What Does Alignment Cost? The Structural Brittleness of Chain-of-Thought Reasoning

Joanna Hao
University of Alberta
jjoannahao@gmail.com

Shanduojiao Jiang
Algoverse AI Research

Asish Nakka
Pennsylvania State University

Abstract

While Chain-of-Thought (CoT) prompting enables Large Language Models to explicitly justify their predictions, the extent to which these textual rationales faithfully reflect internal computation remains unclear. We investigate the circuit-level impact of alignment by performing a strict within-family comparison of the 1B-parameter Llama 3 architecture (Base vs. Instruct). Executing dynamic circuit discovery and dual-direction resample ablation on unconstrained CoT traces across synthetic mathematical primitives and a GSM8K proxy, we find that foundation models possess highly redundant, self-repairing computational networks; completely corrupting their primary reasoning circuits yields a minimal performance drop (2.92%) due to the dynamic compensation of backup heads (the Hydra Effect). In contrast, the instruction-tuned model exhibits reduced structural redundancy, suffering more than double the degradation (6.79%) under identical perturbation. We formalize our observation as an "Alignment Tax on Redundancy": optimizing for human-preference compliance repurposes dormant backup circuits, centralizing mathematical routing and rendering the aligned model's reasoning pathways significantly more vulnerable to internal perturbation.

1 Introduction

Chain-of-Thought (CoT) prompting enables Large Language Models (LLMs) to decompose complex tasks into sequential intermediate steps (Wei et al., 2022). Yet, as these architectures enter high-stakes domains, reasoning trustworthiness has emerged as a critical bottleneck (Wang et al., 2025). Beyond standard hallucinations where models generate statistically likely but factually incorrect tokens, alignment-tuned LLMs exhibit a more insidious failure mode: sycophantic deception (Turpin et al., 2023). In these instances, a model may internally track the correct algorithmic state but actively output a contradictory rationale or target. This behavioral divergence raises a profound question regarding CoT faithfulness: to what extent do generated textual rationales causally determine a model's actual internal prediction?

The debate surrounding CoT faithfulness has heavily relied on behavioral and "hint-based" perturbations (Lanham et al., 2023). While some literature argues that LLMs frequently generate unfaithful, post-hoc rationalizations (Atanasova et al., 2023), recent pushback suggests that CoTs may actually be causally faithful, and that apparent unfaithfulness merely stems from the lossy compression or incomplete verbalization of internal states due to token limits (Zaman and Srivastava, 2025). To resolve this ambiguity, recent research has pivoted to mechanistic interpretability. Notably, Yeo et al. (2025) utilized causal mediation to demonstrate that Reinforcement Learning from Human Feedback (RLHF) increases the macro-level causal overlap between a model's rationale and its output. At the feature level, recent activation patching studies confirm that CoT prompting induces more modular and interpretable internal structures in high-capacity models (Chen et al., 2025). Simultaneously, Lu et al. (2026) proposed the "Decoupling Hypothesis," arguing that unfaithful models rely on parallel computational shortcuts that bypass the CoT entirely.

While these macro-level and theoretical evaluations provide crucial groundwork, our work investigates the exact, circuit-level mechanistic cost of alignment tuning. By pushing beyond observational overlap and testing the functional necessity of reasoning circuits via dynamic activation patching on unconstrained, naturally successful reasoning traces, we bypass the vulnerabilities of verbalization-based metrics (Zaman and Srivastava, 2025). Utilizing a strict within-family comparison of the Llama 3 architecture, we reveal that the apparent faithfulness of instruction-tuned models

Proceedings of the 4th Workshop on Towards Knowledgeable Foundation Models (KnowFM 2026), pages 25–33

masks a profound structural fragility. Ultimately, we provide quantitative evidence of an "Alignment Tax" in how optimizing for instruction compliance smears logical routing across diffuse attention networks, forcing a mechanistic decoupling of the model's explicit working memory from its final target prediction.

2 Related Work

2.1 Faithfulness in Chain-of-Thought Reasoning

The introduction of CoT prompting (Wei et al., 2022) catalyzed a massive subfield dedicated to understanding the reliability of generated rationales. Early investigations into rationale faithfulness frequently framed CoT as an interpretable window into model cognition, yet subsequent behavioral studies revealed significant vulnerabilities. Ye and Durrett (2022) and Atanasova et al. (2023) demonstrated that models often generate post-hoc rationalizations that do not accurately reflect the variables driving the final prediction. This was further expanded by Turpin et al. (2023) and Agarwal et al. (2024), who showed that instruction-tuned models suffer from extreme sycophancy, altering their rationales to match user-injected biases or constraints. Measuring this faithfulness, however, has proven methodologically fragile. Lanham et al. (2023) formalized perturbation-based metrics for CoT, but recent work by Zaman and Srivastava (2025) critiques these behavioral "hint-based" evaluations, arguing that a lack of explicit verbalization does not necessarily equate to unfaithfulness, highlighting the need for deeper, circuit-level interventions to prove causal decoupling.

2.2 Interpretability and Circuit Discovery

To bypass the limitations of behavioral observation, Mechanistic Interpretability seeks to reverse-engineer the computational graph of neural networks. Foundational work on the residual stream and attention mechanisms (Olsson et al., 2022) paved the way for precise circuit discovery. Causal tracing and activation patching techniques have successfully localized specific behaviors, such as factual recall (Meng et al., 2022) and indirect object identification (Wang et al., 2023), to discrete subgraphs of attention heads and MLPs. More recently, feature-level investigations using sparse autoencoders have begun mapping these causal structures during multi-step tasks, revealing that CoT induces modular, interpretable pathways whose causal information is widely distributed across the network (Chen et al., 2025).

Crucially, this distributed nature enables emergent self-repair. McGrath et al. (2023) documented the "Hydra Effect," demonstrating that ablating primary reasoning heads causes dormant, late-layer circuits to dynamically compensate. Evaluating these highly redundant networks requires immense methodological precision; standard zero-ablation frequently pushes the model's residual stream out-of-distribution (Heimersheim and Nanda, 2024), while continuous patching can induce geometric interpretability illusions (Makelov et al., 2024). We adapt our dual-direction resample ablation methodology specifically to account for these structural artifacts while measuring causal load.

2.3 Alignment and Reasoning Topologies

Recent literature has begun bridging mechanistic interpretability with the behavioral artifacts of Reinforcement Learning from Human Feedback (RLHF). While earlier work suggested alignment-tuning generally increases the causal overlap between internal traces and final outputs compared to unaligned baselines (Yeo et al., 2025), the architectural cost of this multi-objective optimization remains underexplored. Foundational alignment theories propose that RLHF acts as a "thin wrapper" over pre-trained capabilities (Zhou et al., 2023), yet enforcing conversational formatting, safety guardrails, and instruction compliance inherently consumes representational capacity.

Our work extends this intersection by mechanistically mapping how alignment training alters the fundamental topology of reasoning. Rather than viewing unfaithfulness purely as learned deception, we utilize targeted activation patching to demonstrate how the capacity constraints of alignment tuning systematically degrade the distributed, self-repairing circuits (the Hydra Effect) native to foundation models, resulting in an "Alignment Tax" that renders the model's explicitly generated working memory structurally brittle.

3 Methodology

Our experimental pipeline consists of eliciting unconstrained, naturally successful Chain-of-Thought (CoT) reasoning traces, performing dynamic circuit discovery via activation patching, and applying resample ablation to quantify the causal load of spe-

cific attention-routing mechanisms.

3.1 Models and Architectural Assumptions

To isolate the mechanistic impact of alignment interventions from general architectural variances, we conduct a controlled within-family comparison. We evaluate a foundation base model alongside its instruction-tuned counterpart (specifically, the Llama-3.2-1B and Llama-3.2-1B-Instruct models). While their attention routing mechanisms utilize Grouped-Query Attention (GQA), the computational backbone remains consistent. At each layer l, attention mechanisms read from and write to a central residual stream. Crucially for our intervention methodology, even when keys and values are shared across heads, the output of each individual head, $O^{(l,h)}$, is computed and projected back into the residual stream independently. This structural uniformity allows our activation patching to target the localized output projections of individual heads, treating them as discrete causal components that can be isolated across both the base and instruction-tuned states.

3.2 Tasks and Dataset Construction

We evaluate mathematical and logical reasoning using a large-scale synthetic dataset of deterministic tasks that isolate sequential updating, boolean aggregation, multi-path routing, and latent feature gating. Let a reasoning task be defined by an input prompt X and a definitive ground-truth target Y. Standard few-shot prompting allows the model to generate an unconstrained sequence of intermediate reasoning tokens $Z = (z_1, z_2, \ldots, z_k)$ prior to predicting Y. To ensure robust statistical power, we scale our dataset to 2,000 total instances (500 instances per task type).

The tasks comprising our dataset include the following four structural primitives:

- **Linear Symbolic**: Sequential arithmetic operations (e.g., "Start with x. Add y. Subtract z.").
- **Parity Computation**: Evaluating the truth value of multiple boolean predicates and determining if the total count of valid predicates is even or odd.
- **Multiway Branching**: Executing conditional logic paths based on the modulo of an initial computation (e.g., "If $S \equiv 0 \pmod 3$, return $x + y$.").
- **Conditional Branching with Latent Gating (CBLG)**: Routing a mathematical operation based on a latent boolean property of a raw input (e.g., "Input: a, b. If a is even, calculate $a/2 + b$. If a is odd, calculate $a - b$.").

For each task instance, we procedurally generate a clean prompt X_c with ground-truth answer Y_c, and a corresponding corrupted prompt X_{corr} (e.g., modifying a starting value to flip a logic gate) which yields a distinct corrupted answer Y_{corr}. Our generation constraints guarantee strict digit-width boundaries, ensuring perfect token-length alignment between X_c and X_{corr}.

These task types serve as a proxy for algorithmic completeness within the transformer architecture and ensure that our faithfulness metric is not biased toward a single cognitive heuristic. If a model's reasoning circuits are proven faithful across all four of these diverse computational primitives, it provides strong causal evidence that the model is actively executing generalizable control-flow operations rather than relying on shallow, task-specific pattern matching. Unlike causal tracing evaluations on unstructured natural language where the variable position of subject tokens limits interventional scope (Meng et al., 2022), our strictly controlled synthetic templates ensure uniform token alignment, providing mathematically sound conditions for resample ablation.

To extend our evaluation beyond isolated primitives, we also employ a mechanistically aligned GSM8K-Proxy dataset. While the original GSM8K dataset (Cobbe et al., 2021) serves as the standard benchmark for multi-step mathematical reasoning, its highly unstructured, free-form nature precludes rigorous resample ablation. Specifically, modifying real GSM8K instances to create counterfactual pairs (X_{corr}) frequently alters the token length of both the prompt and the resulting reasoning trace, leading to catastrophic tensor shape mismatches during activation patching. Our GSM8K-Proxy circumvents this limitation by procedurally generating word problems that mirror the linguistic complexity and multi-step arithmetic of the original dataset, while strictly enforcing the token-width constraints and syntactic symmetry required for clean, bidirectional interventions.

3.3 Eliciting Natural Reasoning Traces

To quantify the causal load of specific attention heads, we evaluate the model's autonomous rea-

soning capabilities. We utilize standard few-shot (8-shot) prompting to elicit unconstrained Chain-of-Thought (CoT) generation. The model is provided with the clean prompt X_c and tasked with generating a sequence of intermediate reasoning tokens $Z_{natural} = (z_1, z_2, \ldots, z_k)$ culminating in a final answer prediction.

Evaluating causal load on mechanically broken rationales yields invalid measurements; ablating a circuit on a task the model already fails to comprehend provides no causal signal. Therefore, we semantically parse each generated trace to extract the model's final predicted answer. We strictly filter the dataset to include only instances where the model's autonomous generation perfectly matches the ground-truth target Y_c. By restricting our analysis to naturally successful traces, we guarantee that the baseline accuracy of our evaluation subset is exactly 100%.

3.4 Dynamic Circuit Discovery via Activation Patching

Prior to conducting resample ablation, we require a rigorous method to identify the specific attention-routing circuit (S) responsible for successfully executing each task. To achieve this without relying on heuristic architectural guesses, we employ exhaustive single-head activation patching on the filtered subset of successful traces.

To ensure our causal measurement occurs at the exact computational moment the model predicts the target, we dynamically append the model's own generated CoT prefix to the prompt. We then record a clean activation cache (A_c) by running a forward pass on this aligned clean prompt. Next, we execute a forward pass on the corrupted prompt to establish a baseline corrupted Logit Difference (the relative probability of the clean target token versus the corrupted target token).

We systematically sweep through the network, intervening on every attention head individually. For each head, we patch its specific output activation (z) in the corrupted run with its corresponding activation from the clean cache A_c. The causal importance of each head is quantified by the magnitude by which patching that single head restores the Logit Difference toward the correct target.

Crucially, rather than selecting an arbitrary, static number of top-k heads, we aggregate the impact scores and apply a Dynamic Thresholding function. We sort the heads by positive causal impact and dynamically select the minimal subset of heads required to achieve 80% of the network's total cumulative restorative impact. This dynamic thresholding intrinsically accounts for architectural differences between the Base and Instruct models, ensuring we isolate the precise functional subgraph regardless of how densely or sparsely the reasoning pathway is distributed.

3.5 Resample Ablation and Faithfulness

Having isolated a 100% mechanically coherent baseline and defined the dynamic reasoning circuit S for each task, we execute the causal intervention. Prior approaches measure faithfulness by perturbing the surface form of chain-of-thought reasoning and observing changes in model outputs (Lanham et al., 2023). In contrast, our method operates directly on internal activations, enabling precise causal analysis of the underlying computation rather than its textual explanation.

To evaluate this causal faithfulness without risking the sequence divergence inherent to open-ended generation or the out-of-distribution manifold collapse associated with zero-ablation (Heimersheim and Nanda, 2024), we employ dual-direction resample ablation.

To ensure exact token-alignment across interventions, we utilize a static prefix alignment strategy. For each successful instance, we append the model's autonomously generated CoT prefix to both the clean (X_c) and corrupted (X_{corr}) prompts. We cache the internal activations of both static passes, yielding A_c and A_{corr}.

Rather than relying on binary accuracy metrics, which can mask subtle internal shifts, we quantify the causal load of the targeted circuit by measuring the Logit Difference of the target prediction token under two continuous intervention states:

- **Noising**: We perform a forward pass on the clean prompt but actively overwrite the targeted reasoning circuit S with activations from A_{corr}:

$$\tilde{A}^{(l,h)} = \begin{cases} A_{corr}^{(l,h)} & \text{if } (l,h) \in S \\ A_c^{(l,h)} & \text{otherwise} \end{cases}$$

  A significant drop in the clean Logit Difference indicates the circuit is causally *necessary*.

- **Denoising**: We perform a forward pass on the corrupted prompt but overwrite circuit S with activations from A_c. A significant recovery

toward the clean target indicates the circuit is causally *sufficient*.

The joint use of noising and denoising allows us to cross-validate causal claims, identifying components that are both necessary and sufficient for the behavior. Crucially, this bidirectional approach helps account for the inherent asymmetries and potential geometric artifacts commonly introduced by activation patching (Makelov et al., 2024).

4 Experiments and Results

4.1 Datasets, Models, and Evaluation Setup

We evaluate our pipeline across two distinct dataset regimes to capture both isolated algorithmic routing and natural language generalization. First, we utilize 2,000 instances of our four synthetic task types: Linear Symbolic, Parity Computation, Multiway Branching, and Conditional Branching with Latent Gating (CBLG). Second, to validate our findings in an open-ended natural language context, we evaluate 500 instances of a Mechanistically Aligned GSM8K-Proxy dataset.

To isolate the impact of alignment tuning on internal representations, we conduct a strict within-family comparison using the 1B parameter class of the Llama architecture, specifically comparing the foundation model (Llama-3.2-1B) against its instruction-tuned counterpart (Llama-3.2-1B-Instruct). Because both models share an identical computational backbone and Grouped-Query Attention (GQA) topology, structural deviations in circuit faithfulness can be directly attributed to post-training alignment interventions.

The evaluation proceeds in three phases:

1. **Natural Baseline Generation**: Establishing the model's autonomous ability to accurately solve the task using unconstrained, 8-shot Chain-of-Thought, yielding our 100% baseline subset.

2. **Dynamic Circuit Discovery**: Identifying the minimal subgraph of attention heads required to account for 80% of the network's restorative causal impact during single-head activation patching.

3. **Dual-Direction Resample Ablation**: Quantifying the causal load of the discovered circuits via Logit Differences, yielding continuous scores for both Necessity (performance drop under Noising) and Sufficiency (performance recovery under Denoising).

4.2 Circuit Discovery and Topological Sparsity

Before executing causal interventions, our Dynamic Circuit Discovery phase revealed topological differences in how reasoning is distributed across the two models. While both models utilize a dedicated subset of attention heads to route mathematical logic, the instruction-tuned model exhibited a higher degree of circuit diffusion on more complex logical tasks.

For instance, to reach the 80% cumulative causal impact threshold on the Parity task, the Base model relied on a highly modular, sparse subgraph of 68 attention heads. In contrast, the Instruct model required a significantly larger network of 181 heads to reach the exact same restorative threshold. This trend held broadly across the multi-step synthetic suite, with the Instruct model requiring an average of 95 heads per task compared to the Base model's 57.

Interestingly, on the natural language GSM8K-Proxy, both architectures converged on highly similar head counts (38 and 37 heads, respectively), suggesting that severe circuit diffusion is most pronounced when the aligned model is forced to execute pure, abstract logical routing without the anchoring of natural language word problems. Ultimately, this structural diffusion across the residual stream necessitated the use of dynamically sized intervention hooks for our subsequent resample ablation.

4.3 Dual-Direction resample ablation

Applying dual-direction resample ablation to these dynamically extracted circuits exposed distinct causal profiles between the foundation and instruction-tuned architectures. Table 1 details the Necessity and Sufficiency scores across all task distributions.

Across both models and all tasks, the discovered circuits demonstrated near-perfect Sufficiency. When the isolated reasoning subgraphs were injected into corrupted contexts, the Llama Base model recovered 100.36% of its baseline logit difference on synthetic tasks and 98.88% on the GSM8K proxy. Similarly, the Instruct model achieved 100.10% and 98.69% recovery, respectively. This confirms that the dynamically thresholded circuits independently possess the full com-

putational capacity required to execute the target mathematical primitives. However, Noising interventions (corrupting the primary reasoning circuits) revealed a stark divergence in Necessity scores.

For the Llama Base model, completely corrupting the targeted reasoning circuits resulted in minimal performance degradation, yielding a Necessity drop of only 2.92% on synthetic primitives and 4.96% on GSM8K. Conversely, the Instruct model proved significantly more sensitive to internal perturbations. While performing identically on the GSM8K proxy (5.10% drop), the Instruct model exhibited a Necessity drop of 6.79% on the synthetic primitives, which is more than double the degradation observed in the Base model under identical noising conditions.

5 Discussion

Our dual-direction resample ablation reveals a topological divergence in how mathematical resilience is distributed between foundation and aligned architectures. In the Llama Base model, completely corrupting the dynamically isolated reasoning circuit resulted in a minimal Necessity drop of only 2.92% on synthetic primitives. Given that this exact same circuit demonstrated near-perfect Sufficiency when evaluated in isolation, this minimal performance degradation initially appears paradoxical.

However, this resilient behavioral plateau perfectly quantifies the Hydra Effect (McGrath et al., 2023) in the context of algorithmic reasoning. The foundation model possesses a highly parallel, distributed mathematical capacity. When early-layer attention heads are corrupted via resample noising, dormant "backup circuits" in later layers dynamically detect the aberrant residual stream and re-compute the missing logical vectors to salvage the generation. Recent feature-level interventions using sparse autoencoders (Chen et al., 2025) have similarly highlighted that this redundant nature of mathematical Chain-of-Thought frequently obscures causal attribution. Our findings confirm that in foundation models, reasoning is not a fragile, linear pathway, but a highly redundant structural web capable of profound self-repair.

Furthermore, our comparison of the Llama-3.2-1B models demonstrates that model self-repair is not a static architectural guarantee and illustrates how alignment tuning can heavily degrade this capability. When subjected to identical noising interventions, the Instruct model exhibited a Necessity drop of 6.79% on the synthetic primitives—more than double the degradation observed in the Base model.

We posit that this increased structural brittleness represents a mechanistic "Alignment Tax." During Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), the network is heavily penalized for raw, textbook-style outputs and rewarded for conversational formatting, tone, and instruction compliance. To satisfy these competing multi-objective constraints without increasing the parameter count, the network is forced to repurpose its attention heads. By reallocating its dormant, redundant capacity toward linguistic and formatting adherence, the aligned model centralizes its mathematical computation. Having sacrificed its distributed backup circuits to the Alignment Tax, the Instruct model's primary reasoning pathway loses its self-repair capabilities, rendering it uniquely vulnerable to internal perturbations.

It is important to contextualize the magnitude of this tax. While an absolute Necessity drop of 6.79% is relatively modest and may not immediately signal catastrophic task failure in real-world deployment, the relative shift—more than doubling the architectural brittleness—serves as a vital mechanistic signal. Furthermore, while this structural degradation intersects with broader debates on CoT faithfulness, our measurements strictly isolate the causal necessity and robustness of internal routing, which serves as the physical prerequisite for faithful verbalization.

This hypothesis is further supported by the models' converging behavior on the GSM8K proxy dataset. While the Instruct model was significantly more brittle on the abstract synthetic tasks, both models exhibited nearly identical Necessity drops on the GSM8K proxy (4.96% for Base, 5.10% for Instruct).

Unlike the isolated synthetic primitives, GSM8K word problems require heavy, multi-step linguistic parsing alongside multi-variable mathematical tracking. We hypothesize that these complex, natural language-grounded tasks push the 1B parameter architecture to its representational capacity ceiling. When evaluating the GSM8K proxy, even the Base model must exhaust its available attention heads to simply route the logic, leaving few dormant heads available to act as a self-repairing "Hydra." Consequently, when task complexity scales to exhaust network capacity, foundation models exhibit the

	Llama-3.2-1B (Base)		Llama-3.2-1B-Instruct	
Task Class	**Necessity** (↓)	**Sufficiency** (↑)	**Necessity** (↓)	**Sufficiency** (↑)
Mechanistically Aligned Synthetic Primitives				
Linear Symbolic	4.29%	99.73%	8.05%	99.37%
Latent Gating (CBLG)	4.13%	100.65%	7.89%	100.59%
Parity Computation	0.47%	100.62%	0.38%	100.54%
Multiway Branching	0.73%	100.32%	0.78%	100.34%
Synthetic Average	**2.92%**	**100.36%**	**6.79%**	**100.10%**
Natural Language Proxy				
GSM8K Proxy	4.96%	98.88%	5.10%	98.69%

Table 1: Dual-Direction Activation Patching Logit Differences. Necessity indicates the relative performance drop under noising interventions (corrupting the targeted reasoning circuit), while Sufficiency indicates performance recovery under denoising interventions. The Instruct model exhibits a necessity drop more than double that of the Base model on synthetic primitives, quantifying the degradation of its structural redundancy.

same structural brittleness as their alignment-taxed counterparts.

6 Limitations

Scale and Capacity Starvation: While our dual-direction resample ablation provides empirical evidence of an alignment-induced brittleness, our experiments are localized to a strict within-family comparison at the 1B parameter scale. At this size, the representational capacity budget may be easily exhausted. It remains an open question whether the "Alignment Tax" is a universal consequence of alignment, or simply an artifact of capacity starvation in smaller architectures. Massive frontier models (e.g., 70B+ parameters) may possess sufficient capacity to absorb multi-objective constraints without sacrificing their distributed backup circuits.

Intervention Granularity: To isolate the specific dynamics of logical routing while maintaining precise intervention boundaries, our resample ablation was restricted to attention head outputs (z). The residual stream sequentially updates via $x^{(l+1)} = x^{(l)} + \text{Attn}^{(l)} + \text{MLP}^{(l)}$. While recent literature highlights the role of Multi-Layer Perceptrons (MLPs) in latent reasoning (Geva et al., 2023), quantifying how alignment independently alters MLP-based representation represents a distinct methodological challenge outside the scope of our attention-focused interventions.

Monolithic Alignment: Our methodology compares a foundational base model directly against its final instruction-tuned counterpart. Modern alignment pipelines consist of multiple distinct phases (e.g., Supervised Fine-Tuning followed by RLHF or DPO). Our current experimental design treats instruction-tuning as monolithic and cannot isolate which specific stage of training is responsible for repurposing the backup circuits.

Dataset Variance: To secure the mathematically airtight token-alignment required for causal scrubbing, our evaluation heavily leverages procedurally generated primitives. While these tasks isolate control-flow operations, they lack the linguistic entropy of free-form queries (Dziri et al., 2023). Additionally, we note that performance degradation was not uniformly distributed across the synthetic suite, suggesting that the brittleness of the aligned model may manifest differently depending on the specific cognitive heuristic being evaluated.

Prompt Limitations: Our dynamic circuit discovery is inherently conditioned on the specific 8-shot prompt template used to elicit the unconstrained baseline. It remains an open question whether these discovered subgraphs represent universal mathematical circuits, or whether they are partially overfit to the stylistic syntax of the provided prompt, a known challenge in automated circuit discovery (Conmy et al., 2023).

7 Conclusion

In this work, we investigated the structural impact of alignment tuning on LLM algorithmic reasoning. Using naturally elicited Chain-of-Thought traces and dual-direction resample ablation on the Llama-3.2-1B architecture, we revealed a notable mechanistic divergence. While foundation models leverage highly redundant, self-repairing computational networks (the Hydra Effect), instruction-tuned models exhibit significantly reduced redundancy. We frame this as the *Alignment Tax on*

Redundancy: managing simultaneous logical and linguistic constraints repurposes dormant backup circuits, centralizing reasoning pathways and increasing their sensitivity to internal perturbation.

Although the absolute magnitude of this measured brittleness is modest at the 1B scale, these findings indicate that evaluating model reliability solely on generated text may be insufficient for high-stakes environments, as alignment optimization can mask underlying mathematical brittleness. Furthermore, while precise circuit mapping carries dual-use risks for adversarial weight-editing, it is essential for transparent evaluation. Future work should scale this methodology to frontier architectures—to disentangle this alignment tax from small-model capacity starvation—and isolate the specific impacts of Supervised Fine-Tuning versus preference optimization, ultimately guiding the development of redundancy-preserving alignment techniques that ensure models remain both human-aligned and mechanistically robust.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *arXiv*.

Pepa Atanasova, Jakob Grue Simonsen, Maria Liakata, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Xi Chen, Aske Plaat, and Niki van Stein. 2025. How does Chain of Thought think? mechanistic interpretability of Chain-of-Thought reasoning with sparse autoencoding. *arXiv*.

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

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*.

Nouha Dziri, Zhou Yu, Siva Reddy, and Danqi Chen. 2023. Faith and fate: Limits of transformers on compositionality. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, and 1 others. 2023. Measuring faithfulness in Chain-of-Thought reasoning. In *Advances in Neural Information Processing Systems*.

Haolang Lu, Hongrui Peng, Weiye Fu, Guoshun Nan, Xinye Cao, Xingrui Li, Hongcan Guo, and Kun Wang. 2026. Disentangling deception and hallucination failures in LLMs. *arXiv*.

Aleksandar Makelov, Georg Lange, and Neel Nanda. 2024. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Thomas McGrath, Jacob Kaplansky, Neel Nanda, and 1 others. 2023. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, and 1 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in Chain-of-Thought prompting. In *Advances in Neural Information Processing Systems*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*.

Yanbo Wang, Yongcan Yu, Jian Liang, and Ran He. 2025. A comprehensive survey on trustworthiness in reasoning with large language models. *arXiv*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Wei Jie Yeo, Ranjan Satapathy, and Erik Cambria. 2025. Towards faithful natural language explanations: A study using activation patching in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10425–10447.

Kerem Zaman and Shashank Srivastava. 2025. Is Chain-of-Thought really not explainability? Chain-of-Thought can be faithful without hint verbalization. *arXiv*.

Chunting Zhou, Pengfei Liu, Puxin Xu, and 1 others. 2023. LIMA: Less is more for alignment. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*.