Deep Minds and Shallow Probes

Which probe families are principled, and what should be aligned to move probes across models?

Su Hyeong Lee[1] **Risi Kondor**[1,2]

[1]Department of Statistics, [2]Department of Computer Science, University of Chicago

sulee@uchicago.edu

NeurIPS 2026

arXiv: 2605.11448

PART I · WHICH PROBES ARE PRINCIPLED?

1 Same behavior ⇒ invertible affine map between final-layer representations → **2** Shallow + symmetric ⇒ polynomial probes; linear is the degree-1 case → **3** Go beyond linear with coordinate-stable low-rank (CP) probes

PART II · HOW DO PROBES MOVE BETWEEN MODELS?

4 Probes see only the quotient $H/K(V)$, shared across equivalent models → **5** Align quotients: zero-label monitor transfer with a coverage signal

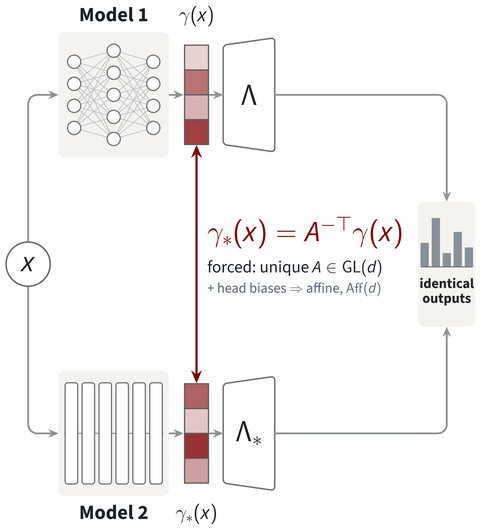
1 Readout symmetry is affine

A probe is a lightweight readout trained on frozen activations. Linear probes are the default because they are **shallow**: success should reflect the representation, not the probe.

- But hidden coordinates are **not unique**: two models can realize the same computation in different coordinates.
- A probe meant to reveal structure *in* the representation should not depend on an arbitrary choice of basis.

Thought experiment

Two LLMs give identical outputs on every input, yet differ arbitrarily in architecture, initialization, and weights. What must their representations share?



Theorem 2.1 (Linear-head equivalence). Let $\gamma, \gamma_* : D \to \mathbb{R}^d$ be final hidden representations and $\Lambda, \Lambda_* \in \mathbb{R}^{n\times d}$ bias-free readouts, with $\gamma(D)$ spanning $\mathbb{R}^d$ and both heads of rank d. If $\Lambda_*\gamma_*(x) = \Lambda\gamma(x)$ for all $x \in D$, there is a **unique** $A \in \mathrm{GL}(d)$ with

$$\gamma_*(x) = A^{-\top}\gamma(x), \qquad \Lambda_* = \Lambda A^\top.$$

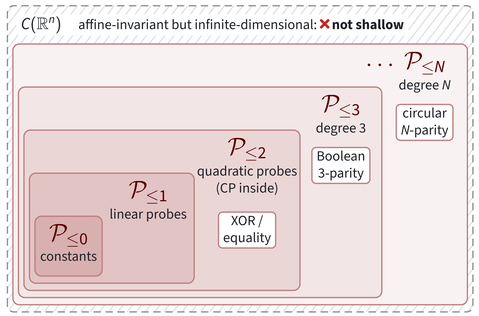
- **Softmax heads:** also a shared shift, $\Lambda_* = \Lambda A^\top + \mathbf{1}c^\top$.
- **Heads with biases:** $h \mapsto A^{-\top}h + b$, the full affine group $\mathrm{Aff}(d)$.

2 Shallow + symmetric = polynomial

Theorem 2.3 (Classification). Let $V \subset C(\mathbb{R}^n)$ be a finite-dimensional linear space with $f \in V$, $g \in \mathrm{Aff}(n) \Rightarrow f \circ g \in V$. Then $V = \{0\}$ or

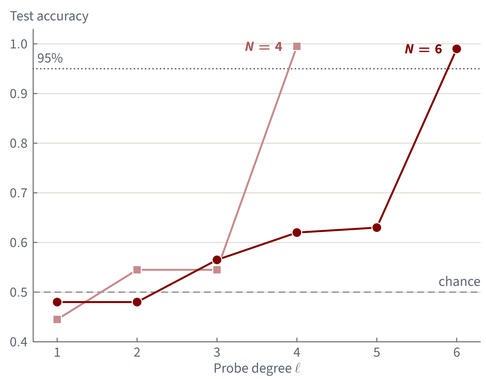
$$V = \mathcal{P}_{\le \ell}(\mathbb{R}^n) \quad \text{for some } \ell \ge 0,$$

the polynomials of total degree $\le \ell$; conversely, each qualifies.



Linear probes are not just an engineering default: they are the degree-1 member of a unique hierarchy.

Degree lower bounds are tight



3 Beyond linear, coordinate-stably

A full degree-ℓ probe with k outputs has $k\binom{n+\ell}{\ell}$ coefficients: over 13 million quadratic features at $n = 5120$ (Qwen-14B). Which structured families respect the symmetry?

Proposition 2.6 (Low-rank structure is stable). Rank-R CP probes $p(x) = \sum_{r=1}^{R} \alpha_r \prod_{j=1}^{m} \langle v_{r,j}, x\rangle$ are closed under precomposition by $\mathrm{GL}(n)$; monomial sparsity is not.

Rank-1 CP $\langle v, x\rangle^2 \xrightarrow{x \mapsto Ax,\ A \in \mathrm{GL}(n)} \langle A^\top v, x\rangle^2$ ✔ still rank 1

1-sparse monomial $x_1^2 \xrightarrow{x \mapsto Ax,\ A = \left[\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right]} x_1^2 + 2x_1x_2 + x_2^2$ ✕ 3 monomials: sparsity lost

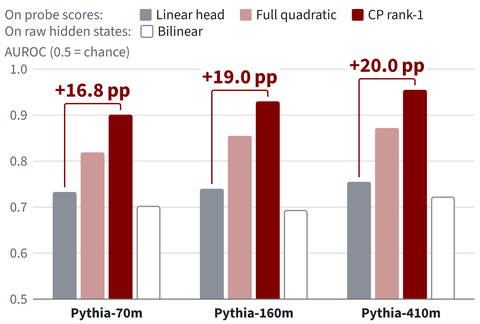
Affine-completed CP (used in the experiments)

$$f(x) = \sum_{r=1}^{R} \alpha_r \big(\langle u_r, x\rangle + u_{r,0}\big)\big(\langle v_r, x\rangle + v_{r,0}\big) + \langle w, x\rangle + w_0$$

is closed under all of $\mathrm{Aff}(n)$.

When the target is degree 2

Subject–verb agreement multiplies number at the subject by number at the verb. Per-token linear probes give a 60-d score vector; a polynomial head on probe scores is itself a polynomial probe.



CP rank-1 beats the linear head by **16.8–20.0 pp AUROC**; full quadratics overfit (train AUROC $\ge$ 0.995).

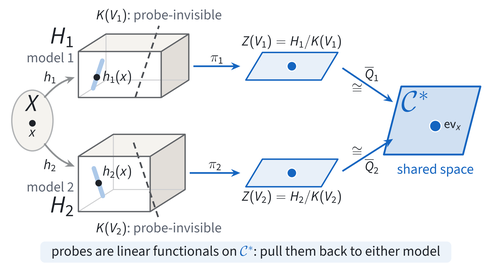
Safety analogue (AUROC). Policy mismatch = prompt risk XOR response refusal: CP rank-1 **0.976**, full quadratic 0.974, linear 0.906.

4 A shared space for transfer

Model i has a representation $h_i : X \to H_i$ and linear probes $V_i \subset H_i^*$, realizing the concepts $\mathcal{C}_i = \{\ell \circ h_i : \ell \in V_i\}$. Probes are blind to $K(V) = \bigcap_{\ell \in V} \ker \ell$ and see only the quotient $Z(V) = H/K(V)$.

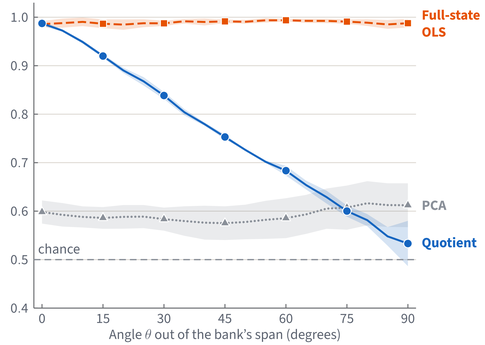
Theorem 3.3 (Common abstract probe space). Suppose both models realize the same finite-dimensional concept family $\mathcal{C}$, with injective probe-to-concept maps. Then their probe-visible spaces are canonically isomorphic:

$$H_1/K(V_1) \;\cong\; \mathcal{C}^* \;\cong\; H_2/K(V_2).$$



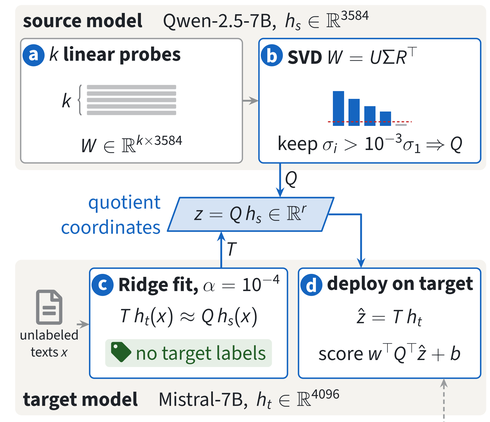
Transfer need not align full hidden states, only the probe-visible quotients (Cor. 3.5).

Synthetic verification



As a held-out concept rotates out of the bank's span (nuisance rank 56), quotient accuracy (blue) falls smoothly from 0.987 to 0.533 (near chance), tracking the lost coverage; full-state OLS stays near 0.99 and gives no sign of the gap. Bands: ±1 std over 5 seeds.

5 Monitors that travel



Safety monitors, zero target labels

AUROC (.5 – 1)	Source	Same family (Qwen)			Cross-arch.	Control
	Qwen-7B	3B	14B	Coder	Mistral	Random
Toxicity	.976	.952	.955	.971	**.926**	.580
Jailbreaking	.974	.932	.926	.838	**.669**	.522
Harmful (BT)	.791	.772	.781	.793	**.773**	.632
Sentiment	.975	.966	.974	.967	**.972**	.511
Moderation	.890	.836	.857	.867	**.820**	.512

On toxicity and sentiment, transfer beats or comes within 0.3 pp of probes trained on 500 target labels.