D Supplementary Details on Real World Experiment for RL Simplicity Bias

For Figure 1b, we use Qwen2.5-7B-Instruct [27] as the base model, Group Relative Policy Optimization with R1-like prompting [31, 15] for the reinforcement learning process, and LeetCode-2K [40] as the training and evaluation dataset. We take the following training configuration by default:

Table 1: Hyperparameter settings for real-world RL experiments with Qwen2.5-instruct models.

Learning Rate	Max Epochs	Rollout Samples	Reverse KL Coefficient	Entropy Loss Coefficient	Effective Batch Size
5e-7	10	16	1e-3	5e-3	256

E Additional Synthetic Experiment Details

E.1 Training details

In default, we train different models (layers ranging from 5 to 9) on the same dataset, which included mixed questions with total operators $T \in [12, 80]$ and random sampled CoT solutions with each step operators $t \in [1, 12]$. All other parameters are kept the same with the huggingface GPT-2 model. During the training process, the CoT indicator token `<t>` is also trained, so that during test-time, we can let the model decide which type of CoT it will use by only prompting the model with the question. For each model, we train 25000 iterations with batch size that equals 256. During test-time, we test 100 questions for each T and t. All experiments can be conducted on one NVIDIA A800 80G GPU.

E.2 Observation of subtask loss

As we observed in training losses, the loss of subtask generation tokens (e.g. $1 + 2$) for the easiest subtask($t = 1$) is about 3 times larger than the hardest subtask ($t = 12$), while the loss ratio for subtask answer tokens is $1e4$. Therefore, it is acceptable for taking the subtask error rate constant with t.

Besides, there is no obvious pattern showing the model sizes affect the subtask loss. Moreover, the smallest model and the largest model have almost the same subtask loss. Therefore, in our settings, we take model size as irrelevant with the subtask error rate.

F Theoretical Results under Broader Scenarios

F.1 General Errors

In the simple case we discussed in Section 4.1, we discussed the trend of overall accuracy with respect to N and the variation of optimal N with M and T, assuming the subtask error rate is a linear function. In the following discussion, we aim to derive conclusions corresponding to more general error rate functions. We find that as long as the error function satisfies some basic assumptions on the **monotonicity** and **convexity** of the error functions, the above conclusions still hold.

Assumption F.1. $E(N, M, T)$ satisfies the following reasonable conditions:

- $0 < E(N = 1, M, T) < 1$
- $\lim_{N \to +\infty} E(N, M, T) = 0$
- $E(N, M, T)$ is monotonically deceasing with N, since more detailed decomposition leads to easier subtask.
- $E(N, M, T)$ is convex with N, since the benefits of further decomposing an already fine-grained problem(N is large) are less than the benefits of decomposing a problem that has not yet been fully broken down(N is small).
- $E(N, M, T)$ is monotonically deceasing with M, since stronger models have less subtask error rate.
- $E(N, M, T)$ is monotonically increasing with T, since harder total task leads to harder subtask while N, M are the same.

Assumption F.2. $\sigma(T)$ is monotonically increasing with T

With Assumption F.1 and F.2), the core insights from the linear case can be generalized.

Theorem F.3. *For a noise function $0 < \sigma(T) < 1$ and a subtask error rate function $0 < E(N, M, T) < 1$ satisfying Assumptions F.1 and F.2, the general final accuracy function $A(N)$ from Proposition 4.2 has the following properties:*

- $\lim_{N \to +\infty} A(N) = 0$. *(Excessively long chains always fail.)*
- *If $A(N)$ has a maximum at $N^* > 1$, then N^* has a lower bound related to M and T:*

$$N^* \geq N_{LB}(M, T) = E_N^{-1}\left(1 - \frac{1}{e^2(1 - \sigma(T))}; M, T\right), \tag{6}$$

where $E_N^{-1}(\cdot; M, T)$ is the inverse of $E(N, M, T)$ with respect to N.

The monotonicity of E_N^{-1} with respect to M (decreasing) and T (increasing, assuming $\sigma(T)$ doesn't dominate adversely) implies that the qualitative scaling laws (Corollaries stemming from Theorem 4.3) still hold under general conditions, supporting the empirically observed Simplicity Bias and the inverted U-shaped performance.

Corollary F.4. *As the model becomes stronger, E^{-1} decreases monotonically with respect to M, which leads to a decrease of $N(M, T)$.*

Corollary F.5. *As the task becomes harder, E^{-1} is monotonically increasing with respect to T, which leads to an increase in $N(M, T)$.*

F.2 Random Error

In Theorem 4.3 and F.3, we make a strong assumption that all sub-question or sub-answer errors are identical, which does not align well with real-world scenarios. In practice, each sub-task may exhibit a different error rate. However, they generally follow a trade-off: the more the task is decomposed, the easier each sub-task becomes. Specifically, we can model the error rate of each sub-task as a random variable with a fixed expectation that monotonically decreases with the number of CoT steps N.

To simplify the problem, here we assume $\sigma_i \sim B(\alpha_1(T), \beta_1(T))$ to be the sub-question error rate, and $e_i \sim B(\alpha_2(N, M, T), \beta_2(N, M, T))$ to be the sub-answer error rate. Then, as a variant of Proposition 4.2, the expectation of final accuracy is $\mathbb{E}\left[\prod_{i=1}^{N}(1 - e_i)(1 - \sigma_i)\right]$.

It is worth noting that each σ_i or e_i is not independent. If most steps are easy (i.e., have low error rates), the remaining steps are more likely to be easy as well. Moreover, if a particular step serves as a self-validation step, its high accuracy can influence the correctness of other steps that depend on it. This also provides an interpretation for reasoning models exhibiting backtracking behavior.

Theorem F.6. *Let $\alpha_1 = T$, $\beta_1 = C - T$, $\alpha_2 = T$, and $\beta_2 = NM - T$. Then the expected error rates for sub-questions and sub-answers are given by $\mathbb{E}[\sigma_i] = \frac{T}{C}$ and $\mathbb{E}[e_i] = \frac{T}{MN}$, respectively. Based on these estimates, we can derive an upper bound $\hat{A}(N)$ on the final accuracy*

$$\mathbb{E}\left[\prod_{i=1}^{N}(1 - e_i)(1 - \sigma_i)\right] \leq \hat{A}(N) = \left[\left(1 - \frac{T}{C + 2N - 1}\right)\left(1 - \frac{T}{NM + 2N - 1}\right)\right]^N,$$

which initially increases and then decreases as the number of CoT steps N grows.

This suggests that even with stochasticity, the fundamental trade-off leading to an optimal CoT length persists.

G Proof

In this section, we provide the proofs for all theorems.

G.1 Proof of Proposition 4.2

Proposition 4.2. *The total accuracy of N-step reasoning is*

$$A(N) = P(a_{\mathit{final}} = a^*_{\mathit{final}} | q, \theta, N) = \alpha \left((1 - E(N, M, T))(1 - \sigma(T)) \right)^N, \tag{1}$$

where α denotes a constant value independent of N.

Proof. In each subtask t_i, which contains t operators, there are $2t + 1$ tokens (as the number of numerical tokens is one more than the number of operators). Therefore, the accuracy of each subtask is given by

$$P(t_i = t^*_i | H_{i-1}, q, \theta) = (1 - \sigma(T))^{2t+1}. \tag{7}$$

In our theoretical analysis, for simplicity, we allow t to be a fraction, defined as $t = \frac{T}{N}$, and assume that each subtask has the same level of difficulty given T and N. Under this assumption, we have the final accuracy:

$$\begin{aligned} A(N) &= P(a_N = a^*_N | q, \theta) & (8) \\ &= \prod_{i=1}^{N} P(t_i = t^*_i | H_{i-1}, q, \theta) P(a_i = a^*_i | t_i, H_{i-1}, q, \theta) & (9) \\ &= \prod_{i=1}^{N} (1 - \sigma(T))^{2t+1} (1 - E(N, M, T)) & (10) \\ &= (1 - \sigma(T))^{N(2t+1)} (1 - E(N, M, T))^N & (11) \\ &= (1 - \sigma(T))^{2T} \left((1 - E(N, M, T))(1 - \sigma(T)) \right)^N & (12) \\ &= \alpha \left((1 - E(N, M, T))(1 - \sigma(T)) \right)^N & (13) \end{aligned}$$

□

G.2 Proof of Theorem 4.3

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.

Proof. Given Eq. (1) that

$$A(N) = \alpha \left(\left(1 - \frac{T}{C} \right) \left(1 - \frac{T}{NM} \right) \right)^N \tag{14}$$

We consider function

$$f(x) \; = \; \Big[\big(1 - \tfrac{T}{Mx}\big) \big(1 - \tfrac{T}{C}\big) \Big]^x. \tag{15}$$

For convenience, define

$$g(x) \; = \; \ln\big(f(x)\big) \; = \; x \; \ln\Big[\big(1 - \tfrac{T}{Mx}\big) \big(1 - \tfrac{T}{C}\big) \Big].$$

Thus,

$$g'(x) = \Big[\ln\big(1 - \tfrac{T}{Mx}\big) \; + \; \frac{T}{Mx\,\big(1 - \tfrac{T}{Mx}\big)} \Big] \; + \; \ln\big(1 - \tfrac{T}{C}\big).$$