CONDITIONAL INTERVENTION SELECTION FOR EFFICIENT ACTIVATION STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering adapts frozen models by intervening on internal representations, yet conventional relevance scores do not reveal whether neurons remain useful when intervened on jointly. We formulate sparse steering-site localization as budgeted conditional selection and introduce SCOUT, which constructs a relevance-ranked working set, accepts neurons according to their measured utility conditioned on the current selection, and calibrates the resulting displacement into static neuron-wise scaling for single-forward deployment. Across the evaluated model scales, SCOUT achieves positive mean accuracy gains of up to 3.44 percentage points while selecting only 11–16 neurons on average. This corresponds to an approximately 84–99.6% reduction in intervention sites relative to baselines using 100 to 4096 sites. Selection ablations further show that conditional utility consistently improves over relevance-only and independent-utility alternatives. These results establish conditional intervention selection as an accurate and highly sparse approach to activation steering.

1 INTRODUCTION

Large language models (LLMs) solve many domain-specific tasks, while task-specific training still incurs substantial data and compute cost. Activation- and representation-level adaptation offers a lightweight alternative by editing internal states while keeping the base model frozen (Turner et al., 2023; Li et al., 2023; Zou et al., 2023; Liu et al., 2024; Wu et al., 2024; Kong et al., 2024). A common activation-steering formulation constructs guidance from positive–negative examples or task demonstrations and applies the resulting direction during inference (Rimsky et al., 2024; Stolfo et al., 2025; Wang et al., 2025; Lee et al., 2025; Sharma & Trivedi, 2026).

Existing methods commonly localize steering targets with contrastive activation differences, activation statistics, discriminative scores, or sparse features (Rimsky et al., 2024; Wang et al., 2025; Feng et al., 2026; Stoehr et al., 2024; Gao et al., 2025; Leask et al., 2025). These quantities rank representation relevance. Downstream accuracy depends on intervening on the chosen sites, and highly ranked neurons can yield zero or negative marginal gain, as illustrated in Figure 1.

We cast steering-target identification as conditional intervention selection under a bounded working set and acceptance budget. Search uses an input-dependent displacement, and held-out calibration

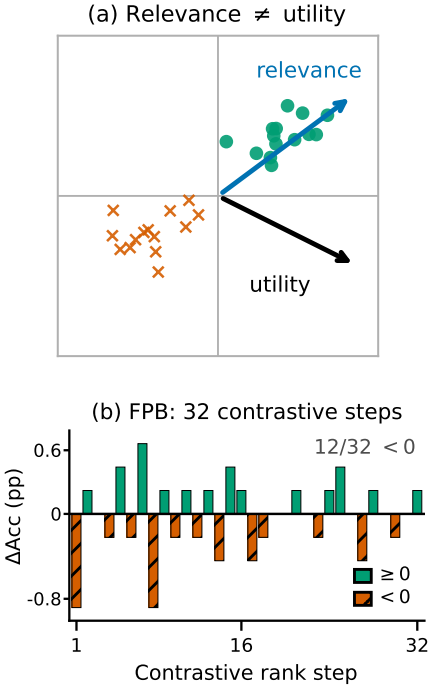


Figure 1: Relevance $\neq$ utility (top); Financial PhraseBank (FPB) contrastive marginals, $12/32{<}0$ (bottom).

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.

yields static neuron-wise scaling for single-forward deployment. On a fixed-pool Qwen3-1.7B confirmatory trio, conditional utility wins $9/9$ dataset–seed cells against contrastive, sensitivity, and random selection, and $6/9$ against independent utility. A Qwen3-0.6B suite reports positive mean SCOUT gains with a sparse selected set, and held-out calibration recovers the search displacement at near-Clean latency.

Contributions.

- **Conceptually,** we separate representation ranking from joint intervention utility and show that highly ranked neurons can have negative marginal steering effects.
- **Methodologically,** we introduce SCOUT, an ordered conditional-acceptance selector with held-out static calibration for single-forward deployment.
- **Empirically,** we provide a Qwen3-0.6B grouped suite, a fixed-pool selection ablation, held-out calibration, and an audited workbook comparison of Clean, CAA, SADI, AUSteer, and SCOUT on Qwen3-{0.6B, 1.7B, 4B, 8B}.

2 Problem Formulation and Motivation

2.1 The budgeted steering problem

Let C be a finite working set of MLP neuron coordinates and let $F(S)$ denote a search-split objective under a joint intervention on set $S \subseteq C$ (we use mean task-label margin). Under acceptance budget B, the budget-optimal target is

$$S^* = \arg \max_{S \subseteq C,\ |S| \leq B} F(S). \tag{1}$$

Exact maximization over subsets is combinatorial. Practical steering therefore needs an explicit evaluation budget, an ordered shortlist of candidates, and a measured acceptance rule that builds S from intervention feedback.

2.2 Why representation relevance misses intervention utility

Given positive and negative examples, let g_i denote a representation score of neuron i (for example a contrastive activation difference), and let S_K denote the top-K neurons under this ranking. Ranking-based steering treats larger g_i as evidence that neuron i is a better intervention target. The downstream objective, however, is task performance after intervention. The marginal change when the K-th ranked neuron is added is

$$\Delta\mathrm{Acc}(K) = \mathrm{Acc}(S_K) - \mathrm{Acc}(S_{K-1}). \tag{2}$$

On Financial PhraseBank under a contrastive ranking with $K{=}32$, twelve of thirty-two rank steps have negative marginal accuracy. Representation relevance and intervention utility therefore define different criteria, consistent with findings in model editing that localization signals and intervention locations can diverge (Hase et al., 2023; Fang et al., 2025).

Table 1: Relevance scores as intervention-utility predictors on the confirmatory trio (Qwen3-1.7B, seed 42, 256-neuron pool). Spearman(R, u) uses signed per-neuron utility; Recall@64 is recovery of the top-16 utility neurons under sensitivity vs. a random baseline.

Dataset	Spearman(R, u)	Sens. Recall@64	Rand. Recall@64
FPB	-0.170	0.188	0.245
CLINC150	-0.282	0.312	0.251
LEDGAR	-0.183	0.125	0.243

Table 1 quantifies the same gap with signed utility. Spearman correlations between sensitivity R and signed utility u are negative on the confirmatory trio, and top-sensitivity Recall@64 does not dominate a random shortlist. Sensitivity remains useful as a budgeted coverage heuristic for working-set construction, while signed utility is obtained from intervention measurement.

2.3 CONDITIONAL UTILITY AS THE SOLUTION PRINCIPLE

Sparse steering applies several interventions jointly, so the relevant quantity depends on the set already selected. Given a current set S, the conditional gain is

$$\text{gain}(c \mid S) = F(S \cup \{c\}) - F(S). \tag{3}$$

SCOUT constructs a feasible path-dependent accepted set toward the budget-optimal target in Eq. 1 by traversing an ordered candidate list and accepting neurons whose measured conditional gain is positive. This shifts the problem from ranking neurons independently to constructing a jointly useful intervention set under an explicit evaluation budget.

Proposition 1 (Telescoping under ordered acceptance)**.** *Fix an ordered working set $C = (c_1, \ldots, c_K)$, budget B, and acceptance threshold 0. Let $S_0 = \varnothing$ and let S_t be the accepted set after considering candidates $c_1, \ldots, c_t$ under the rule "accept iff $\text{gain}(c_t \mid S_{t-1}) > 0$ and $|S_{t-1}| < B$." Whenever c_t is accepted, $\text{gain}(c_t \mid S_{t-1}) > 0$, so $F(S_t) > F(S_{t-1})$. Writing $\widehat{S}$ for the final accepted set and S_c^{pre} for the accepted set immediately before each accepted c,*

$$F(\widehat{S}) - F(\varnothing) = \sum_{c \in \widehat{S}} \text{gain}(c \mid S_c^{\text{pre}}) \geq 0. \tag{4}$$

The identity holds for the measured search objective F under the fixed traversal order and the zero-threshold rule: every accepted update raises F, accepted gains telescope, and the total improvement is non-negative along that path. The proposition concerns monotonicity of accepted updates for this ordered procedure; subset optimality among all $S \subseteq C$ with $|S| \leq B$ remains combinatorial.

3 SCOUT

3.1 FRAMEWORK OVERVIEW

SCOUT has three stages: working-set construction, ordered conditional intervention acceptance, and static steering calibration (Figure 2). It operates on MLP intermediate neurons through forward hooks with frozen model parameters. Disjoint discovery, search, calibration, and evaluation splits keep construction separate from final testing. For each input, we record the MLP intermediate activation at the last non-padding token. Association scores (sensitivity) rank candidates into a working set; intervention measurements govern acceptance (Dai et al., 2022; Qin et al., 2026).

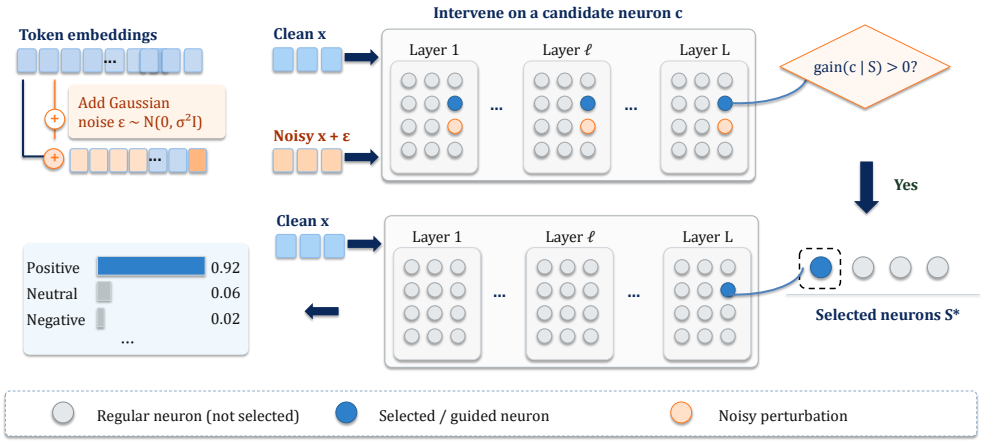


Figure 2: SCOUT overview: clean/noisy embeddings drive layer-wise neuron candidates; ordered conditional acceptance keeps sites with $\text{gain}(c \mid S) > 0$; selected neurons shift output probabilities under a sparse intervention.

3.2 WORKING-SET CONSTRUCTION

We use a lightweight perturbation-sensitivity probe to bound the working set (Qin et al., 2026; Gao et al., 2025; Leask et al., 2025; Dunefsky et al., 2024). Adding Gaussian noise to input embeddings and measuring induced activation displacement yields

$$\Delta a = a^{\text{noise}} - a, \qquad R = \mathbb{E}[|\Delta a|], \tag{5}$$

estimated on the discovery split. Neurons are ordered by R, and a top-K working set $C = (c_1, \ldots, c_K)$ is taken in decreasing sensitivity order. Audited suite, selection-ablation, and dense K/B manifests set `ranking.layer_balanced=false` (global mean-abs ranking). This score limits which candidates are evaluated; acceptance uses measured conditional utility. Protocol values of K, budget B, and strength β are listed in Appendix A.

3.3 ORDERED CONDITIONAL ACCEPTANCE

On the search split, the candidate intervention reverses the noise-induced displacement,

$$a' = a - \beta \Delta a. \tag{6}$$

Let F denote the mean task-label margin on the search split. Starting from $S \leftarrow \varnothing$, SCOUT evaluates candidates in the fixed sensitivity order $c_1, \ldots, c_K$. For each c_i, it measures the conditional gain $\text{gain}(c_i \mid S)$ under Eq. 3 with the cumulative intervention on $S \cup \{c_i\}$, and accepts c_i into S when the gain is positive. Rejected candidates leave S unchanged. Search stops when all K candidates have been considered or $|S|$ reaches budget B.

This procedure is ordered conditional acceptance: a single pass over a fixed order with accept/reject decisions against the current set (Algorithm 1). The implementation iterates indices $i = 1, \ldots, K$ once; acceptance uses $\text{gain}(c_i \mid S)$ against the current accepted set under a fixed traversal (no residual-set re-ranking). Dense K/B and the fixed-pool selection-rule ablation (Qwen3-1.7B) use legacy acceptance $\text{gain} > 0$ with the accuracy guardrail disabled. The Qwen3-0.6B suite instead uses `task_guardrail`: accept iff $\text{gain} > 0$ and $\text{trial_acc} \geq \text{current_acc}$ (accuracy tolerance 0).

Search and online complexity. Ignoring constant factors from batching, offline search performs one clean baseline evaluation plus at most K conditional evaluations of Eq. 6 on the search split. Because candidates are examined in a fixed order, the number of conditional evaluations equals the number of considered candidates (at most K). Online static SCOUT runs a normal forward pass and applies $O(B)$ neuron-site scalar corrections through hooks after coefficients are frozen. Proposition 1 records the telescoping identity for accepted gains under this ordered rule.

3.4 STATIC SCOUT CALIBRATION

The search intervention in Eq. 6 depends on input-specific Δa. SCOUT approximates each selected displacement on the calibration split by a neuron-specific scaling,

$$\Delta a_j \approx \rho_j a_j, \qquad \rho_j = \frac{P_j}{Q_j + \lambda}, \quad P_j = \sum a_j \Delta a_j, \quad Q_j = \sum a_j^2, \tag{7}$$

where the sums are over calibration examples. The deployed intervention is

$$a_j \leftarrow (1 - \beta \rho_j) a_j, \qquad j \in \widehat{S}. \tag{8}$$

Coefficients are fixed before evaluation, so test-time inference uses a single forward pass with frozen site-wise scalars. The online operator applies the frozen multipliers in Eq. 8 at the accepted sites.

Held-out fit quality is reported with an uncentered normalized squared error

$$\text{NSE} = \frac{\sum (\Delta a - \rho a)^2}{\sum (\Delta a)^2}, \tag{9}$$

matching the implementation denominator $\sum (\Delta a)^2$ (uncentered relative squared error). We write explained displacement energy (EDE) as the complementary relative-energy summary $\text{EDE} = 1 - \text{NSE}$.

Algorithm 1 Ordered conditional acceptance and static calibration.

Input: working set $C = (c_1, \ldots, c_K)$ ordered by sensitivity; budget B; strength β; search and calibration splits

Output: accepted set $\widehat{S}$; static coefficients $\{\rho_j\}_{j \in \widehat{S}}$

$S \leftarrow \varnothing$; evaluate clean baseline $F(\varnothing)$ on the search split
for $i = 1$ to K **do**
  if $|S| \geq B$ **then**
    break ▷ budget reached
  end if
  evaluate $F(S \cup \{c_i\})$ under intervention Eq. 6
  if $\mathrm{gain}(c_i \mid S) > 0$ **then**
    $S \leftarrow S \cup \{c_i\}$ ▷ accept positive gain
  end if
end for
$\widehat{S} \leftarrow S$
for all $j \in \widehat{S}$ **do**
  fit ρ_j on the calibration split via Eq. 7 ▷ freeze scale
end for
return $\widehat{S}$, $\{\rho_j\}$

4 Experiments

4.1 Setup

Tasks. Mechanism and selection experiments use Qwen3-1.7B (Yang et al., 2025) on Financial PhraseBank (Malo et al., 2014), CLINC150 (Larson et al., 2019), and LEDGAR (Tuggener et al., 2020) (confirmatory trio), with frozen parameters and hook-based interventions on multiple-choice prompts. Breadth uses a separate Qwen3-0.6B frozen suite over 14 datasets and seeds $\{42, 43, 44\}$. Workbook Acc_XB Clean and Acc_Delta matched method Δs cover Qwen3-$\{$1.7B, 4B, 8B$\}$. Section 4.5 uses a separate Qwen3-1.7B dense K/B sweep on Financial PhraseBank, Twitter, Weibo, and TREC-50. These protocols are distinct and are not interchangeable.

Baselines. The main table reports Clean, CAA (Rimsky et al., 2024), SADI (Wang et al., 2025), AUSteer (Feng et al., 2026), and SCOUT. On 0.6B, Clean and SCOUT share the suite ledger; CAA/SADI/AUSteer Δ use Acc_0.6B's own Clean (not the suite Clean). On 1.7B/4B/8B, Clean uses Acc_XB full-14 Clean(%); CAA/SADI/AUSteer Δs use Acc_Delta matched frozen-source units; SCOUT Δs use the author-provided full-14 ledger. Category columns are within-group dataset macro-means; Overall is the 14-dataset macro-mean (Appendix A.4). The 1.7B selection ablation varies only the selection rule with pool, operator, budget, calibration, and static deploy fixed.

Implementation and hyperparameters. Discovery, search, and calibration use 64, 128, and 128 stratified examples (smaller datasets may yield shorter scout splits). Defaults are K=64, B=16, and β=1. The Qwen3-0.6B suite uses `acceptance_mode=task_guardrail`: accept iff gain> 0 and trial accuracy is at least the current accepted-set accuracy (tolerance 0). The dense K/B sweep and the fixed-pool selection-rule ablation use legacy $\mathrm{gain} > 0$ (guardrail disabled). Sensitivity, Dynamic, and SCOUT share the accepted set when compared under one suite protocol. Coefficients are frozen before test evaluation. Seeds are $\{42, 43, 44\}$ unless noted. Section 4.5 uses seed 42 and grid $\{2, 5, 9, 12, 15, 19, 22, 25, 29, 32\}$ with B=2 when varying K and K=32 when varying B. Details appear in Appendix A–F.

4.2 Main Comparison Across Model Sizes

Table 2 compares Clean and steering methods. On Qwen3-0.6B, suite Clean is 60.34% Overall and SCOUT ΔAcc is $+1.34$ pp (mean $|S|$=11.02); Acc_0.6B CAA/SADI/AUSteer Δs (vs. workbook Clean 59.79%) are $+1.36$/$+0.30$/$+0.89$ pp. On Qwen3-1.7B, full-14 SCOUT is $+3.44$ pp (mean $|S|$=14.21) while matched CAA is $+4.36$ pp; on 4B/8B, full-14 SCOUT is $+0.01$/$+0.26$ pp. Acc_XB

Table 2: Main comparison on Qwen3-0.6B–8B.

Model	Method	#Neurons/sites	Intent	Sentiment	Legal	Science	Overall
Qwen3-0.6B	-	0	72.58	56.49	67.88	36.51	60.34
	CAA	1024	-0.69	+2.28	+0.12	+4.58	+1.36
	SADI	1000	-0.85	+0.98	+0.13	+0.66	+0.30
	AUSteer	100	-0.57	+3.90	-0.57	-1.51	+0.89
	SCOUT	11.02	+0.88	+1.96	+0.62	+1.92	+1.34
Qwen3-1.7B	-	0	75.68	67.45	67.95	49.97	66.86
	CAA	2048	+2.73	+8.45	+1.22	+3.36	+4.36
	SADI	1000	-0.26	+5.89	+0.28	-0.15	+1.94
	AUSteer	100	-0.23	+1.74	-0.20	+1.78	+0.77
	SCOUT	14.21	+1.84	+6.53	+1.21	+2.58	+3.44
Qwen3-4B	-	0	87.71	75.67	71.38	64.20	75.39
	CAA	2560	0.00	+0.72	+0.08	+0.71	+0.38
	SADI	1000	+0.08	+0.89	-0.51	+0.05	+0.20
	AUSteer	100	-0.13	+0.10	+0.12	-0.67	-0.08
	SCOUT	16.07	+0.14	-0.14	+0.08	+0.10	+0.01
Qwen3-8B	-	0	90.50	78.38	69.44	67.29	76.84
	CAA	4096	+0.41	-0.11	-0.17	+0.79	+0.18
	SADI	1000	-0.06	-0.03	+0.22	+0.45	+0.12
	AUSteer	100	+0.09	+0.22	+0.01	+0.01	+0.09
	SCOUT	14.57	+0.06	+0.53	+0.14	+0.14	+0.26

full-14 Clean means are $66.86/75.39/76.84\%$. Accepted SCOUT sets remain sparse versus SADI's 1000-element masks.

4.3 Deployment-Operator Ablation

This subsection isolates the online operator while holding the accepted set fixed on the Qwen3-0.6B suite. Table 3 compares Clean, Sensitivity, Dynamic SCOUT, and Static SCOUT only. Sensitivity applies the shared sites without conditional search filtering beyond the shortlist. Dynamic SCOUT reapplies the input-dependent displacement online, and Static SCOUT uses frozen site-wise scales. Dynamic and Static both improve over Clean on average, and Static closely tracks Dynamic while remaining near Clean latency. The practical deployment operator is therefore the static approximation of the measured search intervention.

4.4 Benefit of Conditional Selection

The fixed-pool experiment evaluates selector effects on the confirmatory trio. Table 4 isolates the selection rule while holding the candidate pool, operator, budget, calibration, and static deployment fixed. Conditional utility improves accuracy on Financial PhraseBank, CLINC150, and LEDGAR, winning $9/9$ dataset–seed cells against contrastive, sensitivity, and random selection, and $6/9$ against independent utility. Independent utility is higher on all three LEDGAR seeds. With operator and pool fixed, the gap is attributable to the selection rule on this Qwen3-1.7B ablation.

4.5 Robustness Across K and B

This subsection measures how Static SCOUT ΔAcc responds to working-set size K and acceptance budget B. We evaluate Financial PhraseBank, Twitter, Weibo, and TREC-50 at seed 42 on the grid $\{2, 5, 9, 12, 15, 19, 22, 25, 29, 32\}$, fixing $B{=}2$ when sweeping K and $K{=}32$ when sweeping B (Figure 3). Acceptance uses the search-split objective, and the plotted quantity is held-out static ΔAcc.

Typical endpoints already diverge. On Financial PhraseBank, $K{=}2$ yields $+0.7$ pp and $K{=}32$ yields $+0.2$ pp at $B{=}2$, while $B{=}32$ at $K{=}32$ reaches $+1.5$ pp. Twitter loses 1.2 pp at $K{=}2$, recovers to $+2.4$ pp at $K{=}32$, and returns to zero at $B{=}32$. Weibo is flat at $K{=}2$, falls to -2.4 pp at $K{=}32$, and

Table 3: Deployment-operator ablation on the Qwen3-0.6B frozen suite (equal weight over dataset×seed cells). Sensitivity, Dynamic SCOUT, and Static SCOUT share the accepted neuron set; only the online operator changes. Clean is Acc (%); others are ΔAcc (pp).

Model	Method	#Neurons	Intent	Sentiment	Legal	Science	Overall
Qwen3-0.6B	Clean	0	72.58	56.49	67.88	36.51	60.34
	Sensitivity	11.02	-0.58	+0.38	-0.01	-0.08	0.00
	Dynamic SCOUT	11.02	+0.82	+2.63	+0.57	+2.15	+1.59
	Static SCOUT	11.02	+0.88	+1.96	+0.62	+1.92	+1.34

Table 4: Matched selection-rule ablation (ΔAcc in pp, mean±std). Identical operator and candidate set; Qwen3-1.7B confirmatory trio (seeds $\{42, 43, 44\}$).

Dataset (n seeds)	Cond. pp	Indep. pp	Sens. pp	Contr. pp	Rand. pp
financial_phrasebank (3)	+5.96±3.55	+5.30±3.88	+1.47±1.93	-1.32±1.01	-6.70±13.51
clinc150 (3)	+3.40±0.46	+2.90±0.50	+1.49±0.22	-4.97±3.72	-5.76±7.51
ledgar (3)	+1.12±0.92	+2.14±0.56	-0.57±0.50	-3.46±3.08	-2.24±4.51
Conditional-Utility W/T/L vs others (dataset×seed): Contrastive-B 9/0/0; Sensitivity-B 9/0/0; Independent-Utility-B 6/0/3					

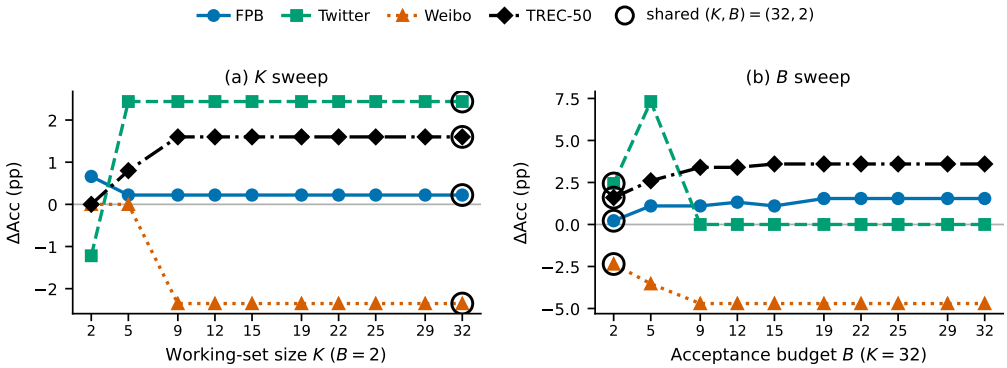


Figure 3: K sweep at B=2 (left) and B sweep at K=32 (right) on FPB, Twitter, Weibo, and TREC-50 (Qwen3-1.7B, seed 42). Open markers mark the shared anchor $(K, B) = (32, 2)$.

reaches −4.7 pp at B=32. TREC-50 accepts no neurons at K=2, then gains +1.6 pp at K=32 and +3.6 pp at B=32.

These shapes are consistent with ordered conditional acceptance under a fixed traversal when the budget changes which candidates enter the accepted path, together with search-to-test transfer of the selected set. Small K can starve the shortlist, as on TREC-50 at K=2. Larger K or B can raise test ΔAcc when later positives remain available, as on TREC-50 and Financial PhraseBank in the B sweep, or leave it non-positive, as on Weibo and on Twitter once B expands past a mid-budget peak. The figure alone does not identify a single cause for negative static gains.

The dense sweep therefore shows budget- and task-dependent heterogeneity rather than a monotone response to larger search. Operating points for K and B remain task-specific under this protocol.

4.6 Static Fidelity and Online Efficiency

This subsection measures how well static calibration recovers Dynamic search displacement and online cost. Held-out median NSE is low on Financial PhraseBank, CLINC150, and LEDGAR, and the complementary EDE remains close to one (Figure 4). On the 0.6B suite, Dynamic and Static ΔAcc agree in sign on almost all valid cells, with a small median absolute gap. Online, Static SCOUT applies frozen site-wise scalars near Clean latency, while Dynamic recomputes an input-dependent

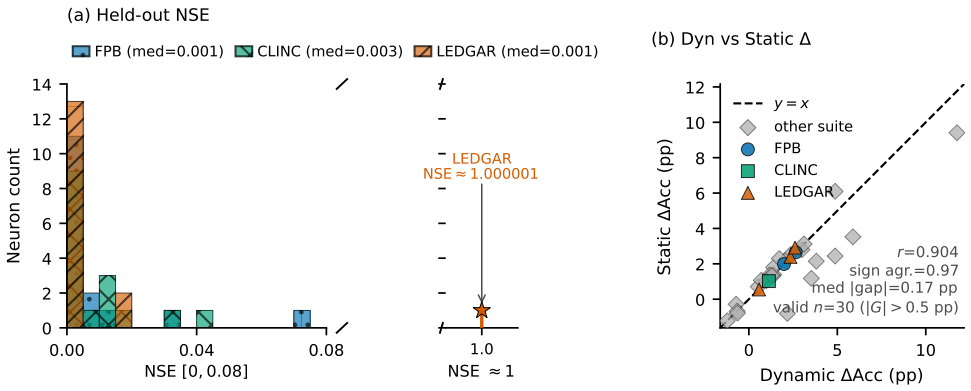


Figure 4: **Left:** held-out uncentered NSE (Qwen3-1.7B, seed 42); broken-x marks one LEDGAR neuron with NSE≈ 1.000001; median EDE $= 1-$NSE stays near one. **Right:** Dynamic vs. Static ΔAcc (pp) on Qwen3-0.6B suite cells with $|G_{\text{dyn}}|>0.5$ pp (n=30).

Table 5: Online latency vs Clean on the Qwen3-0.6B suite (n=42 seed-runs). Static SCOUT applies frozen site-wise scalars; Dynamic recomputes an input-dependent displacement online (measured wall-clock ratio; no online candidate search).

Method	Latency / Clean (mean$\pm$std)	n
Dynamic	3.105$\pm$0.198	42
Static SCOUT	1.039$\pm$0.050	42
Sens.-only (LB)	1.021$\pm$0.041	42

displacement and is substantially slower (Table 5). Aggregate NSE and EDE rows appear in the appendix calibration table.

5 Limitations

SCOUT searches inside a finite sensitivity-ordered working set and leaves rejected neurons unrevisited, so the returned set is path-dependent under Proposition 1; subset optimality among all $S \subseteq C$ with $|S| \leq B$ remains combinatorial (Eq. 1). Matched-n workbook comparisons use a SCOUT-checked frozen-source subset; CAA/SADI/AUSteer Δs are workbook-reported. Matched mean SCOUT Δ is positive at 1.7B and near zero at 4B/8B. The dense K/B sweep covers four Qwen3-1.7B tasks at seed 42. Evidence is concentrated on Qwen3 classifiers with multiple-choice question (MCQ) prompts. Offline search scales with K. Main-table 0.6B Clean+SCOUT use the suite; CAA/SADI/AUSteer Δs retain the Acc_0.6B Clean reference.

6 Related Work

Signal construction. Early methods construct directions from prompt pairs or selected components, including Activation Addition, ITI, Representation Engineering, In-Context Vectors, and Function Vectors (Turner et al., 2023; Li et al., 2023; Zou et al., 2023; Liu et al., 2024; Todd et al., 2024). CAA averages residual-stream activation differences over contrastive prompt pairs (Rimsky et al., 2024). SADI builds semantics-adaptive dynamic steering from contrastive scores and input-dependent scaling (Wang et al., 2025). Instruction-specific and gated controls further modulate intervention (Lee et al., 2025; Stolfo et al., 2025; Sharma & Trivedi, 2026; Nguyen et al., 2026; Kang & Kim, 2026; Rodriguez et al., 2025; Cheng et al., 2025). Along this axis the dominant choice is how to build the steering signal; SCOUT instead measures which sparse sites remain useful after conditioning on sites already accepted.

Site selection. Representation geometry supports controllable directions (Park et al., 2024), while relevance leaves intervention utility unspecified. AUSteer decomposes block activations into fine-grained atomic units and ranks them by activation momenta on contrastive samples (Feng et al., 2026). LayerNavigator searches layers (Sun et al., 2025), and REAL learns activation readouts for localization (Zhan et al., 2026). Perturbing a handful of neurons can impair language ability (Qin et al., 2026). SCOUT uses sensitivity to bound a working set and selects sites by measured set-conditioned utility.

Operator deployment. Activation scaling, spherical rotations that preserve norms along a geodesic, and Stiefel multi-direction optimization modify how interventions are applied (Stoehr et al., 2024; You et al., 2026; Zhu et al., 2026). Related work amortizes steering into parameters or transfers it via weight arithmetic (Yang et al., 2026; Fierro & Roger, 2026), while causal localization and knowledge-neuron analyses identify associated components (Vig et al., 2020; Dai et al., 2022; Meng et al., 2022; Hase et al., 2023; Fang et al., 2025; Jiang et al., 2025). SAEs, Patchscopes, and transcoders provide complementary inspection tools (Gao et al., 2025; Leask et al., 2025; Ghandeharioun et al., 2024; Dunefsky et al., 2024). SCOUT contributes a measured set-conditioned selection criterion with static calibration for deployment.

7 Conclusion

SCOUT reframes sparse activation steering as ordered conditional intervention selection with static calibration for single-forward deployment. Representation ranking identifies associated neurons; measured conditional utility governs which sparse sites enter the accepted set. On the fixed-pool Qwen3-1.7B ablation, conditional utility wins $9/9$ against contrastive, sensitivity, and random selection, and $6/9$ against independent utility. On the Qwen3-0.6B suite, SCOUT yields positive group-mean ΔAcc across data types with a sparse selected set. Held-out calibration recovers the search displacement at near-Clean latency. Within the tested settings, SCOUT selects sparse steering sites by measured conditional utility and deploys them statically.

AI use statement

Generative AI tools assisted with manuscript drafting, editing, and LaTeX formatting. The authors reviewed all AI-assisted content and take responsibility for the final manuscript, including claims and artifacts.

Reproducibility statement

Section 3 specifies the intervention, ordered conditional acceptance, and calibration equations, including the uncentered NSE denominator. Appendix A records protocol settings. The paper package includes audit filters, matched-n aggregates, generated tables and figures, and reproduction materials.

References

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation Addition: Steering Language Models Without Optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023.

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)*, pages 15504–15522, 2024.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-Adaptive Activation Intervention for LLMs via Dynamic Steering Vectors. In *International Conference on Learning Representations*, 2025.

Zixuan Qin, Qingchen Yu, Kunlin Lyu, Zhaoxin Fan, and Yifan Sun. The Achilles' Heel of LLMs: How Altering a Handful of Neurons Can Cripple Language Abilities. In *International Conference on Learning Representations*, 2026.

Zijian Feng, Tianjiao Li, Zixiao Zhu, Hanzhang Zhou, Junlang Qian, Li Zhang, Jia Jim Deryl Chua, Lee Onn Mak, Gee Wah Ng, and Kezhi Mao. Fine-Grained Activation Steering: Steering Less, Achieving More. In *International Conference on Learning Representations*, 2026.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical Steering: Geometry-Aware Activation Rotation for Language Models. In *International Conference on Machine Learning*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and others. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of EMNLP-IJCNLP*, pages 1311–1316, 2019.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, 2020.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving Instruction-Following in Language Models through Activation Steering. In *International Conference on Learning Representations*, 2025.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming Refusal with Conditional Activation Steering. In *International Conference on Learning Representations*, 2025.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems*, 2024.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation Finetuning for Language Models. In *Advances in Neural Information Processing Systems*, 2024.

Lingkai Kong, Haorui Wang, Wenhao Mu, Yuanqi Du, Yuchen Zhuang, Yifei Zhou, Yue Song, Rongzhi Zhang, Kai Wang, and Chao Zhang. Aligning Large Language Models with Representation Editing: A Control Perspective. In *Advances in Neural Information Processing Systems*, 2024.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function Vectors in Large Language Models. In *International Conference on Learning Representations*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 39643–39666, 2024.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 15466–15490, 2024.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and Evaluating Sparse Autoencoders. In *International Conference on Learning Representations*, 2025.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders Find Interpretable LLM Feature Circuits. In *Advances in Neural Information Processing Systems*, 2024.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Mingyang Wan, Guojun Ma, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AnyEdit: Edit Any Knowledge Encoded in Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 27510–27533, 2025.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation Scaling for Steering and Interpreting Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8189–8200, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge Neurons in Pretrained Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, 2022.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. *arXiv preprint arXiv:2004.12265*, 2020.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. In *Advances in Neural Information Processing Systems*, 2023.

Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. In-Context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32287–32307, 2024.

Kartik Sharma and Rakshit S. Trivedi. COLD-Steer: Steering Large Language Models via In-Context One-step Learning Dynamics. In *International Conference on Learning Representations*, 2026.

Dung Viet Nguyen, Yen Pham, Hieu Vu, Lei Zhang, and Tan Nguyen. Activation Steering with a Feedback Controller. In *International Conference on Learning Representations*, 2026.

Minjae Kang and Jaehyung Kim. Enhancing Instruction Following of LLMs via Activation Steering with Dynamic Rejection. In *International Conference on Learning Representations*, 2026.

Dongxuan Zhu, Ly Ho Khanh Tran, Andy Yat-Ming Cheung, Man-Chung Yue, and Viet Anh Nguyen. Exploring Diverse Generation Paths via Inference-time Stiefel Activation Steering. In *International Conference on Learning Representations*, 2026.

Wannan Yang, Xinchi Qiu, Lei Yu, Yuchen Zhang, Aobo Yang, Narine Kokhlikyan, Nicola Cancedda, and Diego Garcia-Olano. Hallucination Reduction with CASAL: Contrastive Activation Steering for Amortized Learning. In *International Conference on Learning Representations*, 2026.

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. LayerNavigator: Finding Promising Intervention Layers for Efficient Activation Steering in Large Language Models. In *Advances in Neural Information Processing Systems*, 2025.

Zifeng Cheng, Jinwei Gan, Zhiwei Jiang, Cong Wang, Yafeng Yin, Xiang Luo, Yuchen Fu, and Qing Gu. Steering When Necessary: Flexible Steering Large Language Models with Backtracking. In *Advances in Neural Information Processing Systems*, 2025.

Pau Rodriguez, Michal Klein, Eleonora Gualdoni, Valentino Maiorca, Arno Blaas, Luca Zappella, Marco Cuturi, and Xavier Suau. LinEAS: End-to-end Learning of Activation Steering with a Distributional Loss. In *Advances in Neural Information Processing Systems*, 2025.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse Autoencoders Do Not Find Canonical Units of Analysis. In *International Conference on Learning Representations*, 2025.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-Space Constrained Model Editing for Language Models. In *International Conference on Learning Representations*, 2025.

Li-Ming Zhan, Bo Liu, Yujie Feng, Chengqiang Xie, Jiannong Cao, and Xiao-Ming Wu. REAL: Reading Out Transformer Activations for Precise Localization in Language Model Steering. In *International Conference on Learning Representations*, 2026.

Tim Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering Evaluation-Aware Language Models To Act Like They Are Deployed. In *International Conference on Learning Representations*, 2026.

Constanza Fierro and Fabien Roger. Steering Language Models with Weight Arithmetic. In *International Conference on Learning Representations*, 2026.

A Experimental Details

A.1 Protocol defaults

Mechanism runs use Qwen3-1.7B with maximum sequence length 1,024 and at most six answer options. Discovery, search, and calibration use 64, 128, and 128 stratified examples, respectively (with protocol-script overrides where noted). Gaussian noise with σ=0.1 is applied to embeddings and scaled relative to the embedding RMS. Activations are read from the last non-padding position of all MLP intermediate layers. Working-set construction uses mean-absolute displacement ranking with default K=64; audited suite/selection/dense manifests set `ranking.layer_balanced=false`. Ordered conditional acceptance evaluates at most 64 candidates in sensitivity order, uses reverse displacement with β=1, and accepts at most B=16 neurons under the default protocol. Qwen3-0.6B suite runs use `acceptance_mode=task_guardrail` (accept iff gain> 0 and $\text{trial_acc} \geq \text{current_acc}$; accuracy tolerance 0). Dense K/B and the fixed-pool selection-rule ablation use $\text{gain} > 0$ under legacy configs (`acceptance_mode=legacy`, `require_accuracy_nondecrease=false`). For the held-out NSE experiment, the 128-example calibration pool is stratified into fit/holdout halves (n_{fit}=62–63, n_{holdout}=64 in the neuron-fit ledger). Static calibration uses λ=10^{-8} in Eq. 7. Selection ablations use seeds $\{42, 43, 44\}$ on all three confirmatory datasets. Sensitivity diagnostics and held-out calibration use seed 42 unless noted.

A.2 Dense K and B grid settings

Main-text Section 4.5 uses the audited dense K/B sweep on Qwen3-1.7B with seed 42. Datasets and test sizes are Financial PhraseBank (453), Twitter (82), Weibo (85), and TREC-50 (500). Both sweeps use the grid $\{2, 5, 9, 12, 15, 19, 22, 25, 29, 32\}$, with B=2 when varying K and K=32 when varying B, and shared anchor $(K, B) = (32, 2)$. Unit manifests set `ranking.layer_balanced=false`, `task.batch_size=4`, and legacy $\text{gain} > 0$. Audit reports 80/80 complete units, empty issues, and no reuse of the coarser confirmatory-trio sweep. That earlier trio sweep at default $(K, B) = (64, 16)$ is not mixed into Section 4.5 numeric claims.

A.3 Data provenance

Main-table Clean Acc for 1.7B/4B/8B uses Acc_XB full-14 Clean(%) means. CAA/SADI/AUSteer Δs use Acc_Delta matched units with $|100 \cdot (\text{Acc} - \text{Clean}) - \Delta| \leq 0.05$ pp and baseline/frozen `SCOUT_source` (tuned exclusions: 7). Table 2 SCOUT for 1.7B/4B/8B uses the author-provided full-14 ledger (Clean/Acc/Δ/K/Source; tuned labels retained); the older Acc_Delta SCOUT matched subset is no longer used for those SCOUT rows. On 0.6B, Clean and SCOUT share the suite

projected/static ledger; CAA/SADI/AUSteer Δs are workbook Acc_0.6B values relative to that sheet's Clean and are not recomputed against the suite Clean. Checkpoint hashes are outside the workbook.

A.4 Dataset groups in the main comparison table

The main comparison table reports four category columns plus Overall. Each category is the unweighted mean over the datasets listed below; Overall is the unweighted mean over all 14 datasets (not the mean of category means). Task names follow the suite experiment-dataset registry.

Table 6 lists the concrete datasets behind Intent/topic, Sentiment/stance, Legal, and Science. CAIL2018 and CJO22 are both accusation-classification tasks; CJO22 has no official validation split and evaluates on CJO22 test with a CAIL2018 train/fit pool. The suite name `lexglue` maps to LexGLUE SCOTUS (LEDGAR is a separate entry).

B Sensitivity Diagnostic (Additional Views)

For a fixed 256-neuron analysis pool from the confirmatory sensitivity-candidate diagnostic, Spearman(R, m) is -0.612, -0.636, and -0.559 on Financial PhraseBank, CLINC150, and LEDGAR; Spearman(R, u) is -0.170, -0.282, and -0.183 (Figures 5–6). Recall@64 for sensitivity ranking is 0.188, 0.312, and 0.125, versus 0.245, 0.251, and 0.243 for random selection. Figure 7 summarizes representation-score versus utility correlations on the same diagnostic setting.

C Additional Calibration

Table 7 aggregates held-out NSE/EDE by dataset. Neuron-level NSE and Dynamic–Static retention are summarized in the main-text calibration subsection. The main-text K/B figure reports only the dense four-dataset sweep described above.

D Full 0.6B Suite Ledger

Table 8 lists per-dataset mean and standard deviation over seeds for the Qwen3-0.6B frozen suite.

E Additional Experiments

E.1 $K\times$pool interaction

Purpose: check whether the Sensitivity advantage persists as K grows. Result: Figure 8 and Table 9 show Sensitivity leading at $K \in \{32, 64, 128\}$ with lower search time per accepted site. Analysis: larger K expands the shortlist but does not invert the ranking under conditional acceptance in these cells. Conclusion: enlarging K does not overturn the Sensitivity prior under conditional acceptance.

E.2 1.7B Dynamic–Static retention

Purpose: measure deployment retention on Qwen3-1.7B (separate from the 0.6B suite latency figure). Result: Table 10 reports 9 cells with 7 valid retention ratios and median retention 1.011. Analysis: valid cells require $|G_{\mathrm{dyn}}|>0$; static coefficients track Dynamic gains while using one forward pass. Conclusion: static coefficients preserve Dynamic gains at one forward pass on this trio.

E.3 Legacy coarse K/B sweep

Purpose: document the earlier confirmatory-trio coarse grid, kept separate from the dense main-text sweep. Result: Figure 9 shows heterogeneous ΔAcc across $K \in \{32, 64, 128\}$ and $B \in \{8, 16, 32\}$. Analysis: coarser endpoints and a different dataset set make these curves non-interchangeable with the dense four-dataset grid. Conclusion: coarse endpoints are not interchangeable with the dense 1.7B four-dataset grid.

Table 6: Dataset membership for the category columns of the main comparison table.

Category	Dataset	Task
Intent/topic	Banking77	banking intent (77 classes)
	CLINC150	assistant intent / out-of-scope
	TREC-50	fine-grained question type (50 classes)
Sentiment/stance	Financial PhraseBank	financial sentiment (3 classes)
	IMDb	movie-review sentiment
	SST-2	binary sentence sentiment
	Twitter	short-text sentiment
	Weibo	Chinese microblog sentiment
Legal	CAIL2018	Chinese accusation classification (CAIL2018 small 5-charge subset)
	CJO22	Chinese accusation classification (CAIL2018 train/fit; CJO22 test)
	LEDGAR	contract provision classification
	LexGLUE	LexGLUE SCOTUS issue-area classification (13 classes)
Science	PubMed-RCT	biomedical abstract sentence role (5 classes)
	SciCite	citation intent

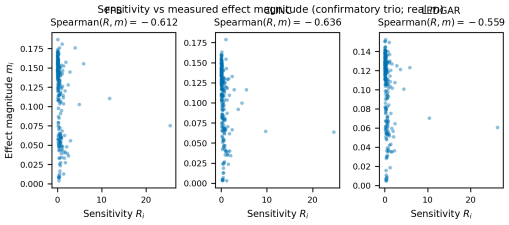


Figure 5: Sensitivity R vs. measured effect magnitude m on the confirmatory trio (FPB/CLINC/LEDGAR; 256-neuron pools; m is the measured effect-magnitude field, not an $|u|$ proxy).

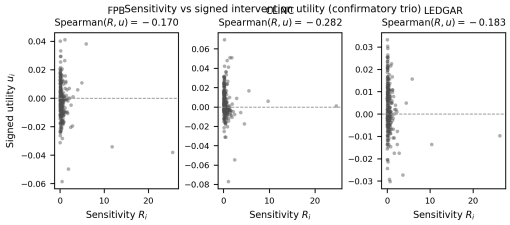


Figure 6: Sensitivity R vs. signed intervention utility u on the same confirmatory pools.

E.4 0.6B F1 AND RESOURCE LEDGER

Purpose: report suite F1 and measured resources for audited suite fields. Result: Tables 11–12 show SCOUT F1 tracking accuracy gains and SCOUT latency near Clean while Dynamic is slower. Analysis: F1 co-moves with Acc under the same accepted sets, and the latency gap is operator-side rather than selection-side. Conclusion: accuracy gains are not an Acc-only artifact, and static deploy remains the cheap online operator.

E.5 MULTI-DATASET MARGINAL HARM

Purpose: extend the FPB contrastive-harm illustration beyond one dataset. Result: Table 13 counts negative Δmargin rank steps (not ΔAcc) at $K{=}32$ across six datasets; Cum. ΔAcc is listed separately. Analysis: on FPB the margin count ($15/32$) differs from the motivation-panel ΔAcc count ($12/32$), so step polarity depends on the local metric. Conclusion: relevance-ranked contrastive steps are not monotone under either step metric.

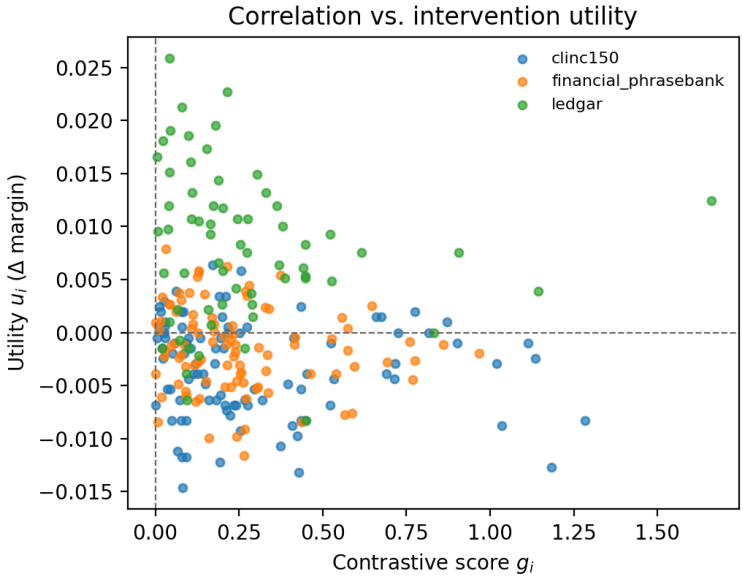


Figure 7: Correlation between representation scores and intervention utility on the confirmatory diagnostic.

Table 7: Appendix: held-out calibration aggregates (Qwen3-1.7B). NSE = SSE/ $\sum(\Delta a)^2$ (uncentered); EDE = 1−NSE (explained displacement energy; not standard R^2). Pearson corr is between Δa and ρa on the holdout split. 0.6B suite retention uses valid cells with $|G_{\text{dyn}}| > 0.5$ pp.

Dataset	median NSE	median corr	median EDE	n neurons
clinc150	0.0033	0.673	0.9967	16
financial_phrasebank	0.0009	0.871	0.9991	16
ledgar	0.0012	0.925	0.9988	16
0.6B suite retention (valid $n = 30$, $\lvert G_{\text{dyn}}\rvert > 0.5$ pp): median 1.000, mean 0.936				

Table 8: Qwen3-0.6B frozen suite per-dataset mean±std over seeds $\{42, 43, 44\}$. Clean Acc (%); other columns ΔAcc (pp). Weibo is a development set. CAA/AUSteer are outside this suite.

Dataset	Clean%	SADIΔ	Sens.Δ	DynΔ	StaticΔ
Banking77	79.42±0.09	-0.30±0.78	-0.37±0.58	+0.22±0.82	+0.31±0.91
CLINC150	87.25±0.70	-0.73±0.97	-0.37±1.07	+0.32±0.72	+0.33±0.64
TREC-50	51.07±1.29	-0.60±1.00	-1.00±2.96	+1.93±0.92	+2.00±0.72
FPB	36.20±0.44	+1.25±0.56	+0.52±0.71	+1.69±1.13	+1.69±1.13
IMDB	88.16±0.03	-0.18±0.45	+0.15±0.14	-0.43±0.45	-0.27±0.41
SST-2	77.68±0.87	+0.08±0.63	+0.08±2.06	+1.99±0.24	+1.22±1.76
Twitter	45.12±4.40	+2.03±3.73	+0.00±1.22	+2.85±3.52	+2.44±3.66
Weibo$^{\text{dev}}$	35.29±1.18	+9.41±2.04	+1.18±3.11	+7.06±4.24	+4.71±4.24
CAIL2018	90.86±0.52	-0.06±0.21	+0.17±0.06	+0.22±0.42	+0.29±0.69
CJO22	82.51±0.31	-0.18±0.21	+0.35±0.24	+0.37±0.18	+0.41±0.26
LEDGAR	79.97±0.18	+1.30±0.86	-0.42±0.70	+1.85±1.10	+1.95±1.24
LexGLUE	18.19±0.94	+0.07±0.40	-0.14±0.26	-0.17±0.41	-0.17±0.55
PubMed	39.33±0.23	-2.21±1.81	-1.11±0.07	+1.57±0.71	+1.63±0.78
SciCite	33.69±0.38	-0.29±0.99	+0.95±0.83	+2.74±1.30	+2.20±0.89
Mean (14)	60.34	+0.69	-0.00	+1.59	+1.34

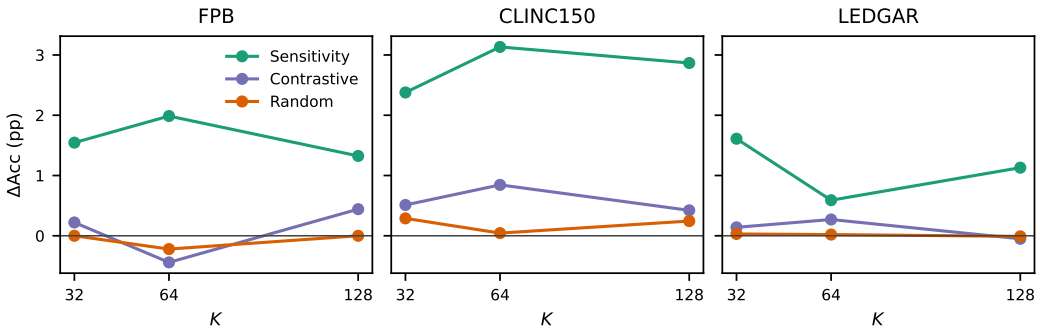


Figure 8: $K\times$pool ΔAcc curves (Qwen3-1.7B, seed 42, B=16).

Table 9: $K\times$candidate-pool interaction (Qwen3-1.7B, seed 42, B=16). Sensitivity remains strongest at $K \in \{32, 64, 128\}$; runtime and accepted counts are reported from the audited interaction ledger.

Dataset	Pool	K	ΔAcc	$\lvert S\rvert$	n_{acc}	search s
CLINC150	Contrastive	32	+0.51	6	6	35.4
CLINC150	Random	32	+0.29	12	12	36.0
CLINC150	Sensitivity	32	+2.38	11	11	36.1
CLINC150	Contrastive	64	+0.84	16	16	58.0
CLINC150	Random	64	+0.04	10	10	71.1
CLINC150	Sensitivity	64	+3.13	16	16	46.1
CLINC150	Contrastive	128	+0.42	16	16	82.1
CLINC150	Random	128	+0.24	16	16	84.9
CLINC150	Sensitivity	128	+2.87	16	16	51.9
FPB	Contrastive	32	+0.22	5	5	38.5
FPB	Random	32	+0.00	6	6	38.6
FPB	Sensitivity	32	+1.55	16	16	32.7
FPB	Contrastive	64	-0.44	12	12	76.4
FPB	Random	64	-0.22	3	3	74.9
FPB	Sensitivity	64	+1.99	16	16	47.5
FPB	Contrastive	128	+0.44	16	16	123.1
FPB	Random	128	+0.00	11	11	155.7
FPB	Sensitivity	128	+1.32	16	16	39.9
LEDGAR	Contrastive	32	+0.14	4	4	112.7
LEDGAR	Random	32	+0.03	6	6	112.7
LEDGAR	Sensitivity	32	+1.61	14	14	112.7
LEDGAR	Contrastive	64	+0.27	9	9	222.1
LEDGAR	Random	64	+0.02	6	6	222.2
LEDGAR	Sensitivity	64	+0.59	16	16	126.6
LEDGAR	Contrastive	128	-0.05	16	16	373.2
LEDGAR	Random	128	-0.01	11	11	441.5
LEDGAR	Sensitivity	128	+1.13	16	16	102.7

Table 10: Qwen3-1.7B Dynamic$\rightarrow$Static deployment retention on the confirmatory trio (9 cells; 7 valid with dynamic ΔAcc > 0.005, i.e. > 0.5 pp). Median retention 1.011. Static uses one forward pass.

Dataset	seed	Dyn Δ	SCOUT Δ	Ret.	$\lvert S\rvert$
CLINC150	44	+3.89	+3.93	1.011	16
CLINC150	42	+3.18	+3.13	0.986	16
CLINC150	43	+3.24	+3.13	0.966	16
FPB	44	+6.62	+7.06	1.067	16
FPB	42	+2.87	+1.99	0.692	16
FPB	43	+7.73	+8.83	1.143	16
LEDGAR	44	+1.91	+2.18	1.141	16
LEDGAR	42	+0.44	+0.59	1.341	16
LEDGAR	43	+0.40	+0.58	1.450	16

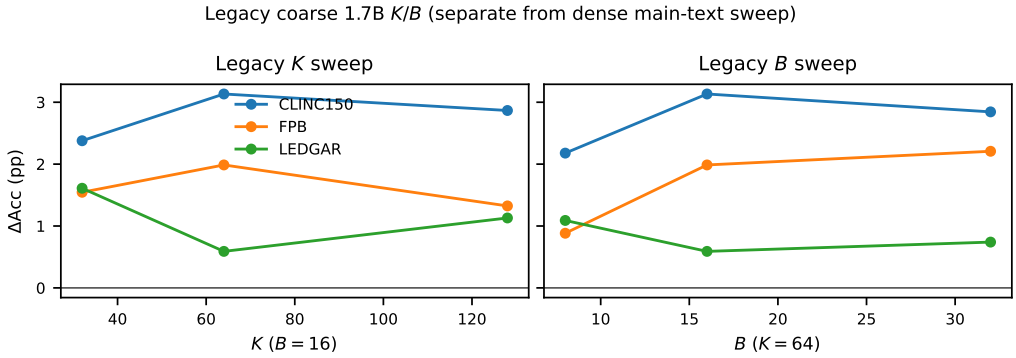


Figure 9: Legacy Qwen3-1.7B coarse K/B sweep (confirmatory trio, seed 42). Not mixed into Section 4.5.

Table 11: Qwen3-0.6B suite F1 companion to the accuracy ledger (mean±std over seeds from the suite aggregate).

Dataset	Clean F1	SADI ΔF1	Dyn ΔF1	SCOUT ΔF1
banking77	0.7886±0.0007	-0.0031±0.0077	0.0016±0.0077	0.0027±0.0090
cail2018	0.7673±0.0168	-0.0028±0.0034	0.0090±0.0142	0.0114±0.0181
cjo22	0.2362±0.0080	-0.0007±0.0008	0.0013±0.0023	0.0025±0.0015
CLINC150	0.8728±0.0070	-0.0072±0.0100	0.0030±0.0071	0.0031±0.0065
FPB	0.3738±0.0016	-0.0004±0.0223	0.0159±0.0131	0.0156±0.0136
imdb	0.8816±0.0003	-0.0018±0.0045	-0.0043±0.0045	-0.0027±0.0042
LEDGAR	0.7427±0.0042	0.0152±0.0104	0.0190±0.0120	0.0207±0.0143
lexglue	0.1343±0.0060	0.0007±0.0045	-0.0005±0.0041	-0.0008±0.0039
pubmed_rct	0.2750±0.0039	-0.0012±0.0004	0.0062±0.0053	0.0064±0.0061
SciCite	0.3474±0.0048	-0.0107±0.0201	0.0267±0.0129	0.0226±0.0092
SST-2	0.7753±0.0090	0.0003±0.0065	0.0198±0.0018	0.0109±0.0206
TREC-50	0.4141±0.0064	0.0047±0.0116	0.0033±0.0147	0.0027±0.0170
Twitter	0.4509±0.0443	0.0121±0.0362	0.0206±0.0438	0.0148±0.0466
Weibo	0.3457±0.0142	0.0977±0.0117	0.0761±0.0475	0.0523±0.0472

Table 12: Qwen3-0.6B suite resource aggregates (n=42 cells per method when present). Latency in seconds; throughput in samples/s; peak memory in MB; search cost in seconds (offline).

Method	Latency	Throughput	Peak mem	Search cost
Clean	56.63	271.17	2372.23	—
SADI	—	—	—	—
Sensitivity	57.44	264.48	2372.23	—
Dynamic	173.19	86.99	2381.19	222.45
SCOUT	57.75	259.95	2372.23	222.45

Table 13: Multi-dataset contrastive rank-step evidence at K=32 (seed-averaged where multiple seeds exist). #Δmargin<0 counts steps with negative Δmargin; Cum. ΔAcc is the separate end-of-prefix accuracy change vs. Clean. Main-text motivation panel (b) instead counts FPB steps with negative ΔAcc (12/32), which need not equal the margin count.

Dataset	Steps	#Δmargin<0	Ratio	Cum. ΔAcc (pp)
FPB	32	15.0	0.469	-0.66
SciCite	32	18.0	0.562	+0.21
SST-2	32	17.0	0.531	+0.23
TREC-50	64	34.0	0.531	+0.59
Twitter	32	15.0	0.469	+0.00
Weibo	32	14.0	0.438	+0.00

F PROMPT TEMPLATES

Task prompts follow `prompt_source=austeer_sadi` with at most six options (letters A–F from the shared option-letter alphabet). Negatives are sampled stably via `md5(seed::example_id)` (first 16 hex digits seed a local RNG). English and Chinese bodies follow the same builder: both terminate with a shared semantic answer cue (language-localized literal). Dataset `input_header`, language, and `max_length` come from the task specification. Option count is $\min(6, |\text{vocab}|)$ with the gold label always retained.

English template

```
Classify the following input. Reply with the option letter only.
{input_header}: {text}
Options:
A. {option_1}
B. {option_2}
...
Answer:
```

Chinese template

```
请根据以下输入选择最合适的类别，只回答选项字母。
{input_header}: {text}
候选项:
A. {option_1}
B. {option_2}
...
答案:
```

Figure 10: Full English and Chinese prompt templates used by the suite (A–F options; stable negative sampling; shared semantic answer cue).

Figure 10 shows the complete templates. Purpose: make the MCQ surface form reproducible without packing protocol detail into the main text. Result: bilingual templates share the lettered option block and a semantic terminal answer cue. Analysis: language only changes the instruction, header punctuation, and cue literal; sampling and option letters are shared. Conclusion: prompt variance across EN/ZH suite cells is limited to the documented template strings.

G SCOUT DETAILED SETTINGS

Models and runtime. 0.6B suite runs use the Qwen3-0.6B checkpoint; dense K/B and confirmatory/selection 1.7B runs use the Qwen3-1.7B checkpoint. Typical dtype bfloat16/float16 on CUDA; max length 1,024. Dense-sweep git commit is recorded as `unknown` in manifests and is disclosed as such.

Batch size (`task.batch_size`). Dense K/B (80/80 units): 4. Fixed-pool selection-rule ablation: 4 (per-run configs). 0.6B suite: per-dataset values in $\{2, 4, 8\}$ from frozen configs—banking77/clinc150/imdb/ledgar/twitter/weibo 4; cail2018/cjo22/lexglue 2; financial_phrasebank/pubmed_rct/scicite/sst-2/trec50 8.

Splits. Strict stratified disjoint discovery/search/scout (projection) splits; official test is evaluation-only. Default sizes 64/128/128. Dense-sweep overrides from unit manifests: Twitter scout 87 (eval 82); Weibo scout 94 (eval 85); FPB/TREC-50 scout 128 (eval 453/500). Suite `scout_examples` may be larger (up to 10,000) when the projection pool permits.

Activations and probe. Target `mlp_intermediate`, all layers, last non-padding position. Gaussian probe σ=0.1, relative embedding RMS, non-padding tokens, resampled per example, one probe.

Ranking and search. Metric mean-abs; working-set size K. Audited suite, selection-ablation, and dense manifests set `ranking.layer_balanced=false`. Search objective mean label margin; fixed sensitivity traversal; budget B; min gain 0. Acceptance formulas: 0.6B suite `task_guardrail` accepts iff $\text{gain} > 0$ and $\text{trial_acc} \geq \text{current_acc}$ (tolerance 0); dense K/B and selection-rule ablation use legacy $\text{gain} > 0$. Intervention direction reverse with β/α=1. Seeds $\{42, 43, 44\}$ unless a unit manifest fixes seed 42.

Calibration and deploy. Independent scout/projection split; $\rho = P/(Q + \lambda)$ with λ=10^{-8}, `clamp_alpha=null`, `drop_invalid=false`, `min_activation_energy`=10^{-10} (suite and dense accepted-neuron configs). Static deploy applies frozen ρ; Dynamic recomputes input-dependent displacement. Metrics: Acc/F1/NLL as logged; random controls in selection ablations; runtimes/hashes from per-run logs and manifests when present.

Exceptions (do not mix). Workbook Acc_0.6B SCOUT Acc/Δ cells are empty (`scout_missing`); 0.6B main-table SCOUT uses the suite projected/static ledger. Acc_Delta matched SCOUT is retained for audit and the matched-workbook scaling table only and is not used for the main-comparison SCOUT rows on 1.7B/4B/8B (those rows use the author-provided full-14 ledger). Tuned Acc_Delta sources remain excluded from matched-n baseline means. Legacy confirmatory-trio coarse K/B is appendix-only and not mixed into Section 4.5.