
Formalizing Latent Thoughts: Four Axioms of Thought Representation in LLMs

Fahd Seddik

Department of Computer Science
University of British Columbia
Kelowna, BC, Canada
fahd.seddik@ubc.ca

Fatemeh Fard

Department of Computer Science
University of British Columbia
Kelowna, BC, Canada
fatemeh.fard@ubc.ca

Abstract

We introduce an axiomatic evaluation framework for latent thought representations in LLMs, comprising metrics that are independent of downstream benchmark scores and reveal representational failures that benchmark accuracy masks. Existing evaluations conflate representation quality with model capacity. Therefore, failures cannot be attributed to the representation rather than to the model that processes it. We formalize four functional axioms (Causality, Minimality, Separability, and Stability) and define a quantitative measure for each, computed directly on the representation independently of downstream accuracy. We audit open-weight LLMs across 23 reasoning tasks (e.g., Spatial Reasoning, Factual QA). We find that no candidate satisfies all four axioms simultaneously, that the representations distinguish task type reliably but cannot distinguish between two questions within the same task, and that the representations encode little information beyond what is already present in the input embedding. The failure is consistent across dense, reasoning-distilled, and RL-trained model families, indicating that the gap is structural rather than a property of model size or training procedure. Code: <https://fard-lab.github.io/formalize-thoughts>.

1 Introduction

Reasoning in Large Language Models (LLMs) has increasingly moved from discrete Chain-of-Thought (CoT) tokens toward continuous latent representations, with a growing body of efficiency-motivated work compressing or replacing explicit CoT steps with continuous vectors [66, 27]. Recent work highlights the limitations of discrete tokenization and decoding for reasoning [31] and replaces discrete steps with continuous representations [9, 81], reporting accuracy gains on complex reasoning benchmarks [73]. However, the field evaluates these “continuous thought representations” almost exclusively through downstream task accuracy [50]. Probes of continuous reasoning tokens find that distinct reasoning paths collapse to a single interpretation in early layers while downstream accuracy remains unchanged [57]. Two prior questions therefore remain open. What functional properties constitute a valid thought representation, and how can we measure those properties independently of any downstream task? Even when models maintain accurate internal representations, they may fail at downstream tasks [76].

We identify three gaps that impede principled progress on continuous reasoning. First, there is no **principled definition** of what a thought representation must do, because existing methods optimize heuristic proxies (step counts, token budgets, imitation of explicit CoT) without a formal statement of the functional requirements. Second, there is no **intrinsic evaluation** that measures representation quality independently of downstream accuracy [58], and therefore failures cannot be attributed to the representation rather than to the decoder or prompt. Third, and consequently, the **status of current methods** is unknown. Recent audits find that latent reasoning tokens are often unnecessary for the predictions they are meant to drive [19], and that latent reasoning does not faithfully implement the structured latent-space search that motivates it [16].

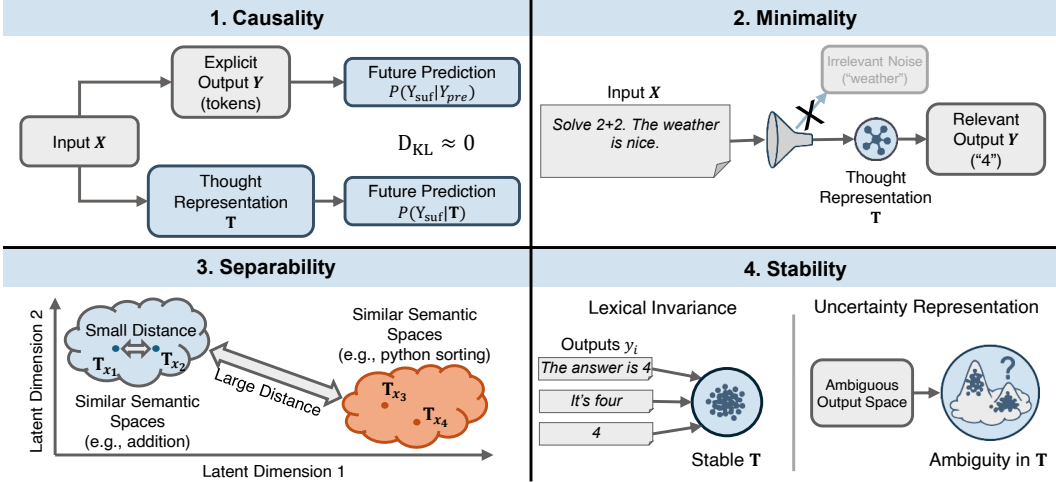


Figure 1: Visualizing the axiomatic properties of a Functional Thought Representation \mathbf{T}

Our approach. We propose an axiomatic characterization of thought representations in terms of four functional properties (*Causality*, *Minimality*, *Separability*, and *Stability*), illustrated in Figure 1. The framework describes the function of a thought representation rather than its form. It applies equally to vectors, tensors, or sets of vectors. For each axiom, we define a quantitative measure evaluated directly on the source LLM without retraining. We use this suite to audit candidate thought representations produced by Soft Thinking with and without Gumbel noise [81, 73] and Latent Thinking [31, 84] across a range of thinking budgets, alongside candidates extracted from last-input-token hidden states. We study five open-weight LLMs spanning dense and sparse mixture-of-experts architectures on the 23 tasks of Big Bench Extra Hard (BBEH) [41].

Across the evaluated source LLMs, the intrinsic protocol exposes failure modes that downstream task accuracy does not. The candidates retain coarse task identity but lose the per-question identity that distinguishes one instance of a task from another, and the input-prompt embedding is itself competitive with the candidates on every axis the framework measures.

Contributions. Each of the three gaps above is addressed by a corresponding contribution.

- **Axiomatic formalization of thought representations.** We provide a definition of thought representations stated in terms of four functional axioms, and prove their logical consistency, independence, and completeness (Appendix B).
- **Intrinsic evaluation protocol.** We introduce a measure for each axiom. KL substitution error measures Causality, the Minimality Gap measures Minimality, same- and cross-task discriminator accuracy measures Separability, and the Distributional Consistency Score (DCS) measures Stability. Each measure is computable on the source LLM without retraining and is decoupled from downstream task accuracy.
- **Empirical audit of candidate thought representations.** Applying the protocol on BBEH exposes a fine-grained *representational collapse* on per-question identity. No evaluated candidate satisfies all four axioms simultaneously, consistent with concurrent reports of silent representational failures in latent reasoning despite unchanged downstream accuracy [58].

The framework provides a principled means to develop and audit candidate thought representations as objects of study, evaluated for what an LLM encodes about a problem rather than for the downstream reasoning quality of any chain that follows. Beyond evaluation, the four measures provide researchers explicit, quantifiable optimization targets and the diagnostic resolution to attribute downstream accuracy changes to specific representational properties rather than to an aggregate benchmark score. The four measures are one realization of the axioms, which admit alternative quantifications under the same principle, allowing new candidates and measurement designs in future work.

2 Related work

Latent representations. Probing studies demonstrate that LLMs build rich internal representations before outputting tokens, with internal states predicting CoT success prior to completion [1, 48], encoding sparse logical features [34], and correlating with task difficulty [35]. These internal signals motivate representations capturing model intent without reliance on explicit outputs. COCONUT [31] and CODI [63] utilize hidden states to represent or compress multiple reasoning paths into continuous vectors. Architectural variants including Tiny Recursive Models (TRMs) [38, 70] and Encode-Think-Decode (ETD) [43] update latent representations via recursive networks or iteration over reasoning-relevant layers, and related models propagate persistent latent state [45] or inject recurrence through a dedicated middle-layer pathway [10]. Similarly, PonderLM-2 [78] and pause-token approaches [29] train models to think via latent pauses or hidden-state interfaces before generation.

Continuous reasoning methods. Soft-token research explores continuous concept-spaces [81, 64], but these methods must be carefully constrained to avoid degenerate greedy behavior [73, 9] and remain competitive with textual CoT only on a narrow subset of tasks [77]. Layer-wise analyses further suggest that last-layer activations are used directly for next-token prediction, while mid-depth representations better balance signal preservation and noise compression [65, 62], and key semantic computations often migrate to these layers [71]. Proposed extraction strategies range from vocab-space constrained optimizers [18] to multi-layer readout modules [14]. These works typically conflate thought representation with reasoning performance. In contrast, we isolate representation quality from reasoning steps, avoiding the discretization errors [31] and heuristic approximations [37] inherent in standard decoding algorithms. Additional related work is in Section A.

3 Formalizing thought

3.1 Semantic equivalence and geometric space

Notation. Calligraphic letters ($\mathcal{X}, \mathcal{Y}, \mathcal{S}, \mathcal{T}$) denote spaces, capital letters (X, Y, Z) denote random variables, lowercase letters (x, y, z) denote specific instances, and bold \mathbf{T} denotes a candidate Thought Representation (TR) throughout the paper.

To rigorously characterize the properties of a thought representation, we first establish a tractable criterion for *semantic equivalence* between output sequences. Let \mathcal{Y} be the space of all possible generated sequences. We posit the existence of a semantic mapping function $\Phi : \mathcal{Y} \rightarrow \mathcal{S}$, where \mathcal{S} is a semantic manifold. Two sequences $y, y' \in \mathcal{Y}$ are defined as semantically equivalent, denoted $y \sim_{sem} y'$, if and only if $\Phi(y) = \Phi(y')$.

We further impose a geometric structure on \mathcal{S} equipped with a metric $d_{\mathcal{S}}(\cdot, \cdot)$. Equivalence implies locality. Non-equivalent sequences nevertheless exhibit varying degrees of proximity. For sequences y_1, y_2, y_3 , if y_1 and y_2 share partial semantic overlap (e.g., distinct numerical answers to the same query) whereas y_3 is conceptually disjoint, we require $d_{\mathcal{S}}(\Phi(y_1), \Phi(y_2)) < d_{\mathcal{S}}(\Phi(y_1), \Phi(y_3))$.

Computationally, we approximate this metric space using high-dimensional embeddings. We utilize the cosine similarity between embedding representations $\mathbf{e}_y, \mathbf{e}_{y'} \in \mathbb{R}^d$ as inversely correlated with the semantic metric $d_{\mathcal{S}}$, consistent with evidence that modern learned text embeddings approximate human semantic judgments across a wide range of similarity tasks [4].

3.2 Thought as a latent functional state

What functional properties must a representation \mathbf{T} satisfy to qualify as a thought, and how can each be measured directly on LLMs?

We define a thought representation not as a communicable linguistic artifact (e.g., Chain-of-Thought) but as a *Functional Thought \mathbf{T}* , a latent state that mediates the transformation from an Input space \mathcal{X} to the semantic Output space \mathcal{S} .

Formally, given an input $x \in \mathcal{X}$, the model induces a probability distribution over the output $P(Y|x)$. The functional thought \mathbf{T} is a representation intended to capture the sufficient statistics of this distribution. We explicitly exclude interpretability from this definition. Interpretability is an observer-dependent property. In contrast, a functional thought \mathbf{T} is constructed to be mathematically optimal in mediating $X \rightarrow Y$, which may render it opaque to human inspection.

Definition 1 (Idealized Thought Representation Mapping). Let \mathcal{X} denote the input space and \mathcal{S} denote the semantic output space of a model $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{S}$. We define a thought representation generator as a function $g : \mathcal{X} \rightarrow \mathcal{T}$, where \mathcal{T} represents the thought space.

The function g is constructed such that it induces an equivalence relation on \mathcal{X} based on the semantic outputs in \mathcal{S} . Specifically, for any pair of inputs $x_i, x_j \in \mathcal{X}$, the mapping satisfies:

$$g(x_i) = g(x_j) \iff \mathcal{M}(x_i) = \mathcal{M}(x_j) \quad (1)$$

This implies that g is a many-to-one mapping effectively compressing \mathcal{X} into \mathcal{T} by preserving distinctness only if the inputs result in semantically distinct outputs in \mathcal{S} .

Remark. Stochastic extraction methods (e.g., Gumbel noise) fix a global random seed, so g remains deterministic in practice. The formal construction is in Section B.

3.3 Quantifying the axioms of Functional Thought

We propose that a robust thought representation \mathbf{T} must satisfy four axiomatic properties, formalized below via information theory and probability and shown in Figure 1. Table 1 maps each axiom to its formal requirement and to the metric used to quantify it. Through Appendix B, we prove consistency that follows from an idealized one-hot semantic bottleneck construction. Independence is established by four counter-models, each preserving three axioms while violating the fourth. Completeness follows from a bijection between \mathcal{T} and the reachable semantic manifold $\mathcal{S}_{\mathcal{M}}$.

Running Example

Take $x = \text{“Is 13 prime, and why?”}$ with output $y = \text{“Yes. 13 is prime because no integer from 2 to 12 divides it.”}$ produced by \mathcal{M}_θ . Let \mathbf{T} denote a candidate thought representation extracted from \mathcal{M}_θ for x . The examples below are simplified for clarity.

1. Causality. Each output y is partitioned into a reasoning prefix y_{pre} and an answer suffix y_{suf} . If \mathbf{T} is a valid thought representation derived from y_{pre} , it must functionally substitute y_{pre} within the computational graph of \mathcal{M}_θ , so that conditioning on \mathbf{T} yields a predictive distribution over y_{suf} indistinguishable from conditioning on the explicit tokens y_{pre} .

Example. Causality requires that replacing the token embeddings of the reasoning prefix $y_{\text{pre}} = \text{“Yes. 13 is prime because no integer from 2 to 12 divides it.”}$ inside \mathcal{M}_θ with the projected \mathbf{T} leaves the distribution over the answer suffix y_{suf} essentially unchanged. Here y_{suf} could be the concluding portion of the same output, such as $\text{“Therefore, 13 is prime.”}$

Quantification. We replace the token embeddings of y_{pre} in the model with the projected \mathbf{T} and measure the resulting divergence on y_{suf} :

$$\text{Causality Error} = D_{\text{KL}}\left(P(y_{\text{suf}} \mid y_{\text{pre}}) \parallel P(y_{\text{suf}} \mid \mathbf{T})\right) \quad (2)$$

A lower value indicates that \mathbf{T} encapsulates the effect of y_{pre} on subsequent generation, consistent with empirical analyses showing causal structure between intermediate latent representations and downstream generation in continuous reasoning [46]. Sensitivity to the answer-window length and number of substituted positions is documented in Section D.6.

2. Minimality. A thought representation satisfies optimality iff it compresses the input and retains maximum relevance to the output distribution. This aligns with the Information Bottleneck principle, a framing that has recently been applied both to characterize how LLM pre-training approaches minimal sufficient compression [15] and to cast chain-of-thought reasoning itself as a bottleneck variable between prompt and answer [49]. Let X and Y denote the random variables over inputs and generated outputs respectively, and let $I(\cdot; \cdot)$ denote mutual information. An optimal \mathbf{T} minimizes $I(X; \mathbf{T})$ subject to a constraint on $I(\mathbf{T}; Y)$:

$$\min_{\mathbf{T}} I(X; \mathbf{T}) - \beta I(\mathbf{T}; Y) \quad (3)$$

This characteristic ensures that \mathbf{T} filters out nuisance variables in X (e.g., irrelevant context or noise) that do not contribute to the generation of the high probability semantic output space.

Example. Suppose the input contains two unrelated topics, e.g., “Hamlet was written by Shakespeare around 1600. Is 13 prime, and why?” If the model’s output addresses only the primality question, the literary detail did not contribute to that output and \mathbf{T} should not encode it. If the output addressed both topics, \mathbf{T} should retain both. Minimality penalises encoding content that did not contribute to y .

Quantification. The IB Lagrangian is intractable because $I(X; \mathbf{T})$ and $I(\mathbf{T}; Y)$ depend on unknown distributions. We construct a cross-entropy surrogate that preserves the Lagrangian’s ranking at $\beta = 2$. The surrogate combines three cross-entropies, $\text{CE}(Y | \mathbf{T})$, $\text{CE}(X | Y)$, and $\text{CE}(X | Y, \mathbf{T})$, and reduces after dropping TR-independent constants to:

$$\Delta_{\text{IB}} = \text{CE}(X | Y, \mathbf{T}) - \text{CE}(Y | \mathbf{T}) \quad (4)$$

A larger Δ_{IB} indicates a representation that is simultaneously relevant (\mathbf{T} predicts Y with low residual entropy) and minimal (\mathbf{T} contributes negligible additional information about X beyond what Y provides). The reduction assumes $I(Y; \mathbf{T} | X) = 0$, which holds when \mathbf{T} is a function of X alone¹.

3. Separability. Separability defines the functional injectivity of the mapping from semantic content to the latent space. Because logical distinctions in high-dimensional representations are encoded in the curvature and flow structure of the underlying manifold rather than in raw point distances [82], we do not adopt fixed geometric distances (e.g., Euclidean margins) and instead rely on functional discriminability. The representation must contain sufficient topological structure to distinguish between semantically nonequivalent output distributions using a bounded capacity projection. Given two inputs $x_1, x_2 \in \mathcal{X}$ that induce disjoint high probability semantic spaces $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, their corresponding thought representations \mathbf{T}_{x_1} and \mathbf{T}_{x_2} must be resolvable by an optimal semantic projection $\phi : \mathcal{T} \rightarrow \mathcal{S}$ drawn from a bounded hypothesis class \mathcal{H} (a linear projection followed by a linear classification head, consistent with the linear representation hypothesis [54]). Using the semantic metric $d_{\mathcal{S}}$ defined over the semantic manifold, we require:

$$d_{\mathcal{S}}(\phi(\mathbf{T}_{x_1}), \phi(\mathbf{T}_{x_2})) > \delta \quad \text{for some } \phi \in \mathcal{H} \quad (5)$$

Conversely, if distinct inputs lead to semantically convergent outputs, their representations should reside on the same functional manifold, rendering them indistinguishable under ϕ . Separability thus ensures that \mathcal{T} contains the necessary decision boundaries to be isomorphic to the semantic space \mathcal{S} .

Example. The same-task setting pairs x with $x' = \text{“Is 14 prime, and why?”}$, which yields $y' = \text{“No. } 14 = 2 \times 7\text{.”}$ The inputs differ by a single token and the answers are semantically opposite, so a bounded classifier acting on \mathbf{T}_x and $\mathbf{T}_{x'}$ must place them on opposite sides of its decision boundary. A cross-task setting pairs x with a medical-domain prompt such as “What are the symptoms of Alzheimer’s?”. Its output occupies a disjoint semantic region and must remain just as resolvable.

Quantification. We instantiate ϕ as a learned binary discriminator $f_{\text{disc}}(\mathbf{T}, Y) \in [0, 1]$ drawn from \mathcal{H} , which scores the alignment between \mathbf{T} and a candidate output sequence Y . Positives pair \mathbf{T} with its corresponding generated sequence. Negatives use two strategies, same-task pairing for fine-grained within-task discrimination and cross-task pairing for cross-domain separability. We realize f_{disc} as a trainable linear projection that maps \mathbf{T} into the embedding space of a frozen LLM backbone, followed by a trained classification head optimized with binary cross-entropy. Classification accuracy is the metric. Concurrent analyses of soft-thinking representations [57] report an analogous superposition failure in which distinct reasoning paths become indistinguishable.

4. Stability. The representation must be invariant to surface level lexical variations in the output space and robust to sampling stochasticity. Rather than encoding a single realization $y \sim P(Y|x)$, \mathbf{T} should encode the parameters of the semantic distribution $P(\mathcal{S}|x)$. This implies two conditions: (1) **Mode Collapse Resistance:** If $P(Y|x)$ represents uncertainty or confusion, \mathbf{T} must reflect this entropy, because LLM generation frequently collapses to a small subset of high probability modes and fails to mirror the underlying predictive distribution [83]. (2) **Lexical Invariance:** For any set of high probability sibling outputs $\{y_1, \dots, y_k\}$ drawn from the same input x that are semantically equivalent ($y_i \sim_{\text{sem}} y_j$), the induced representations should satisfy $\mathbf{T}_{y_i} \approx \mathbf{T}_{y_j}$, a property requiring explicit enforcement against latent divergence under paraphrasing [55].

Example. For lexical invariance, two sibling outputs $y_1 = \text{“Yes. 13 is prime because no integer from 2 to 12 divides it.”}$ and $y_2 = \text{“Yes, 13 has no divisors other than 1 and itself.”}$ are semantically

¹The output embeddings used in experiments violate this assumption by construction and are reported as anchor references rather than IB-Lagrangian estimates; see Section D.5 for the full derivation.

equivalent and must induce $\mathbf{T}_{y_1} \approx \mathbf{T}_{y_2}$. For mode-collapse resistance, suppose the model is imperfect and outputs “Yes” in some generations and “No” in others, so its high-probability outputs do not all agree. Then $P(Y|x)$ has positive entropy and \mathbf{T} must reflect both modes rather than encode just one.

Quantification. For candidates that produce a single representation per input, lexical invariance holds by construction and we probe mode-collapse resistance only. We quantify distributional uncertainty via the semantic entropy H_x of Kuhn et al. [44], computed by binarizing pairwise cosine similarities between K output embeddings at threshold τ to form semantic equivalence classes and setting H_x to the Shannon entropy over class sizes. A question with $H_x = 0$ has all outputs in one class, whereas $H_x > 0$ indicates spread across semantically distinct outputs. To measure whether \mathbf{T} linearly encodes this property, we adopt the difference-of-means probe of Cencerrado et al. [12] and report the cross-validated AUROC for predicting $H_x > 0$. The resulting Distributional Consistency Score (DCS) ranges from 0.5 (random baseline) to 1.0 (perfect discrimination). Further analysis on DCS, input embeddings as a proxy to question difficulty, and sensitivity to τ are in Section C.6.

Table 1: The four axioms with their formal requirement on \mathbf{T} and quantifying measure.

Axiom	Formal requirement	Quantitative measure
1. Causality	$D_{\text{KL}}(P_\theta(Z Y) \ P_\theta(Z \mathbf{T})) \approx 0$	KL substitution error (Equation (2))
2. Minimality	$\min_{\mathbf{T}} I(X; \mathbf{T}) - \beta I(\mathbf{T}; Y)$	IB residual gap Δ_{IB} (Equation (4))
3. Separability	$d_{\mathcal{S}}(\phi(\mathbf{T}_{x_1}), \phi(\mathbf{T}_{x_2})) > \delta, \phi \in \mathcal{H}$	Discriminator accuracy
4. Stability	\mathbf{T} encodes the entropy of $P(\mathcal{S} x)$	DCS AUROC

4 Experimental setup

Candidates. We analyze (1) the Last Input Token (LIT) from all layers and (2) LIT from the final layer. Hidden states in the last position of the prompt is what the language-model head projects to logits, encoding the model’s immediate pre-generation context prior to emitting y . Prior work confirms that linear probes on these pre-generation activations recover non-trivial information about the upcoming generation [48]. We evaluate the two variants as reasoning-relevant computation is often concentrated in middle rather than final layers [10]. We additionally evaluate (3) soft tokens with no noise (ST) [81] and with Gumbel noise (STN) [73], and (4) latent thinking (LT) [84] (see Section C.2). For the soft tokens and latent thinking methods, we test varying thinking steps of 1, 16, 32, 64, and 128. We also evaluate exact and pooled output embeddings and the input prompt embedding.

Table layout. Columns follow Table 2. The Output Embedding (OE) block holds two upper-bound references derived from Y , the Exact variant carrying direct semantic knowledge of Y and the Pooled variant averaging embeddings of possible generations. Neither variant is a reference for Minimality or Stability, where output-based encodings carry their own penalties. The Input Embedding (IE) is the prompt embedding, so a candidate failing to outperform IE adds no information beyond the prompt. The Random Vector (RV) is an information-free reference point.

Table 2: Result-table column groups.

Group	Variants
Output Emb.	Exact, Pooled
Candidates	Hidden states, Think. methods
Baselines	IE, RV

Models. We evaluate on the 23 tasks of BBEH [41] using the original benchmark prompt. We use five open-weight LLMs chosen to cover a range of sizes, training procedures, and architectures (Table 3), specifically Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct [30], DeepSeek-R1-Distill-Qwen-32B [17], Skywork-OR1-32B [32, 33], and GPT-OSS-20B [53]. The selection covers dense, sparse-MoE, reasoning-distilled, and RL-trained paradigms.

Table 3: Source LLMs covered by the audit.

Source LLM	Family	Paradigm
Llama-3.1 8B	Dense	Instruct
Llama-3.3 70B	Dense	Instruct
DS-R1-Qwen 32B	Dense	Reasoning-distill
Skywork-OR1 32B	Dense	Native RL
GPT-OSS 20B	Sparse MoE	Adjust effort

Generation. We assume that \mathbf{T} encodes the sufficient statistics of $P(Y | x)$, so the outputs the model assigns high probability to are those consistent with what \mathbf{T} captures about the input. For each prompt, beam search approximates this high-probability region of $P(Y | x)$, and the eight returned sequences of up to 8192 tokens form an empirical

representative slice on which each axiom is evaluated. Beyond maximizing output probability, beam search guarantees distinct candidate outputs, reduces sampling variance, and exposes an empirical distribution over reasoning paths [24].

Probes. We utilize a frozen LLaMA-3.2-1B [30] backbone with a trainable projection that maps thought representations into its token-embedding space, and the discriminator adds a trained classification head. The frozen backbone serves as a shared decoding surface whose learned representations remain compatible with those of independently trained LLMs [59], while the trainable projection learns features specific to each candidate and source model pair. For the Causality measure specifically, the projection is trained on \mathcal{M}_θ 's own output sequences, thus the KL divergence reflects functional substitution rather than generic transferability. Utilizing a shared backbone ensures that the computational cost of evaluating the metric remains constant, independent of the LLM's size. Semantic similarity between outputs for equivalence classes of DCS are computed with Embed-Nemotron-8B [5], the leading text embedding model on MTEB [51] at the time of writing. Training parameters and auxiliary details are in Sections C, D.1, E.3 and E.4.

5 Results

5.1 Per-axiom analysis

Causality. We measure the divergence between the source LLM's continuation distribution and the distribution induced by substituting a candidate TR, reporting the KL divergence (Equation (2)) for which lower is better. Each table in this subsection shows the family-best variant per source LLM, with cells shaded by the $|z|$ -score of the bootstrap-paired gap to IE (blue: above IE, red: below). Every TR yields KL substantially below the information-free RV baseline (Table 4), establishing that the representations encode continuation-relevant information. None of the TRs consistently exceeds the IE reference, indicating that the thought representations carry no additional causal information beyond the prompt.

Table 4: Causality KL (\downarrow , nats).

LLM	OE	LIT	ST	STN	LT	IE	RV
Llama 8B	5.21	5.01	4.96	4.70	5.32	5.36	9.49
Llama 70B	4.56	5.28	4.65	5.08	4.21	4.71	8.93
DS-R1 32B	4.67	4.79	4.45	4.57	4.62	4.50	9.36
Sky-OR1 32B	4.10	4.09	3.90	4.68	4.34	4.08	9.31
GPT-OSS 20B	3.82	4.19	4.00	4.17	3.90	3.78	9.60

Table 5: Minimality Δ_{IB} (\uparrow , nats).

LLM	OE	LIT	ST	STN	LT	IE	RV
Llama 8B	0.37	0.16	0.25	0.24	0.19	0.22	-0.40
Llama 70B	-0.13	-0.30	-0.24	-0.24	-0.30	-0.23	-0.99
DS-R1 32B	0.07	-0.05	0.10	0.10	0.05	0.04	-0.50
Sky-OR1 32B	-0.16	-0.27	-0.13	-0.14	-0.18	-0.21	-0.59
GPT-OSS 20B	-0.22	-0.25	-0.21	-0.20	-0.17	-0.34	-0.30

construction (see Section 3). Among the remaining candidates, results are mixed (Table 5). LIT falls below IE on most source models, soft-thinking candidates at or above IE, and LT almost the same as IE. This indicates that no candidate consistently encodes more output-relevant compression than the prompt already provides. The absolute scale shifts across source LLMs because the cross-entropy decomposition discards a constant tied to each LLM's entropies. Therefore, ranking is comparable only on each row. Figure 2 applies normalization on each LLM to recover a shared axis.

Separability. We test whether **T** encodes per-question identity (Equation (5) and Table 6). Cross-task accuracy is usually near saturation for every candidate, including the IE reference. This shows that representations encode task-related information required to distinguish one task from another. Same-task (or within-task) accuracy, however, illustrates that every candidate except OE is very close to the random baseline. Panel (a) of Figure 2 shows the joint view of both modes for Separability. It is important to note that the highest same-task accuracy OE achieved is 73%. We view OE as an upper-bound to what the candidate representations can encode about the outputs in this metric. However, the collapse is structural rather than a probe-capacity artifact, with Section D.7 tracing it to too few effective dimensions in the within-task geometry for any probe to recover.

Table 6: Same-task acc. (%), \uparrow .

LLM	OE	LIT	ST	STN	LT	IE	RV
Llama 8B	68.8	53.9	54.7	53.5	54.7	54.5	48.9
Llama 70B	72.6	51.6	52.9	52.8	51.4	52.1	49.7
DS-R1 32B	63.5	52.6	54.8	51.8	50.3	53.5	50.3
Sky-OR1 32B	63.4	53.3	54.2	51.8	51.2	54.0	49.9
GPT-OSS 20B	62.4	50.4	50.7	51.8	51.2	49.5	51.0

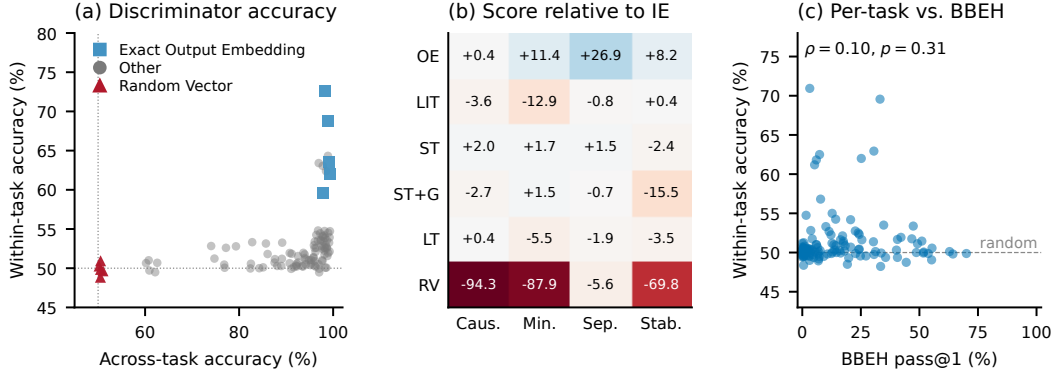


Figure 2: (a) Discriminator accuracy on across- and within-task pairs, one point per (LLM, candidate). (b) Per-axiom score relative to the Input Embedding reference, family-best per LLM averaged across LLMs. (c) Per-task within-task accuracy versus BBEH pass@1 (per-LLM detail in Section D.9).

Table 7: DCS AUROC (\uparrow , $\tau=0.9$).

LLM	OE	LIT	ST	STN	LT	IE	RV
Llama 8B	0.96	0.94	0.94	0.90	0.92	0.93	0.52
Llama 70B	0.89	0.89	0.85	0.84	0.87	0.92	0.50
DS-R1 32B	0.96	0.95	0.95	0.85	0.92	0.94	0.50
Sky-OR1 32B	0.97	0.96	0.95	0.86	0.92	0.93	0.49
GPT-OSS 20B	0.74	0.58	0.55	0.46	0.59	0.59	0.56

Stability. The DCS is the AUROC of a probe that predicts from the candidate whether a question’s beam outputs span more than one semantic equivalence class (Table 7). On the four LLMs that produce non-singleton beam clusters on a sizeable fraction of questions, the candidates clear the Random Vector baseline by a wide margin, the Output Embedding family ranks highest, and the iterative thinking families exhibit decreasing DCS as the step count grows, with the largest decrease for Latent Thinking. The Input Embedding reference matches or exceeds the iterative candidates and is the highest-scoring representation on Llama-70B, reflecting that distributional uncertainty is largely predictable from the question text alone. GPT-OSS-20B is an MoE outlier whose forced beam-generation protocol yields almost no divergent semantic outputs across questions, leaving the DCS estimates uninformative rather than indicative of DCS failure. Per-variant tables across all candidates and step counts, the per-step DCS trajectory, and the projection- and length-control ablations are reported in Sections C.6, D.2, D.4 to D.6 and D.8.

5.2 Per-family analysis

Output Embedding as anchor. The Output Embedding family serves as the upper anchor on axes that reward information about the generated continuation, since these candidates are computed from the output directly. On Causality and Separability the Output Embedding is near the maximum achievable score on every source LLM. In the within-task discrimination setting it is the only family exceeding the random baseline by a meaningful margin, with the Exact variant above the Pooled variant. The Output Embedding is not a meaningful Minimality reference, as its construction violates the chain-rule simplification that the Minimality decomposition assumes. Section D.7 traces this advantage reaching content related to the question encoded along the Output Embedding directions that other candidates sharing its cluster coordinates do not carry.

Cross-axis pattern. Figure 2(b) reports the per-axiom score of each family relative to the Input Embedding reference, averaged across LLMs and computed from the family-best variant per (LLM, axis) pair, with a Random Vector row as a sanity-check anchor. The heatmap recovers the Separability collapse from Section 5.1 and exposes three additional patterns. On within-task Separability, only the Output Embedding clears the random baseline. On Minimality, the iterative thinking families are mostly above the Input Embedding reference whereas Last Input Token is below it, since Last Input Token encodes input-specific information that is not used in the explicit output but is intrinsic to the input. On Stability, the iterative families and Last Input Token are close to the reference, with Soft Thinking with Gumbel noise as the only outlier. The decrease after adding noise is expected and reinforces why noise is a critical step for exploration [18, 73].

The Output Embedding’s Causality cell is within bootstrap noise of the Input Embedding reference, reflecting a substantive empirical observation rather than a normalization artifact. Output Embed-

ding and Input Embedding project into regions of the model’s space that perturb the continuation distribution by comparable amounts, leaving neither dominant on this axis. Taken together, no iterative thinking family achieves the simultaneous strong clustering, wide within-task spread, and high within-question similarity that the framework rewards. The per-family geometric trajectories and the uniform failure pattern across architectures and sizes are reported in Section D.7.

5.3 Joint behavior and synthesis

No candidate thought representation consistently exceeds the Input Embedding reference on any of the four axes when results are averaged across LLMs. The per-LLM advantages reported by Tables 4 to 7 for individual candidates do not generalize (Table 8). In addition, the evaluation does not match what benchmark accuracy alone would predict. A model can score well on a reasoning benchmark while its thought representations fail to discriminate two questions drawn from the same task. The per-task correlation between Separability and downstream accuracy in panel (c) of Figure 2 confirms that this collapse is not explained by task difficulty. The framework’s contribution is a measurement protocol that exposes these gaps at the representation level, without retraining the source model and without dependence on a downstream benchmark.

Takeaway. *No candidate beats the Input Embedding reference on every axis across the tested LLMs, and the iterative thinking variants degrade as the step count grows.*

Table 8: Cells beating IE.

Candidate	Cells/20
Exact	16
Pooled	16
LIT (all, final)	6
ST@1	13
ST@16–128	3–5
STN@1	7
STN@16–128	2–3
LT@1	7
LT@16–128	2–3

6 Conclusion

We introduce an axiomatic evaluation framework for candidate thought representations that runs directly on the source LLM without retraining, instantiated by four measures. We apply this methodology across LLMs that span dense, sparse-MoE, reasoning-distilled, and RL-trained paradigms. The protocol exposes a representational collapse on per-question identity that downstream task accuracy does not reveal. By focusing our evaluation on the representation rather than the subsequent reasoning process, our approach provides a unified foundation. This principled approach readily adapts to new thought representations and axiom measurement designs. Future contributions can propose additional candidate constructions beyond the soft-thinking, latent-thinking, and hidden-state families considered in this work.

Implications. The four axioms serve as explicit optimization targets and diagnostic readouts for thought representations. A new candidate can be scored on each axiom independently, so any change in downstream reasoning accuracy is attributable to the property responsible rather than to an aggregate benchmark score that obscures the cause. An audit identifies which axiom is the binding constraint on an existing representation before any retraining is undertaken, and competing representations are compared directly on a four-metric profile rather than on a single accuracy number. This places future thought representation development on explicit, decomposable, and directly comparable properties that researchers can utilize.

Limitations. The lexical invariance sub-property of Stability is not measured because every candidate evaluated here produces vectors identical across paraphrases by construction, leaving the sub-property trivial in our setting. A measurement protocol awaits candidate constructions that admit non-trivial paraphrase variation. Measurement cost exceeds that of running a single accuracy benchmark, as the protocol requires LLM generations and an additional short probe training step. We position the framework as orthogonal to reasoning benchmark evaluation, where the cost of running the protocol once is small relative to the representation quality information it produces. The empirical audit covers the 23 reasoning tasks of BBEH and five open-weight English-language LLMs spanning a range of architectures, sizes, and training paradigms, and our conclusions do not necessarily hold for multilingual workloads or for generations outside reasoning. Finally, the candidates audited here are all obtainable from a pre-trained LLM without any additional training, and representations satisfying all four axioms may require specifically designed extraction that is trained explicitly to meet those functional requirements. Applying the protocol to such representations is a natural direction for future work.

References

- [1] A. Afzal, F. Matthes, G. Chechik, and Y. Ziser. Knowing before saying: LLM representations encode information about chain-of-thought success before completion. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12791–12806, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.662. URL <https://aclanthology.org/2025.findings-acl.662/>.
- [2] J. Alabi, M. Mosbach, M. Eyal, D. Klakow, and M. Geva. The hidden space of transformer language adapters. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6607, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.356. URL <https://aclanthology.org/2024.acl-long.356/>.
- [3] E. Ameisen, J. Lindsey, A. Pearce, W. Gurnee, N. L. Turner, B. Chen, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. Ben Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson. Circuit Tracing: Revealing Computational Graphs in Language Models. *Transformer Circuits Thread*, 2025. URL <https://transformer-circuits.pub/2025/attribution-graphs/methods.html>.
- [4] A. E. Assadi, I. Chung, R. Solomatin, N. Muennighoff, and K. Enevoldsen. HUME: Measuring the human-model performance gap in text embedding tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=rcmfu1ydAf>.
- [5] Y. Babakhin, R. Osmulski, R. Ak, G. Moreira, M. Xu, B. Schifferer, B. Liu, and E. Oldridge. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks, 2025. URL <https://arxiv.org/abs/2511.07025>.
- [6] L. Bandarkar, B. Muller, P. Yuvraj, R. Hou, N. Singhal, H. Lv, and B. Liu. Layer swapping for zero-shot cross-lingual transfer in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=vQhn4wrQ6j>.
- [7] B. Barak, B. L. Edelman, S. Goel, S. Kakade, E. Malach, and C. Zhang. Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/884baf65392170763b27c914087bde01-Abstract-Conference.html.
- [8] D. Barber and F. Agakov. The IM algorithm: a variational approach to information maximization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems, NIPS’03*, pages 201–208, Cambridge, MA, USA, 2003. MIT Press.
- [9] N. Butt, A. Kwiatkowski, I. Labiad, J. Kempe, and Y. Ollivier. Soft Tokens, Hard Truths. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=9JjKtp8Jmy>.
- [10] Z. Cai, X. Zhu, Y. Dong, Y. He, and S. Arora. T2MLR: Transformer with Temporal Middle-Layer Recurrence. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://openreview.net/forum?id=fQbk1EQWB0>.
- [11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. URL <https://arxiv.org/abs/2104.14294>.
- [12] I. V. M. Cencerrado, A. P. Masdemont, A. G. Hawthorne, D. D. Africa, and L. Pacchiardi. No answer needed: Predicting llm answer accuracy from question-only linear probes, 2026. URL <https://arxiv.org/abs/2509.10625>.

- [13] X. Chen, A. Zhao, H. Xia, X. Lu, H. Wang, Y. Chen, W. Zhang, J. Wang, W. Li, and X. Shen. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning, 2025. URL <https://arxiv.org/abs/2505.16782>.
- [14] D. Chételat, J. Cotnareanu, R. Thompson, Y. Zhang, and M. Coates. InnerThoughts: Disentangling Representations and Predictions in Large Language Models. In Y. Li, S. Mandt, S. Agrawal, and E. Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3862–3870. PMLR, 03–05 May 2025. URL <https://proceedings.mlr.press/v258/chetelat25a.html>.
- [15] H. Conklin, T. Hosking, T. Yi-Chern, J. D. Cohen, S.-J. Leslie, T. L. Griffiths, M. Bartolo, and S. Goldfarb-Tarrant. Learning is Forgetting; LLM Training As Lossy Compression. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=tvD1Qj0GZB>.
- [16] Y. Cui, Z. Dai, B. He, Z. Shi, H. Liu, R. Sun, Z. Liu, Y. Xing, J. Tang, and B. Dumoulin. How Do Latent Reasoning Methods Perform Under Weak and Strong Supervision? In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2602.22441>.
- [17] DeepSeek-AI. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL <https://www.nature.com/articles/s41586-025-09422-z>.
- [18] J. Deng, L. Pang, Z. Wei, S. Xu, Z. Duan, K. Xu, Y. Song, H. Shen, and X. Cheng. Latent Reasoning in LLMs as a Vocabulary-Space Superposition, 2025. URL <https://arxiv.org/abs/2510.15522>.
- [19] C. Dilgren and S. Wiegrefe. Are Latent Reasoning Models Easily Interpretable? In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2604.04902>.
- [20] N. Dragunov, T. Rahmatullaev, E. Goncharova, A. Kuznetsov, and A. Razzhigaev. SONAR-LLM: Autoregressive Transformer that Thinks in Sentence Embeddings and Speaks in Tokens, 2025. URL <https://arxiv.org/abs/2508.05305>.
- [21] C. Du, K. Fu, B. Wen, Y. Sun, J. Peng, W. Wei, Y. Gao, S. Wang, C. Zhang, J. Li, S. Qiu, L. Chang, and H. He. Human-like object concept representations emerge naturally in multimodal large language models. *Nature Machine Intelligence*, 7(6):860–875, June 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01049-z. URL <http://dx.doi.org/10.1038/s42256-025-01049-z>.
- [22] P.-A. Duquenne, H. Schwenk, and B. Sagot. SONAR: Sentence-Level Multimodal and Language-Agnostic Representations, 2023. URL <https://arxiv.org/abs/2308.11466>.
- [23] K. Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, 2019. doi: 10.18653/v1/D19-1006. URL <https://aclanthology.org/D19-1006>.
- [24] E. Fadeeva, M. Goloburda, A. Rubashevskii, R. Vashurin, A. Shelmanov, P. Nakov, M. Sachan, and M. Panov. Don’t Throw Away Your Beams: Improving Consistency-based Uncertainties in LLMs via Beam Search. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=igcQRiVlgu>.
- [25] B. Fatemi, M. Kazemi, A. Tsitsulin, K. Malkan, J. Yim, J. Palowitch, S. Seo, J. Halcrow, and B. Perozzi. Test of Time: A benchmark for evaluating LLMs on temporal reasoning. *arXiv preprint arXiv:2406.09170*, 2024.
- [26] J. Feng, S. Russell, and J. Steinhardt. Monitoring Latent World States in Language Models with Propositional Probes. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=0yvZm2AjUr>.

- [27] S. Feng, G. Fang, X. Ma, and X. Wang. Efficient reasoning models: A survey. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL <https://openreview.net/forum?id=sySqlxj8EB>.
- [28] N. Godey, É. de la Clergerie, and B. Sagot. Anisotropy Is Inherent to Self-Attention in Transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (Volume 1: Long Papers)*, pages 35–48, 2024. URL <https://arxiv.org/abs/2401.12143>.
- [29] S. Goyal, Z. Ji, A. S. Rawat, A. K. Menon, S. Kumar, and V. Nagarajan. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=ph04CRkPdC>.
- [30] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The Llama 3 Herd of Models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- [31] S. Hao, S. Sukhbaatar, D. Su, X. Li, Z. Hu, J. Weston, and Y. Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.
- [32] J. He, J. Liu, C. Y. Liu, R. Yan, C. Wang, P. Cheng, X. Zhang, F. Zhang, J. Xu, W. Shen, S. Li, L. Zeng, T. Wei, C. Cheng, B. An, Y. Liu, and Y. Zhou. Skywork Open Reasoner 1 Technical Report. *arXiv preprint arXiv:2505.22312*, 2025.
- [33] J. He, J. Liu, C. Y. Liu, R. Yan, C. Wang, P. Cheng, X. Zhang, F. Zhang, J. Xu, W. Shen, S. Li, L. Zeng, T. Wei, C. Cheng, Y. Liu, and Y. Zhou. Skywork Open Reasoner Series. <https://capricious-hydrogen-41c.notion.site/Skywork-Open-Reasoner-Series-1d0bc9ae823a80459b46c149e4f51680>, 2025. Notion Blog.
- [34] L. Helff, R. Härle, W. Stammer, F. Friedrich, M. Brack, A. Wüst, H. Shindo, P. Schramowski, and K. Kersting. Activationreasoning: Logical reasoning in latent activation spaces. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=gGJh5AZTG7>.
- [35] V. Herrmann, R. Csordás, and J. Schmidhuber. Measuring In-Context Computation Complexity via Hidden State Prediction. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=X21P8etjwL>.
- [36] J. Hessel, A. Marasović, J. D. Hwang, L. Lee, J. Da, R. Zellers, R. Mankoff, and Y. Choi. Do androids laugh at electric sheep? Humor “understanding” benchmarks from the New Yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.
- [37] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rygGQyrFvH>.
- [38] A. Jolicoeur-Martineau. Less is More: Recursive Reasoning with Tiny Networks, 2025. URL <https://arxiv.org/abs/2510.04871>.
- [39] M. Kazemi, H. Alvari, A. Anand, J. Wu, X. Chen, and R. Soricut. GeomVerse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.
- [40] M. Kazemi, Q. Yuan, D. Bhatia, N. Kim, X. Xu, V. Imbrasaitė, and D. Ramachandran. BoardgameQA: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems*, 36, 2024.
- [41] M. Kazemi, B. Fatemi, H. Bansal, J. Palowitch, C. Anastasiou, S. V. Mehta, L. K. Jain, V. Aglietti, D. Jindal, P. Chen, N. Dikkala, G. Tyen, X. Liu, U. Shalit, S. Chiappa, K. Olszewska, Y. Tay, V. Q. Tran, Q. V. Le, and O. Firat. BIG-bench extra hard. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26473–26501, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1285. URL <https://aclanthology.org/2025.acl-long.1285/>.

- [42] E. Kıcıman, R. Ness, A. Sharma, and C. Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.
- [43] Y. Koishekenov, A. Lipani, and N. Cancedda. Encode, Think, Decode: Scaling test-time reasoning with recursive latent thoughts, 2025. URL <https://arxiv.org/abs/2510.07358>.
- [44] L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=VD-AYtP0dve>.
- [45] Y. Li, J. Chen, F. Wu, J. Yu, H. Qi, W. Xuan, H. Zhao, P. Nie, D. Jin, and X. Tang. Learning Multi-step Reasoning via Persistent Latent State Propagation. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://openreview.net/forum?id=Dcv4B1UCuW>.
- [46] Z. Li, X. Bai, K. Chen, Y. Li, J. Yang, C. Lin, and M. Zhang. Dynamics Within Latent Chain-of-Thought: An Empirical Study of Causal Structure. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2602.08783>.
- [47] A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, and L. F. Abbott. Optimal Degrees of Synaptic Connectivity. *Neuron*, 93(5):1153–1164.e7, 2017. doi: 10.1016/j.neuron.2017.01.030.
- [48] W. Lugoloobi, T. Foster, W. Bankes, and C. Russell. LLMs Encode Their Failures: Predicting Success from Pre-Generation Activations. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2602.09924>.
- [49] F. V. Massoli, A. Kuzmin, and A. Behboodi. Reasoning as Compression: Unifying Budget Forcing via the Conditional Information Bottleneck. In *The 1st Workshop on Scaling Post-training for LLMs*, 2026. URL <https://openreview.net/forum?id=98sbP0T8ck>.
- [50] P. Mondorf and B. Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey. In *First Conference on Language Modeling (COLM)*, 2024. URL <https://openreview.net/forum?id=Lmjgl2n11u>.
- [51] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. MTEB: Massive Text Embedding Benchmark, 2023. URL <https://arxiv.org/abs/2210.07316>.
- [52] A. Nie, Y. Zhang, A. S. Amdekar, C. Piech, T. B. Hashimoto, and T. Gerstenberg. MoCa: Measuring human-language model alignment on causal and moral judgment tasks. *Advances in Neural Information Processing Systems*, 36, 2024.
- [53] OpenAI. gpt-oss-120b & gpt-oss-20b Model Card, 2025. URL <https://arxiv.org/abs/2508.10925>.
- [54] K. Park, Y. J. Choe, and V. Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR, 2024. URL <https://proceedings.mlr.press/v235/park24c.html>.
- [55] Prasanth. Enforcing Logical Invariance in Large Language Models via Symmetry Pair Training. In *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, 2026. URL <https://openreview.net/forum?id=aZFS8rc6Bf>.
- [56] S. Recanatesi, M. Farrell, M. Advani, T. Moore, G. Lajoie, and E. Shea-Brown. Dimensionality compression and expansion in Deep Neural Networks, 2019. URL <https://arxiv.org/abs/1906.00443>.
- [57] M. Rizvi-Martel and M. Mosbach. The Illusion of Superposition in Latent CoT via Soft Thinking. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://openreview.net/forum?id=FvPx9Nzvnw>.
- [58] S. Sahoo, A. Chadha, V. Jain, and D. Chaudhary. When Shallow Wins: Silent Failures and the Depth-Accuracy Paradox in Latent Reasoning. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2603.03475>.

- [59] S. Salhan, E. Zhou, and P. Buttery. Do Monolingual Language Models Learn Cross-Lingual Universal Conceptual Representations? In *ICLR 2026 Workshop on Unifying Concept Representation Learning*, 2026. URL <https://openreview.net/forum?id=frKa6uj0yE>.
- [60] E. Sánchez, B. Alastruey, C. Ropers, P. Stenetorp, M. Artetxe, and M. R. Costa-jussà. Linguini: A benchmark for language-agnostic linguistic reasoning. *arXiv preprint arXiv:2409.12126*, 2024.
- [61] K. Shah, N. Dikkala, X. Wang, and R. Panigrahy. Causal language modeling can elicit search and reasoning capabilities on logic puzzles. *arXiv preprint arXiv:2409.10502*, 2024.
- [62] C. Shani, L. Soffer, D. Jurafsky, Y. LeCun, and R. Shwartz-Ziv. From Tokens to Thoughts: How LLMs and Humans Trade Compression for Meaning, 2025. URL <https://arxiv.org/abs/2505.17117>.
- [63] Z. Shen, H. Yan, L. Zhang, Z. Hu, Y. Du, and Y. He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 677–693, Suzhou, China, Nov. 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.36. URL <https://aclanthology.org/2025.emnlp-main.36/>.
- [64] D. Sheshanarayana, R. S. Pal, M. Sinha, and T. Dasgupta. Thinking in Latents: Adaptive Anchor Refinement for Implicit Reasoning in LLMs. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2603.15051>.
- [65] O. Skean, M. R. Arefin, D. Zhao, N. N. Patel, J. Naghiyev, Y. LeCun, and R. Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=WGXb7UdvTX>.
- [66] Y. Sui, Y.-N. Chuang, G. Wang, J. Zhang, T. Zhang, J. Yuan, H. Liu, A. Wen, S. Zhong, N. Zou, H. Chen, and X. Hu. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL <https://openreview.net/forum?id=HvoG8SxggZ>.
- [67] Q. Sun, M. Pickett, A. K. Nain, and L. Jones. Transformer Layers as Painters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25219–25227, 2025. doi: 10.1609/aaai.v39i24.34708. URL <https://ojs.aaai.org/index.php/AAAI/article/view/34708>.
- [68] L. team, L. Barrault, P.-A. Duquenne, M. Elbayad, A. Kozhevnikov, B. Alastruey, P. Andrews, M. Coria, G. Couairon, M. R. Costa-jussà, D. Dale, H. Elsahar, K. Heffernan, J. M. Janeiro, T. Tran, C. Ropers, E. Sánchez, R. S. Roman, A. Mourachko, S. Saleem, and H. Schwenk. Large Concept Models: Language Modeling in a Sentence Representation Space, 2024. URL <https://arxiv.org/abs/2412.08821>.
- [69] G. Tyen, H. Mansoor, P. Chen, T. Mak, and V. Cărbune. LLMs cannot find reasoning errors, but can correct them! *arXiv preprint arXiv:2311.08516*, 2023.
- [70] W. Wang and F. Reid. Tiny Recursive Reasoning with Mamba-2 Attention Hybrid. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2602.12078>.
- [71] C. Wendler, V. Veselovsky, G. Monea, and R. West. Do llamas work in english? on the latent language of multilingual transformers. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL <https://aclanthology.org/2024.acl-long.820/>.
- [72] C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Naidu, et al. LiveBench: A challenging, contamination-free LLM benchmark. *arXiv preprint arXiv:2406.19314*, 2024.

- [73] J. Wu, J. Lu, Z. Ren, G. Hu, Z. Wu, D. Dai, and H. Wu. LLMs are Single-threaded Reasoners: Demystifying the Working Mechanism of Soft Thinking. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=ASLu0oP78o>.
- [74] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018.
- [75] Y. Yamada, Y. Bao, A. K. Lampinen, J. Kasai, and I. Yildirim. Evaluating spatial understanding of large language models. *arXiv preprint arXiv:2310.14540*, 2023.
- [76] D. Ye, M. Loffgren, O. Kotadia, and L. Wong. Mechanistic Evidence for Faithfulness Decay in Chain-of-Thought Reasoning. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2602.11201>.
- [77] X. Ye, Y. Mao, J. Zhang, Y. Liu, L. Hao, F. Wu, Z. Li, Y. Liao, Z. Wang, Y. Wu, Z. Liu, Z. Yin, L. Yuan, P. Torr, H. Sun, X. Zeng, M. Wang, L. Cong, S. Gao, and X. Tang. LatentChem: From Textual CoT to Latent Thinking in Chemical Reasoning. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2602.07075>.
- [78] B. Zeng, H. Li, S. Song, Y. Wang, Z. He, X. Wang, and Z. Lin. PonderLM-2: Pretraining LLM with Latent Thoughts in Continuous Space, 2025. URL <https://arxiv.org/abs/2509.23184>.
- [79] L.-M. Zhan, B. Liu, Y. Feng, C. Xie, J. Cao, and X.-M. Wu. REAL: Reading Out Transformer Activations for Precise Localization in Language Model Steering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=P38RYdkFLI>.
- [80] J. Zhang, L. Jain, Y. Guo, J. Chen, K. L. Zhou, S. Suresh, A. Wagenmaker, S. Sievert, T. Rogers, K. Jamieson, et al. Humor in AI: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *arXiv preprint arXiv:2406.10522*, 2024.
- [81] Z. Zhang, X. He, W. Yan, A. Shen, C. Zhao, and X. E. Wang. Soft Thinking: Unlocking the Reasoning Potential of LLMs in Continuous Concept Space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL <https://openreview.net/forum?id=ByQdHPGKgU>.
- [82] Y. Zhou, Y. Wang, X. Yin, S. Zhou, and A. Zhang. The Geometry of Reasoning: Flowing Logics in Representation Space. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=ixr5Pcabq7>.
- [83] D. Zhu, L. T. H. Khanh, A. Y.-M. Cheung, M.-C. Yue, and V. A. Nguyen. Exploring Diverse Generation Paths via Inference-time Stiefel Activation Steering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=v0Q0VSVptq>.
- [84] J. Zou, X. Yang, R. Qiu, G. Li, K. Tieu, P. Lu, K. Shen, H. Tong, Y. Choi, J. He, J. Zou, M. Wang, and L. Yang. Latent Collaboration in Multi-Agent Systems, 2025. URL <https://arxiv.org/abs/2511.20639>.
- [85] J. Zou, Y. Xiong, and Y. Liu. The theoretical benefits and limitations of latent chain-of-thought reasoning, 2026. URL <https://openreview.net/forum?id=q7Nhu2Fw11>.

Appendix Table of Contents

A	Extended Related Work	17
B	Formal Analysis of the Axiomatic System	17
B.1	Consistency (Existence)	18
B.2	Independence (Non-Redundancy)	18
B.3	Completeness (Sufficiency)	21
C	Training Details	21
C.1	LLM Data Generation	21
C.2	Soft Thinking and Latent Thinking Generation	21
C.3	Causality Evaluation Protocol	22
C.4	Minimality Probe Architecture and Training	22
C.5	Discriminator Architecture and Training	22
C.6	Stability Sub-Properties and DCS Diagnostics	23
C.7	Reproducibility and Code Release	23
D	Additional Analysis	24
D.1	Bootstrap Confidence Intervals	24
D.2	Detailed per-axiom results	25
D.3	Distributional Views of Causality	27
D.4	Causality with the Output-Reconstruction Projection	27
D.5	Information-Bottleneck Decomposition for the Minimality Metric	28
D.6	Length Sensitivity of the Causality Metric	30
D.7	Geometric Analysis of Representational Collapse	32
D.8	Probe Capacity Ablation	34
D.9	Relationship to Downstream Task Accuracy	35
E	Dataset and Benchmark Details	35
E.1	BBEH Task Coverage	35
E.2	BBEH Answer Extraction	36
E.3	Output-Length Statistics	36
E.4	BBEH Accuracy of the Source LLMs	37
E.5	Qualitative Beam Outputs and Embedder Similarity	38
F	Broader Impacts	40
G	Earlier Formulations	40
G.1	Discriminator-Based DCS Evaluation Protocol	40
G.2	Causality with the Discriminator-Trained Projection	41
G.3	Cross-Entropy Proxy for Minimality	41

G.4 Qualitative Probe Reconstruction	42
G.5 Discriminator-Based DCS Results	43

A Extended Related Work

Surveys and theoretical analyses. A growing set of surveys systematically categorises the emerging space of latent and continuous reasoning methods. Chen et al. [13] provide a comprehensive taxonomy organising approaches along two axes: token-wise horizontal methods that replace discrete tokens with continuous counterparts, and layer-wise vertical methods that propagate latent state across transformer depth. Zou et al. [85] offer a theoretical characterisation of the fundamental exploration–execution trade-off between discrete and continuous reasoning, proving that discrete chain-of-thought is forced into a high-certainty regime while continuous representations enable exploration at the cost of amplified noise on computational tasks; they introduce the Symbolic Index as a scalar measure of decisional certainty that governs this trade-off. Mondorf and Plank [50] survey evaluation practices for reasoning in LLMs, arguing that downstream accuracy conflates reasoning quality with surface-level pattern matching. Sui et al. [66] and Feng et al. [27] survey efficient reasoning from the angle of compute cost and length budgets. Together these works catalogue methods and their downstream trade-offs, whereas our framework evaluates the intrinsic representational quality of the intermediate state independently of the decoding strategy or task accuracy.

Latent world models, monitoring, and localization. Beyond reasoning benchmarks, a parallel line of work demonstrates that LLMs build structured internal models of the world as they process language. Feng et al. [26] show that latent world states can be extracted as structured propositions via propositional probes, providing direct empirical grounding for our Separability axiom: the hidden states must contain linearly decodable information about the semantic state of the world. Du et al. [21] find that human-like object concept representations emerge naturally in multimodal LLMs without explicit supervision, consistent with the view that the internal geometry supports rich semantic structure. Localizing where this structure resides has received substantial attention. Bandarkar et al. [6] show that swapping transformer layer ranges transfers cross-lingual knowledge across models, and Alabi et al. [2] find that adaptation concentrates in early-to-middle adapter layers, both implying semantically meaningful computations are localized at identifiable depths. Sun et al. [67] show that middle layers can be reordered or skipped with minimal accuracy cost, consistent with mid-depth representations sharing structural properties across depth. Ameisen et al. [3] trace computational circuits through residual stream contributions and demonstrate that feature-level computations shift dynamically across layers depending on the prompt, directly motivating why a fixed layer selection strategy for thought representation extraction is insufficient and why our framework tests multiple candidate extraction points.

Sentence-level and continuous-space representations. A separate family of methods pursues meaning representation at the sentence level rather than the token level. Large Concept Models [68] reformulate language modelling as prediction in a sentence embedding space rather than over token vocabularies, leveraging SONAR [22] as the shared multilingual sentence encoder. SONAR-LLM [20] extends this further with an autoregressive transformer that reasons in sentence embedding space and decodes back to tokens, operating at a granularity coarser than individual tokens but finer than document-level representations. These methods design representations at a fixed semantic granularity chosen a priori. Our framework is granularity-agnostic and instead asks whether representations produced at any granularity by an existing model satisfy the four axioms, making it applicable to sentence-level methods as a diagnostic tool as well as to the token-level iterative candidates we evaluate here.

B Formal Analysis of the Axiomatic System

In this section, we provide the theoretical justification for the four axiomatic properties defined in Section 3. We use the same notation as the main text. Appendix-specific symbols are introduced where they first appear. We formally establish the logical framework of the set $\mathcal{P} = \{\text{Causality, Minimality, Separability, Stability}\}$.

B.1 Consistency (Existence)

We first analyze the logical consistency of \mathcal{P} . In empirical, frozen language models trained via next-token prediction, there exists an inherent tension between Causality (which requires \mathbf{T} to perfectly mimic the lexically-entangled predictive state of \mathcal{M}_θ) and Stability/Minimality (which require \mathbf{T} to discard lexical information).

Proposition B.1. *The set of properties \mathcal{P} is logically consistent. There exists an idealized model configuration \mathcal{M}^* and representation \mathbf{T}^* that satisfies all four axioms simultaneously.*

Proof. To prove strict logical consistency, we construct an idealized model \mathcal{M}^* whose intermediate representations perfectly disentangle semantics from syntax. Let $\mathcal{S}_{\mathcal{M}}$ be a discrete semantic space. Let \mathcal{M}^* be constructed such that its forward pass strictly factors through a one-hot semantic bottleneck before projecting to vocabulary logits, a construction closely related to the discrete latent bottlenecks used in recent latent reasoning architectures [18]. Let \mathbf{T}^* be the one-hot activation at this bottleneck.

1. **Causality:** Because \mathcal{M}^* explicitly generates Z conditioned entirely on this semantic bottleneck, substituting the explicit sequence Y with \mathbf{T}^* yields identical downstream logits, so $D_{\text{KL}}(P_{\mathcal{M}^*}(Z | Y) \| P_{\mathcal{M}^*}(Z | \mathbf{T}^*)) = 0$, matching the formal causality criterion of Section 3.
2. **Minimality:** By construction Y depends on X only through \mathbf{T}^* , so $X \rightarrow \mathbf{T}^* \rightarrow Y$ forms a Markov chain and \mathbf{T}^* is sufficient with $I(\mathbf{T}^*; Y) = I(X; Y)$. For any deterministic sufficient statistic $T = T(X)$, $I(X; T) = H(T) \geq I(T; Y) = I(X; Y)$, with equality precisely when T is a bijective function of \mathbf{T}^* . Hence \mathbf{T}^* attains the minimum of $I(X; T)$ over all sufficient T and saturates the information-bottleneck objective. This is an existence claim about the idealized configuration \mathcal{M}^* . The empirical question of how closely a given frozen \mathcal{M}_θ approaches this construction is what Section 5 measures.
3. **Separability:** Distinct semantic intents map to orthogonal one-hot vectors, perfectly resolvable by any linear projection $\phi \in \mathcal{H}$.
4. **Stability:** Semantically equivalent outputs map deterministically to the same one-hot bottleneck, ensuring perfect lexical invariance at the output level. Mode Collapse Resistance holds vacuously in \mathcal{M}^* : by construction, each input x induces a deterministic semantic class, so $P(Y | x)$ is a point mass with $H_x = 0$. There is no distributional uncertainty to represent or collapse, and the one-hot encoding faithfully reflects this zero-entropy output distribution.

Because this configuration satisfies \mathcal{P} without contradiction, the system is logically consistent. \square

Remark B.2. For an arbitrary frozen model \mathcal{M}_θ with lexical-semantic entanglement, the four axioms define a Pareto-optimal frontier rather than a strict intersection. In empirical settings \mathbf{T} acts as an ϵ -optimal approximation that bounds the tradeoff between causal exactness and semantic stability, consistent with prior evidence that intermediate layers balance information compression against signal preservation [65, 62]. We quantify this frontier empirically in Section 5.

B.2 Independence (Non-Redundancy)

Proposition B.3. *The axioms in \mathcal{P} are logically independent. No property can be derived solely from the conjunction of the others.*

Proof. We prove independence by constructing a theoretical counter-model for each property $p \in \mathcal{P}$ that satisfies $\mathcal{P} \setminus \{p\}$ but violates p . Let \mathbf{T}^* be an optimal representation satisfying all axioms in an idealized network.

Case 1: Violation of Causality. Construct $\mathbf{T} = \pi(\mathbf{T}^*)$, where π is a fixed, random permutation of the coordinate dimensions of \mathbf{T}^* . Throughout this case the generation head of \mathcal{M}^* is held frozen, while the Separability probe $\phi \in \mathcal{H}$ may be re-fit on the new representation. This asymmetry reflects how each axiom is operationalised in Section 3 (probes test latent geometry and are trained per representation, whereas Causality tests substitution into a fixed model whose head is never retrained).

- **Satisfies Minimality:** A permutation is a bijection on coordinates, so $I(X; \mathbf{T}) = I(X; \pi(\mathbf{T}^*)) = I(X; \mathbf{T}^*)$ (mutual information is invariant under invertible deterministic transformations), and the information-bottleneck objective is preserved exactly.
- **Satisfies Separability:** Permutations are orthogonal transformations and preserve inner products. Since the Separability probe $\phi \in \mathcal{H}$ is trained on whichever representation is presented, it can absorb π^{-1} in its first linear layer at no cost in norm or expressive power, recovering the separability of \mathbf{T}^* .
- **Satisfies Stability:** The permutation is deterministic and depends only on coordinate indices, not on the output, so $\mathbf{T}_{y_1}^* = \mathbf{T}_{y_2}^* \implies \pi(\mathbf{T}_{y_1}^*) = \pi(\mathbf{T}_{y_2}^*)$, preserving lexical invariance.
- **Violates Causality:** The frozen generation head expects specific semantic features at specific indices. Because the head is not retrained, it cannot absorb π^{-1} , so the permuted representation activates wrong indices and is decoded as a different (typically incorrect) semantic class. Consequently $P_{\mathcal{M}^*}(Z | \mathbf{T})$ diverges from $P_{\mathcal{M}^*}(Z | Y)$ and the Causality KL spikes despite the geometry being preserved.

Case 2: Violation of Minimality. Construct $\mathbf{T} = [\mathbf{T}^*, \xi(X)]$, where $\xi(X)$ is a high-entropy nuisance vector (e.g., a hash of the exact character sequence of X) that is independent of the semantic class. We assume the downstream generation head structurally ignores the dimensions of $\xi(X)$ (e.g., via an attention mask that zeroes them out), so that nuisance information is appended without affecting prediction.

- **Satisfies Causality:** Since the head ignores $\xi(X)$, the conditional distribution $P_{\mathcal{M}^*}(Z | \mathbf{T})$ depends only on the \mathbf{T}^* component and matches $P_{\mathcal{M}^*}(Z | \mathbf{T}^*) = P_{\mathcal{M}^*}(Z | Y)$ exactly.
- **Satisfies Separability:** A linear classifier in \mathcal{H} can place zero weight on the $\xi(X)$ coordinates and recover the same decision boundary as on \mathbf{T}^* , so the Separability margin δ is preserved.
- **Satisfies Stability:** Stability requires lexical invariance among sibling outputs drawn from the same input x . Because ξ is a function of the input alone, all sibling outputs $y_i \sim_{sem} y_j$ generated from the same x inherit an identical $\xi(x)$, so their representations agree.
- **Violates Minimality:** Independence of $\xi(X)$ from \mathbf{T}^* gives $I(X; \mathbf{T}) = I(X; \mathbf{T}^*) + H(\xi(X) | \mathbf{T}^*) = I(X; \mathbf{T}^*) + H(\xi(X))$, so $I(X; \mathbf{T})$ exceeds the bottleneck minimum by $\Theta(H(\xi(X)))$ bits, which can be made arbitrarily large by lengthening the hash. The representation therefore carries strictly more information about X than any sufficient statistic.

Case 3: Violation of Separability. *Scope of \mathcal{H} for this case.* The Independence argument below analyzes Separability under the trainable component of \mathcal{H} and treats the frozen LLM in the empirical realization of f_{disc} (Section C.5) as a fixed feature kernel outside the optimized class. This scoping aligns with the linear representation hypothesis [54], under which high-level semantic concepts are linearly decodable from frozen LLM activations. Because empirical f_{disc} has access to LLM-induced nonlinear features that the theoretical \mathcal{H} does not, the empirical test is more permissive. An empirical Separability failure implies a theoretical violation under \mathcal{H} , but the converse does not hold. The construction below therefore establishes Independence at the theoretical level and a conservative refinement at the empirical level.

Fix two semantic classes $s \in \{0, 1\}$. Encode the representations such that class 0 maps to $R_0 = \{(0, 0, \mathbf{0}), (1, 1, \mathbf{0})\} \subset \{0, 1\}^d$ and class 1 maps to $R_1 = \{(0, 1, \mathbf{0}), (1, 0, \mathbf{0})\}$, where $\mathbf{0}$ is the zero vector in $\{0, 1\}^{d-2}$. The four vertices are the XOR configuration of the unit square padded into the ambient d -cube. The deep layers of \mathcal{M}^* are parameterized to compute the XOR of the first two coordinates before decoding.

- **Satisfies Causality:** The deep network applies the parity function before generation, so \mathbf{T} matches the intermediate distribution the deep head expects and $P_{\mathcal{M}^*}(Z | \mathbf{T}) = P_{\mathcal{M}^*}(Z | Y)$.
- **Approximately satisfies Minimality:** The representation occupies four configurations for two classes, contributing a fixed overhead of one bit above the minimum-sufficient

encoding of the binary class label. This is a constant, content-independent cost: $I(X; \mathbf{T}) = I(X; \mathbf{T}^*) + 1$ bit, where the extra bit encodes within-class position but carries no additional information about X beyond the class label. The violation of strict Minimality is bounded and does not grow with the semantic complexity of X ; for the purposes of the independence argument, this ϵ -deviation from strict Minimality is the same approximation acknowledged in the Remark following Section B.1.

- **Satisfies Stability:** Within-class pairs satisfy $\mathbf{E}_{ij} = 1$ in the DCS equivalence matrix (Section 3) and the corresponding latent representations differ by Hamming distance at most two. The operational \approx tolerance the Stability axiom permits is therefore satisfied.
- **Violates Separability:** The bounded hypothesis class \mathcal{H} defined in Section 3 consists of a linear projection into an embedding space followed by a linear head [54]. The composition of two linear maps is itself linear, so every $\phi \in \mathcal{H}$ realises a single halfspace decision $\text{sign}(w^\top \mathbf{T} + b)$ over \mathbf{T} . We show no such halfspace separates R_0 from R_1 . Requiring positive decisions on R_0 forces $b > 0$ and $w_1 + w_2 + b > 0$, while requiring negative decisions on R_1 forces $w_1 + b < 0$ and $w_2 + b < 0$. Summing the last two yields $w_1 + w_2 < -2b$, which contradicts $w_1 + w_2 + b > 0$ whenever $b > 0$. The classes are XOR-configured and provably outside the linear hypothesis class. Beyond non-realizability, gradient-based training of any such ϕ on parity targets needs sample complexity scaling as $\Omega(d^2)$ before any nontrivial correlation emerges [7], so even the relaxed empirical version of Separability fails. No $\phi \in \mathcal{H}$ therefore attains the required margin δ .

Case 4: Violation of Stability. Under Definition 1, g is a deterministic function of x alone, which means every beam drawn from the same input shares the same representation and Lexical Invariance is satisfied trivially. To show that Stability is not logically entailed by the remaining three axioms, we therefore work in the natural extended domain where the generator may depend on the generation trajectory: $g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{T}$. This extension captures stochastic extraction methods (e.g., beam-specific random seeds) and allows Lexical Invariance to be a nontrivial requirement. The independence argument below shows that even in this richer domain the other three axioms do not force Lexical Invariance. \mathbf{T}_y denotes the representation value associated with output y in this counter-model; \mathbf{T}^* remains the input-determined component from Definition 1, and $\alpha_{lex}(y)$ is the output-dependent augmentation whose role this case isolates.

Construct $\mathbf{T}_y = [\mathbf{T}^*, \alpha_{lex}(y)]$, where $\alpha_{lex}(y) \in \{0, \lambda\}$ is a single massive scalar that flags a trivial lexical feature of the output sequence y , for concreteness the presence of a trailing space. We assume the downstream generation head structurally drops this augmented index. This assumption isolates Stability cleanly. Causality, Minimality, and Separability are all preserved precisely because the offending coordinate is suppressed at inference. The case is informative for representations consumed by models without such suppression which is exactly the empirical setting we evaluate, where the frozen LLM has no architectural mechanism to know which coordinate carries lexical noise. Under those conditions, even a single uncontrolled lexical-flag dimension destroys output-level invariance.

- **Satisfies Causality:** The augmented index is dropped prior to prediction, so the causal logits remain identical to those of \mathbf{T}^* .
- **Satisfies Minimality:** By construction $Y \perp X \mid \mathbf{T}^*$, so $\alpha_{lex}(Y) \perp X \mid \mathbf{T}^*$ and $I(X; \mathbf{T}) = I(X; \mathbf{T}^*)$ exactly. Minimality is preserved without overhead.
- **Satisfies Separability:** Any $\phi \in \mathcal{H}$ can place zero weight on the augmented coordinate and recover the same decision boundary as on \mathbf{T}^* , so the Separability margin δ is preserved.
- **Violates Stability:** Two semantically equivalent outputs $y_i \sim_{sem} y_j$ that differ only in whether they end in a trailing space acquire $\alpha_{lex}(y_i) \neq \alpha_{lex}(y_j)$. Their representations are then separated by Euclidean distance λ , which can be made arbitrarily large, fundamentally destroying the output-level lexical invariance the axiom requires.

Therefore, no axiom logically entails another. □

B.3 Completeness (Sufficiency)

Proposition B.4. *The set of properties \mathcal{P} is complete with respect to the definition of a Functional Thought Representation. Any representation \mathbf{T} strictly satisfying \mathcal{P} establishes a well-defined functional isomorphism between the reachable latent space \mathcal{T} and the reachable semantic manifold $\mathcal{S}_{\mathcal{M}}$.*

Proof. For each semantic class $s \in \mathcal{S}_{\mathcal{M}}$, choose a high-probability output $y \in \mathcal{Y}$ with $\Phi(y) = s$, and define $\psi : \mathcal{S}_{\mathcal{M}} \rightarrow \mathcal{T}$ by $\psi(s) = \mathbf{T}_y$. We verify that ψ is a bijection onto \mathcal{T} .

- 1. Well-Defined (Stability and Minimality):** Fix a representative y with $\Phi(y) = s$ to define $\psi(s) = \mathbf{T}_y$. For any other y' with $\Phi(y') = s$, we need $\mathbf{T}_{y'} = \mathbf{T}_y$. If y' is a sibling output drawn from the same input x as y , strict Stability (Lexical Invariance) gives $\mathbf{T}_{y'} = \mathbf{T}_y$ directly. If y' is drawn from a different input, strict Minimality forces \mathbf{T} to be a function of s alone, so $\mathbf{T}_{y'} = \mathbf{T}_y$ again. Hence $\psi(s)$ is well-defined.
- 2. Injective (Separability):** For distinct classes $s_1 \neq s_2$ with disjoint high-probability semantic spaces, Separability supplies some $\phi \in \mathcal{H}$ that resolves $\psi(s_1)$ from $\psi(s_2)$ with margin $\delta > 0$. In particular $\psi(s_1) \neq \psi(s_2)$, so ψ is injective.
- 3. Surjective (Minimality and the generator definition):** The reachable latent \mathcal{T} is defined as the image $g(\mathcal{X})$, so every $\mathbf{T} \in \mathcal{T}$ equals $g(x)$ for some input x , which induces a semantic class $\Phi(\mathcal{M}(x)) \in \mathcal{S}_{\mathcal{M}}$. Minimality (in the idealized limit clarified in the Remark below) additionally forces \mathbf{T} to carry no information about X beyond a sufficient statistic for the semantic class, because any excess would raise $I(X; \mathbf{T})$ above the bottleneck minimum. Consequently $g(x)$ coincides with \mathbf{T}_y whenever $\Phi(\mathcal{M}(x)) = \Phi(y)$, so $\mathbf{T} = \psi(\Phi(\mathcal{M}(x)))$ and ψ surjects onto \mathcal{T} .
- 4. Functional Equivalence (Causality):** Causality upgrades the set-theoretic bijection to a functional one. Each $\mathbf{T} = \psi(s)$ substitutes for the explicit output sequence Y inside \mathcal{M}_{θ} and induces the same downstream distribution $P_{\theta}(Z | \mathbf{T}) = P_{\theta}(Z | Y)$.

Because ψ is well-defined, injective, surjective onto \mathcal{T} , and functionally equivalent to the explicit generative pathway, it realizes a bijection $\mathcal{T} \cong \mathcal{S}_{\mathcal{M}}$. The four axioms therefore fully determine the structural isomorphism, and no further axiom is required. \square

Remark B.5. The term "completeness" here refers to adequacy of the axiom set in the categoricity sense. No additional axiom is required beyond \mathcal{P} to pin down the functional isomorphism up to relabeling of semantic classes. The bijection is strict in the idealized limit where Stability holds with equality and the Separability margin is unbounded. Under the empirical \approx tolerance of Stability and the finite δ -margin of Separability, the bijection becomes approximate and is quantified in Section 5.

C Training Details

C.1 LLM Data Generation

The generator models, benchmark, beam count, and maximum generation length are defined in Section 4; all models are loaded in their native precision (bfloat16). From each beam we extract hidden states at every decoding step across all layers, and take the last token of the prefill step as the primary thought representation (position -1 , decoding step 0, all layers). The per-model layer count is 33, 81, 65, 65, and 25 for Llama-3.1 8B, Llama-3.3 70B, DeepSeek-R1-Distill-Qwen 32B, Skywork-OR1 32B, and GPT-OSS 20B respectively.

C.2 Soft Thinking and Latent Thinking Generation

Soft Thinking (No Noise) replaces discrete decoding with the weighted combination of token embeddings $\mathbf{T} = \sum_v p_v \mathbf{e}_v$ where $p_v = \text{softmax}(z)_v$ and \mathbf{e}_v is the token embedding [81]. Soft Thinking with Gumbel Noise applies $\hat{p}_v \propto \exp((z_v + \epsilon_v)/\tau)$ with $\epsilon_v \sim \text{Gumbel}(0, 1)$ and temperature $\tau = 1.0$ [73]. Latent Thinking applies recurrent hidden state updates using a protocol similar to

COCONUT [31], as implemented in Zou et al. [84]. For all iterative methods, we evaluate steps $s \in \{1, 16, 32, 64, 128\}$.

Determinism under stochastic extraction. The idealized mapping g in Definition 1 is defined deterministically. Extraction methods that include stochastic components, such as Gumbel noise in Soft Thinking, fix a global random seed across all evaluations, so each input deterministically produces the same \mathbf{T} . The Gumbel perturbation is therefore a property of the extraction procedure rather than a source of representational randomness, and g remains well-defined as a deterministic function of x .

C.3 Causality Evaluation Protocol

The causality evaluation does not involve training. Given a test problem with $K = 8$ beams $\{y_k\}_{k=1}^K$ and corresponding thought representations $\{\mathbf{T}_k\}_{k=1}^K$, we compute:

1. **Prefix–suffix split:** Let y_{suf} denote the last 50 tokens of y_k (the answer suffix) and y_{pre} the preceding tokens (the reasoning prefix), so that $y_k = [y_{\text{pre}}, y_{\text{suf}}]$. Beams shorter than 51 tokens are excluded from this evaluation.
2. **Baseline distribution:** Run the evaluation backbone on the explicit token embeddings of y_{pre} to obtain $P(y_{\text{suf}} | y_{\text{pre}})$ at the y_{suf} positions via teacher forcing.
3. **Intervened distribution:** Replace the prefix embeddings with the projected thought representation \mathbf{T}_k to obtain $P(y_{\text{suf}} | \mathbf{T}_k)$ at the same positions. The projection is taken from an output-reconstruction projection trained on the source LLM’s output sequences via cross-entropy loss (same training splits as the Minimality probe); see Section D.4 for the ablation comparing this choice against the discriminator projection. The evaluation backbone is LLaMA-3.2-1B [30] with its parameters held fixed; the projection learns to map \mathbf{T} into the backbone’s embedding space so that the KL measures whether \mathbf{T} induces the same functional generative effect on y_{suf} as y_{pre} , rather than raw cross-model transferability.
4. **KL divergence:** Compute $D_{\text{KL}}(P(y_{\text{suf}} | y_{\text{pre}}) \| P(y_{\text{suf}} | \mathbf{T}_k))$ averaged over y_{suf} positions.

Position indices under tiling. Because all candidates are tiled to the same substitution length (128 positions), the absolute position indices seen by y_{suf} under LLaMA-3.2-1B’s Rotary Position Embeddings are identical across every candidate for a given problem. Any KL inflation arising from the difference between the tiled length and the natural prefix length $|y_{\text{pre}}|$ is therefore a constant offset shared by every candidate and does not affect their relative ordering. This is performed on the held-out test split (Section E), minus beams excluded by the length gate above. We report the mean KL across all valid beam–problem pairs.

C.4 Minimality Probe Architecture and Training

Two probes are trained under the same architecture, an output-reconstruction probe estimating $\text{CE}(Y | \mathbf{T})$ and a conditional input probe estimating $\text{CE}(X | Y, \mathbf{T})$. Both probes tile \mathbf{T} to 128 positions. The probe maps \mathbf{T} through a parameter-free LayerNorm, a learned linear projection ($d_{\text{thought}} \rightarrow 2048$), a second LayerNorm, and a learned position-independent offset before prepending the result to the token embedding sequence of the target or conditioning text, which is processed by the frozen backbone. Training uses AdamW with a cosine learning-rate schedule ($\text{lr}_{\text{max}} = 5 \times 10^{-5}$, 1 epoch, warmup steps = 20), batch size 64, and cross-entropy loss over the tokenized target sequence.

C.5 Discriminator Architecture and Training

The discriminator $f_{\text{disc}}(\mathbf{T}, Y)$ applies a parameter-free LayerNorm to \mathbf{T} , projects it with a learned linear map ($d_{\text{thought}} \rightarrow 2048$), and passes the result through a second LayerNorm. The input sequence to the frozen backbone is formed by concatenating the token embeddings of Y , a learned separator embedding, the projected thought vectors, and a learned CLS token. The hidden state at the CLS position is classified by a two-layer head consisting of Linear(2048 \rightarrow 1024), LayerNorm, ReLU, Dropout(0.1), and Linear(1024 \rightarrow 1). This pattern of training a binary classifier on projected latent activations as a read-out mechanism for behavioral signals inside transformer models is consistent with recent work on activation read-outs across LLM families [79]. Training uses binary cross-entropy loss with AdamW, learning rate 1×10^{-4} , batch size 64, and 1 epoch.

Hyperparameter provenance. The probe and discriminator configurations reported in this section, including the projection dimensions, the LLaMA-3.2-1B backbone, and the optimizer settings, were selected from a small set of variants explored during initial pilots. The values listed above are fixed across every source LLM and every candidate thought representation reported in this paper. No hyperparameter was tuned on the held-out test split.

C.6 Stability Sub-Properties and DCS Diagnostics

Sub-property coverage. Candidate representations fall into two empirical categories on the lexical-invariance axis. The Last Input Token, Pooled Output Embedding, Input Embedding, and Random Vector candidates, together with the iterative thinking families (Soft Thinking with and without Gumbel noise, Latent Thinking), produce a single vector per question regardless of which beam is drawn, so $\mathbf{T}_i = \mathbf{T}_j$ for every beam pair (i, j) within the same question and lexical invariance holds by construction for these candidates. Mode-collapse resistance is the only Stability sub-property that admits a nontrivial test for them. The Exact Output Embedding is per-beam, with $\mathbf{T}_i = \text{emb}(y_i)$ derived from each beam’s generated text independently. For DCS it is aggregated per-question. Because H_x is computed from pairwise cosine similarities between these same Nemotron output embeddings, the DCS score for the Exact Output Embedding confirms the alignment of the metric with the output distribution rather than a representational property under evaluation.

Input-embedding baseline. We report the embedding of the question text as a question-difficulty baseline alongside all thought representations. Because H_x is computed in the same embedding space, this input embedding benefits from a structural alignment with the label that model-derived representations do not share. Where it matches or exceeds thought representations, distributional uncertainty is largely predictable from the question text alone, which constitutes a finding about model capability rather than a flaw of the metric (see Section 5).

GPT-OSS-20B as an MoE outlier. GPT-OSS-20B yields a non-singleton semantic cluster across beams on only 1.0% of questions at $\tau = 0.9$, against 16–46% for the four dense source LLMs. The Random Vector baseline drifts above 0.5 on the same model because the positive class is extremely small. The pattern is consistent with Top- K MoE routing acting as a re-convergence force on diverging beams, an architectural feature absent in the four dense source LLMs that pulls beam representations back toward each other after they begin to differ. Bespoke generation techniques tailored to MoE architectures may surface divergent semantic outputs but were avoided in this audit to keep the protocol consistent across source LLMs.

Threshold sweep. Figure 3 sweeps the semantic equivalence threshold τ over $\{0.70, 0.80, 0.85, 0.90, 0.95\}$ for one representative per thought-representation family across all five LLMs. All other quantities are held fixed. Only the binarization of the cosine similarity matrix used to compute H_x varies. Rankings are stable across the full range. Families that score above the random-vector baseline at the main-text threshold do so at every other threshold, and families that score near the random baseline do so uniformly.

Cencerrado et al. [12] demonstrate that when H_x is linearly decodable from a representation via a difference-of-means probe, the uncertainty signal is encoded within it. We note that the converse does not necessarily follow, since a representation may encode H_x through a non-linear structure that a linear probe cannot detect and would in that case score near the random baseline despite carrying the relevant information. DCS scores near the random baseline are therefore evidence against linear encoding of distributional uncertainty rather than definitive evidence against encoding in any form.

C.7 Reproducibility and Code Release

The full evaluation pipeline is publicly available at <https://fard-lab.github.io/formalize-thoughts> under the MIT license. The repository is organized as a Hydra project with one runnable script per phase and a shared configuration tree. Every random seed used in the pipeline is pinned in the corresponding Hydra config so that any individual phase can be re-executed without further parameter passing. Source LLMs are downloaded from public HuggingFace repositories at the identifiers given in Section 4, BBEH is taken from the official Google DeepMind release under the Apache 2.0 license, the LLaMA-3.2-1B backbone shared by every probe and discriminator

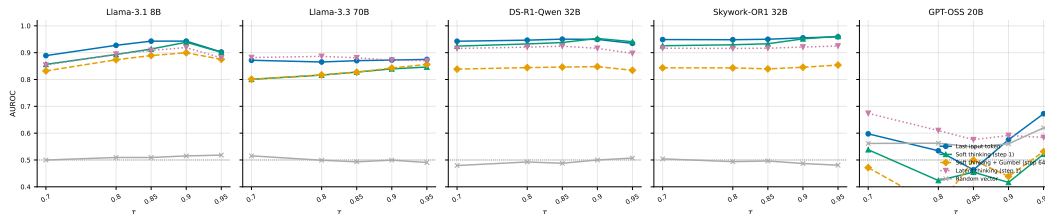


Figure 3: DCS versus the semantic equivalence threshold τ for one representative per thought-representation family across all five LLMs. Rankings are stable across the full range shown.

is a public HuggingFace asset, and the text embedder used is `nvidia/llama-embed-nemotron-8b` on HuggingFace.

Environment. The repository targets Python 3.12 and CUDA 12.6, with dependencies pinned via `pyproject.toml` and resolved by `uv sync`. Optional flash-attention support is documented in the README and is not required to reproduce the reported numbers.

Hardware. The vast majority of compute was performed on NVIDIA H100 SXM 80 GB GPUs, including all source-LLM data generation and all discriminator training. NVIDIA A100 40 GB GPUs were used for a small subset of probe-training and evaluation runs, without changes to the configuration.

Compute totals. All numbers below are H100-equivalent GPU-hours and are reported per source LLM. LLM data generation varies substantially across tasks and source LLMs, ranging from roughly 12 to 48 hours per BBEH task depending on output length and model capacity, and averages near 18 hours per task for an aggregate of approximately 400 hours per LLM. Discriminator training comprises 42 runs per LLM (21 candidates each evaluated under same-task and cross-task pairing), with each run taking roughly 3 to 7 hours and an aggregate of approximately 210 hours per LLM. Minimality probe training comprises 42 runs per LLM (21 candidates each producing the $CE(Y | \mathbf{T})$ and $CE(X | Y, \mathbf{T})$ probes that enter Δ_{IB}), with each run taking roughly 2 to 4 hours and an aggregate of approximately 125 hours per LLM. Causality evaluation runs in roughly 15 to 20 minutes per candidate, for an aggregate near 6 hours per LLM. DCS evaluation runs in roughly 5 minutes total for all five LLMs combined, since it operates only on cached embeddings and a difference-of-means probe. Aggregating across the five source LLMs, the experiments reported in this paper consumed approximately 3,700 H100-equivalent GPU-hours. The full research project consumed additional compute beyond this figure for pilot architecture and hyperparameter exploration, whose final settings underwrite the configurations recorded in Sections C and C.5.

D Additional Analysis

D.1 Bootstrap Confidence Intervals

All main-table metrics are reported as $\mu \pm \hat{\sigma}_B$, where μ is the cell mean over the held-out test split and $\hat{\sigma}_B$ is the bootstrap standard error obtained by the percentile method with $B = 10,000$ resamples and a fixed random seed. The statistical unit of resampling is the *problem*, with replacement from the test split. We then recompute the cell-level mean on the resampled problems, and take $\hat{\sigma}_B$ as the sample standard deviation of the resampled means. The reported \pm values are therefore 1σ bootstrap standard errors, and a 95% interval is recovered as $\mu \pm 1.96 \hat{\sigma}_B$ under the normal approximation, or as the [2.5%, 97.5%] percentiles of the bootstrap distribution when that approximation is undesirable.

The axioms differ in how each problem contributes observations, and we resample at the problem level in every case so that within-problem beam correlations are preserved.

- **Minimality input component $CE(X | \mathbf{T})$:** each problem contributes one per-problem cross-entropy. Resample the test split; the effective sample size equals the test-split problem count from Section E.

- **Causality** $\text{KL}(P(y_{\text{suf}} | y_{\text{pre}}) \| P(y_{\text{suf}} | \mathbf{T}))$: cluster bootstrap by problem, with all valid beams of a resampled problem kept together. Beams excluded by the 51-token length gate are absent from the resampled mean, and problems whose beams are all excluded are absent entirely; on Llama-3.1-8B this reduces the effective support to 360 problems. The paired advantage $\Delta_{\text{KL}}^{\text{RV}}$ is resampled on the same support with each (\mathbf{T}, RV) pair kept glued together (see below).
- **Discriminator-based DCS**: the per-problem discriminator-based DCS score, computed over all off-diagonal beam pairs in the $K \times K$ within-problem matrix, is a single observation per problem. Resample problems, average per-problem score.

Paired advantage versus the noise floor (Causality). Because KL is unbounded above, absolute values are difficult to compare across candidate \mathbf{T} without a shared reference. We therefore compute the paired advantage $\Delta_{\text{KL}}^{\text{RV}}(\mathbf{T})_g = \text{KL}(\mathbf{T})_g - \text{KL}(\text{RV})_g$ per problem g , and average over the Causality support defined above. Pairing on the problem cancels per-problem shifts in the KL scale (the dominant source of dispersion in the raw metric), yielding a sign-interpretable statistic where $\Delta < 0$ indicates \mathbf{T} conveys information about y_{suf} beyond the evaluation backbone’s projection of noise, whereas $\Delta \geq 0$ indicates no detectable advantage. The paired bootstrap resamples problems with each $(\text{KL}(\mathbf{T})_g, \text{KL}(\text{RV})_g)$ pair kept together, so the resulting standard error quantifies the variance of the paired difference rather than that of the two unpaired means. In Table 10, each cell displays the mean $\text{KL}(y_{\text{suf}} | \mathbf{T})$ with its marginal bootstrap standard error.

D.2 Detailed per-axiom results

Tables 9 to 11 report the cell-level mean and bootstrap standard error of the Separability, Causality, and Minimality metrics across the populated source LLMs. Figure 4 reports the corresponding Stability AUROC values. Each candidate occupies one column and each source LLM occupies one row. The cell formatting follows the methodology of Section D.1.

Table 9: Discriminator test accuracy (%) across source LLMs. Each LLM spans two rows, **Same** (within-task instance discrimination) and **Cross** (across-task discrimination). A uniform random classifier scores 50% as data is balanced.

LLM	Reg.	Output Emb.			Last Input Tok.					Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baselines	
		Exc	Pool	All	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE	RV			
Llama-3.1 8B	Same	68.79	64.33	53.39	53.93	53.15	53.50	54.12	54.74	52.49	52.60	52.57	53.55	51.02	54.72	53.57	53.00	52.54	52.65	54.52	48.85	54.52	48.85			
	Cross	98.99	98.89	96.09	99.24	85.70	97.43	97.76	98.17	97.70	77.20	96.42	96.07	97.32	97.64	99.32	98.67	98.41	98.71	98.63	98.94	50.50	98.94	50.50		
Llama-3.3 70B	Same	72.62	54.25	51.56	50.08	52.79	52.93	51.55	51.12	51.30	52.83	51.96	50.26	51.94	51.01	50.50	50.50	51.42	51.01	50.73	52.09	49.67	52.09	49.67		
	Cross	98.29	96.42	98.77	98.55	79.33	95.91	92.53	94.15	94.01	74.05	93.67	93.43	90.76	60.76	96.89	94.18	95.82	94.86	92.57	96.25	51.04	96.25	51.04		
DS-R1-Qwen 32B	Same	63.54	63.05	52.13	52.56	53.26	54.45	54.81	53.86	53.42	50.69	51.84	50.90	51.33	51.06	50.33	50.06	49.94	50.21	50.22	53.53	50.28	53.53	50.28		
	Cross	99.18	97.79	99.24	98.33	89.16	97.25	96.79	98.27	97.90	62.58	91.18	96.03	95.80	95.31	96.09	98.33	97.04	96.28	93.60	96.76	50.22	96.76	50.22		
Skywork-OR1 32B	Same	62.04	63.41	53.26	51.94	53.32	54.19	52.82	52.25	52.64	50.65	51.76	50.58	51.74	50.91	51.23	50.28	50.22	50.12	50.06	54.01	49.85	54.01	49.85		
	Cross	99.41	96.93	99.14	98.95	82.78	98.16	97.57	98.11	98.09	60.20	93.94	95.31	98.16	97.72	74.75	88.97	92.24	91.62	76.92	97.82	50.51	97.82	50.51		
GPT-OSS 20B	Same	59.57	62.38	50.40	50.29	50.46	50.11	50.72	49.71	50.57	49.49	50.57	51.81	51.16	49.97	51.20	50.00	50.64	50.33	49.92	49.47	50.95	49.47	50.95		
	Cross	97.94	98.49	95.19	88.11	76.83	85.11	87.54	60.91	92.22	62.14	82.51	86.06	87.62	82.40	89.10	84.29	90.74	92.12	91.07	97.03	50.53	97.03	50.53		

The Minimality residual Δ_{IB} in Table 11 is read within source LLM. The absolute sign and scale of Δ_{IB} shift across source LLMs because the cross-entropy decomposition discards a TR-independent constant whose value depends on the conditional entropies of X and Y for that LLM, and recovering the absolute IB Lagrangian requires this constant (Section D.5). The within-LLM ranking of candidates is the meaningful comparison. Figure 2 in the main text applies this within-LLM normalisation directly so that all four axioms appear on a comparable within-LLM scale.

Table 10: KL divergence $\text{KL}(P(Z | Y) || P(Z | \mathbf{T}))$ (\downarrow) across source LLMs. Lower values indicate higher predictive sufficiency of \mathbf{T} for the continuation.

LLM	Output Emb.		Last Input Tok.		Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baselines	
	Exc	Pool	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE	RV
Llama-3.1 8B	5.25	5.21	5.26	5.01	5.07	5.69	5.51	5.07	4.96	5.20	5.42	5.06	4.70	4.73	5.32	5.79	6.05	6.49	6.17	5.36	9.49
Llama-3.3 70B	4.56	4.58	6.00	5.28	4.65	5.36	5.63	5.19	5.48	5.08	5.22	5.21	5.34	5.44	4.21	5.17	5.35	5.45	5.88	4.71	8.93
DS-R1-Qwen 32B	4.67	4.77	6.33	4.79	4.45	5.60	5.74	5.34	4.82	4.57	5.42	5.00	4.91	4.98	4.62	5.41	4.84	5.77	5.36	4.50	9.36
Skywork-OR1 32B	4.10	4.37	6.39	4.09	3.90	5.07	4.86	4.62	4.82	4.75	4.91	4.68	4.72	4.89	5.00	4.52	4.34	4.35	4.50	4.08	9.31
GPT-OSS 20B	3.82	4.17	5.61	4.19	4.04	4.00	4.01	4.08	4.23	4.17	4.48	4.48	4.57	4.74	3.90	5.12	4.63	4.68	4.67	3.78	9.60

Table 11: Minimality measure $\Delta_{\text{IB}} = \text{CE}(X | Y, \mathbf{T}) - \text{CE}(Y | \mathbf{T})$ across source LLMs, paired per problem. Larger positive values indicate a representation that is more sufficient for Y and adds less X -information beyond Y .

LLM	Output Emb.		Last Input Tok.		Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baselines	
	Exc	Pool	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE	RV
Llama-3.1 8B	0.37	0.27	0.08	0.16	0.25	0.18	0.21	0.19	0.19	0.24	0.17	0.16	0.17	0.16	0.17	0.19	0.18	0.17	0.17	0.22	-0.40
Llama-3.3 70B	-0.13	-0.20	-0.33	-0.30	-0.24	-0.27	-0.32	-0.32	-0.31	-0.24	-0.31	-0.31	-0.29	-0.32	-0.31	-0.31	-0.30	-0.33	-0.31	-0.23	-0.99
DS-R1-Qwen 32B	0.07	0.05	-0.07	-0.05	0.10	0.00	-0.02	-0.01	-0.01	0.10	0.04	0.04	0.01	0.00	0.05	0.03	0.03	0.02	0.02	0.04	-0.50
Skywork-OR1 32B	-0.16	-0.19	-0.31	-0.27	-0.13	-0.28	-0.29	-0.29	-0.27	-0.14	-0.21	-0.22	-0.23	-0.25	-0.18	-0.20	-0.21	-0.21	-0.21	-0.21	-0.59
GPT-OSS 20B	-0.26	-0.22	-0.27	-0.25	-0.21	-0.21	-0.21	-0.21	-0.21	-0.20	-0.24	-0.21	-0.24	-0.21	-0.23	-0.18	-0.17	-0.18	-0.23	-0.34	-0.30

Table 12: Output-prediction cross-entropy $\text{CE}(Y | \mathbf{T})$ across source LLMs with \mathbf{T} tiled to a common length. The tiled form is the one entering the IB-residual measure of Table 11. Comparison with Table 24 isolates the effect of length normalisation on each representation.

LLM	Output Emb.		Last Input Tok.		Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baselines	
	Exc	Pool	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE	RV
Llama-3.1 8B	0.72	0.79	0.80	0.78	0.82	0.82	0.80	0.82	0.83	0.86	0.88	0.88	0.89	0.90	0.82	0.79	0.79	0.80	0.82	0.82	7.08
Llama-3.3 70B	1.26	1.29	1.16	1.22	1.37	1.19	1.22	1.24	1.25	1.38	1.29	1.27	1.25	1.27	1.26	1.26	1.29	1.29	1.33	7.61	
DS-R1-Qwen 32B	0.82	0.83	0.76	0.80	0.81	0.81	0.81	0.81	0.81	0.85	0.85	0.84	0.83	0.83	0.81	0.81	0.81	0.81	0.81	0.85	6.90
Skywork-OR1 32B	0.96	0.97	0.89	0.94	0.95	0.97	0.94	0.95	0.95	0.97	1.00	0.97	0.97	0.96	0.95	0.93	0.93	0.93	0.94	1.02	6.77
GPT-OSS 20B	0.83	0.80	0.78	0.81	0.82	0.81	0.81	0.81	0.81	0.84	0.81	0.81	0.81	0.80	0.80	0.80	0.80	0.81	0.81	0.84	6.59

Table 13: Conditional input cross-entropy $\text{CE}(X | Y, \mathbf{T})$ across source LLMs with \mathbf{T} tiled to a common length. Higher values indicate that, given the output, \mathbf{T} leaks less residual information about the input. This is the input component of the IB-residual measure in Table 11, paired with $\text{CE}(Y | \mathbf{T})$ from Table 12.

LLM	Output Emb.		Last Input Tok.		Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baselines		
	Exc	Pool	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE	RV	
Llama-3.1 8B	1.09	1.06	0.89	0.94	1.07	1.00	1.01	1.01	1.02	1.11	1.05	1.04	1.05	1.05	0.99	0.98	0.97	0.97	0.99	1.04	6.69	
Llama-3.3 70B	1.13	1.09	0.83	0.92	1.13	0.92	0.91	0.93	0.94	1.14	0.98	0.96	0.96	0.95	0.95	0.95	0.95	0.96	0.96	0.99	1.10	6.62
DS-R1-Qwen 32B	0.90	0.89	0.68	0.75	0.91	0.81	0.79	0.80	0.81	0.95	0.89	0.88	0.84	0.83	0.86	0.83	0.83	0.83	0.83	0.89	6.40	
Skywork-OR1 32B	0.80	0.79	0.58	0.67	0.83	0.69	0.66	0.66	0.69	0.83	0.78	0.76	0.74	0.71	0.77	0.73	0.72	0.72	0.72	0.81	6.17	
GPT-OSS 20B	0.58	0.57	0.50	0.55	0.62	0.60	0.60	0.60	0.60	0.65	0.57	0.60	0.56	0.60	0.56	0.61	0.63	0.63	0.58	0.50	6.29	

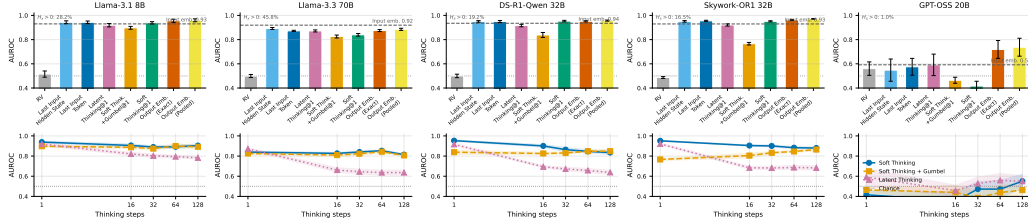


Figure 4: Distributional Consistency Score (DCS) across source LLMs at $\tau = 0.9$. Top row, AUROC of the difference-of-means probe predicting $H_x > 0$ from each thought representation, with the input embedding shown as a question-difficulty baseline. Bottom row, DCS as a function of thinking steps for the iterative thought families.

D.3 Distributional Views of Causality

Table 10 reduces each candidate to a single number per source LLM. We complement that view with two diagnostics that read the same per-problem records bootstrapped in Section D.1, without reweighting or new measurements.

Figure 5 pools all per-beam KLs at the 50-token window into one distribution per representation and reports its CDF, with one panel per representation family. A curve sitting to the left has smaller per-problem KL on average. A steeper curve has a tighter per-problem distribution around the mean reported in Table 10. A curve that approaches the top slowly has a heavy upper tail of high-KL problems that pulls the mean above the per-problem median. With this reading, the anchor panel separates candidates that the mean alone clusters into a single rank, since Exact output embedding lands at a similar mean as Random vector but carries a longer upper tail. Within the thinking families, increasing the step count shifts the entire distribution rather than only its mean, and Soft thinking with Gumbel noise at 32 steps shows a particularly heavy upper tail that is invisible at the mean.

Figure 6 decomposes the KL dispersion into between-problem and within-problem components. Values close to 1 indicate that the per-problem mean is carried by problem-level differences, supporting the cluster-bootstrap design of Section D.1, while values close to 0 indicate that the dispersion lives across beams within a single problem, so the mean averages over a heterogeneous within-problem distribution. The Random Vector and Output Embedding anchors show low ICC values, indicating that their per-beam KL dispersion is concentrated within individual problems rather than driven by problem-level differences. The soft-thinking and latent-thinking families show high ICC, so the scalar mean reported in Table 10 is a stable per-problem property rather than an average over heterogeneous within-problem noise.

D.4 Causality with the Output-Reconstruction Projection

The headline Causality results in Table 10 use the projection trained under the language-modelling objective of the Minimality output probe. The Minimality probe learns a projection of the same shape as the discriminator’s, mapping \mathbf{T} into the embedding space of the frozen 1B model, but optimised to reconstruct Y rather than to solve a binary discrimination task. We document here the ablation that motivated this choice.

We test whether the projection objective is the bottleneck for Causality by substituting the discriminator-trained projection in place of the output-reconstruction projection. Each representation is evaluated at its native substitution length without tiling, so the comparison isolates the projection objective from the length effects examined in Section D.6. The substitution path and KL formula are otherwise unchanged, and the random-vector reference column controls for any fixed projection-induced shift.

Table 14 reports the comparison on Llama-3.3-70B for the four representations of Section 3. The output-reconstruction projection meaningfully lowers KL for the latent representations that already live on the source-model manifold, with soft thinking dropping by roughly half on the KL scale and last input token by a smaller but separable margin. The exact output embedding moves the other way and the random-vector reference is unchanged, confirming the effect is representation-specific rather

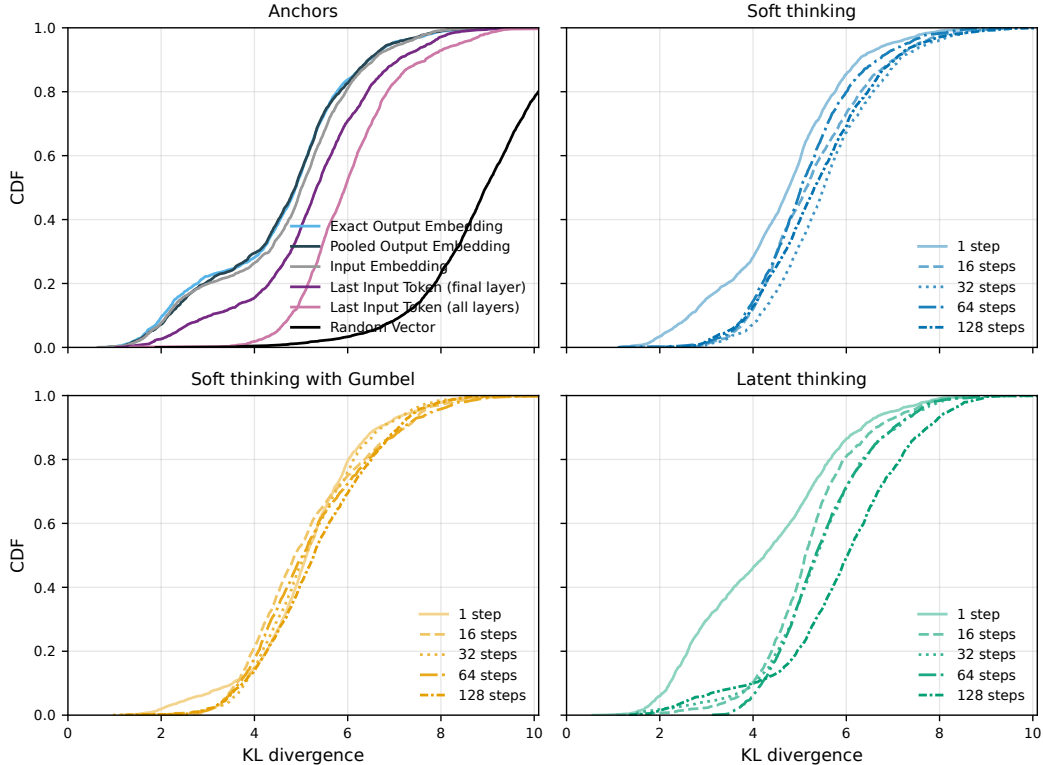


Figure 5: Per-beam KL CDFs at the 50-token averaging window on Llama-3.3-70B, with one panel per representation family (anchor candidates, soft thinking, soft thinking with Gumbel noise, latent thinking). Within each thinking family every step count is shown.

than a fixed projection-induced shift. The discriminator-projection numbers are retained in Table 22 for cell-by-cell comparison.

Table 14: Causality KL (\downarrow) on Llama-3.3-70B comparing the discriminator-trained projection (Table 22; Disc) against the output-reconstruction projection used in the main text (Table 10; LM). The paired column reports the per-problem mean of $KL_{LM} - KL_{Disc}$ with cluster-bootstrap 95% CI; entries with the entire CI on one side of zero are statistically separable from the projection used in the main text.

Representation	KL_{Disc}	KL_{LM}	Paired Δ (95% CI)
Exact output embedding	4.90	5.33	+0.43 [+0.28, +0.57]
Last input token (all layers)	7.41	5.99	-1.42 [-1.55, -1.30]
Soft thinking, 128 steps	10.51	5.49	-5.02 [-5.22, -4.82]
Random vector	4.44	4.39	-0.05 [-0.10, +0.01]

The projection objective is accordingly a substantial contributor to the high KL values in Table 10 for representations native to the source-LLM residual stream, where switching to the output-reconstruction projection yields representation-specific reductions while the random-vector reference remains unchanged.

D.5 Information-Bottleneck Decomposition for the Minimality Metric

The minimality metric Δ_{IB} of Section 3 is the Information Bottleneck Lagrangian at the symmetric weight $\beta = 2$, expressed in cross-entropies a probe can compute. We give the derivation, identify the constant offset between Δ_{IB} and the absolute Lagrangian, and characterise the conditions under which the surrogate is exact.

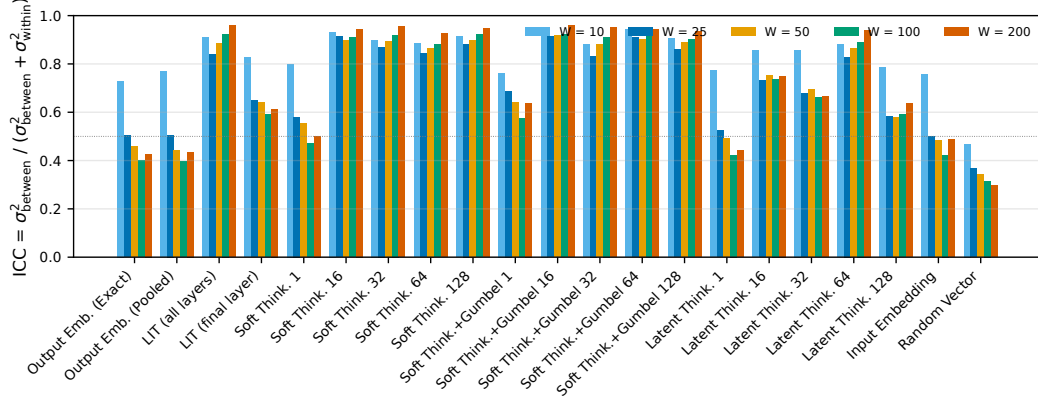


Figure 6: Intra-class correlation $ICC = \sigma_{\text{between}}^2 / (\sigma_{\text{between}}^2 + \sigma_{\text{within}}^2)$ of per-beam causality KL on Llama-3.3-70B, with bars per averaging window. Values above 0.5 indicate that the per-problem mean carries most of the dispersion, validating the cluster bootstrap that resamples problems and keeps within-problem beams glued together (Section D.1).

Chain-rule decomposition. The general chain rule for mutual information reads $I(X; \mathbf{T}) + I(Y; \mathbf{T} | X) = I(\mathbf{T}; Y) + I(X; \mathbf{T} | Y)$. When \mathbf{T} is a deterministic function of X , the term $I(Y; \mathbf{T} | X)$ vanishes because \mathbf{T} has no residual variance once X is known, so the chain rule collapses to $I(X; \mathbf{T}) = I(\mathbf{T}; Y) + I(X; \mathbf{T} | Y)$. Substituting into the IB Lagrangian $L(\beta) = I(X; \mathbf{T}) - \beta I(\mathbf{T}; Y)$ gives $L(\beta) = (1 - \beta)I(\mathbf{T}; Y) + I(X; \mathbf{T} | Y)$. At $\beta = 2$ the two coefficients have equal magnitude with opposite signs, recovering the symmetric trade-off

$$-L(2) = I(\mathbf{T}; Y) - I(X; \mathbf{T} | Y) \quad (6)$$

Barber–Agakov surrogates. Mutual information is intractable because $H(\cdot | \cdot)$ depends on unknown distributions. For any approximate conditional q , Gibbs’ inequality gives $H(A | B) \leq \text{CE}_q(A | B)$, with equality iff q matches the true conditional. Training a probe to minimise empirical negative log-likelihood on (B, A) pairs therefore yields a tight upper bound on $H(A | B)$ in the limit of a sufficiently expressive probe class. We use this to estimate the two terms of $-L(2)$,

$$I(\mathbf{T}; Y) = H(Y) - H(Y | \mathbf{T}) \approx H(Y) - \text{CE}(Y | \mathbf{T}), \quad (7)$$

$$I(X; \mathbf{T} | Y) = H(X | Y) - H(X | Y, \mathbf{T}) \approx \text{CE}(X | Y) - \text{CE}(X | Y, \mathbf{T}). \quad (8)$$

The first follows from the standard Barber–Agakov lower bound on mutual information [8]. The second is its conditional analogue, where two probes share the input Y and differ only in whether \mathbf{T} is appended.

Reduction to Δ_{IB} . Substituting the surrogates into $-L(2)$ and grouping \mathbf{T} -dependent terms,

$$-L(2) \approx [\text{CE}(X | Y, \mathbf{T}) - \text{CE}(Y | \mathbf{T})] + [H(Y) - \text{CE}(X | Y)] = \Delta_{\text{IB}} + C \quad (9)$$

The constant $C = H(Y) - \text{CE}(X | Y)$ depends only on the dataset and the unconditional baseline probe, not on the candidate \mathbf{T} . Comparisons of Δ_{IB} across representations are therefore comparisons of $-L(2)$ shifted by a single offset.

Bounding direction and ranking preservation. By Gibbs’ inequality each CE term upper-bounds the corresponding conditional entropy, so each approximation in Equations (7) and (8) yields a lower bound on the mutual information term it estimates. In Δ_{IB} , however, the two CE approximations are subtracted from each other. Any systematic bias shared across candidates that is absorbed into the constant C cancels in the difference. Within a fixed probe class and a fixed source LLM, the residual approximation error is candidate-independent, so the ranking of Δ_{IB} across thought representations is preserved. This is why the metric is read within a source LLM rather than across LLMs, as C varies with the source LLM’s output distribution and must not be compared across models.

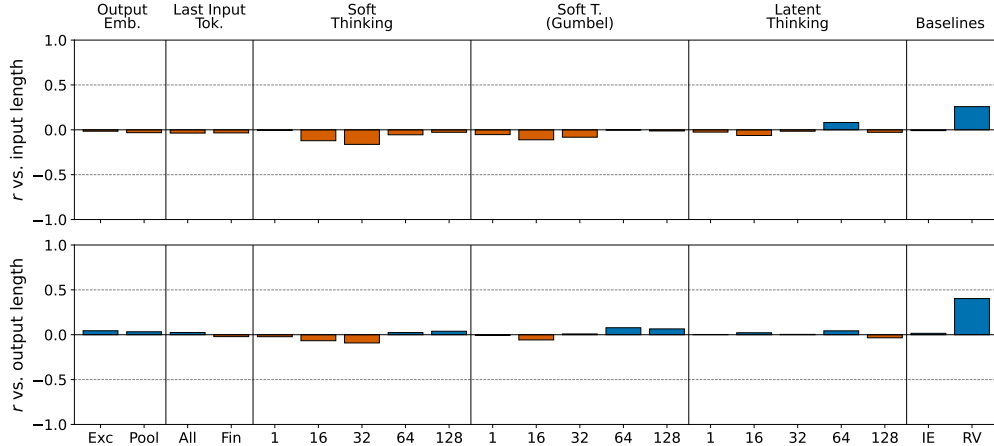


Figure 7: Pearson r between per-example causality KL and input length (top) or output length (bottom), in characters, on Llama-3.1-8B-Instruct. Candidates follow the order of Table 10.

Random Vector cross-entropy interpretation. The elevated $\text{CE}(Y | \mathbf{T})$ observed for the Random Vector anchor is not an out-of-distribution artifact. A random vector carries no information about Y , so a probe conditioned on pure noise cannot predict the output sequence and cross-entropy rises to the level of an unconditional language model. The probe is asked to optimize a signal that does not exist in \mathbf{T} , and the resulting high CE confirms that the metric correctly identifies the absence of output-relevant content rather than a distributional mismatch.

When the chain-rule assumption fails. The reduction relies on $I(Y; \mathbf{T} | X) = 0$. The exact output embedding and pooled output embedding candidates are computed directly from the generated continuation, so \mathbf{T} is a function of Y as well as X and $I(Y; \mathbf{T} | X) > 0$. For these rows, the substitution into the Lagrangian carries a correction term that does not collapse into a TR-independent constant, so Δ_{IB} no longer estimates $-L(2)$ even up to C . The reported value still has a clear empirical reading. For the exact output embedding, \mathbf{T} contains Y by construction, so $\text{CE}(Y | \mathbf{T})$ collapses toward zero and $\text{CE}(X | Y, \mathbf{T})$ approaches $\text{CE}(X | Y)$. The metric therefore flags this row as a trivially sufficient anchor with no residual leakage, which is the correct behavioural diagnosis even though the IB-Lagrangian interpretation no longer applies.

D.6 Length Sensitivity of the Causality Metric

The causality KL in Table 10 averages over the last 50 tokens of each generated beam, which raises two length-based concerns. First, the metric might track prompt size rather than thought content. Second, a longer generation alone could inflate KL for whichever candidate happens to produce it. Figure 7 examines both.

The first concern is ruled out outright. No candidate shows a meaningful coupling between KL and input length, so the main-text ordering does not reflect task-to-task variation in prompt size.

The second concern resolves cleanly. Output length couples with KL only for the Random Vector lower-bound reference, since its content is uninformative by construction and longer generations accumulate more positions that disagree with the explicit prefix and inflate its KL on long beams. The candidates clearly below this reference in Table 10 show near-zero coupling, so their KL is driven by representational fit rather than by how much the source model chose to generate. Output length is therefore not a competing explanation for the main-text ranking but a property of the lower-bound reference against which the ranking is measured.

We further test whether the choice of 50 tokens itself drives the ranking by recomputing the metric across windows of $\{10, 25, 50, 100, 200\}$ tokens at no extra source-LLM cost, since each beam saves a per-window KL during a single causality pass through the discriminator. Figure 8 shows the result. The coarse ordering between candidates is stable across windows, with strong candidates sitting near the bottom of the KL axis and weak candidates near the top at every scale, so the ranking in Table 10

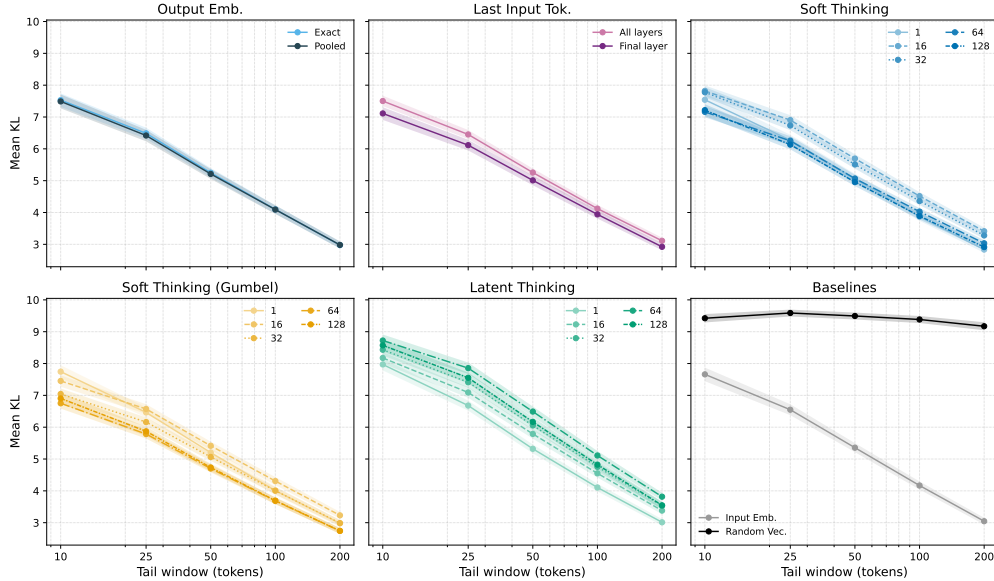


Figure 8: Mean causality KL on Llama-3.1-8B-Instruct across Z -windows of $\{10, 25, 50, 100, 200\}$ tokens, with 95% bootstrap CI bands. One panel per TR family, one line per (TR type, think steps) tuple.

does not depend on the exact window length. Fine-grained orderings within clusters of similarly-performing candidates do shift at the longer windows, where the shared Z context narrows absolute differences to the point that small method-to-method gaps become statistically indistinguishable. At the shortest window the metric averages over very few positions, where per-token stochastic variance dominates and no candidate cleanly separates from the floor. The 50-token window sits between these two extremes, giving the cleanest separation between candidates while still averaging over enough positions to be statistically stable.

A third length-based concern is the number of substituted vector positions itself, which varies across candidates. Last input token (all layers) substitutes one position per source-model layer, Soft thinking substitutes one position per thinking step, while embedding candidates substitute a single position. To isolate the geometric effect of substitution length from any change in information content, we tile each candidate to a common number of positions and recompute KL. Figure 9 reports the result on Llama-3.3-70B for the four candidates from Table 10 that admit clean tiling.

The Random vector curve rises markedly with substitution length, even though its content is by construction uninformative about the prefix, and the same trend appears for Exact output embedding with a smaller dynamic range. Substitution length therefore inflates KL on candidates that carry no instance information, which establishes that absolute KL magnitudes are not commensurable across candidates of different native lengths. The relative ordering between candidates also reorganises once length is matched, with candidates that sat at the strong end of the table at native length sliding toward the weak end once they are tiled out to the length of the longer candidates. The reorganisation does not nullify Table 10, since tiling a one-position candidate produces a rank-one substitution that is structurally different from a natively multi-position candidate, but it does mean that absolute KL gaps in the table reflect a mix of representational quality and substitution geometry.

This finding fixes the methodological choice that all subsequent quantitative comparisons feed \mathbf{T} at a common substitution length, so that no representation is rewarded or penalised by the metric simply for being natively shorter or longer than another. The causality results of Table 10 and the minimality probes underlying Table 11 both adopt the same tiled length, and varying-length comparisons are reserved for the diagnostic in Figure 9.

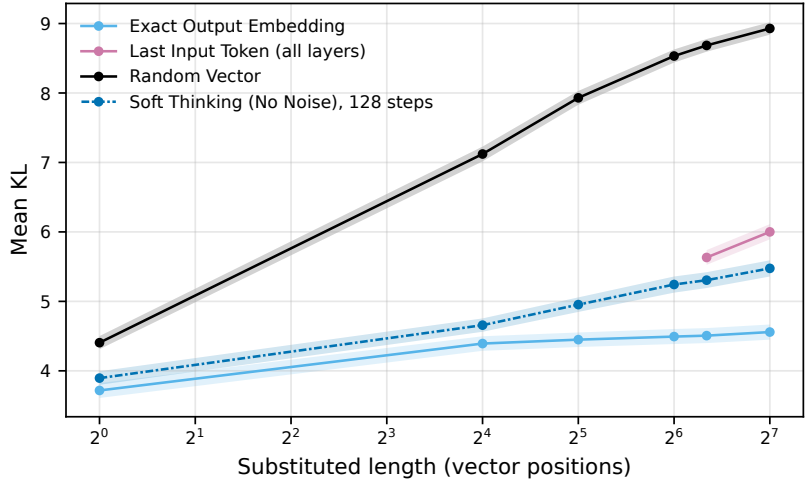


Figure 9: Mean causality KL on Llama-3.3-70B as the substituted length L varies, with 95% cluster-bootstrap CI bands. The single-vector candidates (Exact Output Embedding, Random Vector) are repeated L times. Soft Thinking (No Noise) uses its native 128-step trajectory truncated to its first L vectors at $L \leq 128$, so different L points cover different prefixes of the same trajectory rather than tiled copies of one vector. Last Input Token (all layers) has native length 81 on Llama-3.3-70B and is shown at $L = 81$ (no tiling) and at $L = 128$ (the stack is wrapped to fill the substitution window).

D.7 Geometric Analysis of Representational Collapse

Table 9 reports near-ceiling cross-task discrimination on most candidates and near-random same-task discrimination on almost all of them, which invites the objection that the same-task failure reflects probe capacity rather than a property of the representation. We address this objection by measuring the thought-vector geometry directly, with no trained probe in the loop. Each candidate is flattened to a single vector per problem (the full $L \times d$ hidden-state stack for LIT, the $s \times d$ prefix for s -step thinking candidates), and all comparisons use cosine similarity.

Two quantities, one per discriminator task.

- **k -NN task purity** [74, 11] at $k=10$. For each thought vector, purity is the fraction of its ten nearest neighbours (by cosine similarity) that carry the same task label. The average over all vectors lies in $[0, 1]$, with uniform random neighbour assignment giving ≈ 0.042 , and a value near 1 means the geometry clusters problems by task.
- **Within-task participation ratio** [47, 56] $PR = (\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$, where σ_i are the singular values of the centred matrix of within-task vectors for a given task (so σ_i^2 are the eigenvalues of its sample covariance), averaged over tasks. PR counts how many directions are needed to describe the within-task variance, ranging from 1 when every within-task vector points the same way up to $N-1$ when the vectors spread over as many orthogonal directions as centering allows, where N is the number of within-task problems. On this split $N \approx 20$, so the ceiling is ≈ 19 , and Random Vector reaches $PR = 18.5$ close to that ceiling.

Each clustering quantity tracks one discriminator. Purity must clearly exceed the random-neighbour baseline for cross-task discrimination to succeed, since no probe can recover task structure that is absent from the geometry. PR must clearly exceed 1 for same-task discrimination to succeed, since two within-task vectors that span too few directions are too close together for any learned projection to pull apart. The converse is weaker, because Random Vector achieves the maximum PR by construction and so a high PR alone cannot certify that the within-task directions encode instance content. Geometry therefore detects the two failure modes (low purity or low PR) without relying on probe behaviour, but cannot certify success.

An anisotropy-adjusted similarity scale. We complement the two clustering quantities with a third diagnostic, the within-task self-similarity adjusted for anisotropy, following Ethayarajh [23]. Define $\Delta_{\text{cos}} = \overline{\text{cos}}_{\text{within-task}} - \overline{\text{cos}}_{\text{cross-task}}$, where the cross-task pair mean serves as the anisotropy baseline inherent to self-attention [28]. Subtracting this baseline isolates the within-task scale beyond what transformer geometry alone produces. Where Ethayarajh [23] apply the adjustment at the token level, we compute it on flattened example-level vectors so that a single number summarises each candidate. Δ_{cos} is orthogonal to k -NN purity and PR, since it varies in similarity scale rather than in cluster structure.

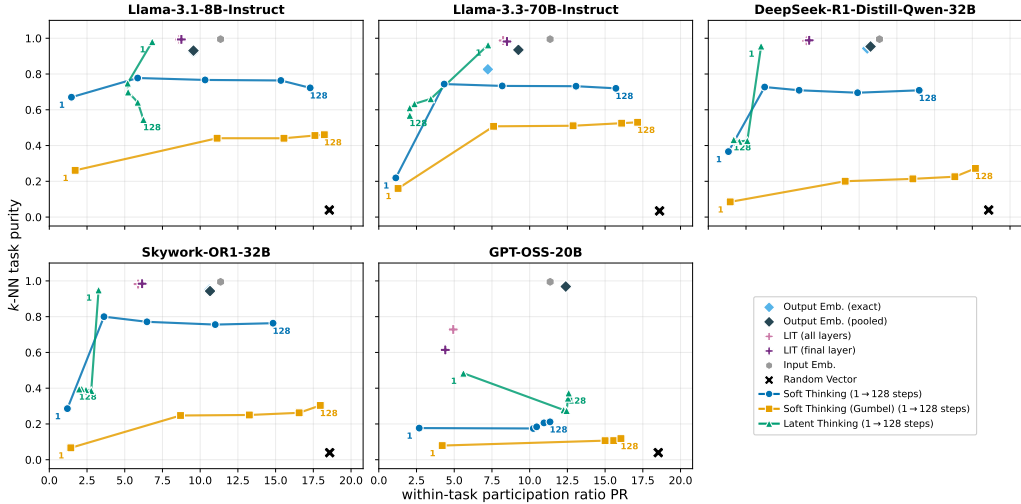


Figure 10: Each candidate placed on the (PR, k -NN purity) plane, one panel per LLM. Solid lines trace thinking-family trajectories as the step count grows from 1 to 128. Candidates further to the upper right are preferable: higher PR indicates a within-task subspace spread across more directions, and higher k -NN purity indicates that nearest neighbours are drawn from the same task.

Figure 10 places every candidate on the (PR, purity) plane for each LLM, and the top-right region remains empty in every panel. LIT, Latent Thinking at one step, Input Embedding, and the two Output Embedding candidates occupy the same region of the plane on four of the five LLMs, with moderate PR below the ceiling and high purity; on GPT-OSS-20B both LIT variants drop to lower purity while the Output and Input Embedding candidates stay in their usual region. The geometry therefore admits cross-task discrimination but constrains within-task variance to a narrow subspace regardless of architecture or size. Output Embedding candidates land in this same cluster yet still succeed at same-task discrimination in Table 9, because their within-task directions encode instance content that LIT’s do not. Geometry alone cannot tell the two cases apart, so the analysis rules out failure without certifying success.

Same-task discrimination fails because the within-task geometry of latent thought representations is itself collapsed, either too narrow (low PR), too noise-drifted (low purity), or both, and this holds across every LLM tested. Probe capacity therefore cannot explain the failure. Latent Thinking begins at one step in the high-purity region near LIT on every LLM, but adding latent steps cuts purity by roughly half on each LLM and additionally sheds PR on the larger Llama and on both distilled and RL-trained reasoning models. Soft Thinking begins in the opposite low-PR corner and gains PR steadily as depth grows, but Δ_{cos} contracts in step with the gain (from ~ 0.30 at one step to ~ 0.09 at 128 on the Llama models), so the additional dimensions opened up by depth carry less, not more, anisotropy-adjusted within-task signal. Soft Thinking with Gumbel noise traces the same rightward path at consistently lower purity and lands at $\Delta_{\text{cos}} \in [0.02, 0.05]$ across all five LLMs, indistinguishable from Random Vector on that axis. No trajectory in any panel enters the high-PR, high-purity, high- Δ_{cos} region occupied by LIT and Output Embedding. The verdict is uniform across architectures and sizes. The distilled and RL-trained reasoning models produce nearly identical panels in Figure 10, with every clustering quantity and similarity scale agreeing to within a few percent across all candidates and only minor offsets in the Latent and Soft Thinking trajectories.

The geometric verdict matches the main-text tables across LLMs. Every candidate apart from Output Embedding has either low PR or low purity in each panel, and all of them fail same-task discrimination in Table 9. The DCS floor in Table 26 shares the same origin, because cosine similarity over a narrow or noise-drifted within-task subspace cannot reproduce a richer semantic equivalence matrix at any threshold. The picture is therefore consistent across architectures and sizes. The failure to discriminate within a task is a property of the geometry that latent computation produces, not an artifact of any single LLM.

Δ_{cos} anchors the same picture on the scale axis. Random Vector is at the lower bound near 0, Input Embedding at the upper bound near 0.52, and LIT and Output Embedding occupy a range between roughly 0.18 and 0.45. All three diagnostics converge on the same failure pattern, and none depend on a trained discriminator.

D.8 Probe Capacity Ablation

Section D.7 ruled out the most direct version of the weak-probe objection at the level of the geometry itself, with no trained classifier in the loop. We add a complementary line of defense within the probe-based metric by varying the same-task discriminator’s capacity on the largest source model, Llama-3.3-70B, and showing that the chance-level same-task verdict for the latent-thinking and last-input-token candidates is invariant under three architectures of strictly increasing capacity.

The Baseline architecture, used throughout the main paper, projects the source hidden state into the input space of the frozen 1B model through a single linear layer. The Deep projection replaces the linear with a two-layer MLP and a LayerNorm. The Deep projection + UF2 architecture additionally unfreezes the last two transformer blocks of the 1B model. Each step strictly enlarges the trainable-parameter budget, and the deepest tier carries roughly an order of magnitude more trainable parameters than the Baseline.

Three thought representations are evaluated as anchors. Last Input Token and Soft Thinking at 128 steps are the focus of the same-task failure mode in the main results. Output Embedding (Exact) acts as the positive control, since it is computed from the generated answer itself. If any architecture has a chance to descend below the random-guess plateau, it should descend on this representation.

Table 15: Probe capacity ablation on Llama-3.3-70B same-task discrimination. Each cell reports test accuracy and the tail-50 mean of the training BCE loss. Columns increase discriminator capacity from a single linear projection (Baseline) through a deep MLP projection (Deep) to additionally unfreezing the last two transformer blocks of the frozen 1B backbone (Deep + UF2). $\ln 2 \approx 0.693$ is the BCE value at uniform 0.5 output and equates to random guess.

Thought representation	Baseline (~ 19M trainable)		Deep (~ 44M trainable)		Deep + UF2 (~ 165M trainable)	
	Acc	BCE	Acc	BCE	Acc	BCE
Last Input Token	0.516	0.685	0.500	0.693	0.501	0.696
Soft Thinking at 128 steps	0.513	0.689	0.511	0.689	0.508	0.692
Output Embedding (Exact)	0.726	0.494	0.617	0.631	0.642	0.659

Table 15 reports test accuracy alongside the tail-50 mean of the training-loss curve, where $\ln 2 \approx 0.693$ is the BCE value at uniform 0.5 output and equates to random-guess accuracy. Output Embedding (Exact) achieves its highest test accuracy at the Baseline tier (0.726), and adding capacity causes overfitting, with test accuracy declining to 0.617 at the Deep tier and 0.642 at Deep+UF2 despite an order-of-magnitude increase in trainable parameters, while BCE worsens toward the random-guess plateau. Last Input Token and Soft Thinking remain at random-guess accuracy at every capacity tier, with no improvement in accuracy or BCE regardless of how many parameters are available. The Baseline architecture was fixed in pilot experiments and applied consistently across all source LLMs and thought representations (see Section C.5). These ablations confirm post-hoc that it avoids the overfitting induced by deeper tiers while retaining the discriminative accuracy of the positive control. Under the larger architectures, the training loss of the candidates carrying no instance-discriminating content decreases transiently during training, but the tail-50 average over held-out steps remains at the plateau, indicating the optimiser learns batch-specific patterns rather than a generalisable structure.

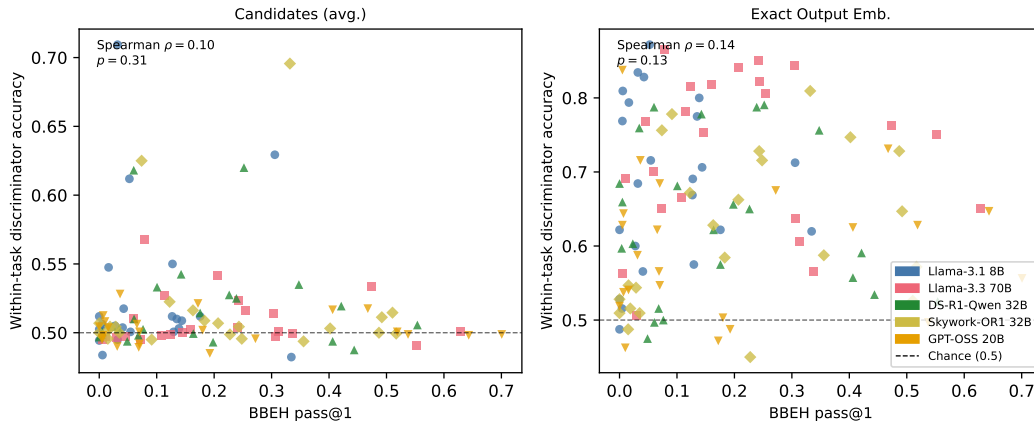


Figure 11: Within-task discriminator accuracy versus BBEH pass@1 for candidates averaged across Last Input Token, soft-thinking, latent-thinking, and last-hidden-state variants (left) and the Exact Output Embedding anchor (right), across 23 BBEH tasks and five LLMs. The dashed line marks the chance baseline. Spearman ρ and p -values are annotated on each panel.

Two independent attacks on the same objection, the geometric one with no probe in the loop and the probe-based one with a probe of increasing capacity, converge on the conclusion that the same-task failure is a property of the representation and not of the discriminator.

D.9 Relationship to Downstream Task Accuracy

A natural concern is that within-task separability collapses on hard tasks because the model’s output distribution itself becomes less structured under difficulty, not because the representation is inadequate. If this were true, the collapse would track downstream task accuracy and tasks with lower pass@1 would show lower within-task discriminability across all candidates.

Figure 11 tests this directly by plotting per-task within-task discriminator accuracy against BBEH pass@1 across the 23 BBEH tasks and five LLMs. For thought-representation candidates averaged over the Last Input Token, soft-thinking, latent-thinking, and last-hidden-state families, the pooled Spearman correlation with downstream accuracy is $\rho = 0.10$ ($p = 0.31$, $n = 115$), and no individual LLM reaches significance. The Exact Output Embedding anchor, which has direct access to the generated continuation, also shows no significant correlation ($\rho = 0.14$, $p = 0.13$). Both series are consistent with near-zero correlation across the full range of task difficulties.

These results support the interpretation in Section 5.1. The within-task separability collapse is a property of the thought representations rather than a consequence of tasks being difficult. The framework surfaces representational failures that downstream accuracy does not register.

E Dataset and Benchmark Details

E.1 BBEH Task Coverage

We evaluate on all 23 tasks from BBEH [41]. Table 16 lists the tasks and their categories. For each task, we use all available examples from the BBEH dataset, yielding 4,520 unique problems across all 23 tasks. Each problem is expanded to $K = 8$ beams during generation, for 36,160 beam–problem pairs in total. These 4,520 problems are further partitioned for probe and discriminator training, stratified per task: within each BBEH task, problems are randomly shuffled and split into train / validation / test at an 0.8/0.1/0.1 ratio, so every task contributes proportionally to every split (3,616 / 452 / 452 problems in aggregate).

Several BBEH tasks build on earlier benchmarks, and we acknowledge the originating work as requested by the BBEH authors. BoardGame QA originates from Kazemi et al. [40]. Causal Understanding draws on Nie et al. [52] and Kıcıman et al. [42]. Dyck Languages and Word Sorting follow the setup of Tyen et al. [69]. Geometric Shapes is based on Kazemi et al. [39], Linguini

Table 16: BBEH tasks used in our evaluation.

TASK	CATEGORY
BOOLEAN EXPRESSIONS	SYMBOLIC REASONING
DYCK LANGUAGES	SYMBOLIC REASONING
MULTISTEP ARITHMETIC	QUANTITATIVE REASONING
OBJECT COUNTING	QUANTITATIVE REASONING
TIME ARITHMETIC	QUANTITATIVE REASONING
SPATIAL REASONING	SPATIAL REASONING
GEOMETRIC SHAPES	SPATIAL REASONING
WORD SORTING	LANGUAGE
HYPERBATON	LANGUAGE
DISAMBIGUATION QA	LANGUAGE
LINGUINI	LANGUAGE
TEMPORAL SEQUENCE	TEMPORAL REASONING
CAUSAL UNDERSTANDING	CAUSAL REASONING
WEB OF LIES	SOCIAL REASONING
SHUFFLED OBJECTS	LOGICAL REASONING
ZEBRA PUZZLES	LOGICAL REASONING
BOARDGAME QA	GAME REASONING
BUGGY TABLES	TABULAR REASONING
MOVIE RECOMMENDATION	RECOMMENDATION
SPORTQA	FACTUAL QA
OBJECT PROPERTIES	COMMONSENSE
NYC COORDINATES (NYCC)	GEOGRAPHY
SARC TRIPLES	SARCASM DETECTION

on Sánchez et al. [60], and NYCC builds on Hessel et al. [36] and Zhang et al. [80]. Spatial Reasoning follows Yamada et al. [75], Time Arithmetic follows Fatemi et al. [25], Web of Lies follows White et al. [72], and Zebra Puzzles follows Shah et al. [61].

E.2 BBEH Answer Extraction

We use the official BBEH answer extraction logic from Kazemi et al. [41], which implements task-specific regular expression patterns to extract final answers from free-form model outputs. Table 17 reports per-LLM parsing statistics over every generation. Per-beam parsing failure varies by more than an order of magnitude across the LLMs we evaluate, and the fraction of problems with at least one parsed beam follows the same ordering. For Llama-3.3-70B-Instruct nearly every beam emits one of the canonical answer prefixes, whereas the other LLMs leave a substantially larger fraction of beams without any of the prefixes the extractor matches against, and the answer cannot be recovered without other methods.

Table 17: Answer-parsing statistics across LLMs over all generations. *Per-beam failure* is the fraction of beams whose response does not contain any of the BBEH answer prefixes recognised by the official extractor of Kazemi et al. [41]. *Recoverable problems* reports the fraction of problems for which at least one of the $K = 8$ beams parses successfully, the headroom a beam-level verifier could realise.

LLM	per-beam failure	recoverable problems
Llama-3.1 8B	42.9%	74.1%
Llama-3.3 70B	4.3%	99.0%
DS-R1-Qwen 32B	34.9%	68.1%
Skywork-OR1 32B	49.6%	51.7%
GPT-OSS 20B	58.3%	43.7%

E.3 Output-Length Statistics

Table 18 and Figures 12 and 13 report character-length statistics of the saved generations across the five LLMs we evaluate, computed over the entire dataset. Two patterns emerge. First, GPT-OSS-

20B, Skywork-OR1-32B and DS-R1-Qwen-32B exceed Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct at the median by roughly an order of magnitude, and their interquartile bands reach into the ten-thousand-character range. Second, individual tasks shift the entire column by a near-constant offset. Hard combinatorial tasks like Web of Lies, Zebra Puzzles, Multistep Arithmetic and Shuffled Objects remain at the top of every column, while short answer-format tasks like NYCC, Linguini, SportQA and SARC Triples remain at the bottom, and each LLM’s relative ordering of tasks is preserved across columns. Llama-3.3-70B-Instruct produces the tightest distribution of the five, with both the narrowest IQR and the lowest mean despite a comparable median to the 8B checkpoint. The maximum of 818,647 characters for GPT-OSS 20B originates from a single pathological example in the Multistep Arithmetic task. The model emits roughly 7,000 characters of valid reasoning, then enters a runaway indentation loop in which each successive opening parenthesis is preceded by an increasing number of leading spaces, growing quadratically until the 8,192-token generation cap is reached. Approximately 816,000 of the 818,647 characters are whitespace. All eight beams degenerate identically due to deterministic beam search. Evaluation operates at the token level and is unaffected by character length.

Table 18: Output-length statistics in characters across LLMs over all generations. Each row reports the median, interquartile range, mean, 95th percentile and maximum of `generated_text` length over every beam.

LLM	Length in characters				
	median	IQR	mean	95th pct.	max
Llama-3.1 8B	2,076	140–6,859	5,826	29,423	42,582
Llama-3.3 70B	2,091	668–3,957	3,152	10,116	39,809
DS-R1-Qwen 32B	7,458	2,377–16,213	10,160	29,259	39,082
Skywork-OR1 32B	17,043	6,447–24,469	15,986	30,990	38,375
GPT-OSS 20B	19,316	5,093–27,129	18,053	33,773	818,647

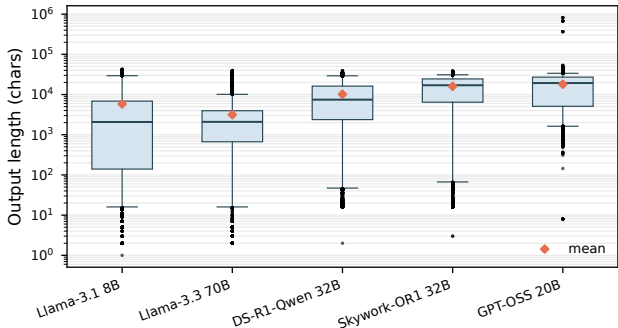


Figure 12: Output-length distribution per source LLM, in characters, pooled across every generation. Reasoning-distilled checkpoints (DS-R1-Qwen, Skywork-OR1) sit roughly an order of magnitude above the instruction-tuned Llama checkpoints at the median.

E.4 BBEH Accuracy of the Source LLMs

Table 19 reports BBEH correctness of the five LLMs scored with the official extraction logic of Kazemi et al. [41] against the canonical task targets. P@1 averages correctness over all beams per problem, while P@8 marks a problem solved if at least one beam scores correct, and the gap between the two columns quantifies the margin a verifier could realise from the existing beam set. Across tasks, Llama-3.3-70B-Instruct, Skywork-OR1-32B and GPT-OSS-20B remain ahead of the other two on P@1 within a few percentage points of each other, and DS-R1-Qwen-32B records the largest P@8 gain relative to its own P@1, while GPT-OSS-20B records the smallest such gain.

The bottom rows aggregate the per-task numbers two ways. The micro average pools all problems across tasks, while the adjusted harmonic mean uses the BBEH paper’s aggregation $N / \sum_i 1/(a_i + 1)$, which down-weights uneven performance across tasks. For the two LLMs that overlap with their Table 2 (Llama-3.1-8B-Instruct and DS-R1-Qwen-32B, reported there at 3.6% and 5.2%), our

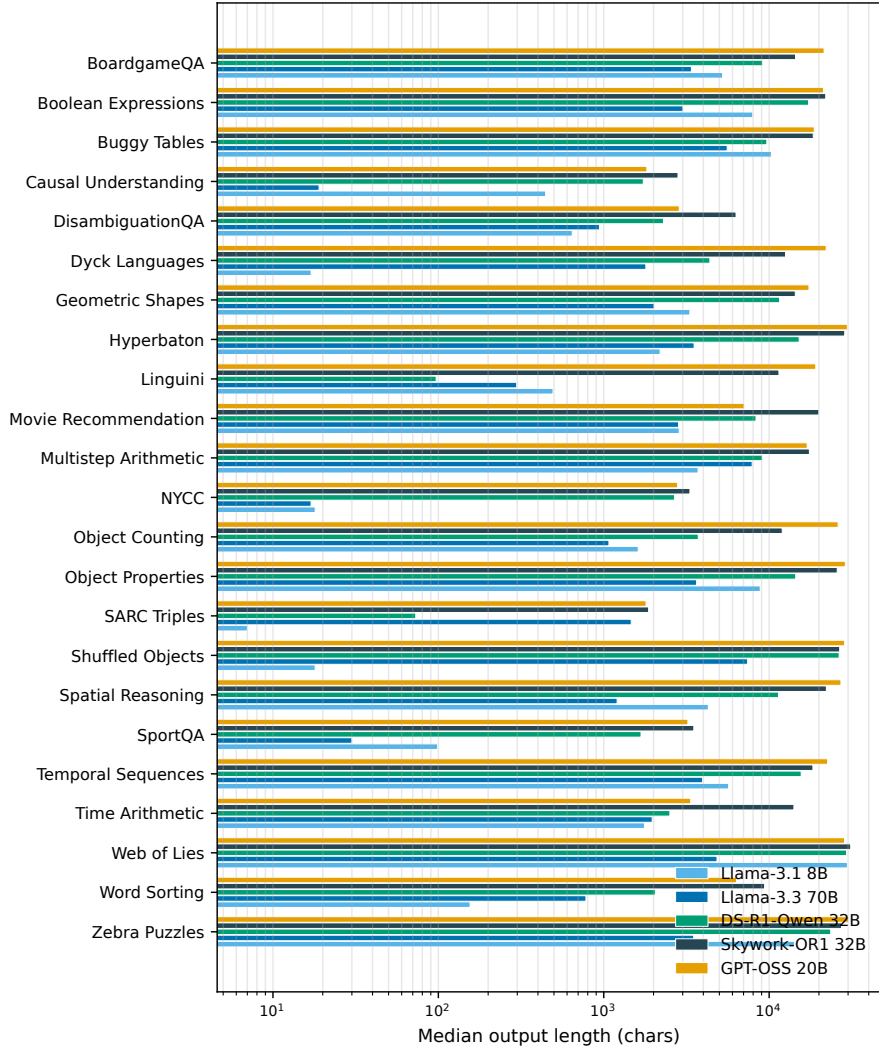


Figure 13: Median output length per BBEH task, in characters, with one bar per source LLM.

P@1 adjusted harmonic means agree to within the rounding granularity of the published per-task accuracies. Llama-3.3-70B-Instruct, Skywork-OR1-32B and GPT-OSS-20B were not evaluated under this aggregation in the original BBEH release, so the corresponding columns extend the benchmark’s leaderboard to those checkpoints.

E.5 Qualitative Beam Outputs and Embedder Similarity

To make the per-LLM beam pool concrete and to motivate why the Output Embedding (*Exact*) candidate of Section 3 captures the available semantic signal, we inspect pairs of beams from the same input prompt and contrast a high-similarity pair against a low-similarity pair. Cosine similarity is taken in the Llama-Embed-Nemotron-8B space, the same space used by the Pooled and Exact Output Embedding candidates throughout the paper. Both prompts are drawn from the held-out test split of Section E.4.

The eight beams of the Causal Understanding example 113 cluster into three regions of the embedder space. These are the long-form chain-of-thought mode and the two short-answer modes (“Yes” and “No”). Within each mode, beams sit at $\cos \geq 0.98$ from one another. Between the long-form mode and either short mode, \cos collapses to the $[0.06, 0.10]$ range regardless of whether the two beams agree on the final verdict, so the low score in Table 21 is driven by format divergence rather than

Table 19: BBEH accuracy (%) per task and source LLM. P@1 averages correctness over all beams per problem and P@8 marks a problem solved if at least one beam scores correct. Bottom rows give the micro average across all problems and the adjusted harmonic mean $N / \sum_i 1/(a_i + 1)$ over the per-task column above.

Task	Llama-8B		Llama-70B		DS-R1-Qwen 32B		Skywork-OR1 32B		GPT-OSS 20B	
	P@1	P@8	P@1	P@8	P@1	P@8	P@1	P@8	P@1	P@8
BoardgameQA	14.4	31.0	31.2	35.0	44.4	53.5	49.2	53.0	51.9	52.5
Boolean Expressions	17.6	30.0	24.1	49.0	16.4	17.5	22.8	24.0	6.6	7.0
Buggy Tables	0.0	0.0	0.6	1.5	0.5	0.5	1.9	2.5	7.0	7.0
Causal Understanding	30.6	68.0	47.4	73.5	55.4	59.0	51.7	58.5	53.8	54.0
DisambiguationQA	33.4	60.0	55.2	68.3	40.6	46.7	51.0	54.2	40.6	41.7
Dyck Languages	4.1	32.0	5.9	16.5	6.8	7.5	20.7	21.0	3.6	4.0
Geometric Shapes	13.5	26.0	30.4	51.5	7.6	11.5	1.5	1.5	1.0	2.0
Hyperbaton	0.4	1.0	4.5	12.5	0.4	0.5	3.5	3.5	0.7	1.0
Linguini	1.6	3.5	7.8	16.5	6.0	10.0	7.4	10.0	6.9	7.0
Movie Recommendation	12.7	29.0	62.9	70.0	34.8	60.5	40.2	47.5	64.3	68.0
Multistep Arithmetic	0.0	0.0	7.2	9.5	22.6	23.5	35.6	37.0	27.2	29.5
NYCC	12.9	47.5	10.8	27.0	10.1	32.0	12.2	24.5	18.0	18.0
Object Counting	0.0	0.0	14.6	19.5	3.4	3.5	9.1	10.0	0.0	0.0
Object Properties	0.5	2.0	1.1	3.0	0.0	0.0	0.0	0.0	0.7	1.0
SARC Triples	13.9	40.0	33.7	44.0	19.8	54.5	24.8	32.5	19.2	21.0
Shuffled Objects	12.8	58.0	12.4	34.0	2.3	5.5	2.9	3.0	1.6	2.5
Spatial Reasoning	5.2	8.5	11.4	17.0	14.2	15.0	16.3	17.5	0.5	0.5
SportQA	2.8	11.5	15.9	28.0	23.9	29.0	24.3	26.5	17.2	22.5
Temporal Sequences	0.6	1.5	2.9	5.5	4.9	5.0	0.0	0.0	0.5	0.5
Time Arithmetic	5.4	11.5	30.6	36.5	42.1	48.5	48.7	52.0	70.1	78.0
Web of Lies	4.2	8.0	24.3	29.5	17.6	18.0	18.3	18.5	7.0	8.0
Word Sorting	3.2	18.0	20.6	34.5	25.2	39.5	33.2	43.0	46.8	56.0
Zebra Puzzles	3.2	11.0	25.4	53.5	6.1	12.0	1.6	3.0	3.1	4.5
Micro avg.	7.9	21.0	20.3	31.3	17.2	23.7	20.2	23.1	19.1	20.8
Adj. HM	3.1	4.4	8.7	15.1	5.5	6.3	5.3	5.7	4.0	4.5

Table 20: High-similarity pair from example 107 of the Time Arithmetic task. Both beams walk through the same arithmetic chain and converge on the same answer, with only the wording of the closing sentence differing.

Prompt (truncated). <i>Question1... If the date for the day after yesterday is 2017-12-02, what is the date 1116 days from now? Question2: A plane leaves Location A at 01 AM UTC+0000 and arrives in Location B X':Y':Z' later. Location B is in IST(+0530). What time is it in Location B when the plane lands?</i>	
Beam 1	Beam 2
<i>To find the date 1116 days from 2017-12-02... $1116 \div 365 = 3$ years with a remainder of 111... $2020-12-02 + 111$ days = 2021-03-23... $X' = 3, Y' = 24, Z' = 9$... UTC+0000 01:00:00 + 03:24:09 = 04:24:09 + 5h + 0.5h = UTC+0530 09:54:09. So, the plane lands on the same day at 09:54:09 in Location B. The answer is: same_day, 09:54:09</i>	<i>To find the date 1116 days from 2017-12-02... $1116 \div 365 = 3$ years with a remainder of 111... $2020-12-02 + 111$ days = 2021-03-23... $X' = 3, Y' = 24, Z' = 9$... UTC+0000 01:00:00 + 03:24:09 = 04:24:09 + 5h + 0.5h = UTC+0530 09:54:09. So, the plane lands on the same day, 09:54:09. The answer is: same_day, 09:54:09</i>
Embedder-space cosine $\cos = 0.997$	

answer disagreement. The showcased pair happens to combine both format divergence and answer disagreement. The long-form beam reaches the correct answer (Yes) while the short beam answers No, but a long-form Yes paired with a short-form Yes would yield a cosine in the same low range. The two short modes sit at $\cos \approx 0.80$ from each other despite contradicting each other on the verdict. The embedder therefore separates beams primarily by response format and only secondarily by the content of the answer itself. This format-first geometry is what lets the Output Embedding reach the high-purity, high- Δ_{\cos} region of Section D.7, and is also why no purely latent candidate inherits the same separation, since none of them preserves the surface form that the embedder keys on.

Table 21: Low-similarity pair from example 113 of the Causal Understanding task. The left beam walks through the rules in a multi-step argument; the right beam returns the verdict in one line. The two beams sit in opposite regions of the embedder space. They also disagree on the final answer: the long-form beam concludes Yes (matching the ground truth), while the short beam answers No.

Prompt (truncated). <i>Jim, Carol, Bob, and Nancy share electricity from 8–9 PM. The breaker fails if three lamps are on at the same time. Jim is not supposed to use his lamp on Thursdays. Carol, Bob, and Nancy may use theirs. This Thursday Jim, Carol, and Bob all turn on their lamps at 8 PM and the breaker fails. Did Jim cause the breaker to fail? Reply Yes / No / Ambiguous.</i>	
Beam 2 (long-form chain of thought)	Beam 7 (short direct answer)
<i>To determine if Jim turning on his lamp at 8 PM caused the circuit breaker to fail, let's analyze the situation step by step. 1. The breaker will fail if three people turn on their lamps at the same time. 2. Jim is not supposed to use his lamp on Thursdays, but he turned it on anyway. 3. Carol and Bob also turned on their lamps at the same time as Jim. 4. Since Carol and Bob are allowed to use their lamps on Thursdays, their actions did not cause the breaker to fail. 5. The breaker failed because three people (Jim, Carol, and Bob) turned on their lamps at the same time. Considering these points, the majority of people would likely say that Jim turning on his lamp at 8 PM caused the circuit breaker to fail. . .</i>	<i>The answer is: No.</i>
<i>Verdict: Yes (correct)</i>	<i>Verdict: No</i>
Embedder-space cosine $\cos = 0.063$	

F Broader Impacts

The work is a diagnostic protocol for representations inside frozen LLMs and produces no new generative capability, no new dataset, and no deployed system. The protocol contributes to better reasoning models and a deeper understanding of functional thought representations in LLMs. We do not identify negative societal effects of the contribution.

G Earlier Formulations

G.1 Discriminator-Based DCS Evaluation Protocol

The earlier formulation of DCS reused the same-task discriminator f_{disc} to score within-question beam pairs. For each test problem x with $K = 8$ beams, two variants of the semantic equivalence matrix were constructed. $\mathbf{E}^{\text{emb}} \in \{0, 1\}^{K \times K}$ sets $\mathbf{E}_{ij}^{\text{emb}} = 1$ when the cosine similarity between the Nemotron embeddings of y_i and y_j exceeds $\tau = 0.9$, and 0 otherwise. $\mathbf{E}^{\text{parse}} \in \{0, 1\}^{K \times K}$ sets $\mathbf{E}_{ij}^{\text{parse}} = 1$ when the extracted final answers from y_i and y_j match exactly using the official BBEH answer extraction logic [41], with beam pairs for which either side has no extractable answer excluded rather than labelled non-equivalent. The functional similarity matrix $\mathbf{M} \in [0, 1]^{K \times K}$ was defined by the symmetric cross score $\mathbf{M}_{ij} = \frac{1}{2}(f_{\text{disc}}(\mathbf{T}_i, y_j) + f_{\text{disc}}(\mathbf{T}_j, y_i))$, and the score was the inverse mean absolute error between \mathbf{M} and \mathbf{E} over off-diagonal pairs,

$$\text{DCS}_{\text{disc}}(x) = 1 - \frac{1}{K(K-1)} \sum_i \sum_{j \neq i} |\mathbf{M}_{ij} - \mathbf{E}_{ij}|. \quad (10)$$

This formulation proved uninformative in practice. The discriminator was trained on cross-question pairs and provided no gradient signal for within-question scoring, so f_{disc} returned values near 0.5 for all same-question beam pairs. This collapsed DCS_{disc} to the random baseline for every representation family and source LLM (see Table 26).

G.2 Causality with the Discriminator-Trained Projection

Table 22 reports the causality KL under the discriminator-trained projection of Section C.5. This was the projection used in our first iteration of the causality protocol, before the projection-swap pilot of Section D.4 motivated adopting the output-reconstruction projection in Table 10. The earlier table is retained so that the effect of the projection swap remains visible cell-by-cell against the current results. Without length normalisation and under the discriminator-trained projection, several causality values are not well calibrated. Random vectors receive lower causality scores than expected because the model can ignore an uninformative input and produce whatever is more aligned with the output. Several other candidate representations produce higher causality than the random vector, which contradicts the behavioural reading the metric should provide.

Table 22: Causality KL (\downarrow) across source LLMs under the discriminator-trained projection of Section C.5. The discriminator dataset expands single-vector representations to a common training length and repeats shorter multi-vector representations to that same length, so the projection itself is learned at fixed length and is internally consistent with feeding \mathbf{T} at the same length here.

LLM	Output Emb.		Last Input Tok.				Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baselines	
	Exc	Pool	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE	RV		
Llama-3.1 8B	4.89	4.69	6.13	4.52	4.43	5.83	9.31	8.93	7.31	4.96	7.46	7.94	8.40	7.15	4.21	8.06	4.37	9.81	8.14	7.46	4.48		
	± 0.09	± 0.09	± 0.06	± 0.07	± 0.06	± 0.08	± 0.09	± 0.08	± 0.08	± 0.16	± 0.10	± 0.08	± 0.06	± 0.07	± 0.08	± 0.09	± 0.09	± 0.10	± 0.13	± 0.07			
Llama-3.3 70B	4.90	8.45	7.41	10.36	4.85	9.34	7.69	9.61	10.51	5.84	6.93	10.27	7.18	8.53	6.85	8.99	9.47	8.09	9.52	4.54	4.44		
	± 0.08	± 0.06	± 0.07	± 0.04	± 0.09	± 0.07	± 0.05	± 0.09	± 0.07	± 0.08	± 0.12	± 0.06	± 0.05	± 0.06	± 0.05	± 0.05	± 0.05	± 0.05	± 0.12	± 0.05			
DS-R1-Qwen 32B	4.60	4.81	9.21	9.18	7.01	6.20	7.24	9.20	7.74	8.15	7.42	6.83	8.90	7.34	4.74	5.52	9.55	7.67	8.02	6.56	4.54		
	± 0.08	± 0.07	± 0.06	± 0.08	± 0.08	± 0.06	± 0.07	± 0.07	± 0.08	± 0.13	± 0.06	± 0.06	± 0.07	± 0.05	± 0.06	± 0.06	± 0.09	± 0.06	± 0.06	± 0.11	± 0.06		

G.3 Cross-Entropy Proxy for Minimality

The minimality formulation we initially proposed approximated the IB Lagrangian as a single cross-entropy gap $\Delta_{CE} = CE(X | \mathbf{T}) - CE(Y | \mathbf{T})$, on the intuition that a high gap reflects a representation that compresses the input while retaining output-relevant information. We retain that analysis here for transparency. Two issues motivated the corrected formulation Δ_{IB} adopted in Section 3 and derived in Section D.5. First, Δ_{CE} corresponds to the IB Lagrangian only at the trade-off weight $\beta = 1$, where the chain-rule decomposition cancels the sufficiency reward and leaves only the redundancy penalty $-I(X; \mathbf{T} | Y)$, so the metric is unbalanced before any empirical concern. Second, the input-reconstruction probe $CE(X | \mathbf{T})$ saturates uniformly across representations because no probe in our class can reconstruct the exact lexical phrasing of X from any latent \mathbf{T} , which collapses the gap into a function of $CE(Y | \mathbf{T})$ alone. The corrected Δ_{IB} replaces $CE(X | \mathbf{T})$ with the conditional $CE(X | Y, \mathbf{T})$, which estimates the residual mutual information $I(X; \mathbf{T} | Y)$ and does not saturate. The components reported below corroborate the saturation diagnosis on every source model.

A high Δ_{CE} driven by high $CE(X | \mathbf{T})$ (input compression) and low $CE(Y | \mathbf{T})$ (output retention) would reflect a minimal sufficient \mathbf{T} , whereas a high gap produced by both conditionals drifting upward reflects a \mathbf{T} that has lost information about both X and Y . Table 23 reports the input component per source LLM, and Table 24 reports the output component. The Random Vector anchor for the input probe converges to $CE(X | RV) \approx 1.87 \pm 0.04$ on every source LLM, since the probe sees the same input distribution and i.i.d. noise regardless of which model produced the candidates, so the column is omitted from Table 23 and reported once here. Every Llama-3.1-8B cell of $CE(X | \mathbf{T})$ has a CI that overlaps this anchor, so the input component alone does not discriminate among candidates on this model. The output component does separate RV from every other candidate, since the Random Vector $CE(Y | \mathbf{T})$ is above the upper CI bound of every non-RV cell.

The two anchor candidates make the rest of the table interpretable. The Output Embedding (*Exact*) achieves the lowest $CE(Y | \mathbf{T})$ on Llama-3.1-8B, since the probe is being asked to predict Y from a representation of Y itself, and the Pooled variant follows behind with the per-beam information collapsed away. The Input Embedding is at the centre of the candidate cloud, since the input prompt alone determines the high-probability output and a probe over its embedding can already predict Y before any latent computation has happened. Any candidate above the Input Embedding’s $CE(Y | \mathbf{T})$ has drifted away from the input without acquiring additional Y -relevant content, while any candidate below it has packed in further output-relevant information beyond what the prompt already carries.

The same anchor reading carries to the implied gap. Subtracting the two components above on Llama-3.1-8B places the Input Embedding’s Δ_{CE} already in the high 0.9 range, so the prompt itself,

before any latent computation, already exhibits the minimality profile that a thought representation is supposed to provide. Most thinking candidates land at or below that input-alone gap, and only the Output Embedding (*Exact*) clearly exceeds it. The proxy therefore separates Random Vector from the rest of the field, but it also reports that latent thinking does not produce a more compressed, output-relevant summary of the problem than directly embedding the input prompt.

Table 23: Input-reconstruction cross-entropy $CE(X | \mathbf{T})$ across source LLMs. Higher values indicate that \mathbf{T} carries less information about the input prompt. The Random Vector baseline is omitted because the probe sees the same X and i.i.d. noise on every source LLM, so $CE(X | RV) \approx 1.87$ uniformly and serves as a single shared anchor referenced in the surrounding prose.

LLM	Output Emb.		Last Input Tok.		Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baseline
	Exc	Pool	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE
Llama-3.1 8B	1.77	1.78	1.74	1.75	1.74	1.74	1.75	1.77	1.78	1.75	1.75	1.76	1.77	1.78	1.74	1.74	1.75	1.76	1.76	1.78
	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04
Llama-3.3 70B	1.75	1.77	1.70	1.73	1.74	1.69	1.72	1.74	1.76	1.74	1.72	1.74	1.75	1.76	1.72	1.72	1.73	1.73	1.74	1.78
	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04
DS-R1-Qwen 32B	1.75	1.78	1.73	1.74	1.74	1.75	1.76	1.78	1.80	1.75	1.78	1.78	1.79	1.81	1.73	1.72	1.73	1.74	1.75	1.78
	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04	± 0.04

Table 24: Output-prediction cross-entropy $CE(Y | \mathbf{T})$ across source LLMs at each representation’s native sequence length. Lower values indicate that \mathbf{T} retains information sufficient to predict the output sequence. The tiled-length companion in Table 12 reports the same quantity with every representation fed at a common length, isolating the effect of length normalisation.

LLM	Output Emb.		Last Input Tok.		Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baselines	
	Exc	Pool	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE	RV
Llama-3.1 8B	0.77	0.84	0.78	0.81	0.85	0.82	0.84	0.86	0.83	0.89	0.84	0.88	0.88	0.90	0.83	0.78	0.79	0.80	0.82	0.83	1.91
	± 0.03	± 0.02	± 0.02	± 0.02	± 0.03	± 0.02	± 0.02	± 0.02	± 0.03	± 0.03	± 0.03	± 0.02	± 0.02	± 0.03	± 0.02	± 0.02	± 0.02	± 0.02	± 0.02	± 0.02	± 0.02
Llama-3.3 70B	1.25	1.31	1.16	1.22	1.38	1.21	1.23	1.25	1.39	1.24	1.27	1.28	1.27	1.28	1.24	1.25	1.27	1.29	1.35	2.35	
	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.07
DS-R1-Qwen 32B	0.83	0.86	0.76	0.78	0.83	0.82	0.83	0.87	0.89	0.86	0.89	0.87	0.88	0.89	0.82	0.83	0.84	0.86	0.87	0.91	1.33
	± 0.03	± 0.03	± 0.02	± 0.02	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.02	± 0.03	± 0.03	± 0.03	± 0.03	± 0.03	± 0.05

G.4 Qualitative Probe Reconstruction

Table 25 shows a concrete example of input probe reconstruction failure. Given the Last Input Token thought representation \mathbf{T} extracted from a BoardGame QA problem (a multi-step logical deduction task from BBEH), the input probe (a linear projection followed by a frozen LLaMA-3.2-1B) attempts to reconstruct the original input sequence X .

Table 25: Qualitative example of input probe reconstruction failure. The probe is conditioned on the Last Input Token thought representation ($\Delta_{CE} = 0.18$, $CE(X|\mathbf{T}) = 1.63$). All 10 sampled problems from this task produce the same boilerplate prediction.

Reference Input X (truncated)	Probe Prediction \hat{X} given \mathbf{T}
<i>A few players are playing a boardgame. The current state of the game is as follows. The akita is watching a movie from 1974. The akita is a high school teacher. The dove has a football with a radius of 30 inches. The elk is 3 years old. The german shepherd unites with the ant. [...] And the rules of the game are as follows. Rule1: ... Rule2: ... Rule3: ...</i>	<i>Question: Which of the following is not a valid argument for the claim that the number of people who are not in the United States is greater than the number of people who are in the United States?</i> A. The number of people who are in the United States is greater than ... B. ... C. ... D. ... Answer: D

Crucially, all 10 sampled BoardGame QA problems yield the same boilerplate prediction, namely a multiple-choice question about a completely unrelated topic. This indicates that the thought representation \mathbf{T} has discarded all problem-specific content. The probe has learned to output a domain-generic template (plausible given the BBEH multiple-choice format) rather than recovering any instance-specific detail. This qualitatively confirms the same-task separability collapse. If the

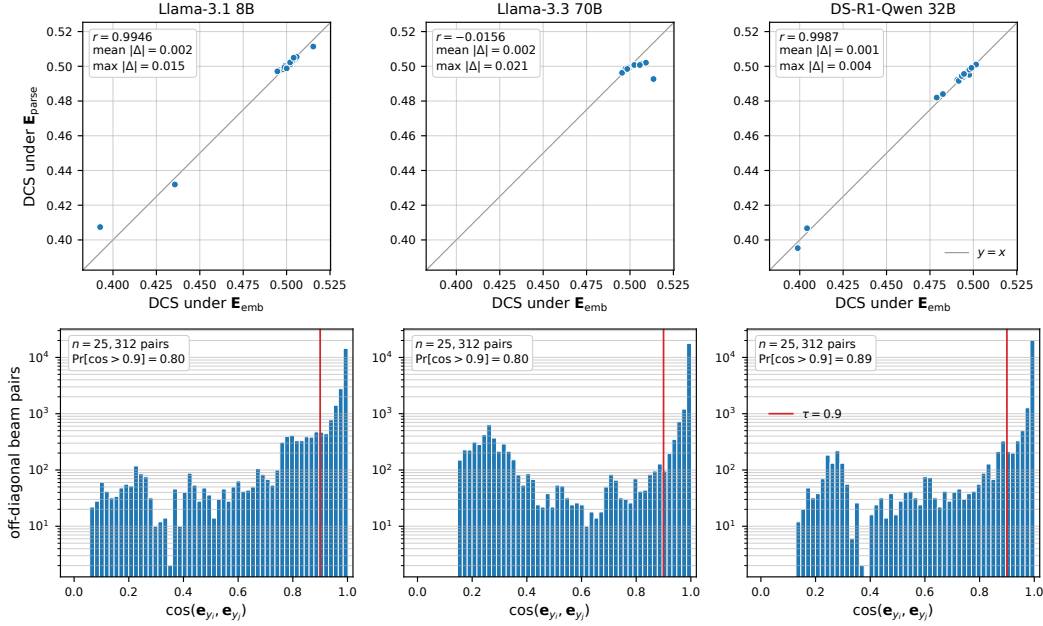


Figure 14: Discriminator-based DCS. The top row shows the score under $\mathbf{E}_{\text{parse}}$ against the same score under \mathbf{E}_{emb} for every representation, one panel per source model. The bottom row shows the empirical distribution of off-diagonal pairwise beam cosine similarities, one panel per source model, with the threshold $\tau=0.9$ overlaid.

probe cannot distinguish between 10 different board game problems when conditioned on \mathbf{T} , neither can the same-task discriminator.

G.5 Discriminator-Based DCS Results

The discriminator-based DCS formulation and its failure mode are documented in Section G.1. The results below retain that analysis for transparency.

Figure 14 (top row) plots discriminator-based DCS under $\mathbf{E}_{\text{parse}}$ against the same score under \mathbf{E}_{emb} for every representation, one panel per source model. The two variants agree closely on every model, so the table reports \mathbf{E}_{emb} . On Llama-3.3-70B-Instruct every representation crowds near 0.5, so the linear correlation collapses to noise even though the absolute disagreement remains small. \mathbf{E}_{emb} also remains defined on beams where BBEH answer extraction fails (see Section E).

Figure 14 (bottom row) shows the cosine-similarity distribution over off-diagonal beam pairs. The distribution is bimodal on every model, with a large right mode near 1 accounting for paraphrase-equivalent pairs and a small left mode below 0.5 for genuine cross-answer pairs.

Table 26: Discriminator-based DCS under \mathbf{E}_{emb} ($\tau=0.9$, \uparrow) across source LLMs, using the formulation documented in Section G.1. Values near 0.5 across representations reflect discriminator saturation on within-question pairs.

LLM	Output Emb.		Last Input Tok.		Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baselines	
	Exc	Pool	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE	RV
Llama-3.1 8B	0.393	0.436	0.505	0.499	0.502	0.504	0.515	0.506	0.502	0.502	0.503	0.504	0.500	0.499	0.505	0.499	0.495	0.498	0.497	0.498	0.499
	± 0.012	± 0.010	± 0.001	± 0.002	± 0.001	± 0.001	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.000	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.003
Llama-3.3 70B	0.514	0.506	0.500	0.498	0.509	0.502	0.499	0.497	0.498	0.506	0.503	0.495	0.498	0.497	0.498	0.498	0.498	0.498	0.498	0.498	0.499
	± 0.012	± 0.001	± 0.001	± 0.000	± 0.001	± 0.001	± 0.000	± 0.001	± 0.000	± 0.001	± 0.000	± 0.001	± 0.001	± 0.001	± 0.000	± 0.000	± 0.000	± 0.001	± 0.000	± 0.000	± 0.001
DS-R1-Qwen 32B	0.399	0.404	0.502	0.496	0.498	0.492	0.482	0.494	0.496	0.499	0.498	0.493	0.495	0.495	0.496	0.497	0.491	0.496	0.497	0.479	0.497
	± 0.010	± 0.010	± 0.000	± 0.001	± 0.001	± 0.003	± 0.004	± 0.003	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.000	± 0.000	± 0.000	± 0.001	± 0.000	± 0.000	± 0.005	± 0.001

Table 27 sweeps the cosine-similarity threshold τ on Llama-3.1-8B-Instruct. All rows share the same test split, trained discriminator, and beam embeddings. Only the binarisation $\mathbf{E}_{\text{emb}}^{ij} = \mathbb{1}[\cos(y_i, y_j) > \tau]$ varies. The ranking is stable across the full range and the chance-floor cluster does not reshuffle at

any threshold, confirming that the saturation is intrinsic to the discriminator-based scoring and not an artefact of the threshold choice.

Table 27: Discriminator-based DCS threshold sensitivity under $E_{\text{emb}}(\uparrow)$ on Llama-3.1-8B-Instruct. The column layout follows Table 26 and the $\tau=0.90$ row reproduces it cell-for-cell.

τ	Output Emb.		Last Input Tok.		Soft Thinking (no noise)					Soft Thinking (Gumbel)					Latent Thinking					Baselines	
	Exc	Pool	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	IE	RV
0.60	0.385	0.438	0.509	0.496	0.510	0.506	0.516	0.507	0.501	0.504	0.502	0.501	0.505	0.498	0.503	0.500	0.496	0.499	0.496	0.499	0.496
	± 0.012	± 0.010	± 0.001	± 0.002	± 0.001	± 0.001	± 0.003	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.000	± 0.002	± 0.001	± 0.002	± 0.001	± 0.001	± 0.001
0.70	0.387	0.437	0.509	0.497	0.509	0.506	0.516	0.507	0.501	0.504	0.503	0.501	0.505	0.498	0.504	0.500	0.496	0.499	0.496	0.499	0.496
	± 0.012	± 0.010	± 0.001	± 0.002	± 0.001	± 0.001	± 0.003	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.000	± 0.002	± 0.001	± 0.002	± 0.001	± 0.001	± 0.001
0.80	0.387	0.436	0.506	0.497	0.506	0.506	0.516	0.506	0.501	0.503	0.503	0.502	0.503	0.498	0.503	0.500	0.495	0.499	0.496	0.498	0.497
	± 0.012	± 0.010	± 0.001	± 0.002	± 0.001	± 0.001	± 0.003	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.000	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001
0.85	0.388	0.435	0.505	0.497	0.504	0.504	0.516	0.506	0.502	0.502	0.503	0.503	0.502	0.498	0.503	0.499	0.494	0.499	0.496	0.497	0.498
	± 0.012	± 0.010	± 0.001	± 0.002	± 0.001	± 0.001	± 0.003	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.000	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001
0.90	0.393	0.436	0.505	0.499	0.502	0.504	0.515	0.506	0.502	0.502	0.503	0.504	0.500	0.499	0.505	0.499	0.495	0.498	0.497	0.498	0.499
	± 0.012	± 0.010	± 0.001	± 0.002	± 0.001	± 0.001	± 0.002	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.000	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001
0.95	0.404	0.437	0.506	0.501	0.501	0.503	0.514	0.505	0.503	0.503	0.503	0.505	0.499	0.499	0.506	0.499	0.496	0.498	0.497	0.497	0.500
	± 0.011	± 0.009	± 0.001	± 0.002	± 0.001	± 0.001	± 0.002	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001	± 0.000	± 0.002	± 0.001	± 0.001	± 0.001	± 0.001	± 0.001