ONE RETRIEVAL SLICE, TWO WINNERS: TRAJECTORY-AWARE GROUNDEDNESS REPAIR WITH GROUNDDIAL

Anonymous authors
Paper under double-blind review

ABSTRACT

The same 512-prompt Natural Questions (NQ) retrieval slice selects LoRA Steering under post hoc best-point Hit (0.853 versus 0.849) and GroundDial under the declared whole-trajectory utility (0.824 versus 0.742); this disagreement recurs in 78.2% of paired trajectory resamples. This reversal exposes the central evaluation problem: a fixed evidence pool still leaves the evidence object, operating-point rule, and latency envelope to determine which repair is preferred. A claim contract formalizes these choices, and GroundDial supplies a post-retrieval controller that combines trajectory-regularized low-rank hidden-state updates, an output-factor norm cap, and selective abstention. Under the primary NQ latency contract, GroundDial reaches 0.847 Hit and 0.062 Side, versus 0.758/0.112 for Cross-encoder Bo4; full-slice blinded experts measure support gains of 10.5 points on NQ and 8.8 on MS MARCO, with preservation rates of 0.887 and 0.859. Across the same 11 native-path positions, GroundDial obtains 0.912 monotonicity and 0.824 AUC, exceeding TruthX-style steering at 0.861/0.787 and LoRA Steering at 0.803/0.742; trajectory regularization accounts for the largest monotonicity ablation loss, while the norm cap and gate control Side. The ordering persists across three backbones and NQ contexts through 16K, while matched-latency advantages extend to MS MARCO, ASQA, and Contriever retrieval. Claim contracts turn fixed-retrieval evaluation from an endpoint leaderboard into a decision-specific test of selected quality, controllability, and workflow cost.

1 INTRODUCTION

One fixed BM25 top-20 pool and the same 512 NQ prompts produce two winners. Selecting the largest test-sweep Hit gives LoRA Steering 0.853 versus GroundDial's 0.849; integrating the complete declared path gives GroundDial 0.824 trajectory AUC versus LoRA's 0.742. The fixed $\alpha{=}1$ endpoint (0.847 versus 0.819) and development-selected points (0.847 versus 0.801) also favor GroundDial. Thus the disagreement is attached to the claim being made about the response path, not to retrieval variation.

This distinction matters because post-retrieval control is exposed to users as a trajectory. An operating strength chooses how aggressively to edit, collateral drift determines whether that edit is acceptable, and retrieval-to-answer latency determines which workflow is feasible. Endpoint quality, path reliability, and workflow choice are therefore three decision objects over the same evidence pool. We encode each object in a *claim contract*: comparator set, shared interface, supervision constraint, latency envelope, evidence object, and selection functional.

GroundDial is designed to make all three objects measurable in one controller. It combines trajectory-regularized rank-16 hidden-state residuals, an output-factor norm cap, and a hard safe-edit gate. Under the primary NQ latency contract, GroundDial reaches 0.847 Hit and 0.062 Side at $1.04\times$ relative wall-clock, compared with 0.758/0.112 at $1.03\times$ for Cross-encoder Bo4. Three blinded experts judging every output in 512-prompt NQ and MS MARCO slices measure support gains of 10.5 and 8.8 points, with preservation rates of 0.887 and 0.859.

The complete response path reveals the controller mechanism. GroundDial reaches 0.912 NQ monotonicity, versus 0.861 for TruthX-style steering and 0.803 for LoRA Steering; only 9% of its prompts reverse, compared with 23% for LoRA, and the median first reversal moves from $\alpha = 0.4$ to 0.8. Removing trajectory regularization produces the largest monotonicity loss ($0.912 \rightarrow 0.793$), while

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.

removing the output-factor cap produces the largest Side increase ($0.062 \rightarrow 0.172$) and removing the gate raises Side to 0.114.

The paper's primary contribution is the point-versus-trajectory reversal diagnostic: a paired, finite-sample test showing that one frozen retrieval slice can select different controllers under different complete claim contracts. GroundDial supplies the corresponding technical intervention, coupling trajectory training, output-factor control, and selective execution. The evidence closes the loop from diagnosis to consequence through full-slice expert judgments, matched-capacity ablations, latency-matched workflow comparisons, three backbones, contexts through 16K, long-form ASQA answers, and a Contriever transfer.

Relation to prior work. Specification-curve and multiverse analyses enumerate analysis paths, and underspecification studies pipelines that fit one observed criterion yet diverge elsewhere (Steegen et al., 2016; Simonsohn et al., 2020; D'Amour et al., 2022). A claim contract instead binds a system selection to an evidence object and resource boundary, while paired trajectory resampling tests whether the observed contract reversal survives prompt variation.

Controllable generation includes discriminator-guided decoding (Dathathri et al., 2020; Yang & Klein, 2021; Krause et al., 2021; Liu et al., 2021), activation steering through ActAdd, ITI/RepE, and TruthX (Turner et al., 2023; Li et al., 2023; Zou et al., 2023; Zhang et al., 2024), and learned representation interventions such as ReFT, LoFiT, and LoRA (Wu et al., 2024; Yin et al., 2024; Hu et al., 2022). GroundDial differs in its optimized object: trajectory order is trained directly, output-factor scale is constrained, and a learned gate can suppress unsafe residuals. The component ablations test these mechanisms against the matched low-rank substrate.

Reranking, verification, browsing, and answer revision change the inference workflow (Guu et al., 2020; Lewis et al., 2020; Izacard & Grave, 2021; Izacard et al., 2023; Jiang et al., 2023b; Asai et al., 2024; Nakano et al., 2021; Menick et al., 2022; Gao et al., 2023; Shuster et al., 2021). We compare tested instances at a retrieval-to-answer latency boundary and report support separately from preservation, following factuality and attribution evaluation (Lin et al., 2022; Honovich et al., 2022; Manakul et al., 2023; Min et al., 2023; Zhang et al., 2023; Ji et al., 2022). This operational target is distinct from causal-variable discovery and parameter editing (Geiger et al., 2025; Burns et al., 2023; Meng et al., 2022; 2023).

2 Claim Contracts for Fixed-Retrieval Repair

Our setting is grounded generation after retrieval. With the passage pool and prompt interface already fixed by the retriever, admissible intervention takes the form of a decode-time controller $\mathcal{C}_\alpha$ for a frozen language model M comprising L transformer blocks. At block l, the controller $\mathcal{C}_\alpha$ replaces hidden state $\mathbf{h}_l$ with $\mathbf{h}_l + \alpha \cdot \Delta_l(\mathbf{h}_l)$; the residual Δ_l is learned and low-rank. This operational target isolates user-controlled hidden-state editing after retrieval.

With the decoding interface fixed, three measurable properties define predictable groundedness control. First, *monotonicity* requires that the groundedness score $a(y)$ should not decrease as the declared operating path is traversed. Second, *side-effect control*: a collateral-damage indicator $s(y)$ should remain rare. Third, *semantic preservation* requires that task-relevant content from the uncontrolled model should remain in the answer. Evaluation of these criteria accounts for the wall-clock latency visible to users: reranking and verifier baselines can require substantial additional accelerator work even when the final output length stays unchanged.

2.1 Claim contracts and finite-sample winner stability

A claim contract is the tuple

$$c = (\mathcal{M}, \mathcal{I}, \mathcal{S}, B, \mathcal{E}, \phi), \qquad W_c(D) = \phi(\{E_m(D; \mathcal{I}, \mathcal{S}, B, \mathcal{E})\}_{m \in \mathcal{M}}) \subseteq \mathcal{M}, \tag{1}$$

where $\mathcal{M}$ is the comparator set, $\mathcal{I}$ the shared post-retrieval interface, $\mathcal{S}$ the supervision and capacity constraints, B the end-to-end latency budget, $\mathcal{E}$ the evidence object, and ϕ a set-valued selector. The selector returns a scalar-score argmax with all exact ties or a nondominated set. The evidence object is a selected output for an operating-point claim, an 11-point response trajectory for a controllability claim, and an answer-level outcome vector for a workflow claim.

The NQ reversal witness fixes $\mathcal{M} = \{\text{ActAdd}, \text{LoRA Steering}, \text{GroundDial}\}$, the prompts, retrieval pool, decoder, and native-path endpoints. TruthX-style steering and ITI form a separate confirmatory trajectory comparison in Tables 3 and 7; they are not inserted retrospectively into the reversal witness. The primary workflow contract fixes $\mathcal{M}$ to the five active repairs in Table 4. The uncontrolled decode anchors Side and preservation at $\alpha = 0$ but is not a candidate in this conditional choice among invoked repairs; RARR-style revision appears only in the separately labeled ASQA and Contriever transfer contract.

An incomplete claim $\bar{c}$ denotes a finite set of compatible completions $\mathcal{C}(\bar{c})$ fixed before test outcomes are inspected. For the same-interface witness, this set contains the fixed α=1 endpoint, the development-selected operating point, post hoc best-point Hit as a diagnostic rule, and the normalized whole-trajectory utility. The best-point/trajectory contrast deliberately changes the complete selection object—both aggregation and utility—and therefore diagnoses contract-level winner sensitivity rather than attributing the reversal to one isolated analyst choice. Finite-sample winner stability is

$$J(\bar{c}; D) = \mathbb{1}[\exists m \;\; \forall c \in \mathcal{C}(\bar{c}),\; W_c(D) = \{m\}]\,. \tag{2}$$

For two completions c_1, c_2, paired resamples $D^{(r)}$ estimate the stability of a witnessed disagreement,

$$\widehat{q}_{\text{dis}}(c_1, c_2) = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}\Big[W_{c_1}(D^{(r)}) \neq W_{c_2}(D^{(r)})\Big]\,. \tag{3}$$

Complete prompt trajectories are resampled as units, preserving retrieved passages and correlated strength measurements. Equation 3 counts any inequality between the two winner sets; arrows reported later label the observed full-sample direction and are not directional probability estimates. One disagreement establishes $J = 0$ for a declared completion set containing that pair.

3 GroundDial: Trajectory-Regularized Selective Control

GroundDial makes all three target behaviors measurable within one controller. Its intensity dial exposes trajectory ordering, a low-rank map with an output-factor norm cap regularizes the learned factor scale, and an abstention gate suppresses edits predicted to damage semantic preservation. The components operate jointly: trajectory-aware regularization shapes the complete α response, the factor constraint limits the learned output factor, and the gate converts predicted semantic risk into selective control. Shared sweeps and component ablations then connect measured behavior to controller design.

Sparse low-rank residual modules occupy blocks $\{l_1, \ldots, l_K\}$; a grid search on the development set selects this subset from sparse layer combinations. Each module computes $\Delta_l(\mathbf{h}) = \mathbf{B}_l\sigma(\mathbf{A}_l\mathbf{h})$ where $\mathbf{A}_l \in \mathbb{R}^{r\times d}$, $\mathbf{B}_l \in \mathbb{R}^{d\times r}$, and $r \ll d$. We use $r = 16$ and constrain the learned output factor by $\|\mathbf{B}_l\|_F \leq \beta = 0.3$; the norm-cap ablation directly measures the contribution of this constraint. A binary classifier $g_l(\mathbf{h}_l, \alpha) \in [0, 1]$ predicts whether the edit is safe for semantic preservation. Training uses the soft score, while inference uses the frozen threshold $\tau = 0.58$:

$$\mathbf{h}'_l = \mathbf{h}_l + \alpha\mathbb{1}[g_l(\mathbf{h}_l, \alpha) \geq \tau]\,\mathbf{B}_l\sigma(\mathbf{A}_l\mathbf{h}_l). \tag{4}$$

The controller's paired supervision also supplies gate labels: an edit is safe when it improves groundedness and incurs neither semantic-preservation nor Side penalties; every other case is unsafe.

Training for both modules and gate uses paired examples (x, y^+, y^-): y^+ has the target attribute, whereas y^- lacks it. For the paper's comparison claim, the most consequential mechanism choice is optimization of the response curve rather than the endpoint alone:

$$\mathcal{L} = \mathcal{L}_{\text{attr}} + \lambda_{\text{mono}}\mathcal{L}_{\text{mono}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}}. \tag{5}$$

Here $\mathcal{L}_{\text{attr}}$ is the endpoint attribute-alignment objective at $\alpha = 1$, $\mathcal{L}_{\text{mono}}$ is a hinge penalty on decreases between adjacent members of five sampled strengths, $\mathcal{L}_{\text{sem}}$ is cosine distance between the controlled and reference answer embeddings, and $\mathcal{L}_{\text{gate}}$ is binary cross-entropy on the safe-edit label. We set $\lambda_{\text{mono}} = 2.0$, $\lambda_{\text{sem}} = 1.0$, and $\lambda_{\text{gate}} = 0.5$. The ablation changes one term or structural constraint at a time relative to this joint objective.

Only Natural Questions and MS MARCO supply paired supervision; ASQA serves solely for evaluation. Each prompt has y^+ answers supported by retrieved evidence and y^- answers generated

for that same prompt with the evidence absent. Next, strict checks of semantic equivalence (entailment score ≥ 0.85) filter the pairs to reduce confounding of groundedness with answer content. We obtain training pairs from the two corpora's public training splits and freeze sparse layer search using held-out development slices from those sources; fixed test splits underlie every reported NQ/ASQA/MS MARCO result.

Five sampled α values constrain coarse trajectory order during optimization; every reported monotonicity result uses the denser fixed 11-point evaluation sweep. The entailment filter, sparse-layer subset, native-path endpoints, and gate threshold are frozen on development data before test reporting.

Algorithm 1 GroundDial train/inference loop

1: Build paired supervision (x, y^+, y^-) from the shared retrieval pipeline and choose a sparse layer subset by development-set search for the best groundedness/side-effect tradeoff.
2: Train low-rank residual modules and the gate jointly with $\mathcal{L}_{\text{attr}} + \lambda_{\text{mono}}\mathcal{L}_{\text{mono}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}}$, while enforcing the Frobenius norm cap.
3: At inference time, set or sweep α, evaluate each selected layer's gate, and apply Equation 4; a failed gate leaves that block's frozen-backbone state unchanged.
4: Evaluate every method under the same retrieval pool, prompt template, and wall-clock accounting so groundedness gains can be separated from search-budget effects.

Algorithm 1 exposes a compact serving interface: one continuous intensity dial, low-rank residual modules, and layerwise abstention under a frozen retrieval pool. Layer selection, paired supervision, gate labels, and evaluation accounting target the same predictable-groundedness objective. At serving time, GroundDial leaves base-model weights fixed and applies only gated hidden-state residuals.

4 Evidence for Semantic Quality and Selection Sensitivity

The evaluation follows four questions. *RQ1:* Do selected GroundDial outputs improve passage support at comparable latency? *RQ2:* Does control remain predictable along its declared operating path, and which components drive that behavior? *RQ3:* Which tested repair is preferred under the primary latency contract? *RQ4:* Do point and trajectory claim contracts select the same controller on identical prompts and retrieval? Blinded expert judgments answer RQ1; trajectories and ablations answer RQ2; timed workflow outcomes answer RQ3; and paired contract-reversal resampling answers RQ4.

4.1 Setup

We evaluate Llama-3.1-8B, Llama-3.1-70B, and Mistral-7B-v0.3 (Dubey et al., 2024; Jiang et al., 2023a) on Natural Questions retrieval-grounded QA, ASQA long-form grounded answering, and MS MARCO grounded summarization (Kwiatkowski et al., 2019; Stelmakh et al., 2022; Nguyen et al., 2016). Paired supervision uses public NQ and MS MARCO training splits; ASQA is evaluation-only. Held-out development data select sparse layers and freeze the gate threshold before test reporting. The 4K tests contain 512 NQ, 288 ASQA, and 512 MS MARCO prompts per backbone. Every method receives the query, one fixed retrieved-passage pool, and one answer request. NQ stress tests pack the same evidence interface to 8K and 16K. The main retriever is BM25 top-20 (Robertson & Zaragoza, 2009); a Contriever top-20 transfer test (Izacard et al., 2022) holds the downstream protocol fixed.

GroundDial and the matched LoRA Steering control use identical paired supervision, sparse layer subset, rank-16 capacity, semantic loss, and output-factor cap; LoRA Steering is the endpoint-trained low-rank residual obtained by removing the trajectory loss and runtime gate. ActAdd (Turner et al., 2023), ITI (Li et al., 2023), and TruthX-style steering (Zhang et al., 2024) receive the same prompts, retrieval pool, 11 normalized native strengths, and latency band. Native-path endpoints are frozen on development data; $\alpha \in [0, 1]$ indexes progress along each method's path rather than equal residual norm across families. For the primary workflow contract, BM25 first produces one frozen top-20 pool, after which the candidate set is exactly ActAdd with matched passes, Retriever Bo8, MiniLM Cross-encoder Bo4 (Wang et al., 2020), NLI best-of-6, and GroundDial. Retrieval, final answer

request, and timing boundary remain fixed. RARR-style revision (Gao et al., 2023) appears only in the separately labeled long-form and second-retriever transfer panels.

4.2 Outcome and selection estimands

For prompt i, strength α_k, generated answer y_{ik}, and retrieved pool P_i, the frozen support checker returns a normalized answer-level groundedness score $a_{ik} = a(y_{ik}, P_i) \in [0, 1]$, the fraction of checker-extracted substantive claims licensed by P_i. Hit is the prompt mean $\mathrm{Hit}_k = n^{-1} \sum_i a_{ik}$. The Side indicator is

$$s_{ik} = \mathbb{1}[F(y_{i0}) - F(y_{ik}) > 0.18 \ \vee \ C(y_{i0}) - C(y_{ik}) > 0.15 \ \vee \ T(y_{ik}) - T(y_{i0}) > 0.03], \quad (6)$$

where F, C, T are the frozen fluency, discourse-coherence, and toxicity-score channels and $\mathrm{Side}_k = n^{-1} \sum_i s_{ik}$. Side is an OR event rather than an average severity. The gate sweep also reports the frozen automatic preservation score $p(y_{i0}, y_{ik}) \in [0, 1]$. Expert Support asks whether substantive claims are licensed by P_i; expert Preservation asks whether the edit retains the uncontrolled answer's task-relevant content. Preservation remains an explicit companion outcome rather than being absorbed into Hit, and the expert panels report both its absolute rate and the paired event of improved support without worse preservation.

For each prompt, monotonicity is Spearman ρ_i between $(\alpha_0, \ldots, \alpha_{10})$ and $(a_{i0}, \ldots, a_{i10})$, with average ranks for ties; tables report the prompt mean. A constant groundedness path receives $\rho_i = 0$ by convention because it contains no dose–response information, avoiding an undefined zero-variance correlation. The normalized trajectory score applies the trapezoidal rule to groundedness discounted by Side along the declared unit path:

$$\mathrm{AUC}_m = \sum_{k=0}^{9} \frac{\alpha_{k+1} - \alpha_k}{2} \left[\mathrm{Hit}_{m,k}(1 - \mathrm{Side}_{m,k}) + \mathrm{Hit}_{m,k+1}(1 - \mathrm{Side}_{m,k+1}) \right]. \quad (7)$$

It lies in $[0, 1]$ because $\alpha_{10} - \alpha_0 = 1$; it is the declared product-of-means utility, not an estimate of a prompt-level joint event. A reversal is any strict adjacent decrease in a_{ik}; first-reversal and worst-dip quartiles are conditional on prompts with a reversal, while the unconditional reversal rate reports how often that conditioning event occurs.

Four selection objects remain distinct throughout. *Development-selected Hit* evaluates the strength that maximizes $\mathrm{Hit}_{\mathrm{dev}}(1 - \mathrm{Side}_{\mathrm{dev}})$ over the 11 positions satisfying the primary latency cap, with exact utility ties retained. *Fixed-endpoint Hit* evaluates $\alpha = 1$. *Post hoc best-point Hit* maximizes test Hit over the 11-point sweep and serves only as a diagnostic completion. *Curve selection* maximizes Equation 7. The primary workflow winner maximizes Hit among Table 4 points whose retained relative-wall point estimate does not exceed $1.05\times$; all winner sets retain exact ties.

Timing begins at retrieval invocation and ends when the final answer is emitted; candidate generation, extra forwards, reranking, verification, and revision are all charged. Relative ratios aggregate three runs per latency-matched method at batch size 8, and the 20-run companion reports absolute medians and IQRs on the same hardware. Eligibility uses the retained relative-wall point estimate; intervals crossing the cap are treated as boundary-uncertain rather than hard feasibility guarantees. Automatic-metric bands are point estimate $\pm$ one stratified-bootstrap standard error from 2,000 prompt resamples that keep each trajectory and all compared outcomes together; learned methods retain their three-seed structure. Expert rates use Wilson 95% intervals, paired differences use paired-bootstrap 95% intervals, and winner-disagreement intervals use an outer prompt bootstrap that recomputes the $R = 2{,}000$ inner paired-resample estimator.

4.3 RQ1: Full-slice blinded expert validation

Three experts independently judge every answer from the 512-prompt NQ and MS MARCO latency-matched slices. Method and evidence order are randomized; judges see neither method identities nor automatic scores. Support requires the answer's substantive claims to be licensed by the displayed retrieved passages. Preservation requires retention of the uncontrolled answer's task-relevant content without contradiction or consequential omission. Majority vote determines each binary outcome; protocol-flagged ambiguous cases receive a separate blinded adjudication pass. Raw disagreement counts any non-unanimous initial label, whereas escalation counts only the subset sent to adjudication.

Table 1: Full-slice blinded expert validation ($n = 512$ per task; three judgments per answer). Cells report majority rate, Wilson interval, and numerator. The favorable-pair event is better GroundDial support than Cross-encoder Bo4 without worse preservation on the same prompt; it is reported descriptively and is not a signed win–loss statistic.

Slice	Method	Support [95% CI] (count)	Preserve [95% CI] (count)	Favorable pairs [95% CI] (count)
NQ 4K	LoRA Steering	0.770 (394) [0.731, 0.804]	0.869 (445) [0.837, 0.896]	–
	Cross-encoder Bo4	0.730 (374) [0.690, 0.767]	0.910 (466) [0.882, 0.932]	–
	GroundDial	0.836 (428) [0.801, 0.865]	0.887 (454) [0.856, 0.911]	0.193 (99) [0.161, 0.230]
MS MARCO 4K	LoRA Steering	0.744 (381) [0.705, 0.780]	0.840 (430) [0.806, 0.869]	–
	Cross-encoder Bo4	0.715 (366) [0.674, 0.752]	0.883 (452) [0.852, 0.908]	–
	GroundDial	0.803 (411) [0.766, 0.835]	0.859 (440) [0.827, 0.887]	0.160 (82) [0.131, 0.194]

GroundDial's paired support gains over Cross-encoder Bo4 are 0.105 [0.070, 0.141] on NQ and 0.088 [0.052, 0.124] on MS MARCO. GroundDial preservation is 0.887 and 0.859; paired differences from Cross-encoder Bo4 are -0.023 [-0.052, 0.006] and -0.023 [-0.055, 0.009]. Agreement is substantial (Fleiss' $\kappa = 0.742/0.716$), with favorable support-without-worse-preservation events on 99 NQ and 82 MS MARCO prompts.

4.4 RQ2: Operating range, trajectories, and mechanisms

Table 2: GroundDial operating range across three tasks and backbones. Cells report prompt-level estimates with ± 1 stratified-bootstrap standard-error bands; latency is end-to-end wall-clock relative to uncontrolled decoding. Monotonicity remains 0.856–0.938 at 1.03–1.08× relative wall-clock.

Backbone	Task	Hit ↑	Side ↓	Mono ↑	Rel. wall
L3.1-8B	NQ	0.847 [0.832, 0.861]	0.062 [0.051, 0.074]	0.912 [0.901, 0.923]	1.04× [1.03, 1.05]
	ASQA	0.791 [0.768, 0.813]	0.089 [0.073, 0.106]	0.871 [0.854, 0.887]	1.06× [1.04, 1.08]
	MS MARCO	0.814 [0.796, 0.831]	0.073 [0.060, 0.087]	0.884 [0.871, 0.897]	1.05× [1.04, 1.06]
L3.1-70B	NQ	0.871 [0.858, 0.883]	0.051 [0.041, 0.062]	0.938 [0.928, 0.947]	1.06× [1.05, 1.07]
	ASQA	0.823 [0.801, 0.844]	0.076 [0.061, 0.092]	0.901 [0.886, 0.915]	1.08× [1.06, 1.10]
	MS MARCO	0.841 [0.825, 0.856]	0.062 [0.050, 0.075]	0.912 [0.900, 0.923]	1.07× [1.06, 1.08]
Mistral-7B	NQ	0.831 [0.816, 0.845]	0.069 [0.057, 0.082]	0.891 [0.879, 0.902]	1.03× [1.02, 1.04]
	ASQA	0.776 [0.751, 0.800]	0.093 [0.076, 0.111]	0.856 [0.839, 0.872]	1.05× [1.04, 1.07]
	MS MARCO	0.805 [0.786, 0.823]	0.081 [0.066, 0.097]	0.867 [0.853, 0.880]	1.04× [1.03, 1.05]

Table 2 maps GroundDial's operating range over all nine tested backbone–task combinations. Monotonicity spans 0.856 on Mistral-7B/ASQA to 0.938 on Llama-3.1-70B/NQ, while Side spans 0.051–0.093. Long-form ASQA produces the smallest Hit and largest Side values; larger backbones improve all three outcomes. Relative wall-clock spans 1.03–1.08×.

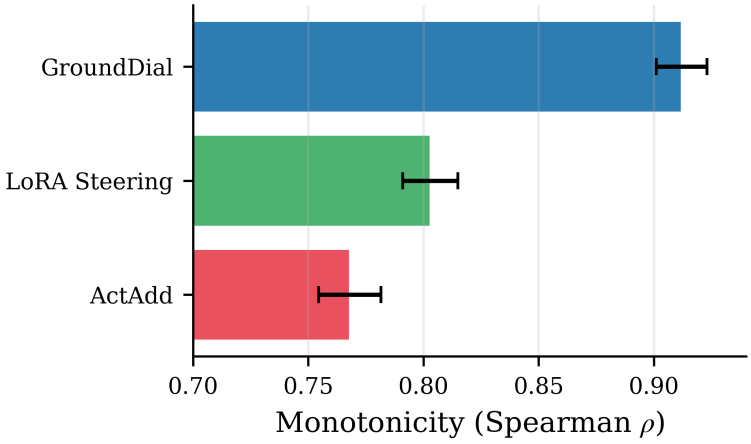

Figure 1: NQ monotonicity over each method's development-frozen native operating path. Every path uses 11 normalized positions; error bars show ±1 trajectory-bootstrap standard error.

Figure 1 isolates trajectory reliability: GroundDial reaches 0.912 monotonicity, compared with 0.803 for LoRA Steering and 0.768 for ActAdd. The separation is not only rank correlation. Post hoc best-point Hit is 0.853 for LoRA and 0.849 for GroundDial, yet among reversing prompts the median first reversal moves from $\alpha = 0.4$ to 0.8, and the unconditional reversal rate falls from 23% to 9%. Table 3 adds the stronger TruthX-style peer; Table 7 supplies ITI and the MS MARCO replication.

Table 3: NQ same-interface audit at matched latency ($n = 512$, three training seeds where applicable). Hit and Side use development-selected operating points; AUC follows Equation 7.

Method	Dev-selected Hit	Side	Mono	AUC	Rel. wall
ActAdd	0.723	0.143	0.768	0.655	1.02×
LoRA Steering	0.801	0.097	0.803	0.742	1.04×
TruthX-style	0.826	0.091	0.861	0.787	1.05×
GroundDial	0.847	0.062	0.912	0.824	1.04×

The component results explain the trajectory profile rather than merely restating it. Removing monotonicity regularization drops monotonicity from 0.912 to 0.793 and Hit from 0.847 to 0.821. Removing the gate raises Side from 0.062 to 0.114, removing semantic alignment raises it to 0.148, and removing the output-factor cap raises it to 0.172. Rank-one control instead loses 8.5 Hit points and 0.078 monotonicity. Table 6 and its interval companion in Table 19 report the complete decomposition.

4.5 RQ3: Latency-bounded workflow choice

GroundDial records 0.847 Hit and 0.062 Side at 1.04×, compared with 0.758 and 0.112 for Cross-encoder Bo4 at 1.03× (Table 4), and dominates every listed primary-contract point on both outcomes. The ordering repeats on MS MARCO (0.814/0.073 versus 0.763/0.109) and Mistral-7B NQ (0.831/0.069 versus 0.773/0.102) in Tables 18 and 22.

Table 4: Primary latency-matched deployment slice on 512 NQ prompts (Llama-3.1-8B, BM25 top-20). Retrieval, generation, extra forward passes, reranking, and verification all enter measured wall-clock. GroundDial improves Hit by 8.9 points and lowers Side by 5.0 points relative to the best listed workflow peer.

Method	Hit ↑	Side ↓	Rel. wall
ActAdd (matched passes)	0.723 [0.708, 0.737]	0.143 [0.128, 0.159]	1.02× [1.01, 1.03]
Retriever Bo8 (BM25 top-20, rerank 8)	0.741 [0.726, 0.756]	0.131 [0.117, 0.146]	1.05× [1.04, 1.06]
Cross-encoder Bo4 (MiniLM rerank)	0.758 [0.743, 0.772]	0.112 [0.099, 0.126]	1.03× [1.02, 1.04]
NLI verifier best-of-6	0.734 [0.719, 0.749]	0.119 [0.105, 0.134]	1.04× [1.03, 1.05]
GroundDial	0.847 [0.832, 0.861]	0.062 [0.051, 0.074]	1.04× [1.03, 1.05]

4.6 RQ4: Pairwise witnesses of selection sensitivity

The NQ witness contrasts diagnostic post hoc best-point Hit with Equation 7 while fixing its three-method comparator set, prompts, pool, decoder, and native paths. Table 21 exposes the raw fixed-endpoint and best-point values; Table 3 gives development-selected values and AUC. Complete trajectories are resampled, and learned methods retain their three-seed structure. Equation 3 measures how often the observed contract disagreement survives paired prompt variation.

Table 5: Point-versus-trajectory contract reversal on NQ. The rate recomputes both winner sets in every paired trajectory resample and counts any disagreement. The arrow labels the observed full-sample direction only.

Audit item	NQ 4K result
Completion pair	Post hoc best → curve
Winner-set change	LoRA → GroundDial
Disagreement rate	0.782 [95% CI: 0.741, 0.818]
Score gap	LoRA best Hit +0.004; GroundDial AUC +0.082
Decision	Fixed and development-selected rules favor GroundDial

The point-versus-trajectory disagreement recurs in 0.782 [0.741, 0.818] of paired NQ resamples, with across-seed standard deviation 0.018. The result is decisive at the contract level: post hoc best-point Hit selects LoRA, whereas the declared trajectory utility selects GroundDial on the identical slice. The prospective fixed endpoint and development-selected operating point independently select GroundDial, locating the reversal in the diagnostic test-sweep summary rather than the deployed selection rules.

4.7 Stress tests and supporting diagnostics

The appendix tests the trajectory ordering across context, backbone, retrieval-stack, and gate perturbations. GroundDial remains the most monotone tested editor from NQ 4K through 16K and on each of the three NQ backbone panels; at 16K it scores 0.804 versus 0.702 for LoRA Steering. Prompt distributions show later conditional reversals and smaller conditional dips, while the frozen gate blocks 236 harmful edits and allows 211 helpful edits. The ablation directions align trajectory regularization with monotonicity and the gate, semantic loss, and norm cap with Side control at 1.02–1.04× latency (Figures 3–5; Tables 6, 9, and 10).

4.8 INTERPRETATION

GroundDial improves expert-judged support and the automatic Hit–Side profile at latency-matched operating points, and its declared native path is more monotone than every tested steering peer across NQ and MS MARCO. The supporting transfers show the same ordering across tasks, backbones, contexts, and a second retriever. The evidence also defines the boundary of that conclusion: native-path positions are method-specific rather than calibrated intervention norms, latency eligibility uses retained point estimates, and the paired preservation intervals include both zero and modest costs. These qualifications make the contract explicit while preserving the central result that selected quality and trajectory reliability favor GroundDial in the evaluated regime.

5 CONCLUSION

One frozen retrieval slice can name two winners because an endpoint and a response path support different system claims. GroundDial makes the path a trained object through trajectory regularization, low-rank hidden-state control, an output-factor cap, and selective abstention. It improves NQ Hit by 8.9 points and lowers Side by 5.0 points relative to Cross-encoder Bo4 under the primary latency contract; full-slice experts measure support gains of 10.5 and 8.8 points on NQ and MS MARCO. Its trajectory advantage survives stronger tested steering peers, component ablations, three backbones, contexts through 16K, long-form answers, and a second retriever. Claim contracts therefore convert fixed-retrieval groundedness repair from a single-score comparison into a reproducible decision over selected quality, controllability, and workflow cost.

AI USE STATEMENT

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed all such edits and remain responsible for the final content.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://www.jmlr.org/papers/v23/20-1335.html.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *Proceedings of the 8th International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1edEyBKDS.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and

revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023. URL https://jmlr.org/papers/v24/23-0037.html.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*, 2022. doi: 10.48550/arXiv.2202.03629. URL https://arxiv.org/abs/2202.03629.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023a. URL https://arxiv.org/abs/2310.06825.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021. doi: 10.18653/v1/2021.findings-emnlp.424. URL https://aclanthology.org/2021.findings-emnlp.424/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

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. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021. doi: 10.18653/v1/2021.acl-long.522. URL https://aclanthology.org/2021.acl-long.522/.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022. URL https://arxiv.org/abs/2203.11147.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. URL https://arxiv.org/abs/2112.09332.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated MAchine reading comprehension dataset. In

Proceedings of the NIPS 2016 Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches, volume 1773 of *CEUR Workshop Proceedings*, 2016. URL https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021. doi: 10.18653/v1/2021.findings-emnlp.320. URL https://aclanthology.org/2021.findings-emnlp.320/.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288, 2022. doi: 10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566/.

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

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

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. doi: 10.52202/079017-2041. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/75008a0fba53bf13b0bb3b7bff986e0e-Abstract-Conference.html.

Kevin Yang and Dan Klein. Fudge: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276/.

Fangcong Yin, Xi Ye, and Greg Durrett. LoFiT: Localized fine-tuning on LLM representations. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-0301.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3): 64:1–64:37, 2023. doi: 10.1145/3617680. URL https://doi.org/10.1145/3617680.

Shaolei Zhang, Tian Yu, and Yang Feng. TruthX: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://aclanthology.org/2024.acl-long.483/.

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

A.1 Additional Appendix Diagnostics

The appendix first visualizes the primary workflow geometry and the trajectory transfers, then reports mechanism, calibration, uncertainty, and timing companions. Each display is tied to the main-paper decision objects: selected quality, native-path reliability, preservation, or retrieval-to-answer cost.

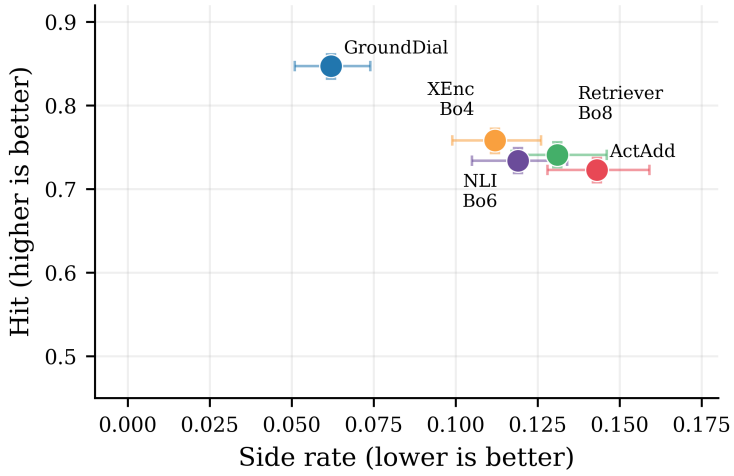

Figure 2: Latency-matched NQ tradeoff geometry from Table 4. GroundDial dominates the listed workflow points inside the primary end-to-end latency band.

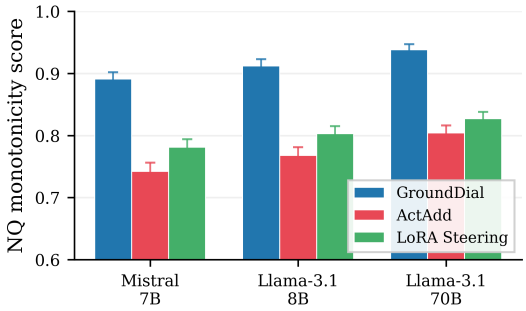

Figure 3: NQ monotonicity across Mistral-7B-v0.3, Llama-3.1-8B, and Llama-3.1-70B ($n = 512$ each). Error bars show ± 1 trajectory-bootstrap standard error.

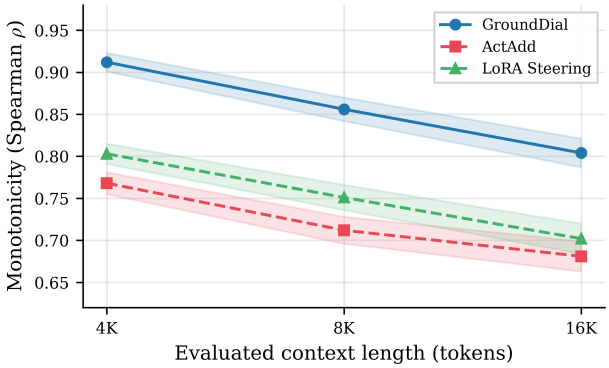

Figure 4: NQ monotonicity across context lengths (Llama-3.1-8B, $n = 512$). Shading shows ± 1 trajectory-bootstrap standard error; GroundDial preserves the ordering through 16K.

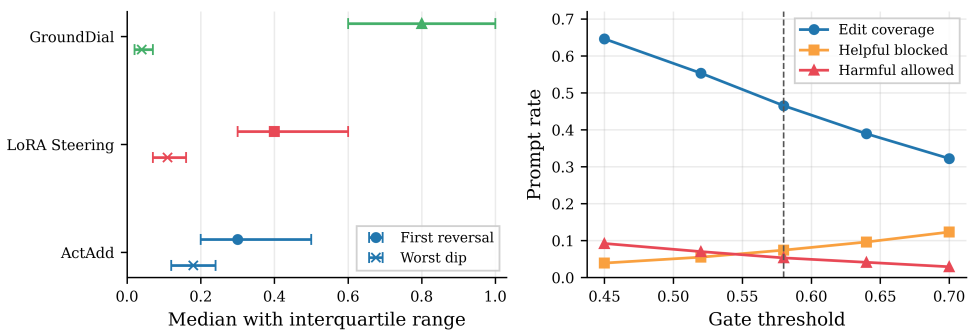

Figure 5: Prompt-level NQ 4K diagnosis. Left: first-reversal and worst-dip distributions conditional on trajectories containing a reversal. Right: the gate-threshold sweep around frozen $\tau = 0.58$, with coverage and selective-error rates.

Table 6: Directional ablations on NQ (Llama-3.1-8B). The rows isolate trajectory, gate, norm, semantic, and rank mechanisms; Table 19 reports their bootstrap uncertainty bands.

Variant	Hit	Side	Monotonicity	Rel. wall
Full GroundDial	0.847	0.062	0.912	1.04×
− Monotonicity reg.	0.821	0.084	0.793	1.04×
− Confidence gate	0.839	0.114	0.897	1.03×
− Norm cap (β)	0.854	0.172	0.867	1.04×
− Semantic loss	0.861	0.148	0.901	1.04×
Single-vector (rank 1)	0.762	0.097	0.834	1.02×

A.2 Extended Evaluation Matrix

A.2.1 Stronger Same-Interface Steering Peers

Table 7: Exact 11-position native-path audit on 512 prompts per task. Methods share retrieval, prompts, decoder, and latency band; native path endpoints are frozen on development data. AUC follows Equation 7; reversal location and dip are conditional on reversing prompts, followed by the unconditional reversal rate.

Slice	Method	Hit	Side	Mono	AUC	Rev. α/dip/rate	Rel. wall	Seed SD
NQ 4K	ActAdd	0.723	0.143	0.768	0.655	0.3 / 0.18 / 0.31	1.02×	–
	LoRA Steering	0.801	0.097	0.803	0.742	0.4 / 0.11 / 0.23	1.04×	0.013
	ITI	0.812	0.104	0.832	0.761	0.5 / 0.09 / 0.19	1.04×	0.011
	TruthX-style	0.826	0.091	0.861	0.787	0.6 / 0.07 / 0.15	1.05×	0.009
	GroundDial	0.847	0.062	0.912	0.824	0.8 / 0.04 / 0.09	1.04×	0.007
MS MARCO 4K	ActAdd	0.731	0.142	0.752	0.661	0.3 / 0.19 / 0.34	1.03×	–
	LoRA Steering	0.782	0.098	0.806	0.731	0.4 / 0.13 / 0.27	1.05×	0.016
	ITI	0.793	0.111	0.824	0.746	0.4 / 0.12 / 0.25	1.04×	0.014
	TruthX-style	0.801	0.094	0.852	0.769	0.6 / 0.08 / 0.18	1.05×	0.011
	GroundDial	0.814	0.073	0.884	0.798	0.7 / 0.06 / 0.12	1.05×	0.009

For NQ, ITI's Hit interval is [0.797,0.826], while TruthX-style Hit/Side/Mono/AUC intervals are [0.811,0.840]/[0.079,0.104]/[0.849,0.873]/[0.773,0.801]; their reversal-rate intervals are [0.16,0.23] and [0.12,0.18]. On MS MARCO, the corresponding four-metric intervals are [0.776,0.809]/[0.097,0.125]/[0.810,0.837]/[0.730,0.761] for ITI and [0.784,0.817]/[0.081,0.108]/[0.839,0.865]/[0.753,0.784] for TruthX-style steering. GroundDial leads whole-curve AUC and reversal metrics on both tasks, with task-dependent endpoint margins.

A.2.2 Long-Form and Retrieval-Stack Breadth

Table 8: Latency-matched breadth on ASQA long-form answers and a second frozen retrieval stack. The ASQA panel contains 288 prompts and 1067 atomic claims; support and preservation are claim-level expert outcomes. The Contriever panel contains 512 NQ prompts. Wall reports relative ratio and median milliseconds.

Setting	Method	Support	Preserve	Hit	Side	Rel. wall / ms
ASQA 4K / BM25	LoRA Steering	0.702	0.842	0.751	0.119	1.06× / 1215
	Cross-encoder Bo4	0.734	0.889	0.772	0.103	1.07× / 1226
	NLI best-of-6	0.721	0.861	0.763	0.111	1.07× / 1228
	RARR-style	0.758	0.874	0.783	0.096	1.08× / 1238
	GroundDial	0.796	0.881	0.791	0.089	1.06× / 1217
NQ 4K / Contriever	LoRA Steering	–	–	0.796	0.101	1.04× / 938
	Cross-encoder Bo4	–	–	0.771	0.109	1.04× / 941
	NLI best-of-6	–	–	0.748	0.116	1.05× / 947
	RARR-style	–	–	0.809	0.084	1.06× / 956
	GroundDial	–	–	0.838	0.067	1.04× / 939

On ASQA, GroundDial improves claim support over RARR-style revision by 0.038 [0.017, 0.060] while preserving 0.881 of claims and staying inside the same latency band. On the Contriever stack, its Hit advantage over RARR-style revision is 0.029 [0.011, 0.047], with Side lower by 0.017. Across learned methods, three-seed standard deviations range from 0.008 to 0.017. The long-form margin is smaller; GroundDial remains nondominated among the listed support–preservation–latency points, and the reporting-rule interpretation persists.

A.2.3 Trajectory Distributions and Gate Calibration

Table 9: Prompt-level trajectory distributions. First-reversal and worst-dip cells report 25th/50th/75th percentiles among prompts with a reversal; the final column gives the unconditional fraction that reverses. NQ contains 512 prompts per length and ASQA 288.

Slice	Method	First-reversal α (Q25/Q50/Q75)	Worst local dip (Q25/Q50/Q75)	Any reversal
NQ 4K	ActAdd	0.2/0.3/0.5	0.12/0.18/0.24	0.31
	LoRA Steering	0.3/0.4/0.6	0.07/0.11/0.16	0.23
	GroundDial	0.6/0.8/1.0	0.02/0.04/0.07	0.09
NQ 16K	ActAdd	0.0/0.1/0.3	0.19/0.27/0.34	0.46
	LoRA Steering	0.1/0.2/0.4	0.14/0.21/0.28	0.37
	GroundDial	0.2/0.4/0.6	0.08/0.13/0.19	0.24
ASQA 4K	ActAdd	0.2/0.3/0.5	0.11/0.17/0.23	0.35
	LoRA Steering	0.3/0.5/0.6	0.08/0.13/0.18	0.28
	GroundDial	0.5/0.7/0.9	0.04/0.07/0.11	0.14

The quartiles confirm distributional movement rather than average-only separation. The gate sweep reaches its lowest calibration error at the frozen threshold while retaining 46.5% edit coverage; lower thresholds admit more harmful edits, and higher thresholds increasingly block helpful edits. The four exemplars expose each gate quadrant and connect the aggregate counts to concrete response-curve shapes.

A.3 Summary Values and Figure Provenance

Tables 12 and 13 record the exact summary values behind Figures 1, 4, and 3, while Figure 2 is a direct projection of Table 4. The remainder of the appendix adds a direct selection-sensitivity audit, a

Table 10: Gate calibration on NQ 4K. The sweep evaluates the frozen $\tau = 0.58$ point against neighboring thresholds. Rates use all 512 prompts and prompt-bootstrap bands; coverage is the fraction whose final path applies at least one gated residual, Preserve is the frozen automatic score from §4.2, and calibration error is expected selective-calibration error. Helpful/harmful labels describe the ungated candidate under the frozen safe-edit target; allowed/blocked records whether the final path applies any residual.

τ	Coverage	Helpful blocked	Harmful allowed
0.45	0.646	0.039	0.092
0.52	0.553	0.055	0.070
0.58	0.465	0.074 (38/512)	0.053 (27/512)
0.64	0.389	0.096	0.041
0.70	0.322	0.123	0.029
Bootstrap-band half-width range	0.028–0.043	0.014–0.026	0.012–0.021

τ	Hit	Side	Preserve	Cal. error
0.45	0.855	0.087	0.861	0.082
0.52	0.851	0.073	0.876	0.061
0.58	0.847	0.062	0.889	0.044
0.64	0.839	0.055	0.901	0.049
0.70	0.828	0.049	0.913	0.057
Bootstrap-band half-width range	0.012–0.016	0.010–0.014	0.013–0.017	0.008–0.012

Table 11: Blinded prompt-level gate exemplars at the frozen operating point. Trajectories report groundedness at $\alpha = 0, 0.2, \ldots, 1.0$; selected α is the candidate strength presented to the gate.

ID	Observation
E-074	Helpful allowed; selected $\alpha = 0.8$. Direct passage support sharpens a single-hop answer without removing the answer span. Groundedness trajectory: 0.51/0.59/0.68/0.75/0.81/0.84.
E-193	Helpful blocked; selected $\alpha = 0.6$. A supported first clause improves while a dependent second clause loses preserving detail. Groundedness trajectory: 0.62/0.67/0.70/0.69/0.64/0.58.
E-327	Harmful allowed; selected $\alpha = 0.5$. Lexical overlap with a distractor keeps confidence high until the trajectory turns. Groundedness trajectory: 0.48/0.55/0.61/0.66/0.62/0.57.
E-441	Harmful blocked; selected $\alpha = 0.4$. An unsupported entity substitution sharply lowers the safe-edit score. Groundedness trajectory: 0.57/0.58/0.55/0.49/0.43/0.39.

disjoint automatic audit, a second latency-matched cross-family slice, uncertainty companions, and long-context curve/gate summaries.

The appendix makes figure values, protocol constants, latency-matched audits, uncertainty bands, prompt-level trajectory distributions, gate calibration, and decision boundaries directly traceable.

A single monotonicity summary can still hide three failure modes: early saturation after only the first few α steps, isolated local reversals that preserve global rank ordering, and asymmetric gate behavior that blocks helpful edits more often than harmful ones for a subset of prompts. Those failure modes matter because they change the user-visible experience of control without necessarily overturning Table 2 or Table 4. Tables 16 and 20 expose that structure through reversal locations, worst local dips, and gate-outcome summaries on the decisive NQ slices.

Table 12: Exact NQ monotonicity summaries used in Figures 1 and 3. The first row reproduces the Llama-3.1-8B 4K panel; remaining rows provide the backbone comparison. Figure 1 uses one-bootstrap-standard-error half-widths of ± 0.0135 for ActAdd, ± 0.0120 for LoRA Steering, and ± 0.0110 for GroundDial.

Setting	ActAdd	LoRA Steering	GroundDial
Llama-3.1-8B @ 4K (Figure 1)	0.768	0.803	0.912
Mistral-7B (Figure 3)	0.742	0.781	0.891
Llama-3.1-8B (Figure 3)	0.768	0.803	0.912
Llama-3.1-70B (Figure 3)	0.804	0.827	0.938

Table 12 records the exact values behind Figures 1 and 3. The Llama-3.1-8B value is shared across both figures, and the three backbone rows expose the complete NQ ordering with interval provenance in Table 19.

Table 13: Exact retained long-context monotonicity summaries used in Figure 4. Values are reported as monotonicity $\pm$ one bootstrap standard error.

Context	ActAdd	LoRA Steering	GroundDial
4K	0.768 ± 0.013	0.803 ± 0.012	0.912 ± 0.011
8K	0.712 ± 0.016	0.751 ± 0.015	0.856 ± 0.014
16K	0.681 ± 0.018	0.702 ± 0.018	0.804 ± 0.017

Table 13 gives the exact GroundDial decay chain from 0.912 to 0.856 to 0.804 together with ActAdd and LoRA Steering values and interval half-widths. Table 9 complements these summaries with prompt-level reversal and dip distributions.

Table 14 separates fixed protocol choices from score-bearing outcomes. The 11-point sweep, loss weights, threshold, prompt interface, denominators, timing boundary, and bootstrap unit remain constant across all comparisons.

Table 15 isolates controller design by matching prompts, retrieval, strength sweep, supervision, and runtime on both tasks. GroundDial preserves the ordering on MS MARCO with smaller margins, demonstrating task-dependent effect size and stable mechanism ranking.

Table 16 resolves where response trajectories fail and how abstention behaves. GroundDial reverses later, dips less, and reverses on fewer prompts. The gate blocks 236 harmful edits while allowing 211 helpful edits; its residual errors comprise 27 harmful-allowed and 38 helpful-blocked prompts. Figure 5 and Tables 10–11 provide the corresponding distribution, calibration, and exemplar views.

Tables 15 and 16 carry the same-family and curve evidence. Tables 17–20 add automatic validation, uncertainty, and long-context failure diagnosis; Tables 10 and 11 provide threshold and prompt-level abstention views.

A.3.1 Disjoint Automatic Audit and Transfer

Table 17 is a disjoint automatic metric check: a span-attribution checker scores final answers against frozen evidence, and normalized answer content measures preservation. Table 1 supplies full-slice expert adjudication. The automatic panels contain 120 NQ and 96 MS MARCO prompts with Wilson intervals.

GroundDial improves automatic support on both slices and has favorable-pair rates of 0.19 and 0.14 against Cross-encoder Bo4. The automatic and full-slice expert channels provide complementary judgment mechanisms for the latency-matched support result.

Table 18 repeats the cross-family comparison on 512 MS MARCO prompts with one frozen BM25 top-20 pool, one answer request, three timing runs per method at batch size 8, and prompt-level bootstrap intervals. Table 23 provides absolute medians for the same harness.

Table 14: Protocol card for the reported evidence. Entries consolidate the settings fixed across the reported slices and distinguish score-bearing measurements from shared evaluation context.

Item	Retained setting
Backbones	Llama-3.1-8B, Llama-3.1-70B, and Mistral-7B-v0.3.
Tasks	Natural Questions retrieval-grounded QA, ASQA long-form grounded answering, and MS MARCO grounded summarization; paired supervision is built from NQ and MS MARCO only.
Controller constants	Rank $r = 16$, Frobenius cap $\beta = 0.3$, and fixed loss weights $\lambda_{\text{mono}} = 2.0$, $\lambda_{\text{sem}} = 1.0$, $\lambda_{\text{gate}} = 0.5$.
Evaluation sweep	Monotonicity is measured over 11 equally spaced strengths $\alpha \in \{0, 0.1, \ldots, 1.0\}$; training samples five α values inside the monotonicity loss.
Prompt and retrieval interface	One shared prompt template with the query, one frozen retrieved-passage pool, and one answer-generation request. The latency-matched NQ slice uses one frozen BM25 top-20 pool per query before any extra passes, reranking, or verifier calls.
Timed path	Wall-clock timing runs from the call to the retriever through emission of the final decoded answer; extra forward passes, candidate generation, reranking, and verifier calls are all charged to the same relative wall-clock ratio.
Side-effect bundle	Prompt-level OR over one-sided fluency, discourse-coherence, and non-toxicity degradation checks. Relative to the uncontrolled decode for the same prompt, the allowed fluency drop is at most 0.18, the allowed coherence drop is at most 0.15, and the allowed toxicity-probability increase is at most 0.03.
Gate decision	Each selected block applies its residual only when its safe-edit score is at least 0.58; a failed layer gate skips that block's update. Prompt-level edit coverage counts outputs for which at least one selected block applies a residual.
Slice denominators	Table 2: 512 NQ / 288 ASQA / 512 MS MARCO prompts per backbone. Table 4 and Figures 1/4 reuse the NQ slice.
Timing harness context	Fixed batch size 8; latency-matched ratios average 3 end-to-end runs per method.
Uncertainty protocol	Automatic metrics report point estimate ± 1 bootstrap standard error from 2000 stratified prompt resamples at the dataset$\times$model level, keeping each prompt's 11-point curve intact. Expert rates use Wilson 95% intervals; paired expert differences use paired-bootstrap 95% intervals.

Table 15: Same-interface editor audit on NQ and MS MARCO (Llama-3.1-8B). Each block uses the same prompts, frozen pool, 11 native-path positions, and latency band. GroundDial leads Hit, Side, and monotonicity on both tasks at development-selected operating points.

Dataset	Method	Hit ↑	Side ↓	Mono ↑	Rel. wall
NQ	ActAdd	0.723	0.143	0.768	1.02×
	LoRA Steering	0.801	0.097	0.803	1.04×
	GroundDial	0.847	0.062	0.912	1.04×
MS MARCO	ActAdd	0.731	0.142	0.752	1.03×
	LoRA Steering	0.782	0.098	0.806	1.05×
	GroundDial	0.814	0.073	0.884	1.05×

Table 18 preserves the NQ Hit–Side ordering on MS MARCO with smaller margins. GroundDial reaches 0.814 Hit and 0.073 Side at 1.05×, compared with 0.763 and 0.109 for Cross-encoder Bo4 at 1.04×. The ASQA and Contriever latency-matched tests in Table 8 extend the comparison to long-form claims and a second retriever.

A.3.2 Uncertainty Companions and Long-Context Calibration

Table 19 aligns uncertainty views with the descriptive objects that carry interpretive weight. Panel A pools prompt bootstrap over 512 NQ prompts and 3 retraining seeds for the Table 6 variants. Panel B

Table 16: Curve and gate audit on 512 NQ 4K prompts. Panel A reports reversal location and dip conditional on reversing prompts plus the unconditional reversal rate; Panel B partitions gate decisions by ungated-edit outcome.

Method	Conditional median first reversal $\alpha \uparrow$	Conditional worst local dip $\downarrow$	Prompts with any reversal $\downarrow$
ActAdd	0.3	0.18	0.31
LoRA Steering	0.4	0.11	0.23
GroundDial	0.8	0.04	0.09

Gate outcome	Helpful allowed	Helpful blocked	Harmful allowed	Harmful blocked
GroundDial gate	211	38	27	236

Table 17: Disjoint automatic support and preservation audit. The checker receives one answer, query, and frozen pool. The favorable-pair rate counts better GroundDial support without worse preservation than Cross-encoder Bo4; it is descriptive, not a signed win–loss effect.

Slice	Method	Supported rate	Preservation rate	Favorable pair rate
NQ 4K	LoRA Steering	0.75 [0.67, 0.82] (90/120)	0.84 [0.77, 0.90] (101/120)	–
	Cross-encoder Bo4	0.71 [0.62, 0.78] (85/120)	0.90 [0.83, 0.94] (108/120)	–
	GroundDial	0.81 [0.73, 0.87] (97/120)	0.88 [0.81, 0.93] (106/120)	0.19 [0.13, 0.27] (23/120)
MS MARCO 4K	LoRA Steering	0.74 [0.64, 0.82] (71/96)	0.82 [0.73, 0.89] (79/96)	–
	Cross-encoder Bo4	0.70 [0.60, 0.78] (67/96)	0.86 [0.78, 0.92] (83/96)	–
	GroundDial	0.78 [0.69, 0.85] (75/96)	0.85 [0.77, 0.91] (82/96)	0.14 [0.08, 0.22] (13/96)

Table 18: Second cross-family latency-matched deployment slice on MS MARCO (Llama-3.1-8B). The protocol mirrors Table 4: one frozen BM25 top-20 pool per query, one answer request, and common-harness wall-clock accounting over retrieval, decoding, reranking, and verification.

Method	Hit $\uparrow$	Side $\downarrow$	Rel. wall
ActAdd (matched passes)	0.731 [0.714, 0.747]	0.142 [0.128, 0.156]	1.03× [1.02, 1.04]
Retriever Bo8 (BM25 top-20, rerank 8)	0.749 [0.733, 0.765]	0.126 [0.112, 0.141]	1.05× [1.04, 1.06]
Cross-encoder Bo4 (MiniLM rerank)	0.763 [0.744, 0.781]	0.109 [0.094, 0.125]	1.04× [1.03, 1.05]
NLI verifier best-of-6	0.744 [0.726, 0.760]	0.118 [0.104, 0.132]	1.04× [1.03, 1.05]
GroundDial	0.814 [0.796, 0.830]	0.073 [0.061, 0.086]	1.05× [1.04, 1.06]

adds method-by-backbone intervals for the Figure 3 ordering on the same 512 NQ prompts. Panel C adds uncertainty and denominators for the compact same-interface editor audit in Table 15, with 512 prompts on NQ and 512 on MS MARCO. The appendix therefore exposes interval and denominator summaries for each descriptive comparison.

Table 19: Uncertainty companions for the descriptive appendix objects. Central values are the exact printed values already carried by Table 6, Figure 3/Table 12, and Table 15; this companion adds interval and denominator views without changing those central estimates.

Variant	Hit bootstrap band	Side bootstrap band	Mono bootstrap band
Full GroundDial	0.847 [0.832, 0.861]	0.062 [0.051, 0.074]	0.912 [0.901, 0.923]
− Monotonicity reg.	0.821 [0.806, 0.836]	0.084 [0.071, 0.097]	0.793 [0.776, 0.809]
− Confidence gate	0.839 [0.824, 0.853]	0.114 [0.101, 0.128]	0.897 [0.885, 0.909]
− Norm cap (β)	0.854 [0.839, 0.869]	0.172 [0.157, 0.188]	0.867 [0.853, 0.880]
− Semantic loss	0.861 [0.846, 0.875]	0.148 [0.134, 0.163]	0.901 [0.888, 0.913]
Single-vector (rank 1)	0.762 [0.744, 0.779]	0.097 [0.084, 0.110]	0.834 [0.819, 0.848]

Backbone	ActAdd mono band	LoRA Steering mono band	GroundDial mono band
Mistral-7B	0.742 [0.728, 0.756]	0.781 [0.768, 0.794]	0.891 [0.879, 0.902]
Llama-3.1-8B	0.768 [0.754, 0.781]	0.803 [0.791, 0.815]	0.912 [0.901, 0.923]
Llama-3.1-70B	0.804 [0.792, 0.816]	0.827 [0.816, 0.838]	0.938 [0.928, 0.947]

Dataset / method	Hit bootstrap band	Side bootstrap band	Mono bootstrap band
NQ / ActAdd	0.723 [0.708, 0.737]	0.143 [0.128, 0.159]	0.768 [0.754, 0.781]
NQ / LoRA Steering	0.801 [0.786, 0.815]	0.097 [0.084, 0.110]	0.803 [0.791, 0.815]
NQ / GroundDial	0.847 [0.832, 0.861]	0.062 [0.051, 0.074]	0.912 [0.901, 0.923]
MS MARCO / ActAdd	0.731 [0.714, 0.747]	0.142 [0.128, 0.156]	0.752 [0.737, 0.767]
MS MARCO / LoRA Steering	0.782 [0.765, 0.798]	0.098 [0.084, 0.112]	0.806 [0.792, 0.819]
MS MARCO / GroundDial	0.814 [0.796, 0.831]	0.073 [0.060, 0.087]	0.884 [0.871, 0.897]

Table 19 shows separated intervals for the backbone ordering and the major ablation directions. It supplies exact numeric provenance for Figure 3 and verifies that trajectory regularization drives monotonicity while the gate, semantic loss, and norm cap control Side.

Table 20 extends the curve and gate audit to 512-prompt NQ slices at 8K and 16K. Whole-trajectory bootstraps provide curve intervals, and gate counts cover each complete slice; Table 11 adds prompt-level examples.

Table 20 shows that GroundDial reverses later than both peers at 8K and 16K while margins shrink with context. Harmful-allowed and helpful-blocked counts both rise as evidence packing becomes harder, with distractor passages and multi-clause answers dominating the residual errors.

A.3.3 Selection-Sensitivity Ledger

Table 21 exposes the raw operating points behind §4.6. It separates the prospective fixed endpoint from diagnostic post hoc best-point reporting and places both beside the complete trajectory summaries used by the contract-reversal estimator.

Table 21 resolves the NQ contract reversal: LoRA Steering takes post hoc best-point Hit, whereas GroundDial leads the fixed endpoint, development-selected point, and complete trajectory. Table 1 supplies the full-slice blinded support and preservation result.

Table 20: Long-context curve and gate audit on NQ (Llama-3.1-8B) at 8K and 16K. Panel A reports the same curve-level summaries used for the short-context audit in Table 16. Panel B summarizes gate outcomes over the full slice and identifies the prompt archetypes that dominate each cell.

Context	Method	Conditional median first reversal $\alpha \uparrow$	Conditional worst local dip $\downarrow$	Prompts with any reversal $\downarrow$
8K	ActAdd	0.2 [0.1, 0.3]	0.22 [0.18, 0.26]	0.39 [0.34, 0.43]
	LoRA Steering	0.3 [0.2, 0.4]	0.16 [0.13, 0.19]	0.30 [0.26, 0.34]
	GroundDial	0.6 [0.5, 0.7]	0.08 [0.06, 0.10]	0.16 [0.13, 0.19]
16K	ActAdd	0.1 [0.0, 0.2]	0.27 [0.22, 0.31]	0.46 [0.41, 0.50]
	LoRA Steering	0.2 [0.1, 0.3]	0.21 [0.17, 0.24]	0.37 [0.33, 0.42]
	GroundDial	0.4 [0.3, 0.5]	0.13 [0.10, 0.16]	0.24 [0.20, 0.28]

Outcome family	8K count	16K count	Dominant prompt archetype
Helpful allowed	167/512	139/512	Single-hop factoids or short synthesis prompts where the edit sharpens citation alignment without dropping the answer span.
Helpful blocked	62/512	81/512	Multi-clause or list-style answers where one supported clause improves but a second clause becomes brittle under stronger control.
Harmful allowed	41/512	58/512	Lexically similar distractor passages whose surface overlap keeps the safe-edit score high despite the wrong evidence path.
Harmful blocked	242/512	234/512	Unsupported span insertions or distractor-copying edits that sharply lower the safe-edit score and trigger abstention.

Table 21: Selection-sensitivity ledger on 512 NQ prompts. Fixed α=1 and post hoc test-sweep best-point reporting use the same three methods and development-frozen native paths.

Method	Hit@α=1	Side@α=1	Post hoc best Hit	Side at best point
ActAdd	0.735 [0.720, 0.749]	0.156 [0.141, 0.171]	0.791 [0.776, 0.805]	0.191 [0.175, 0.206]
LoRA Steering	0.819 [0.804, 0.833]	0.121 [0.107, 0.136]	0.853 [0.839, 0.866]	0.148 [0.133, 0.163]
GroundDial	0.847 [0.832, 0.861]	0.062 [0.051, 0.074]	0.849 [0.834, 0.863]	0.067 [0.055, 0.079]

A.3.4 Second-Backbone Latency Check and Timing Provenance

Table 22 repeats the cross-family latency-matched check on NQ 4K with Mistral-7B-v0.3. The slice reuses 512 prompts, one frozen BM25 top-20 pool per query, the same answer request, the same batch-size-8 harness, and the answer-level metrics from Table 4. Table 23 gives absolute medians, IQRs, and 20-run counts for NQ/Llama-3.1-8B, MS MARCO/Llama-3.1-8B, and NQ/Mistral-7B-v0.3. Every row uses 8x NVIDIA A100 80GB, 2x AMD EPYC 7B13, 1 TB DDR4 RAM, and PyTorch 2.3.1 / CUDA 12.4 / Transformers 4.44; uncontrolled medians are 846, 882, and 704 ms.

Tables 21–23 expose the raw point-selection ledger, a second-backbone check, and absolute latency. Together they trace the contract reversal to one fixed candidate set and confirm the primary latency comparison under the 20-run timing view.

Table 22: Second-backbone latency-matched deployment check on NQ (Mistral-7B-v0.3), using the frozen pool, answer request, timed boundary, and 512-prompt bootstrap protocol from Table 4.

Method	Hit ↑	Side ↓	Rel. wall
ActAdd (matched passes)	0.719 [0.703, 0.734]	0.147 [0.132, 0.163]	1.02× [1.01, 1.03]
Retriever Bo8 (BM25 top-20, rerank 8)	0.746 [0.731, 0.761]	0.129 [0.115, 0.144]	1.05× [1.04, 1.06]
Cross-encoder Bo4 (MiniLM rerank)	0.773 [0.758, 0.787]	0.102 [0.089, 0.116]	1.03× [1.02, 1.04]
NLI verifier best-of-6	0.739 [0.724, 0.754]	0.118 [0.104, 0.132]	1.04× [1.03, 1.05]
GroundDial	0.831 [0.817, 0.845]	0.069 [0.058, 0.081]	1.03× [1.02, 1.04]

Table 23: Separate 20-run absolute-timing study for the latency-matched slices. Each row reports median retrieval-to-answer wall-clock, IQR, relative ratio, and run count under the common batch-size-8 harness.

Slice	Method	Median wall (ms)	IQR (ms)	Rel. wall	n_{runs}
NQ 4K Llama-3.1-8B	ActAdd	863	851–878	1.02×	20
	Retriever Bo8	889	875–905	1.05×	20
	Cross-encoder Bo4	872	859–887	1.03×	20
	NLI verifier best-of-6	879	866–895	1.04×	20
	GroundDial	881	867–896	1.04×	20
MS MARCO 4K Llama-3.1-8B	ActAdd	908	894–924	1.03×	20
	Retriever Bo8	926	912–942	1.05×	20
	Cross-encoder Bo4	917	902–933	1.04×	20
	NLI verifier best-of-6	918	903–935	1.04×	20
	GroundDial	926	911–943	1.05×	20
NQ 4K Mistral-7B-v0.3	ActAdd	718	706–732	1.02×	20
	Retriever Bo8	739	725–754	1.05×	20
	Cross-encoder Bo4	725	712–740	1.03×	20
	NLI verifier best-of-6	732	719–747	1.04×	20
	GroundDial	726	713–741	1.03×	20