STEER RESPONSIBLY: TOWARDS A GROUNDED LIBRARY OF INTERPRETABLE STEERING VECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can concepts grounded in human feedback become reliable activation-steering controls? We build Steer Responsibly to test this path: discover signed preference features, independently confirm their descriptions, extract model-specific vectors, and qualify their behavioral effects against quality and control criteria. Twelve of 32 features pass automated semantic confirmation, yielding 48 candidate tensors across Qwen3-8B and Gemma-3-4B-IT, each tested in both intervention directions. None meets the behavioral qualification criteria. Of 96 signed candidates, 86 select the unsteered fallback. Among ten held-out interventions, five miss the target margin in their means, two more fail random-control specificity in their means, and three meet all mean margins but fail uncertainty bounds. Concrete case studies expose limits on attribution: 36 signed candidates have no baseline score headroom; a random direction better expresses a supportive cue than an empathy vector; and near-identical answers receive opposite labels. A separate public-vector audit also finds order-sensitive judgments. We diagnose the gap between semantic grounding and qualified control. The findings make measurement validity part of the library itself to support future research.

1 INTRODUCTION

Personal LLM assistants must accommodate combinations of preferences, not a single global style. A user may want concise answers for routine questions, step-by-step explanations when learning, and empathetic but direct help with a difficult conversation. Such requirements vary across tasks and users Kirk et al. (2024); Zhang et al. (2025). Fine-grained personalization therefore motivates reusable controls for multiple attributes and, when compatible, their composition. Activation steering makes this design plausible by changing hidden representations without fine-tuning the language model Turner et al. (2024); Rimsky et al. (2024); Zou et al. (2023); Cao et al. (2024).

Supporting these use cases requires more than an isolated direction: it motivates a library from which an assistant can retrieve model-compatible vectors, inspect their meanings and tested effects, and choose candidate combinations. A shared label does not establish compatible coordinates, additive effects, or preserved quality. We therefore treat single-vector audits as prerequisites for future composition tests, not evidence that arbitrary sums implement user preferences. Each entry must preserve its tested conditions, evidence, and limitations so readers can assess successes and failures under the conditions actually tested.

We began with a constructive objective: build responsible steering controls from natural human feedback. Building on WIMHF Movva et al. (2025), we discover signed concepts in preference-pair embeddings, independently confirm their descriptions, and extract corresponding directions inside two language models. This produces 12 confirmed concepts and 48 candidate tensors, each evaluated in both signs. We then ask whether the vectors move responses toward those concepts, preserve quality, and outperform matched random and polarity-shuffled directions on held-out prompts.

The outcome is surprising: **none of the 48 tensors qualifies**. Of 96 vector–sign combinations evaluated during validation, 86 had no tested steering strength that both increased the average intended-behavior score and limited the average quality decrease to 0.05 on a 0–1 scale. The remaining ten advanced to held-out behavioral evaluation. Five of the ten miss the target margin even in their mean effects. Two more exceed that margin but do not beat random directions in mean concept score. Three satisfy all mean margins, but none passes the uncertainty bounds. These

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.

descriptive categories explain the observed decisions; they are not independent causal attributions or a revised testing procedure.

The central question of this paper is therefore *why a grounded construction pipeline can yield no qualified behavioral controls*. Our analysis finds both intervention and evaluation limitations. Baseline scores leave no improvement headroom for 36 signed candidates. A random direction can express a description's surface cue more strongly than the intended vector. Near-identical answers can receive opposite endpoint labels when the judge shifts between relevance and verbosity. Gemma frequently reaches the output limit even without steering. These observations leave two explanations entangled: the intervention may not produce the intended change, or the evaluation may not reliably identify it. The zero-pass result therefore calls for diagnosis of both vectors and measurements.

This measurement concern also appears in our audit of existing vectors: two of 77 released CAA/AxBench directions pass corrected output-description fidelity, none passes both random controls, and judge choices are sensitive to answer order. Three of seven separately evaluated Persona reconstructions pass the original criteria, with the same measurement caveat. We contribute (1) an audit-bearing library retaining all 132 evaluated directions—84 external directions and 48 newly grounded candidates—with model-specific tensors, descriptions, and linked audit traces; (2) a construction and evaluation pipeline separating semantic grounding, description fidelity, and behavioral qualification; and (3) a detailed diagnosis of why no grounded candidate qualifies. Library membership records available evidence, not approval for deployment: unsuccessful entries remain inspectable, and tests not performed are distinguished from failed tests.

2 Related Work

Activation interventions and behavioral control. Activation addition and CAA extract contrasts in hidden activations, while Representation Engineering provides a broader framework for reading and controlling high-level representations Turner et al. (2024); Rimsky et al. (2024); Zou et al. (2023). Instruction vectors demonstrate modular control of response constraints Stolfo et al. (2025); CAST adds context-dependent application of steering rather than intervening indiscriminately Lee et al. (2024). HyperSteer maps textual steering requests to interventions with hypernetworks Sun et al. (2025). We reuse simple paired-difference estimators; our contribution lies in connecting these established interventions to feedback evidence and inspectable qualification records.

Personalization and assistant personas. BiPO directly optimizes vectors using contrastive preference pairs and studies personalized control and composition Cao et al. (2024). Steerable Chatbots evaluates interfaces for exposing activation-strength controls to users Bo et al. (2025). Persona Vectors automatically constructs directions from specified trait descriptions and uses them to monitor and control personality changes Chen et al. (2025). Our library complements these controls with feedback-derived concepts and per-vector audit records. Its immediate question is whether individual controls meet their stated criteria, a prerequisite to testing their usefulness in combination.

Feedback interpretation and sparse grounding. WIMHF learns sparse features of response-embedding differences, interprets preference structure, and already supports personalization and data curation Movva et al. (2025). We retain its signed feature discovery while distinguishing external embedding features from directions inside a target language model. PRISM and Community Alignment provide complementary motivation for contextual and plural preferences Kirk et al. (2024); Zhang et al. (2025); HH-RLHF provides the initial library's feedback evidence Bai et al. (2022). SDCV uses SAEs to denoise concept vectors Zhao et al. (2025). GLASS uses sparse priors and local style vectors for user-history-based personalization Chen et al. (2026), with a different goal from our shared inventory and separate semantic and behavioral audit records.

Benchmarks, reliability, and responsible release. AxBench compares concept detection and steering methods and releases large feature dictionaries, finding strong prompting baselines Wu et al. (2025). AI Steerability 360 already supplies a broad steering toolkit Miehling et al. (2026). Reliability studies find substantial input-dependent variation and associations with activation geometry Tan et al. (2024); Braun et al. (2025). Safety audits document unintended changes in jailbreak susceptibility Li et al. (2026), and deliberate poisoning of steering artifacts is already a studied threat Aidakhmetov

et al. (2026). Together, these studies motivate the evidence-centered design in Table 1: a reusable inventory should expose each vector's provenance, behavior, and limits alongside its label.

Table 1: Closest precedents and our contribution. Rows summarize research focus, not the absence of evaluation in prior work. Our inventory retains audited failures as well as passes.

Work	Established contribution	Our relationship to it
WIMHF Movva et al. (2025)	Sparse concepts from preference data	Extend feedback interpretation to model-specific vectors.
BiPO Cao et al. (2024)	Preference-optimized steering	Alternative extraction method; no superiority claim.
Persona Vectors Chen et al. (2025)	Specified-trait monitoring and control	Discover concepts from natural feedback instead.
GLASS Chen et al. (2026)	Sparse priors and user-history steering	Build a shared inventory with per-entry evidence.
AxBench Wu et al. (2025)	Concept dictionaries and steering evaluation	Connect feedback grounding to intervention validation.
This work	**Audit-bearing vector library**	**Retain every audited vector with meaning, provenance, and test outcomes.**

3 Construction and Evaluation Design

Library contents and per-vector audit records. The local catalog contains 132 model-specific tensors: 7 CAA, 70 AxBench, 7 reconstructed Persona, and 48 feedback-grounded candidates. Controls remain linked comparison artifacts and are excluded from this inventory count. An entry stores *identity* (concept label, model/revision, layer, extraction method, pooling, tensor file, width, norm); *provenance and meaning* (origin, source/evidence hashes, signed description, and semantic evidence where available); and typed *audits* (tested claim, evaluator/rubric, frozen plans, conditions, control comparisons, estimates/bounds, decision, and raw-response/judgment links). Appendix A maps these groups to concrete manifest fields and links reported decisions to their supporting evidence.

Tests attach to operating conditions, not just vector names. External entries record description fidelity and two-random-control specificity; grounded entries record semantic confirmation, sign/dose selection, intended effect, quality, and random/shuffled comparisons. Missing tests are explicitly untested. Thus the three external primary-order passes are not interchangeable with grounded behavioral qualification, and *zero qualified grounded vectors does not mean an empty library*. All entries retain their audit histories; composition is marked untested. Tensor lookup enables reuse, while the audit record exposes why an entry may be unsuitable for a proposed task.

Discover and confirm feedback concepts. We use component-disjoint HH-RLHF preference pairs, grouping exact prompts, responses, and conversation identities before splitting. A signed BatchTopK SAE ($M = 32$, $K = 4$) encodes ModernBERT response-embedding differences, $x_i = e(r_{i,A}) - e(r_{i,B})$. Training uses 41,170 rows and early stopping uses 4,298 separate rows. Following WIMHF, Qwen3-32B generates five signed descriptions per supported feature from training examples. We select a description by its correlation with feature activations on 300 distinct selection components, then confirm it on 300 new components using positive correlation and a two-sided permutation test at $p < 0.05/96$. The denominator retains the planned three-source feature family; the present library uses HH-RLHF only. Four of 32 fitted features lack sufficient support and remain in the inventory denominator. The splits prevent reuse of identical evidence across stages; using the same interpreter and semantic judge still leaves shared-model bias unresolved.

Extract model-specific candidate vectors. For each confirmed concept, orient 32 training pairs by feature sign and extract last-response-token residuals with each target model's chat template. We compare unweighted differences with activation-weighted differences,

$$v_{c,m,\ell} = \mathrm{normalize}\left(\frac{\sum_i |z_{ic}| \left[h_{m,\ell}(r_i^+) - h_{m,\ell}(r_i^-)\right]}{\sum_i |z_{ic}|}\right). \tag{1}$$

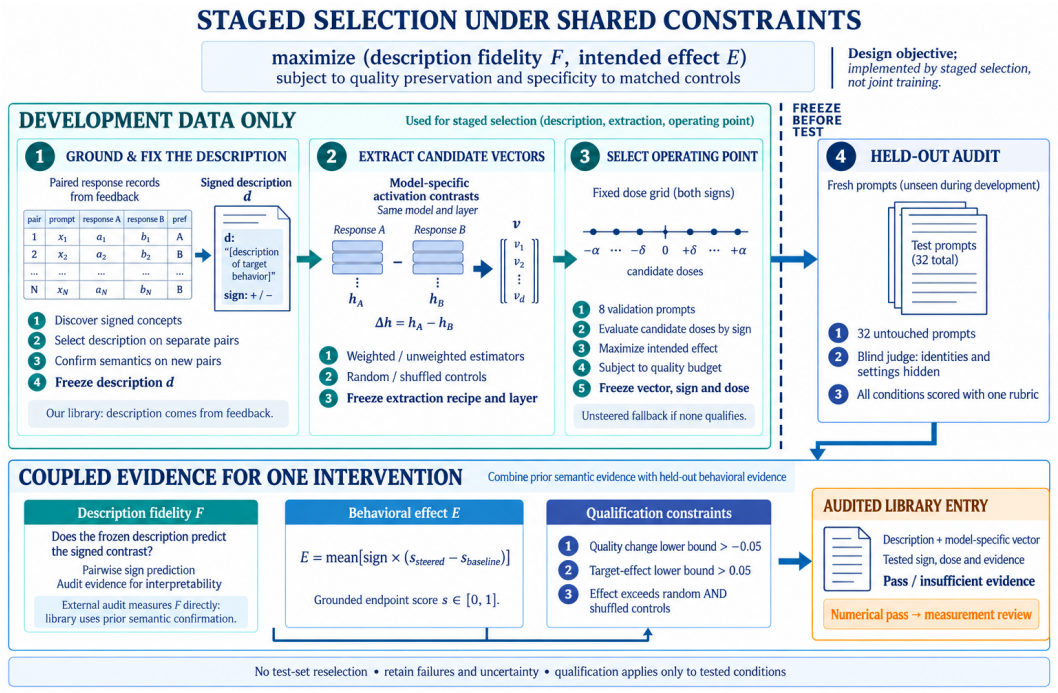

Figure 1: **Staged selection under shared qualification constraints.** Description fidelity and intended behavioral effect define a multi-objective design, subject to quality preservation and matched-control specificity. Our implementation selects and confirms a description, extracts candidate vectors, then selects doses on validation prompts; it does not jointly train these quantities. All choices are frozen before the behavioral test. Prior semantic evidence and held-out behavioral results jointly support an entry's audit record. Numerical passes require measurement review; the held-out responses are used only for evaluation and never to revise the selected description or operating point.

The 12 concepts yield 24 tensors per model at fixed layer 17 for Qwen3-8B and layer 16 for Gemma-3-4B-IT, plus random controls and controls extracted after shuffling contrast signs. Language-model weights remain fixed. The descriptions provide a shared interface; the vectors occupy each model's own residual space and cannot be transferred between models on the basis of shared labels.

Select on validation; qualify on held-out responses. We inject $d\alpha v$ at all token positions, with sign $d \in \{-1, +1\}$ and α equal to 2% or 5% of the training residual norm. On eight validation prompts, each method/sign selects the dose maximizing positive mean intended movement subject to mean quality loss ≤ 0.05; otherwise it selects the unsteered fallback. The chosen dose is frozen before 32 test prompts, grouped into eight fixed topic families. Greedy bfloat16 generation has a 768-token limit. We retain and score truncated text rather than silently dropping it.

Gemini 3 Flash Preview sees the request, frozen positive/negative descriptions, and one response at a time; intervention identity is hidden. Its endpoint and strength map to $s \in \{0, .25, .5, .75, 1\}$, from clear opposite through mixed to clear target. Quality $q \in \{0, .25, .5, .75, 1\}$ is rated independently. Synthetic calibration passes 95/96 fixtures; one ambiguous conciseness fixture was reviewed before scoring and removed from the blocking gate, with its disagreement retained. Unambiguous JSON repairs preserve the raw judgment. These checks establish basic rubric handling, while sensitivity to actual steering-induced differences remains a separate empirical question.

We measure paired changes $\Delta_s = \text{mean}[d(s_{\text{steer}} - s_0)]$ and $\Delta_q = \text{mean}[q_{\text{steer}} - q_0]$. Qualification requires one-sided lower bounds above 0.05 for Δ_s, above -0.05 for Δ_q, and above zero for signed advantages over *both* controls at the selected dose. The frozen analysis uses topic means, Student-t bounds with seven degrees of freedom, and $\alpha = 0.05/(96 \times 4)$. Approximate normality and independence of eight topic means are limitations, not guaranteed by freezing the panel. At least one passing sign qualifies a tensor only at that operating point; real-response measurement review would

Table 2: Completed audit, with cohorts kept separate. CAA/AxBench use the original 77/154-test families; reconstructed Persona uses separate 7/14-test families. Correct and truncated columns concern target-vector pairs; counts share prompts and are descriptive.

Cohort	Directions	Fidelity	Specificity	Correct pairs	Truncated pairs
CAA (7)	7	0	0	107/224	0/224
AxBench anchors (6)	6	1	0	87/192	71/192
AxBench random (64)	64	1	0	724/2048	814/2048
Persona (recon.)	7	3	3	151/224	57/224

still precede promotion. All primary criteria were frozen before test, not externally preregistered. Appendices N and M specify the remaining implementation details and assumptions, including evaluator calibration and the distinct tests and matched controls used in each audit.

4 Motivating Evidence from Public Vectors

The preceding protocol tests intended change in feedback-grounded concepts. For existing vectors, we first ask a narrower question: can an output-derived description predict intervention sign, and is that prediction more reliable than for matched random directions? Eight prompts construct three candidate descriptions, eight select one, and 32 confirm it with a blinded Gemma-3-27B A/B/neither judge. Correct sign predictions require non-identical, untruncated answers; all other cases score zero. CAA/AxBench use exact binomial fidelity tests corrected over 77 directions (26/32 required), and paired exact tests over 154 control comparisons. The separately planned Persona cohort uses seven fidelity tests (24/32 required) and 14 control comparisons, independent of the earlier family.

The cohorts are seven CAA behaviors, six AxBench anchors, 64 uniformly sampled non-anchor AxBench rows, and seven Persona reconstructions. In the original order, only 2/77 CAA/AxBench directions pass fidelity and none passes both controls. Truncation makes the fidelity threshold unreachable for 42/64 sampled AxBench directions. Persona has three passes on both criteria: hallucinating, optimistic, and sycophantic. These are description-fidelity results, not claims of desirable behavior. The cohorts use different models and operating conditions; our CAA token policy and Persona reconstruction also differ from the original papers. These results describe each cohort at its tested operating point and do not rank steering methods. Appendices G and L report every direction and the Persona reconstruction.

Repeating the same response pairs with A/B reversed changes the chosen response in 2,892/6,822 (42.4%) decisive CAA/AxBench judgments and 158/620 (25.5%) Persona judgments, including controls. Repetition and position effects are not separately identified. Thus synthetic calibration and primary-order passes do not establish stable measurement. Appendix H illustrates these external failures with response pairs; Appendix K traces description selection. Appendices I and J report supporting diagnostics and paired reversals. We now separate intervention evidence from measurement reliability for the grounded candidates.

5 Why Did No Grounded Vector Qualify?

5.1 Decompose the zero before interpreting it

Twelve of 32 discovered features pass independent automated semantic confirmation (12/28 tested; 300 distinct components per feature). Correlations range from 0.200 to 0.686; all remain positive after descriptive character-length adjustment (0.187–0.538). This establishes correspondence between descriptions and feedback features, not causal control inside a language model. Appendix F reports every feature. Extraction yields 48 candidate tensors, not 48 independent concepts: weighted/unweighted estimators and two models implement the same 12 meanings.

Semantic confirmation thus succeeds for part of the inventory, but the completed behavioral experiment yields zero numerical passes. To locate that gap, we distinguish **selection rejection** from **held-out failure**. Qwen selects the unsteered fallback for 44/48 signed candidates and Gemma for 42/48. Of these 86 decisions, 85 have no positive validation target movement at either tested dose;

Table 3: Completed library audit. Each model has 24 tensors and 48 signed candidates. "Mean margins" counts selected operating points satisfying all four criteria in point estimates; "qualified" counts tensors passing the corrected bounds. Fallbacks indicate rejection during validation.

Model	Tensors	Fallbacks	Test settings	Mean margins	Qualified
Qwen3-8B	24	44/48	4	2	0/24
Gemma-3-4B-IT	24	42/48	6	1	0/24

one Gemma candidate has positive movement only at a dose violating the quality budget. Fallbacks compare baseline with itself on test, so their recorded zeros cannot be used as evidence that the rejected steering settings have no held-out effect.

Ten signed interventions advance (four Qwen, six Gemma). A descriptive sequence of mean comparisons accounts for all ten: five do not exceed the 0.05 target margin; two more exceed it but do not beat random controls; the remaining three meet every mean margin but fail the lower bounds. All ten meet the quality margin in their means and have positive mean advantage over shuffled controls. This ordering makes the rejection stages explicit. The following analyses explain what each stage establishes and what remains uncertain; the original joint qualification decisions remain unchanged. Appendices D and E report every selected operating point.

5.2 Selection can reject a control with no measurable headroom

The baseline already receives an extreme concept score on 96/96 Qwen and 95/96 Gemma concept–prompt cases. Eight Qwen concepts and ten Gemma concepts are at the same extreme on every validation prompt. For the sign pointing farther toward that extreme, improvement is mathematically impossible under this bounded score: 16 Qwen and 20 Gemma signed candidates across the two estimators. These 36 candidates belong to the 85 without positive validation movement: score saturation explains why this subset cannot advance; the remaining rejections require further explanation.

Figure 2A illustrates this problem. The concept contrasts dialogue/clarification with direct informative answers. Asked why bread dough rises, both baseline and negative-steered Qwen directly explain fermentation; the judge assigns OPPOSITE/CLEAR, $s = 0$, to both. Negative steering is intended to increase this already-maximal endpoint, so its measured gain is necessarily zero. The positive direction has room to move, but this factual request offers little reason to ask clarifying questions. Task appropriateness and improvement headroom are separate design requirements.

Across all candidate doses, 744/768 Qwen and 735/768 Gemma validation responses retain the baseline concept score, although only nine Qwen responses and no Gemma responses are text-identical. This does not show that meaningful change was missed: wording can change without changing the attribute. It does show that our five-level instrument operates almost categorically on this panel. Explicit behavioral prompting produces positive mean movement in 15/24 Qwen signed comparisons, suggesting sensitivity to some stronger changes, without establishing sensitivity to smaller activation interventions. Selection therefore needs both concept-appropriate prompts and anchors that resolve changes in actual responses. Loosening the quality budget alone would not make either tested dose produce positive movement for those 85 candidates.

5.3 A positive baseline difference need not be direction-specific

Advancing beyond selection does not resolve specificity. Gemma's six selected interventions all increase the intended score on average ($+0.023$ to $+0.117$), but four do not outperform the random control even in mean score. Its largest effect, weighted empathetic/supportive steering, is $+0.117$ versus baseline yet -0.008 versus random. In Figure 2B, all answers correctly give 12 pencils. The intended vector receives $s = 0.25$; random receives $s = 1$ because it offers another math problem, matching the broad description's dialogue cue. Quality is 1 for all three. This is a failure of specificity under the chosen operationalization, rather than evidence that this random perturbation provides a general empathy control or that its conversational style is preferable for every task.

The contrast itself bundles empathy, support, dialogue, and constructive advice. The observed random-control success may reflect a generic conversational cue; it does not isolate recognition of

A. No score headroom Qwen, validation
Frozen contrast (abridged): Dialogue/clarification versus direct answers; negative steering seeks direct answers.
Prompt: Explain why bread dough rises.

Baseline OPPOSITE/CLEAR; $s = 0.0$, $q = 1.0$	**Steered** OPPOSITE/CLEAR; $s = 0.0$, $q = 1.0$
Bread dough rises because of a process called fermentation, which is driven by yeast.	Bread dough rises because of a process called fermentation, which is driven by yeast.

Observation: Both full answers directly explain fermentation. Baseline already has the minimum positive-endpoint score; negative steering cannot improve it.

B. A random direction expresses the cue Gemma, test
Frozen contrast (abridged): Empathetic, supportive, dialogue-encouraging versus abrupt/analytical responses.
Prompt: I have three boxes with four pencils each. How many pencils do I have?

Steered OPPOSITE/SOMEWHAT; $s = 0.25$, $q = 1.0$	**Random control** TARGET/CLEAR; $s = 1.0$, $q = 1.0$
You have a total of 12 pencils!	Let me know if you want to try another math problem!

Observation: Baseline: 12 pencils, score 0. The vector gains 0.25, but random gains 1.00; its invitation to continue matches the broad supportive description.

C. Opposite labels for nearly identical answers Qwen, test
Frozen contrast (abridged): Direct, helpful, concise versus indirect, meandering, tangential responses.
Prompt: Explain what a database index does.

Baseline TARGET/CLEAR; $s = 1.0$, $q = 1.0$	**Steered** OPPOSITE/CLEAR; $s = 0.0$, $q = 1.0$
A database index is a data structure that improves the speed of data retrieval operations on a database table.	A database index is a data structure that improves the speed of data retrieval operations on a database table.

Observation: Both include five numbered sections, trade-offs, SQL, and a summary (97% word-token overlap). Judge cues: relevance versus verbosity, respectively.

D. The measured answer ends mid-sentence Gemma, test
Frozen contrast (abridged): Empathetic/supportive versus abrupt/analytical responses; same concept as B.
Prompt: Explain why software version control is useful.

Baseline OPPOSITE/CLEAR; $s = 0.0$, $q = 0.75$	**Steered** TARGET/SOMEWHAT; $s = 0.75$, $q = 0.75$
[https://git-scm.com/ docs/git-tutorial](https://git-scm.com/docs	Without version control, you'd be constantly rewriting the entire document from scratch whenever you make

Observation: These are the exact endings, not editorial omissions. Both reach the 768-token limit. The +0.75 concept change is measured on incomplete answers.

Figure 2: Four purposive diagnostic cases from the grounded library. Blue text is a verbatim excerpt (Markdown emphasis removed); D shows exact output endings. The judge sees one response at a time and the full frozen contrast, not the intervention identity. s scores the positive endpoint; negative steering in A aims to decrease it. Quality q is independent and equals 1 in A–C and 0.75 in D. Cases illustrate mechanisms, not their prevalence; full responses and rationales appear in Appendix C.

emotional context. Feedback grounding confirms a bundled description's association with preference features, whereas reusable control requires testing its relevant components in context. Replacing the rubric after seeing these examples would change the claim, so original scores remain intact.

5.4 Compound descriptions expose inconsistent rubric emphasis

Bundled descriptions can also change which response properties the judge emphasizes. The Qwen task-focus description combines directness, helpfulness, and concision; its opposite includes indirectness and tangents. In Figure 2C, baseline and weighted-steered database-index answers share the same opening, five numbered sections, SQL example, and summary. Their lowercase word-token multiset overlap is 97% (twice the shared-token count divided by the total count). Yet the judge scores baseline TARGET/CLEAR ($s = 1$), citing relevance, and steering OPPOSITE/CLEAR ($s = 0$), citing extensive sections and examples. Neither answer is truncated, and both receive quality 1; the extreme disagreement concerns the concept label rather than task quality.

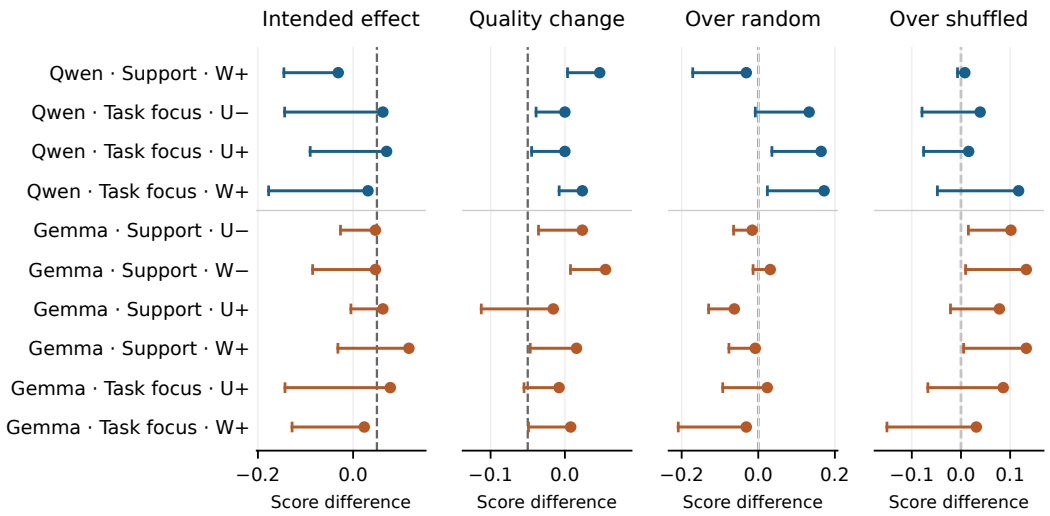

Figure 3: **Why positive means do not suffice.** All ten selected interventions: dots show paired mean effects; horizontal segments extend to unadjusted one-sided 95% topic-t lower bounds. Dashed lines are the qualification margins. Even without multiplicity correction, no setting clears all four margins with its lower bounds. These are pointwise diagnostic bounds, not the frozen corrected analysis. W/U denote weighted/unweighted extraction; signs refer to the frozen concept, and "Support" abbreviates empathetic/supportive versus abrupt responses.

This is evidence of unstable emphasis between components of the rubric on highly similar answers, not a controlled proof of which score is correct. We find three Qwen and two Gemma held-out selected response pairs with at least 80% token overlap and opposite extreme labels among 320 paired comparisons. That post-hoc lexical screen is a case-finding diagnostic, not a judge-error estimator. The external audit's answer-order sensitivity provides a separate reason to review measurement. Neither diagnostic licenses relabeling unfavorable responses. A prospective remedy is to anchor each component and validate judgments on real response pairs with repetition and reversed answer order.

5.5 Small and topic-dependent gains remain uncertain

Even taking the recorded judgments at face value, uncertainty limits the remaining positive results. The three settings meeting all mean margins are both signs of Qwen's unweighted task-focus tensor and the positive sign of Gemma's unweighted task-focus tensor: *two tensors, not three qualified vectors*. Qwen's positive sign improves 4/32 prompt scores, leaves 27 unchanged, and worsens one; its mean gain is $+0.070$. Its eight topic means are $0, 0, 0, -0.250, +0.250, 0, 0, +0.563$. Gemma's corresponding sign improves six scores and worsens five, with mean gain $+0.078$. Average improvement is therefore concentrated in a few prompts and can coexist with opposite changes. The apparent topic variation inherits the limitations of the judgments used to estimate it.

Figure 3 displays all ten settings. The frozen corrected bounds are wide: Qwen's $+0.070$ gain has lower bound -0.503. To test whether correction alone explains rejection, we recompute bounds at unadjusted $\alpha = 0.05$, keeping all selected settings, topic groups, scores, and margins fixed. **None of the ten passes.** For Qwen's positive unweighted direction the target lower bound is still -0.090. Removing the extra four-way penalty for a joint all-criteria decision also cannot change this result. The original correction protects simultaneous component bounds; it is stronger than necessary for just the all-criteria decision. Appendix B reports these post-hoc checks. More independent prompt coverage is needed to distinguish small or context-specific effects from rating variability. This analysis rules out multiplicity correction as the sole reason for rejection, while leaving the presence and size of a reliable behavioral effect unresolved.

5.6 Output limits affect baseline and steering together

The text being scored imposes a further limit on interpretation. Qwen has 6–9 truncated responses per selected 32-prompt intervention, compared with 9–10 in the corresponding baselines. Gemma has 23–27 versus 25 baseline truncations. Thus Gemma's frequent truncation is already present without steering; it cannot be attributed entirely to vector damage. In Figure 2D, both version-control answers stop mid-text. Their quality scores are both 0.75, so the paired quality loss is zero despite incompleteness. A quality noninferiority check asks whether steering worsens the baseline, not whether either answer is complete or absolutely good.

This is materially different from the external audit, where truncation counts as fidelity failure. Here incomplete text is scored and flagged. We cannot infer what longer generations would say or rescue entries by dropping truncated cases. The zero-pass result concerns this fixed token budget and generic prompt panel, not unrestricted generations at optimized layers and doses.

6 Implications and Limitations

Our constructive goal remains a library of usable, interpretable controls. The completed audit shows that semantic grounding is one prerequisite, not a surrogate for intervention qualification. Three requirements emerge: evaluation tasks must admit the intended change; descriptions must distinguish the intended attribute from generic cues; and response-level judgments must be sensitive and stable enough to support effect estimates. More permissive cutoffs do not repair these failures of measurement and experimental coverage.

A follow-up should freeze concept-appropriate prompts and component-level rubrics, validate them on real response contrasts, broaden layer/dose selection on validation, and allocate output length and prompt coverage before a fresh test. This is a proposed design, not an unrun result. Our English HH-RLHF inventory, fixed prompt panel, preview evaluator, one-layer/two-dose search, and automated judgments limit generalization. No human preference, composition, or user-profile study is complete. The supplement provides code and audit records; source licenses and privacy restrictions still govern reuse. Keeping every audited vector with its evidence and limitations makes these follow-ups possible without erasing the original results or overstating current support.

7 Conclusion

We set out to support fine-grained assistant personalization through a grounded vector library. The catalog retains 132 audited directions, including 48 candidates from 12 grounded concepts; none of those 48 qualifies under the tested conditions. The useful result is the decomposition: most settings never advance, some apparent gains are matched by random directions, and the remaining mean effects are uncertain. Saturated scores, inconsistent rubric emphasis, and baseline truncation further limit attribution. A responsible library must therefore expose both the evidence for a vector's meaning and the reliability of the measurements used to judge its behavioral effects.

AI Use Statement

Generative AI tools assisted with refining hypotheses and the evaluation design, implementing and debugging code, constructing synthetic calibration fixtures and prompt panels, analyzing results and qualitative cases, literature search, manuscript drafting and revision, and figure creation. Model-generated concept descriptions and automated judgments are experimental measurements, as specified in Section 3 and the detailed protocols. Code-assisted checks verified report consistency, source hashes, numerical summaries, references, and document layout; these checks do not replace independent human validation.

References

Abzal Aidakhmetov, Donato Crisostomi, Tommaso Mencattini, Adrian Robert Minut, Iacopo Masi, and Emanuele Rodolà. Steering vectors are an adversarial attack surface, 2026.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

Jessica Y. Bo, Tianyu Xu, Ishan Chatterjee, Katrina Passarella-Ward, Achin Kulshrestha, and D Shin. Steerable chatbots: Exploring personalization control interfaces via LLM activation steering, 2025. Version 3, revised July 2026.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models, 2025.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/58cbe393b4254da8966780a40d023c0b-Abstract-Conference.html.

Liuji Chen, Zeyu Zhang, Xinyuan Zhang, Shuai Nie, Qiang Liu, Shu Wu, and Liang Wang. From profiles to steering vectors: Global sparse priors and local semantic calibration for personalized text generation, 2026.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models, 2024.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering, 2024.

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, Bardh Prenkaj, and Gjergji Kasneci. Analysing the safety pitfalls of steering vectors, 2026.

Erik Miehling, Karthikeyan Natesan Ramamurthy, Praveen Venkateswaran, Ching-Yun Ko, Pierre Dognin, Moninder Singh, Tejaswini Pedapati, Avinash Balakrishnan, Matthew Riemer, Dennis Wei, Inge Vejsbjerg, Elizabeth M. Daly, and Kush R. Varshney. AI steerability 360: A toolkit for steering large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 436–444, 2026. doi: 10.18653/v1/2026.acl-demo.43. URL https://aclanthology.org/2026.acl-demo.43/.

Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. What's in my human feedback? learning interpretable descriptions of preference data, 2025.

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. URL https://aclanthology.org/2024.acl-long.828/.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering, 2025.

Jiuding Sun, Sidharth Baskaran, Zhengxuan Wu, Michael Sklar, Christopher Potts, and Atticus Geiger. Hypersteer: Activation steering at scale with hypernetworks, 2025.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024.

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, 2025.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset, 2025.

Haiyan Zhao, Xuansheng Wu, Fan Yang, Bo Shen, Ninghao Liu, and Mengnan Du. Denoising concept vectors with sparse autoencoders for improved language model steering, 2025.

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, 2023.