WHEN REPRESENTATIONS COLLAPSE IN LOGICAL REASONING: DIAGNOSING AND MITIGATING CASCADE FAILURES IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) achieve strong performance on logical reasoning benchmarks, yet remain brittle as proofs become deeper, irrelevant premises accumulate, or evaluation conditions shift. We investigate this brittleness at the level of internal representations. To this end, we introduce **Inferential Information Flow (IIF)**, a reasoning-aware diagnostic framework that measures, layer by layer, whether a model preserves selectivity toward proof-relevant evidence over distractors. Across three model families and three logical reasoning benchmarks, IIF reveals a broad but heterogeneous failure pattern: proof-selective structure often strengthens at intermediate depth and subsequently degrades, while some settings instead fail to form strong selectivity at the query interface. We term the former *cascade IIF collapse*. Causal interventions further show that some degraded value-channel representations can become harmful to downstream reasoning rather than merely uninformative. Motivated by this diagnosis, we introduce FRGI (Fact–Rule Grouped Intervention), a training-free, proof-free inference-time method that uses query-conditioned saliency to select candidate fact and rule spans and an Unlabeled Collapse Proxy (UCP) to locate intervention layers. FRGI applies localized fact–rule and value-channel restoration without retraining, proof annotations, or prompt modification. Across ProverQA, PrOntoQA, and ProofWriter, FRGI consistently improves accuracy over representative prompting, self-consistency, symbolic-reasoning, and activation-intervention baselines. These results show that representation-level failures in logical reasoning can be diagnosed and partially repaired through targeted inference-time intervention.

1 INTRODUCTION

Large language models (LLMs) achieve strong performance on many reasoning tasks, especially with chain-of-thought prompting and related test-time strategies (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023). However, their reasoning remains fragile: performance often degrades with deeper proofs, distractor premises, and shifted evaluation conditions (Saparov & He, 2023; Tafjord et al., 2021; Qi et al., 2025b; Shi et al., 2023; Song et al., 2026).

This brittleness is especially visible in symbolic reasoning, where models must select relevant premises and preserve dependencies across multiple inference steps. Behavioral accuracy reveals that a model failed, but not where the internal reasoning process began to break down. Figure 1 illustrates this gap: although the model derives valid intermediate conclusions, no fact or rule supports the queried claim, whose correct label is therefore *Uncertain*. Several LLMs nevertheless predict *False*, suggesting that valid intermediate deductions can be over-extended beyond what the proof supports.

Prior work mainly studies reasoning failure from two perspectives. Behavioral analyses document reversal errors, compositional breakdowns, and distractibility (Berglund et al., 2024; Press et al., 2023; Shi et al., 2023; Song et al., 2026), while representation-collapse studies track global statistics such as rank, variance, or entropy (Dong et al., 2021; Noci et al., 2022; Barbero et al., 2024; Arefin et al., 2025; Gerasimov et al., 2026; Pagliardini et al., 2024). Neither directly reveals whether proof-relevant information remains selective over distractors as computation proceeds through depth.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We address this gap with **Inferential Information Flow (IIF)**, a reasoning-aware diagnostic of layerwise evidence selectivity. IIF asks whether the relation between the current reasoning state and evidence spans still distinguishes proof-used evidence from unused distractors. Because the underlying conditional mutual information is intractable, we operationalize IIF using complementary MI-inspired and geometric diagnostics.

IIF reveals a *broad but heterogeneous* failure pattern. Proof-selective structure often strengthens at intermediate depth and later degrades, which we term **cascade IIF collapse**; other settings exhibit *formation failure*, where strong query-level selectivity never develops. Rule-interface analysis further shows that local proof structure can still emerge and later deteriorate. Controlled value-channel interventions also indicate that some degraded representations can be harmful rather than merely uninformative.

Guided by this diagnosis, we introduce **FRGI** (Fact–Rule Grouped Intervention), a training-free, proof-free inference-time method. Query-conditioned saliency determines *what* fact and rule spans to modify, an Unlabeled Collapse Proxy (UCP) determines *where* to intervene, and localized cascade and value-channel restoration determine *how* to repair the selected representations. Across LLaMA-3.1-8B, Qwen3-8B, and Mistral-7B on ProverQA, PrOntoQA, and ProofWriter, FRGI achieves the highest overall accuracy in all evaluated model–benchmark settings, while the proof-free UCP approaches offline proof-annotated IIF for layer selection.

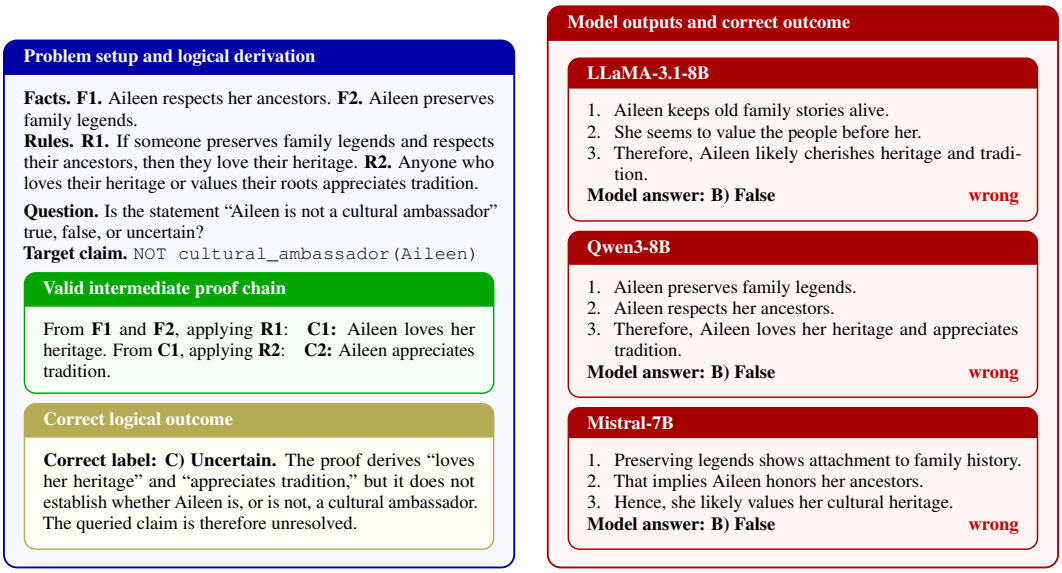

Fig. 1: Case study of a reasoning failure. The derivation contains valid intermediate deductions, but none of the facts or rules provides a proof link to the queried target claim `NOT cultural_ambassador(Aileen)`.

2 Inferential Information Flow Analysis

Each reasoning instance x contains M_x evidence spans $\{e_1, \ldots, e_{M_x}\}$, a query q, and an annotated proof that identifies which evidence participates in the derivation. In symbolic reasoning, evidence spans consist of *facts* and *rules*. The proof partitions them into *proof-used* spans $\mathcal{U}_x$ and *unused distractors* $\mathcal{N}_x$. Throughout the paper, "used" refers only to participation in the annotated proof rather than to general semantic relevance.

At layer l of an L-layer Transformer, evidence span e_j is represented by mean-pooling its token hidden states, $\boldsymbol{r}_{x,j}^{(l)} \in \mathbb{R}^d$, while the query representation $\boldsymbol{z}_x^{(l)} \in \mathbb{R}^d$ is taken from the final query-token position. All diagnostics are computed from a single forward pass over the prompt without generation.

Evidence selectivity. We call *evidence selectivity* the degree to which the relation between the query and evidence representations distinguishes proof-used evidence from unused distractors. Let $X \sim \mathcal{D}$ denote a reasoning instance, $J \mid X = x \sim \mathrm{Unif}(\{1, \ldots, M_x\})$ a sampled evidence-span index, and

$$Y_{X,J} = \mathbf{1}\{e_{X,J} \in \mathcal{U}_X\}$$

its proof-usage label. We define the layer-l query–evidence relation as

$$G_{X,J}^{(l)} = \phi\Big(\boldsymbol{z}_X^{(l)}, \boldsymbol{r}_{X,J}^{(l)}\Big), \tag{1}$$

where ϕ is a similarity function. Unless otherwise stated, we use cosine similarity, $\phi(\boldsymbol{a}, \boldsymbol{b}) = \cos(\boldsymbol{a}, \boldsymbol{b})$.

Definition 1 (Inferential Information Flow) *We define the layer-l* Inferential Information Flow *as*

$$\mathrm{IIF}^{(l)} \triangleq I\Big(G_{X,J}^{(l)}; Y_{X,J} \mid X\Big). \tag{2}$$

Conditioning on X makes IIF instance-conditioned: it asks whether the query–evidence relation within the same reasoning problem retains information that distinguishes proof-used evidence from distractors, rather than exploiting dataset-level regularities.

Exact conditional mutual information is not directly tractable in our setting. We therefore characterize IIF using five complementary layerwise diagnostics rather than interpreting any single score as a calibrated mutual-information estimator or formal bound.

NCE-IIF measures contrastive evidence selectivity: it asks whether proof-used spans receive greater query-conditioned similarity mass than unused spans, with instance-level normalization accounting for different numbers of proof-used premises.

The CLUB-style probe provides a complementary predictability view. At each layer, a one-dimensional logistic probe receives only the query–span similarity and measures how well proof usage can be distinguished from that scalar relation. We use it as a CLUB-inspired layerwise diagnostic rather than as a strict MI upper bound.

ValueIIF applies the same evidence-selectivity analysis to value-channel content. Evidence token values are projected back into residual space before span pooling, allowing us to test whether proof-used information remains selectively represented specifically in the value pathway.

Instance-level Center Alignment (ICA) is a parameter-free geometric diagnostic that compares the query's alignment with the centroid of proof-used evidence against its alignment with the centroid of unused evidence. Positive values indicate stronger attraction toward proof-relevant structure.

Finally, **TripletAcc** measures pairwise ranking: for a proof-used and an unused span from the same instance, it records whether the proof-used span receives the higher query–evidence similarity. Chance-level ranking corresponds to 0.5.

Together, these diagnostics probe complementary aspects of the same underlying quantity: contrastive mass, proof-usage predictability, value-channel selectivity, centroid geometry, and pairwise ranking. Their agreement therefore provides stronger evidence of a genuine layerwise change in proof-relevant selectivity than any single metric alone. Full definitions and robustness analyses are provided in Appendices B and C.

Rule-to-fact specialization. Query-to-evidence selectivity does not directly reveal whether rule representations preserve the facts needed for their application. We therefore construct rule-to-fact (R2F) variants by treating proof-used rules as anchors, proof-participating facts as positives, and other facts as negatives. R2F-IIF tracks whether proof-relevant rule–fact relations emerge and remain selective across layers. Its formal definition is given in Appendix B.5.

3 Cascade IIF Collapse

Having defined IIF and its five operational diagnostics, we now use them as layerwise probes of how proof-relevant information evolves through the network. For each model–benchmark pair, we compute NCE-IIF, the CLUB-style probe, ValueIIF, ICA, and TripletAcc at every layer, smooth and normalize each trajectory as described in §2, and average them into an aggregate evidence-selectivity profile. This layerwise view lets us ask a central question: *once proof-selective structure emerges, is it preserved toward the output?* We first examine this question at the query–evidence interface, and then use a complementary rule-interface analysis to determine whether weak query-level selectivity necessarily implies that proof structure is absent elsewhere in the network.

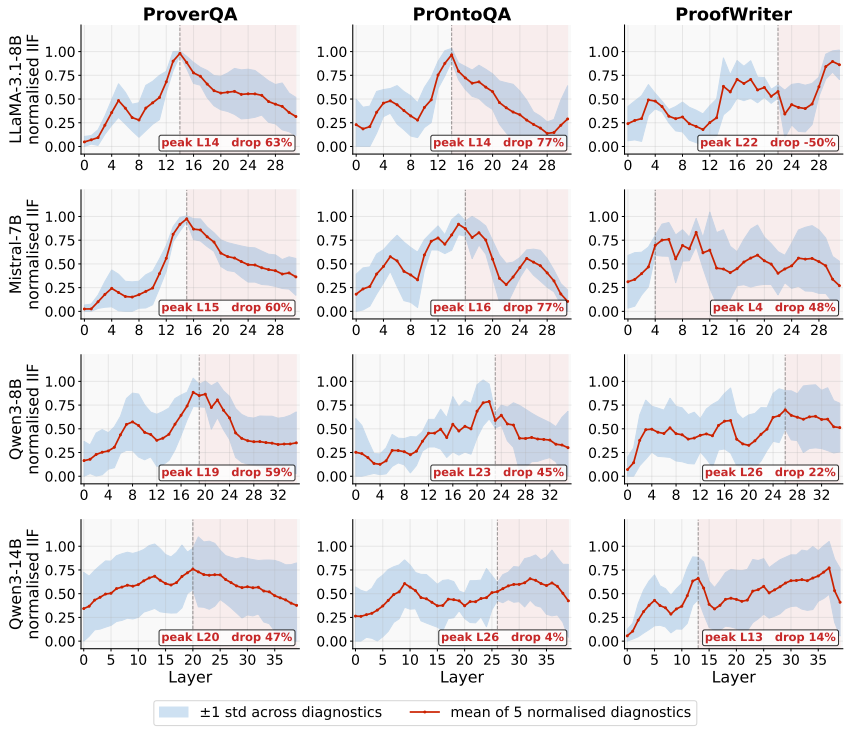

Fig. 2: **Layerwise evidence selectivity across benchmarks and backbones.** Rows denote backbones and columns denote ProverQA, balanced PrOntoQA, and ProofWriter. Each panel aggregates the five IIF diagnostics introduced in §2: the red curve shows their mean after within-cell normalization and the blue band shows variation across diagnostics. ProverQA and PrOntoQA exhibit pronounced intermediate selectivity followed by post-peak degradation across several backbones, whereas some ProofWriter cells remain weakly selective throughout, revealing a distinct formation failure.

3.1 Cross-Benchmark Evidence Selectivity

Figure 2 traces query-anchored evidence selectivity across depth using the five IIF diagnostics introduced in §2, after smoothing, within-cell normalization, and aggregation. On ProverQA, LLaMA-3.1-8B, Mistral-7B, and Qwen3-8B show a clear *rise–peak–fall* trajectory: selectivity strengthens toward an intermediate layer and then declines by 63%, 60%, and 60%, respectively. Balanced PrOntoQA shows the same broad pattern, with substantial post-peak degradation across all three backbones. These results indicate that proof-selective structure can form clearly at intermediate depth but need not remain preserved toward the output.

ProofWriter reveals a complementary failure mode. Several query-anchored profiles remain weak or unstable rather than developing a strong interior maximum. We therefore distinguish *formation failure*, where proof-selective structure never becomes strong at the query interface, from *cascade IIF collapse*, where selectivity first forms and is subsequently degraded. Qwen3-14B further shows that post-peak degradation is not restricted to 7–8B models, although its location and severity remain benchmark dependent. Extended cross-architecture, raw-geometry, and proof-depth analyses are reported in Appendix D.

3.2 Rule-Interface Confirmation

The query-level analysis leaves an important ambiguity. A weak query-anchored profile could mean that proof structure never forms anywhere in the network, or that structured relations emerge locally but fail to remain expressed in the final query state. We therefore probe the rule–fact interface directly. Figure 3 reports three complementary attention-based signals: answer-to-rule selectivity relative to distractor rules, answer-to-fact selectivity, and rule-to-fact coupling relative to background. These signals are not part of the five-diagnostic IIF aggregate; instead, they provide a localized view of whether proof-relevant rule–fact structure forms and persists across depth.

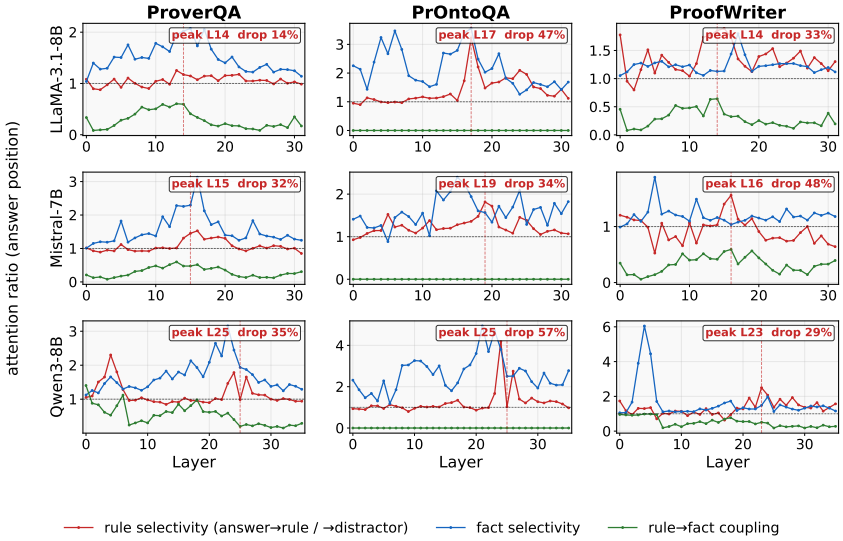

Fig. 3: **Rule-interface structure across benchmarks.** Layerwise attention-based diagnostics of answer-to-rule selectivity relative to distractor rules (red), answer-to-fact selectivity (blue), and rule-to-fact coupling relative to background (green) across ProverQA, PrOntoQA, and ProofWriter.

Across all three benchmarks and 8B backbones, answer-to-rule selectivity develops an interior maximum and subsequently declines. This pattern remains visible on ProofWriter, where the query-anchored IIF profiles are often weak: the post-peak drops are 33%, 48%, and 29% for LLaMA-3.1-8B, Mistral-7B, and Qwen3-8B, respectively. The two analyses therefore expose complementary failure modes. At the query interface, proof selectivity may fail to form strongly at all. At the more local rule interface, structured selectivity can still emerge and later deteriorate. Weak query-level IIF should therefore not be interpreted as evidence that proof structure is absent throughout the network.

Together, Figures 2 and 3 support a broader conclusion: proof-relevant structure is neither uniformly formed nor monotonically preserved across Transformer depth. Its trajectory depends on both the benchmark and the representational interface being measured.

4 FRGI: Fact–Rule Grouped Intervention

Motivated by the cascade IIF collapse identified in §3, we introduce FRGI (Fact–Rule Grouped Intervention), a training-free inference-time method that locally modifies evidence representations without updating model parameters, changing the prompt, or modifying the model architecture. Importantly, FRGI is proof-free at deployment: it uses no proof annotations, proof-used spans, or answer labels for span selection, layer selection, or generation.

As illustrated in Figure 4, FRGI follows a simple *what–where–how* design. A first, no-hook prefill determines *what* evidence should be targeted using query-conditioned attention saliency and *where* the intervention should occur using an Unlabeled Collapse Proxy (UCP). The same prompt is then processed a second time with lightweight hooks that determine *how* the selected representations are modified. The resulting prompt KV cache is retained, after which autoregressive decoding proceeds normally with no further intervention.

Proof-free target selection. Let $\mathcal{E}_x$ denote the evidence spans of an input instance, partitioned into fact spans $\mathcal{F}_x$ and rule spans $\mathcal{R}_x$. During the no-hook prefill, FRGI records attention weights and hidden states without decoding an answer.

To determine *what* to intervene on, we aggregate attention from the query tokens to each evidence span across a small set of diagnostic layers. This produces a query-conditioned saliency score for every span. An adaptive instance-level threshold selects a subset $\mathcal{S}$, which is then separated into selected facts $\mathcal{S}_{\text{fact}}$ and selected rules $\mathcal{S}_{\text{rule}}$. Attention is used only as a lightweight target-selection signal rather than as an explanation of the model's reasoning.

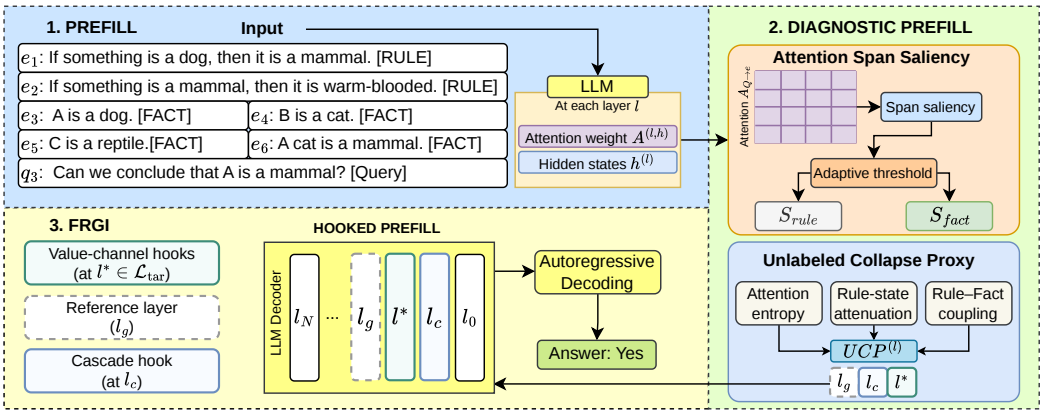

Fig. 4: **FRGI proof-free inference pipeline.** (1) A no-hook prefill records layerwise attention weights and hidden states. (2) Attention saliency selects candidate fact and rule spans, while the Unlabeled Collapse Proxy (UCP) identifies a healthy reference layer l_g, an early cascade layer l_c, a collapse anchor l^*, and sparse restoration layers $\mathcal{L}_{\text{tar}}$. (3) The prompt is prefetched again with a cascade hook at l_c and value-channel hooks at $l \in \mathcal{L}_{\text{tar}}$. Generation then proceeds normally from the resulting repaired KV cache. No proof annotations, proof-used spans, or answer labels are used at inference time.

In parallel, FRGI determines *where* to intervene using the Unlabeled Collapse Proxy, $\text{UCP}^{(l)}$, computed independently at each layer. UCP combines three proof-free signals: query-to-evidence attention entropy, rule-state attenuation, and unlabeled rule–fact coupling. High UCP therefore indicates a layer where evidence attention is more diffuse, rule representations are more attenuated, and rule–fact coupling is weaker. From the layerwise UCP profile, FRGI identifies l_g as a healthy pre-collapse reference, l_c as an early cascade layer, l^* as the collapse anchor, and $\mathcal{L}_{\text{tar}}$ as sparse post-collapse restoration layers. If no salient evidence or sufficiently strong collapse is detected, FRGI abstains and returns the vanilla prediction. Detailed definitions are given in Appendix E.2.2.

Fact–rule grouped intervention. The second prompt prefill applies two localized interventions using the spans and layers selected above.

First, at the early cascade layer l_c, selected rule-token states are nudged toward the centroid of the selected fact representations:

$$\boldsymbol{h}_i^{(l_c)} \leftarrow (1-w)\boldsymbol{h}_i^{(l_c)} + w\,\bar{\boldsymbol{h}}_{\text{fact}}^{(l_c)}, \tag{3}$$

where $\bar{\boldsymbol{h}}_{\text{fact}}^{(l_c)}$ is the mean representation of the selected facts. This hook is applied only when both selected facts and selected rules are available. Its role is to reinforce the interaction between the selected rule and fact context before the diagnosed collapse region.

Second, at each target layer $l \in \mathcal{L}_{\text{tar}}$, FRGI restores the value channel of selected evidence tokens using the healthy representation cached from l_g. Rather than copying a value vector across layers, the cached hidden state is re-projected through the target layer's own value projection:

$$\widetilde{\boldsymbol{v}}_i^{(l)} = W_V^{(l)} \text{Norm}^{(l)}\left(\widehat{\boldsymbol{h}}_i^{(l_g)}\right). \tag{4}$$

The current value is then interpolated with this reference-derived value:

$$\boldsymbol{v}_{i,\text{FRGI}}^{(l)} = (1-\beta_l)\boldsymbol{v}_i^{(l)} + \beta_l \widetilde{\boldsymbol{v}}_i^{(l)}, \tag{5}$$

where β_l adapts the intervention strength to the degree of attenuation observed in the no-hook diagnostic pass. Stronger attenuation therefore induces a larger reference contribution.

Hooked prefill and decoding. The cascade and value-channel hooks are active only during the second prompt prefill. Modified evidence-token values are written into the prompt KV cache as it is constructed. Once prefill is complete, the hooks are removed and the model performs ordinary autoregressive generation from this repaired cache. Thus, FRGI intervenes once on how the prompt is represented and remembered, while leaving all subsequent decoding steps unchanged. Full UCP definitions, layer selection rules, the adaptive β_l schedule, and implementation details are provided in Appendix E.

5 EXPERIMENT

5.1 SETUP

Models. Our main experiments use three instruction-tuned models of comparable scale: LLaMA-3.1-8B-Instruct, Qwen3-8B, and Mistral-7B-Instruct-v0.2. These models span different model families and training pipelines, allowing us to test whether the diagnosed behavior and the resulting intervention generalize across architectures. To study model-scale effects separately from architecture, we further evaluate Qwen3-14B and Qwen3-32B alongside Qwen3-8B on ProverQA. This same-family comparison isolates scaling more cleanly and is used only for the dedicated scale analysis.

Benchmarks. We evaluate on three proof-structured logical reasoning benchmarks: *ProverQA* (Qi et al., 2025a) (1,500 dev instances across easy, medium, and hard difficulty, with FOL annotations and reasoning chains), *PrOntoQA* (Saparov & He, 2023) (5-hop deductive reasoning with ground-truth proof structures), and *ProofWriter* (Tafjord et al., 2021) (depth-5 open-world-assumption problems with proof trees). ProverQA serves as our primary benchmark; PrOntoQA and ProofWriter serve as out-of-distribution (OOD) transfer benchmarks to test generalization.

Metrics. We report overall accuracy and per-difficulty accuracy (easy/medium/hard) on ProverQA, and overall accuracy on PrOntoQA and ProofWriter. All improvements (Δ) are reported in percentage points (pp) relative to the vanilla baseline for each model. We use greedy decoding for the main comparison and report beam-search results (k=4) in Appendix G.1.

Baselines. We compare FRGI against *Vanilla* prompting and three classes of reasoning interventions: prompting-based methods, including Zero-shot CoT (Kojima et al., 2022), Few-shot CoT (Wei et al., 2022), and Self-Consistency (SC, k=5) (Wang et al., 2023); activation-steering methods, including ActAdd (Turner et al., 2025) and SADI (Wang et al., 2025); and task-specific methods, including SymbCoT (Xu et al., 2024) and MTI (Yang et al., 2026). ActAdd and SADI use reasoning-oriented contrastive pairs described in Appendix F.2; SC uses majority voting over five sampled CoT paths.

Additional robustness and diagnostic results. The appendix provides several complementary checks beyond the main greedy-decoding results. Appendix G.1 shows that FRGI's gains persist under beam search ($k = 4$), where it achieves the highest overall accuracy on all three benchmarks for all three model families. Appendix G.3 shows that FRGI incurs only modest inference overhead: it adds one diagnostic prefill and one hooked prefill, while avoiding the repeated generation required by multi-sample reasoning methods. Appendix G.4 further shows that the conclusions remain stable across generation budgets from 256 to 2048 tokens, indicating that the observed gains are not explained by a larger generation allowance. Full per-model and per-dataset results for the FRGI component ablation, proof-free layer selection, and UCP component analysis are provided in Appendix G.2.1.

5.2 EXPERIMENTAL RESULTS & ANALYSIS

Effectiveness of FRGI Table 1 shows that FRGI achieves the highest overall accuracy in all nine model–benchmark pairs. On ProverQA, it improves over the strongest competing baseline by 3.81, 3.75, and 2.18 points for LLaMA-3.1-8B, Qwen3-8B, and Mistral-7B, respectively, while remaining strongest across the easy, medium, and hard subsets. The same pattern holds on PrOntoQA, where the corresponding gains are 5.40, 4.76, and 4.52 points and persist across all proof depths. On ProofWriter, FRGI also gives the best overall accuracy for all three backbones. These results indicate that the intervention generalizes across model families, proof distributions, and reasoning difficulty rather than benefiting only a single benchmark setting.

Component analysis of FRGI. Table 2 isolates the span-targeting and intervention components while keeping UCP-based layer selection fixed. Replacing saliency-selected spans with random spans causes the largest overall degradation, reducing average accuracy from 71.85% to 53.56%, showing that the intervention must act on the appropriate evidence. Among the repair operations, removing value restoration reduces accuracy to 61.88%, whereas removing the cascade hook yields 68.63%. Replacing adaptive intervention strength with a fixed β gives 69.81%. Together, these results support the intended division of labor in FRGI: saliency determines *what* to edit, while value restoration, cascade correction, and adaptive strength determine *how* the selected representations are repaired.

Table 1: Main results comparing FRGI with representative reasoning and intervention baselines. We report greedy-decoding accuracy (%) on ProverQA, PrOntoQA, and ProofWriter.

Model	Method	ProverQA				PrOntoQA					ProofWriter
		Overall	Easy	Medium	Hard	Overall	1-hop	2-hop	3-hop	4-hop	Overall
LLaMA-3.1-8B	Vanilla	36.84	47.23	36.42	26.87	41.28	46.54	42.17	39.43	36.98	52.46
	Zero-shot CoT	43.82	54.46	43.18	33.82	56.12	62.84	58.53	53.27	49.84	63.54
	Few-shot CoT	50.84	65.23	48.45	38.84	67.38	74.45	69.24	64.83	61.02	66.48
	Self-Consistency	59.43	73.64	59.82	44.83	77.14	82.53	78.64	75.23	72.16	75.43
	ActAdd	41.42	55.24	40.63	28.39	57.94	63.82	59.54	55.18	53.22	56.84
	SADI	38.87	52.63	37.44	26.54	51.86	57.43	53.24	49.13	47.64	52.68
	SymbCoT	43.68	56.42	42.64	31.98	72.24	78.53	73.84	69.62	66.97	61.34
	MTI	45.42	58.24	44.63	33.39	73.74	79.83	75.24	71.12	68.77	63.48
	FRGI	**63.24**	**78.64**	**62.43**	**48.65**	**82.54**	**87.43**	**83.84**	**80.23**	**78.66**	**78.14**
Qwen3-8B	Vanilla	37.92	49.43	37.84	26.49	42.64	47.83	43.52	40.24	38.97	48.83
	Zero-shot CoT	44.75	56.63	43.82	33.81	58.23	64.54	60.23	55.42	52.73	52.89
	Few-shot CoT	51.43	67.24	49.62	37.43	70.92	77.63	72.54	68.83	64.68	73.44
	Self-Consistency	61.09	76.43	60.64	46.21	79.42	84.83	80.94	77.43	74.48	77.56
	ActAdd	43.63	57.02	42.44	31.43	60.43	66.24	61.83	57.54	56.11	53.27
	SADI	40.72	54.83	39.54	27.79	53.84	59.42	55.23	51.14	49.57	52.69
	SymbCoT	44.67	57.23	43.54	33.24	74.26	80.43	75.84	71.53	69.24	62.54
	MTI	46.83	60.54	45.83	34.12	75.64	81.83	77.14	72.93	70.66	64.72
	FRGI	**64.84**	**80.83**	**64.24**	**49.45**	**84.18**	**88.64**	**85.34**	**82.13**	**80.61**	**80.53**
Mistral-7B	Vanilla	34.83	44.24	34.43	25.82	42.41	47.84	43.23	40.12	38.33	30.34
	Zero-shot CoT	44.75	55.43	44.22	34.61	55.14	61.43	56.84	52.05	50.24	32.76
	Few-shot CoT	49.56	63.84	47.63	37.21	65.62	72.24	67.53	62.84	59.87	36.57
	Self-Consistency	56.49	70.23	56.64	42.61	74.72	80.14	76.23	72.04	70.47	41.93
	ActAdd	42.36	54.43	41.24	31.41	59.42	65.54	60.83	56.24	55.07	33.98
	SADI	38.48	50.84	37.63	26.97	51.13	56.84	52.43	48.24	47.01	47.43
	SymbCoT	40.73	53.43	39.54	29.22	64.52	70.24	66.13	61.44	60.27	53.05
	MTI	44.33	57.04	43.53	32.42	66.24	72.43	68.04	63.24	61.25	45.83
	FRGI	**58.67**	**73.84**	**58.43**	**43.65**	**79.24**	**84.23**	**80.64**	**77.43**	**74.66**	**55.27**

Table 2: **FRGI component analysis across datasets.** We ablate the span-targeting and intervention components while keeping UCP-based layer selection fixed. Results are mean greedy-decoding accuracy (%) across LLaMA-3.1-8B, Qwen3-8B, and Mistral-7B.

Variant	**FRGI components**				**Mean accuracy across models**			
	Saliency	Cascade	Value	Adaptive β	ProverQA	PrOntoQA	ProofWriter	Avg.
Vanilla	–	–	–	–	36.53	42.11	43.88	40.84
Random spans	Rand.	✓	✓	✓	48.64	58.38	53.65	53.56
w/o cascade	✓	–	✓	✓	59.31	78.08	68.51	68.63
w/o value	✓	✓	–	✓	54.31	69.75	61.58	61.88
Fixed β	✓	✓	✓	Fixed	60.46	79.52	69.44	69.81
Full FRGI	✓	✓	✓	✓	**62.25**	**81.99**	**71.31**	**71.85**

Locating collapse-sensitive layers without proof annotations. Table 3 compares UCP with random, fixed, and non-UCP layer choices while keeping the remaining intervention fixed. UCP reaches 71.85% average accuracy, compared with 58.65% for random selection, 62.74% for a fixed layer, and 64.78% for non-UCP selection. Moreover, UCP is only 2.19 points below the proof-annotated IIF oracle (74.04%), recovering approximately 81% of the improvement from fixed layer selection to the oracle. Thus, effective intervention depends strongly on where it is applied, while the proof-free UCP captures much of the layer information exposed by the offline proof-conditioned IIF diagnostic.

What information does UCP need to select intervention layers? Table 4 shows that no individual UCP signal is sufficient. Among the single-signal selectors, rule–fact coupling performs best at 58.47%, but remains substantially below the full UCP at 71.85%. The leave-one-out analysis further shows that removing rule–fact coupling causes the largest reduction (-6.36 points), followed by rule-state attenuation (-4.66) and attention entropy (-3.12). The full combination performs best, indicating that UCP benefits from complementary signals of attention diffusion, rule-state attenuation, and weakening rule–fact coupling rather than relying on any single proxy.

Table 3: **Proof-free layer selection.** Mean accuracy (%) across ProverQA, PrOntoQA, and ProofWriter.

Selector	LLaMA	Qwen3	Mistral	Avg.
Random	61.32	62.74	51.88	58.65
Fixed	65.41	66.83	55.97	62.74
Non-UCP	67.69	68.99	57.66	64.78
UCP	74.64	76.52	64.39	71.85
Oracle IIF	76.83	78.54	66.76	74.04

Table 4: **UCP component analysis.** Mean greedy-decoding accuracy (%) across the three backbones. Δ is relative to the full UCP.

UCP variant	ProverQA	PrOntoQA	ProofWriter	Avg.	Δ
Single-signal selectors					
Attention entropy only	47.31	63.84	53.34	54.83	−17.02
Rule-state attenuation only	49.42	66.71	55.56	57.23	−14.62
Rule–fact coupling only	50.64	68.12	56.65	58.47	−13.38
Leave-one-out					
w/o attention entropy	59.64	78.43	68.12	68.73	−3.12
w/o rule-state attenuation	58.12	77.14	66.31	67.19	−4.66
w/o rule–fact coupling	56.43	75.42	64.62	65.49	−6.36
Full UCP	**62.25**	**81.99**	**71.31**	**71.85**	–

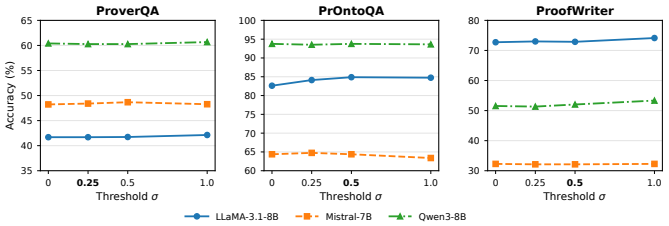

Fig. 5: **Sensitivity to the adaptive saliency threshold.** Accuracy remains stable across σ, indicating limited sensitivity to the exact threshold choice.

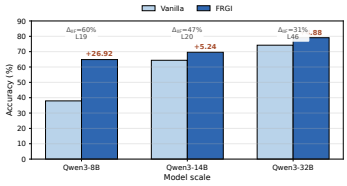

Fig. 6: **FRGI across model scales on ProverQA.** IIF degradation persists from 8B to 32B, while FRGI maintains positive gains over vanilla decoding.

Sensitivity of saliency threshold Figure 5 varies the threshold parameter σ while keeping the remaining FRGI pipeline fixed. Accuracy changes only modestly across the tested range for most model–benchmark pairs, despite larger σ selecting progressively fewer evidence spans. This suggests that the observed gains are not tied to a narrowly tuned threshold value, but persist over a range of reasonable span-selection sparsities.

Scalability Figure 6 evaluates the Qwen3 family from 8B to 32B on ProverQA. The post-peak IIF decline remains visible at all three scales, decreasing from 60% at 8B to 47% at 14B and 31% at 32B. FRGI also remains beneficial, improving vanilla accuracy by 26.92, 5.24, and 4.88 points, respectively. These results distinguish the generality of the diagnostic phenomenon from the magnitude of the recoverable behavioral gain: representation-level degradation persists at larger scales, while the benefit of intervention becomes more model dependent as the underlying model changes.

6 Conclusion

We introduced IIF analysis to track how proof-relevant fact–rule information evolves across Transformer layers. IIF reveals a broad but heterogeneous family of layerwise failures: proof-selective structure often forms and later degrades, while some settings exhibit formation failure or other non-monotonic trajectories. Across ProverQA, PrOntoQA, and ProofWriter, FRGI consistently improves reasoning accuracy over representative prompting, symbolic, and activation-intervention baselines. Our findings show that, within logical reasoning, some failures can be diagnosed and partially repaired at the representation level without retraining or proof annotations at inference time.

Limitations Our study focuses on proof-structured deductive reasoning, where proof-relevant facts and rules can be identified for offline analysis. Although FRGI is proof-free at inference time, the IIF diagnostics rely on proof annotations and may not transfer directly to open-ended reasoning without an equivalent notion of proof relevance. UCP is also an unlabeled proxy whose effectiveness depends on reliable span structure and rule–fact organization. In addition, the benefit of FRGI varies across models, and representation degradation does not guarantee that intervention will help every setting. Finally, FRGI adds an extra diagnostic prefill and several inference-time hyperparameters. Extending the framework to less structured reasoning, longer contexts, and broader model families is an important direction for future work.

AI USE STATEMENT

Generative AI tools were used only for language editing, including grammar correction, vocabulary refinement, and improving the clarity and readability of the manuscript. They were not used to generate research claims, design experiments, analyze experimental results, or make scientific conclusions. All AI-assisted edits were reviewed and revised by the authors to ensure technical accuracy and consistency with the actual experiments. The authors take full responsibility for the final content of the paper, including all text, claims, results, and artifacts.

ETHICS STATEMENT

This work studies the internal representations of language models on publicly available logical reasoning benchmarks and does not involve human subjects, personal data, or private user information. The models and datasets used in our experiments are publicly available and are used only for research purposes. Our intervention method is designed to study and improve logical reasoning behavior, and we do not identify specific foreseeable harms beyond the general risks associated with language-model research. We report the scope and limitations of our claims explicitly and avoid interpreting our diagnostic results as a universal explanation of language-model reasoning failures.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the source code as a supplementary ZIP file accompanying the submission. The supplementary material includes the FRGI implementation, IIF diagnostic code, baseline evaluation scripts, benchmark-preparation instructions, and scripts for the main greedy- and beam-search experiments, ablations, efficiency analysis, and generation-budget checks. It also contains configuration files specifying model checkpoints, decoding settings, hyperparameters, answer parsing, and output formats. All evaluated models and benchmarks are publicly available, and further methodological and experimental details are documented in the main paper and appendix.

REFERENCES

Md Rifat Arefin, Gopeshh Subbaraj, Nicolas Gontier, Yann LeCun, Irina Rish, Ravid Shwartz-Ziv, and Christopher Pal. Seq-VCR: Preventing collapse in intermediate transformer representations for enhanced reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=30oIfmrcFO.

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João G.M. Araújo, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GPKTIktA0k.

Danchun Chen, Qiyao Yan, Chenpeng Wang, and Liangming Pan. Towards a mechanistic understanding of propositional logical reasoning in large language models, 2026. URL https://arxiv.org/abs/2601.04260.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/dong21a.html.

Gleb Gerasimov, Yaroslav Aksenov, Nikita Balagansky, Viacheslav Sinii, and Daniil Gavrilov. You do not fully utilize transformer's representation capacity, 2026. URL https://openreview.net/forum?id=1CoJWNcpmm.

Guan Zhe Hong, Nishanth Dikkala, Enming Luo, Cyrus Rashtchian, Xin Wang, and Rina Panigrahy. A implies b: Circuit analysis in LLMs for propositional logical reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=M0U8wUow8c.

Geonhee Kim, Marco Valentino, and Andre Freitas. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10074–10095, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.525. URL https://aclanthology.org/2025.findings-acl.525/.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: theoretical perspectives and the role of rank collapse. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Matteo Pagliardini, Amirkeivan Mohtashami, Francois Fleuret, and Martin Jaggi. Denseformer: enhancing information flow in transformers via depth weighted averaging. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3806–3824, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.248. URL https://aclanthology.org/2023.findings-emnlp.248/.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Chengwen Qi, Ren Ma, Bowen Li, He Du, Binyuan Hui, Jinwang Wu, Yuanjun Laili, and Conghui He. Large language models meet symbolic provers for logical reasoning evaluation. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=C25SgeXWjE.

Chengwen Qi, Ren Ma, Bowen Li, He Du, Binyuan Hui, Jinwang Wu, Yuanjun Laili, and Conghui He. Large language models meet symbolic provers for logical reasoning evaluation. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=C25SgeXWjE.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=qFVVBzXxR2V.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Peiyang Song, Pengrui Han, and Noah Goodman. Large language model reasoning failures. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=vnX1WHMNmz. Survey Certification.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Eduard Tulchinskii, Laida Kushnareva, Anastasia Voznyuk, Andrei Andriiainen, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. Quantifying logical consistency in transformers via query-key alignment. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35196–35211, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1785. URL https://aclanthology.org/2025.emnlp-main.1785/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2025. URL https://openreview.net/forum?id=2XBPdPIcFK.

Weixuan Wang, JINGYUAN YANG, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8WQ7VTfPTl.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13326–13365, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.720. URL https://aclanthology.org/2024.acl-long.720/.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550/.

Zhen Yang, Mingyang Zhang, Feng Chen, Ganggui Ding, Liang Hou, Xin Tao, and Ying-Cong Chen. Less is more: Improving llm reasoning with minimal test-time intervention, 2026. URL https://arxiv.org/abs/2510.13940.

Yujun Zhou, Jiayi Ye, Zipeng Ling, Yufei Han, Yue Huang, Haomin Zhuang, Zhenwen Liang, Kehan Guo, Taicheng Guo, Xiangqi Wang, and Xiangliang Zhang. Dissecting logical reasoning in

LLMs: A fine-grained evaluation and supervision study. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17075–17098, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.926. URL https://aclanthology.org/2025.findings-emnlp.926/.

APPENDIX CONTENTS

A Appendix: Extended Literature Review **15**

B Appendix: Diagnostic Definitions **16**

B.1 NCE-IIF: Contrastive Evidence Selectivity 16

B.2 CLUB-Style Probe Diagnostic 16

B.3 ValueIIF: Value-Channel Evidence Selectivity 17

B.4 Appendix: Geometric Diagnostics 17

B.5 Formal Rule-to-Fact IIF 17

C Appendix: Robustness of IIF Diagnostics **18**

C.1 Similarity-Function Stability 18

C.2 Temperature Sensitivity 19

C.3 Agreement across Diagnostic Families 19

C.4 Raw Geometric Profiles 20

C.5 Attention-Only Selectivity 20

D Appendix: Extended Cascade IIF Analysis **20**

D.1 Cross-Benchmark Diagnostic Protocol 21

D.2 Query-Anchored Profiles and Formation Failure 22

D.3 Rule-Interface Confirmation 23

D.4 Cross-Architecture Extension 23

D.5 Raw Evidence Separation across Benchmarks and Backbones 25

D.6 Effect of Proof Depth 27

D.7 Architecture-Specific Boundary Cases 28

E Appendix: Full Method Description of FRGI **29**

E.1 Notation and Evidence Structure 30

E.2 Pass 1: Proof-Free Diagnostic Prefill 30

E.3 Pass 2: Fact–Rule Grouped Intervention 32

E.4 Prompt KV-Cache Construction and Decoding 33

E.5 Abstention and Fallback Behavior 33

F Appendix: Baseline Implementation Details **34**

F.1 Prompting Baselines 34

F.2 Activation-Steering Baselines 34

F.3 Structured Inference Baselines 35

G Appendix: Extra Results for Analysis **36**

G.1 Additional Beam-Search Results 36

G.2 Ablation Studies 37
G.3 Efficiency Analysis 39
G.4 Generation Budget and Max-Token Stability 40
G.5 Case studies under label scoring 41

A Appendix: Extended Literature Review

Reasoning failures in LLMs. A growing literature documents systematic weaknesses in LLM reasoning. These include reversal failures (Berglund et al., 2024), compositionality gaps (Press et al., 2023), and sensitivity to irrelevant or disordered context (Shi et al., 2023). More broadly, recent surveys show that such failures persist across tasks, prompting settings, and model families (Song et al., 2026). On deductive reasoning benchmarks such as PrOntoQA, ProofWriter, and ProverQA, model performance often deteriorates as proof depth increases or premise selection becomes more challenging (Saparov & He, 2023; Tafjord et al., 2021; Qi et al., 2025b). In contrast to this behavioral line of work, our goal is to explain *where* the reasoning process breaks inside the model by tracing the layerwise flow of proof-relevant information.

Prompting and solver-based reasoning methods. Many approaches improve logical reasoning by modifying the input, decoding strategy, or external computation available to the model. Chain-of-thought prompting and self-consistency remain standard test-time tools for eliciting stronger reasoning behavior (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023). More task-specific methods incorporate symbolic structure, such as Logic-LM (Pan et al., 2023) and Symbolic Chain-of-Thought (Xu et al., 2024). These methods can substantially improve task accuracy, but they operate primarily at the input-output level and do not directly analyze internal evidence-processing dynamics. By contrast, our work studies the internal mechanism of reasoning failure and uses that diagnosis to motivate FRGI, an inference-time intervention at the representation level.

Representation collapse in Transformers. A separate line of work studies how Transformer representations degrade with depth. Theoretical analyses show that self-attention can suffer rank collapse (Dong et al., 2021), while later work investigates signal propagation and its optimization consequences in deep Transformers (Noci et al., 2022). More recent work links decoder-only language models to information over-squashing and proposes architectural or training-time remedies, such as Seq-VCR, LIMe, and DenseFormer, to preserve richer intermediate representations (Barbero et al., 2024; Arefin et al., 2025; Gerasimov et al., 2026; Pagliardini et al., 2024). Our work differs in both target and measurement: rather than tracking only generic geometric degradation, we ask whether internal reasoning states still preserve selectivity toward proof-relevant evidence over distractors.

Mechanistic analyses of logical reasoning. Recent work has begun to examine logical reasoning directly through internal model computations. Query–key alignment has been used as a lightweight diagnostic of logical consistency, identifying attention-head signals that distinguish valid from invalid inferences (Tulchinskii et al., 2025). Complementary mechanistic studies identify sparse causal circuits for syllogistic and propositional reasoning, including specialized attention heads and modular sub-circuits supporting distinct stages of inference (Kim et al., 2025; Hong et al., 2025). Other work probes whether intermediate entities underlying multi-hop reasoning are represented latently (Yang et al., 2024), evaluates reasoning jointly at the answer, proof-step, and representation levels (Zhou et al., 2025), or characterizes staged information transmission across layers in propositional reasoning (Chen et al., 2026). Our focus is complementary: rather than identifying a task-specific circuit or testing whether a particular intermediate proposition is encoded, IIF measures how selectively proof-relevant facts and rules are preserved relative to unused evidence across depth. This exposes a characteristic rise–peak–fall profile and allows us to study when useful inferential information deteriorates after it has already emerged.

Inference-time activation interventions. Another related direction modifies internal activations at inference time. Activation Addition and related steering methods show that residual-stream interventions can shift model behavior without retraining (Turner et al., 2025; Rimsky et al., 2024; Wang et al., 2025). More recently, Minimal Test-Time Intervention (MTI) improves reasoning by applying selective guidance at uncertain decoding positions (Yang et al., 2026). While these methods demonstrate the promise of test-time control, they are not derived from a reasoning-specific diagnostic

of evidence processing. In contrast, FRGI is motivated by IIF diagnostics but uses a proof-free inference pipeline: UCP selects intervention layers, while attention saliency selects candidate fact and rule spans.

B Appendix: Diagnostic Definitions

This appendix provides the detailed definitions of the five evidence-selectivity diagnostics summarized in §2. These scores provide complementary operational views of the IIF quantity in Eq. equation 2; we do not interpret any individual diagnostic as a calibrated estimator or strict bound on the conditional mutual information.

B.1 NCE-IIF: Contrastive Evidence Selectivity

For evidence span e_j in instance x, define the layer-l critic score

$$f_{x,j}^{(l)} = \frac{s\left(\boldsymbol{z}_x^{(l)}, \boldsymbol{r}_{x,j}^{(l)}\right)}{\tau}, \tag{6}$$

where $\tau > 0$ is a temperature and $s(\boldsymbol{a}, \boldsymbol{b}) = \cos(\boldsymbol{a}, \boldsymbol{b})$ by default.

We compute the instance-normalized multi-positive contrastive score

$$\text{NCE-IIF}^{(l)} = \mathbb{E}_{x\sim\mathcal{D}}\left[\log\frac{M_x}{|\mathcal{I}_{\mathcal{U}_x}|} + \log\frac{\sum_{j\in\mathcal{I}_{\mathcal{U}_x}}\exp(f_{x,j}^{(l)})}{\sum_{j\in\mathcal{I}_{\mathcal{U}_x}\cup\mathcal{I}_{\mathcal{N}_x}}\exp(f_{x,j}^{(l)})}\right]. \tag{7}$$

The first term corrects for differences in the number of proof-used spans across instances. If all spans receive equal critic scores, the normalized quantity is zero; positive values indicate that proof-used evidence receives more contrastive mass than expected from its count alone.

Higher $\text{NCE-IIF}^{(l)}$ therefore indicates stronger used–versus–unused evidence selectivity. Because this is a multi-positive construction with variable numbers of evidence spans, we use NCE-IIF as an MI-inspired contrastive diagnostic rather than as a calibrated mutual-information lower bound. We use $\tau = 0.07$ by default and evaluate sensitivity to this choice in Appendix C.2.

B.2 CLUB-Style Probe Diagnostic

A decline in NCE-IIF alone could reflect properties of the contrastive construction rather than a general loss of evidence selectivity. We therefore introduce a complementary probe-based diagnostic.

For each evidence span, define

$$c_{x,j}^{(l)} = s\left(\boldsymbol{z}_x^{(l)}, \boldsymbol{r}_{x,j}^{(l)}\right), \tag{8}$$

and let $Y_{x,j} = 1$ if $e_j \in \mathcal{U}_x$ and $Y_{x,j} = 0$ otherwise. At each layer, we fit a one-dimensional logistic probe,

$$q_\theta\left(Y_{x,j} = 1 \mid c_{x,j}^{(l)}\right) = \text{sigmoid}\left(\alpha c_{x,j}^{(l)} + \beta\right). \tag{9}$$

Following a CLUB-style construction, we compute

$$I_{\text{CLUB}}^{(l)} = \mathbb{E}_x\left[\frac{1}{M_x}\sum_{j=1}^{M_x}\log q_\theta\left(Y_{x,j} \mid c_{x,j}^{(l)}\right) - \frac{1}{M_x^2}\sum_{j=1}^{M_x}\sum_{j'=1}^{M_x}\log q_\theta\left(Y_{x,j'} \mid c_{x,j}^{(l)}\right)\right]. \tag{10}$$

The second term pairs each query–span similarity with labels from other spans in the same instance, providing a product-of-marginals-style comparison while preserving instance-level label balance.

We use $I_{\text{CLUB}}^{(l)}$ as a CLUB-style probe diagnostic, rather than as a strict upper bound on Eq. equation 2. Because the probe receives only a one-dimensional similarity feature, its role is to compare layerwise trends rather than to serve as a standalone proof-usage classifier. Agreement with NCE-IIF provides complementary evidence of changes in evidence selectivity.

B.3 VALUEIIF: VALUE-CHANNEL EVIDENCE SELECTIVITY

The query and evidence representations above lie in residual-stream space, whereas attention values are computed in value space. To examine whether evidence selectivity is specifically preserved in the value channel, we map value content back into residual space.

For evidence span e_j, define

$$\boldsymbol{u}_{x,j}^{(l)} = \frac{1}{|\mathrm{pos}(e_j)|} \sum_{i \in \mathrm{pos}(e_j)} W_O^{(l)} W_V^{(l)} \mathrm{Norm}^{(l)} \left(\boldsymbol{h}_{x,i}^{(l-1)} \right), \tag{11}$$

where $W_V^{(l)}$ and $W_O^{(l)}$ denote the value and attention-output projections.

We define *ValueIIF* by replacing the residual-stream evidence representation $\boldsymbol{r}_{x,j}^{(l)}$ in NCE-IIF with $\boldsymbol{u}_{x,j}^{(l)}$. ValueIIF therefore measures whether proof-used and unused evidence remain selectively represented in value-channel content before query-dependent attention weighting.

A stronger post-peak decline in ValueIIF than in the full hidden-state diagnostics suggests that evidence-selectivity loss is concentrated in the value channel. We additionally report value-space ranking analyses as robustness checks.

B.4 APPENDIX: GEOMETRIC DIAGNOSTICS

The preceding diagnostics depend on design choices such as temperature or probe fitting. We therefore complement them with two parameter-light geometric measures.

Instance-level Center Alignment (ICA). Let $\boldsymbol{\mu}_{\mathcal{U}_x}^{(l)}$ and $\boldsymbol{\mu}_{\mathcal{N}_x}^{(l)}$ denote the centroids of proof-used and unused evidence representations, respectively. We define

$$\mathrm{ICA}^{(l)} = \mathbb{E}_x \left[s\left(\boldsymbol{z}_x^{(l)}, \boldsymbol{\mu}_{\mathcal{U}_x}^{(l)} \right) - s\left(\boldsymbol{z}_x^{(l)}, \boldsymbol{\mu}_{\mathcal{N}_x}^{(l)} \right) \right]. \tag{12}$$

Positive ICA indicates stronger alignment between the query and the centroid of proof-used evidence than between the query and the distractor centroid.

TripletAcc. TripletAcc measures how often a proof-used span is ranked above an unused span under the same similarity function:

$$\mathrm{TripletAcc}^{(l)} = \mathbb{E}_x \mathbb{E}_{\substack{j^+ \in \mathcal{I}_{\mathcal{U}_x} \\ j^- \in \mathcal{I}_{\mathcal{N}_x}}} \left[\mathbf{1} \left\{ s\left(\boldsymbol{z}_x^{(l)}, \boldsymbol{r}_{x,j^+}^{(l)} \right) > s\left(\boldsymbol{z}_x^{(l)}, \boldsymbol{r}_{x,j^-}^{(l)} \right) \right\} \right]. \tag{13}$$

A value of 0.5 corresponds to chance-level pairwise ranking, whereas larger values indicate stronger evidence selectivity.

B.5 FORMAL RULE-TO-FACT IIF

The query-anchored diagnostics characterize whether the model's current reasoning state distinguishes proof-used evidence from distractors. To analyze the rule–fact interface directly, we define an analogous rule-anchored quantity.

For instance x, let $\mathcal{R}_x$ and $\mathcal{F}_x$ denote the sets of rule-span and fact-span indices. For each proof-used rule $a \in \mathcal{R}_x$, let $\mathcal{F}_x^+(a) \subseteq \mathcal{F}_x$ denote the facts that satisfy the antecedent of a and participate with it in at least one step of the annotated proof. Let

$$\mathcal{F}_x^-(a) = \mathcal{F}_x \setminus \mathcal{F}_x^+(a) \tag{14}$$

denote the remaining facts.

We sample

$$J \mid X = x, A = a \sim \mathrm{Unif}(\mathcal{F}_x)$$

and define

$$Y_{x,a,J} = \mathbf{1}\left\{J \in \mathcal{F}_x^+(a)\right\}. \tag{15}$$

Let $\boldsymbol{r}_{x,a}^{(l)}$ be the representation of the rule anchor and $\boldsymbol{r}_{x,J}^{(l)}$ the representation of the sampled fact. Define the rule–fact relation

$$G_{x,a,J}^{(l)} = \phi\Big(\boldsymbol{r}_{x,a}^{(l)}, \boldsymbol{r}_{x,J}^{(l)}\Big). \tag{16}$$

The corresponding rule-to-fact IIF is

$$\text{R2F-IIF}^{(l)} = \mathbb{E}_x \mathbb{E}_{a \in \mathcal{R}_x : |\mathcal{F}_x^+(a)|>0}\left[I\Big(G_{x,a,J}^{(l)}; Y_{x,a,J} \mid X = x, A = a\Big)\right]. \tag{17}$$

In practice, we apply the same five diagnostics with the rule representation replacing the query anchor. Facts in $\mathcal{F}_x^+(a)$ are treated as positives and facts in $\mathcal{F}_x^-(a)$ as negatives. Rules without an annotated positive fact are excluded because their positive set is empty.

All proof annotations used in this appendix serve only to define offline diagnostic labels. They are not used for deployed inference.

C Appendix: Robustness of IIF Diagnostics

The main analysis uses cosine similarity, a fixed contrastive temperature, and an aggregate over five complementary diagnostics. Here we test whether the observed layerwise behavior depends substantially on these measurement choices. We examine sensitivity to the similarity function and temperature, agreement across diagnostic families, raw unnormalized geometric profiles, and evidence selectivity in attention allocation alone.

C.1 Similarity-Function Stability

Cosine similarity is our default non-parametric relation score. We do not assume that it is uniquely optimal; instead, we ask whether the layerwise R2F-IIF pattern remains stable under alternative definitions of rule–fact alignment.

We repeat the R2F analysis using seven scoring functions:

$\{$COSINE, DOT PRODUCT, EUCLIDEAN, MANHATTAN, PEARSON, JENSEN–SHANNON, MAHALANOBIS$\}$.

All scores are oriented so that larger values indicate stronger alignment. Distance measures are negated. For Jensen–Shannon distance, representations are mapped to probability vectors using a softmax transformation. For Mahalanobis distance, the covariance matrix $\boldsymbol{\Sigma}$ is estimated on the diagnostic split and regularized with $\lambda \boldsymbol{I}$.

For each scoring function, we report three quantities. First, the normalized post-peak drop

$$\Delta_{\downarrow} = \widehat{\text{R2F-IIF}}^{(l_{\text{peak}})} - \frac{1}{3}\sum_{l \in \mathcal{L}_{\text{last3}}} \widehat{\text{R2F-IIF}}^{(l)}, \tag{18}$$

where positive $\Delta_{\downarrow}$ indicates post-peak degradation.

Second, we measure agreement with the cosine profile using

$$\rho_{\cos} = \text{Spearman}\left(\widehat{\text{R2F-IIF}}_{\text{score}}^{(1:L)}, \widehat{\text{R2F-IIF}}_{\cos}^{(1:L)}\right). \tag{19}$$

Third, we recompute $\Delta_{\downarrow}$ over five random subsamples and report its standard deviation.

Table 5 shows that the post-peak decline is preserved across scoring functions. The alternative profiles also have high rank agreement with cosine, particularly for LLaMA and Mistral. Cosine does not always yield the largest decline; rather, it provides a stable, scale-invariant, and parameter-free representative of the broader family of alignment scores.

Table 5: **Similarity-function stability of R2F-IIF.** We report the normalized post-peak drop $\Delta_{\downarrow}$, Spearman agreement with the cosine-based profile $\rho_{\cos}$, and the standard deviation of the drop across five subsamples.

Scoring function	LLaMA-3.1-8B			Mistral-7B			Qwen3-8B		
	$\Delta_{\downarrow}$	$\rho_{\cos}$	Std.	$\Delta_{\downarrow}$	$\rho_{\cos}$	Std.	$\Delta_{\downarrow}$	$\rho_{\cos}$	Std.
Cosine	0.42	1.00	0.018	0.51	1.00	0.021	0.15	1.00	0.019
Dot product	0.39	0.93	0.031	0.47	0.91	0.034	0.13	0.88	0.028
Euclidean	0.41	0.95	0.026	0.49	0.93	0.030	0.14	0.86	0.027
Manhattan	0.38	0.92	0.034	0.45	0.90	0.037	0.12	0.84	0.031
Pearson	0.43	0.97	0.020	0.50	0.95	0.024	0.15	0.89	0.022
Jensen–Shannon	0.35	0.89	0.041	0.42	0.87	0.045	0.10	0.81	0.038
Mahalanobis	0.37	0.91	0.036	0.44	0.89	0.039	0.11	0.83	0.035

Table 6: **Temperature sensitivity of NCE-based R2F-IIF.** ρ measures agreement with the default $\tau = 0.07$ layerwise profile.

Temperature τ	$\boldsymbol{\rho}$ **vs.** $\tau = 0.07$	$\boldsymbol{\Delta_{\downarrow}}$
0.01	0.96	0.40
0.05	0.99	0.42
0.07	1.00	0.42
0.10	0.99	0.41
0.50	0.95	0.37

C.2 Temperature Sensitivity

We next test whether NCE-IIF depends strongly on the contrastive temperature. We compute layerwise R2F-IIF profiles for

$$\tau \in \{0.01, 0.05, 0.07, 0.10, 0.50\}$$

on 200 ProverQA examples using LLaMA-3.1-8B. We compare each normalized trajectory with the default $\tau = 0.07$ profile using Spearman correlation and report the corresponding post-peak drop.

All tested temperatures yield Spearman correlations of at least 0.95 with the default profile, while the post-peak drop remains positive. The observed trajectory is therefore not specific to the choice $\tau = 0.07$.

C.3 Agreement across Diagnostic Families

We next ask whether the aggregate IIF profile is driven by a single diagnostic. Figure 7 shows the individual ValueIIF, NCE-IIF, CLUB-style, ICA, and TripletAcc trajectories, while Figure 8 reports their pairwise correlations across layers.

Mistral-7B. Mistral shows the clearest convergence across diagnostic families. All pairwise correlations are positive and at least 0.71, with strong agreement between ValueIIF and the CLUB-style probe ($\rho = 0.85$), ValueIIF and ICA ($\rho = 0.85$), and ICA and TripletAcc ($\rho = 0.85$). Thus, the observed layerwise profile is not driven by a single metric family.

LLaMA-3.1-8B. For LLaMA, the contrastive and probe-based diagnostics agree strongly: ValueIIF–NCE-IIF reaches $\rho = 0.66$, ValueIIF–CLUB reaches $\rho = 0.77$, and NCE-IIF–CLUB reaches $\rho = 0.73$. The geometric metrics also agree strongly, with ICA–TripletAcc reaching $\rho = 0.81$. Cross-family correlations are weaker, indicating that the diagnostics capture complementary aspects of the same layerwise behavior.

Qwen3-8B. Qwen3 shows greater disagreement across diagnostic families. For example, ValueIIF and ICA have near-zero correlation ($\rho = 0.06$), while ValueIIF remains correlated with the CLUB-style probe ($\rho = 0.68$) and CLUB with TripletAcc ($\rho = 0.59$). We interpret this divergence conservatively: the diagnostics capture different aspects of Qwen3's layerwise representations, and correlation alone does not establish a specific internal routing mechanism.

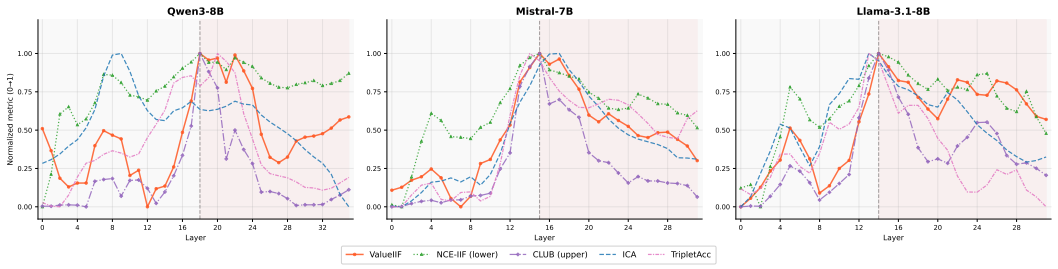

Fig. 7: **Individual IIF diagnostics across layers.** ValueIIF, NCE-IIF, CLUB-style, ICA, and TripletAcc trajectories for Qwen3-8B, Mistral-7B, and LLaMA-3.1-8B. The aggregate layerwise behavior is supported by multiple diagnostic families rather than a single metric.

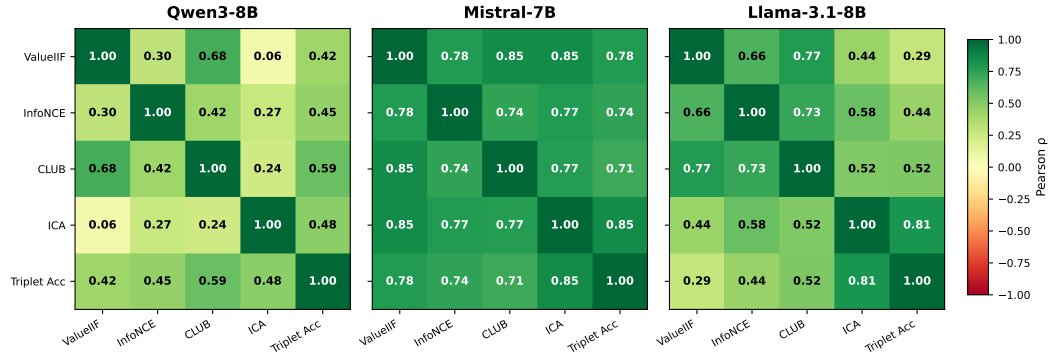

Fig. 8: **Pairwise agreement between IIF diagnostics.** Pearson correlations between ValueIIF, NCE-IIF, the CLUB-style probe, ICA, and TripletAcc across layers. Mistral shows strong agreement across diagnostics, whereas LLaMA and Qwen3 exhibit more structured differences between contrastive/probe-based and geometric diagnostics.

Takeaway. The five diagnostics are complementary rather than redundant. Their shared post-peak behavior provides convergent evidence for degradation, while differences between them indicate whether that degradation is more visible in value content, contrastive selectivity, centroid geometry, or pairwise ranking.

C.4 Raw Geometric Profiles

To verify that the observed post-peak behavior is not introduced by min–max normalization or diagnostic averaging, we additionally report the raw ICA and TripletAcc trajectories on the 5,000-example ProverQA diagnostic split.

C.5 Attention-Only Selectivity

Finally, we ask whether proof-used versus unused selectivity is already visible in raw attention allocation. We compute an AttnIIF diagnostic from the attention mass assigned to proof-used and unused evidence positions.

Across all three models, AttnIIF remains close to zero (order 10^{-5}) throughout the network. Thus, under this diagnostic, raw attention allocation alone exhibits little proof-used versus distractor selectivity; the stronger layerwise structure observed in the main analysis is more visible in residual- and value-content representations.

D Appendix: Extended Cascade IIF Analysis

This appendix expands the representation-level analysis in §3.1. We ask four questions: (i) whether the query-anchored IIF profile persists across benchmarks, (ii) whether the rule interface exhibits the same layerwise degradation when query-level selectivity is weak, (iii) whether the phenomenon

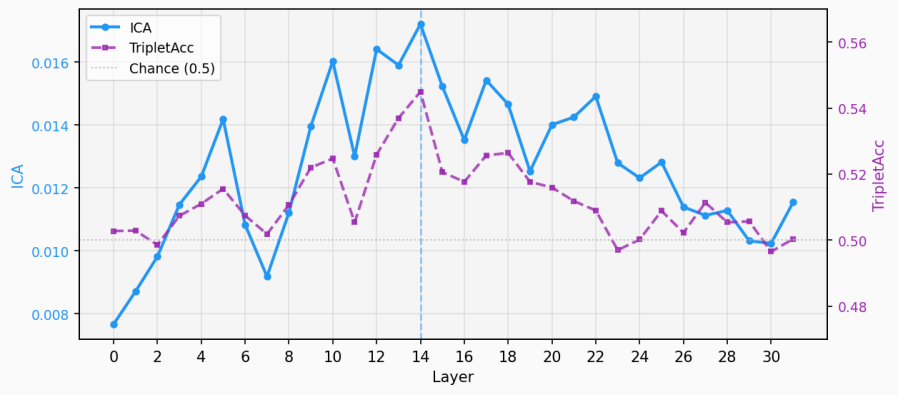

Fig. 9: Raw ICA and TripletAcc profiles for LLaMA-3.1-8B on ProverQA. Both measures strengthen toward intermediate layers and weaken toward the output, with TripletAcc approaching chance-level ranking in later layers.

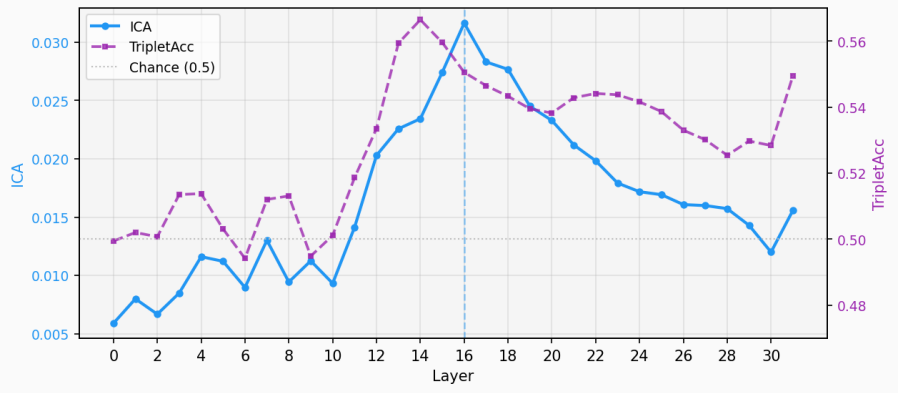

Fig. 10: Raw ICA and TripletAcc profiles for Mistral-7B on ProverQA. Both diagnostics exhibit a clear post-peak decline, consistent with the aggregate layerwise analysis.

extends across architectures and model scales, and (iv) whether the observed post-peak decline can be explained simply by proof depth.

Throughout this appendix, proof annotations are used only to construct offline diagnostic labels. They are never provided to the model during inference.

D.1 Cross-Benchmark Diagnostic Protocol

The main analysis evaluates the layerwise IIF profile on ProverQA. Downstream accuracy gains on another benchmark would not, by themselves, establish that the same representation-level behavior is present there. We therefore recompute the diagnostics directly on PrOntoQA and ProofWriter.

For PrOntoQA, we use a label-balanced split of 800 instances (400 true and 400 false) with proof depths ranging from 4 to 13. For ProofWriter, we use 1,500 balanced depth-5 open-world instances. For each instance, we perform a single forward pass over the prompt without answer generation and extract representations at all layers.

At each layer, we compute the five diagnostics introduced in §2: NCE-IIF, the CLUB-style probe, ValueIIF, ICA, and TripletAcc. Each diagnostic is smoothed with the same three-layer window and min–max normalized within each model–benchmark cell. The normalized trajectories are then averaged.

We use the same peak and post-peak definitions throughout. Let l_{peak} denote the canonical interior maximum of the smoothed aggregate. Its relative post-peak drop is

$$\Delta_{\downarrow} = \frac{S^{(l_{\text{peak}})} - \frac{1}{3}\sum_{l \in \mathcal{L}_{\text{last3}}} S^{(l)}}{S^{(l_{\text{peak}})}} \times 100\%, \tag{20}$$

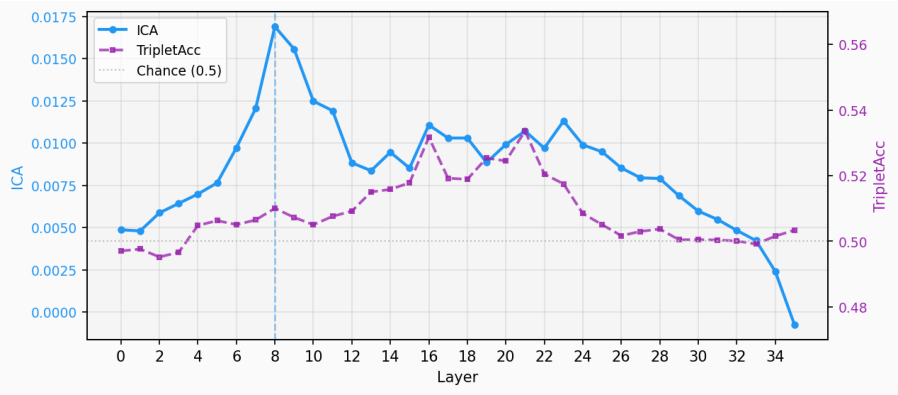

Fig. 11: Raw ICA and TripletAcc profiles for Qwen3-8B on ProverQA. The trajectories are noisier than for LLaMA and Mistral, consistent with the greater cross-diagnostic variation observed for Qwen3.

Table 7: **Attention-only evidence selectivity.** Mean AttnIIF at the peak R2F-IIF layer and the final layer on the 5,000-example ProverQA diagnostic split.

Model	Peak layer	AttnIIF at peak	Final layer
LLaMA-3.1-8B	L13	2.3×10^{-5}	1.8×10^{-5}
Mistral-7B	L18	1.9×10^{-5}	1.4×10^{-5}
Qwen3-8B	L19	2.1×10^{-5}	1.6×10^{-5}

where $S^{(l)}$ denotes the aggregate selectivity score. A positive value indicates degradation after the peak; a negative value indicates that the final-layer aggregate exceeds the selected interior peak.

Keeping this protocol fixed is important: differences across cells therefore reflect differences in the measured representation dynamics rather than changes in the diagnostic procedure.

D.2 Query-Anchored Profiles and Formation Failure

Table 8 gives the exact peak locations and post-peak changes underlying the cross-benchmark profiles reported in the main text.

ProverQA. The clearest query-anchored cascade occurs on ProverQA. LLaMA-3.1-8B, Mistral-7B, and Qwen3-8B reach their maximum aggregate selectivity at L14, L15, and L19, respectively, and subsequently lose approximately 63%, 60%, and 60% of that peak selectivity toward the output. The strongest proof-selective representation therefore occurs well before the final prediction layer.

PrOntoQA. Balanced PrOntoQA exhibits the same qualitative pattern across the three 8B backbones. LLaMA-3.1-8B and Mistral-7B peak at L14 and L16 and show 77% post-peak declines, while Qwen3-8B peaks later at L23 and declines by 45%. Thus, the rise–peak–fall trajectory is not specific to ProverQA, although its location and severity remain architecture dependent.

ProofWriter. The query-anchored picture is qualitatively different on ProofWriter. Several cells remain weak or unstable rather than developing a clean interior maximum. In particular, the LLaMA aggregate rises again near the output, yielding a negative post-peak change. The underlying raw diagnostics are likewise weak in several regions of the network, so we do not interpret a numerical drop in isolation as evidence of a robust query-level collapse.

This motivates a distinction between two failure modes. We call a trajectory *formation failure* when proof-used evidence never becomes strongly selective at the query representation. We reserve *cascade IIF collapse* for the case in which selectivity first becomes strong at an intermediate depth and is subsequently lost. A weak final representation can therefore arise either because useful structure was never formed or because it was formed and later degraded.

Table 8: **Query-anchored IIF statistics across benchmarks and backbones.** Each cell reports the canonical peak layer (out of $L-1$) and the relative post-peak drop. A negative value indicates that the aggregate increases rather than decreases after the selected peak.

Backbone	ProverQA	PrOntoQA	ProofWriter
LLaMA-3.1-8B	L14/31 (63%)	L14/31 (77%)	L22/31 (-50%)
Mistral-7B	L15/31 (60%)	L16/31 (77%)	L4/31 (48%)
Qwen3-8B	L19/35 (60%)	L23/35 (45%)	L26/35 (22%)
Qwen3-14B	L20/39 (47%)	L26/39 (4%)	L13/39 (14%)

Table 9: **Rule-interface selectivity across benchmarks.** Each cell reports the peak layer of the smoothed answer-to-rule selectivity curve and its relative post-peak drop.

Backbone	ProverQA	PrOntoQA	ProofWriter
LLaMA-3.1-8B	L14 (14%)	L17 (47%)	L14 (33%)
Mistral-7B	L15 (32%)	L19 (34%)	L16 (48%)
Qwen3-8B	L25 (35%)	L25 (57%)	L23 (29%)

D.3 Rule-Interface Confirmation

Weak query-level selectivity does not imply that proof structure is absent everywhere in the network. We therefore examine a more local interface using answer-to-rule selectivity, answer-to-fact selectivity, and rule-to-fact coupling.

We sample 300 instances per model–benchmark cell. Rule-specific statistics are computed only on instances containing usable rule structure, yielding 116 rule-bearing ProverQA instances, 200 PrOntoQA instances, and 169 ProofWriter instances.

Table 9 reports the peak layer and post-peak drop of the smoothed answer-to-rule selectivity.

Unlike the query-anchored aggregate, answer-to-rule selectivity forms an interior maximum followed by a post-peak decline in every evaluated 8B model–benchmark pair. Across ProverQA, PrOntoQA, and ProofWriter, the drops are 14%, 47%, and 33% for LLaMA-3.1-8B; 32%, 34%, and 48% for Mistral-7B; and 35%, 57%, and 29% for Qwen3-8B.

ProofWriter is especially informative. Although its query representation can fail to develop a strong proof-used versus distractor separation, the rule interface still develops a selective intermediate state and subsequently loses part of it. Hence, the weak query-level profile does not imply an absence of structured internal degradation.

We therefore interpret query- and rule-anchored diagnostics as complementary views. Query-level analysis reveals whether proof selectivity reaches the final reasoning state, whereas rule-interface analysis tests whether a more local proof structure forms and later deteriorates.

PrOntoQA coupling limitation. The auxiliary rule-to-fact coupling curve is not informative on PrOntoQA under the current surface parser. PrOntoQA expresses many rules as quantified category statements rather than named-entity conditionals, and the parser therefore identifies no governed rule–fact pairs. We do not interpret the resulting zero coupling curve as evidence that rule–fact structure is absent. This limitation affects only the auxiliary coupling diagnostic; answer-to-rule, answer-to-fact, and query-anchored IIF remain available.

D.4 Cross-Architecture Extension

We next test whether the layerwise IIF behavior observed for the three primary backbones extends to a broader set of architectures and model scales. Figure 12 applies the same five-diagnostic analysis to six additional backbones: LLaMA-3.2-3B, Qwen3-14B, Phi-4-mini, DeepSeek-R1-Distill-LLaMA-8B, Qwen2.5-7B, and Mistral-Nemo-12B.

For every available model–benchmark pair, we use exactly the same analysis protocol as in the main experiments. Each diagnostic is smoothed and normalized within the corresponding model–benchmark cell, the five trajectories are averaged to form the aggregate IIF profile, and the canonical

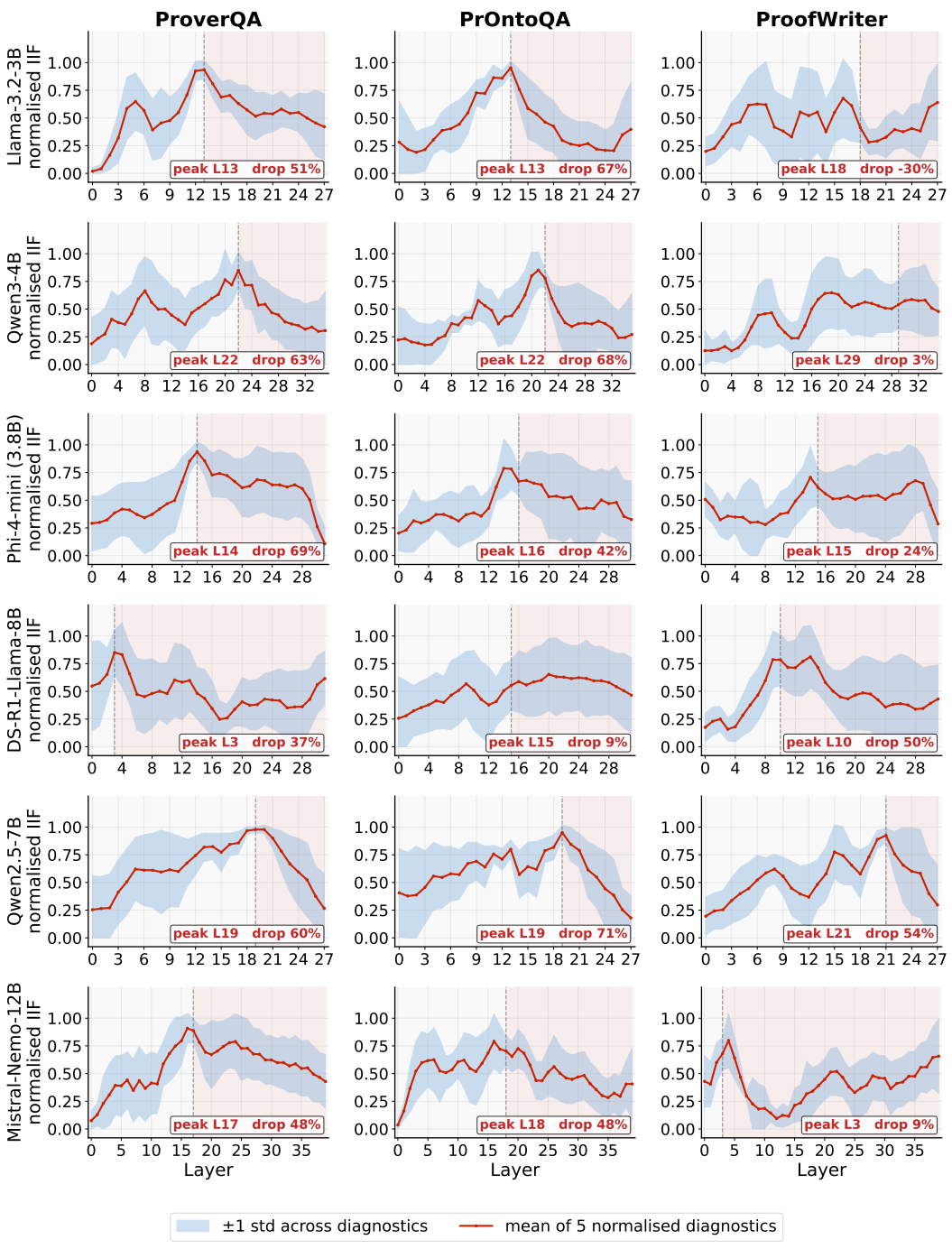

Fig. 12: **Cross-architecture IIF profiles for additional backbones.** Rows denote six additional model backbones and columns denote ProverQA, balanced PrOntoQA, and ProofWriter. In each cell, the red curve is the mean of the five smoothed and normalized IIF diagnostics, while the blue band denotes ± 1 standard deviation across diagnostics. The dashed vertical line marks the canonical aggregate peak, and the annotation reports its layer and subsequent relative post-peak change. Positive values indicate degradation after the peak; negative values indicate that the aggregate later recovers rather than exhibiting sustained post-peak decline.

peak and subsequent post-peak change are computed from this aggregate. We do not introduce architecture-specific peak definitions or normalization choices.

The additional profiles reveal a broad but heterogeneous pattern. Many model–benchmark pairs develop a region of strong evidence selectivity at intermediate depth followed by substantial downstream

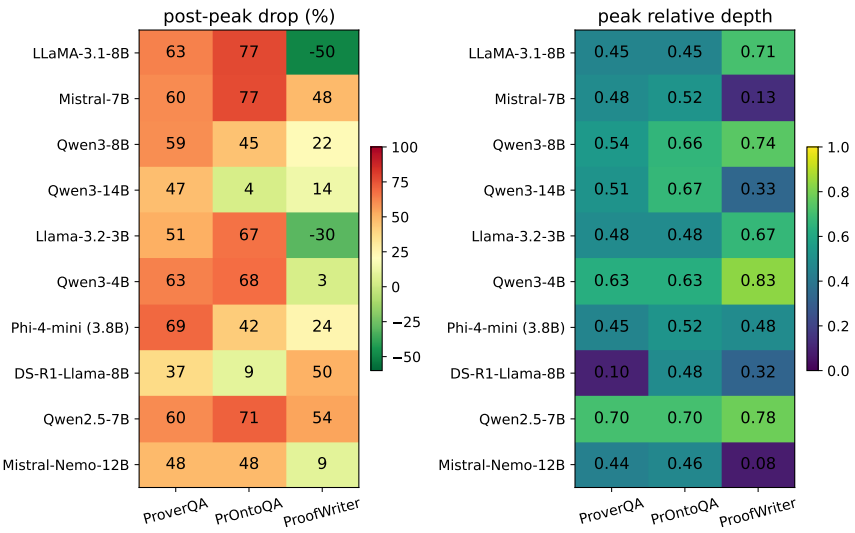

Fig. 13: **Cross-architecture summary of IIF degradation and peak location.** Left: relative post-peak change of the five-diagnostic aggregate; positive values denote downstream degradation, while negative values denote post-peak recovery. Right: canonical peak layer normalized by model depth. The summary shows that both collapse severity and peak location vary substantially across model families, scales, and reasoning benchmarks.

degradation, but neither the location nor the severity of this behavior is fixed across settings. For example, Qwen2.5-7B shows post-peak declines of 60%, 71%, and 54% on ProverQA, PrOntoQA, and ProofWriter, respectively. Phi-4-mini similarly exhibits strong degradation on ProverQA and PrOntoQA, whereas Mistral-Nemo-12B shows much weaker degradation on ProofWriter. Other settings expose qualitatively different regimes: DeepSeek-R1-Distill-LLaMA-8B peaks very early on ProverQA, Qwen3-14B exhibits only weak post-peak change on PrOntoQA and ProofWriter, and LLaMA-3.2-3B on ProofWriter shows post-peak recovery rather than sustained collapse.

To summarize this variation across both the primary and additional backbones, Figure 13 reports the post-peak change and canonical peak location for each model–benchmark pair.

Figure 13 makes two aspects of this heterogeneity particularly clear. First, substantial post-peak degradation appears across multiple architecture families and model scales, especially on ProverQA and PrOntoQA, indicating that the phenomenon is not specific to the three primary backbones. ProofWriter is more heterogeneous: some models exhibit pronounced degradation, whereas others show weak decline or post-peak recovery. Thus, the evidence supports a broadly recurring failure tendency rather than a universal trajectory shared by every model–benchmark pair.

Second, the canonical peak itself spans substantially different relative depths. Some settings peak early, others near the middle of the network, and several ProofWriter configurations peak much later. Consequently, neither a single absolute layer nor a fixed relative depth characterizes the failure across architectures. Instead, the strength, location, and even qualitative form of the IIF trajectory vary with both the model and the reasoning benchmark.

Taken together, these results show that cascade IIF collapse is *broad but heterogeneous*: proof-selective structure frequently forms and subsequently weakens across model families and scales, but its precise trajectory is setting dependent. This variability also motivates the proof-free adaptive layer-selection policy used by FRGI, rather than applying the same intervention depth to all models and inputs.

D.5 Raw Evidence Separation across Benchmarks and Backbones

The normalized IIF aggregate combines five diagnostics with different numerical scales. We therefore test whether the same non-monotonic behavior appears directly in the underlying representation geometry, without cross-diagnostic normalization.

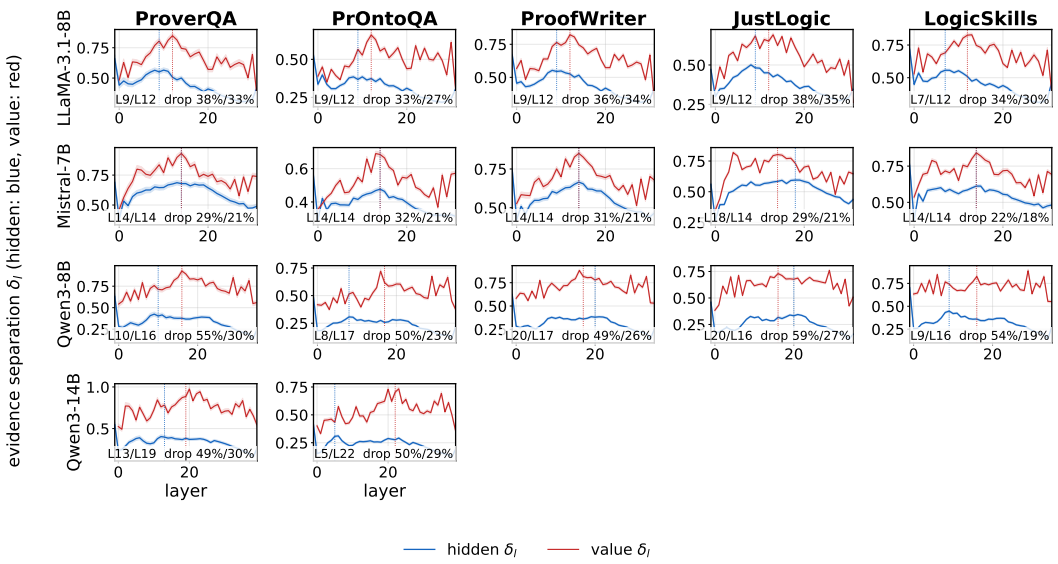

Fig. 14: **Raw evidence separation across benchmarks and backbones.** Layerwise proof-used–versus–distractor separation δ_l in hidden-state space (blue) and value space (red) for LLaMA-3.1-8B, Mistral-7B, and Qwen3-8B across five reasoning benchmarks. Each panel reports the hidden/value peak layers and relative post-peak drops. These curves are measured directly from the representation geometry and do not average or normalize across the five IIF diagnostics.

Table 10: **Raw evidence-separation statistics for the three primary backbones.** Each entry reports hidden/value peak layers followed by hidden/value post-peak drops.

Backbone	ProverQA	PrOntoQA	ProofWriter	JustLogic	LogicSkills
LLaMA-3.1-8B	L9/L12 (38%/33%)	L9/L12 (33%/27%)	L9/L12 (36%/34%)	L9/L12 (38%/35%)	L7/L12 (34%/30%)
Mistral-7B	L14/L14 (29%/21%)	L14/L14 (32%/21%)	L14/L14 (31%/21%)	L18/L14 (29%/21%)	L14/L14 (22%/18%)
Qwen3-8B	L10/L16 (55%/30%)	L8/L17 (50%/23%)	L20/L17 (49%/26%)	L20/L16 (59%/27%)	L9/L16 (54%/19%)

We measure proof-used versus distractor separation independently in hidden-state and value space. Figure 14 evaluates the three primary backbones on five benchmarks: ProverQA, PrOntoQA, ProofWriter, JustLogic, and LogicSkills.

The raw geometry reproduces the broad post-peak degradation without metric aggregation. For LLaMA-3.1-8B, hidden-state separation peaks between L7 and L9 across all five benchmarks and subsequently declines by 33–38%; its value-space separation peaks consistently around L12 and drops by 27–35%.

Mistral-7B shows a similarly stable profile. Hidden-state separation generally peaks at L14 (L18 on JustLogic) with drops of 22–32%, while value-space separation is concentrated near L14 with declines of 18–21%.

Qwen3-8B exhibits larger variation in peak location but stronger hidden-state degradation. Its hidden-state separation peaks around L8–L10 on ProverQA, PrOntoQA, and LogicSkills, but around L20 on ProofWriter and JustLogic. Despite this shift, its post-peak hidden-state drop remains large across all five benchmarks (49–59%). Value-space peaks occur later, around L16–L17, with smaller but consistent declines of 19–30%.

Importantly, ProofWriter is no longer exceptional under this raw geometric view. Its hidden-state drops are 36%, 31%, and 49% for LLaMA-3.1-8B, Mistral-7B, and Qwen3-8B, respectively. Thus, weak query-anchored IIF on ProofWriter does not imply an absence of internal representational degradation; the query anchor and raw evidence geometry capture different aspects of the layerwise dynamics.

Raw geometry across additional architectures.

Table 11: **Raw evidence separation across additional architectures.** Each cell reports hidden/value peak layers followed by hidden/value post-peak drops. Gemma-E2B/E4B are omitted because the raw geometric quantity is degenerate for these architectures.

Backbone	ProverQA	PrOntoQA
Qwen3-14B	L13/L19 (49%/30%)	L5/L22 (50%/29%)
Llama-3.2-1B	L6/L8 (36%/21%)	L0/L15 (36%/11%)
Llama-3.2-3B	L9/L11 (38%/33%)	L12/L11 (35%/25%)
Qwen2.5-1.5B	L1/L16 (55%/13%)	L3/L27 (58%/6%)
Qwen2.5-3B	L8/L20 (53%/25%)	L8/L32 (52%/28%)
Qwen3-4B	L9/L16 (48%/28%)	L8/L17 (45%/20%)
Phi-4-mini (3.8B)	L9/L12 (38%/59%)	L9/L13 (48%/61%)
DS-R1-Llama-8B	L8/L12 (28%/32%)	L9/L12 (24%/19%)
Qwen2.5-7B	L8/L17 (34%/6%)	L7/L27 (37%/5%)
Mistral-Nemo-12B	L8/L10 (44%/31%)	L8/L17 (42%/22%)
DS-R1-Qwen-14B	L18/L22 (52%/13%)	L20/L24 (43%/4%)

We also evaluate raw evidence separation on additional dense decoder backbones for which both ProverQA and PrOntoQA measurements are available. Table 11 reports the resulting hidden/value peak locations and post-peak drops.

Across these additional architectures, hidden-state separation generally exhibits a substantial post-peak decline, although the relationship between hidden-state and value-space degradation varies considerably across families. This supports the existence of a broad layerwise degradation phenomenon while also showing that its precise geometry is architecture dependent.

Qwen2.5-1.5B is an informative boundary case. Its hidden-state separation is already maximal in the first few layers and declines from there, yielding a strong *fall* without a clear preceding *rise*. We therefore do not force every architecture into the same rise–peak–fall template.

D.6 Effect of Proof Depth

A remaining alternative explanation is that the layerwise profile is created by mixing reasoning problems of different proof lengths. Longer proofs may require evidence to remain useful deeper into the network, potentially shifting the apparent peak. We therefore stratify the raw hidden-state evidence separation by annotated proof depth.

For a depth group d, let $l_{\mathrm{peak}}(d)$ denote the layer at which the smoothed hidden-state separation $\delta_l^{(d)}$ is maximal. We define

$$\Delta_{\downarrow}(d) = \frac{\delta^{(d)}_{l_{\mathrm{peak}}(d)} - \frac{1}{3}\sum_{l\in\mathcal{L}_{\mathrm{last3}}} \delta_l^{(d)}}{\delta^{(d)}_{l_{\mathrm{peak}}(d)}} \times 100\%. \tag{21}$$

Thus, $l_{\mathrm{peak}}(d)$ measures *where* evidence separation is strongest at depth d, whereas $\Delta_{\downarrow}(d)$ measures *how much* separation is subsequently lost.

Across the 15 model–benchmark combinations, 12 exhibit peak locations that vary by at most two layers across the available depth groups. The stability is especially clear on ProofWriter for LLaMA and Mistral, on PrOntoQA for Mistral and Qwen3, and across all three backbones on JustLogic and LogicSkills.

Peak location is not strictly invariant, however. Three cells show substantial shifts. Qwen3-8B on ProverQA peaks around L10 for depths 3–7 but moves to L21 at depth 8. On ProofWriter, Qwen3-8B

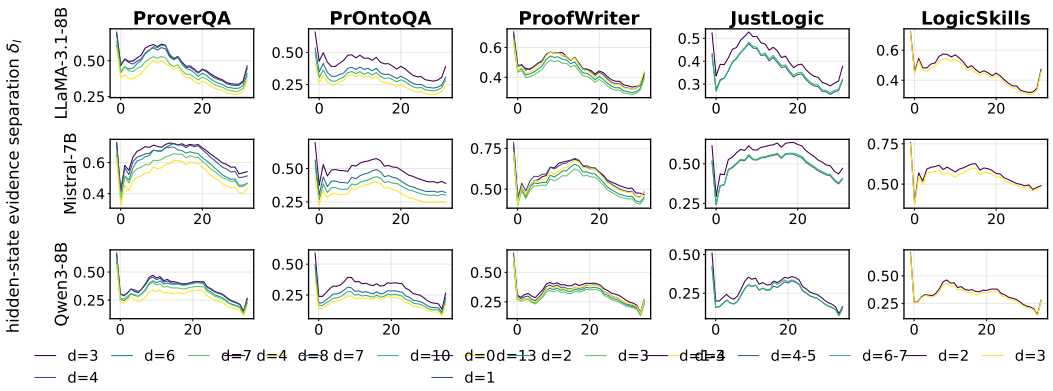

Fig. 15: **Hidden-state evidence separation across proof depths.** Each panel stratifies the raw separation δ_l by proof depth (or depth bin) for a single model–benchmark pair. ProverQA uses depths 3, 4, 6, 7, and 8; PrOntoQA uses depths 4, 7, 10, and 13; ProofWriter uses depths 0–4; JustLogic uses depth bins 1–3, 4–5, and 6–7; and LogicSkills uses depths 2 and 3. Most depth groups follow similar trajectories, although several cells show meaningful shifts in peak location.

moves from L10 at depth 0 to L20 at depths 1–4. LLaMA-3.1-8B on PrOntoQA also contains an atypical depth-7 profile peaking at L0 rather than near L8–L9.

The post-peak *magnitude* is considerably more stable than the peak location. Across all 15 cells, the range of $\Delta_{\downarrow}(d)$ across depths is at most nine percentage points, and in 14 of the 15 cells it is at most seven points. For example, Qwen3-8B retains a 54–56% drop across ProverQA depths despite the L10-to-L21 peak shift, and a 48–50% drop across ProofWriter depths despite the L10-to-L20 shift.

Thus, proof depth can alter *where* the strongest evidence separation occurs, but it does not explain away *whether* that separation subsequently deteriorates. The depth-stratified results therefore rule out the simple explanation that the aggregate rise–peak–fall profile is produced merely by mixing short and long proofs.

D.7 Architecture-Specific Boundary Cases

The broader analysis also exposes settings in which the standard rise–peak–fall description is not appropriate.

Early fall without a clear rise. Qwen2.5-1.5B reaches its strongest hidden-state separation within the first few layers (L1 on ProverQA and L3 on PrOntoQA) and then declines by 55–58%. This is a strong post-peak degradation, but not a canonical rise–peak–fall trajectory because the selective representation is already strongest near the input.

Degenerate Gemma geometry. For Gemma-4-E2B and Gemma-4-E4B, the raw hidden-state separation is already near its maximum from the earliest layers and changes little with depth, while the corresponding value-space statistic is undefined for parts of these architectures under the current extraction protocol. We therefore treat these cells as diagnostic boundary cases and do not use them as evidence for or against the general rise–peak–fall claim.

Incomplete cross-benchmark coverage. Several additional architectures currently have diagnostics for only a subset of the three main benchmarks. We exclude these incomplete rows from the cross-benchmark five-diagnostic comparison rather than interpreting missing cells. Where both ProverQA and PrOntoQA raw geometric measurements are complete, their results are retained in Table 11.

Taken together, the extended analysis supports a heterogeneous but structured picture. Across many model–benchmark pairs, proof-relevant evidence separation strengthens at intermediate depth and subsequently degrades. Other settings exhibit formation failure, an early fall without a preceding rise, or architecture-specific diagnostic degeneracy. The common result is therefore not a universal fixed-shape curve, but a broader failure of monotonic preservation: proof-relevant structure is often formed non-monotonically and need not remain intact through the final layers.

Table 12: **Peak location and post-peak degradation across proof depths.** Each entry reports $l_{\text{peak}}(d)/\Delta_{\downarrow}(d)$. Peak locations are stable within approximately one or two layers in most model–benchmark cells, while several explicit exceptions occur. The magnitude of the post-peak decline is substantially more stable across depth.

ProverQA — depth (n): 3 (22), 4 (38), 6 (8), 7 (32), 8 (19)					
Backbone	depth 3	depth 4	depth 6	depth 7	depth 8
LLaMA-3.1-8B	L9 / 38%	L9 / 37%	L10 / 43%	L9 / 38%	L9 / 39%
Mistral-7B	L12 / 25%	L14 / 29%	L12 / 34%	L14 / 31%	L14 / 32%
Qwen3-8B	L10 / 56%	L10 / 54%	L10 / 56%	L10 / 56%	L21 / 54%
PrOntoQA — depth (n): 4 (50), 7 (50), 10 (50), 13 (50)					
Backbone	depth 4	depth 7	depth 10	depth 13	
LLaMA-3.1-8B	L8 / 32%	L0 / 33%	L9 / 32%	L9 / 38%	
Mistral-7B	L14 / 30%	L14 / 33%	L14 / 30%	L14 / 37%	
Qwen3-8B	L8 / 51%	L8 / 49%	L8 / 49%	L8 / 50%	
ProofWriter — depth (n): 0 (40), 1 (40), 2 (40), 3 (40), 4 (40)					
Backbone	depth 0	depth 1	depth 2	depth 3	depth 4
LLaMA-3.1-8B	L9 / 36%	L9 / 36%	L9 / 37%	L9 / 36%	L9 / 35%
Mistral-7B	L14 / 30%	L14 / 32%	L14 / 32%	L14 / 31%	L14 / 30%
Qwen3-8B	L10 / 48%	L20 / 50%	L20 / 49%	L20 / 49%	L20 / 48%
JustLogic — depth bin (n): 1–3 (87), 4–5 (58), 6–7 (54)					
Backbone	depth 1–3	depth 4–5	depth 6–7		
LLaMA-3.1-8B	L8 / 37%	L8 / 39%	L9 / 39%		
Mistral-7B	L18 / 28%	L18 / 30%	L18 / 30%		
Qwen3-8B	L20 / 58%	L20 / 59%	L20 / 59%		
LogicSkills — depth (n): 2 (100), 3 (100)					
Backbone	depth 2	depth 3			
LLaMA-3.1-8B	L7 / 34%	L7 / 33%			
Mistral-7B	L14 / 23%	L14 / 21%			
Qwen3-8B	L9 / 55%	L9 / 53%			

E Appendix: Full Method Description of FRGI

This appendix provides the complete inference procedure for FRGI (Fact–Rule Grouped Intervention). The method is training-free and operates entirely at inference time through lightweight forward hooks. It does not update model parameters, modify the input prompt, or alter the model architecture. Importantly, FRGI is proof-free at deployment: proof annotations, proof-used spans, and answer labels are never used for span selection, layer selection, intervention, or generation.

The deployed pipeline contains two prompt-prefill passes followed by ordinary autoregressive decoding. The first pass is diagnostic and contains no intervention. It determines *what* evidence should be targeted using query-conditioned attention saliency and *where* the intervention should occur using an Unlabeled Collapse Proxy (UCP). The second pass processes the same prompt with residual- and value-channel hooks that determine *how* the selected representations are modified. The resulting prompt KV cache is retained for generation, after which all hooks are removed.

E.1 Notation and Evidence Structure

For an input instance x, let

$$\mathcal{E}_x = \{e_1, \ldots, e_{M_x}\}$$

denote the evidence spans appearing in the prompt. A lightweight structural parser partitions these spans into fact and rule sets,

$$\mathcal{E}_x = \mathcal{F}_x \cup \mathcal{R}_x, \qquad \mathcal{F}_x \cap \mathcal{R}_x = \emptyset.$$

This partition identifies only the syntactic role of each span; it does not determine whether that span belongs to the annotated proof.

Let $\mathcal{Q}$ denote the query-token positions and $\mathrm{pos}(e_j)$ the token positions belonging to span e_j. At Transformer layer l, $\boldsymbol{h}_i^{(l)}$ denotes the hidden state of token i, and $A_{q,i}^{(l,h)}$ denotes the attention weight from query token q to token i at attention head h.

E.2 Pass 1: Proof-Free Diagnostic Prefill

The first pass processes the original prompt normally and does not decode an answer. No hooks are active during this pass. Instead, FRGI records the layerwise attention statistics and hidden states required for span selection, UCP computation, and healthy-state referencing. The KV cache produced by this pass is not used for final generation.

E.2.1 Query-Conditioned Span Selection

To determine *what* evidence should be modified, FRGI computes query-to-evidence saliency over a small diagnostic layer set $\mathcal{L}_s$. For evidence span e_j, we define

$$a_j = \frac{1}{|\mathcal{L}_s|H|\mathcal{Q}||\mathrm{pos}(e_j)|} \sum_{l\in\mathcal{L}_s} \sum_{h=1}^{H} \sum_{q\in\mathcal{Q}} \sum_{i\in\mathrm{pos}(e_j)} A_{q,i}^{(l,h)}. \tag{22}$$

Averaging over the token positions of each evidence span prevents longer spans from receiving larger scores merely because they contain more tokens.

Rather than using a fixed global threshold, selection is adapted to each input. Let μ_a and $\mathrm{std}(a)$ denote the mean and standard deviation of the span-saliency scores within the current instance. The selected evidence set is

$$\mathcal{S} = \{e_j : a_j \geq \mu_a + \sigma\,\mathrm{std}(a)\}, \tag{23}$$

with the selected fact and rule subsets

$$\mathcal{S}_{\mathrm{fact}} = \mathcal{S} \cap \mathcal{F}_x, \qquad \mathcal{S}_{\mathrm{rule}} = \mathcal{S} \cap \mathcal{R}_x. \tag{24}$$

Larger σ produces a sparser intervention set. As shown in Figure 5, performance remains stable over the evaluated range of threshold values.

Attention is used here only as a lightweight query-conditioned targeting signal. We do not interpret the resulting saliency scores as faithful explanations of the model's reasoning.

If no evidence span satisfies the adaptive threshold, $\mathcal{S} = \emptyset$, FRGI abstains from intervention and returns the vanilla model prediction.

E.2.2 Unlabeled Collapse Proxy

Span saliency identifies *what* to modify but does not determine *where* the modification should occur. We therefore construct a layerwise Unlabeled Collapse Proxy, $\mathrm{UCP}^{(l)}$, from the same no-hook diagnostic pass.

UCP combines three proof-free signals:

1. **Query-to-evidence attention entropy**, which measures how diffusely the query attends over the available evidence;
2. **Rule-state attenuation**, which measures the relative weakening of rule representations across depth;
3. **Unlabeled rule–fact coupling**, which measures the strength of interaction between structurally identified rule and fact representations without using annotated proof edges.

All three components are oriented so that larger UCP values correspond to a less structured layer: more diffuse evidence attention, stronger rule-state attenuation, or weaker rule–fact coupling. The proxy is computed using only the input prompt and first-pass activations; no proof labels enter its construction.

These signals capture complementary properties. As shown in Table 16, no single component matches the full proxy, and removing rule–fact coupling, rule-state attenuation, or attention entropy each reduces downstream accuracy.

E.2.3 Layer-Role Assignment

From the UCP profile, FRGI assigns four distinct layer roles: the collapse anchor l^*, the healthy reference layer l_g, the early cascade layer l_c, and a sparse set of post-collapse value-restoration layers $\mathcal{L}_{\text{tar}}$.

The collapse anchor is selected within the middle portion of the network:

$$l^* = \arg \max_{l \in [\lfloor 0.25L \rfloor, \lfloor 0.75L \rfloor]} \text{UCP}^{(l)}, \tag{25}$$

where L is the total number of Transformer layers.

The healthy reference layer is the lowest-UCP layer before the collapse anchor:

$$l_g = \arg \min_{l < l^*} \text{UCP}^{(l)}. \tag{26}$$

Importantly, l_g is a reference layer only; no intervention hook is applied there.

For the early cascade hook, we explicitly search the first 10–40% of the network for the strongest relative degradation in unlabeled rule–fact coupling:

$$l_c = \arg \max_{l \in [\lfloor 0.1L \rfloor, \lfloor 0.4L \rfloor]} \left(1 - \frac{C_{\text{rf}}^{(l)}}{\max_k C_{\text{rf}}^{(k)} + \epsilon} \right). \tag{27}$$

Thus, l_c identifies an early location where the local fact–rule interface is most weakened relative to its strongest layerwise value.

Finally, value restoration is applied sparsely beginning at the collapse anchor:

$$\mathcal{L}_{\text{tar}} = \{ l^*, l^* + 2, l^* + 4, \ldots, \min(l^* + 10, \lfloor 0.8L \rfloor) \}. \tag{28}$$

The two-layer spacing avoids modifying every downstream Transformer block while covering the region immediately following the detected collapse anchor.

If the diagnostic signal itself is weak,

$$\max_l \text{UCP}^{(l)} < \tau_{\text{collapse}}, \tag{29}$$

FRGI sets a `no_collapse` flag and falls back to the vanilla model. Therefore, the method does not force an intervention when its proof-free diagnostic provides insufficient evidence of a collapse region.

Table 15 evaluates this policy directly. UCP-guided selection substantially outperforms random, fixed, and non-UCP layer choices, while approaching the offline proof-annotated IIF oracle.

E.3 Pass 2: Fact–Rule Grouped Intervention

After Pass 1 determines

$$(\mathcal{S}_{\text{fact}}, \mathcal{S}_{\text{rule}}, l_g, l_c, l^*, \mathcal{L}_{\text{tar}}),$$

the same prompt is processed again from the beginning. During this second prefill, the selections above remain fixed and two localized interventions are applied.

E.3.1 Early Cascade Hook

At layer l_c, FRGI first computes the centroid of the selected fact representations:

$$\bar{\boldsymbol{h}}_{\text{fact}}^{(l_c)} = \frac{1}{|\mathcal{S}_{\text{fact}}|} \sum_{f \in \mathcal{S}_{\text{fact}}} \bar{\boldsymbol{h}}_{f}^{(l_c)}, \tag{30}$$

where $\bar{\boldsymbol{h}}_{f}^{(l_c)}$ is the mean hidden representation of the tokens belonging to fact span f.

For each token i belonging to a selected rule span $r \in \mathcal{S}_{\text{rule}}$, its hidden state is interpolated toward this selected-fact centroid:

$$\boldsymbol{h}_i^{(l_c)} \leftarrow (1-w)\boldsymbol{h}_i^{(l_c)} + w\bar{\boldsymbol{h}}_{\text{fact}}^{(l_c)}. \tag{31}$$

The update is residual rather than substitutive: the original rule representation is retained with weight $1-w$. The cascade hook is applied only when both $\mathcal{S}_{\text{fact}}$ and $\mathcal{S}_{\text{rule}}$ are nonempty.

The purpose of this intervention is limited to strengthening the local interaction between the selected fact and rule context before the diagnosed collapse region. It does not require knowledge of which specific fact satisfies which rule in the annotated proof and does not attempt to reconstruct the latent proof graph.

E.3.2 Healthy-State Reference

The healthy hidden representations at l_g provide the reference signal used by the later value-channel hooks. For selected evidence token i, denote the cached reference state by

$$\widehat{\boldsymbol{h}}_i^{(l_g)}. \tag{32}$$

A key implementation detail is that FRGI does not directly copy value vectors from l_g to later layers. Value spaces are layer-specific, so transferring a raw value vector across depth would ignore the target layer's own normalization and projection.

Instead, the healthy hidden state is re-projected through the target layer's value operator:

$$\widetilde{\boldsymbol{v}}_i^{(l)} = W_V^{(l)} \text{Norm}^{(l)}\left(\widehat{\boldsymbol{h}}_i^{(l_g)}\right), \qquad l \in \mathcal{L}_{\text{tar}}. \tag{33}$$

Thus, the reference originates from a healthier pre-collapse hidden state while remaining expressed in the native value space of the current target layer.

E.3.3 Adaptive Value-Channel Restoration

For each selected evidence position at target layer $l \in \mathcal{L}_{\text{tar}}$, FRGI interpolates the current value representation with the reference-derived value:

$$\boldsymbol{v}_{i,\text{FRGI}}^{(l)} = (1-\beta_l)\boldsymbol{v}_i^{(l)} + \beta_l \widetilde{\boldsymbol{v}}_i^{(l)}. \tag{34}$$

The interpolation strength is adapted separately for each target layer from the no-hook diagnostic states. Let

$$\mathcal{G} = \begin{cases} \mathcal{S}_{\text{rule}}, & \mathcal{S}_{\text{rule}} \neq \emptyset, \\ \mathcal{S}, & \text{otherwise}, \end{cases} \tag{35}$$

and let $\widehat{\bar{\boldsymbol{h}}}_{\mathcal{G}}^{(l)}$ denote the mean no-hook hidden representation of the corresponding group at layer l. We compute

$$\rho_l = \frac{\left\| \widehat{\bar{\boldsymbol{h}}}_{\mathcal{G}}^{(l)} \right\|_2}{\left\| \widehat{\bar{\boldsymbol{h}}}_{\mathcal{G}}^{(l_g)} \right\|_2 + \epsilon}. \tag{36}$$

The intervention strength is then

$$\beta_l = \beta_{\max} \cdot \text{clip}\left(\frac{\tau_\rho - \rho_l}{\tau_\rho}, 0, 1 \right). \tag{37}$$

This schedule makes the restoration conditional on the observed attenuation. When ρ_l remains above the attenuation threshold τ_ρ, $\beta_l = 0$ and no value correction is applied. As the current representation becomes increasingly attenuated relative to the healthy reference, β_l increases smoothly up to $\beta_{\max}$.

The fixed-β ablation in Table 14 confirms that this layer-adaptive strength contributes beyond using a constant interpolation coefficient.

E.4 Prompt KV-Cache Construction and Decoding

Both the cascade hook and value-restoration hooks are active only during the second prompt prefill. Non-selected evidence positions and all other Transformer computations remain unchanged.

At each target layer, the modified selected-token value representations are used while constructing the prompt KV cache. After the second prefill is complete, all hooks are removed. The model then performs standard autoregressive decoding from the resulting cache:

$$p\left(y_t \mid x, y_{<t}; \text{KV}_{\text{FRGI}}\right). \tag{38}$$

No hooks are applied to subsequently generated tokens, and the prompt is not processed a third time. Hence, FRGI modifies how selected prompt evidence is represented and stored during prefill while leaving the autoregressive decoding procedure itself unchanged.

E.5 Abstention and Fallback Behavior

FRGI deliberately retains a vanilla fallback for instances in which its proof-free diagnostic is insufficient. An intervention is skipped when either

$$\mathcal{S} = \emptyset \tag{39}$$

or

$$\max_l \text{UCP}^{(l)} < \tau_{\text{collapse}}. \tag{40}$$

In either case, the final prediction is produced by standard model inference. This design separates failure detection from intervention: FRGI modifies internal representations only when both a candidate evidence target and a sufficiently strong collapse signal are present.

F Appendix: Baseline Implementation Details

Unless otherwise stated, all baselines are evaluated with an effective batch size of 1. This includes prompting baselines, activation-steering baselines, and structured inference baselines. We use batch size 1 to avoid batching artifacts from variable-length reasoning traces, dynamic activation hooks, and multi-stage symbolic prompting. All methods use the same dataset splits, model checkpoints, answer parser, and maximum generation length of 2048 tokens.

F.1 Prompting Baselines

We implement three prompting baselines in a unified evaluation script: Zero-shot Chain-of-Thought (Zero-shot CoT) (Kojima et al., 2022), Few-shot Chain-of-Thought (Few-shot CoT) (Wei et al., 2022), and Self-Consistency (SC) (Wang et al., 2023). These methods are evaluated on ProverQA, PrOntoQA, and ProofWriter using the same base prompting pipeline, with model-specific chat formatting applied for LLaMA-3.1-8B, Qwen3-8B, and Mistral-7B.

For all prompting baselines, the input prompt contains a system message instructing the model to solve a logical reasoning problem and to output the final decision as a single answer letter. The user prompt contains the context, question, and answer options. We explicitly constrain the output format by appending either "Respond with: Final answer: [A/B/C]" for ProverQA or "Respond with: Final answer: [A/B]" for PrOntoQA and ProofWriter. The answer parser first searches for the pattern "Final answer: X" before falling back to looser parsing heuristics.

Zero-shot CoT. Zero-shot CoT (Kojima et al., 2022) is implemented by appending the reasoning trigger "Let's think step by step." after the context, question, and answer options. No exemplars are provided. The prompt therefore asks the model to generate an intermediate reasoning trace before producing the final answer letter.

We evaluate Zero-shot CoT under greedy decoding (`num_beams` $= 1$, `do_sample` $=$ `False`) and beam search (`num_beams` $= 4$, `do_sample` $=$ `False`). All runs use effective batch size 1.

Few-shot CoT. Few-shot CoT (Wei et al., 2022) is implemented as a fixed 2-shot prompting setup. The exemplars are dataset-specific and manually specified. Each exemplar consists of a context, a question, a full chain-of-thought explanation, and a final answer. These two exemplars are prepended to every test prompt for the corresponding dataset.

The exemplars are not sampled from the training set at runtime, are not selected dynamically, and are not stratified by difficulty. After the two exemplars, the actual test instance is appended in the same format as Zero-shot CoT, followed by the reasoning trigger and the final-answer formatting instruction. Few-shot CoT uses the same greedy and beam-search settings as Zero-shot CoT, with effective batch size 1.

Self-Consistency. Self-Consistency (Wang et al., 2023) reuses the Few-shot CoT prompt, including the same two fixed exemplars and reasoning trigger. For each test example, we sample $k = 5$ independent reasoning trajectories using stochastic decoding with temperature $T = 0.7$ and top-$p = 0.95$. The script sets `do_sample` $=$ `True` and `num_beams` $= 1$ for all self-consistency generations. Each sampled output is parsed into an answer letter, and the final prediction is determined by majority vote over the five parsed answers. If none of the five generations yields a parsable answer, the prediction is recorded as missing. Self-consistency is evaluated one instance at a time, i.e., with effective batch size 1.

F.2 Activation-Steering Baselines

We implement activation-steering baselines as inference-time interventions on the residual stream of decoder-only language models. We compare Activation Addition (ActAdd) (Turner et al., 2025) and SADI (Wang et al., 2025). The code also supports Contrastive Activation Addition (CAA) (Rimsky et al., 2024), but CAA is not included in the default experiment grid. All activation-steering baselines use the same task prompt as the Zero-shot CoT baseline rather than the Few-shot CoT prompt. Thus, these baselines should be interpreted as zero-shot prompting plus an internal representation-level intervention. All activation-steering runs use effective batch size 1.

For all steering methods, we use the same maximum generation budget as the other baselines, with `max_new_tokens`=2048. Greedy decoding uses `num_beams = 1`, and beam search uses `num_beams = 4`.

ActAdd. ActAdd (Turner et al., 2025) is implemented using a single generic contrastive prompt pair rather than a calibration set of task-specific examples. The positive prompt is:

> "Carefully apply systematic logical deduction step by step using every given premise to reach the correct conclusion:"

The negative prompt is:

> "Just randomly guess the answer without thinking:"

For a chosen transformer layer ℓ, the script performs one forward pass on each prompt, extracts the last-token residual stream representation immediately before layer ℓ, and computes the steering vector as the difference between the positive and negative residual representations:

$$\mathbf{s}_{\text{ActAdd}} = \frac{\mathbf{h}_\ell^+ - \mathbf{h}_\ell^-}{\|\mathbf{h}_\ell^+ - \mathbf{h}_\ell^-\|_2}.$$

We use a fixed coefficient $\alpha = 6.0$. The default injection layer is the middle transformer layer, computed as `num_hidden_layers // 2`, unless explicitly overridden.

The implementation uses `front` injection mode by default. In this mode, the intervention is applied only during the prefill pass, when the full prompt is processed. It is not added during one-token decode steps:

$$\mathbf{h}_\ell^{(t)} \leftarrow \mathbf{h}_\ell^{(t)} + \alpha \mathbf{s}_{\text{ActAdd}}.$$

SADI. SADI (Wang et al., 2025) is implemented as a dynamic steering method built on a calibration split. For each calibration example, the script constructs a correct-answer and wrong-answer completion, extracts their residual difference vector at the chosen layer, and also extracts a prompt-only residual representation to serve as a query embedding. The prompt-only residuals are clustered with K-means, and each cluster receives its own average steering vector. At test time, the prompt-only residual of the test instance is matched to the nearest centroid, and the corresponding cluster-specific steering vector is applied.

We use calibration size 200, number of clusters $K = 5$, coefficient $\alpha = 6.0$, and `front` injection mode. Formally, if $\mathbf{c}(x)$ denotes the prompt-only residual embedding of input x, K-means produces centroids $\{\boldsymbol{\mu}_k\}_{k=1}^K$ and cluster-specific steering vectors $\{\mathbf{s}_k\}_{k=1}^K$. For a test input x, the assigned cluster is

$$k^* = \arg\max_k \cos(\mathbf{c}(x), \boldsymbol{\mu}_k),$$

and the intervention uses $\mathbf{s}_{k^*}$ as the steering direction. Because the steering vector is selected separately for each input instance, SADI is evaluated with effective batch size 1.

F.3 Structured Inference Baselines

We include two structured inference baselines: Minimal Test-Time Intervention (MTI) (Yang et al., 2026) and Symbolic Chain-of-Thought (SymbCoT) (Xu et al., 2024). Both methods keep model parameters frozen, but they differ from FRGI: MTI operates through decoding-time guidance, while SymbCoT uses a multi-stage symbolic prompting pipeline. We evaluate both methods with effective batch size 1.

MTI. MTI (Yang et al., 2026) is implemented as a classifier-free guidance (CFG) decoding strategy applied entirely at inference time, without parameter updates. At each generation step, the model performs two forward passes: a conditioned pass using the full input prompt, and an unconditioned pass using a stripped prompt that omits the reasoning context. The final next-token logits are computed as a weighted combination of the two distributions, with the guidance scale controlling how strongly the prediction is steered toward the context-conditioned trajectory.

Our implementation builds on the `UnbatchedClassifierFreeGuidanceLogitsProcessor` from the Transformers library. For fairness and consistency with the other baselines, MTI is evaluated

Table 13: Additional beam-search comparison with representative reasoning and intervention baselines. We report accuracy (%) with beam size k=4 on ProverQA, PrOntoQA, and ProofWriter.

Model	Method	ProverQA				PrOntoQA					ProofWriter
		Overall	Easy	Medium	Hard	Overall	1-hop	2-hop	3-hop	4-hop	Overall
LLaMA-3.1-8B	Vanilla	37.40	48.10	36.90	27.20	42.10	47.20	42.80	40.10	37.60	53.10
	Zero-shot CoT	44.60	55.20	43.80	34.60	57.10	63.90	59.40	54.10	50.80	64.40
	Few-shot CoT	51.67	66.10	49.20	39.40	68.23	75.30	70.10	65.70	62.00	67.50
	Self-Consistency	60.13	74.40	60.50	45.80	77.87	83.20	79.40	76.00	73.10	76.20
	ActAdd	42.20	56.00	41.20	29.10	58.80	64.60	60.20	56.10	54.20	57.60
	SADI	39.50	53.40	38.10	27.20	52.60	58.20	54.00	49.90	48.30	53.50
	SymbCoT	44.50	57.20	43.40	32.80	73.10	79.40	74.60	70.50	67.80	62.20
	MTI	46.20	59.00	45.40	34.10	74.60	80.60	76.10	72.00	69.60	64.30
	FRGI	**64.10**	**79.50**	**63.30**	**49.50**	**83.40**	**88.20**	**84.60**	**81.10**	**79.50**	**79.05**
Qwen3-8B	Vanilla	38.70	50.20	38.50	27.30	43.50	48.60	44.30	41.10	39.80	49.60
	Zero-shot CoT	45.50	57.50	44.60	34.50	59.10	65.40	61.10	56.30	53.60	53.70
	Few-shot CoT	52.30	68.10	50.40	38.20	71.80	78.50	73.40	69.70	65.50	74.30
	Self-Consistency	61.80	77.20	61.40	46.90	80.20	85.60	81.70	78.20	75.30	78.40
	ActAdd	44.40	57.80	43.20	32.10	61.20	67.00	62.60	58.30	56.90	54.10
	SADI	41.50	55.60	40.30	28.50	54.60	60.20	56.00	51.90	50.30	53.50
	SymbCoT	45.40	58.10	44.30	33.90	75.10	81.20	76.70	72.40	70.10	63.40
	MTI	47.60	61.40	46.60	34.80	76.50	82.60	78.00	73.80	71.50	65.60
	FRGI	**65.60**	**81.60**	**65.10**	**50.20**	**85.10**	**89.40**	**86.20**	**83.00**	**81.50**	**81.40**
Mistral-7B	Vanilla	35.60	45.10	35.20	26.50	43.20	48.60	44.00	40.90	39.20	31.10
	Zero-shot CoT	45.50	56.30	45.00	35.40	56.00	62.30	57.60	52.90	51.10	33.60
	Few-shot CoT	50.40	64.70	48.50	38.00	66.50	73.10	68.40	63.70	60.70	37.40
	Self-Consistency	57.30	71.10	57.40	43.40	75.50	81.00	77.10	72.80	71.20	42.70
	ActAdd	43.10	55.20	42.00	32.10	60.20	66.40	61.60	57.10	55.80	34.80
	SADI	39.20	51.60	38.40	27.80	52.00	57.70	53.20	49.10	47.90	48.20
	SymbCoT	41.50	54.20	40.30	30.00	65.30	71.10	67.00	62.30	61.10	53.90
	MTI	45.10	57.90	44.30	33.20	67.10	73.30	68.90	64.10	62.10	46.70
	FRGI	**59.50**	**74.60**	**59.30**	**44.40**	**80.10**	**85.10**	**81.50**	**78.30**	**75.50**	**56.10**

one instance at a time with effective batch size 1. Unless otherwise stated, MTI uses greedy decoding (`num_beams` = 1, `do_sample` = `False`) and the same maximum generation length as other baselines.

SymbCoT. SymbCoT (Xu et al., 2024) is implemented as a three-stage inference pipeline following the original symbolic reasoning formulation. In the first stage, the model translates the natural-language context and question into a symbolic representation, such as first-order logic statements or structured rules. In the second stage, the model generates an explicit reasoning plan over the symbolic context. In the third stage, the model executes the plan and produces the final answer.

All three stages use the same frozen language model loaded locally via HuggingFace Transformers. No model parameters are updated. We use dataset-specific prompt templates for ProverQA, PrOntoQA, and ProofWriter, matching the answer format of each dataset. Each stage is run with effective batch size 1. Because SymbCoT requires multiple LLM calls per example, it is more computationally expensive than single-pass prompting baselines, but it provides a strong task-specific symbolic-reasoning comparison.

G Appendix: Extra Results for Analysis

G.1 Additional Beam-Search Results

We additionally evaluate all methods under beam search with beam size k=4. This setting tests whether the gains observed under greedy decoding persist when decoding is allowed to explore multiple candidate continuations. Table 13 reports accuracy on ProverQA, PrOntoQA, and ProofWriter.

Table 13 shows that the main trends observed under greedy decoding persist under beam search. Across the three reported benchmarks and all three model families, FRGI achieves the strongest overall accuracy in this beam-search setting. On ProverQA, FRGI improves over the strongest non-FRGI baseline by $+3.97$ points for LLaMA-3.1-8B, $+3.80$ points for Qwen3-8B, and $+2.20$

points for Mistral-7B. On ProofWriter, the corresponding margins are $+2.85$, $+3.00$, and $+2.20$ points.

Beam search strengthens several prompting baselines, especially Few-shot CoT and Self-Consistency. However, their gains are less uniform across benchmarks. Self-Consistency is competitive on ProverQA and ProofWriter, while MTI is relatively competitive on PrOntoQA but remains weaker on ProverQA and ProofWriter. Activation-intervention baselines such as ActAdd and SADI also show substantial model- and dataset-level variation. In contrast, FRGI consistently maintains top performance across all three datasets under the same beam-search setting, confirming that its improvements are robust and not an artifact of greedy decoding.

The largest margin over the strongest non-FRGI baseline appears on PrOntoQA for LLaMA-3.1-8B, where FRGI reaches 83.40% compared with 77.87% for the strongest competing baseline ($+5.53$ points). Overall, these beam-search results support the same qualitative conclusion as the greedy-decoding experiments: FRGI provides consistent inference-time gains across reasoning benchmarks and decoding strategies.

G.2 Ablation Studies

We conduct two complementary ablation analyses. The first is a component ablation, which removes or replaces deployed components of FRGI. This analysis tests whether the method depends on the intended design choices: saliency-based span selection, the residual-level cascade hook, value-channel restoration, adaptive intervention strength, and UCP-guided layer selection. The second is a hook-level control ablation, which tests whether the gains can be explained by trivial alternatives such as hook registration, arbitrary perturbations, intervention at arbitrary positions, or simple suppression without reference-based restoration.

Unless otherwise stated, all ablations use the same UCP-selected layers and the same proof-label-free saliency-selected spans as the full FRGI method. We report the full component ablation under greedy decoding, where the effect of each intervention on the model's top-1 computation is most direct. Beam-search robustness of the full method is reported in Appendix G.1. We additionally report hook-level controls under beam search (k=4) to verify that the same intervention logic remains valid when decoding explores multiple candidate continuations.

G.2.1 Component Ablation

We provide a fine-grained breakdown of the ablations reported in the main text, separating the effects of span targeting, intervention mechanism, and layer selection across individual backbones and benchmarks. Overall, the trends are consistent across model families: FRGI benefits from jointly identifying *what* evidence to modify, *where* to intervene, and *how* to restore the affected representations.

Span targeting and intervention components. Table 14 shows that each component contributes consistently across the nine model–benchmark pairs. Replacing saliency-selected spans with random spans reduces the overall accuracy from 71.85% to 53.55%, indicating that effective repair depends critically on targeting the appropriate evidence. Among the intervention mechanisms, removing value-channel restoration causes the largest degradation, reducing the overall accuracy to 61.88%, whereas removing the early cascade hook gives 68.63%. Using a fixed intervention strength also underperforms the adaptive schedule (69.81% vs. 71.85%). Importantly, these effects are not driven by a single backbone or dataset: the full method achieves the highest accuracy in every individual cell. This supports the intended factorization of FRGI, where saliency determines *what* to modify and the cascade, value, and adaptive-strength components determine *how* the selected representations are repaired.

Proof-free layer selection. Table 15 isolates the *where* component by holding the remaining FRGI pipeline fixed and changing only the layer-selection policy. UCP achieves 71.85% overall accuracy, substantially exceeding random (58.64%), fixed (62.74%), and non-UCP selection (64.78%). The advantage also holds across all three backbones and all three reasoning benchmarks. Despite using no proof annotations, UCP remains only 2.19 points below the proof-annotated IIF oracle (74.04%), recovering approximately 81% of the improvement from fixed layer selection to the oracle. This

Table 14: **Full fine-grained FRGI component analysis across individual datasets and backbones.** We ablate the span-targeting and intervention components while keeping UCP-based layer selection fixed. Results report greedy-decoding accuracy (%) on ProverQA (PR), PrOntoQA (PO), and ProofWriter (PW).

Variant	LLaMA-3.1-8B			Qwen3-8B			Mistral-7B			Overall
	PR	PO	PW	PR	PO	PW	PR	PO	PW	
Vanilla	36.84	41.28	52.46	37.92	42.64	48.83	34.83	42.41	30.34	40.84
Random spans	50.43	61.24	62.14	48.64	58.43	59.54	46.84	55.47	39.26	53.56
w/o cascade	60.52	78.64	75.43	61.91	80.43	77.84	55.49	75.18	52.27	68.63
w/o value	55.21	70.53	68.94	56.43	71.84	70.13	51.29	66.87	45.67	61.88
Fixed β	61.38	80.14	76.23	63.14	81.84	78.63	56.87	76.57	53.47	69.81
Full FRGI	**63.24**	**82.54**	**78.14**	**64.84**	**84.18**	**80.53**	**58.67**	**79.24**	**55.27**	**71.85**

Table 15: **Full proof-free layer selection across datasets and backbones.** Mean greedy-decoding accuracy (%) broken down by individual reasoning benchmarks and model backbones. Oracle uses proof-annotated IIF and is unavailable at inference time.

Selector	LLaMA-3.1-8B			Qwen3-8B			Mistral-7B			Overall
	PR	PO	PW	PR	PO	PW	PR	PO	PW	
Random	58.43	63.12	62.41	59.12	64.83	64.28	48.34	54.21	53.08	58.64
Fixed	62.14	66.83	67.26	63.42	68.14	68.93	52.14	58.63	57.14	62.74
Non-UCP	57.81	73.43	71.84	59.21	74.53	73.24	53.64	70.47	48.87	64.75
UCP	**63.24**	**82.54**	**78.14**	**64.84**	**84.18**	**80.53**	**58.67**	**79.24**	**55.27**	**71.80**
Oracle IIF	65.42	85.14	79.93	67.14	86.83	81.64	60.24	81.53	58.51	74.04

suggests that the unlabeled signals used by UCP capture much of the layer information exposed by the offline proof-conditioned IIF diagnostic.

Which UCP signals matter? Table 16 further decomposes UCP into its three constituent signals. No individual signal is sufficient: rule–fact coupling is the strongest single-signal selector at 58.47%, followed by rule-state attenuation at 57.23% and attention entropy at 54.83%, all substantially below the full UCP at 71.85%. The leave-one-out results show the same ordering more clearly. Removing rule–fact coupling produces the largest reduction (-6.36 points), followed by rule-state attenuation (-4.66) and attention entropy (-3.12). Thus, rule–fact coupling contributes the strongest individual signal, but the best performance requires all three components, indicating that attention diffusion, representation attenuation, and weakening fact–rule interaction provide complementary evidence for locating the intervention region.

G.2.2 Hook-Level Control Ablation

We next run hook-level controls to rule out trivial explanations for the gains. These controls test whether improvements can be attributed to hook registration alone, arbitrary perturbations, intervention at arbitrary positions, or simple suppression without reference-based restoration.

No-hook (NOHOOK). The hook is registered but returns the value vectors unchanged. This checks whether hook registration itself introduces side effects.

Random reference (RNDREF). The reference-derived value is replaced by a random vector with matched norm before interpolation. This tests whether the content of the reference signal matters.

Wrong position (WRONGPOS). The real reference-based intervention is applied to matched unused distractor positions instead of selected fact/rule spans. This tests whether reasoning-aware target positions are necessary.

Zero suppression (ZEROSUP). Selected value vectors are replaced by zero. This tests whether suppressing selected value-channel content is sufficient without reference-based restoration.

Table 16: **Full UCP component analysis across benchmarks and variants.** We evaluate single-signal and leave-one-out UCP variants. Results report greedy-decoding accuracy (%) across ProverQA, PrOntoQA, and ProofWriter (averaged across backbones). Δ denotes the overall change relative to the full UCP.

UCP variant	ProverQA	PrOntoQA	ProofWriter	Avg.	Δ
Vanilla baseline	36.53	42.11	43.88	40.84	−31.01
Single-signal selectors					
Attention entropy only	47.31	63.84	53.34	54.83	−17.02
Rule-state attenuation only	49.42	66.71	55.56	57.23	−14.62
Rule–fact coupling only	50.64	68.12	56.65	58.47	−13.38
Leave-one-out					
w/o attention entropy	59.64	78.43	68.12	68.73	−3.12
w/o rule-state attenuation	58.12	77.14	66.31	67.19	−4.66
w/o rule–fact coupling	56.43	75.42	64.62	65.49	−6.36
Full UCP	**62.25**	**81.99**	**71.31**	**71.85**	–

Table 17: Hook-level control ablation of FRGI under beam search (k=4). We report overall accuracy (%) on ProverQA, PrOntoQA, and ProofWriter. Avg. denotes the mean accuracy across the three datasets.

Model	Method	Reference signal	Target positions	ProverQA	PrOntoQA	ProofWriter	Avg.
LLaMA-3.1-8B	Vanilla	–	–	37.40	42.10	53.10	44.20
	NOHOOK	none	selected	37.35	42.05	53.00	44.13
	RNDREF	random	selected	45.20	54.60	58.40	52.73
	WRONGPOS	real	unused	53.80	68.50	67.20	63.17
	ZEROSUP	zero	selected	59.60	77.20	73.80	70.20
	FRGI	**real**	**selected**	**64.10**	**83.40**	**79.05**	**75.52**
Qwen3-8B	Vanilla	–	–	38.70	43.50	49.60	43.93
	NOHOOK	none	selected	38.65	43.40	49.50	43.85
	RNDREF	random	selected	46.40	56.20	56.80	53.13
	WRONGPOS	real	unused	55.20	70.40	69.50	65.03
	ZEROSUP	zero	selected	61.20	79.50	76.20	72.30
	FRGI	**real**	**selected**	**65.60**	**85.10**	**81.40**	**77.37**
Mistral-7B	Vanilla	–	–	35.60	43.20	31.10	36.63
	NOHOOK	none	selected	35.50	43.15	31.00	36.55
	RNDREF	random	selected	42.80	53.80	37.50	44.70
	WRONGPOS	real	unused	49.60	65.40	43.80	52.93
	ZEROSUP	zero	selected	55.40	74.20	50.60	60.07
	FRGI	**real**	**selected**	**59.50**	**80.10**	**56.10**	**65.23**

Together, these controls separate the effects of hook registration, reference content, target-position specificity, and the presence of meaningful value-channel restoration.

Table 17 reveals a consistent hierarchy across all three models. Performance with NOHOOK matches the vanilla baseline almost exactly, confirming that the hook registration mechanism itself does not introduce artificial artifacts. Injecting random representations (RNDREF) or intervening at unselected positions (WRONGPOS) substantially underperforms full FRGI, highlighting the necessity of both the semantics of the reference signal and precise position targeting. While zeroing out target activations (ZEROSUP) recovers a notable portion of the performance gain (supporting the hypothesis that toxic representations hinder reasoning), FRGI achieves the highest accuracy overall. This demonstrates that suppressing unfaithful representations must be paired with genuine reference-guided value restoration for optimal reasoning performance.

G.3 Efficiency Analysis

We report per-sample inference latency for all methods in the main comparison. Latency is measured end-to-end and includes prompt processing, method-specific intervention or sampling overhead, and answer generation. For FRGI, latency includes the no-hook diagnostic prefill, the hooked prefill, and

Table 18: Inference time per sample under greedy decoding. We report seconds/sample on PrOntoQA, ProofWriter, and ProverQA. Avg. denotes the mean latency across the three benchmarks.

Model	Method	PrOntoQA	ProofWriter	ProverQA	Avg.
LLaMA-3.1-8B-Instruct	Vanilla	0.46	0.47	0.51	0.48
	Zero-shot CoT	7.23	7.12	5.27	6.54
	Few-shot CoT	3.71	8.38	5.46	5.85
	Self-Consistency	36.77	39.62	30.62	35.67
	ActAdd	7.31	8.09	7.76	7.72
	SADI	20.92	15.86	19.35	18.71
	SymbCoT	23.66	20.08	21.87	21.87
	MTI	5.03	5.27	5.04	5.11
	FRGI	**0.63**	**0.71**	**0.64**	**0.66**
Mistral-7B-Instruct-v0.2	Vanilla	0.96	2.69	1.57	1.74
	Zero-shot CoT	2.21	2.21	2.96	2.46
	Few-shot CoT	3.22	3.86	5.01	4.03
	Self-Consistency	16.34	22.79	28.22	22.45
	ActAdd	7.71	6.94	5.54	6.73
	SADI	19.86	14.64	8.22	14.24
	SymbCoT	20.83	20.84	20.31	20.66
	MTI	4.27	4.27	2.08	3.54
	FRGI	**1.19**	**2.94**	**1.84**	**1.99**
Qwen3-8B	Vanilla	1.49	2.27	1.49	1.75
	Zero-shot CoT	9.35	9.12	8.26	8.91
	Few-shot CoT	7.41	8.02	10.13	8.52
	Self-Consistency	35.41	35.82	40.61	37.28
	ActAdd	10.43	2.63	9.47	7.51
	SADI	28.65	18.41	29.32	25.46
	SymbCoT	19.72	20.26	19.99	19.99
	MTI	5.29	5.27	5.31	5.29
	FRGI	**1.92**	**2.82**	**2.07**	**2.27**

answer generation from the hooked KV cache. There is no separate proof-labeled layer-calibration cost in the deployed pipeline; UCP and span saliency are computed during the no-hook diagnostic prefill and are included in the reported FRGI latency.

Tables 18 and 19 detail the per-sample inference latency under greedy and beam-search decoding, respectively. As expected, Self-Consistency incurs the highest computational cost due to its reliance on multiple sampled generation paths. Similarly, frameworks such as SADI and SymbCoT introduce significant time overhead compared to standard chain-of-thought decoding. In contrast, FRGI is structurally efficient, requiring only a single diagnostic prefill and a hooked intervention pass prior to generation. Consequently, its latency remains strictly competitive with the Vanilla baseline and is fundamentally lower than both multi-sample and iterative symbolic approaches. Minor instances where FRGI executes marginally faster than Vanilla are primarily artifacts of measurement variance and inherently shorter generated output sequences; thus, the macro-average latency provides the most reliable point of comparison. Ultimately, these benchmarks confirm that FRGI delivers significant reasoning improvements without the prohibitive decoding bottlenecks typical of advanced inference methods.

G.4 Generation Budget and Max-Token Stability

Multi-step reasoning methods rely on intermediate generation, making the maximum token budget a potential confounding variable: premature truncation directly degrades measured accuracy. We mitigate this by fixing `max_new_tokens` $= 2048$ uniformly for all baselines and FRGI. Decoding terminates upon reaching an EOS token or emitting a valid label from

Table 19: Inference time per sample under beam-search decoding ($k = 4$). We report seconds/sample on PrOntoQA, ProofWriter, and ProverQA. Avg. denotes the mean latency across the three benchmarks.

Model	Method	PrOntoQA	ProofWriter	ProverQA	Avg.
LLaMA-3.1-8B-Instruct	Vanilla	0.46	0.47	0.51	0.48
	Zero-shot CoT	11.67	8.58	12.57	10.94
	Few-shot CoT	11.95	14.45	15.12	13.84
	Self-Consistency	36.07	39.66	40.94	38.89
	ActAdd	13.25	9.96	12.76	11.99
	SADI	19.98	14.59	24.62	19.73
	SymbCoT	23.66	20.08	21.87	21.87
	MTI	5.03	5.27	5.03	5.11
	FRGI	**0.44**	**0.57**	**0.52**	**0.51**
Mistral-7B-Instruct-v0.2	Vanilla	0.96	2.69	1.57	1.74
	Zero-shot CoT	4.11	4.34	5.26	4.57
	Few-shot CoT	7.73	10.21	8.73	8.89
	Self-Consistency	15.93	21.62	35.26	24.27
	ActAdd	11.39	10.92	10.78	11.03
	SADI	20.54	16.03	13.71	16.76
	SymbCoT	20.83	20.84	20.31	20.66
	MTI	4.27	4.27	2.08	3.54
	FRGI	**2.77**	**2.39**	**1.56**	**2.24**
Qwen3-8B	Vanilla	1.49	2.27	1.49	1.75
	Zero-shot CoT	14.17	13.88	15.78	14.61
	Few-shot CoT	13.02	14.46	14.97	14.15
	Self-Consistency	35.13	36.33	50.82	40.76
	ActAdd	15.53	12.33	16.03	14.63
	SADI	27.58	19.72	34.81	27.37
	SymbCoT	19.72	20.26	19.99	19.99
	MTI	5.31	5.28	5.34	5.31
	FRGI	**2.62**	**1.74**	**2.51**	**2.29**

$\mathcal{Y} = \{\text{TRUE}, \text{FALSE}, \text{UNCERTAIN}\}$, while outputs exceeding the limit without a parsable answer are marked incorrect. Because FRGI intervenes during prompt prefill rather than extending generation, its empirical advantages cannot be attributed to longer reasoning chains.To confirm that our conclusions are robust to the chosen cutoff, we evaluate LLaMA-3.1-8B across constrained budgets: `max_new_tokens` $\in \{256, 512, 1024, 2048\}$ (Table 20). The comparative advantages remain consistent across all settings. Although CoT-style methods show minor gains as the generation ceiling rises, performance largely plateaus by 1024 tokens. Across all tested intervals, FRGI maintains its superior accuracy, retaining a pronounced advantage even under the strictest 256-token limit.

As reported in Table 21, invalid and truncated outputs drop near zero across all methods by 2048 tokens. Notably, FRGI incurs practically negligible truncation even under tighter budgets (0.4% at 256 tokens), demonstrating that its proofs remain compact. This confirms that our main benchmark evaluations reflect genuine logical reasoning quality rather than generation budget bottlenecks.

G.5 Case studies under label scoring

Figures 16–19 show four ProverQA development instances in which the proof-free FRGI intervention changes an incorrect label ranking to the gold label under label scoring, with two examples from LLaMA-3.1-8B and two from Mistral-7B. In all four cases, saliency-based selection is imperfect and includes both proof-used and unused premises, yet the localized intervention is sufficient to recover the correct prediction. FRGI does not use the gold proof or answer; it only modifies selected internal representations during prompt prefill. Each vanilla run also exhibits a post-peak decline in

Table 20: Max-token stability check on LLaMA-3.1-8B under greedy decoding. We report accuracy (%) on ProverQA. These empirical results demonstrate that FRGI's advantage is stable across varying generation budgets.

Method	256 tokens	512 tokens	1024 tokens	2048 tokens
Vanilla	36.60	36.80	36.87	36.87
Zero-shot CoT	43.20	44.10	44.60	44.80
Few-shot CoT	48.60	50.50	51.50	51.80
Self-Consistency	56.80	58.40	59.20	59.47
ActAdd	40.80	41.10	41.30	41.40
SADI	37.90	38.40	38.70	38.85
SymbCoT	40.80	42.50	43.40	43.67
MTI	44.80	45.10	45.30	45.40
FRGI	**62.10**	**62.80**	**63.10**	**63.24**

Table 21: Invalid-or-truncated output rate (%) on LLaMA-3.1-8B ProverQA. An output is defined as invalid if no parsable answer is generated before the token limit.

Method	256	512	1024	2048
Vanilla	0.2	0.1	0.0	0.0
Zero-shot CoT	2.8	0.9	0.2	0.1
Few-shot CoT	4.6	1.3	0.3	0.1
Self-Consistency	3.9	1.1	0.3	0.1
SymbCoT	6.2	2.0	0.6	0.2
FRGI	0.4	0.2	0.1	0.0

the raw hidden-state proof-used–versus–distractor evidence-separation statistic δ_l, consistent with the representation-level degradation analyzed above.

Problem setup and logical derivation

Facts. F1. Lenora does not conduct clinical trials. **F2.** Lenora develops a new treatment. **F3.** Lenora researches pain management. **F4.** Lenora collaborates with experts.
Rules. R1.† If Ailani develops a new treatment, then she either earns medical recognition or publishes a medical breakthrough, but not both. **R2.** If Lenora develops a new treatment, then she either earns medical recognition or publishes a medical breakthrough, but not both. **R3.** If Lenora researches pain management, then she can either improve drug formulations or earn medical recognition, but not both. **R4.**† If Lenora authors influential papers, then she will publish a medical breakthrough. **R5.** If Lenora conducts clinical trials or collaborates with experts, then she can publish a medical breakthrough. **R6.**† If Ailani conducts clinical trials or collaborates with experts, then she can publish a medical breakthrough.
Question. Is the statement "Lenora improves drug formulations." true, false, or uncertain?

Annotated proof chain

From **F1**, **F4**, applying **R5**: Lenora publishes a medical breakthrough. From **F2**, **Lenora publishes a medical breakthrough.**, applying **R2**: Lenora does not earn medical recognition. From **F3**, **Lenora does not earn medical recognition.**, applying **R3**: Lenora improves drug formulations.

Correct logical outcome

Correct label: A) True. 7 proof-used premises, 3 distractors (†).

Proof-free FRGI intervention (LLaMA-3.1-8B, label scoring)

Proof-free selection

Saliency-selected spans (layers 5–8, σ=0.25): R6†, F1, R1† (1 of 3 selected spans is proof-used; † = distractor). UCP layers: l^*=8, l_c=3, targets [8, 10, 12, 14, 16, 18].

Vanilla

Label log-probabilities: A: -5.97, B: -5.79, C: -4.87. **Answer: C) Uncertain** **wrong**

After FRGI hooked prefill

Label log-probabilities: A: -3.80, B: -3.91, C: -3.94. **Answer: A) True** **correct**

Instance diagnostic (vanilla, hidden state)

Evidence separation δ_l between the proof-used and distractor spans of this instance peaks at layer 9 (0.35) and ends at 0.20 over the last three layers (drop 43%).

Fig. 16: **Instance-level FRGI correction on LLaMA-3.1-8B (ProverQA dev instance 566, medium).** Left: the annotated evidence, proof chain, and gold label. Right: the proof-free spans and layers selected by FRGI, together with label log-probabilities before and after the hooked prefill. Although only one of the three selected spans is proof-used, the intervention changes the prediction from *Uncertain* to the correct label *True*. The vanilla hidden states also show a 43% post-peak decline in proof-used–versus–distractor evidence separation.

Problem setup and logical derivation

Facts. **F1.** Taytum eats insects. **F2.**† Taytum either lives in ponds or hibernates seasonally. **F3.** Taytum does not migrate long distances. **F4.** Taytum has webbed feet. **F5.** Taytum is not balanced. **F6.** Taytum regenerates quickly.
Rules. **R1.** If Taytum has either aquatic adaptation or terrestrial adaptation (but not both), then she thrives. **R2.**† If Maximo is resilient, then she explores her environments and regenerates quickly. **R3.** All thriving frogs are resilient. **R4.**† If Duke explores her environment and eats insects, then she is balanced. **R5.**† Inaya either has terrestrial adaptation or hibernates seasonally, but not both. **R6.** If Taytum has webbed feet, then she either migrates long distances or hibernates seasonally, but not both. **R7.**† If Taytum has strong skin, then she jumps far and explores her environments. **R8.** If Taytum is resilient, then she explores her environments and regenerates quickly. **R9.** Taytum either has terrestrial adaptation or hibernates seasonally, but not both. **R10.** If Taytum explores her environment and eats insects, then she is balanced. **R11.**† If Taytum has gills or has lungs, then she can thrive.
Question. Is the statement "Taytum has an aquatic adaptation." true, false, or uncertain?

Annotated proof chain

From **F1**, **F5**, applying **R10**: Taytum does not explore her environments. From **F4**, **F3**, applying **R6**: Taytum hibernates seasonally. From **Taytum does not explore her environments.**, **F6**, applying **R8**: Taytum is not resilient. From **Taytum is not resilient.**, applying **R3**: Taytum does not thrive. From **Taytum hibernates seasonally.**, applying **R9**: Taytum does not have terrestrial adaptation. From **Taytum does not have terrestrial adaptation.**, **Taytum does not thrive.**, applying **R1**: Taytum does not have an aquatic adaptation.

Correct logical outcome

Correct label: B) False. 11 proof-used premises, 6 distractors (†).

Proof-free FRGI intervention (LLaMA-3.1-8B, label scoring)

Proof-free selection

Saliency-selected spans (layers 5–8, σ=0.25): R11†, F1, R3, F5, F2†, R1 (4 of 6 selected spans are proof-used; † = distractor). UCP layers: l^*=8, l_c=3, targets [8, 10, 12, 14, 16, 18].

Vanilla

Label log-probabilities: A: -3.30, B: -3.81, C: -4.34. **Answer: A) True** wrong

After FRGI hooked prefill

Label log-probabilities: A: -4.19, B: -4.10, C: -4.81. **Answer: B) False** correct

Instance diagnostic (vanilla, hidden state)

Evidence separation δ_l between the proof-used and distractor spans of this instance peaks at layer 8 (0.32) and ends at 0.20 over the last three layers (drop 36%).

Fig. 17: **Instance-level FRGI correction on LLaMA-3.1-8B (ProverQA dev instance 1122, hard).** The vanilla model predicts *True*, whereas the proof-free FRGI intervention changes the label ranking and recovers the correct answer *False*. Four of the six saliency-selected spans are proof-used, while the remaining two are distractors. Despite this imperfect selection, the localized intervention is sufficient to reverse the incorrect decision. The vanilla representation exhibits a 36% post-peak decline in evidence separation.

Problem setup and logical derivation

Facts. F1. Caroline seeks change. **F2.** Caroline does not seek a global perspective. **F3.**† Reece appreciates culture or values diversity. **F4.** Caroline appreciates culture or values diversity. **F5.**† Lakelyn does not seek a global perspective. **F6.** Caroline either embraces cultural exchange or seeks global connections. **F7.** Caroline does not appreciate culture. **F8.** Caroline improves her health prospects. **F9.** Caroline fosters social connections. **F10.** Caroline does not value open-mindedness.
Rules. R1.† If Kennedi rejects inbreeding, then she either prioritizes inclusivity or embraces cultural exchange (but not both). **R2.**† If Caroline either respects all backgrounds or values social mixing (but not both), then she prioritizes inclusivity. **R3.**† Caroline will marry outside her family if she rejects tradition. **R4.**† If Lakelyn values diversity and seeks global connections, then she is likely to marry outside her family. **R5.** If Caroline seeks change, then she either values open-mindedness or prioritizes inclusivity (or both). **R6.**† If Reece either seeks a global perspective or fosters social connections (but not both), then she rejects inbreeding. **R7.** If Caroline rejects inbreeding, then she either prioritizes inclusivity or embraces cultural exchange (but not both). **R8.** If Caroline values diversity and seeks global connections, then she is likely to marry outside her family. **R9.** If Caroline either seeks a global perspective or fosters social connections (but not both), then she rejects inbreeding. **R10.**† If Caroline desires personal growth, then she seeks global connections and embraces new experiences. **R11.** For all humans, if they marry outside their family, then they reduce genetic disorders and improve their health prospects.
Question. Is the statement "Caroline reduces genetic disorders." true, false, or uncertain?

Annotated proof chain

From **F2**, **F9**, applying **R9**: Caroline rejects inbreeding. From **F1**, **F10**, applying **R5**: Caroline prioritizes inclusivity. From **F7**, applying **F4**: Caroline values diversity. From **Caroline rejects inbreeding.**, **Caroline prioritizes inclusivity.**, applying **R7**: Caroline does not embrace cultural exchange. From **Caroline does not embrace cultural exchange.**, applying **F6**: Caroline seeks global connections. From **Caroline values diversity.**, **Caroline seeks global connections.**, applying **R8**: Caroline marries outside her family. From **Caroline marries outside her family.**, **F8**, applying **R11**: Caroline reduces genetic disorders.

Correct logical outcome

Correct label: A) True. 13 proof-used premises, 8 distractors (†).

Proof-free FRGI intervention (Mistral-7B, label scoring)

Proof-free selection

Saliency-selected spans (layers 5–8, σ=0.25): F10, R1†, F8, R11, F7, F1, F5†, F9 (6 of 8 selected spans are proof-used; † = distractor). UCP layers: l^*=8, l_c=3, targets [8, 10, 12, 14, 16, 18].

Vanilla

Label log-probabilities: A: -12.30, B: -11.65, C: -0.00. **Answer: C) Uncertain** wrong

After FRGI hooked prefill

Label log-probabilities: A: -0.38, B: -10.82, C: -1.26. **Answer: A) True** correct

Instance diagnostic (vanilla, hidden state)

Evidence separation δ_l between the proof-used and distractor spans of this instance peaks at layer 13 (0.43) and ends at 0.22 over the last three layers (drop 50%).

Fig. 18: **Instance-level FRGI correction on Mistral-7B (ProverQA dev instance 1164, hard).** The vanilla model strongly favors *Uncertain*. After the FRGI hooked prefill, the label scores shift substantially toward the gold answer *True*. Six of the eight selected spans are proof-used, while two are distractors. This instance also shows the largest post-peak decline among the four examples, with evidence separation falling by 50% from its maximum.

Problem setup and logical derivation

Facts. F1. Karson does not collect rare species. **F2.** Karson studies herbalism. **F3.**† Karson enjoys gardening or loves botany. **F4.** Karson has a green thumb. **F5.** Karson appreciates nature.
Rules. R1.† If Karson owns a garden, then he either experiments with plants or grows flowers, but not both. **R2.** If Karson spends time outdoors or appreciates nature, then he can nurture plants. **R3.** If Karson studies herbalism and experiments with plants, then he becomes an expert. **R4.** If someone has a green thumb and loves botany, then they experiment with plants. **R5.**† If someone enjoys gardening and values sustainability, then they nurture plants. **R6.** If Karson either nurtures plants or collects rare species (but not both), then he loves botany.
Question. Is the statement "Karson becomes an expert." true, false, or uncertain?

Annotated proof chain

From **F5**, applying **R2**: Karson nurtures plants. From **Karson nurtures plants.**, **F1**, applying **R6**: Karson loves botany. From **F4**, **Karson loves botany.**, applying **R4**: Karson experiments with plants. From **F2**, **Karson experiments with plants.**, applying **R3**: Karson becomes an expert.

Correct logical outcome

Correct label: A) True. 8 proof-used premises, 3 distractors (†).

Proof-free FRGI intervention (Mistral-7B, label scoring)

Proof-free selection

Saliency-selected spans (layers 5–8, σ=0.25): R1†, F1 (1 of 2 selected spans is proof-used; † = distractor). UCP layers: l^*=8, l_c=3, targets [8, 10, 12, 14, 16, 18].

Vanilla

Label log-probabilities: A: -4.35, B: -14.40, C: -1.81. **Answer: C) Uncertain** **wrong**

After FRGI hooked prefill

Label log-probabilities: A: -0.17, B: -15.56, C: -6.56. **Answer: A) True** **correct**

Instance diagnostic (vanilla, hidden state)

Evidence separation δ_l between the proof-used and distractor spans of this instance peaks at layer 12 (0.42) and ends at 0.30 over the last three layers (drop 29%).

Fig. 19: **Instance-level FRGI correction on Mistral-7B (ProverQA dev instance 541, medium).** The vanilla model predicts *Uncertain*, while the deployed FRGI intervention changes the label ranking and recovers the correct answer *True*. Only one of the two selected spans is proof-used, showing that exact recovery of the annotated proof is not required for the intervention to be effective. The vanilla hidden states show a 29% post-peak decline in proof-used–versus–distractor evidence separation.