INTERPRETING CHAIN-OF-THOUGHT REASONING VIA PARTIAL INFORMATION DECOMPOSITION

Barproda Halder
Department of Electrical and Computer Engineering
University of Maryland, College Park
bhalder@umd.edu

Qiuyi Zhang
Elorian AI
richard@elorian.ai

Sanghamitra Dutta
Department of Electrical and Computer Engineering
University of Maryland, College Park
sanghamd@umd.edu

ABSTRACT

Large reasoning models have generated interest in complex tasks. However, they often generate verbose, repetitive, or incorrect reasoning steps on challenging problems. In this work, we introduce a new interpretability framework SLIDER for evaluating the quality of the reasoning process, assessing consecutive steps in terms of incorrectness and repetitiveness. SLIDER leverages an emerging body of work from information theory called Partial Information Decomposition (PID) to disentangle the information about the target between two consecutive reasoning steps into non-negative components: unique information in a reasoning step S_i or S_{i+1} that is not in the other, redundant information that is common between both steps, and synergistic information which is only meaningful when the steps are considered jointly. Given the responses of a large reasoning model, SLIDER moves across the steps in a sliding-window, projects them onto a meaningful embedding space, and then computes a set of new per-token information-decomposition measures that enables the identification of various failure modes. We demonstrate application of SLIDER to analyze incorrectness and repetitiveness for several use-cases across arithmetic problems and GSM8K word problems.

1 INTRODUCTION

The emergence of large reasoning models (LRMs) such as OpenAI's o1 (OpenAI.), Deepseek's R1 (Guo et al., 2025) and QwQ-32B (Team, 2025) have marked a significant leap in complex problem solving. However, LRMs often generate long, verbose, and repetitive chain-of-thought (CoT) reasoning (Wei et al., 2022), which increases computational cost. When the reasoning steps become very long, it is also harder for humans to follow and judge whether those steps are correct or meaningful. *Thus, there is a pressing need for automated methods that can evaluate the quality of CoT reasoning without human supervision.*

To address this issue, this work proposes a novel unified interpretability framework for evaluating both correctness and repetitiveness of CoT reasoning that we call SLIDER (**SL**iding window **I**nformation **D**ecomposition **E**xplainer for **R**easoning). SLIDER is theoretically-grounded in an information-theoretic technique called Partial Information Decomposition (PID) that disentangles the structure of multivariate information.

Recent works (Ton et al., 2024; Yong et al., 2025) have analyzed CoT reasoning using classical information-theoretic quantities such as mutual information and conditional mutual information. Yet, the structural interactions between various steps, specifically, how consecutive steps jointly and individually contribute to the final outcome, have received limited attention. Previous works aggregate all prior steps together and attempt to quantify whether an additional step is informative or not (Related Works in Appendix B). Such an aggregation prevents a principled isolation of individual step contributions, limiting step-level interpretability. Rethinking previous works, we move beyond

aggregate information gain and instead characterize the structure of task-relevant information across consecutive reasoning steps via a sliding window. Leveraging PID, our framework provides a strictly more expressive characterization than classical information-theoretic metrics by decomposing the information about the answer between consecutive reasoning steps into unique, redundant, and synergistic components. This shift in perspective enables us not only to determine whether a step is informative, but to understand how it contributes: whether it introduces genuinely new information, merely repeats prior content, or complements other steps to produce the final answer.

Contribution Summary: (i) We propose a principled interpretability framework called SLIDER to analyze when and how individual chain-of-thought (CoT) steps contribute towards the final answer. SLIDER is rooted in PID, which decomposes the information about the final answer between two consecutive reasoning steps (S_i, S_{i+1}) into unique, redundant, and synergistic components. (ii) Given the responses of a LRM, we move across the steps in a sliding-window fashion and project them onto a meaningful embedding space. Next, we introduce a new set of per-token information-decomposition measures (Definition 2) that enables the identification of various failure modes in CoT reasoning (see Fig. 1). (iii) Our experiments demonstrate how SLIDER can be used to assess the quality of reasoning steps for different use-cases across arithmetic datasets and GSM8K word problems (Cobbe et al., 2021), identifying incorrectness and repetitiveness.

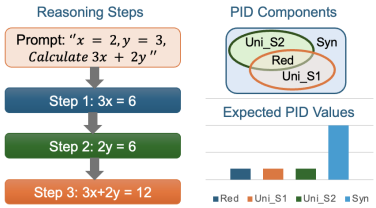

Figure 1: SLIDER in a simple arithmetic problem: Demonstrates high synergistic information between Step 1 (3x) and Step 2 (2y) because they jointly provide complete information about the final answer (3x+2y).

Background on PID: The classical measure of the total information about a target variable A that is contained in two random variables X and Y is given by mutual information $\mathrm{I}(A; X, Y)$ (Cover & Thomas, 2012). However, mutual information $\mathrm{I}(A; X, Y)$ does not disentangle what is uniquely contributed by each or shared by both. An emerging body of work called Partial Information Decomposition (PID) goes beyond classical measures, and decomposes the joint information about a target A shared among multiple random variables X and Y into four non-negative measures (also see Fig. 2):

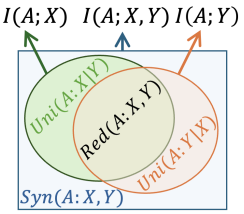

Figure 2: PID of $\mathrm{I}(A; X, Y)$

$$\mathrm{I}(A; X, Y) = \mathrm{Uni}(A{:}Y|X) + \mathrm{Uni}(A{:}X|Y) + \mathrm{Red}(A{:}X, Y) + \mathrm{Syn}(A{:}X, Y). \quad (1)$$

Here, unique information $\mathrm{Uni}(A{:}X|Y)$ and $\mathrm{Uni}(A{:}Y|X)$ captures the information that is exclusively provided by X or Y. Redundant information $\mathrm{Red}(A{:}X, Y)$ is the shared information about A in X and Y, and synergistic information $\mathrm{Syn}(A{:}X, Y)$ emerges only when X and Y are both present together (complementary). We now formally define these quantities.

Definition 1 (Unique information (Bertschinger et al., 2014))**.** *Let Δ be the set of all joint distributions on (A, X, Y) and $\Delta_P = \{Q_{AXY} \in \Delta\colon Q_{AX} = P_{AX}$ and $Q_{AY} = P_{AY}\}$ be the set of joint distributions with same marginals on (A, X) and (A, Y) as the true distribution P_{AXY}. Then, $\mathrm{Uni}(A{:}X|Y) := \min_{Q \in \Delta_P} \mathrm{I}_Q(A; X|Y)$. Here, $\mathrm{I}_Q(A; X|Y)$ is the conditional mutual information under joint distribution Q_{AXY} instead of P_{AXY}.*

Defining any one of the PID components is sufficient to derive the others, due to the following relationships: $\mathrm{I}(A; Y) = \mathrm{Uni}(A{:}X|Y) + \mathrm{Red}(A{:}X, Y)$. Intuitively, $\mathrm{Red}(A{:}X, Y)$ can be interpreted as the overlapping portion between $\mathrm{I}(A; Y)$ and $\mathrm{I}(A; X)$ (see Fig. 2). Hence, $\mathrm{Red}(A{:}X, Y) = \mathrm{I}(A; Y) - \mathrm{Uni}(A{:}X|Y)$. Finally, the synergy corresponds to the remaining information: $\mathrm{Syn}(A{:}X, Y) = \mathrm{I}(A; X, Y) - \mathrm{Uni}(A{:}X|Y) - \mathrm{Uni}(A{:}Y|X) - \mathrm{Red}(A{:}X, Y)$, which can be computed once the unique and redundant information terms have been obtained.

Toy Example: Let $Z{=}(Z_1, Z_2, Z_3)$ with each $Z_i \sim$ i.i.d. Bern(1/2). Let $X = (Z_1, Z_2, Z_3 \oplus N)$, $Y = (Z_2, N)$, and $N \sim$ Bern(1/2) which is independent of Z. Here, $\mathrm{I}(Z; X, Y) = 3$ bits. The unique information about Z only in X and not in Y is effectively in Z_1. Thus, $\mathrm{Uni}(Z{:}X|Y) = \mathrm{I}(Z; Z_1) = 1$ bit. Redundant information about Z that is in both X and Y is effectively in Z_2 and is given by $\mathrm{Red}(Z{:}X, Y) = \mathrm{I}(Z; Z_2) = 1$ bit. Synergistic information about Z that is not in either X or Y alone, but is in both of them together is effectively in the tuple $(Z_3 \oplus N, N)$, and is given by $\mathrm{Syn}(Z{:}X, Y){=}\mathrm{I}(Z; (Z_3 \oplus N, N)) = 1$ bit. This accounts for the 3 bits in $\mathrm{I}(Z; X, Y)$.

2 PROPOSED INTERPRETABILITY FRAMEWORK: SLIDER

Notations: Given an input prompt S_0, we denote the LRM-generated reasoning trajectory by $S = \{S_1, S_2, \ldots, S_T\}$, where each S_i represents an intermediate reasoning step. We segment the reasoning path whenever a double newline token ("\n\n") appears in the model's output. Each segment is considered a distinct reasoning step. The total number of tokens in a step S_i is denoted by $|S_i|$. The final answer is denoted by A.

Proposition 1. *The total information of two consecutive reasoning steps S_i and S_{i+1} about the ground-truth answer A can be decomposed into four non-negative components:*

$$\mathrm{I}(A; S_i, S_{i+1}) = \mathrm{Uni}(A{:}S_i|S_{i+1}) + \mathrm{Uni}(A{:}S_{i+1}|S_i) + \mathrm{Red}(A{:}S_i, S_{i+1}) + \mathrm{Syn}(A{:}S_i, S_{i+1}).$$

Our proposed per-token measures. Each reasoning step comprises a sequence of tokens whose length can vary considerably across examples. As a result, two-step pairs, (S_1, S_2) and (S_1', S_2') may have similar information decomposition with respect to the target A, yet differ significantly in their token lengths. To further account for efficiency, we propose a set of unified per-token measures that capture not only whether synergy/ redundancy/uniqueness exists, but also how efficiently task-relevant information is encoded across the entire step.

Definition 2 (Proposed Per-Token Measures)**.** *Let S_i and S_{i+1} denote two consecutive reasoning steps from a given data distribution, and let A be the final answer. We define the normalized per-token measures for redundancy, uniqueness, and synergy as: (i)* ***Per-token redundancy*** *M_R. The shared task-relevant information between S_i and S_{i+1} is: $M_R = \frac{\mathrm{Red}(A:S_i, S_{i+1})}{|S_i|+|S_{i+1}|}$. (ii)* ***Per-token uniqueness*** *M_U. The information uniquely contributed by each step are $M_U(i) = \frac{\mathrm{Uni}(A:S_i|S_{i+1})}{|S_i|}$ and $M_U(i+1) = \frac{\mathrm{Uni}(A:S_{i+1}|S_i)}{|S_{i+1}|}$. (iii)* ***Per-token synergy*** *M_S. The complementary information that arises only from the joint consideration of both steps is defined as $M_S = \frac{\mathrm{Syn}(A:S_i, S_{i+1})}{|S_i|+|S_{i+1}|}$.*

SLIDER consists of several key steps: We generate multiple samples for each problem instance to reliably estimate the joint distributions required for computing the PID-based measures. Next, we use an encoder-only LLM (all-MiniLM-L6-v2 (Reimers & Gurevych, 2019)) to obtain embeddings of consecutive steps S_i, S_{i+1}, and ground-truth answer A, respectively. We then discretize these continuous embeddings via k-means clustering to obtain categorical variables. Using the resulting cluster assignments, we estimate the joint distribution, $P(S_i = c_i, S_{i+1} = c_{i+1}, A = a)$. Finally, based on Definition 2, we compute the per-token PID measures from the estimated joint distribution.

3 EXPERIMENTAL RESULTS

Identifying Incorrectness in CoT. We consider the following problem: *Arithmetic problem:* "x = {x}, y = {y}. Please calculate the following: 1. $3x$, 2. $2y$, 3. $3x + 2y$." We sample two integers x and y independently and uniformly from $[1, 10^5)$. Using ChatGPT, we generate 500 samples in the exact same format, each containing exactly three reasoning steps. The intermediate reasoning steps are $S_1 = 3x$, $S_2 = 2y$, and $S_3 = 3x + 2y$, and the final answer $A = 3x + 2y$ (see Fig. 3).

Synergistic information arises when consecutive steps together provide complementary information about the target that is not present in either step individually. For the arithmetic problem, individually, neither $S_1 = 3x$ nor $S_2 = 2y$ is sufficient to determine the target A. However, when S_1 and S_2 are considered jointly, their combined information fully specifies A (considering no error in the reasoning steps). This implies that the information about A is predominantly *synergistic/complementary* across S_1 and S_2. Now, suppose that a step S_1 or S_2 contains some errors. Intuitively, the synergistic information between the steps about the final answer should decrease, since the error limits the inference of the final answer, even when considering both steps together. In this scenario, the unique information contributed by the correct step is expected to increase, while that of the erroneous step should decrease. This is because the correct step now carries more task-relevant information toward the final answer. Unique information in one reasoning step captures information that can not be obtained from the other.

Consider the same problem with the last two reasoning steps, $S_2 = 2y$ and $S_3 = 3x + 2y$, and the final answer $A = 3x + 2y$. If S_3 is computed correctly, it contains all the information needed to

Figure 3: Interpretable trends of per-token PID measures under errors: (a,b) For this problem, per-token synergy M_S is dominant when the steps are correct. M_S decreases as the error probability p increases, indicating that synergistic/complementary information in consecutive steps diminishes under erroneous computations. Under errors, per-token uniqueness M_U corresponding to a correct step increases with p, reflecting that correct steps contribute increasingly distinct information relative to their erroneous counterparts. Conversely, M_U associated with the erroneous step decreases as error probability p increases. (c) When error is in S_3, the per-token uniqueness in the erroneous step $M_U(3)$ decreases with higher error, demonstrating that errors in intermediate steps degrade the unique information about the target. (d) Reference example for the reasoning steps.

determine A, and therefore provides more task-relevant information than S_2. Intuitively, S_3 should then contribute more unique information about the target A than S_2. However, if there is an error in the calculation in S_3, its uniqueness decreases, since some examples can no longer provide useful information about A that is not already available from S_2. These interpretations are consistent with the trends that we observe for this problem, as shown in Fig. 3. Thus, SLIDER can help identify incorrectness in reasoning. We can also use SLIDER to identify incorrectness for the GSM8K word problem (Cobbe et al., 2021) (see details in Appendix C.2).

Identifying Repetitiveness in CoT: We use two word problems from GSM8K (Cobbe et al., 2021) to examine how increasing the number of reasoning steps affects model reasoning. *GSM8K word problems:* (1) Easy: "Anthony had {x} pencils. He gave $\frac{1}{2}$ of his pencils to Brandon, and then gave $\frac{3}{5}$ of the remaining pencils to Charlie. He kept the rest. How many pencils did Anthony keep?" (2) Hard: "Jen decides to travel to 3 different countries. He has to pay $x for the supplies he needs, in total. The tickets for travel cost, in total, 50% more than the supplies. How much does travel cost?"

Redundant information arises when step S_i and S_{i+1} share information about the target answer A. Consider the first word problem, which asks how many pencils remain. The corresponding reasoning steps are illustrated in Fig. 4(c). For a simplified interpretation, assume the first two reasoning steps are: $S_1 = \frac{x}{2}$ and $S_2 = \frac{x}{2} \times \frac{3}{5}$ and the final answer $A = x - \frac{x}{2} - \frac{x}{2} \times \frac{3}{5} = \frac{x}{5}$. In this example, both S_1 and S_2 involve the term $\frac{x}{2}$, meaning they contain overlapping information about the target answer, meaning that part of their contribution is redundant. This redundancy reflects that both steps partially convey the same component of the calculation, rather than providing entirely independent contributions toward determining the final answer.

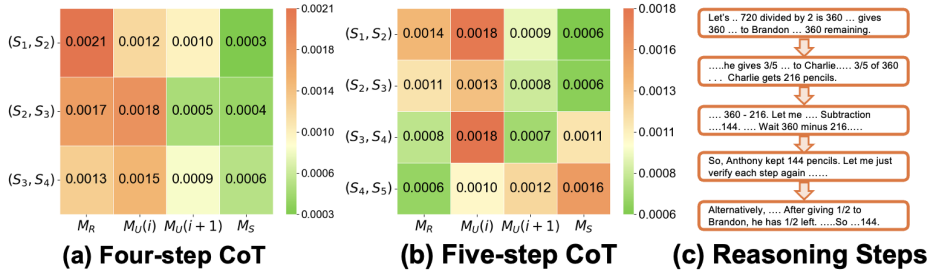

Figure 4: Interpretable trends of per-token PID measures under repetitiveness: (a,b) For this problem, per-token redundancy M_R keeps decreasing while per-token synergy M_S keeps increasing for both four-step and five-step reasoning trajectories. (c) Reference example for the reasoning steps.

For (S_1, S_2) in Fig. 4(a), we observe a high per-token redundancy M_R and a moderate per-token uniqueness M_S in S_1 and S_2. This is because beyond contributing unique task-relevant informa-

tion, the early steps also share overlapping content. The first step computes half of the pencils and determines the remaining quantity, while the second step uses this remaining amount to compute $\frac{3}{5}$ of it. From (S_1, S_2) to (S_3, S_4), we observe that M_R decreases while the per token synergy M_S increases. In Fig. 4(b) per-token synergy becomes prominent in (S_4, S_5). This is because the final step often reconstructs the solution again using an alternative formulation and increases confidence in the calculation: contributes additional complementary task-relevant information. Observing these patterns allows us to identify instances of repetitive reasoning, where steps reiterate previously derived information rather than introducing new insights. Additional results on second word problem are in Appendix C.2 along with discussion and future work, sensitivity analysis, and comparisons to standard PID measures in Appendix A, D, and E, respectively.

REFERENCES

Riccardo Ali, Francesco Caso, Christopher Irwin, and Pietro Liò. Entropy-lens: The information signature of transformer computations. *arXiv preprint arXiv:2502.16570*, 2025.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

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

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

Shaurya Dewan, Rushikesh Zawar, Prakanshul Saxena, Yingshan Chang, Andrew Luo, and Yonatan Bisk. Diffusion pid: Interpreting diffusion via partial information decomposition. *Advances in Neural Information Processing Systems*, 37:2045–2079, 2024.

Pasan Dissanayake, Faisal Hamman, Barproda Halder, Ilia Sucholutsky, Qiuyi Zhang, and Sanghamitra Dutta. Quantifying knowledge distillation using partial information decomposition. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. An information-theoretic quantification of discrimination with exempt features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3825–3833, 2020.

David A Ehrlich, Andreas C Schneider, Michael Wibral, Viola Priesemann, and Abdullah Makkeh. Partial information decomposition reveals the structure of neural representations. *arXiv preprint arXiv:2209.10438*, 2022.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Zeyu Gan, Yun Liao, and Yong Liu. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning. *arXiv preprint arXiv:2501.15602*, 2025.

Chaitanya Goswami, Amanda Merkley, and Pulkit Grover. Computing unique information for poisson and multinomial systems. *arXiv preprint arXiv:2305.07013*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Barproda Halder, Faisal Hamman, Pasan Dissanayake, Qiuyi Zhang, Ilia Sucholutsky, and Sanghamitra Dutta. Towards formalizing spuriousness of biased datasets using partial information decomposition. *Transactions on Machine Learning Research*, 2025.

Faisal Hamman and Sanghamitra Dutta. Demystifying local and global fairness trade-offs in federated learning using partial information decomposition. *arXiv preprint arXiv:2307.11333*, 2023.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. *arXiv preprint arXiv:2306.04539*, 2023.

Aobo Lyu, Andrew Clark, and Netanel Raviv. Explicit formula for partial information decomposition. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 2329–2334, 2024. doi: 10.1109/ISIT57864.2024.10619369.

Salman Mohamadi, Gianfranco Doretto, and Donald A Adjeroh. More synergy, less redundancy: Exploiting joint mutual information for self-supervised learning. *arXiv preprint arXiv:2307.00651*, 2023.

OpenAI. Learning to reason with LLMs. URL `https://openai.com/index/learning-to-reason-with-llms/`.

Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables. *Advances in neural information processing systems*, 34:20295–20307, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Tycho Tax, Pedro Mediano, and Murray Shanahan. The partial information decomposition of generative neural network models. *Entropy*, 19(9):474, 2017.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, 2025.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory. *arXiv preprint arXiv:2411.11984*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Praveen Venkatesh, Corbett Bennett, Sam Gale, Tamina Ramirez, Greggory Heller, Severine Durand, Shawn Olsen, and Stefan Mihalas. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *J. Mach. Learn. Res.*, 24:131–1, 2023.

Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. Think or not? exploring thinking efficiency in large reasoning models via an information-theoretic lens. *arXiv preprint arXiv:2505.18237*, 2025.

A Discussion and Future Work

This work introduces SLIDER a novel interpretability framework for evaluating the quality of reasoning processes using Partial Information Decomposition (PID). Empirical analysis on arithmetic and GSM8K word problems demonstrates that SLIDER effectively identifies failure modes across consecutive reasoning steps. However, the PID estimation in this work depends on dimensionality reduction and discretization, which can lead to information loss; future work will explore continuous PID estimation methods to address this limitation. Additionally, evaluating the quality of reasoning steps with per-token PID measures, whether for correctness or efficiency, requires prior knowledge of the problem structure or an approximate mapping to ground-truth reasoning steps. Developing approaches that reduce these dependencies will enhance the applicability and robustness of SLIDER.

B Related Works

Classical information theoretic measures offer a lens to theoretically analyze and evaluate different aspects of LLMs (Farquhar et al., 2024; Kossen et al., 2024; Ali et al., 2025; Ton et al., 2024; Gan et al., 2025; Yong et al., 2025). Recent works Gan et al. (2025); Ton et al. (2024); Gan et al. (2025); Yong et al. (2025) use information-theoretic measures to evaluate LLM reasoning. Gan et al. (2025) explore slow-thinking in the multi-step reasoning of large language models (LLMs) using information theory. Ton et al. (2024) formalize chain-of-thought reasoning using mutual information. Specifically, they quantify the information gain at each reasoning step using conditional mutual information, enabling the identification of failure modes in CoT reasoning. However, in practice, their method concatenates all prior reasoning steps $S_{\leq j}$ with S_{j+1} and evaluates the difference in next-token likelihood using a language model. Another previous work Yong et al. (2025) also introduce two metrics, namely "InfoGain" and "InfoBias", to quantify stepwise information gain and divergence from the original reasoning path, respectively, utilizing classical information-theoretic measures. Even though their framework accounts for step-level token length, the aggregation of previous reasoning steps fundamentally limits interpretability, as individual step contributions remain entangled with the broader reasoning context.

Partial Information Decomposition (PID) (Williams & Beer, 2010; Venkatesh et al., 2024; Goswami et al., 2023; Pakman et al., 2021; Lyu et al., 2024) is an emerging area of research of information theory, with growing applications across diverse domains, including representation analysis, fairness, and multimodal learning, (Tax et al., 2017; Dutta et al., 2020; Hamman & Dutta, 2023; Ehrlich et al., 2022; Liang et al., 2023; Wollstadt et al., 2023; Mohamadi et al., 2023; Dewan et al., 2024; Dissanayake et al., 2024; Halder et al., 2025). For instance, Liang et al. (2023) employ PID to analyze multimodal interactions, while Dewan et al. (2024) leverage PID to interpret diffusion models.

Despite these advances, the application of PID to reasoning processes, particularly to analyze the informational structure of intermediate reasoning steps, remains largely unexplored. In this work, we extend PID to the study of structured reasoning, introducing a principled framework to quantify the quality and efficiency of chain-of-thought reasoning through the decomposition of information across consecutive reasoning steps.

C Appendix to Experiments

C.1 Setup

We sample x multiple times to create our data distributions required for computing the PID-based measures. For both word problems, we prompt the QwQ-32B reasoning model and collect 5000 responses. For the easier GSM8K word problem, x is sampled randomly from multiples of 10 in the range $[10, 10^{10})$, whereas for the harder problem, x is sampled uniformly from $[10, 10^3)$. Then, we segment each model's response (generated before the `</think>` token) at every occurrence of a double newline token ("`\n\n`") to obtain individual reasoning steps.

Initially, the reasoning steps(S_i) and the final answer(A) are lists of strings, e.g.,S_1 = ['3x = 3', '3x = 6',...], S_2 = ['2y = 2, '2y = 4',...], S_3 =['3x+2y = 5', '3x+2y = 10',...], and A = ['5', '10',...]. For the subsequent calculations in SLIDER , we use an encoder-only LLM (all-MiniLM-L6-v2) to obtain embedding representation of consecutive steps S_i, S_{i+1}, and ground-truth answer A,

respectively. We then discretize these continuous embeddings via k-means clustering to obtain categorical variables. Using the resulting cluster assignments, we estimate the joint distribution, $P(S_i = c_i, S_{i+1} = c_{i+1}, A = a)$. Finally, based on Definition 2, we compute our proposed per-token PID measures from the estimated joint distribution.

C.2 Additional Word Problems

We now consider a challenging problem from the GSM8K dataset to evaluate the effectiveness of our proposed framework, SLIDER , in identifying incorrect and repetitive reasoning: "Jen decides to travel to 3 different countries. He has to pay $x for the supplies he needs, in total. The tickets for travel cost, in total, 50% more than the supplies. How much does travel cost?"

To analyze both errors and repetitive patterns in the reasoning trajectories, we segment each model response into two parts: the text generated before the `</think>` token, which we refer to as *thinking_steps*, and the text generated after the `</think>` token, which we denote as the *final_summary*. The final_summary typically summarizes the logical flow developed in the thinking_steps. We segment each of them independently at every occurrence of a double newline token ("`\n\n`") to obtain individual reasoning steps. We calculate the per-token PID measures on the final_summary to easily identify errors and the thinking_steps to see the effect of larger reasoning steps.

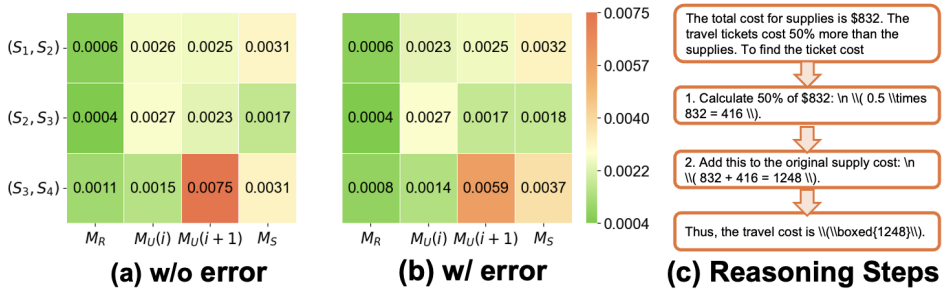

Figure 5: (a,b) Comparison of per-token PID measures for erroneous (w/) and non-erroneous(w/o): per-token uniqueness $M_U(i{+}1)$ decreases for the step pairs (S_2, S_3) and (S_3, S_4) when error occurs. (c) Reference example for the reasoning steps in final_summary (erroneous case).

Identifying Incorrectness in CoT: We can summarize the reasoning flow for the word problem as follows: S_1: compute 50% of the supply cost, x; S_2: compute $0.5x + x$ to obtain the ticket cost; S_3: compute $1.5x + x$ to determine the total travel cost. However, as shown in Fig. 5(c), the final step fails to correctly combine the ticket and supply costs. The model misunderstands the objective of the question, computing only the ticket cost rather than the overall travel cost. Notably, Fig. 5(a,b) shows that when this error occurs, the per-token uniqueness $M_U(i + 1)$ decreases for the step pairs (S_2, S_3) and (S_3, S_4), indicating the presence of erroneous reasoning step.

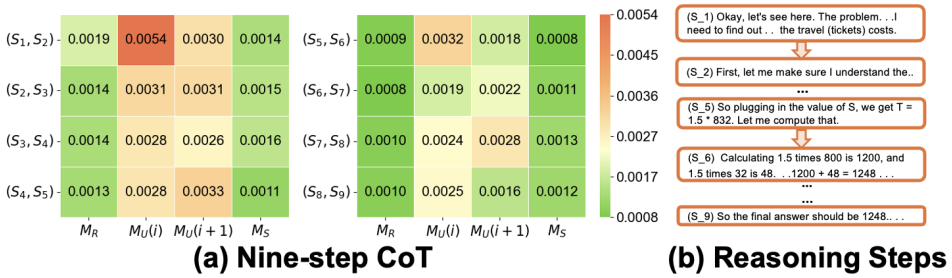

Figure 6: (a) Change in Per-token PID measures over number of reasoning steps (b) Reference example for the reasoning steps in thinking_steps (erroneous case).

Identifying Repetitiveness in CoT: For this challenging problem, the *thinking_steps* are lengthy and predominantly repetitive in nature. As illustrated in Fig. 6(b), the first two steps focus on interpret-

ing and structuring the problem. In contrast, subsequent steps first determine what to compute and then carry out the calculations. This behavior is reflected in the per-token uniqueness M_U shown in Fig. 6(a). For example, at S_5 the model outlines the plan to compute $1.5 \times x$, while at S_6 it performs the actual calculation using additional tokens. Although both steps contain unique information, the per-token normalization highlights the efficiency difference between them (see Fig. 6(a) (S_5, S_6)). The following step again revisits or reformulates the same computation. Such patterns lead to alternating dominance in per-token uniqueness M_U and indicate repetitive reasoning rather than a steady introduction of new task-relevant information.

D SENSITIVITY ANALYSIS

Encoder-Only vs Decoder-Only Representations: One key step of SLIDER is to convert the textual reasoning steps and final answer into vector representations. To this end, we experimented with input embeddings from both decoder-only (Llama-2-7b (Touvron et al., 2023) and encoder-only (all-MiniLM-L6-v2(Reimers & Gurevych, 2019)) LLMs. Interestingly, embeddings from the decoder-only LLM produced counterintuitive PID values (see Table 1). For example, with increasing errors in the reasoning step S_2 of the arithmetic data where $S_1 = 3x$, $S_2 = 2y$, and the final answer $A = 3x + 2y$, the observed synergy increases, whereas intuitively it should decrease, as discussed earlier. A possible explanation is that encoder-only LLMs are trained to produce embeddings that are particularly suitable for classification tasks, making them more cluster-efficient and better aligned for PID estimation.

Table 1: PID components using decoder-only and encoder-only LLMs as embedders.

	Decoder-only				Encoder-only			
Erroneous S_2 w/ prob. p	Red	Uni_S1	Uni_S2	Syn	Red	Uni_S1	Uni_S2	Syn
$p = 0$	0.0194	0.0544	0.0748	0.7100	0.0978	0.1557	0.0817	0.7500
$p = 0.25$	0.0253	0.0485	0.0561	0.7467	0.0692	0.1844	0.0620	0.6872
$p = 0.5$	0.0260	0.0478	0.0649	0.7519	0.0591	0.1945	0.0453	0.6928

Number of clusters: To assess the sensitivity of PID estimation to the number of clusters, we compute PID values using 10 and 20 clusters (see Table 2). We observe that increasing the number of clusters to 20 leads to higher PID values, although the rate of increase varies across the components. Importantly, overall trends remain consistent: redundancy, uniqueness, and synergy patterns are largely preserved, and components that dominate with 10 clusters also dominate with 20 clusters.

Table 2: PID components between consecutive reasoning steps for different numbers of clusters.

(S_i, S_{i+1})	n_{cluster}	$\text{Red}(A{:}S_i, S_{i+1})$	$\text{Uni}(A{:}S_i \mid S_{i+1})$	$\text{Uni}(A{:}S_{i+1} \mid S_i)$	$\text{Syn}(A{:}S_i, S_{i+1})$
(S_1, S_2)	10	0.5496	0.2709	0.0993	0.2045
	20	0.8462	0.4914	0.3134	0.3449
(S_2, S_3)	10	0.4205	0.2298	0.0890	0.1805
	20	0.7081	0.4522	0.2335	0.3779
(S_3, S_4)	10	0.1746	0.3351	0.0501	0.2385
	20	0.4230	0.5193	0.1803	0.6169
(S_4, S_5)	10	0.1328	0.0919	0.1427	0.3408
	20	0.3326	0.2689	0.3088	0.8131

E STANDARD PID-MEASURES VS OUR MEASURES

Fig. 7 demonstrates the effectiveness of our proposed per-token measures. When computing standard PID measures, redundant (Red) information dominates significantly over the other components; however, based on our earlier analysis of the easy word problem, we also expect substantial unique information. In contrast, the proposed per-token formulation captures the prominence of per-token redundancy M_R without overshadowing the contributions of per-token uniqueness, resulting in a more efficient quantification across reasoning steps.

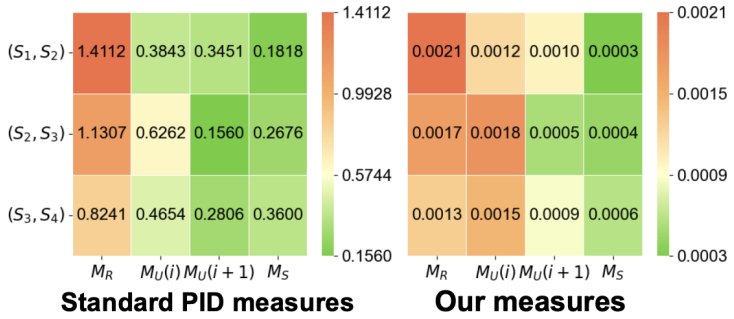

Figure 7: Comparison of PID and per-token measures on the easy GSM8K problem.