BEYOND DECODER PROXIES: CAUSAL FEATURE SELECTION FOR SPARSE AUTOENCODER CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling sparse autoencoder (SAE) dictionaries creates a feature-selection bottleneck: static decoder projections can screen thousands of candidate directions, but do not reveal which cause a desired probability shift in context. We introduce the target-selection margin (TSM), an intervention-based selection method that ranks features by changes in the log ratio of target to distractor probability mass. Across 18 concepts and five public model/SAE settings, TSM correlates more strongly with held-out margin than decoder proxies (Spearman 0.379–0.735 versus 0.155–0.191). In matched 512-feature pools, paired held-out utility gains have concept-clustered 95% confidence intervals entirely above zero in all five settings. Lexical and grammatical audits broaden coverage beyond the original token sets, while severe target replacement exposes specification sensitivity. Two-dictionary audits identify retrieval losses before causal reranking. Separate behavioral audits show that improved contrast-specific decisions need not overcome full-vocabulary competition and do not establish reliable multi-step generation gains. These results support causal feature selection across prompts under a specified target–distractor contrast, distinct from retrieval coverage and generation.

1 INTRODUCTION

Sparse autoencoders (SAEs) decompose language-model activations into dictionaries of feature directions (Bricken et al., 2023; Cunningham et al., 2023), providing candidates for activation steering: modifying internal states at inference time to influence generation without finetuning (Turner et al., 2023; Zou et al., 2023). As dictionaries scale to millions of features (Gao et al., 2024; Templeton, 2024), selecting useful directions becomes a practical challenge: which features produce a desired probability shift, and do their effects persist across prompts?

Static decoder projections offer a low-cost signal of feature relevance, but relevance is not the same as causal utility. A direction whose projection favors concept-related tokens need not produce the desired shift when added to the model's hidden state. Prior work has highlighted the distinction between feature relevance and output utility (Arad et al., 2025; Ding et al., 2025; Marks et al., 2024). Therefore, we ask whether prompt-conditioned intervention effects predict held-out utility better than static projections.

To address this question, we introduce target-selection margin (TSM), which ranks features by their intervention-induced change in the log ratio of target to distractor probability mass. The token sets specify the desired contrast, while the prompts provide the context in which its effect is measured. We add each candidate's SAE decoder vector to the residual stream at the final prompt position using an activation-calibrated scale, then average the resulting margin gains over selection prompts. After selecting the top-K features, we freeze the selected sets and evaluate their individual effects on shared held-out prompts. TSM requires inference-time interventions but no finetuning or gradients. Figure 1 further summarizes the conditional-ranking procedure.

The primary evaluation changes the prompts while preserving the target–distractor contrast, testing cross-prompt transfer of the same margin effect. Separate audits examine sensitivity to the token specification, retrieval coverage, and whether joint feature interventions change decoded outputs. We report four main findings:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

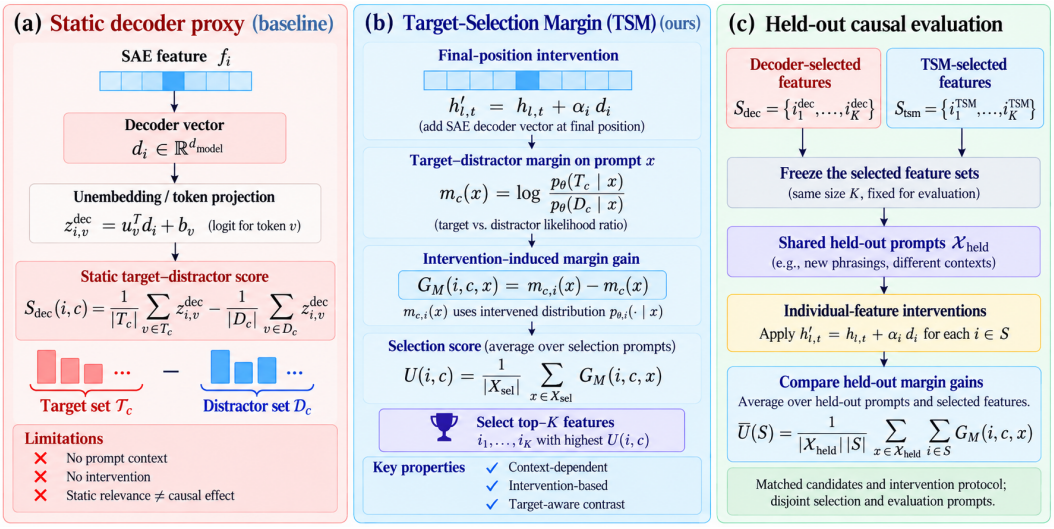


Figure 1: Overview of causal SAE feature selection. (a) The static decoder proxy ranks features by projected target–distractor scores. (b) TSM ranks intervention-induced margin gains on selection prompts. (c) Each selector's top-K set is frozen and evaluated through individual-feature interventions on shared held-out prompts, using matched candidates and calibrated scales. Evaluation changes prompts while preserving the specified contrast.

Causal ranking transfer. TSM measured on selection prompts predicts causal utility on held-out prompts under the same margin objective (0.379–0.735 Spearman, versus 0.155–0.191 for decoder proxies). Within matched pools, this predictiveness yields higher-utility feature selection across five model/SAE settings.

Specification robustness. Complementary token resampling and moderate specification noise preserve the advantage, while severe target replacement and harder semantic-neighbor distractors reveal model-dependent limits. Adjective/verb contrasts extend coverage beyond the original topical concepts.

Retrieval coverage. Exhaustive audits against selection- and held-out-prompt oracles expose decoder-prefilter losses in two dictionaries. Unseen-prompt causal screening improves recall substantially on Gemma, but not reliably on Qwen.

Behavioral limits. Joint-vector tests show model-dependent improvements in candidate-token decisions, without target emission as the full-vocabulary next-token argmax at the tested operating points. The multi-step generation audit does not establish reliable target-mention gains, and low baseline instruction adherence limits interpretation of this null result.

2 Related Work

SAE feature selection and output utility. SAE4Steer distinguishes input-based feature relevance from output utility (Arad et al., 2025); its Output Score and Delta Token Confidence motivate output-utility selectors. Our focus is target-specific utility: TSM ranks intervention-induced changes in a specified target–distractor contrast and tests whether this ranking transfers across prompts and lexical specifications.

Activation steering. Early work extracted latent steering vectors from pretrained language models (Subramani et al., 2022); activation addition, representation engineering, and contrastive activation addition construct directions from contrasting examples or populations (Turner et al., 2023; Zou et al., 2023; Rimsky et al., 2024). Inference-time intervention and function vectors identify directions associated with truthfulness or task behavior (Li et al., 2023; Todd et al., 2024). We select existing SAE dictionary features rather than construct a dense control direction.

Sparse autoencoders for interpretability and control. Motivated by superposition (Elhage et al., 2022), SAE dictionary learning seeks sparse decompositions with interpretable units, from small transformers to large language models (Bricken et al., 2023; Cunningham et al., 2023; Gao et al., 2024; Templeton, 2024). We use public Qwen-Scope (Qwen Team, 2026b), Gemma Scope (Lieberum et al., 2024), and Llama Scope (He et al., 2024) dictionaries to study the utility of learned features without training new SAEs.

SAE evaluation benchmarks. SAE benchmarks evaluate interpretability, concept detection, and downstream utility from complementary perspectives (Gulko et al., 2025; Karvonen et al., 2025; Wu et al., 2025). SAEBench examines multiple capabilities beyond reconstruction, while AxBench compares SAE-based methods with prompting, finetuning, and supervised representation methods. Sparse-probing results likewise caution against equating interpretable features with task utility (Kantamneni et al., 2025). We isolate one component of these broader evaluations: feature selection under matched candidates and a fixed probability contrast. Appendix Table 6 relates the evaluation protocols.

Causal evaluation of representations. Activation patching and causal tracing replace internal states across clean and perturbed runs to localize computations that mediate an output (Zhang & Nanda, 2024; Meng et al., 2022). Causal mediation analysis similarly quantifies effects through model components (Vig et al., 2020). TSM shares this intervention-based perspective but asks which additive SAE directions produce a desired probability contrast across prompts. It evaluates the sufficiency of individual SAE directions at a fixed intervention site, rather than tracing the mechanism underlying a naturally occurring prediction.

3 METHOD

3.1 PRELIMINARIES: SPARSE AUTOENCODERS AND DECODER PROXIES

For a residual-stream state h, an SAE encodes feature activations and reconstructs the state as

$$s_i(h) = (h - b_{\text{dec}})W_{\text{enc},i} + b_{\text{enc},i}, \qquad a_i(h) = \phi_i(s_i(h)), \qquad \hat{h} = b_{\text{dec}} + \sum_i a_i(h)d_i, \tag{1}$$

Here $W_{\text{enc},i}$ is column i of the encoder weight matrix, and $b_{\text{enc},i}$ is the corresponding scalar encoder bias. The decoder bias vector b_{dec} also centers the encoder input. The activation function ϕ_i maps the pre-activation $s_i(h)$ to sparse feature activation $a_i(h)$; d_i is feature i's decoder vector, and $\hat{h}$ is the reconstruction.

For token v, let u_v be its unembedding vector and b_v its output-head bias. Our static decoder baseline projects d_i to $z^{\text{dec}}_{i,v} = u_v^\top d_i + b_v$ and takes the difference of mean projected logits:

$$S_{\text{dec}}(i, c) = \frac{1}{|T_c|} \sum_{v \in T_c} z^{\text{dec}}_{i,v} - \frac{1}{|D_c|} \sum_{v \in D_c} z^{\text{dec}}_{i,v}. \tag{2}$$

Here T_c, D_c are the concept's target and distractor token sets; $b_v = 0$ for a bias-free output head. This score requires no prompt-conditioned forward pass and therefore does not measure the effect of an intervention in context.

3.2 DECODER-VECTOR INTERVENTION

Let $h_{\ell,t}$ be the residual-stream state at layer ℓ and token position t. For SAE feature i with decoder vector d_i, we evaluate an additive intervention

$$h'_{\ell,t} = h_{\ell,t} + \alpha_i d_i, \tag{3}$$

where α_i is a feature-specific intervention scale. In a fixed SAE decomposition, adding $\alpha_i d_i$ increments only coefficient a_i and preserves the reconstruction residual. We use this calibrated increment rather than the prompt's $a_i(h)$ because a useful steering direction may be weakly active or inactive before intervention. This tests direction sufficiency at that scale, not natural occurrence or necessity. Main probes act only at the final prompt position; other positions remain unchanged. Generation applies a joint top-K vector at the current final position at every autoregressive step (Sections 4.6–4.7). Appendix D reports separate prompt-wide diagnostics.

3.3 TARGET-AWARE UTILITY

We evaluate target-specific utility through changes in target probability relative to specified distractors. Each concept c has a target token set T_c and a distractor token set D_c. Let $p_\theta(\cdot \mid x)$ be the unsteered next-token distribution and $p_{\theta,i}(\cdot \mid x)$ the distribution after the feature intervention in Eq. 3. For a token set S, write $p_\theta(S \mid x) = \sum_{v \in S} p_\theta(v \mid x)$. The target-over-distractor margin and its intervention-induced gain are

$$m_c(x) = \log \frac{p_\theta(T_c \mid x)}{p_\theta(D_c \mid x)}, \qquad G_M(i, c, x) = m_{c,i}(x) - m_c(x), \tag{4}$$

where $m_{c,i}$ uses the intervened distribution. TSM averages this gain over selection prompts:

$$U(i, c) = \frac{1}{|X_{\text{sel}}|} \sum_{x \in X_{\text{sel}}} G_M(i, c, x). \tag{5}$$

We rank candidate SAE features by $U(i, c)$, select the top K, and report their mean margin gain on held-out prompts X_{held}. Equivalently, $G_M = G_T - G_D$, where G_T and G_D are changes in the log probability mass of the two sets. Positive margin gain measures relative preference, not absolute target probability. Appendix A gives the log-sum-exp form and derivation.

Why does this differ from a decoder projection? Let $\delta_i(x)$ be the intervention-induced change in output logits, and let $q_S(v \mid x) = p_\theta(v \mid x)\mathbf{1}[v \in S] / \sum_{u \in S} p_\theta(u \mid x)$ be the unsteered distribution restricted to set S. A first-order expansion gives

$$G_M(i, c, x) = \langle q_{T_c}(\cdot \mid x) - q_{D_c}(\cdot \mid x), \delta_i(x) \rangle + O(\|\delta_i(x)\|^2). \tag{6}$$

The shared softmax-normalization term cancels. Unlike a fixed decoder projection, both the contrast weights and the intervention's logit response depend on context. This expansion provides a first-order interpretation; TSM uses the measured margin change, not the approximation (Appendix A).

3.4 CONCEPT SPECIFICATION AND PROMPT SPLIT

The main benchmark specifies 18 topical concepts through curated target and distractor lexicons. Each concept is evaluated with four selection and four held-out prompt templates. These disjoint sets of short prefixes vary the continuation context while keeping the desired contrast fixed. Representative token strings, the complete prompt list, and construction details appear in Appendix Tables 7 and 8.

For the topical benchmark, each lexical entry is mapped to its first subtoken ID. We deduplicate the sets and remove target IDs from the distractors, yielding a next-token contrast rather than a complete-word generation objective. The grammatical-category audit instead requires exact single-token words, while generation tests match whole words or phrases.

3.5 BASELINES AND METRICS

Our primary comparisons are random subsets, static decoder-logit margin (Eq. 2), and target-mass-only selection. The decoder baseline measures static relevance; target-mass selection uses TSM's calibrated interventions but omits distractor subtraction. We also compare contrastive variants. Generic robustness and SAE4Steer-style Delta Token Confidence reuse cached prompt-wide scores as auxiliary transfer diagnostics, not fully matched final-position baselines (Appendix A.1). Held-out oracle rankings serve only as retrieval-recall references (Table 4; Appendix C.8).

Each selector fixes its top-K set without using held-out intervention utilities. For each concept–pool pair, we average the individual-feature margin gains over the selected features and four held-out prompts, using the same calibrated scales. Table 1 aggregates these held-out utilities across pairs; Table 2 reports Spearman correlations between selection scores and mean held-out utility across all 512 candidates.

Selection and evaluation deliberately share the target–distractor margin objective, testing cross-prompt transfer rather than transfer to a different task. This isolates feature selection from decoding and judging but favors objective-aligned selectors. Separate decision and generation tests assess target choices, target mentions, and instruction preservation; global top-1 change measures disruption rather than successful target emission.

3.6 ACTIVATION-CALIBRATED INTERVENTION STRENGTH

To account for different activation magnitudes, we estimate feature scales on fixed calibration states $\mathcal{H}_{\text{cal}}$ shared across selectors. For each feature in candidate pool $\mathcal{F}$, its raw scale is

$$q_i = \text{Quantile}_{0.95}\left(\{a_i(h) : h \in \mathcal{H}_{\text{cal}},\ a_i(h) > 0\}\right), \tag{7}$$

with $q_i = 0$ if no positive activation is observed. We then median-normalize and cap the scales:

$$\tilde{\alpha}_i = \gamma \frac{q_i}{\text{median}_{j \in \mathcal{F}}\, q_j}, \qquad \alpha_i = \min(\alpha_{\max}, \tilde{\alpha}_i). \tag{8}$$

The main experiments use $\gamma = 5.0$, with an intervention cap $\alpha_{\max} = 10$ for Qwen3-1.7B, Qwen3-8B, and Gemma-2-2B, and $\alpha_{\max} = 2$ for Qwen3.5-2B and Llama-3.1-8B. This preserves relative activation scales before clipping; it does not calibrate causal effect sizes across model families. Calibration uses unsteered activations without intervention-utility feedback, but its texts overlap with benchmark prompts (Appendix Table 8).

4 EXPERIMENTS

4.1 SETUP

We evaluate Qwen3-1.7B, Qwen3-8B, Qwen3.5-2B, Gemma-2-2B, and Llama-3.1-8B with public layer-12 SAEs. For each model, three randomly sampled 512-feature pools yield 54 concept–pool comparisons at each K. All selectors use the same candidates, calibration, and evaluation prompts, isolating ranking quality within a supplied pool. The main comparisons use $K = 10$; Appendix C.1 reports $K = 1, 5, 10, 20$.

To test a broader concept source, a Qwen3-1.7B pilot converts 100 AxBench/SAE4Steer descriptions into token contrasts. A 20-description subset supports exhaustive retrieval audits on Qwen3-1.7B and Gemma-2-2B; it is distinct from the main 18-concept benchmark. These experiments reuse external descriptions, not the original behavioral tasks or benchmark scores. The lexical and behavioral extensions are described with their results below; complete protocols appear in Appendix C.

All experiments are inference-only. Comparisons are paired by model, concept, pool, and K. For the matched-pool ranking results, 95% confidence intervals use 10,000 hierarchical bootstrap draws, resampling concepts and then a pool within each sampled concept. This treats the 18 concepts, rather than the 54 rows, as semantic clusters while accounting for pool variation. Intervals are pointwise, without correction for multiple comparisons. Appendix E further lists assets and reproduction details.

4.2 CAUSAL RANKING TRANSFER ACROSS PROMPTS

Do selection-prompt scores predict causal effects on new prompts? Across all 512 candidates, TSM's correlation with held-out margin is 0.379–0.735, versus 0.155–0.191 for the decoder proxy (Table 2). The ranking advantage therefore extends beyond the selected top-K: prompt-conditioned causal scores better predict the same effect in held-out contexts.

This predictiveness improves feature choice. At $K = 10$, TSM exceeds decoder selection in all five settings, with paired 95% intervals above zero and 52–54 wins among 54 comparisons (Table 1). For example, TSM/decoder held-out utilities are 0.388/0.124 on Qwen3.5 and 0.335/0.045 on Llama. Figure 2 shows all five models and paired uncertainty. These are individual-feature effects; their magnitudes should be interpreted within each model because intervention effects are not calibrated across families.

4.3 ABLATIONS OF THE SELECTION CRITERION

The Mass column in Table 1 uses the same interventions but ranks by target log-mass gain without distractor subtraction. TSM has higher mean held-out margin in all five settings. This comparison isolates a practical consequence of the contrast: a feature can increase target mass while also increasing distractor mass, making it less useful for separating the two sets.

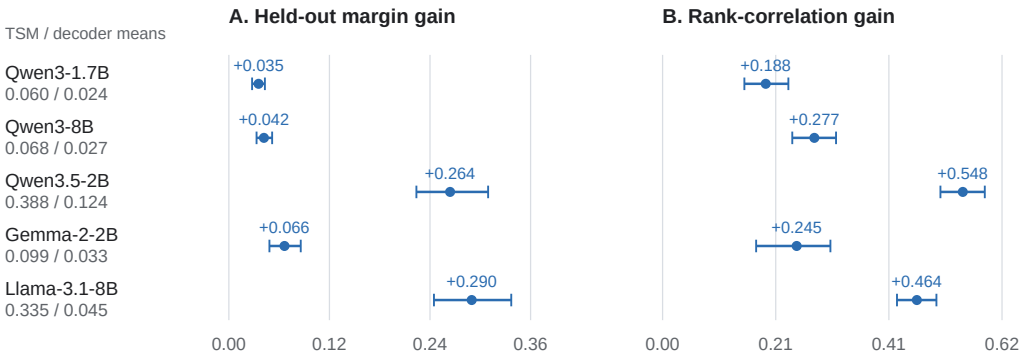


Figure 2: Causal ranking transfer across all five model/SAE settings. Left: paired TSM minus decoder held-out margin at $K = 10$; values under model names are TSM/decoder means. Right: paired Spearman-correlation gain over all 512 candidates. Points and bars show estimates and pointwise 95% concept/pool-bootstrap intervals from Tables 1–2. Margin magnitudes are not calibrated across models.

Table 1: Final-position, matched-pool feature selection ($K = 10$). Values are mean held-out margin gains of individually intervened features, not joint-vector effects. Paired Δ =TSM−decoder; 95% CIs resample concepts and nested pool draws. Mass drops the distractor subtraction. Wins count matched concept–pool rows. Paired differences are computed from unrounded values.

Model	Cap	TSM	Decoder	Mass	Δ [95% CI]	Wins
Qwen3-1.7B	10	0.060	0.024	0.025	0.035 [0.028, 0.043]	52/54
Qwen3-8B	10	0.068	0.027	0.040	0.042 [0.033, 0.052]	54/54
Qwen3.5-2B	2	0.388	0.124	0.225	0.264 [0.224, 0.309]	54/54
Gemma-2-2B	10	0.099	0.033	0.064	0.066 [0.048, 0.086]	53/54
Llama-3.1-8B	2	0.335	0.045	0.216	0.290 [0.245, 0.337]	54/54

The advantage holds across selection budgets: TSM has higher mean held-out utility than both decoder and target-mass selection in all 20 model–K settings, with $K = 1, 5, 10, 20$ (Appendix Tables 9–12). These are individual-feature averages, not joint-steering outcomes.

Related contrastive criteria remain close to TSM: subtracting half the selection-prompt standard deviation gives 0.383 versus 0.388 on Qwen3.5 at $K = 10$; positive-effect-frequency selection is also similar. Together, these results support target-aware contrastive selection across budgets and related criteria, rather than a uniquely optimal mean-margin formula.

4.4 Robustness to concept specification

The preceding comparisons hold the target–distractor specification fixed. We next test how feature selection changes when this specification is resampled or perturbed. We split target and distractor IDs into complementary selection/evaluation halves and edit selection only. Both selectors receive the same edited sets, excluding clean evaluation IDs. Five corruption families add or replace distractors with semantic neighbors, add or replace targets with off-concept tokens, or copy tokens between sets. Here $q = 1, 2, 3, 4$ counts edits, not a percentage; overlap adds q memberships per set ($2q$ total). Five lexical splits and five corruption draws yield 1,350 paired cells per model, family, and level across 18 concepts and three pools.

Table 3(a) reports retention $R = 100(V_{\text{noisy}} - V_{\text{random}})/(V_{\text{clean}} - V_{\text{random}})$, where V is aggregate held-out utility; families receive equal weight. At $q = 2$, all 25 model–family 95% CIs for TSM minus the equally corrupted decoder selector are above zero. Mean retention falls from 85.1–92.2% at $q = 1$ to 58.5–69.4% at $q = 4$. This average conceals severe target-replacement failures: at $q = 4$, retention is only 35.9–38.5% and just one of five paired CIs excludes zero. Appendix C.2 reports every family/level result.

Table 2: Feature-level Spearman correlation with held-out margin across all 512 candidates, averaged over 18 concepts and three pools. TSM's selection score and held-out evaluation use the same margin functional on disjoint prompts; the decoder score is a static proxy. These correlations measure cross-prompt predictiveness, not metric-independent task transfer. Intervals use the main hierarchical bootstrap.

Model	Decoder ρ [95% CI]	TSM ρ [95% CI]	Paired $\Delta\rho$ [95% CI]
Qwen3-1.7B	0.191 [0.165, 0.217]	0.379 [0.346, 0.413]	0.188 [0.149, 0.230]
Qwen3-8B	0.174 [0.152, 0.197]	0.451 [0.414, 0.488]	0.277 [0.237, 0.317]
Qwen3.5-2B	0.186 [0.160, 0.213]	0.735 [0.707, 0.762]	0.548 [0.508, 0.589]
Gemma-2-2B	0.155 [0.134, 0.176]	0.400 [0.330, 0.458]	0.245 [0.171, 0.307]
Llama-3.1-8B	0.156 [0.126, 0.184]	0.620 [0.591, 0.650]	0.464 [0.428, 0.500]

Table 3: Token-interface robustness, $K = 10$, with disjoint clean evaluation tokens and held-out prompts. (a) Five-family mean retention above random (%). (b–c) Paired margin gain, TSM minus decoder [95% hierarchical CI]. Each model/category in (c) has 225 rows: 15 contrasts $\times$ 3 pools $\times$ 5 splits. Complete audits are in Appendix C.2–C.4.

(a) Selection-specification noise: retention R (%)

Model	$q = 1$	$q = 2$	$q = 3$	$q = 4$
Qwen3-1.7B	90.1 [81.6, 98.4]	83.4 [73.1, 94.2]	74.3 [62.9, 85.5]	65.7 [53.4, 77.9]
Qwen3-8B	89.7 [79.9, 99.4]	81.8 [69.9, 93.7]	74.6 [61.7, 86.8]	66.3 [52.6, 79.6]
Qwen3.5-2B	90.3 [80.6, 98.5]	82.6 [69.7, 93.8]	73.1 [58.3, 85.8]	64.3 [48.7, 77.5]
Gemma-2-2B	85.1 [73.0, 98.4]	74.7 [60.1, 91.2]	67.3 [51.0, 84.9]	58.5 [41.1, 77.6]
Llama-3.1-8B	92.2 [85.1, 99.0]	86.5 [77.4, 95.4]	78.1 [67.9, 88.5]	69.4 [58.5, 80.1]

(b) Lexical resampling and whole-set distractor redraw: paired Δ margin

Model	Clean split	Uniform distractors	Neighbor distractors
Qwen3-1.7B	0.0198 [0.0099, 0.0303]	0.0159 [0.0060, 0.0262]	0.0127 [0.0025, 0.0231]
Qwen3-8B	0.0224 [0.0100, 0.0349]	0.0226 [0.0114, 0.0340]	0.0149 [0.0033, 0.0267]
Qwen3.5-2B	0.1586 [0.0941, 0.2265]	0.1630 [0.0969, 0.2336]	0.1025 [0.0320, 0.1752]
Gemma-2-2B	0.0346 [0.0133, 0.0559]	0.0300 [0.0090, 0.0520]	0.0179 [-0.0036, 0.0391]
Llama-3.1-8B	0.1988 [0.1442, 0.2556]	0.1985 [0.1382, 0.2594]	0.1551 [0.0963, 0.2134]

(c) Beyond topical nouns: paired Δ margin

Model	Adjectives	Verbs
Qwen3-1.7B	0.0093 [0.0008, 0.0172]	0.0134 [0.0048, 0.0221]
Gemma-2-2B	0.0135 [-0.0044, 0.0317]	0.0192 [-0.0001, 0.0394]
Llama-3.1-8B	0.0946 [0.0413, 0.1508]	0.1441 [0.0906, 0.1985]

Panel (b) replaces all five selection distractors with uniformly sampled other-concept tokens or semantic neighbors. Under neighbor redraw, four paired 95% CIs are above zero; Gemma's includes zero (Appendix C.3).

Panel (c) tests 15 adjective and 15 base-form verb contrasts across six semantic domains on Qwen3-1.7B, Gemma, and Llama, with grammatical-slot-matched prompts and disjoint lexical halves. All six estimates are positive, but both Gemma CIs include zero: coverage extends beyond topical nouns with model-dependent limits (Appendix C.4).

4.5 Retrieval beyond matched pools

The preceding experiments evaluate ranking within supplied candidate pools. We next examine whether candidate retrieval excludes features with high causal utility. Table 4 audits the full 32,768-feature Qwen and 16,384-feature Gemma dictionaries. Recall measures coverage of each concept's causal top ten, defined separately on selection and held-out prompts; both oracles are audit-only. Static decoder retrieval misses most held-out-oracle features even at $M = 1024$. Averaging four fixed screening prompts disjoint from both oracle prompt sets improves held-out recall substantially

Table 4: Full-dictionary retrieval, $K = 10$, over 20 concepts. D/U denotes decoder / four-unseen-prompt causal screening. Recall is percent of the selection (Sel.) or held-out (Held) oracle's top ten. Paired Δ Held is U minus D in percentage points [95% CI]. Costs estimate screening plus reranking in GPU seconds.

Model	M	Sel. D/U	Held D/U	Δ Held [95% CI]	GPU s D/U
Qwen3-1.7B	128	24.5/27.0	20.5/25.5	5.0 [-4.5, 14.5]	131/2040
	1024	42.5/51.0	38.5/44.0	5.5 [-5.5, 15.5]	791/2393
Gemma-2-2B	128	16.0/60.0	10.0/74.5	64.5 [55.0, 73.0]	138/1149
	1024	42.5/83.0	33.5/91.5	58.0 [49.0, 67.0]	678/1469

on Gemma, whereas Qwen's paired intervals include zero. Appendix C.8 reports the shared-prompt diagnostic, every single/two-template combination, and alternative prefilters.

The dominant cost is an intervened forward pass for each feature–prompt pair, approximately $O(FP)$ for F features and P prompts; the same output scores all concept token sets. Microbatching is memory-limited, while feature sharding reduces elapsed time rather than total GPU work. Exhaustive four-prompt scoring takes about 33.9 GPU-minutes on Qwen and 18.7 on Gemma in the measured H100 setup. Table costs add projected reranking, not measured end-to-end latency. At a per-concept budget of $M = F/8$, decoder candidate unions across concepts already cover about 85% of either dictionary (Appendix C.8), limiting the computational savings from shared causal reranking.

Within a decoder-retrieved union of 2,129 features, the 100-concept Qwen pilot retains a ranking advantage: mean held-out utility is 0.0938 for TSM versus 0.0352 for decoder selection, with wins on 97/100 concepts. This supports conditional reranking on external descriptions; the exhaustive audits separately quantify features lost before reranking (Appendix C.7).

4.6 When margin gains change decisions

The preceding experiments evaluate features individually. To test whether their ranking advantage survives composition and decoding, we combine each selector's top-K set $S_{c,K}$ into a joint intervention:

$$v_{c,K} = \frac{1}{\sqrt{K}} \sum_{i \in S_{c,K}} d_i, \qquad h' = h + \beta v_{c,K}. \tag{9}$$

Directions receive equal weight: α_i calibrates individual probes, not the mixture. Both selectors use $(K, \beta) = (10, 2)$ for Qwen3-1.7B and Gemma, and $(20, 1)$ for Qwen3.5 and Llama, fixed before inspecting multi-step outcomes. Joint effects can include interactions and need not equal mean individual-feature utility. Each model has 216 paired cases: 18 concepts × 3 pools × 4 prompts.

Next-token decisions. Candidate-token top-1 asks whether the highest-logit token within $T_c \cup D_c$ belongs to T_c, measuring contrast-specific choice rather than downstream task accuracy. Table 5(a) reports paired transitions: TSM's net gains over decoder selection are 13.9 and 10.2 percentage points on Qwen3.5 and Llama, with clustered 95% CIs above zero. Qwen3-1.7B and Gemma have mostly unchanged decisions and CIs including zero despite positive margin gains. Neither selector produces a target as the global next-token argmax in any of the 864 paired cases.

Competition outside the specified contrast. Panel (b) reports the best target token's full-vocabulary rank and its logit gap to the strongest token outside $T_c \cup D_c$, neither used for selection. Log-rank improvement is $\log_{10} r_{\text{decoder}} - \log_{10} r_{\text{TSM}}$; positive changes in either endpoint favor TSM. Qwen3.5 and Llama improve on both, yet median target ranks remain 446.5 and 317.5. Target emission requires the best target to outrank every non-target token, which a higher target-to-distractor ratio does not ensure (Appendices A and C.5).

Neither selector makes a target token the full-vocabulary argmax at any of the four tested joint scales; stronger interventions can improve contrast-specific decisions while increasing output disruption (Appendix C.6).

Table 5: Joint-vector effects, with 216 paired cases per model and clustered 95% CIs. (a) Transitions are distractor→target / target→distractor / unchanged class; open-target counts list decoder then TSM. (b) Ranks are medians; Δ columns are means of paired differences, not differences between medians. Dec./TSM denotes decoder/TSM; positive changes favor TSM. Gemma's panel (b) lower bounds round to zero.

(a) Margin gains and candidate-token decisions

Model	Δ margin [95% CI]	Transitions	Net pp [95% CI]	Open target
Qwen3-1.7B	0.009 [0.001, 0.018]	4/1/211	1.4 [-1.4, 5.6]	0/216; 0/216
Gemma-2-2B	0.034 [0.015, 0.052]	2/1/213	0.5 [-1.4, 2.8]	0/216; 0/216
Qwen3.5-2B	0.356 [0.290, 0.418]	30/0/186	13.9 [6.9, 20.8]	0/216; 0/216
Llama-3.1-8B	0.340 [0.278, 0.410]	23/1/192	10.2 [2.8, 18.1]	0/216; 0/216

(b) Full-vocabulary target rank and outside gap

Model	Rank Dec./TSM	Δ $\log_{10}$ rank [95% CI]	Δ outside gap [95% CI]
Qwen3-1.7B	638.5/642	0.0023 [-0.0021, 0.0069]	0.0026 [-0.0182, 0.0234]
Gemma-2-2B	423/408	0.0042 [0.0000, 0.0084]	0.0191 [0.0000, 0.0365]
Qwen3.5-2B	548.5/446.5	0.1449 [0.1120, 0.1804]	0.3712 [0.2821, 0.4653]
Llama-3.1-8B	422/317.5	0.1035 [0.0744, 0.1356]	0.2207 [0.1427, 0.3107]

4.7 Separating ranking utility from generation behavior

We next evaluate the same joint vectors during greedy multi-step generation under four structural instructions that do not name the target concept (up to 40 tokens; 216 paired cases per model). Whole-word/phrase target-mention gains over decoder selection range from 0.0 to 3.7 percentage points, with all 95% CIs including zero; no model meets the prespecified target-mention criterion. The unsteered baseline already has only 1.4–2.8% target mention and zero strict instruction adherence under these raw-text prompts. This floor limits interpretation: the null comparisons establish neither selector equivalence nor a general absence of behavioral utility. Appendix C.9 reports the joint evaluation rule, absolute outcomes, and blinded judgments.

5 Limitations

Concept interface and objective. Concepts, semantic neighbors, and grammatical contrasts remain curated; resampling does not automate specification. First-subtoken and single-token readouts do not capture arbitrary morphology or composition. Selection and headline evaluation share a margin functional, so the ranking claim concerns transfer of that specified effect.

Retrieval and operating points. Exhaustive audits cover two dictionaries, leaving scaling to million-feature dictionaries unresolved. Layer, scale, prompts, and dictionary can affect both ranking and joint interventions.

Sparse and semantic scope. TSM selects existing SAE features, not an optimal dense vector or guaranteed monosemantic units. Supplementary CAA and real-context probes use prompt-wide interventions and remain exploratory.

6 Conclusion

We introduced target-selection margin (TSM), which ranks SAE features by intervention-induced changes in the log ratio of target to distractor probability mass. Across 18 concepts and five model/SAE settings, TSM scores better predict held-out margin gains than static decoder scores, and TSM-selected features achieve higher mean individual-feature utility within matched candidate pools. These results support cross-prompt ranking under a shared margin objective. Lexical and grammatical audits reveal specification sensitivity, while retrieval and behavioral audits distinguish candidate coverage, causal ranking, and decoded outputs. Our multi-step generation audit does not establish reliable target-mention gains, with low baseline instruction adherence limiting interpretation. TSM thus provides a prompt-conditioned selection criterion before downstream steering evaluation.

AI USE STATEMENT

We have used AI to correct grammar, assist with writing and coding, and polish this manuscript. All code and the edited sections of the articles have been checked manually. We take responsibility for the final text, claims, code, figures, and artifacts.

ETHICS STATEMENT

Steering methods can support debugging and interpretability, but may also be misused to alter intended model behavior. We do not evaluate jailbreak resistance, policy compliance, or harmful-content refusal and infer no safety guarantee from topical probability control. Releases should respect model-provider terms; this work is not a turnkey behavioral control system.

REPRODUCIBILITY STATEMENT

Appendices A–C specify the objective, token/prompt construction, and evaluation protocols. Appendix E identifies the model/SAE checkpoints, seeds, fixed inputs, and evaluation and analysis entry points. Detailed implementations and protocols are provided in the supplementary material.

REFERENCES

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering–if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10252–10270, 2025.

Trenton Bricken et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Jianrong Ding, Muxi Chen, Chenchen Zhao, and Qiang Xu. Concept-SAE: Active causal probing of visual model behavior. *arXiv preprint arXiv:2509.22015*, 2025.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Alex Gulko, Yusen Peng, and Sachin Kumar. CE-Bench: Towards a reliable contrastive evaluation benchmark of interpretability of sparse autoencoders. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 1–15, 2025.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama Scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. *arXiv preprint arXiv:2502.16681*, 2025.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *arXiv preprint arXiv:2503.09532*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530, 2023.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen blog and Hugging Face model cards, February 2026a. URL `https://qwen.ai/blog?id=qwen3.5`. Qwen3.5-2B-Base model card accessed 2026-05-05.

Qwen Team. Qwen-Scope: Turning sparse features into development tools for large language models. Technical report and Hugging Face model cards, April 2026b. URL `https://qianwen-res.oss-accelerate.aliyuncs.com/qwen-scope/Qwen_Scope.pdf`. Qwen-Scope Hugging Face collection accessed 2026-05-05.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, 2022.

Adly Templeton. *Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet*. Anthropic, 2024.

Eric Todd, Millicent Li, Arnab Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*. ICLR, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Derivations and Claim Boundaries

This section derives the target-selection margin, its local logit-space interpretation, and the distinction between relative probability control and greedy target emission.

Set-mass identity. For a token set S, write

$$M_S^\theta(x) = \sum_{v \in S} p_\theta(v \mid x). \tag{10}$$

Define $\mathrm{LSE}_{v \in S}\, a_v = \log \sum_{v \in S} \exp(a_v)$. Then $\mathrm{LSE}_{v \in S} \log p_\theta(v \mid x) = \log M_S^\theta(x)$, so the target and distractor log-mass gains are

$$G_T(i, c, x) = \log M_{T_c}^{\theta,i}(x) - \log M_{T_c}^\theta(x), \tag{11}$$

$$G_D(i, c, x) = \log M_{D_c}^{\theta,i}(x) - \log M_{D_c}^\theta(x). \tag{12}$$

Therefore the margin gain is exactly

$$G_M(i, c, x) = \log \frac{M_{T_c}^{\theta,i}(x)}{M_{D_c}^{\theta,i}(x)} - \log \frac{M_{T_c}^\theta(x)}{M_{D_c}^\theta(x)}. \tag{13}$$

Thus TSM ranks features by the change in the log target-to-distractor probability-mass ratio, averaged over selection prompts. Multiplying both masses by the same factor leaves the margin unchanged; equal absolute increases generally do not. Conversely, the ratio can rise even if both masses fall, provided distractor mass falls by a larger proportion. Positive margin gain is therefore a relative effect, not a guarantee of increased absolute target probability.

First-order logit-space view. Let $z_v(x)$ be the unsteered next-token logit and let a feature intervention induce a prompt-dependent logit perturbation $\delta_v(x)$, so that the steered logits are $z_v(x) + \delta_v(x)$. For a set S, define the within-set distribution

$$q_S(v \mid x) = \frac{p_\theta(v \mid x)\mathbf{1}[v \in S]}{M_S^\theta(x)}. \tag{14}$$

A first-order expansion of $\log M_S(x)$ around the unsteered logits gives

$$\Delta \log M_S(x) = \sum_v q_S(v \mid x)\delta_v(x) - \sum_v p_\theta(v \mid x)\delta_v(x) + O(\|\delta(x)\|^2). \tag{15}$$

Subtracting the distractor expansion from the target expansion cancels the full-vocabulary normalization term:

$$G_M(i, c, x) = \langle q_{T_c}(\cdot \mid x) - q_{D_c}(\cdot \mid x), \delta_i(\cdot, x)\rangle + O(\|\delta_i(x)\|^2). \tag{16}$$

Equation 16 is the local geometric interpretation of the contrast. The selector measures the component of the feature's causal logit perturbation along a zero-sum target-versus-distractor contrast vector. Common logit shifts and equal target/distractor log-mass changes cancel; arbitrary confidence changes need not cancel. Target-mass selection lacks this cancellation and can reward features whose main effect is generic probability redistribution rather than concept-specific control.

Why margin control need not imply greedy control. A positive $G_M(i, c, x)$ means that the target-to-distractor probability ratio increases. Greedy target emission instead requires a separate full-vocabulary condition:

$$\max_{t \in T_c} z'_t(x) > \max_{v \notin T_c} z'_v(x). \tag{17}$$

These are different inequalities. A feature can increase $M_{T_c}^{\theta,i}(x)/M_{D_c}^{\theta,i}(x)$ while a high-prior continuation outside both sets remains the largest logit. This formalizes the generation masking effect measured in the experiments: probability-margin control is a causal feature-utility signal, but open-vocabulary top-1 behavior also depends on competition against every other continuation and on autoregressive priors.

A.1 Baseline and Metric Details

The matched comparisons use a common candidate pool and held-out intervention protocol. Random mean averages random top-K subsets. Decoder-logit margin uses Eq. 2, projecting d_i rather than $\alpha_i d_i$; feature-specific calibration enters the shared interventions, not the static ranking. Target-mass selection uses TSM's calibrated final-position interventions but drops distractor subtraction. The retrieval audit uses the full dictionary's top-10 features by mean held-out margin as an oracle for recall, not as a deployable selection baseline (Appendix C.8).

The auxiliary generic robustness selector ranks mean confidence change minus a variance penalty. SAE4Steer-style Delta Token Confidence (DTC) ranks selection-prompt top-k confidence gain; Output Score instead measures gains on a feature's own logit-lens representative tokens (Arad et al., 2025). Generic robustness and DTC reuse prompt-wide selection scores and are evaluated at the final position with the main caps. Output Score was unavailable in that cache and is not evaluated. These auxiliary results do not establish superiority over fully matched output-utility baselines.

For the non-SAE dense baseline, CAA constructs a concept-level residual-stream direction

$$v_c = \mathbb{E}_{x^+ \in X_c^+} h_\ell(x^+) - \mathbb{E}_{x^- \in X_c^-} h_\ell(x^-), \tag{18}$$

where X_c^+ and X_c^- are selection-prompt completions formed from the same target and distractor strings used by TSM. The supplementary CAA study unit-normalizes v_c, selects scale using selection prompts, and evaluates on held-out prompts using prompt-wide interventions (Appendix D). It asks about dense versus sparse control, not SAE feature ranking; no final-position CAA superiority claim follows from it.

The tables report held-out target-versus-distractor log-probability margin gain as the primary endpoint. Win counts compare TSM against decoder-logit selection over matched model, concept, candidate-pool draw, and K rows. Top-1 change reports how often the open-vocabulary next-token argmax changes after intervention; it is a disruption diagnostic, not a success metric.

B Benchmark Construction

These tables document the reference token interface and distinguish the converted concept-description pilot from the external benchmarks' original tasks. Numerical final-position evidence is collected in Appendix C.

Table 6: Positioning relative to external SAE and steering benchmarks. The paper's numerical claim is a matched feature-selector comparison under a probability-margin endpoint, not task-level SOTA against prompting or finetuning.

Benchmark / work	Primary endpoint	How it informs this paper	Boundary of our claim
AxBench (Wu et al., 2025)	Broad concept steering and detection, including behavioral endpoints where prompting and finetuning are strong references.	We convert 100 external concept descriptions into target/distractor token sets for a decoder-prefiltered causal-reranking pilot on Qwen3-1.7B.	This is not the AxBench behavioral evaluation suite, its judging protocol, or a leaderboard comparison.
SAEBench (Karvonen et al., 2025)	SAE quality evaluation across proxy, interpretability, and downstream axes.	Motivates the need to test causal utility directly rather than relying on static SAE quality or interpretability proxies.	Our experiments do not rank SAE dictionaries globally; they rank features within matched candidate pools for each model.
SAE4Steer (Arad et al., 2025)	Evidence that features which explain a concept are not necessarily the features that steer outputs.	Provides the closest motivation and the external concept-description source used in the 100-concept stress test.	We study transfer of a target-aware selector under prompt and target-token shift, rather than reproducing the stock SAE4Steer behavioral pipeline.
This paper	Held-out target-over-distractor log-probability margin under activation-calibrated SAE interventions.	Tests matched-pool ranking, disjoint lexical splits, concept noise, adjective/verb contrasts, dual-oracle retrieval recall, paired decisions, and multi-step generation.	Supports conditional probability-margin feature selection; retrieval and open-generation success remain separate, unresolved requirements.

Table 7: The 18 fixed concepts used in the main benchmark. The table lists every concept and representative target and distractor strings, shown without their invisible leading-space prefix for readability; the artifact contains the complete token-string lists and the tokenizer-derived first-token ids used for each model.

Concept	Representative targets	Representative distractors
military_security	military, weapon, army, soldiers, security, defense	Android, database, insurance, Europe, exception, transport
mobile_android	Android, iPhone, phone, mobile, app, smartphone	military, database, insurance, Europe, exception, transport
software_database	database, SQL, network, deploy, server, API	military, Android, insurance, Europe, exception, transport
insurance_legal	insurance, insurer, policy, claim, legal, court	military, Android, database, Europe, exception, transport
error_exception	error, exception, failed, failure, bug, traceback	military, Android, database, insurance, Europe, transport
europe_transport	Europe, European, France, Germany, transport, train	military, Android, database, insurance, exception, army
medicine_health	medicine, medical, health, hospital, doctor, patient	military, Android, database, insurance, music, climate
finance_business	finance, market, stock, bank, investment, revenue	military, Android, disease, Europe, exception, music
education_learning	education, school, student, teacher, classroom, learning	military, Android, database, insurance, disease, stock
climate_environment	climate, environment, carbon, emissions, pollution, renewable	military, Android, database, insurance, hospital, stock
politics_government	politics, government, election, president, parliament, democracy	Android, database, hospital, stock, classroom, climate
sports_game	sport, football, soccer, basketball, game, team	military, Android, database, insurance, hospital, stock
cooking_food	food, cooking, recipe, kitchen, meal, dinner	military, Android, database, insurance, hospital, stock
music_audio	music, audio, song, album, guitar, concert	military, Android, database, insurance, hospital, stock
science_research	science, research, experiment, laboratory, theory, physics	military, Android, insurance, stock, football, recipe
art_design	art, artist, design, painting, sculpture, museum	military, Android, database, insurance, hospital, stock
space_astronomy	space, planet, star, galaxy, orbit, telescope	military, Android, database, insurance, hospital, stock
religion_philosophy	religion, philosophy, ethics, moral, belief, faith	Android, database, hospital, stock, football, recipe

Table 8: Protocol construction details that affect interpretation. These choices are fixed before scoring and are shared across selectors.

Component	Construction	Interpretation
Fixed-concept token sets	The `expanded18` benchmark encodes manually specified strings and retains their first subtoken IDs, deduplicates each set, and removes target IDs from distractors. Robustness audits split IDs into complementary halves and edit selection only.	A first subtoken is a next-token readout, not necessarily a complete lexical item.
Prompt templates	Selection: `The topic is`; `This passage is about`; `The report mentions`; `A relevant keyword is`. Held-out: `The article discusses`; `The main subject is`; `One important term is`; `The next word should be`.	Four generic short prefixes per split; no selection template is reused for held-out evaluation.
External concept100 manifest	The stress test uses the first 100 Qwen2.5-3B concept descriptions drawn from AxBench and SAE4Steer. We lowercase description words, remove generic stopwords, keep alphabetic words of length at least four, split hyphenated words by their longest part, choose up to eight target words by description-level inverse document frequency, and draw distractors from other concepts' target words. Tokens are evaluated with leading spaces.	This automatically constructed token interface supplies the 100-concept pilot and its 20-concept retrieval subset.
Candidate feature pools	Each model family uses three 512-feature candidate pools. All selectors see the same pool, intervention scale rule, concept labels, and prompt split within a row.	Isolates feature ranking under matched candidates; separate exhaustive two-model audits measure retrieval recall.
Decoder-prefiltered reranking	For the Qwen3-1.7B concept100 retrieval run, every one of the 32,768 SAE features is scored by static decoder-logit margin for each concept. We keep the union of the top 32 features per concept, giving 2,129 candidates, then apply the same activation-calibrated causal selector.	Causal ranking is conditional on this retrieved union; full-dictionary coverage is evaluated separately.
Activation calibration	All token positions of 22 fixed calibration texts supply unsteered states for positive-activation q95 scales, median normalization to 5.0, and capping. These texts include all four selection prompts and sentences extending the four held-out prefixes. Scales are shared across selectors.	Calibration uses no intervention utilities or generated outputs, but its texts are not disjoint from the benchmark prompts.
Statistics	Paired comparisons match model, pool draw, concept, and K. Main intervals resample concepts and nested pools; token audits additionally resample lexical/noise draws. Intervals are pointwise 95%.	Rows are not independent samples from arbitrary user tasks; they are controlled concept-pool evaluations.

C Final-Position Robustness and Endpoint Audits

Held-out interventions in this section use the final position; the cached generic selectors noted below are the exception at selection time. Intervals are pointwise 95% bootstrap intervals, not simultaneous guarantees across all cells. The main comparisons resample concepts and then a pool draw within each sampled concept. Token-interface experiments additionally resample lexical splits and corruption draws; behavioral comparisons preserve paired prompt cases within concept/pool clusters. Tables retain heterogeneous and null results as well as positive ones.

C.1 Matched selector baselines

The following tables evaluate the same pools and final-position caps as the main results. TSM, target mass, and their contrastive variants are scored with the final-position protocol. Generic robustness and Delta Token Confidence (DTC), marked †, reuse selection scores from the prompt-wide cache and are frozen-selector transfer diagnostics. Output Score was not evaluated under this protocol because its selection scores were unavailable in that cache. The alternative contrastive selectors subtract half the selection-prompt standard deviation (Mean–SD) or prioritize positive-effect frequency before mean margin (Pos.–mean). Their proximity to TSM limits any claim of unique formula optimality.

Table 9: Final-position baseline means, $K = 1$. Each entry averages 54 concept–pool rows; raw scales are comparable within, not across, models.

Model	TSM	Dec.	Mass	Random	Robust.†	DTC†	Mean–SD	Pos.–mean
Qwen3-1.7B	0.108	0.060	0.065	-0.001	-0.001	-0.001	0.103	0.104
Qwen3-8B	0.112	0.067	0.085	0.003	0.007	0.002	0.106	0.108
Qwen3.5-2B	0.807	0.402	0.545	-0.006	0.043	0.026	0.823	0.803
Gemma-2-2B	0.155	0.074	0.103	0.001	0.008	0.005	0.159	0.154
Llama-3.1-8B	0.595	0.088	0.499	-0.014	-0.043	-0.068	0.586	0.589

Table 10: Final-position baseline means, $K = 5$. Each entry averages 54 concept–pool rows; raw scales are comparable within, not across, models.

Model	TSM	Dec.	Mass	Random	Robust.†	DTC†	Mean–SD	Pos.–mean
Qwen3-1.7B	0.074	0.032	0.037	-0.001	0.001	0.004	0.073	0.070
Qwen3-8B	0.081	0.036	0.050	0.003	0.001	-0.001	0.077	0.075
Qwen3.5-2B	0.516	0.171	0.337	-0.008	-0.019	-0.024	0.505	0.503
Gemma-2-2B	0.123	0.044	0.082	0.001	0.008	0.008	0.112	0.109
Llama-3.1-8B	0.420	0.054	0.297	-0.013	-0.018	-0.028	0.415	0.409

Table 11: Final-position baseline means, $K = 10$. Each entry averages 54 concept–pool rows; raw scales are comparable within, not across, models.

Model	TSM	Dec.	Mass	Random	Robust.†	DTC†	Mean–SD	Pos.–mean
Qwen3-1.7B	0.060	0.024	0.025	-0.001	0.001	0.001	0.059	0.057
Qwen3-8B	0.068	0.027	0.040	0.003	0.002	0.001	0.067	0.063
Qwen3.5-2B	0.388	0.124	0.225	-0.007	-0.011	-0.011	0.383	0.376
Gemma-2-2B	0.099	0.033	0.064	0.001	-0.000	0.001	0.092	0.086
Llama-3.1-8B	0.335	0.045	0.216	-0.014	-0.019	-0.015	0.325	0.308

Table 12: Final-position baseline means, $K = 20$. Each entry averages 54 concept–pool rows; raw scales are comparable within, not across, models.

Model	TSM	Dec.	Mass	Random	Robust.†	DTC†	Mean–SD	Pos.–mean
Qwen3-1.7B	0.045	0.019	0.017	-0.002	0.001	0.002	0.044	0.040
Qwen3-8B	0.056	0.021	0.032	0.003	0.001	0.003	0.055	0.051
Qwen3.5-2B	0.289	0.083	0.152	-0.007	-0.011	-0.007	0.284	0.273
Gemma-2-2B	0.078	0.023	0.049	0.001	0.004	0.004	0.072	0.068
Llama-3.1-8B	0.245	0.039	0.151	-0.014	-0.016	-0.012	0.237	0.216

C.2 Token resampling and selection-specification corruption

For each of 18 concepts and three matched 512-feature pools, five random complementary splits separate both target and distractor token IDs into selection and clean evaluation halves. Only selection sets are edited; evaluation uses disjoint clean tokens and held-out prompts. The decoder baseline receives exactly the same edited specification. Each corruption family/level uses five draws per split (1,350 paired cells per model). This combines lexical generalization with specification noise rather than evaluating on surviving selection tokens.

Here $q \in \{1, 2, 3, 4\}$ counts token edits, not a percentage. Addition retains all original memberships; replacement removes q original memberships. Overlap makes $2q$ membership additions, so identical q values are not equal corruption fractions across families. Semantic neighbors and far concepts come from a frozen manually defined graph, not an automatic semantic oracle.

Table 13: Selection-side edits. T and D denote the selection target and distractor sets; clean held-out sets never change.

Experiment	Edit
Clean resampling	Random complementary T and D halves; no invalid tokens.
Ambig. D add	Add q target tokens from semantic-neighbor concepts to D.
Ambig. D replace	Replace q D tokens with semantic-neighbor target tokens.
Off-concept T add	Add q tokens from far concepts to T.
Off-concept T replace	Replace q valid T tokens with far-concept tokens.
T/D overlap	Copy q T tokens into D and q D tokens into T ($2q$ edits).

Retention $R = 100(V_{\text{noisy}} - V_{\text{random}})/(V_{\text{clean}} - V_{\text{random}})$ is reported as a percentage, where V is aggregate held-out utility of the selected features and the denominator uses the matched clean-split reference. ρ and overlap compare noisy and clean selection rankings/top-10 lists, not held-out utility. The five-family average weights families equally within each sampled cell. All 25 model–family intervals at $q = 2$ exclude zero for TSM minus decoder. At $q = 4$, target replacement retains only 35.9–38.5% of above-random utility and only one of five model intervals excludes zero. Thus the aggregate advantage does not imply robustness to every severely misspecified target set.

Table 14: Qwen3-1.7B: complete $q = 1$–4 specification audit, $K = 10$. R is percent retention above random; intervals are hierarchical. Clean is lexical resampling without corruption.

Selection edit	q	TSM	Dec.	Δ [95% CI]	R [95% CI]	ρ	Overlap
Clean split	–	0.036	0.016	0.020 [0.010, 0.030]	100.0 [100.0, 100.0]	1.00	1.00
Five-family mean	1	0.032	0.015	0.017 [0.009, 0.027]	90.1 [81.6, 98.4]	0.88	0.77
Five-family mean	2	0.030	0.014	0.016 [0.008, 0.025]	83.4 [73.1, 94.2]	0.79	0.65
Five-family mean	3	0.026	0.012	0.014 [0.007, 0.022]	74.3 [62.9, 85.5]	0.71	0.56
Five-family mean	4	0.023	0.011	0.013 [0.005, 0.020]	65.7 [53.4, 77.9]	0.63	0.48
Ambig. D add	1	0.034	0.016	0.018 [0.008, 0.028]	93.8 [80.1, 106.4]	0.89	0.80
Ambig. D add	2	0.033	0.016	0.017 [0.007, 0.027]	91.8 [75.6, 108.4]	0.84	0.71
Ambig. D add	3	0.032	0.015	0.017 [0.007, 0.027]	90.0 [73.7, 107.0]	0.79	0.65
Ambig. D add	4	0.031	0.016	0.016 [0.006, 0.026]	87.7 [70.2, 105.7]	0.76	0.61
Ambig. D replace	1	0.034	0.015	0.018 [0.008, 0.029]	93.3 [77.7, 110.0]	0.84	0.71
Ambig. D replace	2	0.031	0.015	0.017 [0.007, 0.027]	86.9 [68.7, 106.1]	0.72	0.58
Ambig. D replace	3	0.028	0.013	0.015 [0.005, 0.025]	79.2 [58.6, 101.9]	0.62	0.47
Ambig. D replace	4	0.025	0.013	0.013 [0.002, 0.023]	70.9 [48.6, 94.2]	0.52	0.39
Off-concept T add	1	0.031	0.015	0.016 [0.005, 0.027]	86.8 [68.9, 100.9]	0.91	0.81
Off-concept T add	2	0.029	0.013	0.016 [0.005, 0.027]	81.7 [62.4, 97.2]	0.88	0.75
Off-concept T add	3	0.027	0.012	0.015 [0.004, 0.026]	76.1 [53.6, 94.8]	0.84	0.69
Off-concept T add	4	0.024	0.011	0.013 [0.001, 0.025]	68.4 [44.2, 88.5]	0.81	0.64
Off-concept T replace	1	0.031	0.014	0.018 [0.007, 0.029]	87.5 [69.4, 103.4]	0.88	0.76
Off-concept T replace	2	0.027	0.011	0.016 [0.005, 0.027]	75.2 [52.1, 96.2]	0.78	0.61
Off-concept T replace	3	0.020	0.007	0.012 [0.000, 0.024]	56.3 [29.1, 79.0]	0.69	0.51
Off-concept T replace	4	0.013	0.004	0.008 [-0.003, 0.020]	37.7 [8.8, 64.3]	0.59	0.40
T/D overlap	1	0.032	0.015	0.017 [0.006, 0.028]	89.0 [71.4, 104.3]	0.87	0.74
T/D overlap	2	0.029	0.014	0.015 [0.005, 0.026]	81.3 [61.1, 101.6]	0.74	0.59
T/D overlap	3	0.025	0.012	0.013 [0.003, 0.023]	70.0 [47.1, 91.7]	0.62	0.48
T/D overlap	4	0.023	0.009	0.014 [0.003, 0.024]	63.9 [40.1, 88.2]	0.48	0.37

Table 15: Qwen3-8B: complete $q = 1$–4 specification audit, $K = 10$. R is percent retention above random; intervals are hierarchical. Clean is lexical resampling without corruption.

Selection edit	q	TSM	Dec.	Δ [95% CI]	R [95% CI]	ρ	Overlap
Clean split	–	0.039	0.017	0.022 [0.010, 0.035]	100.0 [100.0, 100.0]	1.00	1.00
Five-family mean	1	0.035	0.016	0.019 [0.009, 0.031]	89.7 [79.9, 99.4]	0.87	0.74
Five-family mean	2	0.032	0.015	0.018 [0.008, 0.028]	81.8 [69.9, 93.7]	0.78	0.62
Five-family mean	3	0.029	0.013	0.016 [0.007, 0.026]	74.6 [61.7, 86.8]	0.70	0.53
Five-family mean	4	0.026	0.012	0.014 [0.006, 0.023]	66.3 [52.6, 79.6]	0.62	0.45
Ambig. D add	1	0.038	0.017	0.021 [0.008, 0.033]	96.8 [82.5, 110.5]	0.90	0.80
Ambig. D add	2	0.037	0.017	0.021 [0.008, 0.033]	95.3 [78.2, 112.8]	0.85	0.72
Ambig. D add	3	0.037	0.017	0.020 [0.008, 0.032]	94.8 [77.0, 113.2]	0.80	0.65
Ambig. D add	4	0.036	0.017	0.019 [0.007, 0.031]	91.6 [73.4, 110.1]	0.77	0.61
Ambig. D replace	1	0.037	0.016	0.021 [0.009, 0.034]	95.7 [80.2, 112.1]	0.84	0.70
Ambig. D replace	2	0.034	0.015	0.019 [0.007, 0.031]	87.5 [67.5, 108.0]	0.73	0.56
Ambig. D replace	3	0.033	0.014	0.018 [0.006, 0.030]	83.2 [60.3, 107.3]	0.61	0.45
Ambig. D replace	4	0.030	0.014	0.016 [0.004, 0.028]	75.2 [50.7, 101.2]	0.50	0.36
Off-concept T add	1	0.033	0.015	0.018 [0.005, 0.031]	84.8 [62.4, 100.7]	0.89	0.79
Off-concept T add	2	0.030	0.014	0.016 [0.003, 0.029]	76.5 [53.7, 94.7]	0.86	0.72
Off-concept T add	3	0.029	0.013	0.015 [0.003, 0.029]	72.5 [47.8, 92.7]	0.82	0.65
Off-concept T add	4	0.026	0.013	0.013 [-0.000, 0.026]	64.6 [37.6, 86.6]	0.78	0.61
Off-concept T replace	1	0.033	0.015	0.018 [0.005, 0.032]	83.1 [59.2, 101.8]	0.86	0.72
Off-concept T replace	2	0.028	0.013	0.015 [0.003, 0.028]	70.6 [44.5, 92.3]	0.76	0.58
Off-concept T replace	3	0.022	0.009	0.013 [-0.000, 0.026]	55.1 [25.3, 80.0]	0.66	0.47
Off-concept T replace	4	0.016	0.007	0.009 [-0.003, 0.022]	38.5 [7.7, 66.8]	0.57	0.36
T/D overlap	1	0.034	0.015	0.019 [0.007, 0.032]	87.9 [69.5, 106.6]	0.85	0.70
T/D overlap	2	0.031	0.014	0.017 [0.005, 0.030]	79.0 [56.1, 103.3]	0.71	0.53
T/D overlap	3	0.027	0.012	0.015 [0.003, 0.026]	67.2 [43.8, 92.3]	0.59	0.41
T/D overlap	4	0.025	0.010	0.015 [0.003, 0.027]	61.4 [35.5, 88.5]	0.46	0.32

Table 16: Qwen3.5-2B: complete $q = 1$–4 specification audit, $K = 10$. R is percent retention above random; intervals are hierarchical. Clean is lexical resampling without corruption.

Selection edit	q	TSM	Dec.	Δ [95% CI]	R [95% CI]	ρ	Overlap
Clean split	–	0.221	0.063	0.159 [0.094, 0.227]	100.0 [100.0, 100.0]	1.00	1.00
Five-family mean	1	0.199	0.056	0.143 [0.084, 0.207]	90.3 [80.6, 98.5]	0.87	0.77
Five-family mean	2	0.182	0.049	0.133 [0.075, 0.194]	82.6 [69.7, 93.8]	0.78	0.66
Five-family mean	3	0.160	0.041	0.119 [0.063, 0.177]	73.1 [58.3, 85.8]	0.69	0.58
Five-family mean	4	0.140	0.034	0.106 [0.054, 0.163]	64.3 [48.7, 77.5]	0.61	0.51
Ambig. D add	1	0.209	0.060	0.149 [0.083, 0.219]	94.6 [81.6, 106.4]	0.89	0.81
Ambig. D add	2	0.200	0.055	0.145 [0.077, 0.214]	90.7 [74.0, 107.2]	0.83	0.73
Ambig. D add	3	0.195	0.052	0.143 [0.075, 0.214]	88.3 [70.6, 105.9]	0.78	0.67
Ambig. D add	4	0.187	0.052	0.136 [0.069, 0.208]	85.1 [66.7, 103.6]	0.75	0.64
Ambig. D replace	1	0.207	0.055	0.152 [0.085, 0.223]	93.7 [77.8, 109.6]	0.83	0.73
Ambig. D replace	2	0.189	0.049	0.140 [0.071, 0.214]	85.8 [65.3, 105.8]	0.73	0.62
Ambig. D replace	3	0.171	0.041	0.130 [0.065, 0.201]	78.0 [56.1, 99.5]	0.61	0.53
Ambig. D replace	4	0.147	0.036	0.111 [0.043, 0.182]	67.3 [42.4, 91.1]	0.52	0.45
Off-concept T add	1	0.188	0.056	0.132 [0.063, 0.206]	85.5 [65.5, 100.0]	0.89	0.82
Off-concept T add	2	0.179	0.050	0.129 [0.056, 0.206]	81.4 [58.6, 97.8]	0.86	0.76
Off-concept T add	3	0.161	0.046	0.115 [0.037, 0.196]	73.7 [46.7, 93.5]	0.81	0.69
Off-concept T add	4	0.146	0.043	0.103 [0.026, 0.183]	66.9 [37.4, 88.1]	0.77	0.65
Off-concept T replace	1	0.191	0.053	0.138 [0.064, 0.212]	86.8 [64.8, 103.3]	0.85	0.75
Off-concept T replace	2	0.154	0.041	0.113 [0.037, 0.192]	70.3 [43.4, 92.4]	0.74	0.60
Off-concept T replace	3	0.110	0.028	0.082 [0.000, 0.162]	51.3 [16.2, 77.2]	0.64	0.50
Off-concept T replace	4	0.075	0.011	0.064 [-0.013, 0.147]	35.9 [2.2, 64.4]	0.53	0.39
T/D overlap	1	0.200	0.055	0.145 [0.075, 0.220]	90.8 [72.0, 106.5]	0.86	0.75
T/D overlap	2	0.187	0.049	0.137 [0.067, 0.211]	84.7 [61.1, 107.7]	0.73	0.61
T/D overlap	3	0.162	0.037	0.125 [0.049, 0.202]	74.0 [46.4, 99.6]	0.62	0.51
T/D overlap	4	0.144	0.029	0.115 [0.044, 0.186]	66.1 [38.5, 94.5]	0.48	0.40

Table 17: Gemma-2-2B: complete $q = 1$–4 specification audit, $K = 10$. R is percent retention above random; intervals are hierarchical. Clean is lexical resampling without corruption.

Selection edit	q	TSM	Dec.	Δ [95% CI]	R [95% CI]	ρ	Overlap
Clean split	–	0.055	0.021	0.035 [0.013, 0.056]	100.0 [100.0, 100.0]	1.00	1.00
Five-family mean	1	0.048	0.019	0.028 [0.012, 0.045]	85.1 [73.0, 98.4]	0.86	0.74
Five-family mean	2	0.042	0.018	0.024 [0.009, 0.040]	74.7 [60.1, 91.2]	0.76	0.60
Five-family mean	3	0.038	0.017	0.022 [0.008, 0.036]	67.3 [51.0, 84.9]	0.68	0.52
Five-family mean	4	0.034	0.015	0.019 [0.006, 0.033]	58.5 [41.1, 77.6]	0.59	0.43
Ambig. D add	1	0.051	0.020	0.031 [0.010, 0.052]	91.8 [73.4, 109.0]	0.90	0.80
Ambig. D add	2	0.048	0.020	0.029 [0.008, 0.050]	86.0 [60.4, 110.3]	0.84	0.69
Ambig. D add	3	0.047	0.020	0.026 [0.006, 0.048]	83.2 [57.6, 108.8]	0.79	0.63
Ambig. D add	4	0.044	0.020	0.025 [0.004, 0.047]	79.0 [49.9, 107.2]	0.75	0.58
Ambig. D replace	1	0.049	0.020	0.029 [0.009, 0.050]	87.3 [64.4, 110.6]	0.83	0.69
Ambig. D replace	2	0.045	0.019	0.026 [0.005, 0.047]	79.6 [50.4, 109.7]	0.71	0.54
Ambig. D replace	3	0.042	0.017	0.024 [0.004, 0.046]	73.6 [41.6, 108.2]	0.61	0.44
Ambig. D replace	4	0.038	0.016	0.022 [0.002, 0.043]	67.2 [33.1, 105.7]	0.51	0.36
Off-concept T add	1	0.045	0.019	0.026 [0.006, 0.045]	80.1 [53.6, 102.1]	0.89	0.79
Off-concept T add	2	0.041	0.018	0.023 [0.004, 0.042]	72.7 [46.3, 98.3]	0.86	0.72
Off-concept T add	3	0.038	0.017	0.021 [0.002, 0.041]	67.4 [36.8, 92.6]	0.81	0.66
Off-concept T add	4	0.034	0.016	0.019 [-0.001, 0.039]	59.8 [29.1, 87.6]	0.77	0.60
Off-concept T replace	1	0.048	0.018	0.030 [0.010, 0.049]	85.2 [60.6, 109.1]	0.86	0.74
Off-concept T replace	2	0.038	0.016	0.022 [0.003, 0.042]	66.9 [37.5, 96.2]	0.74	0.56
Off-concept T replace	3	0.030	0.012	0.018 [-0.001, 0.038]	51.8 [17.5, 84.0]	0.66	0.47
Off-concept T replace	4	0.022	0.009	0.013 [-0.006, 0.033]	36.0 [-0.9, 70.2]	0.55	0.35
T/D overlap	1	0.045	0.019	0.026 [0.007, 0.046]	80.9 [55.8, 106.4]	0.81	0.66
T/D overlap	2	0.039	0.018	0.021 [0.002, 0.040]	68.4 [40.0, 99.1]	0.64	0.48
T/D overlap	3	0.035	0.016	0.018 [-0.001, 0.038]	60.3 [30.1, 94.2]	0.54	0.39
T/D overlap	4	0.030	0.013	0.017 [-0.002, 0.036]	50.7 [18.2, 87.8]	0.38	0.28

Table 18: Llama-3.1-8B: complete $q = 1$–4 specification audit, $K = 10$. R is percent retention above random; intervals are hierarchical. Clean is lexical resampling without corruption.

Selection edit	q	TSM	Dec.	Δ [95% CI]	R [95% CI]	ρ	Overlap
Clean split	–	0.216	0.017	0.199 [0.144, 0.256]	100.0 [100.0, 100.0]	1.00	1.00
Five-family mean	1	0.198	0.014	0.184 [0.133, 0.236]	92.2 [85.1, 99.0]	0.91	0.80
Five-family mean	2	0.185	0.012	0.173 [0.125, 0.222]	86.5 [77.4, 95.4]	0.83	0.70
Five-family mean	3	0.166	0.009	0.157 [0.111, 0.203]	78.1 [67.9, 88.5]	0.75	0.62
Five-family mean	4	0.146	0.007	0.139 [0.098, 0.183]	69.4 [58.5, 80.1]	0.66	0.54
Ambig. D add	1	0.207	0.017	0.190 [0.134, 0.248]	96.1 [86.1, 106.0]	0.93	0.85
Ambig. D add	2	0.205	0.014	0.191 [0.133, 0.249]	95.3 [83.2, 108.3]	0.89	0.77
Ambig. D add	3	0.200	0.013	0.187 [0.131, 0.244]	93.0 [79.4, 106.4]	0.84	0.72
Ambig. D add	4	0.195	0.013	0.181 [0.124, 0.241]	90.8 [76.3, 105.7]	0.81	0.68
Ambig. D replace	1	0.204	0.012	0.192 [0.136, 0.251]	95.0 [82.8, 108.1]	0.89	0.77
Ambig. D replace	2	0.192	0.010	0.183 [0.125, 0.240]	89.6 [73.8, 105.9]	0.77	0.65
Ambig. D replace	3	0.180	0.008	0.172 [0.114, 0.230]	84.3 [66.2, 103.5]	0.67	0.56
Ambig. D replace	4	0.169	0.007	0.162 [0.103, 0.219]	79.6 [60.6, 100.6]	0.53	0.47
Off-concept T add	1	0.191	0.013	0.178 [0.119, 0.237]	89.0 [72.8, 101.3]	0.93	0.84
Off-concept T add	2	0.180	0.012	0.168 [0.108, 0.226]	84.3 [66.6, 98.2]	0.90	0.78
Off-concept T add	3	0.168	0.010	0.158 [0.097, 0.218]	79.2 [60.2, 95.5]	0.86	0.73
Off-concept T add	4	0.151	0.008	0.143 [0.080, 0.204]	71.5 [50.3, 88.8]	0.84	0.69
Off-concept T replace	1	0.191	0.013	0.178 [0.118, 0.239]	89.1 [73.1, 102.2]	0.91	0.79
Off-concept T replace	2	0.164	0.011	0.153 [0.091, 0.215]	77.1 [56.8, 95.2]	0.81	0.65
Off-concept T replace	3	0.115	0.005	0.110 [0.044, 0.174]	56.0 [29.3, 78.5]	0.71	0.53
Off-concept T replace	4	0.075	0.003	0.073 [0.008, 0.138]	38.3 [11.7, 64.2]	0.60	0.41
T/D overlap	1	0.197	0.015	0.182 [0.124, 0.241]	91.9 [77.9, 104.6]	0.89	0.78
T/D overlap	2	0.184	0.011	0.173 [0.113, 0.235]	86.0 [68.6, 103.1]	0.78	0.64
T/D overlap	3	0.165	0.009	0.157 [0.092, 0.222]	77.9 [56.0, 99.7]	0.68	0.55
T/D overlap	4	0.140	0.003	0.137 [0.072, 0.201]	66.8 [41.8, 90.9]	0.54	0.43

C.3 Equal-cardinality whole-set distractor redraw

A separate experiment redraws all five selection distractors, keeping cardinality fixed. Uniform redraw samples without replacement from the other 17 concepts' target-token union; neighbor redraw samples from frozen semantic neighbors. Both exclude the current concept's original target/distractor tokens and preserve zero selection/evaluation token-ID overlap. Each model/family again has 1,350 paired cells. Uniform redraw yields paired 95% CIs entirely above zero in all five models; the harder neighbor redraw does so in four, with Gemma's CI including zero. This directly tests dependence on the original distant distractors, but not arbitrary natural-language concept definitions.

Table 19: Whole-set distractor redraw, $K = 10$. Both families replace all five distractors; this is not an additional value of q in the partial-corruption experiment.

Model	Sampling	TSM	Decoder	Δ [95% CI]	R [95% CI], %
Qwen3-1.7B	Uniform	0.0324	0.0165	0.0159 [0.0060, 0.0262]	90.1 [67.9, 117.6]
Qwen3-1.7B	Neighbor	0.0238	0.0111	0.0127 [0.0025, 0.0231]	67.2 [44.1, 91.0]
Qwen3-8B	Uniform	0.0378	0.0152	0.0226 [0.0114, 0.0340]	97.0 [72.3, 129.9]
Qwen3-8B	Neighbor	0.0277	0.0129	0.0149 [0.0033, 0.0267]	69.9 [44.2, 96.1]
Qwen3.5-2B	Uniform	0.2143	0.0513	0.1630 [0.0969, 0.2336]	96.9 [75.6, 123.1]
Qwen3.5-2B	Neighbor	0.1352	0.0328	0.1025 [0.0320, 0.1752]	62.3 [35.8, 87.1]
Gemma-2-2B	Uniform	0.0498	0.0199	0.0300 [0.0090, 0.0520]	89.3 [55.2, 135.3]
Gemma-2-2B	Neighbor	0.0326	0.0147	0.0179 [-0.0036, 0.0391]	56.2 [18.0, 96.4]
Llama-3.1-8B	Uniform	0.2095	0.0110	0.1985 [0.1382, 0.2594]	97.2 [76.9, 123.3]
Llama-3.1-8B	Neighbor	0.1560	0.0010	0.1551 [0.0963, 0.2134]	73.8 [52.8, 96.8]

C.4 Adjectives and verbs with disjoint lexical halves

We use 15 adjective and 15 base-form verb contrasts spanning six semantic domains, grammatical-slot-matched prompts, three candidate pools, and five lexical splits. Each tokenizer must retain at least

ten exact single-token entries per side before splitting; selection and evaluation use disjoint lexical halves and prompts. The six model-by-POS point estimates are positive, but only four intervals exclude zero. The two Gemma intervals cross zero, so the experiment broadens lexical/grammatical coverage without establishing uniform transfer to arbitrary morphology or multi-token concepts.

Table 20: Expanded grammatical-category audit, $K = 10$, 225 paired rows per model–POS cell (15 contrasts $\times$ 3 pools $\times$ 5 splits). Win rates are descriptive row fractions; uncertainty resamples contrast/pool/split units.

Model	POS	TSM	Decoder	Δ [95% CI]	Wins (%)
Qwen3-1.7B	adjective	0.0115	0.0022	0.0093 [0.0008, 0.0172]	75.1
Qwen3-1.7B	verb	0.0267	0.0133	0.0134 [0.0048, 0.0221]	77.8
Gemma-2-2B	adjective	0.0377	0.0242	0.0135 [-0.0044, 0.0317]	64.0
Gemma-2-2B	verb	0.0344	0.0152	0.0192 [-0.0001, 0.0394]	65.3
Llama-3.1-8B	adjective	0.1435	0.0490	0.0946 [0.0413, 0.1508]	80.4
Llama-3.1-8B	verb	0.1408	-0.0033	0.1441 [0.0906, 0.1985]	92.4

C.5 From margin gain to actual decisions

Joint vectors follow Eq. 9, without feature-specific α_i weights. The controller settings are $(K, \beta) = (10, 2)$ for Qwen3-1.7B and Gemma and $(20, 1)$ for Qwen3.5 and Llama. Each model has 216 paired cases (18 concepts $\times$ 3 pools $\times$ 4 held-out prompts). Candidate-token choice asks whether the largest-logit token in $T \cup D$ belongs to T. Set-mass choice instead asks whether $\log p(T) - \log p(D) > 0$. Transitions compare decoder-selected and TSM-selected outputs; unchanged means the same target/distractor class, not necessarily the same token. Shared operating points use the same endpoint-summary intervals here and in the scale sweep: 10,000 concept/pool-bootstrap draws, preserving the four prompt cases within each sampled pool.

Table 21: Paired decision transitions, final position. Net gain is $100(n_{D\to T} - n_{T\to D})/216$ percentage points. T→T and D→D are unchanged classes.

Model	Decision	D→T	T→D	T→T	D→D	Net pp [95% CI]
Qwen3-1.7B	Set mass	2	1	156	57	0.5 [-1.4, 2.8]
Qwen3-1.7B	Token top-1	4	1	160	51	1.4 [-1.4, 5.6]
Qwen3.5-2B	Set mass	24	0	156	36	11.1 [4.2, 18.1]
Qwen3.5-2B	Token top-1	30	0	147	39	13.9 [6.9, 20.8]
Gemma-2-2B	Set mass	2	2	130	82	0.0 [-2.8, 2.8]
Gemma-2-2B	Token top-1	2	1	123	90	0.5 [-1.4, 2.8]
Llama-3.1-8B	Set mass	22	0	130	64	10.2 [4.2, 18.1]
Llama-3.1-8B	Token top-1	23	1	120	72	10.2 [2.8, 18.1]

Full-vocabulary diagnostics measure the best target token's rank and its logit gap to the highest-logit token outside $T \cup D$. Positive log-rank improvement is $\log_{10} r_{\text{decoder}} - \log_{10} r_{\text{TSM}}$. These endpoints were not used for feature selection. They are associated with, but are not independent replications of, the paired margin changes. Neither selector emits a target token as the global next-token argmax in any of the 864 cases.

Table 22: Full-vocabulary endpoint ladder at the fixed operating points. The outside-gap change compares TSM against decoder; ρ associates paired margin gain and log-rank improvement. Gemma's outside-gap lower bound rounds to zero at the displayed precision.

Model	$\Delta \log_{10}$ rank [95% CI]	Δ outside gap [95% CI]	ρ [95% CI]
Qwen3-1.7B	0.0023 [-0.0021, 0.0069]	0.0026 [-0.0182, 0.0234]	0.536 [0.309, 0.699]
Gemma-2-2B	0.0042 [0.0000, 0.0084]	0.0191 [0.0000, 0.0365]	0.532 [0.312, 0.692]
Qwen3.5-2B	0.1449 [0.1120, 0.1804]	0.3712 [0.2821, 0.4653]	0.653 [0.472, 0.792]
Llama-3.1-8B	0.1035 [0.0744, 0.1356]	0.2207 [0.1427, 0.3107]	0.599 [0.405, 0.744]

C.6 PRESPECIFIED SCALE AND DISRUPTION TRADEOFF

The scale grid is $0.25, 0.5, 1, 2$ times each fixed joint-vector operating scale, with the same saved feature sets and prompts. Open-vocabulary target emission remains zero for both selectors at every scale. Increasing scale can improve candidate-set decisions while changing unrelated global top-1 tokens, particularly on Llama; a top-1 change alone is not target success.

Table 23: Complete four-scale sweep, 216 pairs per row. Net set decisions and global top-1 change are percentage points and percent, respectively. Top-1 change is TSM versus unsteered output; KL lists decoder/TSM mean disruption from the unsteered distribution.

Model	Scale ×	Δ margin [95% CI]	Net set pp [95% CI]	TSM change [95% CI]	KL D/T
Qwen3-1.7B	0.25	0.002 [-0.003, 0.007]	0.5 [0.0, 2.8]	0.0 [0.0, 0.0]	0.001/0.001
Qwen3-1.7B	0.5	0.003 [-0.003, 0.009]	0.0 [-2.8, 2.8]	0.0 [0.0, 0.0]	0.001/0.001
Qwen3-1.7B	1	0.009 [0.001, 0.018]	0.5 [-1.4, 2.8]	0.5 [0.0, 2.8]	0.001/0.001
Qwen3-1.7B	2	0.018 [0.003, 0.034]	1.9 [0.0, 5.6]	0.5 [0.0, 2.8]	0.002/0.002
Gemma-2-2B	0.25	0.007 [-0.006, 0.020]	1.9 [0.0, 5.6]	0.0 [0.0, 0.0]	0.000/0.000
Gemma-2-2B	0.5	0.018 [0.004, 0.032]	0.9 [-1.4, 4.2]	0.0 [0.0, 0.0]	0.000/0.000
Gemma-2-2B	1	0.034 [0.015, 0.052]	0.0 [-2.8, 2.8]	0.0 [0.0, 0.0]	0.001/0.001
Gemma-2-2B	2	0.066 [0.037, 0.095]	1.4 [-1.4, 5.6]	0.0 [0.0, 0.0]	0.001/0.001
Qwen3.5-2B	0.25	0.083 [0.066, 0.099]	2.8 [0.0, 6.9]	0.0 [0.0, 0.0]	0.001/0.002
Qwen3.5-2B	0.5	0.170 [0.139, 0.202]	7.9 [1.4, 15.3]	0.0 [0.0, 0.0]	0.002/0.003
Qwen3.5-2B	1	0.356 [0.290, 0.418]	11.1 [4.2, 18.1]	0.0 [0.0, 0.0]	0.007/0.012
Qwen3.5-2B	2	0.750 [0.618, 0.882]	13.0 [4.2, 25.0]	3.2 [0.0, 6.9]	0.025/0.048
Llama-3.1-8B	0.25	0.070 [0.056, 0.086]	0.9 [0.0, 4.2]	0.0 [0.0, 0.0]	0.001/0.001
Llama-3.1-8B	0.5	0.153 [0.124, 0.184]	3.7 [0.0, 8.3]	4.2 [0.0, 8.3]	0.002/0.005
Llama-3.1-8B	1	0.340 [0.278, 0.410]	10.2 [4.2, 18.1]	14.8 [8.3, 20.8]	0.008/0.019
Llama-3.1-8B	2	0.807 [0.668, 0.967]	25.9 [16.7, 36.1]	21.8 [18.1, 25.0]	0.029/0.087

C.7 EXTERNAL CONCEPT DESCRIPTIONS: CONDITIONAL RERANKING

The 100-concept Qwen3-1.7B pilot statically screens all 32,768 features, takes the union of per-concept top-32 decoder candidates (2,129 features), then causally reranks that union. At $K = 10$, mean held-out individual-feature utility is 0.0938 for TSM versus 0.0352 for decoder selection, with TSM winning 97/100 concepts. Causal scores are measured only within the retrieved union. In the separate joint-vector endpoint audit, the median best-target rank changes from 527 to 513.5, while open target emission remains 0.75% for both selectors. Of TSM's candidate-set target wins, 98.2% are masked by tokens outside the target–distractor sets. Thus the ranking advantage does not translate into more target emissions at this operating point.

Table 24: 100-concept joint-vector diagnostic across all tested K (400 paired prompt cases per row). Ranks are decoder/TSM medians; masking is conditional on a TSM target win within $T \cup D$.

K	Target rank D/T	Token net pp [CI]	Δ $\log_{10}$ rank [CI]	Masking % [CI]	ρ [CI]
1	531/525	0.00 [-0.75, 0.75]	-0.0007 [-0.0030, 0.0016]	98.2 [94.1, 100.0]	0.688 [0.627, 0.739]
5	531.5/521	0.50 [-0.50, 1.50]	0.0015 [-0.0011, 0.0040]	98.2 [94.3, 100.0]	0.693 [0.636, 0.745]
10	527/513.5	0.50 [0.00, 1.25]	0.0017 [-0.0010, 0.0044]	98.2 [94.3, 100.0]	0.699 [0.645, 0.746]
20	524.5/509	0.25 [-0.50, 1.00]	0.0002 [-0.0027, 0.0031]	98.2 [94.2, 100.0]	0.641 [0.573, 0.704]

C.8 TWO-MODEL EXHAUSTIVE DUAL-ORACLE RETRIEVAL AUDIT

We exhaustively score 32,768 Qwen3-1.7B and 16,384 Gemma features on 20 concepts, separately defining the top-10 selection-prompt and held-out-prompt causal oracles. Recall is the fraction of each oracle's ten features recovered by a per-concept candidate set of size M. Oracle rankings are audit-only, not available to the deployed selector. A shared-prompt causal screen reuses one of the selection oracle's four prompts; its recall is an aligned diagnostic, not unseen-prompt transfer. Four additional screening templates are disjoint from both oracle prompt sets. We report every

single/two-template combination as well as their four-template mean, rather than selecting the best template after observing recall.

The bottleneck is model forward computation for each feature–prompt intervention, approximately $O(FP)$ causal forwards. A forward can score all concept token sets, so the causal pass need not be repeated once per concept. Microbatching is memory-limited; feature sharding reduces elapsed time but not total GPU work. Measured exhaustive four-prompt scoring takes about 33.9 GPU-minutes for Qwen and 18.7 for Gemma on H100-80GB. Candidate unions across concepts can approach the full dictionary: at $M = F/8$, decoder unions contain 27,968/32,768 Qwen features and 13,860/16,384 Gemma features. Costs below combine measured screening time with projected reranking cost; they are estimates, not measured end-to-end deployment latency.

Table 25: Qwen3-1.7B: dual-oracle recall (%) at $M = 128, 1024$. Both oracle sets contain ten features per concept. Cost is estimated screening-plus-reranking GPU seconds. Intervals resample the 20 concepts.

Screen	M	Selection recall [CI]	Held-out recall [CI]	Est. GPU s
Decoder	128	24.5 [15.0, 34.0]	20.5 [12.5, 29.5]	131
Decoder	1024	42.5 [30.0, 55.5]	38.5 [26.5, 50.5]	791
Shared prompt	128	53.5 [42.0, 65.5]	41.5 [26.5, 57.0]	608
Shared prompt	1024	75.0 [66.0, 83.5]	62.5 [48.5, 76.0]	1057
Four unseen	128	27.0 [16.0, 39.5]	25.5 [14.0, 39.5]	2040
Four unseen	1024	51.0 [40.5, 62.0]	44.0 [34.0, 55.5]	2393

Table 26: Qwen3-1.7B: candidate-generation alternatives. ΔU is retrieved held-out utility minus decoder-prefilter utility, not recall. All rows use $K = 10$; the shared-prompt row has oracle-prompt overlap.

Screen	M	Sel. %	Held %	ΔU [95% CI]	Est. GPU s
Decoder mean	128	24.5	20.5	0.0000 [0.0000, 0.0000]	131
Decoder mean	1024	42.5	38.5	0.0000 [0.0000, 0.0000]	791
Decoder max token	128	24.5	22.5	0.0023 [-0.0050, 0.0102]	119
Decoder max token	1024	43.5	39.0	0.0032 [-0.0024, 0.0089]	759
Activation (norm.)	128	4.5	5.0	-0.0370 [-0.0538, -0.0216]	105
Activation (norm.)	1024	22.0	25.0	-0.0132 [-0.0223, -0.0046]	517
Activation (raw)	128	10.5	14.0	-0.0176 [-0.0281, -0.0079]	81
Activation (raw)	1024	24.0	29.5	-0.0121 [-0.0203, -0.0035]	420
Decoder + activation	128	22.0	18.5	-0.0030 [-0.0085, 0.0023]	122
Decoder + activation	1024	40.5	40.0	0.0002 [-0.0033, 0.0037]	684
Shared-prompt causal	128	53.5	41.5	0.0106 [-0.0012, 0.0231]	608
Shared-prompt causal	1024	75.0	62.5	0.0042 [-0.0045, 0.0127]	1057
Random	128	1.0	0.0	-0.0696 [-0.0971, -0.0456]	153
Random	1024	4.0	2.5	-0.0578 [-0.0813, -0.0381]	957

Table 27: Qwen3-1.7B: all unseen-template combinations, $K = 10$. Digits identify the four fixed unseen prompts; combinations average their screening scores. Recall is percent; cost is estimated GPU seconds.

Prompts	M	Selection recall [CI]	Held-out recall [CI]	ΔU [95% CI]	Est. GPU s
0	128	25.0 [15.0, 36.5]	24.5 [13.0, 38.5]	-0.0011 [-0.0162, 0.0135]	567
0	1024	52.0 [41.5, 63.0]	44.0 [33.5, 55.5]	-0.0001 [-0.0089, 0.0090]	916
1	128	10.0 [3.5, 18.5]	11.5 [5.0, 19.0]	-0.0232 [-0.0402, -0.0081]	569
1	1024	20.5 [12.0, 30.0]	25.0 [16.5, 34.5]	-0.0154 [-0.0282, -0.0035]	906
2	128	33.0 [21.0, 46.5]	34.5 [21.5, 48.5]	0.0092 [-0.0004, 0.0183]	583
2	1024	59.5 [49.0, 70.0]	54.0 [41.0, 67.0]	0.0042 [-0.0028, 0.0114]	1011
3	128	7.5 [3.5, 12.0]	7.0 [3.5, 11.0]	-0.0391 [-0.0607, -0.0196]	589
3	1024	29.0 [20.5, 38.0]	22.5 [14.5, 31.5]	-0.0181 [-0.0349, -0.0033]	948
0+1	128	15.5 [7.0, 25.0]	15.0 [7.5, 24.5]	-0.0194 [-0.0360, -0.0047]	1053
0+1	1024	33.5 [24.5, 44.0]	35.0 [25.0, 47.5]	-0.0048 [-0.0152, 0.0053]	1399
0+2	128	36.0 [26.0, 47.0]	32.5 [19.5, 46.5]	0.0115 [0.0008, 0.0227]	1058
0+2	1024	64.0 [55.5, 73.0]	55.5 [45.0, 66.5]	0.0055 [-0.0018, 0.0132]	1438
0+3	128	23.5 [15.5, 32.5]	22.0 [12.5, 33.5]	-0.0081 [-0.0239, 0.0062]	1073
0+3	1024	53.5 [43.5, 63.5]	44.0 [34.0, 55.0]	0.0028 [-0.0054, 0.0108]	1434
1+2	128	14.5 [6.0, 24.0]	16.0 [7.0, 27.0]	-0.0126 [-0.0274, 0.0006]	1056
1+2	1024	35.5 [24.0, 48.0]	36.5 [25.5, 49.0]	-0.0048 [-0.0162, 0.0053]	1407
1+3	128	10.5 [4.0, 18.5]	12.5 [7.0, 19.5]	-0.0220 [-0.0383, -0.0068]	1071
1+3	1024	31.0 [22.0, 40.5]	29.0 [20.0, 39.0]	-0.0101 [-0.0235, 0.0017]	1412
2+3	128	27.0 [17.0, 38.0]	25.5 [14.0, 38.5]	0.0004 [-0.0147, 0.0132]	1077
2+3	1024	51.0 [41.5, 61.5]	41.0 [29.0, 53.5]	-0.0014 [-0.0093, 0.0064]	1454
0+1+2+3	128	27.0 [16.0, 39.5]	25.5 [14.0, 39.5]	-0.0005 [-0.0148, 0.0114]	2040
0+1+2+3	1024	51.0 [40.5, 62.0]	44.0 [34.0, 55.5]	-0.0001 [-0.0099, 0.0088]	2393

Table 28: Gemma-2-2B: dual-oracle recall (%) at $M = 128, 1024$. Both oracle sets contain ten features per concept. Cost is estimated screening-plus-reranking GPU seconds. Intervals resample the 20 concepts.

Screen	M	Selection recall [CI]	Held-out recall [CI]	Est. GPU s
Decoder	128	16.0 [12.0, 20.0]	10.0 [6.5, 14.5]	138
Decoder	1024	42.5 [33.5, 52.0]	33.5 [26.5, 40.5]	678
Shared prompt	128	66.5 [54.5, 78.0]	70.0 [60.5, 79.0]	363
Shared prompt	1024	85.5 [77.0, 92.5]	85.0 [78.0, 91.5]	674
Four unseen	128	60.0 [49.5, 69.5]	74.5 [65.0, 83.0]	1149
Four unseen	1024	83.0 [76.0, 89.5]	91.5 [85.0, 97.0]	1469

Table 29: Gemma-2-2B: candidate-generation alternatives. ΔU is retrieved held-out utility minus decoder-prefilter utility, not recall. All rows use $K = 10$; the shared-prompt row has oracle-prompt overlap.

Screen	M	Sel. %	Held %	ΔU [95% CI]	Est. GPU s
Decoder mean	128	16.0	10.0	0.0000 [0.0000, 0.0000]	138
Decoder mean	1024	42.5	33.5	0.0000 [0.0000, 0.0000]	678
Decoder max token	128	18.0	12.0	-0.0039 [-0.0183, 0.0102]	128
Decoder max token	1024	41.5	37.0	-0.0040 [-0.0122, 0.0042]	658
Activation (norm.)	128	7.5	4.0	-0.0484 [-0.0633, -0.0329]	43
Activation (norm.)	1024	35.0	30.5	-0.0116 [-0.0292, 0.0068]	403
Activation (raw)	128	18.0	9.5	-0.0106 [-0.0333, 0.0134]	55
Activation (raw)	1024	37.5	34.5	-0.0057 [-0.0221, 0.0114]	385
Decoder + activation	128	18.0	10.0	-0.0003 [-0.0108, 0.0121]	101
Decoder + activation	1024	43.5	37.5	-0.0020 [-0.0161, 0.0124]	557
Shared-prompt causal	128	66.5	70.0	0.0905 [0.0615, 0.1222]	363
Shared-prompt causal	1024	85.5	85.0	0.0304 [0.0109, 0.0517]	674
Random	128	1.5	1.0	-0.0928 [-0.1131, -0.0749]	163
Random	1024	11.0	6.0	-0.0834 [-0.1097, -0.0584]	819

Table 30: Gemma-2-2B: all unseen-template combinations, $K = 10$. Digits identify the four fixed unseen prompts; combinations average their screening scores. Recall is percent; cost is estimated GPU seconds.

Prompts	M	Selection recall [CI]	Held-out recall [CI]	ΔU [95% CI]	Est. GPU s
0	128	43.0 [32.5, 54.0]	60.5 [48.5, 72.5]	0.0959 [0.0651, 0.1294]	346
0	1024	70.0 [59.5, 79.5]	83.0 [75.5, 90.0]	0.0341 [0.0142, 0.0545]	622
1	128	23.0 [13.5, 34.0]	18.0 [8.5, 28.5]	0.0354 [0.0039, 0.0722]	366
1	1024	39.0 [29.0, 50.0]	45.5 [35.5, 55.0]	0.0029 [-0.0175, 0.0246]	734
2	128	63.5 [52.0, 74.5]	66.0 [57.0, 75.0]	0.0919 [0.0660, 0.1206]	362
2	1024	85.0 [78.0, 91.0]	86.5 [81.0, 91.5]	0.0375 [0.0208, 0.0557]	756
3	128	14.5 [9.0, 21.0]	16.5 [11.0, 23.0]	0.0225 [-0.0037, 0.0518]	338
3	1024	40.5 [32.5, 48.5]	48.0 [39.5, 56.5]	0.0050 [-0.0095, 0.0216]	664
0+1	128	42.0 [32.0, 52.0]	53.5 [40.5, 66.5]	0.0923 [0.0594, 0.1258]	622
0+1	1024	68.5 [58.5, 78.0]	84.0 [76.0, 91.5]	0.0391 [0.0212, 0.0576]	944
0+2	128	65.5 [56.5, 73.5]	80.0 [72.5, 87.5]	0.0995 [0.0720, 0.1288]	617
0+2	1024	89.0 [83.5, 94.0]	94.0 [89.5, 98.0]	0.0366 [0.0196, 0.0551]	933
0+3	128	41.5 [30.5, 52.5]	57.5 [46.0, 69.0]	0.0907 [0.0588, 0.1235]	611
0+3	1024	72.5 [65.0, 80.0]	84.5 [76.5, 91.0]	0.0419 [0.0256, 0.0596]	900
1+2	128	59.0 [48.5, 69.0]	62.5 [53.5, 71.5]	0.0904 [0.0642, 0.1204]	626
1+2	1024	79.5 [70.0, 88.0]	82.5 [76.0, 88.5]	0.0365 [0.0193, 0.0556]	999
1+3	128	23.5 [15.0, 34.0]	28.0 [18.5, 38.5]	0.0491 [0.0193, 0.0824]	615
1+3	1024	49.0 [39.5, 59.0]	58.0 [49.0, 67.0]	0.0219 [0.0049, 0.0411]	942
2+3	128	57.0 [48.5, 65.5]	62.0 [53.0, 70.5]	0.0961 [0.0699, 0.1253]	619
2+3	1024	80.5 [72.5, 87.5]	82.0 [74.5, 89.0]	0.0379 [0.0201, 0.0571]	935
0+1+2+3	128	60.0 [49.5, 69.5]	74.5 [65.0, 83.0]	0.1041 [0.0798, 0.1306]	1149
0+1+2+3	1024	83.0 [76.0, 89.5]	91.5 [85.0, 97.0]	0.0344 [0.0180, 0.0524]	1469

The four-unseen-prompt held-out recall gain at $M = 128/1024$ is 5.0/5.5 percentage points for Qwen (intervals $[-4.5, 14.5]$ and $[-5.5, 15.5]$) versus 64.5/58.0 for Gemma ($[55, 73]$ and $[49, 67]$). The stronger Gemma outcome does not remove Qwen's prompt-sensitivity boundary. Screening causally over all features is itself expensive, so these audits diagnose retrieval rather than solve million-feature search.

C.9 Multi-step generation with blinded judging

The study uses all 18 concepts, three pools, four frozen instructions, and the joint-vector settings in Eq. 9. Instructions request one sentence of 8–20 words, two sentences of at least four words each, one sentence prefixed by `Summary:`, or a numbered two-item list. They do not name the target concept

and are supplied as raw text without a chat template. Unsteered, decoder, and TSM continuations are greedily generated for at most 40 tokens (2,592 continuations; 864 paired selector comparisons), intervening at the current final position at every step. Deterministic endpoints are case-insensitive whole-word/phrase target mention and the fraction of structural checks passed (sentence/word counts, required prefix or numbering, and bullet restrictions). Strict adherence requires all checks. A nondegenerate continuation is nonempty, has unique-word ratio at least 0.55, and repeated-trigram fraction at most 0.15. Qwen2.5-3B-Instruct receives the instruction, concept, and one anonymized continuation without selector identity, then rates relevance, instruction adherence, and fluency through constrained 0/1/2 classifications. Scores are normalized before taking their harmonic mean.

The generation design, rubrics, pairing, bootstrap, and inclusion rule were fixed before inspecting outcomes. Judge output was subsequently changed from free-form generation to constrained classification because earlier versions copied formatting placeholders; those trials are excluded. Positive evidence required a target-mention 95% CI entirely above zero and an instruction-score difference lower bound above -0.05, with judge assessments that did not contradict target improvement or instruction preservation. No model passed the target-mention requirement, so the positive-evidence rule fails independently of the judge results. We report all judge endpoints rather than use them to override this deterministic result.

Even unsteered continuations have low target mention and instruction adherence under these short structural prompts. Unlike the next-token rank diagnostics, this test also depends on instruction following and continuation quality, so its null result cannot be attributed to vocabulary competition alone. The absolute means expose this floor; the paired comparisons establish neither statistical equivalence nor a general absence of behavioral utility.

Table 31: Absolute multi-step outcomes (216 continuations per row). Base is unsteered; target mention and strict adherence are percentages. The low unsteered adherence limits the scope of a null paired result.

Model	Selector	Target (%)	Instruction	Strict (%)	Nondegenerate
Gemma-2-2B	base	2.8	0.333	0.0	0.750
Gemma-2-2B	decoder	2.8	0.333	0.0	0.750
Gemma-2-2B	tsm	3.2	0.333	0.0	0.745
Llama-3.1-8B	base	2.8	0.250	0.0	0.250
Llama-3.1-8B	decoder	1.9	0.255	0.5	0.190
Llama-3.1-8B	tsm	5.6	0.259	0.9	0.278
Qwen3.5-2B	base	1.4	0.333	0.0	0.250
Qwen3.5-2B	decoder	0.5	0.292	0.0	0.310
Qwen3.5-2B	tsm	1.4	0.298	0.5	0.315
Qwen3-1.7B	base	1.4	0.250	0.0	0.250
Qwen3-1.7B	decoder	3.7	0.407	15.7	0.602
Qwen3-1.7B	tsm	3.7	0.421	17.1	0.602

Table 32: Multi-step deterministic endpoints: TSM minus decoder, 216 pairs per model. Target mention and strict instruction adherence are in percentage points; instruction fraction and trigram repetition use their original scales. All target-mention intervals include zero.

Model	Target pp [CI]	Instruction [CI]	Strict pp [CI]	Repetition [CI]
Qwen3-1.7B	0.0 [-2.3, 2.3]	0.014 [-0.042, 0.069]	1.4 [-4.2, 6.9]	-0.007 [-0.035, 0.021]
Qwen3.5-2B	0.9 [-1.4, 3.2]	0.006 [-0.019, 0.031]	0.5 [0.0, 1.9]	-0.007 [-0.057, 0.044]
Gemma-2-2B	0.5 [0.0, 1.9]	0.000 [0.000, 0.000]	0.0 [0.0, 0.0]	-0.000 [-0.004, 0.006]
Llama-3.1-8B	3.7 [-0.5, 8.3]	0.005 [-0.011, 0.020]	0.5 [-1.4, 2.8]	-0.013 [-0.057, 0.032]

Table 33: Blinded Qwen2.5-3B-Instruct judgments on the same continuations. Paired differences and 95% hierarchical intervals are shown for each axis and their harmonic aggregate; these small, uncertain differences do not establish behavioral control.

Model	Concept [CI]	Instruction [CI]	Fluency [CI]	Harmonic [CI]
Qwen3-1.7B	-0.005 [-0.079, 0.065]	0.014 [-0.028, 0.056]	0.000 [-0.065, 0.065]	0.000 [-0.014, 0.014]
Qwen3.5-2B	0.032 [-0.060, 0.125]	0.014 [-0.079, 0.111]	0.019 [-0.139, 0.162]	0.002 [-0.032, 0.038]
Gemma-2-2B	0.014 [0.000, 0.046]	0.005 [-0.014, 0.023]	0.005 [0.000, 0.019]	0.003 [0.000, 0.014]
Llama-3.1-8B	0.009 [-0.083, 0.097]	0.014 [-0.014, 0.046]	0.065 [-0.014, 0.148]	0.005 [-0.005, 0.021]

D Supplementary Prompt-Wide Diagnostics

The following exploratory studies broadcast additive interventions over every prompt position, unlike the final-position protocol in the main text and Appendix C. They provide complementary dense-control, layer, and semantic-readability diagnostics and are interpreted within that distinct intervention protocol.

D.1 Dense-control and layer diagnostics

Under prompt-wide interventions, dense CAA often achieved larger margins than sparse feature combinations. This motivates separating sparse dictionary audit from unrestricted dense control, but does not establish the corresponding final-position frontier. The layer check likewise documents operating-point dependence within the prompt-wide protocol rather than extending the main layer-12 result.

Table 34: Supplementary prompt-wide protocol (not the final-position main experiment). Dense contrastive activation addition (CAA) baseline on the fixed-concept endpoint. CAA directions are built from the same target/distractor token strings and selection prompts as TSM, unit-normalized, and scale-selected using selection margin over a small grid before held-out evaluation. Δ is TSM–CAA with a paired-bootstrap 95% confidence interval over matched seed-concept rows.

Model	K	TSM	Decoder	CAA	Δ [95% CI]	Win	Top-1 change TSM	Top-1 change CAA
Gemma-2-2B	10	0.217	0.070	1.290	-1.073 [-1.138, -1.011]	0/54	0.001	0.000
Gemma-2-2B	20	0.165	0.049	1.290	-1.125 [-1.190, -1.061]	0/54	0.001	0.000
Qwen3-1.7B	10	0.099	0.042	0.512	-0.413 [-0.455, -0.373]	0/54	0.100	0.153
Qwen3-1.7B	20	0.076	0.032	0.512	-0.436 [-0.477, -0.395]	0/54	0.091	0.153

Table 35: Supplementary prompt-wide protocol (not the final-position main experiment). Sparse–dense Pareto points for TSM and CAA. TSM points use K sparse SAE features selected by target-selection margin; CAA points use dense residual-stream directions at fixed scales. The Gemma-2-9B check tests whether the Gemma-2-2B low-disruption sparse point generalizes within Gemma Scope; the scale grid shows isolated sparse competitiveness but no broad sparse dominance.

Model	Method	Setting	Margin	Top-1 change rate
Qwen3-1.7B	TSM	K=1	0.190	0.116
Qwen3-1.7B	TSM	K=10	0.099	0.100
Qwen3-1.7B	TSM	K=40	0.056	0.088
Qwen3-1.7B	CAA	scale=1	0.049	0.014
Qwen3-1.7B	CAA	scale=5	0.251	0.153
Gemma-2-2B	TSM	K=1	0.413	0.000
Gemma-2-2B	TSM	K=10	0.217	0.001
Gemma-2-2B	TSM	K=40	0.120	0.001
Gemma-2-2B	CAA	scale=1	0.134	0.000
Gemma-2-2B	CAA	scale=5	0.653	0.000
Gemma-2-9B	TSM	K=1	0.342	0.005
Gemma-2-9B	TSM	K=10	0.165	0.003
Gemma-2-9B	TSM	K=40	0.081	0.001
Gemma-2-9B	CAA	scale=1	0.108	0.000
Gemma-2-9B	CAA	scale=5	0.538	0.000
Qwen3-8B	TSM	K=1	0.134	0.097
Qwen3-8B	TSM	K=10	0.096	0.087
Qwen3-8B	TSM	K=40	0.078	0.082
Qwen3-8B	CAA	scale=1	0.084	0.000
Qwen3-8B	CAA	scale=5	0.400	0.069
Qwen3.5-2B	TSM	K=1	1.314	0.056
Qwen3.5-2B	TSM	K=10	0.623	0.036
Qwen3.5-2B	TSM	K=40	0.329	0.027
Qwen3.5-2B	CAA	scale=1	1.385	0.000
Qwen3.5-2B	CAA	scale=5	4.752	0.208
Llama-3.1-8B	TSM	K=1	1.379	0.162
Llama-3.1-8B	TSM	K=10	0.721	0.181
Llama-3.1-8B	TSM	K=40	0.359	0.142
Llama-3.1-8B	CAA	scale=1	0.638	0.056
Llama-3.1-8B	CAA	scale=5	3.564	0.306

Table 36: Supplementary prompt-wide protocol (not the final-position main experiment). Qwen3-8B layer-position check with 18 concepts, 512-feature pools, and an intervention cap of 10.

Layer	K	TSM	Decoder	Wins	TSM top-1 change	Held-out oracle
12	10	0.119	0.040	53/54	0.077	0.153
12	20	0.093	0.031	54/54	0.079	0.128
18	10	0.143	0.073	54/54	0.085	0.162
18	20	0.107	0.051	54/54	0.084	0.127
24	10	0.081	0.047	54/54	0.042	0.091
24	20	0.060	0.035	53/54	0.044	0.070

E REPRODUCIBILITY AND ASSETS

The following details complement the protocols in Appendices B–C. Paths are relative to the experiment repository; external assets are listed in Table 37.

Checkpoints and seeds. All main SAEs use residual-stream layer 12. The Qwen-Scope variants are `W32K-L0_100` for Qwen3-1.7B and Qwen3.5-2B, and `W64K-L0_100` for Qwen3-8B. Gemma Scope uses width_16k/average_l0_82; Llama Scope uses `L12R-8x`. The three 512-feature pools use seeds 20260501, 20260502, and 20260503. Reproduction should retain their saved feature IDs in outputs/feature_scores/, shared across selectors. Calibration follows Section 3.6: positive-activation 95th percentiles, median scale 5, and caps 10 for Qwen3-1.7B, Qwen3-8B, and Gemma, or 2 for Qwen3.5-2B and Llama.

Fixed inputs and evaluation. The 18 target/distractor specifications and four selection/four held-out prompts are defined in scripts/eval_qwen3_fixed_concept_utility.py. The main evaluator, scripts/eval_qwen3_fixed_concept_activation_calibrated.py, takes the model/SAE paths and saved feature-summary CSV, using `-concept-set expanded18`, $K \in \{1, 5, 10, 20\}$, 50 random-subset trials, and explicitly `-intervention-position final`. Supplementary specifications are stored in inputs/noisy_concept_neighbors_v1.json, inputs/cross_pos/, and inputs/prompt_sets/multistep_generation_pareto_v1.json; their respective perturbation, grammatical-category, and generation protocols remain separate from the primary ranking protocol. Prompt-wide diagnostics instead follow Appendix D.

Result aggregation. For the main ranking comparisons, scripts/analyze_fixedconcept_finalpos_hierarchical.py aggregates saved selector and per-feature evaluations and computes the $K = 10$ paired intervals with 10,000 hierarchical bootstrap draws and seed 20260825. Numerical summaries are stored in outputs/tables/; scripts/make_master_paper_tables.py assembles the numerical tables from these summaries. Supplementary analyses use the resampling units specified with their corresponding protocols.

Table 37: External assets, sources, terms, and experimental roles. Obtain model and SAE weights from their original sources under the listed terms.

Asset	Source used in experiments	License / terms	Role
Qwen3-1.7B and Qwen3-8B	Qwen3 Hugging Face repositories (Yang et al., 2025)	Apache-2.0 model-card license	Language models
Qwen3.5-2B	Qwen3.5 Hugging Face repository (Qwen Team, 2026a)	Apache-2.0 model-card license	Language model
Qwen-Scope SAEs	Qwen-Scope SAE repositories (Qwen Team, 2026b)	Qwen model-card license and Qwen-Scope caution terms	SAE dictionaries for Qwen models
Gemma-2-2B	Gemma-2-2B repository (Gemma Team et al., 2024)	Gemma Terms of Use	Base language model
Gemma Scope SAE	Gemma Scope 2B residual SAE (Lieberum et al., 2024)	CC-BY-4.0 model-card license	SAE dictionary for Gemma
Llama-3.1-8B	Llama-3.1-8B repository (Grattafiori et al., 2024)	Llama 3.1 Community License and Acceptable Use Policy	Base language model
Llama Scope SAE	Llama Scope SAE repository (He et al., 2024)	Apache-2.0 model-card license	SAE dictionary for Llama
WikiText-2	WikiText-2 raw v1 dataset (Merity et al., 2016)	CC-BY-SA-3.0 and GFDL dataset-card licenses	Real-context activation probe

F SUPPLEMENTARY PROMPT-WIDE FEATURE EXAMPLES

These exploratory examples use features selected by the prompt-wide protocol, not the final-position selections. The activation probes combine decoder-preferred tokens, a compact concept mini-corpus, and open-domain contexts. They illustrate readability and its limits rather than independently

confirming the main ranking estimates. Some selected directions still require a separate semantic-labeling step.

Table 38: Qualitative examples from the representative $K = 10$ target-selection run. Margin reports selected-feature held-out margin versus the same feature's decoder-logit margin. These examples are diagnostic rather than an additional held-out metric.

Model	Feature	Concept(s)	Margin	Qualitative cue
Qwen3-8B	26998	education	0.495 vs 0.092	Decoder tokens include scaffold, Teacher, and classroom; top activation contexts include students and student.
Qwen3-8B	44337	software/error	0.331 vs 0.030	Decoder tokens include Python and PyObject; top activation contexts include exceptions, bug, and exception.
Gemma-2-2B	4924	medicine	0.253 vs 0.002	Decoder tokens include contaminated and infected; the strongest context cue is disease.
Gemma-2-2B	6713	error	0.390 vs -0.012	Decoder tokens include message, getMessage, and messaging; activation contexts are mixed, so this is a causal-utility example rather than a clean semantic-label example.
Gemma-2-2B	9329	education	0.266 vs 0.039	Decoder tokens include skill and skills; the mini-corpus activation probe is sparse.
Llama-3.1-8B	5339	software	4.504 vs -0.007	Top activation contexts include deploy and failed, while decoder tokens are mostly fragments.
Llama-3.1-8B	13574	education, music	6.583 vs 0.006	Top activation contexts include music, songs, and piano; decoder tokens are mixed.
Llama-3.1-8B	19704	music	6.017 vs 0.007	Top activation contexts include sound, song, and music; decoder tokens are not readable enough to serve as labels.

We also run a real-context activation probe on 1500 WikiText passages. Table 39 shows a mixed but useful pattern: some selected features have readable open-domain contexts, while others reveal broader semantics or tokenization artifacts. This reinforces the limitation that target-aware utility is a causal selection criterion, not an automatic semantic-labeling method.

Table 39: Real-context activation inspection on 1500 WikiText passages. The probe is qualitative: it asks whether selected features' highest activating contexts in open-domain text match their fixed-concept labels.

Model	Feature	Concept(s)	Real-context readout
Llama-3.1-8B	19704	music	Strong positive: top contexts repeatedly involve pop, music, song, album, and sound.
Llama-3.1-8B	13574	education/music	Broader positive: top contexts involve chorus, choir, and alto, with one student context.
Gemma-2-2B	9329	education	Skill-related positive: top contexts involve skill, skills, and skilled training/teaching-adjacent passages.
Gemma-2-2B	6713	error	Related but shifted: top contexts involve message and Messages, matching decoder tokens more than the fixed error lexicon.
Qwen3-8B	26998	education	Mild positive: contexts include kindergarten, pupils, materials, and lessons, despite low lexical hit rate.
Qwen3-8B	44337	software/error	Negative artifact: decoder tokens look Python-related, but filtered real-context activations still concentrate on token fragments rather than clean software/error contexts.

Table 40 compares the filtered real-context probe under target-selection and decoder-logit selection. The target selector has the higher mean target-hit rate and a much larger mean held-out margin, while token-boundary fragment artifacts remain common for both selectors. Table 41 repeats the comparison with an expanded concept lexicon over the same local activation windows. The target selector again has a higher artifact-filtered semantic-hit rate, but the absolute rates remain modest. These results characterize prompt-wide-selected features and do not establish semantic enrichment for final-position selections.

Table 40: Filtered real-context probe on 1500 WikiText passages. The comparison uses the same open-domain corpus and the same feature pools, but changes only the feature selector. Mean target-hit is the fraction of top activation contexts whose local text matches the concept target lexicon; artifact-rate additionally counts token-boundary fragments such as subword suffixes inside longer words.

Selector	Target-hit	Distractor-hit	Artifact	Fragment	Held-out margin
target-selection margin	0.037	0.017	0.362	0.362	1.526
decoder-logit margin	0.025	0.014	0.368	0.368	0.087

Table 41: Real-context semantic-purity diagnostic on WikiText activation snippets. Semantic hit uses an expanded concept lexicon over local activation windows; filtered hit removes token-boundary fragment artifacts. The metric is a heuristic appendix diagnostic, not an automatic interpretation benchmark.

Selector	Features	Semantic hit	Filtered hit	Artifact	Held-out margin
decoder-logit margin	35	0.054	0.052	0.368	0.087
target-selection margin	30	0.154	0.152	0.362	1.526

Finally, Table 42 uses two 8B language models as stricter multiple-choice semantic judges over the same real-context windows. Both judges favor target-selected features over decoder-selected features, but the absolute match rates are low and many windows are classified as other. We therefore use this as supporting appendix evidence rather than as a primary benchmark.

Table 42: Real-context multiple-choice semantic diagnostic using two 8B judges. Each judge classifies every local WikiText activation window as one of the six target concepts or other. Filtered match removes token-boundary fragment artifacts. The diagnostic is noisy but less tied to the hand-written target lexicon than Table 41.

Judge	Selector	Choice match	Filtered match	Other top	Held-out margin
Qwen3-8B	decoder-logit margin	0.093	0.102	0.693	0.087
Qwen3-8B	target-selection margin	0.188	0.222	0.688	1.526
Llama-3.1-8B	decoder-logit margin	0.096	0.124	0.675	0.087
Llama-3.1-8B	target-selection margin	0.167	0.216	0.596	1.526