Explainable Chain-of-Thought Reasoning: An Empirical Analysis on State-Aware Reasoning Dynamics

Sheldon Yu[1*] **Yuxin Xiong**[1*] **Junda Wu**[1] **Xintong Li**[1] **Tong Yu**[2]
Xiang Chen[2] **Ritwik Sinha**[2] **Jingbo Shang**[1] **Julian McAuley**[1]
[1]University of California San Diego [2]Adobe Research
{ziy040, y7xiong, juw069, xil240, jshang, jmcauley}@ucsd.edu
{tyu, xiangche, risinha}@adobe.com

Abstract

Recent advances in chain-of-thought (CoT) prompting have enabled large language models (LLMs) to perform multi-step reasoning. However, the explainability of such reasoning remains limited, with prior work primarily focusing on local token-level attribution, such that the high-level semantic roles of reasoning steps and their transitions remain underexplored. In this paper, we introduce a state-aware transition framework that abstracts CoT trajectories into structured latent dynamics. Specifically, to capture the evolving semantics of CoT reasoning, each reasoning step is represented via spectral analysis of token-level embeddings and clustered into semantically coherent latent states. To characterize the global structure of reasoning, we model their progression as a Markov chain, yielding a structured and interpretable view of the reasoning process. This abstraction supports a range of analyses, including semantic role identification, temporal pattern visualization, and consistency evaluation.

1 Introduction

Chain-of-thought (CoT) prompting has become a central technique for eliciting multi-step reasoning in large language models (LLMs) (Wei et al., 2022). By encouraging models to decompose problems into intermediate steps, CoT improves performance on tasks such as arithmetic, logical deduction, and multi-hop question answering. Understanding CoT outputs is, therefore, critical for both evaluation and user comprehension. Recent large-scale studies have shown that even modest increases in textual complexity can hinder human comprehension and increase cognitive load, highlighting the need for more interpretable LLM outputs (Guidroz et al., 2025; Xiao and Yang, 2025; Wasi and Islam, 2024; Wu et al., 2025, 2024a,b). Understanding CoT is even more challenging in this context, given its length, multi-step structure, and abstract semantic progression. However, the explainability of CoT remains limited. Prior work often focuses on token-level attribution(Hou et al., 2023) or heuristic correctness measures(Gan et al., 2025). Focusing on local token-level attribution leaves the high-level semantic roles of reasoning steps and their transitions underexplored. Understanding CoT reasoning benefits from a shift from local analysis to a structured, global perspective.

To address this, we propose a **state-aware transition framework** for CoT reasoning. We segment generated CoTs into discrete steps, embed each step via spectral analysis of token-level representations, and cluster them into semantically coherent latent states. Transitions between these states are modeled as a first-order Markov chain, yielding a structured view of reasoning dynamics. This approach goes beyond surface-level generation by uncovering latent structure in CoT trajectories, offering a non-trivial solution to the challenge of interpreting abstract, multi-step reasoning without access to ground-truth annotations or explicit reasoning labels. Designed as an explainability scaffold, this framework supports diverse analyses, including semantic role identification, temporal transition pattern discovery, and trajectory-level consistency evaluation. We summarize our contributions as follows:

- We propose a framework that abstracts CoT reasoning as structured transitions over latent semantic states, modeled via a Markov Chain.
- We further visualize multi-step CoT reasoning for explainable understanding of reasoning dynamics beyond token-level attribution.
- We enable explainabilities for CoT, including reasoning step abstraction, semantics, and state-aware transition.

[*] Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16660–16667
November 4–9, Suzhou, China, 2025

- Empirical results across multiple tasks and models reveal consistent and recurring latent transition patterns, suggesting that LLM reasoning exhibits structured dynamics beyond surface-level token sequences.

2 Related Work

Chain-of-Thought (CoT) Reasoning in LLMs. Chain-of-Thought (CoT) prompting has emerged as an effective strategy to elicit multi-step reasoning in large language models (LLMs) (Wei et al., 2022). Subsequent work explored improved prompting techniques (Kojima et al., 2023), automatic CoT generation (Zhang et al., 2022), and evaluation of reasoning quality (Wang et al., 2023). While most prior work focuses on generating or enhancing CoT outputs, we aim to abstract and model the latent structure of CoTs for interpretability. Coconut (Hao et al., 2024) introduces a latent reasoning paradigm that operates in continuous space rather than natural language, enabling breadth-first exploration of reasoning paths and highlighting the limitations of token-level CoT representations. This further motivates our structural approach to modeling CoT dynamics.

Explainability in Language Models. Explainability methods for LLMs often focus on token-level attribution, such as attention rollout (Chuang et al., 2024; Wu et al., 2021) or integrated gradients (Zhao et al., 2024). In addition, in-context explanability (Liu et al., 2023; Wei et al., 2023; Liu et al., 2024; Wu et al., 2022) focuses on understanding how to best prompt LLMs with demonstrations. Recent work has also questioned the faithfulness of CoT explanations (Turpin et al., 2023; Atanasova et al., 2023). From the perspective of LLM's latent space, the gradient-based method (Wu et al., 2023) explains more fine-grained CoT behaviors. In contrast, we propose a structural abstraction framework that operates at the step level, capturing global reasoning dynamics rather than local token importance.

3 Explainable Modeling of Chain-of-Thought Reasoning

Step Embedding and Abstraction To capture the evolving semantics of each reasoning step, we segment the CoT output y (elicited from an instruction-tuned LLM given input x) into a sequence of steps $c = (c_1, \ldots, c_T)$. For each step c_t, we extract token embeddings $X_t \in \mathbb{R}^{n_t \times d}$ and compute the local Gram matrix $\tilde{G}_t = X_t^\top X_t$. We define the spectral embedding $E_t \in \mathbb{R}^{k_{\text{eig}}}$ as the vector of top-k_{eig} eigenvalues:

$$E_t = (\lambda_1, \ldots, \lambda_{k_{\text{eig}}}), \quad \lambda_1 \geq \cdots \geq \lambda_{k_{\text{eig}}}.$$

To incorporate contextual accumulation, we recursively update the accumulated Gram matrix across steps:

$$G_t = G_{t-1} + \tilde{G}_t, \quad G_1 = \tilde{G}_1.$$

This yields a trajectory of spectral embeddings $(E_1, \ldots, E_T)$, which we abstract using k-means with k_{clu} clusters to assign each step to a latent state $s_t \in \{1, \ldots, k_{\text{clu}}\}$. The resulting state sequence enables structural abstraction of CoT and supports distribution-level explainability.

Semantics of Reasoning State To assess whether latent clusters correspond to meaningful functional roles, we collect all reasoning steps assigned to each cluster and manually summarize their semantics. To further ground these semantics, we compute the average position of steps assigned to each cluster within the reasoning trajectory. Specifically, for cluster c, the average step index is given by:

$$\bar{t}_c = \frac{1}{|S_c|} \sum_{(i,t) \in S_c} t,$$

where S_c denotes the set of steps assigned to cluster c, and t is the position of the step in its trajectory.

Modeling Transitions via Markov Chains To capture the global structure of CoTs, we model transitions between clusters as a Markov chain. Given the step-wise state sequence $\mathbf{s} = (s_0, \ldots, s_T)$, we estimate a transition matrix $P \in \mathbb{R}^{k \times k}$, where each entry denotes:

$$P_{i,j} = \mathbb{P}(s_{t+1} = j \mid s_t = i) = \frac{C_{i,j}}{\sum_{j'} C_{i,j'}},$$

with $C_{i,j}$ as the number of observed transitions from state i to j. This latent transition structure reveals sequential patterns in CoT and serves as the basis for downstream diagnostics and simulation.

4 Analysis of Reasoning Dynamics

Datasets To comprehensively assess the structural dynamics of chain-of-thought (CoT) reasoning, we consider datasets from three distinct categories: **mathematical** GSM8k (Cobbe et al., 2021),

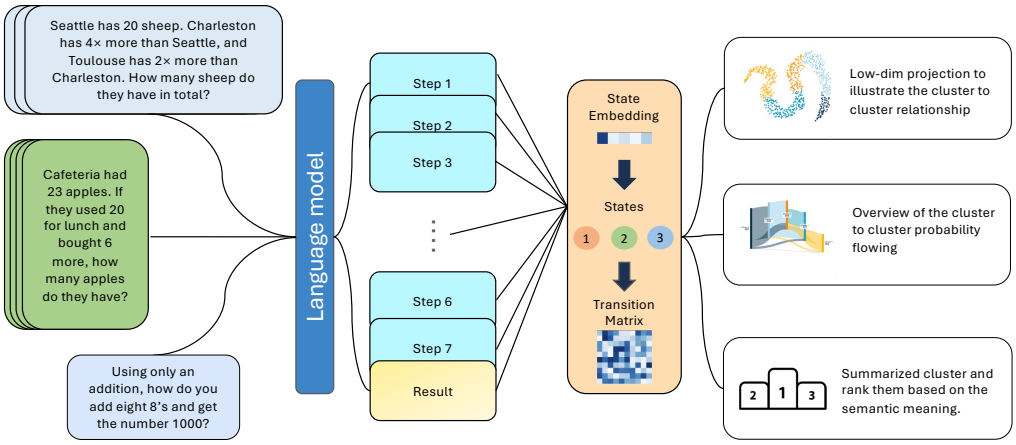


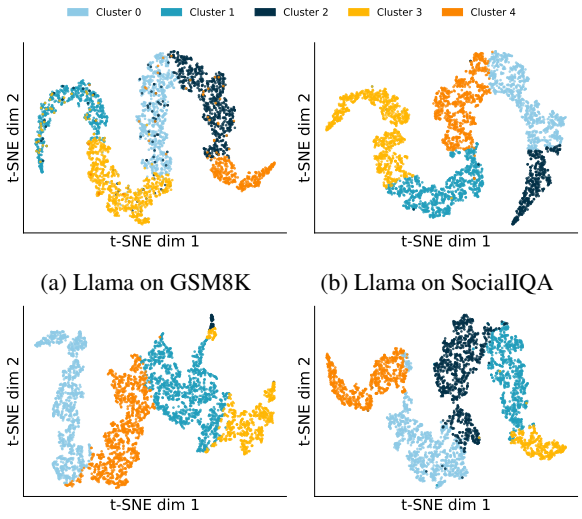
Figure 1: An overview of our CoT abstraction and simulation framework.

MATH (Hendrycks et al., 2021), focusing on symbolic and numerical reasoning; **knowledge-based** (HotpotQA (Yang et al., 2018), MusiQUe (Trivedi et al., 2022)), involving multi-hop factual inference over text; and **commonsense** (CSQA (Talmor et al., 2019), SocialIQa (Sap et al., 2019)), targeting intuitive reasoning about everyday and social scenarios. This diverse selection enables us to evaluate whether latent reasoning structures generalize across quantitative, textual, and intuitive domains.

Implementation Details We analyze CoT reasoning using three instruction-tuned language models: Gemma 2B (Team et al., 2024), LLaMA 3.2B (Liu et al., 2025), and Qwen2.5 7B (Qwen et al., 2025). For each question, we generate a CoT response using a fixed prompt and segment it into reasoning steps via explicit textual markers. We extract token-level hidden states, compute cumulative Gram matrices, and obtain spectral embeddings by taking the top-64 eigenvalues. These embeddings are clustered into k=5 latent states via k-means to estimate a first-order transition matrix P. Monte Carlo rollouts from P yield synthetic reasoning trajectories.

4.1 Explainability of Reasoning Abstraction

We examine whether reasoning steps exhibit structurally coherent organization in the latent embedding space. For each step, we compute its spectral embedding and apply t-SNE for visualization. Figure 2 shows the resulting projections across four datasets, GSM8K, SocialIQA, Math, and MuSiQue, using LLaMA-3B. Coloring each point by its cluster assignment reveals clear separation with minimal overlap, suggesting that the latent representations capture distinct structural modes of reasoning. Notably, the same clustering patterns consistently emerge across diverse domains, indicating that the



(a) Llama on GSM8K (b) Llama on SocialIQA

(c) Llama on on Math (d) Llama on MusiQUe

Figure 2: t-SNE projection of Chain-of-Thought step embeddings from Llama-3B across four datasets.

abstraction is robust and generalizable beyond a single task. These visualizations provide explainability by revealing how reasoning steps naturally fall into functionally distinct groups, even without explicit supervision.

4.2 Explainability of Reasoning Semantics

To investigate whether latent clusters correspond to meaningful reasoning behaviors, we aggregate the texts of all steps within each cluster and manually summarize their functional roles. As shown in Table 1, clusters derived from LLaMA-generated CoT on SocialIQA align with intuitive categories such as scenario description, problem framing, option evaluation, and answer synthesis. These roles were annotated based on step content and ranked by their typical position in the CoT trajectory.

We further validate these interpretations by com-

Cluster ID	Label	Description	Rank in CoT
0	Scenario Description	Summarize the scenario's characters, actions, setting.	4
1	Problem Framing	Extract key facts and define the question and options.	2
2	Detailed Option Evaluation	Integrate reasoning steps into a clear conclusion.	5
3	Option Analysis	Laying out settings before evaluating choices.	1
4	Answer Synthesis	Systematically align each option with the context.	3

Table 1: Interpretation of unsupervised cluster assignments on LLaMA-generated reasoning steps from SocialIQA.

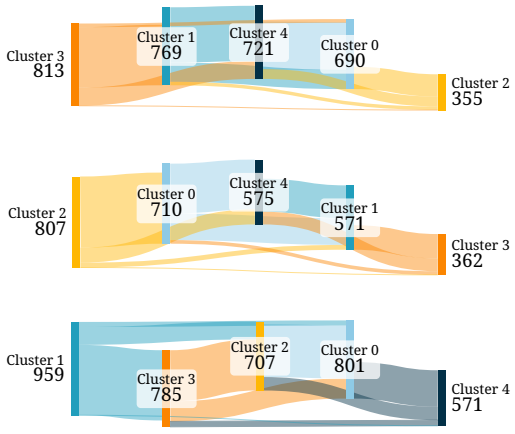
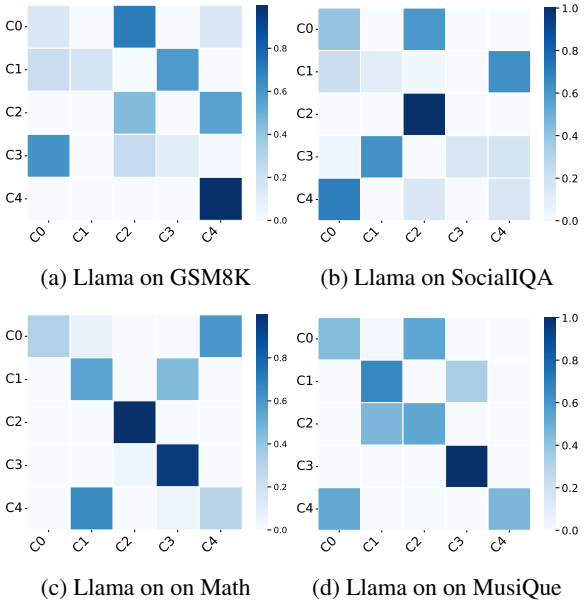


Figure 3: Transition diagrams for the SocialIQA dataset. Each diagram visualizes one model: Llama-3 3B (top), Gemma-2 2B (middle) and Qwen-7B (bottom)



(a) Llama on GSM8K (b) Llama on SocialIQA

(c) Llama on on Math (d) Llama on on MusiQue

Figure 4: Cluster transition probability heatmaps for Llama-3B across four benchmarks.

puting the average step index for each cluster. Table 2 reports the average step index of each cluster based on real CoT trajectories across models and datasets. For LLaMA, we observe strong consistency: clusters with lower average indices (e.g., C3, C1) correspond to early-stage functions like option analysis and problem framing, while those with higher indices (e.g., C2, C4) align with final-stage synthesis. This ordering effect mirrors our manual rank assignment in Table 1, reinforcing the interpretability of the latent abstraction.

4.3 Explainability of Reasoning Transition

To capture the structural dynamics of CoT reasoning, we model the transitions between latent states as a Markov chain. Given the clustered state sequence $\{z_t\}$, we estimate a transition matrix $P \in \mathbb{R}^{k \times k}$, where each entry $P_{i,j}$ reflects the empirical probability of transitioning from state i to j. We visualize P as a heatmap (Figure 4), which reveals structured and asymmetric transition patterns. For instance, certain clusters (e.g., setup-related) predominantly initiate trajectories, while others (e.g., synthesis) absorb transitions at the end, suggesting coherent behaviors across tasks.

To further highlight dominant reasoning trajectories, we render the most frequent transitions using Sankey diagrams (Figure 3) for models like LLaMA, Gemma, and Qwen on SocialIQA. These visualizations show that most trajectories follow consistent paths such as `scenario description` $\rightarrow$ `option evaluation` $\rightarrow$ `answer synthesis`, which aligns well with our semantic interpretation in Table 1. This correspondence provides evidence that the learned transition structure reflects meaningful reasoning phases, rather than arbitrary or noisy transitions.

We also compute the expected step index at which each cluster appears to ground these roles temporally. As shown in Table 2, the observed ordering of clusters is highly correlated with real reasoning positions, particularly for LLaMA. This confirms that the latent transition model captures natural CoT reasoning.

5 Conclusion

We proposed a state-aware transition framework that abstracts CoT trajectories into structured latent dynamics, offering a global perspective on multi-step reasoning in LLMs. Each reasoning step is

embedded via spectral analysis and clustered into semantically coherent states, with their transitions modeled as a first-order Markov chain. This approach moves beyond token-level attribution by uncovering consistent latent structures across models and tasks. It enables explainability applications such as semantic role identification and temporal pattern visualization.Empirical results demonstrate that LLMs exhibit structured reasoning patterns, pointing to underlying strategies beyond surface-level token sequences.

6 Limitation

Our framework assumes access to internal representations of open-source language models. This assumption is common in many existing interpretability and reasoning analysis works (Bharadwaj, 2024; Tang et al., 2025; Orlicki, 2025). While our framework provides interpretable abstractions of CoT reasoning, it focuses primarily on intrinsic structural analysis.

7 Acknowledgment

This work is partially supported by NSF IIS-2432486

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. *arXiv preprint arXiv:2305.18029*.

Aryasomayajula Ram Bharadwaj. 2024. Understanding hidden computations in chain-of-thought reasoning. *arXiv preprint arXiv:2412.04537*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*.

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. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, and 1 others. 2025. Llm-based text simplification and its effect on user comprehension and cognitive load. *arXiv preprint arXiv:2505.01980*.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. *Preprint*, arXiv:2310.14491.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Fuxiao Liu, Paiheng Xu, Zongxia Li, Yue Feng, and Hyemi Song. 2023. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *arXiv preprint arXiv:2307.05052*.

Xu Liu, Tong Yu, Kaige Xie, Junda Wu, and Shuai Li. 2024. Interact with the explanations: Causal debiased explainable recommendation system. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 472–481.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2025. Spinquant: Llm quantization with learned rotations. *Preprint*, arXiv:2405.16406.

José I Orlicki. 2025. Beyond words: A latent memory approach to internal reasoning in llms. *arXiv preprint arXiv:2502.21030*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *Preprint*, arXiv:1904.09728.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Preprint*, arXiv:1811.00937.

Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Wayne Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. 2025. Unlocking general long chain-of-thought reasoning capabilities of large language models via representation engineering. *arXiv preprint arXiv:2503.11314*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu, and Qi Alfred Chen. 2023. Does physical adversarial example really matter to autonomous driving? towards system-level effect of adversarial object evasion attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4412–4423.

Azmine Toushik Wasi and Mst Rafia Islam. 2024. Cogergllm: Exploring large language model systems design perspective using cognitive ergonomics. *arXiv preprint arXiv:2407.02885*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and 1 others. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, and 1 others. 2024a. Ocean: Offline chain-of-thought evaluation and alignment in large language models. *arXiv preprint arXiv:2410.23703*.

Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. 2022. Context-aware information-theoretic causal de-biasing for interactive sequence labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3436–3448.

Junda Wu, Yuxin Xiong, Xintong Li, Zhengmian Hu, Tong Yu, Rui Wang, Xiang Chen, Jingbo Shang, and Julian McAuley. 2025. Ctrls: Chain-of-thought reasoning via latent state-transition. *arXiv preprint arXiv:2507.08182*.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024b. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087.

Junda Wu, Tong Yu, and Shuai Li. 2021. Deconfounded and explainable interactive vision-language retrieval of complex scenes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2103–2111.

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

Chang Xiao and Brenda Yang. 2025. Streaming, fast and slow: Cognitive load-aware streaming for efficient llm serving. *arXiv preprint arXiv:2504.17999*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

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

Haiyan Zhao, Fan Yang, Bo Shen, Himabindu Lakkaraju, and Mengnan Du. 2024. Towards uncovering how large language model works: An explainability perspective. *arXiv preprint arXiv:2402.10688*.

A Model and Processing Details

We use publicly available instruction-tuned checkpoints: Gemma 2B Instruct (fast inference baseline), LLaMA 3.2B Instruct (balanced performance), and Qwen2.5 7B Instruct (strong reasoning quality). For each input, we prepend a standard reasoning prompt (e.g., "Let's solve this step by step.") and extract reasoning steps based on explicit markers (e.g., "Step 1:"). Hidden states from the final transformer layer are projected to 128 dimensions before Gram matrix computation. Top-64 eigenvalues are retained per step to form spectral embeddings.

A.1 Running Example: Llama-3B on SocialIQA

To ground our abstraction and simulation framework, we walk through a single question from the SocialIQA benchmark as answered by Llama-3B.

Prompt *{"id": "sample_33", "prompt": "Context: Cameron flew on a plane because he figured he would get there faster than driving.Question: What will Cameron want to do next? Options:A) buy a ticket B) get back on the plane C) find a hotel Let's solve this problem step-by-step."}*
Model response *{"sample_id": "sample_33", "detected_steps": 4, "steps": [{"step_id": 1, "text": "Step 1: Analyze the situation. Cameron flew on a plane because he thought it would be faster than driving. However, we don't know if he actually arrived faster or if he's now facing a new challenge."}, {"step_id": 2, "text": "Step 2: Consider the options. The options are buying a ticket, getting back on the plane, or finding a hotel. Since Cameron is already on the plane, getting back on the plane is unlikely."}, {"step_id": 3, "text": "Step 3: Evaluate the options. Buying a ticket is not necessary, as Cameron is already on the plane. Finding a hotel is a reasonable option, as it's common for people to stay overnight when they travel."}, {"step_id": 4, "text": "Step 4: Choose the most logical option. Given the context, finding a hotel seems like the most logical next step. The final answer is: C) find a hotel."}]}*

The transition probabilities reveal that Llama-3B most frequently transitions from goal-setting to hazard analysis before arriving at a conclusion, reflecting a two-stage reasoning pattern. A t-SNE plot further confirms that these clusters are well-separated in embedding space. This running example demonstrates how our abstraction captures both the content and dynamics of the model's internal reasoning on a typical SocialIQA question.

A.2 Simulating Reasoning Trajectories via Markov Rollouts

To validate whether the learned transition model faithfully captures the temporal structure of reasoning, we simulate step-wise trajectories using the Markov matrix P. Each trajectory is sampled by recursively drawing the next state from the conditional distribution:

$$s_{t+1} \sim P(\cdot \mid s_t).$$

This sampling process reduces to a standard Markov chain, where the acceptance probability is always 1 due to symmetric proposal and target distributions:

$$\alpha(s_t, s_{t+1}^*) = \min\left\{1, \frac{p(s_{t+1}^* \mid s_t)}{p(s_t \mid s_{t+1}^*)}\right\} = 1.$$

The resulting simulated trajectories $\tau^{(i)} = (s_0^{(i)}, \ldots, s_T^{(i)})$ can be used to estimate trajectory-level statistics via Monte Carlo approximation:

$$\mathbb{E}_{\tau \sim P}[f(\tau)] \approx \frac{1}{N} \sum_{i=1}^{N} f(\tau^{(i)}),$$

where f may represent properties such as average step index or transition counts. These simulations allow us to probe whether the model captures coherent and temporally structured reasoning flows.

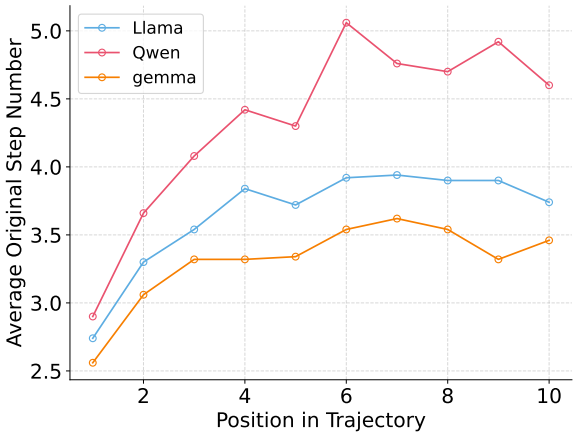


Figure 5: Average original step index at each position in the simulated trajectory on SocialIQA.

Dataset	Model	C0	C1	C2	C3	C4	ρ	p-value
SocialIQA	Gemma	2.26/1.99	3.20/2.74	1.46/1.56	3.56/6.35	2.76/2.27	1.000	0.001
	Llama	3.59/3.09	1.99/1.58	3.88/6.47	1.16/1.16	2.98/2.14	1.000	0.001
	Qwen	4.43/3.06	1.38/1.59	3.65/2.31	2.54/1.95	4.67/6.41	1.000	0.001
CSQA	Gemma	1.73/3.06	3.05/6.47	2.85/2.68	2.54/3.74	1.67/2.11	0.700	0.188
	Llama	2.08/1.65	3.62/3.06	3.11/2.07	3.95/6.46	1.16/1.18	1.000	0.001
	Qwen	1.29/1.31	4.35/3.20	2.41/1.81	4.48/6.49	3.63/2.27	1.000	0.001
GSM8K	Gemma	3.02/3.03	2.69/3.98	1.77/2.16	2.68/3.38	3.02/6.56	0.700	0.188
	Llama	2.44/2.06	1.19/1.17	3.10/3.15	1.69/1.56	3.12/6.50	1.000	0.001
	Qwen	4.30/3.51	1.47/1.47	3.59/2.65	4.52/6.51	2.43/2.04	1.000	0.001

Table 2: Simulated vs. real average cluster positions and Spearman statistics. Columns C0–C4 list mean simulated/real positions. Spearman ρ measures the monotonic agreement between simulated and real rankings ($\rho = 1$ is perfect), and the p-value tests its significance ($p < 0.05$ indicates a non-random correlation).

A.3 Validating Temporal Consistency via Monte Carlo Simulation

To evaluate whether the learned transition matrix captures the temporal structure of CoT reasoning, we perform Monte Carlo rollouts by sampling trajectories from the Markov model. Starting from a common initial cluster, we generate 10-step latent sequences based on P, and analyze the correspondence between simulated states and their real positions in CoT.

Figure 5 shows the average original step index of reasoning steps sampled at each simulated position. The upward trend across all models indicates that our transition model preserves the directional nature of reasoning—from early-stage states (e.g., problem framing) to later stages (e.g., synthesis). This validates the ability of P to approximate real-world temporal dynamics in CoT trajectories.

Beyond position-wise trends, we also examine the expected position of each cluster across simulated trajectories. We estimate $\mathbb{E}_{\tau \sim P}[f(\tau)]$ to obtain the average step index for each cluster and compare it against empirical values. Table 2 reports these results, along with Spearman correlation scores. For all models and datasets, we observe near-perfect alignment between simulated and real rankings, confirming that the learned abstraction preserves both semantic and temporal consistency.