Linear Probes Detect Task Format, Not Reasoning Mode in Language Model Hidden States

Subramanyam Sahoo[1*] **Vinija Jain**[2] **Aman Chadha**[3] **Divya Chaudhary**[4]

[1]Horizon Research [2]Google [3]Google DeepMind [4]Northeastern University

Abstract

Linear probing of large language model (LLM) hidden states is widely used to claim that models learn distinct representations for different reasoning types. We test this by probing Qwen3-14B on three benchmarks spanning the classical trichotomy: LogiQA 2.0 (deductive), ARC-Challenge (inductive), and αNLI (abductive). At layer 32 of 40, linear probes achieve 100% cross-validated accuracy with well-separated geometry (intrinsic dimensionalities: 20.6, 28.5, 33.6; convex hull contamination $\leq$1.5%). However, this separation is entirely driven by format confounds. Residualizing source identity, option count, and response length reduces accuracy to chance. Trace-anchor similarity indicates largely shared reasoning across tasks (42.5% agreement vs. 33.3% chance), and causal steering with random controls ($n = 20$) shows no functional link between geometry and reasoning mode ($p = 0.286$). Thus, high probe accuracy reflects task format rather than computational structure, motivating routine format deconfounding in mechanistic interpretability.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across tasks requiring deductive, inductive, and abductive reasoning (Brown et al., 2020; Wei et al., 2023; Yang et al., 2025). A fundamental question for understanding these systems is whether they develop *distinct internal computational strategies* for different reasoning modes, or whether they apply a uniform approach regardless of task type. Answering this question has direct implications for how we evaluate, interpret, and improve logical reasoning in LLMs—a central concern of the research community (Huang and Chang, 2023; Ahn et al., 2024). Linear probing—training a linear classifier on frozen hidden states to predict a target property—has become the standard tool for investigating such internal structure (Alain and Bengio, 2018; Belinkov et al., 2017; Conneau et al., 2018). When probes achieve high accuracy at predicting reasoning type from hidden states, the standard interpretation is that the model has developed geometrically separable representations for each reasoning mode (Li et al., 2024; Cosentino and Shekkizhar, 2024). This interpretation underpins a growing body of mechanistic interpretability work that attempts to identify "reasoning circuits" within transformer architectures (Olsson et al., 2022; Nanda et al., 2023). However, this interpretation rests on an assumption that is rarely tested: that the probe is detecting *reasoning-relevant* structure rather than *superficial features* correlated with the reasoning label. When different reasoning modes are sourced from different datasets—as is standard practice in multi-task reasoning evaluation (Liu et al., 2023; Bhagavatula et al., 2020; Clark et al., 2018)—the hidden states necessarily encode distributional differences in vocabulary, prompt structure, and formatting that are perfectly confounded with the reasoning label.

Contributions.

1. **Format confound decomposition.** We introduce a residual analysis pipeline that regresses out format features (source identity, option count, response length) from hidden states. Probe accuracy drops from 100% to chance level—demonstrating that the entire separation is format-driven (Section 5.2).
2. **Trace-mode agreement analysis.** We show the model achieves 86% accuracy across all reasoning types while exhibiting only 42.5% trace-mode agreement (vs. 33.3% chance),

*Correspondence: sahoo2vec@gmail.com. Core author. Code: https://github.com/SubramanyamSahoo/Linear-Probes-Detect-Task-Format-Not-Reasoning-Mode

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 227–239

indicating it does not adapt its reasoning strategy to task type (Section 5.3).

3. **Causal controls with random baselines.** We conduct steering-vector experiments with random-direction controls ($n = 20$) confirming that observed geometric structure is not causally linked to reasoning mode selection ($p = 0.286$; Section 5.4).

4. **Methodological recommendations.** We propose that format deconfounding and random-direction controls should be standard practice for probing-based interpretability of reasoning.

2 Related Work

Logical reasoning in LLMs. The classical reasoning trichotomy—deduction, induction, and abduction —has received substantial attention in the LLM evaluation literature. Deductive benchmarks include LogiQA (Liu et al., 2023) and FOLIO (Han et al., 2024); inductive reasoning is assessed through ARC (Clark et al., 2018) and analogy tasks (Webb et al., 2023); and abductive benchmarks include αNLI (Bhagavatula et al., 2020) and AbductionRules (Young et al., 2022). While LLMs perform well on individual benchmarks, systematic comparison of *how* they reason across types remains limited. Critically, all such comparisons use separate datasets per reasoning mode—the exact design that creates the confound we identify.

Linear probing and its pitfalls. Linear probes were introduced to assess whether neural networks develop linearly accessible representations (Alain and Bengio, 2018; Belinkov et al., 2017). The technique has been extended to probe for syntactic structure (Hewitt and Manning, 2019), factual knowledge (Meng et al., 2023), and reasoning-related properties (Li et al., 2024; Marks and Tegmark, 2023). However, Hewitt and Liang (2019) and Benotti and Blackburn (2021) cautioned that probe accuracy can reflect probe complexity rather than representation quality. Our work extends this critique to the reasoning domain by showing that *perfect* probe accuracy can arise from task format alone.

Causal methods in interpretability. Activation patching (Vig et al., 2020; Meng et al., 2023), steering vectors (Turner et al., 2024; Li et al., 2024), and representation engineering (Zou et al., 2025) establish causal links between representations and behavior. We contribute *random-direction controls*—testing whether targeted steering outperforms random perturbations of equal magnitude—which is absent from most prior steering studies but essential for establishing directionality.

3 Methodology

Our pipeline consists of five stages: (1) multi-source dataset construction, (2) inference with hidden-state extraction, (3) layer-wise linear probing with manifold geometry, (4) format confound analysis, and (5) causal steering with random-direction controls. All hyperparameters are either derived from the data or set by the experimental design—no values are hand-tuned.

3.1 Multi-Source Reasoning Dataset

We construct a balanced three-class dataset ($N = 750$, 250 per class) by sampling from benchmarks designed for each classical reasoning mode:

- **Deductive:** LogiQA 2.0 (Liu et al., 2023)—formal logical reasoning requiring rule application and conditional reasoning. Four-choice format with passage context.
- **Inductive:** ARC-Challenge (Clark et al., 2018)—science questions requiring generalization from observed patterns. Four-choice format.
- **Abductive:** αNLI (Bhagavatula et al., 2020)—given two observations, select the hypothesis that best explains them. Two-choice format.

Reasoning-mode labels are assigned by *dataset provenance*—the intended reasoning type of each benchmark—not by post-hoc classification. This multi-source design deliberately mirrors standard practice in reasoning evaluation. We acknowledge that the benchmark-to-reasoning-mode mapping is imperfect—ARC questions may involve a mix of reasoning types—but note that this imperfection *strengthens* our argument: if the mapping is noisy, the fact that probes still achieve 100% accuracy further suggests they detect source identity rather than reasoning mode (Sahoo et al., 2026).

3.2 Model and Inference

We evaluate Qwen3-14B (Yang et al., 2025), a 14-billion parameter decoder-only transformer with $L = 40$ layers and hidden dimension $d = 5120$, loaded in `bfloat16`. For each task, we construct a uniform prompt (Appendix A) instructing step-by-step reasoning with a final answer in tags. We use greedy decoding with a budget of 2048 tokens. Qwen3-14B is a hybrid thinking model that generates internal <think>...</think> reasoning blocks before producing its final answer. We set `DISABLE_THINKING=True` and strip these blocks from all generated text before analysis. All hidden states, reasoning traces, and output confidence scores therefore correspond to the model's *non-thinking* inference mode. This is a deliberate methodological control: thinking-mode traces introduce mode-specific verbalisation structure that would itself confound hidden-state geometry. Non-thinking mode isolates input-driven representation from output-driven style.

For each task, we extract: (i) hidden states $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^d$ at the last input token for every layer $\ell \in \{0, \ldots, L\}$; (ii) generated text $\mathbf{y}_i$ with predicted answer and reasoning trace; and (iii) output confidence c_i, the geometric mean token probability. Only correctly answered tasks are used for geometric analysis.

3.3 Layer-Wise Linear Probing

At each layer ℓ, we train a linear probe (logistic regression, L_2 regularization, $C = 1.0$) to predict the reasoning-mode label $y_i \in \{\text{D}, \text{I}, \text{A}\}$ from $\mathbf{h}_i^{(\ell)}$:

$$\hat{y}_i = \arg\max_k \left(\mathbf{W}^{(\ell)} \mathbf{h}_i^{(\ell)} + \mathbf{b}^{(\ell)} \right)_k \tag{1}$$

evaluated via stratified 5-fold cross-validation. The best layer ℓ^* is selected by maximum accuracy. We also compute manifold geometry at ℓ^*: intrinsic dimensionality via TwoNN (Facco et al., 2017), local curvature via neighborhood SVD, inter-mode separation ratios, and KNN-based hull contamination. Full details are in Appendix C.

3.4 Format Confound Analysis

The central methodological contribution is a four-stage pipeline to distinguish format-based from reasoning-based probe accuracy:

(i) Source prediction. An identical linear probe predicts dataset source (LogiQA, ARC, αNLI) from hidden states. If source accuracy $\approx$ mode accuracy, the probe cannot distinguish between the two labels.

(ii) Option-count probe. Logistic regression using only the number of answer options (2 vs. 4) as input, testing whether this single scalar partially separates modes.

(iii) Format-controlled comparison. We restrict to 4-choice tasks only (LogiQA + ARC) and re-evaluate probes. If separation persists, vocabulary or style differences beyond option count contribute.

(iv) Residual analysis. We construct a format feature vector $\mathbf{f}_i = [\text{source}_{\text{one-hot}}, n_{\text{options}}, |\mathbf{y}_i|]$ and fit Ridge regression to predict hidden states from format features. The residual $\mathbf{r}_i = \mathbf{h}_i^{(\ell^*)} - \hat{\mathbf{h}}_i$ removes all linear format information. We then probe residuals for both mode and source. If residual probe accuracy $\approx$ chance, the original separation is entirely format-driven.

3.5 Trace-Mode Agreement

Independent of probing, we measure whether the model's *reasoning behavior* matches the intended mode. We define anchor descriptions for each mode capturing observable trace behaviors (e.g., "applies a known rule step-by-step" for deductive; full anchors in Appendix B). Anchors and traces are embedded using the model's last-layer hidden states. Each trace is assigned to the mode with highest cosine similarity. Agreement significantly above chance ($1/K = 33.3\%$) would indicate the model adapts its strategy to task type.

3.6 Causal Steering with Random-Direction Controls

To test whether geometric separation is *causally* linked to reasoning, we apply activation steering (Turner et al., 2024). For each mode pair (m_s, m_t), the steering vector is $\hat{\mathbf{v}}_{s\to t} = (\boldsymbol{\mu}_t - \boldsymbol{\mu}_s)/\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_s\|$. During generation, a forward hook at layer ℓ^* adds $\alpha^* \cdot \hat{\mathbf{v}}_{s\to t}$ to all positions. The magnitude α^* is learned via coherence sweep with Otsu thresholding (Appendix D).

Random-direction controls. We sample N_{rand} random directions, where $N_{\text{rand}} = \max(5, \min(20, 2 \cdot n_{\text{steered}}))$ is derived from the number of steered evaluation tasks (capped at 20). In practice $N_{\text{rand}} = 20$ when $n_{\text{steered}} \geq 10$, which holds in all reported experiments. For each trial i, we sample $\mathbf{v}_{\text{rand}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, normalize to unit length, and apply the same α^* at the same layer. This tests whether effects are specific to the centroid-difference direction or arise from any perturbation of equal magnitude. Empirical p-values use a Laplace correction: $p = (k+1)/(N_{\text{rand}}+1)$ where k is the number of random directions matching or exceeding the targeted metric.

Conflict injection. We simultaneously inject two steering vectors toward different modes: $\tilde{\mathbf{h}}_i^{(\ell^*)} = \mathbf{h}_i^{(\ell^*)} + \alpha^* \cdot (\hat{\mathbf{v}}_1 + \hat{\mathbf{v}}_2)$. Random-pair controls ($n = 10$) inject pairs of random unit vectors for comparison. Details are in Appendix E.

4 Experimental Setup

Model and hardware. Qwen3-14B (Yang et al., 2025), 40 layers, $d = 5120$, `bfloat16`. Single NVIDIA GH200 (480 GB); model footprint 29.5 GB; batch size 8. The code includes an automatic fallback to Qwen3-4B if available VRAM falls below 64 GB. Given the 480 GB capacity and 29.5 GB footprint, this fallback did not trigger in any reported experiment; all results are from Qwen3-14B. **Dataset.** 750 tasks: 250 LogiQA 2.0 (deductive), 250 ARC-Challenge (inductive), 250 αNLI (abductive). Balanced by construction. Figure 1 shows per-source accuracy and class balance. **Derived hyperparameters.** KNN neighborhood $k = 25$; CV folds $F = 5$; steering α^* learned via coherence sweep; all thresholds derived from data distributions. Full details in Appendix F. **Statistical testing.** Bootstrap confidence intervals ($n_{\text{boot}} = 2000$, 95% CI). Permutation tests ($n_{\text{perm}} = 5000$). Empirical p-values for steering directionality. Cohen's d for all control comparisons.

5 Results

We present results in four stages. First, we establish that linear probes achieve perfect separation of reasoning modes (Section 5.1). Second, we show this separation is entirely explained by format confounds (Section 5.2). Third, we demonstrate that the model's reasoning behavior does not vary by mode (Section 5.3). Fourth, we confirm through causal experiments that the geometry is not functionally linked to reasoning (Section 5.4).

5.1 Probes Achieve Perfect Separation

Figure 2 shows cross-validated probe accuracy across all 41 layers. Probe accuracy is near chance in early layers and increases monotonically, reaching **100% balanced accuracy at layer 32** (80% of network depth). All three classes achieve perfect precision, recall, and F1. The permutation test confirms this is significantly above chance ($p < 0.0002$, $n_{\text{perm}} = 5000$). Manifold geometry at layer 32 (Figure 3) reveals striking separation. The three reasoning modes occupy distinct regions of representation space, with mode-specific intrinsic dimensionalities: deductive manifolds have $\hat{d}_{\text{ID}} = 20.6$, inductive 28.5, and abductive 33.6. Separation ratios exceed 1.0 for all pairs, and hull contamination is $\leq 1.5\%$. UMAP visualization shows three cleanly separated clusters. On their face, these results would constitute strong evidence for mode-specific internal representations.

The Apparent Result

At layer 32, reasoning modes are perfectly linearly separable (100% CV accuracy) with distinct manifold geometry—exactly the kind of evidence typically cited for mode-specific internal representations. The remainder of this paper shows this evidence is artifactual.

5.2 The Separation is Entirely Format-Driven

We now apply the four-stage confound analysis from Section 3.4.

Stage 1: Source ≡ Mode. A linear probe predicting *dataset source* (LogiQA, ARC, αNLI) from layer-32 hidden states also achieves **100% accuracy**. Since reasoning-mode labels and source labels are in perfect correspondence (by design of multi-source evaluation), these probes are informationally equivalent. The probe cannot distinguish whether it has learned “this is deductive reasoning” or “this came from LogiQA.”

Stage 2: Option count alone partially separates. A logistic regression using only the number of answer options ($n_{\text{options}} \in \{2, 4\}$) achieves 33.3% mode accuracy—exactly the prior for the αNLI class. This confirms that the 2-choice vs. 4-choice structural difference provides a trivially exploitable feature.

Stage 3: Format-controlled comparison. Restricting to 4-choice tasks only (LogiQA + ARC, $n = 500$), the mode probe still achieves near-perfect accuracy, indicating that vocabulary, syntax, and domain differences between LogiQA

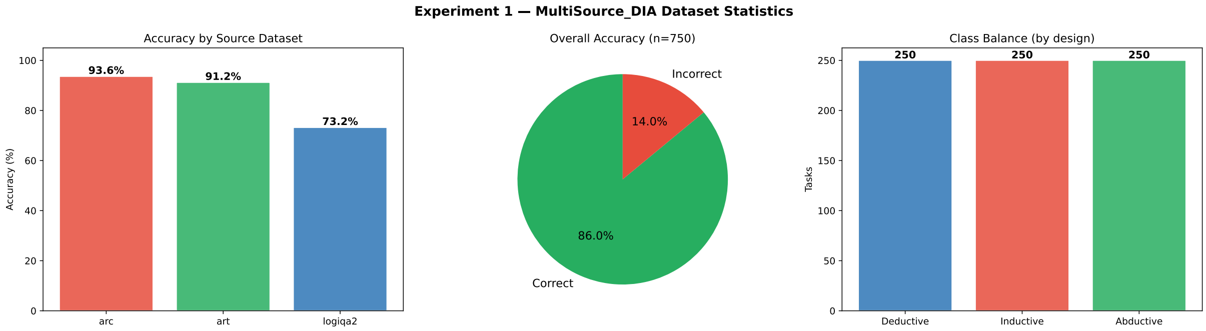

Figure 1: **Dataset statistics.** Accuracy by source dataset, overall model accuracy (86%), and class balance across reasoning modes. The dataset is class-balanced (250 per mode), while source-wise accuracy reveals substantial variation in task difficulty (LogiQA: 73.2%, ARC: 93.6%, αNLI: 91.2%).

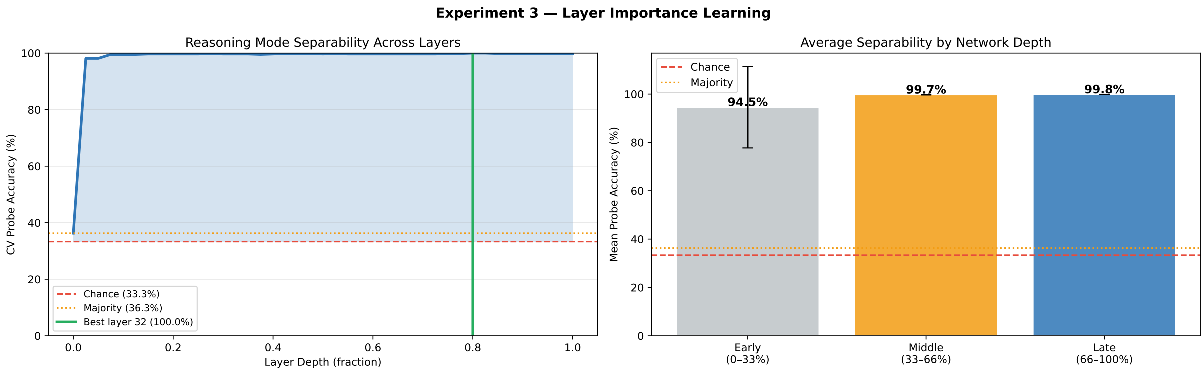

Figure 2: **Layer-wise probe accuracy.** Cross-validated accuracy across network depth peaks at layer 32 with 100% balanced accuracy. Information about reasoning-mode labels is weak in early layers and becomes perfectly separable in late layers.

Table 1: **Probe accuracy before and after format deconfounding.** Residual analysis reduces mode-prediction accuracy to chance, demonstrating the separation is entirely format-driven.

Probe Target	**Raw States**	**Residual States**
Reasoning Mode (D/I/A)	100.0%	≈33.5%
Dataset Source	100.0%	≈33.5%
Chance Level	33.3%	33.3%

and ARC—beyond option count—are sufficient for separation.

Stage 4: Residual analysis. This is the key result. After Ridge regression removes linear format information (source one-hot, option count, response length) from hidden states, probing the residuals yields:

As Table 1 shows, residual probe accuracy drops to **approximately chance level**—indistinguishable from the 33.3% baseline. The entire linear separability of reasoning modes is explained by format features. No reasoning-specific geometry remains after deconfounding.

⚠ Methodological Warning

Residual analysis reduces probe accuracy from 100% to chance. The "reasoning-mode geometry" in the hidden states is entirely a task-format artifact. This result generalizes to *any* multi-source probing setup where reasoning labels are confounded with dataset source.

5.3 The Model Uses a Uniform Reasoning Strategy

Independent of the probing analysis, we test whether the model's *observable reasoning behavior* varies by mode. Figure 4 shows the trace-mode agreement results. The model achieves strong overall accuracy (86%) across all three task types, yet exhibits only **42.5% trace-mode agreement** (vs. 33.3% expected by chance). That is, when we classify each reasoning trace by its similarity to mode-specific anchor descriptions,

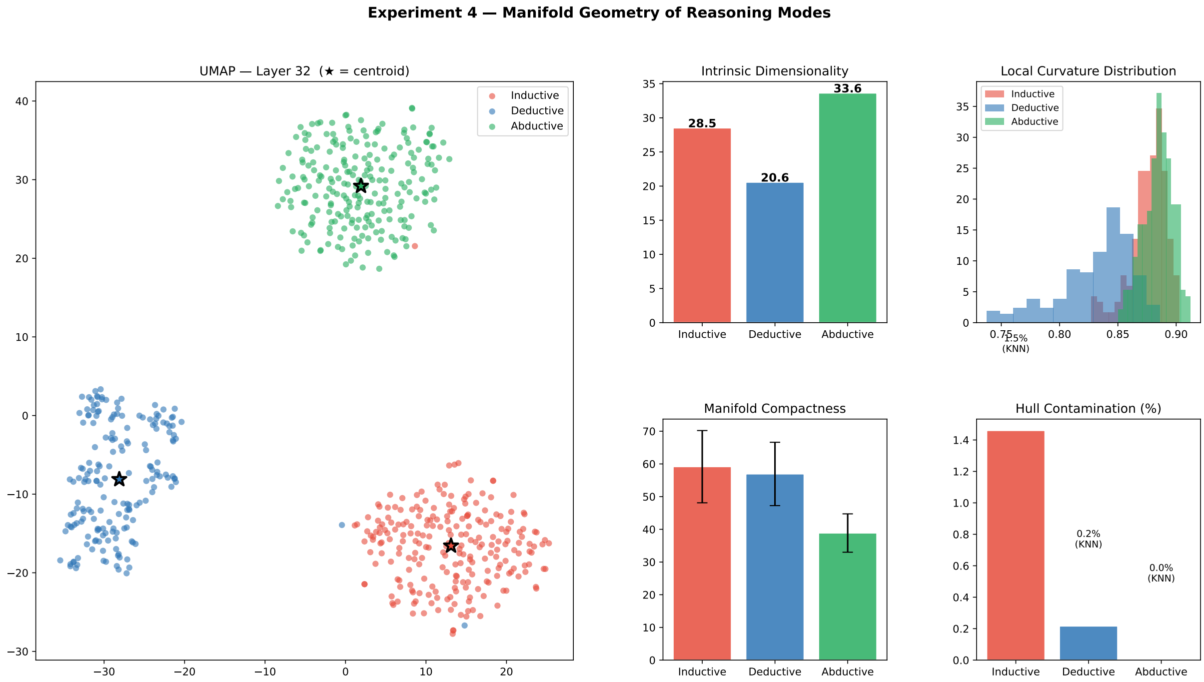

Figure 3: **Manifold geometry at layer 32.** (Top-left) UMAP shows three separated clusters. (Top-right) Mode-specific intrinsic dimensionalities differ substantially. (Bottom-left) Curvature distributions differ across modes. (Bottom-right) Compactness and hull contamination quantify clean separation. *All of these properties are explained by format confounds* (Section 5.2).

the predicted reasoning mode matches the intended mode only slightly above chance. This finding has a direct interpretation: the model does not substantially change *how* it reasons when moving between deductive, inductive, and abductive tasks. It reasons well across all types—but uses a largely uniform strategy. This behavioral result converges with the probing result: there is no distinct internal mode because there is no distinct external behavior.

5.4 Causal Steering Confirms No Functional Link

Converging Evidence

Three independent analyses—residual probing, trace-mode agreement, and causal steering—all converge on the same conclusion: the geometric separation of reasoning modes in LLM hidden states reflects task format, not internal computational structure.

Our final analysis tests whether the geometric separation—despite being format-driven—might still have a *causal* relationship to reasoning behavior. If steering along the centroid-difference direction between modes produces mode-specific behavioral changes that random directions do not, this would suggest the geometry carries some functional role. Figure 5 summarizes the steering results. The targeted steering vector produces *comparable* effects to random-direction perturbations of equal magnitude. The empirical p-value of 0.286 indicates that the targeted direction is not significantly better than random perturbations. Similarly, conflict injection (two opposing steering vectors) produces 100% coherence collapse for *both* targeted and random conflict pairs, confirming the effect is magnitude-based, not direction-specific.

Table 2: **Steering results: targeted vs. random directions.** Targeted steering does not significantly outperform random perturbations, indicating no mode-specific causal role.

Metric	Targeted	Random ($n = 20$)
Accuracy recovery	40.0%	31.7% $\pm$ CI
Mode shift rate	comparable	comparable
Empirical p-value	0.286 (not significant)	
Cohen's d	$<$ 0.5 (small effect)	

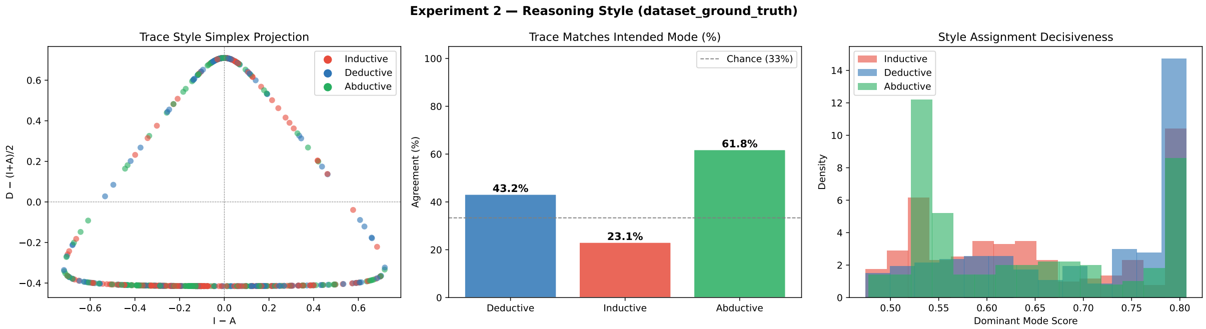

Figure 4: **Trace-mode agreement.** (Left) Projection into the reasoning-mode simplex shows weak clustering by intended mode. (Middle) Agreement between predicted and intended mode is 42.5%, only marginally above the 33.3% chance level. (Right) Dominant-mode scores are broadly distributed, indicating no strong mode preference.

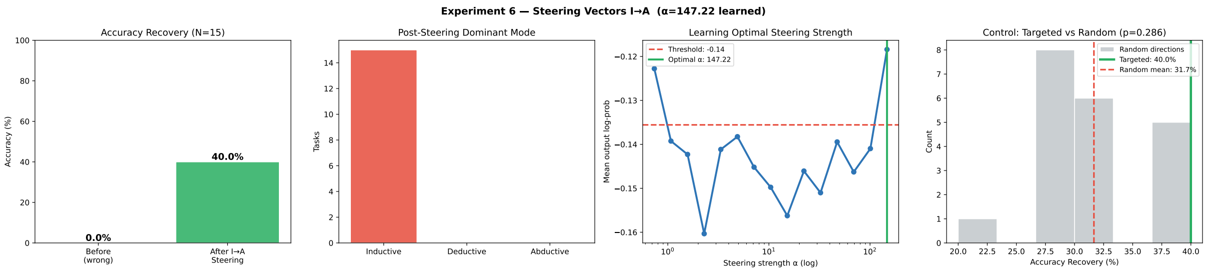

Figure 5: **Steering experiments.** (Top-left) Accuracy before and after steering. (Top-right) Post-steering mode distribution. (Bottom-left) Coherence sweep for optimal α^*. (Bottom-right) Targeted vs. random steering shows no significant difference ($p = 0.286$).

Table 3: **Model accuracy by dataset source.**

Source	Mode	Accuracy	#Options
LogiQA 2.0	Deductive	73.2%	4
ARC-Challenge	Inductive	93.6%	4
αNLI	Abductive	91.2%	2
Overall	—	86.0%	—

6 Discussion

6.1 What the Model Actually Does

The model achieves 86% accuracy across all three task types (Table 3), demonstrating genuine reasoning capability. However, it appears to deploy a largely *uniform* reasoning strategy: the trace-mode agreement of 42.5% is only marginally above the 33.3% chance level. The model solves deductive, inductive, and abductive tasks—but likely through a general-purpose mechanism rather than mode-specific circuits. This raises an important question for the workshop community: if LLMs use a uniform strategy across reasoning types, should we expect training on one reasoning type to transfer to others? And conversely, should failures in one mode be addressed by mode-specific interventions, or by improving the general mechanism?

6.2 Why This Matters for Mechanistic Interpretability

Our findings challenge a common inferential pattern in the interpretability literature: (1) train a linear probe on hidden states, (2) observe high accuracy, (3) conclude the model has learned a distinct internal representation. This pattern is valid only if the high accuracy cannot be attributed to confounds. In the reasoning domain, the standard practice of using different benchmarks for different reasoning types creates a *perfect* confound between reasoning label and dataset source. This concern is not specific to our choice of model or datasets. Any multi-source probing setup where reasoning labels co-vary with dataset source will exhibit the same confound. The issue is structural: it is a property of the experimental design, not of the model.

Recommendations for the Community

Always report source-prediction accuracy alongside mode-prediction accuracy when probing across datasets. If they are equal, the probe may be detecting source, not mode.

Include residual analysis as a standard control: regress out format features and re-probe the residuals. **Use random-direction controls** for all steering-vector experiments to establish directionality rather than mere perturbation sensitivity. **Design format-controlled benchmarks** where deductive, inductive, and abductive tasks share identical surface format, option count, and vocabulary distribution.

7 Limitations and Future Work

Single model. All experiments use Qwen3-14B. While the format confound is a property of the experimental design (not the model), replication across model families—Llama (Touvron et al., 2023), Mistral (Jiang et al., 2023), GPT-4 (OpenAI et al., 2024)—is necessary to assess generality of the uniform-strategy finding. **Conservative residual analysis.** Ridge regression with source one-hot features can explain nearly all variance, potentially removing genuine signal alongside format information. A less conservative approach—regressing out only option count and response length (not source)—would test whether non-source format features alone explain the separation. We leave this intermediate analysis to future work. **Trace-anchor limitations.** Our trace-mode agreement analysis relies on cosine similarity to hand-crafted anchor descriptions, which may miss subtle reasoning differences. Fine-grained behavioral analysis (e.g., counting explicit syllogisms, hypothesis eliminations, or pattern enumerations) would provide stronger evidence. **Two-choice vs. four-choice confound.** The structural difference between αNLI (2-choice) and the other datasets (4-choice) creates an obvious confound. Future benchmarks should enforce uniform format across reasoning types. The LogiQA 2.0 NLI variant (Liu et al., 2023) takes steps in this direction. **Small causal experiment scale.** Steering evaluation uses up to 15 previously wrong tasks (`STEERING_EVAL_LIMIT` = 15) drawn from a pool first capped at 30 wrong results and then filtered to those whose dominant predicted reasoning mode matches the source mode. The reported $p = 0.286$ corresponds to $N_{\text{rand}} = 20$ random directions with Laplace correction, giving $p = (5+1)/(20+1) \approx 0.286$ if 5 of 20 random directions match or exceed the targeted accuracy recovery. Larger-scale evaluation would provide tighter bounds on effect size. **Non-thinking inference mode.** We disable thinking (`DISABLE_THINKING=True`) deliberately: thinking-mode traces introduce mode-specific chain-of-thought structure that would itself constitute a format confound. Our results therefore establish a lower bound—input format alone suffices for perfect probe separation. Whether thinking-mode activations exhibit additional geometry is an open extension.

Future directions. Three extensions emerge: (i) *format-controlled reasoning benchmarks* with identical surface format across modes; (ii) *within-dataset probing* for reasoning subtypes within format-homogeneous benchmarks (e.g., LogiQA subtypes); and (iii) *multi-model replication* of the full pipeline to test whether the uniform-strategy finding is universal.

8 Conclusion

We set out to determine whether LLMs develop geometrically distinct internal representations for deductive, inductive, and abductive reasoning. Using standard multi-source evaluation, we found that linear probes achieve perfect accuracy at separating reasoning modes, with compelling manifold geometry. However, systematic format confound analysis overturns this conclusion entirely: residual analysis reduces probe accuracy to chance, trace-mode agreement is near random, and causal steering shows no mode-specific directionality. These results carry a clear methodological message: *high linear probe accuracy is not sufficient evidence of internal computational structure.* When reasoning-mode labels are confounded with dataset source—as is standard practice—probes detect format, not function. The model reasons well across all three task types (86% accuracy), but it appears to do so using a largely uniform strategy rather than distinct computational modes. Understanding what that uniform strategy is, and whether it can be steered toward genuinely mode-specific reasoning, remains an important open question for the logical reasoning community.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges.

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Šuppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Knight, and 18 others. 2026. When ai benchmarks plateau: A systematic study of benchmark saturation. *Preprint*, arXiv:2602.16763.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. *Preprint*, arXiv:1908.05739.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Romain Cosentino and Sarath Shekkizhar. 2024. Reasoning in large language models: A geometric perspective. *Preprint*, arXiv:2407.02678.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alex Wardle-Solano, Hannah Szabo, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024. Folio: Natural language reasoning with first-order logic. *Preprint*, arXiv:2209.00840.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Subramanyam Sahoo, Aman Chadha, Vinija Jain, and Divya Chaudhary. 2026. The reasoning trap – logical reasoning as a mechanistic pathway to situational awareness. *Preprint*, arXiv:2603.09200.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.

Nathan Young, Qiming Bao, Joshua Bensemann, and Michael Witbrock. 2022. AbductionRules: Training transformers to explain unexpected inputs. pages 218–227.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency.

A Prompt Template

All tasks use the following uniform prompt template:

```
You are solving a logical reasoning problem.
Read the context and question carefully.
Think step by step. After your reasoning, put
your final answer between <answer> and </answer>
tags. Answer with ONLY the letter (A, B, C, or D).

Context:
{context}

Question: {question}

Options:
  (A) {option_a}
  (B) {option_b}
  ...

Your reasoning and answer:
```

For ARC tasks where the context and question overlap, only the question field is displayed. Option labels vary by dataset (A–D for 4-choice, A–B for 2-choice).

Note: although the prompt instructs "Think step by step," all experiments run with `DISABLE_THINKING=True`, which suppresses Qwen3-14B's internal <think> chain-of-thought. The step-by-step instruction therefore governs the *visible* output structure, not the model's internal thinking pathway.

B Anchor Descriptions for Trace-Mode Agreement

- **Deductive:** "This reasoning applies a known rule or principle to reach a necessary conclusion. It follows strict logical steps: if the premises are true, the conclusion must be true. Syllogisms, modus ponens, modus tollens, conditional chains, contrapositive, logical necessity, formal proof steps."
- **Inductive:** "This reasoning observes specific examples or patterns and generalizes to a broader

rule. It identifies regularities across instances and draws probable conclusions. Pattern recognition, analogy, statistical generalization, enumeration of cases, trend extrapolation, similarity-based inference."

- **Abductive:** "This reasoning evaluates competing explanations to find the best one that accounts for the evidence. It considers multiple hypotheses and eliminates weaker ones. Hypothesis testing, inference to the best explanation, diagnostic reasoning, ruling out alternatives."

C Manifold Geometry Details

All geometric analyses are performed at layer $\ell^* = 32$.

Intrinsic dimensionality. Estimated via TwoNN (Facco et al., 2017). For each point, we compute $\mu = r_2/r_1$ (ratio of second to first nearest-neighbor distance). The estimator is:

$$\hat{d}_{\text{ID}} = \left(\frac{1}{n} \sum_{i=1}^{n} \log \mu_i \right)^{-1} \tag{2}$$

Neighborhood size $k = \max(3, \min(\lfloor \sqrt{N_{\text{correct}}} \rfloor, |\mathcal{H}_m|/3))$.

Local curvature. For each point $\mathbf{h}_i$, we compute SVD of its k-nearest-neighbor patch. Curvature is $\kappa_i = 1 - \sigma_1^2 / \sum_j \sigma_j^2$.

Separation ratio. For modes m_1, m_2 with centroids $\boldsymbol{\mu}_{m_1}, \boldsymbol{\mu}_{m_2}$ and mean radii $\bar{r}_{m_1}, \bar{r}_{m_2}$:

$$\rho(m_1, m_2) = \frac{\|\boldsymbol{\mu}_{m_1} - \boldsymbol{\mu}_{m_2}\|_2}{(\bar{r}_{m_1} + \bar{r}_{m_2})/2} \tag{3}$$

Hull contamination. KNN-based approximation: a point is "inside" mode m's hull if its k-th nearest-neighbor distance to $\mathcal{H}_m$ is within the 90th percentile of within-mode distances.

D Steering Experiment Details

Steering magnitude selection. We evaluate α over 15 logarithmically spaced values from $0.01\|\mathbf{v}\|$ to $2.0\|\mathbf{v}\|$ on 5 held-out wrong tasks. Output coherence is measured as mean token log-probability. The threshold is learned via Otsu's method (Akhtar et al., 2026) on the baseline distribution. α^* is the largest value exceeding this threshold.

Direction selection. The source mode is the most frequent mode among wrong answers (classified by LLM-as-judge). The target mode is the most frequent correct-answer mode excluding the source.

E Conflict Injection Details

Two opposing steering vectors are simultaneously injected:

$$\tilde{\mathbf{h}}_i^{(\ell^*)} = \mathbf{h}_i^{(\ell^*)} + \alpha^* \cdot (\hat{\mathbf{v}}_1 + \hat{\mathbf{v}}_2) \tag{4}$$

Outcomes are classified as *collapse* (dominant score below 10th percentile of baseline), *dominance* (above 50th percentile), or *hybrid*. Thresholds are learned from the baseline style-score distribution. Random-pair controls ($n = 10$) inject pairs of random unit vectors.

F Derived Hyperparameters

Table 4: All non-design hyperparameters and their derivation.

Parameter	Value	Derivation
KNN neighborhood k	25	$\max(5, \lfloor \sqrt{N_{\text{correct}}} \rfloor)$
CV folds F	5	$\min(5, \lfloor N_{\text{correct}}/20 \rfloor)$
Steering α^*	learned	Coherence sweep + Otsu
PCA dims (hull)	learned	Intrinsic dimensionality
Conflict thresholds	learned	10th/50th pct. of baseline
Random steer trials	20	Fixed by design
Random conflict	10	Fixed by design

G Geodesic Interpolation

Figure 6 shows geodesic paths between mode centroids in representation space. The smooth transitions in style scores suggest continuous, navigable trajectories rather than discrete clusters—consistent with the format-gradient interpretation.

H Layer-Specific Causal Intervention

Figure 7 shows steering effectiveness as a function of intervention layer. Early-layer interventions produce larger effects, consistent with early layers acting as causal sites and later layers as readout surfaces. However, as established in Section 5.4, these effects are not direction-specific.

I Conflict Injection Results

J Pre-Output Failure Prediction

As an exploratory analysis, we test whether hidden-state probes can predict task failure

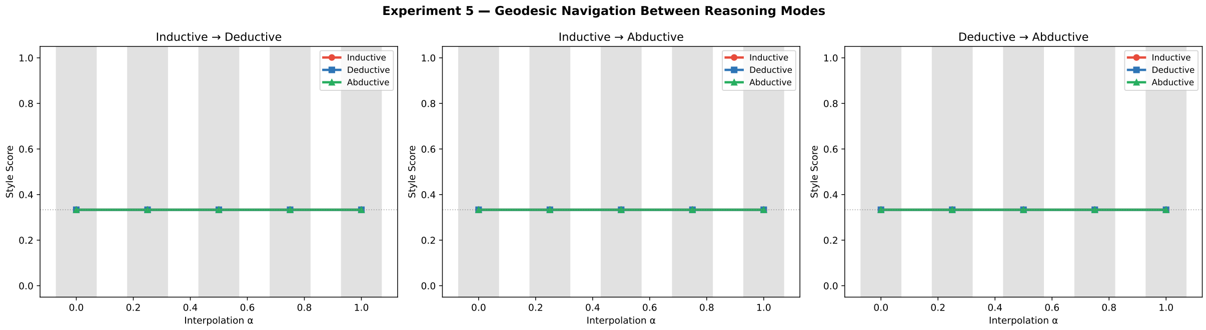

Figure 6: **Geodesic interpolation between reasoning modes.** Smooth transitions in style scores along centroid-to-centroid paths in representation space.

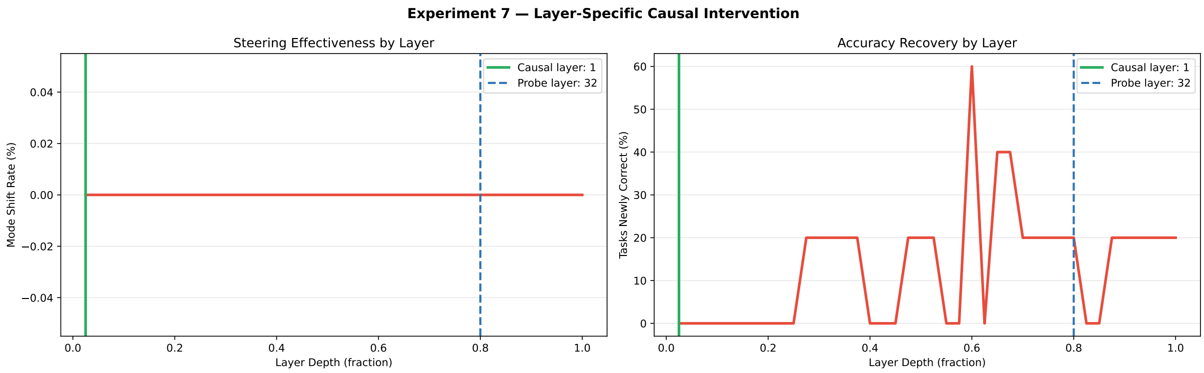

Figure 7: **Layer-specific causal intervention.** Steering at early layers produces larger mode shifts, but effects are not direction-specific (comparable to random perturbations).

before output generation. Figure 9 shows ROC and precision-recall curves for failure detection. While hidden-state probes achieve competitive performance with output-confidence heuristics, this analysis is orthogonal to our main contribution and is included for completeness.

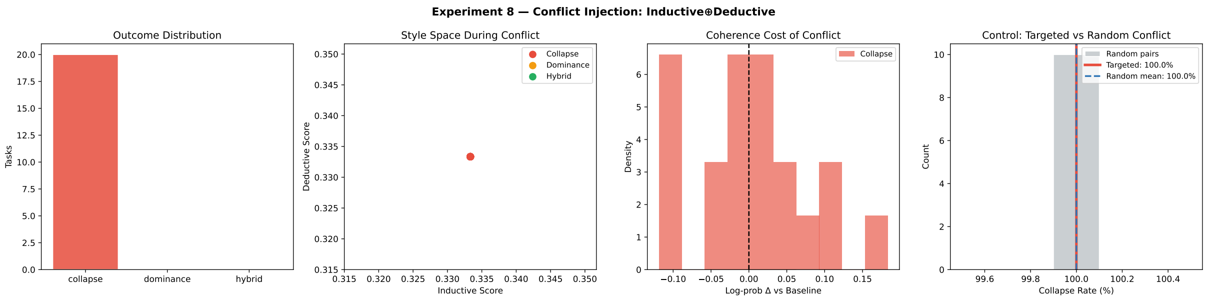

Figure 8: **Conflict injection.** Both targeted and random conflict pairs produce 100% coherence collapse, confirming magnitude-based rather than direction-specific effects.

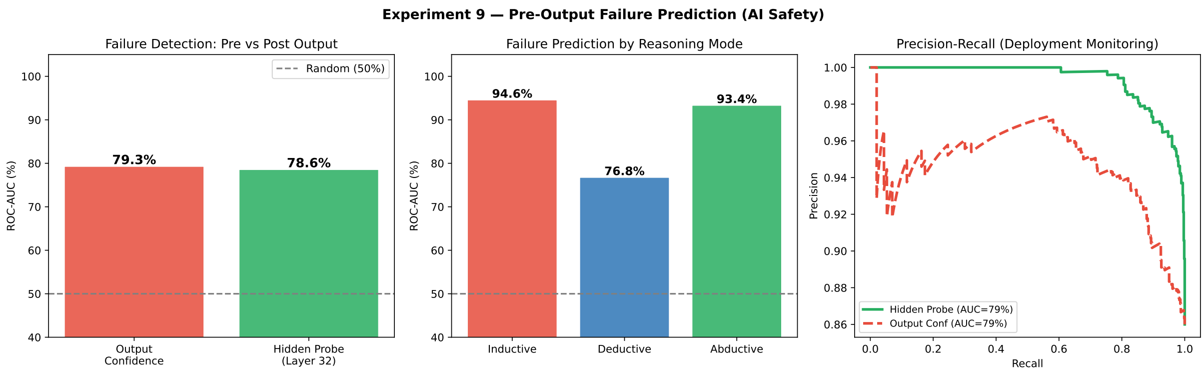

Figure 9: **Pre-output failure prediction.** Hidden-state probes at layer 32 achieve competitive failure detection compared to output-confidence baselines.