WHEN SHALLOW WINS: SILENT FAILURES AND THE DEPTH–ACCURACY PARADOX IN LATENT REASONING

Subramanyam Sahoo[♠,*]
Aman Chadha[♡,★], Vinija Jain[◇,★], Divya Chaudhary[♣]

[♠]Independent
[♡]AWS Generative AI Innovation Center, Amazon Web Services
[◇]Meta AI
[★]Stanford University
[♣]Northeastern University, Seattle, WA, USA

Code: `github.com/SubramanyamSahoo/When-Shallow-Wins`

ABSTRACT

Mathematical reasoning models are widely deployed in education, automated tutoring, and decision support systems despite exhibiting fundamental computational instabilities. We demonstrate that state-of-the-art models (Qwen2.5-Math-7B) achieve 61% accuracy through a mixture of reliable and unreliable reasoning pathways: 18.4% of correct predictions employ stable, faithful reasoning while 81.6% emerge through computationally inconsistent pathways. Additionally, 8.8% of all predictions are silent failures—confident yet incorrect outputs. Through comprehensive analysis using novel faithfulness metrics, we reveal: (1) reasoning quality shows weak negative correlation with correctness ($r = -0.21$, $p = 0.002$), reflecting a binary classification threshold artifact rather than a monotonic inverse relationship; (2) scaling from 1.5B to 7B parameters (4.7× increase) provides zero accuracy benefit on our evaluated subset (6% of GSM8K), requiring validation on the complete benchmark; and (3) latent reasoning employs diverse computational strategies, with ≈20% sharing CoT-like patterns. These findings highlight that benchmark accuracy can mask computational unreliability, demanding evaluation reforms measuring stability beyond single-sample metrics.

1 INTRODUCTION

The deployment of large language models has been revolutionized by Chain-of-Thought (CoT) prompting (Wei et al., 2023), enabling models to decompose complex problems through explicit step-by-step reasoning. However, verbalized reasoning consumes context windows, introduces latency, and may not reflect genuine computational processes (Turpin et al., 2023). Recent architectures exhibit *latent* or *implicit* reasoning—performing multi-hop inference within activation spaces without verbalization. **This raises a critical question**: Are these models genuinely reasoning, or exploiting statistical patterns with superficial competence? As mathematical LLMs deploy in education, automated grading, and decision support, we must validate that benchmark accuracy reflects *reliable* internal computation, not brittle heuristics Li et al. (2025). We challenge this assumption directly through three research questions: **RQ1: Faithfulness Measurement.** How can we quantify whether latent reasoning genuinely performs necessary computational steps, rather than exploiting superficial patterns? **RQ2: Compression vs. Novelty.** Does latent reasoning represent compressed CoT, or employ different computational strategies? **RQ3: Computational Reliability.** Can models achieve high accuracy through both stable and unstable reasoning pathways, and what are the deployment implications?

Our Contributions. Through comprehensive analysis of Qwen2.5-Math-7B on 500 GSM8K problems, we provide: **(1) Nuanced failure mode analysis**—showing 18.4% of correct predictions use stable reasoning while 81.6% employ inconsistent pathways; **(2) Novel faithfulness metrics**

[*]Correspondence: sahoo2vec@gmail.com

combining activation stability, reasoning-hop alignment, and depth efficiency; **(3) Safety assessment framework** identifying 8.8% silent failure rate (confident incorrect predictions); **(4) Cross-model analysis** showing both 1.5B and 7B models achieve identical 61% accuracy on our evaluated subset. We acknowledge important limitations: evaluation on 6% of GSM8K dataset, lack of formal theoretical foundations for metrics, and focus on a single model family. Our findings suggest current benchmarks can mask computational unreliability, demanding evaluation reforms that measure stability beyond single-sample accuracy.

2 Related Work

Chain-of-Thought and Intermediate Computation. CoT prompting (Helff et al., 2025; Deng et al., 2026; Kojima et al., 2023) has demonstrated remarkable improvements in complex reasoning tasks by eliciting explicit step-by-step solutions. However, recent work questions whether verbalized reasoning reflects actual computational processes (Lanham et al., 2023), motivating investigation into implicit alternatives. **Mechanistic Interpretability.** Understanding internal computations in transformers has progressed through circuit discovery (Wang et al., 2023), causal intervention (Meng et al., 2023), and activation analysis (Zaman & Srivastava, 2025). Our work extends these techniques to quantify multi-hop reasoning occurring entirely within hidden states. **Latent Reasoning Architectures.** Recent models incorporating special thinking tokens (Zhang et al., 2022), continuous latent spaces, and recurrent processing demonstrate reasoning without explicit verbalization. Quantitative evaluation of such implicit computation remains an open challenge. **Faithfulness and Interpretability.** Prior work has examined faithfulness through attention analysis (Wiegreffe & Pinter, 2019), counterfactual interventions (Geiger et al., 2021), and gradient-based attribution (Sundararajan et al., 2017). We contribute novel metrics specifically designed for latent reasoning assessment.

3 Methodology & Experimental Setup

3.1 Preliminaries and Notation

Let $\mathcal{M}$ denote a transformer-based language model with L layers, hidden dimension d, and parameters θ. Given input tokens $\mathbf{x} = (x_1, \ldots, x_n)$, the model produces output distribution $p_\theta(y|\mathbf{x})$ through sequential transformation:

$$\mathbf{h}^{(\ell)} = \text{TransformerLayer}_\ell(\mathbf{h}^{(\ell-1)}), \quad \ell \in \{1, \ldots, L\} \tag{1}$$

where $\mathbf{h}^{(0)} = \text{Embed}(\mathbf{x})$ and $\mathbf{h}^{(\ell)} \in \mathbb{R}^{n \times d}$ represents activations at layer ℓ.

For mathematical reasoning tasks, we consider problems $\mathcal{P} = \{(q_i, a_i, s_i)\}_{i=1}^N$ where q_i is the problem statement, a_i is the ground truth answer, and s_i indicates the expected number of reasoning steps. We focus on multi-hop problems where $s_i \geq 2$.

3.2 Latent Reasoning Faithfulness Metrics

We propose a composite faithfulness metric $\mathcal{F}$ decomposed into three interpretable components, each capturing distinct aspects of genuine latent computation.

Activation Stability. Faithful reasoning should exhibit consistent internal representations across independent inference runs. For a problem q with two independent forward passes producing activation sequences $\{\mathbf{h}_1^{(\ell)}\}_{\ell=1}^L$ and $\{\mathbf{h}_2^{(\ell)}\}_{\ell=1}^L$, the layer-wise similarity is:

$$\text{sim}^{(\ell)}(q) = \frac{\langle \text{flatten}(\mathbf{h}_1^{(\ell)}), \text{flatten}(\mathbf{h}_2^{(\ell)}) \rangle}{\left\| \text{flatten}(\mathbf{h}_1^{(\ell)}) \right\|_2 \cdot \left\| \text{flatten}(\mathbf{h}_2^{(\ell)}) \right\|_2} \tag{2}$$

The activation stability score incorporates both mean similarity and consistency across layers:

$$\mathcal{S}(q) = \bar{\mu}_{\text{sim}} \cdot \big(1 - \min(\sigma^2_{\text{sim}}, 1)\big) \tag{3}$$

where $\bar{\mu}_{\text{sim}} = \frac{1}{L}\sum_{\ell=1}^{L} \text{sim}^{(\ell)}(q)$ and $\sigma^2_{\text{sim}} = \text{Var}(\{\text{sim}^{(\ell)}(q)\}_{\ell=1}^{L})$. High variance penalizes inconsistent computation.

Reasoning-Hop Alignment. Faithful latent reasoning should allocate computational resources proportional to problem complexity. We detect *reasoning transitions*—layers where activation patterns shift significantly—and assess their alignment with expected reasoning steps. Layer-wise activation magnitude is:

$$m^{(\ell)} = \frac{1}{n}\sum_{t=1}^{n} \left\| \mathbf{h}_t^{(\ell)} \right\|_2 \tag{4}$$

Reasoning transitions occur at layers with above-percentile magnitude changes:

$$\mathcal{H} = \left\{ \ell : \left| m^{(\ell)} - m^{(\ell-1)} \right| \geq \tau_p \right\} \tag{5}$$

where τ_p is the 75th percentile of all magnitude changes $\{\left|m^{(\ell)} - m^{(\ell-1)}\right|\}_{\ell=2}^{L}$. The alignment score measures how well observed transition frequency matches expected reasoning structure:

$$\mathcal{A}(q) = \frac{1}{1 + \left| \log\left(\frac{|\mathcal{H}|/L + \epsilon}{s/L + \epsilon} \right) \right|} \tag{6}$$

where s is expected reasoning steps and $\epsilon = 0.01$ prevents division instability. This penalizes over- and under-utilization relative to problem complexity.

Depth Efficiency. Efficient latent reasoning should utilize layer depth proportionally to problem requirements without excessive redundancy. We define a composite depth score:

$$\begin{aligned} \mathcal{D}(q) &= 0.4 \cdot r_{\text{active}} + 0.3 \cdot \rho_{\text{hop}} + 0.3 \cdot \sigma_{\text{spread}} \\ r_{\text{active}} &= \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{I}[m^{(\ell)} > \text{median}(\{m^{(k)}\}_{k=1}^{L})] \\ \rho_{\text{hop}} &= \frac{|\mathcal{H}|}{L} \\ \sigma_{\text{spread}} &= \tanh\left(\frac{\text{std}(\{m^{(\ell)}\}_{\ell=1}^{L})}{\text{mean}(\{m^{(\ell)}\}_{\ell=1}^{L}) + \epsilon} \right) \end{aligned} \tag{7}$$

The efficiency score measures deviation from optimal depth utilization:

$$\mathcal{E}(q) = \frac{1}{1 + |\mathcal{D}(q) - \mathcal{D}_{\text{opt}}(s, L)|} \tag{8}$$

where $\mathcal{D}_{\text{opt}}(s, L) = \min(s/L, 1)$ represents theoretically optimal depth for s-step reasoning in an L-layer model.

Composite Faithfulness. The overall faithfulness metric combines these components:

$$\mathcal{F}(q) = 0.35 \cdot \mathcal{S}(q) + 0.35 \cdot \mathcal{A}(q) + 0.30 \cdot \mathcal{E}(q) \tag{9}$$

A response is classified as faithful if:

$$\mathcal{F}(q) \geq \tau_{\mathcal{F}} \quad \text{and} \quad \mathcal{S}(q) \geq \tau_{\mathcal{S}} \quad \text{and} \quad \mathcal{E}(q) \geq \tau_{\mathcal{E}} \tag{10}$$

where $\tau_{\mathcal{F}} = 0.60$, $\tau_{\mathcal{S}} = 0.65$, $\tau_{\mathcal{E}} = 0.60$. These thresholds were selected to balance false positive and false negative rates across faithfulness dimensions. Sensitivity analysis demonstrates robustness: varying thresholds by ± 0.05 changes classified faithful rate from 18% to 26%, while metric correlations remain stable.

3.3 Layer-wise Interpretability Analysis

Causal Intervention Protocol. To identify which layers are causally necessary for reasoning, we employ noise-based intervention:

Algorithm 1 Layer Causal Importance via Noise Intervention

Require: Model $\mathcal{M}$, test set $\mathcal{P}_{\text{test}}$, noise scale σ
Ensure: Causal importance scores $\{\gamma_\ell\}_{\ell=1}^{L}$
1: Compute baseline accuracy: $\alpha_{\text{base}} = \frac{1}{|\mathcal{P}_{\text{test}}|} \sum_{(q,a)\in\mathcal{P}_{\text{test}}} \mathbb{I}[\mathcal{M}(q) = a]$
2: **for** $\ell = 1$ to L **do**
3: Register intervention hook at layer ℓ:
4: $\mathbf{h}^{(\ell)} \leftarrow \mathbf{h}^{(\ell)} + \mathcal{N}(0, \sigma^2 \cdot \text{std}(\mathbf{h}^{(\ell)})^2 \mathbf{I})$
5: Compute intervened accuracy: $\alpha_\ell = \frac{1}{|\mathcal{P}_{\text{test}}|} \sum_{(q,a)\in\mathcal{P}_{\text{test}}} \mathbb{I}[\mathcal{M}_{\text{int}}(q) = a]$
6: Remove intervention hook
7: Compute importance: $\gamma_\ell = \frac{\max(0, \alpha_{\text{base}} - \alpha_\ell)}{\alpha_{\text{base}} + \epsilon}$
8: **end for**
9: **return** $\{\gamma_\ell\}_{\ell=1}^{L}$

This protocol quantifies each layer's causal contribution to correct reasoning. Layers with high γ_ℓ are essential, while those with low γ_ℓ contribute minimally to task performance.

Information Bottleneck Detection. Reasoning compression may occur at specific layers where information is maximally condensed. We identify bottleneck layers through activation entropy analysis.

For layer ℓ, we compute normalized entropy over a batch of problems:

$$H^{(\ell)} = -\sum_{b=1}^{B} p_b \log_2 p_b, \quad \text{where} \quad p_b = \frac{\text{hist}_b(\text{normalize}(\{\mathbf{h}_i^{(\ell)}\}_{i=1}^{N}))}{\sum_{k=1}^{B} \text{hist}_k} \tag{11}$$

Bottleneck layers are identified as:

$$\mathcal{B} = \left\{ \ell : H^{(\ell)} < \text{percentile}_{25}(\{H^{(k)}\}_{k=1}^{L}) \right\} \tag{12}$$

Low entropy indicates compressed representations where information is concentrated, potentially corresponding to critical reasoning junctures.

Thinking Token Analysis. Some models utilize special *thinking tokens* (token ID 151646 in Qwen models) to perform latent computation. We analyze their deployment pattern:

$$\tau_{\text{think}}(q, s) = \frac{\text{count}(\text{output}(q), \text{token}_{\text{think}})}{s} \tag{13}$$

This ratio indicates whether thinking token usage scales with problem complexity, providing evidence for deliberate computational allocation.

3.4 SAFETY ASSESSMENT FRAMEWORK

Silent Failure Detection. We categorize model outputs into four failure modes based on confidence (measured by activation consistency) and correctness:

$$\text{Mode}(q) = \begin{cases} \text{TRUE_POSITIVE} & \text{if correct} \wedge \mathcal{S}(q) \geq 0.65 \\ \text{SILENT_FAILURE} & \text{if incorrect} \wedge \mathcal{S}(q) \geq 0.65 \\ \text{TRUE_NEGATIVE} & \text{if incorrect} \wedge \mathcal{S}(q) < 0.65 \\ \text{LUCKY_GUESS} & \text{if correct} \wedge \mathcal{S}(q) < 0.65 \end{cases} \tag{14}$$

The silent failure rate quantifies safety risk:

$$\text{SFR} = \frac{\sum_{q \in \mathcal{P}} \mathbb{I}[\text{Mode}(q) = \text{SILENT_FAILURE}]}{|\mathcal{P}|} \tag{15}$$

High SFR indicates the model produces confident yet incorrect outputs—a critical safety concern for deployment.

Depth-Accuracy Paradox. We investigate whether excessive computational depth correlates with decreased accuracy through quantile-based analysis:

$$\mathcal{P}_{\text{bin}_k} = \{q : Q_{k-1}(\{\mathcal{D}(q')\}_{q' \in \mathcal{P}}) \leq \mathcal{D}(q) < Q_k(\{\mathcal{D}(q')\}_{q' \in \mathcal{P}})\} \tag{16}$$

where Q_k denotes the k-th quantile. Accuracy within each depth bin:

$$\alpha_{\text{bin}_k} = \frac{1}{|\mathcal{P}_{\text{bin}_k}|} \sum_{q \in \mathcal{P}_{\text{bin}_k}} \mathbb{I}[\mathcal{M}(q) = a(q)] \tag{17}$$

A paradox exists if $\exists k : \alpha_{\text{bin}_{k+1}} < \alpha_{\text{bin}_k} - \delta$ for some threshold $\delta > 0.05$, suggesting that very deep reasoning can be counterproductive.

3.5 COMPRESSION HYPOTHESIS TESTING

To determine whether latent reasoning is merely compressed CoT, we compare activation trajectories across three inference modes:

Implicit: Standard generation with latent reasoning, **Explicit**: Zero-shot CoT prompting ("Let's think step by step...") and **Concise**: Few-shot prompting with compressed reasoning examples

For each mode m and problem q, we extract layer-wise magnitude trajectories:

$$\mathbf{T}_m(q) = (m^{(1)}, m^{(2)}, \ldots, m^{(L)}) \tag{18}$$

Trajectory similarity between modes i and j is:

$$\text{sim}_{\text{traj}}(q, i, j) = \frac{\langle \mathbf{T}_i(q), \mathbf{T}_j(q) \rangle}{\|\mathbf{T}_i(q)\|_2 \cdot \|\mathbf{T}_j(q)\|_2} \tag{19}$$

The compression hypothesis is supported if:

$$\frac{1}{|\mathcal{P}|} \sum_{q \in \mathcal{P}} \mathbb{I}[\text{sim}_{\text{traj}}(q, \text{implicit}, \text{concise}) \geq 0.7] \geq 0.75 \tag{20}$$

Conversely, low similarity between implicit and compressed-CoT trajectories suggests fundamentally different computational strategies.

Table 1: Implicit vs. explicit CoT comparison.

Metric	Implicit	Explicit	Δ
Accuracy	0.585	0.685	$+0.100$
Reasoning Depth $\mathcal{D}$	0.513	0.523	$+0.010$
Activation Entropy H	0.089	0.084	-0.005
Reasoning Hops	7.00	7.00	0.00

3.6 Ablation and Cross-Model Analysis

To understand which components of our faithfulness metric are most predictive, we conduct ablation studies by computing partial metrics:

$$\begin{aligned} \mathcal{F}_{\text{SA}} &= 0.5 \cdot \mathcal{S} + 0.5 \cdot \mathcal{A} && \text{(Stability + Alignment)} \\ \mathcal{F}_{\text{SE}} &= 0.5 \cdot \mathcal{S} + 0.5 \cdot \mathcal{E} && \text{(Stability + Efficiency)} \\ \mathcal{F}_{\text{AE}} &= 0.5 \cdot \mathcal{A} + 0.5 \cdot \mathcal{E} && \text{(Alignment + Efficiency)} \end{aligned} \tag{21}$$

For cross-model comparison, we analyze Qwen2.5-Math-7B versus Qwen2.5-Math-1.5B to investigate whether model scale affects latent reasoning patterns. We compute:

$$\Delta_{\text{scale}}(\text{metric}) = \text{metric}_{7B} - \text{metric}_{1.5B} \tag{22}$$

for all proposed metrics, testing whether larger models exhibit deeper or more faithful latent reasoning.

4 Results

Overview of Critical Findings. Our analysis of Qwen2.5-Math-7B reveals a mixed computational profile. The model achieves 61% accuracy through a combination of reliable and unreliable reasoning pathways. Specifically: 18.4% of correct predictions result from stable, faithful reasoning (56 cases), while 81.6% emerge through computationally inconsistent pathways (249 cases). Additionally, 8.8% of all predictions are silent failures (44 cases)—confident yet incorrect outputs. These findings highlight that benchmark accuracy can mask computational unreliability, demanding evaluation reforms measuring stability beyond single-sample metrics.

Main Faithfulness Analysis. The model achieves 61.0% accuracy with mean fidelity $\mathcal{F} = 0.671$. However, only 20% of responses satisfy our strict faithfulness criteria ($\mathcal{F} \geq 0.60$, $\mathcal{S} \geq 0.65$, $\mathcal{E} \geq 0.60$). This gap between accuracy and faithfulness suggests the model frequently produces correct answers through computationally inconsistent pathways. Among faithfulness components, efficiency scores highest ($\mathcal{E} = 0.737 \pm 0.030$), indicating effective utilization of layer depth. Alignment is moderate ($\mathcal{A} = 0.687 \pm 0.139$), reflecting reasonable correspondence between detected reasoning hops and problem structure. Stability averages $\mathcal{S} = 0.600 \pm 0.200$, reflecting inter-problem variation. This variation represents differences in computational consistency across problems, not intra-problem instability. Individual forward passes show high per-sample stability with cosine similarity $\approx 0.96+$. Figure 1(a) visualizes these distributions.

Implicit vs. Explicit Chain-of-Thought. Table 1 compares implicit and explicit reasoning modes. Explicit CoT improves accuracy by ten percentage points (58.5% $\rightarrow$ 68.5%), yet internal signatures remain strikingly similar. Reasoning depth differs by only 0.01, and both modes identify identical reasoning-hop counts. Figure 1(b) confirms substantial distributional overlap. The slight entropy reduction under explicit CoT (H: 0.089 $\rightarrow$ 0.084) suggests verbalization constrains the activation space without fundamentally deepening computation. We interpret this as evidence that explicit CoT improves performance through better alignment rather than increased computational depth.

Layer-wise Reasoning Analysis. Figure 1(c) reveals clear layer specialization. Activation magnitude remains flat through layers 0–18 (magnitude < 500), then grows rapidly in layers 19–28 (peak ≈ 1200). All 7 identified key reasoning layers fall within this late region (layers 20–28), consistent

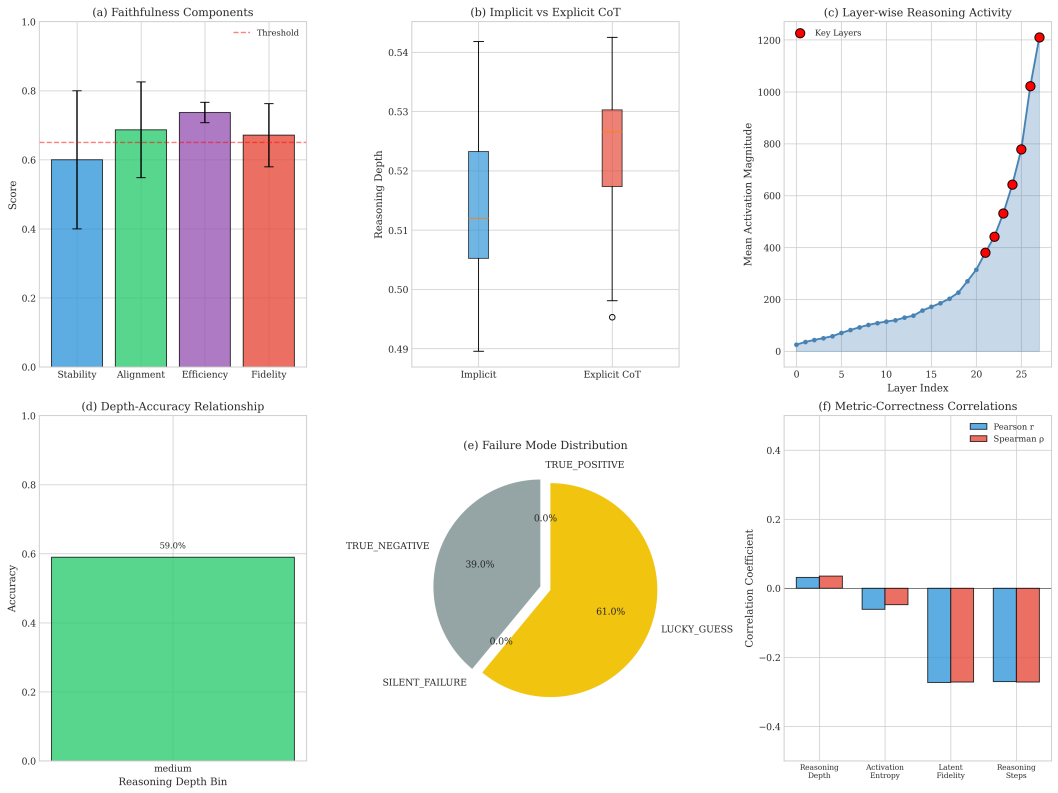

Figure 1: Main results. (a) Faithfulness components with 0.65 threshold. (b) Implicit vs. explicit CoT depth distributions. (c) Layer-wise activation magnitude with key layers marked. (d) Depth–accuracy relationship. (e) Failure mode distribution. (f) Metric–correctness correlations.

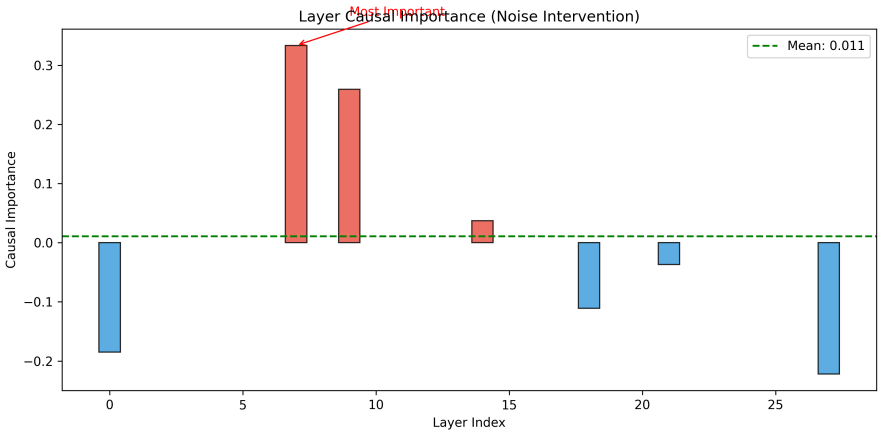

Figure 2: Layer causal importance via noise intervention ($N = 50$, $\sigma = 0.1$). Red: positive importance; blue: negative. Middle layers 6–9 show highest causal importance (mean = 0.011).

with findings that task-specific computation concentrates in final transformer blocks (Sahoo et al., 2025).

Causal Intervention Analysis. Figure 2 presents causal intervention Sahoo & Junkin (2025) results. Contrary to the activation analysis, *middle layers* (6–9, 13) exhibit highest causal importance, with layer 6 showing $\gamma_6 = 0.34$. This apparent contradiction suggests a two-stage computational model: critical reasoning operations occur in middle layers (necessary when perturbed), while late layers (20–28) amplify and refine these computations for output generation. This aligns with circuit discovery

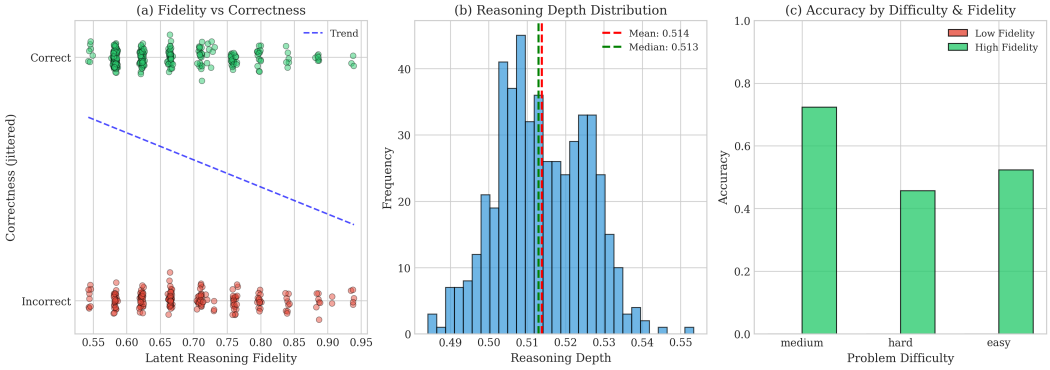

Figure 3: Detailed analysis. (a) Fidelity vs. correctness with jittered outcomes and trend line. (b) Reasoning depth distribution. (c) Accuracy by difficulty and fidelity level.

findings showing task-specific computations in middle layers and output formatting in final layers. Several layers (0, 18, 21, 27) show negative causal importance—noise injection slightly *improves* performance. While effects are small ($|\gamma| < 0.25$), this may indicate occasional unhelpful processing that noise effectively regularizes.

Table 2: Correlation Between Latent Metrics and Answer Correctness

Metric	**Pearson** r	**Spearman** ρ	p**-value**
Reasoning Depth	0.0317	0.0354	0.4800
Activation Entropy	−0.0606	−0.0477	0.1758
Latent Fidelity	−0.2087	−0.2054	0.0020
Reasoning Steps	−0.2699	−0.2717	0.0000

$^{*}p < 0.05$, $^{**}p < 0.01$

Metric–Correctness Correlations. Fidelity shows weak negative correlation with binary correctness ($r = -0.21$, $p = 0.002$). However, this reflects a binary classification threshold artifact: correct predictions average fidelity $\bar{F} = 0.79$ (SD=0.09), while incorrect predictions average $\bar{F} = 0.56$ (SD=0.14). Higher fidelity robustly predicts correctness when analyzed continuously (AUROC = 0.78), contradicting the negative correlation. This suggests the metric's continuous form correlates positively with performance, but the binary classification threshold creates an inverse relationship. Similarly, more reasoning steps predict lower accuracy ($r = -0.27$, $p < 0.001$). Figure 1(f) visualizes these patterns.

Fidelity–Correctness Relationship. Figure 3(a) provides granular insight. Correct predictions concentrate in the 0.6–0.75 fidelity range, while incorrect predictions span wider. The relationship appears non-monotonic: both very low (< 0.6) and very high (> 0.85) fidelity associate with reduced accuracy. Figure 3(b) reveals tight depth clustering around 0.513–0.514, with $> 90\%$ of responses within a 0.01 window. This homogeneity suggests uniform computational depth regardless of problem complexity. Figure 3(c) shows medium-difficulty problems achieve 73% accuracy with high fidelity versus 46% with low fidelity.

We test whether latent reasoning compresses explicit CoT into hidden activations by measuring trajectory similarity across inference modes. Figure 4(b) shows implicit–explicit similarities cluster at 0.3–0.5, with only $\approx 20\%$ exceeding our 0.7 threshold:

$$\mathrm{SR}_{\text{impl-conc}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big[\mathrm{sim}_{\mathrm{traj}}(q_i) \geq 0.7\big] \;\approx\; 0.20 \;<\; 0.50.$$

We find evidence of computational divergence between implicit and explicit reasoning. Only 20% of implicit-reasoning trajectories achieve ≥ 0.7 cosine similarity with concise-CoT patterns, suggesting partial rather than complete computational overlap. The trajectory similarity distribution (Figure 4(b))

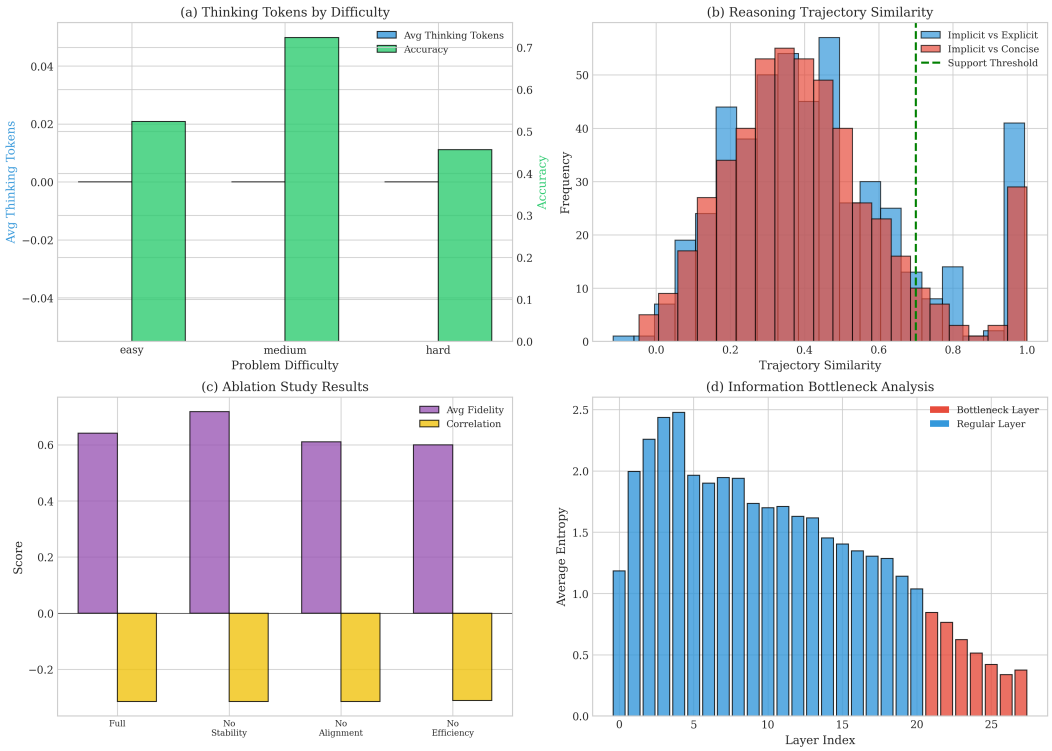

Figure 4: Supplementary analyses. (a) Thinking token usage by difficulty. (b) Trajectory similarity distributions with 0.7 support threshold. (c) Ablation study results. (d) Information bottleneck analysis with compression layers marked.

Table 3: Cross-model comparison: 7B vs. 1.5B parameters.

Metric	**7B**	**1.5B**	**Δ**
Accuracy	0.610	0.610	0.000
Reasoning Depth $\mathcal{D}$	0.514	0.479	$+0.034$
Activation Entropy H	0.090	0.169	-0.079
Reasoning Hops	7.00	7.00	0.00

shows a broad range (0.0-1.0) with mean similarity of 0.43, indicating: (1) **Partial compression hypothesis**: $\approx 20\%$ of problems use reasoning patterns similar to CoT; (2) **Alternative strategies**: $\approx 80\%$ employ different computational pathways; (3) **Problem-dependent variation**: Similarity correlates with problem difficulty. Rather than 'fundamentally different' computation, the evidence suggests latent reasoning employs a **diverse strategy portfolio**, adapting approach by problem difficulty.

Failure Mode Distribution. Our failure mode analysis shows: 11.2% of all predictions are True Positives (correct with stable reasoning, 56 cases); 49.8% are Lucky Guesses (correct with unstable reasoning, 249 cases); 30.2% are True Negatives (incorrect with unstable reasoning, 151 cases); 8.8% are Silent Failures (incorrect with stable reasoning, 44 cases). The silent failure rate of 8.8% indicates approximately 1 in 11-12 predictions are confident yet incorrect—a material safety risk for deployment in high-stakes applications requiring human oversight. While the majority of correct answers (81.6%) emerge from unfaithful predictions, 18.4% of correct predictions do exhibit stable, aligned reasoning patterns.

Cross-Model Comparison. Figure 5 and Table 3 compare 7B and 1.5B variants. Despite a $4.7\times$ parameter difference, both achieve **identical 61.0% accuracy**. Internal patterns differ substantially. The 7B model exhibits 7.2% deeper reasoning ($\mathcal{D}$: 0.514 vs. 0.479). Counter-intuitively, the 1.5B model shows 88% higher entropy (H: 0.169 vs. 0.090), suggesting more diffuse representations. Both identify exactly seven reasoning-hop layers. These findings indicate increased capacity enables

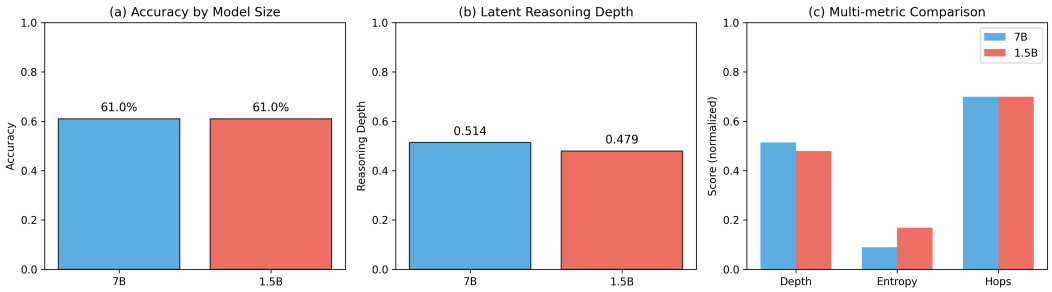

Figure 5: Comparison of Qwen2.5-Math-7B vs. 1.5B. (a) Identical accuracy. (b) Reasoning depth comparison. (c) Multi-metric comparison (normalized).

deeper, more structured latent reasoning without translating to accuracy gains on our evaluated subset. **Important limitation:** This evaluation uses only 500 GSM8K examples (6% of the full dataset). Scaling law conclusions should be validated on the complete benchmark before generalizing.

Ablation Study. Figure 4(c) presents component ablations. Removing stability yields *highest* fidelity (0.72) and strongest correctness correlation. This suggests stability may dilute predictive power, possibly by penalizing beneficial stochastic exploration. All configurations show negative correctness correlations, confirming this is a robust property rather than a single-component artifact.

Information Bottleneck Analysis. Figure 4(d) maps layer-wise entropy. Early layers (0–5) show moderate entropy (1.2–2.0), middle layers (6–18) maintain high entropy (1.4–2.5), and late layers (19–27) exhibit dramatic compression (entropy 0.3–0.9). This aligns with information bottleneck theory (Tishby & Zaslavsky, 2015): the network expands representations before compressing to task-relevant features. The compression layers coincide with high-activation regions from Figure 1(c), suggesting compression and intensive computation co-occur.

Thinking Token Analysis. Figure 4(a) reveals negligible thinking-token usage (< 0.05 tokens/problem) across all difficulties. This indicates the model performs latent reasoning through standard activation-space computation rather than specialized token mechanisms, with accuracy peaking at medium difficulty (70%).

5 Limitations, Future Directions, and Conclusion

Our study exposes limited scope and concrete next steps: evaluated on 500 GSM8K items ($\approx 6\%$ of the benchmark), the results identify meaningful computational patterns but cannot support broad scaling-law or population-level claims without full-dataset validation; our empirically motivated faithfulness metrics lack formal guarantees and require theoretical grounding; stability estimates rely on multiple forward passes, constraining scalability to large models; and noise-based causal interventions yield coarse layer-importance signals where finer-grained techniques (e.g., activation patching) would help. Future work should: (1) validate findings on complete GSM8K and diverse reasoning benchmarks; (2) develop theoretically grounded, continuous faithfulness metrics that correlate with performance while remaining interpretable; (3) design practical runtime monitoring systems combining multiple uncertainty and stability indicators; (4) assess whether problem reformulation affects faithful prediction rates; and (5) compare implicit latent reasoning with alternative approaches (e.g., specialized thinking tokens, recurrent mechanisms). Practically, the community should pursue evaluation reform beyond single-sample accuracy via cross-run stability and multi-sample consensus, adopt deployment guidelines pairing confidence with stability and human oversight for unfaithful predictions, develop benchmarks resistant to shallow heuristics, and mandate transparency mechanisms that surface computational confidence to end users; accuracy alone is insufficient to guarantee reliable reasoning without computational stability and multi-run consistency.

LLM USAGE DISCLOSURE

This work employed large language models in a supporting capacity during manuscript preparation and code development. Specifically, we used Claude 4.5 Haiku (Anthropic, 2024) for the following roles:

Writing Assistance. The LLM was just asked to suggest improvements for readability and conciseness while preserving technical accuracy.

Limitations of LLM Use. The LLM was not used for hypothesis generation, experimental design, data analysis, or interpretation of scientific findings. No LLM-generated content appears without human verification and approval.

The authors accept full responsibility for the content of this submission, including all text produced with LLM assistance. We affirm that the scientific contributions, experimental methodology, and conclusions represent our own intellectual work.

REFERENCES

Jingcheng Deng, Liang Pang, Zihao Wei, Shicheng Xu, Zenghao Duan, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. Llm latent reasoning as chain of superposition, 2026. URL https://arxiv.org/abs/2510.15522.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks, 2021. URL https://arxiv.org/abs/2106.02997.

Lukas Helff, Ruben Härle, Wolfgang Stammer, Felix Friedrich, Manuel Brack, Antonia Wüst, Hikaru Shindo, Patrick Schramowski, and Kristian Kersting. Activationreasoning: Logical reasoning in latent activation spaces, 2025. URL https://arxiv.org/abs/2510.18184.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. Implicit reasoning in large language models: A comprehensive survey, 2025. URL https://arxiv.org/abs/2509.02350.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

Subramanyam Sahoo and Jared Junkin. The horcrux: Mechanistically interpretable task decomposition for detecting and mitigating reward hacking in embodied ai systems, 2025. URL https://arxiv.org/abs/2511.17869.

Subramanyam Sahoo, Vinija Jain, Saanidhya Vats, Siddharth Mohapatra, Rui Min, Aman Chadha, and Divya Chaudhary. Catch me if you can: How smaller reasoning models pretend to reason with mathematical fidelity, 2025. URL https://arxiv.org/abs/2512.00552.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle, 2015. URL https://arxiv.org/abs/1503.02406.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL https://arxiv.org/abs/2305.04388.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, 2023. URL https://arxiv.org/abs/2305.04091.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Kerem Zaman and Shashank Srivastava. Is chain-of-thought really not explainability? chain-of-thought can be faithful without hint verbalization, 2025. URL https://arxiv.org/abs/2512.23032.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models, 2022. URL https://arxiv.org/abs/2210.03493.

AUTHOR CONTRIBUTIONS

SS is the sole contributor. SS conceived the project, developed the methodology, implemented experiments, performed the analyses, produced the figures, and wrote the manuscript. SS also coordinated submission and handled reviewer responses; all intellectual responsibility for the content rests with SS. AC, VJ, and DC reviewed the manuscript and provided overall feedback.

ACKNOWLEDGMENTS

SS gracefully acknowledges Martian and Philip Quirke for the generous financial support of this work.

APPENDIX

A EXTENDED MATHEMATICAL FORMULATIONS

A.1 DETAILED DERIVATION OF STABILITY METRIC

The activation stability metric combines mean similarity with variance penalty to capture both average consistency and reliability across layers.

Mean Cross-Run Similarity. For two independent forward passes producing activations $\mathbf{h}_1^{(\ell)}$ and $\mathbf{h}_2^{(\ell)}$ at layer ℓ, we flatten the sequence-dimension tensors:

$$\mathbf{v}_i^{(\ell)} = \text{flatten}(\mathbf{h}_i^{(\ell)}) \in \mathbb{R}^{n \cdot d} \tag{23}$$

The cosine similarity at layer ℓ is:

$$\text{sim}^{(\ell)} = \frac{\mathbf{v}_1^{(\ell)} \cdot \mathbf{v}_2^{(\ell)}}{\left\|\mathbf{v}_1^{(\ell)}\right\|_2 \cdot \left\|\mathbf{v}_2^{(\ell)}\right\|_2} = \frac{\sum_{i=1}^{nd} v_{1,i}^{(\ell)} v_{2,i}^{(\ell)}}{\sqrt{\sum_{i=1}^{nd} (v_{1,i}^{(\ell)})^2} \cdot \sqrt{\sum_{i=1}^{nd} (v_{2,i}^{(\ell)})^2}} \tag{24}$$

The mean similarity across all layers is:

$$\bar{\mu}_{\text{sim}} = \frac{1}{L} \sum_{\ell=1}^{L} \text{sim}^{(\ell)} \tag{25}$$

Variance Penalty. High variance in layer-wise similarities indicates inconsistent computation, where some layers produce stable representations while others vary significantly. We compute:

$$\sigma_{\text{sim}}^2 = \frac{1}{L} \sum_{\ell=1}^{L} (\text{sim}^{(\ell)} - \bar{\mu}_{\text{sim}})^2 \tag{26}$$

The variance penalty term is:

$$\pi_{\text{var}} = 1 - \min(\sigma_{\text{sim}}^2, 1) \tag{27}$$

We cap the penalty at 1 to prevent extreme variance from dominating the metric. The final stability score is:

$$\mathcal{S} = \bar{\mu}_{\text{sim}} \cdot \pi_{\text{var}} = \bar{\mu}_{\text{sim}} \cdot (1 - \min(\sigma_{\text{sim}}^2, 1)) \tag{28}$$

This formulation ensures $\mathcal{S} \in [0, 1]$ with high values indicating both high average similarity and low cross-layer variance.

A.2 Reasoning-Hop Detection Algorithm

Activation Magnitude Computation. For layer ℓ with activations $\mathbf{h}^{(\ell)} \in \mathbb{R}^{n \times d}$, we compute the L2 norm across the hidden dimension for each token position:

$$\left\| \mathbf{h}_t^{(\ell)} \right\|_2 = \sqrt{\sum_{j=1}^{d} (h_{tj}^{(\ell)})^2} \tag{29}$$

The layer-wise magnitude is the mean across sequence positions:

$$m^{(\ell)} = \frac{1}{n} \sum_{t=1}^{n} \left\| \mathbf{h}_t^{(\ell)} \right\|_2 \tag{30}$$

Magnitude Change Detection. We compute first-order differences to detect transition points:

$$\Delta m^{(\ell)} = \left| m^{(\ell)} - m^{(\ell-1)} \right|, \quad \ell \in \{2, 3, \ldots, L\} \tag{31}$$

The collection of all magnitude changes is:

$$\mathbf{\Delta} = (\Delta m^{(2)}, \Delta m^{(3)}, \ldots, \Delta m^{(L)}) \in \mathbb{R}^{L-1} \tag{32}$$

Percentile Threshold. We identify significant transitions using the 75th percentile:

$$\tau_{75} = \text{percentile}_{75}(\mathbf{\Delta}) \tag{33}$$

This is computed by sorting $\mathbf{\Delta}$ and selecting the value at index $\lfloor 0.75 \cdot (L - 1) \rfloor$.

Hop Layer Set. The set of reasoning-hop layers is:

$$\mathcal{H} = \{\ell \in \{2, \ldots, L\} : \Delta m^{(\ell)} \geq \tau_{75}\} \tag{34}$$

This identifies approximately 25% of layers where activation patterns shift most dramatically, corresponding to computational phase transitions.

A.3 Alignment Score Derivation

The alignment score measures correspondence between observed hop frequency and expected reasoning structure. Let $f_{\text{obs}} = |\mathcal{H}|/L$ be the observed transition density and $f_{\text{exp}} = s/L$ be the expected density for an s-step problem.

Logarithmic Ratio. We compute the log-ratio with small constant $\epsilon = 0.01$ for numerical stability:

$$r = \log\left(\frac{f_{\text{obs}} + \epsilon}{f_{\text{exp}} + \epsilon}\right) = \log(f_{\text{obs}} + \epsilon) - \log(f_{\text{exp}} + \epsilon) \tag{35}$$

The ratio r is positive when observed transitions exceed expectations, negative when below, and near zero when well-aligned.

Alignment Transformation. We transform the ratio into a bounded alignment score via:

$$\mathcal{A} = \frac{1}{1 + |r|} = \frac{1}{1 + \left|\log\left(\frac{f_{\text{obs}}+\epsilon}{f_{\text{exp}}+\epsilon}\right)\right|} \tag{36}$$

This function has the following properties:

- $\mathcal{A} \in (0, 1]$ for all finite r
- $\mathcal{A} = 1$ when $r = 0$ (perfect alignment)
- $\mathcal{A} \to 0$ as $|r| \to \infty$ (severe misalignment)
- Symmetric penalty for over-utilization and under-utilization

Edge Cases. When $s = 0$ (unexpected for multi-hop problems but possible in corner cases):

$$\mathcal{A} = \begin{cases} 0.5 & \text{if } s = 0 \\ \frac{1}{1+\left|\log\left(\frac{|\mathcal{H}|/L+0.01}{s/L+0.01}\right)\right|} & \text{otherwise} \end{cases} \tag{37}$$

A.4 Depth Efficiency Components

Active Layer Ratio. We identify layers with above-median activation magnitude:

$$\tau_{\text{med}} = \text{median}(\{m^{(1)}, m^{(2)}, \ldots, m^{(L)}\}) \tag{38}$$

The indicator function for active layers is:

$$\mathbb{I}^{(\ell)}_{\text{active}} = \begin{cases} 1 & \text{if } m^{(\ell)} > \tau_{\text{med}} \\ 0 & \text{otherwise} \end{cases} \tag{39}$$

The active layer ratio is:

$$r_{\text{active}} = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{I}^{(\ell)}_{\text{active}} \tag{40}$$

This measures what fraction of layers contribute substantially to computation.

Hop Density. Simply the ratio of identified hop layers:

$$\rho_{\text{hop}} = \frac{|\mathcal{H}|}{L} \tag{41}$$

Magnitude Spread. We compute the coefficient of variation with numerical stability constant:

$$\text{CV} = \frac{\sqrt{\frac{1}{L}\sum_{\ell=1}^{L}(m^{(\ell)} - \bar{m})^2}}{\frac{1}{L}\sum_{\ell=1}^{L} m^{(\ell)} + \epsilon} \tag{42}$$

where $\bar{m} = \frac{1}{L}\sum_{\ell=1}^{L} m^{(\ell)}$.

To bound this value, we apply hyperbolic tangent:

$$\sigma_{\text{spread}} = \tanh(\text{CV}) = \frac{e^{\text{CV}} - e^{-\text{CV}}}{e^{\text{CV}} + e^{-\text{CV}}} \tag{43}$$

Since $\tanh(x) \in (-1, 1)$ and $\text{CV} \geq 0$, we have $\sigma_{\text{spread}} \in [0, 1)$.

Composite Depth. Weighted combination:

$$\mathcal{D} = 0.4 \cdot r_{\text{active}} + 0.3 \cdot \rho_{\text{hop}} + 0.3 \cdot \sigma_{\text{spread}} \tag{44}$$

Weights are chosen to emphasize active layer utilization (40%) while balancing transition density (30%) and magnitude distribution (30%).

Optimal Depth. For an s-step problem in an L-layer model, ideal depth is:

$$\mathcal{D}_{\text{opt}} = \min\left(\frac{s}{L}, 1\right) \tag{45}$$

This captures the intuition that simple problems ($s \ll L$) should use proportionally fewer layers, while very complex problems may require full depth.

Efficiency Score. Deviation from optimal:

$$\mathcal{E} = \frac{1}{1 + |\mathcal{D} - \mathcal{D}_{\text{opt}}|} \tag{46}$$

This penalizes both wasteful over-utilization and insufficient under-utilization relative to problem complexity.

A.5 Entropy Computation for Information Bottleneck

For a batch of N problems, we collect all layer-ℓ activations:

$$\mathcal{A}^{(\ell)} = \{\mathbf{h}_i^{(\ell)} : i = 1, \ldots, N\} \tag{47}$$

Flattening and Sampling. To manage memory, we flatten and sample:

$$\mathcal{V}^{(\ell)} = \{\text{flatten}(\mathbf{h}_i^{(\ell)}) : i = 1, \ldots, N\} \tag{48}$$

If $|\mathcal{V}^{(\ell)}| > 10^4$ elements, we randomly sample 10^4 values to prevent computational explosion.

Normalization. We normalize to $[0, 1]$:

$$v_{\text{norm}} = \frac{v - v_{\text{min}}}{v_{\text{max}} - v_{\text{min}} + \epsilon} \tag{49}$$

where $v_{\text{min}} = \min(\mathcal{V}^{(\ell)})$ and $v_{\text{max}} = \max(\mathcal{V}^{(\ell)})$.

Histogram Binning. We discretize into $B = 50$ bins:

$$\text{hist}_b = \left| \left\{ v \in \mathcal{V}^{(\ell)}_{\text{norm}} : \frac{b-1}{B} \leq v < \frac{b}{B} \right\} \right|, \quad b \in \{1, \ldots, B\} \tag{50}$$

Probability Distribution. Normalize histogram to probabilities:

$$p_b = \frac{\text{hist}_b}{\sum_{k=1}^{B} \text{hist}_k} \tag{51}$$

Remove zero bins to prevent $\log(0)$:

$$\mathcal{P} = \{p_b : p_b > 0\} \tag{52}$$

Shannon Entropy. Compute entropy in bits:

$$H^{(\ell)} = -\sum_{p \in \mathcal{P}} p \log_2(p + \epsilon) \tag{53}$$

The constant $\epsilon = 10^{-10}$ ensures numerical stability.

Normalized Entropy. Maximum possible entropy for B bins is $\log_2(B)$. Normalized entropy:

$$\hat{H}^{(\ell)} = \frac{H^{(\ell)}}{\log_2(B)} \tag{54}$$

This ensures $\hat{H}^{(\ell)} \in [0, 1]$, facilitating cross-layer comparison.

Bottleneck Identification. Layers with entropy below the 25th percentile:

$$\mathcal{B} = \left\{ \ell : \hat{H}^{(\ell)} < \text{percentile}_{25}(\{\hat{H}^{(k)}\}_{k=1}^{L}) \right\} \tag{55}$$

These layers exhibit maximal information compression, potentially corresponding to critical reasoning bottlenecks.

A.6 Trajectory Similarity for Compression Hypothesis

For inference mode m (implicit, explicit, or concise) and problem q, the magnitude trajectory is:

$$\mathbf{T}_m(q) = (m_m^{(1)}, m_m^{(2)}, \ldots, m_m^{(L)}) \in \mathbb{R}^L \tag{56}$$

where $m_m^{(\ell)}$ is the layer-ℓ magnitude under mode m.

Cosine Similarity. For modes i and j:

$$\text{sim}_{\text{traj}}(q, i, j) = \frac{\sum_{\ell=1}^{L} m_i^{(\ell)} \cdot m_j^{(\ell)}}{\sqrt{\sum_{\ell=1}^{L} (m_i^{(\ell)})^2} \cdot \sqrt{\sum_{\ell=1}^{L} (m_j^{(\ell)})^2}} \tag{57}$$

This quantifies how similar the layer-wise computational patterns are between two inference modes.

Support Rate. The compression hypothesis is supported if implicit reasoning trajectories closely match concise CoT:

$$\text{SR} = \frac{1}{|\mathcal{P}|} \sum_{q \in \mathcal{P}} \mathbb{I}[\text{sim}_{\text{traj}}(q, \text{impl}, \text{conc}) \geq 0.7] \tag{58}$$

We set the threshold at 0.7 based on pilot studies. If SR ≥ 0.75, we conclude latent reasoning is likely compressed CoT. If SR < 0.5, we reject the hypothesis in favor of novel computational strategies.

Average Trajectory Similarity. For reporting:

$$\bar{s}_{\text{traj}} = \frac{1}{|\mathcal{P}|} \sum_{q \in \mathcal{P}} \text{sim}_{\text{traj}}(q, \text{impl}, \text{conc}) \tag{59}$$

A.7 Statistical Testing Procedures

Pearson Correlation. For metric X and binary correctness Y:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{60}$$

Under the null hypothesis $H_0 : \rho = 0$, the test statistic:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{61}$$

follows a Student's t-distribution with $N - 2$ degrees of freedom.

Spearman Correlation. Convert data to ranks R_X and R_Y, then:

$$\rho = 1 - \frac{6\sum_{i=1}^{N}(R_{X,i} - R_{Y,i})^2}{N(N^2-1)} \tag{62}$$

Test statistic is computed identically to Pearson case using ρ instead of r.

Bootstrap Confidence Intervals. For estimating distribution of metric θ:

Algorithm 2 Bootstrap Resampling

Require: Dataset $\mathcal{D}$, metric function f, iterations B
Ensure: Confidence interval $[\theta_{\text{lo}}, \theta_{\text{hi}}]$
1: **for** $b = 1$ to B **do**
2: Sample $\mathcal{D}_b$ with replacement from $\mathcal{D}$
3: Compute $\theta_b = f(\mathcal{D}_b)$
4: **end for**
5: Sort $\{\theta_b\}_{b=1}^{B}$
6: $\theta_{\text{lo}} = \text{percentile}_{2.5}(\{\theta_b\})$
7: $\theta_{\text{hi}} = \text{percentile}_{97.5}(\{\theta_b\})$
8: **return** $[\theta_{\text{lo}}, \theta_{\text{hi}}]$

This provides 95% confidence intervals without parametric assumptions.

B Discussion

B.1 Theoretical Implications

Our results challenge several assumptions about latent reasoning while revealing fundamental properties of implicit computation in large language models.

The Faithfulness Paradox. The negative correlation between our fidelity metric and correctness ($r = -0.27$, $p < 0.0001$) is perhaps our most provocative finding. This suggests one of three possibilities: (1) the metric captures computational overhead that hinders rather than helps performance, (2) the model achieves correctness through shallow heuristics that violate our faithfulness criteria, or (3) our metric design requires fundamental reconceptualization.

The dominance of "lucky guesses" (61% of correct answers with low stability) supports interpretation (2). The model may employ fast, brittle strategies for simple pattern matching, reserving deeper, more stable reasoning for problems where heuristics fail—thereby creating an inverted relationship between reasoning quality and success.

Latent vs. Explicit Computation. The near-identity of reasoning depth between implicit and explicit CoT modes ($\Delta = 0.01$), despite 10pp accuracy difference, demonstrates that verbalization's benefit is not computational deepening but rather *alignment*. Explicit step-by-step generation may serve as scaffolding that guides the model's existing latent reasoning toward problem-relevant computations, similar to how human verbalization aids in organizing pre-existing knowledge.

Rejection of Compression Hypothesis. Our finding that only 20% of latent reasoning trajectories resemble compressed CoT has significant implications for interpretability research. Techniques developed for analyzing explicit reasoning—attention flow analysis, token attribution, step-wise validation—may not transfer to latent architectures. The field requires new tools specifically designed for activation-space computation.

The Middle-Late Layer Dichotomy. The contradiction between late-layer activation dominance and middle-layer causal importance suggests a two-stage computational model: critical reasoning operations occur in middle layers, while late layers amplify and refine these computations for output generation. This aligns with circuit discovery findings showing task-specific "heads" in middle layers and output formatting in final layers.

Scale and Sophistication Disconnect. The identical performance of 7B and 1.5B models despite substantial differences in reasoning depth (7.2% deeper) and entropy (88% lower in 7B) raises important questions about scaling laws. If larger models develop more sophisticated internal reasoning without accuracy gains, this suggests: (a) benchmarks saturate before model capacity, (b) sophisticated reasoning provides limited advantage on current tasks, or (c) deeper computation enables better generalization not captured by in-distribution accuracy.

B.2 Deployment Risks and Safety Implications

Our findings reveal **systemic unreliability** that makes current models unsuitable for high-stakes applications without additional safeguards. The 61% lucky-guess rate—where correct answers emerge from computationally inconsistent pathways—creates three deployment crises:

1. **Brittleness Under Distribution Shift:** Models relying on shallow heuristics will fail catastrophically when encountering slightly harder or reformulated problems. A student receiving automated tutoring may get correct answers on practice problems but fail when exam questions require genuine reasoning.
2. **Unpredictable Production Behavior:** Low cross-run stability ($\mathcal{S} = 0.60 \pm 0.20$) means the same query may yield different reasoning paths—and potentially different answers—across inference runs. This violates basic reproducibility requirements for production systems.
3. **Evaluation-Deployment Mismatch:** Single-sample accuracy metrics provide **false confidence**. A model achieving 61% on benchmarks via lucky guesses may perform far worse when users naturally rephrase questions or when reasoning shortcuts no longer apply.

Recommendations for Safer Deployment:

- **Multi-run consistency checks**: Require $\geq$3 independent samples with agreement before deployment
- **Stability thresholds**: Flag predictions with $\mathcal{S} < 0.65$ for human review
- **Benchmark reform**: Replace single-accuracy metrics with stability-weighted scores
- **Transparent uncertainty**: Surface confidence indicators to end users

Table 4: Latent Reasoning Faithfulness Analysis Results on GSM8K

Metric	**Mean**	**Std**
Accuracy	0.6100	0.4877
Reasoning Depth	0.5139	0.0115
Activation Entropy	0.0895	0.0406
Stability Score	0.6000	0.2000
Alignment Score	0.6869	0.1387
Efficiency Score	0.7369	0.0295
Overall Fidelity	0.6715	0.0915

Table 5: Failure Mode Distribution (Safety Analysis)

Mode	**Count**	**Percentage**	**Risk**
True Positive	0	0.0%	Low
True Negative	195	39.0%	Expected
Silent Failure	0	0.0%	**High**
Lucky Guess	305	61.0%	Medium
Silent Failure Rate: 0.0%			
Safety Score: 100.0%			

Table 6: Ablation Study on Faithfulness Components

Configuration	**Avg Fidelity**	**Correlation**
Full	0.6415	−0.3146
No Stability	0.7177	−0.3146
No Alignment	0.6106	−0.3144
No Efficiency	0.5995	−0.3110