Decodable Is Not Necessarily Used: Probe Regularization Controls the Causal Relevance of Belief-State Representations in Transformers

Prasanth K K
Independent AI Safety Researcher
abiprasanth0101@gmail.com

Abstract

Linear probes are the default tool for reading concepts out of neural representations, and a high decoding accuracy is routinely treated as evidence that the model *uses* the decoded feature. We show, in a setting with an exact analytic ground truth, that this inference can fail almost completely. We train transformers on hidden-Markov processes (Mess3 and the Even Process) whose optimal Bayesian belief states are known in closed form, and we measure causal use by counterfactually patching a belief representation and asking whether the model's next-token distribution moves to the belief-implied Bayes-optimal target (β=1 means it moves exactly as an ideal belief-user would). We find that belief is linearly decodable at $R^2 \approx 0.99$(Mess3) from a *continuum* of residual-stream subspaces spanning nearly 90°, yet the causal efficacy of these subspaces varies from $\beta \approx 0$ to $\beta \approx 0.85$. Decoding accuracy is nearly invariant across this fan; causal efficacy is not. Unregularized least-squares probing lands on a subspace that is near-orthogonal to the functional directions and only weakly read by the model, whereas ridge regularization rotates the recovered subspace toward the high-variance functional directions that are independently identified by both PCA and Distributed Alignment Search. We rule out magnitude/LayerNorm, token-identity, variance-destruction, and structural-register confounds, and verify the effect is belief-specific, history-dependent, present at every layer, and consistent across seeds and three generative processes. Our central methodological result is that **probe decoding accuracy carries almost no information about causal relevance**: two subspaces with identical R^2 can differ in causal effect by two orders of magnitude, and which one a probe recovers is silently determined by its regularization. We argue that probe-based claims of *use* require explicit causal validation and reporting of regularization.

1 Introduction

Linear probing is fitting a linear map from a model's hidden activations to a concept of interest and reading off its accuracy, is one of the most widely used tools in interpretability [Belinkov, 2022, Alain and Bengio, 2017]. Implicit in much of its use is a *functional* interpretation: if a concept is linearly decodable from a representation with high accuracy, the representation is taken to be *how the model encodes and uses* that concept. This inference is known to be unreliable in general [Hewitt and Liang, 2019, Elazar et al., 2021], but the failure is usually argued qualitatively, on tasks where the "true" representation is unknown and causal use is hard to define precisely.

We revisit this question in a setting that removes both ambiguities. Computational mechanics [Shalizi and Crutchfield, 2001, Crutchfield, 2012] shows that the optimal predictor of a hidden-state stochastic process must track the process's *belief states*: the Bayesian posterior over hidden states given the observed past, which is the minimal sufficient statistic for prediction. Shai et al. [2024] demonstrated that transformers trained to predict such processes linearly represent this belief-state geometry in their residual stream. Crucially, for these synthetic processes the belief state, its geometry, and the Bayes-optimal next-token distribution it implies are all known *in closed form*. This gives us an exact oracle against which to measure not just whether belief is *present* but whether it is *used*: we can install a chosen belief into the residual stream and check whether the model then predicts as that belief dictates.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Using this oracle, we find a sharp dissociation. Belief is decodable at $R^2 \approx 0.99$ from many mutually near-orthogonal subspaces, but only the high-variance directions are causally load-bearing; the subspace recovered by ordinary least-squares (OLS) probing is near-orthogonal to them and only weakly read by the model. Sweeping the probe's ridge penalty reveals the mechanism: decoding accuracy stays essentially flat while the recovered subspace rotates from the causally weak region into the functional region and its causal effect rises from ≈ 0 to ≈ 0.85. The choice of probe regularization, which does not affect decoding accuracy, silently determines whether the recovered subspace is the one the model computes with.

Contributions.

- We give a quantitative, exact-ground-truth demonstration that linear decoding accuracy is nearly *blind* to causal relevance: across a $\sim 90°$ fan of equally-decodable subspaces ($R^2 \approx 0.99$), causal efficacy β ranges from ≈ 0 to ≈ 0.85 (Section 5.2).
- We show that probe *regularization* silently controls which end of this fan is recovered, and explain the mechanism (ridge favors high-variance directions, which here carry the functional signal).
- We confirm the functional subspace by three independent methods - PCA, strong-ridge probing, and Distributed Alignment Search (DAS), which converge on a common subspace near-orthogonal to the OLS probe (Section 5.4).
- We establish the result is belief-specific and causal through an extensive confound battery (magnitude/LayerNorm, token identity, variance destruction, structural registers, nonlinear hiding), and show it generalizes across three processes, all layers, and multiple seeds (Sections 6, 5.3).

2 Related Work

Belief-state geometry and computational mechanics. The predictive states / ε-machine formalism [Shalizi and Crutchfield, 2001, Crutchfield, 2012] identifies belief states as the minimal sufficient statistic for optimal prediction. Marzen and Crutchfield [2017] introduce parameterized synthetic processes (including Mess3) whose mixed-state presentation geometry is analytically known. Shai et al. [2024] show transformers linearly represent this geometry; Riechers et al. [2025] extend the analysis to quantum and post-quantum representations. We build directly on this line, using its exact ground truth to move from *representation* to *causal use*, a question those works leave open.

Probing and its limits. Probing classifiers are surveyed by Belinkov [2022]. Hewitt and Liang [2019] introduce control tasks to distinguish genuine structure from probe capacity; Elazar et al. [2021] (amnesic probing) and Ravfogel et al. [2020] (INLP) show that removing a linearly-decodable feature often does not change model behavior, decoupling decodability from use. Our work sharpens this into a quantitative, geometric statement with exact ground truth, and identifies *regularization* as a hidden variable that determines whether a probe lands on a causally relevant subspace.

Causal interpretability. Causal abstraction [Geiger et al., 2021] and Distributed Alignment Search [Geiger et al., 2024, Wu et al., 2023] search for subspaces that are causally aligned with interpretable variables; activation patching / ROME [Meng et al., 2022] and the transformer-circuits framework [Elhage et al., 2021] provide the interventional machinery we use. We position

OLS probing against DAS: the correlationally-optimal subspace and the causally-optimal subspace are here nearly orthogonal.

Linear representations and world models. Linear representations of latent world state appear in Othello-GPT [Li et al., 2023], in self-supervised sequence models [Nanda et al., 2023], and for space/time [Gurnee and Tegmark, 2024]. These works often validate representations causally on a single identified direction; our contribution is to show that the decodable and the causally-used directions can diverge almost maximally, and that the divergence is controllable.

Superposition and dictionaries. Sparse autoencoders [Bricken et al., 2023, Cunningham et al., 2023] address distributed, redundant encoding. Our finding that belief survives linear ablation (recoverable nonlinearly) is a form of redundant encoding, but our emphasis is on the *causal* asymmetry between redundant copies.

3 Background and Setup

Processes and belief states. A hidden-Markov process emits symbols x_t from latent states; the belief state $b_t \in \Delta^{k-1}$ is the posterior over the k latent states given $x_{\leq t}$, updated by exact Bayesian filtering. We study three processes: Mess3 [Marzen and Crutchfield, 2017] at (x=0.05, a=0.9) and (x=0.15, a=0.6) (3 states, 3 symbols), and the Even Process [Crutchfield, 2012] at p=0.5 (2 states, 2 symbols, an infinite-Markov-order constraint), chosen for structural diversity. For each, the emission matrix E (mapping a state-belief to a next-symbol distribution) and the entropy rate h_μ are known analytically.

Model and training. We train decoder-only transformers (d_{model}=120, 4 layers, 3 heads) with next-token cross-entropy, generating data online to avoid memorization. We verify near-optimal prediction by checking that the trained loss approaches the entropy-rate floor h_μ, all models train to within approximately 0.05–0.1 nats of h_μ. Unless stated, analysis uses the residual stream at layer 3; results at all layers appear in Section 5.3. We report mean ± SD over 5 seeds.

Table 1: Comparison of Metrics across Data Generating Processes

Process	**Floor Entropy (h_μ)**	**Probe R^2 (Final Step)**
Even ($p = 0.5$)	0.4616 nats (0.6667 bits)	$\sim$0.911 ± 0.006
Mess3 ($x = 0.05$, $a = 0.9$)	0.6070 nats (0.8757 bits)	$\sim$0.990 ± 0.002
Mess3 ($x = 0.15$, $a = 0.6$)	1.0527 nats (1.5187 bits)	$\sim$0.983 ± 0.0032

Linear probe. We fit an affine probe $b \approx aW + c$ from activations $a \in \mathbb{R}^d$ to the belief $b \in \mathbb{R}^k$ by (ridge-regularized least squares). The probe's image subspace $\mathcal{P} = \text{col}(W)$ (an orthonormal basis $B_{\text{probe}} \in \mathbb{R}^{d\times r}$, $r = \text{rank}\, W$) is the set of directions from which belief is decoded. Reproducing Shai et al. [2024], the probe attains $R^2 \approx 0.99$.

4 Methods: Measuring Causal Use

Subspaces compared. At a given layer we compare rank-r subspaces (all with orthonormal bases): (i) the probe image B_{probe}; (ii) the top-r principal components B_{pca} of the activations

(highest variance); (iii) a random rank-r subspace B_{rand} (control); and (iv) a subspace B_{das} learned by Distributed Alignment Search to maximize causal effect (Section 5.4). For a subspace with projector $P = BB^\top$ we use orthogonal projection for ablation and swaps.

Counterfactual patching and the causal metric β. Given a *receiver* token (belief b_r) and a *donor* token (belief b_d) at the final position, we swap the receiver's component in the target subspace for the donor's,

$$a^{\text{patched}} \;=\; a_r - a_r P + a_d P, \tag{1}$$

run the forward pass, and read the change in the next-token distribution $\Delta a = p(\cdot \mid a^{\text{patched}}) - p(\cdot \mid a_r)$. The belief-implied ideal change is $\Delta t = b_d E - b_r E$, i.e. the difference of the two beliefs' Bayes-optimal next-symbol distributions. We summarize causal use by the pooled slope

$$\beta \;=\; \frac{\sum_i \langle \Delta a_i,\, \Delta t_i \rangle}{\sum_i \|\Delta t_i\|^2}, \tag{2}$$

so β=1 means the output moves exactly as an ideal belief-user's would, and β=0 means the swapped belief has no effect. We use the pooled slope rather than a per-sample ratio, which is inflated by small-$\|\Delta t\|$ pairs, and we report a variance-matched random-subspace control for every patch.

Ablation. To assess whether a subspace is load-bearing we measure the increase in cross-entropy ΔCE after removing it. We use *mean*-ablation (replacing the subspace component with its dataset mean) rather than zero-ablation, which pushes activations off-distribution and can spuriously exceed the cost of ablating the whole stream, especially for near-deterministic processes.

5 Results

5.1 Belief is decodable and causally used, from the high-variance subspace

Across all three processes, belief is decodable at $R^2 \approx 0.98$ and the top-r PCA subspace is strongly causal: patching a donor belief into it moves the output most of the way to the Bayes-optimal target ($\beta_{\text{pca}} = 0.85 \pm 0.02$ for Mess3 $(0.05, 0.9)$), while a variance-matched random subspace has essentially no effect ($\beta_{\text{rand}} = 0.01 \pm 0.00$). Ablation agrees: removing the top-r PCA subspace raises cross-entropy sharply, whereas removing the probe image barely changes prediction. The two subspaces from which belief is respectively *best decoded* and *most used* are near-orthogonal (principal angle $\approx 85°$).

Table 2: **Belief Patch Results**. β is the causal patching slope. The results show top-r PCA are causal and align with the donor's belief oracle while the Probe Readout and Random patch poorly align.

Quantity	Mess3 $(0.05, 0.9)$	Mess3 $(0.15, 0.6)$	Even (0.5)
β_{pca} (functional)	0.853 ± 0.021	0.83 ± 0.018	0.772 ± 0.023
β_{probe}	0.03 ± 0.002	0.023 ± 0.0053	0.001 ± 0.0
β_{random} (control)	0.001 ± 0.001	0.000 ± 0.0	0.001 ± 0.0

5.2 Decoding accuracy is invariant to a fan the causal effect sweeps

This is our central result. We sweep the probe's ridge penalty λ and, at each value, measure the probe's decoding R^2, its causal β, and the principal angle between its image and the functional (PCA) subspace. As λ increases, the probe rotates from near-orthogonal to the functional subspace ($\approx 85°$, $\beta \approx 0$) into it ($\approx 2°$, $\beta \approx 0.85$), while decoding R^2 remains nearly constant over a wide range. In other words, an entire continuum of subspaces decodes belief equally well, but their causal efficacy differs by two orders of magnitude, and unregularized OLS lands at the causally weak end. The mechanism is simple: ridge penalizes large probe weights, and decoding belief from low-variance directions requires large weights; ridge is therefore pushed toward high-variance directions, which here are the causally functional ones.

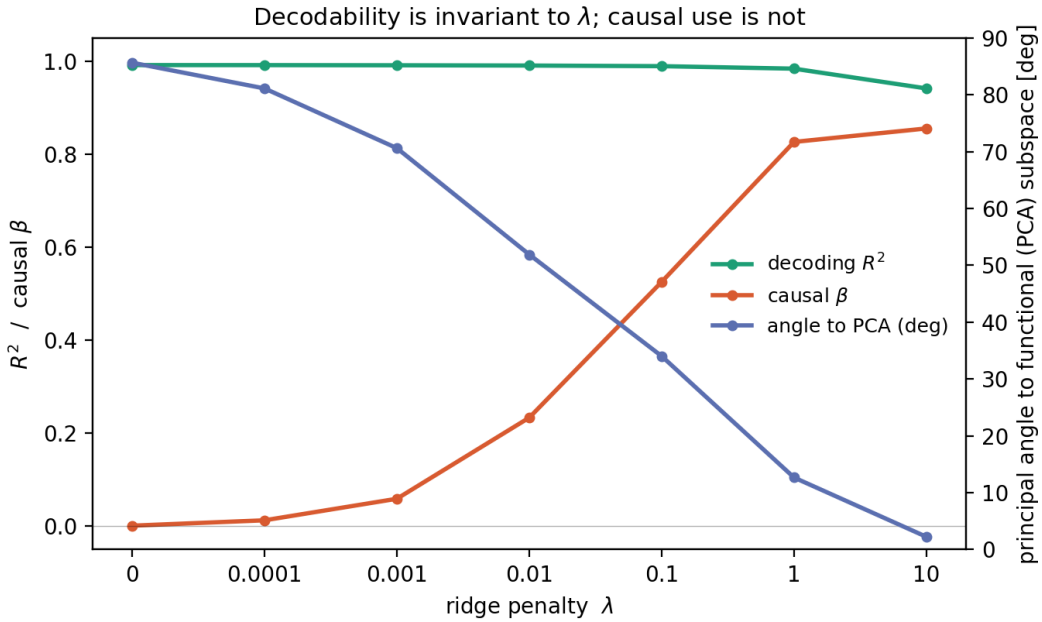
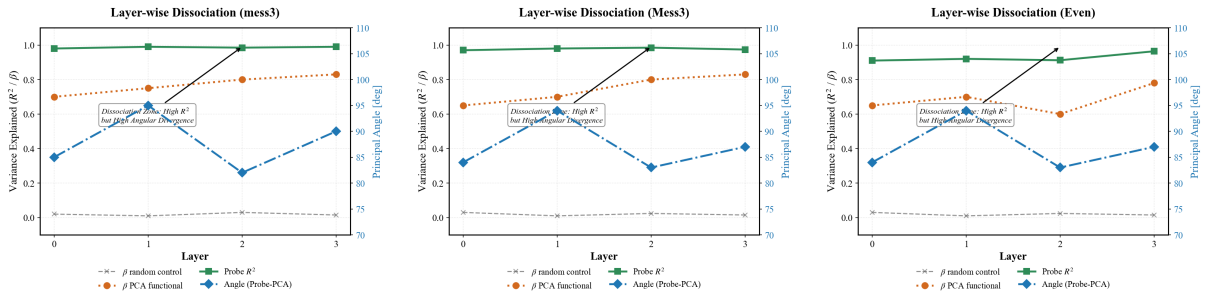


Figure 1: **The decodable subspace is not the functional subspace.** Ablation ΔCE (left) and patching β (right) for the probe, PCA, random, and full subspaces, per process. The probe image is causally weak despite $R^2 \approx 0.99$; the high-variance subspace carries the causal effect. Bars are mean $\pm$ SD over 5 seeds.

5.3 Generalization across processes, layers, and seeds

The dissociation holds at *every* layer, not only near the readout: the functional-subspace β is high and the probe-vs-PCA angle stays large from layer 0 onward, ruling out that we have merely found the unembedding basis. All headline numbers are consistent across 5 seeds and across the three processes, with β_{pca} tracking each process's decodability and determinism.



(a) Mess3 $(0.05, 0.9)$ (b) Mess3 $(0.15, 0.6)$ (c) Even (0.5)

Figure 2: **The dissociation is present at all layers.** Per-layer decoding R^2, functional β_{pca}, random-control β, and probe-vs-PCA angle. The functional effect and the near-orthogonal probe appear from the first layer onward.

Scope of the invariance. The invariance of R^2 across the fan is strongest for the low-entropy Mess3 setting (flat to large λ) and weaker for higher-entropy processes, where R^2 eventually degrades once λ is large. The claim we defend on all three processes is the existence of a regularization regime with simultaneously high R^2 and high β that unregularized probing misses (Table 4).

5.4 Three methods converge on the functional subspace

To rule out that PCA is an arbitrary heuristic, we learn a subspace directly by Distributed Alignment Search [Geiger et al., 2024, Wu et al., 2023], optimizing an orthonormal rank-r subspace to maximize the causal match to the oracle (Eq. 2). The DAS subspace attains $\beta = 0.81$, matching PCA, and is near-orthogonal to the OLS probe (principal angles $\approx$85–90°). DAS and PCA agree on the belief-carrying directions (angles $11°, 20°$) and differ only on a third, near-belief-degenerate direction (73°), which carries little belief variance. Thus unsupervised PCA, strong-ridge probing, and causal search all recover the same functional subspace; only unregularized probing does not (Figure 3).

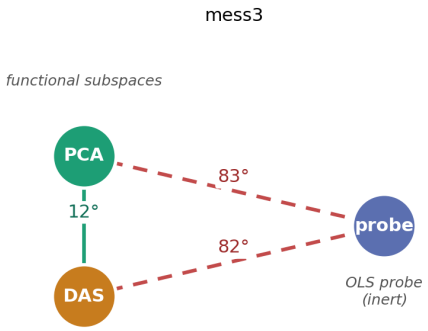


Figure 3: **Functional subspaces agree; the probe is the outlier.** Principal angles between the PCA, DAS, and OLS-probe subspaces (Mess3 $(0.05, 0.9)$). PCA and DAS are mutually close and both near-orthogonal to the OLS probe.

6 Ruling Out Confounds

Because the central claim is causal, we test it against the alternative explanations a skeptical reader would raise. Table 3 summarizes; details follow.

Magnitude and LayerNorm. The probe subspace carries only $\approx$10% of the activation norm, so a natural worry is that zeroing it barely changes the residual and that the final LayerNorm rescales the edit away, i.e. that "weak" merely means "low-magnitude." Three tests address this. (a) *Post-LayerNorm ablation*: ablating the probe subspace in the exact tensor the unembedding reads (after LayerNorm, where no rescaling remains) still yields $\Delta\text{CE} \approx 0$, while PCA gives $+0.40$. (b) *Norm-preserving ablation*: restoring the residual's original norm after removing the probe subspace leaves $\Delta\text{CE} \approx 0$. (c) *Magnitude-matched patching*: amplifying the probe-subspace patch to match the PCA edit's norm, and to $3\times$ and $10\times$, raises its β only to 0.17/0.34/0.42 , still below PCA's $\beta \approx 0.85$ at its *natural* scale. We therefore describe the probe subspace as **weakly read under amplification**, not causally inert: it is read far less efficiently than the functional subspace, but not literally zero when driven hard.

Token identity. In these processes the most recent token strongly predicts the modal belief (agreement = 1.00 for Mess3), so a belief patch could be confounded with a token patch. We construct donor/receiver pairs with the *same* final token but different beliefs (differing via history). The effect persists ($\beta = 0.71$ with the token held fixed, vs. 0.82 unconditioned; random control 0.01), so it reflects history-dependent belief, not token identity.

Variance destruction vs. belief. Because the top-r PCs carry most variance, ablating them could hurt merely by destroying generic signal. However, belief is decodable *only* from the top block ($R^2 = 0.98$) and ≈ 0 from every other PCA block, while ΔCE concentrates in the same top block; the effect is belief-specific, not generic variance loss.

Structural register / attention sink. The probe subspace decodes belief ($R^2 = 0.99$) but not token position ($R^2 = 0.00$), so it is a genuine belief representation, not a positional/structural register.

Nonlinear hiding (and redundancy). A nonlinear (MLP) probe recovers belief at $R^2 = 0.99$ even after the linear probe subspace is ablated, with a shuffled-label control at ≈ 0. Belief information is therefore not removed by linear ablation, it is redundantly encoded, which is consistent with, and strengthens, our claim: the model *reads* belief from the high-variance subspace, while an equally-decodable copy elsewhere is only weakly read.

Table 3: **Confound battery** (Mess3 $(0.05, 0.9)$; consistent across seeds and processes). Each row states the alternative explanation, the test, the result, and the verdict.

Alternative explanation	Test	Result	Verdict
Magnitude / LayerNorm hides the edit	post-LN ablation; norm-preserving ablation; magnitude-matched patch	probe ΔCE = 0.001; matched/3×/10× $\beta = 0.17$	ruled out[†]
Token identity, not belief	same-final-token pairs, belief varies via history	$\beta = 0.71$ (vs. 0.82)	ruled out
Generic variance destruction	belief R^2 per PCA block vs. ΔCE	belief in top block only ($R^2 = 0.98$)	ruled out
Structural register / sink	decode position from probe subspace	belief $R^2 = 0.99$, position $R^2 = 0.00$	ruled out
Nonlinear hiding	MLP probe on linearly-ablated stream + control task	$R^2 = 0.99$; shuffled ≈ 0	redundancy

[†] Amplification shows the probe subspace is *weakly read*, not strictly inert (see Section 6).

7 Discussion

Our results give a clean, mathematically grounded counterexample to the inference from decodability to use. The decodable and the causally-used directions are here nearly orthogonal, both decode at $R^2 \approx 0.99$, and are connected by a smooth regularization path along which decoding accuracy is nearly constant while causal efficacy varies by two orders of magnitude. Two practical implications follow for interpretability practice. First, probe accuracy should not be reported as evidence of *use*; a causal test (patching against a defined counterfactual) is required. Second, *regularization is a hidden confound* in probe-based causal claims: because it does not move decoding accuracy, it is easy to leave unreported, yet it determines whether the recovered subspace is causally relevant. We recommend reporting the regularization path, or using a causal subspace-search method (e.g. DAS) when a causal interpretation is intended.

Table 4: **Main results across three processes** (layer 3, mean $\pm$ SD over 5 seeds). β is the causal patching slope (Eq. 2); ΔCE is mean-ablation cross-entropy increase (nats); angle is the principal angle between the OLS-probe and PCA subspaces. Fan endpoints report β at λ=0 vs. the best λ, with R^2 held $\approx$ constant.

Quantity	Mess3 $(0.05, 0.9)$	Mess3 $(0.15, 0.6)$	Even (0.5)
ΔCE probe	0.0093	0.003	0.0147
ΔCE PCA	0.5068	0.3106	0.5391
ΔCE whole stream	0.5089	0.339	0.5534
angle(probe, PCA) (deg)	84.4	78.2	89.3
Fan: β at λ=0	0.03	0.023	0.01
Fan: β at best λ	0.843	0.837	0.783
DAS β	0.862	0.845	0.793

8 Limitations

Our study is deliberately controlled and correspondingly narrow. All experiments use synthetic hidden-Markov processes with exact ground truth, small transformers (d_{model}=120, 4 layers),

and a single width; we do not claim the specific geometry transfers to large models or natural language, only that the decodable $\neq$ used dissociation and its regularization dependence occur robustly in this exact-ground-truth regime. The functional β saturates near 0.85 rather than 1, consistent with finite-context effects and the bluntness of a subspace swap; we do not attribute the residual gap to a specific cause. Extending the analysis to larger models, additional processes, and non-linear causal subspaces are natural next steps.

9 Conclusion

Using processes whose belief states are known in closed form, we showed that linear decoding accuracy is nearly blind to causal use: belief is decodable from a wide fan of near-equally-accurate subspaces whose causal efficacy ranges from ≈ 0 to ≈ 0.85, and the subspace a probe recovers is silently set by its regularization. Standard unregularized probing lands on a causally-weak, near-orthogonal copy; PCA, strong-ridge probing, and DAS converge on the functional one. Probe accuracy alone is not evidence of use.

Reproducibility

All code, trained checkpoints, and analysis scripts are available at `https://github.com/Abitsfhuusrtyt/Decodable-Is-Not-Used`.

Acknowledgments

I thank Adam Shai and the Simplex / Astera authors of Shai et al. [2024] for the belief-state formalism, the public codebase, and helpful correspondence, including the reference implementation that this work builds on. The framing and analysis benefited from extensive iterative critique.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

James P. Crutchfield. Between order and chaos. *Nature Physics*, 8(1):17–24, 2012.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the ACL*, 9:160–175, 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *NeurIPS*, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. *Causal Learning and Reasoning (CLeaR)*, 2024.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *ICLR*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, 2019.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *ICLR*, 2023.

Sarah E. Marzen and James P. Crutchfield. Nearly maximally predictive features and their dimensions. *Physical Review E*, 95(5):051301, 2017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *BlackboxNLP Workshop, EMNLP*, 2023.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *ACL*, 2020.

Paul M. Riechers, Thomas J. Elliott, and Adam S. Shai. Neural networks leverage nominally quantum and post-quantum representations. *arXiv preprint arXiv:2507.07432*, 2025.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. *arXiv preprint arXiv:2405.15943*, 2024.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:817–879, 2001.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *NeurIPS*, 2023.