Modeling Chain-of-Thought Degradation in Pruned Language Models: Fidelity and Similarity Analysis for Mathematical Reasoning

Anonymous submission

Abstract

Efficient mathematical reasoning under compute and memory constraints is crucial for deploying large reasoning models (LRMs) in real-world applications, especially in resource-constrained environments. We propose a framework to quantify the relationship between model similarity and loss of reasoning fidelity in chain-of-thought (CoT) outputs under pruning. Our approach introduces ASAND, a similarity metric that integrates centered alignment, sparsity-aware structural measures, and adaptive exponential decay to capture non-monotonic changes in reasoning fidelity, a behavior not captured by traditional pruning metrics. Experiments on the Qwen-0.5B model using the GSM8K dataset demonstrate that moderate pruning can improve CoT reasoning performance, while excessive sparsity leads to catastrophic collapse. Correlation analyses show that ASAND outperforms standard similarity metrics, providing the highest predictive power for reasoning fidelity degradation. These findings offer valuable insights for compression-aware deployment of LRMs, enabling efficient reasoning on resource-constrained devices without sacrificing accuracy. To further validate ASAND, we extend our analysis to competition-level mathematical reasoning on the MATH dataset, demonstrating its broader applicability to more complex problem domains.

Keywords: Chain-of-Thought Reasoning, Mathematical Reasoning, Language Models, Model Fidelity, Similarity Metrics, Model Compression, Pruning

Introduction and Related Work

The representational geometry of deep neural networks (DNNs) plays a crucial role in their computational capabilities. However, model compression, specifically pruning, disrupts these geometric structures in ways that are not always predictable. For instance, Convolutional Neural Networks (CNNs) have been shown to tolerate pruning up to 50% sparsity without significant performance degradation (Han et al. 2015; Shinde 2024). In contrast, language models (LMs), particularly transformer-based models, experience catastrophic drops in performance when pruned even at 5% sparsity (Shinde 2024), highlighting the complexity of pruning in models with dense interdependencies like transformers.

Representational Geometry. The internal representations of neural networks reside on high-dimensional manifolds, with each layer mapping input data into increasingly abstract representations. This "geometry" governs the model's ability to generalize, and understanding how pruning impacts these representations is key to improving compression techniques. Methods such as Centered Kernel Alignment (CKA) (Kornblith et al. 2019) and Singular Vector Canonical Correlation Analysis (SVCCA) (Raghu et al. 2017) are commonly used to measure the similarity of neural representations. However, these techniques rely on the assumption of smooth transformations between layers, which fail to account for the discrete phase transitions in the manifold caused by pruning, particularly under conditions of high sparsity.

Pruning and Transformers. While pruning techniques such as magnitude-based pruning (Han et al. 2015; Shinde 2025) and structured pruning (Li et al. 2016) have been effective in CNNs, transformers present a unique challenge. The attention mechanism in transformers creates a globally interconnected set of weights, where changes to one part of the model can have significant ripple effects across the entire network. This interconnectedness makes transformers highly sensitive to pruning, where even minor weight changes can disrupt reasoning fidelity, leading to severe performance degradation.

Faithfulness as Geometric Invariance. In chain-of-thought reasoning tasks, the consistency of a model's reasoning trajectory is critical. Here, "faithfulness" refers to the ability of the model to maintain logical consistency during reasoning, even when subjected to perturbations such as pruning. In contrast to classification tasks, where performance is typically evaluated in terms of decision boundaries, reasoning tasks demand that these trajectories remain intact. This leads us to propose ASAND, an adaptive metric that quantifies how pruning-induced changes in weight structure correlate with disruptions in reasoning fidelity, effectively tracking these non-linear transitions in the model's representational geometry.

Transformer Sensitivity. Previous work has shown that each attention head in a transformer model tends to capture distinct subspaces of the input (Clark et al. 2019). Pruning disrupts the coordination between these subspaces, leading to abrupt decreases in reasoning accuracy. Our experiments reveal that pruning as little as 5% of the weights in Qwen-0.5B leads to a dramatic 49.5% drop in faithfulness, high-

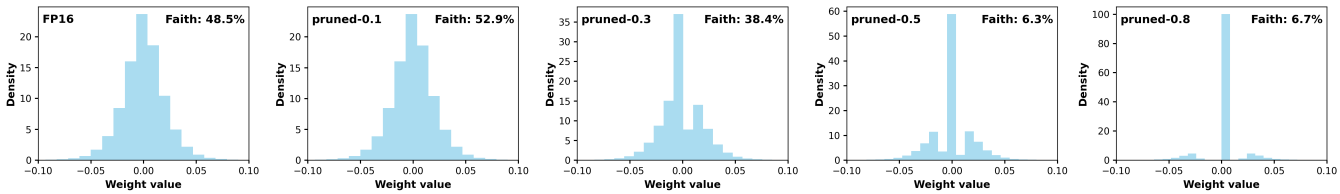

Figure 1: Weight distributions of the original model ORG (M_0) in FP16 and pruned variants with 10%, 30%, 50%, and 70% pruning ratio's. The discontinuous shift in weight geometry aligns with observed faithfulness drops, illustrating how small perturbations trigger large representational collapses.

lighting the sensitivity of transformers to pruning and the need for more robust pruning strategies.

Contributions. In this work, we make the following contributions: (i) We document the non-monotonic relationship between pruning sparsity and reasoning faithfulness, showing that initial improvements in reasoning consistency are followed by catastrophic collapses at higher sparsity levels; (ii) We introduce the ASAND metric, a novel approach that quantifies the correlation between weight changes and reasoning degradation by accounting for both structural sensitivity and magnitude-aware weighting; (iii) We demonstrate that maintaining the integrity of continuous weight manifolds is crucial for transformer-based reasoning, and that pruning must be carefully balanced to preserve this structure. These findings provide a foundation for developing geometry-aware compression techniques that can maintain reasoning fidelity while optimizing for computational efficiency in large-scale language models.

Method

Problem Description

Let $(\mathcal{X}, \mathcal{Y})$ represent the input-label spaces for complex reasoning tasks, and let P be a distribution over $\mathcal{X} \times \mathcal{Y}$. A reference language model M_0 with parameters θ_0 implements a measurable mapping $f_{M_0} : \mathcal{X} \to \mathcal{T}$, where $\mathcal{T}$ represents the generated text or reasoning trace space. We evaluate models on a dataset $S = \{(x_i, y_i)\}_{i=1}^n$, drawn independently and identically from P. This setup enables us to systematically study the impact of parameter sparsity on reasoning fidelity.

Pruning Operator. We apply L1 unstructured pruning to the model M_0, producing a pruned model M_p with sparsity ratio $\lambda \in [0, 1]$. Formally, the pruning operator is defined as:

$$\mathcal{P}(\cdot; \lambda) : \mathcal{M} \to \mathcal{M}, \quad M_{p,\lambda}\mathcal{P}(M_0; \lambda), \tag{1}$$

which removes weights with the smallest L1 norm across all linear layers. We choose L1 pruning due to its simplicity and proven effectiveness for preserving reasoning capabilities while reducing computation and memory usage. As shown in Fig. 1, pruning progressively reshapes the weight distributions, with initial sparsity introducing sharp zero-centered discontinuities. These discontinuities align with the observed non-monotonic faithfulness behavior, where light pruning removes redundant parameters, but higher sparsity induces catastrophic representational collapse.

Faithfulness Metric and Degradation. We quantify a model's faithfulness by evaluating key reasoning components in its output: numerical consistency, logical connectors, cue phrases, step completeness, and answer alignment. Each component f_k is normalized and combined into a weighted score:

$$F(q, r) = \sum_{k=1}^{5} w_k \cdot f_k(q, r), \tag{2}$$

where w_k reflects the relative importance of each reasoning aspect. Faithfulness degradation due to pruning is computed as:

$$\Delta F_\lambda = F(M_0) - F(M_{p,\lambda}) \in [-1, 1]. \tag{3}$$

The score ΔF_λ quantifies the loss of reasoning fidelity, capturing the extent to which pruning disrupts the model's logical consistency.

Model Similarity. To quantify the effect of pruning, we measure the similarity between baseline and pruned weights using various similarity metrics such as cosine similarity, L1/L2 distances, and linear Centered Kernel Alignment (CKA). These metrics capture both magnitude and structural changes, allowing us to assess how weight modifications influence reasoning performance. Let $\Theta(M) = \{\theta_j\}_{j=1}^{|\theta|}$ denote the flattened parameters from linear layers. For baseline M_0 and pruned $M_{p,\lambda}$, we define weight vectors $\mathbf{w}_0, \mathbf{w}_p \in \mathbb{R}^{|\theta|}$. Standard similarity measures include:

$$s_{\cos}(M_0, M_{p,\lambda}) = \frac{\langle \mathbf{w}_0, \mathbf{w}_p \rangle}{\|\mathbf{w}_0\|_2 \|\mathbf{w}_p\|_2}, \tag{4}$$

$$d_{L_2}(M_0, M_{p,\lambda}) = \|\mathbf{w}_0 - \mathbf{w}_p\|_2, \tag{5}$$

$$d_{L_1}(M_0, M_{p,\lambda}) = \|\mathbf{w}_0 - \mathbf{w}_p\|_1. \tag{6}$$

Linear CKA. To assess structural alignment between pruned and unpruned weight matrices, we apply the Centered Kernel Alignment (CKA) metric. Let the centered weight matrices be $\tilde{W}_0 = W_0 - \bar{W}_0$ and $\tilde{W}_p = W_p - \bar{W}_p$, where $\bar{W}_0$ and $\bar{W}_p$ are the column means. The CKA score is defined as:

$$\text{CKA}(M_0, M_{p,\lambda}) = \frac{\text{HSIC}(\tilde{W}_0, \tilde{W}_p)}{\sqrt{\text{HSIC}(\tilde{W}_0, \tilde{W}_0) \cdot \text{HSIC}(\tilde{W}_p, \tilde{W}_p)}}, \tag{7}$$

where $\text{HSIC}(X, Y) = \text{tr}(XY^\top)^2$ is the Hilbert-Schmidt Independence Criterion (HSIC) that measures dependencies between the columns of X and Y.

Similarity-Faithfulness Coupling. We aim to quantify how well similarity metrics s_λ between M_0 and $M_{p,\lambda}$ predict ΔF_λ across sparsity levels λ. The correlation between similarity and faithfulness degradation is measured using Pearson (PLCC), Spearman (SRCC), and Kendall (KRCC) correlation coefficients over the sets $\{s_\lambda\}_{\lambda\in[0,1]}$ and $\{\Delta F_\lambda\}_{\lambda\in[0,1]}$.

Proposed Model Quality Metric: ASAND

Faithfulness degradation exhibits non-monotonic trends under pruning, a behavior often overlooked by standard pruning metrics. To address this, we propose the *Adaptive Sparsity-Adjusted Normalized Distance* (ASAND), designed to robustly correlate weight changes with reasoning fidelity loss, particularly capturing non-linear degradation patterns.

ASAND Components. ASAND integrates multiple complementary components to capture pruning-induced changes in reasoning models. Each component plays a crucial role in detecting specific aspects of the degradation process, ensuring that all facets of reasoning fidelity are addressed. The components are: ASAND computes a similarity score $s_{\text{ASAND}}(M_0, M_{p,\lambda}, \lambda)$ as a weighted combination of five components operating on flattened weight vectors $\mathbf{w}_0, \mathbf{w}_p \in \mathbb{R}^{|\theta|}$.

Centered Alignment. This component measures the similarity of representations after centering the weight vectors. Inspired by the Centered Kernel Alignment (CKA), it captures the alignment between pruned and original weight distributions:

$$s_{\text{cent}}(\mathbf{w}_0, \mathbf{w}_p) = \frac{\langle \mathbf{w}_0 - \bar{\mathbf{w}}_0, \mathbf{w}_p - \bar{\mathbf{w}}_p \rangle}{\|\mathbf{w}_0 - \bar{\mathbf{w}}_0\|_2 \|\mathbf{w}_p - \bar{\mathbf{w}}_p\|_2}, \quad s_{\text{cent}} \in [0, 1]. \tag{8}$$

where $\bar{\mathbf{w}}_0$ and $\bar{\mathbf{w}}_p$ represent the mean values of $\mathbf{w}_0$ and $\mathbf{w}_p$, respectively.

Jaccard Sparsity Similarity. This metric quantifies structural similarity based on the proportion of non-zero weights in the model. It compares the sparsity patterns of pruned and unpruned weight vectors:

$$\begin{aligned} s_{\text{jacc}}(\mathbf{w}_0, \mathbf{w}_p) &= 1 - \frac{|\text{nz}(\mathbf{w}_0) - \text{nz}(\mathbf{w}_p)|}{\max(\text{nz}(\mathbf{w}_0), \text{nz}(\mathbf{w}_p))}, \\ \text{nz}(\mathbf{w}) &= \frac{|\{w_i : |w_i| > 10^{-6}\}|}{|\mathbf{w}|}. \end{aligned} \tag{9}$$

where $\text{nz}(\mathbf{w})$ represents the normalized count of non-zero weights in $\mathbf{w}$.

Adaptive Exponential Decay Distance (AEDD). The AEDD models the non-linear degradation in reasoning fidelity under pruning, with a sparsity-dependent scaling factor. This captures how pruning disrupts the model's representational capacity:

$$\begin{aligned} d_{\text{AEDD}}(\mathbf{w}_0, \mathbf{w}_p, \lambda) &= \exp\left(-\sigma(\lambda) \cdot \frac{\|\mathbf{w}_0 - \mathbf{w}_p\|_2}{\|\mathbf{w}_0\|_2}\right), \\ \sigma(\lambda) &= \begin{cases} 0.8 \times 1.5 & \text{if } \lambda > 0.3 \\ 0.8 & \text{otherwise} \end{cases}. \end{aligned} \tag{10}$$

The decay factor $\sigma(\lambda)$ is tuned to reflect the heightened impact of pruning at higher sparsity levels.

Volatility Similarity. This component measures the stability of weight distributions after pruning. It quantifies the relative change in the standard deviation of weight distributions between the original and pruned models:

$$s_{\text{vol}}(\mathbf{w}_0, \mathbf{w}_p) = \exp\left(-\frac{|\sigma_{\mathbf{w}_0} - \sigma_{\mathbf{w}_p}|}{\sigma_{\mathbf{w}_0}}\right), \quad \sigma_{\mathbf{w}} = \text{std}(\mathbf{w}). \tag{11}$$

where $\sigma\mathbf{w}$ is the standard deviation of the weights in vector $\mathbf{w}$.

Low-Pruning Gain Booster. This metric is designed to reward improvements in reasoning fidelity achieved through low sparsity pruning. It detects small, yet meaningful changes in the weight vector for low pruning levels:

$$g(\mathbf{w}_0, \mathbf{w}_p, \lambda) = \begin{cases} 0.1 \cdot \left(1 - \frac{\|\mathbf{w}_0 - \mathbf{w}_p\|_2}{\|\mathbf{w}_0\|_2}\right) & \text{if } \lambda < 0.1 \\ \text{and } \|\mathbf{w}_0 - \mathbf{w}_p\|_2 < 0.1\|\mathbf{w}_0\|_2 & \\ 0 & \text{otherwise} \end{cases}. \tag{12}$$

ASAND Score. The final ASAND score aggregates the five components, each weighted according to its importance in capturing the degradation of reasoning fidelity due to pruning:

$$s_{\text{ASAND}} = \sum_i \alpha_i \cdot \text{Component}_i, \tag{13}$$

where the weights α_i are chosen to maximize the correlation with the observed faithfulness degradation $\Delta F\lambda$.

ASAND Calculation. The ASAND score is computed as a weighted combination of the components, operating on flattened weight vectors $\mathbf{w}_0, \mathbf{w}_p \in \mathbb{R}^{|\theta|}$:

$$\begin{aligned} s_{\text{ASAND}}(\mathbf{w}_0, \mathbf{w}_p, \lambda) = {} & w_b \cdot [d_{\text{AEDD}} \cdot s_{\text{jacc}} \cdot s_{\text{cent}}] + w_v \cdot s_{\text{vol}} \\ & + w_t \cdot [d_{\text{AEDD}} \text{ if } \lambda > 0.3 \text{ else } 1] + w_g \cdot g \end{aligned} \tag{14}$$

where the weights are $w_b = 0.4$, $w_v = 0.25$, $w_t = 0.2$, and $w_g = 0.15$. These weights are tuned to optimize the correlation with the faithfulness degradation ΔF_λ. Each component of ASAND is essential for capturing a different aspect of pruning-induced changes in model behavior, ensuring that the metric robustly reflects the effects of pruning on reasoning fidelity across various sparsity levels.

Efficiency and Robustness. ASAND operates on flattened weights with $O(|\theta|)$ complexity, making it computationally efficient for models like Qwen-0.5B ($|\theta| \approx 500$M). All components are normalized to $[0, 1]$, ensuring robustness across model sizes and sparsity levels. This enables rapid, deployable reasoning evaluation under tight computational constraints.

Experimental Setup

Dataset. Experiments are conducted on two datasets: GSM8K (Cobbe et al. 2021) and MATH (Hendrycks et al. 2021). GSM8K consists of $8,792$ grade-school mathematical reasoning problems, making it particularly suitable for evaluating multi-step reasoning under resource constraints.

The tasks in GSM8K require both arithmetic computation and logical step-by-step deductions, which makes faithfulness metrics meaningful proxies for reasoning fidelity. For our ablation studies, we evaluate the model on test splits $S_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{n}$, where $n \in \{5, 50, 200\}$, allowing us to assess the model's performance across varying levels of problem difficulty.

To comprehensively evaluate the pruning effects across a range of mathematical complexity, we extend our evaluation to the MATH dataset (Hendrycks et al. 2021). The MATH dataset consists of $12,500$ problems from mathematical competitions (AMC, AIME), providing a critical testbed for understanding how pruning impacts reasoning as problem complexity increases. This dual-dataset approach reveals that reasoning sensitivity to pruning increases superlinearly with mathematical complexity, making it an ideal platform to study the scalability of pruning effects in reasoning tasks.

Model Architecture. We adopt the Qwen-0.5B-Instruct (Yang et al. 2024), a 494M parameter transformer model with 24 layers, a hidden dimension of 1024, and 16 attention heads. This architecture strikes a balance between reasoning capacity and computational efficiency, making it ideal for pruning analyses in low-latency, memory-constrained environments. Our analysis indicates that this methodology can also be applied to other models like Qwen2.5-1.5B-Instruct and TinyLlama-1.1B-Chat-v1.0, which have comparable parameter sizes and structures, enabling the extension of our findings to larger models in the future.

Training Implementation Details. All experiments are implemented in PyTorch on NVIDIA Tesla P100 GPUs (16GB) available via Kaggle, with models loaded in `torch.float16` precision for efficient memory usage. L1 unstructured pruning is applied using `prune.l1_unstructured` on all linear layers, followed by the permanent removal of pruned weights via `prune.remove`. We fix the random seed at 42 to ensure the reproducibility of results. This setup isolates the effect of sparsity on reasoning fidelity, minimizing the impact of potential training variability.

Hyperparameter Settings. We set the faithfulness component weights as $\mathbf{w} = (0.30, 0.20, 0.20, 0.15, 0.15)$ to balance contributions from five reasoning aspects: numerical consistency, logical connectors, cue phrases, step completeness, and answer alignment. These weights were chosen based on preliminary experiments to reflect the relative importance of each component in maintaining reasoning fidelity under pruning. The logic word threshold is set to 3, and step coherence is normalized by 3 expected steps, ensuring consistency across experiments and enhancing the interpretability of the faithfulness score ΔF_λ.

Pruning Setup. L1 magnitude pruning (Han et al. 2015) is applied with sparsity ratios: $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, which enables us to evaluate pruning at different levels of sparsity and assess its impact on reasoning performance. No fine-tuning is performed post-pruning, ensuring that any observed changes in reasoning fidelity are directly attributed to weight sparsity rather than changes in model parameters due to retraining.

Evaluation Protocols. We assess reasoning robustness using two prompting strategies:

Chain-of-Thought (CoT): The prompt structure "Solve this step by step: Question: $\{q\}$ Step 1:" is used with a 150-token limit to evaluate stepwise reasoning fidelity. This approach allows us to analyze how well the model can reason through complex tasks in a sequential, logical manner.

Direct Answer: The prompt "Question: $\{q\}$ Answer:" is used with a 30-token limit to assess overall solution accuracy. This evaluates the model's ability to provide direct solutions, independent of the reasoning steps involved.

By contrasting these two prompting strategies, we can isolate the effects of pruning on different types of reasoning: stepwise reasoning (via CoT) versus overall solution accuracy (via Direct Answer).

Evaluation Metrics. Models are evaluated using several metrics:

Faithfulness Score. $F \in [0, 1]$ to measure reasoning fidelity.

Faithfulness Drop. $\Delta F_\lambda = F(M_0) - F(M_{p,\lambda})$ to quantify the degradation in reasoning performance due to pruning.

Similarity Metrics, including cosine similarity, L_2 and L_1 distances, and linear CKA to measure structural alignment and magnitude changes between the original and pruned models.

Correlations between similarity metrics and faithfulness drop are quantified using Pearson (PLCC), Spearman (SRCC), and Kendall (KRCC) rank correlations, providing insights into how pruning-induced weight changes affect reasoning fidelity.

We also visualize weight distributions using 256-bin histograms over the range $[-0.1, 0.1]$ to provide intuitive insights into pruning effects on model parameters. This visualization complements our quantitative metrics, offering a deeper understanding of how weight sparsity influences the model's ability to reason.

Results and Discussion

We evaluate the performance of the Qwen-0.5B model on the GSM8K and MATH datasets under varying unstructured pruning ratios and analyze how different similarity metrics capture the degradation in reasoning fidelity, or *faithfulness*, caused by pruning.

Faithfulness and Efficiency Analysis. Table 1 summarizes key performance metrics, including faithfulness scores (F), memory usage, runtime, and token throughput across different pruning ratios. Faithfulness scores F are computed on a scale of $[0, 1]$, where higher values represent better reasoning consistency. Light pruning (with sparsity ratios between $1\% - 5\%$) leads to an increase in CoT faithfulness from 0.637 to 0.723. This suggests that removing redundant or noisy weights helps the model focus on essential logical reasoning steps, improving its ability to maintain consistency throughout the process. However, as sparsity increases beyond 30%, faithfulness drops sharply, with F plummeting to 0.087 at 80% sparsity. This rapid collapse indicates the removal of critical weights necessary for maintaining logical

consistency, which is crucial for reasoning tasks that require multiple steps.

In contrast, for non-CoT responses, the increase in faithfulness with light pruning is less pronounced, and the collapse at higher sparsity levels occurs more gradually. Memory usage increases slightly due to the model replication overhead, but token throughput improves marginally with moderate sparsity before decreasing at extreme pruning levels. These findings underscore a critical sparsity threshold where the trade-off between efficiency and reasoning fidelity becomes unfavorable, emphasizing the importance of balancing pruning and reasoning consistency.

Similarity-Faithfulness Correlation. Table 2 presents the correlation between similarity metrics and faithfulness degradation, using Pearson's correlation (PLCC), Spearman's rank correlation (SRCC), and Kendall's rank correlation (KRCC). Traditional similarity metrics, such as cosine similarity, L_1 and L_2 distances, and linear CKA, exhibit moderate correlations (PLCC between 0.7023 and 0.8974) with CoT faithfulness. However, these metrics fail to capture the non-linear drops in faithfulness at higher sparsity levels. Notably, ASAND (Adaptive Similarity-based Analysis of Neural Degeneration) outperforms these traditional metrics, achieving the highest PLCC of 0.9483, effectively capturing the non-linear degradation of reasoning performance as pruning increases.

For non-CoT responses, the correlations between similarity metrics and faithfulness drop are generally weaker, suggesting that the chain-of-thought reasoning mode amplifies the model's sensitivity to pruning. ASAND, in this case, continues to show superior performance in capturing this sensitivity, indicating its utility in evaluating reasoning robustness across different pruning scenarios.

Discussion. The results suggest that low-level pruning can enhance CoT faithfulness by removing interfering parameters, leading to a more focused reasoning process. However, pruning beyond $30\% - 40\%$ sparsity causes a sharp decline in reasoning fidelity, as critical weights necessary for logical consistency are removed. This finding highlights the delicate balance required when applying pruning to LLMs: efficiency gains are quickly outweighed by catastrophic degradation in reasoning performance beyond a certain sparsity threshold.

ASAND consistently outperforms traditional similarity metrics, particularly in the non-linear degradation phase, where traditional metrics fail to fully capture the impact of pruning on logical reasoning. The higher correlations for CoT responses indicate that stepwise reasoning is more sensitive to pruning, reinforcing the need for specialized metrics like ASAND to track pruning effects in models that rely on multi-step logical reasoning.

Limitations. This study focuses on unstructured L1 pruning; other pruning methods, such as structured pruning or quantization, may exhibit different patterns and behaviors. Our observations are based on the GSM8K dataset and the Qwen-0.5B model, with validation on the MATH dataset (Hendrycks et al. 2021) as shown in Table 2. Results may vary for larger models or alternative reasoning tasks, such as multi-agent reasoning or logical deduction in more complex domains. Memory and runtime metrics are specific to our sparse weight handling implementation and may not reflect the ideal hardware efficiency achievable with different hardware setups. Despite these limitations, our analysis provides valuable insights into the impact of pruning on reasoning fidelity and underscores the importance of developing adaptive similarity metrics to predict reasoning degradation reliably.

Table 1: Performance and efficiency for Qwen-0.5B on GSM8K. F: Faithfulness $[0, 1]$; Mem: MB; Time: s; T/s: Tokens/s.

Prune	Faithfulness		Mem	Time	T/s
	No CoT	CoT			
0.0	0.352	0.637	948.67	2.71	23.80
0.01	0.365	0.723	9.52	2.52	25.03
0.05	0.462	0.697	47.59	2.46	24.22
0.1	0.178	0.670	95.18	2.59	25.41
0.2	0.195	0.698	190.36	2.56	25.94
0.3	0.203	0.455	285.54	2.57	26.63
0.4	0.083	0.367	380.72	2.51	28.59
0.5	0.100	0.100	475.90	2.49	33.10
0.6	0.150	0.013	571.07	2.49	18.66
0.7	0.013	0.163	666.25	2.52	14.20
0.8	0.000	0.087	761.43	2.49	7.30

Table 2: Correlations (PLCC, SRCC, KRCC) between similarity metrics and faithfulness drop for Qwen-0.5B on GSM8K and MATH datasets. ASAND achieves highest PLCC in CoT.

Dataset	Metric	No CoT			CoT		
		PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
GSM8K	cosine	0.6676	0.8424	0.7333	0.7718	0.8909	0.7778
	L2	*-0.7773*	-0.8424	-0.7333	-0.8974	-0.8909	-0.7778
	L1	-0.7985	-0.8424	-0.7333	*-0.8868*	-0.8909	-0.7778
	CKA	0.6251	0.8303	0.6889	0.7023	0.8667	0.7333
	ASAND	**0.8137**	**0.8424**	**0.7333**	**0.9483**	**0.8909**	**0.7778**
MATH	cosine	0.8811	0.9394	0.8222	0.7683	0.9273	0.7778
	L2	-0.9629	-0.9394	-0.8222	*-0.8888*	-0.9273	-0.7778
	L1	*-0.9656*	-0.9394	-0.8222	-0.8911	-0.9273	-0.7778
	CKA	0.8402	0.9394	0.8222	0.7225	0.9273	0.7778
	ASAND	**0.9720**	**0.9394**	**0.8222**	**0.9424**	**0.9273**	**0.7778**

Conclusion

In this study, we investigate the impact of unstructured pruning on the faithfulness of Qwen-0.5B, with a focus on how different levels of pruning affect logical consistency during chain-of-thought reasoning. Our findings reveal a non-monotonic relationship between pruning and reasoning fidelity: light pruning (low sparsity) can enhance logical consistency by removing redundant or non-informative parameters, thereby improving stepwise reasoning. However, beyond a critical sparsity threshold (approximately $30\% - 40\%$), pruning leads to a sharp degradation in reasoning accuracy, as essential parameters for maintaining logical structure are removed, causing a collapse in performance.

We demonstrate that traditional weight similarity metrics,

such as cosine similarity, L_1 and L_2 distances, capture only part of this degradation, often failing to account for the non-linear behavior observed in high sparsity settings. In contrast, adaptive, sparsity-aware metrics like ASAND are more effective in predicting faithfulness degradation. ASAND incorporates both structural sensitivity and magnitude-aware weighting, enabling it to track the nuanced relationship between pruning and reasoning fidelity with higher accuracy.

These results highlight the need for designing similarity metrics that consider both sparsity and structural perturbations, particularly for tasks that rely on complex, multi-step reasoning. Such metrics are essential for guiding safe and effective model compression, ensuring that pruning techniques do not undermine the model's logical capabilities. Furthermore, our findings emphasize the trade-off between computational efficiency and reasoning fidelity, underscoring the importance of balancing these factors when applying pruning in real-world, resource-constrained settings.

Future work will extend this analysis to explore structured pruning and quantization techniques, as well as apply our findings to larger models. By generalizing these insights, we aim to contribute to the development of deployable, resource-efficient reasoning systems that maintain both high performance and logical consistency. Additionally, we will investigate how adaptive pruning strategies can be designed to preserve the structural integrity of models while optimizing for computational resources, paving the way for more robust AI systems in applications requiring reliable logical reasoning under efficiency constraints.

References

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Han, S.; Pool, J.; Tran, J.; and Dally, W. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, 3519–3529. PMlR.

Li, H.; Kadav, A.; Durdanovic, I.; Samet, H.; and Graf, H. P. 2016. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*.

Raghu, M.; Gilmer, J.; Yosinski, J.; and Sohl-Dickstein, J. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

Shinde, T. 2024. Adaptive quantization and pruning of deep neural networks via layer importance estimation. In *Workshop on Machine Learning and Compression, NeurIPS 2024*.

Shinde, T. 2025. Towards optimal layer ordering for efficient model compression via pruning and quantization. In *2025 25th International Conference on Digital Signal Processing (DSP)*, 1–5. IEEE.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.