REVEAL: REVERSE ENGINEERING AGENTIC LATENTS FROM FRONTIER-MODEL TRACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier LLM agents perform complex interactive tasks, but their inference costs motivate transferring their experience to smaller open-source agents. Distillation and latent communication typically require target-model training or access to source-model internals. We study how textual frontier-model trajectories can provide reusable local controls for a frozen agent. Inspired by protocol reverse engineering, we introduce *Agentic Latents*, comprising anchored states and dynamic profiles of feedback–response cycles elicited by replay in the target model. Our training-free framework, **REVEAL** (**REV**erse **E**ngineering **A**gentic **L**atents), uses these profiles and trajectory outcomes to match and select local events from successful and failed rollouts of the same task and source model. Their response-transition contrasts yield a target-specific steering direction, applied before each response. Within-benchmark evaluations on Terminal-Bench 2.1 and Humanity's Last Exam with tools report higher accuracy than unmodified agents across three targets. Representation analyses and behavioral case studies support reusing frontier-model experience as local activation control.

1 INTRODUCTION

Large language model (LLM) agents perform complex tasks through repeated interactions with tools and environments. Frontier models offer strong capabilities for navigating these interactions, but repeatedly invoking them over long trajectories can incur substantial inference costs. Smaller open-source models offer a path to lower-cost and more flexible deployment, yet often struggle with complex tool use and sustained task execution (Qin et al., 2024b; Merrill et al., 2026). Frontier-model trajectories provide a record of how stronger agents reason, act, and respond to feedback. Reusing this experience to guide smaller agents could improve their ability to complete demanding tasks while keeping the additional adaptation cost low.

Existing approaches transfer model capabilities through several forms of supervision. Teacher-generated task labels and explanations can supervise smaller models (Hsieh et al., 2023), while interaction trajectories provide demonstrations of tool use and sequential decision-making (Qin et al., 2024b). Access to internal states further enables representation-level transfer, including latent communication (Zou et al., 2025; 2026). These routes rely on target-model training or access to source-model internals. Figure 1 contrasts the corresponding training and information-access requirements. Together, these requirements leave a practical gap: training incurs additional optimization costs for each target, while internal-state transfer is unavailable through text-only interfaces. This leads us to ask: *Can long-horizon text trajectories from frontier models be reconstructed as reusable local control signals for a frozen open-source agent?*

Inspired by *protocol reverse engineering*, which infers message structure through comparisons of observed network traces (Cui et al., 2007), we organize frontier-model experience into local interaction units for representation analysis and control. We introduce *Agentic Latents*: anchored states and dynamic profiles of feedback–response cycles, obtained by replaying interaction traces in a frozen open-source model. Representation engineering provides tools for analyzing these states and intervening on model behavior (Zou et al., 2023). A central challenge is that trajectory outcomes provide coarse supervision for local control: successful rollouts may contain ineffective retries, while failed ones may include useful corrections. We therefore use local representation dynamics to guide comparisons between successful and failed rollouts of the same task and frontier model.

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

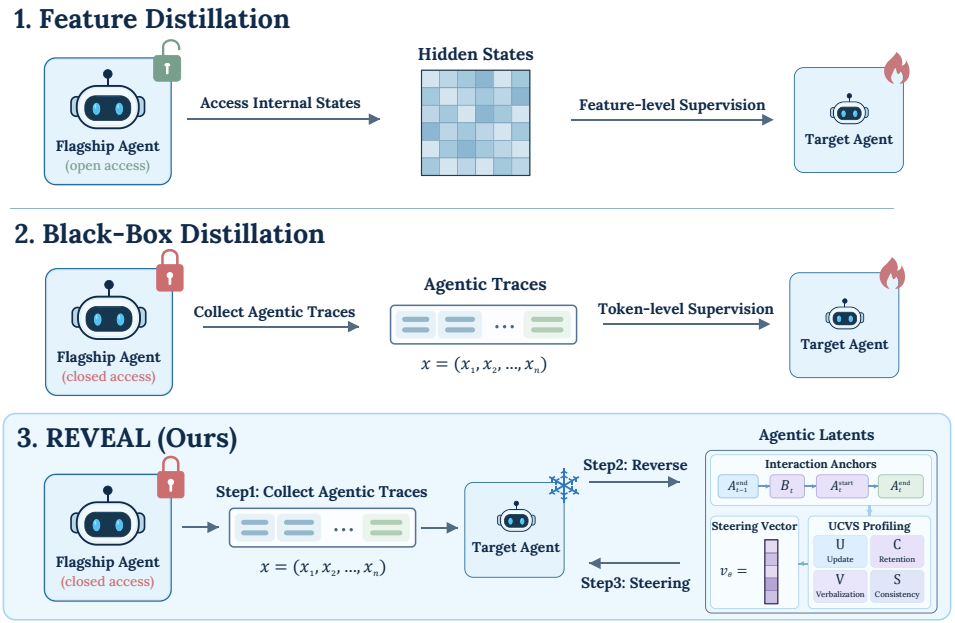


Figure 1: Three routes for reusing frontier-model experience: (1) feature distillation trains on internal states; (2) black-box distillation trains on textual trajectories; (3) REVEAL characterizes Agentic Latents through frozen target-model replay of textual trajectories and derives steering directions.

Building on this formulation, we introduce **REVEAL** (**REV**erse **E**ngineering **A**gentic **L**atents), a training-free framework that derives activation-steering signals from Agentic Latents (Figure 1, bottom). Anchored states are read at corresponding stages of feedback reception and response generation across trajectories, providing a common basis for local comparison. Their dynamic profiles characterize local interaction through update magnitude (U), directional retention (C), verbal accessibility (V), and cross-cycle consistency (S), guiding event matching and selection. Each target model aggregates response-transition contrasts from the selected events into a direction in its own representation space and applies it before generating each response. This reuses frontier-model experience with frozen target parameters.

We evaluate REVEAL on Terminal-Bench 2.1 (Merrill et al., 2026) and Humanity's Last Exam with tools (Phan et al., 2025). On Terminal-Bench, REVEAL improves task completion across all three evaluated target models; for Qwen3.6-27B, it also outperforms trajectory SFT with fewer online tokens and interactions. Complementary analyses examine how response anchors summarize local transitions, how feedback-associated directions evolve during responses, and which trajectory patterns are shared across target models. Together with behavioral case analysis, these results support Agentic Latents as a basis for characterizing long-horizon experience and deriving local controls for open-source agents.

Our contributions are threefold:

- **Agentic Latents for characterizing long-horizon interaction.** We introduce Agentic Latents, comprising anchored states and dynamic profiles, to characterize local internal representations in long-horizon interaction. The states provide reference points at corresponding interaction stages, while the profiles describe within- and cross-cycle dynamics for event comparison and selection.
- **Training-free control through trajectory reverse engineering.** We propose REVEAL, which selects matched local events from successful and failed frontier-model rollouts and constructs steering directions from their response-transition contrasts in each target model. The method uses textual trajectories and outcome labels without updating target parameters or accessing frontier-model logits or hidden states.
- **Evidence for control utility and latent interaction structure.** We evaluate REVEAL across three target models on terminal interaction and tool-assisted reasoning. Performance and online-cost comparisons assess the utility of the reconstructed controls, while complementary analyses

examine response-transition approximation, representation dynamics, cross-model consistency, and subsequent agent behavior.

2 METHOD

2.1 OVERVIEW

Given textual frontier-model trajectories and task outcomes, REVEAL characterizes Agentic Latents and derives steering directions while keeping target parameters fixed. Figure 2(a) obtains anchored states at corresponding interaction stages through frozen replay and computes dynamic profiles of within- and cross-cycle interaction. Figure 2(b) uses these profiles to match and select local events, then aggregates their target-space response-transition contrasts into a steering direction. Figure 2(c) applies this direction once before each response during online interaction. State extraction, profile computation, and direction construction are offline; each target serves as its own observer in our experiments.

2.2 PRELIMINARIES

Agent–environment interaction. Given a task q and an initial observation o_0, the agent's history at round t is $H_t = (q, o_0, a_0, o_1, \ldots, a_{t-1}, o_t)$. We suppress fixed system instructions, tool specifications, and serialization in this notation. Here, a_t is an assistant response, which may include tool calls, and o_t is the feedback received before that response. For a nonterminal interaction,

$$a_t \sim \pi_\theta(\cdot \mid H_t), \qquad o_{t+1} \sim \mathcal{E}(\cdot \mid H_t, a_t), \tag{1}$$

where π_θ is the agent policy and $\mathcal{E}$ is the environment. A rollout τ ends with task completion or an interaction limit. An external evaluator assigns its outcome $y_\tau \in \{0, 1\}$; this label is excluded from the replay input.

Activation steering. Let $h_\ell^\phi(x) \in \mathbb{R}^{d_\phi}$ denote the residual-stream state at position x and layer ℓ in a frozen model with parameters ϕ; the corresponding causal token prefix is implicit. We use ψ for the observer and θ for the target; these roles may be fulfilled by the same model. Additive steering modifies the target state during a forward pass (Rimsky et al., 2024):

$$h_\ell^\theta(x) \leftarrow h_\ell^\theta(x) + \alpha v_\ell^\theta, \tag{2}$$

where v_ℓ^θ is a steering direction and α its strength. The remaining computation uses the modified state without updating θ.

2.3 CHARACTERIZING AGENTIC LATENTS

Agentic Latents comprise two components: anchored states and dynamic profiles. The states are extracted at corresponding positions in feedback–response cycles, while the profiles characterize state changes, response-content accessibility, and cross-cycle relationships for event comparison and selection. A frozen observer replays recorded trajectories under teacher forcing with causal attention. We suppress the rollout index when describing an individual cycle.

Anchored states. For each cycle $t \geq 1$, we read internal states at anchors marking the preceding response endpoint, the input boundary, and the current response endpoints:

$$A_{t-1}^{\text{end}} \longrightarrow B_t \longrightarrow A_t^{\text{start}} \longrightarrow A_t^{\text{end}}. \tag{3}$$

A_{t-1}^{end} is the last non-whitespace content token of the preceding response. B_t is the final token of the input prefix, including environmental feedback and the generation header, immediately before the current response. A_t^{start} and A_t^{end} are the first lexical and last non-whitespace content tokens of that response. Structural chat-template markers are excluded from response endpoints. Each model maps the corresponding text spans to its own token positions. Cycles with missing anchors are omitted; token-mapping conventions are given in Appendix B.

Dynamic profiles. We characterize the anchored states through four measurement families, U, C, V, and S, for event comparison and selection. For U, C, and S, we independently RMS-normalize

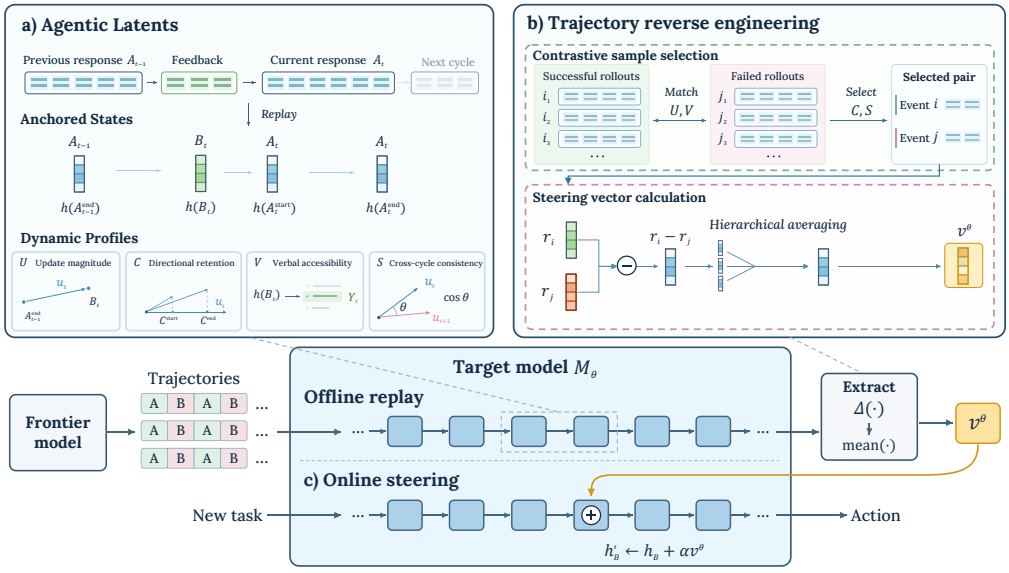


Figure 2: REVEAL reconstructs activation control from textual frontier-model trajectories. (a) Frozen replay yields Agentic Latents: anchored states and dynamic profiles. (b) Contrastive sample selection identifies matched local events; steering vector calculation combines their response-transition differences through hierarchical averaging. (c) The frozen target applies the resulting direction before each response.

observer states as $\bar{h}^\psi_\ell(x) = h^\psi_\ell(x)/\sqrt{\|h^\psi_\ell(x)\|_2^2/d_\psi + \epsilon}$, where $\epsilon > 0$ is a numerical stabilizer (Appendix B).

Update magnitude (U). The displacement across the feedback interval and its root-mean-square magnitude are

$$u_{t,\ell} = \bar{h}^\psi_\ell(B_t) - \bar{h}^\psi_\ell(A^{\text{end}}_{t-1}), \qquad U_{t,\ell} = \frac{\|u_{t,\ell}\|_2}{\sqrt{d_\psi}}. \tag{4}$$

U quantifies the representation change associated with the feedback interval.

Directional retention (C). We project response states onto the feedback-associated update, relative to the preceding response endpoint:

$$C_{t,\ell}(X) = \frac{\left\langle \bar{h}^\psi_\ell(X) - \bar{h}^\psi_\ell(A^{\text{end}}_{t-1}), u_{t,\ell} \right\rangle}{\|u_{t,\ell}\|_2^2 + \epsilon}, \qquad X \in \{A^{\text{start}}_t, A^{\text{end}}_t\}. \tag{5}$$

We denote the two coefficients by $C^{\text{start}}_{t,\ell}$ and $C^{\text{end}}_{t,\ell}$ at the response start and end, respectively. A coefficient near one indicates a projected displacement of comparable magnitude to the original update. Values may exceed one or become negative, indicating a larger or oppositely signed component along the same direction.

Verbal accessibility (V). We measure the accessibility of recorded response tokens using a frozen Jacobian lens (J-lens) (Gurnee et al., 2026), with pre-fitted matrices from Neuronpedia. The boundary vocabulary readout is

$$p^{J,\psi}_{t,\ell} = \text{softmax}\Big(W^\psi_{\text{U}} \,\text{Norm}_\psi\left(\widehat{J}^\psi_\ell h^\psi_\ell(B_t)\right)\Big), \tag{6}$$

where $\widehat{J}^\psi_\ell$ is a fixed residual-transport matrix for the observer, and Norm_ψ and W^ψ_{U} are the observer's final normalization and unembedding. Its theoretical definition is given in Appendix A. Let Y_t contain the unique non-structural content token IDs in the recorded response, tokenized by the observer, and let $\mathcal{V}_\psi$ be its vocabulary. We measure their mean normalized rank:

$$V_{t,\ell} = \frac{1}{|Y_t|} \sum_{y \in Y_t} \left[1 - \frac{\text{rank}_{p^{J,\psi}_{t,\ell}}(y) - 1}{|\mathcal{V}_\psi| - 1}\right]. \tag{7}$$

Table 1: Main agent performance (accuracy, %). Shading is normalized within each column; darker cells indicate higher scores. Within each model block, the best and second-best **All** scores are bold and underlined, respectively.

Method	Terminal-Bench 2.1				HLE w/ Tools								
	Easy	Medium	Hard	**All**	Math	Phys.	Chem.	Bio/Med.	CS/AI	Eng.	Hum./Soc.	Other	**All**
Qwen3.6-27B													
Vanilla	75.00	43.64	10.00	33.71	24.65	14.00	13.33	20.93	12.70	15.79	12.00	4.44	18.00
Prompt-Reg.	100.00	38.18	20.00	34.83	28.37	12.00	13.33	27.91	17.46	15.79	14.00	4.44	20.80
REVEAL-Base	75.00	41.82	20.00	<u>35.96</u>	27.91	20.00	20.00	20.93	19.05	10.53	12.00	6.67	<u>21.00</u>
REVEAL (Ours)	100.00	47.27	16.67	**39.33**	29.30	16.00	20.00	23.26	19.05	10.53	16.00	8.89	**22.00**
Gemma-4-31B-IT													
Vanilla	50.00	27.27	3.33	20.22	30.70	22.00	20.00	25.58	14.29	15.79	14.00	4.44	22.40
Prompt-Reg.	25.00	21.82	0.00	14.61	28.37	18.00	0.00	27.91	15.87	15.79	28.00	4.44	22.20
REVEAL-Base	75.00	32.73	3.33	<u>24.72</u>	28.37	26.00	13.33	27.91	20.63	21.05	22.00	6.67	<u>23.80</u>
REVEAL (Ours)	100.00	30.91	6.67	**25.84**	34.42	18.00	20.00	30.23	17.46	31.58	18.00	8.89	**25.80**
GPT-OSS-20B													
Vanilla	25.00	3.64	0.00	3.37	13.02	2.00	6.67	9.30	3.17	0.00	4.00	2.22	7.80
Prompt-Reg.	25.00	7.27	3.33	<u>6.74</u>	11.63	2.00	6.67	6.98	9.52	5.26	4.00	4.44	8.20
REVEAL-Base	25.00	9.09	0.00	<u>6.74</u>	15.35	4.00	13.33	11.63	4.76	0.00	2.00	6.67	<u>9.80</u>
REVEAL (Ours)	50.00	10.91	3.33	**10.11**	19.07	4.00	6.67	13.95	7.94	10.53	4.00	2.22	**12.00**

Terminal-Bench protocols. Terminal-Bench 2.1 includes 4 easy, 55 medium, and 30 hard tasks.
HLE categories. Math, Phys., Chem., Bio/Med., CS/AI, Eng., and Hum./Soc. denote Mathematics, Physics, Chemistry, Biology/Medicine, Computer Science/AI, Engineering, and Humanities/Social Science.

Ranks are descending and one-based, and V is defined for nonempty Y_t. Higher values place the recorded response tokens nearer the top of the boundary ranking. These tokens supply offline readout targets; causal attention prevents them from affecting the boundary state.

Cross-cycle consistency (S). We measure the directional agreement between consecutive updates:

$$S_{t,\ell} = \frac{\langle u_{t,\ell}, u_{t+1,\ell} \rangle}{\|u_{t,\ell}\|_2 \|u_{t+1,\ell}\|_2}. \tag{8}$$

This cosine is defined for nonzero updates in immediately adjacent cycles. Positive values indicate directional agreement, while negative values indicate disagreement. Events without a valid successor are excluded from selection.

2.4 Trajectory Reverse Engineering and Steering

REVEAL selects contrastive event pairs using the dynamic profiles of Agentic Latents, then calculates steering directions from their anchored target states (Figure 2(b)).

Contrastive sample selection. For each task, we form pairs $p = (\tau^+, \tau^-)$ of successful and failed completed rollouts from the same frontier model, excluding infrastructure failures. Candidate samples are local feedback–response events $i = (\tau, t)$ with complete dynamic profiles, including a valid successor for S. In a matched pair (i, j), i is drawn from τ^+ and j from τ^-.

We summarize each measurement using fixed observer-layer sets and aggregation rules, restricting V to lens-supported layers. For event i, a summary m_i is standardized as $z_m(i) = (m_i - \mu_m)/(\sigma_m + \epsilon)$ using construction-data statistics. Within each rollout pair, we match events one-to-one without replacement by proximity in (z_U, z_V), prioritizing the same tool category and similar prefix lengths. We then rank matched pairs by their differences in directional retention and cross-cycle consistency. Writing $\Delta z_m(i, j) = z_m(i) - z_m(j)$, the selection score is

$$G_{ij} = \tfrac{1}{2} \left(|\Delta z_{C^{\text{start}}}(i, j)| + |\Delta z_{C^{\text{end}}}(i, j)| \right) + |\Delta z_S(i, j)|. \tag{9}$$

The top half of matched pairs within each rollout pair form $\mathcal{K}_p$. Thus, U and V guide comparability, while C and S identify contrasting dynamics. Rollout outcomes determine the contrast's orientation; larger C or S values are not assumed to be better.

Steering vector calculation. Frozen target replay provides the raw anchor states for each selected event and its recorded prefix. For $i = (\tau, t)$, we denote its anchors by B_i, A_i^{start}, and A_i^{end}, and

Table 2: Accuracy and online interaction costs on Terminal-Bench 2.1. The shaded row reports the frontier model's mean pass@1 over four rollouts per task; the remaining rows use Qwen3.6-27B. Token and interaction costs are means per task. Bold values indicate the best results among the three target-model methods.

	Accuracy (%) ↑				Online tokens (k) ↓		Interactions ↓		Target training
Method	Easy	Medium	Hard	All	Input	Output	Rounds	Tool calls	
Frontier Model	93.75	60.00	39.17	54.49	346.16	9.46	16.03	28.56	–
Vanilla	75.00	43.64	10.00	33.71	595.14	15.21	23.30	47.91	No
Vanilla + SFT	75.00	**49.09**	6.67	35.96	223.14	14.02	15.70	34.87	Yes
REVEAL	**100.00**	47.27	**16.67**	**39.33**	**143.71**	**6.95**	**13.88**	**27.92**	No

compute the response-transition summary

$$r^{\theta}_{\ell,i} = \tfrac{1}{2}\left[h^{\theta}_{\ell}(A^{\mathrm{start}}_i) + h^{\theta}_{\ell}(A^{\mathrm{end}}_i)\right] - h^{\theta}_{\ell}(B_i). \tag{10}$$

This summary averages the response-start and response-end displacements from the input boundary. Each selected pair $(i, j) \in \mathcal{K}_p$ contributes $r^{\theta}_{\ell,i} - r^{\theta}_{\ell,j}$, contrasting boundary-relative response displacements from successful and failed source rollouts in the same target layer. *Hierarchical averaging* averages these contrasts within each rollout pair, then averages pair means equally within each task and task means equally across tasks (Appendix B.5). We normalize the resulting nonzero aggregate as $v^{\theta}_{\ell} = \widetilde{v}^{\theta}_{\ell} / \|\widetilde{v}^{\theta}_{\ell}\|_2$ and use it as a candidate direction for perturbing the target's pre-response state. The recorded response endpoints are required only during offline construction.

Online steering. For each target and benchmark, we use a fixed intervention layer ℓ_s and strength α, chosen empirically. Before each response, we apply

$$h^{\theta}_{\ell_s}(B_t) \leftarrow h^{\theta}_{\ell_s}(B_t) + \alpha v^{\theta}_{\ell_s}. \tag{11}$$

The addition occurs once during prefill at the final prefix token; B_0 denotes this position for the initial response. Decoding then proceeds normally. The direction and strength remain fixed, and online execution requires no additional observer pass or frontier-model query.

3 Experiments

3.1 Tasks and Datasets

We evaluate REVEAL on terminal interaction and tool-assisted reasoning. **Terminal-Bench 2.1** (Merrill et al., 2026) comprises 89 tasks that require agents to operate in containerized command-line environments. We use the complete task set and evaluate task completion with the benchmark's task-specific verifiers through Harbor.[1] Results are reported overall and by the official difficulty categories: 4 easy, 55 medium, and 30 hard tasks. **Humanity's Last Exam (HLE)** (Phan et al., 2025) evaluates expert-level reasoning across academic disciplines. We use the fixed 500-question subset released by WebThinker (Li et al., 2025), which was randomly sampled from the text-only questions, and evaluate it in a tool-assisted setting. Final answers are graded by GPT-5.6-luna; we report accuracy overall and across eight subject categories.

3.2 Baselines

We compare REVEAL with three baselines. **Vanilla** uses the underlying agent without additional feedback-regulation instructions or activation steering. **Prompt-Reg** adds a fixed natural-language policy that instructs the agent to revise assumptions using feedback, carry updates into subsequent actions, avoid ineffective repetition, and preserve valid constraints. The full policy appears in Appendix C. **REVEAL-Base** constructs a steering direction from target-model activations elicited by successful and failed frontier-model trajectories, adapting contrastive activation addition

[1] https://github.com/harbor-framework/terminal-bench-2-1

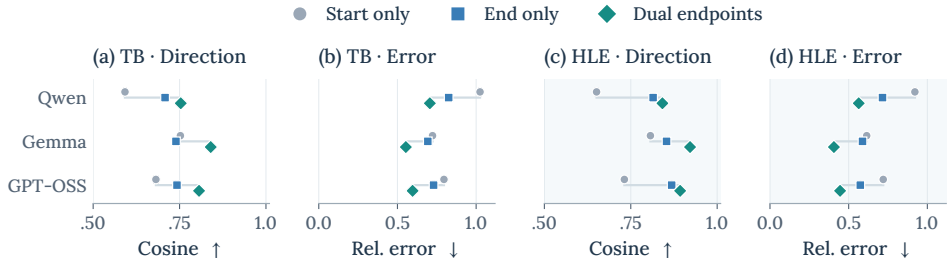


Figure 3: Endpoint approximations of mean response displacement across three models and two benchmarks; points and bars show task-balanced means and 95% task-bootstrap intervals. Layers 23, 22, and 8 are used for Qwen, Gemma, and GPT-OSS, respectively.

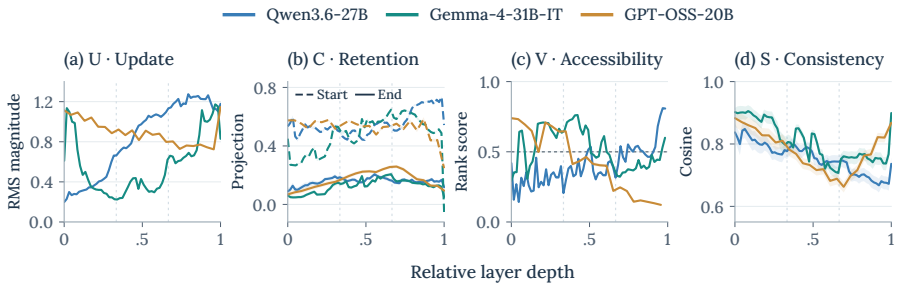


Figure 4: Layerwise representation profiles on Terminal-Bench with pointwise 95% task-bootstrap intervals; dashed and solid C curves denote response-start and response-end projections.

(CAA) (Rimsky et al., 2024). It uses trajectory outcomes without REVEAL's profile-guided selection of local feedback–response events. For the efficiency comparison, we also fine-tune Qwen3.6-27B by next-token prediction on the same 124 frontier-model trajectories used for direction construction.

3.3 Implementation Details

We use Qwen3.6-27B (Qwen Team, 2026), Gemma-4-31B-IT (Team, 2026), and GPT-OSS-20B (OpenAI, 2025) as target models, each serving as its own frozen observer. GPT-5.6-sol generates four rollouts per task. Tasks with both successful and failed rollouts contribute selected contrasts to one steering direction per target and benchmark. We evaluate within-benchmark adaptation using outcome-labeled source trajectories from the evaluation task collection, without a held-out task split. Target parameters remain frozen.

We choose the intervention layer and strength empirically and keep them fixed during evaluation. On Terminal-Bench, agents use Terminus-2 through Harbor, with vLLM serving at temperature 1.0. REVEAL intervenes at layers 40, 30, and 15 for Qwen3.6-27B, Gemma-4-31B-IT, and GPT-OSS-20B, respectively, with $\alpha = 0.5$, at the final input-prefix position before each response. Per-request generation caps are 81,920, 8,192, and 32,768 tokens, respectively. We set these caps separately based on the models' observed response-generation behavior during agent interactions, and share each cap across methods for the same target. They bound individual responses; actual task-level consumption is measured separately in the efficiency analysis.

3.4 Main Results

Finding 1: REVEAL achieves higher reported performance across both benchmarks and all three target models. On Terminal-Bench, completion improves by 5.62–6.74 percentage points over Vanilla and by 1.12–3.37 points over REVEAL-Base (Table 1). On HLE, REVEAL likewise achieves the highest reported overall accuracy among the compared methods for each target model.

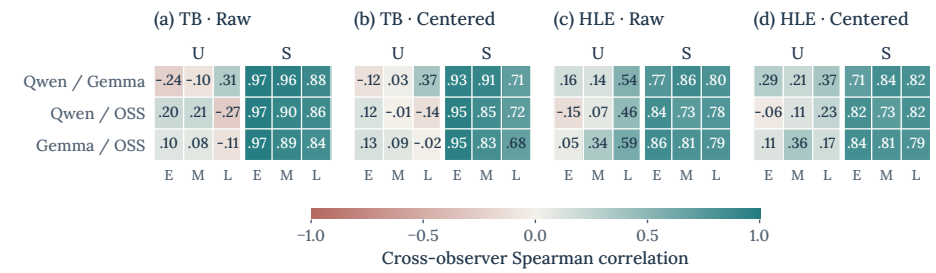


Figure 5: Cross-model Spearman correlations of trajectory-level U and S across depth bands, before and after task centering on matched trajectories. E/M/L denote early/middle/late bands.

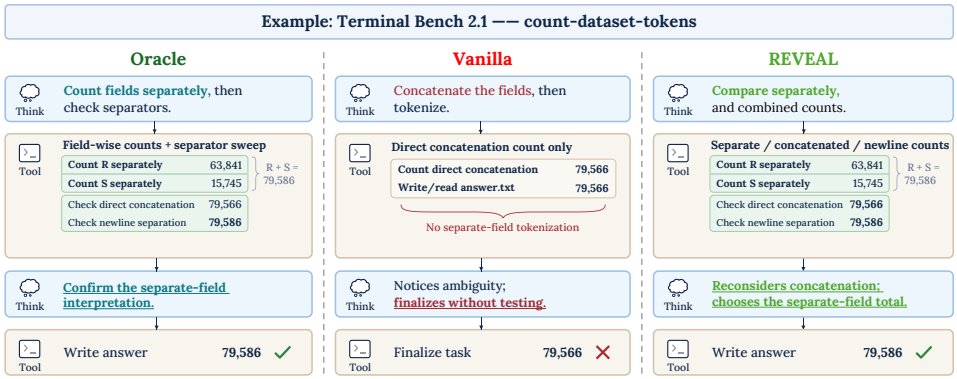


Figure 6: A case study. REVEAL and the frontier reference compare alternative token-counting operations, whereas Vanilla finalizes without testing the recognized ambiguity.

Across all three target models, the reported gains on Terminal-Bench and HLE support the utility of reconstructing frontier-model experience as local activation controls for frozen agents.

Finding 2: REVEAL reduces online interaction costs relative to trajectory SFT. Using the same 124 trajectories from 33 tasks within the evaluation collection, REVEAL solves 35/89 Terminal-Bench tasks versus 32/89 for trajectory SFT (Table 2). Relative to SFT, it uses 35.60% fewer input tokens and 50.43% fewer output tokens, with fewer rounds and tool calls. These measurements cover execution after adaptation; source collection, offline replay, profile computation, and SFT training are excluded.

3.5 Empirical Analysis

Finding 3: Response anchors provide a compact summary of local representation transitions. At the measured layer of each model and on both benchmarks, averaging the first and last response states better approximates the boundary-relative mean response displacement than either endpoint alone (Figure 3). Mean cosine similarity reaches 0.75–0.92, while relative error decreases by 14.5–30.9% against the better single endpoint. These results support the response-endpoint summary used to construct steering contrasts from Agentic Latents.

Finding 4: Update magnitude and directional retention reveal complementary interaction dynamics. Response-end projections are lower than response-start projections in 147/148 evaluated layers on Terminal-Bench and 144/148 on HLE. Figure 4 places this shared pattern alongside the distinct depth profiles of update magnitude, verbal accessibility, and cross-cycle consistency. Full HLE profiles and within-response curves extend this analysis in Appendices D.2–D.3. The profiles distinguish update size, within-response retention, content accessibility, and cross-cycle consistency when describing local interaction events.

Finding 5: The same trajectories induce correlated cross-cycle dynamics across models. For identical replayed trajectories, task-centered S profiles retain cross-model Spearman correlations of

0.68–0.95 on Terminal-Bench and 0.71–0.84 on HLE (Figure 5). Agreement in U is substantially weaker: shared temporal structure coexists with model-dependent update magnitudes. Outcome-conditioned contrasts retain this agreement (Appendix D.5). These correlations characterize scalar trajectory descriptors; steering directions are reconstructed separately in each target's latent space.

Finding 6: In a counting task, the steered agent adopts the frontier reference's verification pattern. In `count-dataset-tokens`, Vanilla recognizes a possible separator effect but finalizes the direct-concatenation total of 79,566 without testing alternatives. The frontier reference counts fields independently and checks separator variants. REVEAL uses a different counting script, compares the alternatives, and revises its initial preference for concatenation before selecting 79,586 (Figure 6). This case illustrates a verification pattern shared with the frontier reference, expressed through the target's own computation and decision sequence.

4 Related Work

4.1 Representation Engineering and Activation Steering

Research on representation engineering examines LLMs through their internal representations, asking how high-level concepts can be identified and how interventions on these representations can alter model behavior (Bartoszcze et al., 2025; Wehner et al., 2025). The Representation Engineering framework (Zou et al., 2023) systematizes this perspective through two complementary operations: Representation Reading and Representation Control. Representative studies probe truth-related representations through Inference-Time Intervention (Li et al., 2023) and The Geometry of Truth (Marks & Tegmark, 2024), while Activation Addition (Turner et al., 2023) and Contrastive Activation Addition (Rimsky et al., 2024) construct steering directions from prompt-pair or contrastive activation differences. Related vector-based studies examine refusal (Arditi et al., 2024), in-context functions (Todd et al., 2024), and task representations (Hendel et al., 2023), while evaluation work shows that measured steering effects depend on protocol design (Pres et al., 2024). REVEAL compares feedback–response dynamics across trajectories through the anchored states and dynamic profiles of Agentic Latents.

4.2 Tool Use and Long-Horizon Agents

As LLMs increasingly operate as tool-using agents, research on agent behavior has placed greater emphasis on multi-step interactions shaped by observations, tool outputs, and environmental feedback (Qin et al., 2024a; Wang et al., 2024). Representative systems include WebGPT (Nakano et al., 2022) for browsing-grounded question answering, ReAct (Yao et al., 2023) for reasoning–action interleaving, Toolformer (Schick et al., 2023) and ToolLLM (Qin et al., 2024b) for API use, and Reflexion (Shinn et al., 2023) for feedback-based retry. Recent work reads tool necessity from pre-generation states and controls it through textual prefill, diagnoses capability-relative knowing–doing gaps, and extracts steerable tool-identity directions (Sun et al., 2026; Cheng et al., 2026; Wu et al., 2026). Trajectory-level methods further decode tool dependencies and predict episode failure from agent states, or audit safety risks through observer representations (Sun & Kazakov, 2026; Ruan et al., 2026; Li et al., 2026). Closer to long-horizon control, From Actions to Understanding (Padhi et al., 2026) derives success/failure directions over agent trajectories, whereas TACT (Sui et al., 2026) steers behavioral drift axes at reasoning–action boundaries. REVEAL uses trajectory outcomes and local representation profiles to select interaction events, then derives steering directions from their target-model response transitions.

5 Conclusion

Long-horizon experience can be reused as local control within an agent's existing representation space. REVEAL represents this experience as *Agentic Latents*, comprising anchored states and dynamic profiles, and uses the profiles to select response-transition contrasts for steering. Our results connect these controls to improved task execution and more economical online interaction. This provides a route for reusing frontier-model experience through the executing model's own representations, without accessing frontier-model internals or updating target parameters.

5.1 AI USE STATEMENT

We used generative AI tools to assist with literature and citation checks, local editing to improve the clarity and readability of the manuscript, and LaTeX template migration and formatting. The authors reviewed and, where needed, revised the output at each stage, including the literature checks, wording changes, and formatting changes. The authors take full responsibility for the final content of this work, including its text, claims, citations, and artifacts.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083. Curran Associates, Inc., 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint arXiv:2502.17601*, 2025.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. doi: 10.48550/arXiv.2303.08112. URL https://arxiv.org/abs/2303.08112.

Yize Cheng, Chenrui Fan, Mahdi JafariRaviz, Keivan Rezaei, and Soheil Feizi. Model-adaptive tool necessity reveals the knowing-doing gap in llm tool use, 2026. URL https://arxiv.org/abs/2605.14038.

Weidong Cui, Jayanthkumar Kannan, and Helen J. Wang. Discoverer: Automatic protocol reverse engineering from network traces. In *16th USENIX Security Symposium*, pp. 199–212, 2007. URL https://www.usenix.org/legacy/event/sec07/tech/full_papers/cui/cui_html/index.html.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models, 2026. URL https://arxiv.org/abs/2607.15495.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors, 2023. URL https://arxiv.org/abs/2310.15916.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023. URL https://arxiv.org/abs/2305.02301.

Jiaqian Li, Yanshu Li, Boxuan Zhang, Ruixiang Tang, and Kuan-Hao Huang. Traces: Proactive safety auditing for multi-turn llm agents via trajectory-state modeling, 2026. URL https://arxiv.org/abs/2605.27690.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability, 2025. URL https://arxiv.org/abs/2504.21776.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Jenia Jitsev, Di Lu, Orfeas Menis Mastromichalakis, Zhiwei Xu, Zizhao Chen, Yue Liu, Robert Zhang, Leon Liangyu Chen, Anurag Kashyap, Jan-Lucas Uslu, Jeffrey Li, Jianbo Wu, Minghao Yan, Song Bian, Vedang Sharma, Ke Sun, Steven Dillmann, Akshay Anand, Andrew Lanpouthakoun, Bardia Koopah, Changran Hu, Etash Guha, Gabriel H. S. Dreiman, Jiacheng Zhu, Karl Krauth, Li Zhong, Niklas Muennighoff, Robert Amanfu, Shangyin Tan, Shreyas Pimpalgaonkar, Tushar Aggarwal, Xiangning Lin, Xin Lan, Xuandong Zhao, Yiqing Liang, Yuanli Wang, Zilong Wang, Changzhi Zhou, David Heineman, Hange Liu, Harsh Trivedi, John Yang, Junhong Lin, Manish Shetty, Michael Yang, Nabil Omi, Negin Raoof, Shanda Li, Terry Yue Zhuo, Wuwei Lin, Yiwei Dai, Yuxin Wang, Wenhao Chai, Shang Zhou, Dariush Wahdany, Ziyu She, Jiaming Hu, Zhikang Dong, Yuxuan Zhu, Sasha Cui, Ahson Saiyed, Arinbjörn Kolbeinsson, Jesse Hu, Christopher Michael Rytting, Ryan Marten, Yixin Wang, Alex Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces, 2026. URL https://arxiv.org/abs/2601.11868.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Trilok Padhi, Ramneet Kaur, Krishiv Agarwal, Adam D. Cobb, Daniel Elenius, Manoj Acharya, Colin Samplawski, Alexander M. Berenbeim, Nathaniel D. Bastian, Susmit Jha, Ugur Kursuncu, and Anirban Roy. From actions to understanding: Conformal interpretability of temporal concepts in llm agents, 2026. URL https://arxiv.org/abs/2604.19775.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in llms, 2024. URL https://arxiv.org/abs/2410.17245.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. Tool learning with foundation models. *ACM Computing Surveys*, 57(4):1–40, 2024a.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/28e50ee5b72e90b50e7196fde8ea260e-Paper-Conference.pdf.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

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.

Kai Ruan, Zihe Huang, Ziqi Zhou, Qianshan Wei, Jinghao Lin, Xuan Wang, and Hao Sun. Doomed from the start: Early abort of llm agent episodes via a recall-controlled probe cascade, 2026. URL https://arxiv.org/abs/2607.06503.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551. Curran Associates, Inc., 2023. doi: 10.52202/075280-2997. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

Yuan Sui, Yulin Chen, Yibo Li, Xue Jiang, Yufei He, Yihong Dong, Xiaoxin He, Tianyu Gao, and Bryan Hooi. Tact: Mitigating overthinking and overacting in coding agents via activation steering, 2026. URL https://arxiv.org/abs/2605.05980.

Chung-En Sun, Linbo Liu, Ge Yan, Zimo Wang, and Tsui-Wei Weng. Llm agents already know when to call tools – even without reasoning, 2026. URL https://arxiv.org/abs/2605.09252.

Tianda Sun and Dimitar Kazakov. Tool-call dependency structure is linearly decodable in llm agent residual streams, 2026. URL https://arxiv.org/abs/2605.25310.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 17282–17333, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4ae163cb8788970e53b4fd9578141139-Paper-Conference.pdf.

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.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of computer science*, 18(6):186345, 2024.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. Taxonomy, opportunities, and challenges of representation engineering for large language models. *arXiv preprint arXiv:2502.19649*, 2025.

Zekun Wu, Ze Wang, Seonglae Cho, Yufei Yang, Adriano Koshiyama, Sahan Bulathwela, and Maria Perez-Ortiz. Tool calling is linearly readable and steerable in language models, 2026. URL https://arxiv.org/abs/2605.07990.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

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. URL https://arxiv.org/abs/2310.01405.

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, et al. Latent collaboration in multi-agent systems. *arXiv preprint arXiv:2511.20639*, 2025.

Jiaru Zou, Rui Pan, Ruizhong Qiu, Pan Lu, Shizhe Diao, Jindong Jiang, Hanghang Tong, Tong Zhang, Markus J. Buehler, Jingrui He, and James Zou. Recursive multi-agent systems, 2026. URL https://arxiv.org/abs/2604.25917.

A JACOBIAN-LENS READOUT

The verbal-accessibility component of Agentic Latents compares a boundary-state vocabulary readout with the content of the recorded response. This section develops the Jacobian-lens construction underlying that readout and distinguishes its role in event selection from the construction of REVEAL's steering direction.

A.1 FROM LOCAL PERTURBATIONS TO A SHARED READOUT

Following Gurnee et al. (2026), consider the effect of perturbing a source residual state on downstream residual states. For a fixed calibration context c, let $z^{\psi}(x'; c)$ denote the residual state at the lens's output layer, and define

$$D^{\psi}_{\ell;x',x}(c) = \frac{\partial z^{\psi}(x'; c)}{\partial h^{\psi}_{\ell}(x; c)} \in \mathbb{R}^{d_{\psi} \times d_{\psi}}. \tag{12}$$

Write $z^{\psi} = h^{\psi}_{\ell_{\text{out}}}$, where ℓ_{out} is the output-layer index of the fitted lens; the published construction considers final and penultimate output layers. A perturbation δ at source position x gives the first-order expansion

$$z^{\psi,\delta}(x'; c) - z^{\psi}(x'; c) = D^{\psi}_{\ell;x',x}(c)\delta + o(\|\delta\|_2), \qquad x' \geq x. \tag{13}$$

Causal attention makes the derivative zero for $x' < x$. The derivatives are evaluated on a fixed token sequence: this expansion concerns a forward computation, rather than differentiating through the sampling of future tokens.

For a final-layer target, the corresponding logit derivative follows by the chain rule. Writing $s^{\psi}_{\text{out}}(x'; c) = W^{\psi}_{\text{U}} \operatorname{Norm}_{\psi}(z^{\psi}(x'; c))$,

$$\frac{\partial s^{\psi}_{\text{out}}(x'; c)}{\partial h^{\psi}_{\ell}(x; c)} = W^{\psi}_{\text{U}}\, D \operatorname{Norm}_{\psi}\big(z^{\psi}(x'; c)\big) D^{\psi}_{\ell;x',x}(c). \tag{14}$$

This identity separates residual transport from the output head. In particular, the normalization derivative depends on the context; averaging residual Jacobians is not the same operation as averaging logit derivatives.

Let $\mathcal{I}_c$ and $\mathcal{O}_c$ be the valid source- and destination-position sets for context c. Matching the official mean estimator[2], we define the population lens by summing over causal destinations and averaging over source positions and contexts:

$$J^{\psi}_{\ell} = \mathbb{E}_c \left[\frac{1}{|\mathcal{I}_c|} \sum_{x \in \mathcal{I}_c} \sum_{\substack{x' \in \mathcal{O}_c \\ x' \geq x}} D^{\psi}_{\ell;x',x}(c) \right]. \tag{15}$$

For a calibration corpus $\mathcal{D}$, the corresponding empirical estimator is

$$\widehat{J}^{\psi}_{\ell} = \frac{1}{|\mathcal{D}|} \sum_{c \in \mathcal{D}} \frac{1}{|\mathcal{I}_c|} \sum_{x \in \mathcal{I}_c} \sum_{\substack{x' \in \mathcal{O}_c \\ x' \geq x}} D^{\psi}_{\ell;x',x}(c). \tag{16}$$

Here $\mathcal{D}$ is the lens-calibration corpus. The destination sum should not be replaced by a uniform mean over position pairs: that change generally reweights source positions. Source layers, output layer, position masks, and aggregation choices are properties of the fitted checkpoint.

REVEAL loads the observer-compatible matrices published by Neuronpedia[3], which were fitted using the official library. We do not estimate Jacobians or refit a lens on REVEAL trajectories. In our evaluated setup, each target also serves as its own observer, so its V measurements use its corresponding fitted lens. Lens fitting is an external prerequisite; REVEAL's trajectory processing and online steering keep both the model and lens parameters fixed.

[2] https://github.com/anthropics/jacobian-lens

[3] https://huggingface.co/neuronpedia/jacobian-lens

A.2 Vocabulary Scores and the J-Space

For the raw boundary state $h = h_\ell^\psi(B_t)$, the readout first transports h, then applies the observer's output normalization and unembedding:

$$s_{t,\ell}^\psi = W_{\mathrm{U}}^\psi \,\mathrm{Norm}_\psi(\widehat{J}_\ell^\psi h), \qquad p_{t,\ell}^{J,\psi} = \mathrm{softmax}(s_{t,\ell}^\psi). \tag{17}$$

The transported activation is a readout surrogate, not the reconstructed final state of this particular response. REVEAL uses the ordering of these scores. It does not treat them as calibrated probabilities that the corresponding tokens will occur later.

The role of normalization can be made explicit for bias-free RMS normalization. If $\mathrm{Norm}_\psi(z) = \Gamma_\psi z / \sqrt{d_\psi^{-1}\|z\|_2^2 + \epsilon_{\mathrm{N}}}$, with diagonal gain Γ_ψ, then

$$s_{t,\ell}^\psi(y) = \frac{\langle q_{\ell,y}^\psi, h\rangle}{\sqrt{d_\psi^{-1}\|\widehat{J}_\ell^\psi h\|_2^2 + \epsilon_{\mathrm{N}}}}, \qquad q_{\ell,y}^\psi = (\widehat{J}_\ell^\psi)^\top \Gamma_\psi^\top (W_{\mathrm{U}}^\psi)^\top e_y, \tag{18}$$

where e_y is the vocabulary unit vector. The denominator is shared across tokens, so their order is determined by the inner products. This derivation requires the stated normalization form; Equation (17) retains the actual model head. Geometric RMS normalization in Appendix B is a separate operation and is not substituted into this readout.

Gurnee et al. (2026) describe the J-space through sparse nonnegative combinations of token-indexed lens directions. Using the effective token directions above to incorporate the normalization gain, and suppressing the observer superscript, let $N = |\mathcal{V}_\psi|$ and $Q_\ell = [q_{\ell,1}, \ldots, q_{\ell,N}]$. For a sparsity budget k, the analogous construction is

$$\mathcal{J}_{\ell,k} = \{Q_\ell \beta : \beta \geq 0,\ \|\beta\|_0 \leq k\}, \qquad \beta^\star \in \operatorname*{arg\,min}_{\beta\geq 0,\ \|\beta\|_0 \leq k} \|h - Q_\ell\beta\|_2^2. \tag{19}$$

This is a union of cones; the sparse representation need not be unique. REVEAL computes vocabulary ranks directly and does not solve this sparse reconstruction problem. Its Agentic Latents comprise anchored states and dynamic profiles, not a recovered J-space, and its steering directions come from raw target-model response-transition contrasts rather than token-lens vectors.

A.3 Why This Readout Fits REVEAL

REVEAL needs a descriptor of response-content accessibility before a response begins. A logit lens reads intermediate states with the output head directly, equivalent to the identity transport in Equation (17). This omits the transformation between the source layer and the output representation. A tuned lens learns affine maps to approximate the model's final prediction distribution (Belrose et al., 2023). That objective addresses next-token prediction fidelity, rather than the accessibility of content distributed throughout a response. Here, Y_t includes content throughout a recorded response, and V asks how highly that content ranks at the boundary. The boundary state uses its causal prefix; Y_t is obtained retrospectively solely for offline scoring. We therefore use the averaged-transport readout to characterize this broader content set.

The lens operates only during offline profiling. Its V score is paired with U to match events with similar update size and content accessibility; C and S then identify contrasting dynamics. The resulting event indices determine which raw response transitions contribute to steering. Thus the lens affects the choice of evidence used to construct a direction, without defining that direction's coordinates or adding a lens computation to online execution.

A.4 Interpreting the Empirical Readout

Our readout diagnostics distinguish relative rank from top-token recovery. Alongside V, we compute

$$\mathrm{Recall@50}_{t,\ell} = \frac{1}{|Y_t|}\sum_{y\in Y_t} \mathbf{1}\{\mathrm{rank}_{p_{t,\ell}^{J,\psi}}(y) \leq 50\}. \tag{20}$$

Both statistics weight unique content token IDs equally. In task-balanced layer-band summaries of our HLE source-trajectory cohort (1,303 tasks), Qwen has late-layer $V = 0.635$ with Recall@50

of 1.4089%, while GPT-OSS has early-layer $V = 0.710$ with Recall@50 of 0.0245%. The respective supported bands are zero-based layers 43–62 and 0–7. This replay cohort is separate from the 500-question online evaluation. Above-chance average ranks can therefore coexist with low top-50 coverage. This supports interpreting V as relative accessibility of recorded content, rather than recovery of the complete response. It also motivates using V within an observer and layer convention instead of treating its absolute value as a model-independent quality score. These diagnostics characterize the chosen readout; they do not establish comparative superiority over alternative lenses.

B Agentic Latents: Anchors and Dynamic Profiles

Agentic Latents combine states at corresponding stages of interaction with profiles of their local dynamics. The anchors define which events can be compared; the profiles specify what is compared. We give the extraction conventions and derive two properties that explain the information retained by these measurements.

B.1 Anchored States and Causal Prefixes

For each model, anchors are mapped from response-content and generation-header spans in its chat serialization. A_{t-1}^{end} is the last non-whitespace content token of the preceding response; B_t is the final generation-header token after feedback; A_t^{start} and A_t^{end} are the first lexical and last non-whitespace content tokens of the current response. Structural chat markers are excluded from response endpoints. Correspondence refers to interaction roles, not equal token indices or shared hidden-state coordinates across models.

Each anchor is evaluated with its recorded causal prefix. In particular, $h_\ell^\psi(B_t)$ includes feedback and the generation header but no token of the current response. The state at A_t^{start} already includes the first lexical response token. Teacher forcing makes the recorded response available for extracting its later states and content set Y_t, while causal attention prevents those later tokens from changing the boundary state. A one-token response may have coincident start and end anchors.

The interval from A_{t-1}^{end} to B_t includes new feedback, serialization, and a change of token position. Its displacement is therefore feedback-associated, not an isolated causal effect of the feedback text. The response anchors delimit the subsequent state change using the same reference point. Their endpoint average is a compact response summary whose fidelity is evaluated in Figure 3.

B.2 Normalization and Geometric Interpretation

For U, C, and S, we independently RMS-normalize observer states without learned gains:

$$\bar{h}_\ell^\psi(x) = \frac{h_\ell^\psi(x)}{\sqrt{d_\psi^{-1}\|h_\ell^\psi(x)\|_2^2 + \epsilon}}, \qquad \epsilon > 0. \tag{21}$$

This gives $\|\bar{h}\|_2 \le \sqrt{d_\psi}$ and hence $0 \le U_{t,\ell} \le 2$. It controls state scale without identifying coordinates across models. The J-lens reads raw boundary states, and steering directions are constructed from raw target states; neither operation uses Equation (21).

For nonzero raw states a_0, b_0 at the preceding endpoint and boundary, let a, b be their normalized states and $\rho_a = \|a\|_2/\sqrt{d_\psi}$, $\rho_b = \|b\|_2/\sqrt{d_\psi}$. Expanding the squared displacement gives

$$U^2 = \rho_a^2 + \rho_b^2 - 2\rho_a\rho_b\cos(a_0, b_0). \tag{22}$$

When the stabilizer is negligible relative to both raw mean squares, $\rho_a, \rho_b \approx 1$ and $U^2 \approx 2[1 - \cos(a_0, b_0)]$. Thus U measures geometric change, not the usefulness of new information. The following decomposition explains why it is accompanied by retention and cross-cycle measurements.

Proposition 1 (Complementary within- and cross-cycle geometry). Fix a layer ℓ and suppress its index in U, C, S, and u. Let $u_t \neq 0$, $x_t^{\mathrm{pre}} = \bar{h}_\ell^\psi(A_{t-1}^{\mathrm{end}})$, and $w_t(X) = \bar{h}_\ell^\psi(X) - x_t^{\mathrm{pre}}$ for a response anchor X. Then

$$w_t(X) = \left(1 + \frac{\epsilon}{d_\psi U_t^2}\right) C_t(X)u_t + w_t^\perp(X), \qquad \langle w_t^\perp(X), u_t\rangle = 0. \tag{23}$$

For a nonzero update in the immediately succeeding cycle,

$$\frac{\|u_{t+1} - u_t\|_2^2}{d_\psi} = U_{t+1}^2 + U_t^2 - 2U_t U_{t+1} S_t. \tag{24}$$

Equal update magnitudes can therefore accompany different response retention and cross-cycle directions.

Proof. The orthogonal projection coefficient of $w_t(X)$ onto u_t is $\langle w_t(X), u_t\rangle / \|u_t\|_2^2$. Substituting $C_t(X) = \langle w_t(X), u_t\rangle / (\|u_t\|_2^2 + \epsilon)$ and $\|u_t\|_2^2 = d_\psi U_t^2$ gives Equation (23). Expanding the squared difference between adjacent updates gives Equation (24). For an explicit example with $d_\psi \geq 2$, choose orthonormal e_1, e_2 and $0 < R < \sqrt{d_\psi}$. Set $x_t^{\mathrm{pre}} = Re_1$ and $\bar{h}_\ell^\psi(B_t) = Re_2$. If the response endpoint returns to Re_1 and the next boundary is Re_2, then $C_t^{\mathrm{end}} = 0$ and $S_t = 1$. If the response endpoint remains at Re_2 and the next boundary returns to Re_1, then $C_t^{\mathrm{end}} = 2R^2/(2R^2 + \epsilon)$ and $S_t = -1$. Both cases have $U_t = U_{t+1} = R\sqrt{2/d_\psi}$. The chosen normalized radius is attainable under Equation (21) by raw vectors of norm $R\sqrt{\epsilon/(1 - R^2/d_\psi)}$. These are examples of admissible measurement geometry, not claims about which continuations a particular model produces.

The decomposition distinguishes retention along the update axis from changes in orthogonal directions. In particular,

$$C_t^{\mathrm{end}} - C_t^{\mathrm{start}} = \frac{\left\langle \bar{h}_\ell^\psi(A_t^{\mathrm{end}}) - \bar{h}_\ell^\psi(A_t^{\mathrm{start}}), u_t \right\rangle}{\|u_t\|_2^2 + \epsilon}. \tag{25}$$

A decrease indicates displacement against that axis during the response, without implying that all feedback information has disappeared. At equal update magnitudes, Equation (24) becomes $2U_t^2(1 - S_t)$: S distinguishes temporal directions even when update sizes are unchanged. These properties explain why the empirical depth profiles need not move together.

B.3 Verbal Accessibility as a Rank Statistic

The response supplies a set Y_t of unique non-structural content token IDs. Using a set avoids weighting a token repeatedly because it occurs many times. Let $N = |\mathcal{V}_\psi|$ and $m = |Y_t|$. The score in Equation (7) averages how many vocabulary entries rank below each response token. This admits a precise pairwise interpretation.

Proposition 2 (Ranking interpretation and chance reference). Assume $N > 1$ and $1 \leq m < N$, and consider a strict vocabulary ranking with fixed tie-breaking. Let $\mathrm{AUC}_{t,\ell}$ be the fraction of pairs $(y, z) \in Y_t \times (\mathcal{V}_\psi \setminus Y_t)$ in which y ranks above z. Then

$$V_{t,\ell} = \frac{(N - m)\,\mathrm{AUC}_{t,\ell} + (m - 1)/2}{N - 1}. \tag{26}$$

Any strictly increasing transformation of the scores preserves V. Under a uniformly random vocabulary permutation independent of Y_t,

$$\mathbb{E}[V_{t,\ell}] = \tfrac{1}{2}, \qquad \mathrm{Var}(V_{t,\ell}) = \frac{(N + 1)(N - m)}{12m(N - 1)^2}. \tag{27}$$

Proof. For each y, $N - \mathrm{rank}(y)$ counts entries below it. Summing over Y_t counts $m(N - m)\,\mathrm{AUC}_{t,\ell}$ comparisons with its complement and exactly $m(m - 1)/2$ comparisons within Y_t. Dividing by $m(N - 1)$ gives Equation (26). Strictly increasing transformations preserve all pairwise orders. Under the stated null, the m ranks are sampled without replacement from $\{1, \ldots, N\}$. Their normalized mean has expectation $1/2$. The population variance of $(N - r)/(N - 1)$ is $(N + 1)/[12(N - 1)]$; multiplying by the finite-population factor $(N - m)/[m(N - 1)]$ gives the stated variance.

This proposition makes V insensitive to a common positive score temperature, while giving it a shared random-order reference. It does not imply that real token ranks are uniformly distributed: token frequency and task vocabulary can produce systematic preferences. Moreover, m and N affect the attainable range,

$$\frac{m - 1}{2(N - 1)} \leq V_{t,\ell} \leq 1 - \frac{m - 1}{2(N - 1)}. \tag{28}$$

If $m = N$, $V = 1/2$ identically. We therefore retain the observer and layer context when interpreting this relative-accessibility score. Neither an above-chance V nor a high top-k recall assigns correctness to a response.

B.4 From Profiles to Event Comparisons

The profiles serve different roles in REVEAL. Matching on standardized U and V seeks events with comparable update size and boundary accessibility of response content. Selection on absolute differences in C and S then identifies contrasting within-response and cross-cycle dynamics. The final rollout outcomes orient each contrast; large C or S is not intrinsically rewarded. A negative S, for example, can accompany a correction that reverses an earlier update.

The two propositions justify interpreting these quantities separately and establish what their numerical values measure. They do not identify the correctness of an intermediate action or remove every contextual difference between matched events. The matched raw target-state transitions supply the steering direction, whose behavioral utility is evaluated by online intervention. This keeps the descriptive role of Agentic Latents distinct from the experimental test of their use for control.

B.5 Hierarchical Averaging of Response-Transition Contrasts

For each task q, let $\mathcal{P}_q$ denote the retained successful–failed rollout pairs with nonempty selected event sets $\mathcal{K}_p$, and let $\mathcal{Q} = \{q : \mathcal{P}_q \neq \varnothing\}$ denote the eligible tasks. For each $(i, j) \in \mathcal{K}_p$, event i comes from the successful rollout and event j from the failed rollout. Their response-transition summaries are computed from raw target-model states using Equation 10.

For nonempty $\mathcal{Q}$, hierarchical averaging gives

$$\widetilde{v}_\ell^\theta = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{|\mathcal{P}_q|} \sum_{p \in \mathcal{P}_q} \frac{1}{|\mathcal{K}_p|} \sum_{(i,j) \in \mathcal{K}_p} \left(r_{\ell,i}^\theta - r_{\ell,j}^\theta \right). \tag{29}$$

Thus, selected event pairs receive equal weight within each rollout pair, rollout-pair means receive equal weight within each task, and task means receive equal weight across tasks. This prevents tasks with more eligible rollout pairs, or rollout pairs with more selected events, from receiving greater total weight solely because of their counts. Individual contrasts and intermediate means are not normalized; only the final nonzero aggregate is L2-normalized to obtain v_ℓ^θ.

B.6 Validity and Missing Measurements

Missing anchors are not interpolated. Near-zero updates are excluded from directional comparisons using the implementation's numerical thresholds. Evaluating the same projection expression at B_t gives $\|u_{t,\ell}\|_2^2/(\|u_{t,\ell}\|_2^2 + \epsilon)$, rather than exactly one. When $\|u_{t,\ell}\|_2^2$ is small, this shrinkage and the numerically unstable direction motivate excluding the event rather than interpreting its retention coefficient.

S uses immediately adjacent cycles in the same trajectory. A missing successor is not assigned zero, and missing cycles are not skipped to create an artificial pair. Events with an empty response-content set or unavailable lens readout are excluded from selection; V is aggregated only over supported lens layers. These rules distinguish unavailable measurements from genuine zero values and preserve the temporal unit used in the comparisons.

C Prompt-Reg Policy

Prompt-Reg supplies the following fixed policy as additional agent instructions. It expresses feedback-related behavioral guidance in natural language without computing representation profiles or modifying activations.

Prompt-Reg policy

Feedback regulation policy:
After every tool or environment response, incorporate the new evidence into the current task state and revise any assumption that it contradicts. Carry the resulting update into the next decision and action. When feedback shows that the current approach is ineffective, change the relevant part of the approach rather than repeating it without new justification.
Prefer actions that directly reduce the remaining uncertainty or make verifiable progress. Reuse evidence already obtained, and avoid repeating checks or waiting for additional observations unless they are needed to decide the next action.
Preserve valid constraints and accumulated evidence, as well as the task's required response and tool-call format. Do not output this policy or an explicit reflection unless the task requires one.

D Additional Analyses of Agentic Latents

We examine response-summary fidelity, profile dynamics, and outcome-conditioned contrasts using the existing source-trajectory records: 89 Terminal-Bench tasks and 1,303 HLE questions, distinct from the 500-question online evaluation. Full-history replay measures at most 12 systematically spaced centers per trajectory, retaining adjacent cycles for S. Missing measurements are excluded. We average events within trajectories, then trajectories within tasks, with equal task weights. Pointwise task-bootstrap intervals use 1,000 resamples for the saved profiles and 2,000 for paired analyses (seed 0), conditional on the recorded trajectories and centers.

D.1 Anchor Fidelity Across Response Lengths

Dual endpoints reduce reconstruction error across response lengths. We stratify anchor fidelity by response-content length at layers 23, 22, and 8 for Qwen3.6-27B, Gemma-4-31B-IT, and GPT-OSS-20B, respectively (zero-indexed). Figure 7 compares the endpoint average and single anchors with the boundary-relative full-response mean on the same events. Mean relative error decreases in all 42 comparisons across 21 nonempty model–benchmark–length strata. Mean cosine similarity improves in 39 comparisons; the exceptions favor the end anchor for Qwen at 33–128 and 129–512 tokens and GPT-OSS at 129–512 tokens on HLE. Thus, the compact response-transition summary reduces reconstruction error across lengths, while its directional advantage depends on the stratum.

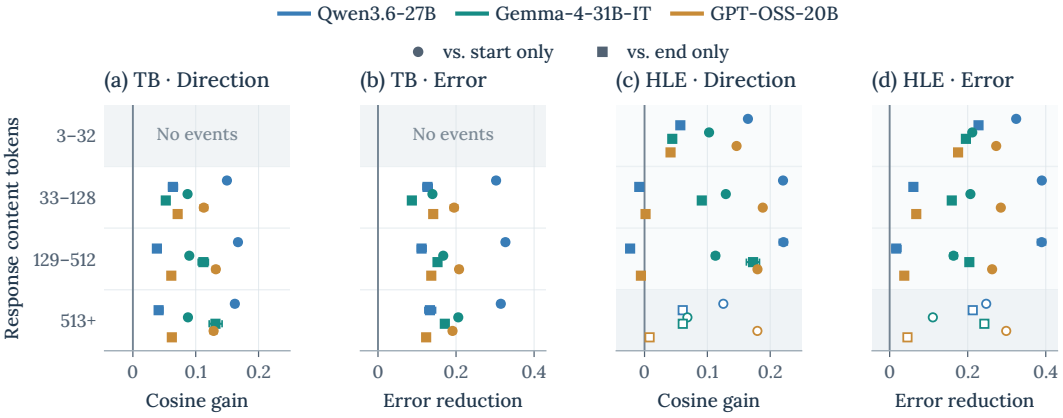


Figure 7: Dual endpoints versus individual response anchors, stratified by response length. Positive values indicate higher cosine similarity or lower relative error against the full-response mean displacement. Bars show paired 95% task-bootstrap intervals; hollow markers denote strata with fewer than three tasks.

D.2 Representation Dynamics on HLE

Lower response-end retention persists on HLE despite different update profiles. Figure 8 extends the full-depth analysis to tool-assisted reasoning. Qwen's update magnitude generally in-

creases with depth, Gemma exhibits a middle-layer trough, and GPT-OSS declines before a final-layer rise. Despite these differences, response-end retention is below response-start retention in 144/148 layers; the four exceptions occur in Gemma's early layers. Verbal accessibility and cross-cycle consistency also follow distinct depth profiles. Together, the profiles distinguish update size, retention, accessibility, and consistency when comparing events.

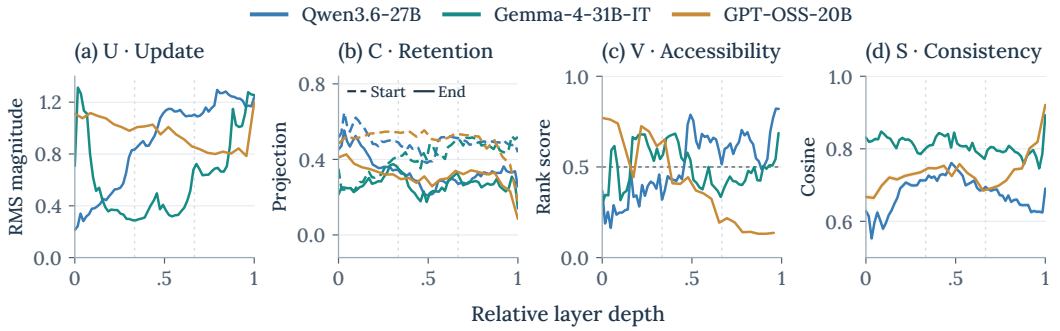


Figure 8: Full-depth HLE profiles with pointwise 95% task-bootstrap intervals. Dashed and solid C curves mark response start and end. $U/C/V$ use 1,303 tasks; S uses 1,091 with valid adjacent updates. Only horizontal depth is rescaled; V excludes unsupported final layers.

D.3 Retention Within a Response

The same endpoint ordering can arise from different response dynamics. At the diagnostic layers used above, mean end-anchor retention is below start-anchor retention in all six model–benchmark groups (Figure 9). The intermediate paths nevertheless differ. On Terminal-Bench, the final token lies 0.126–0.157 below the last-bin mean; on HLE, it lies 0.007–0.083 above it. Gemma shows the clearest HLE recovery, reaching $C^{\text{end}} = 0.394$ after lower middle-response projections. Endpoint attenuation therefore does not imply uniform decay throughout generation. Distinct anchors preserve these stage-specific changes when local response transitions are compared.

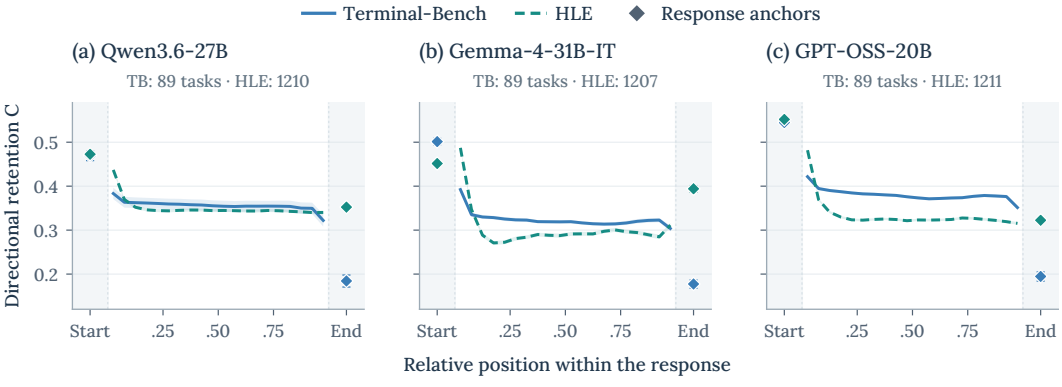


Figure 9: Within-response retention at layers 23/22/8 for Qwen/Gemma/GPT-OSS. Lines show 20 relative-position bin means; separate diamonds mark the first and last tokens. All recorded events with at least 20 content tokens are included. Shading and bars show 95% task-bootstrap intervals.

The length criterion avoids empty bins and keeps event support identical across positions: 2,935/2,935/2,923 Terminal-Bench events and 9,524/9,062/9,254 HLE events for Qwen/Gemma/GPT-OSS. Single-token anchors remain distinct from bin means.

D.4 Readout Rank and Content Coverage

High relative content rank need not imply concentrated content recovery. Figure 10 compares V with Recall@50, the fraction of unique response-content tokens among the boundary J-lens's top 50 vocabulary entries. GPT-OSS reaches $V = 0.710$ in early HLE layers, but Recall@50 is only 0.0245%; the largest band-level recall across all groups is 1.59%. Relative vocabulary rank and top-ranked content coverage thus capture different properties. In REVEAL, V describes accessibility for matching events within each observer, rather than action correctness or recovery of the full response.



Figure 10: Task-balanced V and Recall@50 across early/middle/late depth bands on Terminal-Bench and HLE, using supported J-lens layers. Panels use separate linear scales; uniformly random vocabulary ranking has expected $V = 0.5$.

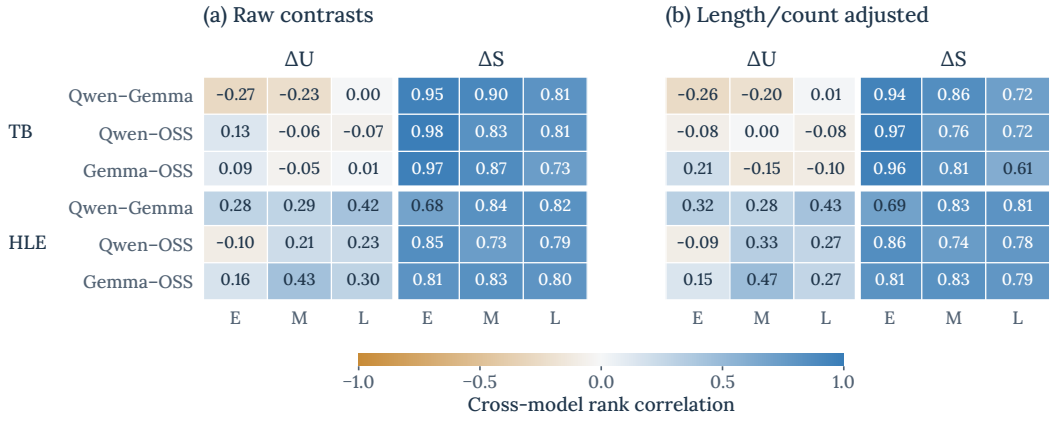


Figure 11: Cross-model correlations of same-task success–failure contrasts before and after adjustment for response length and analysis-center count. E/M/L denote early/middle/late depth bands. Every panel uses the same 34 Terminal-Bench and 208 HLE tasks; ΔU provides the update-magnitude comparison.

Throughout the band-level analyses, early/middle/late denote zero-based layers 0–21/22–42/43–63 for Qwen, 0–19/20–39/40–59 for Gemma, and 0–7/8–15/16–23 for GPT-OSS. Readout summaries exclude each model's final layer, where the fitted lens is unavailable.

D.5 Outcome-Conditioned Structure Across Observers

Success–failure contrasts in cross-cycle consistency remain aligned across observers. For each task with successful and failed frontier-model rollouts, we subtract the two groups' mean profiles and correlate these taskwise contrasts across models. Across 34 Terminal-Bench and 208 HLE tasks, ΔS correlations range from 0.729–0.977 and 0.684–0.855, respectively (Figure 11). After rank-based adjustment for response length and analysis-center count, they remain 0.609–0.971

and 0.688–0.856. The sign of ΔS varies across tasks: successful rollouts do not uniformly have higher consistency. This supports comparing task-specific dynamics by contrast magnitude, as in REVEAL's selection score, while using outcomes to orient the resulting contrasts.

Paired analysis. We use known-outcome source trajectories with complete three-band U/S profiles, average trajectories within each task/outcome class, and retain tasks shared by all observers. Adjustment regresses ranked profile contrasts on an intercept and four ranked covariates: each observer's success–failure differences in mean $\log(1+\text{response characters})$ and $\log(1+\text{analysis centers})$. Character counts are trajectory means before transformation. Correlations use the residual ranks; bootstrap resamples refit ranks and regressions. All 18 adjusted ΔS intervals remain above zero. These task-level associations complement local-event selection; they do not establish its causal contribution.