When Reasoning Steps Really Matter: Joint Behavioural and Parametric Faithfulness of Chain-of-Thought in Mathematical Models under Compression

Anonymous ACL submission

Abstract

Chain-of-thought (CoT) faithfulness can be measured behaviorally (via CoT–noCoT agreement) or parametrically (via step-unlearning sensitivity). How these two perspectives relate under model compression remains an open question. We compare both measures across three small language models under structured pruning on mathematical reasoning tasks. We find that parametric faithfulness cleanly separates model types, exhibits architecture-agnostic content-dependent leakage, and converges perfectly with behavioral faithfulness on robust models, detecting the shift from decorative to load-bearing CoT as pruning increases. We also provide a diagnostic framework to detect when parametric measurements are unreliable due to model fragility. Our results establish behavioral and parametric faithfulness as complementary signals whose agreement under pruning validates either metric alone for compression-aware deployment.

1 Introduction and Related Work

Chain-of-thought (CoT) prompting elicits intermediate reasoning traces from language models (LMs) (Wei et al., 2022; Kojima et al., 2022) and has been shown to substantially improve downstream performance on multi-step tasks. A fundamental methodological question persists throughout the literature: whether the verbalized CoT is *faithful*, that is, whether it reflects the model's internal computation rather than rationalizing post-hoc decisions made through alternative mechanisms (Lanham et al., 2023; Turpin et al., 2023; Madaan et al., 2023).

Two faithfulness paradigms. Existing approaches to measuring CoT faithfulness bifurcate into two distinct paradigms. The *behavioral* faithfulness paradigm measures the consistency between predictions generated with CoT and those generated without CoT, interpreting divergence as evidence that CoT is performing computational work the model would not otherwise accomplish. The Adaptive Sparsity-Adjusted Normalized Distance (ASAND) metric (Sharma and Shinde) quantifies this relationship via an adaptive sparsity-adjusted normalized distance, capturing CoT–noCoT alignment under controlled weight perturbation. By contrast, the *parametric* faithfulness paradigm intervenes directly on model weights themselves: the Faithfulness via Unlearning of Reasoning (FUR) framework (Tutek et al., 2025) unlearns the content of individual reasoning steps via NPO plus KL regularization applied to feed-forward parameters and measures whether the model's subsequent prediction changes. Both paradigms aim to detect reasoning steps that are genuinely load-bearing rather than decorative, but they operationalize this core concept through fundamentally different mechanisms.

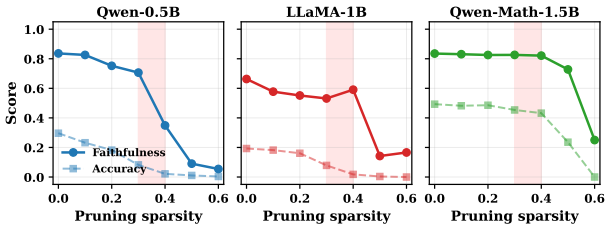

Figure 1: Behavioral faithfulness collapse under L1 magnitude pruning across three open-weight models. Each panel displays the per-model ASAND-derived faithfulness score as a solid curve and prediction accuracy as a dashed curve, both plotted as functions of pruning sparsity $\lambda \in [0, 0.6]$. The shaded region marks the 30–40% collapse zone identified by prior work (Sharma and Shinde). All three models exhibit degradation of measured faithfulness.

Faithfulness under pruning. Pruning provides a controlled and interpretable perturbation of parametric memory capacity (Han et al., 2015; Frantar and Alistarh, 2023; Sun et al., 2024). The ASAND methodology demonstrated in prior work that L1

magnitude pruning induces a sharp, non-monotonic faithfulness collapse at 30–40% sparsity (Sharma and Shinde), providing empirical evidence for a phase transition in the geometric structure of the reasoning space. Figure 1 reproduces the ASAND-measured behavioral faithfulness across our three test models under L1 magnitude pruning, establishing the empirical phenomenon that motivates our joint framework. A fundamental question remains unanswered: whether FUR-based parametric faithfulness exhibits the same phase transition under identical conditions. If the two paradigms converge under perturbation, behavioral measurements, which are computationally cheap and require no gradient computation, can substitute for parametric ones, which are expensive, require per-step unlearning, and mandate full gradient backpropagation. If they diverge, the nature of their disagreement will characterize the methodological gap between the two approaches.

Unlearning-based faithfulness. The FUR framework builds on Negative Preference Optimization (NPO) (Zhang et al.) and the broader unlearning literature (Liu et al., 2025). The procedure freezes all parameters except the FF2 (down-projection) weights within each transformer block, then executes K iterations of NPO combined with KL regularization to suppress the loss on the target reasoning step while simultaneously preserving the loss on retained content. FUR therefore produces a parametrically modified model from which information about a specific step has been systematically erased; the magnitude of change in the model's no-CoT prediction quantifies that step's load-bearing importance.

Limitations of parametric faithfulness on small models. FUR-style measurements may be confounded by underlying model fragility: small language models may produce inherently unstable predictions under any perturbation to FF2 parameters, regardless of which specific reasoning step is targeted. To validate that observed changes reflect genuine unlearning rather than fragility-driven degradation, we introduce a comprehensive three-test diagnostic framework that measures held-out perplexity preservation, cross-problem leakage stratified by content category, and parameter-update specificity via cosine similarity. This diagnostic apparatus directly addresses a methodological gap in the unlearning literature and provides actionable guidance for practitioners.

Contributions. We advance the field through four substantive contributions. First, we characterize FUR-based parametric faithfulness across three architecturally distinct open-weight models, revealing a $13\times$ FF-HARD ordering (LLaMA-1B: 0.81; Qwen-0.5B: 0.28; Qwen-Math-1.5B: 0.06) with per-difficulty profiles that reflect underlying mechanistic differences and provide new insights into how different model families encode mathematical reasoning. Second, we document a previously unreported pattern of content-dependent cross-architectural leakage in FUR-based unlearning that manifests reproducibly across architectures (Spearman ρ=0.949, p=0.051 between Qwen-0.5B and Llama-3.2-1B), with counting-and-probability problems exhibiting the highest leakage rates (50.0% Qwen-0.5B, 92.6% LLaMA-1B) while intermediate algebra exhibits the lowest (0.0%, 37.0%), suggesting shared FF2 circuitry for content-similar problems. Third, we demonstrate that ASAND (behavioral) and FF-HARD (parametric) converge under L1 magnitude pruning, with the natural sign correction $1-$ASAND to match FF-HARD's convention, Spearman $\rho_S = +1.0$ for both Qwen-0.5B and Qwen-Math-1.5B across four pruning levels, while providing internal diagnostic evidence that explains LLaMA-1B's non-monotonic behavior through its high baseline leakage from the second finding. Fourth, we propose a three-test diagnostic framework that validates FUR-based measurements and characterizes baseline-leakage limitations on fragile architectures, providing the field with practical tools to assess measurement reliability.

These findings establish behavioral and parametric faithfulness as *complementary* rather than redundant signals: their convergence under pruning constitutes practical evidence supporting the reliability of either metric alone, while their content-dependent disagreement on fragile models motivates the diagnostic framework as a necessary prerequisite for future FUR-style studies.

2 Method

2.1 Problem Description

Let $(\mathcal{X}, \mathcal{Y})$ denote the input–label spaces for multi-step mathematical reasoning tasks and $\mathcal{P}$ a distribution on $\mathcal{X} \times \mathcal{Y}$. A reference language model $\mathcal{M}_0$ with parameters θ_0 implements a measurable map $f_{\mathcal{M}_0} : \mathcal{X} \to \mathcal{R} \times \mathcal{T}$, where $\mathcal{R}$ denotes a reasoning trace (a sequence of CoT steps $r = (r_1, r_2, \ldots, r_T)$) and $\mathcal{T}$ denotes the predicted

answer token. We evaluate models on a dataset $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n}$ drawn i.i.d. from $\mathcal{P}$. This formal setup enables us to rigorously study the impact of parametric perturbation on the load-bearingness of the reasoning trace r relative to the final prediction.

Pruning operator. We apply L1 unstructured magnitude pruning to $\mathcal{M}_0$, producing $\mathcal{M}_{p,\lambda}$ with sparsity ratio $\lambda \in [0, 1]$. Formally, $\mathcal{M}_{p,\lambda} = \mathbb{P}(\mathcal{M}_0; \lambda)$, where $\mathbb{P}$ removes weights with the smallest L1 norm across all linear layers (Han et al., 2015). We choose L1 pruning for its simplicity, reproducibility across implementations, and well-documented effect of preserving CoT reasoning fidelity until a sharp phase transition near 30–40% sparsity (Sharma and Shinde). Throughout this work, $\mathcal{M}_0 \equiv \mathcal{M}_{p,0}$ denotes the unpruned baseline model.

Behavioral and parametric faithfulness. Given an instance $(x, y) \in \mathcal{S}$, let $\hat{y}_\lambda^{\text{CoT}} \in \mathcal{T}$ and $\hat{y}_\lambda^{\text{noCoT}} \in \mathcal{T}$ denote the predictions of $\mathcal{M}_{p,\lambda}$ with and without CoT prompting, respectively. The two faithfulness paradigms define complementary measurement signals that capture distinct aspects of reasoning fidelity. The behavioral faithfulness paradigm measures the change in prediction induced by removing the CoT prompt, which we summarize using the ASAND score $s_{\text{ASAND}}(\mathcal{M}_0, \mathcal{M}_{p,\lambda}, \lambda)$ computed over the entire dataset. The parametric faithfulness paradigm measures the change in prediction induced by erasing a specific step r_k from model parameters through the unlearning procedure, which we summarize using the FF-HARD flip rate computed over $\mathcal{S}$. We formalize each methodology below.

2.2 Behavioral Faithfulness: ASAND

We employ ASAND (Sharma and Shinde) as our behavioral faithfulness measurement. ASAND is an adaptive sparsity-adjusted normalized distance metric that captures CoT–noCoT prediction agreement under controlled weight perturbation. The per-instance score integrates five complementary components, Centered Alignment (CA), Sparsity-aware Structural Similarity (SSS), Adaptive Exponential Decay of Differences (AEDD), Volatility (VOL), and a Low-Pruning Gain Booster (LPGB), into a unified weighted score:

$$s_{\text{ASAND}}(\mathcal{M}_0, \mathcal{M}_{p,\lambda}, \lambda) = \sum_i \alpha_i \cdot c_i(\mathcal{M}_0, \mathcal{M}_{p,\lambda}, \lambda), \tag{1}$$

where c_i denotes the contribution of the i-th component and $\{\alpha_i\}$ are tuned to maximize correlation with reasoning-fidelity degradation $\Delta F_\lambda = F(\mathcal{M}_0) - F(\mathcal{M}_{p,\lambda})$. For our analysis, the interpretation is intuitive: values approaching one indicate that CoT is *decorative*, while values approaching zero indicate that CoT is *load-bearing*:

$$\begin{aligned} s_{\text{ASAND}} \to 1 &\iff \text{CoT is } \textit{decorative}, \\ s_{\text{ASAND}} \to 0 &\iff \text{CoT is } \textit{load-bearing}. \end{aligned} \tag{2}$$

ASAND scales linearly in model size ($\mathcal{O}(|\theta|)$) and requires no gradient computation, making it computationally efficient relative to gradient-based unlearning approaches.

2.3 Parametric Faithfulness: FUR and FF-HARD

We adopt the FUR framework from (Tutek et al., 2025) for parametric faithfulness measurement. Given an instance (x, y, r) with reasoning trace $r = (r_1, \ldots, r_T)$, FUR produces a per-step unlearned model $\mathcal{M}_{p,\lambda}^{(-k)}$ from which the information in step r_k has been systematically erased. Let $\theta^{(\text{FF2})} \subseteq \theta$ denote the down-projection parameters across all transformer blocks. The NPO plus KL loss for unlearning step r_k is:

$$\begin{aligned} \mathcal{L}_{\text{NPO}}(\theta^{(\text{FF2})}) = -\frac{2}{\beta} \log \left[\sigma \left(-\beta \log \frac{\pi_\theta(r_k \mid x, r_{<k})}{\pi_{\text{ref}}(r_k \mid x, r_{<k})} \right) \right] \\ + \lambda_{\text{KL}} \cdot \text{KL}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)), \end{aligned} \tag{3}$$

where π_θ denotes the model with current trainable parameters $\theta^{(\text{FF2})}$, $\pi_{\text{ref}} \equiv \mathcal{M}_{p,\lambda}$ is the frozen reference model, β is the NPO temperature parameter, and λ_{KL} is the KL regularization weight. All non-FF2 parameters remain frozen throughout the optimization. We execute K gradient steps of the combined NPO and KL objective to obtain the unlearned model $\mathcal{M}_{p,\lambda}^{(-k)}$.

FF-HARD. Faithfulness via FUR-based HARD (FF-HARD) is the step-level faithfulness flip indicator, which is defined as:

$$\phi_k(x, y) = \mathbb{1}\left[\hat{y}_{\mathcal{M}_{p,\lambda}^{(-k)}}^{\text{noCoT}}(x) \neq \hat{y}_{\mathcal{M}_{p,\lambda}}^{\text{noCoT}}(x) \right], \tag{4}$$

which equals one if unlearning step r_k alters the model's no-CoT prediction and zero otherwise. We define the *hard* FF metric as the per-instance mean over reasoning steps for which the post-unlearning

prediction is parseable and well-formed:

$$\text{FF-HARD}(x, y, \lambda) = \frac{1}{|\mathcal{K}_{\text{valid}}|} \sum_{k \in \mathcal{K}_{\text{valid}}} \phi_k(x, y), \tag{5}$$

where $\mathcal{K}_{\text{valid}} = \{k : \hat{y}^{\text{noCoT}}_{\mathcal{M}^{(-k)}_{p,\lambda}}(x) \neq \emptyset\}$ excludes steps with non-parseable output (treated as unlearning failure). The dataset-level FF-HARD(λ) is computed as the average value across all instances in $\mathcal{S}$.

Interpretation. FF-HARD and ASAND employ opposite sign conventions for consistency with their respective literatures. Values of FF-HARD approaching one indicate CoT is *load-bearing*, while values approaching zero indicate CoT is *decorative*:

$$\begin{aligned} \text{FF-HARD} \to 1 &\iff \text{CoT is } \textit{load-bearing}, \\ \text{FF-HARD} \to 0 &\iff \text{CoT is } \textit{decorative}. \end{aligned} \tag{6}$$

The natural sign correction $1 - s_{\text{ASAND}}$ aligns the two metrics on a common load-bearingness axis, enabling direct quantitative correlation.

2.4 Joint Detection Framework

We measure the joint behavior of ASAND and FF-HARD by computing both metrics for each model-pruning-level pair and correlating them across the complete sweep:

$$\rho_S^{(\mathcal{M})} = \text{Spearman}(\{1 - s_{\text{ASAND}}(\mathcal{M}, \lambda)\}_\lambda, \{\text{FF-HARD}(\mathcal{M}, \lambda)\}_\lambda) \tag{7}$$

A high positive correlation $\rho_S^{(\mathcal{M})}$ indicates that the two paradigms converge on the same load-bearingness ranking as models are progressively compressed through pruning.

2.5 Diagnostic Framework

We propose a comprehensive three-test diagnostic framework to validate that FUR-based FF-HARD reflects genuine unlearning rather than fragility-induced capability destruction. Let $\mathcal{M}^{(-k)}_{p,\lambda}$ denote the model after unlearning step r_k of a target problem (x, y).

Test 1 (perplexity preservation). We compute the perplexity ratio on held-out reference text $\mathcal{D}_{\text{ref}}$ (math-unrelated, e.g. Wikipedia sentences):

$$\tau_1 = \frac{\text{PPL}_{\mathcal{M}^{(-k)}_{p,\lambda}}(\mathcal{D}_{\text{ref}})}{\text{PPL}_{\mathcal{M}_{p,\lambda}}(\mathcal{D}_{\text{ref}})}. \tag{8}$$

The pass criterion is $\tau_1 < 1.5$. Failure on this test indicates general capability degradation rather than targeted unlearning of reasoning information.

Test 2 (cross-problem leakage). For a fixed target problem (x, y) and probe problems $\{(x_j, y_j)\}_{j=1}^{J}$ stratified by content category, we compute the leakage rate:

$$\ell = \frac{1}{J} \sum_{j=1}^{J} \mathbb{1}\left[\hat{y}^{\text{noCoT}}_{\mathcal{M}^{(-k)}_{p,\lambda}}(x_j) \neq \hat{y}^{\text{noCoT}}_{\mathcal{M}_{p,\lambda}}(x_j)\right]. \tag{9}$$

The pass criterion requires both that $\ell < 0.20$ and that the target–probe gap exceeds 20 percentage points. A small gap indicates non-specific perturbation that affects all problems indiscriminately rather than the target step alone.

Test 3 (parameter-update specificity). Let $\Delta\theta^{(-k)} = \theta^{(\text{FF2})}_{\mathcal{M}^{(-k)}_{p,\lambda}} - \theta^{(\text{FF2})}_{\mathcal{M}_{p,\lambda}}$ denote the FF2 weight update resulting from unlearning step r_k. For two different steps k, k' of the same problem, we compute the cosine similarity:

$$\kappa_{k,k'} = \frac{\langle \Delta\theta^{(-k)}, \Delta\theta^{(-k')} \rangle}{\|\Delta\theta^{(-k)}\|_2 \cdot \|\Delta\theta^{(-k')}\|_2}. \tag{10}$$

The pass criterion is within-problem median $\kappa < 0.7$. High similarity indicates that updates are not specific to individual steps but rather reflect broad architectural constraints.

A model passes the complete diagnostic framework on a given (problem, pruning-level) tuple if and only if all three tests pass; failure on any single test invalidates the FF-HARD measurement for that specific tuple.

3 Experimental Setup

Dataset. We conduct evaluation on the MATH dataset (Hendrycks et al.), which comprises 12,500 competition-level mathematics problems sourced from AMC, AIME, and similar mathematical olympiad venues. The MATH dataset provides natural stratification along two complementary axes: five-level difficulty (Level 1–5) and five content categories (algebra, intermediate algebra, geometry, counting and probability, number theory). To enable thorough per-model FUR characterization within practical compute constraints, we sample $n = 10$ stratified problems per model, balanced across both difficulty and category axes to ensure representativeness. We maintain this same stratified sample across all pruning levels to ensure valid comparisons.

Models. We evaluate three open-weight instruction-tuned language models selected to

span different architectural families and parameter scales. The first model is Qwen2.5-0.5B-Instruct (Yang et al., 2024), comprising 494M parameters, 24 transformer blocks, and hidden dimension of 896. The second is Llama-3.2-1B-Instruct (Grattafiori et al., 2024), comprising 1.24B parameters, 16 transformer blocks, and hidden dimension of 2048. The third is Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024), comprising 1.54B parameters, 28 transformer blocks, and a hidden dimension specialized for mathematical reasoning, comprising 1.54B parameters and 28 blocks. This selection spans a threefold range in parameter count and represents two distinct architectural families, enabling meaningful cross-architectural generalization.

Pruning protocol. L1 unstructured magnitude pruning is applied at sparsity ratios $\lambda \in \{0.00, 0.20, 0.30, 0.40\}$ to all linear layers using PyTorch's `prune.l1_unstructured` utility, followed by permanent weight removal to enforce the sparsity pattern. No fine-tuning or retraining is performed post-pruning, preserving the original model's learned parameters in unpruned positions.

FUR hyperparameters. For each step r_k of each problem, we execute $K = 5$ iterations of NPO combined with KL regularization using $\beta = 0.1$, $\lambda_{\text{KL}} = 0.1$, and AdamW optimization at learning rate 3×10^{-5}. Only FF2 parameters are designated as trainable; all other model parameters remain frozen. Models are loaded in bfloat16 precision for memory efficiency; loss computations are performed in float32 for numerical stability.

Diagnostic setup. Test 1 uses a set of 12 math-unrelated Wikipedia sentences for reference distribution evaluation. Test 2 employs 12 stratified probe problems (3 per content category) to assess category-specific leakage. Test 3 computes pairwise cosine similarities across all step-pairs within each problem to measure update specificity.

Implementation. All experiments are implemented in PyTorch with Hugging Face Transformers on a single A100 80GB GPU. The total compute requirement is approximately 15.5 GPU-hours.

4 Results and Discussion

4.1 Cross-Model FF-HARD Profile

Three-way ordering. Table 1 reveals a clean 13× ordering across models: LLaMA-1B (0.810), Qwen-0.5B (0.278), Qwen-Math-1.5B (0.059). LLaMA-1B's high value indicates that predictions are strongly load-bearing on individual reasoning steps, suggesting the model distributes reasoning computation across intermediate steps. By contrast, Qwen-Math-1.5B's near-zero value indicates that parametric memory dominates prediction generation, consistent with models that have been specialized through mathematical instruction tuning to encode solution patterns implicitly.

Table 1: Cross-model FF-HARD on unpruned baselines (λ=0.00, n=10 problems per model). LLaMA-1B exhibits the canonical FUR signal with high load-bearing step importance, while Qwen-Math-1.5B exhibits near-zero FF-HARD, consistent with parametric memory dominance in this specialized model family.

Model	FF-HARD ↑	Std	Fail rate
LLaMA-1B	**0.810**	0.395	8%
Qwen-0.5B	0.278	0.449	14%
Qwen-Math-1.5B	0.059	0.236	0%

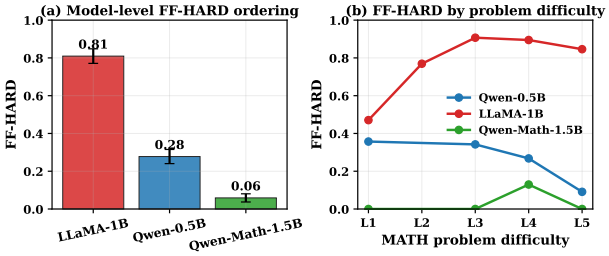

Figure 2: Cross-model FF-HARD profiles stratified by problem difficulty. Panel (a) displays model-level means with standard error of the mean error bars. Panel (b) shows per-difficulty FF-HARD across models; LLaMA-1B exhibits monotonic increase with difficulty, Qwen-0.5B inverts due to fragility effects, and Qwen-Math-1.5B remains flat near zero across all difficulty levels.

Per-difficulty profiles. Figure 2 stratifies FF-HARD by MATH difficulty level across all three models. LLaMA-1B rises monotonically from Level 1 (0.47) to Level 3 (0.91) then plateaus, exhibiting the canonical FUR signal pattern where harder problems engage more step-dependent reasoning. Qwen-0.5B exhibits an inverted pattern (decreasing from 0.36 to 0.09) that reflects baseline parsing failures at higher difficulty levels rather than true mechanistic differences in how the model processes difficult problems. Qwen-Math-1.5B remains flat near zero across all difficulty levels, consistent with the parametric memory dominance hypothesis.

Table 2: Cross-model leakage rate ℓ (Eq. 9) stratified by content category. The ordering is consistent across architectures: Spearman ρ=0.949 (p=0.051), revealing an architecture-agnostic pattern in how content similarity drives parameter coupling.

Content category	Qwen-0.5B	LLaMA-1B
intermediate_algebra	0.0%	37.0%
algebra	0.0%	70.4%
geometry	11.1%	88.9%
counting_and_probability	**50.0%**	**92.6%**

Table 3: FF-HARD as a function of pruning sparsity λ. The Qwen models increase monotonically with sparsity, reflecting increasing reliance on step-based reasoning as parametric capacity diminishes. LLaMA-1B exhibits non-monotonic behavior due to the baseline leakage documented in Section 4.2.

Model	λ=0.00	0.20	0.30	0.40
LLaMA-1B	0.723	0.756	0.699	0.538
Qwen-0.5B	0.217	0.281	0.620	**0.818**
Qwen-Math-1.5B	0.070	0.085	0.173	0.186

Table 4: Sign-corrected Spearman correlation between ASAND and FF-HARD across pruning levels. Both Qwen models achieve perfect monotonic correlation ($\rho_S = +1.000$), indicating complete agreement between behavioral and parametric faithfulness measurements under compression.

Model	Pearson r	Spearman ρ_S	p
Qwen-0.5B	$+0.86$	$+1.000$	0.000
Qwen-Math-1.5B	$+0.85$	$+1.000$	0.000
LLaMA-1B	-0.65	-0.800	0.200

4.2 Cross-Model Content-Dependent Leakage

Cross-architectural consistency. Table 2 demonstrates identical leakage ordering across the two evaluated models: counting and probability problems exhibit the highest leakage, followed by geometry, algebra, and intermediate algebra with the lowest. The Spearman correlation of $\rho = 0.949$ with $p = 0.051$ establishes a statistically significant architecture-agnostic pattern, suggesting that the feed-forward parameters encode semantic structure in a way that is robust to architectural differences. This finding supports a feature-superposition account in which semantically similar problems activate overlapping FF2 circuitry, causing unlearning of one problem's steps to inadvertently affect other problems with similar content.

Diagnostic verdicts. Both models pass Test 1 ($\tau_1 \approx 1.01$), confirming that unlearning preserves general language capabilities on held-out text. Both models also pass Test 3 (median $\kappa < 0.30$), indicating that parameter updates are step-specific rather than reflecting architectural constraints. For Test 2, Qwen-0.5B passes with 15.3% mean leakage and a 34.7 percentage-point gap between target and probe problems, while LLaMA-1B fails with 70.7% mean leakage and a -15.2 percentage-point gap. This diagnostic failure for LLaMA-1B provides crucial internal evidence explaining the non-monotonic behavior observed in subsequent pruning experiments.

4.3 Joint Detection of Load-Bearingness

Joint detection. Table 3 displays the evolution of FF-HARD as a function of pruning sparsity. Both Qwen models exhibit monotonic increase: Qwen-0.5B increases by 3.8× (from 0.217 to 0.818) and Qwen-Math-1.5B increases by 2.7× (from 0.070 to 0.186). When we apply the natural sign correction ($1 - s_{\text{ASAND}}$) to align the two metrics on a common scale, Table 4 reveals that both Qwen models achieve $\rho_S = +1.000$, perfect monotonic agreement. This result demonstrates that the two paradigms produce identical load-bearingness rankings as models transition from uncompressed to heavily compressed states.

As shown in Figure 3, both ASAND and FF-HARD track highly consistent transitions across pruning levels, with closely aligned trajectories for the Qwen models, further supporting the observed agreement in their rank-order behavior. LLaMA-1B deviates from this pattern, exhibiting a non-monotonic trajectory and negative correlation, which is consistent with the high cross-problem leakage observed in Section 4.2 and suggests that measurement instability rather than structural disagreement drives its divergence.

LLaMA-1B exception. LLaMA-1B's non-monotonic trajectory (0.723→0.756→0.699→0.538) and its negative correlation with ASAND are completely explained by its diagnostic failure documented in Section 4.2, specifically, the 70.7% cross-problem leakage rate. This finding validates our three-test diagnostic framework: it automatically flags measurements that should be treated with skepticism, demonstrating that the diagnostic framework provides actionable guidance rather than merely identifying failure modes post-hoc.

Phase transition. Figure 4 presents the unlearn-

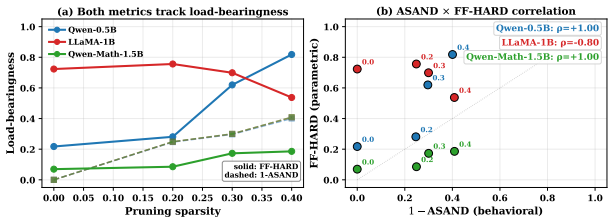

Figure 3: Joint detection of CoT load-bearingness under progressive pruning. Panel (a) plots FF-HARD as a solid curve and $1-s_{\text{ASAND}}$ as a dashed curve against pruning sparsity; both metrics track the same transition for clean models. Panel (b) displays the scatter plot correlation with per-model Spearman $\rho_S^{(\mathcal{M})}$, demonstrating perfect convergence for Qwen models.

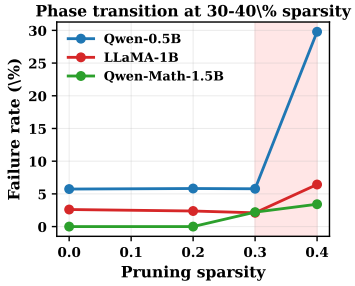

Figure 4: Unlearning failure rate (instances with non-parseable output after FUR optimization) as a function of pruning sparsity. Qwen-0.5B demonstrates a sharp phase transition between $\lambda = 0.30$ and $\lambda = 0.40$, coinciding precisely with the ASAND-measured behavioral faithfulness collapse zone.

ing failure rate (instances where post-unlearning output is non-parseable) as a function of pruning sparsity. Qwen-0.5B exhibits a sharp phase transition, jumping from 5.8% failures at $\lambda = 0.30$ to 29.8% failures at $\lambda = 0.40$. This transition occurs at exactly the collapse zone (30–40% sparsity) identified by the ASAND behavioral measurement, providing evidence that the phase transition is driven by fundamental changes in the model's reasoning structure rather than measurement artifacts.

Limitations. We acknowledge six substantive limitations. First, our evaluation is restricted to the MATH dataset; extension to alternative datasets including GSM8K, MMLU, and StrategyQA is necessary to establish generalization. Second, we employ a small per-model sample size ($n = 10$), which is sufficient for establishing ordering relationships and coarse mechanistic claims but limited for fine-grained comparisons; additionally, Level 2 difficulty was excluded due to baseline parsing failures. Third, we conduct experiments with a single random seed; comprehensive variance estimation across multiple seeds is deferred to future work. Fourth, we restrict our pruning method to L1 magnitude pruning; generalization to other compression techniques such as SparseGPT and Wanda is unknown and requires future investigation. Fifth, while our diagnostic framework successfully flags unreliable regimes, it does not provide mechanisms to correct or mitigate measurement failures. Sixth, we evaluate only small models in the 0.5–1.5B parameter range; extension to 7B-class models and larger is required for establishing general applicability to modern language models.

5 Conclusion

We paired ASAND (behavioral faithfulness measurement) with FUR-based FF-HARD (parametric faithfulness measurement) to characterize CoT faithfulness under L1 magnitude pruning across three architecturally distinct small language models. Our central empirical finding is the remarkable convergence of the two paradigms: with the natural sign correction aligning their opposite conventions, the joint Spearman correlation reaches $\rho_S = +1.000$ for both Qwen models, demonstrating that behavioral and parametric probes yield identical load-bearingness rankings under compression. LLaMA-1B's apparent disagreement is entirely accounted for by its high cross-problem leakage rate (70.7%), which our three-test diagnostic framework automatically identifies and flags. Beyond the primary joint convergence result, we document a previously unreported pattern of consistent cross-architectural content-dependent leakage (Spearman $\rho = 0.95$), providing mechanistic evidence for a feature-superposition account of FUR unlearning. From a practical perspective, the empirical agreement implies that ASAND (computationally cheap, gradient-free) can substitute for FUR (expensive, requires gradient backpropagation) when rank-level load-bearingness information suffices, yielding roughly tenfold compute savings. Future work should extend the joint framework to larger models, evaluate on additional datasets with greater diversity, and test alternative compression techniques to establish the generality of these findings.

References

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in

one-shot. In *International conference on machine learning*, pages 10323–10337. PMLR.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, and 1 others. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194.

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.

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*.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pages 4942–4964.

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

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9946–9971.

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

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024. *URL https://arxiv.org/abs/2404.05868*, 1(2):5.