

What Makes a Latent Thought Valid?

Defining and Auditing Thought Representations in LLMs

Fahd Seddik¹ Fatemeh Fard¹

¹ FARD Lab, University of British Columbia, Okanagan, Canada

Correspondence: fahd.seddik@ubc.ca, fatemeh.fard@ubc.ca

 fard-lab.github.io/valid-thoughts

Abstract. Chain-of-Thought (CoT) reasoning improves the performance of Large Language Models (LLMs), yet it writes steps as text and follows one reasoning path at a time, which limits the capabilities of such models. Latent reasoning methods replace these steps with continuous thought representations and claim that a thought representation substitutes for the written reasoning, compresses trajectories, distinguishes semantic content, and achieves superposition of multiple exploration paths at once. These claims concern the content of the thought representation, whereas the evaluations of these methods rely on downstream task accuracy, token counts, and analyses specific to each method. We formalize these claims as a valid thought representation space and prove that a space is valid if and only if its thought representations attain four properties, namely causality, minimality, separability, and stability, each of which follows from one claim. Translating these properties into four measures and auditing three training-free and three trained methods shows gaps in these claims that accuracy may not necessarily reveal. Training-free methods fail all four properties and do not encode the reasoning they are meant to replace, whereas trained methods generally improve substitution. The four properties and measures therefore provide a theoretically grounded evaluation beyond downstream task accuracy and expose failure modes that may guide future development of latent reasoning methods.

1 INTRODUCTION

A language model may reason without writing its intermediate steps, by conditioning each step on a continuous thought representation instead of a token, a step called a latent step. Latent reasoning methods construct such thought representations to replace part or all of the written reasoning (Hao et al., 2025; Zhang et al., 2026). The replacement is motivated by substitution, in that a thought representation encodes the content of the reasoning steps that the model would otherwise write as text (Cheng & Van Durme, 2024; Deng et al., 2025). A second motivation is superposition, in that one thought representation holds several candidate reasoning paths together, as in a breadth-first search (Hao et al., 2025; Deng et al., 2025). These methods report a maintained or improved downstream task accuracy with fewer decoded tokens (Zhang et al., 2026; Hao et al., 2025; Tan et al., 2025; Mondorf & Plank, 2024). Both motivations are claims about the content of the thought representation, whereas the reported results measure task accuracy and the number of tokens.

Each latent reasoning method introduces a dedicated mechanism for constructing thought representations. Coconut uses the last hidden state as the input embedding of the next position (Hao et al., 2025), Soft Thinking weights the token embeddings by the next-token distribution (Zhang et al., 2026), and Cache-to-Cache fuses a projection of the key-value cache of one model into another (Fu et al., 2026). Each method further optimizes its mechanism toward an objective of its own design, such as the number of reasoning steps that condition on a thought representation, a token budget, or the imitation of written reasoning (Hao et al., 2025; Shen et al., 2025; Wei et al., 2026). These differences in objectives and evaluation criteria hinder comparison across methods and the identification of stable design principles (Yu et al., 2026). Therefore, which qualities a good thought representation should attain remains an open question. Audits find downstream task accuracy unchanged when the model bypasses the thought representations (Rizvi-Martel & Mosbach, 2026; Sahoo et al., 2026; Dilgren & Wiegrefe, 2026). Since thought representations are opaque to direct human inspection, outcome-level performance is insufficient to guide progress in latent reasoning (Yu et al., 2026). Downstream task accuracy

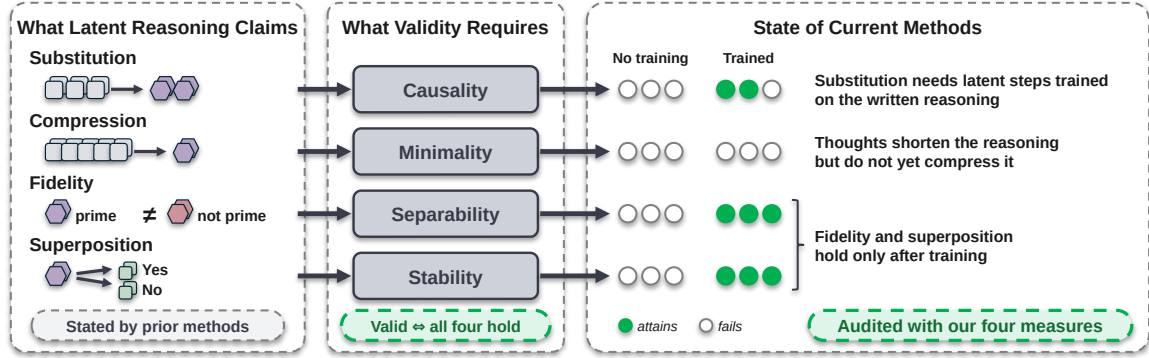


Figure 1: Latent reasoning methods claim that a thought representation substitutes for written reasoning, compresses it, keeps distinct semantic content distinct, and holds several answers together. Each claim corresponds to one property, and no current method attains all four.

thus cannot attribute a gain or a loss to a specific quality of the thought representation.

Prior work nevertheless applies individual requirements on a thought representation implicitly. Audits treat an answer that bypasses the thought representation, or that follows an inconsistent pathway, as a failure of latent reasoning (Cui et al., 2026; Dilgren & Wiegrefe, 2026; Sahoo et al., 2026). Methods that supervise thought representations against the written reasoning steps expect a thought representation to preserve the content of the replaced steps (Shen et al., 2025; Wei et al., 2026). Compression of the reasoning chain is the stated aim of thought representations that replace written reasoning (Cheng & Van Durme, 2024; Tan et al., 2025). Thought representations that lose their semantic diversity (Wei et al., 2026), and distinct reasoning paths that collapse to one interpretation (Rizvi-Martel & Mosbach, 2026), count as failures as well. These expectations recur across the construction, supervision, and audit of thought representations, which points to a common set of requirements on a thought representation. Furthermore, a systematic theory of latent representations remains underdeveloped (Yu et al., 2026).

What must a thought representation attain to substitute for written reasoning?

We answer this question by defining when a thought representation space is valid. The space holds the thought representations that a model assigns to its contexts. In a valid thought representation space, the thought representation of a context substitutes for the generated text of that context while preserving the distribution over the subsequent tokens. Two contexts further share a thought representation if and only if the model assigns the two contexts the same distribution over outputs. Validity thus states the claims of latent reasoning literature (in Figure 1), namely substitution, compression, fidelity, and superposition, as requirements on the thought representation space. We show that validity is equivalent to four properties of a thought representation, namely causality, minimality, separability, and stability, which unify the individual requirements of prior work. Hence, a thought representation that lacks any of the four properties would not satisfy the claims that motivate its method. We translate each property into a measure computed on the thought representation. The measures identify the property that a thought representation lacks and quantify it, independently of downstream task accuracy and of human inspection.

The main contributions of this paper are as follows.

- **Definition of a valid thought representation.** Four properties that a thought representation may attain, namely causality, minimality, separability, and stability (Sections 2 and 3). We prove that each property is attainable, independent of the other three, and that, over a finite set of inputs, a model and a generator attain all four properties in their strongest form if and only if their thought representation space is valid (Section 3.2).
- **Measures independent of downstream task accuracy.** The lack of standardized evaluation protocols is a central open challenge of latent reasoning (Yu et al., 2026). We supply such a protocol by translating each property into a measure computed on the thought representation (Section 3).
- **Audit of existing thought representations.** We audit six latent reasoning methods, three without training on five open-weight models over the 23 reasoning tasks of Big-Bench Extra Hard (BBEH) (Kazemi et al.,

2025) and three trained on GSM8K (Cobbe et al., 2021) (Section 4). Every training-free method fails all four properties. On the trained methods, the measures track the questions answered incorrectly and distinguish methods that reach similar accuracy.

2 VALID THOUGHT REPRESENTATIONS

2.1 SEMANTIC EQUIVALENCE

Let f_θ denote a language model that generates an output $y \in \mathcal{Y}$ for an input $x \in \mathcal{X}$. Let $\Phi : \mathcal{Y} \rightarrow \mathcal{S}$ map each generated output to the element of the semantic output space $\mathcal{S}$ that the output expresses.

Definition 1 (Semantic Equivalence). Two outputs $y, y' \in \mathcal{Y}$ are semantically equivalent, written $y \sim_{\text{sem}} y'$, when

$$\Phi(y) = \Phi(y'). \quad (1)$$

2.2 VALID THOUGHT REPRESENTATION SPACE

A thought representation is a latent state that mediates the transformation from the input space $\mathcal{X}$ to the semantic output space $\mathcal{S}$. Given an input x , the model induces a distribution $P(Y \mid x)$ over possible outputs. A thought representation encodes the sufficient statistics of that distribution. We explicitly exclude interpretability, which is a property of an observer. Optimality in mediating the transformation may render a thought representation opaque to human understanding. Let $\mathcal{T}$ denote the thought representation space, the set of thought representations that the model assigns to its inputs.

Definition 2 (Thought Generator). Let $g : \mathcal{X} \rightarrow \mathcal{T}$ assign a thought representation to each input, and let y_i denote the output that f_θ generates for x_i under a fixed decoding rule. The generator induces an equivalence relation on $\mathcal{X}$ through the semantic output space, in that for any pair of inputs $x_i, x_j \in \mathcal{X}$,

$$g(x_i) = g(x_j) \quad \text{if and only if} \quad y_i \sim_{\text{sem}} y_j. \quad (2)$$

Running Example

Let x denote the question “Is 13 prime, and why?”, and assume that f_θ generates the outputs

$y_1 = \text{“Yes. 13 is prime because no integer from 2 to 12 divides it.”}$

$y'_1 = \text{“Yes, 13 has no divisors other than 1 and itself.”}$

$y_2 = \text{“No. 13 is divisible by 3.”}$

The pair y_1 and y'_1 encodes one answer in different wording and satisfies Definition 1, whereas y_2 answers the same question with a different value.

The generator is many-to-one, which compresses $\mathcal{X}$ into $\mathcal{T}$ and preserves distinctness only between inputs whose outputs differ in semantic content. Let x' denote the input “Determine whether 13 is a prime number and justify the answer.”, and assume that f_θ generates an output y' for x' with $y' \sim_{\text{sem}} y_1$. Definition 2 then gives $g(x') = g(x)$.

A valid thought representation space maps every thought representation to a semantic output. Conversely, every semantic output that the model can generate is mapped to a thought representation. We therefore posit a bijection between the thought representation space $\mathcal{T}$ and the semantic output space $\mathcal{S}$. However, before generation ends, the output is undetermined. In this case, a thought representation encodes the distribution over the outputs that may follow its context. Once generation ends, the output is fixed and it encodes the semantic content of that output. Therefore, the bijection is stated on the distribution induced by a context.

Let $\mathcal{C}$ denote the set of contexts, where a context $c = [x; u]$ is an input $x \in \mathcal{X}$ together with the segment u generated after that input. The segment u is empty before generation begins. The generated output is then $y = [u; v]$, where v denotes the tokens generated after the context. The generator applies to every context, in that $\mathbf{T}_c = g(c)$ denotes the thought representation of a context and $\mathcal{T} = g(\mathcal{C})$ denotes the thought representation space. Let $E(u)$ denote the token embeddings of the segment u in f_θ . Let $[x; \mathbf{T}]$ denote the

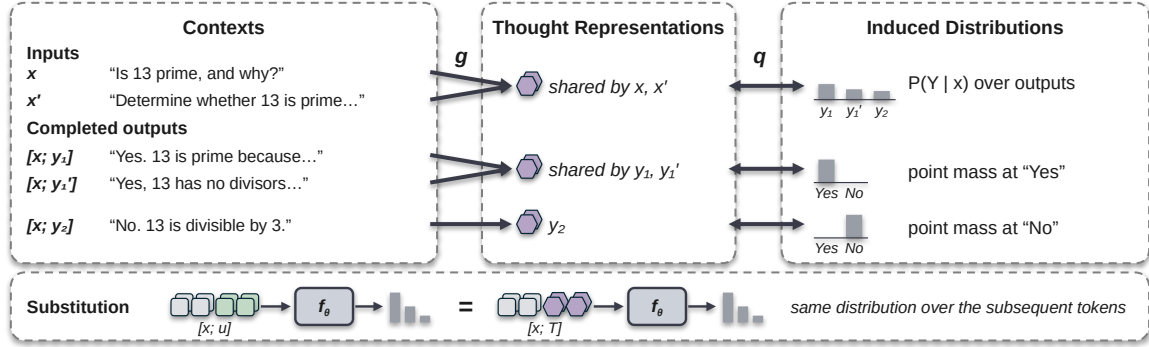


Figure 2: **Valid thought representation space.** The generator g maps the inputs x and x' , and maps the completed outputs y_1 and y_1' , to one representation, where each corresponds to exactly one induced distribution. Substituting a thought representation for its text preserves the distribution.

context in which $\mathbf{T}$ replaces the embeddings $E(u)$ of the generated segment. Let $\Delta(\mathcal{Y})$ and $\Delta(\mathcal{S})$ denote the probability distributions on the output space and on the semantic output space,

$$\Delta(\mathcal{A}) = \left\{ \tau : \mathcal{A} \rightarrow [0, 1] \text{ with } \sum_{a \in \mathcal{A}} \tau(a) = 1 \right\}, \quad \mathcal{A} \in \{\mathcal{Y}, \mathcal{S}\}. \quad (3)$$

Definition 3 (Valid Thought Representation Space). Let $\mathcal{C}_{\text{in}}$ and $\mathcal{C}_{\text{out}}$ denote the contexts whose generated segment is empty and the contexts whose generated segment is a complete output, let $\mathcal{T}_{\text{in}}$ and $\mathcal{T}_{\text{out}}$ denote their thought representations, and let μ_c denote the distribution that a context induces,

$$\mathcal{C}_{\text{in}} = \{ x : x \in \mathcal{X} \}, \quad \mathcal{C}_{\text{out}} = \{ [x; y] : x \in \mathcal{X}, y \in \mathcal{Y} \}, \quad (4)$$

$$\mathcal{T}_{\text{in}} = g(\mathcal{C}_{\text{in}}), \quad \mathcal{T}_{\text{out}} = g(\mathcal{C}_{\text{out}}), \quad (5)$$

$$\mu_x = P(Y | x) \in \Delta(\mathcal{Y}), \quad \mu_{[x; y]} = P(\Phi(Y) | [x; y]) \in \Delta(\mathcal{S}). \quad (6)$$

The space $\mathcal{T}$ is valid when the thought representation substitutes for the text at every context $[x; u]$ and every sequence v of subsequent tokens,

$$P(v | [x; u]) = P(v | [x; \mathbf{T}_{[x; u]}]), \quad (7)$$

and when, for each $i \in \{\text{in}, \text{out}\}$, two contexts in $\mathcal{C}_i$ have one thought representation if and only if the two contexts induce one distribution,

$$g(c_1) = g(c_2) \iff \mu_{c_1} = \mu_{c_2}, \quad c_1, c_2 \in \mathcal{C}_i. \quad (8)$$

The map $q_i(g(c)) = \mu_c$ is then a bijection from $\mathcal{T}_i$ onto the distributions $\{\mu_c : c \in \mathcal{C}_i\}$ induced by the contexts in $\mathcal{C}_i$.

On completed outputs, the bijection of Definition 3 pairs each semantic output that the model generates with one thought representation (Corollary A.8). Figure 2 illustrates Definition 3 on the running example.

3 CHARACTERIZING VALID THOUGHT REPRESENTATIONS

3.1 THE FOUR PROPERTIES AND MEASURES

Causality. As thought representations are used to substitute explicit text reasoning, the thought representation should encode the necessary information about that reasoning. Therefore, this motivates the first property to compare such substitution of text with a thought representation.

Definition 4 (Causality). Causality holds when substituting $\mathbf{T}$ for the embeddings $E(u)$ inside f_θ leaves the distribution over v unchanged,

$$P(v | [x; u]) = P(v | [x; \mathbf{T}]). \quad (9)$$

In the notation of Section 2.2, $\mathbf{T} = \mathbf{T}_{[x;u]}$ is the thought representation of the context $[x; u]$. Let $u = y_1$ of the running example, and let v denote the segment “Therefore, 13 is prime.” generated after u . Causality requires the distribution over v to be the same whether f_θ conditions on the tokens of y_1 or on $\mathbf{T}$.

Measure. Since the equality of Definition 4 cannot be computed on a finite sample, the causality error reports the divergence between the two distributions as a surrogate. The causality error takes v as the final n tokens of a generated output $y = [u; v]$ and averages the divergence between the next-token distributions P of f_θ over these tokens,

$$\text{KL}_{\text{caus}}(\mathbf{T}) = \frac{1}{n} \sum_{t=1}^n D_{\text{KL}} \left(P(\cdot \mid [x; u], v_{<t}) \parallel P(\cdot \mid [x; \mathbf{T}], v_{<t}) \right). \quad (10)$$

Let $\text{KL}_{\text{caus}}(\emptyset)$ denote the causality error omitting $\mathbf{T}$. The reasoning recovered by $\mathbf{T}$ would be,

$$R(\mathbf{T}) = 1 - \frac{\text{KL}_{\text{caus}}(\mathbf{T})}{\text{KL}_{\text{caus}}(\emptyset)}, \quad (11)$$

which equals 1 for $\mathbf{T} = E(u)$ and 0 when $\mathbf{T} = \emptyset$.

Minimality. Latent reasoning methods compress reasoning trajectories into fewer steps than the written reasoning. The number of steps limits how much information a thought representation may hold, but not which information it holds. Therefore, the second property states compression on information, in that a thought representation keeps only what the output needs and discards the rest of the input. The mutual information $I(X; \mathbf{T})$ measures the content of the input that $\mathbf{T}$ encodes. The mutual information $I(\mathbf{T}; Y)$ measures the content of the generated output that $\mathbf{T}$ encodes.

Definition 5 (Minimality). Let I_0 be a threshold on the information that $\mathbf{T}$ retains about Y . Minimality holds when $\mathbf{T}$ minimizes $I(X; \mathbf{T})$ subject to a constraint on $I(\mathbf{T}; Y)$,

$$\min_{\mathbf{T}} I(X; \mathbf{T}) \quad \text{subject to} \quad I(\mathbf{T}; Y) \geq I_0, \quad (12)$$

which the information bottleneck expresses as

$$\min_{\mathbf{T}} J(\beta_{\text{IB}}), \quad J(\beta_{\text{IB}}) = I(X; \mathbf{T}) - \beta_{\text{IB}} I(\mathbf{T}; Y). \quad (13)$$

The weight $\beta_{\text{IB}} > 0$ replaces the threshold I_0 by a penalty on $I(\mathbf{T}; Y)$. Let x'' denote the input “Hamlet was written by Shakespeare around 1600. Is 13 prime, and why?”, and assume that f_θ generates the output y_1 of the running example for it. Since the literary sentence did not contribute to y_1 , minimality requires $\mathbf{T}$ to discard the literary sentence.

Measure. Since the two mutual informations of Definition 5 cannot be computed, the information bottleneck residual reports a surrogate built from two cross-entropies of f_θ . Let $\text{CE}(Y \mid \mathbf{T})$ denote the cross-entropy of f_θ on the output after a prompt that supplies $\mathbf{T}$ and asks for the answer. Let $\text{CE}(X \mid Y, \mathbf{T})$ denote the cross-entropy of f_θ on the input after a prompt that also holds the output and asks to restate the question. We compute the residual $\Delta_{\text{IB}} = \text{CE}(X \mid Y, \mathbf{T}) - \text{CE}(Y \mid \mathbf{T})$, derived in Appendix B. The output information r_Y and the extra input information r_X normalize the two cross-entropies,

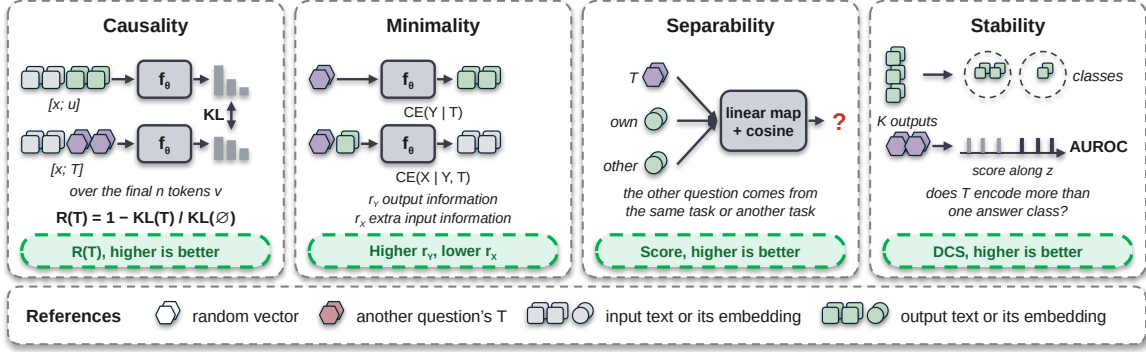
$$r_Y = \frac{\text{CE}(Y \mid \emptyset) - \text{CE}(Y \mid \mathbf{T})}{\text{CE}(Y \mid \emptyset) - \text{CE}(Y \mid Y)}, \quad r_X = \frac{\text{CE}(X \mid Y, \emptyset) - \text{CE}(X \mid Y, \mathbf{T})}{\text{CE}(X \mid Y, \emptyset) - \text{CE}(X \mid Y, X)}. \quad (14)$$

Each ratio equals 0 for $\mathbf{T} = \emptyset$ and 1 for the corresponding text. The output text attains $r_Y = 1$ and $r_X = 0$, the values of a minimal thought representation.

Separability. As semantically distinct outputs should have distinct thought representations, we should enforce separability among thought representations. A model with a finite vocabulary and a bounded output length generates finitely many outputs, which bounds the number of semantic outputs as well. To compare thought representations by the semantic content of their outputs, we assume that the semantic output space would exist in a vector space with a metric d_S , which at $d_S = 0$ denotes semantic equivalence,

Assumption 3.1 (Semantic Output Space). A finite subset of a vector space, $\mathcal{S} \subset \mathbb{R}^d$, where d_S is a metric on $\mathbb{R}^d$,

$$d_S(\Phi(y), \Phi(y')) = 0 \quad \text{if and only if} \quad y \sim_{\text{sem}} y'. \quad (15)$$


 Figure 3: **Measures.** What enters each measure and what it returns, with the references.

The metric is further assumed to order semantic outputs by subject, in that two answers to one question are closer to each other than to an answer to another question.

Assumption 3.2 (Semantic Proximity). For two outputs y_1 and y_2 that answer one question with different values, and an output y_3 that answers a question on another subject,

$$d_S(\Phi(y_1), \Phi(y_2)) < d_S(\Phi(y_1), \Phi(y_3)). \quad (16)$$

Definition 6 (Separability). Let c_1 and c_2 be contexts whose outputs are not semantically equivalent. Separability holds when some projection $\phi : \mathcal{T} \rightarrow \mathbb{R}^d$ separates the two representations by a margin $\delta \geq 0$,

$$d_S(\phi(\mathbf{T}_{c_1}), \phi(\mathbf{T}_{c_2})) > \delta. \quad (17)$$

Inputs whose outputs are semantically equivalent share one representation under Definition 2 and are indistinguishable under every projection. Let x''' denote the input “Is 14 prime, and why?”, and assume that f_θ generates the output $y_2' = \text{“No. } 14 = 2 \times 7\text{.”}$, which is not semantically equivalent to y_1 . The two inputs differ by one token. Separability nevertheless requires a projection ϕ that has $d_S > \delta$ for them.

Measure. Since neither Φ nor d_S can be computed on a trained model, the separability score replaces $\Phi(y)$ with the embedding of the output y under a text embedding model, and d_S with the cosine similarity between embeddings. Let $\bar{e}_c$ denote the mean embedding of the outputs generated for a context c . The projection ϕ is a linear map fitted by ridge regression from the thought representations of training contexts to their mean output embeddings. On two held-out contexts c_1 and c_2 , the projection separates c_1 from c_2 at margin δ when the projected thought representation of c_1 is closer to the outputs of c_1 than to the outputs of c_2 ,

$$\cos(\phi(\mathbf{T}_{c_1}), \bar{e}_{c_1}) - \cos(\phi(\mathbf{T}_{c_1}), \bar{e}_{c_2}) > \delta. \quad (18)$$

Since contexts from different tasks already differ in their text, the separability score is computed over pairs of held-out contexts from the same task.

Stability. As one input admits many outputs and decoding samples one of them, a thought representation should encode the distribution over semantic outputs and remain invariant to the wording sampled. Therefore, the fourth property states one condition for the wording and one condition for the distribution.

Definition 7 (Stability). Let $\mathbf{T}_y$ denote the representation associated with a generated output y . Stability holds when semantically equivalent outputs give equal representations, which is *lexical invariance*,

$$\mathbf{T}_{y_i} \approx \mathbf{T}_{y_j} \quad \text{whenever} \quad y_i \sim_{\text{sem}} y_j, \quad (19)$$

and when the semantic distribution that the input induces is recoverable from the representation, which is *mode collapse resistance*, in that a map q from $\mathcal{T}$ to the distributions over $\mathcal{S}$ satisfies, for every input x ,

$$q(\mathbf{T}_x) = P(\Phi(Y) | X = x). \quad (20)$$

In the notation of Section 2.2, $\mathbf{T}_y = \mathbf{T}_{[x;y]}$ is the thought representation of the completed context $[x; y]$.

Lexical invariance removes the dependence of $\mathbf{T}$ on the wording of the sampled output. Mode collapse resistance keeps the uncertainty of the semantic distribution in $\mathbf{T}$, hence a thought representation extracted from one sample reports the semantic distribution of the input. The outputs y_1 and y'_1 of the running example are equivalent and therefore give approximately equal representations. If the model answers “Yes” in some samples and “No” in others, stability then requires $\mathbf{T}$ to encode both classes.

Measure. Since the semantic distribution of Definition 7 cannot be computed, the distributional consistency score approximates the distribution by K outputs generated for the input. Binarizing the cosine similarity between the embeddings of these outputs at a threshold τ partitions the outputs into classes $\mathcal{E}_1, \mathcal{E}_2, \dots$, whose sizes give the semantic entropy of the input $\mathbb{H}_x$ (Kuhn et al., 2023). A difference-of-means (Cencerrado et al., 2026) direction recovers whether the input has more than one class,

$$z = \frac{1}{|\mathcal{D}_+|} \sum_{\mathbf{T} \in \mathcal{D}_+} \mathbf{T} - \frac{1}{|\mathcal{D}_-|} \sum_{\mathbf{T} \in \mathcal{D}_-} \mathbf{T}, \quad (21)$$

where $\mathcal{D}_+$ holds the thought representations of the inputs with $\mathbb{H}_x > 0$ and $\mathcal{D}_-$ the rest. The Distributional Consistency Score (DCS) is the Area Under the Receiver Operating Characteristic curve (AUROC) of the score $\langle z, \mathbf{T} \rangle$ for separating $\mathcal{D}_+$ from $\mathcal{D}_-$. Figure 3 summarizes the four measures.

3.2 UNIFIED FRAMEWORK

We prove consistency, independence, completeness, and necessity of the four properties. Details in Appendix A.

Consistency. Let $\mathcal{P}$ denote the set of the four properties. Since properties that contradict one another would not describe a thought representation, we first study whether the properties in $\mathcal{P}$ are consistent, that is, whether one model and one thought representation may attain all of them. The construction takes the thought representation of a context to be the distribution that the context induces over the semantic output space. The model conditions the tokens generated after each context only on that distribution.

Proposition 3.3 (Consistency). *The properties in $\mathcal{P}$ are jointly achievable, in that one model and one thought representation may attain all four requirements.*

Independence. Consistency, however, does not imply that the properties in $\mathcal{P}$ are free of redundancy. A consistent set may still hold a property that the remaining three imply, which would make that property redundant. Therefore, we prove that no property in $\mathcal{P}$ is implied by the conjunction of the other three. The four properties are requirements on a pair (f_θ, g) of a model and a generator. The generator g of such a pair is any map from contexts to a thought representation space, which may violate Definition 2. For each property, Appendix A edits the thought representation of the consistency model once, such that the edited pair attains the other three properties and lacks that property.

Corollary 3.4 (Independence). *No property in $\mathcal{P}$ is implied by the conjunction of the other three.*

Completeness. Only afterwards do we prove completeness, in the sense that $\mathcal{P}$ yields a valid thought representation space. Definition 3 states the target space of the four properties. Causality gives the substitution condition (7). Separability and lexical invariance give the bijection (8) on completed outputs. Minimality gives the bijection on inputs.

Theorem 3.5 (Completeness). *Let $\mathcal{X}$ be finite, and let a pair (f_θ, g) attain the four properties with the threshold of (12) at its largest attainable value, $I_0 = I(X; Y)$, and with lexical invariance (19) holding with equality. Then its thought representation space is valid.*

Necessity. Completeness shows only that the four properties suffice for validity. Therefore, we also prove the converse, that validity implies each of the four properties. The substitution condition of Definition 3 is causality. The bijection on inputs gives minimality and mode collapse resistance. Since a projection resolves any two distinct thought representations, both bijections give separability. The bijection on completed outputs gives lexical invariance with equality.

Proposition 3.6 (Necessity). *Let $\mathcal{X}$ be finite, and let the thought representation space of a pair (f_θ, g) be valid. Then (f_θ, g) attains causality at every context, minimality with the threshold $I_0 = I(X; Y)$, separability on $\mathcal{C}_{\text{in}}$ and on $\mathcal{C}_{\text{out}}$, and stability with lexical invariance (19) holding with equality.*

Together, Theorem 3.5 and Proposition 3.6 characterize validity, in that a thought representation space is valid if and only if the pair that generates it attains the four properties under the conditions of Proposition 3.6.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

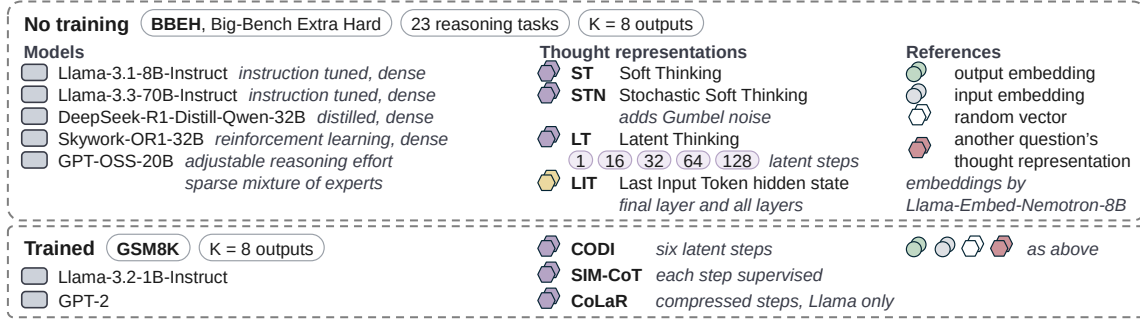


Figure 4: **Experimental setup.** The two audits, with the models, thought representations, and references.

Thought representations. Figure 4 summarizes the two audits. The audit without training takes the hidden state of the last input token, which encodes information about the tokens generated after the prompt (Lugoloobi et al., 2026), from the final layer and from all layers (Cai et al., 2026). The training-free methods are Soft Thinking (ST) (Zhang et al., 2026), Stochastic Soft Thinking (STN), which adds Gumbel noise (Wu et al., 2026), and Latent Thinking (LT) (Hao et al., 2025; Zou et al., 2026a), each at 1, 16, 32, 64, and 128 latent steps. The trained audit takes CODI (Shen et al., 2025), SIM-CoT (Wei et al., 2026), and CoLaR (Tan et al., 2025), which fine-tune Llama-3.2-1B-Instruct (Grattafiori et al., 2024) and GPT-2 (Radford et al., 2019) to reason in latent space (details in Appendix C).

Baselines. To establish baseline comparisons for thought representations, we utilize reference representations with known information content to evaluate specific properties. (i) *The input embedding* encodes only the question. A thought representation attains separability and stability when its measure exceeds this baseline, indicating it captures information about the outputs that the question itself lacks. (ii) *A random vector* replaces the latent steps to ensure no question-specific information is present. A thought representation attains causality if the reasoning recovered from it is closer to that of the written reasoning ($R = 1$) than to the reasoning recovered from this random baseline (confirming the effect stems from the content of the steps, not just their presence). (iii) *The output and question texts* serve as reference points in the (r_Y, r_X) space relative to $\text{CE}(Y | \emptyset)$ and $\text{CE}(X | Y, \emptyset)$. A thought representation attains minimality when its coordinate is closer to the output text than to either the question text or the origin. Finally, to control for general task information that is not specific to the prompt, we utilize the thought representation of a different question from the same task as an additional baseline.

Models and data. The five models used for the training-free audit span four model families, four training paradigms, two architectures, and four sizes (Grattafiori et al., 2024; DeepSeek-AI, 2025; He et al., 2025a;b; OpenAI, 2025). Each model generates $K = 8$ outputs per problem of the 23 reasoning tasks of BBEH (Kazemi et al., 2025). The trained methods generate $K = 8$ outputs per question of GSM8K (Cobbe et al., 2021). To compare the effect of training on one model and one task, the three training-free methods also run on Llama-3.2-1B-Instruct over GSM8K.

4.2 RESULTS

Table 1 marks the properties that each method attains. Every training-free method fails all four properties, whereas CODI and SIM-CoT attain every property except minimality.

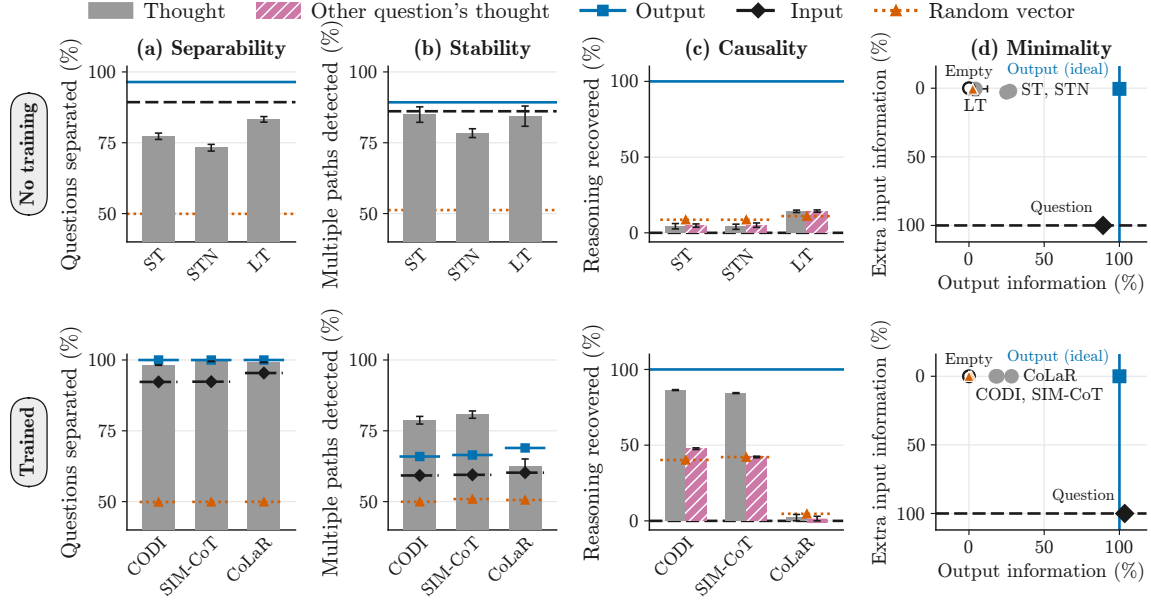


Figure 5: **Performance.** Top, no training at best step, averaged over models. Bottom, trained methods.

Causality. As shown in Figure 5(c), training-free methods fail to attain causality. They are generally unable to recover the written reasoning, and the part they do recover is not specific to the question (performing similarly to the control representation of a different question). In contrast, the trained thought representations of CODI and SIM-CoT successfully attain causality as they recover the majority of the reasoning and exceed the performance of both baseline controls.

Minimality. Regardless of whether a method is trained or training-free, none successfully attain minimality (Figure 5(d)). Averaged over models, $CE(Y | \mathbf{T})$ of every thought representation remains closer to $CE(Y | \emptyset)$ than to $CE(Y | Y)$. Furthermore, all methods exhibit near-zero values for r_X . Because r_Y is also consistently low, this near-zero r_X does not imply that the thought representation is successfully discarding irrelevant input; rather, it indicates that the model recovers only a limited portion of either kind of information.

Separability and stability. While training-free thought representations separate questions and identify reasoning paths more effectively than a random vector, their capabilities are bounded by the input embedding (Figure 5(a,b)). The hidden state of the last input token shows the same bound (Appendix D). Consequently, these representations merely encode the question itself, lacking the reasoning that distinguishes specific outputs or the uncertainty surrounding them. Trained representations, however, achieve separability that surpasses the input embedding baseline. Furthermore, the thought representations of CODI and SIM-CoT detect multiple reasoning paths more accurately than even the output embedding, demonstrating that training successfully embeds both the generated reasoning and its associated uncertainty into the representation.

Table 1: Properties attained by each method.

Property	No training			Trained		
	ST	STN	LT	CODI	SIM-CoT	CoLaR
Causality	×	×	×	✓	✓	×
Minimality	×	×	×	×	×	×
Separability	×	×	×	✓	✓	✓
Stability	×	×	×	✓	✓	✓

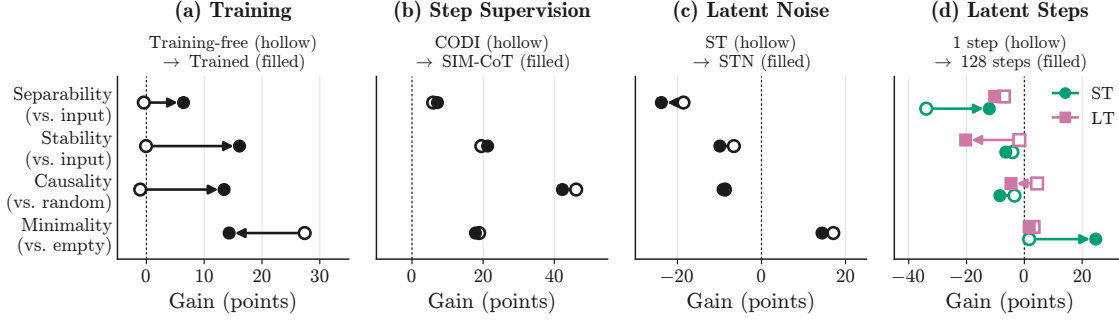


Figure 6: **Design choices.** Each arrow runs from the gain of one variant to that of the other.

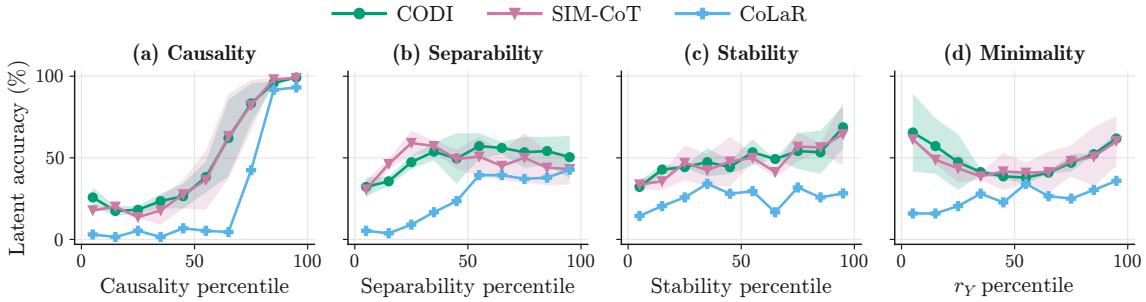


Figure 7: **Measures and accuracy.** Accuracy of each trained method on GSM8K test questions ranked by a property of their thought representation, with minimality ranked by its output information r_Y . CODI and SIM-CoT are averaged over Llama-3.2-1B-Instruct and GPT-2, and the band spans the two models.

4.3 IMPLICATIONS FOR METHOD DESIGN

Design choices. Figure 6 isolates the impact of individual design choices by comparing pairs of variants that differ by a single factor. We evaluate these choices based on their gains over our established baselines.

The analysis reveals distinct trade-offs for each modification. Training improves causality, separability, and stability, though it comes at the cost of reduced minimality (Figure 6a). Supervising each latent step, as SIM-CoT does over CODI, lowers the causality gain and leaves separability, stability, and minimality almost unchanged (Figure 6b). Gumbel noise yields a small effect (Figure 6c). The impact of adding more latent steps depends on the method (Figure 6d). For Soft Thinking, additional steps improve separability and minimality at the expense of stability and causality, whereas for Latent Thinking, more steps degrade separability, stability, and causality. Every design choice thus negatively impacts at least one property, yet no single property is harmed by all choices. The cost of any design choice therefore becomes visible only when all four measures are evaluated together.

The measures and accuracy. To relate our proposed measures to the downstream accuracy of latent reasoning methods, Figure 7 ranks the evaluated questions for each trained method according to the specific properties of their thought representations. This analysis demonstrates that incorrect answers consistently correlate with low scores across every measure. Specifically, downstream accuracy rises in tandem with causality (Figure 7a), and questions exhibiting the lowest separability and stability correspondingly yield the lowest accuracy (Figure 7b, c). Accuracy similarly increases alongside r_Y whenever $CE(Y | T)$ is below $CE(Y | \emptyset)$ (Figure 7d). However, because r_Y remains low for most representations, this r_Y ranking primarily spans the lower end of the scale. Beyond simply predicting correctness, these measures provide a deeper diagnostic tool that distinguishes between methods of similar accuracy. For example, CODI and SIM-CoT reach similar accuracy on Llama-3.2-1B-Instruct, yet they differ in causality and minimality (Table 6).

5 RELATED WORK

Constructing and supervising thought representations. Latent reasoning methods construct thought representations in a single model, through recurrence in depth, or across models. In a single model, a thought representation is the last hidden state (Hao et al., 2025), a mixture of token embeddings (Zhang et al., 2026; Butt et al., 2026), or a compressed token (Cheng & Van Durme, 2024; Tan et al., 2025). Recurrent-depth models iterate a block of layers on a latent state before decoding (Jolicoeur-Martineau, 2025; Wang & Reid, 2026; Koishekenov et al., 2025; Zeng et al., 2026). Across models, a thought representation may be a key-value cache passed from one model to another (Zou et al., 2026a; Fu et al., 2026). Prior work trains these thought representations against the hidden states of a teacher that conditions on the written reasoning (Shen et al., 2025), against the text of each reasoning step (Wei et al., 2026), toward a compressed latent chain (Tan et al., 2025; Deng et al., 2025), or through a reward on the answer (Butt et al., 2026). Every mechanism and training signal above is specific to its method. In contrast, Definition 3 states validity on the thought representation, independently of its construction and training.

Evaluating thought representations. Audits of latent reasoning intervene on latent steps, decode latent tokens, or trace reasoning paths (Dilgren & Wiegrefe, 2026; Cui et al., 2026; Rizvi-Martel & Mosbach, 2026; Sahoo et al., 2026; Zhang et al., 2025; Liang & Pan, 2026). Probing studies recover task correctness, uncertainty, or future tokens from hidden states (Afzal et al., 2025; Cencerrado et al., 2026; Lugoloobi et al., 2026; Du et al., 2026). A benchmark of latent-space reasoning measures these abilities through task performance (Hagendorff & Fabi, 2025). Each audit and each probe examines one method on its own data. In contrast, the measures of Section 3 are computed on the thought representation of any method and reported on one scale against shared references. Figure 8 contrasts the objects of these evaluations.

Theory of latent reasoning. Latent reasoning lacks a systematic theory (Yu et al., 2026). Theoretical analyses characterize what latent thought can compute, including the expressiveness of latent thought relative to chain of thought (Xu & Sato, 2026), the parallel search that continuous thoughts support through superposition (Zhu et al., 2025), and the benefits and limits of latent chain of thought (Zou et al., 2026b). These analyses state what latent thought can compute, whereas Section 3 unifies the requirements that a thought representation must attain.

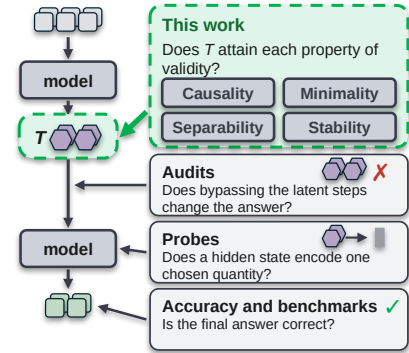


Figure 8: Positioning of this work.

6 CONCLUSION

Latent reasoning methods claim that a thought representation substitutes for written reasoning, compresses it, separates distinct outputs, and achieves superposition, whereas their evaluations report accuracy and token counts. We state these claims as the validity of a thought representation space, prove that validity is equivalent to four properties, namely causality, minimality, separability, and stability, and translate each property into a measure computed on these representations for any method. Training-free thought representations encode the question but not the reasoning they replace and consequently support none of the four claims. Training improves substitution, fidelity, and superposition, yet no audited method compresses the reasoning beyond simply shortening the trajectory. Incorrect answers correlate with a low score on each measure, which points to the property that a thought representation lacks, and methods of similar accuracy can be compared across the four measures. Therefore, accuracy and token counts do not specify which claims a method satisfies, and the four measures answer that question.

Implications. The results imply that (i) an evaluation of a latent reasoning method should report the four measures alongside its accuracy, (ii) the evaluation should include the thought representation of another question and a random vector as controls, (iii) a method that claims substitution should train its latent steps on the written reasoning, and (iv) its step supervision, noise, and number of latent steps should be selected against all four measures.

Limitations. Each measure reports a surrogate for its property, which the audit compares with references of known information content. Causality and minimality are computed in the model, hence they apply to thought representations in the input embedding space, whereas separability and stability apply to every thought representation. The audit approximates the output distribution of a model by K generated outputs, hence it

measures stability on those outputs. Since every audited method gives one thought representation per input, the audit measures stability through mode collapse resistance, and lexical invariance has no measure yet. The audited tasks are English reasoning tasks, hence the findings may differ for multilingual workloads and for generations outside reasoning. The trained audit covers the released checkpoints of each method, and extending it to further models is future work.

REFERENCES

- Anum Afzal, Florian Matthes, Gal Chechik, and Yftah Ziser. Knowing before saying: LLM representations encode information about chain-of-thought success before completion. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 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/>.
- Jesujoba Alabi, Marius Mosbach, Matan Eyal, Dietrich Klakow, and Mor Geva. The hidden space of transformer language adapters. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6588–6607, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.356. URL <https://aclanthology.org/2024.acl-long.356/>.
- Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit Tracing: Revealing Computational Graphs in Language Models. *Transformer Circuits Thread*, 2025. URL <https://transformer-circuits.pub/2025/attribution-graphs/methods.html>.
- Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks, 2025. URL <https://arxiv.org/abs/2511.07025>.
- Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing 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>.
- David Barber and Felix Agakov. The IM algorithm: a variational approach to information maximization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS’03, pp. 201–208, Cambridge, MA, USA, 2003. MIT Press.
- Natasha Butt, Ariel Kwiatkowski, Ismail Labiad, Julia Kempe, and Yann Ollivier. Soft Tokens, Hard Truths. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=9JjKTp8Jmy>.
- Ziyang Cai, Xingyu Zhu, Yihe Dong, Yinghui He, and Sanjeev Arora. T2MLR: Transformer with Temporal Middle-Layer Recurrence. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://openreview.net/forum?id=fQbk1EQWBO>.
- Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. No answer needed: Predicting LLM answer accuracy from question-only linear probes, 2026. URL <https://arxiv.org/abs/2509.10625>.
- Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning, 2025. URL <https://arxiv.org/abs/2505.16782>.
- Jeffrey Cheng and Benjamin Van Durme. Compressed Chain of Thought: Efficient Reasoning Through Dense Representations. *arXiv preprint arXiv:2412.13171*, 2024.
- Karl Cobbe, Vin Kosaraju, Mojtava Bavarian, John Schulman, Joshua Moss, Miles Brundage, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Yingqian Cui, Zhenwei Dai, Bing He, Zhan Shi, Hui Liu, Rui Sun, Zhiji Liu, Yue Xing, Jiliang Tang, and Benoit 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>.
- 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>.

- Jingcheng Deng, Liang Pang, Zihao Wei, Shicheng Xu, Zenghao Duan, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. LLM Latent Reasoning as Chain of Superposition. *arXiv preprint arXiv:2510.15522*, 2025.
- Connor Dilgren and Sarah Wiegrefe. Are Latent Reasoning Models Easily Interpretable? In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2604.04902>.
- Nikita Dragunov, Temurbek Rahmatullaev, Elizaveta Goncharova, Andrey Kuznetsov, and Anton Razzhigaev. SONAR-LLM: Autoregressive Transformer that Thinks in Sentence Embeddings and Speaks in Tokens, 2025. URL <https://arxiv.org/abs/2508.05305>.
- Changde Du, Kaicheng Fu, Bincheng Wen, Yi Sun, Jie Peng, Wei Wei, Ying Gao, Shengpei Wang, Chuncheng Zhang, Jinpeng Li, Shuang Qiu, Le Chang, and Huiguang 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>.
- Hanwen Du, Yuxin Dong, and Xia Ning. Latent Thinking Optimization: Your Latent Reasoning Language Model Secretly Encodes Reward Signals in Its Latent Thoughts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=2jkAk3EP0v>.
- Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. SONAR: Sentence-Level Multimodal and Language-Agnostic Representations, 2023. URL <https://arxiv.org/abs/2308.11466>.
- Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring Latent World States in Language Models with Propositional Probes. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL <https://openreview.net/forum?id=0yvZm2AjUr>.
- Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856. URL <https://openreview.net/forum?id=sySqlxj8EB>.
- Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 43130–43158, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/474ada926b331d78f06d95e8913111cc-Paper-Conference.pdf.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Thilo Hagendorff and Sarah Fabi. Beyond Chains of Thought: Benchmarking Latent-Space Reasoning Abilities in Large Language Models, 2025. URL <https://arxiv.org/abs/2504.10615>.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=ltxz7S4lp3>.
- Jujie He, Jiakai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork Open Reasoner 1 Technical Report. *arXiv preprint arXiv:2505.22312*, 2025a.
- Jujie He, Jiakai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Yang Liu, and Yahui Zhou. Skywork Open Reasoner Series. <https://capricious-hydrogen-41c.notion.site/Skywork-Open-Reasoner-Series-1d0bc9ae823a80459b46c149e4f51680>, 2025b. Notion Blog.
- Alexia Jolicoeur-Martineau. Less is More: Recursive Reasoning with Tiny Networks, 2025. URL <https://arxiv.org/abs/2510.04871>.
- Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V Le, and Orhan Firat. BIG-bench extra hard. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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/>.
- Yeskendir Koishikenov, Aldo Lipani, and Nicola Cancedda. Encode, Think, Decode: Scaling test-time reasoning with recursive latent thoughts, 2025. URL <https://arxiv.org/abs/2510.07358>.

- Lorenz Kuhn, Yarin Gal, and Sebastian 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>.
- LCM team, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R. Costa-jussà, David Dale, Hady Elsahar, Kevin Heffernan, João Maria Janeiro, Tuan Tran, Christophe Ropers, Eduardo Sánchez, Robin San Roman, Alexandre Mourachko, Safiyyah Saleem, and Holger Schwenk. Large Concept Models: Language Modeling in a Sentence Representation Space, 2024. URL <https://arxiv.org/abs/2412.08821>.
- Jia Liang and Liangming Pan. Do Latent-CoT Models Think Step-by-Step? A Mechanistic Study on Sequential Reasoning Tasks, 2026. URL <https://arxiv.org/abs/2602.00449>.
- William Lugoloobi, Thomas Foster, William Bankes, and Chris Russell. LLMs Encode Their Failures: Predicting Success from Pre-Generation Activations. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2602.09924>.
- Philipp Mondorf and Barbara 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>.
- OpenAI. gpt-oss-120b & gpt-oss-20b Model Card, 2025. URL <https://arxiv.org/abs/2508.10925>.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI, 2019.
- Michael Rizvi-Martel and Marius Mosbach. The Illusion of Superposition in Latent CoT via Soft Thinking. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://openreview.net/forum?id=FvPx9Nzvnw>.
- Subramanyam Sahoo, Aman Chadha, Vinija Jain, and Divya 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>.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, Suzhou, China, November 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/>.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia 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>.
- Qi Sun, Marc Pickett, Aakash Kumar Nain, and Llion Jones. Transformer Layers as Painters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25219–25227, 2025. doi: 10.1609/aaai.v39i24.34708. URL <https://ojs.aaai.org/index.php/AAAI/article/view/34708>.
- Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of LLM reasoning chains. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 4646–4668. Curran Associates, Inc., 2025. doi: 10.52202/085713-0164. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0706261aedab63814a2b73c32564b4c4-Paper-Conference.pdf.
- Wenlong Wang and Fergal Reid. Tiny Recursive Reasoning with Mamba-2 Attention Hybrid. In *LIT Workshop @ ICLR 2026*, 2026. URL <https://arxiv.org/abs/2602.12078>.
- Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. SIM-CoT: Supervised implicit chain-of-thought. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 56721–56742, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5d087955ee13fe9a7402eedec879b9c3-Paper-Conference.pdf.
- Junhong Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, and Hua 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=ASLuOoP78o>.
- Kevin Xu and Issei Sato. A Formal Comparison Between Chain of Thought and Latent Thought. In *International Conference on Machine Learning (ICML)*, 2026. URL <https://arxiv.org/abs/2509.25239>.

- Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Guanting Dong, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, Jie Xu, Guibin Zhang, Jiale Tao, Jiayi Zhang, Siyuan Ma, Kaituo Feng, Haojie Huang, Youxing Li, Ronghao Chen, Huacan Wang, Chenglin Wu, Zikun Su, Xiaogang Xu, Kelu Yao, Kun Wang, Chen Gao, Yue Liao, Ruqi Huang, Tao Jin, Zhucun Xue, Cheng Tan, Jiangning Zhang, Wenqi Ren, Yanwei Fu, Yong Liu, Yu Wang, Xiangyu Yue, Yu-Gang Jiang, and Shuicheng Yan. The latent space: Foundation, evolution, mechanism, ability, and outlook, 2026. URL <https://arxiv.org/abs/2604.02029>.
- Boyi Zeng, He Li, Shixiang Song, Yixuan Wang, Zitong Wang, Ziwei He, Xinbing Wang, and Zhouhan Lin. PonderLM-2: Pretraining LLM with latent thoughts in continuous space. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti (eds.), *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pp. 153459–153482. PMLR, 06–11 Jul 2026. URL <https://proceedings.mlr.press/v306/zeng26l.html>.
- Yuyi Zhang, Boyu Tang, Tianjie Ju, Sufeng Duan, and Gongshen Liu. Do Latent Tokens Think? A Causal and Adversarial Analysis of Chain-of-Continuous-Thought, 2025. URL <https://arxiv.org/abs/2512.21711>.
- Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Eric 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>.
- Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Reasoning by Superposition: A Theoretical Perspective on Chain of Continuous Thought. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL <https://arxiv.org/abs/2505.12514>.
- Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti (eds.), *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pp. 168813–168847. PMLR, 06–11 Jul 2026a. URL <https://proceedings.mlr.press/v306/zou26k.html>.
- Jiaxuan Zou, Yaozhong Xiong, and Yong Liu. The theoretical benefits and limitations of latent chain-of-thought reasoning, 2026b. URL <https://openreview.net/forum?id=q7Nhu2Fw11>.

APPENDIX CONTENTS

A	Proofs	17
A.1	Consistency	17
A.2	Independence	19
A.3	Completeness	24
A.4	Necessity	26
B	Derivation of the Information Bottleneck Residual	28
C	Implementation Details	30
D	Additional Results	31
E	Supporting Diagnostics	35
F	Dataset Details	35
G	Extended Related Work	36

A PROOFS

Let X denote the random input, with distribution $P(X = x)$ over $\mathcal{X}$, and let Y denote the output that f_θ generates for it. As in Definition 3, μ_c denotes the distribution that a context c induces (6), hence $\mu_x = P(Y | x)$ is the output distribution of an input x . Let $\mathbb{H}$ denote the Shannon entropy, and let $\mathbf{1}[\cdot]$ take the value one when its condition holds and zero otherwise. Let the thought representation space $\mathcal{T}$ lie in $\mathbb{R}^m$. The projection ϕ of Definition 6 belongs to the class $\mathcal{H}$ of projections that factor through an intermediate space $\mathbb{R}^k$ of some dimension k ,

$$\mathcal{H} = \{ \phi = \phi_2 \circ \phi_1 : \phi_1 : \mathcal{T} \rightarrow \mathbb{R}^k \text{ and } \phi_2 : \mathbb{R}^k \rightarrow \mathbb{R}^d \text{ linear, } \|\phi_1\| \leq 1, \|\phi_2\| \leq 1 \}, \quad (22)$$

where the norm of a linear map is the largest factor by which it stretches a vector,

$$\|\phi\| = \sup_{a \neq 0} \frac{\|\phi(a)\|_2}{\|a\|_2}, \quad \text{hence} \quad \|\phi(a)\|_2 \leq \|\phi\| \|a\|_2. \quad (23)$$

The bound on $\mathcal{H}$ makes separability at a margin $\delta > 0$ a property of $\mathbf{T}$. An unrestricted class of projections stretches the difference of any two distinct representations without bound, which leads to a trivial solution that enforces no property on $\mathbf{T}$.

A.1 CONSISTENCY

This section proves Proposition 3.3 by constructing one model and one thought representation that attain the four properties. The thought representation of a context is the distribution that the context induces over the semantic output space. The model conditions each continuation only on that thought representation.

Lemma A.1 (Decomposition of Mutual Information). *Let γ be a map defined on $\mathcal{X}$ and let $Z = \gamma(X)$. Then*

$$I(X; Y) = I(Z; Y) + I(X; Y | Z). \quad (24)$$

Proof of Lemma A.1. Expanding $I(X, Z; Y)$ by the chain rule in the two orders,

$$I(X, Z; Y) = I(Z; Y) + I(X; Y | Z), \quad (25)$$

$$I(X, Z; Y) = I(X; Y) + I(Z; Y | X). \quad (26)$$

Since $Z = \gamma(X)$ is a function of X , the last term vanishes,

$$I(Z; Y | X) = 0, \quad (27)$$

hence (26) reduces to $I(X, Z; Y) = I(X; Y)$. Substituting this into (25) gives (24).

Proposition 3.3 (Consistency). *The properties in $\mathcal{P}$ are jointly achievable, in that one model and one thought representation may attain all four requirements.*

Proof of Proposition 3.3. Construction. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite, let $M \geq 2$ denote the number of semantic outputs, and let the semantic output space be the standard basis of $\mathbb{R}^M$,

$$\mathcal{S} = \{s_1, \dots, s_M\} \subset \mathbb{R}^M, \quad s_i = e_i, \quad d_{\mathcal{S}}(s, s') = \|s - s'\|_2, \quad (28)$$

where e_i denotes the i -th standard basis vector, which meets Assumption 3.1. Let U denote the prefix of an output. Define the generator of Definition 2 as the conditional distribution of the semantic class, and let $\mathbf{T}_c = g(c)$ denote the representation of a context c ,

$$g(x) = P(\Phi(Y) | X = x), \quad g([x; u]) = P(\Phi(Y) | X = x, U = u). \quad (29)$$

Let every output be a single token. Let class i hold two tokens, $y^{(i)}$ and $y_{\square}^{(i)}$, the second being the first followed by a space, with $\Phi(y^{(i)}) = \Phi(y_{\square}^{(i)}) = s_i$. Let every input give positive probability to at least two classes, and let one input give two classes different probabilities. The only contexts are then the inputs, whose representations lie off the vertices $e_1, \dots, e_M$, and the completed contexts. Construct f_θ such that generation is conditioned on the representation of its context,

$$P(v | c) = P(v | \mathbf{T}_c), \quad (30)$$

$$P(v | [x; \mathbf{T}]) = P(v | \mathbf{T}), \quad (31)$$

where the model generates one token of each class with half of its probability off the vertices, and ends generation at a vertex,

$$P(v \mid \tau) = \sum_{i=1}^M \frac{\tau[i]}{2} \left(\mathbf{1}[v = y^{(i)}] + \mathbf{1}[v = y_{\perp}^{(i)}] \right), \quad P(v \mid e_i) = \mathbf{1}[v = \emptyset], \quad (32)$$

with $\tau \in \Delta(\mathcal{S}) \setminus \{e_1, \dots, e_M\}$, $\tau[i]$ the i -th entry of τ , and $\emptyset$ the empty sequence. The construction is self-consistent. At an input x , the model conditions on $\mathbf{T}_x = g(x)$, which lies off the vertices, and (32) gives class i the probability that (29) assigns to it,

$$P(\Phi(Y) = s_i \mid X = x) = \frac{g(x)[i]}{2} + \frac{g(x)[i]}{2} = g(x)[i]. \quad (33)$$

At a completed context $[x; y]$ with $\Phi(y) = s_i$, the class is fixed, hence $g([x; y]) = e_i$, and the model ends generation, hence the output is complete. Since $\mathbf{T}_c = g(c)$ is a function of c , conditioning on the pair $(c, \mathbf{T}_c)$ equals conditioning on c ,

$$P(v \mid c, \mathbf{T}_c) = P(v \mid c), \quad (34)$$

and substituting (30) into (34),

$$P(v \mid c, \mathbf{T}_c) = P(v \mid \mathbf{T}_c), \quad (35)$$

which is the Markov chain $X \rightarrow \mathbf{T} \rightarrow Y$.

Causality. By (30) with $c = [x; u]$, and by (31) with $\mathbf{T} = \mathbf{T}_{[x; u]}$,

$$P(v \mid [x; u]) = P(v \mid \mathbf{T}_{[x; u]}), \quad (30) \quad (36)$$

$$P(v \mid [x; \mathbf{T}_{[x; u]}]) = P(v \mid \mathbf{T}_{[x; u]}), \quad (31) \quad (37)$$

hence the two left sides are equal, which is (9).

Minimality. Let $\mathbf{T}$ denote $g(X)$. By Lemma A.1 with $Z = \mathbf{T}$, together with (35),

$$I(X; Y) = I(\mathbf{T}; Y) + \underbrace{I(X; Y \mid \mathbf{T})}_{=0}, \quad \text{hence} \quad I(\mathbf{T}; Y) = I(X; Y). \quad (38)$$

Since $I(\mathbf{T}'; Y) \leq I(X; Y)$ holds for every $\mathbf{T}' = g'(X)$ by the data processing inequality, the constraint in (12) is met at its largest attainable value $I_0 = I(X; Y)$. Let g' be a generator whose representation $\mathbf{T}' = g'(X)$ meets that constraint. By Lemma A.1 with $Z = \mathbf{T}'$, together with $I(\mathbf{T}'; Y) = I(X; Y)$,

$$I(X; Y \mid \mathbf{T}') = 0, \quad (39)$$

which states that X and Y are conditionally independent given $\mathbf{T}'$,

$$P(Y \mid X = x) = P(Y \mid \mathbf{T}' = g'(x)). \quad (40)$$

Summing both sides of (40) over the outputs with $\Phi(y) = s$,

$$P(\Phi(Y) = s \mid X = x) = P(\Phi(Y) = s \mid \mathbf{T}' = g'(x)), \quad s \in \mathcal{S}, \quad (41)$$

hence, by (29), the representation $\mathbf{T}_x$ is a function of $\mathbf{T}'$,

$$\mathbf{T}_x = q'(g'(x)), \quad q'(\tau) = P(\Phi(Y) \mid \mathbf{T}' = \tau). \quad (42)$$

By (42) the chain $X \rightarrow \mathbf{T}' \rightarrow \mathbf{T}$ is Markov, and the data processing inequality gives

$$I(X; \mathbf{T}) \leq I(X; \mathbf{T}'), \quad (43)$$

which is the minimum in (12).

Separability. Let $k = M$, and let ϕ_1 and ϕ_2 be the identity on $\mathbb{R}^M$, which is linear and of unit operator norm, hence $\phi \in \mathcal{H}$ and $\phi(\mathbf{T}) = \mathbf{T}$. Let the decoding rule of Definition 2 generate, at a context c , an output y_c of the class of largest probability under $\mathbf{T}_c$, with ties broken toward the smallest index,

$$\Phi(y_c) = s_{j(\mathbf{T}_c)}, \quad j(\tau) = \min \left\{ i : \tau[i] = \max_k \tau[k] \right\}. \quad (44)$$

The semantic output of a context is a function of its representation, hence two contexts whose outputs are not semantically equivalent have distinct representations,

$$\Phi(y_{c_1}) \neq \Phi(y_{c_2}) \implies \mathbf{T}_{c_1} \neq \mathbf{T}_{c_2}. \quad (45)$$

Since $\mathcal{X}$ and $\mathcal{Y}$ are finite, the contexts are finite in number. The smallest distance between two such representations is therefore positive,

$$\delta_0 = \min \left\{ \|\mathbf{T}_{c_1} - \mathbf{T}_{c_2}\|_2 : \Phi(y_{c_1}) \neq \Phi(y_{c_2}) \right\} > 0. \quad (46)$$

For every margin $\delta < \delta_0$,

$$d_{\mathcal{S}}(\phi(\mathbf{T}_{c_1}), \phi(\mathbf{T}_{c_2})) = \|\mathbf{T}_{c_1} - \mathbf{T}_{c_2}\|_2 \geq \delta_0 > \delta, \quad (47)$$

which is (17).

Stability. For an output y sampled for the input x , let $i(y)$ denote the index with $s_{i(y)} = \Phi(y)$. Conditioning on $Y = y$ fixes the value of $\Phi(Y)$, hence (29) gives

$$\mathbf{T}_y = g([x; y]) = P(\Phi(Y) \mid X = x, Y = y) = e_{i(y)}. \quad (48)$$

For outputs with $y_i \sim_{\text{sem}} y_j$, Definition 1 gives $\Phi(y_i) = \Phi(y_j)$ and therefore $i(y_i) = i(y_j)$, hence

$$\mathbf{T}_{y_i} = e_{i(y_i)} = e_{i(y_j)} = \mathbf{T}_{y_j}, \quad (49)$$

which is (19). Every value of g is a distribution over $\mathcal{S}$ by (29), hence the identity on $\mathcal{T}$ is a map from $\mathcal{T}$ to the distributions over $\mathcal{S}$. Taking q to be that map,

$$q(\mathbf{T}_x) = \mathbf{T}_x = g(x) = P(\Phi(Y) \mid X = x), \quad (50)$$

which is (20).

The four requirements hold for the model and the representation of (29), which proves the proposition.

A.2 INDEPENDENCE

Each proof exhibits one pair, constructed from the model of Proposition 3.3 by a single edit to its thought representation. Each of the four attains the remaining three properties.

Throughout this section the semantic output space, the generator and the model are those of Proposition 3.3, where $M \geq 2$,

$$\mathcal{S} = \{s_1, \dots, s_M\} \subset \mathbb{R}^M, \quad s_i = e_i, \quad d_{\mathcal{S}}(s, s') = \|s - s'\|_2, \quad (51)$$

$$g(c) = P(\Phi(Y) \mid c) \in \Delta(\mathcal{S}), \quad (52)$$

$$\Delta(\mathcal{S}) = \left\{ \tau \in \mathbb{R}^M : \tau[i] \geq 0 \text{ for every } i, \sum_{i=1}^M \tau[i] = 1 \right\}, \quad (53)$$

$$P(v \mid c) = P(v \mid \mathbf{T}_c), \quad (54)$$

$$P(v \mid [x; \mathbf{T}]) = P(v \mid \mathbf{T}). \quad (55)$$

The display (53) states (3) in the coordinates of this construction. Each construction below replaces g by a map g' , states the conditional P' of its model through (54) and (55), and leaves the rest of the construction fixed. Let $\mathbf{T} = g(X)$ and $\mathbf{T}' = g'(X)$ denote the two representations of the random input, and let $\mathbf{T}'_c = g'(c)$ denote the edited representation of a context c .

Lemma A.2 (Reparameterization). *Let $B \in \mathbb{R}^{M \times M}$ be invertible, let $g'(c) = Bg(c)$, and let the model generate as in (54), under which (X, Y) keeps its distribution. Then g' attains minimality and stability whenever g attains them.*

Proof of Lemma A.2. Since an invertible map leaves mutual information unchanged,

$$I(X; \mathbf{T}') = I(X; B\mathbf{T}) = I(X; \mathbf{T}), \quad (56)$$

$$I(\mathbf{T}'; Y) = I(B\mathbf{T}; Y) = I(\mathbf{T}; Y), \quad (57)$$

hence the objective and the constraint of (12) take the same values at $\mathbf{T}'$ as at $\mathbf{T}$, and $\mathbf{T}'$ attains the minimum whenever $\mathbf{T}$ attains it. For two outputs with $y_a \sim_{\text{sem}} y_b$,

$$\mathbf{T}_{y_a} = \mathbf{T}_{y_b} \implies B\mathbf{T}_{y_a} = B\mathbf{T}_{y_b} \implies \mathbf{T}'_{y_a} = \mathbf{T}'_{y_b}, \quad (58)$$

which is (19). Let q denote the map of (20) for g , and let q' denote $q \circ B^{-1}$, which maps the thought representation space to the distributions over $\mathcal{S}$. Then

$$q'(\mathbf{T}'_x) = q(B^{-1}B\mathbf{T}_x) = q(\mathbf{T}_x) = P(\Phi(Y) \mid X = x), \quad (59)$$

which is (20).

Lemma A.3 (Augmentation). *Let ω map contexts to $\mathbb{R}^r$, let*

$$g'(c) = [g(c); \omega(c)] \in \mathbb{R}^{M+r}, \quad (60)$$

and let the model of the pair condition on the first block, for every value τ of g and every $\zeta \in \mathbb{R}^r$,

$$P'(v | c) = P(v | \mathbf{T}_c), \quad (61)$$

$$P'(v | [x; [\tau; \zeta]]) = P(v | \tau). \quad (62)$$

Then g' attains causality and separability whenever g attains them.

Proof of Lemma A.3. At a context $[x; u]$, let τ denote $g([x; u])$ and let ζ denote $\omega([x; u])$. Then

$$P'(v | [x; u]) = P(v | \mathbf{T}_{[x; u]}) = P(v | \tau), \quad (61)$$

$$P'(v | [x; \mathbf{T}'_{[x; u]}]) = P'(v | [x; [\tau; \zeta]]) = P(v | \tau), \quad (62)$$

hence the two left sides are equal, which is (9). Let $\rho : \mathbb{R}^{M+r} \rightarrow \mathbb{R}^M$ denote the projection $\rho([\tau; \zeta]) = \tau$, whose operator norm is at most one,

$$\|\rho([\tau; \zeta])\|_2 = \|\tau\|_2 \leq \|[\tau; \zeta]\|_2, \quad \text{hence} \quad \|\rho\| \leq 1. \quad (65)$$

Let $\phi = \phi_2 \circ \phi_1$ attain the margin of (17) for g , and let $\phi' = \phi_2 \circ (\phi_1 \rho)$. Its first factor meets the bound of (22), hence $\phi' \in \mathcal{H}$,

$$\|\phi_1 \rho\| \leq \|\phi_1\| \|\rho\| \leq 1, \quad (66)$$

$$\phi'(\mathbf{T}'_c) = \phi_2 \phi_1 \rho([g(c); \omega(c)]) = \phi_2 \phi_1 g(c) = \phi(\mathbf{T}_c), \quad (67)$$

hence ϕ' attains the margin of Proposition 3.3,

$$d_S(\phi'(\mathbf{T}'_{c_1}), \phi'(\mathbf{T}'_{c_2})) = d_S(\phi(\mathbf{T}_{c_1}), \phi(\mathbf{T}_{c_2})) > \delta, \quad (68)$$

which is (17).

Proposition A.4 (Coordinate Permutation). *There exists a pair that attains minimality, separability and stability, and that does not attain causality.*

Proof of Proposition A.4. Construction. Let x_0 be an input and $i \neq j$ two classes with $g(x_0)[i] \neq g(x_0)[j]$, let $\Pi \in \mathbb{R}^{M \times M}$ be the permutation matrix that swaps the coordinates i and j , let

$$g'(c) = \Pi g(c), \quad (69)$$

and let the model be that of (54) and (55), unchanged.

Minimality and stability. By Lemma A.2 with $B = \Pi$, which is invertible. The model is unchanged, hence (X, Y) keeps its distribution.

Separability. A permutation matrix reorders the entries of a vector, hence

$$\|\Pi a\|_2 = \|a\|_2 \quad \text{for every } a \in \mathbb{R}^M, \quad \|\Pi\| = \|\Pi^{-1}\| = 1. \quad (70)$$

Let $\phi = \phi_2 \circ \phi_1$ attain the margin of (17) for g , and let $\phi' = \phi_2 \circ (\phi_1 \Pi^{-1})$. Its first factor meets the bound of (22), hence $\phi' \in \mathcal{H}$,

$$\|\phi_1 \Pi^{-1}\| \leq \|\phi_1\| \|\Pi^{-1}\| \leq 1, \quad (71)$$

$$\phi'(\mathbf{T}'_c) = \phi_2 \phi_1 \Pi^{-1} \Pi g(c) = \phi_2 \phi_1 g(c) = \phi(\mathbf{T}_c), \quad (72)$$

hence ϕ' attains the margin of Proposition 3.3.

Causality. By (54), (55) and (69),

$$P'(v | x_0) = P(v | g(x_0)), \quad (73)$$

$$P'(v | [x_0; \mathbf{T}'_{x_0}]) = P(v | \Pi g(x_0)). \quad (74)$$

By (32) at $v = y^{(i)}$,

$$P'(y^{(i)} | x_0) = \frac{g(x_0)[i]}{2} \neq \frac{g(x_0)[j]}{2} = P'(y^{(i)} | [x_0; \mathbf{T}'_{x_0}]), \quad (75)$$

hence (9) fails at the context x_0 .

Proposition A.5 (Input Hash). *There exists a pair that attains causality, separability and stability, and that does not attain minimality.*

Proof of Proposition A.5. Construction. Let $\mathcal{X}$ be finite, and let $\kappa : \mathcal{X} \rightarrow \{1, \dots, r\}$ be a bijection with $r = |\mathcal{X}|$. Let the block hold an injective encoding of the input while the generated segment is empty,

$$\xi(x) = e_{\kappa(x)} \in \mathbb{R}^r, \quad \omega(x) = \xi(x), \quad \omega([x; u]) = 0 \text{ for } u \neq \emptyset, \quad (76)$$

and let g' and the model be those of Lemma A.3.

Causality and separability. By Lemma A.3.

Stability. Let y_a and y_b be generated outputs with $y_a \sim_{\text{sem}} y_b$, and let $i(y)$ denote the index of (48). The generated segment of their contexts is not empty, hence

$$\begin{aligned} \mathbf{T}'_{y_a} &= g'([x_a; y_a]) = [g([x_a; y_a]); 0] & (76) \\ &= [e_{i(y_a)}; 0] & (48) \\ &= [e_{i(y_b)}; 0] & \Phi(y_a) = \Phi(y_b) \\ &= [g([x_b; y_b]); 0] = \mathbf{T}'_{y_b}, & (48) \end{aligned} \quad (77)$$

which is (19). Let q' denote the projection onto the first block, $q'([\tau; \zeta]) = \tau$, which maps the thought representation space to the distributions over $\mathcal{S}$ by (52). Then

$$q'(\mathbf{T}'_x) = q'([g(x); \xi(x)]) = g(x) = P(\Phi(Y) | X = x), \quad (78)$$

which is (20).

Minimality. Let $\mathbf{T}$ denote $g(X)$, and let $\mathbf{T}'$ denote $g'(X) = [\mathbf{T}; \xi(X)]$. By the chain rule for mutual information,

$$I(X; \mathbf{T}') = I(X; [\mathbf{T}; \xi(X)]) = I(X; \mathbf{T}) + I(X; \xi(X) | \mathbf{T}). \quad (79)$$

Expanding the second term, and using that $\xi(X)$ is a function of X ,

$$I(X; \xi(X) | \mathbf{T}) = \mathbb{H}(\xi(X) | \mathbf{T}) - \mathbb{H}(\xi(X) | X, \mathbf{T}), \quad (80)$$

$$\mathbb{H}(\xi(X) | X, \mathbf{T}) = 0. \quad (81)$$

Since ξ is injective, it has a left inverse ξ^{-1} on $\xi(\mathcal{X})$,

$$\xi^{-1}(\xi(x)) = x \quad \text{for every } x \in \mathcal{X}, \quad (82)$$

hence X is a function of $\xi(X)$. The entropy of a function of a random variable is at most the entropy of that variable, which bounds the conditional entropy in both directions,

$$\mathbb{H}(\xi(X) | \mathbf{T}) \leq \mathbb{H}(X | \mathbf{T}), \quad \xi(X) \text{ is a function of } X \quad (83)$$

$$\mathbb{H}(X | \mathbf{T}) \leq \mathbb{H}(\xi(X) | \mathbf{T}). \quad X = \xi^{-1}(\xi(X)) \text{ by (82)} \quad (84)$$

Substituting the resulting equality and (81) into (80), and then into (79),

$$I(X; \xi(X) | \mathbf{T}) = \mathbb{H}(X | \mathbf{T}), \quad I(X; \mathbf{T}') = I(X; \mathbf{T}) + \mathbb{H}(X | \mathbf{T}). \quad (85)$$

Let the construction hold two inputs $x \neq x'$ of positive probability that induce one distribution over $\mathcal{S}$, as the two inputs of Section 2.2 do, hence one representation by (29),

$$g(x) = g(x'), \quad x \neq x', \quad P(X = x) > 0, \quad P(X = x') > 0. \quad (86)$$

Given $\mathbf{T} = g(x)$, the input then takes at least two values of positive probability, hence the added term is positive,

$$\mathbb{H}(X | \mathbf{T}) > 0, \quad I(X; \mathbf{T}') > I(X; \mathbf{T}). \quad (87)$$

Since $\mathbf{T} = \rho(\mathbf{T}')$ is a function of $\mathbf{T}'$ under the projection ρ of (65), both representations meet the constraint of (12),

$$I(\mathbf{T}'; Y) \geq I(\mathbf{T}; Y) = I(X; Y) \geq I_0, \quad (88)$$

by the data processing inequality and by Proposition 3.3. A representation that meets the constraint therefore reaches a smaller value of the objective than $\mathbf{T}'$, and $\mathbf{T}'$ does not attain the minimum in (12).

Proposition A.6 (Collapse). *For every margin $\delta \geq 0$, there exists a pair that attains causality, minimality and stability, and that does not attain separability.*

Proof of Proposition A.6. Construction. Every input gives positive probability to at least two classes, hence two completed contexts hold outputs of different semantic content,

$$[x_1; y_1], [x_2; y_2] \in \mathcal{C}_{\text{out}}, \quad \Phi(y_1) \neq \Phi(y_2). \quad (89)$$

Let the reserved vector be the origin of $\mathbb{R}^M$. Every value of g lies in $\Delta(\mathcal{S})$, whose entries sum to one, whereas the entries of the origin sum to zero,

$$\tau^* = 0 \in \mathbb{R}^M, \quad \sum_i g(c)[i] = 1 \neq 0 = \sum_i \tau^*[i], \quad \text{hence} \quad \tau^* \notin g(\mathcal{C}). \quad (90)$$

Let the generator keep g off the completed contexts and map every completed context to τ^* , and let the model end generation at τ^* ,

$$g'(c) = g(c) \text{ for } c \notin \mathcal{C}_{\text{out}}, \quad g'(c) = \tau^* \text{ for } c \in \mathcal{C}_{\text{out}}, \quad (91)$$

$$P'(v | c) = P'(v | g'(c)), \quad (92)$$

$$P'(v | [x; \mathbf{T}]) = P(v | \mathbf{T}) \text{ for } \mathbf{T} \neq \tau^*, \quad P'(v | [x; \tau^*]) = \mathbf{1}[v = \emptyset]. \quad (93)$$

Off the completed contexts the model generates as before, and at a completed context it generates nothing,

$$P'(v | c) = P(v | g(c)) = P(v | c) \text{ for } c \notin \mathcal{C}_{\text{out}}, \quad P'(v | c) = \mathbf{1}[v = \emptyset] \text{ for } c \in \mathcal{C}_{\text{out}}, \quad (94)$$

hence the pair keeps the inputs, the outputs, and the output distributions of Proposition 3.3.

Causality. At a context $[x; u] \notin \mathcal{C}_{\text{out}}$, let $\mathbf{T}' = g'([x; u]) = g([x; u]) \neq \tau^*$. Then

$$P'(v | [x; u]) = P(v | g([x; u])), \quad (92), (93)$$

$$P'(v | [x; \mathbf{T}']) = P(v | \mathbf{T}') = P(v | g([x; u])). \quad (93)$$

At a completed context $[x; y]$, where $\mathbf{T}' = \tau^*$,

$$P'(v | [x; y]) = P'(v | \tau^*) = \mathbf{1}[v = \emptyset] = P'(v | [x; \tau^*]).$$

In both cases the two distributions are equal, which is (9) at every context.

Minimality and mode collapse resistance. An input has an empty generated segment and is not a completed context, hence

$$g'(x) = g(x), \quad P'(Y | X = x) = P(Y | X = x). \quad (95)$$

The triples $(X, g'(X), Y)$ and $(X, g(X), Y)$ therefore have one joint distribution,

$$P'(X = x, \mathbf{T}' = \tau, Y = y) = P(X = x) \mathbf{1}[g(x) = \tau] P(Y = y | X = x) = P(X = x, \mathbf{T} = \tau, Y = y).$$

Mutual information is a function of the joint distribution, hence the objective and the constraint of (12) take the values at which Proposition 3.3 attains the minimum. Let q' equal the map q of Proposition 3.3 on $g(\mathcal{C})$, and let $q'(\tau^*)$ be any element of $\Delta(\mathcal{S})$. Then

$$q'(\mathbf{T}'_x) = q(g(x)) = P(\Phi(Y) | X = x),$$

which is (20).

Stability. For outputs with $y_i \sim_{\text{sem}} y_j$ generated for the input x , both completed contexts lie in $\mathcal{C}_{\text{out}}$,

$$\mathbf{T}'_{y_i} = g'([x; y_i]) = \tau^* = g'([x; y_j]) = \mathbf{T}'_{y_j},$$

which is (19) with equality. Together with mode collapse resistance, the pair attains stability.

Separability. The two completed contexts of (89) share τ^* , and every $\phi \in \mathcal{H}$ maps them to one point,

$$d_{\mathcal{S}}(\phi(\mathbf{T}'_{[x_1; y_1]}), \phi(\mathbf{T}'_{[x_2; y_2]})) = d_{\mathcal{S}}(\phi(\tau^*), \phi(\tau^*)) = 0 \leq \delta,$$

where the second equality holds because a metric assigns distance zero to a point and itself. No $\phi \in \mathcal{H}$ therefore meets (17) at any margin $\delta \geq 0$, and the pair does not attain separability.

Proposition A.7 (Lexical Flag). *There exists a pair that attains causality, minimality and separability, and that does not attain stability.*

Proof of Proposition A.7. Construction. Let $\lambda > 0$, let $r = 1$, and let

$$\omega([x; u]) = \lambda \cdot \mathbf{1}[u \text{ ends with a space}]. \quad (96)$$

The map g' and the model are those of Lemma A.3.

Causality and separability. By Lemma A.3.

Minimality. The context x has no generated segment, hence u is empty and

$$\omega(x) = 0, \quad \mathbf{T}'_x = g'(x) = [g(x); 0] = [\mathbf{T}_x; 0]. \quad (97)$$

The map $\tau \mapsto [\tau; 0]$ is injective, and an injective map leaves mutual information unchanged,

$$I(X; \mathbf{T}') = I(X; [\mathbf{T}; 0]) = I(X; \mathbf{T}), \quad (98)$$

$$I(\mathbf{T}'; Y) = I([\mathbf{T}; 0]; Y) = I(\mathbf{T}; Y) = I(X; Y) \geq I_0, \quad (99)$$

where the last two equalities are those of Proposition 3.3. The objective and the constraint of (12) take the values at which Proposition 3.3 attains the minimum.

Stability. Let y_a and y_b be generated outputs for one input x that differ in a trailing space,

$$\Phi(y_a) = \Phi(y_b), \quad \text{hence} \quad y_a \sim_{\text{sem}} y_b. \quad (100)$$

Their representations differ in the appended coordinate,

$$\mathbf{T}'_{y_a} = g'([x; y_a]) = [g([x; y_a]); \lambda], \quad (101)$$

$$\mathbf{T}'_{y_b} = g'([x; y_b]) = [g([x; y_b]); 0], \quad (102)$$

and their first blocks are one vector, by (48) and (100),

$$g([x; y_a]) = e_{i(y_a)} = e_{i(y_b)} = g([x; y_b]), \quad (103)$$

hence the appended coordinate holds the whole difference,

$$\mathbf{T}'_{y_a} - \mathbf{T}'_{y_b} = [0; \lambda], \quad \|\mathbf{T}'_{y_a} - \mathbf{T}'_{y_b}\|_2 = \lambda. \quad (104)$$

The scalar λ is arbitrary, hence the two representations are at an arbitrarily large distance and (19) fails. Definition 7 states stability as the conjunction of (19) with (20), hence the pair does not attain stability.

Corollary 3.4 (Independence). *No property in $\mathcal{P}$ is implied by the conjunction of the other three.*

Proof of Corollary 3.4. Let $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4\}$ hold the four properties in the order of Definitions 4 to 7, and let C_i denote the pair constructed in the proposition that does not attain $\mathcal{P}_i$. Proposition A.4 constructs C_1 , Proposition A.5 constructs C_2 , Proposition A.6 constructs C_3 , and Proposition A.7 constructs C_4 . Each pair attains three of the four properties,

$$C_i \text{ attains every property in } \mathcal{P} \setminus \{\mathcal{P}_i\}, \quad C_i \text{ does not attain } \mathcal{P}_i. \quad (105)$$

If the conjunction of $\mathcal{P} \setminus \{\mathcal{P}_i\}$ implied $\mathcal{P}_i$, then C_i would attain $\mathcal{P}_i$, which contradicts (105). Therefore

$$\bigwedge (\mathcal{P} \setminus \{\mathcal{P}_i\}) \not\Rightarrow \mathcal{P}_i \quad \text{for every } i \in \{1, \dots, 4\}, \quad (106)$$

and the four properties are independent.

A.3 COMPLETENESS

Causality gives the substitution, separability and lexical invariance give the bijection on completed outputs, and minimality gives the bijection on inputs.

Theorem 3.5 (Completeness). *Let $\mathcal{X}$ be finite, and let a pair (f_θ, g) attain the four properties with the threshold of (12) at its largest attainable value, $I_0 = I(X; Y)$, and with lexical invariance (19) holding with equality. Then its thought representation space is valid.*

Proof of Theorem 3.5. Substitution. At a context $[x; u] \in \mathcal{C}$, the thought representation derived from u is the value of the generator,

$$\mathbf{T} = \mathbf{T}_{[x; u]} = g([x; u]). \quad (107)$$

Since Definition 4 holds for every partition $y = [u; v]$ of a generated output, it applies at every context in $\mathcal{C}$,

$$P(v \mid [x; u]) = P(v \mid [x; \mathbf{T}]) = P(v \mid [x; \mathbf{T}_{[x; u]}]), \quad (108)$$

which is (7).

Completed outputs. Conditioning on a complete output fixes its semantic output, hence the distribution of a completed context assigns probability one to one element of $\mathcal{S}$,

$$\mu_{[x; y]}(s) = P(\Phi(Y) = s \mid [x; y]) = \mathbf{1}[s = \Phi(y)], \quad s \in \mathcal{S}. \quad (109)$$

Two completed contexts therefore induce one distribution if and only if their outputs are semantically equivalent,

$$\mu_{[x_1; y_1]} = \mu_{[x_2; y_2]} \iff \Phi(y_1) = \Phi(y_2). \quad (110)$$

For the forward arrow of (8), let $\Phi(y_1) \neq \Phi(y_2)$. Separability at the two contexts gives some $\phi \in \mathcal{H}$ with

$$\begin{aligned} d_{\mathcal{S}}(\phi(\mathbf{T}_{[x_1; y_1]}), \phi(\mathbf{T}_{[x_2; y_2]})) &> \delta \geq 0, & (17) \\ \phi(\mathbf{T}_{[x_1; y_1]}) &\neq \phi(\mathbf{T}_{[x_2; y_2]}), & d_{\mathcal{S}} \text{ is a metric} \\ \mathbf{T}_{[x_1; y_1]} &\neq \mathbf{T}_{[x_2; y_2]}, & \phi \text{ is a map} \end{aligned} \quad (111)$$

which is the contrapositive of the forward arrow. For the backward arrow,

$$\begin{aligned} \mu_{[x_1; y_1]} = \mu_{[x_2; y_2]} &\implies \Phi(y_1) = \Phi(y_2) & (110) \\ &\implies y_1 \sim_{\text{sem}} y_2 & (1) \\ &\implies \mathbf{T}_{y_1} = \mathbf{T}_{y_2} & (19) \\ &\implies g([x_1; y_1]) = g([x_2; y_2]), & \mathbf{T}_y = g([x; y]) \end{aligned} \quad (112)$$

where the third line applies lexical invariance with equality.

Inputs, forward arrow. Let $\mathbf{T}$ denote $g(X)$. The constraint of minimality at $I_0 = I(X; Y)$ gives a lower bound on $I(\mathbf{T}; Y)$. Since $\mathbf{T}$ is a function of X , the data processing inequality gives an upper bound,

$$\begin{aligned} I(\mathbf{T}; Y) &\geq I(X; Y), & (12) \\ I(\mathbf{T}; Y) &\leq I(X; Y), & \text{data processing inequality} \\ I(\mathbf{T}; Y) &= I(X; Y). \end{aligned} \quad (113)$$

Lemma A.1 with $Z = \mathbf{T}$ then gives

$$I(X; Y) = I(\mathbf{T}; Y) + I(X; Y \mid \mathbf{T}), \quad I(X; Y \mid \mathbf{T}) = 0. \quad (114)$$

Conditional mutual information is zero if and only if the two variables are independent given the third. Since $X = x$ fixes $\mathbf{T} = g(x)$, every input x of positive probability satisfies

$$P(Y \mid X = x) = P(Y \mid X = x, \mathbf{T} = g(x)) = P(Y \mid \mathbf{T} = g(x)). \quad (115)$$

Two inputs with one thought representation τ therefore induce one distribution,

$$g(x_1) = g(x_2) = \tau \implies \mu_{x_1} = P(Y \mid \mathbf{T} = \tau) = \mu_{x_2}. \quad (116)$$

Inputs, backward arrow. Let W denote μ_X , the distribution over outputs that the random input induces, which is a function of X with value μ_x at $X = x$. Conditioning on $W = w$ averages $P(Y | X = x)$ over the inputs with $\mu_x = w$,

$$\begin{aligned}
 P(Y | W = w) &= \sum_{x: \mu_x=w} P(X = x | W = w) \mu_x && \text{total probability} \\
 &= \sum_{x: \mu_x=w} P(X = x | W = w) w && \mu_x = w \text{ in each term} \\
 &= w \sum_{x: \mu_x=w} P(X = x | W = w) && \text{distributivity} \\
 &= w, && \text{probabilities sum to one} \quad (117)
 \end{aligned}$$

hence the distribution of Y given the input equals its distribution given W , which is the independence of X and Y given W ,

$$P(Y | X = x) = \mu_x = P(Y | W = \mu_x), \quad I(X; Y | W) = 0. \quad (118)$$

Lemma A.1 with $Z = W$ gives

$$I(W; Y) = I(X; Y) - I(X; Y | W) = I(X; Y) \geq I_0, \quad (119)$$

hence W meets the constraint of (12). Since $\mathbf{T}$ attains the minimum of (12) over the representations that meet the constraint, W gives an upper bound on its objective,

$$I(X; \mathbf{T}) \leq I(X; W). \quad (120)$$

For a function Z of X , the conditional entropy $\mathbb{H}(Z | X)$ is zero, hence

$$I(X; Z) = \mathbb{H}(Z) - \mathbb{H}(Z | X) = \mathbb{H}(Z). \quad (121)$$

Substituting $Z = \mathbf{T}$ and $Z = W$ into (120) gives

$$\mathbb{H}(\mathbf{T}) \leq \mathbb{H}(W). \quad (122)$$

By (115), W is a function of $\mathbf{T}$, with $q_{\text{in}}(\tau) = P(Y | \mathbf{T} = \tau)$,

$$W = \mu_X = P(Y | \mathbf{T} = g(X)) = q_{\text{in}}(\mathbf{T}). \quad (123)$$

Expanding the joint entropy of $\mathbf{T}$ and W by the chain rule in both orders,

$$\begin{aligned}
 \mathbb{H}(\mathbf{T}, W) &= \mathbb{H}(\mathbf{T}) + \mathbb{H}(W | \mathbf{T}) = \mathbb{H}(\mathbf{T}), && (123) \\
 \mathbb{H}(\mathbf{T}, W) &= \mathbb{H}(W) + \mathbb{H}(\mathbf{T} | W), && \text{chain rule} \\
 \mathbb{H}(\mathbf{T}) &= \mathbb{H}(W) + \mathbb{H}(\mathbf{T} | W). && (124)
 \end{aligned}$$

A conditional entropy is nonnegative, hence (124) bounds $\mathbb{H}(\mathbf{T})$ from below, and (122) bounds it from above,

$$\begin{aligned}
 \mathbb{H}(W) &\leq \mathbb{H}(\mathbf{T}) \leq \mathbb{H}(W), && (124), (122) \\
 \mathbb{H}(\mathbf{T} | W) &= \mathbb{H}(\mathbf{T}) - \mathbb{H}(W) = 0. && (124) \quad (125)
 \end{aligned}$$

For a value w of positive probability, let $p_\tau = P(\mathbf{T} = \tau | W = w)$. The conditional entropy averages the entropies given each value of W , and each of these is a sum of nonnegative terms,

$$\begin{aligned}
 0 &= \sum_w P(W = w) \mathbb{H}(\mathbf{T} | W = w), && (125) \\
 \mathbb{H}(\mathbf{T} | W = w) &= - \sum_\tau p_\tau \log p_\tau, \quad -p_\tau \log p_\tau \geq 0, && \text{definition of } \mathbb{H}
 \end{aligned}$$

hence every term vanishes, and $-p_\tau \log p_\tau = 0$ holds only at $p_\tau \in \{0, 1\}$. Given $W = w$, the thought representation therefore takes a single value. Two inputs of positive probability with one distribution both lie in that event,

$$\begin{aligned}
 \mu_{x_1} &= \mu_{x_2} = w, \\
 X = x_1 &\implies W = w \text{ and } \mathbf{T} = g(x_1), \\
 X = x_2 &\implies W = w \text{ and } \mathbf{T} = g(x_2),
 \end{aligned}$$

hence

$$g(x_1) = g(x_2), \quad (126)$$

which is the backward arrow of (8) on $\mathcal{C}_{\text{in}}$.

Both arrows of (8) hold on $\mathcal{C}_{\text{out}}$ by (111) and (112), and on $\mathcal{C}_{\text{in}}$ by (116) and (126). These arrows and (108) make the thought representation space valid.

Corollary A.8 (Bijection with the Semantic Output Space). *Under the conditions of Theorem 3.5, let $\Phi(\mathcal{Y}) = \{ \Phi(y) : y \in \mathcal{Y} \}$ denote the semantic outputs of the generated outputs. The map Ψ with $\Psi(\Phi(y)) = \mathbf{T}_{[x;y]}$ is a bijection from $\Phi(\mathcal{Y})$ onto $\mathcal{T}_{\text{out}}$.*

Proof of Corollary A.8. Two point masses are equal if and only if their supports coincide, hence the map $s \mapsto \mathbf{1}[\cdot = s]$ is injective,

$$\mathbf{1}[\cdot = s] = \mathbf{1}[\cdot = s'] \iff s = s'. \quad (127)$$

By (109), it maps $\Phi(\mathcal{Y})$ onto the distributions of the completed contexts,

$$\{ \mu_{[x;y]} : [x;y] \in \mathcal{C}_{\text{out}} \} = \{ \mathbf{1}[\cdot = s] : s \in \Phi(\mathcal{Y}) \}. \quad (128)$$

Theorem 3.5 makes q_{out} a bijection from $\mathcal{T}_{\text{out}}$ onto the same set, and the composition of two bijections is a bijection,

$$\Psi(s) = q_{\text{out}}^{-1}(\mathbf{1}[\cdot = s]), \quad \Psi(\Phi(y)) = q_{\text{out}}^{-1}(\mu_{[x;y]}) = \mathbf{T}_{[x;y]}. \quad (129)$$

The bijection that Section 2.2 posits between the thought representation space and the semantic output space follows on the completed outputs. When lexical invariance holds up to a tolerance, the backward arrow on completed outputs and the bijection of Corollary A.8 hold up to the same tolerance.

A.4 NECESSITY

The substitution condition of a valid space is causality, and the two bijections give minimality, separability, and stability.

Proposition 3.6 (Necessity). *Let $\mathcal{X}$ be finite, and let the thought representation space of a pair (f_θ, g) be valid. Then (f_θ, g) attains causality at every context, minimality with the threshold $I_0 = I(X; Y)$, separability on $\mathcal{C}_{\text{in}}$ and on $\mathcal{C}_{\text{out}}$, and stability with lexical invariance (19) holding with equality.*

Proof of Proposition 3.6. Causality. Let $[x; u] \in \mathcal{C}$ be a context, and let $y = [u; v]$ be a partition of a generated output. The substitution condition of Definition 3 and the generator of Section 2.2 give

$$\begin{aligned} P(v \mid [x; u]) &= P(v \mid [x; \mathbf{T}_{[x;u]}]), & (7) \\ \mathbf{T} &= \mathbf{T}_{[x;u]} = g([x; u]), & \mathbf{T}_c = g(c) \\ P(v \mid [x; u]) &= P(v \mid [x; \mathbf{T}]), & \text{substitution} \end{aligned}$$

which is (9). The context and the partition are arbitrary, hence causality holds at every context.

Minimality, the constraint. Let $\mathbf{T} = g(X)$, and fix a value τ with $P(\mathbf{T} = \tau) > 0$ and an input x with $g(x) = \tau$ and $P(X = x) > 0$. Conditioning on $\mathbf{T} = \tau$ averages the output distributions of the inputs that map to τ ,

$$\begin{aligned} P(Y \mid \mathbf{T} = \tau) &= \sum_{x': g(x') = \tau} P(X = x' \mid \mathbf{T} = \tau) \mu_{x'} & \text{total probability} \\ &= \sum_{x': g(x') = \tau} P(X = x' \mid \mathbf{T} = \tau) \mu_x & (8) \text{ on } \mathcal{C}_{\text{in}} \\ &= \mu_x \sum_{x': g(x') = \tau} P(X = x' \mid \mathbf{T} = \tau) & \text{distributivity} \\ &= \mu_x. & \text{probabilities sum to one} \end{aligned} \quad (130)$$

Since $X = x$ fixes $\mathbf{T} = g(x)$,

$$P(Y \mid X = x, \mathbf{T} = g(x)) = P(Y \mid X = x) = \mu_x = P(Y \mid \mathbf{T} = g(x)),$$

which is the independence of X and Y given $\mathbf{T}$. A conditional mutual information is zero if and only if the two variables are independent given the third, hence Lemma A.1 with $Z = \mathbf{T}$ gives

$$I(X; Y) = I(\mathbf{T}; Y) + \underbrace{I(X; Y \mid \mathbf{T})}_{=0}, \quad \text{hence} \quad I(\mathbf{T}; Y) = I(X; Y) = I_0. \quad (131)$$

Minimality, the minimum. Let g' be a generator whose representation $\mathbf{T}' = g'(X)$ meets the constraint. The constraint and the data processing inequality give

$$\begin{aligned} I(\mathbf{T}'; Y) &\geq I(X; Y), & (12) \\ I(\mathbf{T}'; Y) &\leq I(X; Y), & \text{data processing inequality} \\ I(\mathbf{T}'; Y) &= I(X; Y). \end{aligned}$$

Lemma A.1 with $Z = \mathbf{T}'$ then gives, as in (131),

$$I(X; Y | \mathbf{T}') = 0, \quad \text{hence} \quad \mu_x = P(Y | \mathbf{T}' = g'(x)) \quad (132)$$

for every input x of positive probability. Let W denote μ_X , the distribution over outputs that the random input induces. Let q' and q_{in} map a value of $\mathbf{T}'$ and a value of $\mathbf{T}$ to the distribution of Y given that value, and let h map a distribution w to the thought representation of the inputs that induce it,

$$q'(\tau') = P(Y | \mathbf{T}' = \tau'), \quad q_{\text{in}}(\tau) = P(Y | \mathbf{T} = \tau), \quad h(w) = g(x) \text{ for any input } x \text{ with } \mu_x = w. \quad (133)$$

The map h is well defined, since the bijection (8) on $\mathcal{C}_{\text{in}}$ gives $g(x_1) = g(x_2)$ whenever $\mu_{x_1} = \mu_{x_2}$. The three maps express W as a function of $\mathbf{T}'$ and of $\mathbf{T}$, and $\mathbf{T}$ as a function of W ,

$$W = q'(\mathbf{T}'), \quad (132)$$

$$W = q_{\text{in}}(\mathbf{T}), \quad (130)$$

$$\mathbf{T} = h(W). \quad (133) \quad (134)$$

Let V be a random variable, and let $Z = \gamma(V)$ for a map γ . Since $\mathbb{H}(Z | V) = 0$ and a conditional entropy is nonnegative, the chain rule in the two orders gives

$$\mathbb{H}(V) = \mathbb{H}(V) + \mathbb{H}(Z | V) = \mathbb{H}(V, Z) = \mathbb{H}(Z) + \mathbb{H}(V | Z) \geq \mathbb{H}(Z). \quad (135)$$

Applying (135) to the three maps of (134),

$$\begin{aligned} \mathbb{H}(W) &\leq \mathbb{H}(\mathbf{T}'), & \mathbb{H}(W) &\leq \mathbb{H}(\mathbf{T}), & \mathbb{H}(\mathbf{T}) &\leq \mathbb{H}(W), \\ \mathbb{H}(\mathbf{T}) &= \mathbb{H}(W) \leq \mathbb{H}(\mathbf{T}'). \end{aligned}$$

Both $\mathbf{T}$ and $\mathbf{T}'$ are functions of X , hence (121) turns each mutual information with X into an entropy,

$$I(X; \mathbf{T}) = \mathbb{H}(\mathbf{T}) \leq \mathbb{H}(\mathbf{T}') = I(X; \mathbf{T}'), \quad (136)$$

hence $g(X)$ attains the minimum of (12) at $I_0 = I(X; Y)$.

Separability on completed outputs. Let $[x_1; y_1], [x_2; y_2] \in \mathcal{C}_{\text{out}}$ with $\Phi(y_1) \neq \Phi(y_2)$. Conditioning on a complete output fixes its semantic output, and two point masses are equal if and only if their supports coincide,

$$\begin{aligned} \mu_{[x; y]}(s) &= P(\Phi(Y) = s | [x; y]) = \mathbf{1}[s = \Phi(y)], & s \in \mathcal{S}, \\ \mu_{[x_1; y_1]} &\neq \mu_{[x_2; y_2]}, & \Phi(y_1) \neq \Phi(y_2) \\ g([x_1; y_1]) &\neq g([x_2; y_2]). & (8) \text{ on } \mathcal{C}_{\text{out}} \end{aligned} \quad (137)$$

Separability on inputs. Let $x_1, x_2 \in \mathcal{C}_{\text{in}}$, and let y_1 and y_2 denote the outputs that f_θ generates for them under the fixed decoding rule of Definition 2, with $\Phi(y_1) \neq \Phi(y_2)$. Since Φ is a map, $y_1 \neq y_2$. Let y_t denote the t -th token of an output y , let $y_{<t}$ and $y_{\leq t}$ denote its tokens before and up to position t , and let $Y_{<t}$ and $Y_{\leq t}$ denote the same prefixes of Y . Each next-token distribution is a ratio of two marginals of μ_x ,

$$\begin{aligned} P(y_t | x, y_{<t}) &= \frac{P(Y_{\leq t} = y_{\leq t} | x)}{P(Y_{<t} = y_{<t} | x)}, & \text{conditional probability} \\ P(Y_{\leq t} = y_{\leq t} | x) &= \sum_{y': y'_{\leq t} = y_{\leq t}} \mu_x(y'). & \text{marginalization} \end{aligned}$$

We assume that the fixed decoding rule chooses each token as a deterministic function of the next-token distributions of f_θ , as greedy decoding and beam search do. The decoded output is then a function of μ_x . Let D denote this function. Since the two outputs differ, so do the two distributions and the two thought representations,

$$\begin{aligned} y &= D(\mu_x), \\ \mu_{x_1} &\neq \mu_{x_2}, & y_1 \neq y_2, D \text{ is a map} \\ g(x_1) &\neq g(x_2). & (8) \text{ on } \mathcal{C}_{\text{in}} \end{aligned} \quad (138)$$

Separability, the projection. Let $\mathbf{T}_1 \neq \mathbf{T}_2$ denote the two thought representations of (137) or (138) in $\mathcal{T} \subset \mathbb{R}^m$, and let $a = \mathbf{T}_1 - \mathbf{T}_2$ and $e = a/\|a\|_2$. Let $b \in \mathbb{R}^d$ be a unit vector, and let ϕ factor through $\mathbb{R}^k$ with $k = 1$,

$$\phi_1(z) = \langle e, z \rangle, \quad \phi_2(t) = t b, \quad \phi = \phi_2 \circ \phi_1.$$

By the Cauchy–Schwarz inequality and the operator norm of (23),

$$\begin{aligned} \|\phi_1\| &= \sup_{z \neq 0} \frac{|\langle e, z \rangle|}{\|z\|_2} \leq \|e\|_2 = 1, & \text{Cauchy–Schwarz} \\ \|\phi_2\| &= \sup_{t \neq 0} \frac{\|t b\|_2}{|t|} = \|b\|_2 = 1, \end{aligned}$$

hence $\phi \in \mathcal{H}$ by (22). Applying ϕ to both thought representations,

$$\begin{aligned} \phi(\mathbf{T}_1) - \phi(\mathbf{T}_2) &= \langle e, \mathbf{T}_1 \rangle b - \langle e, \mathbf{T}_2 \rangle b && \text{definitions of } \phi_1, \phi_2 \\ &= \langle e, a \rangle b && \text{linearity} \\ &= \|a\|_2 b, && e = a/\|a\|_2 \end{aligned}$$

whose norm is $\|a\|_2 > 0$, hence $\phi(\mathbf{T}_1) \neq \phi(\mathbf{T}_2)$. A metric is zero if and only if its two arguments are equal, hence

$$d_S(\phi(\mathbf{T}_1), \phi(\mathbf{T}_2)) > 0 = \delta,$$

which is (17) at margin $\delta = 0$ on $\mathcal{C}_{\text{in}}$ and on $\mathcal{C}_{\text{out}}$.

Lexical invariance. For outputs with $y_i \sim_{\text{sem}} y_j$ generated for the input x , the two point masses coincide,

$$\mu_{[x; y_i]} = \mathbf{1}[\cdot = \Phi(y_i)] = \mathbf{1}[\cdot = \Phi(y_j)] = \mu_{[x; y_j]}, \quad (1)$$

$$\mathbf{T}_{y_i} = g([x; y_i]) = g([x; y_j]) = \mathbf{T}_{y_j}, \quad (8) \text{ on } \mathcal{C}_{\text{out}}$$

which is (19) with equality.

Mode collapse resistance. Grouping the outputs by their semantic output gives the induced distribution over $\mathcal{S}$,

$$P(\Phi(Y) = s \mid X = x) = \sum_{y: \Phi(y)=s} \mu_x(y), \quad s \in \mathcal{S}.$$

Let σ_0 be any element of $\Delta(\mathcal{S})$, which is nonempty by Assumption 3.1, and let q map each $\tau \in \mathcal{T}$ to a distribution over $\mathcal{S}$,

$$q(\tau)(s) = \begin{cases} \sum_{y: \Phi(y)=s} \mu_x(y) & \text{for any input } x \text{ with } g(x) = \tau, \\ \sigma_0(s), & \tau \notin \mathcal{T}_{\text{in}}. \end{cases} \quad \begin{matrix} \tau \in \mathcal{T}_{\text{in}}, \\ \tau \notin \mathcal{T}_{\text{in}}. \end{matrix}$$

The first case is independent of the choice of x , since

$$g(x_1) = g(x_2) \implies \mu_{x_1} = \mu_{x_2}, \quad (8) \text{ on } \mathcal{C}_{\text{in}}$$

and at every input,

$$q(\mathbf{T}_x) = P(\Phi(Y) \mid X = x),$$

which is (20). Together with lexical invariance, the pair attains stability, which concludes the proof.

B DERIVATION OF THE INFORMATION BOTTLENECK RESIDUAL

Definition 5 states minimality as the information bottleneck objective (13), whose two mutual informations depend on unknown distributions. At the symmetric weight, the objective becomes a difference of two terms, each estimated by a cross-entropy of the model. The measure is that difference.

Given a minimal thought representation, the output has low uncertainty, and the input has high uncertainty once the output is known. At the symmetric weight $\beta_{\text{IB}} = 2$, the negative objective $-J(2)$ equals $I(\mathbf{T}; Y) - I(X; \mathbf{T} \mid Y)$, where the second term is the information about the input that $\mathbf{T}$ encodes and Y does not encode. Each term is a difference of entropies. The cross-entropy of a language model gives an upper bound on each conditional entropy. The measure Δ_{IB} keeps the two cross-entropy terms that depend on $\mathbf{T}$ and removes a constant that every candidate shares.

The chain rule of mutual information, applied to the pair (X, Y) , gives two decompositions of one quantity,

$$I(\mathbf{T}; X, Y) = I(X; \mathbf{T}) + I(\mathbf{T}; Y | X) = I(\mathbf{T}; Y) + I(X; \mathbf{T} | Y). \quad (139)$$

A thought representation computed from the input is a deterministic function of X , hence

$$I(\mathbf{T}; Y | X) = 0, \quad I(X; \mathbf{T}) = I(\mathbf{T}; Y) + I(X; \mathbf{T} | Y). \quad (140)$$

Substituting (140) into the objective,

$$\begin{aligned} J(\beta_{\text{IB}}) &= I(X; \mathbf{T}) - \beta_{\text{IB}} I(\mathbf{T}; Y) \\ &= I(\mathbf{T}; Y) + I(X; \mathbf{T} | Y) - \beta_{\text{IB}} I(\mathbf{T}; Y) \\ &= (1 - \beta_{\text{IB}}) I(\mathbf{T}; Y) + I(X; \mathbf{T} | Y). \end{aligned} \quad (141)$$

At $\beta_{\text{IB}} = 2$ the two coefficients have equal magnitude and opposite sign,

$$-J(2) = I(\mathbf{T}; Y) - I(X; \mathbf{T} | Y). \quad (142)$$

Both terms of (142) are entropy differences,

$$I(\mathbf{T}; Y) = \mathbb{H}(Y) - \mathbb{H}(Y | \mathbf{T}), \quad I(X; \mathbf{T} | Y) = \mathbb{H}(X | Y) - \mathbb{H}(X | Y, \mathbf{T}), \quad (143)$$

and for random variables A and B , the cross-entropy under an approximating conditional $Q(A | B)$ exceeds the conditional entropy by a divergence,

$$\text{CE}(A | B) = \mathbb{H}(A | B) + \mathbb{E}_B [D_{\text{KL}}(P(A | B) \| Q(A | B))] \geq \mathbb{H}(A | B), \quad (144)$$

with equality when Q matches the true conditional (Barber & Agakov, 2003). The minimality measure takes Q as the next-token distribution of the model under its two prompts, which gives

$$\begin{aligned} I(\mathbf{T}; Y) &\approx \mathbb{H}(Y) - \text{CE}(Y | \mathbf{T}), \\ I(X; \mathbf{T} | Y) &\approx \text{CE}(X | Y, \emptyset) - \text{CE}(X | Y, \mathbf{T}), \end{aligned} \quad (145)$$

where the prompts of the two cross-entropies in the second line both include the output Y and differ in whether they include $\mathbf{T}$.

Substituting (145) into (142),

$$\begin{aligned} -J(2) &\approx [\mathbb{H}(Y) - \text{CE}(Y | \mathbf{T})] - [\text{CE}(X | Y, \emptyset) - \text{CE}(X | Y, \mathbf{T})] \\ &= \mathbb{H}(Y) - \text{CE}(Y | \mathbf{T}) - \text{CE}(X | Y, \emptyset) + \text{CE}(X | Y, \mathbf{T}) \\ &= [\text{CE}(X | Y, \mathbf{T}) - \text{CE}(Y | \mathbf{T})] + [\mathbb{H}(Y) - \text{CE}(X | Y, \emptyset)]. \end{aligned} \quad (146)$$

The first bracket is the residual Δ_{IB} of the minimality measure, and the second is the constant

$$C_{\text{IB}} = \mathbb{H}(Y) - \text{CE}(X | Y, \emptyset). \quad (147)$$

Neither term of C_{IB} depends on the candidate. Since C_{IB} changes with the output distribution of the model, Δ_{IB} is compared among the candidates of one model.

Relation (144) gives the excess of each cross-entropy over the corresponding conditional entropy,

$$\text{CE}(X | Y, \mathbf{T}) = \mathbb{H}(X | Y, \mathbf{T}) + \varepsilon_X(\mathbf{T}), \quad \text{CE}(Y | \mathbf{T}) = \mathbb{H}(Y | \mathbf{T}) + \varepsilon_Y(\mathbf{T}), \quad (148)$$

with $\varepsilon_X(\mathbf{T}) \geq 0$ and $\varepsilon_Y(\mathbf{T}) \geq 0$ by Gibbs' inequality. Substituting (148) into the residual,

$$\Delta_{\text{IB}} = [\mathbb{H}(X | Y, \mathbf{T}) - \mathbb{H}(Y | \mathbf{T})] + [\varepsilon_X(\mathbf{T}) - \varepsilon_Y(\mathbf{T})]. \quad (149)$$

Both cross-entropies come from one model under one pair of prompts for every candidate, hence the audit assumes that the second bracket, the difference between the two errors, is constant across the candidates, and attributes the differences in Δ_{IB} among them to the first bracket.

Without (140), the chain rule (139) instead gives

$$\begin{aligned} I(X; \mathbf{T}) &= I(\mathbf{T}; Y) + I(X; \mathbf{T} | Y) - I(\mathbf{T}; Y | X), \\ -J(2) &= I(\mathbf{T}; Y) - I(X; \mathbf{T} | Y) + I(\mathbf{T}; Y | X). \end{aligned} \quad (150)$$

The correction term $I(\mathbf{T}; Y | X)$ depends on the candidate, hence it cannot be absorbed into C_{IB} . A representation computed from the generated output satisfies $I(\mathbf{T}; Y | X) > 0$, and its cross-entropies approach

$$\text{CE}(Y | \mathbf{T}) \rightarrow 0, \quad \text{CE}(X | Y, \mathbf{T}) \rightarrow \text{CE}(X | Y, \emptyset), \quad \Delta_{\text{IB}} \rightarrow \text{CE}(X | Y, \emptyset), \quad (151)$$

which is the largest value of the measure. The output text is therefore an anchor rather than an estimate of the objective.

C IMPLEMENTATION DETAILS

Generation. Every model runs in its native precision and generates K outputs per input by beam search (Section 4.1). To save compute, causality and minimality on Llama-3.3-70B run with FP8 weights on a subset of questions that spans the range of output lengths and show similar trends as other models. The experiments run on NVIDIA H100 GPUs with 80GB of memory. The last input token candidate is the hidden state at the final position of the prompt before the first decoding step, taken from the final layer or from every layer.

Candidates. Soft Thinking replaces each decoded token by the token embeddings $\mathbf{e}_v$ weighted by the next-token distribution $p = \text{softmax}(z)$ over the logits z , $\mathbf{T} = \sum_v p_v \mathbf{e}_v$ (Zhang et al., 2026). Stochastic Soft Thinking perturbs the logits with Gumbel noise before the softmax, $\hat{p}_v \propto \exp(z_v + \epsilon_v)$ with $\epsilon_v \sim \text{Gumbel}(0, 1)$ (Wu et al., 2026). Latent Thinking supplies the last hidden state as the next input embedding, as in Coconut (Hao et al., 2025), with the implementation of Zou et al. (2026a). Since Definition 2 requires a deterministic generator, every stochastic extraction runs under one fixed random seed, which gives each input one thought representation. The output embedding and the input embedding are computed with Llama-Embed-Nemotron-8B (Babakhin et al., 2025).

Trained methods. CODI and SIM-CoT use their released checkpoints on Llama-3.2-1B-Instruct and GPT-2, in which each of six latent steps is a learned projection of the last hidden state. CoLaR samples each latent step from a learned distribution and ends the latent steps at a learned end token. Each method samples eight outputs per question under the default decoding configuration of Llama-3.2-1B-Instruct. The causality error takes the answer segment of each output as v . The random vector matches the norm of each latent step. The minimality prompts enclose $\mathbf{T}$ in the latent delimiters of each method.

Causality measure. The causality error splits each output into the reasoning prefix u and the final $n = 32$ tokens v , and excludes the outputs shorter than $n + 1$ tokens. The model computes the next-token distributions over v by teacher forcing under its chat template, once after $[x; u]$ and once after $[x; \mathbf{T}]$. The random vector has as many positions as $\mathbf{T}$ and the mean norm of the token embeddings of the model. The thought representation of another question is taken from a question of the same task. The measure omits the hidden states of the last input token, which lie outside the input embedding space.

Minimality measure. Each cross-entropy is the mean over the tokens of its target. Since a thought representation that raises $\text{CE}(X | Y, \mathbf{T})$ above $\text{CE}(X | Y, \emptyset)$ does not encode additional input information, r_X is clipped at zero. When $\text{CE}(X | Y, \emptyset) - \text{CE}(X | Y, X) < 0.1 \text{CE}(X | Y, \emptyset)$, the denominator of r_X is near zero, hence r_X is set to zero for that model.

Separability measure. Each candidate is reduced to one vector per input by averaging over its positions. Each output embedding is normalized to unit length before the output embeddings of an input are averaged. The ridge regression is fitted in closed form with standardized inputs, and its penalty is selected from a logarithmic grid by the same-task score on held-out problems. Separability holds out problems by an 80/10/10 split of each BBEH task and by five-fold cross-validation on GSM8K. The penalty bounds the norm of the linear map, as the class $\mathcal{H}$ of Appendix A bounds the norm of its projections. The same-task score compares each held-out input with every other held-out input of its task, and the cross-task score compares it with every held-out input of the other tasks. A gap of exactly zero counts as half a separated pair. The pooled output embedding equals the mean output embedding of its input, hence it separates every pair of questions whose outputs differ.

Stability measure. The semantic entropy of an input is computed from the sizes of the classes that the threshold $\tau = 0.9$ forms among its K outputs,

$$\mathbb{H}_x = - \sum_r \frac{|\mathcal{E}_r|}{K} \log \frac{|\mathcal{E}_r|}{K}. \quad (152)$$

As for separability, every candidate is reduced to the mean over its positions, which for a training-free method averages its latent steps and for the exact output embedding averages the K outputs. The difference-of-means direction and its AUROC are computed under five-fold stratified cross-validation over the inputs. The DCS is the mean AUROC over the folds, reported with the standard deviation over the folds.

Uncertainty. The uncertainty of every measure is estimated by a bootstrap that resamples problems with replacement, each together with all of its outputs. The bootstrap draws 200 resamples for separability and 1,000 for causality and minimality. All settings were fixed before any evaluation.

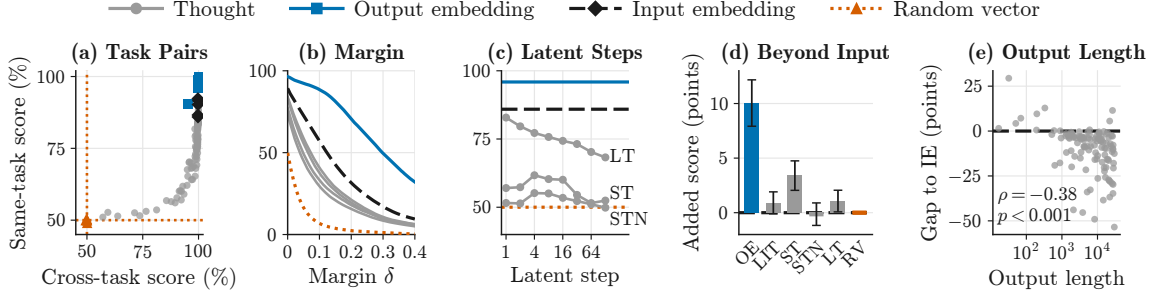


Figure 9: **Separability for training-free methods.** (a) Cross-task against same-task score. (b) Same-task pairs separated by more than δ . (c) Same-task score of a single latent step. (d) Same-task score that each method adds given the input embedding. (e) Gap between input embedding against mean output length.

D ADDITIONAL RESULTS

Tables 2 and 3 report the separability and the DCS of every candidate at every number of latent steps on every model. Tables 4 and 5 report the reasoning recovered, the output information, and the extra input information of every training-free method at every number of latent steps. These four tables also report the training-free methods on Llama-3.2-1B-Instruct over GSM8K, and Table 6 reports the four measures of every trained method. The trained audit covers the train and test splits of GSM8K, with the changes between them reported in Figure 10. A negative value in Tables 4 to 6 means that the thought representation or random vector raises the causality error or the cross-entropy above its value at $\emptyset$. Figure 5 clips such values at zero.

Separability without training. Nearly every representation other than the random vector separates questions from different tasks (Figure 9(a)). Only questions from the same task distinguish the representations, which motivates the same-task pairs of the separability score. At every margin, the thought representations separate fewer same-task pairs than the input embedding (b). The gap to the input embedding is consequently independent of the choice of δ . A single latent step of Soft Thinking separates same-task questions only slightly better than the random vector (c). The separability of Soft Thinking thus arises from the mean over its latent steps rather than from any single step. In contrast, the first latent step of Latent Thinking separates questions nearly as well as the input embedding. Each further step lowers this separability. When combined with the input embedding, the output embedding raises the same-task score substantially, whereas each thought representation adds a fraction of this gain (d). Hence the outputs hold the information that separates same-task questions and that the input embedding lacks. The thought representations encode only a limited portion of this information. The gap to the input embedding also widens on tasks with longer outputs (e). Tasks with long reasoning accordingly offer a new method the largest margin for improvement.

Table 2: Separability score (%) on same-task (Same) and cross-task (Cross) question pairs.

Model	Pairing	Output Emb.		Last Input Tok.		Soft Thinking					Stochastic Soft Thinking					Latent Thinking					References	
		Exact	Pooled	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	Input Emb.	Random Vec.
Llama-3.1 8B	Same	95.90	99.98	82.48	83.12	57.03	75.23	78.72	77.07	78.23	51.55	65.31	68.02	71.20	72.24	82.86	79.27	77.90	76.43	75.62	85.92	50.01
	Cross	± 0.60	± 0.01	± 1.14	± 1.14	± 1.49	± 1.38	± 1.35	± 1.28	± 1.22	± 1.45	± 1.54	± 1.41	± 1.38	± 1.38	± 1.18	± 1.29	± 1.33	± 1.35	± 1.30	± 1.21	± 1.39
Llama-3.3 70B	Same	99.89	100.00	99.68	99.74	92.25	98.90	99.48	99.54	99.37	73.67	98.46	98.74	99.27	99.19	99.71	99.68	99.67	99.65	99.58	99.75	50.01
	Cross	± 0.04	± 0.00	± 0.10	± 0.08	± 0.88	± 0.30	± 0.14	± 0.15	± 0.14	± 1.34	± 0.34	± 0.25	± 0.14	± 0.16	± 0.10	± 0.09	± 0.09	± 0.10	± 0.11	± 0.08	± 1.20
DS-R1-Qwen 32B	Same	90.47	99.96	87.12	86.99	53.07	72.96	75.66	79.18	81.75	52.57	68.77	72.20	77.85	79.72	86.72	84.57	83.38	82.27	81.35	86.60	50.02
	Cross	± 1.00	± 0.02	± 1.10	± 1.08	± 1.35	± 1.31	± 1.29	± 1.25	± 1.15	± 1.33	± 1.29	± 1.33	± 1.27	± 1.21	± 1.05	± 1.19	± 1.21	± 1.20	± 1.22	± 1.01	± 1.35
Skywork-OR1 32B	Same	95.17	100.00	99.72	99.66	76.50	98.66	98.62	98.46	98.36	71.35	97.77	98.45	98.75	98.87	99.62	99.63	99.60	99.59	99.60	99.73	49.43
	Cross	± 0.55	± 0.00	± 0.10	± 0.09	± 1.20	± 0.21	± 0.26	± 0.28	± 0.29	± 1.18	± 0.32	± 0.29	± 0.27	± 0.26	± 0.11	± 0.11	± 0.11	± 0.11	± 0.11	± 0.07	± 1.35
GPT-OSS 20B	Same	97.10	99.95	89.68	90.12	56.45	74.71	77.21	79.72	80.94	52.72	64.07	69.20	73.40	75.65	88.47	84.29	83.54	84.16	84.04	91.57	49.07
	Cross	± 0.56	± 0.02	± 0.87	± 0.85	± 1.37	± 1.43	± 1.42	± 1.31	± 1.25	± 1.43	± 1.48	± 1.40	± 1.39	± 1.33	± 0.93	± 1.06	± 1.17	± 1.15	± 1.14	± 0.86	± 1.44
Llama-3.2 1B (GSM8K)	Same	99.65	100.00	99.84	99.86	87.76	99.42	99.63	99.73	99.50	59.38	96.81	99.22	99.57	99.55	99.92	99.92	99.90	99.87	99.72	99.94	50.09
	Cross	± 0.09	± 0.00	± 0.06	± 0.05	± 0.98	± 0.19	± 0.17	± 0.09	± 0.16	± 1.35	± 0.59	± 0.17	± 0.11	± 0.12	± 0.04	± 0.04	± 0.05	± 0.05	± 0.18	± 0.03	± 1.25
GPT-OSS 20B	Same	98.73	99.96	89.47	89.96	56.81	74.54	77.57	81.90	82.90	51.05	64.15	71.07	74.25	78.32	87.97	83.71	83.27	83.70	83.97	92.30	50.00
	Cross	± 0.33	± 0.02	± 0.94	± 0.90	± 1.34	± 1.33	± 1.27	± 1.18	± 1.16	± 1.50	± 1.48	± 1.50	± 1.37	± 1.28	± 0.89	± 1.19	± 1.20	± 1.18	± 1.25	± 0.81	± 1.51
Llama-3.2 1B (GSM8K)	Same	99.88	100.00	99.95	99.91	85.63	99.15	99.24	99.71	99.58	57.13	97.32	98.52	99.03	99.76	99.95	99.90	99.79	99.66	99.64	99.72	50.01
	Cross	± 0.04	± 0.00	± 0.02	± 0.04	± 1.07	± 0.35	± 0.30	± 0.14	± 0.19	± 1.42	± 0.50	± 0.40	± 0.35	± 0.08	± 0.02	± 0.04	± 0.09	± 0.20	± 0.17	± 0.21	± 1.25
GPT-OSS 20B	Same	99.99	100.00	70.05	67.59	53.57	60.09	60.04	61.80	62.12	51.42	56.03	57.78	59.04	60.27	65.27	68.04	68.29	69.87	70.30	90.25	50.44
	Cross	± 0.01	± 0.00	± 1.51	± 1.56	± 1.55	± 1.41	± 1.41	± 1.49	± 1.45	± 1.58	± 1.44	± 1.52	± 1.50	± 1.53	± 1.35	± 1.55	± 1.40	± 1.47	± 1.43	± 1.15	± 1.59
Llama-3.2 1B (GSM8K)	Same	100.00	100.00	95.77	94.97	81.22	91.33	92.35	92.94	94.85	63.80	89.51	90.31	93.47	91.81	95.19	96.70	97.23	96.69	95.88	99.78	50.02
	Cross	± 0.00	± 0.00	± 0.71	± 0.83	± 1.14	± 0.96	± 0.95	± 0.95	± 0.76	± 1.46	± 0.93	± 1.00	± 0.80	± 0.96	± 0.84	± 0.57	± 0.51	± 0.57	± 0.72	± 0.10	± 1.41
Llama-3.2 1B (GSM8K)	Same	99.92	100.00	—	—	86.73	99.77	99.85	99.79	99.59	65.48	97.90	98.79	98.69	98.26	99.99	99.97	99.96	99.93	99.87	99.98	49.97
	Cross	± 0.01	± 0.00	—	—	± 0.22	± 0.03	± 0.02	± 0.02	± 0.03	± 0.31	± 0.09	± 0.07	± 0.08	± 0.09	± 0.00	± 0.01	± 0.01	± 0.01	± 0.02	± 0.01	± 0.30

Table 3: Distributional consistency score (%) of every candidate.

Model	Output Emb.		Last Input Tok.		Soft Thinking					Stochastic Soft Thinking					Latent Thinking					References	
	Exact	Pooled	All	Final	1	16	32	64	128	1	16	32	64	128	1	16	32	64	128	Input Emb.	Random Vec.
Llama-3.1 8B	95.53 ±1.30	95.96 ±1.20	94.30 ±1.08	94.35 ±1.18	93.88 ±0.96	90.53 ±1.44	88.99 ±1.25	89.18 ±1.74	90.41 ±1.58	89.74 ±1.17	88.70 ±1.33	87.43 ±1.39	89.96 ±1.59	89.17 ±1.71	91.91 ±1.37	82.15 ±1.56	80.27 ±1.69	79.47 ±1.29	78.26 ±1.19	92.99 ±0.97	51.52 ±2.57
Llama-3.3 70B	87.48 ±0.81	88.51 ±0.87	87.25 ±0.47	89.19 ±0.82	83.95 ±1.08	82.69 ±0.87	84.06 ±1.06	85.32 ±0.71	81.42 ±1.09	82.57 ±1.20	80.99 ±1.58	82.32 ±1.83	84.28 ±1.23	80.63 ±1.11	87.19 ±0.85	66.15 ±2.24	64.52 ±2.52	63.65 ±2.46	63.65 ±2.47	91.95 ±1.44	49.94 ±1.27
DS-R1-Qwen 32B	95.18 ±0.49	95.74 ±0.45	94.95 ±0.81	95.06 ±0.65	95.30 ±0.62	90.06 ±1.30	86.39 ±2.19	84.65 ±1.58	83.50 ±1.97	83.84 ±2.06	82.57 ±1.15	83.06 ±1.14	84.80 ±0.85	84.83 ±1.80	91.62 ±1.01	69.40 ±1.10	67.23 ±1.70	65.72 ±1.96	63.82 ±2.11	93.52 ±0.80	49.99 ±1.44
Skywork-OR1 32B	96.29 ±0.42	97.14 ±0.25	95.52 ±0.47	95.11 ±0.64	95.14 ±0.45	90.42 ±0.65	90.04 ±1.04	88.29 ±1.00	87.99 ±1.00	76.63 ±1.04	80.48 ±2.31	83.29 ±1.05	84.55 ±1.03	86.47 ±1.47	92.10 ±0.88	68.38 ±1.38	68.23 ±1.69	68.65 ±2.01	68.34 ±2.27	92.95 ±1.21	48.68 ±0.75
GPT-OSS 20B	71.89 ±7.37	73.73 ±7.41	57.55 ±6.95	54.79 ±9.18	41.65 ±3.94	35.46 ±9.00	47.30 ±7.35	47.19 ±8.81	55.10 ±6.82	46.40 ±2.51	43.98 ±6.74	38.60 ±6.93	43.87 ±7.41	46.24 ±4.84	59.13 ±8.87	45.78 ±5.26	52.86 ±6.25	55.88 ±6.82	54.99 ±5.48	59.19 ±11.96	56.03 ±5.55
Llama-3.2 1B (GSM8K)	75.92 ±0.75	-	-	-	66.14 ±2.03	61.07 ±2.54	59.57 ±2.91	60.34 ±2.97	56.18 ±2.35	59.19 ±0.60	57.33 ±1.35	56.17 ±1.40	57.90 ±1.41	56.21 ±2.69	72.40 ±1.76	60.56 ±3.49	59.41 ±3.46	58.79 ±3.45	58.45 ±3.42	65.97 ±2.35	50.15 ±2.35

Table 4: Reasoning recovered R (%) of every training-free method.

Model	Arm	Soft Thinking					Stochastic Soft Thinking					Latent Thinking				
		1	16	32	64	128	1	16	32	64	128	1	16	32	64	128
Llama-3.1 8B	Thought	2.5 ±1.0	-2.2 ±1.4	0.8 ±1.6	0.7 ±1.8	2.1 ±1.7	2.9 ±1.0	-2.5 ±1.4	-1.5 ±1.6	-2.3 ±1.8	-1.7 ±1.9	16.0 ±1.5	-215 ±18.1	-283 ±20.7	-318 ±22.4	-358 ±23.3
	Other question	3.0 ±1.0	1.6 ±1.4	8.3 ±1.5	9.4 ±1.5	6.5 ±1.4	3.2 ±1.1	2.8 ±1.5	6.3 ±1.4	5.5 ±1.5	4.5 ±1.6	16.2 ±1.5	-212 ±17.7	-279 ±20.1	-316 ±21.3	-352 ±22.5
	Random	12.3 ±1.0	22.1 ±0.9	17.5 ±1.1	17.2 ±1.2	17.1 ±1.1	12.3 ±1.0	22.1 ±0.9	17.5 ±1.1	17.2 ±1.2	17.1 ±1.1	12.3 ±1.0	22.1 ±0.9	17.5 ±1.1	17.2 ±1.2	17.1 ±1.1
Llama-3.3 70B	Thought	0.7 ±4.2	1.9 ±8.3	-9.6 ±8.5	-8.7 ±7.9	-11.0 ±8.0	0.0 ±4.1	3.2 ±7.9	-6.2 ±9.5	-9.7 ±9.4	-16.6 ±12.1	26.4 ±4.1	34.2 ±3.3	34.1 ±3.6	32.7 ±3.6	30.6 ±4.0
	Other question	1.5 ±3.9	4.4 ±4.6	3.5 ±3.9	7.8 ±8.0	6.1 ±9.4	3.0 ±4.7	8.0 ±6.4	10.0 ±5.2	9.4 ±5.1	8.2 ±8.1	24.9 ±5.8	34.1 ±3.0	33.5 ±2.9	31.9 ±2.2	30.8 ±2.1
	Random	21.1 ±3.7	31.0 ±2.7	32.6 ±3.2	32.1 ±3.4	32.4 ±3.7	21.1 ±3.7	31.0 ±2.7	32.6 ±3.2	32.1 ±3.4	32.4 ±3.7	21.1 ±3.7	31.0 ±2.7	32.6 ±3.2	32.1 ±3.4	32.4 ±3.7
DS-R1-Qwen 32B	Thought	-16.7 ±0.9	-32.9 ±1.6	-33.4 ±1.8	-32.8 ±1.9	-33.8 ±2.0	-17.1 ±0.9	-31.0 ±1.5	-32.9 ±1.6	-32.8 ±1.8	-32.8 ±1.9	5.1 ±0.8	9.3 ±0.7	-9.3 ±2.0	-158 ±10.3	-244 ±13.5
	Other question	-16.0 ±0.9	-27.1 ±1.6	-26.5 ±1.6	-24.0 ±1.6	-22.0 ±1.7	-16.7 ±1.0	-26.0 ±1.6	-26.0 ±1.5	-24.4 ±1.6	-23.1 ±1.8	5.4 ±0.8	9.4 ±0.7	-9.8 ±0.7	-162 ±10.8	-246 ±13.3
	Random	-13.2 ±0.9	5.9 ±0.8	9.5 ±0.8	12.0 ±0.9	-2.0 ±1.5	-13.2 ±0.9	5.9 ±0.8	9.5 ±0.8	12.0 ±0.9	-2.0 ±1.5	-13.2 ±0.9	5.9 ±0.8	9.5 ±0.8	12.0 ±0.9	-2.0 ±1.5
Skywork-OR1 32B	Thought	-13.5 ±0.7	-31.2 ±1.4	-34.7 ±1.4	-36.7 ±1.4	-42.0 ±1.9	-13.9 ±0.8	-29.6 ±1.3	-32.7 ±1.5	-36.3 ±1.5	-40.4 ±1.9	5.2 ±0.7	8.8 ±0.6	-7.8 ±1.5	-142 ±9.1	-200 ±10.9
	Other question	-13.3 ±0.8	-26.0 ±1.4	-26.9 ±1.5	-26.0 ±1.4	-27.2 ±1.6	-13.8 ±0.8	-24.3 ±1.4	-25.2 ±1.4	-25.2 ±1.4	-26.3 ±1.6	5.5 ±0.7	8.7 ±0.6	-8.0 ±1.7	-140 ±9.3	-199 ±11.0
	Random	-8.5 ±0.7	6.0 ±0.7	9.0 ±0.7	10.6 ±1.0	4.2 ±1.3	-8.5 ±0.7	6.0 ±0.7	9.0 ±0.7	10.6 ±1.0	4.2 ±1.3	-8.5 ±0.7	6.0 ±0.7	9.0 ±0.7	10.6 ±1.0	4.2 ±1.3
GPT-OSS 20B	Thought	13.1 ±1.7	17.5 ±2.5	15.2 ±2.6	11.7 ±2.9	9.2 ±3.1	10.9 ±2.1	14.0 ±2.9	12.0 ±3.0	10.8 ±3.2	10.2 ±3.3	2.7 ±1.4	0.4 ±2.1	2.6 ±1.9	0.8 ±2.1	-2.0 ±2.3
	Other question	12.7 ±1.6	16.6 ±2.3	14.8 ±2.6	11.3 ±2.9	9.0 ±3.2	10.2 ±1.7	14.0 ±3.0	12.0 ±3.1	10.7 ±3.3	10.0 ±3.3	3.5 ±1.4	0.5 ±2.1	2.1 ±2.1	0.9 ±2.1	-1.6 ±2.1
	Random	-55.3 ±4.4	-48.1 ±4.3	-46.9 ±4.5	-48.0 ±4.6	-49.2 ±4.5	-55.3 ±4.4	-48.1 ±4.3	-46.9 ±4.5	-48.0 ±4.6	-49.2 ±4.5	-55.3 ±4.4	-48.1 ±4.3	-46.9 ±4.5	-48.0 ±4.6	-49.2 ±4.5
Llama-3.2 1B (GSM8K)	Thought	1.3 ±0.2	2.3 ±0.3	2.3 ±0.4	-9.8 ±0.6	-33.4 ±0.6	1.2 ±0.2	-2.1 ±0.3	-4.7 ±0.4	-13.8 ±0.6	-33.5 ±0.6	14.2 ±0.2	-15.9 ±0.6	-25.3 ±1.0	-36.3 ±1.3	-46.8 ±1.6
	Random	17.7 ±0.1	11.9 ±0.2	9.5 ±0.2	5.3 ±0.2	1.1 ±0.2	17.7 ±0.1	11.9 ±0.2	9.5 ±0.2	5.3 ±0.2	1.1 ±0.2	17.7 ±0.1	11.9 ±0.2	9.5 ±0.2	5.3 ±0.2	1.1 ±0.2

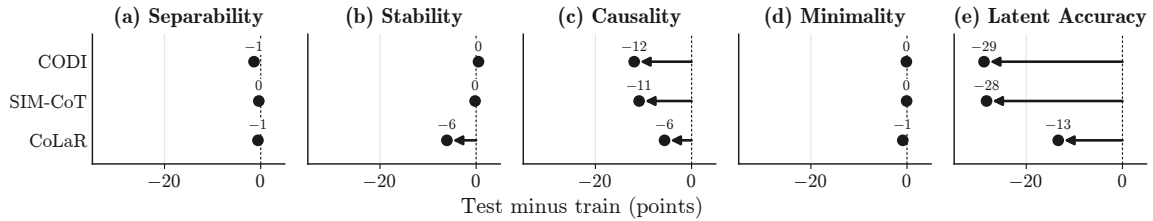
Figure 10: **Train and test.** Change of each measure from the GSM8K train questions to the test questions.

Table 5: Output information r_Y (%) and extra input information r_X (%) of every training-free method.

Model	Arm	Soft Thinking					Stochastic Soft Thinking					Latent Thinking				
		1	16	32	64	128	1	16	32	64	128	1	16	32	64	128
Output information r_Y																
Llama-3.1 8B	Thought	6.6	36.6	43.7	49.1	54.0	6.7	32.3	37.2	43.3	49.2	1.1	-686	-800	-898	-963
		± 1.4	± 3.8	± 4.1	± 3.8	± 3.5	± 1.3	± 3.6	± 3.8	± 4.0	± 3.7	± 3.2	± 86.1	± 94.6	± 102	± 105
	Random	-4.5	8.7	9.8	11.1	4.7	-4.5	8.7	9.8	11.1	4.7	-4.5	8.7	9.8	11.1	4.7
Llama-3.3 70B	Thought	± 1.4	± 1.9	± 2.0	± 2.1	± 2.2	± 1.4	± 1.9	± 2.0	± 2.1	± 2.2	± 1.4	± 1.9	± 2.0	± 2.1	± 2.2
		2.0	23.5	23.1	26.4	27.5	1.9	19.9	16.0	22.7	22.9	14.3	9.5	8.9	7.8	9.2
	Random	± 1.9	± 15.6	± 13.7	± 13.5	± 11.1	± 2.0	± 12.5	± 12.9	± 12.0	± 12.3	± 8.7	± 5.9	± 5.1	± 4.5	± 5.1
DS-R1-Qwen 32B	Thought	1.8	7.1	6.2	6.2	7.7	1.8	7.1	6.2	6.2	7.7	1.8	7.1	6.2	6.2	7.7
		± 3.6	± 4.2	± 4.0	± 4.8	± 6.2	± 3.6	± 4.2	± 4.0	± 4.8	± 6.2	± 3.6	± 4.2	± 4.0	± 4.8	± 6.2
	Random	0.2	14.4	16.8	21.4	27.8	-0.1	9.3	12.3	16.6	24.2	0.7	3.0	-10.2	-65.3	-99.8
Skywork-OR1 32B	Thought	± 0.1	± 4.0	± 4.1	± 3.8	± 3.7	± 0.2	± 3.3	± 3.0	± 3.0	± 3.1	± 0.2	± 0.8	± 2.2	± 8.4	± 9.2
		3.4	5.6	4.6	4.1	-4.6	3.4	5.6	4.6	4.1	-4.6	3.4	5.6	4.6	4.1	-4.6
	Random	± 0.6	± 1.0	± 0.8	± 0.9	± 1.4	± 0.6	± 1.0	± 0.8	± 0.9	± 1.4	± 0.6	± 1.0	± 0.8	± 0.9	± 1.4
GPT-OSS 20B	Thought	-0.2	11.9	15.1	18.4	24.5	-0.4	8.9	13.0	17.0	24.3	0.8	3.3	-10.0	-67.5	-101
		± 0.2	± 5.1	± 5.1	± 4.9	± 5.0	± 0.2	± 3.9	± 3.9	± 4.5	± 5.8	± 0.2	± 0.8	± 2.0	± 8.4	± 9.7
	Random	2.2	5.2	3.7	2.9	-6.7	2.2	5.2	3.7	2.9	-6.7	2.2	5.2	3.7	2.9	-6.7
Llama-3.2 1B (GSM8K)	Thought	± 0.5	± 1.1	± 0.8	± 0.6	± 2.0	± 0.5	± 1.1	± 0.8	± 0.6	± 2.0	± 0.5	± 1.1	± 0.8	± 0.6	± 2.0
		-0.7	2.4	1.0	1.0	-0.6	1.6	4.0	1.8	3.1	-1.4	-316	-884	-937	-964	-993
	Random	± 1.1	± 2.5	± 2.3	± 2.8	± 4.0	± 1.6	± 3.3	± 4.2	± 4.9	± 5.9	± 38.7	± 98.9	± 105	± 108	± 114
Llama-3.1 8B	Thought	-341	-221	-194	-177	-169	-341	-221	-194	-177	-169	-341	-221	-194	-177	-169
		± 40.7	± 37.7	± 35.7	± 35.0	± 35.1	± 40.7	± 37.7	± 35.7	± 35.0	± 35.1	± 40.7	± 37.7	± 35.7	± 35.0	± 35.1
	Random	0.7	28.0	39.5	47.6	46.2	1.2	25.8	36.4	43.7	42.9	1.6	-20.6	-28.3	-37.9	-48.9
Llama-3.2 1B (GSM8K)	Thought	± 0.1	± 0.4	± 0.4	± 0.4	± 0.4	± 0.2	± 0.4	± 0.4	± 0.4	± 0.4	± 0.1	± 0.6	± 1.1	± 1.6	± 2.2
		-0.8	-2.2	-3.5	-4.9	-4.5	-0.8	-2.2	-3.5	-4.9	-4.5	-0.8	-2.2	-3.5	-4.9	-4.5
	Random	± 0.1	± 0.2	± 0.2	± 0.2	± 0.2	± 0.1	± 0.2	± 0.2	± 0.2	± 0.2	± 0.1	± 0.2	± 0.2	± 0.2	± 0.2
Extra input information r_X																
Llama-3.1 8B	Thought	-0.2	-0.5	0.4	1.5	2.8	-0.1	-0.1	0.7	2.0	3.9	-2.8	-938	-1064	-1171	-1252
		± 0.1	± 0.3	± 0.4	± 0.6	± 0.8	± 0.1	± 0.2	± 0.4	± 0.6	± 0.9	± 0.4	± 92.5	± 104	± 113	± 114
	Random	0.0	-0.4	-0.4	-0.4	-0.2	0.0	-0.4	-0.4	-0.4	-0.2	0.0	-0.4	-0.4	-0.4	-0.2
Llama-3.3 70B	Thought	± 0.1	± 0.1	± 0.2	± 0.2	± 0.2	± 0.1	± 0.1	± 0.2	± 0.2	± 0.2	± 0.1	± 0.1	± 0.2	± 0.2	± 0.2
		0.4	0.5	0.1	0.0	0.8	0.5	0.5	0.2	-0.4	0.8	0.8	1.6	1.5	1.4	0.9
	Random	± 0.1	± 0.5	± 0.5	± 0.6	± 1.1	± 0.2	± 0.6	± 0.6	± 0.6	± 1.6	± 0.3	± 0.5	± 0.6	± 0.6	± 1.0
DS-R1-Qwen 32B	Thought	0.8	1.4	1.4	1.3	1.2	0.8	1.4	1.4	1.3	1.2	0.8	1.4	1.4	1.3	1.2
		± 0.3	± 0.3	± 0.4	± 0.4	± 0.6	± 0.3	± 0.3	± 0.4	± 0.4	± 0.6	± 0.3	± 0.3	± 0.4	± 0.4	± 0.6
	Random	0.0	0.4	1.0	1.9	3.3	0.1	0.7	1.5	3.1	5.1	0.0	0.5	-1.1	-6.6	-20.4
Skywork-OR1 32B	Thought	± 0.0	± 0.2	± 0.3	± 0.5	± 1.0	± 0.0	± 0.2	± 0.4	± 0.7	± 1.1	± 0.0	± 0.1	± 0.3	± 1.0	± 2.4
		0.1	1.0	1.0	-0.1	-6.5	0.1	1.0	1.0	-0.1	-6.5	0.1	1.0	1.0	-0.1	-6.5
	Random	± 0.0	± 0.1	± 0.1	± 0.1	± 0.3	± 0.0	± 0.1	± 0.1	± 0.1	± 0.3	± 0.0	± 0.1	± 0.1	± 0.1	± 0.3
GPT-OSS 20B	Thought	0.1	0.6	1.3	1.8	3.3	0.0	0.6	1.2	2.1	4.4	-0.1	0.9	-0.8	-7.1	-18.2
		± 0.0	± 0.2	± 0.3	± 0.6	± 0.8	± 0.0	± 0.2	± 0.4	± 0.7	± 1.0	± 0.0	± 0.1	± 0.5	± 1.6	± 2.9
	Random	-0.1	1.3	1.3	0.5	-9.2	-0.1	1.3	1.3	0.5	-9.2	-0.1	1.3	1.3	0.5	-9.2
Llama-3.2 1B (GSM8K)	Thought	± 0.0	± 0.1	± 0.1	± 0.2	± 0.6	± 0.0	± 0.1	± 0.1	± 0.2	± 0.6	± 0.0	± 0.1	± 0.1	± 0.2	± 0.6
		-1.0	-0.5	-3.5	-3.3	-4.1	-0.4	0.9	-0.7	0.0	-0.3	-178	-626	-690	-753	-802
	Random	± 0.4	± 0.9	± 1.1	± 1.1	± 1.4	± 0.6	± 1.1	± 1.3	± 1.5	± 1.7	± 23.8	± 85.2	± 95.9	± 103	± 110
Llama-3.1 8B	Thought	-114	-206	-180	-184	-172	-114	-206	-180	-184	-172	-114	-206	-180	-184	-172
		± 19.6	± 35.3	± 33.3	± 34.1	± 34.8	± 19.6	± 35.3	± 33.3	± 34.1	± 34.8	± 19.6	± 35.3	± 33.3	± 34.1	± 34.8
	Random	-1.3	-1.5	1.3	4.6	7.0	-1.6	-0.4	2.7	6.2	8.9	-0.5	-9.5	-17.2	-26.1	-38.6
Llama-3.2 1B (GSM8K)	Thought	± 0.0	± 0.1	± 0.2	± 0.2	± 0.2	± 0.0	± 0.1	± 0.2	± 0.2	± 0.2	± 0.0	± 0.6	± 1.1	± 1.7	± 2.1
		-0.6	-0.1	-0.5	-1.3	-0.3	-0.6	-0.1	-0.5	-1.3	-0.3	-0.6	-0.1	-0.5	-1.3	-0.3
	Random	± 0.0	± 0.1	± 0.1	± 0.1	± 0.1	± 0.0	± 0.1	± 0.1	± 0.1	± 0.1	± 0.0	± 0.1	± 0.1	± 0.1	± 0.1

Table 6: The four measures (%) and latent accuracy (%) of every trained method on GSM8K.

Arm	<i>Llama-3.2 1B</i>			<i>GPT-2</i>	
	CODI	SIM-CoT	CoLaR	CODI	SIM-CoT
<i>Separability</i>					
Thought	98.75 ± 0.06	99.55 ± 0.04	99.18 ± 0.04	97.65 ± 0.07	99.46 ± 0.03
Output Emb.	99.99 ± 0.00	99.98 ± 0.01	99.99 ± 0.00	99.98 ± 0.01	99.98 ± 0.01
Input Emb.	91.25 ± 0.18	91.56 ± 0.17	95.39 ± 0.13	93.27 ± 0.15	93.11 ± 0.15
Random Vec.	49.78 ± 0.29	49.85 ± 0.28	49.94 ± 0.30	49.96 ± 0.28	49.98 ± 0.29
<i>Distributional consistency score</i>					
Thought	78.97 ± 1.18	80.84 ± 1.00	62.55 ± 1.28	78.50 ± 0.72	80.60 ± 0.85
Output Emb.	65.03 ± 1.24	65.31 ± 1.44	68.94 ± 1.87	66.81 ± 0.21	67.62 ± 1.34
Input Emb.	56.42 ± 1.51	57.36 ± 2.45	60.22 ± 0.82	62.05 ± 1.03	61.56 ± 1.97
Random Vec.	48.71 ± 1.21	49.96 ± 1.05	50.57 ± 1.81	51.20 ± 1.02	51.86 ± 1.86
<i>Reasoning recovered R</i>					
Thought	89.0 ± 0.5	92.2 ± 0.3	2.1 ± 2.2	83.8 ± 0.4	76.5 ± 0.5
Other question	39.9 ± 0.9	45.8 ± 0.6	-104 ± 3.0	55.5 ± 0.6	38.6 ± 0.8
Random	64.2 ± 0.5	74.1 ± 0.3	4.7 ± 0.8	16.3 ± 0.3	10.1 ± 0.3
<i>Output information r_Y</i>					
Thought	4.3 ± 0.3	10.4 ± 0.2	28.3 ± 0.3	33.4 ± 0.5	25.3 ± 0.4
Random	3.5 ± 0.2	1.8 ± 0.1	-1.3 ± 0.1	-17.2 ± 0.6	-1.1 ± 0.6
<i>Extra input information r_X</i>					
Thought	0.0 ± 0.0	0.0 ± 0.0	-4.9 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
Random	0.0 ± 0.0	0.0 ± 0.0	-2.6 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
<i>Latent accuracy</i>					
Thought	55.5	55.7	25.5	42.5	39.3

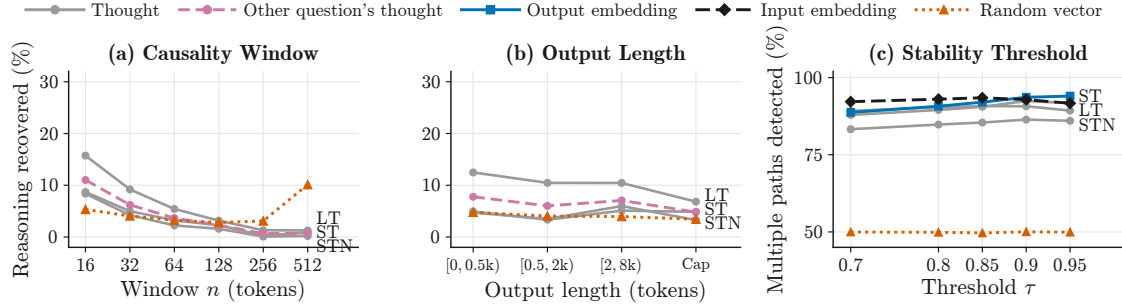


Figure 11: **Robustness of the measures.** Training-free methods at their best step, averaged over models. (a) Reasoning recovered at each window of n tokens. (b) Reasoning recovered in bins of output length. (c) DCS at each threshold τ on the four dense models.

E SUPPORTING DIAGNOSTICS

Robustness of the measures.. The causality measure averages the divergence over the last n tokens of each output. The output lengths of the models differ by an order of magnitude (Appendix F). Figure 11(a) recomputes the reasoning recovered at windows from 16 to 512 tokens. Figure 11(b) pools it in bins of output length, with the outputs that reach the generation cap in a separate bin. The reasoning recovered of every training-free method decreases as the window grows. At every window and in every length bin, the training-free methods recover at most a limited portion of the reasoning, about as much as the thought representation of another question, as in Figure 5(c). The stability measure binarizes the cosine similarity between the output embeddings at the threshold τ . Figure 11(c) recomputes the DCS at thresholds from 0.7 to 0.95. At every threshold, every training-free method exceeds the random vector and at most matches the input embedding, as in Figure 5(b). Hence the verdicts of Section 4.2 hold at every window, output length, and threshold.

Stability on GPT-OSS-20B.. The difference-of-means direction separates the inputs whose outputs form more than one semantic class from the rest, hence the measure requires inputs on which the model generates more than one class. At the threshold of the audit, GPT-OSS-20B generates more than one class on one percent of the inputs, compared with 16 to 46 percent on the four dense models. The DCS of the random vector exceeds one half on this model, since the positive class contains few inputs.

Linear encoding of the semantic entropy.. The semantic entropy is computed from the spread of the output embeddings, whereas a difference-of-means direction scores their mean position, hence the output embedding is a reference informed by the outputs rather than an upper bound. A difference-of-means direction recovers the semantic entropy from a representation that encodes it linearly (Cencerrado et al., 2026). A representation that encodes the semantic entropy through a nonlinear structure scores near the random vector under this direction. A DCS near the random vector is therefore evidence against a linear encoding of the semantic entropy.

F DATASET DETAILS

Tasks.. BBEH replaces each of the 23 tasks of BIG-Bench Hard with a harder task that tests a similar reasoning skill, such as many-hop deduction, tracking of objects and time over long contexts, and learning of new rules from examples (Kazemi et al., 2025). Every task contains 200 problems except Disambiguation QA, which contains 120. Every model receives the prompt of the benchmark.

Accuracy and output length.. Table 7 reports the accuracy and the output length of each model. The adjusted harmonic mean of Llama-3.1-8B and of DeepSeek-R1-Distill-Qwen-32B differs by less than one percentage point from the value that BBEH reports, hence the generation reproduces the published accuracy of these models. A failed extraction lowers the accuracy without changing any measure, since every measure is computed on the generated tokens. Output length differs by an order of magnitude across the models. Figure 11(b) reports the reasoning recovered in bins of output length.

Table 7: BBEH accuracy (%), failed extraction (%), and median output length in characters of each model.

Model	Pass@1	Pass@8	Harmonic mean	Failed extraction	Median length
Llama-3.1 8B	7.9	21.0	3.1	42.9	2,076
Llama-3.3 70B	20.3	31.3	8.7	4.3	2,091
DS-R1-Qwen 32B	17.2	23.7	5.5	34.9	7,458
Skywork-OR1 32B	20.2	23.1	5.3	49.6	17,043
GPT-OSS 20B	19.1	20.8	4.0	58.3	19,316

Qualitative examples.. Two outputs of the same input share one input embedding, yet their output embeddings may be far apart. The boxes below give two outputs that Llama-3.3-70B generates for one input, with the cosine similarity of their output embeddings under Llama-Embed-Nemotron-8B. The two outputs differ both in their reasoning and in their verdict, since one output derives the correct answer over several steps and the other states an incorrect answer in one line.

Output 2 of Llama-3.3-70B-Instruct (Causal Understanding, example 113)

Question (truncated):

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.

Output 2:

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...

Verdict: Yes *Correct Answer* ✓

Output 7 of Llama-3.3-70B-Instruct (Causal Understanding, example 113)

Question (truncated):

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.

Output 7:

The answer is: No.

Verdict: No *Wrong Answer* ✗

Cosine similarity to Output 2: 0.063

The eight outputs of this input form three groups under the embedder, and the outputs within a group are nearly identical. One input thus yields outputs that differ in their reasoning and in their verdict, whereas all eight share one input embedding. A thought representation that exceeds the input embedding on separability and stability therefore encodes information about the outputs that the input lacks.

G EXTENDED RELATED WORK

Surveys and theoretical analyses.. Surveys organize latent reasoning methods into token-wise methods, which replace discrete tokens with continuous vectors, and layer-wise methods, which propagate a latent state across the depth of the transformer (Chen et al., 2025). Discrete chain of thought restricts reasoning to high-certainty decisions, whereas continuous representations support exploration at the cost of amplified noise on computational tasks, a trade-off governed by a scalar index of decisional certainty (Zou et al., 2026b). Evaluations of reasoning in language models report downstream accuracy, which conflates the quality of

the reasoning with pattern matching (Mondorf & Plank, 2024). Surveys of efficient reasoning compare methods by their compute cost and their length budget (Sui et al., 2025; Feng et al., 2025b). These works catalogue methods and their downstream trade-offs, whereas the measures of Section 3 evaluate the thought representation independently of the decoding strategy and of downstream task accuracy.

Internal world models and the localization of computation.. Propositional probes extract the states of the internal world models of language models as structured propositions (Feng et al., 2025a). Object concept representations that resemble those of humans emerge in multimodal language models without explicit supervision (Du et al., 2025). Swapping ranges of transformer layers transfers cross-lingual knowledge between models (Bandarkar et al., 2025). Adaptation concentrates in the early and middle adapter layers (Alabi et al., 2024). Semantically meaningful computation is thus localized at identifiable depths. Middle layers can be reordered or skipped at a small cost in accuracy, which indicates that representations at middle depths share their structure (Sun et al., 2025). Circuit tracing through the contributions to the residual stream shows that feature-level computation shifts across layers depending on the prompt (Ameisen et al., 2025). Since the depth of the relevant computation changes with the prompt, the audit examines candidates taken from the final layer, from every layer, and from latent steps (Section 4.1).

Sentence-level representations.. Large Concept Models (LCM team et al., 2024) predict the next sentence embedding instead of the next token, with SONAR (Duquenne et al., 2023) as the shared multilingual sentence encoder. SONAR-LLM (Dragunov et al., 2025) reasons autoregressively in the sentence embedding space and decodes back to tokens, at a granularity coarser than a token and finer than a document. These methods fix the semantic granularity of their representations in advance. Since Definition 3 states validity for a thought representation of any granularity, the four measures apply to sentence-level representations as well as to the token-level candidates of Section 4.