# Appendix of the paper "Concise Thoughts: Impact of Output Length on LLM Reasoning and Cost"

# Appendices

## A Additional analysis of the conciseness

### A.1 Analysis of the RMS

Figures 7,8, and 9 present the RMS values for Llama2-70b and Falcon-40b across the three datasets analyzed in the paper. These results specifically highlight scenarios where the model successfully answers questions with the CCoT prompt but not with the CoT prompt. The purpose of these plots is to check whether the behavior of CCoT in terms of redundancy is coherent with that observed in Section 7.4. In addition, these plots help to investigate the potential correlation between the accuracy of the model's responses and the reduction in redundancy. The plots in the left column correspond to Llama2-70b, while those in the right column correspond to Falcon-40b. In each plot, we also highlight the interval between Q1 and Q3 for CoT and CCoT answers, along with the overlap range of these intervals.

As we can see from these plots, the redundancy behavior for CCoT and CoT closely resembles the trend observed in Section 7.4 calculated across all questions. This suggests that the reduction in redundancy (for Llama2-70b) and number of steps (for Falcon-40b) achieved by CCoT is consistent and provides more accurate and concise answers compared to CoT. In fact, the reduction in terms of redundancy or number of steps in the cases where CoT fails indicates that the CCoT prompt provides better guidance to the model, even in complex or ambiguous scenarios where CoT tends to struggle.

### A.2 Information Flow

Since the main paper examined the information flow in answers with 8 steps based on the CCoT answer distribution, we further demonstrate the effectiveness of CCoT by analyzing step-by-step information flow for another median step count, specifically 9 steps. Tables 5 and 6 present the scores for Llama2-70b on GSM8K and SVAMP, respectively.

In Table 5, for GSM8K, all steps show a reduced repetition of semantic information for CCoT, except for CCoT-45 at the initial step and CCoT-100 at the final step. A different behavior is observed at the final step for all CCoTs in the other dataset (SVAMP), as shown in Table 6, although semantic information is retained throughout the steps.

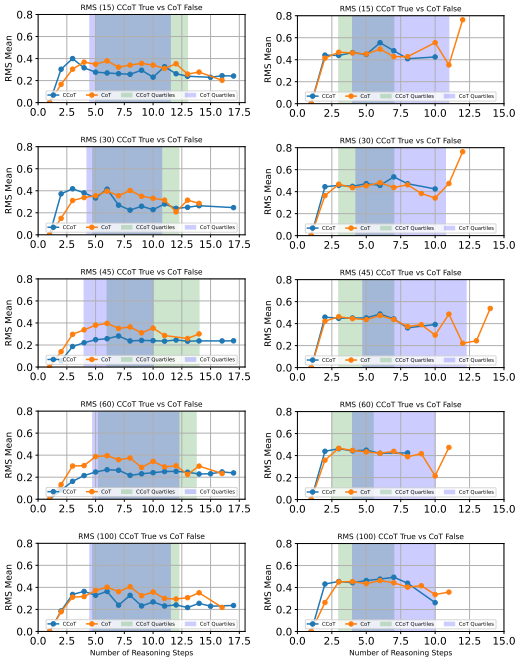



Figure 7: RMS mean score comparison between CCoT-true and CoT-false on the GSM8K dataset, with Llama2-70b (left side) and Falcon-40b (right side). The overlapped highlighted area illustrates the answer distribution between Q1 and Q3.

We also calculated the semantic information flow for answers with 8 and 9 steps, based on the median step distribution of CCoT answers generated by Falcon-40b. The results are presented in Tables 7 and 8 for the GSM8K and SVAMP datasets, respectively. Interestingly, Falcon-40b exhibits contrasting behavior in terms of information scores, often displaying higher values, which suggest greater repetition of semantic information across steps. This behavior is likely influenced by its medium-scale architecture and the nature of its training dataset, which may not be well-suited for handling constrained reasoning tasks.

However, as also highlighted in the main text of the paper, it is important to note that Falcon-40b

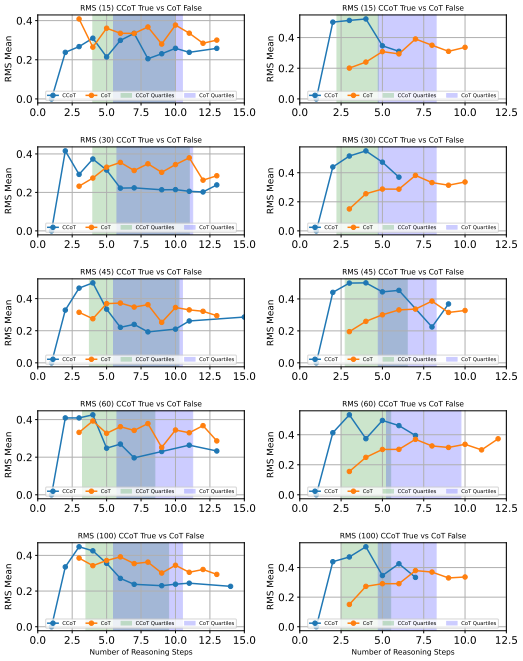



Figure 8: RMS mean score comparison between CCoT-true and CoT-false on the SVAMP dataset, with Llama2-70b (left side) and Falcon-40b (right side). The overlapped highlighted area illustrates the answer distribution between Q1 and Q3.

| Steps | CoT | CCoT-15 | CCoT-30 | CCoT-45 | CCoT-60 | CCoT-100 |
|---|---|---|---|---|---|---|
| 1 ⇒ 2 | 0.48 | 0.29 | 0.32 | 0.48 | 0.45 | 0.33 |
| 2 ⇒ 3 | 0.54 | 0.38 | 0.40 | 0.35 | 0.33 | 0.36 |
| 3 ⇒ 4 | 0.49 | 0.36 | 0.37 | 0.40 | 0.38 | 0.35 |
| 4 ⇒ 5 | 0.49 | 0.35 | 0.36 | 0.38 | 0.37 | 0.34 |
| 5 ⇒ 6 | 0.51 | 0.38 | 0.40 | 0.39 | 0.38 | 0.36 |
| 6 ⇒ 7 | 0.49 | 0.37 | 0.37 | 0.39 | 0.36 | 0.36 |
| 7 ⇒ 8 | 0.50 | 0.36 | 0.37 | 0.43 | 0.40 | 0.35 |
| 8 ⇒ 9 | 0.55 | 0.50 | 0.54 | 0.45 | 0.51 | 0.61 |

Table 5: GSM8K Llama2-70b Information Flow Mean Values comparison across answers with 9 setps

achieves improved conciseness due to the significantly lower average number of steps it produces. This allows for quicker reasoning decisions in many samples. This observation clarifies that the analysis reported in Tables 7 and 8 is inherently unbalanced, as the number of Falcon-40b answers with 8 and 9 steps under CCoT is smaller than those under CoT. Please note that in Table 8, there are no information scores for CCoT-15 because Falcon-40b does not generate answers with 8 or 9 steps when constrained to a reasoning length of up to 15 tokens.

## B Testing CCoT with smaller LLMs.

In this experiments, we investigate the capability of larger set of LLMs to handle the CCoT prompting. Specifically, in Figure 10, we present an evaluation conducted on five LLMs, including small

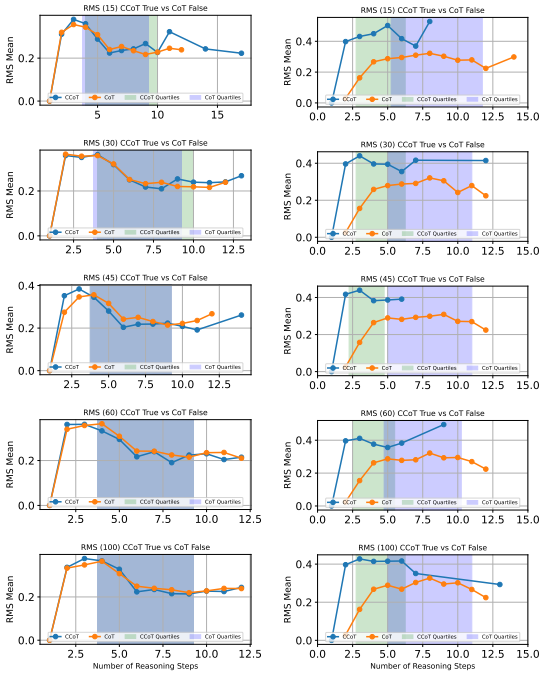



Figure 9: RMS mean score comparison between CCoT-true and CoT-false on the ASDIV dataset, with Llama2-70b shown (left side) and Falcon-40b (right side). The overlapped highlighted area illustrates the answer distribution between Q1 and Q3.

| Steps | CoT | CCoT-15 | CCoT-30 | CCoT-45 | CCoT-60 | CCoT-100 |
|---|---|---|---|---|---|---|
| 1 ⇒ 2 | 0.46 | 0.29 | 0.32 | 0.30 | 0.29 | 0.32 |
| 2 ⇒ 3 | 0.49 | 0.31 | 0.36 | 0.33 | 0.34 | 0.34 |
| 3 ⇒ 4 | 0.46 | 0.33 | 0.34 | 0.31 | 0.30 | 0.32 |
| 4 ⇒ 5 | 0.45 | 0.32 | 0.34 | 0.32 | 0.29 | 0.31 |
| 5 ⇒ 6 | 0.49 | 0.32 | 0.39 | 0.37 | 0.36 | 0.34 |
| 6 ⇒ 7 | 0.51 | 0.31 | 0.37 | 0.41 | 0.34 | 0.33 |
| 7 ⇒ 8 | 0.47 | 0.32 | 0.38 | 0.38 | 0.37 | 0.39 |
| 8 ⇒ 9 | 0.49 | 0.54 | 0.63 | 0.68 | 0.62 | 0.58 |

Table 6: SVAMP Llama70b Information Flow Mean Values comparison across answers with 9 setps

and medium-sized models such as Falcon-7b (Almazrouei et al., 2023) and LLama2-7b (Touvron et al., 2023). The results acknowledge some difficulties in addressing the CCoT prompt when considering smaller models.

We believe that such different outcomes of CCoT prompting can be attributed to various factors, such as the training data, the approach used to train the model, the model size, and the technique adopted during training. For instance, Llama2-70b is an autoregressive large-scale language model fine-tuned with human feedback, trained on a diverse combination of generic and open-source datasets. Such technical measures contribute to making CCoT effective in controlling the output length while improving the model accuracy. The Falcon-40b model, in contrast, is smaller than Llama2-70b and trained

Table 7: GSM8K Falcon-40b Information Flow Mean Values Comparison

| Steps | CoT | CCoT-15 | CCoT-30 | CCoT-45 | CCoT-60 | CCoT-100 |
|---|---|---|---|---|---|---|
| 1 ⇒ 2 | 0.58 | 0.65 | 0.64 | 0.65 | 0.61 | 0.59 |
| 2 ⇒ 3 | 0.62 | 0.69 | 0.68 | 0.70 | 0.68 | 0.64 |
| 3 ⇒ 4 | 0.60 | 0.74 | 0.76 | 0.74 | 0.68 | 0.63 |
| 4 ⇒ 5 | 0.59 | 0.66 | 0.63 | 0.64 | 0.64 | 0.66 |
| 5 ⇒ 6 | 0.61 | 0.62 | 0.72 | 0.67 | 0.66 | 0.64 |
| 6 ⇒ 7 | 0.60 | 0.61 | 0.66 | 0.68 | 0.68 | 0.62 |
| 7 ⇒ 8 | 0.59 | 0.50 | 0.55 | 0.54 | 0.56 | 0.59 |

(a) Answers with 8 steps

| Steps | CoT | CCoT-15 | CCoT-30 | CCoT-45 | CCoT-60 | CCoT-100 |
|---|---|---|---|---|---|---|
| 1 ⇒ 2 | 0.55 | 0.62 | 0.69 | 0.61 | 0.58 | 0.67 |
| 2 ⇒ 3 | 0.58 | 0.62 | 0.65 | 0.65 | 0.57 | 0.73 |
| 3 ⇒ 4 | 0.54 | 0.62 | 0.67 | 0.63 | 0.61 | 0.73 |
| 4 ⇒ 5 | 0.54 | 0.62 | 0.64 | 0.46 | 0.55 | 0.63 |
| 5 ⇒ 6 | 0.54 | 0.69 | 0.74 | 0.67 | 0.64 | 0.76 |
| 6 ⇒ 7 | 0.55 | 0.70 | 0.59 | 0.48 | 0.56 | 0.66 |
| 7 ⇒ 8 | 0.55 | 0.65 | 0.77 | 0.68 | 0.65 | 0.75 |
| 8 ⇒ 9 | 0.59 | 0.52 | 0.60 | 0.46 | 0.48 | 0.68 |

(b) Answers with 9 steps

Table 8: SVAMP Falcon-40b Information Flow Mean Values Comparison

| Steps | CoT | CCoT-15 | CCoT-30 | CCoT-45 | CCoT-60 | CCoT-100 |
|---|---|---|---|---|---|---|
| 1 ⇒ 2 | 0.26 | 0.60 | 0.58 | 0.55 | 0.58 | 0.61 |
| 2 ⇒ 3 | 0.68 | 0.58 | 0.53 | 0.53 | 0.60 | 0.52 |
| 3 ⇒ 4 | 0.60 | 0.68 | 0.52 | 0.60 | 0.49 | 0.54 |
| 4 ⇒ 5 | 0.61 | 0.56 | 0.51 | 0.56 | 0.63 | 0.60 |
| 5 ⇒ 6 | 0.63 | 0.51 | 0.65 | 0.55 | 0.68 | 0.64 |
| 6 ⇒ 7 | 0.63 | 0.95 | 0.76 | 0.52 | 0.64 | 0.62 |
| 7 ⇒ 8 | 0.55 | 0.73 | 0.57 | 0.60 | 0.56 | 0.65 |

(a) Answers with 8 Steps

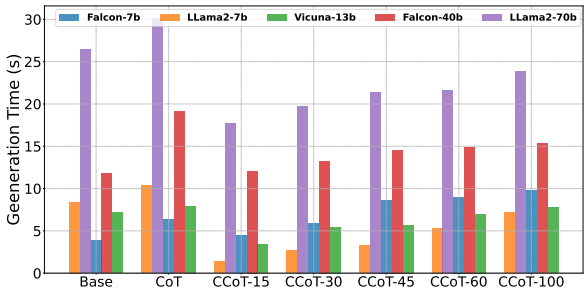
| Steps | CoT | CCoT-15 | CCoT-30 | CCoT-45 | CCoT-60 | CCoT-100 |
|---|---|---|---|---|---|---|
| 1 ⇒ 2 | 0.26 | - | 0.64 | 0.62 | 0.56 | 0.61 |
| 2 ⇒ 3 | 0.65 | - | 0.59 | 0.55 | 0.53 | 0.66 |
| 3 ⇒ 4 | 0.57 | - | 0.66 | 0.52 | 0.62 | 0.63 |
| 4 ⇒ 5 | 0.58 | - | 0.74 | 0.67 | 0.55 | 0.54 |
| 5 ⇒ 6 | 0.60 | - | 0.67 | 0.47 | 0.67 | 0.61 |
| 6 ⇒ 7 | 0.57 | - | 0.86 | 0.47 | 0.75 | 0.64 |
| 7 ⇒ 8 | 0.63 | - | 0.77 | 0.69 | 0.63 | 0.62 |
| 8 ⇒ 9 | 0.54 | - | 0.61 | 0.56 | 0.57 | 0.43 |

(b) Answers with 9 Steps

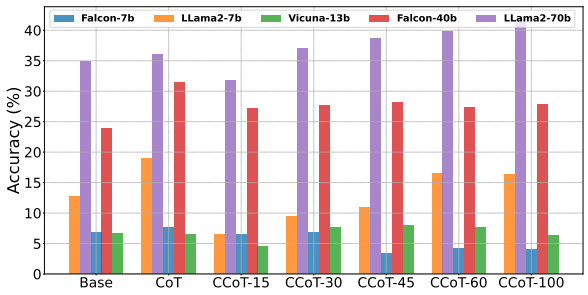
on a different dataset (the dedicated RefinedWeb data (Penedo et al., 2023)). While CCoT does not improve the accuracy of the model with respect to CoT, it still performs better than the base plain prompting, offering a trade-off by reducing generation times compared to CoT. Vicuna-13b also provides competitive results across different prompts, as it is a fine-tuned version of Llama2 and smaller than the previous Llama2-70b.

Conversely, small-scale LLMs, such as Falcon-7b and Llama2-7b, are not capable of properly handling the constrained prompting conditions in CCoT, resulting in higher generation times (as shown for Falcon-7b with large length values in CCoT) or incorrect answers with short CCoT values in Llama2-7b. This suggests that model size and training strategies severely impact the effectiveness of CCoT.

Considering the observations presented here, we believe that future directions could face potential training and fine-tuning strategy to integrate a better awareness and capability fo thandlign lengths also for smaller-sized models.




(a) Generation time




(b) Accuracy

Figure 10: Generation time (a) and accuracy (b) of five LLMs (Llama2-7b, Llama2-70b, Falcon-7b, Falcon-40b, and Vicuna-13b) on the GSM8K test dataset. Each model is evaluated using plain promt (base), CoT, and CCoT with different length constraints.