Proposition 4.2 establishes the quantitative relationship between CoT's final performance A and reasoning length N. Once we obtain estimates for $E(N, M, T)$ and $\sigma(T)$, we can determine the optimal CoT length N^* as a function of the model capability M and task complexity T.

Simple Case with Linear Error. To gain intuition, we first consider a simple case where the sub-question error scales linearly with T, i.e., $\sigma(T) = T/C$, where C is a constant representing the maximum task difficulty models are trained to handle. Throughout this analysis, we assume $\sigma(T) \leq 0.9$ to restrict our discussion to tasks that are within the model's training regime. Otherwise, it would be unreasonable to claim the model has learned to solve such problems. We also assume the sub-answer error rate scales linearly with harder tasks, fewer steps, and weaker models as $E(N, M, T) = (T/N)/M = T/(NM)$. Under these simplified conditions, we can derive the following closed-form expression for the optimal CoT length.

Theorem 4.3 (Optimal CoT Length)**.** *For a given model capability M and task difficulty T, the total accuracy $A(N) = \alpha[(1 - T/C) \cdot (1 - T/(NM))]^N$ (Eq. (1)) initially increases and then decreases as N increases (forming an inverted U-shape). Thus, there exists an optimal CoT length:*

$$N^*(M, T) = \frac{TZ}{M(Z + 1)}, \tag{2}$$

that maximizes $A(N)$, where $Z = W_{-1}(-1 - T/(Ce))$, and $W_{-1}(x)$ is the smaller real branch of the Lambert W function satisfying $we^w = x$, and e is the natural number.

This theorem formally establishes the inverted U-shaped curve and provides an explicit form for N^*. From this, we can formally prove the first three scaling behaviors characterized in Section 3.2.

Corollary 4.4 (Scaling laws of Optimal CoT Length)**.** *Based on Theorem 4.3, one can derive:*

- *$N^*(M, T)$ increases monotonically with T, i.e., harder tasks require more reasoning steps to attain the optimal performance.*
- *The optimal number of operators per step $t^* = T/N^*(M, T) = M(1 + 1/Z)$ increases monotonically with T. This aligns with the envelope curve result (Figure 3a).*
- *$N^*(M, T)$ decreases monotonically with M, i.e., more capable models require fewer reasoning steps to attain the optimal performance, reflecting the simplicity bias.*

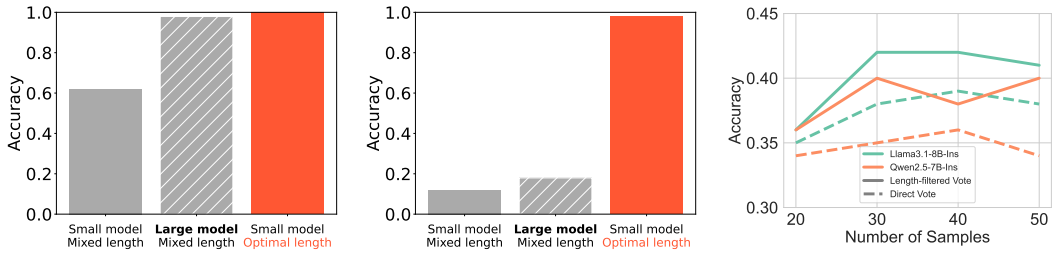
Extension to Broader Scenarios. Here, we adopt a simple linear model to facilitate intuitive understanding. However, this analysis can be extended to more general settings, including general error functions (*only* with mild assumptions of monotonicity and convexity) and stochastic error models, where *each subtask may exhibit a different error rate*. These extensions introduce additional technical subtleties but follow the same underlying principles. We defer this part to Appendix F.

4.2 Why does RL Exhibit Simplicity Bias?

The analysis above also provides a natural understanding of RL's simplicity bias (Section 3.2).As in the arithmetic task, we generate samples within a finite discrete action space $\mathcal{A} = \{N_1, N_2, \ldots, N_k\}$ during RL that receive binary outcome rewards. This reduces to a stateless bandit: each N_i yields reward $r \in \{0, 1\}$ with probability $A(N_i)$ (from Proposition 4.2). Let us parameterize a softmax policy $\pi_\theta(N_i) = \frac{e^{\theta_i}}{\sum_j e^{\theta_j}}$, and define the RL objective as $J(\theta) = \sum_{i=1}^k \pi_\theta(N_i)\, A(N_i)$. As a result, the policy-gradient becomes $\nabla_{\theta_i} J = \sum_{j=1}^k A(N_j)\, \pi_\theta(N_j)\big(\delta_{ij} - \pi_\theta(N_i)\big)$.

Corollary 4.5 (RL Converges to Optimal CoT Length)**.** *For gradient ascent on $J(\theta)$ with sufficiently small step size, the policy converges to a deterministic policy $\pi_\theta(N_i) = 1$ iff $i = \arg\max_j A(N_j)$. Thus, RL training converges to the optimal CoT length $N^* = \arg\max_{N \in \mathcal{A}} A(N)$.*

This corollary shows that RL will automatically discover the optimal length (usually shorter length) through optimizing the reward function and exhibit a decreasing CoT length as in the simplicity bias phenomenon. In this way, our theory offers an explanation of the optimal CoT length, its scaling behavior and RL's simplicity bias within a unified framework.



(a) Influence of CoT Data ($T = 32$) (b) Influence of CoT Data ($T = 64$) (c) Length-Filtered Vote

Figure 4: (a) and (b) compare model performance under different pretraining data distributions: Mixed Length (uniform over all lengths) vs. Optimal Length (only optimal-length solutions). Despite its smaller size, the small (6 layer) model trained on optimal-length data outperforms the large (9 layer) model trained on mixed-length data, with the performance gap widening as task difficulty increases. (c) Our Length-Filtered Vote method consistently outperforms vanilla majority vote on the GPQA dataset, maintaining strong performance even as the number of samples increases.

5 Practical Applications of Optimal CoT Length

Guided by the understanding above, in this section, we illustrate via some proof-of-concept experiments that adapting LLM training and inference configurations to the optimal CoT length can improve the model's reasoning performance.

5.1 Training with Data of Optimal CoT Length

Training with Optimal-Length CoT Data: The existence of an adaptive, optimal CoT length suggests that one should design the CoT training data adaptively to fully optimize the model's reasoning performance. To examine the influence of the CoT length of the training data, we train a model on a specialized dataset that contains CoT solutions with lengths known to be optimal for the given model size and task difficulty (T). We compare this model against a baseline model trained on a dataset of CoT solutions with uniformly distributed step lengths t. During testing, models were allowed to freely choose their CoT strategy.

Results. As shown in Figures 4a and 4b, the model trained on optimal-length CoTs significantly outperforms the models trained on mixed-length solutions. Remarkably, a smaller model (e.g., 6 layers) trained on optimal-length data can even outperform a larger model (e.g., 9 layers) trained on randomly chosen CoT lengths. This proof-of-concept experiment underscores the critical influence of the suitability of the CoT length in training data for the model. While it is generally hard to exactly estimate optimal CoT lengths in real-world problems, our theoretical and empirical studies provide valuable guidelines for a coarse estimate. We leave more in-depth studies to future work.

5.2 Adaptive Length-Filtered Vote at Inference Time

The observation that CoTs of optimal length yield higher accuracy suggests that inference-time strategies could benefit from this insight. Standard approaches like majority voting over multiple sampled CoTs, such as self-consistency [37], treat all valid reasoning paths equally, regardless of their length. However, paths that are too short (underthinking) or too long (overthinking and error-prone) may contribute noisy or incorrect answers to the voting pool.

Inspired by our findings, we propose **Length-Filtered Vote**, an adaptive method that enhances standard majority voting by preferentially weighting or exclusively considering answers derived from CoTs whose lengths fall within a proper range. Specifically, in majority vote, given a model f_θ, a question q, a ground truth answer a^*, we first sample a set of answer candidates $c_1, \ldots, c_n \overset{i.i.d.}{\sim} f_\theta(q)$ independently. After that, instead of a direct vote, we group the answers by their corresponding CoT length $\ell(c_i)$ (discussed in Appendix C) into groups with equal bin size D (by default, we set $D = 2$), denoted as $\{L_j\}_{j=1}^m$. As our theory suggests that the prediction accuracy is peaked around a certain range of CoT length, we identify such groups through the prediction uncertainty of the answers within each group, based on the intuition that lower uncertainty implies better predictions. Specifically, we

calculate the Shannon entropy $H(L_i)$ of the final answers given by the CoT chains in each group L_i. We use a majority vote only for the K (by default, we set $K = 3$) groups with the smallest entropy. A detailed description of the algorithm is in Appendix H.

Results. We evaluate the proposed method against vanilla majority vote (i.e., self-consistency [37]) on a randomly chosen subset of 100 questions from the GPQA dataset [29], a more challenging collection of multiple-choice questions. The results in Figure 4c show that our filtered vote consistently outperforms vanilla majority vote at different sample numbers and shows little performance degradation as the sample number increases. This further underlines the importance of considering CoT length in the reasoning process.

6 Related Work

Chain-of-Thought for LLM Reasoning. CoT has become a core technique for LLMs to solve complex reasoning tasks by generating intermediate steps [38]. Numerous variants arise to enhance CoT reasoning with more structural substeps, such as least-to-most prompting [45], tree of Thoughts [42], and divide-and-conquer methods [44, 25]. These methods fundamentally treat CoT as a framework for task decomposition and subtask solving that falls in our analysis in Section 4.

Overthinking in CoT Reasoning. With the rise of powerful reasoning models like OpenAI o1, scaling test-time compute with long CoT has gained prominence [33, 7, 39, 2]. These studies often suggest that more computation like longer CoT can lead to better results. However, this is not always true. With similar interests as ours, a few concurrent works also investigated the "overthinking" phenomenon [6] where reasoning models generate excessively long CoTs for simple problems and proposed some mitigation strategies [16, 23, 24, 34]. Our analysis not only reveals the inverted-U curve of CoT length and the existence of optimal CoT length, but also provides a in-depth understanding on the scaling behaviors and simplicity bias of the optimal CoT length, as supported by both controlled experiments and theoretical analysis. This establishes a systematic explanation of overthinking and points out principled guidelines for better CoT designs.

Simplicity Bias and Occam's Razor in Machine Learning. The simplicity bias of CoT identified in our work resonates with broader principles like Occam's Razor, which favors simpler explanations or models. In machine learning, a 'simplicity bias' often refers to neural networks learning simpler functions first or being biased towards solutions with lower intrinsic complexity [1, 19]. Our findings extend this understanding to the realm of generated reasoning paths: we reveal that even structured, multi-step reasoning processes like CoT, as produced by LLMs, exhibit such simplicity bias by favoring concise reasoning paths, particularly as model capability increases.

Theoretical Understanding of CoT. Numerous studies aim to theoretically formalize the Chain-of-Thought (CoT) process and understand its effectiveness. They include analyzing CoT's computational advantages via circuit complexity [11, 22], demonstrating how coherent reasoning paths enhance error correction and accuracy [10], and quantifying step-wise information gain from an information-theoretic standpoint [35]. Further research has shown that detailed CoT improves learning stability by affecting gradient dynamics [21], while controlled synthetic experiments have helped uncover underlying problem-solving mechanisms in LLMs [43]. Distinct from these varied theoretical explorations, our theory characterizes how CoT length influences final performance and explains its scaling behaviors through the interplay of task decomposition and error accumulation. Furthermore, our findings on CoT scaling behaviors and the consequent need for model-specific CoT structures (as discussed in Section 3.2) resonate with the concept of algorithmic alignment [41], which suggests that models perform best when the problem structure aligns with their computation structure.

7 Conclusion

In this paper, we challenged the notion that longer Chain-of-Thought (CoT) processes are invariably superior, demonstrating through extensive experiments and theoretical analysis that CoT length and accuracy typically follow an inverted U-shaped curve, implying an optimal length that balances task decomposition against error accumulation. We discovered the simplicity bias of CoT, where more capable models prefer shorter effective reasoning paths, and formally derived scaling laws for this optimal length relative to model capability and task difficulty. Practically, we showed that reinforcement learning can guide models towards this optimal CoT length, that training on