

**What Prediction Forces and What Training Chooses:
Necessary, Optional and Unreachable Capabilities in a
Predictor Small Enough to Enumerate**

PIATRA . INSTITUTE

September 2026

Abstract

Systems trained only to predict the next symbol can come to perform arithmetic, translation or board-state tracking, capabilities usually described as emergent. We ask a narrower question: for a given prediction objective and resource budget, which secondary capabilities are required, which are permitted, and which are excluded? We answer it by exhaustive enumeration in a system small enough to allow one. With three input bits, one predicted bit and an encoder limited to K internal states, there are 4140 distinct encoders, and each is evaluated for its exact population loss and for the best possible recovery of each input bit. Three results follow. Better prediction does not imply better representation: going from two to three internal states lowers the loss and lowers recovery of the most predictive input bit from 1.0 to 0.75, at a unique optimum and with no training or noise. The distinction between required and optional capabilities holds at exact optimality and disappears with 0.05 bits of slack, so a claim that an objective requires a capability depends on a stated tolerance. A capability can be fully encoded but unusable: at full capacity the irrelevant bit is recovered perfectly by an unrestricted decoder and not at all through the predictor’s output. A hidden-Markov construction shows the same objective requiring a latent distinction at a two-symbol prediction horizon that it does not require at one symbol.

Introduction

A next-token objective does not specify syntax, geography, arithmetic or board states, and replication does not specify an eye. That functions can arise which no one specified was noted for morphogenesis and complex systems long before language models (Turing 1952; Anderson 1972), and it became pressing when trained systems began to show capabilities that had not been listed in advance (Wei et al. 2022).

Calling a capability emergent records that it was not expected, which is a fact about the observer. The term does not distinguish four situations: the capability was requested directly; the data contained enough to support it; good performance on the objective is impossible without it; or a particular training procedure happens to find it. Only the third is a claim about the objective, and only the third would justify saying that an objective requires a capability.

Two literatures supply most of what is needed to pose the question precisely. Computational mechanics gives the exact condition under which any optimal predictor must preserve a distinction between histories, and constructs the minimal such predictor (Crutchfield and Young 1989; Shalizi and Crutchfield 2001). Work on Rashomon sets and underspecification studies the set of models that perform almost equally well on a primary task and shows that they can differ substantially in other respects (Semenova et al. 2022; D’Amour et al. 2020). Combining them gives the object computed here: the set of resource-bounded models that predict well enough, and the range a secondary capability takes across that set. Computing the range exactly requires a model set small enough to enumerate.

Construction

Three independent uniform bits (a, b, c) are observed and one bit Y is predicted, with

$$P(Y = 1 \mid a, b, c) = 0.05 + 0.60a + 0.30b.$$

The first input bit has the largest effect, the second a smaller one, and the third none. Before predicting, the system maps its eight possible inputs into at most K internal states, so an encoder is a partition of the eight inputs; encoders that differ only in the labelling of states are identical.

The number of such encoders is the eighth Bell number, 4140. For each, the population loss is exact, because inputs are uniform and the optimal prediction for each internal state is the mean of $P(Y = 1)$ over its members. The best possible recovery of an input bit is also exact: a decoder reading the internal state answers with the majority value of that bit within each state, so recovery is a count. No training, sampling, initialisation or random seed enters the calculation, and every reported value is a property of the objective and the capacity.

Following the Rashomon construction, let $\mathcal{R}_{\varepsilon, K}$ be the set of encoders with at most K states whose loss is within ε of the best achievable at that capacity, and record the lowest and highest recovery of each input bit across that set. When both are high, good performance requires the capability; when the lowest is low and the highest high, the capability is permitted but not required; when both are low, it is absent. As shown below, which category a capability falls into depends on ε .

Loss and recovery by capacity

K	best loss, bits	optimal encoders	recover a	recover b	recover c
1	1.000000	1	0.50	0.50	0.50
2	0.721928	1	1.00	0.50	0.50
3	0.643198	1	0.75	0.75	0.50
4	0.610233	1	1.00	1.00	0.50
8	0.610233	16	1.00	1.00	0.50 to 1.00

Going from two internal states to three improves prediction, from 0.721928 to 0.643198 bits, and reduces the recoverability of a , the most predictive input bit, from 1.00 to 0.75. With two states the best encoder splits the inputs on a . With three it isolates the two extreme predicted probabilities and merges the two intermediate cases, which differ in a . The merge gains 0.078730 bits and loses a quarter of the recoverability of a .

The optimum at three states is unique: one of the 966 three-state encoders attains the minimum. The decrease in recovery is therefore a property of the objective at that capacity and does not depend on how ties between optima are broken.

Beyond four states, additional capacity does not improve prediction but increases the number of optimal encoders. The loss is 0.610233 at four states and does not fall further, while the number of encoders attaining it rises to 16 at full capacity, across which the recovery of the irrelevant bit ranges from 0.50 to 1.00. At that capacity the objective places no constraint on c .

The recoverability of the input from the representation is thus not a monotone function of predictive performance, even without noise, optimisation or architectural choices. Changes in capacity reorganise what a predictor retains. Interpreting probe results as a measure of how well a model predicts assumes a monotonicity that fails in the smallest system in which it can be checked.

Tolerance and the required–permitted distinction

At exact optimality with four states the classification is clear: a and b are required, each recovered at 1.00 by every optimal encoder, and c is absent, recovered at 0.50.

With a tolerance of 0.05 bits the classification disappears. At four states 26 encoders fall within that tolerance, and across them the recovery of a ranges from 0.75 to 1.00, of b from 0.75 to 1.00, and of c from 0.50 to 0.75. All three bits become permitted, and no bit is required or absent. The requirement on a first fails at a tolerance of 0.022015 bits, 3.6 percent of the optimal loss.

A sufficiently good predictor must therefore represent a , but a predictor within a few hundredths of a bit of optimal need not, and trained systems do not reach the exact optimum. Claims that prediction requires a world model therefore need a stated tolerance, on which the conclusion largely depends. Computational mechanics states its results for exact optimal prediction, which suits a theory but not a claim about a trained system (Shalizi and Crutchfield 2001).

The cost of retaining a capability can be computed in the same way. At four states, retaining the irrelevant bit with full recovery costs 0.111696 bits of excess loss, about 18 percent above the optimum. At eight states it costs nothing, since the identity encoder is among the optima. Whether a capability is costly, free or required therefore depends jointly on the objective and the budget, and the same capability moves between these categories as the budget changes.

Recovery from the state and through the output

The predictor exposes only its predicted probability, and a capability the system can use must be recoverable through that output. Recomputing recovery for a decoder restricted to thresholding the state's predicted probability separates two quantities that probing studies usually report as one.

At four states the second predictive bit b is fully recovered by an unrestricted decoder, at 1.00, and at 0.75 through the output. The four states have predicted probabilities 0.05, 0.35, 0.65 and 0.95, and the states with $b = 1$ are the second and fourth, which no single threshold separates from the others.

At full capacity the difference is larger. In the encoders that retain the irrelevant bit, an unrestricted decoder recovers it perfectly, while recovery through the output is 0.50, chance level. A probe

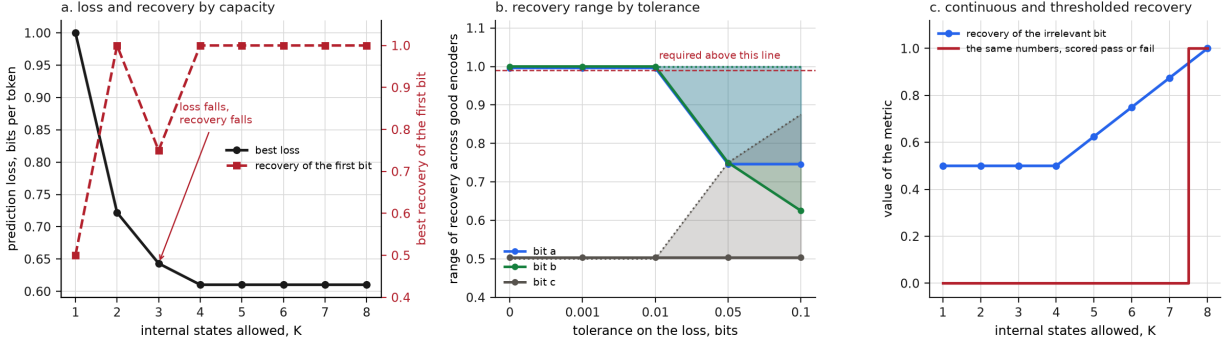


Figure 1: Exhaustive enumeration of 4140 encoders. (a) The best achievable loss at each capacity against the best achievable recovery of the most predictive input bit; at three states the two move in opposite directions. (b) The range of recovery across all encoders within a given tolerance of the optimum at four states; no capability is required or absent at 0.05 bits. (c) The best recovery of the irrelevant bit at each capacity, and the same values scored as pass or fail at a threshold of 0.95.

would report that the model represents c , but the model cannot use c , because its only output is a probability that c does not affect.

Probing therefore measures whether a representation contains information, which differs from whether the system can use that information, and in this construction the first is 1.00 and the second 0.50. Studies that test capabilities by intervention, in addition to decoding, address this difference (Li et al. 2022), and the observation that a particular linear encoding is not mathematically required even when the information is present concerns the same distinction (Shai et al. 2024).

Prediction horizon

The preceding sections vary capacity with the objective fixed. This section fixes capacity and varies only how far ahead the system must predict.

Four latent states, A , B , C and D , each emit one bit. From A and B the emitted bit is uniform, so the one-step predictive distribution is identical from both. A always moves to C and B to D , and C and D emit zero with probabilities 0.9 and 0.1 respectively. Whether the process is in A or in B therefore does not affect the next-symbol distribution and is fully determined by the distribution over the next two symbols.

At a one-symbol horizon the four states have 3 distinct predictive signatures and the latent question is recovered at 0.75, by separating C from D and guessing between A and B . At a two-symbol horizon there are 4 distinct signatures and the latent question is recovered at 1.00. A third symbol adds nothing. By predictive sufficiency, histories with the same signature must receive the same answer, and at one step the states A and B share a signature.

The objective, the generator and the capacity are unchanged across these cases. Only the prediction horizon differs, and that change alone makes a latent distinction required where it was previously

absent. For language models, the relevant form of a next-token objective is the joint distribution over a long context, which differs from the one-step conditional in the same way.

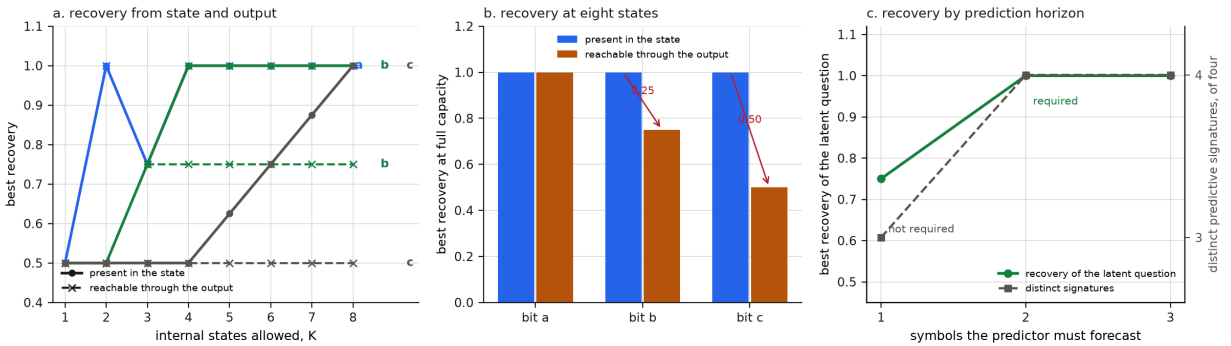


Figure 2: The interface and the horizon. (a) Recovery of each input bit from the internal state against recovery through the predictor’s own output, at each capacity. (b) Recovery at full capacity: the irrelevant bit is fully recoverable from the state and at chance through the output. (c) The latent question in the hidden-Markov construction, not required at a one-symbol horizon and required at two, with the count of distinct predictive signatures alongside.

Implications and limitations

The results are properties of a constructed system and are not evidence about any trained network. The construction was kept small enough that no effect specific to trained networks could enter it. Its use is to test general claims, which does not require scale.

The results contradict the claim that better performance on a prediction objective implies a richer internal representation, since a single increase in capacity in a noiseless three-bit system produces the opposite. They contradict the claim that an objective requires a capability without qualification, since here a capability is required only within 0.022015 bits of the exact optimum. And they contradict the inference from decodability to capability, since a fully decodable bit is at chance through the output.

They also bear on sharp capability onsets. The best recovery of the irrelevant bit rises from 0.50 to 1.00 in four equal increments of 0.125 as capacity increases from four to eight states. Measured continuously it rises linearly; scored as whether recovery exceeds 0.95 it is 0 for seven capacities and 1 at the eighth. The system has no noise, optimiser or benchmark effects, so the step is produced entirely by the choice of metric. This supports the argument of Schaeffer et al. (2023) in a setting where alternative explanations are excluded by construction, although it does not show that any particular reported onset in a large model has this cause.

The results do not establish how often these phenomena occur at scale, which depends on training procedures and data. Extrapolation would require that trained systems lie close enough to their achievable optimum for the tolerance argument to apply, that their output interfaces are as restrictive relative to their internal states as a scalar threshold is here, and that the relevant capacity limits are as sharp. None of these conditions is known to hold, and all three can be measured.

The question belongs to a wider study of which capabilities a generative process requires, permits or excludes, in both learning and evolution. In biology, a function that outlasts the context that selected it is called an exaptation (Gould and Vrba 1982), and digital evolution has documented many adaptations that no one specified (Lehman et al. 2020; Lenski et al. 2003; Agüera y Arcas et al. 2024), alongside work on the definition of open-endedness (Hughes et al. 2024). On the computational side the relevant results are the constraint arguments of computational mechanics, the information bottleneck and work on emergence (Shalizi and Crutchfield 2001; Tishby et al. 2000; Rosas et al. 2024). An open question is whether the necessity, contingency and reachability of a secondary function can be predicted from the structure of the generating process before the process is run.

The results provide a practical check. A claim that a system must have acquired a capability because it predicts well should state the excess loss it tolerates, the interface through which the capability would be used, and the horizon over which prediction is scored, since in this construction each of the three alone can reverse the conclusion. Instruction-tuned systems have also been trained directly toward behaviours later described as consequences of prediction (Ouyang et al. 2022), and this must be separated as well before such a claim can be evaluated.

References

- Agüera y Arcas, Blaise, Jyrki Alakuijala, James Evans, Ben Laurie, et al. 2024. “Computational Life: How Well-Formed, Self-Replicating Programs Emerge from Simple Interaction.” In *arXiv*, 2406.19108. Preprint.
- Anderson, P. W. 1972. “More Is Different.” *Science* 177 (4047): 393–96. <https://doi.org/10.1126/science.177.4047.393>.
- Crutchfield, James P., and Karl Young. 1989. “Inferring Statistical Complexity.” *Physical Review Letters* 63 (2): 105–8. <https://doi.org/10.1103/PhysRevLett.63.105>.
- D’Amour, Alexander, Katherine Heller, Dan Moldovan, et al. 2020. “Underspecification Presents Challenges for Credibility in Modern Machine Learning.” In *arXiv*, 2011.03395. Preprint.
- Gould, Stephen Jay, and Elisabeth S. Vrba. 1982. “Exaptation: A Missing Term in the Science of Form.” *Paleobiology* 8 (1): 4–15. <https://doi.org/10.1017/S0094837300004310>.
- Hughes, Edward, Michael D. Dennis, Jack Parker-Holder, et al. 2024. “Open-Endedness Is Essential for Artificial Superhuman Intelligence.” *Proceedings of the 41st International Conference on Machine Learning*.
- Lehman, Joel, Jeff Clune, Dusan Misevic, et al. 2020. “The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities.” *Artificial Life* 26 (2): 274–306. https://doi.org/10.1162/artl_a_00319.

- Lenski, Richard E., Charles Ofria, Robert T. Pennock, and Christoph Adami. 2003. "The Evolutionary Origin of Complex Features." *Nature* 423 (6936): 139–44. <https://doi.org/10.1038/nature01568>.
- Li, Kenneth, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task." In *arXiv*, 2210.13382. Preprint.
- Ouyang, Long, Jeff Wu, Xu Jiang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." In *arXiv*, 2203.02155. Preprint.
- Rosas, Fernando E., Bernhard C. Geiger, Andrea I. Luppi, Anil K. Seth, Daniel Polani, and Michael Gastpar. 2024. "Software in the Natural World: A Computational Approach to Hierarchical Emergence." In *arXiv*, 2402.09090. Preprint.
- Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023. "Are Emergent Abilities of Large Language Models a Mirage?" In *arXiv*, 2304.15004. Preprint.
- Semenova, Lesia, Cynthia Rudin, and Ronald Parr. 2022. "On the Existence of Simpler Machine Learning Models." 2022 *ACM Conference on Fairness, Accountability, and Transparency*, 1827–58. <https://doi.org/10.1145/3531146.3533232>.
- Shai, Adam S., Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. 2024. "Transformers Represent Belief State Geometry in Their Residual Stream." In *arXiv*, 2405.15943. Preprint.
- Shalizi, Cosma Rohilla, and James P. Crutchfield. 2001. "Computational Mechanics: Pattern and Prediction, Structure and Simplicity." *Journal of Statistical Physics* 104 (3): 817–79. <https://doi.org/10.1023/A:1010388907793>.
- Tishby, Naftali, Fernando C. Pereira, and William Bialek. 2000. "The Information Bottleneck Method." In *arXiv*, Physics/0004057. Preprint.
- Turing, Alan Mathison. 1952. "The Chemical Basis of Morphogenesis." *Philosophical Transactions of the Royal Society of London. Series B* 237 (641): 37–72. <https://doi.org/10.1098/rstb.1952.0012>.
- Wei, Jason, Yi Tay, Rishi Bommasani, et al. 2022. "Emergent Abilities of Large Language Models." In *arXiv*, 2206.07682. Preprint.