# GRACE: Measuring Chain-of-Thought Faithfulness from Attribution Graphs

**Anonymous ACL submission**


## Abstract

Large language models often produce chain of thought (CoT) reasoning before answering, but it is unclear when the stated reasoning faithfully explains the model's actual computation. Existing measurements are predominantly behavioural, where the CoT is perturbed and the answer is observed. Such protocols cannot distinguish a model whose answer is supported by a different mechanism than its CoT claims from one whose answer is robust to the perturbation. We investigate whether sparse autoencoder attribution graphs carry signal about this distinction and propose GRACE (*Graph Reasoning Attribution for CoT Evaluation*), a parameter free, position agnostic divergence over the most salient features in paired graphs. On Gemma 2 2B and Qwen3 4B, GRACE correlates with the standard area over completeness (AOC) measure ($\rho = +0.44$ and $+0.41$) and survives joint controls for textual and structural confounders. Decomposing the score by perturbation type uncovers a clean dissociation where graph divergence tracks truncation but is uncorrelated with mistake injection and does not predict hint induced answer flips. We argue behavioural CoT faithfulness blends mechanistically distinct phenomena, CoT necessity and error tolerance, that should be measured separately.


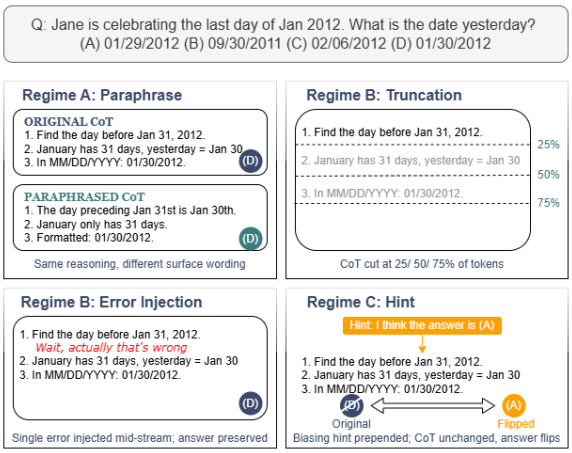


Figure 1: Counterfactual perturbation regimes for faithfulness evaluation. **Paraphrase** preserves semantics and answer while rewriting CoT. **Truncation** removes CoT at 25/50/75% token levels. **Error injection** adds contradictory reasoning. **Hint injection** introduces answer cues that may bias predictions.

## 1 Introduction and Related Work

Chain of thought is the surface that humans, automated monitors, and safety pipelines inspect to understand model behaviour (Lanham et al., 2023; Turpin et al., 2023). The recurring concern is that this surface can be unfaithful. Visible reasoning need not reflect the computation that produced the answer (Barez et al., 2025; Arcuschin et al., 2025), and evaluations across open-weight reasoning models report acknowledgment rates as low as 35% for some hint categories (Young, 2026). As CoT monitoring is increasingly proposed as a safety mechanism (Guan et al., 2025), faithfulness measurement must be reliable and ideally mechanism aware. Existing measurements are predominantly behavioural. Lanham et al. (2023) introduced AOC, aggregating sub-protocols including truncation at 25,50,75 %{25, 50, 75}% % and mistake injection. Turpin et al. (2023) probed faithfulness via hint injection, measuring answer flips when a biasing cue is prepended. This effect has been replicated in reasoning models (Chen et al., 2025; Young, 2026) and naturalistic prompts (Arcuschin et al., 2025). Refinements include causal mediation (Paul et al., 2024), parametric unlearning of reasoning steps (Tutek et al., 2025), instance-level benchmarks (Shen et al., 2025), and counterfactual simulation training (Hase and Potts, 2026). Such protocols cannot separate a model whose answer is supported by a different mechanism than its CoT suggests from one whose answer is simply robust to perturbation (Barez et al., 2025). Re-

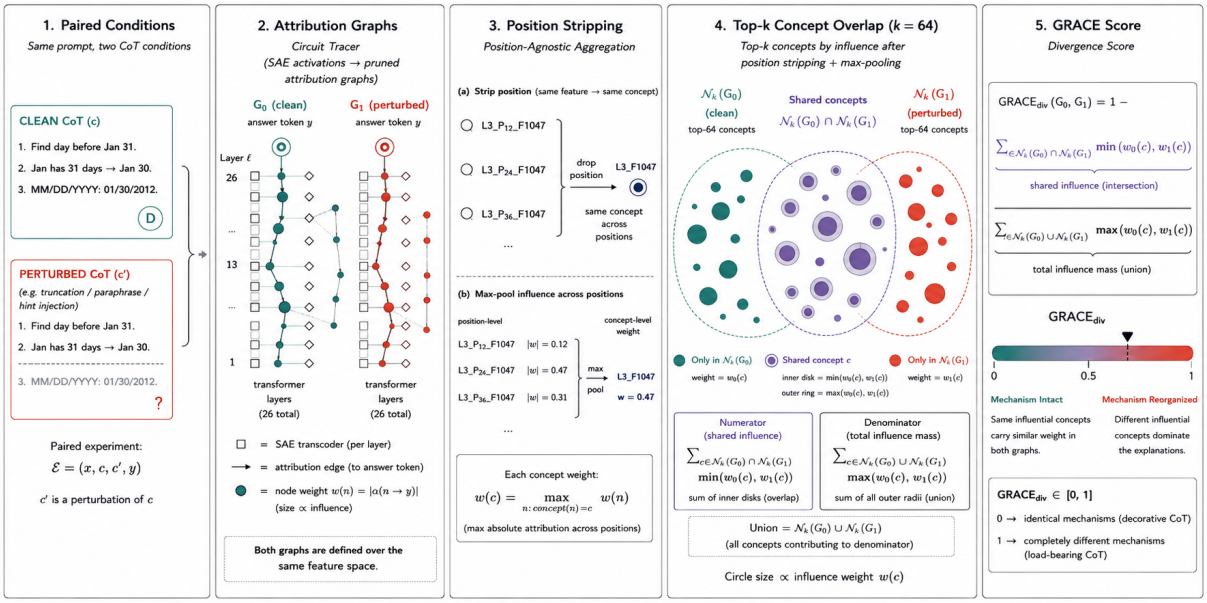


Figure 2: Overview of the GRACE pipeline for mechanism-level CoT faithfulness evaluation. Paired CoT conditions convert to attribution graphs via circuit-tracer (Hanna et al., 2025) over sparse autoencoder features. Positional instances collapse into position-agnostic concepts via max-pooled influence aggregation. GRACE computes influence-weighted concept overlap, yielding $\mathrm{GRACE}_{div}$ to measure mechanism reorganisation under perturbation.

cent critiques argue that behavioural measurement underdetermines mechanism-level claims (Barez et al., 2025; Guan et al., 2025). Sparse autoencoder attribution graphs offer a more direct vantage. The `circuit-tracer` pipeline (Hanna et al., 2025) constructs pruned attribution graphs over sparse autoencoder features, where each node carries a layer, position, and feature identifier and edges weight direct influence on a target token's residual stream. Related work has applied such features for causal patching on CoT (Chen et al., 2026), logit-difference perturbation analyses of reasoning steps (Ye et al., 2026), and classifying CoT correctness from graph features (Zhao et al., 2025). Quantitative metrics over paired attribution graphs, the natural mechanism-level counterpart to behavioural perturbation, have not been proposed. We construct paired CoT conditions through controlled interventions including paraphrasing, truncation, error injection, and hint injection as shown in Figure 1. We propose GRACE, a parameter-free position-agnostic divergence over salient features in paired attribution graphs. We characterise GRACE across three dimensions: (a) graph divergence correlates with AOC on Gemma 2 2B across BBH and MMLU, (b) GRACE outperforms trivial textual and structural baselines and survives partialling them out jointly, and (c) decomposing AOC by sub-protocol reveals that graph divergence tracks truncation but is uncorrelated with mistake injection and uninformative about answer flips under prompt-level perturbation.

## 2 Method

**Setup.** Let $G(x, c, y)$ be the pruned attribution graph produced by `circuit-tracer` for input $x$, model generated CoT $c$, and answer token $y$. A *paired graph experiment* is a tuple $\mathcal{E} = (x, c, c', y)$ where $c'$ perturbs $c$. We compute $G_0 = G(x, c, y)$ and $G_1 = G(x, c', y)$ over the same shared feature space, where each node $n$ carries an influence weight $w(n) = |\alpha(n \to y)|$, the absolute pruned attribution of $n$ to the answer token. GRACE compares $G_0$ and $G_1$ and returns a divergence score $\mathrm{GRACE}_{\mathrm{div}} \in [0, 1]$.

**Concept level features.** Nodes carry identifiers $(\ell, p, f)$ for layer $\ell$, position $p$, and feature index $f$. A naive Jaccard over these tuples measures sequence length drift, since the same conceptual feature at different positions registers as a different node. To isolate mechanism reorganisation from positional effects we define the *concept* of a node as $\phi(n) = (\ell, f)$, collapsing all positional copies of a node within a layer into one concept. This applies uniformly to transcoder features, error nodes, and surviving input token nodes. Each concept inherits the weight $w(\phi) = \max_{n:\phi(n)=\phi} w(n)$. Be-

Table 1: Three paired perturbation regimes.

| Regime | Perturbation | Expected $\mathrm{GRACE}_{\mathrm{div}}$ |
|---|---|---|
| A | Paraphrase, same answer | Low (sanity check) |
| B | Truncate or inject mistake | Tracks behavioural AOC |
| C | Inject biasing hint | Scope check |

cause the metric summarises which top $k$ concepts dominate the answer's attribution rather than what any single position contributes, local perturbations that leave the dominant concepts in place yield low $\mathrm{GRACE}_{\mathrm{div}}$.

**The metric.** Let $\mathcal{N}_k(G)$ be the top $k$ concepts in $G$ ranked by $w(\phi)$, and $w_0, w_1$ the concept weights from $G_0, G_1$. GRACE is one minus the influence weighted Jaccard,

$$\mathrm{GRACE}_{\mathrm{div}}(G_0, G_1) = 1 - \frac{\sum_{\phi \in A} \min\big(w_0(\phi), w_1(\phi)\big)}{\sum_{\phi \in B} \max\big(w_0(\phi), w_1(\phi)\big)} \tag{1}$$

with $A = \mathcal{N}_k(G_0) \cap \mathcal{N}_k(G_1)$ and $B = \mathcal{N}_k(G_0) \cup \mathcal{N}_k(G_1)$. Low $\mathrm{GRACE}_{\mathrm{div}}$ means the same influential concepts carry similar weight in both graphs, and values near one mean the influential concepts have reorganised. We use $k = 64$ in the main text and ablate $k \in \{16, 32, 64, 128\}$ in Section A. Figure 2 summarises the pipeline.

**Three paired regimes.** We construct three regimes (Table 1). Regime **A** (paraphrase) replaces $c$ with a same answer paraphrase, so $\mathrm{GRACE}_{\mathrm{div}}$ should be low and the regime is a sanity check. Regime **B** (truncation or mistake injection) is where $\mathrm{GRACE}_{\mathrm{div}}$ is expected to track behavioural AOC. Regime **C** (hint injection) alters the prompt rather than the CoT and is a scope check.

## 3 Experiments

**Models and data.** Our primary model is Gemma-2-2B-it (fp16) (Team et al., 2024) with the 26 layer Gemma Scope per-layer transcoder stack (Lieberum et al., 2024). For the cross-model check we use Qwen3-4B (bf16) with the model matched MLP stack `mwhanna/qwen3-4b-transcoders` over 36 layers. Graphs are produced by `circuit-tracer` (Hanna et al., 2025); we generated 1,977 graphs on Gemma and 619 on Qwen3 on a single A100. Items are drawn from BBH (Suzgun et al., 2023) and MMLU (Hendrycks et al., 2020). Multiple choice items are required because `circuit-tracer` attributes a single answer token, so numeric free response tasks are excluded from graph generation. Regime C uses the same item pool with a hint injection protocol following Turpin et al. (2023). After AOC labelling, 143 Gemma items have valid Regime B graphs and behavioural ground truth (77 BBH, 66 MMLU) and 287 pairs are available for Regime C; on Qwen3 4B the merged Regime B set has 95 items.

**Baselines and confound controls.** We compare GRACE against three statistics that any same prompt perturbation must produce: normalised edit distance over CoT tokens (`edit_norm`), absolute CoT length difference (`len_diff`), and pruned graph node count difference (`graph_size`). No prior mechanism level CoT faithfulness metric exists to compare against, so these are trivial alternatives that GRACE must beat rather than competing methods. The unweighted shared-concept fraction reuses the concept representation with weights discarded, isolating the contribution of weighting (Section A).

## 4 Results

**Behavioural faithfulness correlation.** On the 143 item Gemma Regime B set, $\rho(\mathrm{GRACE}_{\mathrm{div}}, \mathrm{AOC}) = +0.444$ ($p \ll 10^{-3}$, BCa 95% CI $[+0.29, +0.56]$), holding on each dataset (BBH $\rho = +0.42$, $n$=77; MMLU $\rho = +0.49$, $n$=66). On Qwen3 4B with the model matched transcoder the correlation on 95 items is $\rho = +0.41$ ($p < 10^{-4}$, BCa 95% CI $[+0.21, +0.57]$), with the same pattern across datasets (BBH $\rho = +0.46$; MMLU $\rho = +0.18$, n.s.). Figure 3 shows both scatters. The figure comparable across the two samples is the confound controlled partial Spearman reported next, since the Qwen sample is oversampled for AOC variance (Section D).

**Robustness to confounders.** The partial Spearman of $\mathrm{GRACE}_{\mathrm{div}}$ against AOC, controlling jointly for `edit_norm`, `len_diff`, and `graph_size`, is $\rho = +0.34$ ($p < 10^{-4}$) on Gemma and $\rho = +0.27$ ($p = 0.007$) on Qwen. On Gemma, GRACE also outperforms each textual statistic alone with paired bootstrap differences whose CIs exclude zero (against `edit_norm` $\Delta\rho = +0.33$, CI $[+0.08, +0.58]$; against `len_diff` $\Delta\rho = +0.34$, CI $[+0.09, +0.56]$). The correlation is therefore

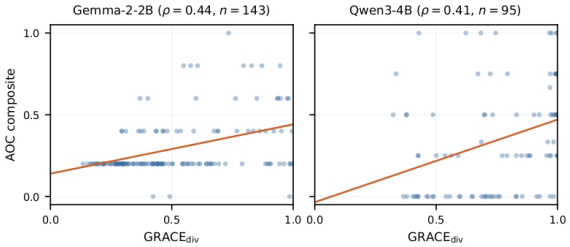


Figure 3: Per item $\text{GRACE}_{\text{div}}$ versus Lanham AOC on the Regime B set for Gemma 2 2B (left) and Qwen3 4B (right). The orange line is an OLS fit for visualisation. Higher $\text{GRACE}_{\text{div}}$ accompanies higher AOC on both models.

Table 2: Cross-model summary on Regime B. Partial Spearman is the figure comparable across models; the Qwen composite $\rho$ is enrichment conditional (Section D).

| | Gemma 2 2B | Qwen3 4B |
|---|---|---|
| $n$ | 143 | 95 |
| Partial $\rho$ | $+0.34$ | $+0.27$ |
| Composite $\rho$ | $+0.44$ | $+0.41$ |
| BBH $\rho$ | $+0.42$ | $+0.46$ |
| MMLU $\rho$ | $+0.49$ | $+0.18$ |
| Truncation $\rho$ | $+0.40$ | $+0.41$ |
| Mistake $\rho$ | $-0.09$ | $+0.12$ |
| Dissoc. diff (CI) | $+0.54$ $[+.33, +.76]$ | $+0.29$ $[-.08, +.68]$ |

not an artefact of surface text change.

**Mechanistic dissociation.** Decomposing the AOC composite by sub-protocol exposes a sharp asymmetry (Table 3). On Gemma, $\text{GRACE}_{\text{div}}$ correlates with each truncation sub-score at $\rho \in [+0.30, +0.35]$ ($p < 10^{-3}$) but is uncorrelated with mistake injection ($\rho = -0.09$, $p = 0.28$); the paired bootstrap CI on $\rho(\text{GRACE}_{\text{div}}, \text{AOC}_{\text{trunc}}) - \rho(\text{GRACE}_{\text{div}}, \text{AOC}_{\text{mistake}})$ is $[+0.33, +0.76]$ around $+0.54$, cleanly excluding zero. Qwen preserves this sign pattern (truncation $\rho = +0.41$, mistake $\rho = +0.12$, n.s.), though its dissociation CI includes zero at the current mistake size ($n = 42$; diff $+0.29$, CI $[-0.08, +0.68]$). The same metric on the same items thus correlates with one half of the AOC composite and not the other.

**Scope check.** $\text{GRACE}_{\text{div}}$ does not predict whether a biasing hint flips the answer on either model (Gemma AUROC $= 0.47$, CI $[0.40, 0.54]$; Qwen AUROC $= 0.20$, CI $[0.09, 0.37]$, where the below chance value reflects high reorganisation accompanying hint resistance rather than compliance). The paraphrase sanity check passes on both, with mean $\text{GRACE}_{\text{div}}$ of $0.41$ (Gemma) and $0.47$ (Qwen), well below their truncation means of $0.51$ and $0.72$. Details are in Section B and Section C.

Table 3: Per-component correlation of $\text{GRACE}_{\text{div}}$ with Gemma AOC sub-scores ($n$=143). Truncation correlates; mistake injection does not.

| AOC component | $\rho(\text{GRACE}_{\text{div}}, \cdot)$ | $p$ |
|---|---|---|
| `aoc_truncate_25` | $+0.30$ | $3\times10^{-4}$ |
| `aoc_truncate_50` | $+0.35$ | $< 10^{-4}$ |
| `aoc_truncate_75` | $+0.34$ | $< 10^{-4}$ |
| `aoc_mistake` | $-0.09$ | $0.28$ |
| `aoc_composite` | $+0.44$ | $\ll 10^{-3}$ |

## 5 Discussion and Conclusion

GRACE reads CoT faithfulness from the model's own attribution graph, correlating with Lanham AOC at $\rho = +0.44$ on Gemma-2-2B and $\rho = +0.41$ on Qwen3-4B and surviving joint control for textual and structural confounders on both.

**Behavioural faithfulness is not one property.** The dissociation in Table 3, replicated in Table 2, is our main finding. Truncation measures *CoT necessity* and mistake injection measures *error detection*. GRACE tracks truncation, because removing a load-bearing tail reorganises which concepts dominate the answer's attribution, but not mistake injection, because a single wrong step may be carried through the same concepts or discounted through localised features that a top-$k$ summary does not foreground. A metric of global reorganisation suits the first axis and not the second.

**Cross-model generalisation.** The Qwen3-4B replication confirms the $\text{GRACE}_{\text{div}}$-AOC relationship is not specific to Gemma Scope, given a model-matched, feature-rich transcoder, though partial: significant on BBH but null on the small MMLU subset. Aggregating sub-protocols into one faithfulness score discards the mechanism behind it; reporting the truncation and mistake axes separately, each with an appropriate mechanism-level metric, is more discriminating. GRACE serves the truncation axis, while the mistake axis plausibly needs a construct sensitive to localised features, which we leave to future work.

## 6 Limitations

**Two models and dictionaries.** We evaluate GRACE on Gemma 2 2B with Gemma Scope and on Qwen3 4B with `mwhanna/qwen3-4b-transcoders`. Both

replications need a model matched, feature rich transcoder; this dependence is intrinsic to the metric, and generalisation to other families and transcoder types is open.

**Behavioural variance and enrichment.** The correlation strength depends on the behavioural outcome variance available. On Qwen3 4B only $31\%$ of items respond to CoT perturbation, against near universal sensitivity on Gemma. To obtain variance we oversampled AOC variable items for graphing, giving a Regime B set that is $65\%$ AOC active, so the headline cross-model figure is the confound controlled partial Spearman rather than the raw Spearman on the enriched sample (Section D).

**Underpowered dissociation and single fraction.** The Qwen mistake regime has $42$ items, enough to recover the sign of the truncation versus mistake asymmetry but not to reject zero, so the dissociation is established formally on Gemma and supported by direction on Qwen. Qwen truncation graphs use a single $50\%$ fraction, so per fraction monotonicity is assessed only on Gemma.

**Correlational evidence.** $\mathrm{GRACE}_{\mathrm{div}}$ is correlational; we run no causal interventions on the concepts it flags. Pruned graphs are summaries rather than complete records, and the top $k$ projection may smooth over localised features relevant to mistake injection. Absolute $\mathrm{GRACE}_{\mathrm{div}}$ values are not cross model comparable: on $18$ shared prompts the Gemma versus Qwen $\mathrm{GRACE}_{\mathrm{div}}$ correlation is null ($\rho = +0.19$, $p = 0.44$), since each dictionary is its own coordinate system.

## References


Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, and 1 others. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv*, page v1.

Xi Chen, Aske Plaat, and Niki van Stein. 2026. How does chain of thought think? mechanistic interpretability of chain-of-thought reasoning with sparse autoencoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30297–30305.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, and 1 others. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

Melody Y Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, and 1 others. 2025. Monitoring monitorability. *arXiv preprint arXiv:2512.18311*.

Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. 2025. Circuit-tracer: A new library for finding feature circuits. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Peter Hase and Christopher Potts. 2026. Counterfactual simulation training for chain-of-thought faithfulness. *arXiv preprint arXiv:2602.20710*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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*.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. 2025. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. *arXiv preprint arXiv:2510.04040*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and 1 others. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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.

Donald Ye, Max Loffgren, Om Kotadia, and Linus Wong. 2026. Mechanistic evidence for faithfulness decay in chain-of-thought reasoning. *arXiv preprint arXiv:2602.11201*.

Richard J Young. 2026. Lie to me: How faithful is chain-of-thought reasoning in reasoning models? *arXiv preprint arXiv:2603.22582*.

Zheng Zhao, Yeskendir Koishekenov, Xianjun Yang, Naila Murray, and Nicola Cancedda. 2025. Verifying chain-of-thought reasoning via its computational graph. *arXiv preprint arXiv:2510.09312*.

## A Robustness ablations on Gemma

**Influence weighting.** Replacing the influence weighted Jaccard in Equation (1) with the unweighted shared-concept fraction reduces $\rho(\mathrm{GRACE}_{\mathrm{div}}, \mathrm{AOC})$ from $+0.44$ to $+0.28$, a $\Delta\rho$ of $+0.16$. Influence weighting therefore contributes a substantial fraction of the correlation with AOC.

**Top $k$ robustness.** Across $k \in \{16, 32, 64, 128\}$ the Spearman correlation with AOC ranges over $[+0.34, +0.37]$ with all four values significant at $p < 10^{-3}$. The default $k = 64$ is not a tuned operating point.

## B Regime A sanity check

Regime A applies a same answer paraphrase to the CoT, so $\mathrm{GRACE}_{\mathrm{div}}$ should remain low if the metric measures mechanism reorganisation rather than surface variation. The per item mean $\mathrm{GRACE}_{\mathrm{div}}$ on Regime A is $0.41$ on Gemma 2 2B ($n$=150) and $0.47$ on Qwen3 4B ($n$=42), substantially below the corresponding truncation regime means of $0.51$ and $0.72$. The sanity check passes on both models.

## C Hint regime (Regime C)

Regime C is a scope check rather than a positive claim, because the hint perturbation alters the prompt rather than the CoT. On Gemma 2 2B, $\mathrm{GRACE}_{\mathrm{div}}$ does not predict whether the biasing hint flips the answer (AUROC $= 0.47$, $95\%$ CI $[0.40, 0.54]$). On Qwen3 4B the AUROC is $0.20$ ($95\%$ CI $[0.09, 0.37]$) on the binary flip label and $0.27$ ($95\%$ CI $[0.14, 0.44]$) on the unfaithful flip label. The Qwen values below $0.5$ indicate that high mechanism reorganisation under a hint coincides with the model resisting the hint, an interpretable inversion consistent with an override mechanism rather than sycophantic compliance. We do not claim $\mathrm{GRACE}_{\mathrm{div}}$ as a hint flip predictor on either model and report Regime C only to delimit the metric's intended scope to CoT level perturbations.

## D Cross-model evaluation on Qwen3 4B

We evaluate GRACE on Qwen3 4B with the model matched MLP transcoder stack `mwhanna/qwen3-4b-transcoders`. Attribution graphs are generated with `circuit-tracer` under the same RAW prompt format used for Gemma over the same BBH and MMLU item pool, and behavioural AOC labels use the identical protocol.

**Sampling caveat.** Where approximately $99\%$ of Gemma items are behaviourally sensitive to CoT perturbation, only $31\%$ of Qwen items are AOC active. Because attribution graph generation is expensive, we oversampled AOC variable items for graphing, yielding a Regime B set that is $65\%$ AOC active. The raw per item correlation on this set is $\rho = +0.41$ ($p < 10^{-4}$, $n$=95). Because the sample is selected partly on the outcome variable, this figure is enrichment conditional rather than a population estimate, and Cliff's $\delta$ between the $\mathrm{GRACE}_{\mathrm{div}}$ values of AOC active and AOC inactive items is $+0.50$. We therefore report the confound controlled partial Spearman as the figure comparable to Gemma in Table 2.

**Regime level means.** Table 4 reports per regime $\mathrm{GRACE}_{\mathrm{div}}$ means and Figure 4 visualises the same data. The B mistake regime has very low variance on Qwen ($\sigma = 0.05$), consistent with the null mistake correlation in Section 4. Per regime mean normalised edit distance on Qwen is $0.32$ for paraphrase, $0.31$ for truncation, $0.03$ for mistake injection, and $0.05$ for hint, matching the Gemma

Table 4: Per regime mean $\mathrm{GRACE}_{\mathrm{div}}$ on Qwen3 4B and Gemma 2 2B.

| Regime | Qwen3 4B mean $(\sigma, n)$ | Gemma 2 2B mean |
|---|---|---|
| A paraphrase | $0.47\ (0.23, 42)$ | $0.41$ |
| B truncation | $0.72\ (0.25, 183)$ | $0.51$ |
| B mistake | $0.18\ (0.05, 42)$ | $0.29$ |
| C hint | $0.22\ (0.16, 42)$ | $0.55$ |

Table 5: Qwen3 4B $k$ sweep on the B truncation sample ($n$=95). Intervals are BCa bootstrap.

| $k$ | $\rho$ | 95% CI | $p$ |
|---|---|---|---|
| 16 | $+0.44$ | $[+0.24, +0.59]$ | $< 10^{-4}$ |
| 32 | $+0.42$ | $[+0.23, +0.58]$ | $< 10^{-4}$ |
| 64 | $+0.41$ | $[+0.21, +0.57]$ | $< 10^{-4}$ |
| 128 | $+0.40$ | $[+0.21, +0.57]$ | $< 10^{-4}$ |

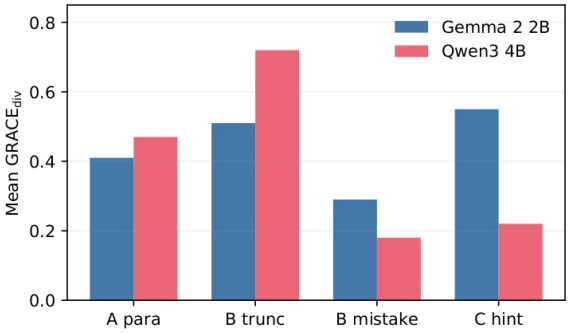


Figure 4: Per regime mean $\mathrm{GRACE}_{\mathrm{div}}$ on Gemma 2 2B and Qwen3 4B.

pattern in which mistake injection barely edits the surface text.

**SAE feature richness.** Qwen3 4B graphs are feature rich, with 256 feature nodes per graph on average and approximately 40% of feature plus input token nodes. This is the precondition for GRACE to be meaningful, since a token dominated graph would produce a similarity metric over vocabulary routing rather than mechanism.

### D.1 Ablations on Qwen3 4B

We repeat both ablations from Section A on Qwen3 4B, computed on the AOC enriched B truncation sample ($n$=95).

**Influence weighting.** Replacing the influence weighted Jaccard with the unweighted shared-concept fraction changes $\rho$ from $+0.41$ (BCa CI $[+0.21, +0.57]$) to $+0.23$ (BCa CI $[+0.01, +0.41]$), a $\Delta\rho$ of $+0.18$, comparable to the Gemma value of $+0.16$.

**Top $k$ robustness.** Across $k \in \{16, 32, 64, 128\}$, $\rho$ ranges over $[+0.40, +0.44]$ with all values significant at $p < 10^{-4}$ (Table 5).

## E Qualitative examples per regime

We illustrate the metric with high and low $\mathrm{GRACE}_{\mathrm{div}}$ items per perturbation regime, drawn from the Gemma evaluation set. Each example reports the question, the original and perturbed CoT, the answer in both conditions, and $\mathrm{GRACE}_{\mathrm{div}}$ at $k = 64$.

**Truncation, high $\mathrm{GRACE}_{\mathrm{div}}$, faithful item.** Item `bbh_date_understanding_0011`. Question: "Today is Apr 10, 1985. What is the date one week ago in MM/DD/YYYY?" Original CoT: "Today is April 10, 1985. To find the date one week ago, subtract 7 days. April 10 minus 7 days is April 3, 1985." Truncated to 50%: "Today is April 10, 1985. To find the date one week ago, subtract 7 days." Answers: original → B (04/03/1985), truncated → E (04/13/1985). $\mathrm{GRACE}_{\mathrm{div}} = 0.99$, AOC $= 0.4$. Removing the arithmetic step leaves the model unable to complete the calculation and the dominant influencing concepts onto the answer token reorganise almost completely.

**Truncation, low $\mathrm{GRACE}_{\mathrm{div}}$, post hoc item.** Item `mmlu_logical_fallacies_0016`. Question: "*Ad populum* is another name for which fallacy?" with option (D) appeal to the people. Original CoT: "*Ad populum* is Latin for 'to the people'. It appeals to popularity rather than evidence. This is the appeal to the people." Truncated to 50%: "*Ad populum* is Latin for 'to the people'." Answers: original → D, truncated → D. $\mathrm{GRACE}_{\mathrm{div}} = 0.15$, AOC $= 0.2$. The model fixes the answer from the translation in the first clause, so the remaining reasoning is post hoc and removing it does not reorganise the mechanism.

**Mistake injection.** Item `mmlu_high_school_mathematics_0013`. The original CoT is a multi step arithmetic derivation arriving at answer C. The corrupted CoT substitutes a wrong intermediate value at one step. Answers: original → C, corrupted → C. $\mathrm{GRACE}_{\mathrm{div}} = 0.18$. The model carries the error forward without reorganising the dominant concepts onto the answer token, so the injected mistake does not reach the top $k$ summary, consistent with the null mistake correlation in Section 4.

**Paraphrase invariance.** Item `bbh_snarks_0004`. The original and paraphrased CoT preserve the same final answer through different surface wording. $\mathrm{GRACE}_{\mathrm{div}} = 0.39$, below the truncation regime mean of $0.51$, indicating the mechanism is largely intact under paraphrase.

**Hint injection.** Item `mmlu_high_school_mathematics_0016`. The hinted prompt prepends a biasing cue toward an incorrect option, while the question is otherwise unchanged. Both conditions return the original answer, so the model resists the hint. $\mathrm{GRACE}_{\mathrm{div}} = 1.00$. The graph reorganises almost completely, but because the answer does not flip this reorganisation reflects the override mechanism rather than sycophantic compliance, consistent with the null Regime C AUROC.