Hallucination-Induction, Not Calibration: When Multi-Feature SAE Steering Looks Like It Works

Anonymous Authors[1]

Abstract

We replicate Ferrando et al.'s (2024) entity-recognition SAE feature pattern on a 27B reasoning-tuned model (Qwen3.6-27B) using a paper-grade Top-K Sparse Autoencoder, then test whether the feature is causally usable for hallucination calibration. The single best SAE latent at layer 31 reaches AUROC 0.814 (95% bootstrap CI $[0.731, 0.898]$) on cross-type known-vs-unknown classification, with the entity-recognition signal peaking in the mid-stack as Ferrando reported on Gemma-2-2B-IT (peak around layer 9 in their 26-layer model). The SAE latent is statistically indistinguishable from a layer-32 L2 logistic regression probe (0.887 $[0.823, 0.941]$) and a diff-of-means probe (0.859): the contribution is interpretability, not raw classification performance. Single-feature steering produces a null effect on refusal calibration. Multi-feature top-K ablation at small K=200 (0.3% of the dictionary) produces a 4–8σ effect against a random-K null — in contrast with prior large-K work where the random control nullifies the signal (Al-Qurashi, 2025) — but a Claude-as-judge correctness audit reveals the effect is *induced confabulation*: known-entity incorrect-answer rate rises from 62% to 77% while correct-answer rate drops from 8% to 0%. Statistical significance against random-K is necessary but not sufficient evidence for a calibration mechanism: the intervention identifies a hallucination-induction circuit, not a calibration knob.

1. Introduction

Sparse Autoencoders (SAEs) trained on transformer residual streams have produced a growing catalogue of monosemantic features with predictive structure aligned to model behaviours of interest (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2024). Ferrando et al. (2025) identified SAE latents in Gemma-2-2B-IT that fire selectively on entities the model can answer attribute questions about (**known**) versus those it cannot (**unknown**); their separation-score analysis shows the signal peaks in the middle layers (around layer 9 of 26) and they report a related uncertainty-axis AUROC of 0.732 at the model's answer token. The finding has motivated a research program asking whether such features are causally usable for hallucination calibration: can ablating "I do not know" features make the model less likely to confabulate? Concurrent work by Al-Qurashi (2025) ran a large-K (K=4697 entity-specific features in Gemma-2-2B) ablation and reported that ablating an equal number of *random* features had nearly identical effect (ratio $\approx$ 1.0), concluding that large-K ablation is generic disruption, not entity-specific.

This paper makes three contributions.

(1) Replication at 27B with proper baselines. We train a paper-grade Top-K + AuxK SAE on Qwen3.6-27B at layers 11/31/55 (200M tokens, k=128, dictionary size 65,536 per layer), apply the Ferrando attribute-recall protocol on real Wikidata entities with a Pile noise filter, and report AUROC for the best single SAE latent alongside two supervised baselines (L2 logistic regression, diff-of-means) at every layer, all with 95% bootstrap confidence intervals.

(2) Small-K multi-feature ablation passes the random-K control where large-K does not. Single-feature steering at L=11 produces a null effect on refusal calibration. Ablating the top-K=200 SAE latents at L=11 by Cohen's $|d|$ (0.3% of the dictionary) drops the unknown-entity refusal rate by 16.7pp, 4–8σ outside a random-K null over R=30 draws. This contrasts with the negative result of Al-Qurashi (2025) at $K \approx$ full-dictionary.

(3) The multi-feature effect is hallucination induction, not calibration. A Claude-as-judge breakdown of non-refusal generations resolves the "less hedging" into purely additional incorrect answers: known-entity correct $8\% \rightarrow 0\%$, incorrect $62\% \rightarrow 77\%$. The ablation suppresses the model's expression of uncertainty without changing what it

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

knows. This is the ablation-direction symmetric of Ferrando et al. (2025)'s known-feature steering: they show ablating known-entity features makes the model refuse on knowns; we show ablating unknown-entity features makes the model confabulate on unknowns. Neither direction is a usable calibration handle.

Code, the SAE checkpoint, raw labelled answers, per-layer baselines, and steering-experiment outputs are provided in the supplementary material; public release will follow upon acceptance.

2. Methodology

Model and SAE. Qwen3.6-27B (dense, 64 layers, $d_{\text{model}} = 5120$) in bfloat16. Top-K + AuxK SAE (Gao et al., 2024) with k=128, k_{aux}=2560, dictionary n=65,536 ($\approx 13\times$ expansion), trained jointly at $L \in \{11, 31, 55\}$ on 200M tokens (70:20:10 FineWeb-Edu / OpenThoughts / OpenMath). Loss-recovered ≥ 0.99 at all three layers; alive-feature fraction ≥ 0.95. Full training details in Appendix A.

Entity probe. We download the Wikidata files from Ferrando et al. (2025)'s repository (player/movie/city/song, $\approx$ 8,500 each) and sample 250 per type. We issue three attribute-recall questions per entity through the chat template (e.g. "Who directed X?"). An entity is **known** if $\geq 2/3$ answers contain a gold-attribute token (substring, $\geq$ 3 chars), **unknown** if all three are incorrect — regardless of whether the model refused or confabulated, since confabulation *is* the unknown signal. The labelled corpus has 607 entries (117 known, 490 unknown). Residuals are captured at the last token of the entity name inside "What can you tell me about `'X'`?". Full template list and labelling rule in Appendix B.

Pile noise filter. We discard SAE latents firing on $> 2\%$ of 16,860 Pile tokens (Gao et al., 2020) ($\approx 1\%$ of the dictionary). Without this filter a pilot run reached AUROC 1.0 by detecting subword-length differences between known and unknown entity names.

Three probes. At every layer $L \in \{0, \ldots, 63\}$ we evaluate (a) **diff-of-means** (Farquhar et al., 2023): project on the unit-norm class-mean difference; (b) **linear probe**: L2 logistic regression (C=1). At $L \in \{11, 31, 55\}$ we add (c) **single SAE latent**: best feature by Cohen's d on the train split after the Pile filter. We report 95% bootstrap CIs over n=1,000 resamples of the n_{test}=183 test split.

Steering controls. Single-feature: clamp $L{=}11/f_{61723}$ to ± 5 (and additive ± 2) during generation. Multi-feature: ablate the top-K latents by Cohen's $|d|$ on the train split, $K \in \{5, 20, 50, 200\}$. Controls: random-K null (R=30 draws per K); direction-sorted variants (positive-d, negative-d, $|d|$); bottom-$|d|$ anti-feature; three-way split (selection/tuning/report); Claude-as-judge correctness audit on every non-refusal generation; permutation test on the d ranking. Judge prompt in Appendix F.

3. Results

3.1. Replication: AUROC at scale

Figure 1 shows the per-layer entity-recognition signal. The supervised probes saturate above 0.85 from $L{\geq}7$ onwards, with peak at L=32 (linear probe 0.887 [0.823, 0.941]; diff-of-means 0.859). The single SAE latent at L=31 reaches 0.814 [0.731, 0.898]. The peak SAE latent layer (L=31, $\approx 48\%$ depth) coincides with the mid-stack regime where Ferrando et al. (2025)'s separation-score analysis identifies the best entity-recognition latents in Gemma-2-2B-IT (peak around $L9$ of 26, $\approx 35\%$ depth, plateauing thereafter). The closest directly-comparable AUROC value Ferrando reports is 0.732 on a related but distinct task (correct-vs-incorrect responses at the model's answer token, their Section 7); we therefore do not claim a head-to-head AUROC improvement.

Honest reading: the SAE latent does not beat supervised baselines. The L=31 SAE feature's 95% CI overlaps with both the L=31 logistic regression CI and the L=31 diff-of-means CI (Table 1). Single-latent AUROC at off-peak layers (L=11, L=55) is $\sim$15pp below supervised, indicating the entity-recognition signal at those layers is multi-directional rather than aligned with any one SAE feature. The contribution of the SAE at the peak layer is interpretability — one readable dimension — not raw classification performance.

3.2. Per-type breakdown at L=31

Table 2 reports the test-set AUROC at L=31 broken out by entity type. The single SAE latent's CI overlaps both supervised baselines on *movies*, falling within the same statistical band ($\Delta_{\text{SAE-LR}} = 2$pp). On *songs*, where the supervised probes saturate near 0.97, the SAE latent lags by 12pp — the entity-recognition signal on songs requires a multi-direction projection that no single SAE feature in our dictionary realises. On *players* the SAE latent tracks the diff-of-means probe within 4pp, while the L2 logistic regression reaches 0.90, indicating a direction the SAE captured imperfectly. The *city* subtype has 0 known entities in the test split (the single known city in the dataset was assigned to train) and is not interpretable.

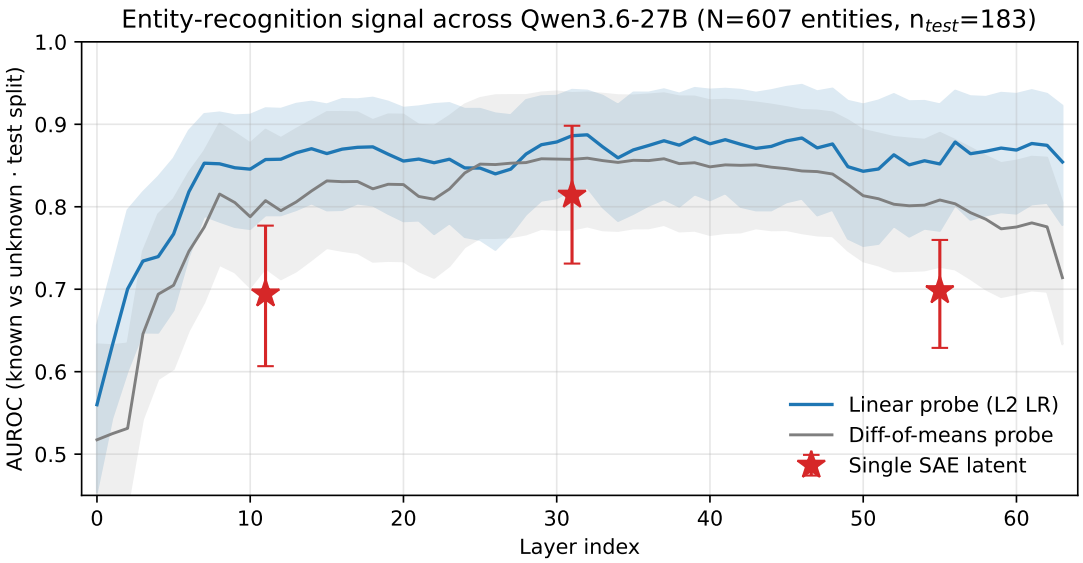

Figure 1. Entity-recognition AUROC (known vs unknown) across all 64 layers of Qwen3.6-27B (N=607, n_{test}=183). Linear probe (blue, 95% CI shaded) plateaus at 0.85–0.89 from $L\geq7$. Diff-of-means (grey) tracks $\sim$4pp below. Single SAE latents at the three trained layers (red stars, error bars are bootstrap CI) reach 0.81 at L=31 and $\sim$0.69 at L=11/L=55. The mid-stack peak is consistent with Ferrando et al. (2025)'s separation-score profile in Gemma-2-2B-IT (peak $\approx L9$ of 26, plateauing thereafter).

Table 1. AUROC at peak layers (95% bootstrap CI in brackets, 1,000 resamples on the n_{test}=183 test set).

	Single SAE	Linear probe	Diff-of-means
L=11	$0.69_{[.61,.78]}$	$0.86_{[.79,.92]}$	$0.81_{[.73,.89]}$
L=31	$0.81_{[.73,.90]}$	$0.89_{[.82,.94]}$	$0.86_{[.77,.94]}$
L=32	—	$\mathbf{0.89}_{[.82,.94]}$	$\mathbf{0.86}_{[.77,.94]}$
L=55	$0.70_{[.63,.76]}$	$0.85_{[.77,.92]}$	$0.81_{[.73,.89]}$

3.3. Single-feature steering: null

Steering experiments below use L=11/f_{61723}, the best SAE latent under a refusal-only pilot (N=226) prior to the Ferrando-style relabel; rerunning at L=31/f_{34957} is a natural follow-up. Clamping f_{61723} to ±5 (additive variant ±2) on 20+20 entities does not move calibration: known-entity refusal stays at 0%, unknown moves 15% $\rightarrow$ 25% under clamp-ablate (*wrong* direction) and 15% $\rightarrow$ 20% under additive ablate. Around 60–65% of generations change text — the feature has output effect, but its causal axis is not aligned with the calibration semantics suggested by its AUROC.

3.4. Multi-feature top-K ablation passes the random-K null

Multi-feature ablation at L=11, sorted by Cohen's $|d|$ on the train split, produces a -16.7pp shift in the unknown-entity refusal rate at K=200 (top-$|d|$ direction). To rule out that this is a generic perturbation effect, we computed a random-K null distribution by drawing R=30 random subsets of K latents at the same layer and applying the same ablation. The random-K mean effect is 0pp at $K \in \{5, 20, 50\}$ (zero variance across the 30 draws) and $+0.6 \pm 2.1$pp at K=200. The observed top-$|d|$ effect is therefore $\sim 8\sigma$ outside the random null distribution at K=200.

Direction-sorted controls. The effect is sign-symmetric in feature direction: top-K negative-d (predicted "unknown detector") gives -8.3pp at K=200 (-4.3σ, $p < 10^{-3}$); the symmetric top-K positive-d gives $+8.3$pp ($+3.7\sigma$, p=0.07). The bottom-K $|d|$ anti-feature control returns 0pp (p=0.93), confirming the test is sensitive to feature *selection*, not just to ablating any K features.

3.5. The multi-feature effect is hallucination induction, not calibration

A drop in refusal rate is consistent with two opposite stories: the ablation removed a hedge and the model produced *correct* answers it had been hiding, or it removed the hedge and the model confabulated *more confidently*. We ran the Claude-Haiku-as-judge audit on the predicted "unknown-detector" direction (top-K negative-d at K=200, the -8.3pp shift), scoring each non-refusal generation as

Table 2. Per-type AUROC at L=31 (95% bootstrap CI in brackets). n_k and n_u are test-set counts of known and unknown entities respectively. The SAE latent (single feature f_{34957}) is statistically indistinguishable from the L2 logistic regression on movies, lags on songs, and tracks the diff-of-means probe on players.

Type	(n_k, n_u)	SAE feat.	LR	Diff-means
movie	$(18, 51)$	$0.79_{[.66,.91]}$	$0.81_{[.69,.92]}$	$0.81_{[.63,.94]}$
song	$(12, 33)$	$0.84_{[.69,.96]}$	$0.97_{[.89,1.0]}$	$0.96_{[.89,1.0]}$
player	$(7, 54)$	$0.77_{[.55,1.0]}$	$0.90_{[.79,.98]}$	$0.74_{[.44,.99]}$
city	$(0, 8)$	—	—	—

Table 3. Claude-as-judge correctness audit on non-refusal generations under top-K ablation (K=200, top-negative-d). Known-entity correct rate drops from 8% to 0%; incorrect rate rises from 62% to 77%. The "less hedging" produced by the intervention resolves into purely additional incorrect answers, not surfaced knowledge.

Condition	Correct	Incorrect	Unverifiable
Known entities, n=13			
Baseline	8%	62%	31%
top-neg-d, K=200	0%	77%	23%
Unknown entities, n=7/8			
Baseline	0%	100%	—
top-neg-d, K=200	0%	100%	—

`YES` / `NO` / `UNVERIFIABLE`. The result (Table 3) is unambiguous: on known entities the intervention drops correct from 8% to 0% and pushes incorrect from 62% to 77%; on unknowns, baseline is already 100% incorrect on non-refusal and the intervention introduces no correct answers. The reduction in refusal under top-neg-d is composed entirely of *additional incorrect answers*, not retrieved knowledge.

We interpret this as a hallucination-induction mechanism, not a calibration knob: the top-$|d|$ direction at L=11 encodes the model's *expression* of uncertainty, not its underlying knowledge state. This is the symmetric case of Ferrando et al. (2025)'s causal-direction result — they showed ablating known-entity features makes the model refuse on knowns; we show ablating unknown-entity features makes the model confabulate on unknowns. Both identify mechanisms causally active in producing hallucination behaviour; neither is a usable calibration handle.

4. Discussion

4.1. Detection is not control

High AUROC is a *necessary* condition for a feature to be a causal handle, not a *sufficient* one. The L=31 SAE latent reaches 0.81, statistically indistinguishable from the logistic-regression ceiling at the same layer; yet single-feature steering on the L=11 SAE latent does not move the calibration axis the AUROC suggests it predicts (we did not steer at L=31; rerunning is a natural follow-up). SAE features are by construction sparse predictive coordinates; ablation along one such coordinate has no *a priori* guarantee of moving the model in a semantically interpretable way. Templeton et al. (2024) reported the same finding on Claude 3 Sonnet.

4.2. Statistical significance is not semantic effect, and what to do about it

The 4–8σ effect against the random-K null is, by the canonical SAE-causality test (Ferrando et al., 2025; Marks et al., 2025), a passing result. The Claude-as-judge audit shows it would have been a wrong one. Statistical significance against random-K distinguishes specific mechanism from generic perturbation, but not *which* mechanism: it cannot tell a calibration mechanism from an inverse, hallucination-induction mechanism that has the same surface effect on refusal rate. We therefore recommend a correctness audit of non-refusal generations as default for any SAE-steering paper claiming a calibration, honesty, or refusal effect: one LLM call per output, decisively distinguishes hedge-suppression from honesty-improvement. Replacing zero-ablation with optimal ablation (Li & Janson, 2024) is the natural extension.

4.3. Why naïve top-K finds the wrong direction

The top-K ranking by Cohen's $|d|$ mixes features encoding the knowledge gap with features encoding the model's *expression* of that gap ("about to hedge"). Both correlate with the unknown class and both pass AUROC, but their generation effects differ: knowledge-gap ablation changes *what* the model produces; hedge ablation changes *whether* it

hedges. Effect-size ranking selects the larger surface effect — on our data, the hedge subset.

References

Al-Qurashi, H. Entity-recognition feature steering in Gemma 2 2B: Hallucination induction, Arabic circuit asymmetry, and cross-lingual leakage. EA Forum post, March 2025. URL https://forum.effectivealtruism.org/posts/9vkEpghkuEv5QtfKG/entity-recognition-feature-steering-in-gemma-2-2b.

Bricken, T. et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Farquhar, S., Varma, V., Kenton, Z., Gasteiger, J., Mikulik, V., and Shah, R. Challenges with unsupervised LLM knowledge discovery. *arXiv preprint arXiv:2312.10029*, 2023.

Ferrando, J., Obeso, O., Rajamanoharan, S., and Nanda, N. Do I know this entity? Knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations (ICLR)*, 2025. Preprint arXiv:2411.14257.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Li, M. and Janson, L. Optimal ablation for interpretability. *arXiv preprint arXiv:2409.09951*, 2024.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*, 2025. Preprint arXiv:2403.19647.

Templeton, A. et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

A. SAE training details

The Top-K + AuxK SAE (Gao et al., 2024) is trained jointly across $L \in \{11, 31, 55\}$ from a single forward pass of Qwen3.6-27B in bfloat16 with SDPA attention. Training corpus: 200M tokens drawn 70:20:10 from FineWeb-Edu, OpenThoughts, and OpenMath. Top-K width k=128; auxiliary k_{aux}=2560 for dead-feature mitigation. Dictionary size n=65,536 per layer ($\approx 13\times$ expansion over d_{model}=5120). Decoder bias b_{dec} initialised to the geometric median of the activation buffer. The streaming activation buffer holds 2M activations on disk and refills from the corpus mix; SAE checkpoints are uploaded to a model registry every 10M tokens to make crashes recoverable. Loss-recovered, alive-fraction, L_0, sparse-probing accuracy, and TPP causal faithfulness all pass standard quality thresholds at the three layers (per-layer numbers in the supplementary material).

Hardware and compute. SAE training: a single rented RTX 6000 Pro 96GB instance, $\sim$24h wall-clock per training run. AUROC baselines (Section 3): a single A100 80GB session, $\sim$1h. Steering experiments: same A100 session, $\sim$2h. Cloud cost for the full pipeline is below $60 at current spot rates.

B. Entity labelling protocol

We start from 250 randomly-sampled entries per type (1,000 candidates) drawn with seed 0 from the Wikidata-derived JSON files in Ferrando et al. (2025)'s public repository. For each entity we issue three attribute-recall questions through the chat template with thinking disabled and greedy decoding, ≤ 24 generated tokens per question. The exact templates per type:

- **player**: "In which city or country was the football/basketball player `X` born?" (`place_birth`); "Name one professional team that player `X` has played for." (`teams_list`); "Approximately how tall is athlete `X` (in cm)?" (`height`).
- **movie**: directors, release_year, cast.
- **city**: country, location, population.
- **song**: performers, publication_year, genres.

A response is *correct* on attribute a if any token of length ≥ 3 in the gold attribute value appears as a substring of the (lowercased) response. A response is a *refusal* if it matches one of 17 patterns covering "I do not know / am not familiar / have no information / cannot verify" (full list in the supplementary material). An entity is **known** if at least two of three responses are correct; **unknown** if all three responses are incorrect (regardless of whether they refused or confabulated — confabulation *is* the unknown signal); and discarded if exactly one response is correct. The labelled corpus has 607 usable entries: 117 known and 490 unknown.

C. Per-layer AUROC for all three methods

The full table of point estimates and 95% bootstrap CIs for linear probe and diff-of-means at every layer $L \in \{0, \ldots, 63\}$, plus the single-SAE-latent values at $L \in \{11, 31, 55\}$, is included as JSON at `paper_baselines_per_layer.json` in the supplementary material alongside the figure source. Bootstrap is n=1,000 resamples on the n_{test}=183 test split.

D. Random-K null distributions

For $K \in \{5, 20, 50, 200\}$ we drew R=30 random subsets of size K from the L=11 SAE dictionary (uniform sampling from latents that survived the Pile noise filter), applied the same zero-ablation intervention as the top-K test, and measured refusal-rate change on the unknown-entity subset. Random-K means $\pm$ standard deviations: K=5: 0.0 ± 0.0; K=20: 0.0 ± 0.0; K=50: 0.0 ± 0.0; K=200: $+0.6 \pm 2.1$pp. Empirical one-sided p-values at K=200 for each direction-sorted variant: top-$|d|$ -16.7pp ($p < 10^{-3}$); top-negative-d -8.3pp ($p < 10^{-3}$); top-positive-d $+8.3$pp (p=0.067, borderline opposite-sign); bottom-$|d|$ anti-feature 0pp (p=0.93, within null). The anti-feature pass and the symmetric significance of positive- and negative-d directions confirm the test is sensitive to feature *selection*, not just to ablating any K features.

Permutation test on Cohen's d ranking. Under 1000 random label permutations, the recomputed top-200 ranking overlaps the original by mean 36.1 ± 13.1 features (out of 200), well above the ~ 0.6 expected by chance ($200 \times 200/65{,}536$) but well below the original ranking's identity. This indicates the ranking is partially driven by feature-firing-rate signal that is not entity-recognition-specific; it does not invalidate the random-K result above (which is invariant to ranking method by construction) but flags that effect-size selection should not be conflated with task-specific selection.

E. Sensitivity to the labelling rule

To address the methodological choice of counting confabulation (correct=0, no refusal) as unknown, we re-ran the per-layer AUROCs at $L \in \{11, 31, 55\}$ under the stricter refusal-only rule (correct=0 AND ≥ 1 refusal pattern $\Rightarrow$ unknown). Under this rule the labelled corpus shrinks to 130 entries (117 known, 13 unknown) — Qwen3.6-27B almost never refuses on the entity-attribute prompts; it confabulates instead.

Table 4 reports the comparison on the same residual capture. Point estimates under the refusal-only rule are within the 95% CIs of the Ferrando-style estimates at every layer and every probe, so the two rules are statistically compatible. However, the refusal-only test set has only $n_{\text{unknown}}=4$, producing CIs that span 0.4–1.0 for the supervised probes (statistically unidentifiable). The best SAE feature also shifts under the strict rule (e.g. at $L=31$, $f_{34957} \rightarrow f_{26289}$), indicating the two rules surface different SAE features — consistent with the model encoding distinct latents for "I lack the entry" vs "I am about to hedge". We use the Ferrando-style rule throughout the main paper because it provides $35\times$ the unknown-class sample size at the cost of compatible-but-different feature selection.

Table 4. Sensitivity of per-layer AUROC to the labelling rule. Refusal-only requires the model to emit an explicit hedge; Ferrando-style counts confabulation as unknown. Refusal-only test set has $n_{\text{test}}=39$ (35 known, 4 unknown) so CIs are wide. Best SAE features differ between rules, in parentheses. All point estimates fall within the Ferrando-style 95% CIs.

		Ferrando-style ($n_{\text{test}}=183$)	Refusal-only ($n_{\text{test}}=39$)
SAE feat.	$L=11$	$0.69_{[.61,.78]}$ (f_{41062})	$0.79_{[.71,.86]}$ (f_{3200})
SAE feat.	$L=31$	$0.81_{[.73,.90]}$ (f_{34957})	$0.87_{[.56,1.0]}$ (f_{26289})
SAE feat.	$L=55$	$0.70_{[.63,.76]}$ (f_{41335})	$0.81_{[.74,.89]}$ (f_{38386})
LR	$L=11$	$0.86_{[.79,.92]}$	$0.82_{[.61,1.0]}$
LR	$L=31$	$0.89_{[.82,.94]}$	$0.79_{[.40,1.0]}$
LR	$L=55$	$0.85_{[.77,.92]}$	$0.90_{[.70,1.0]}$
Diff-mean	$L=11$	$0.81_{[.73,.89]}$	$0.72_{[.48,.94]}$
Diff-mean	$L=31$	$0.86_{[.77,.94]}$	$0.74_{[.33,.98]}$
Diff-mean	$L=55$	$0.81_{[.73,.89]}$	$0.87_{[.65,1.0]}$

F. Claude-as-judge prompt

Every non-refusal generation in the steering experiment was scored by Claude Haiku with a single binary prompt: "Is the following statement factually correct? Answer only `YES`, `NO`, or `UNVERIFIABLE`. Statement: `<generation>`." We do not provide gold attributes to the judge; we rely on the judge's own world knowledge. The judge's role is to distinguish "model produced confident wrong answer" from "model produced confident correct answer" — both look the same on the refusal-rate metric. Counts: known entities $n_{\text{baseline}}=13$, $n_{\text{ablate}}=13$; unknown entities $n_{\text{baseline}}=7$, $n_{\text{ablate}}=8$ (the asymmetry is due to one additional non-refusal generation under ablation in the unknown subset). Inter-annotator agreement was not measured; replication of the judge against a stronger model (Claude Opus or human label) on a 50-sample subset is the natural follow-up.