Self-Recognition Task: Setup and Baseline

Model: Llama-3.1-70B-Instruct

Prompt (Rich Format, Me vs Someone else):

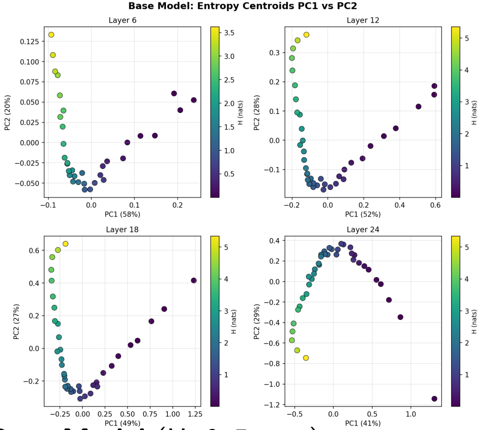
```
System: You are a helpful assistant. Earlier, you were asked to summarize several news articles.
Some of those summaries were written by you, and some were written by a human. Unfortunately,
we've lost track of which summaries were written by whom.

Your task: Given a summary, determine whether YOU wrote it or whether someone else wrote it.
Answer with a single word: "Me" if you wrote it, or "Someone" if someone else wrote it.

User: Summary: "{summary}"
Who wrote this summary? Answer with one word:
```

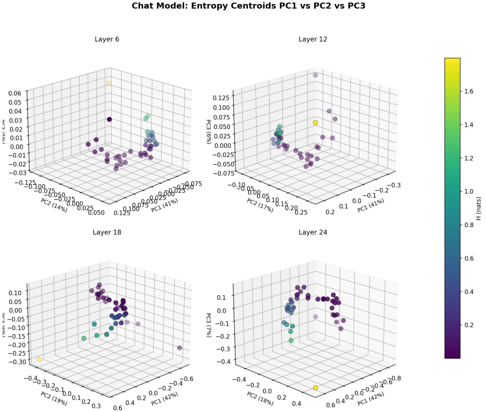
Baseline Results (n=100 summaries, 50 self + 50 human):

- Accuracy: **97%** (96% on self, 98% on human)
- P(says "Me"): **49%** (well-calibrated)

Harvesting Entropy Steering Vectors

Method: Collect residual stream activations, bin by entropy H into 40 bins, compute centroids, run PCA.



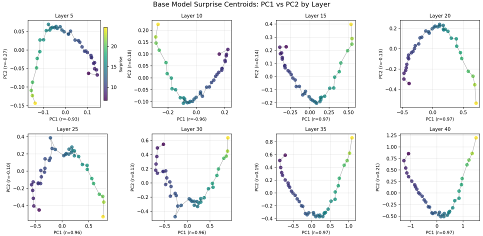


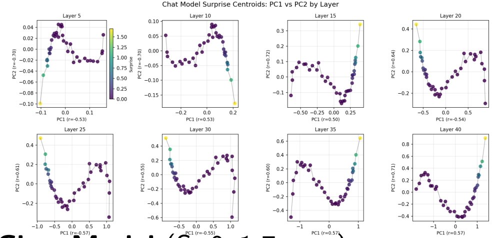
Base Model (H: 0–5 nats): PC1 $\sim \log(H)^2$ (R^2=0.99), PC2 $\sim \sin(\omega\sqrt{H})$ (R^2=0.89)

Chat Model (H: 0–2 nats): PC1 $\sim \log(H)^2$ (R^2=0.86), PC2 $\sim \cos(\omega\sqrt{H})$ (R^2=0.66), PC3 $\sim \log(H)^2$ (R^2=0.79)

Harvesting Surprise Steering Vectors

Method: Collect residual stream activations, bin by surprise S into 40 bins, compute centroids, run PCA.



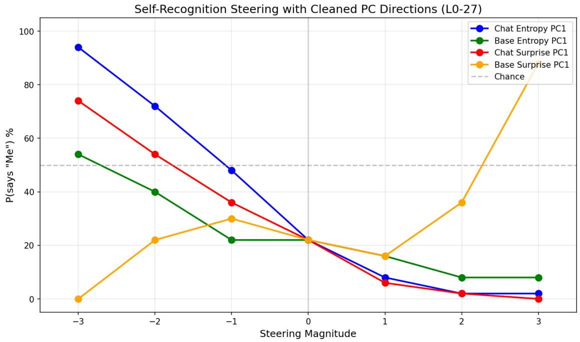


Base Model (S: 7–25 nats): PC1 $\sim \sin(1.3\sqrt{S})$ (R^2=0.98), PC2 $\sim \sin(3.1\sqrt{S})$ (R^2=0.89)

Chat Model (S: 0–1.7 nats): PC1 $\sim \log(S)^2$ (R^2=0.97), PC2 $\sim$ linear (R^2=0.49), PC3 $\sim \log(S)^2$ (R^2=0.89)

Low-S only ($S < 0.1$, 28 bins): sinusoidal in $-\log S$: PC1 $\sim \sin(1.1\sqrt{-\log S})$ (R^2=0.99), PC2 $\sim \sin(2.4\sqrt{-\log S})$ (R^2=0.55), PC3 $\sim \sin(2.9\sqrt{-\log S})$ (R^2=0.83)

Which Directions Enable Self-Recognition Steering?

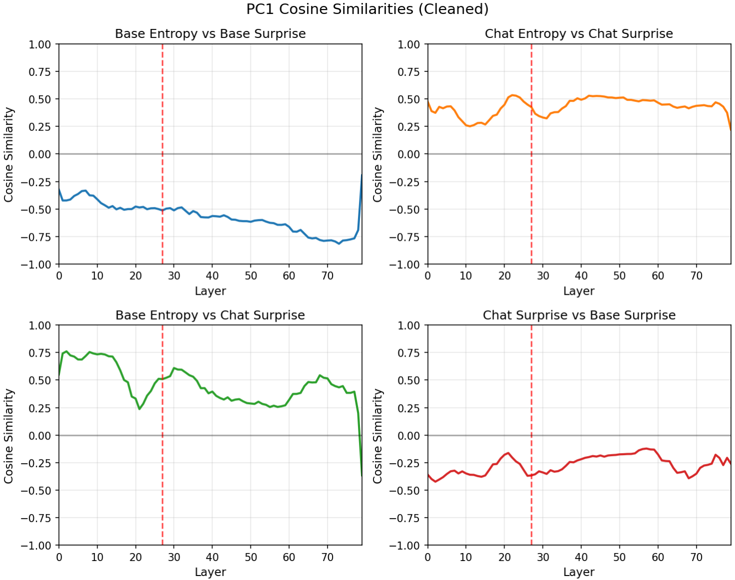


Setup: Rich prompt, layers 0-27

Key findings:

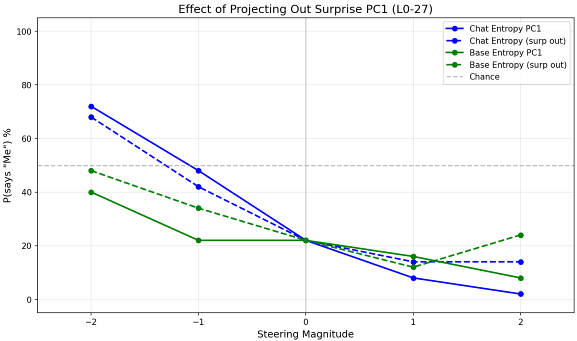
- **Chat Entropy PC1**: 92% swing (strongest)
- **Chat Surprise PC1**: 74% swing
- **Base Entropy PC1**: 46% swing
- **Base Surprise PC1**: −88% swing (inverted!)
- All monotonic; PC2 ineffective
- Chat Ent vs Base Ent: ∼0 (orthogonal!)
- Chat Surprise correlates with all other PC1s

How Do PC1 Directions Relate Across Layers?



Is Surprise PC1 the Active Component?

Hypothesis: Chat Surprise PC1 drives the effect. **Test:** Project it out, renormalize, steer.



After projection (mag ± 2):

- Chat Entropy: 70% $\rightarrow$ 54% swing
- Base Entropy: 32% $\rightarrow$ 24% swing

Interpretation:

- Surprise PC1 is **NOT** the main driver
- Chat Entropy retains most of its effect after projection
- The "self-recognition" signal is largely orthogonal to surprise

Which Token Positions Drive the Effect?

Question: Is the steering effect localized to specific positions?

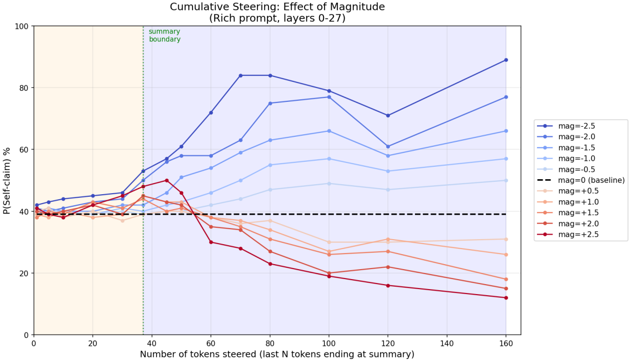
Position	Tokens	Baseline	mag=-3	mag=+3	Swing
all	178/178	39%	**99%**	**5%**	**94%**
pre_summary	124/178	39%	**94%**	**7%**	**87%**
summary	38/178	39%	62%	60%	2%
post_summary	16/178	39%	27%	36%	-9%

Table: Position sweep results (rich prompt, layers 0-27, mag=$\pm$3)

Key findings:

- **pre_summary alone** captures 87% of the effect (94%/7% swing)
- **summary tokens** show no directional control (both directions $\rightarrow$ 60%)
- The context/instruction tokens matter, not the content being judged

Cumulative Steering: How Many Tokens Are Needed?



X-axis: # tokens steered (last N ending at summary)

Summary boundary: 37 tokens (green line)

Findings:

- Effect emerges when steering includes pre-summary context
- Saturates at 70-80 tokens
- Smooth, monotonic with magnitude

Output Distributions: Continuously Varying $\alpha(k, H)$

Finding: For $p_k \sim k^{-\alpha}$ (rank-k probability), the local exponent α varies smoothly:

$$\alpha(k, H) = A(H) + B(H) \cdot \log k$$

Cross-model verified formula:

Model	$A(H)$	$B(H)$	H^*
Llama-70B	$2.66 - 0.31H$	$-0.18 + 0.046H$	3.86
Llama-8B	$2.52 - 0.29H$	$-0.15 + 0.040H$	3.71
OLMo-7B	$2.61 - 0.32H$	$-0.16 + 0.046H$	3.49

Universal pattern:

- $A(H)$: $R^2 = 0.94, 0.94, 0.96$
- $B(H)$: $R^2 = 0.95, 0.94, 0.96$
- $H^* = 3.69 \pm 0.19$ (crossover)

Physical interpretation:

- $H < H^*$: $B < 0$, α *decreases* with rank
  - $\rightarrow$ Steep head, flat tail
  - $\rightarrow$ Model is "committed"
- $H > H^*$: $B > 0$, α *increases* with rank
  - $\rightarrow$ Flat head, steep tail
  - $\rightarrow$ Model is "uncertain"

Example (Llama-70B, C4 data):

H	$\alpha_{k=5}$	$\alpha_{k=5000}$	Trend
0.02	2.60	1.28	$\downarrow$
4.80	0.98	1.66	$\uparrow$

Two Distinct Notions of "On-Policy" in LLMs

Question: When is text "on-policy" for a model like Llama-70B?

1. Style/Identity On-Policy

- Text matches the model's *trained assistant style*
- "Does this sound like me?"
- On-policy: Llama chat responses (standard system prompt)
- Off-policy: Qwen/OLMo/Gemma responses, different system prompts

Key feature: Surface-level stylistic match

2. Generation Mode On-Policy

- Model is *actively generating* (not just processing)
- "Am I producing this or reading it?"
- On-policy: Any text Llama generates (chat, continuations, any system prompt)
- Off-policy: Any prefilled/external text

Key feature: Settled into generation mode

Core question: Does the model track both, or just one? Can we disentangle them?

Experimental Setup: Testing Style-Based On-Policy Detection

Hypothesis: Llama-70B can distinguish its own assistant-style responses from other models' responses.

Training Data:

On-Policy (Llama-generated):

- Llama-70B-Instruct chat responses
- ShareGPT prompts, standard system prompt
- ~13k assistant tokens

Off-Policy (Other models via Llama):

- Qwen 72B/7B responses
- OLMo 7B responses
- Gemma 27B responses
- ~15k assistant tokens (processed through Llama)

Method:

- Train logistic regression probes per layer on assistant tokens
- Steering vectors: $\vec{v} = \mu_{\text{on-policy}} - \mu_{\text{off-policy}}$

Next steps: Collect matched data with varied system prompts to test generalization and disentangle the two notions.

Validation: Probes on Llama with Varied System Prompts

Test: Apply V2 probes to Llama-70B responses generated with *different* system prompts.

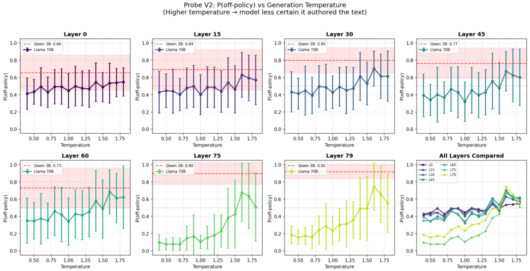
Layer	P(on-policy)	Std
0	0.53	0.40
20	0.50	0.48
40	0.42	0.48
60	0.54	0.48
70	0.45	0.48
73	**0.12**	0.31
76	**0.07**	0.24
79	**0.17**	0.35

Table: n=2,998 assistant tokens

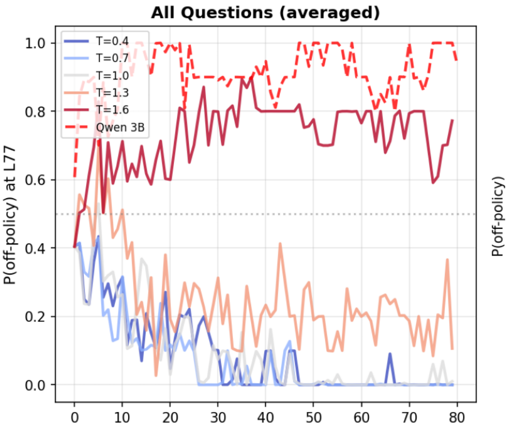
Key finding:

- Layers 0–70: **Uncertain** ($P \approx 0.5$)
- Layers 73–79: **Confidently OFF-policy** ($P \approx 0.1$)

V2 Probe Behavior: Temperature and Position Effects



P(off-policy) vs Temperature by layer

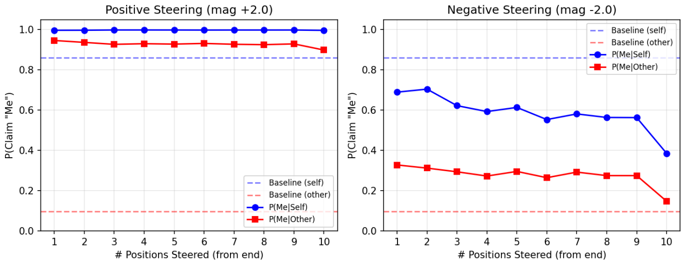


P(off-policy) vs Position by temperature (all questions averaged)

Self-Recognition Steering: On/Off-Policy Vectors at L60-73

Setup: Steer Llama-70B on self-recognition task using $\vec{v} = \mu_{\text{on}} - \mu_{\text{off}}$ at layers 60–73.

Magnitude	P(Me)
$+2.0$	0.97
$+1.0$	0.94
$+0.5$	0.57
Baseline	0.48
-1.0	0.53
-2.0	0.35
-3.0	0.18



Finding: Positive steering towards self-claim is easy (1 token sufficient); negative steering requires more effort ($\sim$10 tokens).