompt

In zero-shot prompting, the standard approach relies on a single prompt, *"Therefore, the answer is"*, to directly extract the answer. In contrast, zero-shot EchoPrompt introduces a two-stage prompting process. The language model is initially instructed to rephrase the query using a task-agnostic prompt, *"Let's repeat the question. "*, and then the answer is extracted using the same prompt as in zero-shot prompting.

Similarly, in Zero-shot-CoT (Kojima et al., 2022), the conventional approach involves using the prompt *"Let's think step by step."* to guide the model in generating its reasoning steps before producing the final answer. However, in Zero-shot-CoT with EchoPrompt, we introduce a query-rephrasing subtask by employing prompts like *"Let's repeat the question and also think step by step."*. This modification encourages the model to generate the query in its own words and then engage in reasoning. The prompt used for answer extraction remains consistent across all zero-shot methodologies. Figure 1 shows an example, highlighting the key differences between the two approaches.

2.2 Few-shot EchoPrompt

In few-shot learning, we teach the language model to rephrase the test query in a particular structure before answering the query. We do this by providing exemplars demonstrating the rephrase structure (Figure 2) and corresponding responses to example queries. Figure 4 in appendix shows an example with all the rephrasing formats, for a given query. We examine three distinct rephrasing structures in addition to teaching the model to repeat the exact query in the following formats:

- **Rephrased to *Compound Sentences***: Queries are formulated using compound sentences incorporating multiple clauses or phrases.

- **Rephrased to *Question First***: Queries are structured to present the final question at the beginning, followed by contextual information.

Datasets	Arithmetic	GSM8K, SVAMP, SingleOp, MultiArith, AQuA, MMLU-h, GSMIC-4k
	Logical Reasoning	Date Understanding, Shuffled Objects, LogiQA, Coin flipping
	Reading Comprehension	DROP (Football, Non-football, Census, Break), SQuAD
	Commonsense Reasoning	StrategyQA, Winogrande
Language Models	-	Codex, GPT-3.5-Turbo, CodeLlama-Instruct, StarCoder-15B, GPT-J-6B

Table 1: Evaluation setup used for Echoprompt

- **Rephrased to *Short and Simple Sentences***: Queries are constructed by breaking down the original problem's context into simpler and shorter sentences.
- ***Repetition***: Repeating the original query itself can serve as a fundamental form of rephrasing, and we consider it one of the rephrase structures.

3 Evaluation Setup

Our evaluation setup ensures a comprehensive analysis of EchoPrompt's performance across diverse benchmarks and language models. We assess the performance of EchoPrompt on various tasks, spanning across numerical reasoning, logical reasoning, reading comprehension, and commonsense reasoning. Table 1 summarizes the datasets and language models used as part of our evaluation. Please refer to Appendix B for additional details regarding the datasets and prompts used.

4 Results

We conduct an extensive comparison of our approach against zero-shot, Zero-shot-CoT, few-shot, and few-shot-CoT prompting strategies. Figure 3 (and Table 10 in the appendix) summarizes the overall performance of EchoPrompt, while detailed analysis and extended results are available in Appendix A and D respectively. The findings on individual models are summarized below.

Code-davinci-002 Overall, we observe that EchoPrompt performs well regardless of the baseline prompting strategy. Notably, EchoPrompt shows significant improvements in zero-shot scenarios, especially for tasks with longer query contexts, such as different DROP and SQuAD subsets containing extraneous information. For example, we observe an 18.5% improvement in accuracy on the DROP (Census) for zero-shot prompting. Similarly, EchoPrompt with Zero-shot-CoT on SVAMP achieves a 7.4% improvement in accuracy, making the overall accuracy comparable to few-shot-CoT prompting. However, it is worth noting that EchoPrompt does not yield any improvements in cases where the baseline method cannot solve the task, as it is not a standalone strategy but rather an approach to reduce errors in the reasoning process. For example, in the Shuffled Objects task, EchoPrompt shows a slight drop in zero-shot performance (36.4% to 35.2%), which is close to random choice (33.3%). Nevertheless, it considerably improves the accuracy of Zero-shot-CoT (42.4% to 58.2%) when the model can partially solve the task. We also do not observe significant improvements in commonsense reasoning tasks, where the ability to answer relies more on the model's inherent capabilities rather than deducing information explicitly from the queries. Similarly, in tasks involving multiple-choice questions, such as AQuA-RAT, MMLU, and LogiQA, EchoPrompt does not exhibit substantial gains as these tasks necessitate selecting the correct option among several choices rather than an unbiased answer generation.

GPT-3.5-Turbo and CodeLlama-Instruct-34B To evaluate the effectiveness of EchoPrompt on models with performance comparable to Code-davinci-002, we conduct experiments using GPT-3.5-Turbo and CodeLlama-Instruct models on a selected subset of tasks. Detailed results can be found in Table 10 in the appendix. Overall, our findings align with our previous experiments on code-davinci-002. Notably, EchoPrompt led to significant accuracy improvements, with a 6.5% increase for GPT-3.5-Turbo and a 4.6% increase for CodeLlama-Instruct models in few-shot-CoT settings. However, it is worth noting that manual qualitative analysis of false negatives was required for GPT-3.5-Turbo due to its tendency to generate descriptive responses rather than instruction-based, extractable answers.

StarCoder-15B, GPT-J-6B Similarly, we evaluate EchoPrompt on smaller and publicly available models: StarCoder-15B and GPT-J-6B. Since these models are less capable of challenging reasoning tasks, we evaluate our approach on coin-flipping, SingleOp, SVAMP, and date-understanding tasks. The results are summarized in Table 10 in the ap-

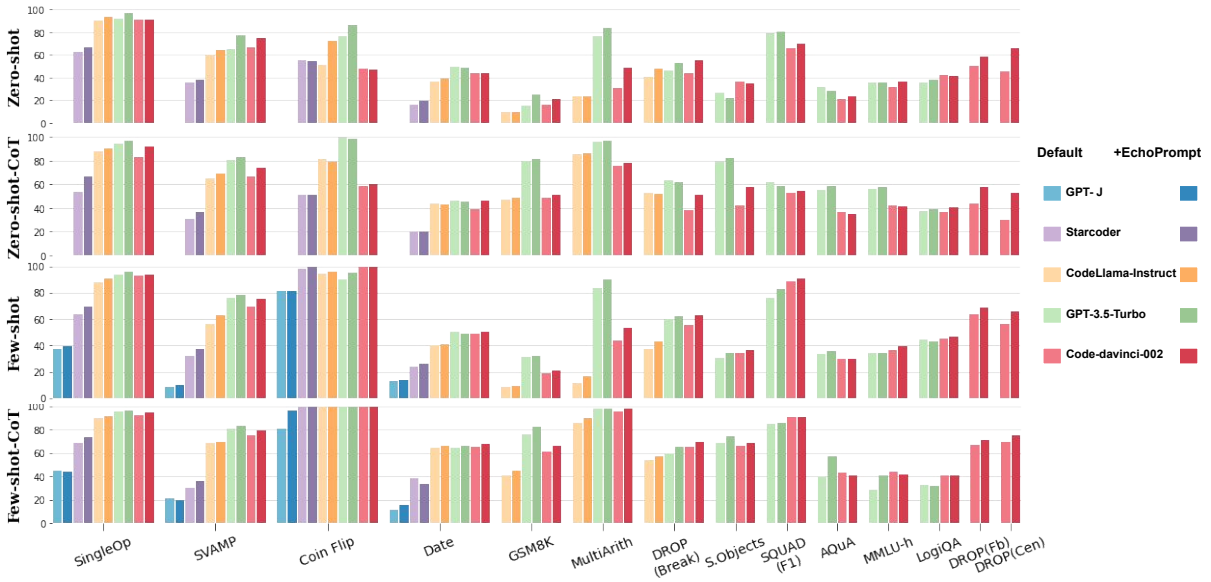

Figure 3: Performance summary of EchoPrompt with repetition in zero-shot and compound sentence rephrasing in few-shot settings. Experiments on some datasets were limited to certain models, to manage computational costs. EchoPrompt consistently achieves performance gains across different prompting strategies, particularly in zero-shot scenarios. For details, see Table 10.

pendix. We observe that EchoPrompt improves performance with standard prompting, although we observe inconsistent results with chain-of-thought reasoning. This finding is not surprising, as chain-of-thought is considered an emergent phenomenon in larger language models (Wei et al., 2023).

5 Related Work

Prompting The success of LLMs has sparked interest in improving task performance through prompting techniques (Brown et al., 2020). Recent studies focus on task-based instruction tuning, by fine-tuning the entire model (Raffel et al., 2020; Wei et al., 2021; Sanh et al., 2021; Wang et al., 2022b; Huang et al., 2022), maintaining task-specific parameters (Li and Liang, 2021; Lester et al., 2021), and optimization strategies (Yang et al., 2023; Prasad et al., 2022; Xu et al., 2022; Zhou et al., 2023). Our work is a general prompting approach that improves in-context learning abilities and falls into the domain of works inspired by human cognitive processes (Yasunaga et al., 2023; Webb et al., 2023; Li et al., 2023a)

Intermediate steps The concept of employing language models to generate intermediate steps for process supervision has been extensively examined in the context of solving reasoning tasks, whether through training (Nye et al., 2021; Zelikman et al., 2022), zero-shot (Kojima et al., 2022), few-shot prompting (Wei et al., 2022) or action planning (Yao et al., 2022). Recent works focus on problem decomposition and teaching the language model to answer the subtasks to eventually answer complex problems (Zhou et al., 2022a; Dua et al., 2022; Wang et al., 2023a, 2022a; Zhou et al., 2022b). EchoPrompt is orthogonal to these approaches, augmenting the input query rather than rationale generation. Consequently, it can be easily extended with any of these prompting strategies. Please refer to Appendix C for additional related works

6 Conclusion and Future Work

We have proposed EchoPrompt, a simple yet effective approach that builds upon existing prompting approaches and integrates query-rephrasing as a subtask in the in-context learning process. Despite its simplicity, understanding the mechanisms that make EchoPrompt effective requires further exploration. We conducted ablation studies (detailed in Section A in the Appendix) to unravel its working, yet the intricate reasons behind its consistent performance improvements, especially in standard prompting scenarios, remain elusive. Our analysis, including saliency and perplexity studies of the complete prompts provided to the language model, did not yield any conclusive results. This challenge opens up several avenues for future work, aiming to gain valuable insights and potentially develop more effective prompting techniques.

7 Reproducibility Statement

Our primary results are on Code-davinci-002 and GPT-3.5-Turbo[1], which are publicly accessible OpenAI models. For other models employed in our experiments, we make use of transformer models available through HuggingFace[2]. The football subset of the DROP dataset was curated by applying keyword-based filtering with the keyword "yard" (Zhou et al., 2022a), and the census subset was created by selectively filtering passages that contained the terms "population" and "census." For all the experiments, we employed greedy decoding with top_p=1 and no presence or repetition penalties. The number of tokens generated varies, ranging from 50 to 600 tokens from the models, depending on the task and considering the query length at each stage. To increase reproducibility, we have included prompts for all the tasks in the appendix. We also plan to release the code soon.

References

2019. Winogrande: An adversarial winograd schema challenge at scale.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Zieglxer, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2022. Codet: Code generation with generated tests. *arXiv preprint arXiv:2207.10397*.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *arXiv preprint arXiv:2212.04092*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback.

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

Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. 2013. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023. Self-evolve: A code evolution framework via large language models.

[1] https://openai.com/blog/chatgpt/. We use gpt-3.5-turbo-0301 snapshot from March 2023

[2] Source code at https://huggingface.co/

Laurice M Joseph, Sheila Alber-Morgan, Leigh Ann Amspaugh, Kelsey Ross, Maria Helton, Moira Konrad, and Carrie Davenport. 2019. Stop to ask and respond: Effects of a small-group self-questioning intervention on reading comprehension performance. *Research and Practice in the Schools: The Official Journal of the Texas Association of School Psychologists*, 6(1):27.

Laurice M Joseph and Kelsey M Ross. 2018. Teaching middle school students with learning disabilities to comprehend text using self-questioning. *Intervention in School and Clinic*, 53(5):276–282.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Discriminator-guided multi-step reasoning with language models.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. 2023a. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023b. Starcoder: may the source be with you!

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2021. A diverse corpus for evaluating and developing english math word problem solvers.

Ansong Ni, Srini Iyer, Dragomir Radev, Ves Stoyanov, Wen tau Yih, Sida I. Wang, and Xi Victoria Lin. 2023. Lever: Learning to verify language-to-code generation with execution.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems?

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Boshi Wang, Xiang Deng, and Huan Sun. 2022a. Iteratively prompt pre-trained language models for chain of thought. *arXiv preprint arXiv:2203.08383*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. Decomposition enhances reasoning via self-evaluation guided decoding.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. Gps: Genetic prompt search for efficient few-shot learning. *arXiv preprint arXiv:2210.17041*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2023. Large language models as analogical reasoners.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2022a. Least-to-most prompting enables complex reasoning in large language models.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022b. Teaching algorithmic reasoning via in-context learning. *arXiv preprint arXiv:2211.09066*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers.

EchoPrompt?	Stage-1 Prompt	GSM8K	SVAMP	MultiArith	SingleOp
Zero-shot					
✗	-	16.4	66.8	31.0	91.6
✓	Let's repeat the question. "	**20.7**(+4.3)	**74.7**(+7.9)	48.5(+17.5)	91.8(+0.2)
✓	Let's reiterate the question. "	19.7(+3.3)	73.4(+6.6)	**51.0**(+20.0)	93.0(+1.4)
✓	Let's restate the question. "	19.2(+2.8)	74.6(+7.8)	47.7(+16.7)	89.6(−2.0)
✓	Let's summarize the question. "	20.6(+4.2)	73.2(+6.4)	48.8(+17.8)	**93.7**(+2.1)
Zero-shot-CoT					
✗	Let's think step by step.	49.3	66.5	76.0	82.9
✓	Let's repeat the question and also think step by step.	44.6(−4.7)	**74.7**(+8.2)	70.9(−5.1)	92.3(+9.4)
✓	Let's reiterate the question and also think step by step.	**51.1**(+1.8)	73.9(+7.4)	78.7(+2.7)	**92.4**(+9.5)
✓	Let's repeat the question and also think step by step. "	42.0(−7.3)	60.4(−6.1)	78.1(+2.1)	88.3(+5.4)
✓	Let's restate the question and also think step by step.	47.0(−2.3)	73.9(+7.4)	**79.3**(+3.3)	90.2(+7.3)

Table 2: **code-davinci-002: Arithematic reasoning** Evaluation of EchoPrompt on various prompt templates. All the prompts improve the performance in zero-shot setting. However, we find that only the prompt "Let's reiterate the question and also think step by step." consistently outperforms baseline Zero-shot-CoT.

	EchoPrompt	GSM8K	SVAMP	MultiArith	DROP (Census)	DROP (Break)	DROP (Football)	SQuAD(F1)
Standard	-	19.2	69.8	44.0	56.8	55.5	63.7	88.7
	Repeat	21.4(+2.2)	75.8(+6.6)	53.8(+9.8)	65.9(+9.1)	**63.1**(+7.6)	**69.2**(+5.5)	91.3(+2.6)
	Compound	20.8(+1.6)	75.1(+5.3)	54.0(+10.0)	**67.3**(+10.5)	62.7(+6.9)	67.7(+4.0)	90.6(+1.9)
	Question First	20.9(+1.7)	75.0(+5.2)	53.6(+9.6)	65.2(+8.4)	59.7(+3.9)	63.1(−0.6)	**92.2**(+3.5)
	Simple	**21.5**(+2.3)	**76.6**(+6.8)	**55.6**(+11.6)	65.1(+8.3)	**63.1**(+7.6)	67.1(+3.4)	90.9(+2.2)
CoT	-	61.1	75.2	96.1	70.0	65.3	67.3	90.5
	Repeat	63.5(+2.4)	77.6(+2.4)	98.8(+2.7)	71.6(+1.6)	**70.0**(+4.7)	71.3(+4.0)	-
	Compound	**65.9**(+4.8)	**79.0**(+3.8)	97.8(+1.7)	**75.4**(+5.4)	69.6(+4.3)	70.8(+3.5)	90.8(+0.3)
	Question First	64.4(+3.3)	77.0(+1.8)	98.3(+2.2)	75.3(+5.3)	68.1(+2.8)	**72.0**(+4.7)	-
	Simple	63.6(+2.5)	76.9(+1.7)	**99.0**(+2.9)	73.5(+3.5)	67.7(+2.4)	71.2(+3.9)	-
LTM	-	63.2(+2.1)	**82.2**(+7.0)	93.7(−2.4)	73.8(+3.8)	61.2(+4.1)	66.2(+1.1)	-

Table 3: **code-davinci-002** Evaluation of EchoPrompt using the proposed rephrase structures and query-repetition. We compare these approaches with baseline methods in arithmetic and reading comprehension tasks. The results showcase improvements across all rephrase structures, with no single structure consistently outperforming the others.

	Query Structure	GSM8K	SVAMP	DROP (Census)	DROP (Break)	DROP (Football)
Standard	Original	19.2	69.8	56.8	55.5	63.7
	Compound	19.9(+0.7)	71.8(+2.0)	59.1(+2.3)	54.1(−1.4)	65.1(+1.4)
	Question First	14.6(−4.6)	58.5(−11.3)	28.2(−28.6)	36.2(−19.3)	48.8(−14.9)
	Simple	19.7(+0.5)	70.9(+1.1)	56.5(−0.3)	55.5(+0.0)	62.7(−1.0)
Standard+ Repeat	-	**21.5**	**76.6**	**65.1**	**63.1**	**67.1**
CoT	Original	61.1	75.2	69.6	65.3	67.3
	Compound	62.1(+1.0)	78.0(+2.8)	71.9(+2.3)	66.7(+1.4)	68.2(+0.9)
	Question First	55.1(−6.0)	66.6(−8.6)	48.1(−21.5)	64.5(−0.8)	57.8(−9.5)
	Simple	61.3(+0.2)	75.8(+0.6)	70.3(+0.7)	67.3(+2.0)	67.1(−0.2)
CoT+ Compound	-	**65.9**	**79.0**	**74.3**	**69.6**	**70.8**

Table 4: **Standalone Rephrases: code-davinci-002 Standalone Rephrases** Compound Sentence rephrasing performs better than the original queries, while Question-First rephrasing performs worse. We observe information loss in the rephrases for certain tasks, indicating that the performance gains of EchoPrompt are due to the combination of rephrasing and having multiple versions.

		GSM8K	SVAMP	DROP
Repeat	SubTask	63.5	77.6	70.0
	Augment	63.4	76.3	69.3
Compound	SubTask	65.9	79.0	69.6
	Augment	64.2	77.2	69.7

Table 5: **code-davinci-002** A comparison between EchoPrompt and query augmentation, indicating similar performance improvements for both approaches.

	times	GSM8K	SVAMP	DROP
Repeat	1	**63.5**	**77.6**	**70.3**
	3	59.8	77.8	69.3
	5	59.9	76.9	67.5
Compound	1	**65.9**	**79.0**	**69.6**
	2	63.7	77.9	68.8
	3	63.2	78.9	67.9

Table 6: **code-davinci-002** The accuracies drop as the number of rephrases/repetitions increases when generating multiple rephrases with EchoPrompt.

A Analysis

To gain a deeper understanding of the factors that contribute to the success of EchoPrompt, we perform a series of ablation studies in the following sections:

Effect of prompts on zero-shot EchoPrompt To investigate the impact of prompts used to instruct the language model for rephrasing queries in zero-shot settings, we experiment with a variety of prompts on arithmetic tasks, including both standard and chain-of-thought prompting. The results shown in Table 2 indicate that EchoPrompt consistently enhances performance when compared to the baseline method, regardless of the chosen prompt. However, we observe a difference in performance with various prompt selections in the Zero-shot-CoT setting. The prompt "Let's reiterate the question and also think step by step." achieves the best results.

Effect of rephrases on few-shot EchoPrompt In the few-shot setting, we assess the performance of the proposed rephrase structures compared to baseline techniques, focusing on arithmetic and reading comprehension tasks that require explicit answer generation. The results, as shown in Table 3, reveal that although there is variance among the performances, all the rephrase structures outperform the standard and chain-of-thought prompting, highlighting the effectiveness of EchoPrompt. Notably, no single rephrase structure consistently outperforms the others.

Are rephrased queries self-sufficient? To assess whether the EchoPrompt performance gains are solely due to the rephrased queries or if both the original and rephrased queries are essential, we isolate the LM-generated rephrases. This process involves two steps. First, we generate the rephrased query using the same method as before and with the same exemplars through in-context learning. Then, we prompt the language model with the revised exemplars that match the rephrased query structure. We only provide the rephrased queries for the model to answer. The results in Table 4 show that standalone rephrases yield lower accuracy than EchoPrompt. Although rephrased queries can improve accuracy compared to baseline prompting (compound sentence rephrases), the improvements are still considerably lower than those achieved with EchoPrompt. This suggests that the primary source of improvement in EchoPrompt lies in the provision of two query versions.

Generating vs Augmenting the rephrases To study whether EchoPrompt can be considered as an augmentation technique, we compare the performance of EchoPrompt with directly augmenting the original question using a rephrase. In EchoPrompt, the model generates both the rephrase and the answer, while in query augmentation, the query is provided to the language model beforehand, and the model only generates the answer. The result of this experiment is summarized in Table 5, demonstrating that both approaches yield comparable improvements. This result indicates that although we introduce EchoPrompt as a subtask within in-context learning, it can also be considered a query-augmentation technique.

Stacking multiple rephrases for EchoPrompt The benefits observed with query-rephrasing in EchoPrompt naturally prompts us to investigate the effects of having the language model generate multiple rephrases. The summarized results in Table 6 show a drop in performance as the number of rephrases increases. When manually examining the generated answers, we observe a tendency towards repetition in the CoT reasoning despite successfully generating the desired number of rephrases. This repetition phenomenon is prominent when the question requires longer multi-hop reasoning. This observation aligns with expectations since the task's focus shifts from reasoning to rephrase generation when the number of rephrases is increased

	Standard		CoT		LTM	
EchoPrompt?	✗	✓	✗	✓	✗	✓
Zero-shot	23.7	30.1 (+6.4)	46.7	52.8 (+6.1)	N/A	N/A
1-shot	27.1	29.1 (+2.0)	72.6	77.2 (+4.6)	73.8	81.3 (+7.5)
4-shot	25.2	31.0 (+5.8)	77.4	81.8 (+4.4)	84.3	85.4 (+1.1)

Table 7: **code-davinci-002** Performance of EchoPrompt on GSMIC-4k(which contains irrelevant context in queries). EchoPrompt improves performance on both chain-of-thought and least-to-most prompting, even though it repeats the perturbation sentence in the rephrase.

in EchoPrompt. Consequently, the model prioritizes generating the requested number of rephrases rather than the reasoning process.

Robustness to irrelevant text Recent work (Shi et al., 2023) has shed light on the sensitivity of large language models (LLMs) to irrelevant information using various prompting methods, including the CoT reasoning. Intuitively, EchoPrompt could be particularly prone to such distractions, given that it rephrases or regenerates the query, including the distractions. To evaluate if EchoPrompt technique works even in the presence of such perturbations, we study the performance of EchoPrompt on GSMIC-4k dataset (Shi et al., 2023). The evaluation results in Table 7 demonstrate that EchoPrompt maintains its improvements, even in the presence of perturbations.

B Evaluation Setup

B.1 Benchmarks

We evaluate EchoPrompt across a range of natural language processing tasks, specifically focusing on four types, including fourteen widely recognized benchmarks. We experiment with four categories of causal language models to ensure a broad and thorough evaluation. In this section, we delve into the details of our evaluation setup.

Numerical Reasoning We evaluate numerical reasoning tasks from (Wei et al., 2023) for a fair comparison between the methods, including **GSM8K** (Cobbe et al., 2021), **SVAMP** (Patel et al., 2021), **AQUA-RAT** (Ling et al., 2017), **SingleOp**, and **MultiArith** subsets from (Roy and Roth, 2016) . Additionally, we examine the performance of EchoPrompt on the high school mathematics subset of the **MMLU** dataset (Hendrycks et al., 2021a,b) and the **GSMIC-4k** dataset (Shi et al., 2023), which focuses explicitly on queries containing perturbations.

Logical Reasoning For logical reasoning, we assess the **Date Understanding**, **Shuffled Objects** (tracking three objects) tasks from bigBench (Ghazal et al., 2013), **LogiQA** (Liu et al., 2020) and generate 1000 random samples with two trials of flipping for the **Coin Flipping** task (Wei et al., 2023).

Reading Comprehension While we evaluate multiple numerical subsets of **DROP** (Dua et al., 2019), (including Football, Non-football, Census, and Break(Wolfson et al., 2020) from the **QDMR** dev subset) that could also be included in the arithmetic benchmarks, we group them with **SQuAD** (Rajpurkar et al., 2016) based on the query style. We evaluate EchoPrompt on **DROP** (Dua et al., 2019) and **SQuAD** (Rajpurkar et al., 2016) as two standard reading comprehension benchmarks.

Commonsense Reasoning For commonsense reasoning, we use **StrategyQA** (Geva et al., 2021), **Winogrande** (ai2, 2019) datasets to assess the performance of EchoPrompt on tasks that involve simpler queries but require factual knowledge.

B.2 Language models

For our experiments, we use code-davinci-002 (Chen et al., 2021) as the primary model for all tasks since this model is free to evaluate and has strong in-context learning ability. Additionally, we present the results on a subset of datasets on GPT-3.5-Turbo and CodeLlama-Instruct-34B (Roziere et al., 2023), models that are comparable in performance to Codex and are also instruction tuned. We also experiment with the smaller and publicly available models, such as StarCoder-15B (Li et al., 2023b), and GPT-J-6B (Wang and Komatsuzaki, 2021), specifically on synthetic and simpler tasks.

B.3 Prompts

Few-shot Exemplars For a fair comparison of methods, we use the same exemplars introduced in (Wei et al., 2023) for the GSM8K, SVAMP, SingleOp, MultiArith, Date Understanding, and Coin-Flipping tasks across all models. Additionally, we evaluate with the prompts suggested by

(Zhou et al., 2022a) for GSM8K, SVAMP, MultiArith, and DROP subsets. Furthermore, we provide a new set of prompts specifically for the DROP Census subset since no prior proposals exist.

Zero-shot-CoT Prompts As proposed in (Kojima et al., 2022), we employ the prompt *"Let's think step by step."* in stage 1. In stage 2, we extract the answer using different prompts depending on the type of task. For multiple-choice tasks, we utilize prompts like "From (a) through (e), the answer is." For other tasks, we use the phrase "Therefore, the answer is."

C Additional Related works

Interpretability, consistency and outcome correction Another related research direction involves exploring interpretability and consistency in the rationale generated by large-scale models. Recent works (Imani et al., 2023; Miao et al., 2023; Madaan and Yazdanbakhsh, 2022) help improve the interpretability of arithmetic and reasoning tasks through validation. Although these approaches are not directly tied to the EchoPrompt technique, they utilize chain-of-thought prompting, where we have shown that EchoPrompt exhibits promising results, particularly in zero-shot scenarios. In the domain of outcome correction, approaches such as (Jung et al., 2022; Wang et al., 2023b; Yao et al., 2023; Miao et al., 2021; Xie et al., 2023) leverage consistency among multiple generated rationales, while (Weng et al., 2023; Khalifa et al., 2023; Yang and Klein, 2021; Ni et al., 2023; Chen et al., 2022) prioritize the ranking of plausible generations to enhance performance across arithmetic, reasoning, and code-generation tasks. Building upon these foundations, self-correction methodologies like (Madaan et al., 2023; Jiang et al., 2023; Hao et al., 2023; Shinn et al., 2023), which employ feedback loops for refinement and multi-agent debating strategies (Du et al., 2023; Cohen et al., 2023; Fu et al., 2023) have evolved. EchoPrompt distinguishes itself from these approaches by focusing on single rationale generation rather than considering multiple generated responses.

D Additional details

Original Question:
If Pam is currently twice as young as Rena is, and in 10 years Rena will be 5 years older than her, how old is Pam now?
Compound Sentence Rephrase
Given that Pam is currently twice as young as Rena and that in 10 years Rena will be 5 years older than Pam, how old is Pam now?
Question First Rephrase
What is Pam's current age if Rena is twice as old as Pam and in 10 years Rena will be 5 years older than Pam?
Simple Sentence Rephrase
Currently, Pam is twice as young as Rena. In 10 years, Rena will be 5 years older than Pam. So, how old is Pam now?
Repeatition
If Pam is currently twice as young as Rena is, and in 10 years Rena will be 5 years older than her, how old is Pam now?

Figure 4: Example of rephrases used for the proposed rephrase structures in few-shot exemplars. The Rephrases of exemplars are generated using ChatGPT based on prompts in Table 8.

Rephrase	Prompt
Compound	Rephrase the following query using compound sentences without loss of details, starting with "Given that" and ending with the question in the query: **<Question>**
Question First	Rephrase the following query by asking the question in the query first, without loss of details: **<Question>**
Short and simple sentences	Rephrase the following query using short and simple sentences, without loss of details: **<Question>**

Table 8: Prompts used to create rephrases for exemplars, using ChatGPT

		GSM8k	SVAMP	DROP (Break)
Standard	-	17.0	67.6	56.1
	Repeat	16.7	72.0	**63.3**
	Compound	**19.2**	72.2	61.3
	Question First	18.4	71.1	56.9
	Simple	18.2	**72.4**	59.9
CoT	-	66.9	74.7	70.6
	Repeat	68.2	75.4	72.0
	Compound	**69.3**	**76.4**	**67.9**
	Question First	68.4	76.2	**72.2**
	Simple	68.2	75.3	71.6

Table 9: **code-davinci-002: Effect of exemplar selection**: While Table 3 utilized exemplars proposed in (Wei et al., 2023; Zhou et al., 2022a) that showcase simpler reasoning, this table employs exemplars demonstrating longer reasoning chains. The results indicate that although EchoPrompt provides higher gains with simpler exemplars; choosing better exemplars achieves higher overall accuracies, highlighting the significance of exemplar selection.

Model	Dataset	zero-shot				few-shot			
		Standard		CoT		Standard		CoT	
	EchoPrompt ?	✗	✓	✗	✓	✗	✓	✗	✓
Code-davinci-002	GSM8K	16.4	20.7(+4.3)	49.3	51.1(+1.8)	19.2	21.4(+2.2)	61.1	65.9(+4.8)
	SVAMP	66.8	74.7(+7.9)	66.5	73.9(+7.4)	69.8	75.8(+6.0)	75.2	79.0(+3.8)
	MultiArith	31.0	48.5(+17.5)	76.0	78.7(+2.7)	44.0	53.8(+9.8)	96.1	97.8(+1.7)
	SingleOp	91.6	91.8(+0.2)	82.9	92.4(+9.5)	93.2	94.2(+1.0)	92.8	94.7(+1.9)
	Shuffled Objects	36.4	35.2(−1.2)	42.4	58.2(+15.8)	34.8	36.7(+1.9)	66.0	68.9(+2.9)
	Coin Flip	47.7	47.2(−0.5)	58.5	60.1(+1.6)	99.6	100.0(+0.4)	100.0	100.0(+0.0)
	Date	44.2	43.8(−0.4)	39.0	46.8(+7.8)	49.3	50.4(+1.1)	65.6	68.1(+2.5)
	DROP(Football)	50.8	58.3(+7.5)	44.1	58.0(+13.9)	63.7	69.2(+5.5)	67.3	70.8(+3.5)
	DROP(Nonfootball)	43.2	57.1(+13.9)	39.7	52.6(+12.9)	57.1	63.3(+6.2)	69.2	72.2(+3.0)
	DROP(Census)	45.9	66.3(+20.4)	30.0	53.3(+23.3)	56.8	65.9(+9.1)	69.6	75.4(+5.8)
	DROP(Break)	43.7	55.8(+12.1)	38.2	51.2(+13.0)	55.5	63.1(+7.6)	65.3	69.6(+4.3)
	SQuAD(F1)	65.7	69.8(+4.1)	52.6	54.4(+1.8)	88.7	91.3(+2.6)	90.5	90.8(+0.3)
	AQUA-RAT	21.2	23.3(+2.1)	37.0	35.4(−1.6)	30.7	31.4(+0.7)	46.1	43.3(−2.8)
	MMLU-h	31.8	36.7(+4.9)	42.5	41.7(−0.8)	36.7	39.3(+2.6)	43.7	40.2(−3.5)
	logiQA	42.5	41.6(−0.9)	37.0	40.9(+3.9)	45.3	46.6(+1.3)	40.9	41.0(+0.1)
	Winogrande	-	-	-	-	71.9	70.8(−1.1)	68.5	67.4(−1.1)
	StrategyQA	-	-	-	-	74.8	74.5(−0.3)	79.4	79.3(−0.1)
GPT-3.5 (Turbo)	GSM8K	15.6	24.8(+9.2)	80.2	81.5(+1.3)	31.3	32.1(+0.8)	76.1	82.6(+6.5)
	SVAMP	65.5	77.5(+12.0)	80.5	83.0(+2.5)	76.1	78.4(+2.3)	80.8	83.3(+2.5)
	MultiArith	76.5	83.7(+7.2)	96.2	97.3(+1.1)	83.4	90.5(+7.1)	97.8	98.5(+0.7)
	SingleOp	92.6	96.8(+4.2)	94.3	96.8(+2.5)	93.9	96.2(+2.3)	95.7	96.5(+0.8)
	Shuffled Objects	26.9	21.6(−5.3)	79.5	82.2(+2.7)	30.6	34.6(+4.0)	68.8	74.3(+5.5)
	Coin Flip	76.7	86.8(+10.1)	99.8	98.6(−1.2)	90.0	95.6(+5.6)	100.0	100.0(+0.0)
	Date	50.1	48.6(−1.5)	53.1	55.5(+2.4)	50.4	49.3(−1.1)	64.5	66.2(+1.7)
	DROP(Break)	46.1	52.7(+6.6)	63.4	62.3(−1.1)	56.2	64.3(+8.1)	59.7	65.2(+5.5)
	SQuAD(F1)	79.1	80.6(+1.5)	62.1	59.4(−2.7)	76.4	83.2(+6.8)	85.3	86.1(+0.8)
	AQUA-RAT	31.5	28.4(−3.1)	55.8	58.9(+3.1)	33.4	35.8(+2.4)	42.9	57.1(+14.2)
	MMLU-h	35.5	35.9(+0.4)	55.9	57.7(+1.8)	34.1	34.8(+0.7)	37.0	50.8(+13.8)
	logiQA	36.2	38.2(+2.0)	37.6	39.0(+1.4)	45.1	43.3(−1.8)	32.5	32.3(−0.2)
CodeLlama-Instruct-(34B)	GSM8K	9.4	9.7(+0.3)	47.1	49.3(+2.2)	8.6	9.7(+1.1)	40.8	45.4(+4.6)
	SVAMP	60.6	64.5(+3.9)	65.2	69.3(+4.1)	56.3	62.9(+6.6)	68.8	69.3(+0.5)
	MultiArith	23.8	22.9(−0.7)	85.3	86.1(+0.8)	11.3	16.5(+5.2)	86.3	90.3(+4.0)
	Singleop	90.9	93.6(+2.7)	88.4	90.8(+2.4)	88.4	90.9(+2.5)	90.0	92.0(+2.0)
	Coin Flip	51.7	72.5(+20.8)	82.0	79.2(−2.8)	94.8	96.2(+1.4)	100.0	100.0(+0.0)
	Date	36.3	39.2(+2.9)	43.8	43.1(−0.7)	40.1	40.9(+0.8)	56.6	57.4(+0.8)
	DROP(Break)	40.5	48.3(+7.8)	53.3	52.1(−1.2)	37.4	43.3(+5.9)	53.7	57.6(+3.9)
Starcoder (15B)	SingleOp	63.1	66.9(+3.8)	53.5	66.5(+13.0)	64.0	70.1(+6.1)	68.8	73.6(+4.8)
	SVAMP	35.6	37.9(+2.3)	30.9	36.7(+5.8)	32.4	37.2(+4.8)	30.2	36.2(+6.0)
	Coin Flip	55.4	54.3(−1.1)	51.6	51.0(−0.6)	98.6	99.8(+1.2)	100.0	100.0(+0.0)
	Date	15.9	19.2(+3.3)	20.6	19.9(−0.7)	24.4	26.6(+2.2)	38.4	33.8(−4.6)
GPT-J (6B)	SingleOp	-	-	-	-	37.2	39.9(+2.7)	45.3	44.5(−0.8)
	SVAMP	-	-	-	-	8.9	10.1(+1.2)	21.1	19.8(−1.3)
	Coin Flip	-	-	-	-	81.3	81.3(+0.0)	80.6	96.4(+15.8)
	Date	-	-	-	-	13.2	13.6(+0.4)	11.1	15.8(+4.7)

Table 10: Performance summary of EchoPrompt on all models. EchoPrompt consistently improves performance across different prompting strategies, showing significant improvements in zero-shot prompting scenarios. It outperforms the prior state-of-the-art in numerical reasoning and reading comprehension tasks. However, we do not see consistent improvements in multiple-choice tasks.

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Rewriting in simple words, the question is: "Given that there are currently 15 trees in the grove, and after the grove workers plant some more trees, there will be 21 trees, how many trees did the grove workers plant today?"
Now, to answer the rewritten question, There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Rewriting in simple words, the question is: "Given that there are currently 3 cars in the parking lot, and 2 more cars arrive, how many cars are in the parking lot now?"
Now, to answer the rewritten question, There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. The answer is 5.
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Rewriting in simple words, the question is: "Given that Leah had 32 chocolates and her sister had 42, and they ate a total of 35 pieces, how many pieces of chocolate do they have left in total?"
Now, to answer the rewritten question, Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. The answer is 39.
Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Rewriting in simple words, the question is: "Given that Jason had 20 lollipops, and after giving some to Denny, he ended up with 12 lollipops, how many lollipops did Jason give to Denny?"
Now, to answer the rewritten question, Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. The answer is 8.
Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Rewriting in simple words, the question is: "Given that Shawn had five toys and received two toys each from both his mom and dad for Christmas, how many toys does Shawn have now?"
Now, to answer the rewritten question, Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9. The answer is 9.
Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: Rewriting in simple words, the question is: "Given that there were initially 9 computers in the server room, and 5 more computers were installed each day from Monday to Thursday, how many computers are now in the server room?"
Now, to answer the rewritten question, There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 is 29. The answer is 29.
Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Rewriting in simple words, the question is: "Given that Michael had 58 golf balls initially, he lost 23 golf balls on Tuesday, and then 2 more on Wednesday, how many golf balls did he have at the end of Wednesday?"
Now, to answer the rewritten question, Michael started with 58 golf balls. After losing 23 on tuesday, he had 58 - 23 = 35. After losing 2 more, he had 35 - 2 = 33 golf balls. The answer is 33.
Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Rewriting in simple words, the question is: "Given that Olivia has $23 and bought five bagels for $3 each, how much money does she have left?"
Now, to answer the rewritten question, Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 x 3 = 15 dollars. So she has 23 - 15 dollars left. 23 - 15 is 8. The answer is 8.

Table 11: **GSM8K, SVAMP, MultiArith, SingleOp** Few-shot exemplars used for EchoPrompt with compound sentence rephrasing

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Rewriting in simple words, the question is: "How many trees did the grove workers plant today if there are 15 trees in the grove and after they are done, there will be 21 trees?"
Now, to answer the rewritten question, There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Rewriting in simple words, the question is: "How many cars are in the parking lot if there are 3 cars in the parking lot and 2 more cars arrive?"
Now, to answer the rewritten question, There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. The answer is 5.
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Rewriting in simple words, the question is: "How many chocolates do Leah and her sister have left if they had 32 and 42 chocolates respectively, and they ate 35 chocolates in total?"
Now, to answer the rewritten question, Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. The answer is 39.
Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Rewriting in simple words, the question is: "How many lollipops did Jason give to Denny if Jason had 20 lollipops and now has 12 lollipops?"
Now, to answer the rewritten question, Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. The answer is 8.
Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Rewriting in simple words, the question is: "How many toys does Shawn have now if he had five toys and for Christmas he got two toys each from his mom and dad?"
Now, to answer the rewritten question, Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9. The answer is 9.
Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: Rewriting in simple words, the question is: "How many computers are now in the server room if there were nine computers in the server room and five more computers were installed each day from Monday to Thursday?"
Now, to answer the rewritten question, There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 is 29. The answer is 29.
Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Rewriting in simple words, the question is: "How many golf balls did Michael have at the end of wednesday if he had 58 golf balls and lost 23 golf balls on tuesday and 2 more on wednesday?"
Now, to answer the rewritten question, Michael started with 58 golf balls. After losing 23 on tuesday, he had 58 - 23 = 35. After losing 2 more, he had 35 - 2 = 33 golf balls. The answer is 33.
Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Rewriting in simple words, the question is: "How much money does Olivia have left if she has $23 and buys five bagels for $3 each?"
Now, to answer the rewritten question, Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 x 3 = 15 dollars. So she has 23 - 15 dollars left. 23 - 15 is 8. The answer is 8.

Table 12: **GSM8K, SVAMP, MultiArith, SingleOp** Few-shot exemplars used for EchoPrompt with question first rephrasing

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: Rewriting in simple words, the question is: "There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?"
Now, to answer the rewritten question, There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: Rewriting in simple words, the question is: "There are three cars in the parking lot. Another two cars arrive. How many cars are in the parking lot now?"
Now, to answer the rewritten question, There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. The answer is 5.
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
A: Rewriting in simple words, the question is: "Leah and her sister had a total of 32 and 42 chocolates. They ate 35 chocolates. How many chocolates do they have left in total?"
Now, to answer the rewritten question, Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. The answer is 39.
Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Rewriting in simple words, the question is: "Jason had 20 lollipops. He gave some lollipops to Denny. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?"
Now, to answer the rewritten question, Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. The answer is 8.
Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: Rewriting in simple words, the question is: "Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?"
Now, to answer the rewritten question, Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9. The answer is 9.
Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?
A: Rewriting in simple words, the question is: "There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?"
Now, to answer the rewritten question, There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 is 29. The answer is 29.
Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?
A: Rewriting in simple words, the question is: "Michael started with 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls does he have at the end of wednesday?"
Now, to answer the rewritten question, Michael started with 58 golf balls. After losing 23 on tuesday, he had 58 - 23 = 35. After losing 2 more, he had 35 - 2 = 33 golf balls. The answer is 33.
Q: Olivia has $23. She bought five bagels for 3$ each. How much money does she have left?
A: Rewriting in simple words, the question is: "Olivia has $23. She bought 5 bagels for 3$ each. How much money does she have left?"
Now, to answer the rewritten question, Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 x 3 = 15 dollars. So she has 23 - 15 dollars left. 23 - 15 is 8. The answer is 8.

Table 13: **GSM8K, SVAMP, MultiArith, SingleOp** Few-shot exemplars used for EchoPrompt with simple sentence rephrasing

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?
Answer Choices: (a) 50 (b) 45 (c) 65 (d) 78 (e) 64
A: Rewriting in simple words, the question is: "Given that the average of 15 numbers is 40, and 10 is added to each number, what is the mean of the numbers?"
Now, to answer the rewritten question, If 10 is added to each number, then the mean of the numbers also increases by 10. So the new mean would be 50. The answer is (a).
Q: If a / b = 3/4 and 8a + 5b = 22,then find the value of a.
Answer Choices: (a) 1/2 (b) 3/2 (c) 5/2 (d) 4/2 (e) 7/2
A: Rewriting in simple words, the question is: "Given that a / b = 3/4 and 8a + 5b = 22, what is the value of a?"
Now, to answer the rewritten question, If a / b = 3/4, then b = 4a / 3. So 8a + 5(4a / 3) = 22. This simplifies to 8a + 20a / 3 = 22, which means 44a / 3 = 22. So a is equal to 3/2. The answer is (b).
Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km
A: Rewriting in simple words, the question is: "Given that a person is traveling at 20 km/hr and reached his destiny in 2.5 hr, what is the distance?"
Now, to answer the rewritten question, The distance that the person traveled would have been 20 km/hr * 2.5 hrs = 50 km. The answer is (e).
Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788
A: Rewriting in simple words, the question is: "Given that the numbers from 1 to 500 need to be typed, how many keystrokes are needed?"
Now, to answer the rewritten question, There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (b).

Table 14: **AQuA, MMLU-h** Few-shot exemplars used for EchoPrompt with compound sentence rephrasing

Q: Four years ago, Kody was only half as old as Mohamed. If Mohamed is currently twice 30 years old, how old is Kody?
A: Rewriting in simple words, the question is: "Given that Mohamed is twice 30 years old currently, and four years ago, Kody was only half Mohamed's age, what is the current age of Kody?"
Now, to answer the rewritten question, We were told that Mohamed is currently twice 30 years old, so he is currently 30 * 2 = 60 years old. That means that four years ago he must have been 60 - 4 = 56 years old. Four years ago, Kody was half as old as Mohamed, so Kody must have been 56 / 2 = 28 years old then. Since Kody was 28 years old four years ago, she must now be 28 + 4 = 32 years old. The answer is 32.
Q: Carla bought 2 bags of mini peanut butter cups on clearance. Each bag was $6.00 but was 75% off. How much did she spend on 2 bags of candy?
A: Rewriting in simple words, the question is: "Given that Carla bought two bags of mini peanut butter cups on clearance for $6.00 each and each bag was discounted by 75%, what was the total amount she spent on both bags of candy?"
Now, to answer the rewritten question, Each bag was $6.00 but was 75% off. So each bag cost $6.00 * (1 - 0.75) = $6.00 * 0.25 = $1.50. Carla bought 2 bags. So she spent $1.50 * 2 = $3.00. The answer is 3.
Q: If Pam is currently twice as young as Rena is, and in 10 years Rena will be 5 years older than her, how old is Pam now?
A: Rewriting in simple words, the question is: "Given that Pam is currently twice as young as Rena and that in 10 years Rena will be 5 years older than Pam, how old is Pam now?"
Now, to answer the rewritten question, Since Rena will be 5 years older than Pam in 10 years, she must be 5 years older than Pam now as well. If Pam is currently twice as young as Rena, that means that Rena is currently twice as old as Pam is. So if P stands for Pam's age now and R stands for Rena's age now, then we know that R = 2 * P And since Rena is 5 years older than Pam now, we know that R = P + 5. By substitution, we have P + 5 = 2 * P, which means that P = 5. The answer is 5.
Q: Cappuccinos cost $2, iced teas cost $3, cafe lattes cost $1.5 and espressos cost $1 each. Sandy orders some drinks for herself and some friends. She orders three cappuccinos, two iced teas, two cafe lattes, and two espressos. How much change does she receive back for a twenty-dollar bill?
A: Rewriting in simple words, the question is: "Given that Sandy ordered three cappuccinos at $2 each, two iced teas at $3 each, two cafe lattes at $1.5 each, and two espressos at $1 each, and paid with a twenty-dollar bill, what amount of change will she receive?"
Now, to answer the rewritten question, Sandy ordered three cappuccinos, which cost $2 each, so she spent $2 * 3 = $6 on cappuccinos. She ordered two iced teas, which cost $3 each, so she spent $3 * 2 = $6 dollars on ice teas. She ordered two cafe lattes, which cost $1.5 each, so she spent $1.5 * 2 = $3 on cafe lattes. She ordered two espressos, which cost $1 each, so she spent $1 * 2 = $2 on espressos. So altogether, Sandy spent $6 + $6 + $3 + $2 = $17 on drinks, which means that sandy will get $20 - $17 = $3 as change. The answer is 3.

Table 15: **GSM8K, SVAMP, MultiArith, SingleOp** Few-shot exemplars demonstrating long reasoning chains, used for EchoPrompt with compound sentence rephrasing

Q: Four years ago, Kody was only half as old as Mohamed. If Mohamed is currently twice 30 years old, how old is Kody?
A: Rewriting in simple words, the question is: "What is Kody's current age if four years ago he was half as old as Mohamed, who is currently twice 30 years old?"
Now, to answer the rewritten question, We were told that Mohamed is currently twice 30 years old, so he is currently 30 * 2 = 60 years old. That means that four years ago he must have been 60 - 4 = 56 years old. Four years ago, Kody was half as old as Mohamed, so Kody must have been 56 / 2 = 28 years old then. Since Kody was 28 years old four years ago, she must now be 28 + 4 = 32 years old. The answer is 32.
Q: Carla bought 2 bags of mini peanut butter cups on clearance. Each bag was $6.00 but was 75% off. How much did she spend on 2 bags of candy?
A: Rewriting in simple words, the question is: "How much did Carla spend on two bags of mini peanut butter cups that were on clearance for 75% off, if each bag cost $6.00?"
Now, to answer the rewritten question, Each bag was $6.00 but was 75% off. So each bag cost $6.00 * (1 - 0.75) = $6.00 * 0.25 = $1.50. Carla bought 2 bags. So she spent $1.50 * 2 = $3.00. The answer is 3.
Q: If Pam is currently twice as young as Rena is, and in 10 years Rena will be 5 years older than her, how old is Pam now?
A: Rewriting in simple words, the question is: "What is Pam's current age if Rena is twice as old as Pam and in 10 years Rena will be 5 years older than Pam?"
Now, to answer the rewritten question, Since Rena will be 5 years older than Pam in 10 years, she must be 5 years older than Pam now as well. If Pam is currently twice as young as Rena, that means that Rena is currently twice as old as Pam is. So if P stands for Pam's age now and R stands for Rena's age now, then we know that R = 2 * P And since Rena is 5 years older than Pam now, we know that R = P + 5. By substitution, we have P + 5 = 2 * P, which means that P = 5. The answer is 5.
Q: Cappuccinos cost $2, iced teas cost $3, cafe lattes cost $1.5 and espressos cost $1 each. Sandy orders some drinks for herself and some friends. She orders three cappuccinos, two iced teas, two cafe lattes, and two espressos. How much change does she receive back for a twenty-dollar bill?
A: Rewriting in simple words, the question is: "How much change will Sandy receive back for a twenty-dollar bill if she orders three cappuccinos, two iced teas, two cafe lattes, and two espressos, and the cappuccinos cost $2, iced teas cost $3, cafe lattes cost $1.5, and espressos cost $1 each?"
Now, to answer the rewritten question, Sandy ordered three cappuccinos, which cost $2 each, so she spent $2 * 3 = $6 on cappuccinos. She ordered two iced teas, which cost $3 each, so she spent $3 * 2 = $6 dollars on ice teas. She ordered two cafe lattes, which cost $1.5 each, so she spent $1.5 * 2 = $3 on cafe lattes. She ordered two espressos, which cost $1 each, so she spent $1 * 2 = $2 on espressos. So altogether, Sandy spent $6 + $6 + $3 + $2 = $17 on drinks, which means that sandy will get $20 - $17 = $3 as change. The answer is 3.

Table 16: **GSM8K, SVAMP, MultiArith, SingleOp** Few-shot exemplars demonstrating long reasoning chains, used for EchoPrompt with question first rephrasing

Q: Four years ago, Kody was only half as old as Mohamed. If Mohamed is currently twice 30 years old, how old is Kody?
A: Rewriting in simple words, the question is: "Four years ago, Kody's age was half of Mohamed's age. Currently, Mohamed's age is twice 30 years old. How old is Kody now?"
Now, to answer the rewritten question, We were told that Mohamed is currently twice 30 years old, so he is currently 30 * 2 = 60 years old. That means that four years ago he must have been 60 - 4 = 56 years old. Four years ago, Kody was half as old as Mohamed, so Kody must have been 56 / 2 = 28 years old then. Since Kody was 28 years old four years ago, she must now be 28 + 4 = 32 years old. The answer is 32.
Q: Carla bought 2 bags of mini peanut butter cups on clearance. Each bag was $6.00 but was 75% off. How much did she spend on 2 bags of candy?
A: Rewriting in simple words, the question is: "Carla noticed a sale on mini peanut butter cups. She decided to buy two bags of the candy. The candy was discounted at 75% off the original cost of $6.00 per bag. How much did Carla spend on the two bags of candy?"
Now, to answer the rewritten question, Each bag was $6.00 but was 75% off. So each bag cost $6.00 * (1 - 0.75) = $6.00 * 0.25 = $1.50. Carla bought 2 bags. So she spent $1.50 * 2 = $3.00. The answer is 3.
Q: If Pam is currently twice as young as Rena is, and in 10 years Rena will be 5 years older than her, how old is Pam now?
A: Rewriting in simple words, the question is: "Currently, Pam is twice as young as Rena. In 10 years, Rena will be 5 years older than Pam. So, how old is Pam now?"
Now, to answer the rewritten question, Since Rena will be 5 years older than Pam in 10 years, she must be 5 years older than Pam now as well. If Pam is currently twice as young as Rena, that means that Rena is currently twice as old as Pam is. So if P stands for Pam's age now and R stands for Rena's age now, then we know that R = 2 * P And since Rena is 5 years older than Pam now, we know that R = P + 5. By substitution, we have P + 5 = 2 * P, which means that P = 5. The answer is 5.
Q: Cappuccinos cost $2, iced teas cost $3, cafe lattes cost $1.5 and espressos cost $1 each. Sandy orders some drinks for herself and some friends. She orders three cappuccinos, two iced teas, two cafe lattes, and two espressos. How much change does she receive back for a twenty-dollar bill?
A: Rewriting in simple words, the question is: "Sandy got three cappuccinos, two iced teas, two cafe lattes, and two espressos. Cappuccinos cost $2, iced teas cost $3, cafe lattes cost $1.5 and espressos cost $1 each. She paid with a twenty-dollar bill and received some change back. How much change did Sandy receive?"
Now, to answer the rewritten question, Sandy ordered three cappuccinos, which cost $2 each, so she spent $2 * 3 = $6 on cappuccinos. She ordered two iced teas, which cost $3 each, so she spent $3 * 2 = $6 dollars on ice teas. She ordered two cafe lattes, which cost $1.5 each, so she spent $1.5 * 2 = $3 on cafe lattes. She ordered two espressos, which cost $1 each, so she spent $1 * 2 = $2 on espressos. So altogether, Sandy spent $6 + $6 + $3 + $2 = $17 on drinks, which means that sandy will get $20 - $17 = $3 as change. The answer is 3.

Table 17: **GSM8K, SVAMP, MultiArith, SingleOp** Few-shot exemplars demonstrating long reasoning chains, used for EchoPrompt with simple sentence rephrasing

Q: The Seahawks played the San Francisco 49ers. In the first quarter, the Hawks RB Julius Jones got a 27-yard TD run, along with DT Craig Terrill returning a fumble 9 yards for a touchdown. In the third quarter, the 49ers almost rallied as RB H. J. Torres made a 12-yard TD pass to Lucas Nelly, along with Mare kicking a 32-yard field goal. In the final quarter, Julius Jones got another 11-yard TD. How many yards do the shortest touchdown run and the longest touchdown pass combine for?
A: Rewriting in simple words, the question is: "Given that during the game between the Seattle Seahawks and San Francisco 49ers, Julius Jones, the Seahawks' RB, scored a 27-yard TD run in the first quarter, and also got an 11-yard TD run in the final quarter, while DT Craig Terrill also contributed to the score by returning a fumble for a 9-yard touchdown in the first quarter, and in the third quarter, the 49ers rallied as RB H.J. Torres threw a 12-yard TD pass to Lucas Nelly, and Mare kicked a 32-yard field goal, what is the sum of the shortest touchdown run and the longest touchdown pass in terms of yards?"
Now, to answer the rewritten question, All the touchdown runs are: a 27-yard touchdown run, a 9-yard touchdown run, a 11-yard touchdown run. The smallest number among 27, 9, 11 is 9. So the shortest touchdown run was 9 yards. All the touchdown passes are: a 12-yard touchdown pass. So the longest touchdown pass was 12 yards. So the shortest touchdown run and the longest touchdown pass combine for 9 + 12 = 21 yards. So the answer is 21.
Q: The Steelers went home for a duel with the Baltimore Ravens. Pittsburgh would deliver the opening punch in the first quarter with a 1-yard touchdown from running back Rashard Mendenhall. The Ravens would make it even as running back Willis McGahee got a 9-yard TD. The Ravens kicker Billy Cundiff got a 45-yard field goal in the second quarter, concluding the first half with a 10-7 lead. The Steelers brought the game into overtime with a 38-yard field goal by Andrew Foster. The Ravens Billy Cundiff pulled off a winning 33-yard field goal in overtime. How many points did the Ravens have at halftime?
A: Rewriting in simple words, the question is: "Given that during the game between the Steelers and Ravens, Rashard Mendenhall scored a 1-yard TD for the Steelers, while Willis McGahee scored a 9-yard TD for the Ravens to tie the game, and in the second quarter, the Ravens' Billy Cundiff kicked a 45-yard field goal, the sum of which gave the Ravens a 10-7 lead at halftime, and the game was later tied in regulation time with a 38-yard field goal by Andrew Foster for the Steelers, follwed by the Ravens ultimately winning on a 33-yard field goal by Cundiff in overtime. What was the Ravens' halftime score?"
Now, to answer the rewritten question, The Ravens kicker Billy Cundiff got a 45-yard field goal in the second quarter, concluding the first half with a 10-7 lead. So the Ravens had 10 points at halftime. So the answer is 10.
Q: The Vikings flew to Bank of America Stadium to face the Carolina Panthers. After a scoreless first quarter, Carolina got on the board with quarterback Matt Moore finding fullback Brad Hoover on a 1-yard TD pass. After yet another scoreless quarter, Carolina sealed the game as Matt Moore completed a 42-yard touchdown pass to wide receiver Steve Smith. How many scoreless quarters were there?
A: Rewriting in simple words, the question is: "Given that during the game between the Vikings and Panthers at Bank of America Stadium, the first quarter was scoreless, followed by Carolina scoring a 1-yard TD pass from Matt Moore to Brad Hoover in the second quarter, and another scoreless quarter followed, and in the fourth quarter, Moore completed a 42-yard touchdown pass to Steve Smith, ultimately securing the win for Carolina, then how many scoreless quarters were there in the game?"
Now, to answer the rewritten question, The first and third quarters were the scoreless quarters. So there are 2 scoreless quarters. So the answer is 2.

Table 18: **DROP-Football** Few-shot exemplars used for EchoPrompt with compound sentence rephrasing

Q: The Seahawks played the San Francisco 49ers. In the first quarter, the Hawks RB Julius Jones got a 27-yard TD run, along with DT Craig Terrill returning a fumble 9 yards for a touchdown. In the third quarter, the 49ers almost rallied as RB H. J. Torres made a 12-yard TD pass to Lucas Nelly, along with Mare kicking a 32-yard field goal. In the final quarter, Julius Jones got another 11-yard TD. How many yards do the shortest touchdown run and the longest touchdown pass combine for?
A: Rewriting in simple words, the question is: "What is the combined total of the shortest touchdown run and the longest touchdown pass in the Seahawks vs. 49ers game, where Julius Jones got a 27-yard TD run, DT Craig Terrill returned a fumble 9 yards for a touchdown, RB H. J. Torres made a 12-yard TD pass to Lucas Nelly, and Mare kicked a 32-yard field goal, and Julius Jones got another 11-yard TD?"
Now, to answer the rewritten question, All the touchdown runs are: a 27-yard touchdown run, a 9-yard touchdown run, a 11-yard touchdown run. The smallest number among 27, 9, 11 is 9. So the shortest touchdown run was 9 yards. All the touchdown passes are: a 12-yard touchdown pass. So the longest touchdown pass was 12 yards. So the shortest touchdown run and the longest touchdown pass combine for 9 + 12 = 21 yards. So the answer is 21.
Q: The Steelers went home for a duel with the Baltimore Ravens. Pittsburgh would deliver the opening punch in the first quarter with a 1-yard touchdown from running back Rashard Mendenhall. The Ravens would make it even as running back Willis McGahee got a 9-yard TD. The Ravens kicker Billy Cundiff got a 45-yard field goal in the second quarter, concluding the first half with a 10-7 lead. The Steelers brought the game into overtime with a 38-yard field goal by Andrew Foster. The Ravens Billy Cundiff pulled off a winning 33-yard field goal in overtime. How many points did the Ravens have at halftime?
A: Rewriting in simple words, the question is: "What was the score of the Ravens at halftime in the football game against the Pittsburgh Steelers, where the Steelers scored a 1-yard touchdown in the first quarter through Rashard Mendenhall, the Ravens tied it with a 9-yard TD run by Willis McGahee, and Billy Cundiff of the Ravens made a 45-yard field goal in the second quarter to give the Ravens a 10-7 lead going into halftime, before the Steelers tied it with a 38-yard field goal by Andrew Foster and eventually lost in overtime to the Ravens' 33-yard field goal by Billy Cundiff?"
Now, to answer the rewritten question, The Ravens kicker Billy Cundiff got a 45-yard field goal in the second quarter, concluding the first half with a 10-7 lead. So the Ravens had 10 points at halftime. So the answer is 10.
Q: The Vikings flew to Bank of America Stadium to face the Carolina Panthers. After a scoreless first quarter, Carolina got on the board with quarterback Matt Moore finding fullback Brad Hoover on a 1-yard TD pass. After yet another scoreless quarter, Carolina sealed the game as Matt Moore completed a 42-yard touchdown pass to wide receiver Steve Smith. How many scoreless quarters were there?
A: Rewriting in simple words, the question is: "How many scoreless quarters were there in the football game between the Minnesota Vikings and the Carolina Panthers, where Carolina scored a 1-yard TD pass from quarterback Matt Moore to fullback Brad Hoover in the second quarter and sealed the game with a 42-yard touchdown pass to wide receiver Steve Smith in the fourth quarter?"
Now, to answer the rewritten question, The first and third quarters were the scoreless quarters. So there are 2 scoreless quarters. So the answer is 2.

Table 19: **DROP-Football** Few-shot exemplars used for EchoPrompt with question first rephrasing

Q: The Seahawks played the San Francisco 49ers. In the first quarter, the Hawks RB Julius Jones got a 27-yard TD run, along with DT Craig Terrill returning a fumble 9 yards for a touchdown. In the third quarter, the 49ers almost rallied as RB H. J. Torres made a 12-yard TD pass to Lucas Nelly, along with Mare kicking a 32-yard field goal. In the final quarter, Julius Jones got another 11-yard TD. How many yards do the shortest touchdown run and the longest touchdown pass combine for?
A: Rewriting in simple words, the question is: "During a game between the Seattle Seahawks and San Francisco 49ers, Julius Jones, the Seahawks' RB, scored a 27-yard TD run in the first quarter, and also got an 11-yard TD run in the final quarter. DT Craig Terrill also contributed to the score by returning a fumble for a 9-yard touchdown in the first quarter. In the third quarter, the 49ers rallied as RB H.J. Torres threw a 12-yard TD pass to Lucas Nelly, and Mare kicked a 32-yard field goal. What is the sum of the shortest touchdown run and the longest touchdown pass in terms of yards?"
Now, to answer the rewritten question, All the touchdown runs are: a 27-yard touchdown run, a 9-yard touchdown run, a 11-yard touchdown run. The smallest number among 27, 9, 11 is 9. So the shortest touchdown run was 9 yards. All the touchdown passes are: a 12-yard touchdown pass. So the longest touchdown pass was 12 yards. So the shortest touchdown run and the longest touchdown pass combine for 9 + 12 = 21 yards. So the answer is 21.
Q: The Steelers went home for a duel with the Baltimore Ravens. Pittsburgh would deliver the opening punch in the first quarter with a 1-yard touchdown from running back Rashard Mendenhall. The Ravens would make it even as running back Willis McGahee got a 9-yard TD. The Ravens kicker Billy Cundiff got a 45-yard field goal in the second quarter, concluding the first half with a 10-7 lead. The Steelers brought the game into overtime with a 38-yard field goal by Andrew Foster. The Ravens Billy Cundiff pulled off a winning 33-yard field goal in overtime. How many points did the Ravens have at halftime?
A: Rewriting in simple words, the question is: "During a game between the Steelers and Ravens, Rashard Mendenhall scored a 1-yard TD for the Steelers, while Willis McGahee scored a 9-yard TD for the Ravens to tie the game. In the second quarter, the Ravens' Billy Cundiff kicked a 45-yard field goal, giving the Ravens a 10-7 lead at halftime. The game was tied in regulation time with a 38-yard field goal by Andrew Foster for the Steelers. In overtime, the Ravens ultimately won on a 33-yard field goal by Cundiff. What was the Ravens' halftime score?"
Now, to answer the rewritten question, The Ravens kicker Billy Cundiff got a 45-yard field goal in the second quarter, concluding the first half with a 10-7 lead. So the Ravens had 10 points at halftime. So the answer is 10.
Q: The Vikings flew to Bank of America Stadium to face the Carolina Panthers. After a scoreless first quarter, Carolina got on the board with quarterback Matt Moore finding fullback Brad Hoover on a 1-yard TD pass. After yet another scoreless quarter, Carolina sealed the game as Matt Moore completed a 42-yard touchdown pass to wide receiver Steve Smith. How many scoreless quarters were there?
A: Rewriting in simple words, the question is: "During a game between the Vikings and Panthers at Bank of America Stadium, the first quarter was scoreless. In the second quarter, Carolina scored a 1-yard TD pass from Matt Moore to Brad Hoover. Another scoreless quarter followed, and in the fourth quarter, Moore completed a 42-yard touchdown pass to Steve Smith, securing the win for Carolina. How many scoreless quarters were there in the game?"
Now, to answer the rewritten question, The first and third quarters were the scoreless quarters. So there are 2 scoreless quarters. So the answer is 2.

Table 20: **DROP-Football** Few-shot exemplars used for EchoPrompt with simple sentence rephrasing

Q: For the period 2011-15, the estimated median annual income for a household in the town was $48,059, and the median income for a family was $96,005. Male full-time workers had a median income of $64,750 versus $39,278 for females. The per capita income for the town was $18,905. About 8.7% of families and 34.7% of the population were below the poverty line, including 18.2% of those under age 18 and 4.5% of those age 65 or over. How many percent of the population is not living below the poverty line?
A: Rewriting in simple words, the question is: "Given that for the period 2011-15, a household in the town had an estimated median annual income of $48,059, and a family had a median income of $96,005, while male full-time workers had a median income of $64,750 versus $39,278 for females, and the per capita income for the town was $18,905, with about 8.7% of families and 34.7% of the population living below the poverty line, including 18.2% of those under age 18 and 4.5% of those age 65 or over, what percent of the population is not living below the poverty line?"
Now, to answer the rewritten question, The population is 34.7% below the poverty line. So the population is 100 - 34.7 = 65.3% not below the poverty line. The answer is 65.3.
Q: As of the census of 2010, there were 2,213 people, 870 households, and 563 families residing in the town. There were 902 housing units. The racial makeup of the town was 82.0% White (U.S. Census), 6.1% African American (U.S. Census), 0.1% Native American (U.S. Census), 5.7% Asian (U.S. Census), 0.1% Race (U.S. Census), 2.8% from Race (U.S. Census), and 3.2% from two or more races. Hispanic (U.S. Census) or Latino (U.S. Census) of any race were 6.4% of the population. How many more people are there than housing units?
A: Rewriting in simple words, the question is: "Given that according to the 2010 census, there were 2,213 individuals residing in the town and living in 870 households, which consisted of 563 families, while there were 902 housing units in total, and the town's ethnic composition was 82.0% White, 6.1% African American, 0.1% Native American, 5.7% Asian, 0.1% from another race, 2.8% from two or more races, and 3.2% identifying as Hispanic or Latino, regardless of their race, how many individuals exceed the number of housing units in the town?"
Now, to answer the rewritten question, There are 2,213 and 902 housing units. So the difference is 2,213 - 902 = 1,311. The answer is 1,311.
Q: In the city, the population was spread out with 28.1% under the age of 18, 9.2% from 18 to 24, 26.2% from 25 to 44, 21.4% from 45 to 64, and 15.1% who were 65 years of age or older. The median age was 36 years. For every 100 females, there were 83.5 males. For every 100 females age 18 and over, there were 77.7 males. For every 100 females age 18 and over, how many fewer males were there?
A: Rewriting in simple words, the question is: "Given that in the city, 28.1% of the population was under the age of 18, 9.2% were aged 18 to 24, 26.2% were aged 25 to 44, 21.4% were aged 45 to 64, and 15.1% were 65 years of age or older, with a median age of 36 years, and for every 100 females, there were 83.5 males, how many fewer males were there for every 100 females age 18 and over, considering that there were 77.7 males for every 100 females age 18 and over?"
Now, to answer the rewritten question, For every 100 females age 18 and over, there were 77.7 males. So for every 100 females, there were 77.7 males. There were 83.5 males. So 100 - 77.7 = 22.3 males. The answer is 22.3.

Table 21: **DROP Census** Few-shot exemplars used for EchoPrompt with compound sentence rephrasing

Q: For the period 2011-15, the estimated median annual income for a household in the town was $48,059, and the median income for a family was $96,005. Male full-time workers had a median income of $64,750 versus $39,278 for females. The per capita income for the town was $18,905. About 8.7% of families and 34.7% of the population were below the poverty line, including 18.2% of those under age 18 and 4.5% of those age 65 or over. How many percent of the population is not living below the poverty line?
A: Rewriting in simple words, the question is: "What percentage of the population is not living below the poverty line if the estimated median annual income for a household in the town was $48,059, and the median income for a family was $96,005. Male full-time workers had a median income of $64,750 versus $39,278 for females. The per capita income for the town was $18,905. About 8.7% of families and 34.7% of the population were below the poverty line, including 18.2% of those under age 18 and 4.5% of those age 65 or over?"
Now, to answer the rewritten question, The population is 34.7% below the poverty line. So the population is 100 - 34.7 = 65.3% not below the poverty line. The answer is 65.3.
Q: As of the census of 2010, there were 2,213 people, 870 households, and 563 families residing in the town. There were 902 housing units. The racial makeup of the town was 82.0% White (U.S. Census), 6.1% African American (U.S. Census), 0.1% Native American (U.S. Census), 5.7% Asian (U.S. Census), 0.1% Race (U.S. Census), 2.8% from Race (U.S. Census), and 3.2% from two or more races. Hispanic (U.S. Census) or Latino (U.S. Census) of any race were 6.4% of the population. How many more people are there than housing units?
A: Rewriting in simple words, the question is: "How many people surpass the number of housing units in the town if as of the census conducted in 2010, the town accommodated 2,213 individuals, who dwelled in 870 households, comprising 563 families. Meanwhile, the housing units were measured at 902. For the town's ethnic composition, 82.0% identified as White, 6.1% as African American, 0.1% as Native American, 5.7% as Asian, 0.1% as belonging to another race, 2.8% were from two or more races, and 3.2% identified as Hispanic or Latino, regardless of their race?"
Now, to answer the rewritten question, There are 2,213 and 902 housing units. So the difference is 2,213 - 902 = 1,311. The answer is 1,311.
Q: In the city, the population was spread out with 28.1% under the age of 18, 9.2% from 18 to 24, 26.2% from 25 to 44, 21.4% from 45 to 64, and 15.1% who were 65 years of age or older. The median age was 36 years. For every 100 females, there were 83.5 males. For every 100 females age 18 and over, there were 77.7 males. For every 100 females age 18 and over, how many fewer males were there?
A: Rewriting in simple words, the question is: "How many fewer males were there in the city for every 100 females age 18 and over, if the population was spread out with 28.1% under the age of 18, 9.2% from 18 to 24, 26.2% from 25 to 44, 21.4% from 45 to 64, and 15.1% who were 65 years of age or older, and the median age was 36 years, and for every 100 females, there were 83.5 males?"
Now, to answer the rewritten question, For every 100 females age 18 and over, there were 77.7 males. So for every 100 females, there were 77.7 males. There were 83.5 males. So 100 - 77.7 = 22.3 males. The answer is 22.3.

Table 22: **DROP Census** Few-shot exemplars used for EchoPrompt with question first rephrasing

Q: For the period 2011-15, the estimated median annual income for a household in the town was $48,059, and the median income for a family was $96,005. Male full-time workers had a median income of $64,750 versus $39,278 for females. The per capita income for the town was $18,905. About 8.7% of families and 34.7% of the population were below the poverty line, including 18.2% of those under age 18 and 4.5% of those age 65 or over. How many percent of the population is not living below the poverty line?
A: Rewriting in simple words, the question is: "For the period between 2011-15, the estimated median annual income for a household in the town was $48,059, and the median income for a family was $96,005. Male full-time workers had a median income of $64,750, while females had a median income of $39,278. The per capita income for the town was $18,905. Approximately 8.7% of families and 34.7% of the population were living below the poverty line, including 18.2% of those under the age of 18 and 4.5% of those over the age of 65. How many percent of the population is not living below the poverty line?"
Now, to answer the rewritten question, The population is 34.7% below the poverty line. So the population is 100 - 34.7 = 65.3% not below the poverty line. The answer is 65.3.
Q: As of the census of 2010, there were 2,213 people, 870 households, and 563 families residing in the town with 902 housing units. The racial makeup of the town was 82.0% White (U.S. Census), 6.1% African American (U.S. Census), 0.1% Native American (U.S. Census), 5.7% Asian (U.S. Census), 0.1% Race (U.S. Census), 2.8% from Race (U.S. Census), and 3.2% from two or more races. Hispanic (U.S. Census) or Latino (U.S. Census) of any race were 6.4% of the population. How many more people are there than housing units?
A: Rewriting in simple words, the question is: "According to the census of 2010, the town had a population of 2,213 people, with 902 housing units available. The racial makeup of the town was 82.0% White, 6.1% African American, 0.1% Native American, 5.7% Asian, 0.1% of another race, 2.8% from two or more races, and 3.2% from some other race. Hispanic or Latino of any race were 6.4% of the population. How many more people are there than housing units?"
Now, to answer the rewritten question, There are 2,213 and 902 housing units. So the difference is 2,213 - 902 = 1,311. The answer is 1,311.
Q: In the city, the population was spread out with 28.1% under the age of 18, 9.2% from 18 to 24, 26.2% from 25 to 44, 21.4% from 45 to 64, and 15.1% who were 65 years of age or older. The median age was 36 years. For every 100 females, there were 83.5 males. For every 100 females age 18 and over, there were 77.7 males. For every 100 females age 18 and over, how many fewer males were there?
A: Rewriting in simple words, the question is: "The city's population was spread out with 28.1% under 18, 9.2% between 18-24, 26.2% between 25-44, 21.4% between 45-64, and 15.1% who were 65 or older. The median age was 36 years. For every 100 females, there were 83.5 males, and for every 100 females 18 and over, there were 77.7 males. For every 100 females age 18 and over, how many fewer males were there?"
Now, to answer the rewritten question, For every 100 females age 18 and over, there were 77.7 males. So for every 100 females, there were 77.7 males. There were 83.5 males. So 100 - 77.7 = 22.3 males. The answer is 22.3.

Table 23: **DROP Census** Few-shot exemplars used for EchoPrompt with simple sentence rephrasing

Q: Since the 1970s, U.S. governments have negotiated managed-trade agreements, such as the North American Free Trade Agreement in the 1990s, the Dominican Republic-Central America Free Trade Agreement in 2006, and a number of bilateral agreements. In Europe, six countries formed the European Coal and Steel Community in 1951 which became the European Economic Community in 1958. Two core objectives of the EEC were the development of a common market, subsequently renamed the single market, and establishing a customs union between its member states. How many years did the European Coal and Steel Community exist?
A: Rewriting in simple words, the question is: "Given that U.S. governments have negotiated managed-trade agreements since the 1970s, such as the North American Free Trade Agreement in the 1990s and the Dominican Republic-Central America Free Trade Agreement in 2006, and that in Europe, six countries formed the European Coal and Steel Community in 1951 which later became the European Economic Community in 1958 with the objectives of developing a common market and establishing a customs union between its member states, how long did the European Coal and Steel Community exist?" Now, to answer the rewritten question, According to the passage, the European Coal and Steel Community was established in 1951 and became the EEC in 1958. 1958 - 1951 = 7. So the answer is 7.
Q: In the county, the population was spread out with 23.50% under the age of 18, 8.70% from 18 to 24, 29.70% from 25 to 44, 24.70% from 45 to 64, and 13.30% who were 65 years of age or older. How many more percent are under the age of 18 compared to the 18 to 24 group?
A: Rewriting in simple words, the question is: "Given that in the county, the population was spread out with 23.50% under the age of 18, 8.70% from 18 to 24, 29.70% from 25 to 44, 24.70% from 45 to 64, and 13.30% who were 65 years of age or older, what is the difference between the percentage of people under the age of 18 and the percentage of people from 18 to 24?" Now, to answer the rewritten question, According to the passage, 23.5% are under the age of 18, and 8.7% are from ages 18 to 24. 23.5% - 8.7% = 14.8%. So the answer is 14.8.
Q: Playing in their second straight Thanksgiving game, the Eagles struggled especially on defense, where they were unable to stop the much-hyped Lions offense. The worst of it all was how unproven rookie Eric Rowe was tasked with covering wide receiver Calvin Johnson, leading to Johnson catching 3 touchdowns. Stafford's five passing touchdowns, including three of them to Johnson was too much for the Eagles to overcome and for the second consecutive time this season, the Eagles gave up 45 points in a game. With the loss, the Eagles drop to 4-7 on the season and 6-1 when playing on Thanksgiving. How many TD passes did Stafford throw other than to Jhonson?
A: Rewriting in simple words, the question is: "Given that playing in their second straight Thanksgiving game, the Eagles struggled especially on defense, where they were unable to stop the much-hyped Lions offense, and that the worst of it all was how unproven rookie Eric Rowe was tasked with covering wide receiver Calvin Johnson, leading to Johnson catching 3 touchdowns, and that Stafford's five passing touchdowns, including three of them to Johnson was too much for the Eagles to overcome and for the second consecutive time this season, the Eagles gave up 45 points in a game, and that with the loss, the Eagles drop to 4-7 on the season and 6-1 when playing on Thanksgiving, how many TD passes did Stafford throw other than to Jhonson?" Now, to answer the rewritten question, According to the passage, Stafford threw 5 TD passes, 3 of which were to Johnson. 5 - 3 = 2. So the answer is 2.

Table 24: **DROP (Break, Non-football) subsets** Few-shot exemplars used for EchoPrompt with compound sentence rephrasing

Q: Since the 1970s, U.S. governments have negotiated managed-trade agreements, such as the North American Free Trade Agreement in the 1990s, the Dominican Republic-Central America Free Trade Agreement in 2006, and a number of bilateral agreements. In Europe, six countries formed the European Coal and Steel Community in 1951 which became the European Economic Community in 1958. Two core objectives of the EEC were the development of a common market, subsequently renamed the single market, and establishing a customs union between its member states. How many years did the European Coal and Steel Community exist?
A: Rewriting in simple words, the question is: "How many years did the European Coal and Steel Community exist, if it was formed in 1951 by six European countries and subsequently became the European Economic Community in 1958, with two core objectives being the establishment of a customs union between member states and the development of a common market? Also, noting that since the 1970s, the U.S. government has negotiated managed-trade agreements, including the North American Free Trade Agreement in the 1990s, the Dominican Republic-Central America Free Trade Agreement in 2006, and several bilateral agreements?"
Now, to answer the rewritten question, According to the passage, the European Coal and Steel Community was established in 1951 and became the EEC in 1958. 1958 - 1951 = 7. So the answer is 7.
Q: In the county, the population was spread out with 23.50% under the age of 18, 8.70% from 18 to 24, 29.70% from 25 to 44, 24.70% from 45 to 64, and 13.30% who were 65 years of age or older. How many more percent are under the age of 18 compared to the 18 to 24 group?
A: Rewriting in simple words, the question is: "How much higher is the percentage of the population under the age of 18 compared to the percentage of the population aged between 18 to 24, in the county where the population is spread out with 23.50% under the age of 18, 8.70% from 18 to 24, 29.70% from 25 to 44, 24.70% from 45 to 64, and 13.30% who were 65 years of age or older?"
Now, to answer the rewritten question, According to the passage, 23.5% are under the age of 18, and 8.7% are from ages 18 to 24. 23.5% - 8.7% = 14.8%. So the answer is 14.8.
Q: Playing in their second straight Thanksgiving game, the Eagles struggled especially on defense, where they were unable to stop the much-hyped Lions offense. The worst of it all was how unproven rookie Eric Rowe was tasked with covering wide receiver Calvin Johnson, leading to Johnson catching 3 touchdowns. Stafford's five passing touchdowns, including three of them to Johnson was too much for the Eagles to overcome and for the second consecutive time this season, the Eagles gave up 45 points in a game. With the loss, the Eagles drop to 4-7 on the season and 6-1 when playing on Thanksgiving. How many TD passes did Stafford throw other than to Jhonson?
A: Rewriting in simple words, the question is: "How many TD passes did Stafford throw other than to Jhonson, in a game where the Eagles struggled especially on defense, where they were unable to stop the much-hyped Lions offense, and where the worst of it all was how unproven rookie Eric Rowe was tasked with covering wide receiver Calvin Johnson, leading to Johnson catching 3 touchdowns? Also, noting that Stafford's five passing touchdowns, including three of them to Johnson was too much for the Eagles to overcome and for the second consecutive time this season, the Eagles gave up 45 points in a game. With the loss, the Eagles drop to 4-7 on the season and 6-1 when playing on Thanksgiving?"
Now, to answer the rewritten question, According to the passage, Stafford threw 5 TD passes, 3 of which were to Johnson. 5 - 3 = 2. So the answer is 2.

Table 25: **DROP (Break, Non-football) subsets** Few-shot exemplars used for EchoPrompt with question first rephrasing

Q: Since the 1970s, U.S. governments have negotiated managed-trade agreements, such as the North American Free Trade Agreement in the 1990s, the Dominican Republic-Central America Free Trade Agreement in 2006, and a number of bilateral agreements. In Europe, six countries formed the European Coal and Steel Community in 1951 which became the European Economic Community in 1958. Two core objectives of the EEC were the development of a common market, subsequently renamed the single market, and establishing a customs union between its member states. How many years did the European Coal and Steel Community exist?
A: Rewriting in simple words, the question is: "The US government negotiated managed-trade agreements since the 1970s. Some examples include the North American Free Trade Agreement in the 1990s, the Dominican Republic-Central America Free Trade Agreement in 2006, and various bilateral agreements. The European Coal and Steel Community was established in 1951 by six countries in Europe. It evolved into the European Economic Community in 1958, with two primary goals of creating a common market and a customs union between member states. What is the duration of the existence of the European Coal and Steel Community?"
Now, to answer the rewritten question, According to the passage, the European Coal and Steel Community was established in 1951 and became the EEC in 1958. 1958 - 1951 = 7. So the answer is 7.
Q: In the county, the population was spread out with 23.50% under the age of 18, 8.70% from 18 to 24, 29.70% from 25 to 44, 24.70% from 45 to 64, and 13.30% who were 65 years of age or older. How many more percent are under the age of 18 compared to the 18 to 24 group?
A: Rewriting in simple words, the question is: "The county's population was distributed into different age groups. 23.50% under 18 years, 8.70% aged between 18 to 24, 29.70% between 25 to 44, 24.70% between 45 to 64, and 13.30% were 65 years or older. What is the percentage difference between the under-18 group and the 18 to 24 age group?"
Now, to answer the rewritten question, According to the passage, 23.5% are under the age of 18, and 8.7% are from ages 18 to 24. 23.5% - 8.7% = 14.8%. So the answer is 14.8.
Q: Playing in their second straight Thanksgiving game, the Eagles struggled especially on defense, where they were unable to stop the much-hyped Lions offense. The worst of it all was how unproven rookie Eric Rowe was tasked with covering wide receiver Calvin Johnson, leading to Johnson catching 3 touchdowns. Stafford's five passing touchdowns, including three of them to Johnson was too much for the Eagles to overcome and for the second consecutive time this season, the Eagles gave up 45 points in a game. With the loss, the Eagles drop to 4-7 on the season and 6-1 when playing on Thanksgiving. How many TD passes did Stafford throw other than to Jhonson?
A: Rewriting in simple words, the question is: "The Eagles played in their second straight Thanksgiving game but struggled, especially on defense. They were unable to stop the Lions' much-hyped offense, and rookie Eric Rowe struggled to cover wide receiver Calvin Johnson, who caught three touchdowns. Stafford threw five passing touchdowns, three of them to Johnson. The Eagles gave up 45 points for the second consecutive time this season, resulting in a loss, bringing their record to 4-7 for the season. They have a 6-1 record when playing on Thanksgiving. What is the number of TD passes thrown by Stafford, other than to Johnson?"
Now, to answer the rewritten question, According to the passage, Stafford threw 5 TD passes, 3 of which were to Johnson. 5 - 3 = 2. So the answer is 2.

Table 26: **DROP (Break, Non-football) subsets** Few-shot exemplars used for EchoPrompt with simple sentence rephrasing

Q: 2015 is coming in 36 hours. What is the date one week from today in MM/DD/YYYY?
A: Rewriting in simple words, the question is: "Given that 2015 is coming in 36 hours, what is the date one week from today in MM/DD/YYYY?"
Now, to answer the rewritten question, If 2015 is coming in 36 hours, then it is coming in 2 days. 2 days before 01/01/2015 is 12/30/2014, so today is 12/30/2014. So one week from today will be 01/05/2015. So the answer is 01/05/2015.
Q: The first day of 2019 is a Tuesday, and today is the first Monday of 2019. What is the date today in MM/DD/YYYY?
A: Rewriting in simple words, the question is: "Given that the first day of 2019 is a Tuesday and that today is the first Monday of 2019, what is the date today in MM/DD/YYYY?"
Now, to answer the rewritten question, If the first day of 2019 was Tuesday, then 01/01/2019 was a Tuesday. Today is the first monday, would be six days later. So today is 01/07/2019. So the answer is 01/07/2019.
Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?
A: Rewriting in simple words, the question is: "Given that the concert was scheduled to be on 06/01/1943, but was delayed by one day to today, what is the date 10 days ago in MM/DD/YYYY?"
Now, to answer the rewritten question, One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.
Q: It is 4/19/1969 today. What is the date 24 hours later in MM/DD/YYYY?
A: Rewriting in simple words, the question is: "Given that it is 4/19/1969 today, what is the date 24 hours later in MM/DD/YYYY?"
Now, to answer the rewritten question, Today is 04/19/1969. 24 hours later is one day after today, which would be 04/20/1969. So the answer is 04/20/1969.
Q: Jane thought today is 3/11/2002, but today is in fact Mar 12, which is 1 day later. What is the date 24 hours later in MM/DD/YYYY?
A: Rewriting in simple words, the question is: "Given that Jane thought today is 3/11/2002, but today is in fact Mar 12, which is 1 day later, what is the date 24 hours later in MM/DD/YYYY?"
Now, to answer the rewritten question, Today is 03/12/2002. So the date 24 hours later will be 03/13/2002. So the answer is 03/13/2002.
Q: Jane was born on the last day of Feburary in 2001. Today is her 16-year-old birthday. What is the date yesterday in MM/DD/YYYY?
A: Rewriting in simple words, the question is: "Given that Jane was born on the last day of Feburary in 2001 and that today is her 16-year-old birthday, what is the date yesterday in MM/DD/YYYY?"
Now, to answer the rewritten question, The last day of February is the 28th, so Jane was born on 02/28/2001. Today is her 16-year old birthday, so today is 02/28/2017. So yesterday was 02/27/2017. So the answer is 02/27/2017.

Table 27: **Date Understanding** Few-shot exemplars used for EchoPrompt with compound sentence rephrasing

Q: From the German point of view, March 1941 saw an improvement. The Luftwaffe flew 4,000 sorties that month, including 12 major and three heavy attacks. The electronic war intensified but the Luftwaffe flew major inland missions only on moonlit nights. Ports were easier to find and made better targets. To confuse the British, radio silence was observed until the bombs fell. X- and Y-Ger¨at beams were placed over false targets and switched only at the last minute. Rapid frequency changes were introduced for X-Ger¨at, whose wider band of frequencies and greater tactical flexibility ensured it remained effective at a time when British selective jamming was degrading the effectiveness of Y-Ger¨at. How many sorties were flown in March 1941?
A: Rewriting in simple words, the question is: "Given that from the German point of view, March 1941 saw an improvement, and that the Luftwaffe flew 4,000 sorties that month, including 12 major and three heavy attacks, and that the electronic war intensified but the Luftwaffe flew major inland missions only on moonlit nights, and that ports were easier to find and made better targets, and that to confuse the British, radio silence was observed until the bombs fell, and that X- and Y-Ger¨at beams were placed over false targets and switched only at the last minute, and that rapid frequency changes were introduced for X-Ger¨at, whose wider band of frequencies and greater tactical flexibility ensured it remained effective at a time when British selective jamming was degrading the effectiveness of Y-Ger¨at, how many sorties were flown in March 1941?"
Now, to answer the rewritten question, the answer is "4,000".
Q: At the 52nd Annual Grammy Awards, Beyoncé received ten nominations, including Album of the Year for I Am... Sasha Fierce, Record of the Year for "Halo", and Song of the Year for "Single Ladies (Put a Ring on It)", among others. She tied with Lauryn Hill for most Grammy nominations in a single year by a female artist. In 2010, Beyoncé was featured on Lady Gaga's single "Telephone" and its music video. The song topped the US Pop Songs chart, becoming the sixth number-one for both Beyoncé and Gaga, tying them with Mariah Carey for most number-ones since the Nielsen Top 40 airplay chart launched in 1992. "Telephone" received a Grammy Award nomination for Best Pop Collaboration with Vocals. Beyonce received how many nominations at the 52nd Annual Grammy Awards?
A: Rewriting in simple words, the question is: "Given that at the 52nd Annual Grammy Awards, Beyoncé received ten nominations, including Album of the Year for I Am... Sasha Fierce, Record of the Year for "Halo", and Song of the Year for "Single Ladies (Put a Ring on It)", among others, and that she tied with Lauryn Hill for most Grammy nominations in a single year by a female artist, and that in 2010, Beyoncé was featured on Lady Gaga's single "Telephone" and its music video, and that the song topped the US Pop Songs chart, becoming the sixth number-one for both Beyoncé and Gaga, tying them with Mariah Carey for most number-ones since the Nielsen Top 40 airplay chart launched in 1992, and that "Telephone" received a Grammy Award nomination for Best Pop Collaboration with Vocals, how many nominations did Beyonce receive at the 52nd Annual Grammy Awards?"
Now, to answer the rewritten question, the answer is "ten".

Table 28: **SQuAD** Few-shot exemplars used for EchoPrompt with compound sentence rephrasing