Chain-of-Thought Reasoning for Math: Theoretical Foundation and Applications

Jessica E. Liang [1]

Abstract

Chain-of-Thought (CoT) prompting improves the reasoning capabilities of large language models (LLMs), but its theoretical basis remains poorly understood. We propose an information-theoretic framework to analyze and improve CoT through two complementary lenses. First, we model CoT as a Markov process $X \to Z \to Y$, where intermediate steps Z mediate information from inputs X to outputs Y. By applying the Data Processing Inequality and Fano's inequality, we show that explicit reasoning lowers the bound on prediction error. Second, we use Partial Information Decomposition (PID) to quantify how CoT rationales contribute to task performance. Our analysis reveals strong synergy, i.e., reasoning and answers together, provide more information than either alone. Building on this insight, we introduce a PID-guided loss that promotes synergy during CoT distillation. On the e-SNLI dataset, this approach outperforms standard fine-tuning and mutual information baselines. To validate CoT's benefits in structured domains, we also study few-shot arithmetic reasoning. CoT prompting boosts accuracy from 4% to 70% with just one example and up to 90% with four, far surpassing regular prompting. Overall, our findings offer a theoretical foundation for CoT and suggest new strategies for improving reasoning in LLMs.

1. Introduction

Recent advances in large language models (e.g., GPT-4.5, GPT-4o, PaLM) have demonstrated remarkable capabilities in tasks ranging from text generation to question-answering and language understanding. However, these models often rely on implicit representations when transforming input to output in a single pass. Chain-of-Thought (CoT) prompting has been proposed as a technique that induces the model to break down complex tasks into step-by-step reasoning, often leading to significant improvements in accuracy, especially in mathematical, commonsense, and multi-step reasoning tasks (Wei et al., 2023; Kojima et al., 2022).

CoT prompting is the practice of instructing a model to think through each step of a solution explicitly before arriving at an answer. Instead of directly producing an answer from the prompt, the model generates intermediate reasoning steps. For example,

- Mathematical problem-solving: The model writes down intermediate calculations (e.g., partial sums, factoring steps, etc.).
- Logical puzzle solving: The model enumerates possible scenarios, discards contradictions, and narrows down to the final conclusion.

Empirical results suggest that when CoT is used, the final answers are more accurate. However, these improvements are primarily observed empirically, prompting the question: *What is the theoretical underpinning?*

Despite CoT's empirically demonstrated success, a thorough theoretical rationale for *why* chain-of-thought helps remains elusive. In this work, we propose a perspective rooted in information theory, particularly the Data Processing Inequality (DPI) and the Partial Information Decomposition (PID), to shed light on how intermediate "explanations" or "reasoning steps" might preserve and highlight the most relevant information for a final prediction.

Large language models often rely on CoT reasoning, where the model generates intermediate reasoning steps (Z) before producing a final answer (Y). Empirically, this "think out loud" approach has been shown to improve performance on complex tasks compared to providing a direct answer.

The main contributions of this paper are as follows:

1. *An information-theoretic explanation of CoT using DPI.* While empirical studies have demonstrated the effec-

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA. Correspondence to: Jessica E. Liang <jeliang@seas.upenn.edu>.

The second AI for MATH Workshop at the 42^{nd} International Conference on Machine Learning, Vancouver, Canada.

tiveness of CoT prompting in LLMs, its theoretical underpinnings remain underdeveloped. We address this gap by formally modeling CoT reasoning as a Markov process $X \to Z \to Y$, enabling a principled analysis through the DPI. Additionally, we invoke Fano's inequality to explain why CoT can theoretically improve prediction accuracy. Specifically, if the intermediate reasoning steps (Z) capture more relevant information about the target output (Y) than the input (X) alone, the lower bound on the prediction error decreases. Thus, a model trained to generate explicit CoT steps can achieve improved empirical performance. We validate this theory empirically on four diverse NLP benchmarks—e-SNLI (textual entailment), ANLI (adversarial NLI), CommonsenseQA (commonsense reasoning), and SVAMP (arithmetic reasoning)—consistently observing increased mutual information and higher predictive accuracy under CoT prompting (Hsieh et al., 2023).

2. *Interpretability of CoT via PID.* We introduce a novel application of PID (Williams & Beer, 2010) to analyze the informational roles of CoT rationales. PID enables a fine-grained decomposition of mutual information into unique, redundant, and synergistic components, providing deeper insights into how different sources contribute to a model's predictions. Drawing inspiration from structured multi-step reasoning puzzles (e.g., hat puzzles (Winkler, 2003)), we demonstrate how PID reveals interactions between inputs, rationales, and outputs. Our analysis across four datasets reveals varied informational patterns: in e-SNLI, rationales provide significant unique information beyond the input; in ANLI, rationales add little beyond what the input already provides; and in CommonsenseQA and SVAMP, synergy dominates-neither the input nor rationale alone suffices, but their combination is informative. We also show that this insight can guide CoT distillation, with PID-enhanced training leading to improved performance.

2. Related Work

2.1. Foundations and Theoretical Insights

CoT reasoning has emerged as a pivotal approach for enhancing the problem-solving capabilities of LLMs. By decomposing complex tasks into intermediate reasoning steps, CoT enables LLMs to excel in mathematical, logical, and commonsense reasoning. This approach was popularized by (Wei et al., 2023), who demonstrated that CoT achieves state-of-the-art results on benchmarks like GSM8K, often outperforming fine-tuned models with verifiers. Subsequent theoretical analyses, such as those by (Feng et al., 2023), have examined CoT from a model-capacity perspective, underscoring the role of intermediate reasoning steps in leveraging the latent capabilities of LLMs.

A major research focus has been ensuring the reliability of CoT reasoning. (Lyu et al., 2023) introduced the notion of faithful CoT reasoning, emphasizing the need for logical coherence and alignment with ground truth. Surveys by (Chu et al., 2024) have explored CoT's potential extensions to multimodal and multilingual contexts, while (Zhang et al., 2022) proposed automated prompting techniques to reduce reliance on handcrafted CoT prompts.

2.2. Interpreting CoT

Interpreting the reasoning processes within CoT is crucial for understanding its mechanisms and improving its reliability. (Lanham et al., 2023) proposed metrics to assess the faithfulness of CoT-generated reasoning chains, ensuring that the model's reasoning aligns with its outputs. Additionally, (Saparov & He, 2022) highlighted pitfalls in reasoning coherence, showing that LLMs often generate plausible yet unfaithful reasoning paths that do not accurately reflect their internal computations.

Information-theoretic perspectives have further contributed to CoT interpretability. (Ton et al., 2024) explored how information gain at each reasoning step can be quantified, aiding in the identification of failure modes. Meanwhile, (Wang & Zhou, 2024) proposed methods for extracting CoT reasoning paths without explicit prompts, facilitating a more autonomous generation of intermediate reasoning steps. (Chen et al., 2024) also applied mutual information to CoT distillation, enhancing the efficiency of knowledge transfer.

To improve CoT interpretability, researchers have introduced logical frameworks. (Zhao et al., 2024) developed logical CoT strategies that integrate structured reasoning principles, guiding the generation of coherent reasoning chains. These efforts highlight the growing emphasis on aligning CoT reasoning with human expectations and enhancing transparency.

2.3. Applications and Extensions

CoT has demonstrated versatility across a wide range of domains. In medicine, (Miao et al., 2024) explored its application in nephrology, improving diagnostic decision-making. In broader problem-solving contexts, (Suzgun et al., 2022) evaluated CoT on complex BIG-Bench tasks, where it frequently surpassed human-level reasoning.

Recent innovations have further optimized CoT prompting strategies. Active prompting (Diao et al., 2024) improves prompt selection, while deductive verification (Ling et al., 2023) enhances logical consistency. Contrastive prompting (Chia et al., 2023) refines CoT-generated reasoning

processes by explicitly contrasting correct and incorrect reasoning paths.

Generalized CoT frameworks have expanded its capabilities. Tree-of-Thought (ToT) and Graph-of-Thought (GoT) models explore multiple reasoning paths in parallel, improving problem-solving efficiency (Xia et al., 2025). Additionally, multimodal extensions such as EmbodiedGPT (Mu et al., 2023) integrate visual inputs with CoT reasoning, broadening its applicability across diverse data modalities.

2.4. Robustness, Optimization, and Advancements

The robustness of CoT prompting has been critically examined to understand when it enhances performance and when it may be detrimental. (Chen et al., 2023) analyzed optimal conditions for CoT application, while (Liu et al., 2024) identified scenarios where CoT can degrade performance, underscoring the need for task-specific designs.

Several optimization techniques have been introduced to refine CoT reasoning. Chain of Preference Optimization (Zhang et al., 2024) leverages feedback to improve accuracy and user alignment. Meanwhile, emerging security concerns have also been addressed. (Xiang et al., 2024) revealed vulnerabilities to backdoor attacks, emphasizing the need for robust safeguards against adversarial manipulation.

Advancements such as graph-based reasoning (Jin et al., 2024) and compositional CoT prompting (Mitra et al., 2024) underscore the potential of CoT to tackle increasingly complex tasks. Ensuring the faithfulness of CoT reasoning remains a significant challenge. Logical CoT frameworks (Zhao et al., 2024) have been developed to enhance zero-shot reasoning capabilities, while analyses of transformer expressivity in CoT settings (Merrill & Sabharwal, 2023) have demonstrated the model's enhanced reasoning capabilities.

3. Interpreting CoT via DPI

3.1. CoT as a Markov Chain

In a CoT reasoning, assume the following

- X: The initial input (or problem statement).
- Z: An intermediate (latent) representation or "thought" derived from X.
- Y: The final output (or solution) that is ultimately derived from Z.

In CoT reasoning, we often explicitly or implicitly generate a series of intermediate steps or "thoughts" (summarized as Z) from the input X. We then derive the final answer Y based on those intermediate steps.

A Markov chain $X \to Z \to Y$ implies the Markov property, namely (Cover & Thomas, 2006):

$$p(Y \mid X, Z) = p(Y \mid Z). \tag{1}$$

Equivalently, one can say that once Z is known, Y is conditionally independent of X. In other words, any knowledge about X beyond what is already encoded in Z does not affect our predictions of Y. This conditional independence can also be written in terms of joint distributions:

$$p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z). \tag{2}$$

Here's why this factorization makes sense in a CoT context:

1. X Given Z
   Once we know Z, which is assumed to capture the essential information derived from X, how X relates to Z can be treated on its own. Formally, $p(X \mid Z)$ quantifies how the original input X might have led to that intermediate representation Z.
2. Y Given Z
   Similarly, once Z is known, Y depends only on Z. In the chain-of-thought framework, Z is supposed to encapsulate all necessary reasoning or knowledge to arrive at Y. Hence, $p(Y \mid Z)$ captures how we generate the final answer from the latent thought process.

Putting these two pieces together leads to:

$$p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z). \tag{3}$$

This equality expresses conditional independence of X and Y once Z is known.

In CoT prompting, the intermediate steps Z are explicitly generated by the model, though in practice it is often kept "internal." The assumption is that once the chain-of-thought (i.e., Z) is fixed, the final answer Y is conditionally determined. This perspective helps us break down complex reasoning tasks into more manageable chunks: first encode or derive Z from X, and then use Z to produce Y. Based on the above analysis, in CoT reasoning,

$$X \longrightarrow Z \longrightarrow Y \tag{4}$$

3.2. Interpreting CoT via DPI

For Markov Chain in CoT reasoning,

$$X \longrightarrow Z \longrightarrow Y, \tag{5}$$

we can get

$$I(X; Y) \leq I(Z; Y) \tag{6}$$

based on DPI (Cover & Thomas, 2006).

Let $\hat{Y}$ be a predictor of the correct label Y. We define

$$P_e = \Pr\{\hat{Y} \neq Y\}. \tag{7}$$

Our goal is to understand *the best possible* P_e that any estimator can achieve, given some variable(s) from which it decodes. We can use Fano's inequality (Cover & Thomas, 2006) to make analysis.

In its simplest form (with base-2 logarithms), Fano's inequality says that (Cover & Thomas, 2006),

$$H(A \mid \hat{A}) \leq P_e \log_2 |\mathcal{A}| + 1, \tag{8}$$

where $P_e = \Pr\{\hat{A} \neq A\}$. We can easily show that

$$P_e \geq 1 - \frac{I(A; \hat{A}) + 1}{\log_2 |\mathcal{A}|}. \tag{9}$$

The proof is in Appendix A.

To study predicting Y (the true answer) from either X or Z, we identify $A \equiv Y$ in Fano's setup. Thus:

- Direct (no CoT): Predict Y from X.
- Chain-of-Thought: Predict Y from Z.

Fano's inequality tells us the minimum achievable error (among all possible predictors) is bounded *below* by a function of the mutual information between the predictor's input and the true label Y.

Let $\hat{Y}_X = f(X)$ be a predictor that uses X alone. By Fano's inequality:

$$P_e(X) = \Pr[\hat{Y}_X \neq Y] \geq 1 - \frac{I(X; Y) + 1}{\log_2 |\mathcal{Y}|}. \tag{10}$$

Now let $\hat{Y}_Z = g(Z)$ be a predictor that uses the chain-of-thought Z. Then:

$$P_e(Z) = \Pr[\hat{Y}_Z \neq Y] \geq 1 - \frac{I(Z; Y) + 1}{\log_2 |\mathcal{Y}|}. \tag{11}$$

Based on the DPI in (6),

$$1 - \frac{I(Z; Y) + 1}{\log_2 |\mathcal{Y}|} \leq 1 - \frac{I(X; Y) + 1}{\log_2 |\mathcal{Y}|}. \tag{12}$$

This shows that the Fano lower bound for predicting Y from Z is *smaller* than the Fano lower bound for predicting Y from X. Hence, CoT performs better. The lower bound on error from X is *larger* than the lower bound from Z. This means that using Z *permits* a smaller achievable error rate, consistent with empirical findings.

4. Interpreting CoT via PID

4.1. PID for CoT

We consider the following variables in CoT reasoning:

- X: The *input* or *question* given to the model.
- C: The *chain-of-thought (rationale)* generated by the model.
- Y: The *final label* or *answer* produced by the model.
- T: A *target variable* of interest (e.g., "Is the solution both correct and interpretable?").

PID provides a finer analysis of mutual information by dividing it into four distinct components: redundancy, synergy, and the unique contributions of C and Y (Williams & Beer, 2010). The breakdown of mutual information between C and Y concerning T is given by (Williams & Beer, 2010):

$$\begin{aligned} I(C, Y; T) \quad = \quad & R(C, Y; T) + S(C, Y; T) \\ & + U(C; T|Y) + U(Y; T|C), \end{aligned} \tag{13}$$

where:

- $R(C, Y; T)$ denotes the redundant information about T that is captured simultaneously by both C and Y.
- $S(C, Y; T)$ represents the synergistic information about T that arises only when C and Y are jointly considered.
- $U(C; T|Y)$ captures the unique contribution of C to T that is not shared with Y.
- $U(Y; T|C)$ reflects the unique information that Y provides about T independently of C.

Figure 1 visualizes the decomposition of $I(C, Y; T)$. The overlapping red region corresponds to redundancy $R(C, Y; T)$, while the blue section highlights synergy $S(C, Y; T)$. The orange and green portions indicate the unique contributions of the vision and text modalities, $U(C; T|Y)$ and $U(Y; T|C)$, respectively (Williams & Beer, 2010).

The individual components of the decomposition are computed as follows (Williams & Beer, 2010):

$$R(C, Y; T) = I(C; Y), \tag{14}$$

$$\begin{aligned} S(C, Y; T) = & I(C, Y; T) - I(C; T|Y) \\ & - I(Y; T|C) - I(C; Y), \end{aligned} \tag{15}$$

$$U(C; T|Y) = I(C; T|Y), \tag{16}$$

$$U(Y; T|C) = I(T; T|C). \tag{17}$$

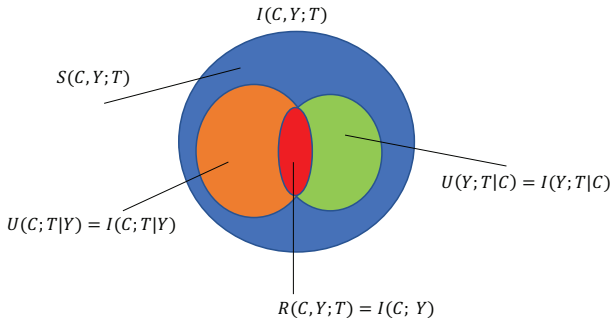

Figure 1. Partial Information Decomposition (PID) for CoT reasoning. The diagram illustrates the redundant component $R(C, Y; T)$, synergistic component $S(C, Y; T)$, and the unique contributions $U(C; T|Y)$ and $U(Y; T|C)$ (Williams & Beer, 2010).

4.2. An Example of Arithmetic-Type CoT

Question (X): "What is the sum of the digits of 194?"

A language model (or any reasoner) is prompted to provide:

1. A *Chain-of-Thought (rationale)* C.
2. A *Final Answer (label)* Y.

Rationale (C): "The digits are 1, 9, and 4. Summing them gives 1 + 9 + 4 = 14."

Final Answer (Y): 14

Define the target variable: T = "Is the solution correct and interpretable to a human?"

- *Looking only at* Y: We know the final numerical result is 14, so we can check *numerical correctness* but not necessarily how the answer was obtained.
- *Looking only at* C: We see how digits are summed, which suggests 14 is the total. But without seeing the final *explicit* answer, we might not be sure that the model's final concluded answer was indeed 14 (in more complex scenarios, rationales might contain steps that get overruled or misapplied).
- *Looking at* (C, Y) *jointly*: We confirm the process *and* the final result. This combination often provides *synergistic* information about whether the solution is *both correct and transparent*.

From a PID perspective, the synergy term captures how C and Y *together* convey information about T that neither can fully convey alone. If we want to *maximize synergy*, we would aim to produce a rationale C and a final label Y such that T is only fully resolved by viewing C and Y in tandem. In this arithmetic example, the CoT C and the final label Y each provide partial (and sometimes overlapping) information about whether the solution is correct and interpretable (T). The synergetic information is that which arises only when C and Y are considered together. While this example is simple, it illustrates how PID can bring a structured lens to analyzing—and potentially optimizing—CoT explanations.

4.3. An Example of Logic Reasoning CoT

To clearly illustrate the synergy between CoT rationales (C) and final labels (Y), we consider a logical puzzle involving three individuals—Alice, Bob, and Charlie—each wearing a hat that is either red or blue. The details of this example are provided in Appendix B. Each person can see the hats of the other two but not their own. They know there are exactly two hats of one color and one hat of the other, but initially do not know which color is in the majority.

The reasoning unfolds as follows: Alice sees Bob and Charlie's hats. If both hats were blue, she would immediately deduce her own hat is red. Her initial uncertainty implies Bob and Charlie cannot both be blue. Bob, aware of Alice's uncertainty and seeing Charlie's blue hat, concludes he himself must have a red hat; otherwise, Alice would have already deduced her hat color. Finally, Charlie realizes that Bob's confident deduction implies his own hat must be the single blue one.

An LLM-generated CoT rationale (C) explicitly articulates each of these logical inference steps, while the final label (Y) succinctly states, "Charlie is wearing a blue hat." If we consider each component separately, the rationale alone provides logical steps without explicitly verifying the final solution, and the final label alone provides a conclusion without justification. Only by jointly examining C and Y can we confidently verify the correctness and logical validity of the solution. Thus, this puzzle effectively demonstrates how synergy, as defined by PID, is crucial in fully validating both correctness and interpretability within CoT reasoning tasks.

5. Experiments

We evaluate our approaches on four widely-used datasets, covering three distinct NLP tasks (Hsieh et al., 2023): for textual entailment, we utilize the Explainable SNLI (e-SNLI) dataset (Camburu et al., 2018) alongside the Adversarial NLI (ANLI) dataset (Nie et al., 2020); for commonsense reasoning tasks, we employ CommonsenseQA (CQA) (Talmor et al., 2019; Rajani et al., 2019); and for arithmetic reasoning, we leverage the SVAMP dataset (Patel et al., 2021). We obtained the datasets and their corresponding CoT reasoning from the GitHub repository provided by (Hsieh et al., 2023). The CoT reasoning was generated using

Table 1. Mutual information and Fano lower bounds using direct input (X) vs. CoT explanations (Z) across datasets. In all cases, $I(Z;Y) > I(X;Y)$, supporting the Data Processing Inequality and demonstrating the informativeness of CoT reasoning.

Dataset	$I(X;Y)$ (bits)	$I(Z;Y)$ (bits)	Fano Bound (X)	Fano Bound (Z)
e-SNLI	0.0012	0.0017	0.3683	0.3680
ANLI	0.0039	0.0059	-0.0039	-0.0059
CQA	0.3633	0.4174	-0.3633	-0.4174
SVAMP	0.3933	0.3968	-0.3933	-0.3968

T5 models as described in (Hsieh et al., 2023). We analyze the four datasets and their CoT reasoning using DPI and PID. In Appendix C, we provide another set of experiments based on synthetic data.

5.1. DPI for CoT Reasoning

We evaluate the DPI in the context of CoT reasoning across four diverse NLP datasets. Under the Markov chain $X \to Z \to Y$, where X is the direct input, Z is the CoT explanation, and Y is the target label, DPI states that $I(X;Y) \leq I(Z;Y)$. To assess this, we compute the mutual information between input/explanation and the label using dit package in Python (James et al., 2018), and use Fano's inequality to estimate the minimum achievable prediction error.

On the e-SNLI dataset, we observe that the mutual information between the CoT explanation (Z) and the true label (Y), $I(Z;Y) = 0.0017$ bits, exceeds the mutual information between the direct input (X) and the label, $I(X;Y) = 0.0012$ bits. The corresponding Fano lower bounds indicate that CoT permits a slightly smaller minimum prediction error (0.3680 vs. 0.3683), consistent with the expectations of DPI.

For the ANLI dataset, we use a simulated binary correctness label. Here, $I(Z;Y_{\text{bin}}) = 0.0059$ bits is greater than $I(X;Y_{\text{bin}}) = 0.0039$ bits. The Fano lower bounds, although negative due to the low mutual information values, still show a lower theoretical error when using CoT (-0.0059) compared to the direct input (-0.0039). This reinforces the advantage of CoT reasoning even in adversarial entailment settings.

In the CQA dataset, CoT explanations provide $I(Z;Y_{\text{bin}}) = 0.4174$ bits, significantly more than $I(X;Y_{\text{bin}}) = 0.3633$ bits. This difference is also reflected in the Fano bounds (-0.4174 for Z vs. -0.3633 for X), demonstrating that CoT enhances the amount of useful information in commonsense reasoning tasks.

Finally, in the SVAMP arithmetic reasoning dataset, we again find that CoT provides higher mutual information ($I(Z;Y_{\text{bin}}) = 0.3968$ bits vs. $I(X;Y_{\text{bin}}) = 0.3933$ bits) and a lower Fano error bound (-0.3968 vs. -0.3933). Although the difference is marginal, it still supports the theoretical benefits of CoT.

Across all datasets, the DPI condition holds: $I(Z;Y) > I(X;Y)$, and CoT consistently leads to lower theoretical error bounds. These findings, summarized in Table 1, provide strong empirical evidence that CoT reasoning improves informativeness in a range of reasoning tasks.

5.2. PID for CoT Analysis

To analyze the informational contribution of CoT explanations, we applied Partial Information Decomposition (PID) using the PID_WB method from the `dit` Python package (James et al., 2018). PID decomposes mutual information into unique, redundant, and synergistic components, providing a fine-grained view of how input sources contribute to target prediction. We conducted experiments across several NLP datasets, using input components (e.g., premise, hypothesis, rationale) as sources and the label as the target.

On the e-SNLI dataset, the hypothesis alone provided nearly all predictive information (1.5680 bits), while the premise added no unique contribution. Replacing the hypothesis with a gold rationale resulted in slightly higher unique information (1.5753 bits), suggesting rationales offer stronger predictive signals than hypotheses for NLI.

For ANLI, the combined input (premise + hypothesis) was solely responsible for predictive information (1.5509 bits), with no unique or synergistic value added by the rationale. This contrasts with e-SNLI and may reflect differences in task difficulty or rationale quality.

In contrast, both CQA and SVAMP exhibited fully synergistic information patterns. In CQA, all 6.6439 bits of information arose jointly from the question and rationale, while in SVAMP, all 3.3219 bits came from the combination of problem and CoT explanation. Neither input alone held predictive value, indicating the necessity of contextual reasoning.

Table 2 summarizes the results. Collectively, these findings highlight the varying roles CoT rationales play—from unique information sources in classification to synergistic components in complex reasoning tasks.

Table 2. Summary of PID experiments across datasets.

Dataset	Source 1	Source 2	Primary PID Component	Information (bits)
e-SNLI	Premise	Hypothesis	Unique (Hypothesis)	1.5680
e-SNLI	Premise	Rationale	Unique (Rationale)	1.5753
ANLI	Premise+Hypothesis	Rationale	Unique (Premise+Hypothesis)	1.5509
CQA	Question	Rationale	Synergy	6.6439
SVAMP	Problem	Rationale	Synergy	3.3219

5.3. PID for CoT Distillation

This work interprets CoT reasoning through the lens of DPI and PID, which has broad applicability across explainable AI. As a concrete use case, we apply PID to enhance CoT distillation.

We propose a loss function that simultaneously optimizes label prediction and rationale generation while incorporating synergy-based regularization. The objective is defined as:

$$\begin{aligned}\mathcal{L} \quad = \quad & \mathcal{L}_{\text{label}} + \alpha \mathcal{L}_{\text{rationale}} - \beta S(p, h; y) \\ & -\gamma S(p, h; r), \end{aligned} \tag{18}$$

where p is the premise, h the hypothesis, y the label, and r the rationale. Here, $\mathcal{L}_{\text{label}}$ is the standard cross-entropy loss for predicting labels (Hsieh et al., 2023), and $\mathcal{L}_{\text{rationale}}$ is an auxiliary cross-entropy loss for generating a coherent rationale explaining the prediction (Hsieh et al., 2023). The terms $S(p, h; y)$ and $S(p, h; r)$ represent synergy terms from PID with respect to the label and rationale, respectively.

We use the `T5-small` model (Raffel et al., 2019) as the student in our CoT distillation experiments, trained on the e-SNLI dataset (Camburu et al., 2018), which includes premise-hypothesis pairs annotated with labels and human-written rationales. The data is tokenized using the `T5` tokenizer to maintain consistency. Training is performed for five epochs using the Hugging Face `Trainer` API with a learning rate of 1.0×10^{-4}, batch size of 32, and Adam optimizer. We selected the regularization parameters through hyperparameter tuning, setting $\alpha = 1.0$, $\beta = 0.1$, and $\gamma = 0.1$.

Experiments were conducted on a Google Colab T4 GPU. The e-SNLI training set includes 549,367 examples, and the test set contains 9,824 examples. Training took approximately six hours, and inference required about two hours.

Table 3 compares the performance of our PID-based method with prior approaches. The fine-tuning (FT) baseline (Dodge et al., 2020) achieves 82.90% accuracy. DSS (Hsieh et al., 2023), which structures the distillation process, improves performance to 83.43%. An MI-based approach (Chen et al., 2024) that aligns CoT rationales with labels obtains 83.23%.

Our PID-based method achieves the highest accuracy at 83.71%, indicating that modeling and maximizing informational synergy can enhance both reasoning and prediction in CoT learning. These results underscore the promise of PID in guiding more interpretable and effective distillation strategies.

Method	Accuracy (%)
FT (Dodge et al., 2020)	82.90
DSS (Hsieh et al., 2023)	83.43
MI (Chen et al., 2024)	83.23
Our PID	**83.71**

Table 3. Comparison of different approaches on the e-SNLI test dataset.

5.4. CoT Prompting for Arithmetic Reasoning

To evaluate the effectiveness of different prompting strategies in arithmetic reasoning, we constructed a synthetic dataset consisting of mathematical expressions that involve a combination of addition, subtraction, and multiplication operations. Each expression is randomly generated in the following format: `num1 operator1 num2 operator2 num3`

The operators `operator1` and `operator2` are independently sampled from the set $\{+, -, *\}$, and the operands `num1`, `num2`, and `num3` are integers randomly drawn from the range $[0, 9]$. Operator precedence is respected (i.e., multiplication is evaluated before addition or subtraction).

The dataset is composed of:

- 50 evaluation examples, generated with a fixed random seed of 42.
- 32 training examples, generated using a different seed (43), which are used for constructing few-shot prompt demonstrations.

We employed the instruction-tuned language model `google/gemma-2b-it`, which is designed to follow task instructions and produce appropriate completions. This model is suitable for evaluating general-purpose prompt-

ing strategies on reasoning tasks. Two types of prompting methods were compared:

1. **Regular Few-Shot Prompting:** Each demonstration in the prompt is presented as a question-answer pair, e.g., "`3 + 5 * 2?  13`", without any explanation or reasoning steps.
2. **Chain-of-Thought (CoT) Prompting:** Each demonstration includes a full decomposition of the arithmetic operations leading to the final result. For example, "`3 + 5 * 2?  5 * 2 = 10.  3 + 10 = 13.  13`".

We varied the number of few-shot examples in the prompt ($N \in \{1, 2, 4, 8, 16, 32\}$) and recorded the model's accuracy over the 50 evaluation expressions for both prompting strategies. Results are summarized in Table 4.

N	Few-Shot Accuracy (%)	CoT Accuracy (%)
1	4.0	70.0
2	4.0	86.0
4	24.0	90.0
8	20.0	86.0
16	20.0	88.0
32	28.0	86.0

Table 4. Arithmetic accuracy of the model under few-shot and CoT prompting strategies across different values of N.

Figure 2 visualizes the accuracy trends for both prompting strategies.

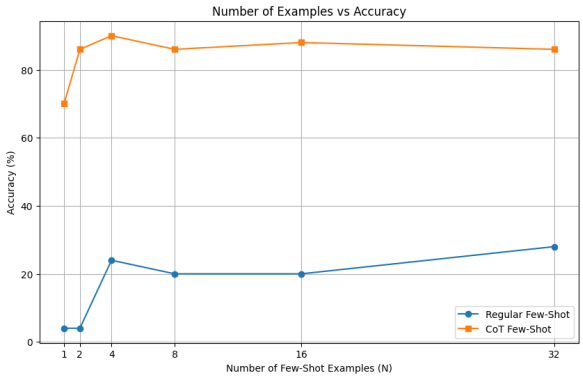

Figure 2. Comparison of Few-Shot and Chain-of-Thought (CoT) Prompting Accuracy as a Function of N.

As observed in Table 4 and Figure 2, CoT prompting consistently outperforms regular few-shot prompting across all values of N. With just a single example, CoT achieves 70% accuracy—demonstrating strong inductive generalization even in low-shot settings. Accuracy peaks at 90% with four CoT examples and remains stable thereafter. In contrast, regular few-shot prompting yields limited gains, reaching a maximum of only 28% with 32 examples. These results highlight the importance of intermediate reasoning in arithmetic tasks and underscore the benefits of structured thought processes in improving model performance.

6. Conclusions and Future Work

We presented an information-theoretic perspective on CoT reasoning in LLMs, offering two complementary frameworks. First, we modeled CoT as a Markov process $X \to Z \to Y$, where Z denotes intermediate reasoning steps. By applying the DPI, we showed that CoT can preserve or increase mutual information between inputs and outputs. Second, we used PID to analyze the contributions of CoT explanations (C) and final answers (Y) to task performance (T). We found consistent *synergy*, where CoT and predictions together conveyed more information than either alone—offering a theoretical basis for CoT's empirical success.

We also demonstrated that PID can enhance CoT distillation. Our proposed loss jointly optimizes rationales and predictions while encouraging high synergy. Experiments on e-SNLI showed that this approach outperforms standard distillation methods, confirming the practical value of information-theoretic objectives.

To further validate our framework, we conducted a focused study on arithmetic reasoning. We found that CoT prompting dramatically improves accuracy over regular prompting—achieving 70% with just one example and up to 90% with four. In contrast, standard few-shot prompting plateaued at 28%. These results support the idea that intermediate reasoning enables stronger generalization, especially in low-shot settings.

In future work, mathematics remains a promising and underexplored frontier for CoT prompting. Future research could examine symbolic manipulation, multi-step algebra, and formal proofs to better understand how reasoning steps interact in highly structured domains. Extending PID to quantify the information flow in multi-hop or hierarchical mathematical reasoning could yield deeper insights into how LLMs generalize beyond surface patterns. Scaling PID to broader tasks and modalities may also reveal richer dynamics in reasoning. Additionally, optimizing prompting strategies for synergy—via Tree-of-Thoughts, graph-based inference, or adaptive prompting—offers a promising direction. Finally, the DPI and PID tools developed here can generalize to other interpretability frameworks, including retrieval-augmented generation (RAG), multi-hop inference, and policy learning in reinforcement learning.

References

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, 2018.

Chen, J., Chen, L., Huang, H., and Zhou, T. When do you need chain-of-thought prompting for chatgpt?, 2023. URL https://arxiv.org/abs/2304.03262.

Chen, X., Huang, H., Gao, Y., Wang, Y., Zhao, J., and Ding, K. Learning to maximize mutual information for chain-of-thought distillation. *arXiv preprint arXiv:2403.03348*, 2024.

Chia, Y. K., Chen, G., Tuan, L. A., Poria, S., and Bing, L. Contrastive chain-of-thought prompting. *arXiv preprint arXiv:2301.11903*, 2023.

Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B., and Liu, T. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future, 2024. URL https://arxiv.org/abs/2309.15402.

Cover, T. and Thomas, J. *Elements of information theory*. Wiley-Interscience, 2006.

Diao, S., Wang, P., Lin, Y., Pan, R., Liu, X., and Zhang, T. Active prompting with chain-of-thought for large language models, 2024. URL https://arxiv.org/abs/2302.12246.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5235–5247. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.468. URL https://aclanthology.org/2020.acl-main.468/.

Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., and Wang, L. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qHrADgAdYu.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

James, R. G., Ellison, C. J., and Crutchfield, J. P. dit: a Python package for discrete information theory. *The Journal of Open Source Software*, 3(25):738, 2018. doi: https://doi.org/10.21105/joss.00738.

Jin, B., Xie, C., Zhang, J., Roy, K. K., Zhang, Y., Li, Z., Li, R., Tang, X., Wang, S., Meng, Y., and Han, J. Graph chain-of-thought: Augmenting large language models by reasoning on graphs, 2024. URL https://arxiv.org/abs/2404.07103.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., and Su, H. Deductive verification of chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2306.03872.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., and Griffiths, T. L. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse, 2024. URL https://arxiv.org/abs/2410.21333.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., and Rao, D. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.11903*, 2023.

Merrill, W. and Sabharwal, A. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*, 2023.

Miao, J., Thongprayoon, C., Suppadungsuk, S., Krisanapan, P., Radhakrishnan, Y., and Cheungpasitporn, W. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1), 2024. ISSN 1648-9144. doi: 10.3390/medicina60010148. URL https://www.mdpi.com/1648-9144/60/1/148.

Mitra, C., Huang, B., Darrell, T., and Herzig, R. Compositional chain-of-thought prompting for large multimodal models, 2024. URL https://arxiv.org/abs/2311.17076.

Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., and Luo, P. EmbodiedGPT: Vision-language pre-training via embodied chain

of thought. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=IL5zJqfxAa.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Patel, A., Bhattamishra, S., and Goyal, N. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, 2021. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942, Florence, Italy, 2019. Association for Computational Linguistics.

Saparov, A. and He, H. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*, 2022.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. URL https://arxiv.org/abs/2210.09261.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

Ton, J.-F., Taufiq, M. F., and Liu, Y. Understanding chain-of–thought in llms through information theory. *arXiv preprint arXiv:2411.11984*, 2024.

Wang, X. and Zhou, D. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Williams, P. L. and Beer, R. D. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010. URL https://arxiv.org/abs/1004.2515.

Winkler, P. *Mathematical Puzzles: A Connoisseur's Collection*. A K Peters/CRC Press, 2003. ISBN 9781568812014. URL https://openlibrary.org/books/OL25435493M/Mathematical_puzzles_a_connoisseur%E2%80%99s_collection?utm_source=chatgpt.com.

Xia, Y., Wang, R., Liu, X., Li, M., Yu, T., Chen, X., McAuley, J., and Li, S. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms, 2025. URL https://arxiv.org/abs/2404.15676.

Xiang, Z., Jiang, F., Xiong, B., Ramasubramanian, B., et al. Badchain: Backdoor chain-of-thought prompting for large language models. *arXiv preprint arXiv:2403.11903*, 2024.

Zhang, X., Du, C., Pang, T., Liu, Q., Gao, W., and Lin, M. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *arXiv preprint arXiv:2403.14312*, 2024.

Zhang, Z., Zhang, A., Li, M., and Smola, A. Automatic chain of thought prompting in large language models, 2022. URL https://arxiv.org/abs/2210.03493.

Zhao, X., Li, M., Lu, W., Weber, C., Lee, J. H., Chu, K., and Wermter, S. Enhancing zero-shot chain-of-thought reasoning in large language models through logic, 2024. URL https://arxiv.org/abs/2309.13339.

A. Fano's Inequality

In its simplest form (with base-2 logarithms), Fano's inequality says that (Cover & Thomas, 2006),

$$H(A \mid \hat{A}) \leq P_e \log_2 |\mathcal{A}| + 1, \tag{19}$$

where $P_e = \Pr\{\hat{A} \neq A\}$. Rearranging, one obtains:

$$P_e \geq \frac{H(A \mid \hat{A}) - 1}{\log_2 |\mathcal{A}|}. \tag{20}$$

If A is uniformly distributed, we can rewrite $H(A \mid \hat{A})$ in terms of mutual information $I(A; \hat{A})$,

$$\begin{aligned} H(A|\hat{A}) &= H(A) - I(A; \hat{A}) && (21) \\ &= \log_2 |\mathcal{A}| - I(A; \hat{A}) && (22) \end{aligned}$$

which yields

$$P_e \geq 1 - \frac{I(A; \hat{A}) + 1}{\log_2 |\mathcal{A}|}. \tag{23}$$

B. An Example of Logic Reasoning in PID for CoT

Whereas a simple arithmetic example may only need one or two steps of reasoning, here we present a short logic puzzle where synergy between the rationale and final label is more pronounced. The puzzles requiring reasoning about other agents' knowledge (such as the hat puzzle in (Winkler, 2003)) serve as motivating examples for multi-step reasoning tasks.

Consider a puzzle involving three people—Alice, Bob, and Charlie—each wearing a hat. The hats can be either red or blue. All three participants know:

- There are exactly three hats in total, one per person.
- The distribution of colors is either: (2 hats of one color, 1 hat of the other color), i.e., it could be $(2\text{R}, 1\text{B})$ or $(2\text{B}, 1\text{R})$.
- Each person can *see* the hats of the other two, but not their own.
- Alice, Bob, and Charlie each must *guess* their own hat color.

They do *not* know *which* color is in the majority (red or blue), only that one color has 2 hats and the other color has 1 hat.

Suppose we pose the question to an LLM:

$$\begin{aligned} X = &\text{"Which color is Charlie's hat,} \\ &\text{and how can you prove it?"} \end{aligned}$$

We then instruct the model to produce:

$$\begin{aligned} C = &\textit{CoT rationale}, \\ Y = &\textit{Final Label (the model's answer)}. \end{aligned}$$

Here is the logical reasoning scenario that leads to the conclusion *(2 red, 1 blue)*, with Charlie as blue:

1. **Alice's perspective.** She sees Bob and Charlie:

   "If Bob and Charlie both had blue hats, then that would use up both blue hats. Hence I (Alice) must be wearing red. In that case, I would know my hat color immediately. But I am initially uncertain. Therefore, Bob and Charlie are not *both blue."*

2. **Bob's perspective.** Bob notices Alice is uncertain and sees Charlie's hat is *blue*:

"If I (Bob) were also blue, then Alice, upon seeing two blues, would know she is the single red. But Alice was not certain. Therefore, I must be red."

3. **Charlie's perspective.** Once Bob concludes he is red, Charlie thinks:

   "If I were also red, Bob's reasoning might have unfolded differently. Given Bob's announcement, I must be the lone blue hat."

Below is a hypothetical CoT *rationale* C an LLM might produce:

```
 ``Alice sees Bob and Charlie.  If Bob and Charlie both had blue hats,
Alice would immediately know she is wearing red (since there would be 2
blues in view).
But Alice does not immediately know her hat color, so it's unlikely both
are blue.
Bob sees that Alice is unsure, and Bob notices Charlie's hat is blue.
If Bob himself were also blue, then Alice would have known she was red.
Hence Bob deduces he must be wearing red.
Charlie realizes if he were also red, Bob might have shown a different
logic or timing.
Given Bob's confidence, Charlie concludes he is the single blue hat.''
```

The final label Y is succinct:

$$Y = \text{"Charlie is wearing a blue hat."}$$

Define the *target variable* T to be:

$$T = \text{"Is the puzzle's conclusion both } \textit{correct} \text{ and } \textit{logically justified}\text{?"}$$

- Only looking at Y: We learn the claim "Charlie is wearing a blue hat," which is correct if the puzzle indeed yields that distribution. But we do *not* see the reasoning that supports it.
- Only looking at C: We see logical steps involving Alice, Bob, and Charlie's knowledge states, but we do *not* see the final official answer. If the chain-of-thought is consistent, we *suspect* Charlie is blue, but we're not 100% sure the model's *declared* final color is blue.
- Looking at (C, Y) jointly: By combining the step-by-step logic (C) and the final claim (Y), we can confirm that the conclusion is both *(a)* correct under the puzzle constraints and *(b)* logically justified by the reasoning.

This is a more complex puzzle. Multi-step logic (even if it's just a few steps) shows how synergy between C (the reasoning) and Y (the conclusion) can be more evident than in a trivial puzzle. Seeing both the rationale and final label lets a human or another system verify *both* correctness and internal consistency. One could use an objective that *maximizes synergy* in PID terms, ensuring C and Y jointly provide more clarity about T than either does alone.

This puzzle demonstrates how synergy—in the sense of PID—applies to CoT reasoning. The final label Y alone reveals the conclusion ("Charlie is wearing a blue hat"), but the chain-of-thought C supplies the crucial logical steps. Only by considering both R and Y do we fully validate the conclusion is *both correct and justified*, thus illustrating the synergy term in PID.

C. Experimental Results Based on Synthetic Data

C.1. DPI for CoT

We perform a small-scale empirical study to illustrate how the Markov chain assumption $X \to Z \to Y$ in a CoT setting obeys the DPI. Concretely, we generate a synthetic dataset in which:

- X is sampled from a simple Bernoulli or categorical distribution,
- Z is determined by transforming X with some noise or probabilistic rule (e.g., "flip" with a small probability),
- Y is then generated from Z according to a similar probabilistic mapping.

Under this construction, once Z is known, the final label Y is conditionally independent of X. Formally, we have $X \to Z \to Y$ as a Markov chain. We draw a large number of samples (we used 10^5) to form an empirical joint distribution $\hat{p}(X, Z, Y)$ and then estimate the marginal distributions $\hat{p}(X, Y)$ and $\hat{p}(Z, Y)$. Next, we use standard discrete formulas to compute the mutual information $I(X; Y)$ and $I(Z; Y)$.

Table 5 shows a run result of this experiment:

Mutual Information	**Value (bits)**
$I(X; Y)$	0.1715
$I(Z; Y)$	0.5334

Table 5. Empirical estimates of $I(X; Y)$ and $I(Z; Y)$ from a synthetic CoT-style Markov chain.

Because Z incorporates and refines information from X, it is unsurprising that $I(Z; Y)$ exceeds $I(X; Y)$. This outcome agrees with the Data Processing Inequality, which states that: $I(X; Y) \leq I(Z; Y)$ whenever $X \to Z \to Y$ is a Markov chain.

Intuitively, knowledge of the intermediate variable Z reduces uncertainty about Y more effectively than observing X alone. In the context of CoT reasoning, one can interpret Z as the "internal" or "latent" chain-of-thought representation that guides the final output Y. This result thus supports the view that having access to the CoT (i.e., Z) can, in principle, improve our ability to infer or predict Y.

Empirically, even in a simplified discrete scenario, we see that $I(X; Y) \approx 0.1715$ bits, while $I(Z; Y) \approx 0.5334$ bits, which is over three times larger. These findings highlight how intermediate reasoning steps, when modeled as Z, encode crucial information about the final answer. Consequently, in any chain-of-thought style approach, revealing or utilizing Z (where feasible) naturally yields stronger predictive power about correctness or final responses than relying solely on the original input X.

C.2. Experiment on PID for Arithmetic-Type CoT

We conduct a small-scale experiment to illustrate how PID can analyze the roles of the CoT and the final answer in determining solution correctness and interpretability. Specifically, we construct a simple discrete distribution over the variables (C, Y, T), where:

- $C \in \{\text{coherent}, \text{partial}, \text{none}\}$ represents the chain-of-thought,
- $Y \in \{14, 13, 9\}$ denotes the final answer proposed by the model, and
- $T \in \{0, 1\}$ indicates whether the overall solution is *correct and interpretable* ($T = 1$) or not ($T = 0$).

An example assignment in our synthetic dataset is $\Pr\{C = \text{coherent}, Y = 14, T = 1\} = 0.30$, with other triplets receiving smaller or larger probabilities depending on correctness and coherence. After ensuring all probabilities sum to 1.0, we compute both the marginal mutual information values and the PID measures:

- $I(C; T)$ and $I(Y; T)$, respectively capturing how much the chain-of-thought alone and the final answer alone reveal about correctness and interpretability,
- $I(C, Y; T)$, the total information that (C, Y) jointly carry about T,
- The four PID components: *redundancy* R, *synergy* S, and the *unique* contributions $U(C; T \mid Y)$ and $U(Y; T \mid C)$.

Table 6. A representative outcome from the PID procedure, showing the decomposition of $I(C, Y; T)$ into redundancy $R(C, Y; T)$, synergy $S(C, Y; T)$, and unique information $U(\cdot; T \mid \cdot)$. All values are in bits.

Quantity	Value
$I(C; T)$	0.5568
$I(Y; T)$	0.3958
$I(C, Y; T)$	0.8813
$R(C, Y; T)$	0.1958
$S(C, Y; T)$	0.6855
$U(C; T \mid Y)$	0.3610
$U(Y; T \mid C)$	0.2000

A representative outcome from this procedure is summarized below:

The values highlight several interesting properties:

1. Chain-of-Thought Alone vs. Final Answer Alone. We observe that $I(C; T) > I(Y; T)$, suggesting that the chain-of-thought itself provides more information about correctness and interpretability than the final numeric answer alone. This can reflect scenarios where seeing the reasoning process clarifies possible mistakes or confirms logical consistency.

2. Synergy Dominates. The synergy $S(C, Y; T) \approx 0.6855$ is relatively large compared to the total $I(C, Y; T) \approx 0.8813$. This implies that the combination of seeing *both* the chain-of-thought and the final label provides substantially more information about whether the solution is correct and interpretable than either variable in isolation.

3. Unique Contributions. Both C and Y have nontrivial unique parts ($U(C; T \mid Y) = 0.3610$ and $U(Y; T \mid C) = 0.2000$), indicating that each variable reveals some aspect of correctness and interpretability that the other does not fully capture.

Overall, these simple results align with intuition that the *reasoning process* (C) and the *final answer* (Y) each carry partial, sometimes overlapping information about solution quality. Most notably, the large synergy term underscores how observing the *rationale along with* the final label can dramatically clarify correctness and interpretability. Even in this simplified discrete example, PID helps reveal the structure of how C and Y jointly determine the target T.

C.3. Experiment on PID for Logical Reasoning CoT

In this section, we illustrate how PID can be used to analyze CoT explanations in logical reasoning tasks. Specifically, we examine how the correctness of the chain-of-thought (C) and the correctness of the final answer (Y) together inform a target variable T, which indicates whether the model's solution is both *correct* and *logically justified*.

We constructed a small, synthetic dataset of logical puzzles where:

- $C \in \{0, 1\}$ captures whether the chain-of-thought is coherent or logically consistent;
- $Y \in \{0, 1\}$ reflects whether the final answer is correct;
- $T \in \{0, 1\}$ indicates whether the puzzle's solution (including the final answer and explanation) is *both* correct and justified.

For each puzzle, we recorded a frequency count of how often the model produced each combination of (C, Y, T), and a probability distribution $p(C, Y, T)$ by normalizing these frequency counts. We then applied the PID to decompose the mutual information $I(C, Y; T)$. Table 7 summarizes the PID components obtained from our simple dataset.

Observe Table 7, these values suggest that there is moderate *redundant* information R (0.549 bits) shared by the chain-of-thought correctness (C) and final answer correctness (Y) about whether the solution is both correct and justified (T). However, the *synergy* component $S = 0.428$ is also substantial, implying that one must consider both C and Y together to gain additional insight into whether the solution is truly valid and justified.

Table 7. PID results for logical reasoning with CoT on a synthetic dataset.

Quantity	Value
$I(C;T)$	0.774
$I(Y;T)$	0.674
$I(C,Y;T)$	0.977
$R(C,Y;T)$	0.549
$S(C,Y;T)$	0.428
$U(C;T \mid Y)$	0.226
$U(Y;T \mid C)$	0.125

In other words, while either the chain-of-thought or the final answer alone conveys meaningful information about T, looking at both variables jointly reveals extra information (0.428 bits) that neither one alone can fully capture. This aligns with the intuition that logical reasoning tasks often require both a correct final answer and a coherent explanatory chain-of-thought to ensure that a solution is truly well-justified.

Our experiment highlights that synergistic information can be substantial in logical reasoning: a coherent chain-of-thought and a correct final answer jointly determine whether a solution is fully justified.

D. Limitations

While our study provides valuable theoretical insights into CoT reasoning, several limitations should be acknowledged. First, our analyses primarily focused on theoretical and controlled experimental settings with limited dataset sizes and simplified tasks. Real-world tasks often involve more complex reasoning patterns and noisy data, potentially affecting the generalizability of our results. Second, our PID experiments relied on discretization techniques to manage textual data, which may result in information loss or overly coarse approximations of the underlying information-theoretic relationships. Moreover, the computational complexity of mutual information and PID calculations grows rapidly with larger datasets and richer vocabularies, potentially limiting scalability. Finally, our analyses assume accurate and high-quality rationales or reasoning steps. In practice, LLM-generated rationales may sometimes be plausible yet incorrect or incomplete, necessitating further research on robustness and reliability of CoT outputs.

E. Ethics Statement

This research involves the theoretical analysis of LLMs, specifically focusing on interpretability and information processing through CoT reasoning. Although our work does not involve direct collection or usage of sensitive personal data, ethical considerations regarding LLMs still apply. LLMs may inadvertently generate biased, misleading, or harmful outputs depending on training data and usage context. Our methodologies aim to improve interpretability, potentially helping identify and mitigate harmful biases or inaccuracies. However, enhancing interpretability does not inherently solve issues of fairness, accountability, and transparency. Researchers and practitioners must remain aware that interpretability methods alone cannot fully address ethical risks associated with model deployment. We encourage ongoing attention to responsible AI practices and transparency in methodologies.

During the writing of this paper, we used ChatGPT 4o with editing (e.g., grammar, spelling, word choice) and facilitating the experiments.