arXiv:2502.07266v3 [cs.AI] 27 May 2025

When More is Less: Understanding Chain-of-Thought Length in LLMs

Yuyang Wu[*]
Peking University

Yifei Wang[*]
MIT

Ziyu Ye
University of Chicago

Tianqi Du
Peking University

Stefanie Jegelka
TUM[†] and MIT[‡]

Yisen Wang[§]
Peking University

Abstract

Large Language Models (LLMs) employ Chain-of-Thought (CoT) reasoning to deconstruct complex problems. While longer CoTs are often presumed superior, this paper challenges that notion, arguing that **longer is not always better**. Drawing on combined evidence from real-world observations, controlled experiments, and theoretical analysis, we demonstrate that task accuracy typically follows an inverted U-shaped curve with CoT length, where performance initially improves but eventually decreases as the number of CoT steps increases. With controlled experiments, we further uncover the **scaling behaviors of the optimal CoT length**: it increases with task difficulty but decreases with model capability, exposing an inherent **simplicity bias** where more capable models favor shorter, more efficient CoT reasoning. This bias is also evident in Reinforcement Learning (RL) training, where models gravitate towards shorter CoTs as their accuracy improves. To have a deep understanding of these dynamics, we establish a simple theoretical model that formally proves these phenomena, including the optimal length's scaling laws and the emergence of simplicity bias during RL. Guided by this framework, we demonstrate significant practical benefits from training with optimally-lengthed CoTs and employing length-aware filtering at inference. These findings offer both a principled understanding of the "overthinking" phenomenon and multiple practical guidelines for CoT calibration, enabling LLMs to achieve optimal reasoning performance with adaptive CoTs tailored to task complexity and model capability.

1 Introduction

> *"Everything should be made as simple as possible, but not simpler."* — Albert Einstein

Large language models (LLMs) have demonstrated impressive capabilities in solving complex reasoning tasks [3, 36]. A key technique for its success is Chain-of-Thought (CoT) reasoning [38]. By generating explicit intermediate reasoning steps, CoT allows models to break down complex problems into simpler, more manageable sub-problems, akin to a divide-and-conquer strategy [44].

A common intuition, supported by some research [12, 20], is that longer and more detailed CoT processes generally lead to better performance, especially for difficult tasks. Meanwhile, recent observations also suggest that concise CoTs can sometimes be effective, albeit with potential performance

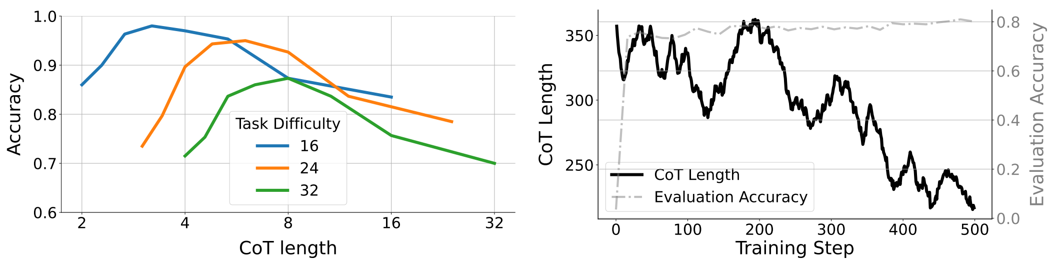
[*] Equal Contribution

[†] School of CIT, MCML, MDSI

[‡] EECS and CSAIL

[§] Corresponding Author: Yisen Wang (yisen.wang@pku.edu.cn)

Preprint. Under review.



(a) Reasoning Accuracy vs. CoT length

(b) Evolution of LLMs' CoT lengths during RL

Figure 1: (a) The performance of a 6-layer GPT2 model (Section 3) follows inverted U-shaped curves on arithmetic tasks at different difficulty levels. As task difficulty increases, the accuracy peak progressively shifts toward longer CoT lengths. (b) As RL training progresses and model accuracy on reasoning tasks improves, the average length of the generated Chain-of-Thought can decrease. This hints at the model learning more efficient, concise reasoning paths (*i.e.*, simplicity bias). We conduct this experiment using Qwen2.5-7B-Instruct trained with GRPO on the LeetCode-2K dataset.

trade-offs on complex problems [26]. This raises a crucial question: does reasoning performance consistently improve as CoTs grow longer and longer?

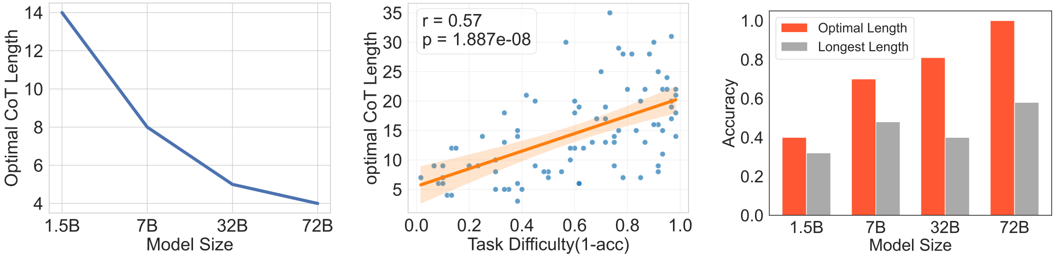
In this paper, through a comprehensive combination of evidence from theoretical analysis, controlled synthetic experiments, and real-world observations, we show that for CoT length, **longer is not always better**. As illustrated by the trend in Figure 1a , when plotting task accuracy against measures related to the CoT length, performance typically follows an **inverted U-shaped** curve. Performance initially improves as the CoT appropriately decomposes the task, but eventually deteriorates if the CoT becomes excessively long (increasing error accumulation) or too short (steps are too complex). This indicates the existence of an **optimal CoT length** that balances these competing factors.

Further, we discover scaling behaviors of this optimal CoT length with respect to model capability and task difficulty: harder tasks tend to have longer optimal CoTs, while more capable models often achieve peak performance with shorter optimal CoTs. This latter point interestingly implies an inherent **simplicity bias** in LLM reasoning, where models favor shorter, more efficient reasoning paths as their capabilities increase. Moreover, we observe this simplicity bias during LLMs' reinforcement learning (RL) training. As shown in Figure 1b, RL-trained models exhibit a gradual shift towards using shorter CoTs compared to the base model, indicating an acquired preference for shorter CoTs as a result of the simplicity bias of optimal CoT length. This surprising phenomenon parallels humans' natural preference for simplest possible reasoning processes, as evident in Einstein's quote.

To gain a deeper understanding of the rise of optimal CoT length and its simplicity bias, we focus on a controlled study using a synthetic arithmetic task that allows us to ablate nuanced factors present in practical LLM training. In this controlled setting, we not only successfully replicate these phenomena but also theoretically derive the existence of the optimal CoT length and its scaling behaviors with respect to task complexity and model capability. Intuitively, task decomposition into more steps yields easier subtask but also accumulate errors exponentially, leading to an optimal tradeoff at an intermediate CoT length. Notably, this theory also explains the emergence of the simplicity bias as observed during RL training. Thus, although simple, our theory provides valuable characterization of LLMs' behaviors during CoT. Translating this understanding into practice, we show significant benefits from training with optimally-lengthed CoTs and employing *Length-aware Vote* to filter out excessively long CoTs at inference.

To summarize, this paper makes the following main contributions:

- We demonstrate the existence of an optimal CoT length and the simplicity bias of CoT on both real-world LLMs (Section 2) and synthetic arithmetic experiments (Section 3).
- We establish a theoretical model of CoT that allows to formally characterize and prove the existence of an optimal CoT length as well as its scaling laws and simplicity bias (Section 4).
- We explore the implications of these findings, showing how training with optimal-length CoT data can significantly boost performance, and how filtering excessively long CoTs with entropy measures can benefit reasoning performance at inference (Section 5).



(a) Optimal CoT length vs. Model size (Qwen2.5 series).

(b) Optimal CoT length vs. Task difficulty (with the 1.5B model).

(c) Optimal vs. Longest CoT length accuracy on MATH Level 5.

Figure 2: Real-world CoT length observations. (a) Larger models tend to achieve optimal performance with shorter CoTs. (b) More difficult tasks (as measured by lower accuracy on the x-axis) tend to require longer optimal CoTs (with a positive correlation of significance $p \ll 0.05$). (c) Accuracy for CoTs of optimal length is significantly higher than that of the longest CoTs.

Our findings offer a fresh perspective for calibrating CoT generation, moving beyond the assumption that longer is always better. By understanding and adapting to the optimal CoT length, we can develop LLMs that reason more effectively, avoiding both underthinking and counterproductive overthinking.

2 Optimal CoT Length and Simplicity Bias in Real-World LLMs

To ground our investigation in practical scenarios, we first explore the relationship between CoT length and reasoning performance using publicly available LLMs.

2.1 Scaling Behaviors of Optimal CoT Length in Real-World LLMs

Setup. To assess how model capability interacts with CoT length. We evaluate Qwen2.5 series of Instruct models [27] on Level 5 questions in MATH dataset composed of challenging competition mathematics problems [18]. For each question, we generate 60 solutions with as much variation in length as possible. The CoT length is determined by the number of intermediate reasoning steps generated by the model. The optimal CoT length is the one that yields the highest average accuracy. See Appendix C for additional experiments (MMLU STEM dataset [17], different models) and implementation details on step segmentation and solution length control.

Optimal Length Decreases with Stronger Model Capabilities: For each model, we randomly select 30 questions since our focus lies in exploring different lengths of solutions for the same problem rather than evaluating the whole dataset. As depicted in Figure 2a, there is a clear trend where the optimal CoT length decreases as the model size increases. For instance, the optimal length shifts from 14 steps for the 1.5B parameter model to 4 steps for the 72B parameter model. This suggests that more capable models can consolidate reasoning into fewer, more potent steps, aligning with the Simplicity Bias concept where stronger models prefer shorter effective paths.

Optimal Length Grows with Harder Tasks: We also investigate how task difficulty influences the optimal CoT length. For this, we consider 100 randomly selected questions and compute the accuracy of an LLM on each question from 60 sampled solutions. We use (1 - accuracy) on these questions as a proxy for the difficulty. Figure 2b shows a statistically significant positive correlation (notably $p = 1 \times 10^{-8} \ll 0.05$) between task difficulty and the optimal CoT length of Qwen1.5B-Instruct model. This indicates that more challenging problems will significantly benefit from a longer CoT with more extended decomposition steps. Similar trends for other models are provided in Appendix C.2.

Excessively Long CoTs Lead to Significant Degradation: The above scaling behaviors of CoT suggest that one should adaptively select the optimal CoT length w.r.t. the given model and task. Here, we illustrate the significance of this choice by compare the performance of using the optimal and the longest CoT lengths. As shown in Figure 2c, there is a large gap between the two that grows larger as the models become more capable. For a 72B model, the gap can be as large as 40% accuracy, showing great potential gains of adapting CoT length to attain optimal reasoning performance.