A Theoretical Understanding of Chain-of-Thought: Coherent Reasoning and Error-Aware Demonstration

Yingqian Cui
Michigan State University

Pengfei He
Michigan State University

Xianfeng Tang
Amazon

Qi He
Amazon

Chen Luo
Amazon

Jiliang Tang
Michigan State University

Yue Xing
Michigan State University

Abstract

Few-shot Chain-of-Thought (CoT) prompting has demonstrated strong performance in improving the reasoning capabilities of large language models (LLMs). While theoretical investigations have been conducted to understand CoT, the underlying transformer used in these studies isolates the CoT reasoning process into separated in-context learning steps (Stepwise ICL). In this work, we theoretically show that, compared to Stepwise ICL, the transformer gains better error correction ability and more accurate predictions if the reasoning from earlier steps (Coherent CoT) is integrated. Given that this coherent reasoning changes the behavior of the transformer, we further investigate the sensitivity of the transformer with Coherent CoT when the demonstration examples are corrupted at the inference stage. Our theoretical results indicate that the transformer is more sensitive to errors in intermediate reasoning steps than the final outcome. Building upon this observation, we propose an improvement on CoT by incorporating both correct and incorrect reasoning paths in the demonstration. Our experiments validate the effectiveness of the proposed approach.

1 INTRODUCTION

Few-shot Chain-of-Thought (CoT) prompting has emerged as a highly effective technique to enhance the reasoning capabilities of large language models (LLMs) (Wei et al., 2022). Given a few examples of step-by-step reasoning at the *inference stage*, the model can generalize the reasoning process to new tasks, demonstrating significant improvement in solving complex problems, particularly in mathematical reasoning and commonsense inference (Wei et al., 2022; Fu et al., 2023; Lyu et al., 2023).

Aside from these empirical successes, recent efforts have provided valuable theoretical insights into CoT. In the literature, two main approaches are typically used to analyze the ability of transformers in CoT tasks (and more broadly, ICL). In the first approach, to show the existence of transformers that are capable of performing CoT, people explicitly construct a transformer by specifying its parameters and assigning each component to perform a specific task (Feng et al., 2023). The second approach defines a specific data format to organize different reasoning steps and then trains a transformer model to learn a step-by-step prediction process. Subsequent analysis focuses on the properties of this trained model. Using these approaches, previous studies are able to explain the expressiveness power of the transformer in CoT tasks and understand how different components of the model contribute to the multi-step reasoning (Li et al., 2023, 2024b; Tutunov et al., 2023).

However, a key limitation of these analyses is that they often overlook the connections among multiple reasoning steps. In the approach where a transformer is directly constructed to perform CoT, predictions at each reasoning step are based solely on the result from the previous step (Feng et al., 2023). Similarly, when training a transformer from scratch on CoT tasks, to simplify the analysis, a "filtering" process is used to retain only the information from the most recent step (Li et al., 2023). In summary, when constructing/training the transformer from either way

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

mentioned above, the prediction process for a training sample (a CoT prompt) involves multiple reasoning steps, and each step only focuses on the immediate task at hand without incorporating earlier reasoning steps. We refer to this prediction process as "Stepwise ICL" (formally defined in Definition 3.1).

However, in real-world scenarios, e.g., next token prediction, the LLM takes into account all the previous tokens in the context window, rather than treating each step in isolation (referred to as "Coherent CoT", defined in Definition 3.2).[1] Observing the discrepancy between Stepwise ICL and Coherent CoT, we extend the theoretical framework in Zhang et al. (2023) and study the properties of a model trained using Coherent CoT and compare it with the model trained with Stepwise ICL. Based on our result, using Coherent CoT instead of Stepwise ICL during the *training stage* provides better prediction performance. Intuitively, when treating CoT as a holistic process – where later steps integrate the reasoning from earlier steps – the transformer will consider the potential errors in previous predictions and adjust subsequent predictions accordingly, which provides a form of self-correction and enhances the prediction performance.

While Coherent CoT potentially outperforms Stepwise ICL as elaborated in our first contribution, it requires training the model on the entire reasoning chain, which alters the model's optimal parameters and changes the behavior of the model. This change introduces uncertainty about which steps in the reasoning chain are most sensitive to errors, highlighting the necessity for sensitivity analysis. This leads to our **second contribution**, which focuses on the sensitivity of the trained Coherent CoT model to perturbations in the demonstration examples at the *inference stage*. To quantify the sensitivity, we examine how the Coherent CoT model reacts with random perturbations at different reasoning steps. We reveal that, during inference, the Coherent CoT model is more sensitive to noise in the intermediate reasoning steps of the demonstration examples than to inaccuracies in their final outcomes.

Inspired by the sensitivity analysis, our **third contribution** is to propose a prompt composing method to enhance CoT performance at the *inference stage*. Based on our sensitivity result, CoT is more sensitive to possible incorrectness at the intermediate reasoning steps, thus improving the accuracy of these steps can better enhance the overall CoT performance. We propose to incorporate both correct and incorrect reasoning paths in the demonstrations to enhance the accuracy of the intermediate reasoning steps, which we define as "Error-Aware Demonstration". We provide theoretical results to demonstrate the effectiveness of Error-Aware Demonstration in improving the final prediction of CoT and validate it through experiments on real text data.

[1] A real-life example can be found in Appendix A.

2 RELATED WORKS

Empirical Insights in CoT. Chain-of-Thought (CoT) prompting, introduced by Wei et al. (2022), has proven to be highly effective in enhancing the reasoning capabilities of LLMs by breaking complex tasks into step-by-step processes. This technique has been expanded into various variants and extensions, including Zero-shot CoT (Kojima et al., 2022), Self-Consistency (Wang et al., 2022b), Auto-CoT (Zhang et al., 2022), Tree-of-Thought (Yao et al., 2023), and Graph-of-Thought (Besta et al., 2023), to further improve efficiency or model performance.

Building upon the success of CoT, recent studies seek to deepen the understanding behind CoT by empirically exploring its mechanisms. For example, Wu et al. (2023) finds that CoT enables the model to maintain the attention robust and stable on the semantically relevant tokens in the prompt. Madaan and Yazdanbakhsh (2022) defines the key components of a prompt as symbols, patterns, and text, and investigates how each of these elements and their interaction contribute to the superior performance of CoT. Additionally, Wang et al. (2022a) validates that the relevance of the demonstration example to the query and the correct ordering of reasoning steps are key factors for the effectiveness of CoT, and Jin et al. (2024) examines the relationship between CoT's effectiveness and reasoning step length. Based on Jin et al. (2024), simpler tasks require fewer reasoning steps, while more complex tasks benefit greatly from more detailed inference sequences.

Theories in ICL and CoT. The mechanism of ICL has been extensively studied in the theoretical literature. For example, studies such as Von Oswald et al. (2023); Ahn et al. (2023); Akyürek et al. (2022); Zhang et al. (2023); Huang et al. (2023) have explained how ICL learns to perform linear regression using gradient descent. The work by Cheng et al. (2023) extends these analyses by investigating how transformers can apply ICL to non-linear functions, while Bai et al. (2023) focuses on generalized linear models, ridge regression, and LASSO. Some other works study ICL from a Bayesian perspective (He et al., 2024). Additionally, Cui et al. (2024); Chen et al. (2024) explains why multi-head attention is preferred than single-head attention when performing ICL.

Besides ICL, recent research also starts to establish theoretical frameworks to understand CoT. For example, Li et al. (2023) offers a specific framework that views CoT as a series of ICL components, each addressing a smaller subproblem. Feng et al. (2023) constructs a transformer that solves arithmetic and linear equation tasks using CoT. Prystawski et al. (2024) offers a Bayesian perspective on how intermediate steps improve reasoning. Additionally, Li et al. (2024b) provides a TC° upper bound for the expressiveness of constant-precision transformer, and Tutunov et al. (2023) introduces a two-level hierarchical graphical model to explain how LLMs generate sequences of reasoning steps. Li et al. (2024a) analyzes generalization capability of CoT considering nonlinear attentions, and Wen et al. (2024) suggests that a key factor in the improvement brought by CoT is the sparsity in attention layers.

3 THEORETICAL RESULTS

We extend the existing theory of ICL in transformers, e.g., Zhang et al. (2023), to a CoT scenario to conduct our theoretical investigation. Briefly speaking, there are two key observations: (1) Compared to Stepwise ICL, using Coherent CoT at the training stage results in a model with better inference performance. (2) When noises exist in the demonstration examples at the inference stage, the model with Coherent CoT is more sensitive to perturbations in the reasoning steps than the inaccuracies in the final response.

In the following, we introduce some setups in Section 3.1 for data generation and the transformer architecture, then present the theoretical results for (1) and (2) respectively in Section 3.2 and 3.3.

3.1 Model Setup

Data generation process. Since demonstration examples are needed in the prompt in CoT, we define how the examples as well as the query data are generated as follows.

Assumption 3.1 (Data Generation Process) *In each prompt, the examples (x_i, z_i, y_i) and the query data (x_q, z_q, y_q) are i.i.d sampled from the following "two-layer" noisy regression:*

- *The independent variable $x \in \mathbb{R}^d \in N(0, I_d)$.*
- *The intermediate response $z = \beta^\top x$.*
- *The final response $y = z + \epsilon$, $\epsilon \in N(0, \sigma^2)$.*

For each prompt, all the examples and the query data share the same β. In different prompts, β is i.i.d. uniformly sampled from unit sphere, i.e., $\|\beta\|_2^2 = 1$.

Assumption 3.1 is primarily based on Li et al. (2023) with some modifications about the relation between y and z. Following Li et al. (2023), we assume x follows a Gaussian distribution to simplify the analysis. In terms of the final response y, we assume $y = \beta^\top x + \epsilon$, a linear mapping of x with added noise. To improve the prediction of y, an intermediate response $z = \beta^\top x$ is introduced. As a noise-free representation of the relationship between x and y, introducing z helps to mitigate the impact of noise ϵ and guide a more accurate prediction of y. An additional discussion on potential relaxations of the assumptions can be found in Remark 3.2.

Coherent CoT and Stepwise ICL. We define "Coherent CoT" and "Stepwise ICL" as follows:

Definition 3.1 (Stepwise ICL) *There are two separate ICL steps involved in the Stepwise ICL process, and each is performed by a different model. First, the intermediate response $\hat{z}_q$ is predicted using the model f_1 (to be defined later) as $\hat{z}_q = f_1(E^{(1)}_{ICL})_{d+1,D+1}$, where the input prompt is formatted as*

$$E^{(1)}_{ICL} = \begin{pmatrix} x_1 & x_2 & \dots & x_D & x_q \\ z_1 & z_2 & \dots & z_D & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(D+1)}. \tag{1}$$

Obtaining $\hat{z}_q$, the input prompt of the second step is

$$E^{(2)}_{ICL} = \begin{pmatrix} z_1 & z_2 & \dots & z_D & \hat{z}_q \\ y_1 & y_2 & \dots & y_D & 0 \end{pmatrix} \in \mathbb{R}^{2\times(D+1)}. \tag{2}$$

The final prediction is obtained using the model f_2 as $\hat{y}_q = f_2(E^{(2)}_{ICL})_{d+1,D+1}$.

Definition 3.2 (Coherent CoT) *An Coherent CoT process involves two steps. In the first step, the input prompt is formatted as:*

$$E^{(1)}_{CoT} = \begin{pmatrix} x_1 & x_2 & \dots & x_D & x_q \\ z_1 & z_2 & \dots & z_D & 0 \\ y_1 & y_2 & \dots & y_D & 0 \end{pmatrix} \in \mathbb{R}^{(d+2)\times(D+1)}. \tag{3}$$

The intermediate response $\hat{z}_q$ is predicted as $\hat{z}_q = f(E^{(1)}_{CoT})_{d+1,D+1}$. After $\hat{z}_q$ is obtained, it is plugged into the input prompt, forming:

$$E^{(2)}_{CoT} = \begin{pmatrix} x_1 & x_2 & \dots & x_D & x_q \\ z_1 & z_2 & \dots & z_D & \hat{z}_q \\ y_1 & y_2 & \dots & y_D & 0 \end{pmatrix} \in \mathbb{R}^{(d+2)\times(D+1)}. \tag{4}$$

The final prediction of the query input is made by $\hat{y}_q = f(E^{(2)}_{CoT})_{d+2,D+1}$ using the same transformer.

For both Stepwise ICL and Coherent CoT, we consider that the input prompt follows a structured data format. While some existing literature attempts to relax the assumption on the restrictive structured data format condition (e.g., Xing et al. (2024)), it is observed that the performance of ICL with structured data serves as a lower bound for the best possible performance.

The main difference between Coherent CoT and Stepwise ICL is that, in the second step of Stepwise ICL, the input prompt only includes the intermediate and final responses, z_is and y_is, of the in-context examples. In contrast, Coherent CoT plugged the predicted $\hat{z}_q$ back into the original input prompt to form the input prompt for the second step and retains the initial inputs x_is. This allows Coherent CoT to leverage both the intermediate reasoning and the original input when making the final prediction.

Remark 3.1 *Although our analysis primarily focuses on two-step CoT, the theoretical framework can be extended to multi-step reasoning. We provide additional results on multi-step CoT in Appendix D. However, compared to the two-step case, multi-step reasoning does not introduce additional insights beyond further demonstrating the superiority of Coherent CoT over Stepwise ICL. Therefore, our main analysis is centered on the two-step setting for clarity and interpretability.*

Model architecture. We follow the work of Zhang et al. (2023) and use transformers with one single-head linear attention layer as the model, which is defined as

$$f(E) = W_{out}W^V E \cdot \left((W^K E)^\top W^Q E\right), \tag{5}$$

where E denotes the input prompt, W^Q, W^K, $W^V \in \mathbb{R}^{m\times m}$ refer to the key, query and value matrix of the attention node and $W_{out} \in \mathbb{R}^{m\times m}$ refers to a fully-connected layer conducted to the output of the attention node. For $f_1(\cdot)$ in Stepwise ICL, $m = d+1$; for $f_2(\cdot)$, $m = 2$. For CoT, $m = d+2$.

Training objectives. For both Stepwise ICL and Coherent CoT, to train a model, we minimize the following loss function:

$$L(\Theta) = \mathbb{E}_{\{x_i\},x_q}(\widehat{y}_q - y_q)^2, \tag{6}$$

which represents the mean squared error (MSE) between the predicted response $\widehat{y}_q$ and the true response y_q of the query example. Here, Θ denotes the set of parameters in the transformer.

3.2 Coherent CoT vs Stepwise ICL

This section presents the main results of comparing Coherent CoT and Stepwise ICL. To simplify the derivation, we assume a specific format for the optimal attention parameters as follows:

Assumption 3.2 *We considerthe following specific formulations of the matrices W^K, W^Q, W_{out} and W^V for Stepwise ICL and Coherent CoT.*

- *For Stepwise ICL, the specific format for the parameters of $f_1(\cdot)$ is: $(W^K)^\top W^Q = \begin{bmatrix} u_x I_d & 0 \\ 0 & 0 \end{bmatrix}$ and $(W_{out}W^V)_{d+1,:} = (0,\dots,0,1/u_x)$. For $f_2(\cdot)$, we assume $(W^K)^\top W^Q = \begin{bmatrix} u_z & 0 \\ 0 & 0 \end{bmatrix}$ and $(W_{out}W^V)_{2,:} = (0, 1/u_z)$.*
- *For Coherent CoT, we assume*

$$(W^K)^\top W^Q = \begin{bmatrix} v_x I_d & & 0 \\ & v_z & \\ 0 & & v_y \end{bmatrix},$$

$$(W_{out}W^V)_{d+1,:} = (0,\dots,0,1/v_x,0)$$
$$(W_{out}W^V)_{d+2,:} = (0,\dots,0,1/v_y).$$

Assumption 3.2 is built upon the observations in the works of Zhang et al. (2023) and Cui et al. (2024). Based on these studies, in order to optimize the task of $x \to z$ in ICL, the corresponding off-diagonal elements of $(W^K)^\top W^Q$ and the first d elements in the $(W_{out}W^V)_{d+1,:}$ vector should be all zero when x follows $N(0, I_d)$. Since each step of the Coherent CoT process utilizes the same transformer model, we consider the same format for the $z \to y$ process, and assume all off-diagonal elements of $(W^K)^\top W^Q$ are zero and $(W_{out}W^V)_{d+1,:} = (0,\dots,0,v_y)$. With Assumption 3.2, we can rewrite $L(\Theta)$ to $L(u_x, u_z)$ for Stepwise ICL or $L(v_x, v_y, v_z)$ for Coherent CoT to highlight these parameters.

Given Assumption 3.2, we figure out the optimal solution of Stepwise ICL in Theorem 3.1.

Theorem 3.1 *Under Assumption 3.1 and 3.2, the optimal expected loss of Stepwise ICL is achieved when $u_x \neq 0$ and $u_z \neq 0$ are satisfied. The corresponding optimal loss is*

$$L(u_x^*, u_z^*) = \sigma^2 + \frac{7+d}{D} + \frac{\sigma^2}{D} + o\left(\frac{1}{D}\right).$$

The proof of Theorem 3.1 can be found in Appendix E.1. In short, we separate the MSE loss of the prediction into multiple terms and taking the expectation of each term considering $D \to \infty$.

Remark 3.2 *In Assumption 3.1, we assume that $x \sim N(0, I_d)$. The exact Gaussian distribution is used to derive the closed-form expression of the loss. If we relax it to other distributions, there will be no closed-form expression of the loss to exactly compare Coherent CoT and Stepwise ICL. Nonetheless, the high-level* *intuition on why Coherent CoT outperforms Stepwise ICL still holds: Stepwise ICL misses information on the previous reasoning steps, thus the prediction is worse than Coherent CoT.*

To compare with the optimal results of Stepwise ICL, we derive the optimal solution for Coherent CoT and present the results in Theorem 3.2.

Theorem 3.2 *Under Assumption 3.1 and 3.2, the optimal parameters v_x, v_y and v_z that minimize the Coherent CoT's loss satisfying $v_y = (v_x + v_z)$. The corresponding loss for Coherent CoT becomes*

$$\begin{aligned} L(v_x, v_y, v_z) \quad &= \quad \sigma^2 + \frac{d\sigma^2}{D} + \frac{1+d}{D} + \frac{4}{D}\frac{v_x v_z}{v_y^2} + \frac{6}{D}\frac{v_z^2}{v_y^2} \\ &\quad -\frac{v_z(2v_x + v_z)}{Dv_y^2}(d-1)\sigma^2 + o\left(\frac{1}{D}\right). \end{aligned}$$

To minimize the above loss, it is require that $v_y = \frac{(d-1)\sigma^2+2}{(d-1)\sigma^2-2}v_z = \frac{(d-1)\sigma^2+2}{4}v_x$ (where $v_z, v_x, v_y \neq 0$). Then the optimal expected loss of Coherent CoT is

$$\begin{aligned} L(v_x^*, v_y^*, v_z^*) \quad &= \quad \sigma^2 + \frac{d\sigma^2}{D} + \frac{1+d}{D} \\ &\quad -\frac{\left((d-1)\sigma^2 - 2\right)^2}{D((d-1)\sigma^2 + 2)} + o\left(\frac{1}{D}\right). \end{aligned}$$

The proof of Theorem 3.2 is similar to that of Theorem 3.1, whose details are shown in Appendix E.2.

Building on the above results, Proposition 3.1 below directly compares the expected loss of Coherent CoT and Stepwise ICL.

Proposition 3.1 *Given that $d \geq 2$, the expected loss of Coherent CoT equals to the expected loss of Stepwise ICL when $v_x = 0$ and $v_z = v_y$. In addition, the minimal expected loss of Coherent CoT is smaller than the one of Stepwise ICL.*

The proof of Proposition 3.1 is in Appendix E.3. Proposition 3.1 indicates that, regardless of the values of d and σ, the optimal expected loss of Coherent CoT is smaller than that of Stepwise ICL. To explain this, when $v_x = 0$ and $v_z = v_y$, Coherent CoT is equivalent to Stepwise ICL, and this condition also aligns with the optimal solution of Stepwise ICL. On the other hand, since $(v_x = 0, v_z = v_y)$ is not the optimal solution of Coherent CoT, Coherent CoT can achieve a smaller loss given its corresponding optimal solution.

Insights from the theory. To further investigate how the x_is and z_is contribute to the final prediction of $\hat{y}_q$ in CoT, we present the following proposition:

Proposition 3.2 *For any σ and d, v_z and v_y always share the same sign for optimal Coherent CoT. In addition, the ratio $\frac{v_z^*}{v_y^*}$ for the optimal Coherent CoT is consistently smaller 1. Meanwhile, when v_x, v_z and v_y are set such that Coherent CoT reduces to Stepwise ICL, the ratio satisfies $\frac{v_z}{v_y} = 1$.*

Based on Theorem 3.2 and Proposition 3.2, the prediction of the final response can be formulated as

$$\hat{y}_q = \frac{1}{D}\sum y_i \left(\frac{v_x}{v_y}x_i^\top x_q + \frac{v_z}{v_y}z_i\hat{z}_q\right). \tag{7}$$

The above formulation indicates how x_is and y_is impact the final prediction of $\hat{y}_q$. According to Proposition 3.2, in Coherent CoT, the initial inputs x_is always positively contribute to the final prediction $\hat{y}_q$. Additionally, the final prediction $\hat{y}_q$ in Coherent CoT relies less on the z_is and $\hat{z}_q$ compared to Stepwise ICL, indicating a reduced dependency on these intermediate values when using the optimal Coherent CoT.

Based on Proposition 3.2, the consequence of leveraging Coherent CoT to train a model is that, when there is an error in the prediction of z_q, Coherent CoT's reduced reliance on $\hat{z}_q$, along with its attention to x_q, ensures that the error in the prediction of z_q has a smaller impact on the final prediction. Moreover, since the final prediction of y_q incorporates both x_is and z_is, when $\hat{z}_q$ is inaccurate, the model can better adjust its prediction using the x_is values. This inherently provides a form of self-correction by leveraging the combined information from both x_is and z_is values.

Notably, the theorems in Section 3.2 and the later Section 3.3 focus on different stages of the model's application. The theorem in Section 3.2 highlights that compared with Stepwise ICL, using Coherent CoT during the *training stage* provides a better inference performance of the model. The theorem in Section 3.3, which examines the sensitivity of Coherent CoT to random noise, focuses on the *inference stage*. We also conduct simulations to demonstrate the advantage of Coherent CoT compared with Stepwise ICL, whose details can be found in Appendix B.

3.3 Sensitivity against Random Perturbation

While Section 3.2 investigates how Coherent CoT gains a better prediction performance compared to Stepwise ICL by considering all the previous steps during the reasoning process, this holistic process may also lead to a different sensitivity to potential errors/corruptions in each reasoning step. Therefore, in Section 3.3, with a model trained with Coherent CoT, we investigate the model's sensitivity and quantify the impact of random perturbations at different reasoning steps y_i, x_i, and z_i at the inference stage. The results are summarized in the following theorems respectively.

Theorem 3.3 *Under Assumption 3.1 and 3.2, when there is random perturbation $\delta_i \sim N(0, \sigma_\epsilon^2)$ added to y_i, we denote the loss for Coherent CoT as $L'_y(v_x, v_y, v_z)$. When v_x, v_y and v_z takes the optimal values, we have*

$$L'_y(v_x^*, v_y^*, v_z^*) - L(v_x^*, v_y^*, v_z^*) = o(1/D).$$

The proof of Theorem 3.3 is postponed to Appendix E.4. As indicated by the theorem, when the number of in-context examples D increases, the additional loss resulting from adding noise to y decreases at a rate of $o(\frac{1}{D})$. This suggests that adding noise to y leads to a negligible influence on CoT performance.

While the above theorem discusses the sensitivity of CoT to random perturbation added to y_is, the following theorem explores the cases when the random noise is added to x_is.

Theorem 3.4 *Under Assumption 3.1 and 3.2, when a random perturbation $\delta_i \sim N(0, \sigma_\epsilon^2)$ is added to x_i* ***in inference****, and taking $v_x + v_x = v_y$ as in Theorem 3.2, the expected loss for CoT becomes*

$$\begin{aligned} & L'_x(v_x, v_y, v_z) \\ = \; & \frac{1+d}{D} + \frac{4}{D}\frac{v_x v_z}{v_y^2} + \frac{6}{D}\frac{v_z^2}{v_y^2} + \frac{d}{D}\sigma_\epsilon^2 + \sigma^2 + \frac{d\sigma^2}{D} \\ & - \frac{v_z(2v_x + v_z)}{Dv_y^2}(d-1)\sigma^2 + \frac{v_x^2}{v_y^2}\sigma^2\sigma_\epsilon^2 + o\left(\frac{1}{D}\right). \end{aligned}$$

The theorem below is for the corruptions in z_is.

Theorem 3.5 *Under Assumption 3.1 and 3.2, when a random perturbation $\delta_i \sim N(0, \sigma_\epsilon^2)$ is added to z_i* ***in inference****, and taking $v_x + v_x = v_y$ as in Theorem 3.2, the expected loss for CoT becomes,*

$$\begin{aligned} & L'_z(v_x, v_y, v_z) \\ = & \frac{1+d}{D} + \frac{4}{D}\frac{v_x v_z}{v_y^2} + \frac{6}{D}\frac{v_z^2}{v_y^2} + \frac{3+d}{D}\frac{v_z^2}{v_y^2}\sigma_\epsilon^2 + \sigma^2 + \frac{d\sigma^2}{D} \\ & - \frac{v_z(2v_x + v_z)}{Dv_y^2}(d-1)\sigma^2 + \frac{v_z^2}{v_y^2}\sigma^2\sigma_\epsilon^2 + o\left(\frac{1}{D}\right). \end{aligned}$$

The proof of Theorem 3.4 and 3.5 are similar to that of Theorem 3.3, whose details are shown in Appendix E.5 and E.6.

Given the results in Theorem 3.4 and 3.5, the following proposition provides a clearer comparison:

Proposition 3.3 *When v_x, v_y and v_z takes the optimal values, we have (1) $L'_x(v_x^*, v_y^*, v_z^*) > L'_y(v_x^*, v_y^*, v_z^*)$ and $L'_z(v_x^*, v_y^*, v_z^*) > L'_y(v_x^*, v_y^*, v_z^*)$ for any values of d and σ; (2) $L'_x(v_x^*, v_y^*, v_z^*) > L'_z(v_x^*, v_y^*, v_z^*)$ for $\sigma^2 \in [0, a) \cup (b, \infty)$ and $L'_x(v_x^*, v_y^*, v_z^*) < L'_z(v_x^*, v_y^*, v_z^*)$ for $\sigma^2 \in (a, b)$, where $b > a$ are the positive roots of $f(\theta) = (d-1)^2\theta^3 + (3d-7)(d-1)\theta^2 - (4d + 8d^2)\theta + 12 = 0$.*

The proof of Proposition 3.3 is shown in Appendix E.7.

Insights from the theory. The results in Proposition 3.3 indicate that when a certain level of random noise is introduced to CoT, whether adding noise to the x_is or the z_is leads to a higher expected loss depends on the scale of the variance σ^2. Compared the results with that of adding noise to y_i, adding noise to x_i or z_i leads to a greater influence on the performance of CoT. An important insight from this result is that, compared to the noise in the final label (mislabeling), errors occurring in the reasoning steps of CoT have a greater influence on the accuracy of the final prediction. From this perspective, CoT is more sensitive to mistakes made during the reasoning process than to the noise in the final response. As a result, when asking the model to account for potential errors during reasoning, it is crucial to pay more attention to error in the intermediate reasoning steps rather than label noise. We also provide simulation results illustrating the sensitivity of CoT to random perturbations at different reasoning steps, whose details can be found in Appendix B.

3.4 Effectiveness of Error-Aware Demonstration

The analysis in Section 3.3 reveals that, at the inference stage, CoT is more sensitive to errors in the intermediate reasoning steps of the demonstration than to errors in the final outcome. Inspired by this, we hypothesize that it is beneficial for the model to learn to adjust its predictions by considering the possibility of earlier mistakes. In this subsection, we provide theoretical results to verify this conjecture and show that considering potential mistakes and their corresponding corrections within the CoT process (which we define as "Error-Aware Demonstration") yields better final predictions.

The results of "Error-Aware Demonstration" are presented in the following theorem:

Theorem 3.6 *Under Assumption 3.1 and 3.2, when there is a random perturbation $\delta_i \sim N(0, \sigma_\epsilon^2)$ added to z_i* ***at the inference stage*** *with the following format:*

$$\begin{pmatrix} x_1 & x_1 & \cdots & x_{D/2} & x_{D/2} & x_q \\ z_1 + \delta_1 & z_1 - \delta_1 & \cdots & z_{D/2} + \delta_{D/2} & z_{D/2} - \delta_{D/2} & \hat{z}_q \\ y_1 & y_1 & \cdots & y_{D/2} & y_{D/2} & 0 \end{pmatrix}, \tag{8}$$

where $y_i = \beta^\top x_i + \epsilon_i$ and $\epsilon_i \sim N(o, \sigma^2)$. When taking $v_x + v_x = v_y$ as in Theorem 3.2, the expected loss for CoT becomes

$$\begin{aligned} & L'_{z,\ new}(v_x, v_y, v_z) \\ = & \sigma^2 + \frac{2d\sigma^2}{D} + \frac{2+2d}{D} - \frac{2v_z(2v_x + v_z)}{Dv_y^2}(d-1)\sigma^2 \end{aligned}$$

$$+\frac{8}{D}\frac{v_x v_z}{v_y^2}+\frac{12}{D}\frac{v_z^2}{v_y^2}+o\left(\frac{1}{D}\right). \tag{9}$$

The proof of the theorem is shown in Appendix E.8. In the above theorem, when considering the input prompt during the inference stage following the format (8) (and no corruption in the training prompts), we mimic Error-Aware Demonstration, where there are error in the intermediate step z, e.g., $(x_1, z_1+\delta_1, y_1)$, and correction for the error, e.g., $(x_1, z_1-\delta_1, y_1)$. We derive the loss of the Error-Aware Demonstration in Eq.(9). In comparison, the standard demonstration (without error correction) having the following format:

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_D & x_q \\ z_1+\delta_1 & z_2+\delta_2 & \cdots & z_D+\delta_D & \hat{z}_q \\ y_1 & y_2 & \cdots & y_D & 0 \end{pmatrix}. \tag{10}$$

In the following proposition, we demonstrate that although the number of distinct (x_i, z_i, y_i) pairs in the Error-Aware Demonstration is less than the standard demonstration, the loss $L'_{\text{z, new}}(v_x, v_y, v_z)$, is smaller than that of standard demonstration, i.e., the $L'_{\text{z}}(v_x, v_y, v_z)$ in Theorem 3.5.

Proposition 3.4 *When v_x, v_y and v_z takes the optimal values to minimize the pretraining loss of Coherent CoT we have $L'_z(v_x^*, v_y^*, v_z^*) - L'_{z,\ new}(v_x^*, v_y^*, v_z^*) = c\delta_\epsilon^2 - O(\frac{1}{D})$ for some constant c, which indicates that $L'_{z,\ new}(v_x^*, v_y^*, v_z^*) < L'_z(v_x^*, v_y^*, v_z^*)$ when $\delta_\epsilon^2 \gg O(\frac{1}{D})$.*

The above proposition verify the superiority of the Error-Aware Demonstration, which provide a theoretical justification for our proposed Error-Aware Demonstration method.

On the other hand, since there are only $D/2$ distinct (x_i, z_i, y_i) pairs in the Error-Aware Demonstration, one may question that whether directly using D pairs of (x_i, y_i) as one-step ICL is better or not. To investigate this, we take

$$E_{ICL}^{(1)} = \begin{pmatrix} x_1 & x_2 & \dots & x_D & x_q \\ y_1 & y_2 & \dots & y_D & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(D+1)}. \tag{11}$$

and denote $L_{icl}(u_x)$ as the one-step ICL loss with $(W_{out}W^V)_{d+1,:} = (0, \dots, 0, 1/u_x)$. With the optimal choice u_x^*, the following proposition hold:

Proposition 3.5 *When v_x, v_y and v_z takes the optimal values to minimize the pretraining loss of ICl, considering that d_1 and d_2 are the roots of $(2\sigma^4-\sigma^2)d^2 + (-4\sigma^4 - 9\sigma^2 - 2)d + (\sigma^4 + 10\sigma^2 + 8) = 0$ and $d_1 < d_2$, and $d_0 = \frac{\sigma^4+10\sigma^2+8}{4\sigma^4+9\sigma^2+2}$*

- *If $\sigma^2 > \frac{1}{2}$, $L_{icl}(u_x^*) < L'_{z,new}$ for $d \in (d_1, d_2)$ and $L_{icl}(u_x^*) > L'_{z,new}$ for $d \in (d_2, \infty) \cup (0, d_1)$.*
- *If $0 < \sigma^2 < \frac{1}{2}$, $L_{icl}(u_x^*) > L'_{z,new}$ for $d \in (d_1, d_2)$ and $L_{icl}(u_x^*) < L'_{z,new}$ for $d \in (d_2, \infty) \cup (0, d_1)$.*
- *If $\sigma^2 = \frac{1}{2}$, $L_{icl}(u_x^*) > L'_{z,new}$ for $d \in (0, d_0)$ and $L_{icl}(u_x^*) < L'_{z,new}$ for $d \in (d_0, \infty)$.*

According the above proposition, whether it is better to use error-aware CoT or one-step ICL depends on the value of d, which is the dimension of the input x. The detailed formula of L_{icl} can be found in Theorem C.1 in Appendix C.

4 IMPROVING COT THROUGH ERROR-AWARE DEMONSTRATIONS

The theoretical analysis in Section 3.4 reveals that incorporating error demonstrations with corresponding corrections leads to improved model performance. Building on this insight, in this section, we propose a method to incorporate both correct and incorrect reasoning paths when composing demonstrations in practical applications of CoT.

4.1 Methodology

We use a date understanding problem as an example to explain the proposed method in detail, and present the demonstration format in Figure 1. In standard CoT prompting, few-shot examples with correct reasoning paths are provided in the prompt demonstration. In contrast, in the proposed demonstration format, after presenting the question, we first provide a potentially incorrect reasoning path, clearly labeled as a wrong solution. We then provide a detailed explanation of why this reasoning is incorrect, pointing out specific missteps or logical errors. Both the incorrect reasoning path and the analysis of why it is flawed are created manually. After analyzing the incorrect path, we provide a step-by-step correct reasoning process that leads to the correct answer.

By incorporating both correct and incorrect reasoning paths in the demonstrations, the proposed method teaches the model not only the correct reasoning path but also how to identify and handle potential reasoning errors. This error-aware approach helps improve the model's ability to adjust predictions and enhances its overall performance in reasoning tasks.

Notably, there have been some paper consider leveraging incorrect reasoning paths. However, our method introduces novel aspects that distinguish it from prior approaches. Specifically, while Tong et al. (2024) suggests to use interactive prompting to guide the model to recognize and correct mistakes from its initial predictions, and An et al. (2023) proposes to use a Mistake-Correction Dataset to fine-tune LLMs, our work did not involve interactive prompting or LLM fine-tuning. Instead, we operates with a single query, relying on a demonstration that incorporate both correct and incorrect reasoning paths. This one-time demonstration requires less additional computational overhead while demonstrating effectiveness in improving the CoT performance. While Mo et al. (2024) also proposes including potential errors in demonstrations, it focuses specifically on ICL rather than CoT reasoning. Furthermore, our work emphasizes the importance of not only including incorrect paths but also explaining the errors to help the model understand and utilize the demonstration more effectively. This necessity will be further discussed in Section 4.4.

4.2 Experiments

In this section, we evaluate our proposed method with various LLMs on multiple benchmarks. Analysis on the influence of the number of examples and the effect of Error-Aware Demonstrations on token usage and efficiency are postponed to Appendix F.1 and F.2.

4.2.1 Experimental setup

Language Models. Our main experiments involve four LLMs: GPT-3.5-Turbo (Brown, 2020), GPT-4o-mini (Achiam et al., 2023), Gemini Pro (Team et al., 2023) and DeepSeek 67B (Bi et al., 2024)[2]. We provide additional results using GPT-4-Turbo (Achiam et al., 2023) and LLaMA-3.1-8B (Instruct) (Grattafiori et al., 2024) in Appendix F.3. For generation, we set the temperature to 0 to ensure deterministic outputs.

Benchmarks. We use five datasets from two benchmarks for the experiments: the BBH benchmark (Srivastava et al., 2022) and the GSM8K benchmark (Cobbe et al., 2021)[3]. The BBH benchmark focuses on reasoning tasks, and we select Disambiguation QA, Tracking Shuffled Objects (7 objects), Date Understanding, and Penguins in a Table from this benchmark. Each of these datasets contains 250 examples. Besides the BBH benchmark, we also use the GSM8K benchmark, which consists of 1,319 examples of arithmetic and grade-school math problems.

4.3 Main Results

In Table 1, we demonstrate the performance of various LLMs across different datasets when using standard CoT prompting (w/o IR) and our proposed method, which incorporates Incorrect Reasoning (IR) in the demonstrations (w/ IR). From Table 1, we can see that in most cases, adding handcrafted incorrect reasoning paths to CoT demonstrations improves the models' performance. In some settings, our proposed method brings a significant improvement exceeding 5%. For example, in the Tracking Shuffled Objects dataset, Gemini Pro shows a 6.60% improvement (from 58.20% to 64.80%), and in Penguins in a Table, DeepSeek 67B demonstrates an increase of 6.17% (from 73.97% to 80.14%). These results highlight the positive impact of exposing models to incorrect reasoning paths.

Table 1: Performance of CoT with and without handcrafted Incorrect Reasoning (IR) across different datasets and models.

Dataset	Method	GPT-3.5-Turbo	GPT-4o-mini
Disambiguation QA	w/o IR	68.00%	70.00%
	w/ IR	**72.00%**	**71.60%**
Tracking Shuffled Objects	w/o IR	56.53%	88.00%
	w/ IR	**61.20%**	**88.80%**
Date understanding	w/o IR	82.27%	90.80%
	w/ IR	**85.07%**	**91.47%**
Penguins in a table	w/o IR	81.34%	91.10%
	w/ IR	**82.19%**	**92.92%**
GSM8K	w/o IR	81.03%	92.72%
	w/ IR	**83.38%**	**93.03%**
Dataset	**Method**	**Gemini Pro**	**DeepSeek 67B**
Disambiguation QA	w/o IR	68.80%	81.20%
	w/ IR	**76.80%**	81.20%
Tracking Shuffled Objects	w/o IR	58.20%	71.20%
	w/ IR	**64.80%**	**72.40%**
Date understanding	w/o IR	88.80%	80.80%
	w/ IR	88.80%	**83.20%**
Penguins in a table	w/o IR	82.19%	73.97%
	w/ IR	**83.56%**	**80.14%**
GSM8K	w/o IR	80.82%	**83.85%**
	w/ IR	**81.27%**	83.40%

4.4 Additional Experiments

Table 2: Performance of CoT with GPT-3.5-Turbo with and without model error explanation (EE).

Method	Disambiguation QA	Tracking Shuffled Objects	Date understanding	Penguins in a table	GSM8K
w/o IR	68.00%	56.53%	82.27%	81.34%	81.03%
w/ EE	72.00%	61.20%	85.07%	82.19%	83.38%
w/o EE	70.80%	57.60%	82.00%	80.82%	82.87%

Necessity of including error explanation. We present an ablation study to evaluate our proposed method when only the incorrect reasoning path is provided, without including an explanation of the error in that reasoning path. The experiments were conducted using GPT-3.5-Turbo, and the results on different datasets are shown in Table 2. We compare the results with the case where both the incorrect reasoning paths and error explanations are included, as well as the case where neither of them is provided.

The results demonstrate that, in general, the performance of our method is reduced if the demonstration

[2]All models are used in accordance with their respective licensing terms.

[3]Both datasets are distributed under the MIT license.

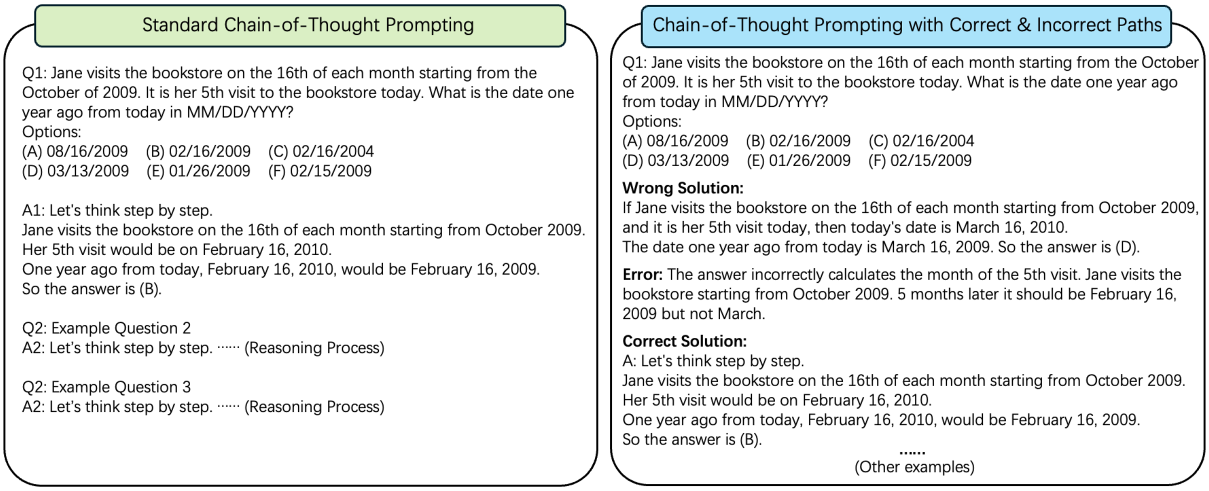

Figure 1: Standard CoT prompting vs the proposed method: CoT prompting with correct & incorrect paths.

lacks an explanation of the error in the incorrect reasoning path. In some cases, such as with the Date Understanding dataset, the performance is even worse than when neither incorrect reasoning paths nor explanations are provided (from 82.27% to 82.2%). This suggests that merely presenting incorrect reasoning without explaining why it is wrong may confuse the model, making it harder for it to discern the correct logic from the incorrect one. These results highlight the necessity of including error explanations when providing incorrect reasoning paths in demonstrations.

Model-generated incorrect reasoning paths. As a variant of the proposed method, we further consider an implementation when the incorrect reasoning paths provided in the demonstration are not handcrafted but model-generated. Specifically, we select incorrect reasoning paths from LLM-generated answers and include them in the demonstration. Intuitively, while handcrafting the wrong reasoning paths is generally more efficient, using model-generated incorrect paths has greater potential for improvement. Since the incorrect solutions are generated by the model itself, it allows the model to better recognize and avoid similar mistakes in future tasks. Notably, in this variant, the corresponding explanations of the errors in the reasoning still need to be provided manually.

The results when using model-generated incorrect reasoning paths are shown in Table 3 as "IR(M)", with a comparison to the handcrafted incorrect reasoning paths, denoted as "IR (H)". The experiments are conducted on the Penguins in a Table Dataset.

From the results in Table 3 we can observe that, across most models, using model-generated incorrect reasoning paths consistently leads to better performance compared to using handcrafted incorrect reasoning paths. In some cases, the improvement is substantial: for DeepSeek 67B, the improvement is from 82.88% to 88.36%. This indicates that model-generated incorrect paths allow the models to better recognize and learn from their own errors, resulting in enhanced reasoning capabilities.

Remark 4.1 *The results for the w/o IR case in Table 3 are different from those in Table 1. This is because the corresponding example questions used in the demonstrations are different. This discrepancy suggests that, beyond incorporating incorrect reasoning paths, selecting appropriate example questions in the demonstration of CoT is also crucial.*

Table 3: Performance of CoT in the Penguins in a Table Dataset with or without handcrafted/model generated Incorrect Reasoning (IR(H)/IR(M)) across different models.

Method	GPT-3.5-turbo	GPT-4o-mini	Gemini Pro	DeepSeek 67B
w/o IR	79.45%	99.32%	81.51%	81.51%
w/ IR(H)	87.67%	99.32%	82.88%	82.88%
w/ IR(M)	**89.04%**	99.32%	**85.62%**	**88.36%**

5 CONCLUSION

This paper provides a theoretical analysis of Coherent CoT by investigating its advantages over Stepwise ICL, showing that treating CoT as a holistic process leads to improved error correction and prediction accuracy. In addition, we examine Coherent CoT's sensitivity to errors in different reasoning steps of the demonstration examples during the inference stage. We observe that Coherent CoT is more sensitive to the error in the intermediate reasoning process than the inaccuracies in the final response. Inspired by this result, we propose to incorporate both correct and incorrect reasoning paths in demonstrations to improve the accuracy of the intermediate steps to enhance CoT performance. Experimental results validate the effectiveness of this approach.

Acknowledgements

Yingqian Cui, Pengfei He and Jiliang Tang are supported by the National Science Foundation (NSF) under grant numbers CNS2321416, IIS2212032, IIS2212144, IOS2107215, DUE2234015, CNS2246050, DRL2405483 and IOS2035472, the Army Research Office (ARO) under grant number W911NF-21-1-0198, Amazon Faculty Award, JP Morgan Faculty Award, Meta, Microsoft and SNAP.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Learning from mistakes makes llm better reasoner. *arXiv preprint arXiv:2310.20689*, 2023.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2023.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. *arXiv preprint arXiv:2312.06528*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yingqian Cui, Jie Ren, Pengfei He, Jiliang Tang, and Yue Xing. Superiority of multi-head attention in in-context linear regression. *arXiv preprint arXiv:2401.17426*, 2024.

Guhao Feng, Yuntian Gu, Bohang Zhang, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *arXiv preprint arXiv:2305.15408*, 2023.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Pengfei He, Yingqian Cui, Han Xu, Hui Liu, Makoto Yamada, Jiliang Tang, and Yue Xing. Towards the effect of examples on in-context learning: A theoretical case study. *arXiv preprint arXiv:2410.09411*, 2024.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

Mingyu Jin, Qinkai Yu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, Mengnan Du, et al. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis. *arXiv preprint arXiv:2410.02167*, 2024a.

Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. Dissecting chain-of-thought: A study on compositional in-context learning of mlps. *arXiv preprint arXiv:2305.18869*, 2023.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 2024b.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*, 2023.

Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.

Ying Mo, Jiahao Liu, Jian Yang, Qifan Wang, Shun Zhang, Jingang Wang, and Zhoujun Li. C-icl: contrastive in-context learning for information extraction. *arXiv preprint arXiv:2402.11254*, 2024.

Ben Prystawski, Michael Li, and Noah Goodman. Why think step by step? reasoning emerges from the locality of experience. *Advances in Neural Information Processing Systems*, 36, 2024.

Rylan Schaeffer, Kateryna Pistunova, Samar Khanna, Sarthak Consul, and Sanmi Koyejo. Invalid logic, equivalent gains: The bizarreness of reasoning in language model prompting. *arXiv preprint arXiv:2307.10573*, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. Can llms learn from previous mistakes? investigating llms' errors to boost for reasoning. *arXiv preprint arXiv:2403.20046*, 2024.

Rasul Tutunov, Antoine Grosnit, Juliusz Ziomek, Jun Wang, and Haitham Bou-Ammar. Why can large language models generate correct chain-of-thoughts? *arXiv preprint arXiv:2310.13571*, 2023.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: unveiling how chain-of-thought enhances transformer sample efficiency. *arXiv preprint arXiv:2410.05459*, 2024.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *arXiv preprint arXiv:2307.13339*, 2023.

Yue Xing, Xiaofeng Lin, Namjoon Suh, Qifan Song, and Guang Cheng. Benefits of transformer: Incontext learning in linear regression tasks with unstructured data. *arXiv preprint arXiv:2402.00743*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.

[Yes] A clear description of the mathematical setting, assumptions and models are presented in Section 3.1 and 3.2

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable] The paper not propose new algorithm.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No] The method we proposed involves only modifying prompt demonstration and does not require code implementation.

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes] The statements of the full set of assumptions are presented in Assumption 3.1 and 3.2.

   (b) Complete proofs of all theoretical results. [Yes] The complete proofs of all theoretical results are presented in the Appendix.

   (c) Clear explanations of any assumptions. [Yes] The explanations of the assumptions are included in Section 3.1 and 3.2.

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] We have provided sufficient details about the setting of the experiments to ensure the reproducibility.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] We have provided sufficient details about the setting of our experiments in Section 4.2.1.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] We have clearly defined the measure and provided prediction intervals when reporting the results in figures.

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] We have provide a description of the computing infrastructure in the appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes] We have properly cited the creator of the existing assets.

   (b) The license information of the assets, if applicable. [Yes] We have provided the license information of the assets.

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] We have provided examples of our proposed demonstration format in the appendix.

   (d) Information about consent from data providers/curators. [Yes] The license information we included indicates the consent from the data provider.

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable] The paper does not include sensible content.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable] The paper does not involve crowdsourcing nor research with human subjects.

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Real-life Examples of Stepwise ICL and Coherent CoT

To explain the difference between Stepwise ICL and Coherent CoT, in Table 4, we consider a real example and present the input provided to the LLM for each step in Stepwise ICL and Coherent CoT. From the table, we can see that in Stepwise ICL, the input for each step consists only of the result from the previous step. In contrast, Coherent CoT aggregates the question and the outputs from all previous steps as the input for each step.

Table 4: The input provided to LLM / The Output of LLM for each step in Step-wise ICL and Coherent CoT. The question is: $7 \times 2 \div (6 + 3 \times 8 - 4 \times 7)$.

Step	**Step-wise ICL Input**	**Coherent CoT Input**	**Output**
Step 1	$7 \times 2 \div (6 + 3 \times 8 - 4 \times 7)$	$7 \times 2 \div (6 + 3 \times 8 - 4 \times 7)$	$= 14 \div (6 + 24 - 4 \times 7)$
Step 2	$= 14 \div (6 + 24 - 4 \times 7)$	$7 \times 2 \div (6 + 3 \times 8 - 4 \times 7)$ $= 14 \div (6 + 24 - 4 \times 7)$	$= 14 \div (6 + 24 - 28)$
Step 3	$= 14 \div (6 + 24 - 28)$	$7 \times 2 \div (6 + 3 \times 8 - 4 \times 7)$ $= 14 \div (6 + 24 - 4 \times 7)$ $= 14 \div (6 + 24 - 28)$	$= 14 \div (30 - 28)$
Step 4	$= 14 \div (30 - 28)$	$7 \times 2 \div (6 + 3 \times 8 - 4 \times 7)$ $= 14 \div (6 + 24 - 4 \times 7)$ $= 14 \div (6 + 24 - 28)$ $= 14 \div (30 - 28)$	$= 14 \div 2$
Step 5	$= 14 \div 2$	$7 \times 2 \div (6 + 3 \times 8 - 4 \times 7)$ $= 14 \div (6 + 24 - 4 \times 7)$ $= 14 \div (6 + 24 - 28)$ $= 14 \div (30 - 28)$ $= 14 \div 2$	$= 7$

B Simulation Results

Settings. In simulations, the transformers are trained from scratch to learn from the data generated from Assumption 3.1. We modify the implementation of Li et al. (2023) to conduct the experiments. Specifically, the transformer consists of a single softmax attention layer without read-in or read-out layers. During the training process, we set the learning rate to 1e-4, the batch size to 256, and the number of training steps to 2e5. For the inference process, the batch size was set to 64, with a total of 128,000 inference examples.

Coherent CoT vs Stepwise ICL. To ensure stable convergence in training the Stepwise ICL, we separately trained the two models involved in the Stepwise ICL process. Based on Theorem 3.1, when minimizing Eq. 6, Stepwise ICL indeed optimizes both transformers. The results of the comparison are shown in Figure 2. From the figure, we can observe that as the number of in-context examples increases, Coherent CoT converges to a lower error than Stepwise ICL. These results align with our theoretical findings, confirming the superiority of Coherent CoT in achieving more accurate predictions.

Sensitivity of CoT to random perturbations. The simulation illustrating the sensitivity of CoT to random perturbations at different reasoning steps is shown in Figure 3[4]. The training setup is similar to the experiment in Figure 2, with the variance of the noise, σ_ϵ^2, set to 1. From the figure, we observe that when random perturbations are added to the y_is, the loss remains close to the case where there is no noise. However, when noise is introduced to the x_is or z_is, the loss increases significantly. This result is consistent with our theoretical findings, confirming that CoT is more sensitive to perturbations in the earlier reasoning steps than label noise.

C Theoretical Results on One-step ICL

To further investigate whether it is better to use Error-Aware Demonstration or directly skip the intermediate

[4]In both Figure 2 and 3 we plot the lower and upper bounds of the 90% prediction interval for the mean loss values. However, when the bounds are very close to the mean, they may not be clearly visible.

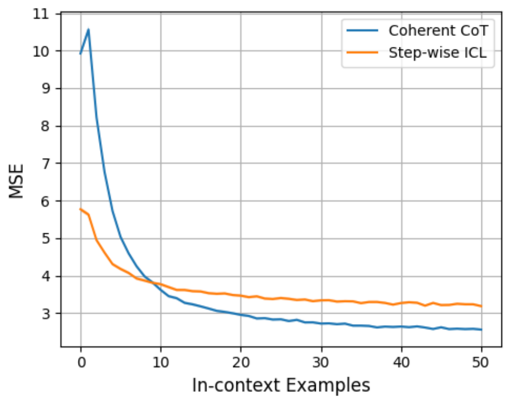

Figure 2: Comparison of Coherent CoT and Stepwise ICL

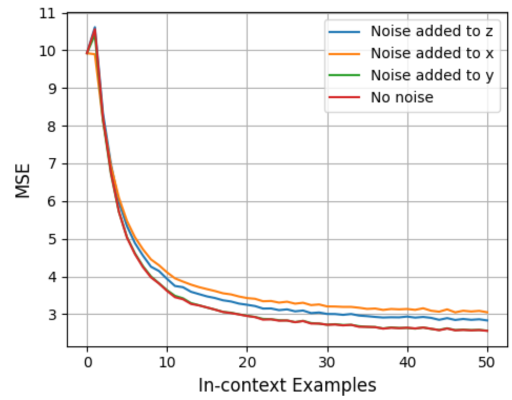

Figure 3: CoT's sensitivity to different steps (The green and red lines overlap because the results for adding noise to y and the case of no noise are very close to each other.)

step to predict y only based on x (use one-step ICL instead of CoT), we provide the following theorem about the performance of one-step ICL.

Theorem C.1 *Considering that the input prompt is formatted as*

$$E_{ICL}^{(1)} = \begin{pmatrix} x_1 & x_2 & \dots & x_D & x_q \\ y_1 & y_2 & \dots & y_D & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(D+1)}. \tag{12}$$

Under Assumption 3.1, the optimal configurations of the transformer satisfies $(W^K)^\top W^Q = \begin{bmatrix} u_x I_d & 0 \\ 0 & 0 \end{bmatrix}$ *and* $(W_{out}W^V)_{d+1,:} = (0,\dots,0,1/u_x)$*. The corresponding optimal loss for one-stage ICL is given by*

$$L_{icl}(u_x^*) = \sigma^2 + \frac{d\sigma^2}{D} + \frac{1+d}{D}.$$

Proof. Given that $E_{ICL}^{(1)} = \begin{pmatrix} x_1 & x_2 & \dots & x_D & x_q \\ y_1 & y_2 & \dots & y_D & 0 \end{pmatrix}$, the MSE loss is expressed as

$$\begin{aligned}
\mathbb{E}\left(y_q - \hat{y_q}\right)^2 &= \mathbb{E}\left(y_q - (W_{out}W^V_{d-1:d,:})^\top E_{ICL}^{(1)}\left(\frac{1}{D}E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \\ 0 \end{bmatrix}\right)\right)^2 \\
&= \mathbb{E}_{\{x_q,\epsilon_q\}}\mathbb{E}_{\{x_i,\epsilon_i\}_{i\in[D]}}\left(y_q - (\frac{1}{Dv_x})\left[y_1,y_2,\dots,y_D,0\right]\begin{bmatrix} v_x x_1^\top x_q \\ v_x x_2^\top x_q \\ \vdots \\ v_x x_q^\top x_q \end{bmatrix}\right)^2 \\
&= \mathbb{E}_{\{x_q,\epsilon_q\}}\mathbb{E}_{\{x_i,\epsilon_i\}_{i\in[D]}}\left((\beta^\top x_q + \epsilon_q) - \frac{1}{D}\sum(\beta^\top x_i+\epsilon_i)x_i^\top x_q\right)^2 \\
&= \underbrace{\mathbb{E}_{\{x_q,\epsilon_q\}}\left(\beta^\top x_q x_q^\top \beta + \epsilon_q^2 + 2\beta^\top x_q\epsilon_q\right)}_{A_1} - \underbrace{\mathbb{E}_{\{x_q,\epsilon_q\}}\mathbb{E}_{\{x_i,\epsilon_i\}_{i\in[D]}}\left(\frac{2}{D}(\beta^\top x_q+\epsilon_q)\sum(\beta^\top x_i+\epsilon_i)x_i^\top x_q\right)}_{A_2} \\
&\quad + \underbrace{\mathbb{E}_{\{x_q,\epsilon_q\}}\mathbb{E}_{\{x_i,\epsilon_i\}_{i\in[D]}}\left(\frac{1}{D}\sum(\beta^\top x_i+\epsilon_i)x_i^\top x_q\right)^2}_{A_3}
\end{aligned}$$

Because $x_i \in N(0,I_d)$ and $x_q \in N(0,I_d)$, we have $\mathbb{E}_{\{x_i\}_{i\in[D]}}(x_i x_i^\top) = I$ and $\mathbb{E}_{x_q}(x_q x_q^\top) = I$. Because $\epsilon_i \in N(0,\sigma^2)$ and $\epsilon_q \in N(0,\sigma^2)$, we have $\mathbb{E}_{\epsilon_1}(\epsilon_i^2) = \sigma^2$ and $\mathbb{E}_{\epsilon_q}(\epsilon_q^2) = \sigma^2$.

Therefore,

$$A_1 = \|\beta\|^2 + \sigma^2$$

$$\begin{aligned}
A_2 &= \mathbb{E}_{x_q}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{2}{D}(\beta^\top x_q \sum(\beta^\top x_i x_i^\top x_q)\right) \\
&= \frac{2}{D}\mathbb{E}_{x_q}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left((\beta^\top(\sum x_i x_i^\top)x_q x_q^\top \beta)\right) \\
&= 2\|\beta\|^2,
\end{aligned}$$

and

$$\begin{aligned}
A_3 &= \mathbb{E}_{x_q}\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{1}{D^2}\left(\sum(\beta^\top x_i + \epsilon_i)x_i^\top\right)x_q x_q^\top\left(\sum(\beta^\top x_i + \epsilon_i)x_i^\top\right)^\top\right) \\
&= \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{1}{D^2}\left(\sum(\beta^\top x_i + \epsilon_i)x_i^\top\right)\left(\sum(\beta^\top x_i + \epsilon_i)x_i^\top\right)^\top\right) \\
&= \mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{1}{D^2}(\sum \epsilon_i x_i^\top)(\sum \epsilon_i x_i) + \mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{1}{D^2}(\sum \beta^\top x_i x_i^\top)(\sum \beta^\top x_i x_i) \\
&= \frac{1}{D}\mathbb{E}_{x_1}(\epsilon_1^2 x_1^\top x_1) + \frac{1}{D}\mathbb{E}_{x_1}\left(\beta^\top(x_1 x_1^\top)^2\beta\right) + \frac{D(D-1)}{D^2}\mathbb{E}_{x_1}\mathbb{E}_{x_2}\left(\beta^\top(x_1 x_1^\top)(x_2 x_2^\top)\beta\right) \\
&= \frac{d}{D}\sigma^2 + \frac{d+2}{D}\|\beta\|^2 + \frac{D(D-1)}{D}\|\beta\|^2
\end{aligned}$$

Summing A_1, $-A_2$ and A_3, we have

$$\begin{aligned}
\mathbb{E}\left(y_q - \hat{y_q}\right)^2 &= \|\beta\|^2 + \sigma^2 - 2\|\beta\|^2 + \frac{d}{D}\sigma^2 + \frac{d+2}{D}\|\beta\|^2 + \frac{D(D-1)}{D}\|\beta\|^2 \\
&= \sigma^2 + \frac{d}{D}\sigma^2 + \frac{d+1}{D}\|\beta\|^2
\end{aligned}$$

Recalling that $\|\beta\|^2 = 1$, we have

$$\mathbb{E}\left(y_q - \hat{y_q}\right)^2 = \sigma^2 + \frac{d}{D}\sigma^2 + \frac{d+1}{D}$$

D Theoretical Results on Multi-Step Reasoning

In this section, we provide a sketch of the theorem about Stepwise ICL vs Coherent CoT when our analysis with two step reasoning is generalized to multi-step reasoning.

For **the data generation process**, we consider the following:

- The independent variable $x \in \mathbb{R}^d \in N(0, I_d)$.
- $n-1$ intermediate response $z_i = W_i z_{i-1}$, $i \in [2, n-1]$ and $z_1 = W_1 x$.
- The final response $y = z_n + \epsilon$, $\epsilon \in N(0, \sigma^2)$.

$W_1 \in \mathbb{R}^{d_1 \times d}$, $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ for $2 \le i \le n-1$, $W_n \in \mathbb{R}^{1 \times d_{i-1}}$ and W_i satisfies $W_i^\top W_i = I$

For the Definition of Stepwise ICL and Coherent CoT: Similar to Definition 3.1, the input format for different steps in Step-wise ICL is formatted by concatenating the current inputs and outputs of the in-context examples and the current input of the query example, where the current input of the query example is derived from the output of the last step. The input format for Coherent CoT is formatted by concatenating all the input, intermediate steps and outputs of the in-context example, and provide all the obtained intermediate prediction in the last column of the matrix. After every intermediate prediction, the result is plugged into the input prompt. Both process involves $n+1$ steps of prediction.

For the assumption on the optimal format of the model parameters, considering that:

- For Stepwise ICL: the specific format for the parameters of $f_1(\cdot)$ is: $(W^K)^\top W^Q = \begin{bmatrix} u_x I_d & 0 \\ 0 & 0 \end{bmatrix}$ and $(W_{out}W^V)_{d+1,:} = (0,\dots,0,1/u_x)$. For $2 \le i \le n-1$, the specific format for the parameters of $f_i(\cdot)$ is: $(W^K)^\top W^Q = \begin{bmatrix} u_{i-1} I_{d_{i-1}} & 0 \\ 0 & 0 \end{bmatrix}$ and $(W_{out}W^V)_{d+1,:} = (0,\dots,0,1/u_i)$. $f_{n+1}(\cdot)$ satisfies $(W^K)^\top W^Q = \begin{bmatrix} u_y & 0 \\ 0 & 0 \end{bmatrix}$ and $(W_{out}W^V)_{2,:} = (0, 1/u_y)$

- For Coherent CoT, we assume

$$(W^K)^\top W^Q = \begin{bmatrix} v_x I_d & & & & 0 \\ & v_{z_1} I_{d_1} & & & \\ & & \cdots & & \\ & & & v_{z_n} I_{d_n} & \\ 0 & & & & v_y \end{bmatrix},$$

and

$$W_{out}W^V = \begin{bmatrix} & & & & 0 \\ & \frac{1}{v_{w_1}} I_{d_1} & & & \\ & & \cdots & & \\ & & & \frac{1}{v_{w_n}} I_{d_n} & \\ 0 & & & & \frac{1}{v_{w_{n+1}}} \end{bmatrix},$$

Then we have the following theorems regarding the optimal loss of Stepwise ICL and Coherent CoT.

Theorem D.1 *Under the assumption on the data generation process and optimal model configurations, the optimal expected loss of Stepwise ICL is achieved when $u_x \neq 0$, $u_y \neq 0$ and $u_i \neq 0 \quad (i \in \{2,\dots,n\})$ are satisfied. The corresponding optimal loss is*

$$L(u_x^*, u_y^*, \{u_i^*\}_{i\in\{2,\dots,n-1\}}) = \sigma^2 + \frac{1}{D} d\sigma^2 + \frac{n^2(1+d)d}{D} + o\left(\frac{1}{D}\right).$$

Theorem D.2 *Under the assumption on the data generation process and optimal model configurations, with some derivation, the prediction of $\hat{y}_q$ can be reparameterized using $\{a_i\}_{i\in[n]}$ as:*

$$\hat{y}_q = \left(W_n W_{n-1} \dots W_2 W_1 \left(\sum_{k=1}^n a_k \frac{1}{D^k} (\sum x_i x_i^\top)^k \right) x_q \right) + \left(\sum_{k=1}^n \left(a_k \frac{1}{D^k} \sum \epsilon_i x_i (\sum x_i x_i^\top)^{k-1} \right) x_q \right),$$

where $a_1 = \frac{v_x}{v_{w_{n+1}}}$, $a_n = \frac{v_x v_{z_1} \dots v_n}{v_{w_1} v_{w_2} v_{w_{n+1}}}$ and $a_i = \frac{v_{z_{i-1}}}{v_{w_n}} a_{i-1} + \frac{v_x v_{z_1} \dots v_{z_{i-1}}}{v_{w_1} \dots v_{w_{i-1}} v_{w_n}}$ for $i \in [2, n-1]$.
The optimal expected loss of Coherent CoT is achieved when $\sum_{i=1}^n a_i$=1. Taking a special value that $a_n + a_1 = 0$ and $a_m = 0$ for $m \in [2, n-1]$, the corresponding loss becomes

$$L(\{a_i^*\}_{i\in[n]}) = \sigma^2 + \frac{1}{D} d\sigma^2 + ((1-n)a^2 + (n^2+n)a + 1 - 2a)\frac{(1+d)d}{D} + + \frac{n^2(1+d)d}{D} + o\left(\frac{1}{D}\right).$$

The proof of the two theorems can be found in Appendix D.1 and E.10. Notably, in Theorem D.2, we take special values that $a_n + a_1 = 0$ and $a_m = 0$ and derive the corresponding loss of Coherent CoT. Due to the complexity of the representation, obtaining an exact analytical expression for the minimal loss for Coherent CoT is challenging. Thus, we simplify the calculation using these specific values and leverage the following proposition to show that, under these conditions, Coherent CoT achieves a lower loss than the minimal loss of Stepwise ICL. This confirms the existence of parameter configurations where Coherent CoT outperforms optimal Stepwise ICL.

Proposition D.1 *For $n \ge 1$, there is always a specific value for a_n and a_1 such that the optimal loss of Coherent CoT is smaller than that of Stepwise ICL.*

Proposition D.1 indicates that the superiority of Coherent CoT over Stepwise ICL remain consistent when considering multi-step reasoning. Furthermore, comparing the optimal loss of Stepwise ICL and Coherent CoT presented in the two theorems, it can be observed that, when taking different values of n, the extent to which Coherent CoT outperforms Stepwise ICL depends on the specific values of d (the dimension of the data) and σ (the noise level).

Similar to the above comparison between Stepwise ICL and Coherent CoT, for the sensitivity analysis of CoT steps, we recognize that the same approach can be extended to derive results for multi-step reasoning and the insights will be the consistent to that of the two-step case.

E Proofs

E.1 Proof of Theorem 3.1:

The proof begins by decomposing the mean square loss into distinct components involving attention scores. Each term's expectation is then calculated separately, are combined to represent the total expected loss.

Recalling that for Step-Wise ICL, we first obtain $\hat{z}_q$ by

$$\begin{aligned}
\hat{z}_q &= f_1(E_{ICL}^{(1)})_{d+1,D+1} \\
&= (W_{out}W_{d+1,:}^V)E\left(\frac{1}{D}E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \end{bmatrix}\right) \\
&= (\frac{1}{Du_x})\left[z_1, z_2, \ldots, z_D, 0\right]\left(\begin{bmatrix} u_x x_1^\top x_q \\ u_x x_2^\top x_q \\ \vdots \\ u_x x_q^\top x_q \end{bmatrix}\right) \\
&= \frac{1}{D}\sum z_i(x_i^\top x_q)
\end{aligned}$$

Then $\hat{y}_q$ is obtained by

$$\begin{aligned}
\hat{y}_q &= f_2(E_{ICL}^{(2)})_{2,D+1} \\
&= (W_{out}W_{d+1,:}^V)E\left(\frac{1}{D}E^\top (W^K)^\top W^Q \begin{bmatrix} \hat{z}_q \\ 0 \end{bmatrix}\right) \\
&= (\frac{1}{Du_z})\left[y_1, y_2, \ldots, y_D, 0\right]\left(\begin{bmatrix} u_z z_1^\top \hat{z}_q \\ u_z z_2^\top \hat{z}_q \\ \vdots \\ u_z \hat{z}_q^\top \hat{z}_q \end{bmatrix}\right) \\
&= \frac{1}{D}\sum y_i(z_i^\top \hat{z}_q) = \frac{1}{D^2}\sum y_i z_i^\top \left(\sum z_i(x_i^\top x_q)\right)
\end{aligned}$$

Since $z_i = \beta^\top x_i$ and $y_i = \beta^\top x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. Then the expectation of the MSE loss can be expressed as:

$$\begin{aligned}
L(u_x, u_z) &= \mathbb{E}\left(y_q - \hat{y}_q\right)^2 \\
&= \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\left(\beta^\top x_q + \epsilon_q\right) - \frac{1}{D^2}\left(\sum\left(\beta^\top x_i + \epsilon_i\right)\left(\beta^\top x_i\right)\right)\left(\sum \beta^\top x_i(x_i^\top x_q)\right)\right)^2 \qquad (13) \\
&= \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\beta^\top x_q + \epsilon_q\right)^2 + \underbrace{\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{1}{D^2}\left(\sum\left(\beta^\top x_i + \epsilon_i\right)\left(\beta^\top x_i\right)\right)\left(\sum \beta^\top x_i(x_i^\top x_q)\right)\right)^2}_{A_1}
\end{aligned}$$

$$
\underbrace{-\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{2}{D^2}\left(\beta^\top x_q+\epsilon_q\right)\sum\left(\beta^\top x_i+\epsilon_i\right)\left(\beta^\top x_i\right)\left(\sum\beta^\top x_i(x_i^\top x_q)\right)\right)}_{A_2}
$$

$$
\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\beta^\top x_q+\epsilon_q\right)^2=\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\beta^\top x_q x_q^\top\beta+\epsilon_q^2+2\beta^\top x_q\epsilon_q\right)=\|\beta\|^2+\sigma^2
$$

$$
\begin{aligned}
A_1 &= \frac{(D-1)(D-2)(D-3)}{D^3}\mathbb{E}\left(\left(\beta^\top x_1x_1^\top\beta+\beta^\top x_1\epsilon_1\right)\left(\beta^\top x_2x_2^\top\beta+\beta^\top x_2\epsilon_2\right)\left(\beta^\top x_3x_3^\top x_q\right)\left(\beta^\top x_4x_4^\top x_q\right)\right)\\
&\quad+\frac{(D-1)(D-2)}{D^3}\mathbb{E}\left(\left(\beta^\top x_1x_1^\top\beta+\beta^\top x_1\epsilon_1\right)\left(\beta^\top x_2x_2^\top\beta+\beta^\top x_2\epsilon_2\right)\left(\beta^\top x_3x_3^\top x_q\right)^2\right)\\
&\quad+\frac{(D-1)(D-2)}{D^3}\mathbb{E}\left(\left(\beta^\top x_1x_1^\top\beta+\beta^\top x_1\epsilon_1\right)^2\left(\beta^\top x_3x_3^\top x_q\right)\left(\beta^\top x_4x_4^\top x_q\right)\right)+o\left(\frac{1}{D}\right)\\
&\quad+\frac{4(D-1)(D-2)}{D^3}\mathbb{E}\left(\left(\beta^\top x_1x_1^\top\beta+\beta^\top x_1\epsilon_1\right)\left(\beta^\top x_2x_2^\top\beta+\beta^\top x_2\epsilon_2\right)\left(\beta^\top x_1x_1^\top x_q\right)\left(\beta^\top x_3x_3^\top x_q\right)\right)\\
&= \frac{(D-1)(D-2)(D-3)}{D^3}\mathbb{E}\left(\left(\beta^\top x_1x_1^\top\beta\right)\left(\beta^\top x_2x_2^\top\beta\right)\left(\beta^\top x_3x_3^\top x_q\right)\left(\beta^\top x_4x_4^\top x_q\right)\right)\\
&\quad+\frac{(D-1)(D-2)}{D^3}\mathbb{E}\left(\left(\beta^\top x_1x_1^\top\beta\right)\left(\beta^\top x_2x_2^\top\beta\right)\left(\beta^\top x_3x_3^\top x_q\right)^2\right)\\
&\quad+\frac{(D-1)(D-2)}{D^3}\mathbb{E}\left(\left((\beta^\top x_1x_1^\top\beta)^2+\beta^\top x_1x_1^\top\beta\epsilon_1^2\right)\left(\beta^\top x_3x_3^\top x_q\right)\left(\beta^\top x_4x_4^\top x_q\right)\right)+o\left(\frac{1}{D}\right)\\
&\quad+\frac{4(D-1)(D-2)}{D^3}\mathbb{E}\left(\left(\beta^\top x_1x_1^\top\beta\right)\left(\beta^\top x_2x_2^\top\beta\right)\left(\beta^\top x_1x_1^\top x_q\right)\left(\beta^\top x_3x_3^\top x_q\right)\right)\\
&= \frac{(D-1)(D-2)(D-3)}{D^3}\|\beta\|^6+\frac{(D-1)(D-2)}{D^3}(2+d)\|\beta\|^6+\frac{(D-1)(D-2)}{D^3}\sigma^2\|\beta\|^4\\
&\quad+\frac{15(D-1)(D-2)}{D^3}\|\beta\|^6+o\left(\frac{1}{D}\right)\\
&= \|\beta\|^6+\frac{11}{D}\|\beta\|^6+\frac{d}{D}\|\beta\|^6+\frac{1}{D}\sigma^2\|\beta\|^4+o\left(\frac{1}{D}\right)
\end{aligned}
$$

$$
\begin{aligned}
A_2 &= \frac{2}{D}\mathbb{E}\left(\beta^\top x_q\beta^\top x_1x_1^\top\beta\beta^\top x_1x_1^\top x_q\right)+\frac{2D(D-1)}{D^2}\mathbb{E}\left(\beta^\top x_1x_1^\top\beta\beta^\top x_2x_2^\top\beta\right)\\
&= \frac{4}{D}\|\beta\|^4+2\|\beta\|^4
\end{aligned}
$$

Based on the results of A_1 and A_2, we have

$$
L(u_x^*,u_z^*) = \|\beta\|^2+\sigma^2+\|\beta\|^6+\frac{11}{D}\|\beta\|^6+\frac{d}{D}\|\beta\|^6+\frac{1}{D}\sigma^2\|\beta\|^4-\left(\frac{4}{D}\|\beta\|^4+2\|\beta\|^4\right)+o\left(\frac{1}{D}\right).
$$

Since $\|\beta\|^2=1$, we have

$$
L(u_x^*,u_z^*) = \sigma^2+\frac{7+d}{D}+\frac{\sigma^2}{D}+o\left(\frac{1}{D}\right).
$$

As the parameters u_x and u_z are canceled out after taking the expectation, the optimal values u_x^* and u_z^* can be any non-zero values.

E.2 Proof of Theorem 3.2.

The proof begins by decomposing the mean square loss into distinct components involving attention scores. Each term's expectation is then calculated separately, are combined to represent the total expected loss. To minimize this loss, optimal values for the parameters v_x, v_z and v_y are derived. These optimal parameters are subsequently substituted back into the expression to further refine the loss representation.

The prediction of the final response y_q can be expressed as:

$$\begin{aligned}
\hat{y}_q &= (W_{out}W^V_{d+1,:})E\left(\frac{1}{D}E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\ \hat{z}_q\\ 0\end{bmatrix}\right)\\
&= (\frac{1}{Dv_y})\left[y_1,y_2,\ldots,y_D,0\right]\left(\begin{bmatrix}v_x x_1^\top x_q+v_z z_1^\top \hat{z}_q\\ v_x x_2^\top x_q+v_z z_2^\top \hat{z}_q\\ \vdots\\ v_x x_q^\top x_q+v_z z_q^\top \hat{z}_q\end{bmatrix}\right)\\
&= (\frac{1}{Dv_y})\sum y_i(v_x x_i^\top x_q+v_z z_i^\top \hat{z}_q)
\end{aligned}$$

where

$$\begin{aligned}
\hat{z}_q &= (W_{out}W^V_{d,:})E\left(\frac{1}{D}E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\ 0\\ 0\end{bmatrix}\right)\\
&= (\frac{1}{Dv_x})\left[z_1,z_2,\ldots,z_D,0\right]\left(\begin{bmatrix}v_x x_1^\top x_q\\ v_x x_2^\top x_q\\ \vdots\\ v_x x_q^\top x_q\end{bmatrix}\right)\\
&= \frac{1}{D}\sum z_i x_i^\top x_q
\end{aligned}$$

Therefore, the expectation of the MSE loss can be expressed as:

$$\begin{aligned}
&\mathbb{E}\left(y_q-\hat{y}_q\right)^2\\
=\;&\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\left(\beta^\top x_q+\epsilon_q\right)-\frac{1}{D}\frac{v_x}{v_y}\sum(\beta^\top x_i+\epsilon_i)x_i^\top x_q-\frac{1}{D^2}\frac{v_z}{v_y}\left(\sum(\beta^\top x_i+\epsilon_i)(\beta^\top x_i)\right)\left(\sum\beta^\top x_i x_i^\top x_q\right)\right)^2\\
=\;&\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\beta^\top x_q+\epsilon_q\right)^2+\underbrace{\frac{1}{D^2}\frac{v_x^2}{v_y^2}\left(\sum(\beta^\top x_i+\epsilon_i)x_i^\top x_q\right)^2}_{A_1}\\
&+\underbrace{\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{1}{D^2}\frac{v_z}{v_y}\left(\sum\left(\beta^\top x_i+\epsilon_i\right)\left(\beta^\top x_i\right)\right)\left(\sum\beta^\top x_i(x_i^\top x_q)\right)\right)^2}_{A_2}\\
&-\underbrace{\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{2}{D}\frac{v_x}{v_y}\left(\beta^\top x_q+\epsilon_q\right)\sum\left(\beta^\top x_i+\epsilon_i\right)x_i^\top x_q\right)}_{A_3}\\
&-\underbrace{\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{2}{D^2}\left(\beta^\top x_q+\epsilon_q\right)\left(\sum\left(\beta^\top x_i+\epsilon_i\right)\left(\beta^\top x_i\right)\right)\left(\sum\beta^\top x_i(x_i^\top x_q)\right)\right)}_{A_4}\\
&+\underbrace{\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{2}{D^3}\frac{v_xv_z}{v_y^2}\left(\sum\left(\beta^\top x_i+\epsilon_i\right)x_i^\top x_q\right)\left(\sum\left(\beta^\top x_i+\epsilon_i\right)\left(\beta^\top x_i\right)\right)\left(\sum\beta^\top x_i(x_i^\top x_q)\right)\right)}_{A_5}
\end{aligned}$$

$$\mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\beta^\top x_q+\epsilon_q\right)^2=\|\beta\|^2+\sigma^2$$

$$\begin{aligned}
A_1 &= \frac{v_x^2}{v_y^2}\frac{1}{D}\mathbb{E}\left((\beta^\top x_1 x_1^\top x_q)^2 + \epsilon_i^2 (x_i^\top x_q)^2\right) + \frac{v_x^2}{v_y^2}\mathbb{E}\left(\frac{D(D-1)}{D^2}\beta^\top x_1 x_1^\top x_q x_q^\top x_2 x_2^\top \beta\right) \\
&= \frac{v_x^2}{v_y^2}\frac{(d+2)}{D}\|\beta\|^2 + \frac{v_x^2}{v_y^2}\frac{d}{D}\sigma^2 + \frac{v_x^2}{v_y^2}\frac{D(D-1)}{D^2}\|\beta\|^2 = \frac{v_x^2}{v_y^2}\frac{d}{D}\sigma^2 + \frac{v_x^2}{v_y^2}\frac{D(D+1+d)}{D^2}\|\beta\|^2
\end{aligned}$$

$$A_2 = \frac{v_z^2}{v_y^2}\|\beta\|^6 + \frac{11}{D}\frac{v_z^2}{v_y^2}\|\beta\|^6 + \frac{d}{D}\frac{v_z^2}{v_y^2}\|\beta\|^6 + \frac{1}{D}\frac{v_z^2}{v_y^2}\sigma^2\|\beta\|^4 + o\left(\frac{1}{D}\right)$$

$$A_4 = \frac{4}{D}\frac{v_z}{v_y}\|\beta\|^4 + 2\frac{v_z}{v_y}\|\beta\|^4$$

$$A_3 = \mathbb{E}\left(2\frac{v_x}{v_y}\left(\beta^\top x_q\right)\beta^\top x_1 x_1^\top x_q\right) = 2\frac{v_x}{v_y}\|\beta\|^2$$

$$\begin{aligned}
A_5 &= \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(\frac{2}{D^3}\frac{v_x v_z}{v_y^2}\left(\sum\left(\beta^\top x_i + \epsilon_i\right)x_i^\top x_q\right)\left(\sum\left(\beta^\top x_i + \epsilon_i\right)\left(\beta^\top x_i\right)\right)\left(\sum \beta^\top x_i (x_i^\top x_q)\right)\right) \\
&= \frac{2(D-1)(D-2)}{D^2}\frac{v_x v_z}{v_y^2}(\beta^\top x_1 x_1^\top x_q)(\beta^\top x_2 x_2^\top \beta)(\beta^\top x_3 x_3^\top x_q) \\
&\quad + \frac{2(D-1)}{D^2}\frac{v_x v_z}{v_y^2}(\beta^\top x_1 + \epsilon_1)^2 x_1^\top x_q \beta^\top x_1 \beta^\top x_2 x_2^\top x_q + \frac{2(D-1)}{D^2}\frac{v_x v_z}{v_y^2}(\beta^\top x_1 x_1^\top x_q)^2 \beta^\top x_2 x_2^\top \beta \\
&\quad + \frac{2(D-1)}{D^2}\frac{v_x v_z}{v_y^2}(\beta^\top x_1 x_1^\top x_q)(\beta^\top x_2 x_2^\top \beta)(\beta^\top x_2 x_2^\top x_q) + o\left(\frac{1}{D}\right) \\
&= \frac{2v_x v_z}{v_y^2}\|\beta\|^4 - \frac{3}{D}\frac{2v_x v_z}{v_y^2}\|\beta\|^4 + \frac{2}{D}\frac{v_x v_z}{v_y^2}\sigma^2\|\beta\|^2 + \frac{3}{D}\frac{2v_x v_z}{v_y^2}\|\beta\|^4 + \frac{2d+4}{D}\frac{v_x v_z}{v_y^2}\|\beta\|^4 + \frac{3}{D}\frac{2v_x v_z}{v_y^2}\|\beta\|^4 \\
&\quad + o\left(\frac{1}{D}\right) \\
&= \frac{2v_x v_z}{v_y^2}\|\beta\|^4 + \frac{2}{D}\frac{v_x v_z}{v_y^2}\sigma^2\|\beta\|^2 + \frac{2d+4}{D}\frac{v_x v_z}{v_y^2}\|\beta\|^4 + \frac{3}{D}\frac{2v_x v_z}{v_y^2}\|\beta\|^4 + o\left(\frac{1}{D}\right)
\end{aligned}$$

Based on the results of A_1 to A_5, we have

$$\begin{aligned}
\mathbb{E}\left(y_q - \hat{y}_q\right)^2 &= \|\beta\|^2 + \sigma^2 + \frac{v_x^2}{v_y^2}\frac{d}{D}\sigma^2 + \frac{v_x^2}{v_y^2}\frac{D(D+1+d)}{D^2}\|\beta\|^2 + \frac{v_z^2}{v_y^2}\|\beta\|^6 + \frac{11}{D}\frac{v_z^2}{v_y^2}\|\beta\|^6 + \frac{d}{D}\frac{v_z^2}{v_y^2}\|\beta\|^6 \\
&\quad + \frac{1}{D}\frac{v_z^2}{v_y^2}\sigma^2\|\beta\|^4 - \frac{4}{D}\frac{v_z}{v_y}\|\beta\|^4 - 2\frac{v_z}{v_y}\|\beta\|^4 - 2\frac{v_x}{v_y}\|\beta\|^2 + \frac{2v_x v_z}{v_y^2}\|\beta\|^4 + \frac{2}{D}\frac{v_x v_z}{v_y^2}\sigma^2\|\beta\|^2 \\
&\quad + \frac{2d+4}{D}\frac{v_x v_z}{v_y^2}\|\beta\|^4 + \frac{3}{D}\frac{2v_x v_z}{v_y^2}\|\beta\|^4 + o\left(\frac{1}{D}\right) \\
&= \sigma^2 + \mathbb{E}\left(\frac{\left(v_x - v_y + v_z\|\beta\|^2\right)^2\|\beta\|^2}{v_y^2} + \frac{v_x^2 d + v_z^2\|\beta\|^4 + 2v_x v_z\|\beta\|^2}{D v_y^2}\sigma^2\right) \\
&\quad + \mathbb{E}\left(\frac{1}{D}\frac{(v_x + v_z\|\beta\|^2)^2}{v_y^2}(1+d)\|\beta\|^2 + \frac{10}{D}\frac{v_z^2}{v_y^2}\|\beta\|^6 + \frac{8}{D}\frac{v_x v_z}{v_y^2}\|\beta\|^4 - \frac{4}{D}\frac{v_z}{v_y}\|\beta\|^4\right) + o(\frac{1}{D})
\end{aligned}$$

If we have $\|\beta\|^2 = 1$, it becomes

$$\mathbb{E}\left(y_q - \hat{y}_q\right)^2$$

$$= \left(\frac{(v_x - v_y + v_z)^2}{v_y^2} + \sigma^2 + \frac{(v_x + v_z)^2}{v_y^2} (\frac{1+d}{D}) - \frac{4}{D} \frac{v_z v_y}{v_y^2} + \frac{8}{D} \frac{v_x v_z}{v_y^2} + \frac{10}{D} \frac{v_z^2}{v_y^2} + \frac{v_x^2 d + v_z^2 + 2 v_x v_z}{D v_y^2} \sigma^2 \right) + o(\frac{1}{D})$$

To minimize the loss, we need to let $(v_x - v_y + v_z)^2 = 0$, meaning that $v_y = (v_x + v_z)$

Then the loss becomes

$$\begin{aligned} \mathbb{E}\left(y_q - \hat{y}_q\right)^2 &= \sigma^2 + \frac{d\sigma^2}{D} + \frac{1+d}{D} - \frac{v_z(2v_x + v_z)}{D v_y^2}(d-1)\sigma^2 + \frac{4}{D}\frac{v_x v_z}{v_y^2} + \frac{6}{D}\frac{v_z^2}{v_y^2} \\ &= \sigma^2 + \frac{d\sigma^2}{D} + \frac{1+d}{D} - \frac{(2v_y - v_z)v_z}{D v_y^2}(d-1)\sigma^2 + \frac{4}{D}\frac{(v_y - v_z)v_z}{v_y^2} + \frac{6}{D}\frac{v_z^2}{v_y^2} \end{aligned}$$

Letting θ represents $\frac{v_x}{v_y}$, the expectation can be further expressed as a function of θ:

$$\mathbb{E}\left(y_q - \hat{y}_q\right)^2 = f(\theta) = \sigma^2 + \frac{d\sigma^2}{D} + \frac{1+d}{D} - \frac{(2\theta - \theta^2)}{D}(d-1)\sigma^2 + \frac{4}{D}(\theta - \theta^2) + \frac{6}{D}\theta^2$$

Then we have

$$f'(\theta) = \frac{(2\theta - 2)}{D}(d-1)\sigma^2 + \frac{1}{D}(4\theta + 4)$$

Letting $f'(\theta) = 0$, we have

$$\theta = \frac{(d-1)\sigma^2 - 2}{(d-1)\sigma^2 + 2}$$

Therefore, the optimal $\frac{v_z}{v_y}$ satisfies $\frac{v_z^*}{v_y^*} = \frac{(d-1)\sigma^2 - 2}{(d-1)\sigma^2 + 2}$. Then we further have $\frac{v_x^*}{v_y^*} = 1 - \frac{v_z}{v_y} = \frac{4}{(d-1)\sigma^2 + 2}$. Substituting the optimal parameters back to the formula of the expected loss, we have

$$\begin{aligned} L(v_x^*, v_y^*, v_z^*) &= \mathbb{E}\left(y_q - \hat{y}_q\right)^2 = \sigma^2 + \frac{d\sigma^2}{D} + \frac{1+d}{D} - \frac{\left((d-1)^3\sigma^6 - 2(d-1)^2\sigma^4 - 4(d-1)\sigma^2 + 12\right)}{((d-1)\sigma^2 + 2)^2} + o\left(\frac{1}{D}\right) \\ &= \sigma^2 + \frac{d\sigma^2}{D} + \frac{1+d}{D} - \frac{\left((d-1)\sigma^2 - 2\right)^2}{D((d-1)\sigma^2 + 2)} + o\left(\frac{1}{D}\right). \end{aligned}$$

E.3 Proof of Proposition 3.1.

Recalling that the loss of Coherent CoT is written as:

$$\begin{aligned} & L(v_x, v_z, v_y) \\ = \;& \mathbb{E}_{\{x_i\}_{i\in[D]}} \left((\beta^\top x_q + \epsilon_q) - \frac{1}{D}\frac{v_x}{v_y} \sum (\beta^\top x_i + \epsilon_i) x_i^\top x_q - \frac{1}{D^2}\frac{v_z}{v_y} \left(\sum (\beta^\top x_i + \epsilon_i)(\beta^\top x_i)\right) \left(\sum \beta^\top x_i x_i^\top x_q\right) \right)^2 \end{aligned}$$

When setting $v_x = 0$, $v_y = v_z$, we have

$$L(v_x, v_z, v_y) = \mathbb{E}_{\{x_i\}_{i\in[D]}} \left((\beta^\top x_q + \epsilon_q) - \frac{1}{D^2} \left(\sum (\beta^\top x_i + \epsilon_i)(\beta^\top x_i)\right) \left(\sum \beta^\top x_i x_i^\top x_q\right) \right)^2$$

which is equivalent to the Eq.(13) in the proof of Theorem 3.1. This indicates that when setting $v_x = 0$, $v_y = v_z$, Coherent is equivalent to Step-wise CoT. Moreover, since $(v_x = 0, v_z = v_y)$ is not the optimal solution of Coherent CoT, Coherent CoT can achieve a smaller loss given its corresponding optimal solution. Therefore, we can conclude that the minimal expected loss of Coherent CoT must be smaller than the one of Stepwise ICL.

E.4 Proof of Theorem 3.3.

The proof of Theorem 3.3 to Theorem 3.5 is similar to that of Theorem 3.2, with the only difference being the presence of additional terms resulting from the added noise. To save space, we present only the additional terms introduced by the perturbation and denote the remaining terms as $L(v_x, v_y, v_z)$.

When there is random perturbation $\delta_i \sim N(0, \sigma_\epsilon^2)$ added to y_i, the loss becomes

$$
\begin{aligned}
& L'_{\mathrm{y}}(v_x, v_y, v_z) = \mathbb{E}\,(y_q - \hat{y}_q)^2 \\
= \quad & \mathbb{E}_{\{x_i\}_{i\in[D]}}\Bigg((\beta^\top x_q + \epsilon_q) - \frac{1}{D}\frac{v_x}{v_y}\sum(\beta^\top x_i + \epsilon_i + \delta_i) x_i^\top x_q \\
& \qquad -\frac{1}{D^2}\frac{v_z}{v_y}\left(\sum(\beta^\top x_i + \epsilon_i + \delta_i)(\beta^\top x_i)\right)\left(\sum \beta^\top x_i x_i^\top x_q\right)\Bigg)^2 \\
= \quad & L(v_x, v_z, v_y) + \mathbb{E}\left(\frac{1}{D^2}\frac{v_x^2}{v_y^2}(\delta_i x_i^\top x_q)^2 + \frac{1}{D^4}\left(\sum \delta_i \beta^\top x_i \sum \beta^\top x_i x_i^\top x_q\right)^2\right) \\
& + \mathbb{E}\left(\frac{2}{D^3}\frac{v_x v_z}{v_y^2}\sum \delta_i x_i^\top x_q \sum \delta_i \beta^\top x_i \sum \beta^\top x_i x_i^\top x_q\right) + o\left(\frac{1}{D}\right) \\
= \quad & L(v_x, v_z, v_y) + \mathbb{E}\left(\frac{1}{D}\frac{v_x^2}{v_y^2}\delta_1^2 x_i^\top x_q x_q^\top x_i + \frac{(D-1)(D-2)}{D^3}\frac{v_z^2}{v_y^2}\delta_1^2 \beta^\top x_1 x_1^\top \beta \beta^\top x_2 x_2^\top x_q \beta^\top x_3 x_3^\top x_q\right) \\
& + 2\mathbb{E}\left(\frac{(D-1)}{D^2}\frac{2 v_x v_z}{v_y}\delta_1^2 x_1^\top x_q \beta^\top x_1 \beta^\top x_2 x_2^\top x_q\right) + o\left(\frac{1}{D}\right) \\
= \quad & L(v_x, v_z, v_y) + \frac{1}{D}\left(\frac{v_x - v_y + v_z}{v_y}\right)^2 \sigma_\epsilon^2 + o\left(\frac{1}{D}\right) \\
= \quad & L(v_x, v_z, v_y) + o\left(\frac{1}{D}\right)
\end{aligned}
$$

E.5 Proof of Theorem 3.4.

When there is random perturbation $\delta_i \sim N(0, \sigma_\epsilon^2)$ added to x_i, the loss becomes

$$
L'_{\mathrm{x}}(v_x, v_y, v_z) = \mathbb{E}\,(y_q - \hat{y}_q)^2 \quad = \quad \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q - \frac{1}{D}\frac{v_x}{v_y}\sum y_i (x_i + \delta_i)^\top x_q - \frac{1}{D}\frac{v_z}{v_y}\sum y_i z_i \hat{z}_q\right)^2,
$$

where

$$
\hat{z}_q = \frac{1}{D}\sum z_i (x_i^\top + \delta_i^\top) x_q.
$$

Therefore, the loss can be written as:

$$
\begin{aligned}
& L'_{\mathrm{x}}(v_x, v_y, v_z) = \mathbb{E}\,(y_q - \hat{y}_q)^2 \\
= \quad & \mathbb{E}_{\{x_i\}_{i\in[D]}}\Bigg((\beta^\top x_q + \epsilon_q) - \frac{1}{D}\frac{v_x}{v_y}\sum(\beta^\top x_i + \epsilon_i)(x_i^\top + \delta_i^\top) x_q \\
& \qquad -\frac{1}{D^2}\frac{v_z}{v_y}\left(\sum(\beta^\top x_i + \epsilon_i)\beta^\top x_i\right)\left(\sum \beta^\top x_i (x_i^\top + \delta_i^\top) x_q\right)\Bigg)^2 \\
= \quad & L(v_x, v_z, v_y) + \mathbb{E}\left(\frac{1}{D}\frac{v_x}{v_y}\sum \beta^\top x_i \delta_i^\top x_q\right)^2 + \mathbb{E}\left(\frac{1}{D}\frac{v_x}{v_y}\sum \epsilon_i \delta_i^\top x_q\right)^2 + \left(\frac{1}{D}\frac{v_z}{v_y}\sum \beta^\top x_i z_i^\top (\frac{1}{D}\sum \beta^\top x_i \delta_i^\top x_q)\right)^2 \\
& + 2\mathbb{E}\left[\left(\frac{1}{D^2}\frac{v_x v_z}{v_y}\sum \beta^\top x_i \delta_i^\top x_q\right)\left(\sum \beta^\top x_i z_i^\top (\frac{1}{D}\sum \beta^\top x_i \delta_i^\top x_q)\right)\right]^2 + o(\frac{1}{D})
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \left(\frac{1+d}{D}+\frac{4}{D}\frac{v_xv_z}{v_y^2}+\frac{6}{D}\frac{v_z^2}{v_y^2}+\sigma^2+\frac{d\sigma^2}{D}-\frac{v_z(2v_x+v_z)}{Dv_y^2}(d-1)\sigma^2\right)\\
& +\mathbb{E}\left(\frac{1}{D}\frac{v_x^2}{v_y^2}(\beta^\top x_1\epsilon_1^\top x_q)^2+\frac{1}{D}\frac{v_x^2}{v_y^2}(\delta_1\epsilon_1^\top x_q)^2+\frac{2}{D}\frac{v_xv_z}{v_y^2}\beta^\top x_1x_1^\top\beta(\beta^\top x_2\epsilon_2^\top x_q)^2\right)\\
& +\mathbb{E}\left(\frac{1}{D}\frac{v_z^2}{v_y^2}(\beta^\top x_1x_1^\top\beta)(\beta^\top x_2x_2^\top\beta)(\beta^\top x_3\epsilon_3^\top x_q)^2\right)+o(\frac{1}{D})\\
= \quad & \frac{1+d}{D}+\frac{4}{D}\frac{v_xv_z}{v_y^2}+\frac{6}{D}\frac{v_z^2}{v_y^2}+\sigma^2+\frac{d\sigma^2}{D}-\frac{v_z(2v_x+v_z)}{Dv_y^2}(d-1)\sigma^2+\frac{d}{D}\frac{(v_x+v_z)^2}{v_y^2}\sigma_\epsilon^2+\frac{v_x^2}{v_y^2}\sigma^2\sigma_\epsilon^2+o(\frac{1}{D})\\
= \quad & \frac{1+d}{D}+\frac{4}{D}\frac{v_xv_z}{v_y^2}+\frac{6}{D}\frac{v_z^2}{v_y^2}+\frac{d}{D}\sigma_\epsilon^2+\sigma^2+\frac{d\sigma^2}{D}-\frac{v_z(2v_x+v_z)}{Dv_y^2}(d-1)\sigma^2+\frac{v_x^2}{v_y^2}\sigma^2\sigma_\epsilon^2+o\left(\frac{1}{D}\right)
\end{aligned}
$$

E.6 Proof of Theorem 3.5.

When there is random perturbation $\delta_i \sim N(0,\sigma_\epsilon^2)$ added to z_i, the loss becomes

$$
L_z'(v_x,v_y,v_z)=\mathbb{E}\left(y_q-\hat{y}_q\right)^2 \quad = \quad \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(y_q-\frac{1}{D}\frac{v_x}{v_y}\sum y_ix_i^\top x_q-\frac{1}{D}\frac{v_z}{v_y}\sum y_i(z_i+\delta_i)\hat{z}_q\right)^2,
$$

where

$$
\hat{z}_q=\frac{1}{D}\sum(z_i+\delta_i)x_i^\top x_q.
$$

Therefore, the loss can be written as:

$$
\begin{aligned}
& L_z'(v_x,v_y,v_z)=\mathbb{E}\left(y_q-\hat{y}_q\right)^2\\
= \quad & \mathbb{E}_{\{x_i\}_{i\in[D]}}\Bigg((\beta^\top x_q+\epsilon_q)-\frac{1}{D}\frac{v_x}{v_y}\sum(\beta^\top x_i+\epsilon_i)x_i^\top x_q\\
& \qquad -\frac{1}{D^2}\frac{v_z}{v_y}\left(\sum(\beta^\top x_i+\epsilon_i)(\beta^\top x_i+\delta_i)\right)\left(\sum(\beta^\top x_i+\delta_i)x_i^\top x_q\right)\Bigg)^2\\
= \quad & L(v_x,v_y,v_z)-2y_q\left[\frac{1}{D^2}\frac{v_z}{v_y}\sum\left(\beta^\top x_i\delta_i(\sum\delta_ix_i^\top x_q)\right)\right]\\
& +2\left[\frac{1}{D^2}\frac{v_xv_z}{v_y^2}(\sum\beta^\top x_ix_i^\top x_q)\sum\left(\beta^\top x_i\delta_i(\sum\delta_ix_i^\top x_q)\right)\right]+\frac{1}{D^4}\frac{v_z^2}{v_y^2}\left(\left(\sum\epsilon_i\delta_i\right)^2\left(\sum(\beta^\top x_i+\delta_i)x_i^\top x_q\right)^2\right)\\
& +\frac{1}{D^2}\frac{v_z^2}{v_y^2}\left(\mathbb{E}\left[\sum\left(\beta^\top x_i(z_i+\delta_i)^\top\sum(z_i+\delta_i)x_i^\top x_q\right)\right]^2-\mathbb{E}\left[\sum\left(\beta^\top x_iz_i\sum z_ix_i^\top x_q\right)\right]^2\right)+o(\frac{1}{D})\\
= \quad & \left(\sigma^2+\frac{d\sigma^2}{D}+\frac{1+d}{D}-\frac{(2v_y-v_z)v_z}{Dv_y^2}(d-1)\sigma^2+\frac{4}{D}\frac{v_xv_z}{v_y^2}+\frac{6}{D}\frac{v_z^2}{v_y^2}\right)\\
& -\mathbb{E}\left(\frac{2}{D}\frac{v_zv_y}{v_y^2}\delta_1^2\beta^\top x_q\beta^\top x_1x_1^\top x_q+\frac{2}{D}\frac{v_xv_z}{v_y^2}\beta^\top x_1x_1^\top x_q\beta^\top x_2x_2^\top x_q\delta_2^2\right)\\
& +\mathbb{E}\left(\frac{1}{D}\frac{v_z^2}{v_y^2}\epsilon_1^2\delta_1^2\beta^\top x_2x_2^\top x_q\beta^\top x_3x_3^\top x_q+\frac{1}{D}\frac{v_z^2}{v_y^2}\beta^\top x_1x_1^\top\beta\delta_1^2\beta^\top x_2x_2^\top x_q\beta^\top x_3x_3^\top x_q\right)\\
& +\mathbb{E}\left(\frac{1}{D}\frac{v_z^2}{v_y^2}\beta^\top x_1x_1^\top\beta\beta^\top x_2x_2^\top\beta\delta_3^2x_3^\top x_qx_q^\top x_3+\frac{4}{D}\frac{v_z^2}{v_y^2}\delta_1^2\beta^\top x_1x_1^\top x_q\beta^\top x_2x_2^\top\beta\beta^\top x_3x_3^\top x_q\right)+o\left(\frac{1}{D}\right)\\
= \quad & \sigma^2+\frac{d\sigma^2}{D}+\frac{1+d}{D}-\frac{(2v_y-v_z)v_z}{Dv_y^2}(d-1)\sigma^2+\frac{4}{D}\frac{v_xv_z}{v_y^2}+\frac{6}{D}\frac{v_z^2}{v_y^2}-\frac{2}{D}\frac{v_zv_y}{v_y^2}\sigma_\epsilon^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{2}{D}\frac{v_z v_x}{v_y^2}\sigma_\epsilon^2+\frac{1}{D}\frac{v_z^2}{v_y^2}\sigma_\epsilon^2+\frac{4}{D}\frac{v_z^2}{v_y^2}\sigma_\epsilon^2+\frac{d}{D}\frac{v_z^2}{v_y^2}\sigma_\epsilon^2+\frac{v_z^2}{v_y^2}\sigma^2\sigma_\epsilon^2+o\left(\frac{1}{D}\right)\\
=\;&\frac{1+d}{D}+\frac{4}{D}\frac{v_x v_z}{v_y^2}+\frac{6}{D}\frac{v_z^2}{v_y^2}+\frac{3+d}{D}\frac{v_z^2}{v_y^2}\sigma_\epsilon^2+\sigma^2+\frac{d\sigma^2}{D}-\frac{v_z(2v_x+v_z)}{Dv_y^2}(d-1)\sigma^2+\frac{v_z^2}{v_y^2}\sigma^2\sigma_\epsilon^2+o\left(\frac{1}{D}\right).
\end{aligned}
$$

E.7 Proof of Proposition 3.3.

Because $L'_{\mathrm{x}}(v_x,v_y,v_z)=O\left(\frac{1}{D}\right)$, $L'_{\mathrm{z}}(v_x,v_y,v_z)=O\left(\frac{1}{D}\right)$ and $L'_{\mathrm{y}}(v_x,v_y,v_z)=o\left(\frac{1}{D}\right)$. We can conclude that

- $L'_{\mathrm{x}}(v_x,v_y,v_z)>L'_{\mathrm{y}}(v_x,v_y,v_z)$
- $L'_{\mathrm{z}}(v_x,v_y,v_z)>L'_{\mathrm{y}}(v_x,v_y,v_z)$

To compare $L'_{\mathrm{z}}(v_x,v_y,v_z)$ and $L'_{\mathrm{x}}(v_x,v_y,v_z)$, we calculate:

$$
\begin{aligned}
&L'_{\mathrm{z}}(v_x,v_y,v_z)-L'_{\mathrm{x}}(v_x,v_y,v_z)\\
&=\frac{3+d}{D}\frac{v_z^2}{v_y^2}\sigma_\epsilon^2-\frac{d}{D}\frac{v_x^2}{v_y^2}\sigma_\epsilon^2+\frac{v_z^2-v_x^2}{v_y^2}\sigma^2\sigma_\epsilon^2
\end{aligned}
$$

Substituting $v_y^*=\frac{(d-1)\sigma^2+2}{(d-1)\sigma^2-2}v_z^*$ and $v_y^*=v_x^*+v_z^*$, we have

$$
\begin{aligned}
&L'_{\mathrm{z}}(v_x^*,v_y^*,v_z^*)-L'_{\mathrm{x}}(v_x^*,v_y^*,v_z^*)\\
&=\frac{3+d}{D}\frac{v_z^2}{v_y^2}\sigma_\epsilon^2-\frac{d}{D}\sigma_\epsilon^2+\frac{v_z^2-v_x^2}{v_y^2}\sigma^2\sigma_\epsilon^2\\
&=\left[\frac{3+d}{D}\frac{\left((d-1)\sigma^2-2\right)^2}{\left((d-1)\sigma^2+2\right)^2}-\frac{d}{D}+\frac{\left((d-1)\sigma^2-2\right)^2\sigma^2-16\sigma^2}{\left((d-1)\sigma^2+2\right)^2D}\right]\sigma_\epsilon^2\\
&=\frac{\sigma_\epsilon^2}{D}\frac{(d-1)^2\sigma^6+(3d-7)(d-1)\sigma^4-\left(4d+8d^2\right)\sigma^2+12}{\left((d-1)\sigma^2+2\right)^2}
\end{aligned}
$$

Given that the numerator is always positive, we can conclude that when $\left((d-1)^2\sigma^6+(3d-7)(d-1)\sigma^4-\left(4d+8d^2\right)\sigma^2+12\right)$ is positive, $L'_{\mathrm{z}}(v_x,v_y,v_z)>L'_{\mathrm{x}}(v_x,v_y,v_z)$. Therefore, considering that $b>a$ are the positive roots of $f(\theta)=(d-1)^2\theta^3+(3d-7)(d-1)\theta^2-\left(4d+8d^2\right)\theta+12=0$. We have:

- $L'_{\mathrm{x}}(v_x^*,v_y^*,v_z^*)>L'_{\mathrm{z}}(v_x^*,v_y^*,v_z^*)$ for $\sigma^2\in[0,a)\cup(b,\infty)$
- $L'_{\mathrm{x}}(v_x^*,v_y^*,v_z^*)<L'_{\mathrm{z}}(v_x^*,v_y^*,v_z^*)$ for $\sigma^2\in(a,b)$.

E.8 Proof Theorem 3.6

Given that

$$
E=\begin{pmatrix} x_1 & x_1 & \cdots & x_{D/2} & x_{D/2} & x_q\\ z_1+\delta_1 & z_1-\delta_1 & \cdots & z_{D/2}+\delta_{D/2} & z_{D/2}-\delta_{D/2} & \hat{z}_q\\ y_1 & y_1 & \cdots & y_{D/2} & y_{D/2} & 0 \end{pmatrix}, \tag{14}
$$

The prediction of the final response y_q can be expressed as:

$$
\begin{aligned}
\hat{y}_q\;&=\;(W_{out}W^V_{d+1,:})E\left(\frac{1}{D}E^\top(W^K)^\top W^Q\begin{bmatrix}x_q\\ \hat{z}_q\\ 0\end{bmatrix}\right)\\
&=\;\left(\frac{1}{Dv_y}\right)\left[y_1,y_1,\ldots,y_{D/2},y_{D/2},0\right]\left(\begin{bmatrix}v_x x_1^\top x_q+v_z(z_1+\delta_1)^\top\hat{z}_q\\ v_x x_1^\top x_q+v_z(z_1-\delta_1)^\top\hat{z}_q\\ \vdots\\ v_x x_q^\top x_q+v_z z_q^\top\hat{z}_q\end{bmatrix}\right)
\end{aligned}
$$

$$= \quad (\frac{2}{Dv_y})\sum_{i=1}^{D/2} y_i(v_x x_i^\top x_q + v_z z_i^\top \hat{z}_q)$$

where

$$\begin{aligned}
\hat{z}_q &= (W_{out}W^V_{d,:})E\left(\frac{1}{D}E^\top (W^K)^\top W^Q \begin{bmatrix} x_q \\ 0 \\ 0 \end{bmatrix}\right) \\
&= (\frac{1}{Dv_x})\left[(z_1+\delta_1),(z_1-\delta_1),\dots,(z_{D/2}+\delta_{D/2}),(z_{D/2}-\delta_{D/2}),0\right]\left(\begin{bmatrix} v_x x_1^\top x_q \\ v_x x_1^\top x_q \\ \vdots \\ v_x x_q^\top x_q \end{bmatrix}\right) \\
&= \frac{2}{D}\sum_{i=1}^{D/2}\beta^\top x_i x_i^\top x_q
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left(y_q-\hat{y}_q\right)^2 \\
= \;& \mathbb{E}_{\{x_i\}_{i\in[D]}}\left((\beta^\top x_q+\epsilon_q)-\left(\frac{2}{Dv_y}\right)\sum_{i=1}^{D/2} y_i\left(v_x x_i^\top x_q + v_z z_i^\top\left(\frac{2}{D}\sum_{i=1}^{D/2}\beta^\top x_i x_i^\top x_q\right)\right)\right)^2
\end{aligned}$$

The formula is similar to the loss derived in the proof of Theorem 3.2, with two key differences: (1) the summation now runs from $i=1$ to $i=D/2$ instead of from $i=1$ to $i=D$, and (2) the fraction $\frac{1}{D}$ is replaced by $\frac{2}{D}$. Using a similar idea to the proof of Theorem 3.2, we can obtain that

$$\mathbb{E}\left(y_q-\hat{y}_q\right)^2 = \sigma^2+\frac{2d\sigma^2}{D}+\frac{2+2d}{D}-\frac{2v_z(2v_x+v_z)}{Dv_y^2}(d-1)\sigma^2+\frac{8}{D}\frac{v_xv_z}{v_y^2}+\frac{12}{D}\frac{v_z^2}{v_y^2}+o\left(\frac{1}{D}\right).$$

E.9 Proof of Theorem D.1

Given that $y_q = W_n...W_2W_1x_q+\epsilon_q$ and $\hat{y}_q = \frac{1}{D^n}\left(W_n...W_2W_1(\sum x_ix_i^\top)^n x_q + \sum \epsilon_i x_i^\top(\sum x_ix_i^\top)^{n-1}x_q\right)$, the MSE loss of Step-wise ICL can be written as:

$$\begin{aligned}
&\mathbb{E}\left(y_q-\hat{y}_q\right)^2 \\
= \;& \mathbb{E}_{\{x_i\}_{i\in[D]}}\left((W_n...W_2W_1x_q+\epsilon_q)-\frac{1}{D^n}\left(W_n...W_2W_1(\sum x_ix_i^\top)^n x_q+\sum\epsilon_ix_i^\top(\sum x_ix_i^\top)^{n-1}x_q\right)\right)^2 \\
= \;& \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(W_n...W_2W_1x_q+\epsilon_q\right)^2+\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{1}{D^{2n}}\left(W_n...W_2W_1(\sum x_ix_i^\top)^n x_q\right)^2 \\
&+\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{1}{D^{2n}}\left(\sum\epsilon_ix_i^\top(\sum x_ix_i^\top)^{n-1}x_q\right)^2 \\
&-\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{2}{D^n}\left(W_n...W_2W_1x_q+\epsilon_q\right)\left(W_n...W_2W_1(\sum x_ix_i^\top)^n x_q+\sum\epsilon_ix_i^\top(\sum x_ix_i^\top)^{n-1}x_q\right) \\
= \;& \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(x_q^\top W_1^\top W_2^\top...W_n^\top W_n...W_2W_1x_q+\epsilon_q^2+2W_n^\top W_n...W_2W_1x_q\epsilon_q\right) \\
&+\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{1}{D^{2n}}\left(x_q^\top(\sum x_ix_i^\top)^n W_1^\top W_2^\top...W_n^\top W_n...W_2W_1(\sum x_ix_i^\top)^n x_q\right) \\
&+\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{1}{D^{2n}}\left((\sum\epsilon_ix_i^\top)(\sum x_ix_i^\top)^{n-1}x_qx_q^\top(\sum x_ix_i^\top)^{n-1}(\sum\epsilon_ix_i)\right) \\
&-\mathbb{E}_{\{x_i\}_{i\in[D]}}\frac{2}{D^n}\left(x_q^\top W_1^\top W_2^\top...W_n^\top W_n...W_2W_1(\sum x_ix_i^\top)^n x_q\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(x_q^\top x_q + \epsilon_q^2 + \frac{1}{D^{2n}} x_q^\top (\sum x_i x_i^\top)^{2n} x_q\right) + \left(\frac{1}{D^{2n}} (\sum \epsilon_i x_i^\top)^2 (\sum x_i x_i^\top)^{2n-2}\right) \\
&\quad - \mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{2}{D^n}\left(x_q^\top (\sum x_i x_i^\top)^n x_q\right)
\end{aligned}
$$

Considering that

$$
\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i\in[D]}}(\frac{1}{D^n}\sum x_i x_i^\top)^n \\
=\ & \frac{D^2 - (n(n-1)/2)\, D}{D^2}\left(\mathbb{E}_{x_1}\left(x_1 x_1^\top\right)\right)^n + \frac{(n(n-1)/2)\, D}{D^2}\left(\mathbb{E}_{x_1} x_1 x_1^\top\right)^2 \left(\mathbb{E}_{x_1}\left(x_2 x_2^\top\right)\right)^{n-1} + o\left(\frac{1}{D}\right) \\
=\ & d + \frac{n(n-1)}{2}\frac{d(1+d)}{D} + o\left(\frac{1}{D}\right)
\end{aligned}
$$

The loss can be simplified to

$$
\begin{aligned}
\mathbb{E}\left(y_q - \hat{y}_q\right)^2 &= d + \sigma^2 + \left(d + \frac{(2n-1)n}{D}(1+d)d\right) + \frac{d}{D}\sigma^2 - 2\left(d + \frac{(n-1)n}{2D}(1+d)d\right) + o\left(\frac{1}{D}\right) \\
&= \sigma^2 + \frac{1}{D} d\sigma^2 + \frac{n^2(1+d)d}{D} + o\left(\frac{1}{D}\right)
\end{aligned}
$$

E.10 Proof of Theorem D.2

In multi-step Coherent CoT,

$$
\begin{aligned}
\hat{y}_q &= \frac{1}{D^2}\left(\sum y_i z_{ni}^\top \left(\cdots \left(\sum z_{3i} z_{2i}^\top \left(\sum z_{2i} z_{1i}^\top \left(\sum z_{1i}(x_i^\top x_q)\right)\right)\right)\right)\right) \\
&= \frac{1}{D^2}\left(\sum W_n x_i x_i^\top W_{n-1}\left(\cdots\left(\sum W_3 W_2 W_1 x_i x_i^\top W_1^\top W_2^\top \left(\sum W_2 W_1 x_i x_i^\top W_1^\top \left(\sum W_1 x_i x_i^\top x_q\right)\right)\right)\right)\right) \\
&= \left(W_n W_{n-1} ... W_2 W_1 \left(a_n \frac{1}{D^n}(\sum x_i x_i^\top)^n + \cdots + a_2 \frac{1}{D^2}(\sum x_i x_i^\top)^2 + a_1 \frac{1}{D}(\sum x_i x_i^\top)\right) x_q\right) \\
&\quad + \left(\left(a_n \frac{1}{D^n}\sum \epsilon_i x_i (\sum x_i x_i^\top)^{n-1} + \cdots + a_2 \frac{1}{D^2}\sum \epsilon_i x_i (\sum x_i x_i^\top) + a_1 \frac{1}{D}\sum \epsilon_i x_i^\top\right) x_q\right).
\end{aligned}
$$

Take a special case that $a_m = 0$ for $m \in [2, n-1]$, it becomes

$$
\begin{aligned}
\hat{y}_q &= \left(W_n W_{n-1} ... W_2 W_1 \left(a_n \frac{1}{D^n}(\sum x_i x_i^\top)^n + a_1 \frac{1}{D}(\sum x_i x_i^\top)\right) x_q\right) \\
&\quad + \left(\left(a_n \frac{1}{D^n}\sum \epsilon_i x_i (\sum x_i x_i^\top)^{n-1} + a_1 \frac{1}{D}\sum \epsilon_i x_i\right) x_q\right)
\end{aligned}
$$

Then the MSE loss is written as

$$
\begin{aligned}
&\mathbb{E}\left(y_q - \hat{y}_q\right)^2 \\
=\ & \mathbb{E}_{\{x_i\}_{i\in[D]}}\left(W_n ... W_2 W_1^\top x_q + \epsilon_q\right)^2 + \mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{a_n^2}{D^{2n}}\left(W_n ... W_2 W_1 (\sum x_i x_i^\top)^n x_q\right)^2 \\
&+ \mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{a_n^2}{D^{2n}}\left(\sum \epsilon_i x_i^\top (\sum x_i x_i^\top)^{n-1} x_q\right)^2 \\
&+ \mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{a_1^2}{D^2}\left(W_n ... W_2 W_1 (\sum x_i x_i^\top) x_q\right)^2 + \frac{a_1^2}{D^2}\left(\sum \epsilon_i x_i^\top x_q\right)^2 \\
&+ \mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{2 a_1 a_n}{D^{n+1}}\left(W_n ... W_2 W_1 (\sum x_i x_i^\top)^n x_q\right)\left(W_n ... W_2 W_1 (\sum x_i x_i^\top) x_q\right) \\
&+ \mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{2 a_1 a_n}{D^{n+1}}\left(\epsilon_i x_i^\top (\sum x_i x_i^\top)^{n-1} x_q\right)\left(\sum \epsilon_i x_i^\top x_q\right) \\
&- \mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{2 a_n}{D^n}\left(W_n ... W_2 W_1 x_q\right)\left(W_n ... W_2 W_1 (\sum x_i x_i^\top)^n x_q\right)
\end{aligned}
$$

$$-\mathbb{E}_{\{x_i\}_{i\in[D]}} \frac{2a_1}{D^n} (W_n...W_2W_1x_q) \left(W_n...W_2W_1(\sum x_i x_i^\top)x_q\right)$$

Considering that $a_1 + a_n = 1$ (i.e., $a_1 = 1 - a_n$), similar to the Proof of Theorem D.1, the loss can be simplified to

$$\mathbb{E}\left(y_q - \hat{y}_q\right)^2 = \sigma^2 + \frac{1}{D}d\sigma^2 + ((1-n)a^2 + (n^2+n)a + 1 - 2a)\frac{(1+d)d}{D} + + \frac{n^2(1+d)d}{D} + o\left(\frac{1}{D}\right).$$

F Additional Experiments

F.1 Impact of Error-Aware Demonstrations Beyond Increased Reasoning Chains

In this subsection, we conduct additional experiments to demonstrate that the effectiveness of the proposed Error-Aware Demonstration Method stems from explicitly incorporating incorrect reasoning paths with corrections, rather than merely increasing the number of reasoning chains. Specifically, we perform experiments where we double the number of examples in the W/ IR scenario so that the total number of reasoning paths in both w/o IR and w IR settings are the same. The results, denoted as "w/o IR+", are presented in the following Table 5, with a comparison of the results of the original settings, 'w/o IR" and 'w IR". From the results, we observe that simply doubling the number of reasoning paths (w/o IR+) is insufficient to improve the performance of CoT reasoning. In some cases, such as the GSM8K dataset, increasing the number of examples and reasoning paths even led to decreased accuracy. In contrast, incorporating incorrect reasoning paths in demonstrations (w IR) consistently improved the model's performance and outperformed the w/o IR+ scenario. These findings validate that the proposed Error-Aware Demonstration method derives its effectiveness from explicitly including incorrect reasoning paths, rather than merely increasing the quantity of reasoning chains. Therefore, while Schaeffer et al. (2023) indicates that feeding a model with incorrect reasoning chains can already yield performance improvements, our results show that combining correct and incorrect reasoning paths can provide additional benefits.

Table 5: Comparison of model performance under different demonstration settings.

Dataset	Method	GPT-3.5-Turbo	Gemini Pro
Disambiguation QA	w/o IR	68.00%	68.80%
	w/o IR+	68.40%	66.40%
	w/ IR	**72.00%**	**76.80%**
Tracking Shuffled Objects	w/o IR	56.50%	58.80%
	w/o IR+	58.00%	58.40%
	w/ IR	**61.20%**	**64.80%**
Penguins in a table	w/o IR	81.34%	82.19%
	w/o IR+	81.51%	81.51%
	w/ IR	**82.19%**	**83.56%**
GSM8K	w/o IR	81.03%	80.82%
	w/o IR+	77.56%	78.23%
	w/ IR	**83.38%**	**81.27%**

F.2 Effect of Demonstration Quantity and Impact on Token Usage and Efficiency

In this subsection, we provide additional experiments to examine (1) the impact of demonstration quantity on reasoning performance, and (2) the impact of Error-Aware Demonstrations on token usage and efficiency.

Impact of Demonstration Quantity. In Table 6, we provide results of "w/o IR", "w/o IR+" and "w IR" when there are different numbers of examples (we also show the number of reasoning paths for each setting). From the table, we can observe that, across different settings, "w IR" consistently demonstrates the best reasoning capability, which indicates the benefit of combining correct and incorrect reasoning paths in CoT demonstration.

Impact of Error-Aware Demonstration on Token Usage and Efficiency. In Table 6, we also show the number of input and output tokens when using different demonstrations for conducting CoT. According to the results in the table, when using different demonstrations, the number of generated tokens is comparable. While the proposed w IR involves more input tokens than w/o IR, the increase is not excessive. Furthermore, w IR achieves both higher accuracy and requires fewer input tokens compared to w/o IR+. These results demonstrate that the proposed method provides a favorable trade-off between computational cost and performance.

Table 6: Performance of CoT with and without Incorrect Reasoning (IR) across settings.

				Accuracy (Acc)		Tokens	
Dataset	**Method**	**Examples**	**Reasoning Paths**	**GPT-Turbo-3.5**	**Gemini Pro**	**Input**	**Output**
Disambiguation QA	w/o IR	2	2	60.40%	71.60%	713	205.9
	w/o IR+	4	4	64.80%	71.60%	1189	203.1
	w IR	2	4	**65.20%**	**72.40%**	854	205.2
	w/o IR	3	3	68.00%	68.80%	992	201.6
	w/o IR+	6	6	68.40%	66.40%	1775	202.9
	w IR	3	6	**72.00%**	**76.80%**	1198	205.3
Penguins in a Table	w/o IR	2	2	81.34%	82.19%	945.3	86.2
	w/o IR+	4	4	81.51%	81.51%	1547	94.2
	w IR	2	4	**82.19%**	**83.56%**	1126.3	103.7
	w/o IR	3	3	80.14%	80.82%	743.3	89.8
	w/o IR+	6	6	80.82%	82.88%	1146.3	93.2
	w IR	3	6	**84.93%**	**85.62%**	838.3	111.3

F.3 Results with Additional Models

To further demonstrate the generalization of the proposed method across different models, we conduct experiments using GPT-4-Turbo and Llama-3.1-8B (Instruct). The results are shown in the following Table. (As GPT-4-Turbo can already perform perfectly in the Penguins in a Table dataset with w/o IR, we consider the Word Sorting dataset from BBH for the experiments of GPT-4-Turbo.) From the results, we can see that across different models and datasets, our proposed w IR consistently demonstrates better performance in reasoning compared with w/o IR and w/o IR+.

Table 7: Performance of CoT when using LLaMA-3.1-8B and GPT4-Turbo.

Models	**Methods**	**Number of Examples**	**Penguins in a Table**	**Disambiguation QA**
LLaMA-3.1-8B	w/o IR	3	71.92%	74.80%
	w/o IR +	6	66.44%	72.40%
	w IR	3	**73.29%**	**76.40%**
GPT4-Turbo	w/o IR	3	90.00%	85.60%
	w/o IR +	6	88.00%	84.00%
	w IR	3	**92.40%**	**88.00%**

G More examples for the proposed demonstration format.

In this section, we provide some additional examples for the proposed demonstration format.

G.1 Disambiguation QA Dataset

Clarify the meaning of sentences with ambiguous pronouns.
Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous. Sentence: The chief told the counselor that they took the day off.
Options:
(A) The chief took the day off
(B) The counselor took the day off
(C) Ambiguous
Wrong Answer: The incorrect reasoning would choose (B) and state “The psychologist took the day off, and that is why the chief is explaining something to them.”
Error: This reasoning is wrong because there is no clear indicator in the sentence that implies the psychologist (counselor) took the day off.
Correct Answer: Let’s think step by step.
Here we need to determine who the pronoun ”they” might be referring to. There are two possible referents for ”they”, namely the chief and the counselor. The verb ”told” might be able to help us determine which one is more likely (if either). Let X be the chief and Y the counselor. The sentence is then of the form ”X told Y that (X or Y) did something.”
Let’s consider Y first: ”X told Y that Y did something.” This case does not make much sense, as Y would already have the information that Y did something, because it is information about themselves.
Now, consider X: ”X told Y that X did something.” This makes sense, because X would be sharing some information about themselves that Y might not have known before.
Because in this context, X is the chief and Y is the counselor, the answer should be the chief.
So the answer is (A).

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous. Sentence: The manager sent a message to the secretary, but he didn’t reply yet.
Options:
(A) The secretary didn’t reply yet
(B) The manager didn’t reply yet
(C) Ambiguous
Wrong Answer: The incorrect reasoning would choose (B) and state “The manager didn’t reply yet, because perhaps the manager is busy or forgot to respond.”
Error: This is incorrect because the sentence clearly establishes that the manager was the one who initially sent the message.
Correct Answer: Let’s think step by step.
Here we need to determine who the pronoun ”he” might be referring to. There are two possible referents for ”he”, namely the manager and the secretary. The verbs ”sent” and ”reply” might be able to help us determine which one is more likely (if either). Let X be the manager and Y the secretary. The sentence is then of the form ”X sent a message to Y, but (X or Y) didn’t reply yet.”
Let’s consider Y first: ”X sent a message to Y, but Y didn’t reply yet.” This case makes sense, because of the implicit causality of the sentence. Y was the receiver of the message, but Y didn’t get back to X yet. Now, consider X: ”X sent a message to Y, but X didn’t reply yet.” This case doesn’t make sense, because X was the initial sender of the message, so it is now Y’s turn to write back to X.
Because in this context, X is the manager and Y is the secretary, the answer should be the secretary.
So the answer is (A).

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous. Sentence: Bailey will plan to meet the director at his office
Options:
(A) It will be Bailey’s office
(B) It will be the director’s office
(C) Ambiguous
Wrong Answer: The incorrect reasoning would choose (B) and state ”The meeting will take place at the director’s office, as it makes sense for a director to conduct meetings in their own office.”
Error: This is incorrect because the sentence does not explicitly indicate whose office the meeting will take place in.
Correct Answer: Let’s think step by step.
Here we need to determine who the pronoun ”his” might be referring to. There are two possible referents for ”his”, namely Bailey’s and the director’s. The verb phrase ”plan to meet” might be able to help us determine which one is more likely (if either). Let X be Bailey and Y the director. The sentence is then of the form ”X will plan to meet Y at (X or Y)’s office.”
Let’s consider Y first: ”X will plan to meet Y at Y’s office.” This case makes sense, because X might want to meet up with Y at Y’s office.
Now, consider X: ”X will plan to meet Y at X’s office.” This case also makes sense, because X might want to meet up with Y at X’s own office.
Because both X and Y are possible at the same time, we conclude that the antecedent of the pronoun is ambiguous.
So the answer is (C).

G.2 Penguins in a Table Dataset

Answer questions about a table of penguins and their attributes.
Q: Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm. We now add a penguin to the table:
James, 12, 90, 12
How many penguins are less than 8 years old?
Options:
(A) 1
(B) 2
(C) 3
(D) 4
(E) 5
Wrong Answer: There are 3 penguins less than 8 years old: Louis, Bernard, and Gwen. So, the answer is (C).
Error: This answer incorrectly includes Gwen, who is exactly 8 years old and should not be counted.
Correct Answer:
A: Let's think step by step.
This question focuses on age. We know the following: Louis is 7 years old, Bernard is 5 years old, Vincent is 9 years old, and Gwen is 8 years old.
Now, we add James to this table: James is 12 years old.
The penguins that are less than 8 years old are Louis and Bernard.
There are 2 penguins less than 8 years old. So the answer is (B).

Q: Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm. Which is the youngest penguin?
Options:
(A) Louis
(B) Bernard
(C) Vincent
(D) Gwen
(E) James
Wrong Answer: The youngest penguin is Louis because he is 7 years old. The answer is (A).
Error: This answer overlooks the fact that Bernard is 5 years old, making him the youngest.
Correct Answer:
A: Let's think step by step.
This question focuses on age. We know the following: Louis is 7 years old, Bernard is 5 years old, Vincent is 9 years old, and Gwen is 8 years old.
According to the table, Bernard (5) is the youngest amongst them.
The youngest penguin is Bernard. So the answer is (B).

Q: Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm. What is the name of the second penguin sorted by alphabetic order?
Options:
(A) Louis
(B) Bernard
(C) Vincent
(D) Gwen
(E) James
Wrong Answer: The names alphabetically are Bernard, Gwen, James, Louis, Vincent. The second penguin is Louis. The answer is (A).
Error: This answer incorrectly skips Gwen, which should be the second name alphabetically.
Correct Answer:
A: Let's think step by step.
This question focuses on the name. We know the following: The names of the penguin in the table are Louis, Bernard, Vincent, and Gwen.
When we sort their names alphabetically, we get Bernard, Gwen, Louis, Vincent.
The name of the second penguin sorted by alphabetical order is Gwen.
The name of the second penguin sorted by alphabetic order is Gwen. So the answer is (D).

G.3 Tracking Shuffled Objects Dataset

A task requiring determining the final positions of a set of objects given their initial positions and a description of a sequence of swaps.

Q: Alice, Bob, and Claire are playing a game. At the start of the game, they are each holding a ball: Alice has a yellow ball, Bob has a blue ball, and Claire has a pink ball.
As the game progresses, pairs of players trade balls. First, Claire and Alice swap balls. Then, Alice and Bob swap balls. Finally, Claire and Bob swap balls. At the end of the game, Bob has the
Options:
(A) yellow ball (B) blue ball (C) pink ball
Wrong Answer: Let's think step by step.
(0) At the start: Alice: yellow, Bob: blue, Claire: pink.
(1) Claire and Alice swap balls: Alice: pink, Bob: blue, Claire: yellow.
(2) Alice and Bob swap balls: Alice: blue, Bob: pink, Claire: yellow.
At the end of the game, Bob has the pink ball. So the answer is (C).
Error: This answer forgets the last swap between Claire and Bob, incorrectly concluding that Bob holds the pink ball from the second swap, when in fact, Bob trades with Claire again.
Correct answer:
A: Let's think step by step.
(0) At the start: Alice: yellow, Bob: blue, Claire: pink.
(1) Claire and Alice swap balls: Alice: pink, Bob: blue, Claire: yellow.
(2) Alice and Bob swap balls: Alice: blue, Bob: pink, Claire: yellow.
(3) Claire and Bob swap balls: Alice: blue, Bob: yellow, Claire: pink.
At the end of the game, Bob has the yellow ball. So the answer is (A).

Q: Alice, Bob, and Claire are playing a game. At the start of the game, they are each holding a ball: Alice has a white ball, Bob has a purple ball, and Claire has a pink ball. As the game progresses, pairs of players trade balls. First, Bob and Alice swap balls. Then, Bob and Claire swap balls. Finally, Bob and Alice swap balls. At the end of the game, Alice has the
Options:
(A) white ball (B) purple ball (C) pink ball
Wrong Answer: Let's think step by step.
(0) At the start: Alice: white, Bob: purple, Claire: pink.
(1) Bob and Alice swap balls: Alice: purple, Bob: white, Claire: pink.
(2) Bob and Alice swap balls: Alice: white, Bob: pink, Claire: pink.
At the end of the game, Bob Alice has the white ball. So the answer is (A).
Error: This answer assumes Alice gets her original ball back, ignoring the correct sequence of swaps where Alice ends up with the pink ball after the final exchange.
Correct answer:
A: Let's think step by step.
(0) At the start: Alice: white, Bob: purple, Claire: pink.
(1) Bob and Alice swap balls: Alice: purple, Bob: white, Claire: pink.
(2) Bob and Claire swap balls: Alice: purple, Bob: pink, Claire: white.
(3) Bob and Alice swap balls: Alice: pink, Bob: purple, Claire: white.
At the end of the game, Alice has the pink ball. So the answer is (C).

Q: Alice, Bob, and Claire are dancers at a square dance. At the start of a song, they each have a partner: Alice is dancing with Lola, Bob is dancing with Rodrigo, and Claire is dancing with Patrick. Throughout the song, the dancers often trade partners. First, Alice and Bob switch partners. Then, Claire and Bob switch partners. Finally, Bob and Alice switch partners. At the end of the dance, Alice is dancing with
Options:
(A) Lola (B) Rodrigo (C) Patrick
Wrong Answer: Let's think step by step.
(0) At the start: Alice: Lola, Bob: Rodrigo, Claire: Patrick.
Error: This answer mistakenly assumes that Alice ends up with her original partner, overlooking the fact that Alice has switch partners for twice.
Correct answer:
A: Let's think step by step.
(0) At the start: Alice: Lola, Bob: Rodrigo, Claire: Patrick.
(1) Alice and Bob switch partners: Alice: Rodrigo, Bob: Lola, Claire: Patrick.
(2) Claire and Bob switch partners: Alice: Rodrigo, Bob: Patrick, Claire: Lola.
(3) Bob and Alice switch partners: Alice: Patrick, Bob: Rodrigo, Claire: Lola.
At the end of the dance, Alice is dancing with Patrick. So the answer is (C).

G.4 Date Understanding Dataset

Infer the date from context.

Q: Today is Christmas Eve of 1937. What is the date 10 days ago in MM/DD/YYYY?
Options:
(A) 12/14/2026
(B) 12/14/1950
(C) 12/14/2007
(D) 12/14/1937
(E) 07/14/1938
(F) 12/14/1988
Wrong Answer: The date 10 days ago from Christmas Eve of 1937 would be 12/14/1950.
Error: This answer incorrectly chooses a future year. The correct answer is 12/14/1937, as 10 days before Christmas Eve of 1937 is December 14, 1937.
Correct Answer:
Let's think step by step.
If today is Christmas Eve of 1937, then today's date is December 24, 1937. 10 days before today is December 14, 1937, that is 12/14/1937. So the answer is (D).

Q: Tomorrow is 11/12/2019. What is the date one year ago from today in MM/DD/YYYY?
Options:
(A) 09/04/2018
(B) 11/11/2018
(C) 08/25/2018
(D) 11/02/2018
(E) 11/04/2018
Wrong Answer: The date one year ago from today would be 09/04/2018.
Error: This answer selects an incorrect month and day. Since today is 11/11/2019, one year ago would be 11/11/2018, making the correct answer 11/11/2018.
Correct Answer:
Let's think step by step.
If tomorrow is 11/12/2019, then today is 11/11/2019. The date one year ago from today is 11/11/2018. So the answer is (B).

Q: Jane and John married on Jan 2, 1958. It is their 5-year anniversary today. What is the date tomorrow in MM/DD/YYYY?
Options:
(A) 01/11/1961
(B) 01/03/1963
(C) 01/18/1961
(D) 10/14/1960
(E) 01/03/1982
(F) 12/03/1960
Wrong Answer: Since it is their 5-year anniversary today, the date tomorrow would be 01/18/1961.
Error: This answer incorrectly calculates the year. Since Jane and John were married on January 2, 1958, and today is their 5-year anniversary, the correct date tomorrow would be 01/03/1963.
Correct Answer:
Let's think step by step.
If Jane and John married on Jan 2, 1958, then and if it is their 5-year anniversary today, then today's date is Jan 2, 1963. The date tomorrow is Jan 3, 1963, that is 01/03/1963. So the answer is (B).

G.5 GSM8K Dataset

Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
Wrong Answer: The incorrect reasoning might be, "There were 15 trees, and there will be 21 after planting. Adding these gives 15 + 21 = 36 trees planted today."
Error: This is wrong because the correct method is to subtract the initial 15 trees from the final 21. The correct answer is 6 trees planted today.
Correct Answer:
Let's think step by step
There are 15 trees originally.
Then there were 21 trees after some more were planted.
So there must have been 21 - 15 = 6.
The answer is 6.

Question: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot? Wrong Answer: The incorrect reasoning might be, "There are 3 cars, and 2 more arrive. Multiplying 3 by 2 gives 6, so there are 6 cars."
Error: This is wrong because the correct operation is addition, not multiplication. You should add the 2 new cars to the 3 already there, giving 5 cars.
Correct Answer:
Let's think step by step
There are originally 3 cars.
2 more cars arrive.
3 + 2 = 5.
The answer is 5.

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?
Wrong Answer: The incorrect reasoning might be, "Leah had 32 chocolates and her sister had 42. Subtracting the 35 they ate from 42 gives 7 chocolates left."
Error: This is incorrect because the total pieces they had initially is 32 + 42 = 74, and subtracting 35 from 74 gives the correct answer of 39 pieces left.
Correct Answer:
Let's think step by step
Originally, Leah had 32 chocolates.
Her sister had 42.
So in total they had 32 + 42 = 74.
After eating 35, they had 74 - 35 = 39.
The answer is 39.

Question: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
Wrong Answer: The incorrect reasoning might be, "Jason had 20 lollipops, and now he has 12. Adding 12 to 20 gives 32 lollipops given to Denny."
Error: This is wrong because the correct operation is subtraction, not addition. Subtracting 12 from 20 gives the correct answer of 8 lollipops given to Denny.
Correct Answer:
Let's think step by step
Jason started with 20 lollipops.
Then he had 12 after giving some to Denny.
So he gave Denny 20 - 12 = 8.
The answer is 8.

H Details about the computing infrastructure.

Experiments for the simulations in Figure 2 and 3. We conducted the experiments on a single A6000 GPU with 48GB of memory. Training the transformer took approximately 30 minutes, while inference took around 40 seconds.

Experiments in Section 4.2. For these experiments, we called all models directly using their APIs, except for DeepSeek 67, for which we downloaded the checkpoint and ran the model on four A6000 GPUs. For the BBH benchmark datasets (each containing 250 examples), it took approximately 3–6 hours to generate solutions for all questions. For the GSM8K dataset (which contains 1,319 examples), the total generation time was around 14 hours.