

Concise Thoughts: Impact of Output Length on LLM Reasoning and Cost

Sania Nayab[◊] Giulio Rossolini[◊] Marco Simoni^{*†} Andrea Saracino^{◊*}
Giorgio Buttazzo[◊] Nicolamaria Manes[†] Fabrizio Giacomelli[†]

[◊]Department of Excellence in Robotics and AI, Scuola Superiore Sant’Anna, Pisa, Italy

^{*}Institute of Informatics and Telematics, National Research Council of Italy

[†]Sapienza Università di Roma, [†]Mediavoice Srl - Roma e Napoli, Italy

Abstract

Today’s large language models (LLMs) can solve challenging question-answering tasks, and prompt engineering techniques, such as chain-of-thought (CoT), have gained attention for enhancing the explanation and correctness of outputs. However, many models and techniques tend to produce excessively verbose and lengthy answers, leading to issues with both conciseness and generation time. To address this, this paper analyzes the impact of output lengths on LLM inference pipelines by introducing and proposing novel metrics to evaluate the *correct conciseness* of a model and related prompting techniques. Then, we examine the impact of controlling output length through a refined prompt engineering strategy, Constrained-CoT (CCoT), which encourages the model to produce more concise outputs. To better understand the effects of such a prompt, we also introduce two additional scores for analyzing the conciseness, measured in terms of redundancy and information flow in generated answers. Experiments on pretrained LLMs and multiple datasets demonstrate the benefits of the proposed metrics and the effectiveness of CCoT across different models.

1 Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in tackling complex question-answering tasks, making significant strides in natural language understanding and generative AI (Taori et al., 2023; Chiang et al., 2023; Dolly, 2023; Geng et al., 2023). The continuous advancements made in architectures and training methods played a crucial role in enhancing the performance of these models. Alongside these developments, prompt techniques have also seen substantial evolution. One such technique that has attracted considerable attention is chain-of-thought (CoT) prompting (Wei et al., 2022; Fu et al., 2023),

which enhances the explanation and correctness of the output by encouraging the LLM to articulate its answer through intermediate reasoning steps.

Despite its advantages, the CoT prompting can lead to long outputs, increasing the time required for the model to generate a response. This is due to the nature of autoregressive transformers, which decode text word by word (Vaswani et al., 2017; Shekhar et al., 2024), which implies that the time required to generate a response is unbounded and heavily influenced by the length of the reasoning provided, as demonstrated in Section 3. Such lengthy and variable delays in responses can be undesirable when an LLM interacts with a user in an interactive conversation. Furthermore, especially for complex models, long answers imply a loss of conciseness and even of precision in the answer, with a performance degradation which is not only bound to computation time. This issue highlights the need to consider (i) metrics for evaluating the conciseness of the outputs and (ii) solutions to avoid excessively long chains of reasoning.

To address this, the first part of this work emphasizes the importance of accounting for the length of an answer in its correctness evaluation, as an indicator of computational cost. This is achieved by introducing three novel metrics (HCA, SCA, and CCA) that assess both the brevity and correctness of a generated answer. The proposed metrics aim to reweight the accuracy of a model by considering aspects related to output length that impact inference time and time predictability.

Then, to address the significant increase in output length caused by CoT techniques, the second part of this work explores how to leverage the benefits of CoT advances while getting control over the length of CoT reasoning through specific prompt designs. To this end, we introduce a refined prompt strategy called Constrained-CoT (CCoT), which encourages LLMs to generate concise outputs by explicitly limiting the reasoning length. The ap-

proach allows users to set a flexible length constraint that serves as a tunable parameter, balancing the strictness of brevity in the answers. The objective is to enable controlled reasoning, ensuring that outputs are concise and computationally efficient without sacrificing accuracy.

To better assess the ability of LLMs to follow such instructions and gain deeper insights into conciseness, we also introduce additional scores that analyze the level of redundancy and the information flow in the generated answers. These scores help demonstrate, through experimental analysis, that large models (such as Llama2-70b and Falcon-40b) can effectively leverage the proposed prompt to produce more concise yet accurate responses, while still retaining useful information. This allows achieving an enhanced trade-off between accuracy and brevity, measured comprehensively using the proposed metrics. For instance, using LLaMA2 on three datasets (GSM8K, SVAMP, ASDIV), constraining the reasoning length to 30 words (CCoT-30) increases the average accuracy by 4.41% and reduces computational costs by 5.12s. These accuracy and cost improvements are better remarked and unified in the proposed metrics, showing an improvement of 10%, 9% and 9% for HCA, SCA, and CCA, respectively.

To summarize, this work provides the following main contributions:

- Introduction of the concept of Constrained-CoT (CCoT), a prompt engineering strategy designed to limit the length of answers generated by LLMs, thereby enhancing the trade-off between generation time and correct conciseness.
- Three novel metrics to evaluate the correctness of LLM outputs while accounting for the conciseness (HCA, SCA and CCA).
- Introduction of an analysis of the conciseness in terms of redundancy and information flow for a given answer, thus offering an understanding of the effects of constraining output length.
- We conducted multiple experiments to analyze the impact of CCoT across different datasets and LLMs, demonstrating its benefits in terms of inference time, accuracy and conciseness with respect to the original CoT. Furthermore, we show the benefits of adopting the proposed metrics and scores in terms of conciseness.

The rest of the paper is organized as follows: Section 2 discusses the literature related to this work; Section 3 motivates the addressed study; Section 4 presents a set of metrics that account for conciseness; Section 5 introduces the proposed CCoT approach; Section 6 introduces analysis for evaluating the conciseness of a given answer; Section 7 reports the results of a set of experiments carried out on pre-trained models with three arithmetic reasoning datasets; and Section 8 states the conclusions and discusses some future directions.

2 Related work

To the best of our knowledge, most recent works on LLMs focused on increasing their accuracy (Jiang et al., 2020; Kaplan et al., 2020; Zhu et al., 2023). However, as models scale up, they tend to generate more extensive and articulated responses (Bhar-gava et al., 2023), which can introduce other problems, such as hallucinations (where the model produces information that appears plausible but not grounded (Kadavath et al., 2022), or unnecessarily long explanations (Qiu et al., 2024; Azaria and Mitchell, 2023)), which can obscure key information, making it difficult for users to extract relevant content efficiently (Khashabi et al., 2021; Wang et al., 2024b). To filter out useless reasoning, Li et al. (2021) proposed a multi-hop processing technique, where an extraction task on the encoder to obtain the rationale for an answer, which is the most relevant piece of text in an input prompt to a given question.

To further improve the accuracy of LLMs, several prompt engineering approaches have been presented in recent years (Qin and Eisner, 2021). Prompt engineering involves the strategic design of input patterns to guide the model toward generating more accurate and relevant responses (Reynolds and McDonell, 2021; Marvin et al., 2023). However, most of these approaches have been conceived to enhance model accuracy, increasing the output length. For instance, Lo (2023) and Strobelt et al. (2022) introduced prompt-based approaches by adding task-specific patterns to frame the input data. While these methods allow boosting accuracy, they can also produce longer outputs due to the additional context and detail introduced by the prompt, making it challenging to provide factual and concise answers (Shi et al., 2023).

Another form of prompt engineering was proposed to improve reasoning within the conclusive

answer. In this context, Chain-of-Thought (CoT) prompting (Wei et al., 2022) is one of the most notable methods, showing significant benefits in QA tasks by requiring the model to provide a step-by-step explanation along with the final response. However, as also highlighted in Section 3, answers generated with CoT tend to be lengthy, hence increasing the generation time (Liu et al., 2018; Takase and Okazaki, 2019).

Given the substantial amount of work focused on improving the accuracy of LLMs, it is not surprising that most of the adopted metrics (Lin, 2004; Stallings and Gillmore, 1971) and benchmarks (Clark et al., 2018; Lin et al., 2021) only address the correctness of the responses, without paying attention to conciseness and response times (Bhargava et al., 2023; Chiang and Lee, 2024). In other tasks too, such as reasoning in control engineering, the focus has been primarily on correctness rather than consistency or conciseness (Kevian et al., 2024). Additionally, several studies have addressed computational cost challenges but not the conciseness. For example, Wang et al. (2024a) evaluated the budget-aware reasoning capabilities of LLMs, while Zheng et al. (2024b) proposed an inference pipeline to improve processing speed. Other works have explored similar optimization approaches, including (Hao et al., 2024) and (Bi et al., 2020). In addition, Chiang and Lee (2024) proposed a benchmark to study LLMs accuracy while incorporating a manual redundancy assessment.

Despite recent advancements, several key aspects remain not sufficiently explored: (i) the impact of concise answers on inference cost and time predictability; (ii) the integration of such aspects into unified metrics that evaluate LLMs not only in terms of correctness but also conciseness; and (iii) an understanding of the analysis of conciseness through the conciseness based on the embeddings content extracted by the generated answers.

This work. To address these challenges, this work introduces novel metrics that jointly account for the conciseness and correctness of generated responses. Additionally, two new scores are proposed to assess conciseness from the embeddings produced by the model. These scores focus on the importance of reasoning steps by analyzing redundancy and information flow. Then, to evaluate the ability of LLMs to control the length of reasoning in their outputs, this work introduces a refined version of

the CoT prompting (Wei et al., 2022), termed Constrained Chain-of-Thought (CCoT). This approach explicitly guides the model to limit the length of its reasoning while preserving the quality of its answers and improving the inference time. This is achieved by improving the conciseness of the responses, which is analyzed and assessed using the proposed scores.

3 Motivations

The output generation time of an LLM depends on various factors, including the model architecture, the pre-and post-processing steps, the answer decoding process, and the question posed, also considering the use of prompt engineering approaches. While the computational cost due to the architecture is well understood, the influence of the other aspects on the overall generation time is less clear and requires further investigation. More formally, an LLM can be represented as a function f that takes as input a prompt x with $\mathcal{N}(x)$ tokens¹ and generates an output $\hat{y} = f(x)$, having $\mathcal{N}(\hat{y})$ tokens, where \mathcal{N} is a length operator that simply counts the number of tokens. The input x can be considered as composed of the original user input x_{us} and a prompt engineering text x_p , depending on the technique used. For instance, in a zero-shot CoT setting, the prompt can be computed as $x = \text{concat}(x_{us}, x_p)$, where x_p is an explicit request for providing reasoning steps in the answer and $\text{concat}(a, b)$ is the concatenation operator that merges two vectors a and b into a single one.

In an encoder-decoder architecture, as the one used by Transformers (Vaswani et al., 2017), let $f_e(x)$ and $f_d(x)$ denote the functions associated with the encoder and the decoder, respectively. Then, the output \hat{y} is a list of tokens $[a^{(1)}, \dots, a^{(\mathcal{N}(\hat{y}))}]$, where each $a^{(i)}$ is computed based on the previously generated tokens and the encoder’s embedding representation $f_e(x)$. That is,

$$a^{(i)} = f_d(f_e(x), [a^{(0)}, \dots, a^{(i-1)}]), \quad i > 0. \quad (1)$$

From Equation (1), it is clear that the larger the set of output tokens in the answer, the higher the time the model takes to generate the answer due to the increased number of times the decoder is invoked. The same consideration could also be

¹Even though ‘tokens’ and ‘words’ refer to different items in the sentence, for simplicity, in this work we will refer to both indistinguishably.