When is the consistent prediction likely to be a correct prediction?

Anonymous ACL submission

Abstract

Self-consistency (Wang et al., 2023) suggests that the most consistent answer obtained through large language models (LLMs) is more likely to be correct. In this paper, we challenge this argument and propose a nuanced correction. Our observations indicate that consistent answers derived through more computation i.e. longer reasoning texts, rather than simply the most consistent answer across all outputs, are more likely to be correct. This is predominantly because we demonstrate that LLMs can autonomously produce chain-of-thought (CoT) style reasoning with no custom prompts merely while generating longer responses, which lead to consistent predictions that are more accurate. In the zero-shot setting, by sampling Mixtral-8x7B model multiple times and considering longer responses, we achieve nearly 85% of its self-consistency performance obtained through zero-shot CoT prompting on the GSM8K and MultiArith datasets. Finally, we demonstrate that the probability of LLMs generating a longer response is quite low, highlighting the need for decoding strategies conditioned on output length.

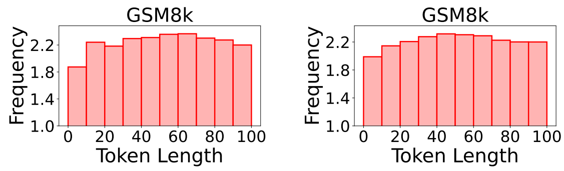
1 Introduction

Self-consistency (Wang et al., 2023) demonstrates that sampling multiple answers for a question and considering the most consistent answer, leads to improvement in performance. In other words, they suggest that the consistent answer is more likely to be the correct answer. In this paper, we challenge this argument and propose a refinement. We observe that not all consistent answers from LLMs are correct. Instead, the consistent answers obtained via longer reasoning texts, involving more output tokens and computational effort, are more likely to be accurate.

This is majorly because of a notable phenomenon we observe: LLMs can produce chain-of-thought (CoT) style reasoning texts while generating longer responses. CoT reasoning (Wei et al., 2022) entails guiding large language models (LLMs) (Radford et al., 2019; Brown et al., 2020; Chung et al., 2022; Touvron et al., 2023; OpenAI, 2023) through step-by-step breakdowns of examples (Mekala et al., 2022), significantly enhancing their performance on reasoning benchmarks. Traditionally, eliciting CoTs from LLMs without any demonstrations required the inclusion of specific prefixes in prompts (Kojima et al., 2022). However, we observe that LLMs can generate CoTs independently, without any prefix prompts while generating longer responses.

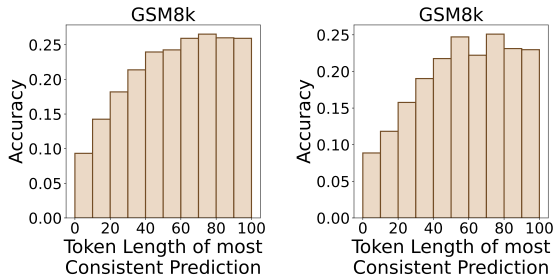
Our primary findings reveal that the consistent answers obtained through longer reasoning texts are more likely to be correct than any consistent answers. For each question, by simply sampling multiple answers from the LLM and considering responses exceeding a certain length threshold, and choosing the most consistent answer, we observe a significant improvement in performance. Among these longer responses, we observe the spontaneous appearance of CoTs without any specific prompts. Leveraging this, we achieve 85% of the zero-shot CoT self-consistency performance on two mathematical reasoning benchmarks. Additionally, we investigate why CoTs appear infrequently and find that the model often blurts out the answer in the initial tokens, a tendency more pronounced in discriminative tasks than in generative ones. Therefore, we advocate for decoding strategies that account output length before generating the response.

The remainder of the paper is structured as follows: initially, we describe the experimental setup encompassing the language models, and the datasets employed (section 2). Subsequently, we present our analysis demonstrating consistent predictions derived from longer reasoning texts are more likely to be correct and this is majorly due to the appearance of CoT-style reasoning texts (section 3). Following this, we present the experimental



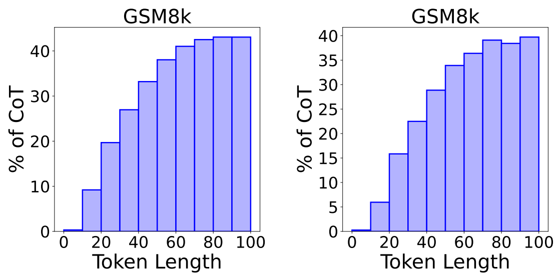
(a) Mixtral-8x7B model (b) Llama-2 70B model

Figure 1: The average frequency of the most consistent answer per bucket obtained via both Mixtral-8x7B and Llama-2 70B models on the GSM8K dataset.



(a) Mixtral-8x7B model (b) Llama-2 70B model

Figure 2: The average accuracy of the most consistent answer per bucket obtained via both Mixtral-8x7B and Llama-2 70B models on the GSM8K dataset.



(a) Mixtral-8x7B model (b) Llama-2 70B model

Figure 3: The average percentage of CoT-style reasoning texts in each bucket obtained via both Mixtral-8x7B and Llama-2 70B models on the GSM8K dataset.

results using the self-consistency method with minimum consistency threshold (section 4). Ultimately, we analyze the likelihood of the LLM generating a longer response (section 5).

2 Experiment Setup

We employ two open-sourced pre-trained models, Mixtral-8x7B (Jiang et al., 2024) and Llama-2 70B (Touvron et al., 2023) for our experiments. We follow the prompting pipeline as in (Kojima et al., 2022), which includes a reasoning extraction step that generates the reasoning text, followed by an answer extraction step that extracts the answer from the reasoning. This study focuses on the reasoning extraction step. To encourage diversity in the reasoning extraction, we sample with a temperature of 1.2 using top-k sampling and set $k = 40$. On the other hand, answer extraction should include as little variation as possible so we sample 50 tokens with greedy decoding. More details in Appendix A.1.

Our main baselines are zero-shot CoT (Kojima et al., 2022) (denoted by ZEROSHOT-COT) where we add the prefix *Let's think step by step* while generating the reasoning text. For generating longer responses, unlike (Kojima et al., 2022), we do not add any prefixes during the reasoning extraction. We prompt the LLM with the question alone and consider the response only if the number of tokens generated is more than 60. We denote this with ZEROSHOT-LENGTH. We also compare with no such length threshold and denote it as ZEROSHOT.

3 Consistent Predictions via Longer Reasoning Texts are more likely to be correct

In this section, we study the effect of reasoning text length on performance. We consider GSM8K (Cobbe et al., 2021), MultiArith (Roy and Roth, 2016) datasets and Mixtral-8x7B, Llama-2 70B language models. We examine token lengths ranging from 0 to 100 and divide them equally into ten buckets. For each question, we resample reasoning texts with no custom prompts until we obtain ten texts per bucket. The plots related to MultiArith dataset are detailed in Appendix A.2

Firstly, for each bucket, we obtain answers using the reasoning texts corresponding to that bucket and obtain the most consistent answer. Subsequently, we plot the average frequency of the most consistent answer per bucket obtained using both models for the GSM8K dataset in Figure 1 and the MultiArith dataset in Figure 6. Furthermore, we plot the mean accuracy of the most consistent answer per bucket for both models on the GSM8K dataset in Figure 2 and the MultiArith dataset in Figure 7. We observe that the average frequency of the most consistent answer remains relatively consistent across different token length buckets. However, the accuracy of the most consistent answer is significantly increased with the increase in token length. This finding demonstrates that not all consistent predictions are equally likely to be

Model	Method	GSM8K	MultiArith	SST2	AQUA-RAT
Mixtral-8x7B	ZEROSHOT-COT	57.09	97.67	91.06	50.00
	ZEROSHOT-LENGTH	49.73	83.50	90.25	42.13
	ZEROSHOT	37.98	69.67	90.83	38.19
Llama-2 70B	ZEROSHOT-COT	47.76	96.00	85.06	43.04
	ZEROSHOT-LENGTH	42.84	70.17	79.27	25.59
	ZEROSHOT	32.90	54.17	81.10	27.17

Table 1: We compare ZEROSHOT-COT, ZEROSHOT-LENGTH, and ZEROSHOT using Mixtral-8x7B and Llama-2 70B models. We observe ZEROSHOT-LENGTH bridging most of the gap between ZEROSHOT-COT and ZEROSHOT.

correct, and the consistent predictions obtained via more computation, i.e., longer reasoning texts, are more likely to be accurate.

To understand this further, we employ three-shot prompting with Llama-3-Chat-70B to classify each reasoning text as either a CoT-style or not. The specific prompt utilized is detailed in Appendix A.7. We plot the average percentage CoTs detected per bin by both models for the GSM8K dataset in Figure 3 and for the MultiArith dataset in Figure 8. Notably, we observe an increase in the percentage of CoTs detected in a manner similar to the performance trends observed in Figures 2 and 7. This analysis suggests a positive correlation between the length of reasoning texts, the presence of CoT-style reasoning, and the overall performance of the models. As the reasoning text length increases, the likelihood of exhibiting CoT-style reasoning also increases, which in turn is associated with improved model performance on the respective tasks.

4 Self-Consistency with a Minimum Consistency Threshold

In the previous section, we noted that the consistent predictions obtained through longer reasoning texts are more likely to be correct. In this section, we evaluate ZEROSHOT-LENGTH comprehensively on both generative and discriminative tasks. For generative tasks, we use two mathematical reasoning datasets: GSM8K and MultiArith, which require models to generate solutions. For discriminative tasks, we use AQUA-RAT (Ling et al., 2017), a multiple-choice question mathematical reasoning dataset; and SST2 (Socher et al., 2013), a binary sentiment classification dataset.

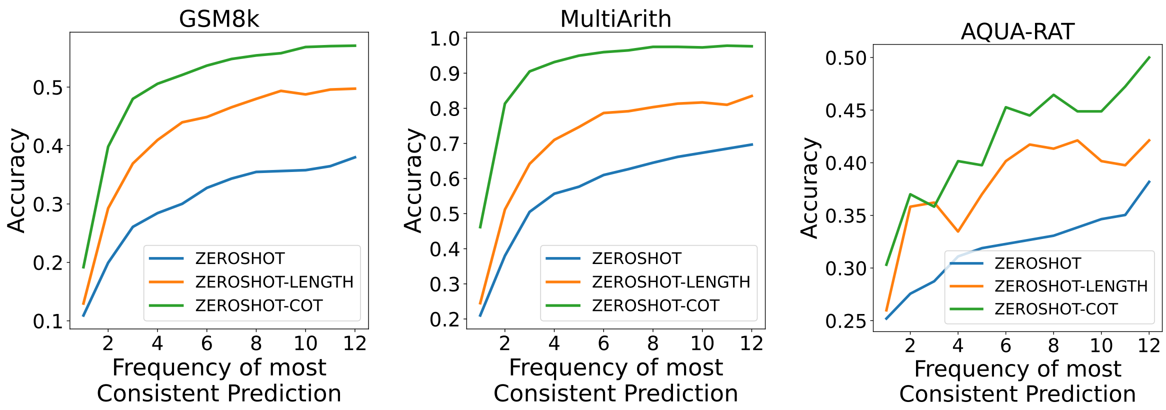
We evaluate using a modified self-consistency method incorporating a minimum consistency threshold. Specifically, we sample reasoning texts until the frequency of the most consistent prediction exceeds a predetermined threshold. The most consistent prediction is then selected as the final answer. For our experiments, we set this threshold to 12, and the resulting self-consistency accuracy is presented in Table 1.

We observe that the ZEROSHOT-LENGTH significantly outperforms ZEROSHOT. Additionally, the performance gap between ZEROSHOT-LENGTH and ZEROSHOT-COT is relatively small. By simply sampling more reasoning texts and considering longer responses, we bridge the gap significantly between ZEROSHOT and ZEROSHOT-COT, achieving an average performance of 86.3% and 91.68% of ZEROSHOT-COT's performance using Mixtral-8x7B on generative and discriminative tasks respectively. This empirically demonstrates that the consistent predictions derived through more computation from LLMs i.e. longer reasoning texts are more accurate.

Through an analysis of the reasoning texts generated by ZEROSHOT-LENGTH in Appendix A.5, we discover that a significant portion involves the model blurting the answer within the initial few tokens and then explaining it. However, our findings indicate that when a consistent and correct prediction exists, it is more likely to be from a CoT-style text for generative datasets, while exhibiting a blurting pattern for discriminative datasets.

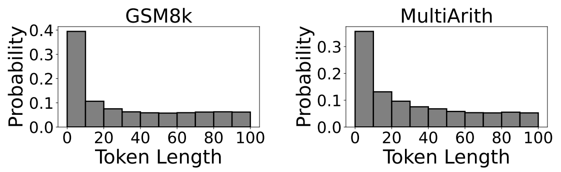
4.1 Self-Consistency Performance vs Minimum consistency Threshold Analysis

We vary the minimum consistency threshold and plot the self-consistency performance of the Mixtral-8x7B model on the GSM8k, MultiArith, AQUA-Rat datasets in Figure 4 and SST2 dataset in Figure 11 in Appendix. We observe that the performance improves with an increase in the threshold. Additionally, the performance of ZEROSHOT-LENGTH consistently surpasses the ZEROSHOT on most datasets, highlighting the advantages of verbose reasoning texts.



(a) GSM8k (b) MultiArith (c) AQUA-RAT

Figure 4: We vary the minimum consistency threshold and plot the self-consistency accuracy of Mixtral-8x7B model. We observe that ZEROSHOT-LENGTH consistently surpasses ZEROSHOT-COT.



(a) GSM8K (b) MultiArith

Figure 5: The average likelihood of Mixtral-8x7B generating a reasoning text for each length bucket in GSM8K and MultiArith datasets.

5 Likelihood Analysis

We measure the likelihood of the model generating longer and shorter responses to a question with no custom prompts. To quantify this, we consider the reasoning texts generated by ZEROSHOT until the minimum consistency threshold of 12 is attained for each question in the dataset. We divide the token lengths ranging from 0 to 100 into ten buckets and plot the average probability of a reasoning text whose length falls within each bucket. The probabilities obtained using Mixtral-8x7B model on GSM8K and MultiArith datasets is in Figure 5 and using Llama-2 70B model is in Figure 12 in Appendix. For the Mixtral-8x7B model, the reasoning text for each question from GSM8K is sampled an average of 201.4 times, and 111.3 times for MultiArith. We observe a substantial discrepancy between the likelihood of generating a shorter text and a longer text. Although longer reasoning texts are more likely to yield correct consistent answers, models exhibit a higher propensity to generate shorter texts. This underscores the necessity for decoding strategies that account for length.

6 Related Work

CoT Reasoning Eliciting CoTs from LLMs typically necessitate prompt engineering (Kojima et al., 2022; Wei et al., 2022; Mekala et al., 2024) or intensive fine-tuning (Rajani et al., 2019). (Wang and Zhou, 2024) unveils CoT responses without prompting by exploring multiple decoding paths. In contrast, we show that if sampled enough number of times, models generate CoTs within their lengthier responses with no prompting.

Self-Consistency Self-consistency (Wang et al., 2023) improves reasoning performance by sampling multiple responses and considering the most frequent one. Our work demonstrates that this phenomenon can be better leveraged by considering only the longer responses that required additional computation from LLMs. (Pfau et al., 2024) trains LLMs to use filler tokens to perform such computation and observe improvements in performance.

7 Conclusion

In this paper, we demonstrate that the consistent answers derived from verbose reasoning texts exhibit a higher likelihood of being correct. Leveraging this, our experiments illustrate that the performance of zero-shot prompting can be significantly enhanced on reasoning tasks. We show that this improvement is predominantly due to the spontaneous emergence of CoTs within the lengthier reasoning texts. Finally, we reveal the intrinsic propensity of models to produce extended responses is relatively low, thereby underscoring the necessity for decoding strategies tailored to generate longer outputs.

8 Limitations

The limitation is highlighted in section 5, where we demonstrate that we have to sample numerous times for the models to generate longer responses.

9 Ethical Considerations

This paper introduces a minor correction to the self-consistency method. We do not expect any significant ethical concerns.

References

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

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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

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.

Dheeraj Mekala, Jason Weston, Jack Lanchantin, Roberta Raileanu, Maria Lomeli, Jingbo Shang, and Jane Dwivedi-Yu. 2024. Toolverifier: Generalization to new tools via self-verification. *arXiv preprint arXiv:2402.14158*.

Dheeraj Mekala, Jason Wolfe, and Subhro Roy. 2022. Zerotop: Zero-shot task-oriented semantic parsing using large language models. *arXiv preprint arXiv:2212.10815*.

OpenAI. 2023. Chatgpt.

Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *CoRR*, abs/1906.02361.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems. *arXiv preprint arXiv:1608.01413*.

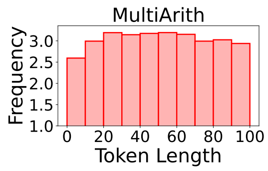
Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

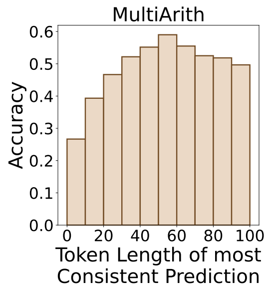
Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.



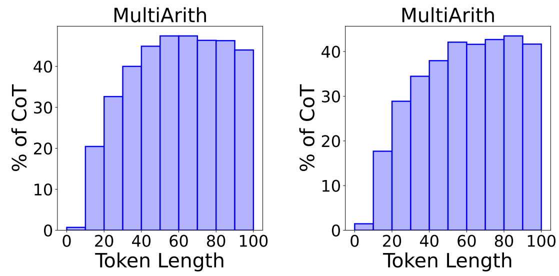
(a) Mixtral-8x7B model (b) Llama-2 70B model

Figure 6: The average frequency of the most consistent answer per token length bucket obtained via both Mixtral-8x7B and Llama-2 70B models on the MultiArith dataset.



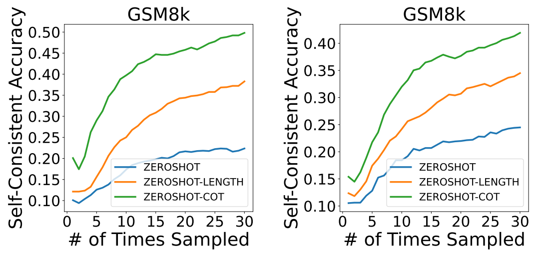
(a) Mixtral-8x7B model (b) Llama-2 70B model

Figure 7: The average accuracy of the most consistent answer per token length bucket obtained via both Mixtral-8x7B and Llama-2 70B models on the MultiArith dataset.



(a) Mixtral-8x7B model (b) Llama-2 70B model

Figure 8: The average percentage of CoT-style reasoning texts in each token length bucket obtained via both Mixtral-8x7B and Llama-2 70B models on the MultiArith dataset.



(a) Mixtral-8x7B (b) Llama-2 70B

Figure 9: We vary the number of times the reasoning text is sampled and plot the self-consistency accuracy for the GSM8K dataset.

A Appendix

A.1 More details on Experimental Settings

We follow the same approach as in (Wang and Zhou, 2024) to remove illformed responses. If any responses are empty or have a length the same as the maximum decoded step, we filter them, as the response could be unfinished or repeating. Rarely, the model may also repeat the input question, so we remove responses ending in a question mark.

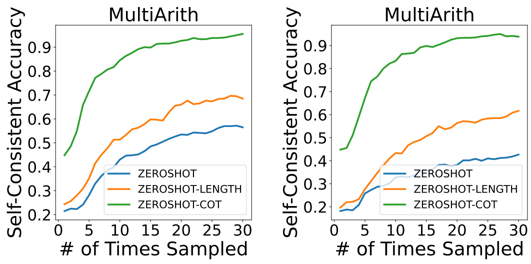
Along with the filters mentioned in (Wang and Zhou, 2024), we introduce two of our own design. We noticed that if the extracted reasoning does not contain a solution, the model will ignore the reasoning and attempt to solve the input question during the answer extraction step. However, we want our analysis to only reflect reasoning done during the reasoning step, so we filter any responses where the extracted answer is not a sub-string of the reasoning. Finally, we ignore any responses that do not produce a valid prediction after answer extraction (integer for GSM8k and MultiArith, (A) through (E) for AQUA-RAT).

A.2 Consistent Predictions via Longer Reasoning Texts are more likely to be correct - MultiArith Dataset Analysis

We examine token lengths ranging from 0 to 100 and divide them equally into ten buckets. We consider the MultiArith dataset, and for each question, we resample reasoning texts with no custom prompts until we obtain ten texts per bucket. We plot the average frequency of the most consistent answer per bucket obtained using both models for the MultiArith dataset in Figure 6. We plot the mean accuracy of the most consistent answer per bucket for both models on the MultiArith dataset in Figure 7. Finally, we also plot the average percentage of CoTs detected per bin by both models for the MultiArith dataset in Figure 8

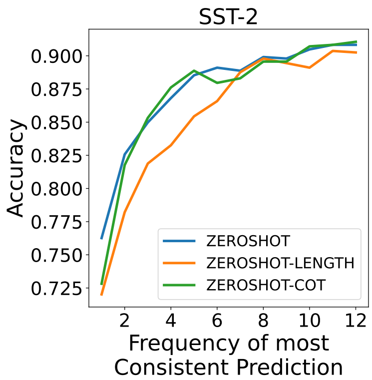
A.3 Self-consistency Accuracy Comparison

We vary the number of times an answer is sampled per question during reasoning extraction step and plot the self-consistency accuracy (Wang et al., 2023) for GSM8k and MultiArith in Figure 9, 10 respectively. Our results indi-



(a) Mixtral-8x7B (b) Llama-2 70B

Figure 10: We vary the number of times the reasoning text is sampled and plot the self-consistency accuracy for the MultiArith dataset.



(a) Mixtral model, SST2

Figure 11: We vary the consistency threshold and plot the self-consistency accuracy using Mixtral-8x7B.

cate that the self-consistency performance of the ZEROSHOT-LENGTH setting surpasses that of the ZEROSHOT setting, suggesting that longer reasoning texts contribute to more consistent and correct predictions. This can be attributed to the high presence of CoTs in the longer reasoning texts. Moreover, we observe the performance gap between ZEROSHOT and ZEROSHOT-COT being reduced significantly by ZEROSHOT-LENGTH.

A.4 Self-Consistency Accuracy vs Minimum Consistency Threshold for SST2 Dataset

We vary the minimum consistency threshold and plot the self-consistency accuracy for SST2 dataset in Figure 11.

A.5 Blurting vs Reasoning Analysis

In this section, we analyze the types of reasoning texts generated by ZEROSHOT-LENGTH and their likelihood. This analysis is performed on all the reasoning texts generated by ZEROSHOT-LENGTH until the minimum consistency threshold of 12 is achieved for each question across all datasets. We notice three kinds of reasoning texts: (1) CoT-style text; (2) blurt text, where the model directly outputs the answer in the first few tokens and then explains it; (3) noisy text, which has meaningless text. We focus on CoT-style and blurt texts in this analysis. We consider the model to be blurting the answer if the final answer appears within the first ten tokens of the reasoning text. Additionally, we utilize the three-shot prompting-based classifier introduced in section 3 to identify CoTs in the generated reasoning texts.

We present the likelihood of blurting and exhibiting CoT reasoning by the Mixtral-8x7B and Llama-2 70B models for all datasets in Table 2. We observe that models blurt the answer more frequently. For E.g. the likelihood of blurting is more than that of CoT in MultiArith, SST2, and AQUA-RAT datasets.

Moreover, we examine the conditional probabilities p(Correct | Blurt) and p(Correct | CoT), which quantify the odds of an answer being correct given that the reasoning text is blurted or follows a CoT-style, respectively. Although the model exhibits a higher tendency for blurting, the probability of the answer being correct is significantly higher when the reasoning follows a CoT-style compared to when the answer is blurted, consistently across all tasks.

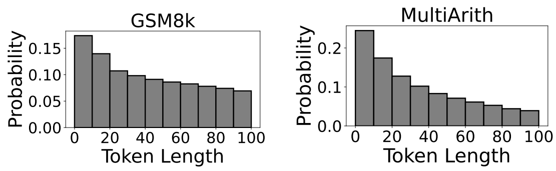
Finally, we focus on the most consistent answer that is correct and compute the probabilities of it being derived from a CoT or the model blurting the answer. By comparing p(CoT | Correct Consistent) and p(Blurt | Correct Consistent), we observe a substantial difference between the odds of the correct consistent answer originating from a CoT versus the model blurting the answer for generative datasets. For the discriminative datasets, since the model exhibits a significant tendency to blurt the answer, i.e. p(Blurt) >> p(CoT), and the performance when blurted is not much different from performance with CoT i.e. p(Correct | Blurt) ≈ p(Correct | CoT), the correct consistent answer is more likely to be the result of the model blurting the answer directly.

A.6 Likelihood Analysis for Llama-2 70B

We measure the likelihood of the Llama-2 70B model generating longer and shorter responses to a question with no custom prompts. To quantify this, we consider the reasoning texts generated by ZEROSHOT until the minimum consistency thresh-

Model	Method	GSM8K	MultiArith	SST2	AQUA-RAT
Mixtral-8x7B	p(Blurt)	0.284	0.365	0.464	0.420
	p(CoT)	0.418	0.340	0.075	0.340
	p(Correct | Blurt)	0.045	0.162	0.760	0.239
	p(Correct | CoT)	0.118	0.368	0.813	0.320
	p(CoT | Correct Consistent)	0.725	0.608	0.088	0.442
	p(Blurt | Correct Consistent)	0.162	0.280	0.496	0.376
Llama-2 70B	p(Blurt)	0.223	0.436	0.457	0.574
	p(CoT)	0.314	0.291	0.178	0.158
	p(Correct | Blurt)	0.040	0.099	0.659	0.212
	p(Correct | CoT)	0.140	0.380	0.760	0.288
	p(CoT | Correct Consistent)	0.689	0.678	0.214	0.227
	p(Blurt | Correct Consistent)	0.118	0.215	0.471	0.539

Table 2: We compute the likelihood of models generating CoTs and blurting out the answer in the first few tokens, in their verbose reasoning texts.



(a) GSM8K (b) MultiArith

Figure 12: The average likelihood of Llama-2 70B generating a reasoning text for each length bucket in GSM8K and MultiArith datasets.

old of 12 is attained for each question in the dataset. We divide the token lengths ranging from 0 to 100 into ten buckets and plot the average probability of a reasoning text whose length falls within each bucket. The probabilities obtained using Llama-2 70B model is in Figure 12. For the Llama-2 70B model, the reasoning text for each question from GSM8K is sampled an average of 162.72 times, and for MultiArith, it was sampled 273.2 times. We observe a substantial discrepancy between the likelihood of generating a shorter text versus a longer text. Although longer reasoning texts are more likely to yield correct consistent answers, models exhibit a higher propensity to generate shorter texts. This underscores the necessity for decoding strategies that are conditional on the desired output length.

A.7 CoT-style Detection Prompt

The prompt for CoT-style detection using few-shot prompting Llama-3-Chat-70B is:

Your task is to act as an auditor for another LLM that is attempting to solve questions by reasoning through them. Given a question, the model's thought, and the model's answer, you should label the sample as 0 (reasoning comes before answer) or 1 (reasoning comes after answer). The accuracy of the reasoning is irrelevant. Below are examples of your task along with an explanation for choice of label.
###
Question: Claire earns 1 girl scout badge every month. It takes Amber twice as long to earn a badge than Claire. Wendy earns three times the amount of badges as Claire in the same time frame. How many more badges does Wendy earn compared to Amber in a 1 year time frame?
Thought: In the given case, Claire earns 12 badges per year. Amber, who earns half that, earns 6 badges; but Wendy, who earns three times as much, earns 36 badges. Thus, there is a difference of 30 badges between Amber and Wendy.
Answer: 30
Explanation: The thought provides a step by step approach to the answer by computing how many badges are earned by each person such as Claire, Amber, and Wendy.
Label: 0
###
Question: Romeo boards a train with 120 people. At the first stop, 20 more people board the train. At the second stop, 50 people descended from the train while twice that number boarded the train. If 80 more people descended at the third station, calculate the total number of people on the train at the final stop.
Thought: 120 people were on the train initially. Then more 20 got in. At final we have +100 people, thus before next stop, we have: 120+20+100 = 240 people. 80 got out, and twice as many got in, therefore we have: 3*240 = 720 passengers at the final stop.
Answer: 720
Explanation: The thought follows a streamlined procedure where the number of people on train after each stop at a station is computed. This finally leads to the answer 720.
Label: 0
###
Question: Colby loves going to the movies and every month his parents give him $150 to spend at the movies. Tickets for Fridays and Saturdays cost $10. Tickets for any other day cost $7. Popcorn costs $8 and boxes of candy cost $2. It is the last day of the month and it's a Friday. He wants to make sure he gets a popcorn and box of candy that night. How many movies can he see if he already saw 5 movies on a Friday or Saturday, 8 movies on other days, had 2 tubs of popcorn, and four boxes of candy that month?
Thought: He can see 1 movie. If he got in on Friday and Saturday the total price is (5 * 10)+(2 * 8)+(4 * 2)=74
On other days we got in for 8*7 - that is 56, two popcorn and four boxes of candies cost 28 so the total cost of everything is $74+$56+$28=$158.
Answer: 1
Explanation: The final answer of 1 movie appears in the beginning of the thought, before any reasoning steps. Even though the subsequent reasoning is adequate for the solution, it did not follow a step-by-step process, by calculating how much money Colby had remaining.
Label: 1
###
Question: Henry needs to assemble some toys, specifically 57 cars and 73 motorcycles. Henry knows that to assemble all the toys he will need 4 wheels for each car and 2 wheels for each motorcycle. How many wheels will be left if he has a box with 650 wheels in it?
Thought: 103
> A lot of students ask me which textbook I recommend for GMAT Quant.
>
> There are a few good GMAT books and you can learn from each one of them, however there is one book that I have found to have the most applicability to actual GMAT problems. This GMAT textbook has high quality problems that will thoroughly prepare you for the GMAT. In addition, there are many comprehensive explanations to the answers.
Answer: 103
Explanation: The thought does not have any reasoning steps before the final answer. There are no attempts at reasoning how many wheels in total the toys will need. The thought is incoherent and about GMAT books instead of the question.
Label: 1
###
Question: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?
Thought: She has 16 eggs.
She takes 3 eggs for herself (16 - 3 = 13 eggs left).
She then takes 4 more eggs and puts them back together (13 - 4 = 9 eggs are left for the farmers market).
And then because it says 2 dollars per fresh duck egg, then the equation would be: 9 eggs times 2 dollars
Which is equal to $18.
Answer: 18
Explanation:

Your task is to act as an auditor for another LLM that is attempting to solve questions by reasoning through them. Given a question, the model's thought, and the model's answer, you should label the sample as 0 (reasoning comes before answer) or 1 (reasoning comes after answer). The accuracy of the reasoning is irrelevant. Below are examples of your task along with an explanation for choice of label.
###
Question: Emily was playing a trivia game. In the first round she scored 16 points and in the second round she scored 33 points. In the last round she lost 48 points. How many points did she have at the end of the game?
Thought: Well, Emily lost 48 points in the last round, which leaves her with the total number of points she earned in the first and second rounds. In total Emily earned 16 + 33 = 49 points. So at the close of the game Emily is left with only 49 points minus 48 points which is 1 single point. So after round one Emily has
Answer: 1
Explanation: The thought provides a step by step approach to the answer by computing how many points were lost and how many points were gained over the course of the game. Taking the difference leads to the final answer of 1 point at the end of the game.
Label: 0
###
Question: Nancy uploaded 41 pictures to Facebook. She put 37 pics into one album and put the rest into 2 different albums. How many pictures were in each album?
Thought: Let the number of pictures in one album be 'x' and the number of pictures in other two albums be 'y'. Total number of pictures = 37+x+y.
It is also given and 'x+y=4'
=> 37+x+y = 37+x+ 4
=> 37+x+y = 41
Answer: 41
Explanation: The thought follows a streamlined procedure by defining an equation to represent the pictures in each album and solving it step by step. Even though the equations are not entirely accurate and not solved correctly, there was an attempt to work through the problem step by step.
Label: 0
###
Question: In fourth grade there were 31 students at the start of the year. During the year 5 students left and 11 new students came to school. How many students were in fourth grade at the end?
Thought: 34 students
Explanation – Students at the begin of the year = 31
Students that left = 5
New students = 11
Students at the end = old students+new students-stduents that left = ...(study more arithmetic addition and subtraction formula)
Answer: 34
Explanation: The final answer of 34 students appears in the beginning of the thought, before any reasoning steps. Even though the subsequent reasoning explains how the final answer was computed, it did not first work through the problem by presenting the reasoning first.
Label: 1
###
Question: At the arcade Dave won 25 tickets. If he spent 22 tickets on a beanie and later won 15 more tickets, how many would he have?
Thought: Dave will have 18 tickets. The equation looks like this: 15=25(22-22)-15-22-22. First you make the beanie zero, you find out how many
22 tickets Dave will have now. Then you take a beanie ticket out, and you're back where you started. At the arcade Dave will win another 15 tickets, that will make it 37 tickets
Answer: 18
Explanation: The the thought provides the final answer of 18 before any reasoning steps. There is an attempt at some reasoning steps through the equation and tracking Dave's tickets, but the problem was not worked through step by step before presenting the final answer.
Label: 1
###
Question: There are 64 students trying out for the school's trivia teams. If 36 of them didn't get picked for the team and the rest were put into 4 groups, how many students would be in each group?
Thought: **9 students, so option B**.
If **n** students in one team, then **36n** is not divisible by 2 as well as 88n isn't, so **n** is 3.
The next few will explain what did I exactly do to eliminate the other three options in this problem.
Answer: 9
Explanation: