Milestone 013 — Accessible does not imply used

Tyler J.H.G.

Abstract

A frozen-representation probe found a natural-source distinction that the model's native consumer barely used.

Research period: September 8–9, 2026
Public article source revision: September 29, 2026
Historical anchor: #328
Primary source: Milestone notebook 013

1. The question

I had a compact unit-hypersphere state and a revision-branch task that required a specific distinction. The next question was where the predictive failure occurred: had the state lost the distinction, or did the model's own consumer fail to use information still present in it?

Those cases require different changes. If the distinction has been erased, changing the final readout alone cannot recover it. If a small readout can recover it from a frozen state, we have a reason to investigate how the native consumer turns that state into a prediction.

I froze the representation and compared the frozen native consumer with a separately trained affine-softmax probe. Keeping the representation fixed made the external readout a test of recoverability under a deliberately small consumer family.

2. What the probe measured

For a frozen hidden state h, the probe computes logits

$$z = Wh + b, \qquad p(y \mid h) = \mathrm{softmax}(z). \tag{1}$$

This is an affine map followed by a probability readout. It can combine and reweight existing coordinates, but it adds no nonlinear representation layer. Success on held-out examples therefore supports a narrow statement: the tested natural-source branch distinction was accessible to this simple external consumer.

The reported quantity is evaluation log-loss gain over a constant baseline,

$$G = L_{\text{constant}} - L_{\text{consumer}}, \tag{2}$$

so positive values indicate lower evaluation loss than the baseline. The source notebook reports the following comparison:

Published Sep 29, 2026

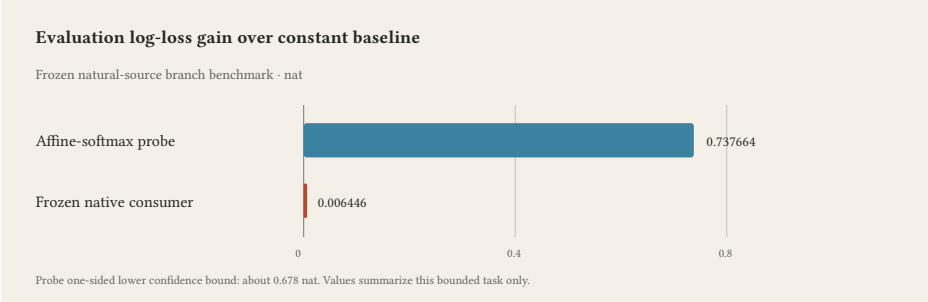


Figure 1: On the frozen natural-source branch benchmark, a small external readout recovered substantially more of the tested distinction than the model's own consumer.

Consumer	Evaluation log-loss gain over constant baseline	Readout
Affine-softmax probe	0.737664 nat; one-sided lower confidence bound about 0.678 nat	External, separately trained probe on the frozen representation
Frozen native consumer	0.006446 nat	The model's original downstream consumer on the same bounded branch task

The tangent-coordinate analysis reached the same qualitative conclusion, so the probe did not depend on an obvious radial coordinate. Selectivity controls also passed. The result was classified as **ACCESSIBLE-BUT-UNDERUSED**.

3. Inspect the geometry

The representation is constrained to a unit hypersphere. A point on that sphere has a local tangent space, and a probe selects a direction along which it reads out variation. The toy plot below shows this relationship in two dimensions.

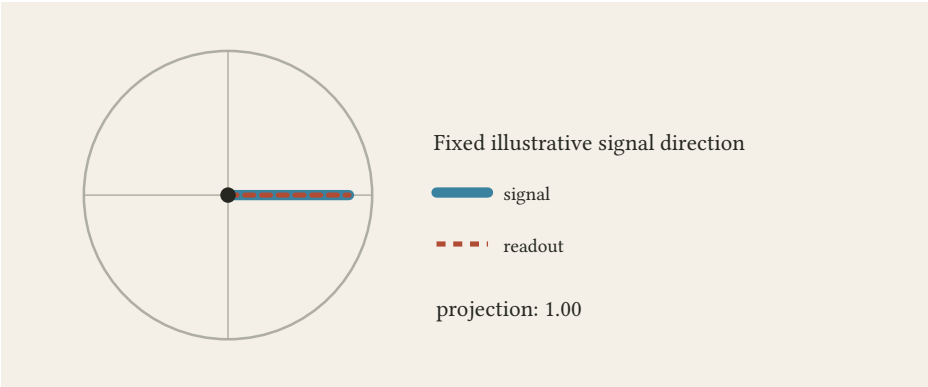


Figure 2: In a 2D slice, the signal and readout directions align, so the projection is 1.00. When JavaScript is available, a slider appears below the figure so readers can rotate the readout and inspect the projection.

For a unit vector h, the tangent projector is $P_h = I - hh^\top$. More generally, for nonzero h,

$$P_h = I - \frac{hh^\top}{\| h \|^2}. \tag{3}$$

This projects onto directions tangent to the sphere at h. The frozen-model analysis tested what those directions carried empirically. **Formal checkpoint — LIB-SPH-002 / FRM-000020:** the derivative of nonzero normalization annihilates radial perturbations and scales tangent perturbations by inverse radius. This is the local geometry used to formulate the empirical control. See the formal reference and the theorem source at the reviewed revision `14ff466c4dabb3a9a09717d65236832b7ff3a26e`.

4. Why a successful probe is not enough

A probe needs controls because several procedures can produce an impressive score. It might recover the intended distinction, memorize labels, exploit source leakage, or succeed on accidental structure. The score alone leaves those explanations open.

Control-task selectivity asks whether the probe succeeds on the intended structure and fails when that structure is destroyed. Information-theoretic and minimum-description-length approaches address related questions about recoverable information and the cost of encoding the probe's solution (Hewitt & Liang, 2019; Pimentel et al., 2020; Voita & Titov, 2020).

The analysis used label shuffles, representation shuffles, matched random features, random rotations, fixed page groups, and learning-efficiency checks. These challenged accidental label structure, memorization, coordinate artifacts, source leakage, and a procedure that would find apparent signal almost anywhere. The tangent-coordinate comparison addressed the radial-versus-tangent explanation raised by the earlier sphere work.

Together, the controls support recovery of the tested branch distinction through the small external readout.

5. Try a small example

A toy set of six states carries a class signal along the first coordinate. Change `probe_degrees` to rotate the readout and see how classification changes.

```
import math

states = [
    (0.9, 0.1), (0.7, -0.3), (0.6, 0.2),
    (-0.9, -0.1), (-0.7, 0.3), (-0.6, -0.2),
]
labels = [1, 1, 1, -1, -1, -1]
probe_degrees = 0
angle = math.radians(probe_degrees)
```

```
weights = (math.cos(angle), math.sin(angle))
correct = sum(
    ((1 if state[0] * weights[0] + state[1] * weights[1] >= 0 else -1)
== label)
    for state, label in zip(states, labels)
)
print(f"synthetic teaching accuracy: {correct / len(labels):.0%}")
```

Saved output: with `probe_degrees = 0`, all six toy states are classified correctly (100%).

Run the cell in your browser to try other angles. New output appears under **Your session**.

6. Interpretation

The probe recovered the tested branch distinction from frozen states, while the native consumer barely used it. That gives the investigation a concrete place to intervene: keep the state fixed and change the consumer.

The comparison is specific to this task and probe family. A binary revision-branch decision is easier than 256-way next-byte prediction, and external recoverability leaves causal use by the native model unresolved. I need to test whether the recovered signal improves the full predictive distribution.

The public notebook records the benchmark analysis. An exact frozen-model replay package is not currently published.

7. The next experiment

Keep the representation frozen and train a ladder of very small consumers on the 256-way next-byte task. How much of the signal recovered by the probe can those consumers turn into useful full-distribution prediction?

8. Sources and chronology

- Milestone notebook 013 — primary historical account
- Milestone 012 — Natural-source task distinctions
- Milestone 008 — What does the sphere actually do?
- Visual Intuition Atlas — probe recoverability and normalization geometry
- Representation and Readout reference
- Research chronology

9. Publication map

See the Research Notes publication dispositions for the complete article/source classification.

References

Hewitt, J., & Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2733–2743. https://doi.org/10.18653/v1/D19-1275

Pimentel, T., Valvoda, J., Maudslay, R. H., Zmigrod, R., Williams, A., & Cotterell, R. (2020). Information-Theoretic Probing for Linguistic Structure. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4609–4622. https://doi.org/10.18653/v1/2020.acl-main.420

Voita, E., & Titov, I. (2020). Information-Theoretic Probing with Minimum Description Length. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 183–196. https://doi.org/10.18653/v1/2020.emnlp-main.14