How Chain-of-Thought Works? Tracing Information Flow from Decoding, Projection, and Activation

Anonymous ACL submission

Abstract

Chain-of-Thought (CoT) prompting significantly enhances model reasoning, yet its internal mechanisms remain poorly understood. We analyze CoT's operational principles by reversely tracing information flow across decoding, projection, and activation phases. Our quantitative analysis suggests that CoT may serve as a decoding space pruner, leveraging answer templates to guide output generation, with higher template adherence strongly correlating with improved performance. Furthermore, we surprisingly find that CoT modulates neuron engagement in a task-dependent manner: reducing neuron activation in open-domain tasks, yet increasing it in closed-domain scenarios. These findings offer a novel mechanistic interpretability framework and critical insights for enabling targeted CoT interventions to design more efficient and robust prompts. We released our code and data at `https://anonymous.4open.science/r/cot-D247`.

1 Introduction

Large language models (LLMs) excel in diverse tasks but falter in multi-step reasoning. Chain-of-Thought (CoT) prompting, introduced by (Wei et al., 2022), mitigates this by guiding models through step-by-step reasoning, markedly improving performance in arithmetic, commonsense, and symbolic reasoning tasks (Wang et al., 2023; Jin et al., 2024; Prabhakar et al., 2024; Nayab et al., 2025). Despite its effectiveness, CoT's underlying mechanisms remain poorly understood.

Prior studies propose hypotheses for CoT's efficacy. For instance, Madaan et al. (2023) argue that CoT reduces task complexity, making tasks more manageable for models. Similarly, Madaan and Yazdanbakhsh (2022) propose that models imitate answer templates in CoT prompts, and Schaeffer et al. (2023) suggest that prompt features unrelated to logical reasoning drive performance gains. Although these insights are intuitively compelling, they lack direct experimental support linking models' internal states to observed outcomes.

This work advances our understanding of CoT's mechanisms by adopting a mechanistic interpretability approach (Dumas et al., 2024; Lee et al., 2024). Mechanistic interpretability aims to "open" the model's "black box", investigating how its internal "*parts*" (computational units) and "*wiring*" (connection units) enable information to be sequentially processed and flow. It is crucial for CoT that significantly alters model behavior. Consistent with prior work (Chia et al., 2023; Madaan et al., 2023; Jin et al., 2024), we focus on vanilla CoT (Wei et al., 2022), as its internal mechanisms remain underexplored. Figure 7 shows our framework.

To investigate CoT's mechanisms, we select six models (ranging from 3B to 70B) and nine datasets spanning three core reasoning types: arithmetic, commonsense, and symbolic reasoning, covering a wide range of task complexities, input formats, and answer spaces (e.g., open-domain GSM8K and closed-domain AQuA). Our multi-faceted analysis yields several key mechanistic insights:

- ○ We propose a novel mechanism by which CoT may constrain the decoding space by leveraging answer templates. While previous work has hinted at the importance of prompt structure, we *quantify* this by demonstrating a strong correlation between reasoning structure adherence and performance.
- ◇ Our analysis demonstrates that CoT yields concentrated probability distributions, potentially minimizing prediction uncertainty, *providing new evidence* for CoT's role in enhancing model confidence.
- ⋆ We identify a task-dependent modulation of neuron engagement by CoT, which may vary across datasets: reducing activation in open-domain tasks, while unexpectedly increasing it in closed-domain tasks, a phenomenon *not previously reported*.

These findings provide a more nuanced understanding of CoT's mechanisms and offer a novel mechanistic interpretability and critical insights for designing more efficient and robust prompts.

2 Related Work

This section reviews mechanistic interpretability for LLM analysis and CoT prompting with its influencing factors. We then position our work, which applies mechanistic interpretability to investigate CoT's internal mechanisms.

Mechanistic Interpretability Mechanistic interpretability endeavors to reverse-engineer neural network computations. Unlike approaches focused solely on input-output correlations, it constructs mappings from inputs via internal states to outputs through an analysis of the model's internal structure (Nanda et al., 2023). Specific components have been investigated. Transformer *feed-forward networks* (FFNs), for instance, are interpreted as key-value memories linking textual patterns to output distributions (Geva et al., 2021b). Studies also characterize neuron properties and functions, such as identifying universal neurons (Gurnee et al., 2024), analyzing activation patterns of reasoning neurons in FFNs (Rai and Yao, 2024), and investigating activation sparsity (Voita et al., 2024).

Other mechanistic interpretability approaches analyze the flow and *representation* of information through intermediate layers. Techniques include patching intermediate representations to localize computations (Fierro et al., 2025), tracking the trajectory of embeddings through input, concept, and output spaces (Wendler et al., 2024), and extending methods like the logit lens on vocabulary embeddings (Cancedda, 2024). Collectively, these methods offer tools to dissect LLM internal processing and investigate mechanisms.

Chain-of-Thought CoT prompting enhances LLM reasoning by generating intermediate steps before the final answer (Wei et al., 2022). This is effective across diverse reasoning tasks (Tanneru et al., 2024) and leads to numerous extensions (Li and Qiu, 2023; Bi et al., 2024; Chen et al., 2024b).

Subsequent research on factors influencing CoT indicates that prompt format may be more crucial than the specific content or logical validity of reasoning steps. For instance, models perform effectively even with irrelevant (Webson and Pavlick, 2022) or logically invalid steps (Wang et al., 2023; Schaeffer et al., 2023), or when keywords are absent, provided the overall reasoning structure is preserved (Li et al., 2025). Supporting this notion, counterfactual analyses have revealed that consistent patterns and common text formats, rather than specific symbols or grammatical details, are key to CoT prompting (Madaan et al., 2023).

Furthermore, other factors influencing CoT efficacy include the length of reasoning steps, where longer rationales boost performance while shorter ones diminish it (Jin et al., 2024). Other works suggest CoT primarily enhances imitation of style and instructions rather than factuality or problem-solving (Gudibande et al., 2024), influences the robustness of feature attribution scores (Wu et al., 2023), benefits from diverse reasoning skills shown across exemplars (Ye et al., 2023), and is affected by factors like probability, memorization, and noisy reasoning in specific tasks (Prabhakar et al., 2024).

Despite these insights, a key gap remains in understanding how CoT alters a model's internal states to produce these effects, which drives our investigation.

3 Experimental Setup

Following (Wei et al., 2022; Chia et al., 2023; Madaan et al., 2023; Jin et al., 2024), we selected nine datasets spanning three core reasoning task types: arithmetic, commonsense, and symbolic reasoning, where CoT demonstrates substantial performance gains. Examples are shown in Table 2.

Arithmetic reasoning tasks. These tasks require to solve mathematical problems through multi-step calculations. We used three widely adopted datasets: GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021), and AQuA (Ling et al., 2017).

Commonsense reasoning tasks. These tasks involve answering questions based on commonsense knowledge. We employed four commonly used datasets: Bamboogle (Press et al., 2023), StrategyQA (Geva et al., 2021a), Date, and Sports (bench authors, 2023).

Symbolic reasoning tasks. These tasks involve processing symbolic sequences using logical rules. We considered two available datasets: the Coin Flip (Wei et al., 2022) and the Last Letters Concatenation (Kojima et al., 2022).

3.1 Models & Parameters Settings

Consistent with prior work (Chia et al., 2023; Madaan et al., 2023; Jin et al., 2024), we focus on vanilla CoT (Wei et al., 2022), as its internal mechanisms remain underexplored. We used 4-shot prompts derived from (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023). We evaluated six pretrained models from diverse families and scales: LLaMA3.1 (8B, 70B) (Grattafiori and etc.,

Task	Dataset (Answer Space)	Example
Arithmetic	GSM8K / SVAMP (Open, Numerical)	**Question**: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking? **Answer**: There are originally 3 cars. 2 more cars arrive. $3+2=5$. So the answer is 5.
	AQuA (Closed, Options)	**Question**: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance? **Choices**: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km **Answer**: The distance that the person traveled would have been 20 km/hr $\times$ 2.5 hrs $=$ 50 km. So the answer is (e).
Common sense	Bamboogle (Open, Text)	**Question**: Who lived longer, Theodor Haecker or Harry Vaughan Watkins? **Answer**: Theodor Haecker was 65 years old when he died. Harry Vaughan Watkins was 69 years old when he died. So the answer is Harry Vaughan Watkins.
	StrategyQA / Sports (Binary, Yes/No)	**Question**: Do hamsters provide food for any animals? **Answer**: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals. So the answer is Yes.
	Date (Formatted)	**Question**: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY? **Answer**: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.
Symbolic	Coin Flip (Binary, Yes/No)	**Question**: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up? **Answer**: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is No.
	Last Letters (Open, Text)	**Question**: Take the last letters of each words in "Lacey Nora Debra Ashleigh" and concatenate them. **Answer**: The answer is yaah.

Table 2: Part Dataset *Examples* (Full examples is in Appendix B).

2024), Gemma2 (2B, 9B, 27B) (Team, 2024), and LLaMA3.2-3B (AI, 2024). Greedy decoding with a 300-token limit ensured deterministic output. Each dataset's test set was evaluated, with performance measured by accuracy, extracted via regular expressions. CoT prompt exemplars are shown in Table 2, with full prompts in Appendix L. This study compared LLM behaviors under CoT and standard prompts. Due to space constraints, representative results are presented here, with comprehensive results in Appendix E $\sim$ J.

4 Methodology

CoT prompts elicit step-by-step generation, so our analysis starts with decoding (generated tokens), tracing backward through projection (probability projections) to activation (FFN neurons). By analyzing these interconnected layers, we aim to provide a comprehensive understanding of CoT.

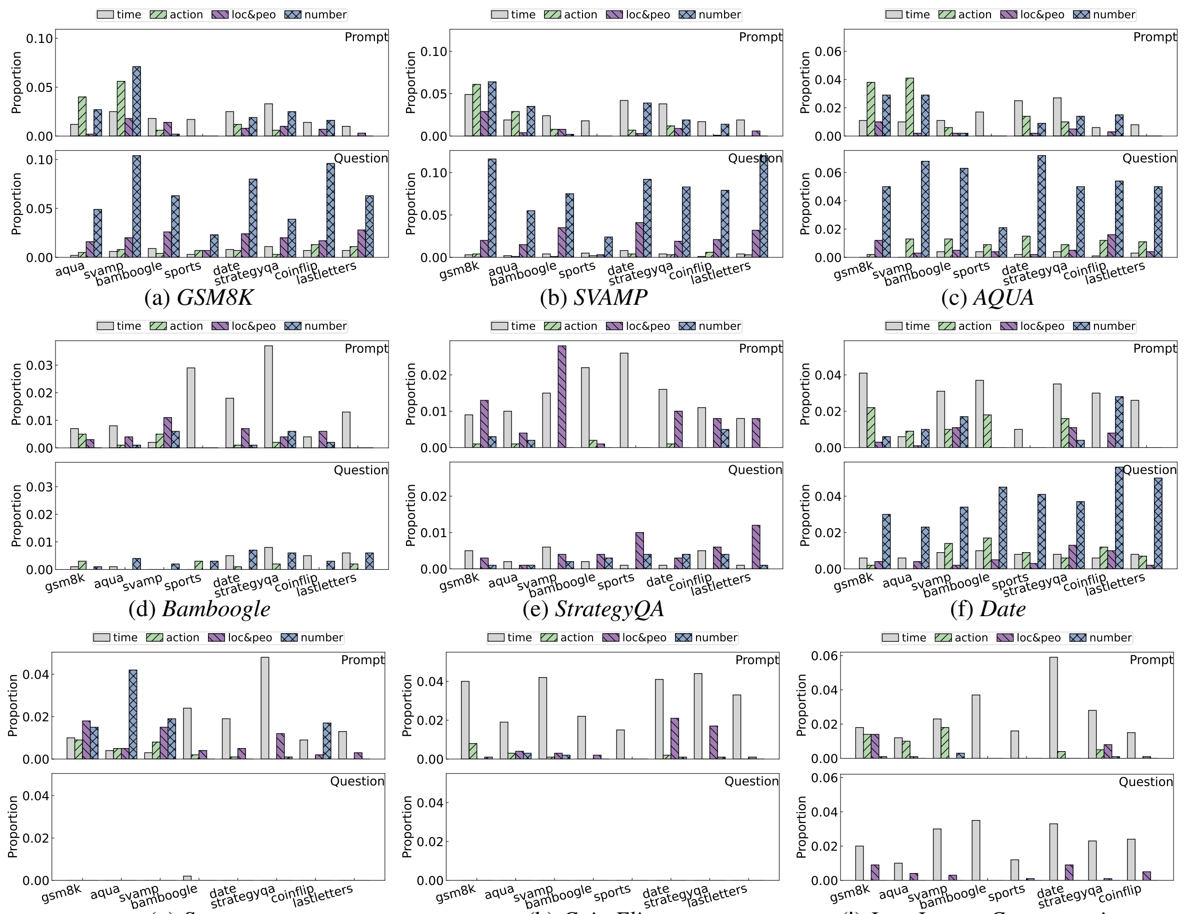
4.1 Decoding

CoT prompting reshapes model outputs by guiding the generation of intermediate reasoning steps before the final answer. We hypothesize that this guidance narrows the decoding space, yielding more structured, task-relevant outputs. To investigate this, we analyzed the characteristics of the generated text from two perspectives: the imitation of specific keywords present in prompts and questions, and the adherence to a answer structure defining CoT reasoning.

Keyword Imitation Analysis Prior work (Madaan and Yazdanbakhsh, 2022; Madaan et al., 2023; Gudibande et al., 2024) suggest that CoT improves reasoning by prompting models to mimic prompt formats, but the degree of imitation remains unquantified. To bridge this gap, we *first* sought to measure imitation by identifying specific keywords and comparing their presence in model outputs against the input (prompts and questions). To this end, we introduce the concept of "*test points*", which are keywords reflecting key reasoning aspects as observed in the generated CoT steps.

Test Points. We classified test points into four types based on observed patterns in input imitation: 1) *time* (e.g., "before", "therefore), indicating temporal order, logical sequence, or causality; 2) *action* (e.g., "add", "increase"), representing operations; 3) location and people (*loc & peo*) (e.g., "there", "someone"); and 4) *numbers* (e.g., "1", "two"). Keywords were extracted by analyzing prompts and questions across all test datasets. Using the *Spacy* library, we performed part-of-speech tagging to identify candidate words (e.g., verbs for actions, adverbs for time), followed by manual verification for accurate categorization. The keyword list is shown in Table 6. Imitation is measured as the proportion of test points in the model's generated text that match those in the prompt or question. Additionally, to assess the generalizability of CoT prompts and models' adaptability to diverse reasoning structures, we conducted cross-dataset prompt transfer experiments. CoT prompts designed for dataset X were applied to dataset Y, and we evaluated changes in generated content characteristics.

Analysis. Figure 1 illustrates the average propor-



(a) *GSM8K* (b) *SVAMP* (c) *AQUA* (d) *Bamboogle* (e) *StrategyQA* (f) *Date* (g) *Sports* (h) *Coin Flip* (i) *Last Letters Concatenation*

Figure 1: Keyword imitation results (Gemma2-27B is reported).

tion of generated tokens that correspond to the four defined test points for both prompts and questions across different datasets. Our analysis of keyword imitation reveals two primary patterns, shedding light on how CoT influences the structure and content of generated reasoning steps. First, we observe **varied imitation patterns by source**: Models tend to imitate "time" and "action" keywords more frequently from the CoT prompts, suggesting that prompts effectively convey sequential and operational structuring cues for the reasoning process. In contrast, "number" keywords are often imitated more from the input questions, particularly in arithmetic datasets like GSM8K, SVAMP, AQUA, and Date (i.e., Figure 1a $\sim$ 1c, and 1f). This indicates that while CoT prompts provide the structural framework, models primarily extract specific content and entities directly from the problem statement to populate this structure.

Second, we find **varied imitation patterns across datasets**: Imitation of question test points is notably *lower* for tasks requiring significant external commonsense knowledge, such as Bamboogle, StrategyQA, and Sports (Figure 1d, 1e, and 1g). This is likely because the necessary reasoning information is not contained within the input question, compelling models to rely more heavily on their internal knowledge. Similarly, in Coin Flip (Figure 1h), the reasoning involves concepts not explicit in the question, reducing direct keyword imitation from the input. Conversely, the Last Letters Concatenation (Figure 1i) exhibits balanced imitation, reflecting the high relevance of both prompt and question keywords to its structured symbolic task.

These findings demonstrate that CoT prompting effectively steers models to produce text that conforms to a specified reasoning structure by selectively imitating structural keywords (e.g., "time", "action") from the prompt, while incorporating task-relevant content (e.g., "number") by mimicking keywords from the input question. This differential keyword imitation underpins CoT's ability to impose order on model outputs. We posit that such structural guidance narrows the decoding space, enhancing the focus and accuracy of token predictions, as further explored in the subsequent analysis of probability projections.

Reasoning Structure Adherence Analysis Building upon the keyword imitation findings, we further analyze the model's adherence to a higher-level, abstract reasoning structure commonly observed in effective CoT generations. Our observation that "time", "action", "loc&peo", and "number" frequently fulfill specific syntactic roles within the reasoning process motivates the formalization of a CoT Reasoning Structure:

$$\texttt{Reasoning Structure:}\ \mathcal{E}_p \xrightarrow{\mathcal{O}} \mathcal{E}_g + \mathcal{S}_l, \quad (1)$$

where $\mathcal{E}_p$ represents input entities (often corresponding to "number" or "loc&peo"), $\mathcal{O}$ represents reasoning operations or predicates (i.e., "action"), $\mathcal{E}_g$ signifies derived intermediate entities, and $\mathcal{S}_l$ is the final answer statement (i.e., "the answer is..."). For example, in a step like "3 + 2 = 5. So the answer is 5.", "3" and "2" are $\mathcal{E}_p$, "+" is $\mathcal{O}$, "5" is $\mathcal{E}_g$, and "So the answer is 5." is $\mathcal{S}_l$. By quantifying the extent to which generated samples adhere to this structure (measured as "Imitation Count" based on keyword patterns and their sequence), we aim to assess the model's ability to capture the structural properties of CoT reasoning. We hypothesize that this ability to generate structurally coherent reasoning steps is essential for multi-step problem-solving, as it provides a clear path towards the correct answer, thereby helping the model navigate the problem space more effectively.

Analysis. We examined the relationship between the adherence level ("Imitation Count") and task performance (Accuracy) on the GSM8K dataset, utilizing the original CoT prompt and several transferred CoT prompts as shown in Figure 2.

We first observe a **consistent strong positive correlation** between Imitation Count and Accuracy. As more samples align with the CoT reasoning structure, performance on GSM8K improves, providing evidence that the performance benefits of CoT are strongly coupled with the model's ability to generate content adhering to the reasoning structure. These findings quantitatively validate our hypothesis: CoT-induced coherent reasoning structures are critical for performance gains in multi-step reasoning tasks.

Second, scatter plots reveal differences between prompt types. Prompts with reasoning patterns aligning with the target task's structural needs, such as sequential or arithmetic structures (e.g., Date or GSM8K's native prompt), are effective at inducing high adherence to the defined reasoning structure and lead to better performance. Conversely, incompatible prompts, like the commonsense-focused

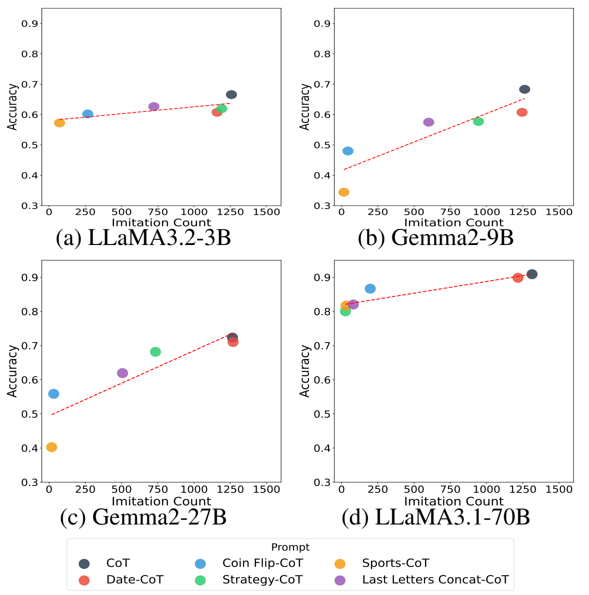


Figure 2: Correlation between reasoning structure adherence and accuracy on GSM8K dataset.

Sports-CoT applied to arithmetic GSM8K, result in low structural adherence and poorer performance. These findings reveal that CoT's effectiveness hinges on this structural alignment – its ability to guide the model's decoding process towards generating a coherent sequence of reasoning steps that follows the expected format. This structural guidance, rather than the strict logical correctness of the content within steps, appears to be the key factor. This provides insight into why logically incorrect steps in CoT can still be effective: as long as these generated steps maintain the overall reasoning structure and adhere to the expected format (which our "Imitation Count" measures), the model can follow this structural template during decoding, even if the step's content is logically flawed. The adherence to the structural template is prioritized over the logical validity of the step's content.

4.2 Projection

To examine how CoT reshapes model behavior within the information flow framework, we analyzed the projection phase, where internal states are mapped to probability distributions over the vocabulary. We study this phase from two perspectives: the probability of generated sequences and the probability distribution of individual tokens.

Probability of Generated Token *Sequences*. To evaluate the model's generation confidence, we analyzed the probability of generated token sequences, focusing on the common phrase "answer is ..." across all datasets and prompt types. We select this phrase as it marks the final deci-

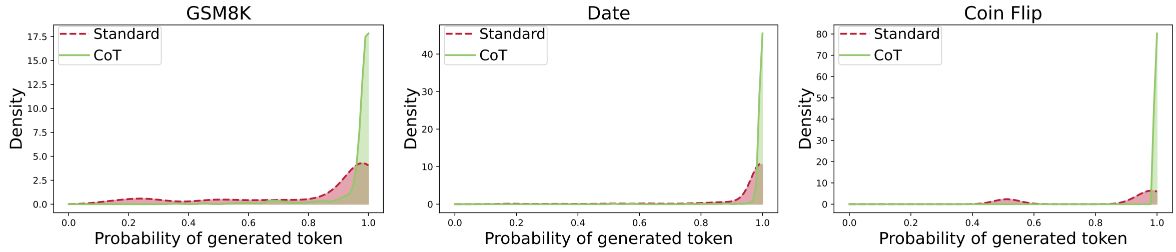
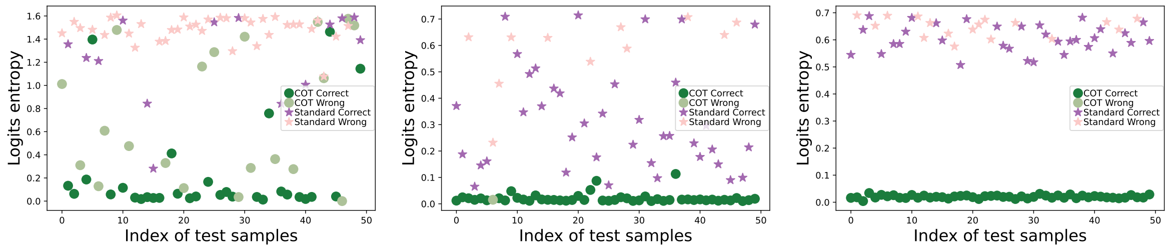


Figure 3: Kernel density estimation (Gemma2-9B is reported, see Appendix H for more details).



(a) *AQuA* (b) *Sports* (c) *Coin Flip*

Figure 4: Entropy (Gemma2-27B is reported, see Appendix I for more details)

sion point, revealing CoT's impact on decision certainty. We model the probability sequence as $\mathcal{P} = [p(\text{“answer”}), p(\text{“is”}), \ldots]$ and compute kernel density estimates (KDE) to visualize the probability distribution's *density*. See Appendix H for details about KDE.

Analysis. As shown in Figure 3, the generated probabilities of CoT ($\mathcal{P}_{\texttt{CoT}}$) are *consistently higher and more concentrated* than those of standard prompts ($\mathcal{P}_{\texttt{Standard}}$). This implies that with CoT, the model is assigning significant probability mass to a smaller and more specific set of next tokens at each step of generation. This shift in probability distribution indicates that CoT's structural guidance, provided through the intermediate steps (as discussed in Section 4.1), effectively constrains the decoding space. By limiting the set of plausible next tokens and narrowing down the possible continuations, this guidance leads to reduced uncertainty and increased model conviction, facilitating more decisive generation of the concluding tokens.

Probability Distribution of *Individual* Tokens. Beyond specific token probabilities, we analyzed the entire probability distribution across the vocabulary at each generation step. This allows us to understand how the model's attention was divided across vocabulary. We used entropy, calculated as $H(\mathcal{P}) = -\sum_{i=1}^{n} p_i \log(p_i)$ for a probability distribution $H(\mathcal{P})$, to measure the uncertainty in the distribution, where lower entropy indicates a more concentrated distribution. Additionally, to provide a controlled setting allowing us to verify the probabilities correspond to legitimate answer choices, we strategically selected closed-domain datasets with fixed answer options, including AQuA (answer space: “*a, b, c, d, e*”), Sports, and Coin Flip (answer space: “*yes, no*”).

Analysis. Continuing our investigation into how CoT shapes information flow, Figure 4 presents the entropy values of the probability distribution over tokens, reflecting the model's predictive uncertainty during the Projection phase. Our analysis reveals two key findings. First, we observe that correct answers consistently exhibit lower entropy than incorrect ones. This finding aligns with previous work (Li et al., 2024), suggesting that the model is more certain when its output is correct. Second, and more significantly for understanding CoT's mechanism, CoT prompts consistently lead to substantially lower entropy values compared to standard prompts. This implies that CoT effectively narrows the model's predictive focus to a smaller set of more relevant tokens at decision points, thereby sharpening its decision boundaries in the probability landscape. n essence, CoT appears to foster a more concentrated distribution and reduced predictive uncertainty. This reduction in uncertainty provides strong evidence that CoT fundamentally alters the model's output generation process by reducing the ambiguity in token prediction. This complements our findings on structural adherence in the Decoding phase.

4.3 Activation

Motivated by the findings of Yi et al. (2024); Rai and Yao (2024); Chen et al. (2024a); Voita et al. (2024), which showed that different neurons are activated by different types of information, we analyzed the activated neurons in FFNs, which is formulated as:

$$\mathbf{h}^{(l)} = \mathbf{W}_{\text{down}}^{(l)}(\text{Act}(\tilde{\mathbf{h}}^{(l)}\mathbf{W}_{\text{up}}^{(l)}),$$

where $\tilde{\mathbf{h}}^{(l)} \in \mathbb{R}^d$ is the hidden state output by attention module, serving as the input information flow to the FFN layer. $\mathbf{W}_{\text{up}}^{(l)} \in \mathbb{R}^{d \times d_1}$ is the upward projection weight, projecting $\tilde{\mathbf{h}}^{(l)}$ into higher-dimensional space (typically $d_1 \gg d$), enabling the FFN to represent a far greater number of features than its neuron count. $\text{Act}(\cdot)$ denotes the activation function (e.g., SwiGLU) and $\mathbf{W}_{\text{down}}^{(l)} \in \mathbb{R}^{d_1 \times d}$ maps the processed high-dimensional features back to the original space, reflecting the outflow of information. $\text{Act}(\tilde{\mathbf{h}}^{(l)}\mathbf{W}_{\text{up}}^{(l)}) \in \mathbb{R}^{d_1}$ represents the neurons. For GeLU or SwiGLU activation functions, a neuron is considered activated if its output is greater than zero, i.e., $\text{Act}(\tilde{\mathbf{h}}^{(l)}\mathbf{W}_{\text{up}}^{(l)}) > 0$. This definition aligns with prior work by Geva et al. (2021b); Voita et al. (2024)

Overall Neuron Activation. At the generation step $t \in [1, \cdots, T]$, the amount of activated neurons is calculated by:

$$A_t^{(l)} = \sum_{j=1}^{d_1} \mathbb{I}\left[\text{Act}(\tilde{\mathbf{h}}_t^{(l)}, \mathbf{W}_{\text{up}}^{(l)})_j > 0\right],$$

where $\mathbb{I}[\cdot]$ is the indicator function and $\tilde{\mathbf{h}}_t^{(l)}$ represents the hidden state for token x_t at layer l. The average Neuron Activation Count during generation is $\bar{\mathcal{A}} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{A}_t$ and $\mathcal{A}_t = \sum_{l=1}^{L} A_t^{(l)}$.

Analysis. Figure 5 presents the average Neuron Activation Count for CoT and standard prompts across various datasets for the LLaMA3.1-70B model. Across datasets, the distribution of CoT generally exhibits a downward shift compared to standard prompts, resulting in a *lower* overall number of activated neurons during the generation process. For instance, on the AQuA dataset, standard prompts engage $\sim 820K$ neurons on average, whereas CoT prompts induce a lower average activation of $\sim 790K$ neurons.

This observation suggests that CoT facilitates a more focused processing regimen. Drawing upon our earlier findings, a potential mechanism is that CoT's structured decomposition of complex problems into sequential steps guides the model's processing attention. This focused guidance may allow FFNs to operate more selectively, activating a smaller, task-relevant subset of neurons at each step. This selective processing is hypothesized to be related to the reduced predictive uncertainty observed in the Projection phase.

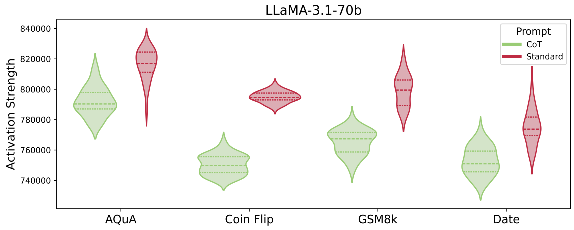


Figure 5: Average Neuron Activation Count (Full results are in Appendix K).

Furthermore, we observed that the distribution for Coin Flip appears relatively "short and wide", indicating a more concentrated range of activation counts, while that for other datasets (e.g. GSM8K) is "tall and narrow", suggesting a tighter cluster around the mean. These differences likely reflect the varying activation patterns associated with different reasoning tasks. Tasks requiring discrete state tracking (e.g., "head up", "even/odd") might involve a different distribution of neural activity compared to tasks requiring continuaus processing of information (e.g, $25 - 9 = 16, 16 + 3 = 19$).

While these task-dependent variations in the distribution of overall activation counts are observed, the precise mechanistic implications of these specific distribution shapes and their relation to different reasoning strategies warrant further detailed investigation in future work. While this analysis of overall activation provides insight into the general efficiency changes induced by CoT, a more nuanced and task-dependent picture of how CoT modulates neural activity emerges when examining activation differences at the layer level.

Layer-wise Activation Differences. To understand precisely where in the model CoT influences neuron activation, we analyzed the difference in activated neuron count between CoT and Standard prompts at each layer. This layer-wise activation difference is defined as:

$$\mathcal{A}^{(l)} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{A}^{(l)}t, \quad \Delta\mathcal{A}^{(l)} = \mathcal{A}\texttt{CoT}^{l} - \mathcal{A}_{\texttt{Standard}}^{(l)}$$

Analysis. Figure 6 illustrates the evolution of this layer-wise activation difference ($\Delta\mathcal{A}^{(l)}$) across model layers (l) for representative datasets, with more intense colors represent larger $|\Delta\mathcal{A}^{(l)}|$. This visualization reveals that:

Concentration in Later Layers. The most pronounced differences consistently occur in the final $1/3$ of the model's layers. This implies that CoT primarily influences the later stages of processing within the model, which are typically associated

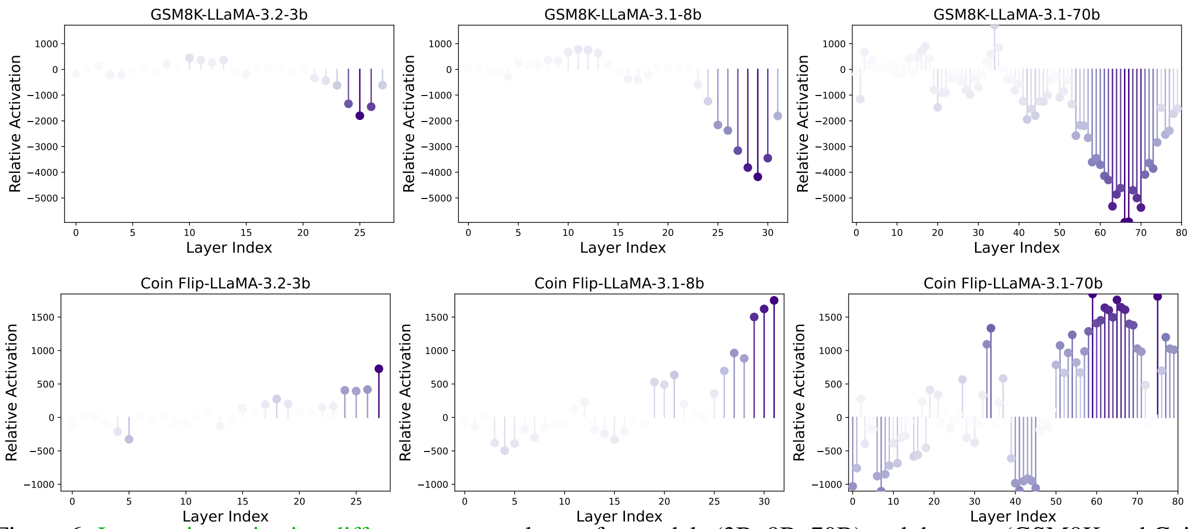


Figure 6: Layer-wise activation differences across layers for models (3B, 8B, 70B) and datasets (GSM8K and Coin Flip). For full datasets results, please refer to Appendix J.

with higher-level semantic processing, reasoning, and output generation. This observed concentration of CoT's impact in later layers may support that CoT's guidance is most relevant during the structuring of final reasoning steps and the formulation of the answer, processes that are predominantly handled by these later layers.

Model Size Influence. Larger models (e.g., 70B) exhibit generally *larger and more widespread layer-wise activation differences* compared to smaller models (e.g., 3B). While the precise reasons are complex, these larger differences might reflect how CoT allows larger models to better utilize their greater representational capacity or interact differently with their hierarchical structures, potentially influencing a broader range of neuron functions or abstraction levels.

Task-Specific Patterns. A key finding is the emergence of contrasting task-specific patterns in the layer-wise activation differences $\Delta\mathcal{A}^{(l)}$, particularly in later layers. Specifically, for open-domain tasks like GSM8K and Bamboogle, we consistently observe $\Delta\mathcal{A}^{(l)} < 0$ in the final layers, conversely, for closed-domain tasks such as Coin Flip, AQUA, and Sports, we find $\Delta\mathcal{A}^{(l)} > 0$ in the later layers. These contrasting patterns reflect how CoT structures the model's processing in response to different task demands. We hypothesize that in open-domain tasks, which often require navigating a vast solution space, CoT's step-by-step guidance facilitates a more focused processing mode. By explicitly laying out a reasoning path, CoT could allow the model to selectively engage relevant features or knowledge, leading to the observed reduction in later-layer activation. Conversely, for closed-domain tasks, where the primary challenge is discriminating among limited options, CoT's guidance might encourage the model to thoroughly evaluate features associated with all plausible choices.

This comprehensive consideration of options, potentially by activating relevant FFNs encoding those features, could explain the observed increase in later-layer activation. These distinct modes of modulation suggest interpretive analogies for CoT's function at the neural level: akin to a "pruner" reducing activity in open-domain tasks, and an "amplifier" boosting relevant features within the predefined answer space for closed-domain task, revealing that CoT's influence on model processing (observed here at the activation level) *maybe not uniform but dynamically tailored to the demands of the task*, providing novel insights into its operational principles.

5 Conclusion

In this work, we provided a comprehensive mechanistic interpretability analysis of CoT prompting by tracing information flow from decoding to activation. Our findings suggest that CoT may constrain the decoding space, reduces predictive uncertainty, and modulates neuron engagement in a task-dependent manner. These findings provide a deeper understanding of CoT's internal mechanisms and offer a novel mechanistic interpretability framework for analyzing and comparing prompting techniques for LLMs. Future research should investigate the interplay between task properties (including difficulty and type) in shaping activation patterns and information flow.

Limitations

This study offers valuable mechanistic insights into the effects of CoT prompting on LLMs by analyzing information flow across different internal stages. However, it is crucial to contextualize our findings within the broader limitations inherent to the current state of LLM interpretability research and the fundamental opacity of these complex systems.

Modern LLMs, with their immense scale and intricate architectures, largely function as "black boxes". Our understanding of the precise roles of individual modules, the nature of high-dimensional representations, and how billions of parameters collectively give rise to complex behaviors like CoT-guided reasoning remains nascent. This fundamental lack of transparency makes it exceedingly difficult to establish definitive causal links between external inputs (like a CoT prompt) and specific internal computational processes or neural activities that drive observed outputs and performance gains. The problem of attribution, i.e., pinpointing why a model behaves in a certain way, is a significant challenge in this nascent field.

Consequently, research in LLM interpretability, including this work, is often necessarily fragmented. Studies tend to focus on isolated aspects of the model (e.g., specific layers, activation patterns, or particular types of inputs/outputs) or employ specific analytical techniques to probe individual phenomena. While each such piece contributes to the overall puzzle, assembling these disparate findings into a complete, unified, and fully causal mechanistic account of complex behaviors like CoT's influence is a significant undertaking that lies at the frontier of the field and is beyond the scope of any single study at this time.

Furthermore, a common challenge in current interpretability research, stemming from the difficulty in establishing clear causal chains within the black box, is that the relationship between experimental observations and derived conclusions is often heuristic or suggestive rather than strictly conclusive. We can observe strong correlations (e.g., between structural adherence and performance, or CoT and certain activation patterns) and propose plausible mechanistic hypotheses based on these correlations. However, definitively proving causality through direct intervention or formal verification within large-scale models remains a complex methodological hurdle for the field. Our findings, while empirically supported and offering valuable insights, should be interpreted as strong evidence supporting particular mechanistic hypotheses about CoT's operation within the constraints of current interpretability methods.

This work contributes significantly to empirically grounded investigations of CoT mechanisms under these prevailing conditions. A comprehensive and ultimately causal understanding of how CoT fundamentally alters LLM processing will require continued advancements across the entire field of mechanistic interpretability.

References

Meta AI. 2024. Llama 3.2: Connect 2024 — vision on the edge for mobile devices. Accessed: 2025-02-16.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Zhen Bi, Ningyu Zhang, Yinuo Jiang, Shumin Deng, Guozhou Zheng, and Huajun Chen. 2024. When do program-of-thought works for reasoning? In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*.

Nicola Cancedda. 2024. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4792–4808, Bangkok, Thailand. Association for Computational Linguistics.

Ruizhe Chen, Tianxiang Hu, Yang Feng, and Zuozhu Liu. 2024a. Learnable privacy neurons localization in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 256–264, Bangkok, Thailand. Association for Computational Linguistics.

Sijia Chen, Baochun Li, and Di Niu. 2024b. Boosting of thoughts: Trial-and-error problem solving with large language models. In *The Twelfth International Conference on Learning Representations*.

Yew Ken Chia, Guizhen Chen, and 1 others. 2023. Contrastive chain-of-thought prompting. *arXiv preprint arXiv:2311.09277*.

Karl Cobbe, Vineet Kosaraju, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Clément Dumas, Veniamin Veselovsky, Giovanni Monea, Robert West, and Chris Wendler. 2024. How do llamas process multilingual text? a latent exploration through activation patching. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How do multilingual language models remember facts? *Preprint*, arXiv:2410.14387.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021a. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021b. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori and Abhimanyu Dubey etc. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Arnav Gudibande, Eric Wallace, and 1 others. 2024. The false promise of imitating proprietary language models. In *The Twelfth International Conference on Learning Representations*.

Wes Gurnee, Theo Horsley, and 1 others. 2024. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*.

Mingyu Jin, Qinkai Yu, and 1 others. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842, Bangkok, Thailand. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: a case study on dpo and toxicity. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Bo Li, Qinghua Zhao, and Lijie Wen. 2024. Rome: Memorization insights from text, logits and representation. *arXiv preprint arXiv:2403.00510*.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Shishir G Patil, Matei Zaharia, Joseph E Gonzalez, and Ion Stoica. 2025. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*.

Xiaonan Li and Xipeng Qiu. 2023. MoT: Memory-of-thought enables ChatGPT to self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374, Singapore. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? a counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore. Association for Computational Linguistics.

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

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Sania Nayab, Giulio Rossolini, and 1 others. 2025. Concise thoughts: Impact of output length on llm reasoning and cost. *Preprint*, arXiv:2407.19825.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. 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*, pages 2080–2094, Online. Association for Computational Linguistics.

Akshara Prabhakar, Thomas L. Griffiths, and R. Thomas McCoy. 2024. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3710–3724, Miami, Florida, USA. Association for Computational Linguistics.

Ofir Press, Muru Zhang, and 1 others. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Daking Rai and Ziyu Yao. 2024. An investigation of neuron activation as a unified lens to explain chain-of-thought eliciting arithmetic reasoning of LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7174–7193, Bangkok, Thailand. Association for Computational Linguistics.

Rylan Schaeffer, Kateryna Pistunova, and 1 others. 2023. Invalid logic, equivalent gains: The bizarreness of reasoning in language model prompting. In *ICML 2023 Workshop on Knowledge and Logical Reasoning in the Era of Data-driven Learning*.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1288–1301, Bangkok, Thailand. Association for Computational Linguistics.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

Biao Yi, Sishuo Chen, and 1 others. 2024. BadActs: A universal backdoor defense in the activation space. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5339–5352, Bangkok, Thailand. Association for Computational Linguistics.

A Framework

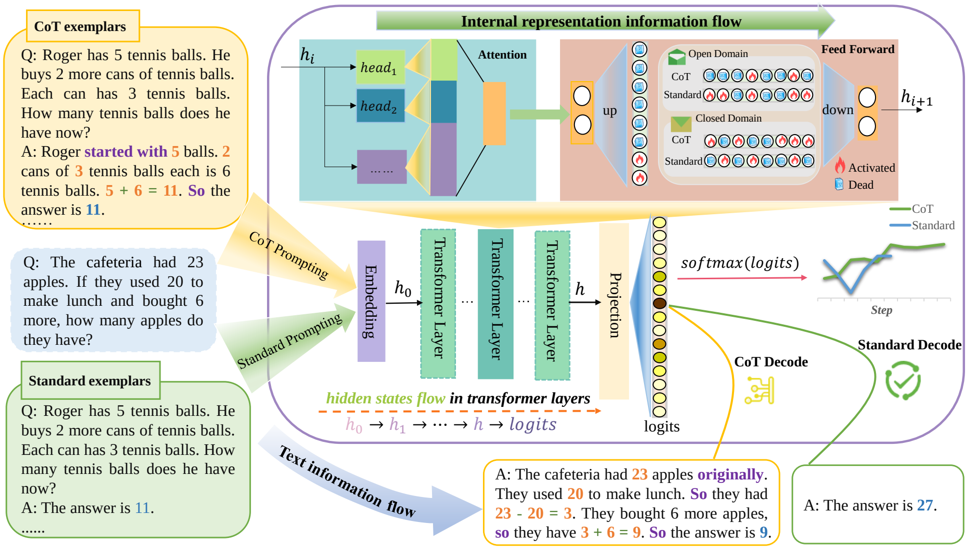


Figure 7: Framework of our method

B Examples of Tested Datasets

Task	Dataset (Answer Space)	Example
Arithmetic	GSM8K (Open, Numerical)	**Question**: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking? **Answer**: There are originally 3 cars. 2 more cars arrive. $3 + 2 = 5$. So the answer is 5.
	SVAMP (Open, Numerical)	**Question**: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny? **Answer**: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny $20 - 12 = 8$. So the answer is 8.
Arithmetic	AQuA (Closed, Options)	**Question**: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance? **Choices**: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km **Answer**: The distance that the person traveled would have been 20 km/hr $\times$ 2.5 hrs $=$ 50 km. So the answer is (e).
Common sense	Bamboogle (Open, Text)	**Question**: Who lived longer, Theodor Haecker or Harry Vaughan Watkins? **Answer**: Theodor Haecker was 65 years old when he died. Harry Vaughan Watkins was 69 years old when he died. So the answer is Harry Vaughan Watkins.
	StrategyQA (Binary, Yes/No)	**Question**: Do hamsters provide food for any animals? **Answer**: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals. So the answer is Yes.
	Sports (Binary, Yes/No)	**Question**: Is the following sentence plausible? "Kyle Palmieri was called for slashing" **Answer**: Kyle Palmieri is a hockey player. Being called for slashing is part of hockey. So the answer is Yes.
	Date (Formatted)	**Question**: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY? **Answer**: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.
Symbolic	Coin Flip (Binary, Yes/No)	**Question**: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up? **Answer**: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is No.
	Last Letters Concatenation (Open, Text)	**Question**: Take the last letters of each words in "Tim Candace Cecil Misael" and concatenate them. **Answer**: The last letter of "Tim" is "m". The last letter of "Candace" is "e". The last letter of "Cecil" is "l". The last letter of "Misael" is "l". Concatenating them is "mell". So the answer is mell.

Table 4: Full Dataset Examples

C Test Points

Test points	Contained words
time	originally, then, after, so, start, first, next, last, finally, before, later, afterwards, subsequently, meanwhile, during, while, when, once, as, since, because, due, hence, therefore, thus, consequently, accordingly, result, resulting, resulted, initially, earlier, until, at the same time
action	+, -, *, /, =, >, <, add, subtract, multiply, divide, average, increase, decrease, equal, calculate, total, square, root, cube, prime
loc&peo	there, location, site, area, spot, venue, someone, somebody, anyone, nobody, everyone, person, individual, participant, operator, handler, 's, his, her, their, its, he, she, they, it
number	"it was obtained by regular expression"

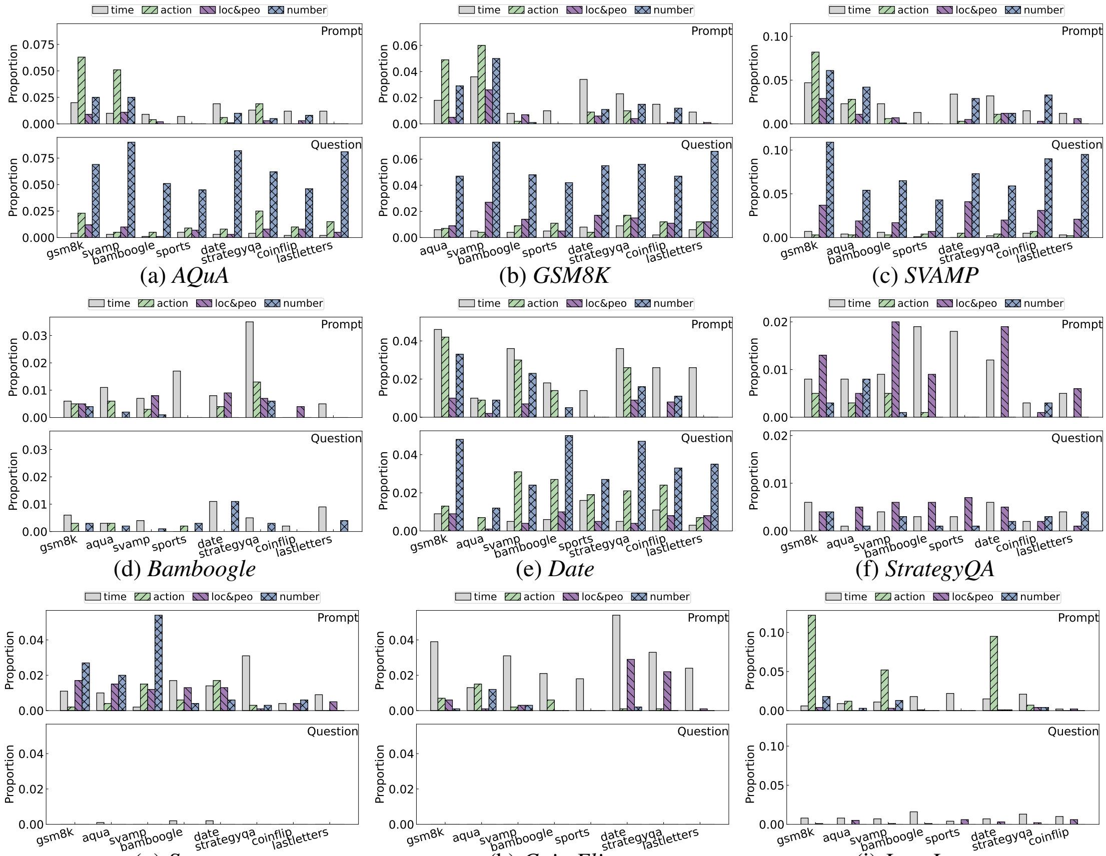
Table 6: Test points and their associated words.

D Results on Tested Datasets

Table 7: LLaMA3.1-8B

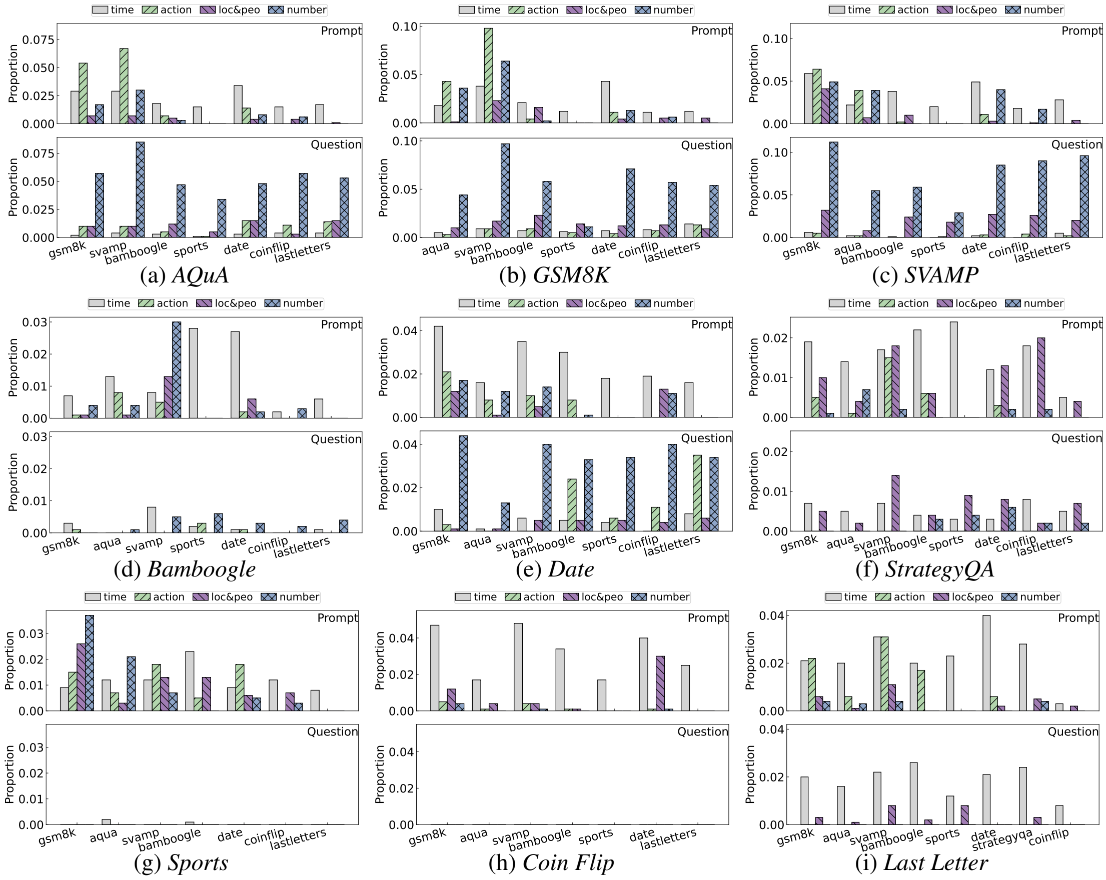
	Closed-domain			Open-domain		
Metric	AQuA	Sports	Coin Flip	GSM8K	Date	Last Letter Concat
Standard Acc	0.3110	0.7497	0.4580	0.1774	0.4417	0.0
CoT Acc	0.4961	0.9395	1.000	0.7771	0.7100	0.4496
Relative Improvement	+59.49%	+25.30%	+118.34%	+338.03%	+67.4%	$+\infty$

E Resutls of Keyword Imitation



(a) *AQuA* (b) *GSM8K* (c) *SVAMP* (d) *Bamboogle* (e) *Date* (f) *StrategyQA* (g) *Sports* (h) *Coin Flip* (i) *Last Letter*

Figure 8: Results of Gemma2-2b.

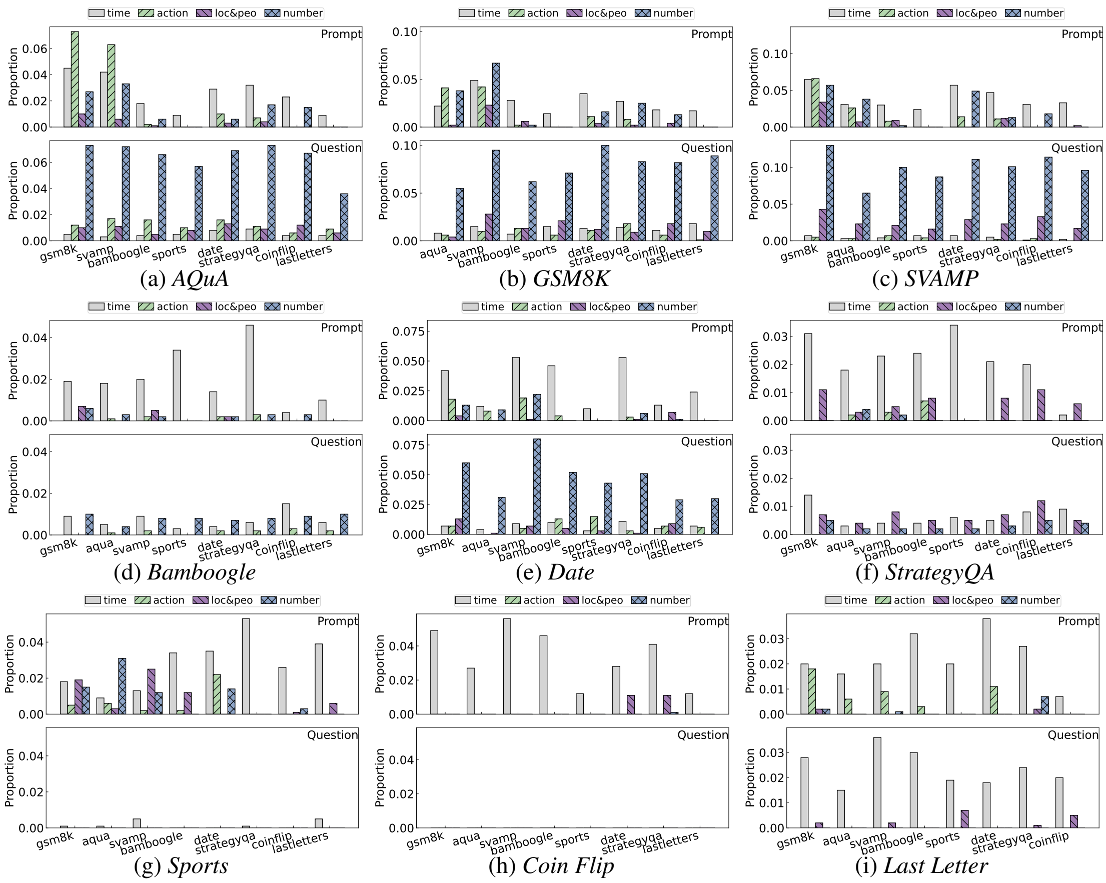


(a) *AQuA* (b) *GSM8K* (c) *SVAMP*

(d) *Bamboogle* (e) *Date* (f) *StrategyQA*

(g) *Sports* (h) *Coin Flip* (i) *Last Letter*

Figure 9: Results of Gemma2-9b.



(a) *AQuA* (b) *GSM8K* (c) *SVAMP*

(d) *Bamboogle* (e) *Date* (f) *StrategyQA*

(g) *Sports* (h) *Coin Flip* (i) *Last Letter*

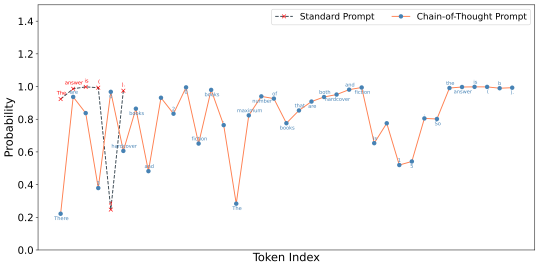
Figure 10: Results of LLaMA2-13b.

F Quantifying Structure Adherence

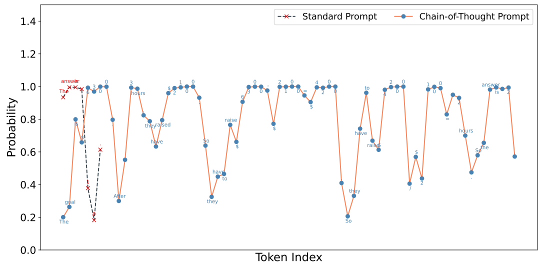
Evaluating the fidelity of the model's generated content to the CoT reasoning format involved a three-stage assessment. Initially, entities were extracted from the input prompts. The definition and extraction methodology varied by task domain: numbers for arithmetic tasks; names, locations, and temporal entities using SpaCy for commonsense benchmarks (e.g., Bamboogle, Sports), regular expressions for the Date dataset, names for symbolic reasoning like Coin Flip, and target words for the Last Letter Concatenation. Subsequently, we determined the presence of intermediate reasoning steps. For most tasks, this involved identifying the generation of new entities not present in the initial prompt. However, for tasks less prone to explicit entity generation (e.g., Coin Flip, Last Letter), a verb-based heuristic was employed: counting occurrences of key process verbs such as "flips, is, was, are, be, were". A count exceeding four for these tasks was taken as an indicator of reasoning activity. Finally, the completion of the reasoning process with a conclusive answer was verified by locating the explicit phrase "the answer is" at the terminal position of the generated text. Content was deemed to exhibit CoT adherence if it satisfied all three aforementioned conditions.

G Probability Distribution of Individual Tokens

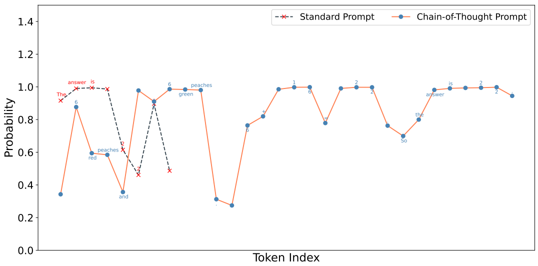
We conducted experiments on all test datasets across various models, obtaining the token sequences and their corresponding generation probability sequences under both CoT and Standard prompting methods. Figure 11 presents the test results for the Gemma2-2b model, Figure 12 shows the results for the LLaMA2-13b model, and Figure 13 illustrates the results for the Gemma2-27b model. The experimental results indicate that the probability trends for token generation under both CoT and Standard prompts are consistent across models of different sizes and series.



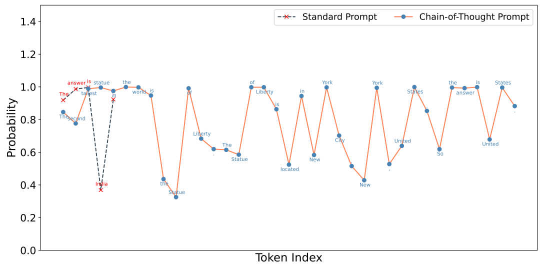
(a) *AQuA*



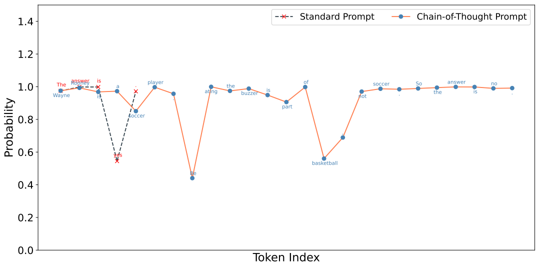
(b) *GSM8K*



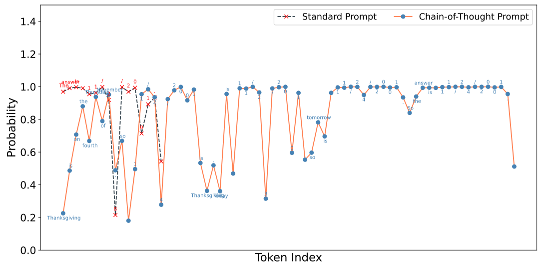
(c) *SVAMP*



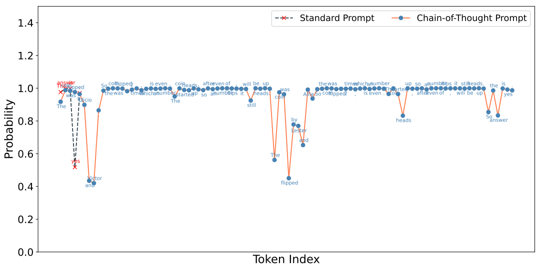
(d) *Bamboogle*



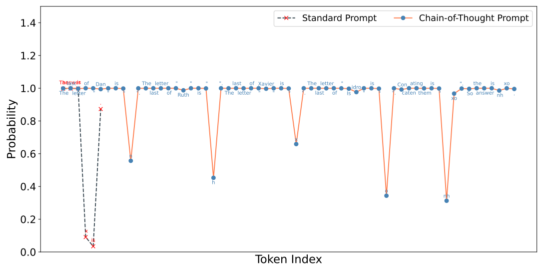
(e) *Sports*



(f) *Date*

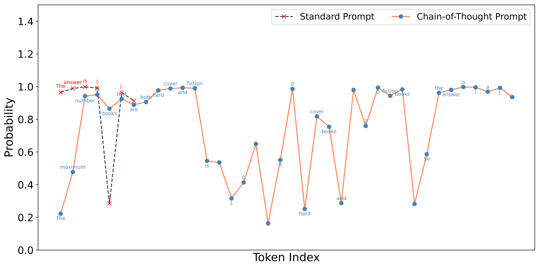


(g) *Coin Flip*

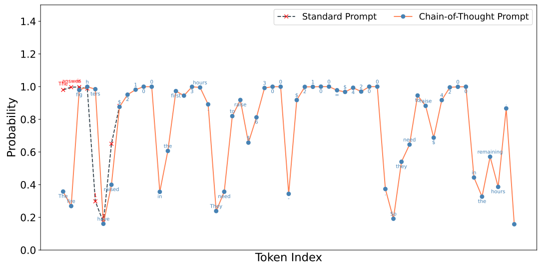


(h) *Last Letter Concatenation*

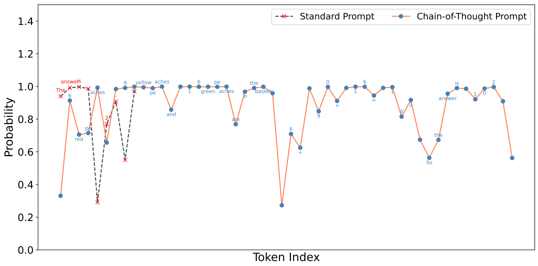
Figure 11: Probability value of each generated token (results of Gemma2-2b is shown).



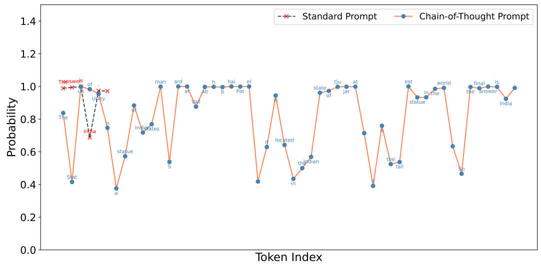
(a) *AQuA*



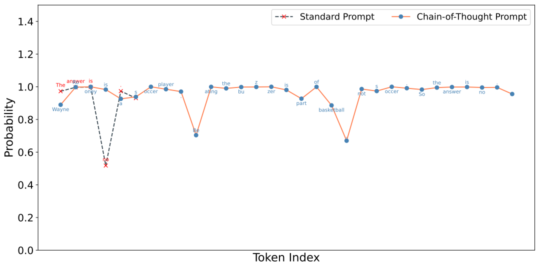
(b) *GSM8K*



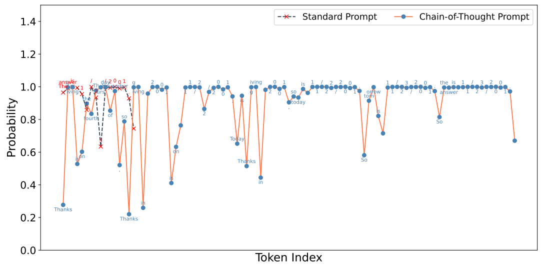
(c) *SVAMP*



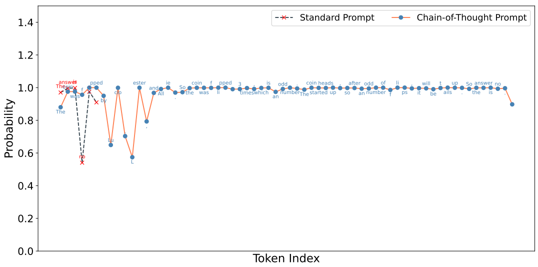
(d) *Bamboogle*



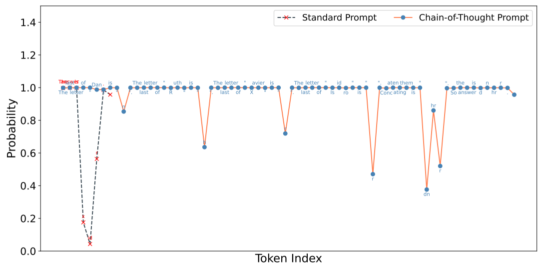
(e) *Sports*



(f) *Date*

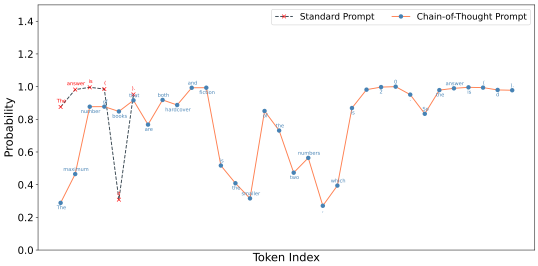


(g) *Coin Flip*

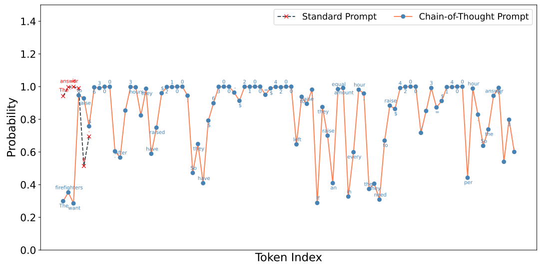


(h) *Last Letter Concatenation*

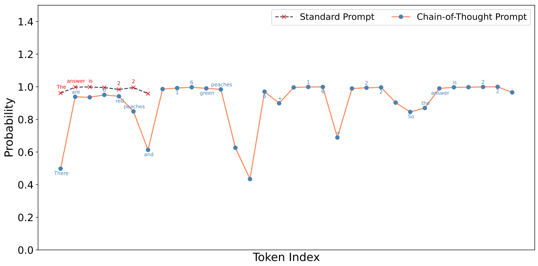
Figure 12: Probability value of each generated token (the results of LLaMA2-13b is shown).



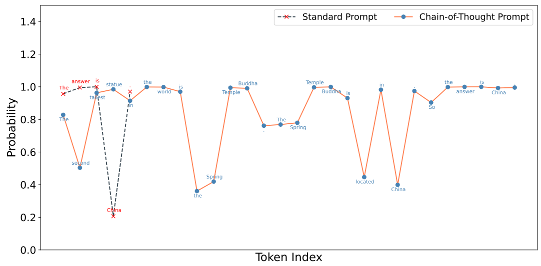
(a) *AQuA*



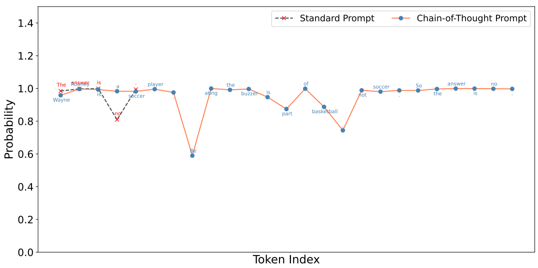
(b) *GSM8K*



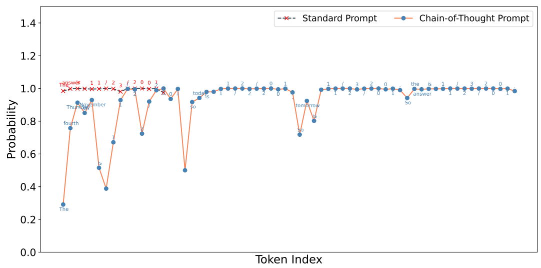
(c) *SVAMP*



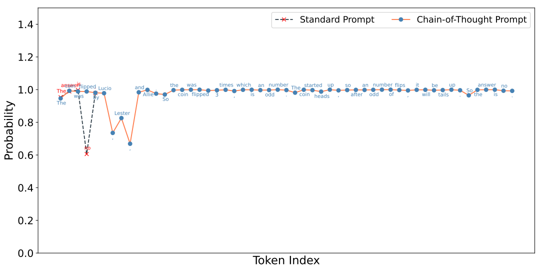
(d) *Bamboogle*



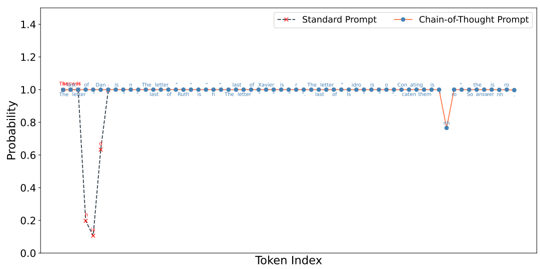
(e) *sports*



(f) *Date*



(g) *Coin Flip*



(h) *Last Letter Concatenation*

Figure 13: Probability value of each generated token (the results of Gemma2-27b is shown).

H Kernel Density Estimate

Kernel Density Estimation (KDE) is a non-parametric method for estimating the probability density function (PDF) of a dataset. It provides a smooth, continuous approximation of the underlying data distribution by summing localized kernel functions centered at each observed data point. The mathematical formulation of KDE is given by Equation 2:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2}$$

where $\hat{f}(x)$ is the estimated density at point x, n is the number of data points, h is the bandwidth (a smoothing parameter), x_i represents the observed data points, and $K(\cdot)$ is a kernel function. In this work, we employed the standard Gaussian kernel, defined as shown in Equation 3:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{3}$$

Given that our data are confined to a bounded range [0, 1], the standard Gaussian kernel was applied, and the resulting density estimates are plotted within this domain. We conducted experiments using four different models: Gemma2-2b, Gemma2-9b, LLaMA2-13b, and Gemma2-27b. The KDE plots illustrating the experimental results for these models are presented in Figure 14 (Gemma2-2b), Figure 3 (Gemma2-9b), Figure 15 (LLaMA2-13b), and Figure 16 (Gemma2-27b).

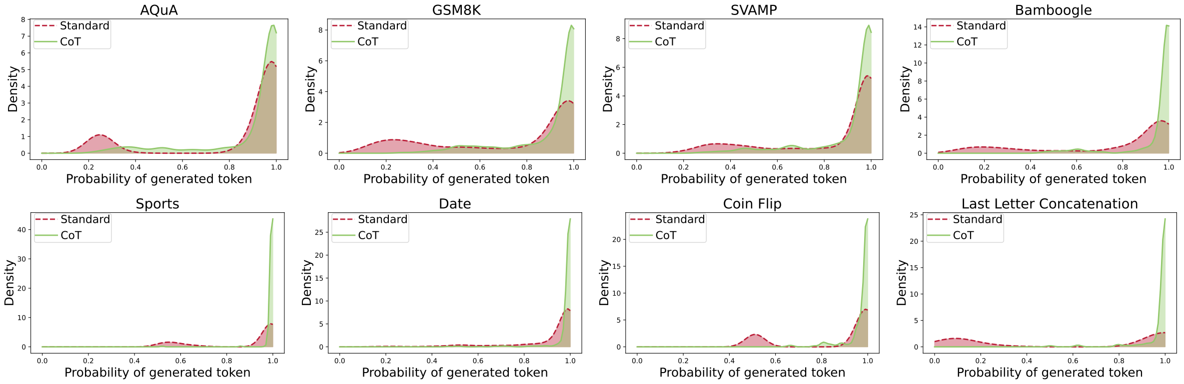


Figure 14: Kernal density of Gemma2-2B.

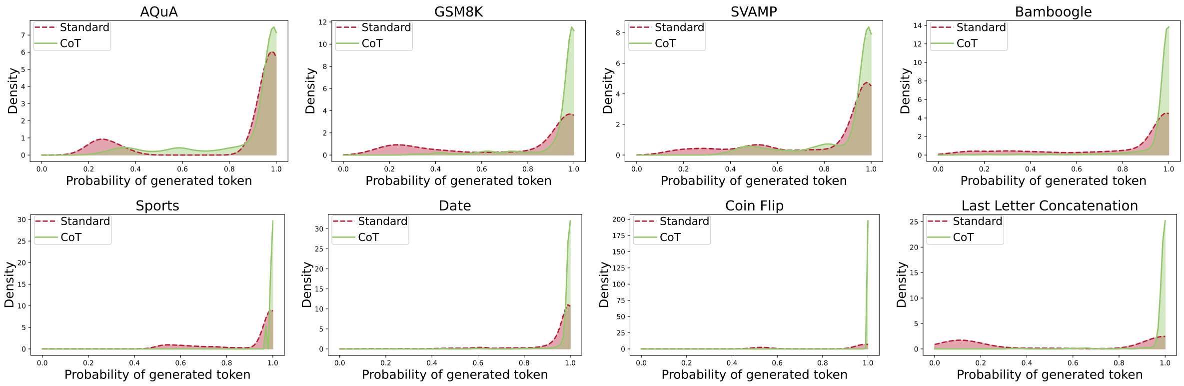


Figure 15: Kernal density of LLaMA2-13B.

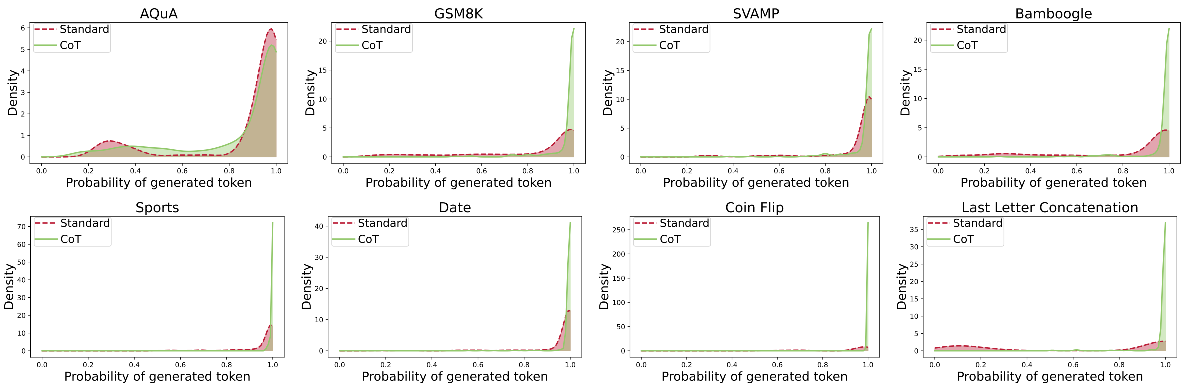
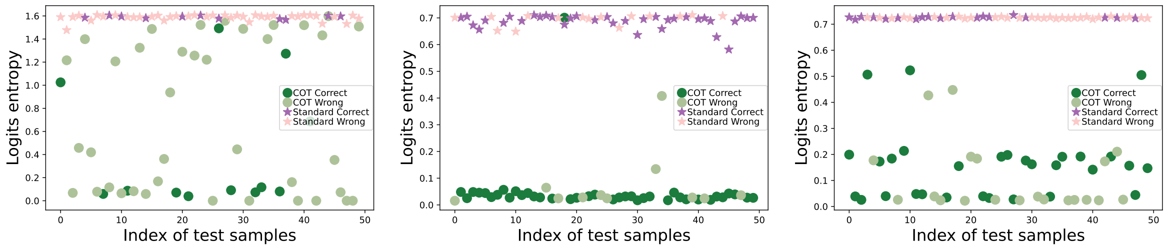


Figure 16: Kernal density of Gemma2-27B.

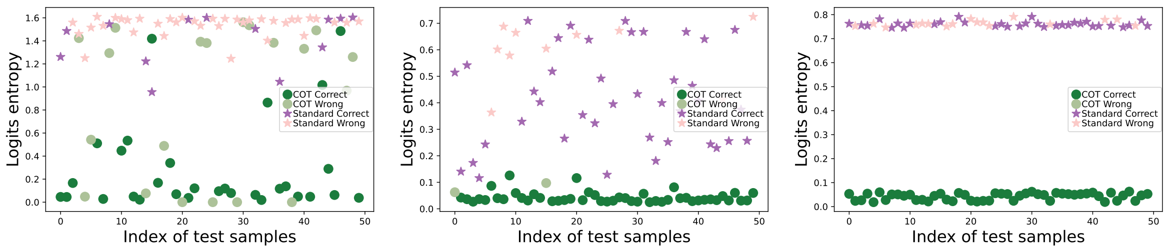
I Entropy of Probability Distribution

To investigate the model's confidence when predicting answers, we examined the probability distribution over candidate answers for the token generated at the answer prediction step. We focused on datasets AQuA, Sports, and Coin Flip, where the number of answer tokens is 1, and the answer space is finite. For datasets with multiple answer tokens and open-domain questions, it is not feasible to retrieve all potential answers and their corresponding probabilities. Surprisingly, for the AQuA dataset, the top five probabilities at the answer prediction step corresponded exactly to the answer space options: "a ", "b ", "c ", "d ", and "e". Similarly, for the Sports and Coin Flip datasets, the top two probabilities corresponded to their answer spaces, "yes" and "no". We selected the top k probabilities at the answer prediction step (where k is the size of the answer space), normalized them, calculated the entropy, and plotted a scatter plot to compare the entropy of answer probabilities generated using the Standard and CoT methods. Figure 17 shows the experimental results for Gemma2-2b, Figure 19 for LLaMA2-13b, and Figure 18 for Gemma2-9b.



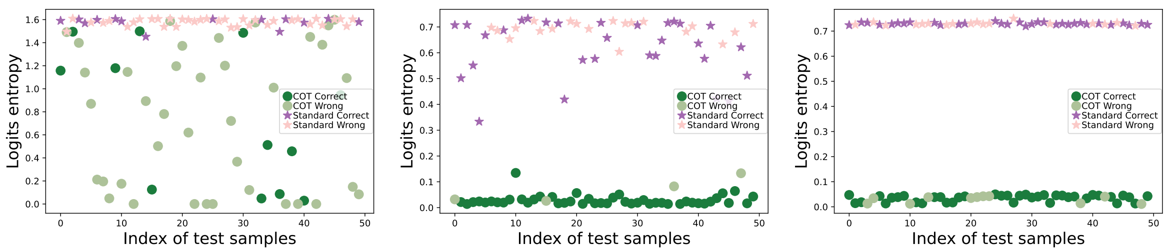
(a) *AQuA* (b) *Sports* (c) *Coin Flip*

Figure 17: Results of Gemma2-2b.



(a) *AQuA* (b) *Sports* (c) *Coin Flip*

Figure 18: Results of Gemma2-9b.



(a) *AQuA* (b) *Sports* (c) *Coin Flip*

Figure 19: Results of LLaMA2-13b.

J Activation

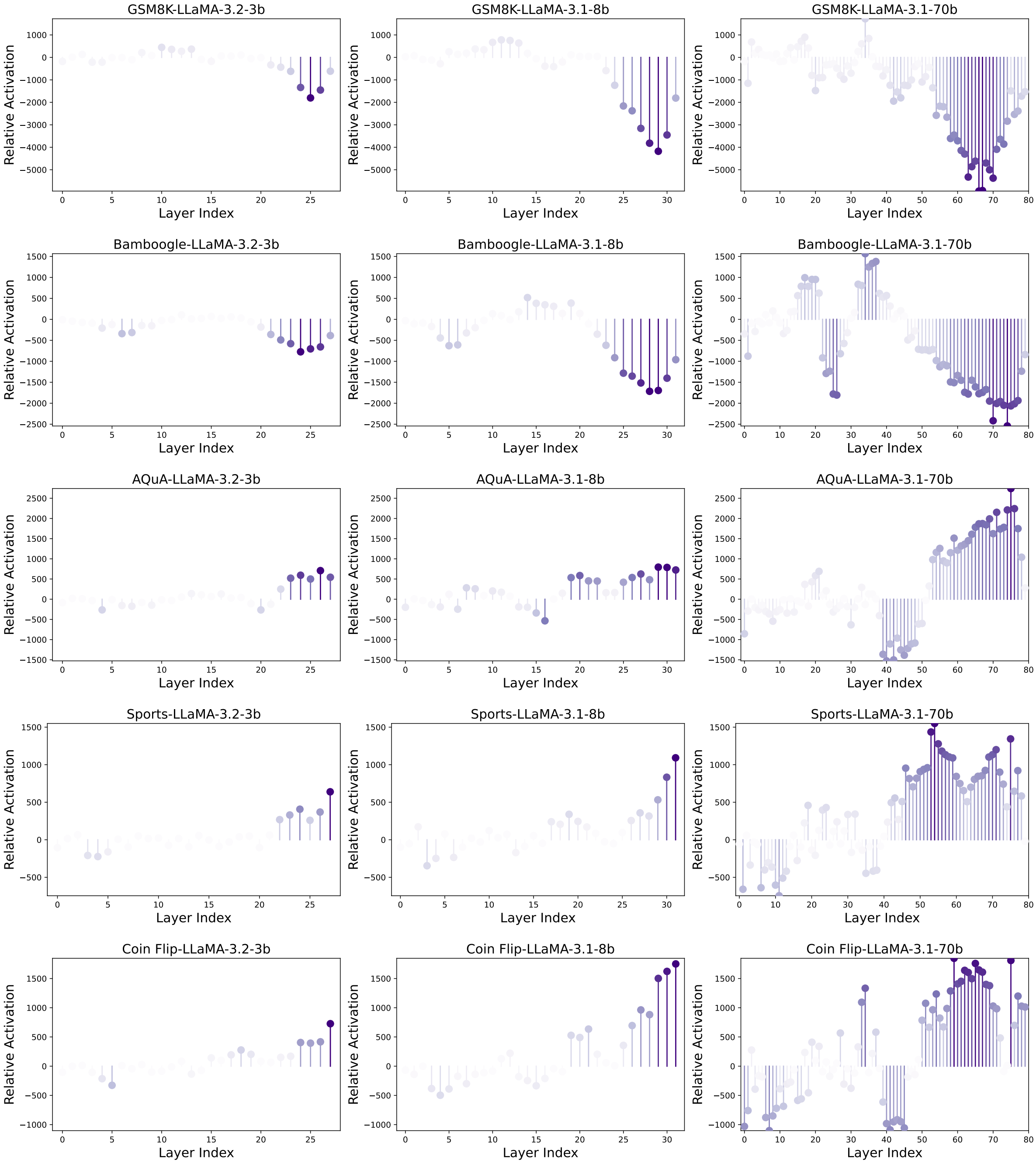
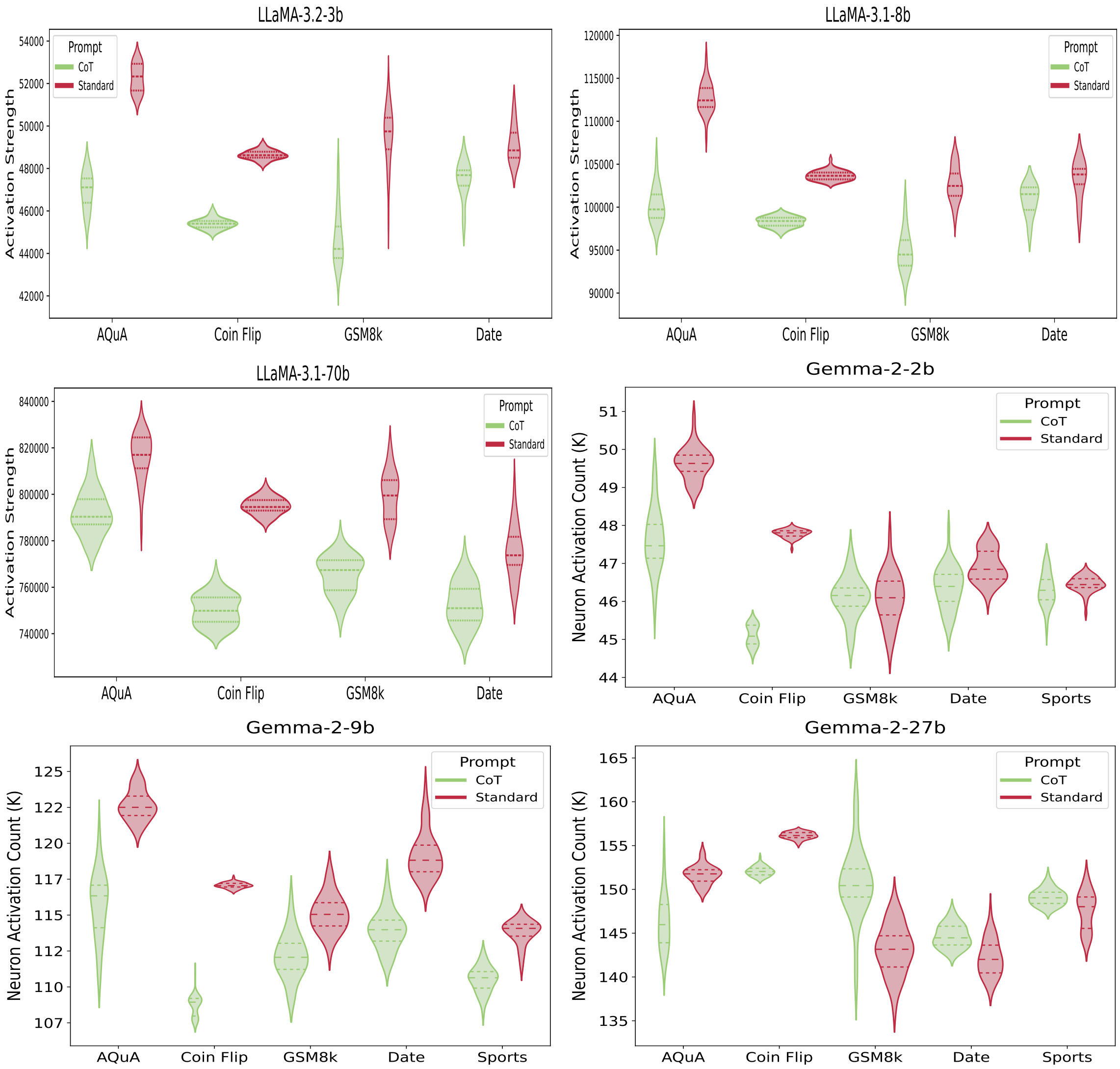


Figure 20: Layer-wise activation differences across layers for models (3B, 8B, 70B) .

K Activation Count



L Prompts Used in This Study

This study employs two primary prompt types: Standard prompts and Chain-of-Thought (CoT) prompts, adapted from prior work (Wei et al., 2022; Wang et al., 2023; Kojima et al., 2022). Specifically, the Standard and CoT prompts for the AQuA, GSM8K, SVAMP, Sports, Date, and Coin Flip datasets are derived from Wei et al. (2022), those for the Bamboogle dataset are sourced from Wang et al. (2023), and those for the Last Letter Concatenation dataset are based on Kojima et al. (2022). Minor modifications were applied to these prompts to ensure consistency; for instance, CoT prompts conclude with the phrase *"So the answer is..."* while Standard prompts use *"The answer is..."* to present the final response.

For each dataset, we provide four exemplars per prompt type. The following subsections detail the specifications of the Standard and CoT prompts for each dataset.

L.1 Standard Prompts

The Standard Prompt supplies the model with multiple question-answer pairs, enabling the model to directly generate the final answer without producing intermediate reasoning steps. The specific Standard Prompts for each dataset are presented as follows: Table 8 for the GSM8K dataset, Table 10 for the AQuA dataset, Table 11 for the SVAMP dataset, Table 9 for the Bamboogle dataset, Table 12 for the Date dataset, Table 13 for the Sports dataset, Table 14 for the Coin Flip dataset, and Table 15 for the Last Letter Concatenation dataset.

GSM8K (Standard)
Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today? **A**: The answer is 6.
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total? **A**: The answer is 39.
Q: There were nine computers in the server room. Five more computers were installed each day, from Monday to Thursday. How many computers are now in the server room? **A**: The answer is 29.
Q: Michael had 58 golf balls. On Tuesday, he lost 23 golf balls. On Wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday? **A**: The answer is 33.

Table 8: Standard Prompt exemplars for the GSM8K dataset, adapted from Wei et al. (2022).

Bamboogle (Standard)
Q: Who lived longer, Theodor Haecker or Harry Vaughan Watkins? **A**: The answer is Harry Vaughan Watkins.
Q: Why did the founder of Versus die? **A**: The answer is shot.
Q: Who is the grandchild of Dambar Shah? **A**: The answer is Rudra Shah.
Q: Are both the director of the film FAQ: Frequently Asked Questions and the director of the film The Big Money from the same country? **A**: The answer is no.

Table 9: Standard Prompt exemplars for the Bamboogle dataset, adapted from Wang et al. (2023).

AQuA (Standard)

Q: John found that the average of 15 numbers is 40. If 10 is added to each number, then the mean of the numbers is?
Answer Choices: (a) 50 (b) 45 (c) 65 (d) 78 (e) 64
A: The answer is (a).

Q: If a / b = 3/4 and 8a + 5b = 22, then find the value of a.
Answer Choices: (a) 1/2 (b) 3/2 (c) 5/2 (d) 4/2 (e) 7/2
A: The answer is (b).

Q: A person is traveling at 20 km/hr and reached his destination in 2.5 hr. Then find the distance?
Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km
A: The answer is (e).

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788
A: The answer is (b).

Table 10: Standard Prompt exemplars for the AQuA dataset, adapted from Wei et al. (2022).

SVAMP (Standard)

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: The answer is 5.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?
A: The answer is 9.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: The answer is 8.

Table 11: Standard Prompt exemplars for the SVAMP dataset, adapted from Wei et al. (2022).

Date (Standard)

Q: 2015 is coming in 36 hours. What is the date one week from today in MM/DD/YYYY?
A: The answer is 01/05/2015.

Q: The first day of 2019 is a Tuesday, and today is the first Monday of 2019. What is the date today in MM/DD/YYYY?
A: The answer is 01/07/2019.

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?
A: The answer is 05/23/1943.

Q: Jane was born on the last day of February in 2001. Today is her 16th birthday. What is the date yesterday in MM/DD/YYYY?
A: The answer is 02/27/2017.

Table 12: Standard Prompt exemplars for the Date dataset, adapted from Wei et al. (2022).

Sports (Standard)
Q: Is the following sentence plausible? "Kyle Palmieri was called for slashing." **A**: The answer is yes.
Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship." **A**: The answer is no.
Q: Is the following sentence plausible? "Carson Wentz set the pick and roll." **A**: The answer is no.
Q: Is the following sentence plausible? "Malcolm Brogdon banked the shot in." **A**: The answer is yes.

Table 13: Standard Prompt exemplars for the Sports dataset, adapted from Wei et al. (2022).

Coin Flip (Standard)
Q: A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up? **A**: The answer is yes.
Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up? **A**: The answer is no.
Q: A coin is heads up. Millicent does not flip the coin. Conception flips the coin. Is the coin still heads up? **A**: The answer is no.
Q: A coin is heads up. Ryan flips the coin. Shaunda flips the coin. Is the coin still heads up? **A**: The answer is yes.

Table 14: Standard Prompt exemplars for the Coin Flip dataset, adapted from Wei et al. (2022).

Last Letter Concatenation (Standard)
Q: Take the last letters of each word in "Tim Candace Cecil Misael" and concatenate them. **A**: The answer is mell.
Q: Take the last letters of each word in "Alina Alessandra Amina Bianca" and concatenate them. **A**: The answer is aaaa.
Q: Take the last letters of each word in "Felipe Heidi Nino Bradley" and concatenate them. **A**: The answer is eioy.
Q: Take the last letters of each word in "Lacey Nora Debra Ashleigh" and concatenate them. **A**: The answer is yaah.

Table 15: Standard Prompt exemplars for the Last Letter Concatenation dataset, adapted from Kojima et al. (2022).

L.2 Chain-of-Thought Prompts

The Chain-of-Thought Prompt incorporates intermediate reasoning steps within the exemplars provided to the model, guiding it to derive the final answer through a step-by-step process. The CoT Prompts for each dataset are presented as follows: Table 16 for the GSM8K dataset, Table 18 for the AQuA dataset, Table 17 for the SVAMP dataset, Table 19 for the Bamboogle dataset, Table 21 for the Date dataset, Table 20 for the Sports dataset, Table 22 for the Coin Flip dataset, and Table 23 for the Last Letter Concatenation dataset.

GSM8K (CoT)
Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today? **A**: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. So the answer is 6.
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total? **A**: Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. So the answer is 39.
Q: There were nine computers in the server room. Five more computers were installed each day, from Monday to Thursday. How many computers are now in the server room? **A**: There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 = 29. So the answer is 29.
Q: Michael had 58 golf balls. On Tuesday, he lost 23 golf balls. On Wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday? **A**: Michael started with 58 golf balls. After losing 23 on Tuesday, he had 58 - 23 = 35. After losing 2 more, he had 35 - 2 = 33. So the answer is 33.

Table 16: Chain-of-Thought Prompt exemplars for the GSM8K dataset, adapted from Wei et al. (2022).

SVAMP (CoT)
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot? **A**: There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. So the answer is 5.
Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny? **A**: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. So the answer is 8.
Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now? **A**: Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9. So the answer is 9.
Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left? **A**: Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 × 3 = 15 dollars. So she has 23 - 15 dollars left. 23 - 15 = 8. So the answer is 8.

Table 17: Chain-of-Thought Prompt exemplars for the SVAMP dataset, adapted from Wei et al. (2022).

AQuA (CoT)
Q: John found that the average of 15 numbers is 40. If 10 is added to each number, then the mean of the numbers is? **Answer Choices**: (a) 50 (b) 45 (c) 65 (d) 78 (e) 64 **A**: If 10 is added to each number, then the mean of the numbers also increases by 10. So the new mean would be 40 + 10 = 50. So the answer is (a).
Q: If a / b = 3/4 and 8a + 5b = 22, then find the value of a. **Answer Choices**: (a) 1/2 (b) 3/2 (c) 5/2 (d) 4/2 (e) 7/2 **A**: If a / b = 3/4, then b = 4a / 3. So 8a + 5(4a / 3) = 22. This simplifies to 8a + 20a / 3 = 22, which means 44a / 3 = 22. So a = (22 × 3) / 44 = 3/2. So the answer is (b).
Q: A person is traveling at 20 km/hr and reached his destination in 2.5 hr. Then find the distance? **Answer Choices**: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km **A**: The distance that the person traveled would be 20 km/hr × 2.5 hr = 50 km. So the answer is (e).
Q: How many keystrokes are needed to type the numbers from 1 to 500? **Answer Choices**: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788 **A**: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 × 1 + 90 × 2 + 401 × 3 = 9 + 180 + 1203 = 1392. So the answer is (b).

Table 18: Chain-of-Thought Prompt exemplars for the AQuA dataset, adapted from Wei et al. (2022).

Bamboogle (CoT)
Q: Who lived longer, Theodor Haecker or Harry Vaughan Watkins? **A**: Theodor Haecker was 65 years old when he died. Harry Vaughan Watkins was 69 years old when he died. So the answer is Harry Vaughan Watkins.
Q: Why did the founder of Versus die? **A**: The founder of Versus was Gianni Versace. Gianni Versace was shot and killed on the steps of his Miami Beach mansion on July 15, 1997. So the answer is shot.
Q: Who is the grandchild of Dambar Shah? **A**: Dambar Shah (? - 1645) was the king of the Gorkha Kingdom. He was the father of Krishna Shah. Krishna Shah (? - 1661) was the king of the Gorkha Kingdom. He was the father of Rudra Shah. So the answer is Rudra Shah.
Q: Are both the director of the film FAQ: Frequently Asked Questions and the director of the film The Big Money from the same country? **A**: The director of the film FAQ: Frequently Asked Questions is Carlos Atanes, who is Spanish. The director of the film The Big Money is John Paddy Carstairs, who is British. So the answer is no.

Table 19: Chain-of-Thought Prompt exemplars for the Bamboogle dataset, adapted from Wang et al. (2023).

Sports (CoT)
Q: Is the following sentence plausible? “Kyle Palmieri was called for slashing.” **A**: Kyle Palmieri is a hockey player. Being called for slashing is a common penalty in hockey. So the answer is yes.
Q: Is the following sentence plausible? “Joao Moutinho caught the screen pass in the NFC championship.” **A**: Joao Moutinho is a soccer player. The NFC championship is an American football event, and catching a screen pass is specific to football, not soccer. So the answer is no.
Q: Is the following sentence plausible? “Carson Wentz set the pick and roll.” **A**: Carson Wentz is an American football player. The pick and roll is a basketball play, not used in football. So the answer is no.
Q: Is the following sentence plausible? “Malcolm Brogdon banked the shot in.” **A**: Malcolm Brogdon is a basketball player. Banking the shot in is a common basketball term. So the answer is yes.

Table 20: Chain-of-Thought Prompt exemplars for the Sports dataset, adapted from Wei et al. (2022).

Date (CoT)
Q: 2015 is coming in 36 hours. What is the date one week from today in MM/DD/YYYY? **A**: If 2015 is coming in 36 hours, then it is coming in 2 days. 2 days before 01/01/2015 is 12/30/2014, so today is 12/30/2014. So one week from today will be 01/05/2015. So the answer is 01/05/2015.
Q: The first day of 2019 is a Tuesday, and today is the first Monday of 2019. What is the date today in MM/DD/YYYY? **A**: If the first day of 2019 was Tuesday, then 01/01/2019 was a Tuesday. Today is the first monday, would be six days later. So today is 01/07/2019. So the answer is 01/07/2019.
Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY? **A**: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.
Q: Jane was born on the last day of Feburary in 2001. Today is her 16-year-old birthday. What is the date yesterday in MM/DD/YYYY? **A**: The last day of February is the 28th, so Jane was born on 02/28/2001. Today is her 16-year old birthday, so today is 02/28/2017. So yesterday was 02/27/2017. So the answer is 02/27/2017.

Table 21: Chain-of-Thought Prompt exemplars for the Date dataset, adapted from Wei et al. (2022).

Coin Flip (CoT)
Q: A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up? **A**: The coin was flipped by Ka and Sherrie. So the coin was flipped 2 times, which is an even number. The coin started heads up, so after an even number of flips, it will still be heads up. So the answer is yes.
Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up? **A**: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.
Q: A coin is heads up. Millicent does not flip the coin. Conception flips the coin. Is the coin still heads up? **A**: The coin was flipped by Conception. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.
Q: A coin is heads up. Ryan flips the coin. Shaunda flips the coin. Is the coin still heads up? **A**: The coin was flipped by Ryan and Shaunda. So the coin was flipped 2 times, which is an even number. The coin started heads up, so after an even number of flips, it will still be heads up. So the answer is yes.

Table 22: Chain-of-Thought Prompt exemplars for the Coin Flip dataset, adapted from Wei et al. (2022).

Last Letter Concatenation (CoT)
Q: Take the last letters of each words in "Tim Candace Cecil Misael" and concatenate them. **A**: The last letter of "Tim" is "m". The last letter of "Candace" is "e". The last letter of "Cecil" is "l". The last letter of "Misael" is "l". Concatenating them is "mell". So the answer is mell.
Q: Take the last letters of each words in "Alina Alessandra Amina Bianca" and concatenate them. **A**: The last letter of "Alina" is "a". The last letter of "Alessandra" is "a". The last letter of "Amina" is "a". The last letter of "Bianca" is "a". Concatenating them is "aaaa". So the answer is aaaa.
Q: Take the last letters of each words in "Felipe Heidi Nino Bradley" and concatenate them. **A**: The last letter of "Felipe" is "e". The last letter of "Heidi" is "i". The last letter of "Nino" is "o". The last letter of "Bradley" is "y". Concatenating them is "eioy". So the answer is eioy.
Q: Take the last letters of each words in "Lacey Nora Debra Ashleigh" and concatenate them. **A**: The last letter of "Lacey" is "y". The last letter of "Nora" is "a". The last letter of "Debra" is "a". The last letter of "Ashleigh" is "h". Concatenating them is "yaah". So the answer is yaah.

Table 23: Chain-of-Thought Prompt exemplars for the Last Letter Concatenation dataset, adapted from Kojima et al. (2022).