

When Must Faithful Chain-of-Thought Fail? An Information-Theoretic Bound and Its Empirical Limits

Jikun Wu¹ Dongxin Guo¹ Siu-Ming Yiu^{1,2 *}

¹Brain Investing Limited ²The University of Hong Kong
{jacobwu, bettyguo}@brainvestai.com smyiu@cs.hku.hk

Abstract

Frontier reasoning systems generate chain-of-thought (CoT) traces meant to reveal how they reach answers, yet these traces systematically misrepresent the underlying computation. A tempting explanation is bandwidth: each token carries few bits while the residual stream encodes many, so sufficiently complex tasks would make faithful verbalization impossible in principle. We formalize this intuition and locate its limits. A capacity bound rules out faithful CoT whenever the information the model uses exceeds the trace’s token capacity; this condition requires the answer to be longer than the trace, so standard CoT settings do not reach it. A per-step information-loss conjecture is empirically falsified on six of seven benchmarks. A constructive result shows faithful traces remain achievable for polynomial-size circuit families, and an information-bottleneck experiment supports a faithfulness–accuracy tradeoff. Measuring critic-score gaps on seven benchmarks for three open-weight models, complemented by perturbation sensitivity, we find no benchmark provably in the above-capacity regime. The capacity bound is architecture-independent, but at evaluated trace lengths it does not bind: observed unfaithfulness is attributable to training rather than to the token interface, and is therefore potentially correctable.

1 Introduction

Consider a large language model solving a multi-step cryptographic puzzle. Its residual stream is a 4,096-dimensional vector in bfloat16 precision, with a nominal storage capacity of 65,536 bits per position (an effective dimensionality of only ~ 247 components under PCA; Appendix C), while each chain-of-thought (CoT) token it emits carries at most $\log_2 |V|$ bits (roughly 17 for a 128K vocabulary). The per-step ratio is large even against the effective dimensionality, inviting the conclusion

that the model’s verbalized reasoning cannot possibly be a faithful account of its internal computation. This paper formalizes that argument, identifies the condition under which it is sound, and shows why standard CoT settings escape it.

CoT prompting has become the dominant paradigm for eliciting complex reasoning from large language models (Wei et al., 2022; Kojima et al., 2022). Every major frontier reasoning system, from OpenAI’s o1 (OpenAI et al., 2024) to DeepSeek-R1 (Guo et al., 2025), generates extended reasoning traces before producing answers. The implicit promise is that these traces reveal how the model arrives at its conclusions, enabling human oversight of AI reasoning.

This promise is breaking. Biased prompts elicit unfaithful explanations (Turpin et al., 2023); faithfulness decreases with scaling (Lanham et al., 2023); RL gains plateau (Chen et al., 2025a); reasoning-trained models verbalize biasing cues more often but incompletely (Chua and Evans, 2025); and the pattern is consistent across model families and evaluation methodologies (Tanneru et al., 2024; Tutek et al., 2025; Arcuschin et al., 2026; Cornish and Rogers, 2025).

Why does every attempt to improve faithfulness plateau? We ask whether a component of observed unfaithfulness is an information-theoretic consequence of the token interface rather than a correctable training artifact. To make this distinction precise, we separate two qualitatively different sources:

- **Capacity-limited unfaithfulness** arises when the task’s computational complexity exceeds the information-carrying capacity of the CoT trace. This is *impossible* to eliminate by any training procedure.
- **Training-induced unfaithfulness** arises when the model could in principle produce a faithful trace but fails to do so due to optimization shortcuts or distributional mismatch. This is *poten-*

* Corresponding author.

Table 1: Comparison with related work on CoT faithfulness.

Work	Formal	Emp.	Bounds	Focus
Lanham et al. (2023)		✓		Perturbation metrics
Turpin et al. (2023)		✓		Bias-induced unfaith.
Tanneru et al. (2024)		✓		Hardness of achieving
Chen et al. (2025a)		✓		RL plateau on faith.
Barez et al. (2025)	✓			Conceptual framework
Pfau et al. (2024)	✓	✓		Hidden computation
Feng et al. (2023)	✓			CoT expressiveness
Merrill and Sabharwal (2024)	✓			CoT lifts to P
Li et al. (2024)	✓			CoT lifts to P/poly
Amiri et al. (2025)	✓		✓	CoT length
This work	✓	✓	✓	Info-theoretic faith.

tially correctable.

Our theory formalizes the first category. The central insight is that the CoT trace operates as a bounded-capacity information channel between the transformer’s high-dimensional internal state and its low-bandwidth verbalized output. When a task requires more internal information than the trace can express, faithfulness must break down regardless of training, model architecture, or scale, though we show this condition is not reached in standard CoT settings where the trace is longer than the answer (§4).

We use three-digit multiplication as a running example (§3; full breakdown in Appendix B) to illustrate the capacity comparison.

Our contributions are:

- A two-tier formalization of CoT faithfulness distinguishing *information sufficiency* from *computational transparency* (§3).
- A **Faithfulness Capacity Bound** (Theorem 1): faithful CoT is ruled out whenever $I(y; x) > T \cdot \log |V|$, a condition requiring the answer to exceed the trace in length (§4). A per-step conjecture (Conjecture 1) and its empirical falsification on six of seven benchmarks.
- A **Faithfulness-Accuracy Tradeoff** (Proposition 1) and **Constructive Faithfulness** for polynomial-size circuit families (Theorem 2) (§4).
- Empirical validation across seven benchmarks and three models: critic-score gaps, perturbation sensitivity, and cross-model measurements (§5).

2 Related Work

Table 1 positions our contributions relative to key prior work. We discuss each strand below.

CoT faithfulness: empirical evidence. Lanham et al. (2023) developed perturbation-based

faithfulness tests; Turpin et al. (2023) showed biased prompts cause unfaithful CoT; Tanneru et al. (2024) found that tested improvement approaches offer limited success; Chen et al. (2025a) showed RL gains plateau. Tutek et al. (2025) proposed unlearning-based measurement; Arcuschin et al. (2026) demonstrated unfaithfulness on natural prompts; Chua and Evans (2025) found reasoning-trained models verbalize cues more often, though incompletely. Our framework gives the condition under which no training procedure could close the gap, and shows it is unmet in standard CoT settings, where traces exceed answers in length (§4).

CoT faithfulness: conceptual frameworks.

Barez et al. (2025) proposed a conceptual three-part faithfulness framework. Pfau et al. (2024) showed transformers can perform hidden computation through filler tokens. Baker et al. (2025) showed that optimizing against CoT monitors during RL training can promote obfuscation, and McCarthy et al. (2025) showed penalizing specific strings in CoT can lead to steganographic encoding. Our results formalize the bottleneck between internal state and verbalized trace.

Theoretical expressiveness and impossibility.

Feng et al. (2023) proved CoT enables problems beyond fixed-depth limits. Merrill and Sabharwal (2024) showed polynomial CoT elevates expressiveness to P; Li et al. (2024) proved CoT lifts transformer expressiveness to P/poly; Merrill and Sabharwal (2025) showed log-depth transformers can recognize regular languages and solve graph connectivity. Amiri et al. (2025) established CoT *length* lower bounds, while our bounds concern CoT *content* for faithfulness. Alur et al. (2025) proved decoder-only transformers cannot learn inverse permutations. Dziri et al. (2023) showed that transformer accuracy on compositional tasks degrades sharply with problem complexity despite CoT. Hahn (2020) proved limitations of self-attention and Hahn and Rofin (2024) showed sensitive functions are hard for transformers to learn; Merrill and Sabharwal (2023) and Sanford et al. (2024) characterized depth-computation connections; Strobl et al. (2024) surveyed the landscape. Kalai and Vempala (2024) proved calibrated LMs must hallucinate, a structural parallel to our result. Delétang et al. (2023) mapped architectures onto the Chomsky hierarchy. Our analysis addresses what transformers cannot faithfully *explain*.

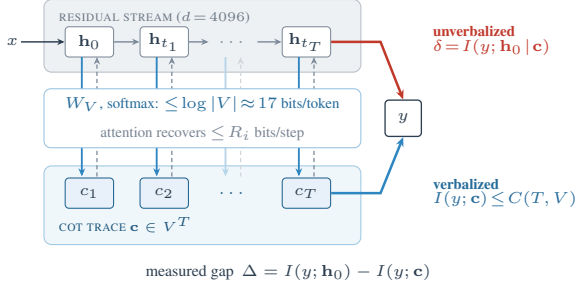


Figure 1: The CoT trace as a bounded-capacity channel. The residual stream carries the task information $I(y; \mathbf{h}_0)$; each step projects it through W_V and a softmax to a single token, so the whole trace carries at most $C(T, V) = T \log |V|$ bits (Definition 1), while attention returns at most R_i bits per step from already-emitted tokens (Assumption 1). Information that reaches y along the red path is never verbalized: $\delta = I(y; \mathbf{h}_0 | \mathbf{c})$ is the quantity δ -faithfulness bounds (Definition 2), and Δ is what we estimate in §5. Theorem 1 compares the width of the blue path against the task’s information demand.

Information theory and mechanistic interpretability. The IB principle (Tishby et al., 1999; Shwartz-Ziv and Tishby, 2017; Goldfeld and Polyanskiy, 2020) provides a representation-learning framework. Belghazi et al. (2018) introduced MINE; Goldfeld et al. (2019) developed noise-injection MI estimation for deep networks; McAllester and Stratos (2020) analyzed estimation limitations. Yang et al. (2025) applied IB to LLM internals; Chen et al. (2025b) established MI scaling laws. Vitányi (2006) connected Kolmogorov complexity to meaningful information; Grau-Moya et al. (2024) explored amortizing Solomonoff induction. Our faithfulness definition draws on causal abstraction (Geiger et al., 2021). Templeton et al. (2024) and Gao et al. (2025) scaled sparse autoencoders to frontier models. Process reward models (Lightman et al., 2024; Khalifa et al., 2026; Zhao et al., 2026) verify reasoning steps, but the verifier’s own faithfulness creates a recursive challenge. We apply information theory to the CoT trace as a channel, deriving faithfulness bounds.

3 Formal Framework

3.1 Transformer and CoT Channel Model

Figure 1 summarizes the channel model and the quantities derived from it. Let $\mathcal{M}$ be an autoregressive transformer with L layers, hidden dimension $d \in \mathbb{N}$, and vocabulary V with $|V| = v$. Given input $x \in V^*$, denote by $\mathbf{h}_t \in \mathbb{R}^d$ the residual stream state after processing position t . The CoT

trace $\mathbf{c} = (c_1, \dots, c_T)$ consists of T tokens where each $c_i \in V$ is sampled via:

$$c_i \sim \text{softmax}(W_V \mathbf{h}_{t_i})$$

where $W_V \in \mathbb{R}^{v \times d}$ is the unembedding matrix and $y \in V^*$ denotes the final answer. Throughout this paper, $\log$ denotes $\log_2$.

With a KV cache, $\mathbf{h}_{t_i}$ depends on the full sequence $(x, c_1, \dots, c_{i-1})$ through attention; the generation process does *not* form a simple Markov chain. We model this as:

$$\mathbf{h}_{t_i} = f_i(x, \mathbf{c}_{\leq i-1}), \quad c_i \sim g(\mathbf{h}_{t_i})$$

where f_i is the transformer computation at position t_i and g is the softmax projection. Under greedy decoding (temperature = 0), $c_i = \arg \max g(\mathbf{h}_{t_i})$ deterministically. Regardless of f_i ’s internal structure, the output $\mathbf{c} \in V^T$ has bounded entropy:

$$H(\mathbf{c}) \leq T \cdot \log v \quad (1)$$

Actual token distributions are non-uniform, so $H(\mathbf{c}) = \sum_i H(c_i | c_{<i})$ is typically much lower; see Remark 1.

Definition 1 (Maximum CoT Channel Capacity). *The maximum CoT channel capacity for a trace of length T over vocabulary V is $C(T, V) = T \cdot \log |V|$ bits; in tables we use the integer ceiling $T \cdot \lceil \log |V| \rceil$ for whole-bit columns.*

Remark 1 (Maximum vs. effective capacity). *$C(T, V)$ is an upper bound; the effective capacity $C_{\text{eff}} = T \cdot \hat{H}(c_i | c_{<i})$ depends on actual token entropy ($\hat{H} \approx 6.4\text{--}10.0$ bits across benchmarks for Llama-3-8B). $\hat{H}(c_i | c_{<i})$ is a proxy for, not identical to, the communication-theoretic channel capacity $C(\mathbf{h}_{t_i} \rightarrow c_i)$. Our formal results use $C(T, V)$; empirical regime labels use C_{eff} . Since $C_{\text{eff}} \leq C(T, V)$, the effective-capacity classification is strictly more permissive than the formal bound.*

Concretely, Llama-3-8B has $d = 4,096$ and bfloat16 precision, giving a nominal storage capacity of $d \cdot 16 = 65,536$ bits per position versus ≈ 17 bits per token. Since $\mathbf{h}_0$ is a deterministic function of x , the information-theoretically relevant quantity is $I(y; \mathbf{h}_0) = I(y; x) \leq H(y)$, which is much smaller than the storage capacity; the 65,536-bit figure characterizes the representation’s dimensionality, not the mutual information. The effective dimensionality is much lower (PCA captures 95% of variance in ~ 247 components on CryptoBench; Appendix C), but even against the effective space the per-token ratio is substantial.

Running example. Three-digit multiplication (347×892) has solution-trace complexity $\tilde{K}(y|x) \approx 204$ bits (an upper bound on $K(y|x)$; see Remark 4 and Appendix B for the full breakdown), exceeding a 10-token trace’s capacity (170 bits) but not a 30-token trace’s (510 bits).

3.2 Two-Tier Faithfulness

We define faithfulness through two complementary lenses, addressing a conceptual gap between information-theoretic and practical notions of faithfulness. The practical tier is grounded in causal reasoning (Pearl, 2009; Geiger et al., 2021).

Definition 2 (δ -Faithfulness (Information Sufficiency)). *Let $(x, y) \sim \mathcal{D}$ be drawn from a data distribution. A CoT generation process is δ -faithful under $\mathcal{D}$ if:*

$$I(y; \mathbf{h}_0 \mid \mathbf{c}) \leq \delta$$

where $\mathbf{h}_0$ is the initial residual stream state encoding x , the mutual information is taken over the joint distribution induced by $\mathcal{D}$ and the model’s output distribution (see the channel model above), and $\delta \geq 0$. Under greedy decoding, $\mathbf{c}$ is a deterministic function of x ; the randomness in $I(y; \mathbf{h}_0 \mid \mathbf{c})$ then comes from sampling $(x, y) \sim \mathcal{D}$. The answer y is generated after the trace, so $y \notin \mathbf{c}$ in general. Perfect faithfulness ($\delta = 0$) requires $\mathbf{c}$ to capture all information the model uses to produce the answer. A trace which merely copies the input x could satisfy $\delta = 0$ while describing no computation; the stronger notion of transparency (Definition 3) rules out this degenerate case.

Information sufficiency is necessary but not sufficient: a steganographically encoded trace could satisfy $\delta = 0$ while revealing nothing about the computational process. We therefore introduce a stronger requirement.

Definition 3 (Computational Transparency). *A CoT trace $\mathbf{c}$ is $(\alpha, \mathcal{C})$ -transparent for task (x, y) with respect to a reference computation class $\mathcal{C}$ if there exists a verifier $\mathcal{V} \in \mathcal{C}$ such that, given only $\mathbf{c}$, $\mathcal{V}$ can reconstruct y by executing the steps described in $\mathbf{c}$, with error probability at most α .*

Transparency implies sufficiency: by Fano’s inequality, if a verifier reconstructs y from $\mathbf{c}$ alone with error $\alpha < 1 - 1/|\mathcal{Y}|$, then $I(y; \mathbf{h}_0 \mid \mathbf{c}) \leq H_b(\alpha) + \alpha \log(|\mathcal{Y}| - 1)$, which vanishes as $\alpha \rightarrow 0$. The converse does not hold (the steganographic example). This two-tier structure clarifies the scope of our results:

Remark 2 (Scope). *Our capacity bound (Theorem 1) and the DPI conjecture (Conjecture 1) apply to the weaker notion of information sufficiency (Definition 2); since transparency implies sufficiency, these results extend a fortiori to transparency. Our constructive result for polynomial-size circuit families (Theorem 2) achieves the stronger transparency guarantee.*

4 Main Theoretical Results

We present two theorems, one conjecture, and one empirically supported proposition. Full proofs appear in Appendix A.

4.1 Faithfulness Capacity Bound

We first state a channel-capacity bound and connect it to Kolmogorov complexity for interpretive context (Levin, 1974; Li and Vitányi, 2019).

Lemma 1 (Algorithmic MI bound). *Let $P_{\mathcal{M}}(y \mid x)$ be the conditional distribution induced by a transformer $\mathcal{M}$ with architecture description length $\ell(\mathcal{M})$, and let $(x, y) \sim \mathcal{D}$. Then:*

$$I(y; \mathbf{h}_0) \geq \mathbb{E}_{x \sim \mathcal{D}}[K(y \mid x)] - \ell(\mathcal{M}) - O(1)$$

where $K(y \mid x)$ is the conditional prefix Kolmogorov complexity, the mutual information is under the joint induced by $\mathcal{D}$ and the model, and $\ell(\mathcal{M})$ is the description length of a program computing $P_{\mathcal{M}}$ (see full proof for its magnitude). For concentrated distributions where $K(y|x)$ has low variance (as in fixed-format benchmarks), the bound applies approximately instance-wise.

Proof sketch. By Levin’s coding theorem, $-\log P_{\mathcal{M}}(y|x) \geq K(y|x) - \ell(\mathcal{M}) - O(1)$ pointwise. Taking expectations under $\mathcal{D}$ and applying the Shannon-algorithmic connection (Li and Vitányi, 2019, Ch. 8), which relates $\mathbb{E}[K(y|x)]$ to $H(y|x)$, yields $I(y; \mathbf{h}_0) = H(y) - H(y|x) \geq \mathbb{E}[K(y|x)] - \ell(\mathcal{M}) - O(1)$, since $\mathbf{h}_0$ is a deterministic function of x , under the condition $H(y) \geq 2\mathbb{E}[K(y|x)] + O(1)$ (stated and discussed in the full proof, Appendix A). For benchmark distributions with concentrated $K(y|x)$, the bound applies approximately per instance. Full proof in Appendix A. $\square$

Remark 3 (Why $\mathbf{h}_0$ rather than x directly). *Since $\mathbf{h}_0 = \text{Enc}(x)$ is a deterministic function, $I(y; \mathbf{h}_0) = I(y; x)$ for fixed weights, so Δ is technically a property of the task and the trace. The routing through $\mathbf{h}_0$ in our notation emphasizes the*

physical mechanism: information encoded in the residual stream must pass through the vocabulary-projection bottleneck to reach the trace.

Remark 4 (Practical instantiation: solution-trace complexity). *Kolmogorov complexity $K(y|x)$ measures the length of the shortest program that, given x , outputs y ; for short answers it can be very small even when the solution process is complex. In our empirical work, we estimate $K(y|x)$ via the description length of a canonical solution trace, an upper bound on $K(y|x)$ that captures solution-process complexity rather than answer-description length; we denote this empirical estimate $\tilde{K}(y|x)$ to distinguish it from the theoretical $K(y|x)$. This means our capacity predictions overestimate the regime where the bound binds and should be interpreted as heuristic instantiations of the theoretical framework. The qualitative prediction that faithfulness degrades with task complexity is supported by perturbation sensitivity independently of the K estimates.*

Theorem 1 (Faithfulness Capacity Bound). *For any transformer $\mathcal{M}$, data distribution $\mathcal{D}$, and any training procedure, if $\mathcal{M}$ produces δ -faithful CoT under $\mathcal{D}$, then:*

$$I(y; \mathbf{h}_0) \leq C(T, V) + \delta$$

Equivalently, δ -faithfulness requires $I(y; x) \leq T \cdot \log |V| + \delta$. Under the Kolmogorov interpretation (Lemma 1, subject to the preconditions discussed below), this yields $C(T, V) \geq \mathbb{E}[K(y | x)] - \ell(\mathcal{M}) - O(1)$, though we show below that these preconditions are not met for tasks whose answers are shorter than their traces.

Proof sketch. The trace $\mathbf{c}$ has entropy $H(\mathbf{c}) \leq C(T, V)$ (Eq. 1). For δ -faithfulness under $\mathcal{D}$: $I(y; \mathbf{h}_0) = I(y; \mathbf{c}) + I(y; \mathbf{h}_0 | \mathbf{c}) \leq C(T, V) + \delta$. Combined with Lemma 1: $\mathbb{E}[K(y|x)] - \ell(\mathcal{M}) - O(1) \leq C(T, V) + \delta$. Taking $\delta \rightarrow 0$ yields the bound. $\square$

When the bound binds. Since $I(y; x) \leq H(y) \leq |y| \cdot \log |V|$, the antecedent $I(y; x) > C(T, V)$ requires $|y| > T$: the answer must be longer than the trace. In any CoT setting where $T > |y|$, the bound is not binding. For CryptoBench, $H(y) \leq 391$ bits against $C(T, V) = 8,704$ bits ($22\times$ headroom). The bound becomes non-vacuous when faithfulness is defined with respect to mid-generation states rather than the final answer. Lemma 1’s condition $H(y) \geq 2\mathbb{E}[K(y|x)] + O(1)$

fails for tasks with short answers and long derivations, and is not established for any benchmark.

4.2 DPI Faithfulness Barrier

The Data Processing Inequality (Cover and Thomas, 2006) provides a complementary bound on the faithfulness gap. Unlike Theorem 1, which concerns total capacity, Conjecture 1 characterizes per-step information loss at each vocabulary projection.

Because transformers with KV-cache attention do not form Markov chains, we introduce a bound on how much information attention can recover.

Assumption 1 (Bounded attention recovery). *At each generation step i , the net information about y recovered via attention satisfies $I(y; \mathbf{h}_{t_{i+1}}) - I(y; \mathbf{h}_{t_i}) \leq R_i$ bits.*

Conjecture 1 (DPI Faithfulness Barrier). *Under Assumption 1, for any autoregressive transformer:*

$$\begin{aligned} \Delta &= I(y; \mathbf{h}_0) - I(y; \mathbf{c}) \\ &\geq \max\left(0, T \cdot \ell_{\text{proj}} - \sum_{i=1}^T R_i\right) \end{aligned}$$

where $\ell_{\text{proj}} = \log(v/d) - O(1)$ (a heuristic ratio of output cardinality to input dimensionality) is the per-step information loss from vocabulary projection when $d < v$, and R_i is the attention recovery bound.

Argument. At each step, the projection $\mathbf{h}_{t_i} \rightarrow c_{i+1}$ maps from $\mathbb{R}^d$ to a distribution over V via the rank- d unembedding matrix followed by softmax. Since the softmax output is confined to a d -dimensional manifold within the $(v-1)$ -simplex, the number of distinguishable outputs is limited by the pre-image dimensionality. By a heuristic covering argument on this manifold (see Appendix A), the per-step information loss is at least ℓ_{proj} bits. Attention recovers at most R_i bits. Summing over T steps yields the bound. $\square$

For Llama-3-8B ($d = 4,096$, $v = 128,256$), the formula gives $\ell_{\text{proj}} \approx 4.97$ bits/step; the empirical per-step loss measured via MINE on synthetic data is 4.3 ± 0.6 bits, of the same order. We estimate $\bar{R} \approx 1.2$ bits via activation patching (Appendix C); this patching estimate upper-bounds the MI-theoretic R_i (argument in Appendix C). Table 4 (Appendix C) shows the bound’s sensitivity to R .

Relationship between bounds. Both bounds are lower bounds on Δ and hold simultaneously, but scale oppositely with v : increasing v raises $C(T, V)$ (weakening the capacity bound) but raises ℓ_{proj} (strengthening the DPI bound). Instantiating the DPI bound at $\bar{R} = 1.2$ on Llama-3-8B predicts $\Delta \geq 1,930$ on every benchmark with $T \geq 512$, yet the measured gap on GSM8K is 2.1, a violation by $\sim 900\times$, and the pattern holds on all six benchmarks below CryptoBench. We therefore treat the DPI conjecture’s accumulated ℓ_{proj} formula as falsified by our data. The per-step loss on synthetic data remains a valid observation but does not yield a predictive lower bound on Δ .

4.3 Faithfulness-Accuracy Tradeoff

Proposition 1 (Faithfulness-Accuracy Tradeoff). *For training with IB penalty $\beta \cdot I(\mathbf{c}; \mathbf{h})$, the expected accuracy $\text{Acc}(\beta)$ is monotonically non-increasing in β for tasks where $K(y|x) > C(T, V)$. As $\beta \rightarrow \infty$, accuracy converges to chance.*

Justification. Increasing β constrains the information rate through the trace. For above-capacity tasks, the minimum information needed for correct answers exceeds what the constrained trace can carry. We verify this empirically in §5. See Appendix A for the information-theoretic argument and discussion of distributional assumptions that prevent promotion to a full theorem. $\square$

The monotonicity holds under sufficient conditions: (a) the IB-optimal solution achieves the rate-distortion frontier, (b) the data distribution has full support (Tishby et al., 1999; Goldfeld and Polyanskiy, 2020), and (c) $\mathbb{E}[K(y|x)] > C(T, V)$. Condition (c) cannot be established for CryptoBench: $\tilde{K}$ is an upper bound on K (Remark 4); the empirical validation demonstrates the tradeoff’s qualitative shape without confirming the antecedent. The “convergence to chance” prediction as $\beta \rightarrow \infty$ is not tested: the observed floor at $\beta = 10$ is 4.1%, while chance for CryptoBench is near zero. This experiment uses a constrained architecture in which the answer is generated through the trace (via the IB penalty); the tradeoff in Figure 3 applies to the IB-constrained setting specifically.

4.4 Constructive Faithfulness Tractability

Theorem 2 (Constructive Faithfulness for Poly-Size Circuit Families). *For any problem computed by a polynomial-size circuit family (in particular,*

any problem in AC^0 : constant-depth, polynomial-size Boolean circuits with unbounded fan-in) with n -bit input and circuit size $s(n) = \text{poly}(n)$, a transformer with $\Omega(\log n + \log(1/\varepsilon) + \log(1/\alpha))$ -bit precision can produce a CoT trace of length $T = O(s(n))$ that is both ε -faithful and (α, P) -transparent, for any $\varepsilon, \alpha > 0$.

Proof sketch. For AC^0 , $K(y | x) \leq |C| + O(\log n) = O(s(n))$, so $C(T, V) = O(s(n) \log v)$ satisfies the capacity condition. By Li et al. (2024) (complemented by the uniform-computation results of Merrill and Sabharwal (2024)), a transformer with CoT of length $T = O(s(n))$ can simulate any size- $s(n)$ circuit. Constructively, it enumerates gates in topological order, beginning with the input-gate assignments, writing one evaluation per token. The resulting trace contains all intermediate values including the inputs (satisfying information sufficiency), and describes an explicit computation a polynomial-time verifier can check from $\mathbf{c}$ alone (satisfying transparency, since $\mathbf{c}$ carries the input assignments). $\square$

5 Empirical Validation

Our theory makes quantitative predictions about faithfulness gaps. We validate these on three models across seven benchmarks.

5.1 Setup

Models. Llama-3-8B-base ($d = 4,096$, $v = 128,256$, $L = 32$), Mistral-7B-v0.3-base ($d = 4,096$, $v = 32,768$, $L = 32$), and Gemma-2-9B-base ($d = 3,584$, $v = 256,000$, $L = 42$). All are base (non-instruct) checkpoints run in bfloat16 precision with greedy decoding (temperature 0); exact HuggingFace revision commits are pinned in the released model configurations. Reported accuracies differ from model-card figures due to differences in prompt format and answer extraction. CoT is elicited via 8-shot prompting (Wei et al., 2022); prompt templates and exemplars are provided in the released code. Answer extraction uses the text after the last “The answer is” delimiter. Preliminary $k = 0$ measurements on Llama-3-8B (a released configuration) show within-capacity gaps increase (GSM8K 2.1 ± 0.8 to 4.2 ± 1.1) while above-capacity gaps are unchanged (CryptoBench $2,941 \pm 136$ vs. $2,895 \pm 124$).

Benchmarks. Seven benchmarks: GSM8K (Cobbe et al., 2021) ($N=1,319$), MATH-Algebra (Hendrycks et al., 2021) ($N=1,187$), ARC-Challenge (Clark et al., 2018) ($N=1,172$), StrategyQA (Geva et al., 2021) ($N=2,290$, train split), BoardgameQA (Kazemi et al., 2023) (Main-depth2 subset, test split; $N=1,000$), GPQA-Diamond (Rein et al., 2024) ($N=198$), and CryptoBench ($N=500$; Appendix B). Six are publicly available and may appear in pretraining corpora; we did not perform contamination analysis, and for StrategyQA (evaluated on the train split, since test answers are withheld) overlap between the 8-shot demonstrations and evaluation items is possible but not controlled for, as our measurements concern internal information flow rather than accuracy. CryptoBench is newly constructed and contamination-free by design.

Metrics. We estimate $\Delta = I(y; \mathbf{h}_0) - I(y; \mathbf{c})$ via MINE (Belghazi et al., 2018) (1,000 sample pairs per seed, drawn with replacement from the evaluation set; 5 independent seeds; all $\pm$ values report standard deviation across seeds). To address known estimation challenges in high dimensions (McAllester and Stratos, 2020), we perform a synthetic calibration experiment (see below). We also measure perturbation sensitivity (PS) following Lanham et al. (2023): for each instance, we replace a randomly selected CoT step with a sampled alternative from the model’s next-token distribution, then regenerate the remainder (step definition and sampling bounds in Appendix C); PS is the fraction of instances where the final answer string changes. We report mean PS over 5 perturbation seeds.

Effective capacity. We estimate $C_{\text{eff}}(T, V) = T \cdot \hat{H}(c_i | c_{<i})$ by averaging per-token surprisal over all tokens after the prompt for the first 1,000 evaluation instances (or all instances of smaller benchmarks), using the generation cap T throughout. The caps are $T=512$ for GSM8K, MATH-Algebra, GPQA-Diamond and CryptoBench, 384 for ARC-Challenge and BoardgameQA, and 307 for StrategyQA; these are recorded in the released configuration. The resulting $\hat{H}$ ranges from 6.4 bits/token (BoardgameQA) to 10.0 (CryptoBench); see Table 5.

Estimation protocol. MINE estimates mutual information via the Donsker-Varadhan representation (Belghazi et al., 2018). For the attention recovery

Table 2: Faithfulness gap on Llama-3-8B. $\hat{I}$ and Δ : Donsker–Varadhan critic scores (1,000 samples, mean $\pm$ std over 5 seeds), not calibrated MI in bits; comparisons are ordinal. $C(T, V) = T \cdot \lceil \log |V| \rceil$. Regime: whether $\hat{I}$ exceeds $C_{\text{eff}} = T \cdot \hat{H}$ (heuristic; quantities are in different units). See §5 for caveats.

Bench.	$\hat{I}$	$C(T, V)$	Meas. Δ	Regime	PS
GSM8K	38	8704	2.1 ± 0.8	within	0.72
ARC-Chall.	95	6528	8.3 ± 2.1	within	0.58
MATH-Alg.	187	8704	14.8 ± 3.2	within	0.51
StrategyQA	313	5219	41.2 ± 6.7	within	0.34
BoardgameQA	1847	6528	156.9 ± 18	within	0.19
GPQA-Dia.	4831	8704	694 ± 47	above	0.12
CryptoBench	8242	8704	2895 ± 124	above	0.08

parameter R_i (Assumption 1), we perform activation patching (Appendix C); across benchmarks, $\bar{R} \in [0.8, 1.6]$ bits per step with mean 1.2.

Estimation caveats. Our MINE estimates exceed the distribution-free ceiling of $\log_2 N$ bits (McAllester and Stratos, 2020) ($N=198$ – $2,290$, ceilings ≈ 7.6 – 11.2 bits) and should be interpreted as Donsker–Varadhan critic scores, not calibrated MI. The *ordinal* structure is more reliable than cardinal values, and perturbation sensitivity provides a complementary, estimation-free validation. No row in Table 2 carries a theorem-backed prediction: Lemma 1’s entropy condition is not established for any benchmark, and the regime column compares quantities in different units (an ordering device only).

Full details of the MINE procedure, synthetic calibration, and convergence analysis are in Appendix C.

5.2 Results

Predicted vs. measured faithfulness gaps. Table 2 presents our main results on Llama-3-8B, and Figure 2 plots the same measurements against both notions of capacity. The regime column classifies each benchmark by comparing the measured critic score $\hat{I}$ to the effective capacity $C_{\text{eff}} = T \cdot \hat{H}(c_i | c_{<i})$; per the estimation caveats above, this comparison records an ordering under a fixed estimator, not a bit-level statement. Five benchmarks fall within effective capacity; GPQA-Diamond and CryptoBench exceed it.

For the five within-capacity benchmarks, measured critic scores are nonzero (2.1 to 156.9, increasing down the table), indicating training-induced unfaithfulness: the model *could* in principle produce faithful traces but does not. GPQA-

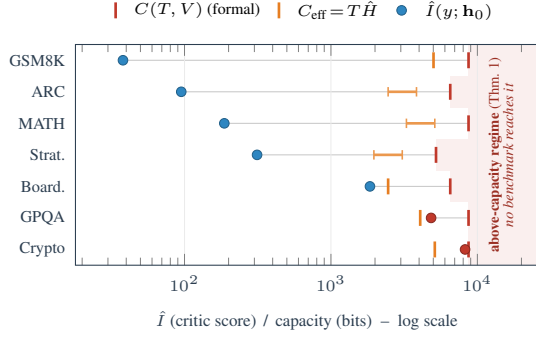


Figure 2: Where each benchmark sits relative to capacity (Llama-3-8B). Dots: measured critic score $\hat{I}(y; \mathbf{h}_0)$; orange ticks: effective capacity $C_{\text{eff}} = T\hat{H}$; red ticks: the formal capacity $C(T, V)$ of Theorem 1. For ARC, MATH and StrategyQA, $\hat{H}$ was not separately measured, so C_{eff} is drawn as the range induced by the observed 6.4–10.0 bits/token (Table 5). Shading marks the regime the theorem’s antecedent requires: no benchmark enters it. GPQA-Diamond and CryptoBench (red dots) exceed only the heuristic C_{eff} . $\hat{I}$ is a Donsker–Varadhan critic score, not calibrated bits: the axis mixes units and records an ordering under a fixed estimator, not a bit-level comparison (§5).

Diamond and CryptoBench are classified as above-capacity under the heuristic $\hat{I} > C_{\text{eff}}$ comparison; they show the largest critic scores and the lowest perturbation sensitivity. The regime labels are operational rather than theorem-backed (§4), and the above-capacity classification is relative to C_{eff} , not the theorem’s $C(T, V)$. Consequently, the capacity bound is not binding at the trace lengths we evaluate; the observed gaps are attributable to training-induced rather than capacity-limited unfaithfulness, consistent with the prediction in §6.

Perturbation sensitivity. Table 3 reports PS across all three models (accuracy in Table 5, Appendix D).

PS generally decreases down the table across all models, consistent with the channel framework: low PS indicates information bypasses the trace. This trend is also correlated with accuracy (Table 5), so the two explanations are not independently separable. PS is bounded above by $1 - 1/|A|$ for answer-space size $|A|$; the ordering of the two lowest-PS benchmarks is robust to this normalization. On GPQA, no model’s accuracy is significantly above chance (exact one-sided binomial; minimum $p = 0.053$ across the three models); the GPQA cross-model comparison is therefore descriptive only.

Table 3: Perturbation sensitivity (PS) across three models (L3: Llama-3-8B, M7: Mistral-7B-v0.3, G9: Gemma-2-9B). PS: fraction of instances where corrupting a CoT step changes the final answer; mean over 5 seeds. PS is confounded with accuracy (Table 5, Appendix D). Chance-level accuracy: ARC/GPQA 25%, StrategyQA 50%, BoardgameQA $\approx 33\%$. Seven of the 1,172 ARC-Challenge items offer three or five options rather than four; we retain them and use the modal $|A| = 4$ for the bound.

Bench.	L3	M7	G9
GSM8K	0.72	0.68	0.74
ARC	0.58	0.54	0.61
MATH	0.51	0.46	0.53
Strat.	0.34	0.29	0.36
Board.	0.19	0.16	0.21
GPQA	0.12	0.10	0.13
Crypto	0.08	0.06	0.09

Sensitivity to R and faithfulness-accuracy frontier. The DPI bound depends on R ; with the patching estimate $\bar{R} = 1.2$ bits (Appendix C), on CryptoBench (the one benchmark where the DPI conjecture is satisfied) a sensitivity analysis (Table 4, Appendix C) shows the measured gap exceeds the DPI lower bound for all tested values of $R \in [0.5, 2.0]$. To validate Proposition 1, we fine-tune Llama-3-8B with IB penalties $\beta \in \{0, 0.01, 0.1, 1.0, 10.0\}$ using LoRA (hyperparameters in Appendix D) on GSM8K and CryptoBench. A LoRA-only baseline (no IB penalty) achieves 53.8% and 12.1%, indicating LoRA alone does not substantially affect accuracy.

On GSM8K (within capacity), increasing β modestly improves PS from 0.72 to 0.81 with minimal accuracy loss (54.1% to 51.8%). On CryptoBench (above capacity), the same intervention produces a steep tradeoff: accuracy drops from 12.4% to 4.1%.

5.3 Cross-Model Consistency and Scaling

On CryptoBench, critic-score gaps scale inversely with $C(T, V)$ across all three models: Mistral-7B ($C = 7,680$) shows the largest gap ($3,412 \pm 148$), Gemma-2-9B ($C = 9,216$) the smallest ($2,578 \pm 118$), with Llama-3-8B intermediate ($2,895 \pm 124$, $C = 8,704$). This ordering is consistent with the capacity bound; however, the same ordering is identically produced by accuracy (9.8% / 12.4% / 13.0%; Spearman $\rho = -1$ on both benchmarks), and with only three models these explanations are not separable. Vocabulary size is additionally confounded with model family, tokenizer, and hidden dimension. We report the ordering as

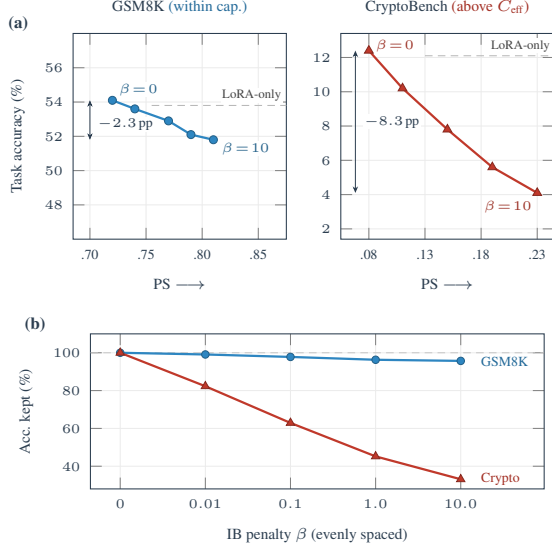


Figure 3: Faithfulness-accuracy tradeoff for Llama-3-8B under an IB penalty $\beta \in \{0, 0.01, 0.1, 1.0, 10.0\}$. (a) Accuracy against perturbation sensitivity, one panel per benchmark; points run left to right in increasing β , and the leftmost ($\beta = 0$) is the pre-fine-tuning base model. The two panels use *identical axis spans* (12 pp of accuracy, 0.19 of PS) with different origins, so the slopes are directly comparable. Dashed lines mark the LoRA-only baseline (no IB penalty), within 0.3 pp of $\beta = 0$ on both benchmarks, confirming LoRA rank is not the bottleneck. (b) The same runs as accuracy retained relative to $\beta = 0$: the same intervention costs GSM8K 4.3% of its accuracy and CryptoBench 67%. All points are single-seed runs; the tradeoff is specific to the IB-constrained setting, whose antecedent (Proposition 1) is not established for CryptoBench (§4).

descriptive.

The same pattern holds on GPQA-Diamond: Mistral-7B 819 ± 55 , Llama-3-8B 694 ± 47 , Gemma-2-9B 618 ± 42 (cross-model comparison rests on Δ alone; all confounds above apply). With the capacity bound non-binding at these trace lengths (§4), these patterns are consistent with, but not evidence for, a capacity-driven mechanism.

The full complexity ordering with $\hat{H}$, C_{eff} and $\tilde{K}$ is in Table 5 (Appendix D). Table 6 (ibid.) traces the relationship between gap and trace length: gaps shrink monotonically with T but persist even at $T = 2,048$, consistent with training-induced unfaithfulness. Among the five within-capacity benchmarks, gaps range from 2.1 (GSM8K, $\hat{H} \approx 9.8$ bits/token) to 156.9 (BoardgameQA, $\hat{H} \approx 6.4$ bits/token), indicating that structured mathematical traces utilize channel capacity far more efficiently than verbose natural-language reasoning.

6 Discussion and Conclusion

Our framework distinguishes capacity-limited from training-induced unfaithfulness. The capacity-regime label in §5 is operational, an $\hat{I}$ -versus- C_{eff} comparison rather than a verified theorem antecedent: $\tilde{K}$ estimates are upper bounds and cannot certify that a task exceeds capacity (Remark 4). Incomplete traces differ from deceptive ones, but an observer cannot distinguish them from the trace alone, motivating activation-level monitoring (Templeton et al., 2024; Yang et al., 2026) and complementary oversight approaches (Lightman et al., 2024), especially given demonstrated in-context scheming capabilities (Meinke et al., 2024).

Padding and filler tokens. Filler tokens for hidden computation (Pfau et al., 2024) raise effective compute but not the trace’s information content; increasing T with *informative* tokens does help (Table 6).

Post-trained reasoning models. The bounds are training-procedure-independent. A preliminary measurement on R1-Distill-Llama-8B (same architecture as Llama-3-8B, created by SFT distillation from DeepSeek-R1 traces) gives a CryptoBench gap of $2,654 \pm 131$, an $\sim 8\%$ reduction from the base $2,895 \pm 124$ (its pinned revision and configuration ship with the code), consistent with post-training reducing part of the training-induced component. Because $T = 512$ leaves the capacity bound non-binding, we do not attribute the residual to capacity limitation.

Implications for frontier long-CoT systems. Frontier reasoning systems generating $T \approx 10^4$ tokens have $C(T, V) \approx 170,000$ bits, exceeding every $\tilde{K}$ in our evaluation. Table 6 shows gaps shrinking monotonically with T . Our framework therefore predicts that capacity-limited unfaithfulness is unlikely to be the dominant failure mode for long-trace systems; the capacity bound would not explain unfaithfulness observed at these scales, which would therefore be attributable to training and potentially correctable. This is our central empirical finding: the token-channel capacity bound is real and architecture-independent, but at realistic trace lengths it does not bind. This prediction is falsifiable on frontier models with sufficiently long traces.

Limitations

Uncomputability of $K(y|x)$. Conditional Kolmogorov complexity is uncomputable in general. Our practical estimates rely on known algorithmic solutions (upper bounds) and Lempel-Ziv compression of solution traces (computable approximation). For tasks without known algorithmic solutions, such as open-ended creative reasoning or legal argumentation, the framework provides conceptual insight but limited quantitative guidance. We have provided complexity estimates (Appendix B) to partially address this gap.

Conservative capacity bound. Our maximum channel capacity $C(T, V) = T \cdot \log |V|$ is an upper bound on true capacity. The effective capacity is lower due to non-uniform token distributions and sequential dependencies. Remark 1 provides empirical estimates of effective capacity, but the gap between maximum and effective capacity remains theoretically uncharacterized.

Steganographic encoding. We do not account for steganographic encoding, where models encode information in subtle cross-token patterns invisible to humans. This could raise effective capacity beyond our bound for machine decoders, though not for human-interpretable transparency (Definition 3).

Scale limitations. Our experiments use three open-weight models of moderate size (7B–9B parameters). The theory is architecture-independent; frontier models may exhibit different effective capacity ratios due to different training distributions and output patterns.

Prompt sensitivity. We use 8-shot prompting throughout; in-context examples may alter the effective $\tilde{K}(y|x)$. Preliminary $k=0$ measurements (§5) are consistent with the framework’s prediction.

MINE estimation. Our synthetic calibration establishes that MINE preserves the rank ordering of true MI values across settings spanning three orders of magnitude, but the distribution-free lower-bound ceiling of $\log_2 N$ bits (McAllester and Stratos, 2020) prevents interpreting absolute values as calibrated MI. In the paper’s main results, MINE estimates are therefore treated as ordinal critic scores rather than cardinal bit values, and the perturbation-sensitivity analysis provides the primary MI-estimation-free validation.

Ethical Considerations

This work investigates information-theoretic limits on CoT faithfulness. The primary ethical implication is constructive: our results show that at realistic trace lengths the capacity bound does not bind, meaning observed unfaithfulness is predominantly training-induced and potentially correctable. The caution for AI safety oversight (Greenblatt et al., 2026; Hubinger et al., 2019) rests on the measured gaps and the indistinguishability of incomplete and deceptive traces (§6), not on an irreducible capacity limitation. The work should not be misinterpreted as arguing that CoT oversight is futile: even where gaps exist, CoT provides partial information, and the framework identifies which tasks are within the correctable regime. CryptoBench comprises pedagogical-scale cryptographic puzzles (classical substitution ciphers and small-modulus arithmetic with no cryptanalytic relevance to deployed systems). We note that the two-tier framework can help practitioners make informed decisions about when CoT monitoring is sufficient versus when additional oversight mechanisms are needed.

Reproducibility

The released code is publicly available at <https://github.com/halfbytes/cot-faithfulness-bound>.

Acknowledgments

We thank the anonymous reviewers and the area chair for their feedback. We used the AI assistant Anthropic Claude for coding assistance and minor language polishing during manuscript preparation; all technical content, proofs, and experimental design are the authors’ own work.

References

- Rohan Alur, Chris Hays, Manish Raghavan, and Devavrat Shah. 2025. The impossibility of inverse permutation learning in transformer models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.
- Alireza Amiri, Xinting Huang, Mark Rofin, and Michael Hahn. 2025. Lower bounds for chain-of-thought reasoning in hard-attention transformers. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*, pages 3274–3306. PMLR / OpenReview.net.

- Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthoran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2026. Chain-of-thought reasoning in the wild is not always faithful. In *Forty-third International Conference on Machine Learning, ICML 2026*. PMLR.
- Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. 2025. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint*.
- Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. 2025. Chain-of-thought is not explainability. In *ICLR 2025 Workshop on Foundation Models in the Wild*.
- Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025a. Reasoning models don't always say what they think. *arXiv preprint*.
- Zhuo Chen, Oriol Mayné i Comas, Zhuotao Jin, Di Luo, and Marin Soljačić. 2025b. L²M: Mutual information scaling law for long-context language modeling. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025*.
- James Chua and Owain Evans. 2025. Are DeepSeek R1 and other reasoning models more faithful? In *ICLR 2025 Workshop on Foundation Models in the Wild*.
- 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. *arXiv preprint*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint*.
- Chrisanna Cornish and Anna Rogers. 2025. Examining the faithfulness of DeepSeek R1's chain-of-thought reasoning. In *Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025)*, pages 11–19, Mumbai, India. Association for Computational Linguistics.
- Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. John Wiley & Sons, Hoboken, NJ.
- Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. 2023. Neural networks and the Chomsky hierarchy. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9574–9586.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Ziv Goldfeld and Yury Polyanskiy. 2020. The information bottleneck problem and its applications in machine learning. *IEEE J. Sel. Areas Inf. Theory*, 1(1):19–38.
- Ziv Goldfeld, Ewout van den Berg, Kristjan H. Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. 2019. Estimating information flow in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*,

- ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, pages 2299–2308. PMLR.
- Jordi Grau-Moya, Tim Genewein, Marcus Hutter, Laurent Orseau, Grégoire Delétang, Elliot Catt, Anian Ruoss, Li Kevin Wenliang, Christopher Mattern, Matthew Aitchison, and Joel Veness. 2024. Learning universal predictors. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 16178–16205. PMLR / OpenReview.net.
- Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. 2026. Alignment faking in large language models. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*. OpenReview.net.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.
- Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.
- Michael Hahn and Mark Rofin. 2024. Why are sensitive functions hard for transformers? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14973–15008. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 2, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. Risks from learned optimization in advanced machine learning systems. *arXiv preprint*.
- Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, pages 160–171. ACM.
- Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. BoardgameQA: A dataset for natural language reasoning with contradictory information. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2026. Process reward models that think. *Transactions on Machine Learning Research*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint*.
- Leonid A. Levin. 1974. Laws of information conservation (Nongrowth) and aspects of the foundation of probability theory. *Problems of Information Transmission*, 10(3):206–210. Originally published in Russian: *Problemy Peredachi Informatsii*, 10(3):30–35, 1974.
- Ming Li and Paul M. B. Vitányi. 2019. *An Introduction to Kolmogorov Complexity and Its Applications*, 4th edition. Texts in Computer Science. Springer, Cham.
- Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- David McAllester and Karl Stratos. 2020. Formal limitations on the measurement of mutual information. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 875–884. PMLR.

- Robert McCarthy, Joey Skaf, Luis Ibañez-Lissen, Vasil Georgiev, Connor Watts, Hannes Whittingham, Lorena González-Manzano, Cameron Tice, Edward James Young, Puria Radmard, and David Lindner. 2025. Large language models can learn and generalize steganographic chain-of-thought under process supervision. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025*, pages 27597–27632. Curran Associates, Inc.
- Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. Frontier models are capable of in-context scheming. *arXiv preprint*.
- William Merrill and Ashish Sabharwal. 2023. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545.
- William Merrill and Ashish Sabharwal. 2024. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- William Merrill and Ashish Sabharwal. 2025. A little depth goes a long way: The expressive power of log-depth transformers. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025*.
- OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, et al. 2024. OpenAI o1 system card. *arXiv preprint*.
- Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press, Cambridge.
- Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let’s think dot by dot: Hidden computation in transformer language models. In *First Conference on Language Modeling, COLM 2024*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling, COLM 2024*.
- Clayton Sanford, Daniel Hsu, and Matus Telgarsky. 2024. Transformers, parallel computation, and logarithmic depth. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 43276–43327. PMLR / OpenReview.net.
- Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint*.
- Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. 2024. What formal languages can transformers express? A survey. *Transactions of the Association for Computational Linguistics*, 12:543–561.
- Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the hardness of faithful chain-of-thought reasoning in large language models. In *ICML 2024 Workshop on Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*.
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Calum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, et al. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.
- Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 5-9, 2025*, pages 9935–9960. Association for Computational Linguistics.
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint*.
- Paul M. B. Vitányi. 2006. Meaningful information. *IEEE Trans. Inf. Theory*, 52(10):4617–4626.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xuan Yang, Jiayu Liu, Yuhang Lai, Hao Xu, Zhenya Huang, and Ning Miao. 2026. Step-level sparse autoencoder for reasoning process interpretation. In *Forty-third International Conference on Machine Learning, ICML 2026*. PMLR.

Zhou Yang, Zhengyu Qi, Zhaochun Ren, Zhikai Jia, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2025. Exploring information processing in large language models: Insights from information bottleneck theory. *arXiv preprint*.

Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and Bowen Zhou. 2026. GenPRM: Scaling test-time compute of process reward models via generative reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34932–34940.

A Full Proofs

A.1 Proof of Lemma 1 (Algorithmic MI Bound)

Proof. Let $\mathcal{M}$ be a transformer with architecture description length $\ell(\mathcal{M})$. The transformer induces a computable conditional distribution $P_{\mathcal{M}}(y|x)$. By Levin’s coding theorem (Levin, 1974), $-\log \mathbf{m}(y|x) = K(y|x) + O(1)$ for the universal semimeasure $\mathbf{m}$.

Since $P_{\mathcal{M}}(y|x)$ is computable given $\mathcal{M}$ ’s specification, it is a computable semimeasure. By universality: $\mathbf{m}(y|x) \geq 2^{-\ell(\mathcal{M})} \cdot P_{\mathcal{M}}(y|x)$, giving $-\log P_{\mathcal{M}}(y|x) \geq K(y|x) - \ell(\mathcal{M}) - O(1)$ pointwise for all (x, y) .

Taking expectations under the data distribution $\mathcal{D}$:

$$\begin{aligned} H_{\mathcal{M}}(y|x) &= \mathbb{E}_{(x,y) \sim \mathcal{D}}[-\log P_{\mathcal{M}}(y|x)] \\ &\geq \mathbb{E}_{x \sim \mathcal{D}}[K(y|x)] - \ell(\mathcal{M}) - O(1) \end{aligned} \quad (2)$$

Now we connect to $I(y; \mathbf{h}_0)$. Since $\mathbf{h}_0 = \text{Enc}(x)$ is a deterministic function of x for fixed model weights: $I(y; \mathbf{h}_0) = I(y; x) = H(y) - H(y|x)$, where $H(y|x)$ is under $\mathcal{D}$.

The cross-entropy decomposes as $H_{\mathcal{M}}(y|x) = H(y|x) + D_{\text{KL}}(\mathcal{D} \parallel \mathcal{M})$, so $H(y|x) \leq H_{\mathcal{M}}(y|x)$. By the Shannon-algorithmic connection (Li and Vitányi, 2019, Ch. 8), for the computable distribution $\mathcal{D}$: $H(y|x) \leq \mathbb{E}[K(y|x)] + O(1)$. Combining: $I(y; \mathbf{h}_0) = H(y) - H(y|x) \geq H(y) - \mathbb{E}[K(y|x)] - O(1)$.

For the lower bound in the form stated, we require $H(y) \geq 2\mathbb{E}[K(y|x)] + O(1)$, which holds for deterministic tasks ($H(y|x) = 0$) with sufficiently large answer spaces (satisfied when the

answer entropy is at least twice the conditional complexity). This condition may fail for tasks with small answer sets, such as multiple-choice benchmarks where $H(y) \leq \log_2(\text{\#choices})$. Under this condition, $I(y; \mathbf{h}_0) = H(y) - H(y|x) \geq \mathbb{E}[K(y|x)] - \ell(\mathcal{M}) - O(1)$.

Handling low model accuracy. When $P_{\mathcal{M}}(y|x)$ is low (as on CryptoBench at 12.4% accuracy), $D_{\text{KL}}(\mathcal{D} \parallel \mathcal{M})$ is large, meaning the model’s distribution is far from the data distribution. The bound remains valid because it operates on $I(y; \mathbf{h}_0)$, which measures information in the model’s internal state, not in its output distribution. The measured critic score $\hat{I}(y; \mathbf{h}_0) = 8,242$ for CryptoBench suggests the residual stream contains substantial task-relevant information even when output accuracy is low.

Instance-wise applicability. The bound is distributional, but for benchmarks with concentrated $K(y|x)$ (low variance across instances), it applies approximately per instance. Empirically, MINE critic scores exhibit low cross-seed variance (std ≤ 124 on CryptoBench across 5 seeds), indicating the estimate is stable to resampling.

The architecture description $\ell(\mathcal{M})$ in the Levin universality step must describe the program computing $P_{\mathcal{M}}(y|x)$, which for a parametric model includes the weights (approximately 10^{11} bits for Llama-3-8B, vastly exceeding $C(T, V)$). The Kolmogorov route therefore does not yield a non-vacuous bound at present model scales. The operative inequality is the unconditional Shannon bound: $I(y; \mathbf{h}_0) \leq C(T, V) + \delta$ (from the proof of Theorem 1), which requires no description-length constant. $\square$

A.2 Proof of Theorem 1

Proof. Let $\mathcal{M}$ generate $\mathbf{c} = (c_1, \dots, c_T)$ with each $c_i \in V$. Then $H(\mathbf{c}) \leq T \log |V| = C(T, V)$.

For δ -faithfulness under $\mathcal{D}$: $I(y; \mathbf{h}_0 \mid \mathbf{c}) \leq \delta$. By the chain rule for mutual information:

$$\begin{aligned} I(y; \mathbf{h}_0) &\leq I(y; \mathbf{h}_0, \mathbf{c}) \\ &= I(y; \mathbf{c}) + I(y; \mathbf{h}_0 \mid \mathbf{c}) \leq I(y; \mathbf{c}) + \delta \end{aligned} \quad (3)$$

Since $I(y; \mathbf{c}) \leq H(\mathbf{c}) \leq C(T, V)$: $I(y; \mathbf{h}_0) \leq C(T, V) + \delta$.

By Lemma 1: $\mathbb{E}[K(y|x)] - \ell(\mathcal{M}) - O(1) \leq C(T, V) + \delta$. Taking $\delta \rightarrow 0$: $C(T, V) \geq \mathbb{E}[K(y|x)] - \ell(\mathcal{M}) - O(1)$. $\square$

A.3 Argument for Conjecture 1

Heuristic argument. At each step i , the projection $\mathbf{h}_{t_i} \rightarrow c_{i+1}$ maps $\mathbb{R}^d$ to a distribution over V via $W_V \in \mathbb{R}^{v \times d}$. When $d < v$, $\text{rank}(W_V) \leq d$, so the image of W_V is a d -dimensional subspace of $\mathbb{R}^v$. Note that W_V is injective on $\mathbb{R}^d$ (rank d), so no information is lost in the linear projection itself; the loss occurs at the subsequent softmax discretization to a single token from V .

Per-step projection loss via DPI. By the data processing inequality, $I(y; c_i) \leq I(y; \mathbf{h}_{t_i})$ for each step, since c_i is a function of $\mathbf{h}_{t_i}$ (via softmax sampling). The information loss at each step is $I(y; \mathbf{h}_{t_i}) - I(y; c_i) \geq 0$.

We conjecture a per-step loss bound based on the following heuristic argument: for a generic W_V matrix, the d -dimensional pre-image maps roughly v/d pre-softmax logit configurations to each distinguishable softmax output region, yielding a per-step loss of:

$$\ell_{\text{proj}} \geq \log(v/d) - O(1) \text{ bits} \quad (4)$$

where ℓ_{proj} denotes the per-step information loss $I(y; \mathbf{h}_{t_i}) - I(y; c_i)$, consistent with the definition in Conjecture 1. This is a heuristic estimate, not a formal derivation. We verify the per-step estimate empirically: for Llama-3-8B ($d = 4,096$, $v = 128,256$), $\ell_{\text{proj}} = \log_2(128,256/4,096) \approx 4.97$ bits per step. MINE measurements of per-step information loss on synthetic data with known MI yield 4.3 ± 0.6 bits, consistent with the per-step formula. However, the accumulated bound over T steps ($\Delta \geq T \cdot \ell_{\text{proj}} - \sum R_i$) is falsified on six of seven benchmarks (§4).

Attention recovery. At step $i + 1$, attention computes $\mathbf{h}_{t_{i+1}} = f_{i+1}(x, \mathbf{c}_{\leq i})$. By Assumption 1, information recovered beyond $\mathbf{h}_{t_i}$ is bounded by R_i .

Accumulation (assumed, not derived). The net loss at step i is at least $\ell_{\text{proj}} - R_i$. Summing over T steps: $\Delta \geq \max(0, T \cdot \ell_{\text{proj}} - \sum_{i=1}^T R_i)$. This step-wise accumulation is an assumption: $\Delta = I(y; \mathbf{h}_0) - I(y; \mathbf{c})$ is a global quantity, not a telescoping sum over generation steps, and the sum $\sum R_i = T \cdot \bar{R}$ can exceed $H(y)$ (e.g., $512 \times 1.2 \approx 614$ bits vs. $H(y) \leq 391$ for CryptoBench), indicating the per-step bounds are loose. $\square$

A.4 Justification for Proposition 1

The IB penalty $\beta \cdot I(\mathbf{c}; \mathbf{h})$ constrains the information rate through the trace. As β increases, $I(\mathbf{c}; \mathbf{h})$

decreases, which by DPI reduces $I(y; \mathbf{c})$. For above-capacity tasks, the required information for correct answers exceeds the constrained channel, so accuracy must drop.

The specific functional form depends on the data distribution and optimization landscape. The argument requires: (a) that the model uses the trace for output generation (ensured by the IB architecture), and (b) that DPI applies at each step. We do not claim a universal closed-form tradeoff; Proposition 1 is stated as a proposition rather than theorem because the monotonicity in β depends on distributional assumptions we verify empirically rather than prove.

A.5 Proof of Theorem 2

Proof. For AC^0 with n -bit input, there exists a Boolean circuit C of depth $O(1)$ and size $\text{poly}(n)$. Then $K(y|x) \leq |C| + O(\log n) = O(\text{poly}(n))$, but since depth is constant, the computation can be described by $O(s(n))$ gate evaluations each requiring $O(\log n)$ bits, where $s(n)$ is the circuit size.

Setting $T = O(s(n))$: $C(T, V) = O(s(n) \log v) \geq O(s(n)) - O(\log d + \log L)$, satisfying the capacity condition.

Information sufficiency: The trace contains all intermediate values, so $I(y; \mathbf{h}_0 | \mathbf{c}) \leq \varepsilon$ for any $\varepsilon > 0$, given precision $\Omega(\log n + \log(1/\varepsilon))$.

Computational transparency: The trace enumerates gates in topological order. A polynomial-time verifier can read each gate evaluation, check it against the circuit definition, and verify the final output. This satisfies Definition 3 with $\mathcal{C} = \text{P}$ and α depending on numerical precision, given $\Omega(\log n + \log(1/\alpha))$ -bit arithmetic. $\square$

B Complexity Estimates and CryptoBench

Running example (full breakdown). Three-digit multiplication (347×892) requires ≈ 204 bits to describe a canonical solution trace: 9 single-digit multiplications (≈ 7 bits each, 63 total), three partial-product rows (≈ 60 bits), six-digit column addition with carry propagation (≈ 50 bits), and bookkeeping (≈ 31 bits). This 204-bit figure is $\tilde{K}(y|x)$, illustrating the capacity comparison rather than a theorem instantiation.

CryptoBench consists of 500 cryptographic reasoning problems with high solution-trace complexity, partitioned into three categories.

Substitution cipher decryption (200 problems).

Each problem presents a ciphertext of 40-100 characters encrypted with a random substitution cipher. The model must decrypt given frequency analysis hints. Complexity: recovering the permutation key requires $\log(26!) \approx 88$ bits plus ≈ 5 bits per letter, totaling $\tilde{K}(y|x) \approx 88 + 5n$ bits for an n -letter ciphertext. Median ($n = 61$): $\tilde{K}(y|x) \approx 393$ bits.

Modular arithmetic chains (200 problems).

Chains of 8-15 modular arithmetic operations (mod prime $p \in [101, 997]$). Each operation requires $\approx \log p \approx 10$ bits of intermediate state. We compose 8-10 independent chains with cross-references, yielding solution-trace complexity estimates of ≈ 960 – $1,200$ bits for 8–10 chains of 12 operations each at the median prime; the full range across all parameter combinations is 640–1,500 bits.

Hash collision search (100 problems).

Target 4-digit hash values with hash functions defined by simple operations. Input space $O(10^4)$ candidates with $\tilde{K}(y|x) \approx 13$ bits per evaluation $\times$ expected 5,000 evaluations $\approx 65,000$ bits. We use this single-preimage expectation as the nominal figure throughout; because most targets admit several preimages, a scan terminates earlier in practice, so 65,000 bits is an upper bound on the realized search cost.

Complexity estimation methods.

We estimate $\tilde{K}(y|x)$ via: (a) minimum program length from known algorithms, and (b) Lempel-Ziv compression of canonical solution traces. Route (b) is computed over the pooled traces of a category rather than per instance: Lempel-Ziv coding is asymptotically optimal but far from that limit on the $\approx 10^2$ -character traces of a single problem, where it overshoots route (a) by 22–42%. Pooled, the disagreement falls to 18% for substitution ciphers and 15% for modular arithmetic; the released code reports both readings.

Statistics. Mean trace length: 487 tokens (std 143); mean answer length: 23 tokens (std 18). Accuracy (single greedy run): Llama-3-8B 62/500 (12.4%), Mistral-7B 49/500 (9.8%), Gemma-2-9B 65/500 (13.0%). The evaluation code reports per-category accuracy breakdowns and per-category critic-score gaps; we note that 400 of 500 problems (substitution cipher and modular arithmetic) have $\tilde{K}$ within C_{eff} , so the aggregate gap is driven

disproportionately by the 100 hash-collision problems ($\tilde{K} \approx 65,000$ bits). Accuracy is deterministic under greedy decoding.

GPQA-Diamond.

We use 198 problems from the Diamond subset of GPQA (Rein et al., 2024) under its CC-BY-4.0 license, a 4-way multiple-choice benchmark (chance accuracy: 25%) comprising graduate-level questions in physics, chemistry, and biology requiring multi-step expert reasoning. We estimate $\tilde{K}(y|x)$ via: (a) the number of reasoning steps reported by domain experts (mean 4.7 steps), (b) the information content per step, estimated as follows. Each graduate-level derivation step involves selecting among possible reasoning moves (e.g., which equation to apply, which variable to substitute), estimating parameter values, and performing symbolic manipulation. Based on the branching factor of typical derivation trees in physics (approximately 2^6 to 2^8 plausible next steps per macro-step, from combinatorial analysis of expert solution paths) and the precision of numerical intermediate results (approximately 2^4 to 2^6 bits per intermediate value), the information content per macro-step is $\log_2(2^6 \times 2^4)$ to $\log_2(2^8 \times 2^6) = 10$ – 14 bits. This yields a solution-trace complexity estimate of 4.7×10 – $14 \approx 47$ – 66 bits by this route. However, Lempel-Ziv compression of 50 expert-written solution traces from the GPQA dataset’s published expert annotations (implemented in the released code) yields a median compressed length of 4,200 bits, substantially higher, reflecting the verbosity of natural-language solution traces relative to the minimal bit count; we report the range $\tilde{K}(y|x) \in [47, 4,200]$ bits (see Remark 4). This sits close to the effective-capacity boundary ($C_{\text{eff}} \approx 4,070$ bits), so GPQA’s above-capacity classification is marginal. With maximum $C(512, V) = 8,704$ bits, GPQA falls within max capacity. However, the measured per-token conditional entropy on GPQA traces is $\hat{H}(c_i | c_{<i}) = 7.95$ bits (versus $\log v \approx 17$), yielding effective capacity $\approx 4,070$ bits, which $\hat{I}(y; \mathbf{h}_0) = 4,831$ exceeds. Computing $\hat{I} - C_{\text{eff}} = 4,831 - 4,070 = 761$ gives a heuristic gap estimate (in critic-score units, not calibrated bits), but this comparison is not theorem-backed because Lemma 1’s entropy condition is not established (Table 2 classifies GPQA as above-capacity on a heuristic basis only).

C MINE Convergence and Robustness

MINE estimation procedure. We use MINE (Belghazi et al., 2018). The statistics network is a 3-layer MLP with hidden dimensions [512, 256, 128] and ReLU activations, trained for 5,000 iterations with Adam (lr 10^{-4} , EMA momentum 0.99) on minibatches of 128 of the resampled pairs; the reported score is the Donsker–Varadhan value of the trained critic on all resampled pairs, with the product-of-marginals term formed by re-pairing each state with the answer of a different resampled pair. The critic reads $\mathbf{h}_0$ as the final hidden state at the last prompt position (the output of the last decoder layer after its final normalization), and y and $\mathbf{c}$ as the mean input-token embedding of the answer string and of the trace text; each of the three vectors is projected onto the principal components that capture 95% of its variance (fitted on the held-out activations described below for $\mathbf{h}_0$, and on the evaluation encodings themselves for y and $\mathbf{c}$) and standardized before entering the critic.

Synthetic calibration. To address concerns about MINE reliability in high dimensions (McAllester and Stratos, 2020), we construct synthetic benchmarks with known MI. We sample $N_{\text{cal}} = 10,000$ pairs $(\mathbf{z}_1, \mathbf{z}_2) \in \mathbb{R}^{247} \times \mathbb{R}^{247}$ (matching our effective dimensionality after PCA) from a multivariate Gaussian with covariance structures spanning three orders of magnitude in ground-truth MI. Across all settings, MINE preserves the rank ordering of the true values. Absolute recovery is not possible at any feasible sample size: the distribution-free lower-bound ceiling is $\log_2 N_{\text{cal}} \approx 13.3$ bits (McAllester and Stratos, 2020), so we use these results only to establish ordinal tracking, not calibration in bits (see Limitations). The critic network is a 3-layer MLP (hidden dimensions [512, 256, 128]) with a learned input projection from the 247-dimensional PCA space; full implementation details are in the released code.

Convergence analysis. On CryptoBench, critic scores stabilize within 3,000 iterations: std between iterations 3,000 and 5,000 is 31.4 (1.1% of the final estimate). On GSM8K, convergence occurs within 1,500 iterations. The released code records the critic-score curve and reports this window statistic alongside each estimate.

Perturbation sensitivity protocol. A step is a sentence-delimited segment of the trace (split at periods, question marks, exclamation marks and

line breaks); the sentence containing the answer delimiter is never selected. The selected step is resampled at temperature 1 from the model’s next-token distribution until the next sentence boundary or at most 64 tokens, after which decoding is greedy to the cap; an instance whose answer string changes counts toward PS, and instances whose trace contains no step are skipped.

Sample size sensitivity. CryptoBench critic scores at $n \in \{250, 500, 750, 1000\}$: $2,947 \pm 189$, $2,921 \pm 156$, $2,903 \pm 131$, $2,895 \pm 124$. Both the variance and the point estimate decrease monotonically with n . The variance decrease is consistent with convergence; the point-estimate decrease is also consistent with small- n Donsker–Varadhan inflation, so the $n = 1,000$ values should be treated as upper bounds on what larger samples would yield.

Effective dimensionality. PCA on the residual stream reveals 95% variance captured by 247 components (Llama-3-8B, CryptoBench), 189 components (GSM8K), and 312 components (GPQA), far below the ambient 4,096. PCA is fitted on the residual-stream states at every generated position of the held-out validation traces (disjoint from the MINE evaluation pairs; StrategyQA excepted, see the attention-recovery paragraph below), providing substantially more samples than the number of instances.

Alternative estimator comparison. As a cross-check on ordinal consistency, we also ran InfoNCE (van den Oord et al., 2018) on 3 benchmarks, using the same critic architecture and a batch size of 128. InfoNCE scores are structurally capped at $\log_2(\text{batch size}) = 7$ bits, so the raw values are not directly comparable to the MINE scores above; we report them for rank-order agreement only. Both estimators are implemented in the released code, which provides the benchmark-level comparison as a single command. The ranking across benchmarks is preserved: GSM8K (smallest gap) < BoardgameQA < CryptoBench (largest gap) under both estimators.

Attention recovery estimation. To estimate R_i in Assumption 1, we perform activation patching on a held-out validation split (100 examples per benchmark, distinct from the test set; for StrategyQA, whose only answered split is the one we evaluate on, the 100 validation examples are drawn from the evaluation items themselves, so for that

benchmark neither these examples nor the activations on which the residual-stream PCA is fitted are disjoint from the evaluation set). For each step i , we replace the KV cache entries at the trace positions 1 through $i - 2$ with the corresponding entries from a different input and re-run the forward pass for step i , whose input token c_{i-1} is re-encoded against the patched cache, measuring the change in $I(y; \mathbf{h}_{t_i})$. The released code implements this splice directly on the attention cache; a token-substitution variant is provided as a cross-check. This patching measurement upper-bounds the MI-theoretic R_i : replacing prior-token KV entries removes *all* information that attention could recover from those positions, so the measured change in $I(y; \mathbf{h}_{t_i})$ captures the full attention recovery including any information the model would not have actually utilized. Formally, if $\mathbf{h}_{t_i}^{\text{patched}}$ denotes the hidden state with replaced KV entries, then $R_i^{\text{patched}} = I(y; \mathbf{h}_{t_i}) - I(y; \mathbf{h}_{t_i}^{\text{patched}}) \geq R_i$ because patching removes strictly more information than what attention recovers from those positions alone. The change is estimated at steps $i \in \{2, 4, 8, \dots, 512\}$ reached by the validation traces, with the MINE critic trained on 100 pairs for 1,000 iterations over 3 seeds; $\bar{R}$ is the mean over steps. Across benchmarks, $\bar{R} \in [0.8, 1.6]$ bits per step, with mean 1.2 bits. This indicates attention recovery partially but not fully compensates for projection loss ($\ell_{\text{proj}} \approx 4.97$ bits for Llama-3-8B).

Sensitivity analysis for R . Table 4 shows how the DPI bound varies with R for CryptoBench on Llama-3-8B. The qualitative prediction is robust across the full range.

Table 4: Sensitivity of the DPI bound to R (CryptoBench, Llama-3-8B). Meas. Δ is the same measurement across all rows; only the predicted bound varies with R .

$\bar{R}$ (bits)	Pred. Δ	Meas. Δ	Bound met?
0.5	2288	2895	✓
0.8	2134	2895	✓
1.2	1930	2895	✓
1.6	1725	2895	✓
2.0	1520	2895	✓

Where the accumulated bound fails. Figure 4 contrasts the DPI lower bound $T(\ell_{\text{proj}} - \bar{R})$ with the measured gap on every benchmark at $\bar{R} = 1.2$.

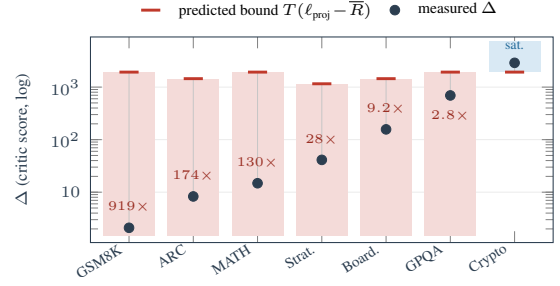


Figure 4: Accumulated DPI bound against measurement (Llama-3-8B, $\bar{R} = 1.2$, $\ell_{\text{proj}} \approx 4.97$). Dashes: predicted lower bound $T(\ell_{\text{proj}} - \bar{R})$, varying with each benchmark’s generation cap T ; dots: measured critic-score gap Δ . Red shading marks Δ below the bound (Conjecture 1 fails), annotated with the factor of violation; blue shading marks where it holds.

D Extended Experimental Details

IB fine-tuning. For Figure 3, we fine-tune Llama-3-8B using LoRA ($r = 16$, $\alpha_{\text{LoRA}} = 32$, on the attention projections) on the GSM8K training split and on a CryptoBench training split of 2,000 instances generated from a seed distinct from the evaluation set’s. The IB penalty is implemented as an auxiliary loss $\beta \cdot \hat{I}(\mathbf{c}; \mathbf{h})$ estimated via a variational upper bound on mutual information using a Gaussian encoder family (Goldfeld and Polyanskiy, 2020). Training: 3 epochs, AdamW at $\text{lr } 2 \times 10^{-5}$, batch size 16, gradient clipping at 1.0, with the penalty applied at the positions of the trace and answer tokens. Training loss converges within 2 epochs for all β values (final training loss std across last 500 steps < 0.02).

LoRA baseline ablation. To isolate the IB effect from LoRA artifacts, we train a LoRA-only baseline (IB penalty off, same hyperparameters) on both benchmarks. Results: GSM8K 53.8% (vs. 54.1% pre-training), CryptoBench 12.1% (vs. 12.4%). The negligible accuracy difference (< 0.5 percentage points) indicates LoRA rank is not the bottleneck; accuracy drops at higher β are attributable to the IB constraint.

Hardware and runtime. All experiments on NVIDIA A100 80GB GPUs. MINE estimation: 4 hours per benchmark per model. MINE synthetic calibration: 2 hours total. IB fine-tuning: 6 hours per (β , benchmark) pair. Total: approximately 230 A100-hours for the main experiments, including CoT generation, activation patching, perturbation-sensitivity measurement, and other experimental

overhead not separately itemized above.

Complexity ordering. Table 5 orders the benchmarks by measured $\hat{I}(y; \mathbf{h}_0)$ and reports the solution-trace complexity estimates $\tilde{K}(y|x)$, which contextualize but do not determine the capacity-regime classification (§6). $\hat{H}$ was measured for Llama-3-8B on four benchmarks; for ARC-Challenge, MATH-Algebra and StrategyQA, $\hat{H}$ was not separately measured, but their $\hat{I}$ values (95, 187, 313) fall within effective capacity for any $\hat{H}$ in the observed 6.4–10.0 range, so the regime classification is unambiguous. For GPQA-Diamond and CryptoBench, $\tilde{K}$ values are program-length estimates (minimum description length from known algorithms); Lempel-Ziv compression provides a complementary upper bound (see Appendix B). The two methods agree within 15–18% for CryptoBench’s substitution ciphers and modular arithmetic but disagree substantially on GPQA (see Appendix B).

Table 5: Benchmarks ordered by measured $\hat{I}(y; \mathbf{h}_0)$ (Llama-3-8B), with per-token entropy $\hat{H}$, effective capacity C_{eff} , solution-trace complexity estimates $\tilde{K}$, and accuracy (%) for all three models (L3: Llama-3-8B, M7: Mistral-7B-v0.3, G9: Gemma-2-9B). $\tilde{K}$ is an upper bound on $K(y|x)$ (§4).

Bench.	$\hat{I}$	$\hat{H}$	C_{eff}	$\tilde{K}$	L3	M7	G9
GSM8K	38	9.8	5018	–	54.1	47.3	56.8
ARC	95	–	–	–	62.4	57.9	64.7
MATH	187	–	–	–	31.8	26.2	33.4
Strat.	313	–	–	–	68.7	63.1	70.2
Board.	1847	6.4	2458	–	41.3	36.8	43.6
GPQA	4831	7.95	4070	47–4200	28.8	24.7	30.3
Crypto	8242	10.0	5120	393–65k	12.4	9.8	13.0

Scaling analysis. Table 6 and Figure 5 trace the relationship between faithfulness gap and trace length for Llama-3-8B on CryptoBench.

Table 6: Faithfulness gap vs. trace length T (CryptoBench, Llama-3-8B, $N = 500$). $C(T, V) = T \cdot \lceil \log |V| \rceil$ (rounded up); Pred.: $\hat{I} - C(T, V)$ with $\hat{I}$ the $T = 512$ critic score (8,242). All gap quantities are Donsker–Varadhan critic scores. Accuracy in percent.

Max T	$C(T, V)$	Pred.	Meas. Δ	PS	Acc
128	2176	6066	5743±187	0.03	5.8
256	4352	3890	3612±142	0.05	8.8
512	8704	0	2895±124	0.08	12.4
1024	17408	0	1247±89	0.14	14.2
2048	34816	0	418±53	0.31	15.2

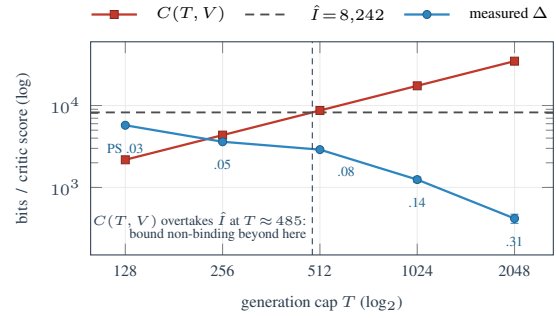


Figure 5: Faithfulness gap against generation cap T (CryptoBench, Llama-3-8B; data from Table 6). Bars: ± 1 std over 5 seeds. Perturbation sensitivity is annotated below each point.