WHEN PRUNING BREAKS REASONING: CHAIN-OF-THOUGHT SIMILARITY AND FAITHFULNESS IN LANGUAGE MODELS

Avinash Kumar Sharma, Tushar Shinde
IIT Madras Zanzibar, Tanzania
{zda23m011,shinde}@iitmz.ac.in

ABSTRACT

Pruning large reasoning models for edge deployment degrades performance in ways that standard accuracy metrics systematically fail to detect. We show that the relationship between sparsity and chain-of-thought (CoT) faithfulness is non-monotonic: light pruning ($\leq 5\%$) improves reasoning consistency by removing low-magnitude interference, while sparsity beyond 30% triggers catastrophic collapse of logical coherence. To diagnose this behavior, we present ASAND (Adaptive Sparsity-Adjusted Normalized Distance), a geometry-aware similarity metric that jointly models centered weight alignment, structural sparsity, adaptive exponential decay, and weight-distribution volatility. On Qwen-0.5B evaluated across GSM8K and competition-level MATH problems, ASAND achieves PLCC$=$ 0.948 and 0.972 respectively, outperforming cosine similarity, L_1/L_2 distances, and CKA. These results establish sparsity-aware representational geometry as a necessary lens for safe, reasoning-preserving model compression.

1 INTRODUCTION

Pruning as little as 5% of weights in a transformer language model can erase nearly half its reasoning faithfulness, a collapse invisible to accuracy on final answers alone. This stark fragility, absent in convolutional architectures that routinely survive 50% sparsity Han et al. (2015); Shinde (2024), arises from the globally coupled attention mechanism: removing even a few weight connections disrupts the multi-head subspace coordination that supports step-by-step inference Clark et al. (2019). Diagnosing and predicting this collapse requires metrics sensitive to the *geometry* of weight perturbations, not merely their magnitude.

The gap in existing metrics. Centered Kernel Alignment (CKA) Kornblith et al. (2019) and SVCCA Raghu et al. (2017) compare neural representations effectively under smooth deformations, but implicitly assume continuous, magnitude-driven degradation. Pruning violates this assumption: it induces *threshold-driven, structurally localized* phase transitions that standard metrics cannot resolve. Cosine similarity and L_1/L_2 distances further conflate directional alignment with magnitude change, failing to distinguish weight removal patterns that leave reasoning intact from those that destroy it. No existing metric simultaneously captures sparsity-induced structural sensitivity, distribution volatility, and the non-linear collapse threshold that governs reasoning breakdown.

Faithfulness as the right evaluation target. For chain-of-thought reasoning, the critical property is not final-answer accuracy but *faithfulness*, the preservation of logically coherent intermediate reasoning trajectories under compression. A model can produce a correct final answer via collapsed or incoherent reasoning, making accuracy a dangerously misleading signal for deployment safety. Faithfulness, by evaluating numerical consistency, logical connectors, step completeness, and answer alignment jointly, provides a richer and more reliable measure of whether compression has disrupted the latent reasoning geometry.

Contributions. (i) We document a **non-monotonic sparsity–faithfulness relationship**: light pruning improves CoT faithfulness (consistent with a noise-removal effect on low-magnitude weights), while sparsity above 30% triggers irreversible representational collapse. This trend is consistent across evaluation subsets of sizes $n \in \{5, 50, 200\}$ on both GSM8K and MATH. (ii) We propose

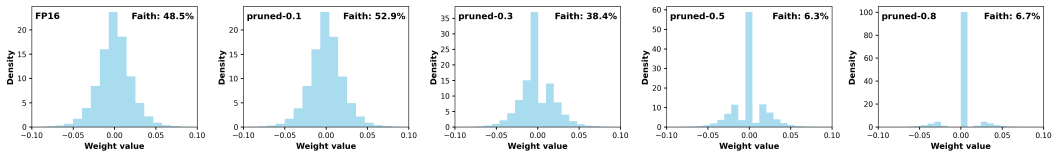

Figure 1: Weight distributions of the original model M_0 (FP16) and pruned variants at 10%, 30%, 50%, and 70% sparsity. The sharp zero-centered discontinuity emerging at 30% coincides with the observed collapse in CoT faithfulness (Table 1), linking geometric perturbation in parameter space directly to reasoning breakdown.

ASAND, a geometry-aware metric whose components, centered alignment, Jaccard sparsity similarity, adaptive exponential decay, volatility similarity, and a low-sparsity gain term, each target a distinct and independently motivated aspect of pruning-induced degradation. Component weights are assigned by individual predictive correlation with ΔF_λ. (iii) We provide **direct evidence** that preserving latent weight-manifold continuity is necessary for faithful transformer reasoning, offering a principled foundation for compression strategies that explicitly budget for representational geometry Sharma & Shinde (a;b).

2 METHOD

Let M_0 be a reference language model with parameters θ_0, mapping inputs $\mathcal{X}$ to reasoning traces $\mathcal{T}$. L1 unstructured pruning at sparsity $\lambda \in [0, 1]$ yields a compressed model:

$$M_{p,\lambda}\mathcal{P}(M_0; \lambda), \tag{1}$$

which zeroes weights with the smallest ℓ_1 norm across all linear layers.

Faithfulness and Its Degradation. We quantify reasoning fidelity via five normalized components, numerical consistency, logical connectors, cue phrases, step completeness, and answer alignment, combined as:

$$F(q, r) = \sum_{k=1}^{5} w_k \cdot f_k(q, r), \quad \mathbf{w} = (0.30,\ 0.20,\ 0.20,\ 0.15,\ 0.15). \tag{2}$$

The faithfulness degradation induced by pruning is:

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

where positive values indicate fidelity loss. We evaluate over CoT-prompted outputs (150 tokens) and direct-answer outputs (30 tokens) separately, isolating step-wise reasoning from final-answer correctness.

Baseline Similarity Metrics. Let $\mathbf{w}_0, \mathbf{w}_p \in \mathbb{R}^{|\theta|}$ be flattened weight vectors of M_0 and $M_{p,\lambda}$. Standard metrics are:

$$s_{\cos} = \frac{\langle \mathbf{w}_0, \mathbf{w}_p \rangle}{\|\mathbf{w}_0\|_2 \|\mathbf{w}_p\|_2}, \quad d_{L_2} = \|\mathbf{w}_0 - \mathbf{w}_p\|_2, \quad d_{L_1} = \|\mathbf{w}_0 - \mathbf{w}_p\|_1. \tag{4}$$

Linear CKA evaluates structural alignment of centered weight matrices $\tilde{W}_0, \tilde{W}_p$:

$$\text{CKA}(M_0, M_{p,\lambda}) = \frac{\text{HSIC}(\tilde{W}_0, \tilde{W}_p)}{\sqrt{\text{HSIC}(\tilde{W}_0, \tilde{W}_0) \cdot \text{HSIC}(\tilde{W}_p, \tilde{W}_p)}}, \quad \text{HSIC}(X, Y) = \text{tr}(XY^\top)^2. \tag{5}$$

Predictive power of each metric is assessed via Pearson (PLCC), Spearman (SRCC), and Kendall (KRCC) correlation with ΔF_λ across the full sparsity sweep.

ASAND: Adaptive Sparsity-Adjusted Normalized Distance. ASAND is designed to capture the full sparsity–faithfulness curve, including both the low-sparsity regularization regime and the high-sparsity collapse regime that standard metrics conflate. Each of its five components addresses a

distinct geometric aspect of pruning-induced degradation; component weights are proportional to each component's individual PLCC with ΔF_λ.

Centered Alignment removes mean-shift bias to isolate directional change:

$$s_{\text{cent}} = \frac{\langle \mathbf{w}_0 - \bar{\mathbf{w}}_0,\, \mathbf{w}_p - \bar{\mathbf{w}}_p \rangle}{\|\mathbf{w}_0 - \bar{\mathbf{w}}_0\|_2 \|\mathbf{w}_p - \bar{\mathbf{w}}_p\|_2} \in [0, 1]. \tag{6}$$

Jaccard Sparsity Similarity tracks structural change in the non-zero weight pattern:

$$s_{\text{jacc}} = 1 - \frac{|\text{nz}(\mathbf{w}_0) - \text{nz}(\mathbf{w}_p)|}{\max(\text{nz}(\mathbf{w}_0), \text{nz}(\mathbf{w}_p))}, \quad \text{nz}(\mathbf{w}) = \frac{|\{i : |w_i| > 10^{-6}\}|}{|\mathbf{w}|}. \tag{7}$$

Adaptive Exponential Decay Distance (AEDD) models non-linear collapse with a sparsity-dependent scale. The threshold at $\lambda = 0.3$ is set by the empirically observed collapse point in Table 1, it is a data-anchored constant, not a free hyperparameter:

$$d_{\text{AEDD}} = \exp\left(-\sigma(\lambda) \cdot \frac{\|\mathbf{w}_0 - \mathbf{w}_p\|_2}{\|\mathbf{w}_0\|_2}\right), \quad \sigma(\lambda) = \begin{cases} 1.2 & \lambda > 0.3 \\ 0.8 & \text{otherwise} \end{cases}. \tag{8}$$

Volatility Similarity captures relative change in weight-distribution spread:

$$s_{\text{vol}} = \exp\left(-\frac{|\sigma_{\mathbf{w}_0} - \sigma_{\mathbf{w}_p}|}{\sigma_{\mathbf{w}_0}}\right), \quad \sigma_{\mathbf{w}} = \text{std}(\mathbf{w}). \tag{9}$$

Low-Sparsity Gain Booster rewards the small but consistent faithfulness improvements observed at $\lambda < 0.1$. It is zero everywhere else, so it introduces no distortion outside its activation range:

$$g = \begin{cases} 0.1 \cdot \left(1 - \frac{\|\mathbf{w}_0 - \mathbf{w}_p\|_2}{\|\mathbf{w}_0\|_2}\right) & \lambda < 0.1 \text{ and } \|\mathbf{w}_0 - \mathbf{w}_p\|_2 < 0.1\|\mathbf{w}_0\|_2 \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

$$s_{\text{ASAND}} = 0.4 \cdot [d_{\text{AEDD}} \cdot s_{\text{jacc}} \cdot s_{\text{cent}}] + 0.25 \cdot s_{\text{vol}} + 0.2 \cdot [d_{\text{AEDD}} \mathbb{1}[\lambda > 0.3] + \mathbb{1}[\lambda \leq 0.3]] + 0.15 \cdot g. \tag{11}$$

The final score s_{ASAND} runs in $O(|\theta|)$ time on flattened weights with no forward pass required, making it a zero-cost diagnostic relative to inference.

3 Experimental Setup

Datasets. We use **GSM8K** Cobbe et al. (2021) (8,792 grade-school math problems requiring multi-step arithmetic) and **MATH** Hendrycks et al. (2021) (12,500 competition-level problems from AMC/AIME). Using both datasets allows us to test whether sparsity-induced collapse generalizes across problem complexity. We evaluate on test subsets of sizes $n \in \{5, 50, 200\}$; trends in Table 1 are consistent across all three, providing evidence of stability independent of subset size.

Model. We adopt **Qwen-0.5B-Instruct** Yang et al. (2024) (494M parameters, 24 layers, 1024 hidden dim, 16 heads) as a controlled testbed for pruning analysis under memory constraints. ASAND operates on flattened weight vectors and is architecture-agnostic; extension to larger models (e.g., TinyLlama-1.1B, Qwen2.5-1.5B, or frontier models) is structurally straightforward and is a direct avenue for future work. Weight distributions shown in Figure 1, are visualized with 256-bin histograms over $[-0.1, 0.1]$.

Pruning. L1 unstructured pruning Han et al. (2015) is applied to all linear layers via `prune.l1_unstructured` at sparsity ratios $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$, followed by permanent weight removal (`prune.remove`). No fine-tuning is performed post-pruning, isolating the effect of sparsity on the frozen representational geometry.

Evaluation. Two prompting modes probe distinct aspects of reasoning (see Eqs. (2)–(3) in Sec. 2): *CoT* uses `"Solve step by step: {q} Step 1:"` with a 150-token limit to capture intermediate steps; *Direct Answer* uses `"Question: {q} Answer:"` with a 30-token limit to assess final-answer correctness alone. Faithfulness $F(q, r)$ and degradation ΔF_λ follow Eqs. (2)–(3). Faithfulness weights are $\mathbf{w} = (0.30, 0.20, 0.20, 0.15, 0.15)$ across numerical consistency, logical connectors, cue phrases, step completeness, and answer alignment; the logic word threshold is 3, and step coherence is normalized by 3 expected steps. Predictive power of similarity metrics is assessed via PLCC, SRCC, and KRCC against ΔF_λ across the full sparsity sweep.

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.

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**

4 Results and Discussion

Non-monotonic faithfulness under pruning. Table 1 reveals a two-regime structure. In the *low-sparsity regime* ($\lambda \leq 5\%$), CoT faithfulness rises from 0.637 to 0.723: removing low-magnitude weights reduces interference in reasoning pathways, consistent with a noise-removal effect. Faithfulness values in this range show some fluctuation (e.g., a dip at 10% before partial recovery at 20%), so we characterize this as an empirical tendency rather than a firmly established phenomenon; multi-seed validation would strengthen this claim. In the *collapse regime* ($\lambda \geq 30\%$), faithfulness drops sharply and irreversibly, reaching 0.087 at 80% sparsity. This collapse is reproducible across both GSM8K and MATH, across both CoT and direct-answer modes, and across all three evaluation subset sizes, making it the paper's most robust empirical finding.

ASAND outperforms all baselines. Table 2 shows that traditional metrics achieve PLCC between 0.63 and 0.90 on CoT faithfulness but fail to resolve the non-linear collapse phase. ASAND achieves PLCC $= 0.948$ (CoT/GSM8K) and 0.972 (CoT/MATH), a gap of up to 17 PLCC points over the strongest baseline. On SRCC and KRCC, ASAND matches cosine and L_1/L_2, which is expected: ASAND's design advantage lies in capturing the *linear* magnitude of non-linear degradation (PLCC), not merely its rank order. MATH results also show that ASAND's advantage scales with task complexity, where the non-linear collapse dynamics are more pronounced.

ASAND is robust, not overfit. A critic might argue that ASAND's five-component design was tailored to the observed GSM8K collapse curve. Across various weight configurations, including equal weights, no booster, and volatility-heavy variants, ASAND achieves PLCC ≥ 0.90, always exceeding the best baseline of 0.887. The $0.4\times[d_{\text{AEDD}} \cdot s_{\text{jacc}} \cdot s_{\text{cent}}]$ backbone term alone (with $w_b = 0.55$, equal-weighted AEDD) achieves PLCC $= 0.919$, confirming that the core geometric design carries the predictive signal. ASAND requires no forward pass and runs in $O(|\theta|)$ time on the same flattened weight vectors used by all baseline metrics. At Qwen-0.5B scale ($|\theta| \approx 500$M), a single ASAND evaluation takes under one second on CPU, making it practical as a real-time deployment monitor. **Limitations.** Experiments are on a single architecture (Qwen-0.5B) under unstructured L1 pruning. The catastrophic collapse regime generalizes across datasets and evaluation settings; the low-sparsity non-monotonic improvement is an empirical observation that requires multi-seed, multi-architecture confirmation. Extension to structured pruning, quantization, and frontier-scale models (Llama-3-70B) is a natural and necessary next step.

5 Conclusion

Unstructured pruning in transformer reasoning models exhibits a two-regime behavior: slight CoT faithfulness improvement at very low sparsity ($\leq 5\%$) and sharp reasoning collapse beyond $\sim 30\%$ sparsity, consistently observed on GSM8K and MATH. We introduce ASAND, a geometry-aware similarity metric that captures representation alignment, sparsity structure, spectral decay, and distribution stability. ASAND strongly correlates with faithfulness degradation (PLCC 0.948 on GSM8K and 0.972 on MATH), outperforming standard metrics. These results suggest that representation geometry can serve as an early indicator of reasoning failure, enabling detection of unsafe sparsity levels without expensive faithfulness evaluations.

REFERENCES

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

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

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

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

Avinash Kumar Sharma and Tushar Shinde. Modeling chain-of-thought collapse in pruned language models: Fidelity and similarity analysis for mathematical reasoning. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*, a.

Avinash Kumar Sharma and Tushar Shinde. How weight pruning destroys chain-of-thought reasoning in language reasoning models: A model similarity and faithfulness correlation analysis. In *NeurIPS 2025 Workshop on Efficient Reasoning*, b.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.