Data Mining and Machine Learning Lab

Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens

Chengshuai Zhao[1], **Zhen Tan**[1], **Pingchuan Ma**[1], **Dawei Li**[1], **Bohan Jiang**[1], **Yancheng Wang**[1], **Yingzhen Yang**[1] **and Huan Liu**[1]

[1]Arizona State University, USA

Chain-of-Thought (CoT) prompting has been shown to improve Large Language Model (LLM) performance on various tasks. With this approach, LLMs appear to produce human-like reasoning steps before providing answers (a.k.a., CoT reasoning), which often leads to the perception that they engage in deliberate inferential processes. However, some initial findings suggest that CoT reasoning may be more superficial than it appears, motivating us to explore further. In this paper, we study CoT reasoning via a data distribution lens and investigate if CoT reasoning reflects a structured inductive bias learned from in-distribution data, allowing the model to conditionally generate reasoning paths that approximate those seen during training. Thus, its effectiveness is fundamentally bounded by the degree of distribution discrepancy between the training data and the test queries. With this lens, we dissect CoT reasoning via three dimensions: *task*, *length*, and *format*. To investigate each dimension, we design DataAlchemy, an isolated and controlled environment to train LLMs from scratch and systematically probe them under various distribution conditions. Our results reveal that CoT reasoning is a brittle mirage that vanishes when it is pushed beyond training distributions. This work offers a deeper understanding of *why* and *when* CoT reasoning fails, emphasizing the ongoing challenge of achieving genuine and generalizable reasoning. Our code is available at GitHub: https://github.com/ChengshuaiZhao0/DataAlchemy.

1. Introduction

Recent years have witnessed Large Language Models' (LLMs) dominant role in various domains (Li et al., 2025b; Ting et al., 2025; Zhao et al., 2025, 2023) through versatile prompting techniques (Kojima et al., 2022; Wei et al., 2022; Yao et al., 2023). Among these, Chain-of-Thought (CoT) prompting (Wei et al., 2022) has emerged as a prominent method for eliciting structured reasoning from LLMs (a.k.a., CoT reasoning). By appending a simple cue such as "Let's think step by step," LLMs decompose complex problems into intermediate steps, producing outputs that resemble human-like reasoning. It has been shown to be effective in tasks requiring logical inference(Xu et al., 2024), mathematical problem solving (Imani et al., 2023), and commonsense reasoning (Wei et al., 2022). The empirical successes of CoT reasoning lead to the perception that LLMs engage in deliberate inferential processes (Ling et al., 2023; Yu et al., 2023; Zhang et al., 2024a,c).

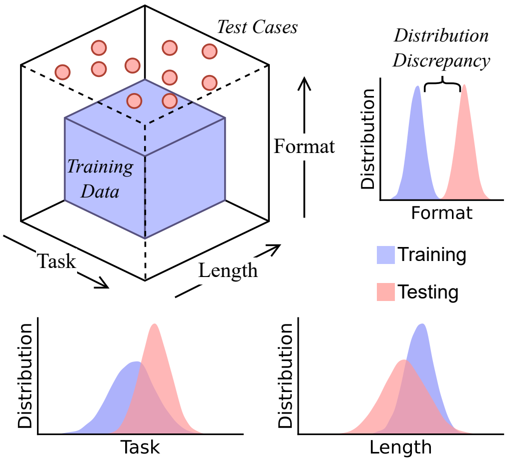


Figure 1 | The data perspective lens. CoT reasoning's effectiveness is fundamentally bounded by the degree of distribution discrepancy between the training data and the test queries. Guided by this lens, we dissect CoT reasoning via three dimensions: *task*, *length*, and *format*.

Corresponding author(s): {czhao93, ztan36, pingchua, daweili5, bjiang14, yancheng.wang, yingzhen.yang, huanliu}@asu.edu

However, a closer examination reveals inconsistencies that challenge this optimistic view. Consider this straightforward question: *"The day the US was established is in a leap year or a normal year?"* When prompted with the CoT prefix, the modern LLM Gemini responded: *"The United States was established in 1776. 1776 is divisible by 4, but it's not a century year, so it's a leap year. Therefore, the day the US was established was in a normal year."* This response exemplifies a concerning pattern: the model correctly recites the leap year rule and articulates intermediate reasoning steps, yet produces a logically inconsistent conclusion (i.e., asserting 1776 is both a leap year and a normal year). Such inconsistencies suggest that there is a distinction between human-like inference and CoT reasoning.

An expanding body of analyses reveals that LLMs tend to rely on surface-level semantics and clues rather than logical procedures (Chen et al., 2025b; Kambhampati, 2024; Lanham et al., 2023; Stechly et al., 2024). LLMs construct superficial chains of logic based on learned token associations, often failing on tasks that deviate from commonsense heuristics or familiar templates (Tang et al., 2023). In the reasoning process, performance degrades sharply when irrelevant clauses are introduced, which indicates that models cannot grasp the underlying logic (Mirzadeh et al., 2024). This fragility becomes even more apparent when models are tested on more complex tasks, where they frequently produce incoherent solutions and fail to follow consistent reasoning paths (Shojaee et al., 2025). Collectively, these pioneering works deepen the skepticism surrounding the true nature of CoT reasoning.

In light of this line of research, we question the CoT reasoning by proposing an alternative lens through data distribution and further investigating *why* and *when* it fails. We **hypothesize** that CoT reasoning reflects a structured inductive bias learned from in-distribution data, allowing the model to conditionally generate reasoning paths that approximate those seen during training. As such, its effectiveness is inherently limited by the nature and extent of the distribution discrepancy between training data and the test queries. Guided by this data distribution lens, we dissect CoT reasoning via three dimensions: (i) *task*—To what extent CoT reasoning can handle tasks that involve transformations or previously unseen task structures. (2) *length*—how CoT reasoning generalizes to chains with length different from that of training data; and (3) *format*—how sensitive CoT reasoning is to surface-level query form variations. To evaluate each aspect, we introduce DataAlchemy, a controlled and isolated experiment that allows us to train LLMs from scratch and systematically probe them under various distribution shifts.

Our findings reveal that CoT reasoning works effectively when applied to in-distribution or near in-distribution data but becomes fragile and prone to failure even under moderate distribution shifts. In some cases, LLMs generate fluent yet logically inconsistent reasoning steps. The results suggest that what appears to be structured reasoning can be a mirage, emerging from memorized or interpolated patterns in the training data rather than logical inference. These insights carry important implications for both practitioners and researchers. For practitioners, our results highlight the risk of relying on CoT as a plug-and-play solution for reasoning tasks and caution against equating CoT-style output with human thinking. For researchers, the results underscore the ongoing challenge of achieving reasoning that is both faithful and generalizable, motivating the need to develop models that can move beyond surface-level pattern recognition to exhibit deeper inferential competence. Our contributions are summarized as follows:

- ⋆ **Novel perspective.** We propose a data distribution lens for CoT reasoning, illuminating that its effectiveness stems from structured inductive biases learned from in-distribution training data. This framework provides a principled lens for understanding *why* and *when* CoT reasoning succeeds or fails.
- ⋆ **Controlled environment.** We introduce DataAlchemy, an isolated experimental framework that enables training LLMs from scratch and systematically probing CoT reasoning. This controlled setting allows us to isolate and analyze the effects of distribution shifts on CoT reasoning without

interference from complex patterns learned during large-scale pre-training.

⋆ **Empirical validation.** We conduct systematic empirical validation across three critical dimensions—*task*, length, and format. Our experiments demonstrate that CoT reasoning exhibits sharp performance degradation under distribution shifts, revealing that seemingly coherent reasoning masks shallow pattern replication.

⋆ **Real-world implication**. This work reframes the understanding of contemporary LLMs' reasoning capabilities and emphasizes the risk of over-reliance on COT reasoning as a universal problem-solving paradigm. It underscores the necessity for proper evaluation methods and the development of LLMs that possess authentic and generalizable reasoning capabilities.

2. Related Work

2.1. LLM Prompting and CoT

Chain-of-Thought (CoT) prompting revolutionized how we elicit reasoning from Large Language Models by decomposing complex problems into intermediate steps (Wei et al., 2022). By augmenting few-shot exemplars with reasoning chains, CoT showed substantial performance gains on various tasks (Imani et al., 2023; Wei et al., 2022; Xu et al., 2024). Building on this, several variants emerged. Zero-shot CoT triggers reasoning without exemplars using instructional prompts (Kojima et al., 2022), and self-consistency enhances performance via majority voting over sampled chains (Wang et al., 2023). To reduce manual effort, Auto-CoT generates CoT exemplars using the models themselves (Zhang et al., 2023). Beyond linear chains, Tree-of-Thought (ToT) frames CoT as a tree search over partial reasoning paths (Yao et al., 2023), enabling lookahead and backtracking. SymbCoT combines symbolic reasoning with CoT by converting problems into formal representations (Xu et al., 2024). Recent work increasingly integrates CoT into the LLM inference process, generating long-form CoTs (Guo et al., 2025; Jaech et al., 2024; Team et al., 2025; Team, 2024). This enables flexible strategies like mistake correction, step decomposition, reflection, and alternative reasoning paths (Chen et al., 2025a; Yeo et al., 2025). The success of prompting techniques and long-form CoTs has led many to view them as evidence of emergent, human-like reasoning in LLMs. In this work, we challenge that viewpoint by adopting a data-centric perspective and demonstrating that CoT behavior arises largely from pattern matching over training distributions.

2.2. Discussion on Illusion of LLM Reasoning

While Chain-of-Thought prompting has led to impressive gains on complex reasoning tasks, a growing body of work has started questioning the nature of these gains (Kambhampati et al., 2025; Stechly et al., 2024, 2025). One major line of research highlights the fragility of CoT reasoning. Minor and semantically irrelevant perturbations such as distractor phrases or altered symbolic forms can cause significant performance drops in state-of-the-art models (Mirzadeh et al., 2024; Tang et al., 2023). Models often incorporate such irrelevant details into their reasoning, revealing a lack of sensitivity to salient information. Other studies show that models prioritize the surface form of reasoning over logical soundness; in some cases, longer but flawed reasoning paths yield better final answers than shorter, correct ones (Bentham et al., 2024). Similarly, performance does not scale with problem complexity as expected—models may overthink easy problems and give up on harder ones (Shojaee et al., 2025). Another critical concern is the faithfulness of the reasoning process. Intervention-based studies reveal that final answers often remain unchanged even when intermediate steps are falsified or omitted (Lanham et al., 2023), a phenomenon dubbed the illusion of transparency (Bentham et al., 2024; Chen et al., 2025b). Together, these findings suggest that LLMs are not principled reasoners but rather sophisticated simulators of reasoning-like text. However, a systematic understanding of

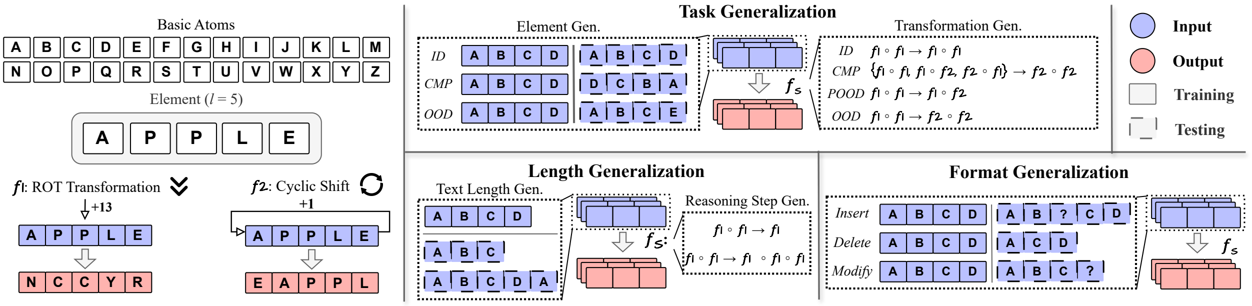


Figure 2 | Framework of DataAlchemy. It creates an isolated and controlled environment to train LLMs from scratch and probe the task, length, and format generalization.

why and when CoT reasoning fails is still a mystery.

2.3. OOD Generalization of LLMs

Out-of-distribution (OOD) generalization, where test inputs differ from training data, remains a key challenge in machine learning, particularly for large language models (LLMs)(Budnikov et al., 2025; Yang et al., 2024, 2023; Zhang et al., 2024b). Recent studies show that LLMs prompted to learn novel functions often revert to similar functions encountered during pretraining (Garg et al., 2022; Wang et al., 2024). Likewise, LLM generalization frequently depends on mapping new problems onto familiar compositional structures (Song et al., 2025). CoT prompting improves OOD generalization (Wei et al., 2022), with early work demonstrating length generalization for multi-step problems beyond training distributions (Shen et al., 2025; Yao et al., 2025). However, this ability is not inherent to CoT and heavily depends on model architecture and training setups. For instance, strong generalization in arithmetic tasks was achieved only when algorithmic structures were encoded into positional encodings (Cho et al., 2024). Similarly, finer-grained CoT demonstrations during training boost OOD performance, highlighting the importance of data granularity (Wang et al., 2025a). Theoretical and empirical evidence shows that CoT generalizes well only when test inputs share latent structures with training data; otherwise, performance declines sharply (Li et al., 2025a; Wang et al., 2025b). Despite its promise, CoT still struggles with genuinely novel tasks or formats. In the light of these brilliant findings, we propose rethinking CoT reasoning through a data distribution lens: decomposing CoT into *task*, *length*, and *format* generalization, and systematically investigating each in a controlled setting.

3. The Data Distribution Lens

We propose a fundamental reframing to understand what CoT actually represents. We **hypothesize** that the underlying mechanism is better understood through the lens of data distribution: rather than executing explicit reasoning procedures, CoT operates as a pattern-matching process that interpolates and extrapolates from the statistical regularities present in its training distribution. Specifically, we posit that CoT's success stems not from a model's inherent reasoning capacity, but from its ability to generalize conditionally to out-of-distribution (OOD) test cases that are structurally similar to in-distribution exemplars.

To formalize this view, we model CoT prompting as a conditional generation process constrained by the distributional properties of the training data. Let $\mathcal{D}_{\text{train}}$ denote the training distribution over input-output pairs (x, y), where x represents a reasoning problem and y denotes the solution sequence (including intermediate reasoning steps). The model learns an approximation $f_\theta(x) \approx y$ by minimizing empirical risk over samples drawn from $\mathcal{D}_{\text{train}}$.

Let the *expected training risk* be defined as:

$$R_{\text{train}}(f_\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}}[\ell(f_\theta(x), y)], \tag{1}$$

where ℓ is a task-specific loss function (e.g., cross-entropy, token-level accuracy). At inference time, given a test input a_{test} sampled from a potentially different distribution $\mathcal{D}_{\text{test}}$, the model generates a response y_{test} conditioned on patterns learned from $\mathcal{D}_{\text{train}}$. The corresponding *expected test risk* is:

$$R_{\text{test}}(f_\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}}[\ell(f_\theta(x), y)]. \tag{2}$$

The degree to which the model generalizes from $\mathcal{D}_{\text{train}}$ to $\mathcal{D}_{\text{test}}$ is governed by the *distributional discrepancy* between the two, which we quantify using divergence measures:

Definition 3.1 (Distributional Discrepancy). *Given training distribution $\mathcal{D}_{train}$ and test distribution $\mathcal{D}_{test}$, the distributional discrepancy is defined as:*

$$\Delta(\mathcal{D}_{train}, \mathcal{D}_{test}) = \mathcal{H}(\mathcal{D}_{train} \parallel \mathcal{D}_{test}) \tag{3}$$

where $\mathcal{H}(\cdot \parallel \cdot)$ is a divergence measure (e.g., KL divergence, Wasserstein distance) that quantifies the statistical distance between the two distributions.

Theorem 3.1 (CoT Generalization Bound). *Let f_θ denote a model trained on $\mathcal{D}_{train}$ with expected training risk $R_{train}(f_\theta)$. For a test distribution $\mathcal{D}_{test}$, the expected test risk $R_{test}(f_\theta)$ is bounded by:*

$$R_{test}(f_\theta) \leq R_{train}(f_\theta) + \Lambda \cdot \Delta(\mathcal{D}_{train}, \mathcal{D}_{test}) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right) \tag{4}$$

where $\Lambda > 0$ is a Lipschitz constant that depends on the model architecture and task complexity, n is the training sample size, and the bound holds with probability $1 - \delta$, where δ is the failure probability.

The proof is provided in Appendix A.1

Building on this data distribution perspective, we identify three critical dimensions along which distributional shifts can occur, each revealing different aspects of CoT's pattern-matching nature: ❶ *Task generalization* examines how well CoT transfers across different types of reasoning tasks. Novel tasks may have unique elements and underlying logical structure, which introduces distributional shifts that challenge the model's ability to apply learned reasoning patterns. ❷ *Length generalization* investigates CoT's robustness to reasoning chains of varying lengths. Since training data typically contains reasoning sequences within a certain length range, test cases requiring substantially longer or shorter reasoning chains represent a form of distributional shift along the sequence length dimension. This length discrepancy could result from the reasoning step or the text-dependent solution space. ❸ *Format generalization* explores how sensitive CoT is to variations in prompt formulation and structure. Due to various reasons (e.g., sophistical training data or diverse background of users), it is challenging for LLM practitioners to design a golden prompt to elicit knowledge suitable for the current case. Their detailed definition and implementation are given in subsequent sections.

Each dimension provides a unique lens for understanding the boundaries of CoT's effectiveness and the mechanisms underlying its apparent reasoning capabilities. By systematically varying these dimensions in controlled experimental settings, we can empirically validate our hypothesis that CoT performance degrades predictably as distributional discrepancy increases, thereby revealing its fundamental nature as a pattern-matching rather than reasoning system.

4. DataAlchemy: An Isolated and Controlled Environment

To systematically investigate the influence of distributional shifts on CoT reasoning capabilities, we introduce DataAlchemy, a synthetic dataset framework designed for controlled experimentation. This environment enables us to train language models from scratch under precisely defined conditions, allowing for rigorous analysis of CoT behavior across different OOD scenarios. The overview is shown in Figure 2.

4.1. Basic Atoms and Elements

Let $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \ldots, \mathtt{Z}\}$ denote the alphabet of 26 basic atoms. An element $\mathbf{e}$ is defined as an ordered sequence of atoms:

$$\mathbf{e} = (a_0, a_1, \ldots, a_{l-1}) \quad \text{where} \quad a_i \in \mathcal{A}, \quad l \in \mathbb{Z}^+ \tag{5}$$

This design provides a versatile manipulation for the size of the dataset $\mathcal{D}$ (i.e., $|\mathcal{D}| = |\mathcal{A}|^l$) by varying element length l to train language models with various capacities. Meanwhile, it also allows us to systematically probe text length generalization capabilities.

4.2. Transformations

A transformation is an operation that operates on elements $F : \mathbf{e} \to \hat{\mathbf{e}}$. In this work, we consider two fundamental transformations: the *ROT Transformation* and the *Cyclic Position Shift*. To formally define the transformations, we introduce a bijective mapping $\phi : \mathcal{A} \to \mathbb{Z}_{26}$, where $\mathbb{Z}_{26} = \{0, 1, \ldots, 25\}$, such that $\phi(c)$ maps a character to its zero-based alphabetical index.

Definition 4.1 (ROT Transformation)**.** *Given an element $\mathbf{e} = (a_0, \ldots, a_{l-1})$ and a rotation parameter $n \in \mathbb{Z}$, the ROT Transformation f_{rot} produces an element $\hat{\mathbf{e}} = (\hat{a}_0, \ldots, \hat{a}_{l-1})$. Each atom $\hat{a}_i$ is:*

$$\hat{a}_i = \phi^{-1}((\phi(a_i) + n) \pmod{26}) \tag{6}$$

This operation cyclically shifts each atom n positions forward in alphabetical order. For example, if $\mathbf{e} = (A, P, P, L, E)$ and $n = 13$, then $f_{rot}(\mathbf{e}, 13) = (N, C, C, Y, R)$.

Definition 4.2 (Cyclic Position Shift)**.** *Given an element $\mathbf{e} = (a_0, \ldots, a_{l-1})$ and a shift parameter $n \in \mathbb{Z}$, the Cyclic Position Shift f_{pos} produces an element $\hat{\mathbf{e}} = (\hat{a}_0, \ldots, \hat{a}_{l-1})$. Each atom $\hat{a}_i$ is defined by a cyclic shift of indices:*

$$\hat{a}_i = a_{(i-n) \pmod{l}} \tag{7}$$

This transformation cyclically shifts the positions of the atoms within the sequence by n positions to the right. For instance, if $\mathbf{e} = (A, P, P, L, E)$ and $n = 1$, then $f_{pos}(\mathbf{e}, 1) = (E, A, P, P, L)$.

Definition 4.3 (Generalized Compositional Transformation)**.** *To model multi-step reasoning, we define a compositional transformation as the successive application of a sequence of operations. Let $S = (f_1, f_2, \ldots, f_k)$ be a sequence of operations, where each f_i is one of the fundamental transformations $\mathcal{F} = \{f_{rot}, f_{pos}\}$ with its respective parameters. The compositional transformation f_S for the sequence S is the function composition:*

$$f_S = f_k \circ f_k \circ \cdots \circ f_1 \tag{8}$$

The resulting element $\hat{\mathbf{e}}$ is obtained by applying the operations sequentially to an initial element $\mathbf{e}$:

$$\hat{\mathbf{e}} = f_k(f_{k-1}(\ldots(f_1(\mathbf{e}))\ldots)) \tag{9}$$

This design enables the construction of arbitrarily complex transformation chains by varying the type, parameters, order, and length of operations within the sequence. At the sample time, we can naturally acquire the COT reasoning step by decomposing the intermediate process:

$$\underbrace{f_S(\mathbf{e}):}_{\text{Query}} \quad \underbrace{\mathbf{e} \xrightarrow{f_1} \mathbf{e}^{(1)} \xrightarrow{f_2} \mathbf{e}^{(2)} \cdots \xrightarrow{f_{k-1}} \mathbf{e}^{(k-1)} \xrightarrow{f_k}}_{\text{COT reasoning steps}} \underbrace{\boxed{\hat{\mathbf{e}}}}_{\text{Answer}} \tag{10}$$

4.3. Environment Setting

Through systematic manipulation of elements and transformations, DataAlchemyoffers a flexible and controllable framework for training LLMs from scratch, facilitating rigorous investigation of diverse OOD scenarios. Without specification, we employ a decoder-only language model GPT-2 (Radford et al., 2019) with a configuration of 4 layers, 32 hidden dimensions, and 4 attention heads. We utilize a Byte-Pair Encoding (BPE) tokenizer. Both LLMs and the tokenizer follow the general modern LLM pipeline. During the inference time, we set the temperature to 1e-5. For rigor, we also study LLMs with various parameters, architectures, and temperatures in Section 8. Details of the implementation are provided in the Appendix B. We consider that each element consists of 4 basic atoms, which produces 456,976 samples for each dataset with varied transformations and token amounts. We initialize the two transformations $f_1 = f_{\text{rot}}(e, 13)$ and $f_2 = f_{\text{pos}}(e, 1)$. We consider the exact match rate, Levenshtein distance (i.e., edit distance) (Yujian and Bo, 2007), and BLEU score (Papineni et al., 2002) as metrics and evaluate the produced reasoning step, answer, and full chain. Examples of the datasets and evaluations are shown in Appendix C

5. Task Generalization

Task generalization represents a fundamental challenge for CoT reasoning, as it directly tests a model's ability to apply learned concepts and reasoning patterns to unseen scenarios. In our controlled experiments, both transformation and elements could be novel. Following this, we decompose task generalization into two primary dimensions: element generalization and transformation generalization.

Task Generalization Complexity. Guided by the data distribution lens, we first introduce a measure for generalization difficulty:

Proposition 5.1 (Task Generalization Complexity). *For a reasoning chain f_S operating on elements $\mathbf{e} = (a_0, \ldots, a_{l-1})$, define:*

$$\text{TGC}(C) = \alpha \sum_{i=1}^{m} \mathbb{I}\left[a_i \notin \mathcal{E}^i_{train}\right] + \beta \sum_{j=1}^{n} \mathbb{I}\left[f_j \notin \mathcal{F}_{train}\right] + \gamma \mathbb{I}\left[(f_1, f_2, \ldots, f_k) \notin \mathcal{P}_{train}\right] + C_T \tag{11}$$

as a measurement of task discrepancy Δ_{task}, where α, β, γ are weighting parameters for different novelty types and C_T is task specific constant. $\mathcal{E}^i_{train}$, $\mathcal{F}_{train}$, and $\mathcal{P}_{train}$ denote the bit-wise element set, relation set and the order of relation set used during training.

We establish a critical threshold beyond which CoT reasoning fails exponentially:

Theorem 5.1 (Task Generalization Failure Threshold). *There exists a threshold τ such that when $TGC(C) > \tau$, the probability of correct CoT reasoning drops exponentially:*

$$P(correct|C) \leq e^{-\delta(TGC(C)-\tau)} \tag{12}$$

The proof is provided in Appendix A.2.

5.1. Transformation Generalization

Transformation generalization evaluates the ability of CoT reasoning to effectively transfer when models encounter novel transformations during testing, which is an especially prevalent scenario in real-world applications.

Experimental Setup. To systematically evaluate the impact of transformations, we conduct experiments by varying transformations between training and testing sets while keeping other factors constant (e.g., elements, length, and format). Guided by the intuition formalized in Proposition 5.1, we define four incremental levels of distribution shift in transformations as shown in Figure 2: (i) In-Distribution (ID): The transformations in the test set are identical to those observed during training, e.g., $f_1 \circ f_1 \rightarrow f_1 \circ f_1$. (ii) Composition (CMP): Test samples comprise novel compositions of previously encountered transformations, though each individual transformation remains familiar, e.g., $f_1 \circ f_1, f_1 \circ f_2, f_2 \circ f_1 \rightarrow f_2 \circ f_2$. (iii) Partial Out-of-Distribution (POOD): Test data include compositions involving at least one novel transformation not seen during training, e.g., $f_1 \circ f_1 \rightarrow f_1 \circ f_2$. (iv) Out-of-Distribution (OOD): The test set contains entirely novel transformation types that are unseen in training, e.g., $f_1 \circ f_1 \rightarrow f_2 \circ f_2$.

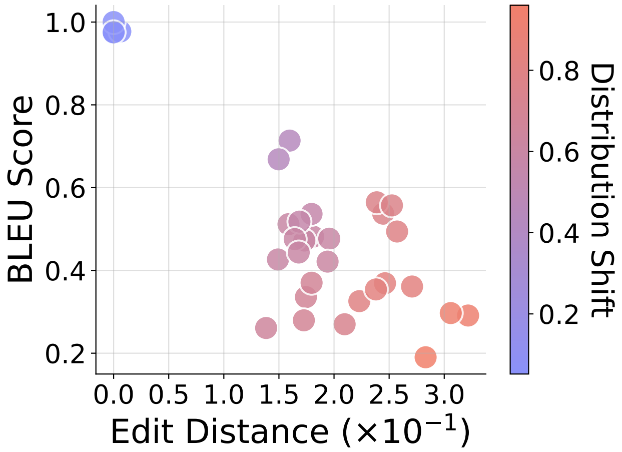


Figure 3 | Performance of CoT reasoning on transformation generalization. Efficacy of CoT reasoning declines as the degree of distributional discrepancy increases.

Table 1 | Full chain evaluation under different scenarios for transformation generalization.

Transformation (Train → Test)	Scenario	Exact Match	Edit Distance	BLEU Score
$f_1 \circ f_1 \rightarrow f_1 \circ f_1$	ID	100.00%	0	1
$\{f_2 \circ f_2, f_1 \circ f_2, f_2 \circ f_1\} \rightarrow f_1 \circ f_1$	CMP	0.01%	0.1326	0.6867
$f_1 \circ f_2 \rightarrow f_1 \circ f_1$	POOD	0.00%	0.1671	0.4538
$f_2 \circ f_2 \rightarrow f_1 \circ f_1$	OOD	0.00%	0.2997	0.2947

Findings. Figure 3 illustrates the performance of the full chain under different distribution discrepancies computed by *task generalize complexities* (normalized between 0 and 1) in Definition 5.1. We can observe that, in general, the effectiveness of CoT reasoning decreases when distribution discrepancy increases. For the instance shown in Table 1, from in-distribution to composition, POOD, and OOD, the exact match decreases from 1 to 0.01, 0, and 0, and the edit distance increases from 0 to 0.13, 0.17 when tested on data with transformation $f_1 \circ f_1$. Apart from ID, LLMs cannot produce a correct full chain in most cases, while they can produce correct CoT reasoning when exposed to some composition and POOD conditions by accident. As shown in Table 2, from $f_1 \circ f_2$ to $f_2 \circ f_2$, the LLMs can correctly answer 0.1% of questions. A close examination reveals that it is a coincidence, e.g., the query element is A, N, A, N, which happened to produce the same result for the two operations detailed in the Appendix D.1. When further analysis is performed by breaking the full chain into reasoning steps and answers, we observe strong consistency between the reasoning steps and answers. For example, under the composition generalization setting, the reasoning steps are entirely correct on test data distribution $f_1 \circ f_1$ and $f_2 \circ f_2$, but with wrong answers. Probe these insistent cases in Appendix D.1, we can find that when a novel transformation (say $f_1 \circ f_1$) is present, LLMs try to generalize the reasoning paths based on the most similar ones (i.e., $f_1 \circ f_2$) seen during training, which leads to correct reasoning paths, yet incorrect answer, which echo the example in the introduction. Similarly, generalization from $f_1 \circ f_2$ to $f_2 \circ f_1$ or vice versa allows LLMs to produce correct answers that are attributed to the commutative property between the two orthogonal transformations with unfaithful reasoning paths. Collectively, the above results indicate that the CoT reasoning fails to generalize to novel transformations, not even to novel composition transforms. Rather than demonstrating a true understanding of text, CoT reasoning under task transformations appears to reflect a replication of patterns learned during training.

Table 2 | Evaluation on different components in CoT reasoning on transformation generalization. CoT reasoning shows inconsistency with the reasoning steps and answers.

Transformation (Train → Test)	**Exact Match**			**Edit Distance**		
	Reason	Answer	Full Chain	Reason	Answer	Full Chain
$\{f_1 \circ f_1, f_1 \circ f_2, f_2 \circ f_1\} \rightarrow f_2 \circ f_2$	100.00%	0.01%	0.01%	0.000	0.481	0.133
$\{f_1 \circ f_2, f_2 \circ f_1, f_2 \circ f_2\} \rightarrow f_1 \circ f_1$	100.00%	0.01%	0.01%	0.000	0.481	0.133
$f_1 \circ f_2 \rightarrow f_2 \circ f_1$	0.00%	100.00%	0.00%	0.373	0.000	0.167
$f_2 \circ f_1 \rightarrow f_1 \circ f_2$	0.00%	100.00%	0.00%	0.373	0.000	0.167

Experiment settings. To further probe *when* CoT reasoning can generalize to unseen transformations, we conduct supervised fine-tuning (SFT) on a small portion λ of unseen data. In this way, we can decrease the distribution discrepancy between the training and test sets, which might help LLMs to generalize to test queries.

Findings. As shown in Figure 4, we can find that generally a very small portion ($\lambda = 1.5e^{-4}$) of data can make the model quickly generalize to unseen transformations. The less discrepancy between the training and testing data, the quicker the model can generalize. This indicates that a similar pattern appears in the training data, helping LLMs to generalize to the test dataset.

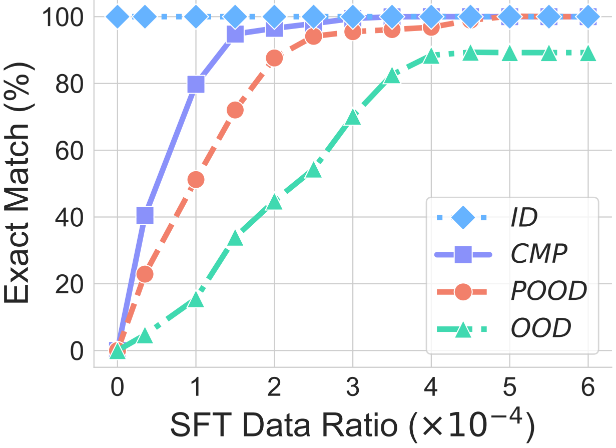


Figure 4 | Performance on unseen transformation using SFT in various levels of distribution shift. Introducing a small amount of unseen data helps CoT reasoning to generalize across different scenarios.

5.2. Element Generalization

Element generalization is another critical factor to consider when LLMs try to generalize to new tasks.

Experiment settings. Similar to transformation generalization, we fix other factors and consider three progressive distribution shifts for elements: ID, CMP, and OOD, as shown in Figure 2. It is noted that in composition, we test if CoT reasoning can be generalized to novel combinations when seeing all the basic atoms in the elements, e.g., $(\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}) \rightarrow (\mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{A})$. Based on the atom order in combination (can be measured by edit distance n), the CMP can be further developed. While for OOD, atoms that constitute the elements are totally unseen during the training.

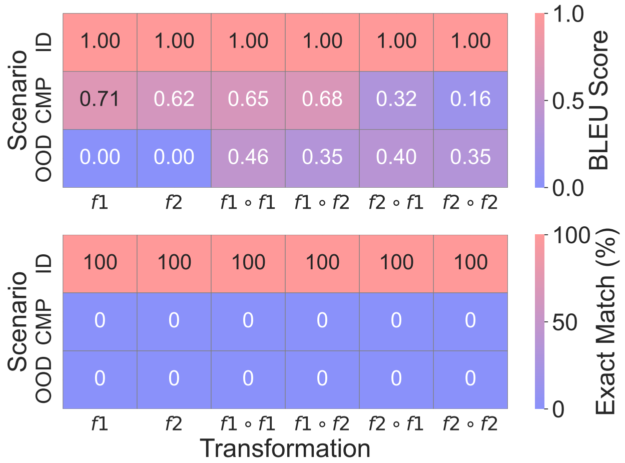
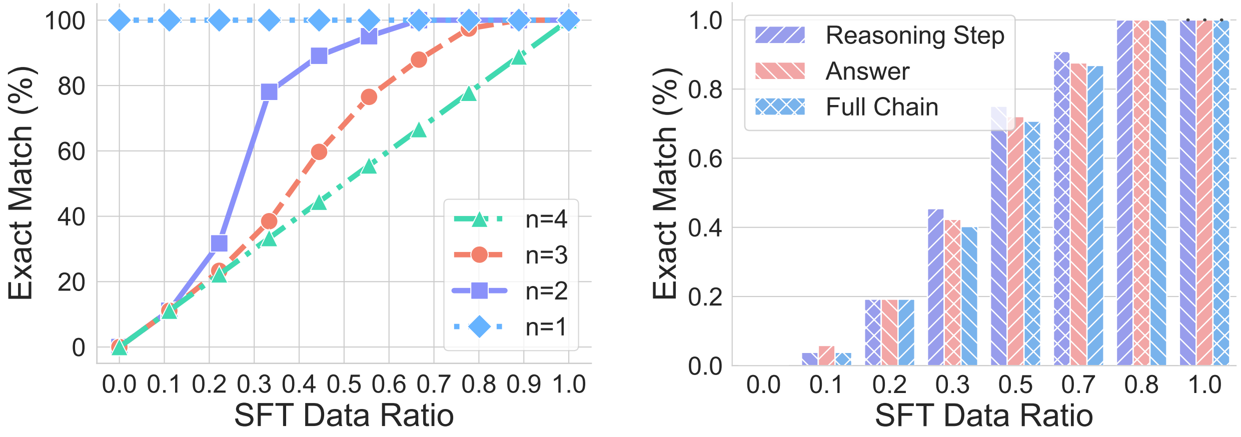


Figure 5 | Element generalization results on various scenarios and relations.

Findings. Similar to transformation generalization, the performances degrade sharply when facing the distribution shift consistently across all transformations, as shown in Figure 5. From ID to CMP and OOD, the exact match decreases from 1.0 to 0 and 0, for all cases. Most strikingly, the BLEU score is 0 when transferred to f_1 and f_2 transformations. A failure case in Appendix D.1 shows that the models cannot respond to any words when novel elements are present. We further explore when CoT reasoning can generalize to novel elements by conducting SFT. The results are summarized in Figure 6. We evaluate the performance under three exact matches for the full chain under three scenarios, CMP based on the edit distance n. The result is similar to SFT on transformation. The performance increases rapidly when presented with similar (a small n) examples in the training data. Interestingly, the exact match rate for CoT reasoning aligns with the lower bound of performance when $n = 3$, which might suggest the generalization of CoT reasoning on novel elements is very limited, even SFT on the downstream task. When we further analyze the exact match of reasoning, answer, and token during the training for $n = 3$, as summarized in Figure 6b. We find that there is a mismatch of accuracy between the answer and the reasoning step during the training process, which somehow might provide an explanation regarding why CoT reasoning is inconsistent in some cases.



(a) Performance on unseen element via SFT in various CMP scenarios.

(b) Evaluation of CoT reasoning in SFT.

Figure 6 | SFT performances for element generalization. SFT helps to generalize to novel elements.

6. Length Generalization

Length generalization examines how CoT reasoning degrades when models encounter test cases that differ in length from their training distribution. The difference in length could be introduced from the text space or the reasoning space of the problem. Therefore, we decompose length generalization into two complementary aspects: text length generalization and reasoning step generalization. Guided by instinct, we first propose to measure the length discrepancy.

Length Extrapolation Bound. We establish a power-law relationship for length extrapolation:

Proposition 6.1 (Length Extrapolation Gaussian Degradation)**.** *For a model trained on chain-of-thought sequences of fixed length L_{train}, the generalization error at test length L follows a Gaussian distribution:*

$$\mathcal{E}(L) = \mathcal{E}_0 + (1 - \mathcal{E}_0) \cdot \left(1 - \exp\left(-\frac{(L - L_{train})^2}{2\sigma^2}\right)\right) \tag{13}$$

where $\mathcal{E}_0$ is the in-distribution error at $L = L_{train}$, σ is the length generalization width parameter, and L is the test sequence length

The proof is provided in Appendix A.3.

Table 3 | Evaluation for text length generalization.

Length	Exact Match (%)			Edit Distance			BLEU Score		
	Full Chain	Reason	Answer	Full Chain	Reason	Answer	Full Chain	Reason	Answer
2	0.00%	0.00%	0.00%	0.3772	0.4969	0.5000	0.4214	0.1186	0.0000
3	0.00%	0.00%	0.00%	0.2221	0.3203	0.2540	0.5471	0.1519	0.0000
4	100.00%	100.00%	100.00%	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
5	0.00%	0.00%	0.00%	0.1818	0.2667	0.2000	0.6220	0.1958	0.2688
6	0.00%	0.00%	0.00%	0.3294	0.4816	0.3337	0.4763	0.1174	0.2077

6.1. Text Length Generalization

Text length generalization evaluates how CoT performance varies when the input text length (i.e., the element length l) differs from training examples. Considering the way LLMs process long text, this aspect is crucial because real-world problems often involve varying degrees of complexity that manifest as differences in problem statement length, context size, or information density.

Experiment settings. We pre-train LLMs on the dataset with text length merely on $l = 4$ while fixing other factors and evaluate the performance on a variety of lengths. We consider three different padding strategies during the pre-training: (i) None: LLMs do not use any padding. (ii) Padding: We pad LLM to the max length of the context window. (iii) Group: We group the text and truncate it into segments with a maximum length.

Findings. As illustrated in the Table 3, the CoT reasoning failed to directly generate two test cases even though those lengths present a mild distribution shift. Further, the performance declines as the length discrepancy increases shown in Figure 7. For instance, from data with $l = 4$ to those with $l = 3$ or $l = 5$, the BLEU score decreases from 1 to 0.55 and 0.62. Examples in Appendix D.1 indicate that LLMs attempt to

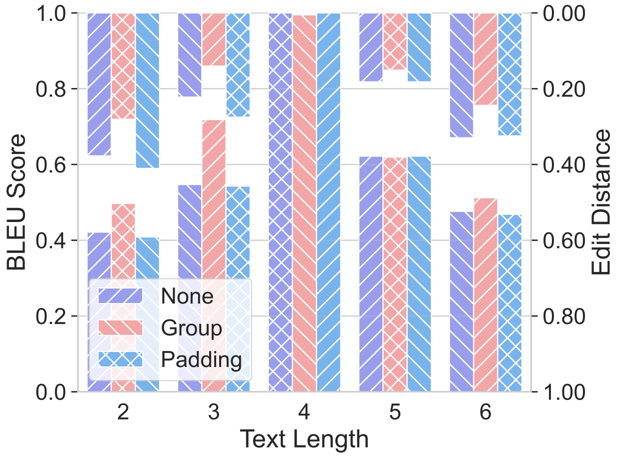


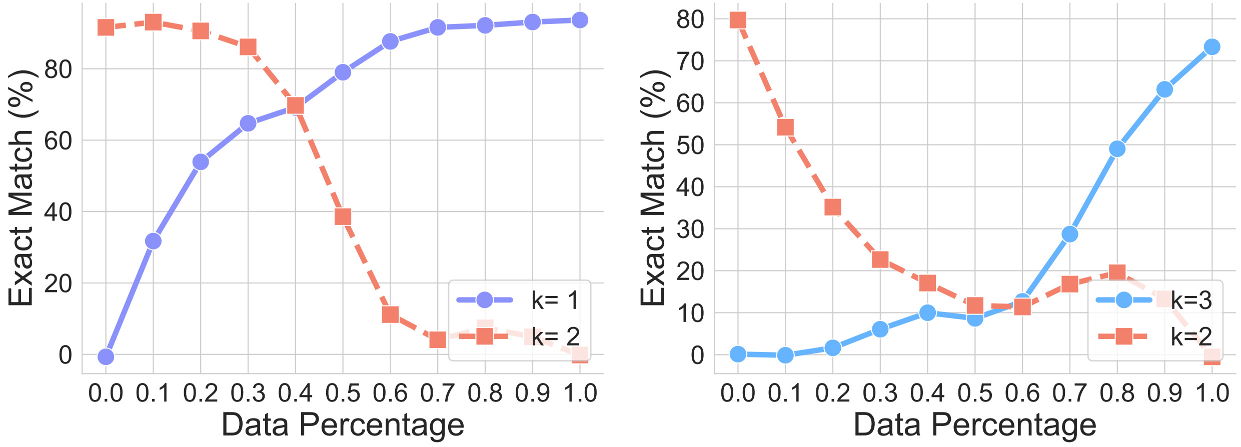
Figure 7 | Performance of text length generalization across various padding strategies. Group strategies contribute to length generalization.

produce CoT reasoning with the same length as the training data by adding or removing tokens in the reasoning chains. The efficacy of CoT reasoning length generalization deteriorates as the discrepancy increases. Moreover, we consider using a different padding strategy to decrease the divergence between the training data and test cases. We found that padding to the max length doesn't contribute to length generalization. However, the performance increases when we replace the padding with text by using the group strategy, which indicates its effectiveness.

6.2. Reasoning Step Generalization

The reasoning step generalization investigates whether models can extrapolate to reasoning chains requiring different steps k from those observed during training. which is a popular setting in multi-step reasoning tasks.

Experiment settings. Similar to text length generalization, we first pre-train the LLM with reasoning step $k = 2$, and evaluate on data with reasoning step $k = 1$ or $k = 3$.



(a) Reasoning step. From k=2 to k=1

(b) Reasoning step. From k=2 to k=3

Figure 8 | Test performance for reasoning-step generalization across varying training data compositions. Performance varies systematically with changes in the distribution of training data components.

Findings. As showcased in Figure 8, CoT reasoning cannot generalize across data requiring different reasoning steps, indicating the failure of generalization. Then, we try to decrease the distribution discrepancy introduced by gradually increasing the ratio of unseen data while keeping the dataset size the same when pre-training the model. And then, we evaluate the performance on two datasets. As we can observe, the performance on the target dataset increases along with the ratio. At the same time, the LLMs can not generalize to the original training dataset because of the small amount of training data. The trend is similar when testing different-step generalization, which follows the intuition and validates our hypothesis directly.

7. Format Generalization

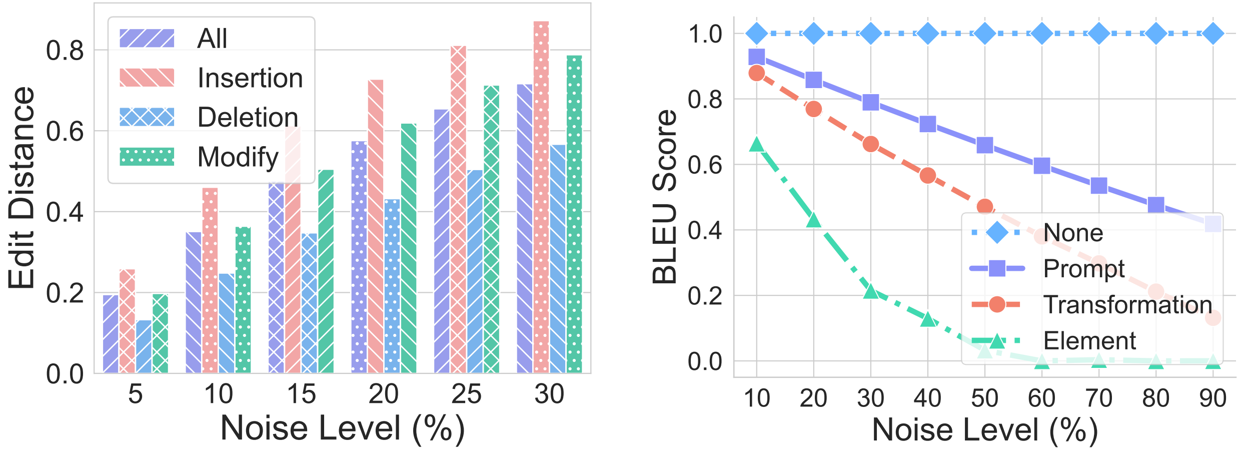
Format generalization assesses the robustness of CoT reasoning to surface-level variations in test queries. This dimension is especially crucial for determining whether models have internalized flexible, transferable reasoning strategies or remain reliant on the specific templates and phrasings encountered during training.

Format Alignment Score. We introduce a metric for measuring prompt similarity:

Definition 7.1 (Format Alignment Score)**.** *For training prompt distribution P_{train} and test prompt p_{test}:*

$$PAS(p_{test}) = \max_{p \in P_{train}} \cos(\phi(p), \phi(p_{test})) \tag{14}$$

where ϕ is a prompt embedding function.



(a) Format generalization. Performance under various perturbation methods.

(b) Format generalization. Performance vs. various applied perturbation areas.

Figure 9 | Performance of format generalization. Testing performance varies with different noise levels and areas where the noise is applied.

Experiment settings. To systematically probe this, we introduce four distinct perturbation modes to simulate scenario in real-world: (i) insertion, where a noise token is inserted before each original token; (ii) deletion: it deletes the original token; (iii) modification: it replaces the original token with a noise token; and (iv) hybrid mode: it combines multiple perturbations. Each mode is applied for tokens with probabilities p, enabling us to quantify the model's resilience to increasing degrees of prompt distribution shift.

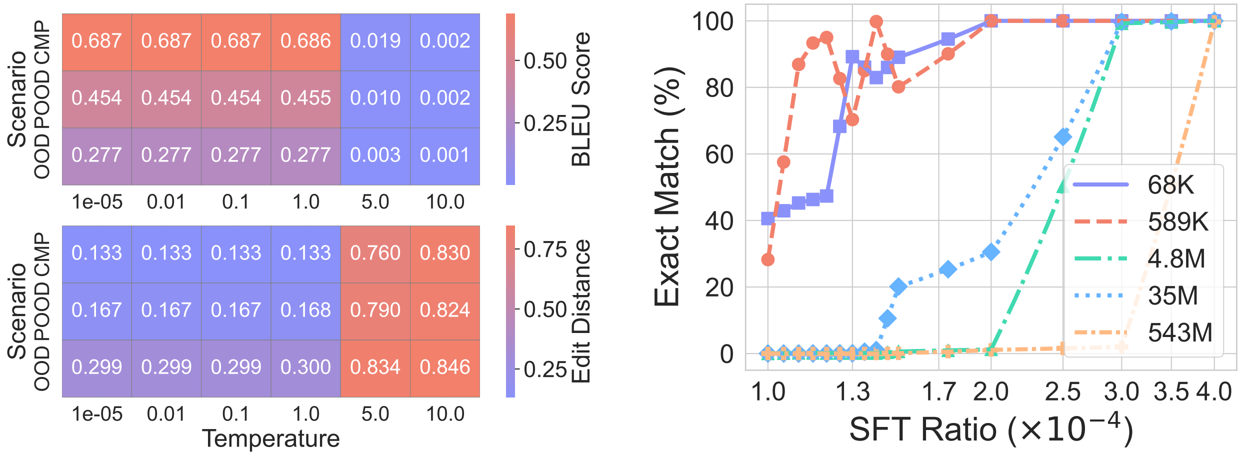
Findings. As shown in Figure 9a, we found that generally CoT reasoning can be easily affected by the format changes. No matter insertion, deletion, modifications, or hybrid mode, it creates a format discrepancy that affects the correctness. Among them, the deletion slightly affects the performance. While the insertions are relatively highly influential on the results. We further divide the query into several sections: elements, transformations, and prompt tokens. As shown in Figure 9b, we found that the elements and transformation play an important role in the format, whereas the changes to other tokens rarely affect the results.

8. Temperature and Model Size

Temperature and model size generalization explores how variations in sampling temperature and model capacity can influence the stability and robustness of CoT reasoning. For the sake of rigorous evaluation, we further investigate whether different choices of temperatures and model sizes may significantly affect our results.

Experiment settings. We explore the impact of different temperatures on the validity of the presented results. We adopt the same setting in the transformation generalization.

Findings. As illustrated in Figure 10a, LLMs tend to generate consistent and reliable CoT reasoning across a broad range of temperature settings (e.g., from 1e-5 up to 1), provided the values remain within a suitable range. This stability is maintained even when the models are evaluated under a variety of distribution shifts.



(a) Influences of various temperatures.

(b) Influences of various sizes.

Figure 10 | Temperature and model size. The findings hold under different temperatures and model sizes.

Experiment settings. We further examine the influence of model size by employing the same experimental configuration as used in the novel relation SFT study. In particular, we first pretrain models of different sizes using the transformation $f_1 \circ f_1$, and subsequently perform SFT on $f_2 \circ f_2$ while varying the SFT ratios.

Finding. Fig. 10b shows the accuracy of models with different sizes using different SFT ratios, which closely matches the result of our default model size across all evaluated settings and configurations.

9. Discussion and Implication

Our investigation, conducted through the controlled environment of DataAlchemy, reveals that the apparent reasoning prowess of Chain-of-Thought (CoT) is largely a **brittle mirage**. The findings across task, length, and format generalization experiments converge on a conclusion: CoT is not a mechanism for genuine logical inference but rather a sophisticated form of **structured pattern matching**, fundamentally bounded by the data distribution seen during training. When pushed even slightly beyond this distribution, its performance degrades significantly, exposing the superficial nature of the "reasoning" it produces.

While our experiments utilized models trained from scratch in a controlled environment, the principles uncovered are extensible to large-scale pre-trained models. We summarize the implications for practitioners as follows.

Guard Against Over-reliance and False Confidence. CoT should not be treated as a "plug-and-play" module for robust reasoning, especially in high-stakes domains like medicine, finance, or legal analysis. The ability of LLMs to produce **"fluent nonsense"**—plausible but logically flawed reasoning chains—can be more deceptive and damaging than an outright incorrect answer, as it projects a false aura of dependability. Sufficient auditing from domain experts is indispensable.

Prioritize Out-of-Distribution (OOD) Testing. Standard validation practices, where the test set closely mirrors the training set, are insufficient to gauge the true robustness of a CoT-enabled system. Practitioners must implement rigorous **adversarial and OOD testing** that systematically probes for vulnerabilities across task, length, and format variations.

Recognize Fine-Tuning as a Patch, Not a Panacea. Our results show that Supervised Fine-Tuning (SFT) can quickly "patch" a model's performance on a new, specific data distribution. However, this should not be mistaken for achieving true generalization. It simply expands the model's "in-distribution" bubble slightly. Relying on SFT to fix every OOD failure is an unsustainable and reactive strategy that fails to address the core issue: the model's lack of abstract reasoning capability.

10. Conclusion

In this paper, we critically examine the COT reasoning of LLMs through the lens of data distribution, revealing that the perceived structured reasoning capability largely arises from inductive biases shaped by in-distribution training data. We propose a controlled environment, DataAlchemy, allowing systematic probing of CoT reasoning along three crucial dimensions: task structure, reasoning length, and query format. Empirical findings consistently demonstrate that CoT reasoning effectively reproduces reasoning patterns closely aligned with training distributions but suffers significant degradation when faced with distributional deviations. Such observations reveal the inherent brittleness and superficiality of current CoT reasoning capabilities. We provide insights that emphasize real-world implications for both practitioners and researchers.

References

O. Bentham, N. Stringham, and A. Marasovic. Chain-of-thought unfaithfulness as disguised accuracy. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ydcrP55u2e. Reproducibility Certification.

M. Budnikov, A. Bykova, and I. P. Yamshchikov. Generalization potential of large language models. *Neural Computing and Applications*, 37(4):1973–1997, 2025.

Q. Chen, L. Qin, J. Liu, D. Peng, J. Guan, P. Wang, M. Hu, Y. Zhou, T. Gao, and W. Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025a.

Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025b.

H. Cho, J. Cha, P. Awasthi, S. Bhojanapalli, A. Gupta, and C. Yun. Position coupling: Improving length generalization of arithmetic transformers using task structure. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=5cIRdGM1uG.

S. Garg, D. Tsipras, P. S. Liang, and G. Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

S. Imani, L. Du, and H. Shrivastava. Mathprompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, 2023.

A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

S. Kambhampati. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18, 2024.

S. Kambhampati, K. Stechly, and K. Valmeekam. (how) do reasoning models reason? *Annals of the New York Academy of Sciences*, 1547(1):33–40, 2025.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

H. Li, S. Lu, P.-Y. Chen, X. Cui, and M. Wang. Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=n7n8McETXw.

Y. Li, Z. Lai, W. Bao, Z. Tan, A. Dao, K. Sui, J. Shen, D. Liu, H. Liu, and Y. Kong. Visual large language models for generalized and specialized applications. *arXiv preprint arXiv:2501.02765*, 2025b.

Z. Ling, Y. Fang, X. Li, Z. Huang, M. Lee, R. Memisevic, and H. Su. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 36:36407–36433, 2023.

I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Z. Shen, H. Yan, L. Zhang, Z. Hu, Y. Du, and Y. He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025.

P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

J. Song, Z. Xu, and Y. Zhong. Out-of-distribution generalization via composition: a lens through induction heads in transformers. *Proceedings of the National Academy of Sciences*, 122(6):e2417182122, 2025.

K. Stechly, K. Valmeekam, and S. Kambhampati. Chain of thoughtlessness? an analysis of cot in planning. *Advances in Neural Information Processing Systems*, 37:29106–29141, 2024.

K. Stechly, K. Valmeekam, A. Gundawar, V. Palod, and S. Kambhampati. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens. *arXiv preprint arXiv:2505.13775*, 2025.

X. Tang, Z. Zheng, J. Li, F. Meng, S.-C. Zhu, Y. Liang, and M. Zhang. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv:2305.14825*, 2023.

K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Q. Team. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face*, 2024.

L. P.-Y. Ting, C. Zhao, Y.-H. Zeng, Y. J. Lim, and K.-T. Chuang. Beyond rag: Reinforced reasoning augmented generation for clinical notes. *arXiv preprint arXiv:2506.05386*, 2025.

Q. Wang, Y. Wang, Y. Wang, and X. Ying. Can in-context learning really generalize to out-of-distribution tasks? *arXiv preprint arXiv:2410.09695*, 2024.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Y. Wang, F.-C. Chang, and P.-Y. Wu. Chain-of-thought prompting for out-of-distribution samples: A latent-variable study. *arXiv e-prints*, pages arXiv–2504, 2025a.

Y. Wang, F.-C. Chang, and P.-Y. Wu. A theoretical framework for ood robustness in transformers using gevrey classes. *arXiv preprint arXiv:2504.12991*, 2025b.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

J. Xu, H. Fei, L. Pan, Q. Liu, M.-L. Lee, and W. Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, 2024.

J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

L. Yang, Y. Song, X. Ren, C. Lyu, Y. Wang, J. Zhuo, L. Liu, J. Wang, J. Foster, and Y. Zhang. Out-of-distribution generalization in natural language processing: Past, present, and future. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, 2023.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822, 2023.

X. Yao, R. Ren, Y. Liao, and Y. Liu. Unveiling the mechanisms of explicit cot training: How chain-of-thought enhances reasoning generalization. *arXiv e-prints*, pages arXiv–2502, 2025.

E. Yeo, Y. Tong, M. Niu, G. Neubig, and X. Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

Z. Yu, L. He, Z. Wu, X. Dai, and J. Chen. Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint arXiv:2310.04959*, 2023.

L. Yujian and L. Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

X. Zhang, C. Du, T. Pang, Q. Liu, W. Gao, and M. Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356, 2024a.

Y. Zhang, H. Wang, S. Feng, Z. Tan, X. Han, T. He, and Y. Tsvetkov. Can llm graph reasoning generalize beyond pattern memorization? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2289–2305, 2024b.

Z. Zhang, A. Zhang, M. Li, and A. Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5NTt8GFjUHkr.

Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024, 2024c.

C. Zhao, Z. Tan, C.-W. Wong, X. Zhao, T. Chen, and H. Liu. Scale: Towards collaborative content analysis in social science with large language model agents and human intervention. *arXiv preprint arXiv:2502.10937*, 2025.

W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

A. Proof of Theorems

A.1. Proof of CoT Generalization Bound

Proof. Let f_θ be a model trained on samples from the distribution $\mathcal{D}_{\text{train}}$ using a loss function $\ell(f_\theta(x), y)$ that is Λ-Lipschitz and bounded. The expected test risk is given by

$$R_{\text{test}}(f_\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}}\left[\ell(f_\theta(x), y)\right]. \tag{15}$$

We can decompose the test risk as

$$R_{\text{test}}(f_\theta) = R_{\text{train}}(f_\theta) + \left(R_{\text{test}}(f_\theta) - R_{\text{train}}(f_\theta)\right). \tag{16}$$

To bound the discrepancy between R_{test} and R_{train}, we invoke a standard result from statistical learning theory. Given that ℓ is Λ-Lipschitz and the discrepancy measure $\Delta(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$ is an integral probability metric (e.g., Wasserstein-1 distance), we have

$$|R_{\text{test}}(f_\theta) - R_{\text{train}}(f_\theta)| \leq \Lambda \cdot \Delta(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}). \tag{17}$$

Therefore, the test risk satisfies

$$R_{\text{test}}(f_\theta) \leq R_{\text{train}}(f_\theta) + \Lambda \cdot \Delta(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}). \tag{18}$$

We next account for the generalization gap between the empirical training risk $\hat{R}_{\text{train}}(f_\theta)$ and the expected training risk $R_{\text{train}}(f_\theta)$. By applying a concentration inequality (e.g., Hoeffding's inequality), with probability at least $1 - \delta$, we have

$$R_{\text{train}}(f_\theta) \le \hat{R}_{\text{train}}(f_\theta) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{19}$$

where n is the number of training samples.

Combining the above, we obtain that with high probability,

$$R_{\text{test}}(f_\theta) \le \hat{R}_{\text{train}}(f_\theta) + \Lambda \cdot \Delta(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{20}$$

This concludes the proof. □

A.2. Proof of Task Generalization Failure Threshold

We establish the exponential decay bound through a probabilistic analysis of reasoning failure modes in the presence of task generalization complexity.

Let Ω denote the sample space of all possible reasoning configurations, and let $C \in \Omega$ represent a specific configuration. We define the following events: A_i as the event that element a_i is novel, i.e., $a_i \notin \mathcal{E}^i_{\text{train}}$; F_j as the event that transformation f_j is novel, i.e., $f_j \notin \mathcal{F}_{\text{train}}$; and Q as the event that the transformation sequence $(f_1, f_2, \ldots, f_k)$ is novel, i.e., $(f_1, f_2, \ldots, f_k) \notin \mathcal{P}_{\text{train}}$.

Here we make the assumption that the reasoning failures induced by novel arguments, functions, and patterns contribute independently to the overall failure probability and hence we model the success probability as a product of component-wise success rates:

$$P(\text{correct}|C) = P_0 \prod_{i=1}^{m} \rho_a^{\mathbb{I}[A_i]} \prod_{j=1}^{n} \rho_f^{\mathbb{I}[F_j]} \rho_p^{\mathbb{I}[Q]} \rho_c^{C_T}$$

where $P_0 \in (0, 1]$ represents the baseline success probability when all components are within the training distribution, and $\rho_a, \rho_f, \rho_p, \rho_c \in (0, 1)$ are the degradation factors associated with novel arguments, functions, patterns, and task-specific complexity, respectively.

$$\ln P(\text{correct} \mid C) \;=\; \ln P_0 + \sum_{i=1}^{m} \mathbb{I}[A_i]\; \ln \rho_a + \sum_{j=1}^{n} \mathbb{I}[F_j]\; \ln \rho_f + \mathbb{I}[Q]\; \ln \rho_p + C_T\; \ln \rho_c \tag{21}$$

For notational convenience, we define the positive constants:

$$\xi_a := -\ln \rho_a > 0, \xi_f := -\ln \rho_f > 0, \xi_p := -\ln \rho_p > 0, \xi_c := -\ln \rho_c > 0$$

hence we have:

$$\ln P(\text{correct}|C) = \ln P_0 - \xi_a \sum_{i=1}^{m} \mathbb{I}[A_i] - \xi_f \sum_{j=1}^{n} \mathbb{I}[F_j] - \xi_p \mathbb{I}[Q] - \xi_c C_T \tag{22}$$

Lemma: Relationship to TGC. The expression in equation above can be bounded in terms of TGC(C) as follows:

$$\ln P(\text{correct}|C) \le \ln P_0 - \delta \cdot \text{TGC}(C) \tag{23}$$

where $\delta = \min(\frac{\xi_a}{\alpha}, \frac{\xi_f}{\beta}, \frac{\xi_p}{\gamma}, \xi_c) > 0$.

Proof of Lemma: From the definition of $\text{TGC}(C)$ in Eq. (11), we have:

$$\text{TGC}(C) = \alpha \sum_{i=1}^{m} \mathbb{I}[A_i] + \beta \sum_{j=1}^{n} \mathbb{I}[F_j] + \gamma \mathbb{I}[Q] + C_T \tag{24}$$

By the definition of δ, each term in Eq. (22) satisfies:

$$\xi_a \sum_{i=1}^{m} \mathbb{I}[A_i] \geq \delta\alpha \sum_{i=1}^{m} \mathbb{I}[A_i] \tag{25}$$

$$\xi_f \sum_{j=1}^{n} \mathbb{I}[F_j] \geq \delta\beta \sum_{j=1}^{n} \mathbb{I}[F_j] \tag{26}$$

$$\xi_p \mathbb{I}[Q] \geq \delta\gamma \mathbb{I}[Q] \tag{27}$$

$$\xi_c C_T \geq \delta C_T \tag{28}$$

Summing these inequalities establishes Eq. (23).

We now define the threshold $\tau := \frac{\ln P_0}{\delta}$. From Eq. (23), when $\text{TGC}(C) > \tau$, we have:

$$\begin{aligned} \ln P(\text{correct} \mid C) &\leq \ln P_0 - \delta \cdot \text{TGC}(C) && (29) \\ &= \delta(\tau - \text{TGC}(C)) && (30) \\ &= -\delta(\text{TGC}(C) - \tau) && (31) \end{aligned}$$

Exponentiating both sides yields the desired bound: $P(\text{correct} \mid C) \leq e^{-\delta(\text{TGC}(C)-\tau)}$

A.3. Proof of Length Extrapolation Bound

Proof. Consider a transformer model f_θ processing sequences of length L. The model implicitly learns position-dependent representations through positional encodings $\text{PE}(i) \in \mathbb{R}^d$ for position $i \in \{1, \ldots, L\}$ and attention patterns $A_{ij} = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d}}\right)$.

During training on fixed length L_{train}, the model learns a specific distribution:

$$p_{\text{train}}(\mathbf{h}) = p(\mathbf{h} \mid L = L_{\text{train}}) \tag{32}$$

where $\mathbf{h} = \{h_1, \ldots, h_L\}$ represents hidden states.

For sequences of length $L \neq L_{\text{train}}$, we encounter distribution shift in two forms: (1) positional encoding mismatch, where the model has never seen positions $i > L_{\text{train}}$ if $L > L_{\text{train}}$, and (2) attention pattern disruption, where the learned attention patterns are calibrated for length L_{train}.

The KL divergence between training and test distributions can be bounded:

$$D_{KL}(p_{\text{test}} \| p_{\text{train}}) \propto |L - L_{\text{train}}|^2 \tag{33}$$

This quadratic relationship arises from linear accumulation of positional encoding errors and quadratic growth in attention pattern misalignment due to pairwise interactions.

Let $\mathcal{E}(L)$ be the prediction error at length L. We decompose it as:

$$\mathcal{E}(L) = \mathcal{E}_{\text{inherent}}(L) + \mathcal{E}_{\text{shift}}(L) \tag{34}$$

where $\mathcal{E}_{\text{inherent}}(L) = \mathcal{E}_0$ is the inherent model error (constant) and $\mathcal{E}_{\text{shift}}(L)$ is the error due to distribution shift.

The distribution shift error follows from the Central Limit Theorem. As the error accumulates over sequence positions, the total shift error converges to:

$$\mathcal{E}_{\text{shift}}(L) = (1 - \mathcal{E}_0) \cdot \left(1 - \exp\left(-\frac{(L - L_{\text{train}})^2}{2\sigma^2}\right)\right) \tag{35}$$

This form ensures that $\mathcal{E}_{\text{shift}}(L_{\text{train}}) = 0$ (no shift at training length) and $\lim_{|L-L_{\text{train}}|\to\infty} \mathcal{E}_{\text{shift}}(L) = 1 - \mathcal{E}_0$ (maximum error bounded by 1).

The width parameter σ depends on:

$$\sigma = \sigma_0 \cdot \sqrt{\frac{d}{L_{\text{train}}}} \tag{36}$$

where σ_0 is a model-specific constant, d is the model dimension, and the $\sqrt{d/L_{\text{train}}}$ factor captures the concentration of measure in high dimensions.

Therefore, the total error follows:

$$\mathcal{E}(L) = \mathcal{E}_0 + (1 - \mathcal{E}_0) \cdot \left(1 - \exp\left(-\frac{(L - L_{\text{train}})^2}{2\sigma^2}\right)\right) \tag{37}$$

This Gaussian form naturally emerges from the accumulation of position-dependent errors and matches the experimental observation of near-zero error at $L = L_{\text{train}}$ with symmetric increase in both directions. □

B. Experiment Details

The model is trained using the AdamW optimiser in mixed precision (FP16). The default learning rate is 3×10^{-3}, and the schedule follows a cosine decay with a 10% warm-up ratio. Training is conducted for 10 epochs, using a batch size of 1024. A weight decay of 0.01 is applied, and gradient norms are clipped at 1.0.

C. Illustration of Datasets

Below are the examples of transformation f_1 and f_2:

Transformation[F1]: A A F Q [F1] <answer> N N S D
Transformation[F2]: A A L P [F2] <answer> A L P A

aside from single transformation, we can composite transformations arbitrarily:

Transformation[F1F2]: A C I A [F1] [F2] <think>
N P V N [F2] <answer>
P V N N
Transformation[F2F2]: N O V S [F2] [F2] <think>
O V S N [F2] <answer>
V S N O

we use exact match, edit distance and BELU score to measure the discrepancy between generated tokens and the labels. For more than one transformation examples, we can further measure the discrepancy for reasoning and answering seperately.

D. Additional Experimental Results

D.1. Additional Qualitative Analysis

D.1.1. Orthognal Transformation Caused Coincidence

The following case shows that even if the transformation is different, the model that trained on transformation $f_2 \circ f_1$ can still provide correct answer through incorrect reasoning:

```
Prompt:     ’A A A B [F1] [F2] <answer>’
Generated:’B A A A [F1] <answer> O N N N’
Expected:   ’O N N N’
```

D.1.2. Correct reasoning but failed in final answer

The following case shows that the model pretrained on the union of three transformation $f_1 \circ f_2, f_2 \circ f_1, f_2 \circ f_2$ and test on $f_1 \circ f_1$

```
Prompt:  ’A A A D [R1] [R1] <think>’
Generated:  ’N N N Q [R1] <answer> N N Q N’
Expected:  ’N N N Q [R1] <answer> A A A D’
```

D.1.3. Failure to generalize to novel element

The following case shows that the model trained on element set $a_i \in [A, M]$ can not generalize to unseen elements such as N or O

```
Prompt:  ’N N N O [F1] [F1] <think>’
Generated:  ’R V Q S [F1] <answer> E I D F’
Expected:  ’A A A B [F1] <answer> N N N O’
```

D.1.4. LLM reproduces CoT reasoning at seen lengths

The following case shows that model trained under $f_1 \circ f_1$ tried to reproduce the length in training data by adding tokens in the reason chain even prompted with seen transformation f_1

```
Prompt:  ’A A B D [f1] <answer>’
Generated:  ’N O A Z N N O Q [f1]
             <answer> A A B D’
Expected:  ’N N O Q’
```