

Multi-View Structural Interpretability of Agentic Reasoning Traces

Ritika Hooda^{1*}, Sravanthi Machcha^{1*}, Jackson MacDonald¹
Lavanika Srinivasaraghavan¹, Atharva Nijasure¹, Rui Wang², Lei Shi²
Junda Wu², Tong Yu^{2#}

¹University of Massachusetts Amherst ²Adobe Research
{rhooda, smachcha, jrmacdonald, lsrinivasara, anijasure}@umass.edu
{ruiwan, leis, jundaw, tyu}@adobe.com

Abstract

Agentic language models solve tasks through reasoning, tool calls, observations, and terminal answers, exposing rich process information beyond the final answer. For practitioners, knowing which process patterns lead to success or failure is essential to diagnose agent errors and build better process reward models without relying on final-answer supervision. However, existing analyses often evaluate only final outcomes or, when examining traces, flatten them into homogeneous step sequences, fragmenting each reasoning round into disconnected steps. Our key insight is that an agent’s reasoning, action, and observation within a round are better treated as typed segments of a single underlying process state, rather than independent steps. We propose MULTI-VIEW-RA, a reasoning abstraction that represents each round through two complementary views: accumulated state and within-round change. Using unlabeled clustering and binary outcome labels, we apply MULTI-VIEW-RA to two practical use cases: diagnosing failures by surfacing the process states that separate correct from incorrect trajectories, and providing a transition-level correctness signal that is a candidate source of process supervision. Across Math, HotpotQA, and FEVER with multiple Qwen and Llama models, MULTI-VIEW-RA outperforms step-level explainability baselines in cases with sufficient success rates, with particularly strong gains on longer, tool-rich traces.

1 Introduction

Language models are increasingly used as agents for text-based reasoning tasks, where they reason, call tools, read observations, and commit to an answer over many steps (Yao et al., 2023; Shinn et al., 2023). This makes their outputs qualitatively different from ordinary single-turn generation: the final

answer is only one part of the record, and the trajectory itself is a structured record of how the model attempted the task. For practitioners, that record is valuable precisely because knowing which process patterns precede success or failure is what lets them diagnose agent errors and build better process reward models without relying solely on final-answer supervision. If an agent succeeds, which patterns tend to come first? If it fails, does it fall into recognizable loops, error-recovery states, or premature finalization? Answering such questions requires treating the trace as a structured object rather than only as text or as a final prediction. In this work, we focus on the open-weight setting, where final-layer hidden states are accessible and serve as our substrate for structural analysis.

Most prior analyses consider only the final outcome (Schick et al., 2023; Qin et al., 2024; Li et al., 2023), and those that examine the trace itself typically represent it as an undifferentiated sequence of homogeneous steps (Yu et al., 2025). While convenient, this representation discards structure that is essential in agentic traces. In REACT-style settings (Yao et al., 2023), thoughts, actions, observations, and final answers serve distinct functional roles, and a given action is frequently interpretable only with respect to the thought that produced it and the observation that followed. A failure mode, likewise, rarely corresponds to an isolated step; it more often manifests as a transition between process states, such as from an invalid tool call to repeated repair attempts, or from evidence integration to final-answer synthesis (Section 6.3).

We uncover a simple but consequential structure: within a single round, the reasoning, action, and observation are not separate steps but a single underlying process state. Building on this, we propose MULTI-VIEW-RA, a trace abstraction that draws on multi-view learning (Yang et al., 2025; Gao et al., 2025) to represent each round through two complementary views: accu-

*Equal contribution.

#Corresponding author.

mulated state and within-round change. This is not straightforward: the three typed segments occupy different functional roles, so collapsing a round into a single flat step or simply concatenating its parts can obscure how information accumulates and changes within the round. Our hypothesis is that these round-level abstractions expose correctness-relevant structure more clearly than flat step-level clustering (Yu et al., 2025). We contrast four representations of a trajectory: FLAT clusters each typed segment independently, MERGED-AO merges action-observation segments before clustering, RAW ROUND-CONCAT represents each round by concatenating its thought, action, and observation embeddings, and MULTI-VIEW-RA represents each round through accumulated-state and within-round-change views. Each abstraction turns a trajectory into a sequence of cluster assignments.

With unlabeled clustering and a binary correct/incorrect label per trajectory, MULTI-VIEW-RA supports two practical use cases: surfacing the process states that distinguish correct from incorrect runs to diagnose failures, and providing a transition-level correctness signal that is a candidate source of process supervision (Lightman et al., 2024; Wang et al., 2024). Concretely, we estimate transition matrices separately for correct and incorrect training traces and give each transition a reward equal to the log-likelihood ratio between its correct and incorrect probabilities. Held-out trajectories are then scored by the average reward of their transitions. A single quantity therefore does double duty: it measures how much correctness-relevant structure an abstraction carries, while its highest- and lowest-reward transitions can be inspected and, in particular on the failure side, tied back to qualitative process patterns. Our contributions are:

- **Key insight.** We show that the reasoning, action, and observation of a single round are better treated as one process state than as independent steps, and represent this state with two complementary views constructed from their cumulative-Gram embeddings, framing agentic trace analysis as structural interpretability over the resulting states and transitions.
- **MULTI-VIEW-RA.** We propose MULTI-VIEW-RA, a reasoning trace abstraction inspired by multi-view learning, and compare it against FLAT step-level, MERGED-AO action-observation abstractions and RAW ROUND-CONCAT round-concatenation abstractions of

REACT traces under a shared transition-based correctness-scoring diagnostic.

- **Two use cases.** We apply MULTI-VIEW-RA to failure diagnosis and to a transition-based correctness signal that can serve as a candidate source of process supervision, and establish that its transition dynamics predict trajectory correctness better than step-level abstractions while surfacing states and transitions that map to interpretable process patterns.

2 Related Work

Latent structure in chain-of-thought reasoning.

We build on work treating a trace as latent states whose transitions, not surface text, predict correctness. Yu et al. (2025) spectrally embed each step, cluster them, and model the trace as a first-order Markov chain over states; Wu et al. (2025) show such states are actionable for reinforcement learning. Others read correctness from segment roles: Li et al. (2026) label math reasoning as Schoenfeld episodes linking exploration to success, and Sun et al. (2026) predict correctness from late-stage subspace divergence. Most assume a flat, homogeneous, tool-free trace unlike agentic traces.

Multi-view learning and clustering. Multi-view learning treats an entity through complementary views that are individually incomplete but jointly informative, fusing them before clustering rather than concatenating or collapsing them (Qin et al., 2025). Classically these views are given by the data, such as distinct modalities or sensors (Yang et al., 2025), or feature sets of independent samples (Gao et al., 2025), and the goal is to reconcile information across them (Sun et al., 2024). We instead apply this principle *within* a single trajectory: we derive two complementary views—accumulated state and within-round change—from the thought, action, and observation of each round and align them into one round-level state rather than treat as independent steps.

Agentic reasoning and latent trajectory structure.

Tool-use and agent benchmarks (Schick et al., 2023; Qin et al., 2024; Li et al., 2023) evaluate task success, tool selection, or efficiency rather than latent transition structure, while REACT (Yao et al., 2023) and Reflexion (Shinn et al., 2023) interleave deliberation with tool calls and observations. Option and skill-discovery methods (Sutton et al., 1999; Bacon et al., 2017; Eysenbach et al., 2019)

seek reusable structure in agent trajectories but operate over sequences of actions and observations rather than language-model hidden states. Continuous latent-reasoning methods (Hao et al., 2024; Zhang et al., 2025; Zou et al., 2025) intervene at generation time, whereas we analyze existing traces post-hoc without altering agent behavior.

Process reward models and reward-guided reasoning. Our scoring use case, providing a transition-level correctness signal that is a candidate source of process supervision, is related to process reward modeling, which values intermediate steps over final answers (Lightman et al., 2024; Wang et al., 2024). Recent work pushes this signal in related directions: Liu et al. (2025) add interpretable reward dimensions; Gu and Liang (2025) prune steps with an external step-level reward; Li et al. (2025) reward tool-use completeness over actions; Zhang and Xiong (2025) add step-wise process supervision; and Zhang (2025) score reasoning blocks by their contribution to confidence. These signals are attached to individual steps or actions; we instead require only a single binary outcome label per trace and score the *transitions* between latent states (Section 6.2).

3 Preliminaries

Agentic traces. Let x denote an input question. An agentic trace is organized into *rounds*. Each round r_t ($t = 1, \dots, T$) comprises of atmost three typed segments: a thought τ_t , an action a_t , and an observation o_t . The trace terminates when the agent emits a `final_answer`, yielding a binary outcome label $\ell \in \{+, -\}$ (correct / incorrect).

Step embedding and spectral abstraction. Following Yu et al. (2025), we treat each typed segment as one step in the trace and index segments by their linear position $t = 1, \dots, N$. For a segment with token-level hidden states $X_t \in \mathbb{R}^{n_t \times d}$ from the last transformer layer, we compute the local Gram matrix $\tilde{G}_t = X_t^\top X_t$ and the spectral embedding $E_t \in \mathbb{R}^{k_{\text{eig}}}$ as the vector of top- k_{eig} eigenvalues of the Gram matrix:

$$E_t = (\lambda_1, \dots, \lambda_{k_{\text{eig}}}), \quad \lambda_1 \geq \dots \geq \lambda_{k_{\text{eig}}}.$$

To incorporate contextual accumulation, we maintain a running Gram matrix updated as each segment is processed:

$$G_t = G_{t-1} + \tilde{G}_t, \quad G_1 = \tilde{G}_1,$$

and take E_t from the top- k_{eig} eigenvalues of G_t . We define $E_{\tau_t}, E_{a_t}, E_{o_t}$ as the top- k_{eig} eigenvalues of the cumulative Gram matrices $G_{\tau_t}, G_{a_t}, G_{o_t}$, updated sequentially within each round as $G_{\tau_t} = G_{o_{t-1}} + \tilde{G}_{\tau_t}, G_{a_t} = G_{\tau_t} + \tilde{G}_{a_t}, G_{o_t} = G_{a_t} + \tilde{G}_{o_t}$, with $G_{o_0} = \mathbf{0}$.

4 Method

Figure 1 gives an overview of how MULTI-VIEW-RA abstracts REACT trajectories into latent process states and uses their transitions for diagnosis and trajectory correctness scoring.

4.1 Multi-View Round-Level Abstraction

The nested-view problem. A naive application of multi-view learning to agentic traces fails because the cumulative Gram matrix construction makes within-round embeddings nearly identical. Since each G_t is a sum of Gram matrices that are positive semi-definite, eigenvalues can only increase as segments are added, yielding the component-wise ordering:

$$E_{\tau_t} \leq E_{a_t} \leq E_{o_t}.$$

As a result, the three within-round embeddings are nearly collinear: simply concatenating or averaging them adds negligible new information over the thought embedding alone.

State and delta views. We resolve this by decomposing each round r_t into two views. Let $E_{\text{end}}(r_t)$ denote the spectral embedding at the end of the round, E_{o_t} for a complete round (thought $\rightarrow$ action $\rightarrow$ observation) and E_{final} for a terminal round (thought $\rightarrow$ final_answer). Recall from Section 3 that $E_{\tau_t}, E_{a_t}, E_{o_t}$ are the cumulative spectral embeddings evaluated after the thought, action, and observation segments of round r_t . We define:

$$\begin{aligned} \mathbf{v}_t^{(s)} &= E_{\tau_t}, \\ \mathbf{v}_t^{(\delta)} &= E_{\text{end}}(r_t) - E_{\tau_t}, \end{aligned}$$

where $E_{\text{end}}(r_t) = E_{o_t}$ for a complete round and $E_{\text{end}}(r_t) = E_{\text{final}}$ for a terminal round ending in `final_answer`. The state view $\mathbf{v}_t^{(s)}$ captures the accumulated context at the start of the round, while the delta view $\mathbf{v}_t^{(\delta)}$ captures the net spectral change the round produces. Intuitively, the state view captures *where the model is*; the delta view captures *what this round contributed*. The views are non-redundant by construction: delta is by definition

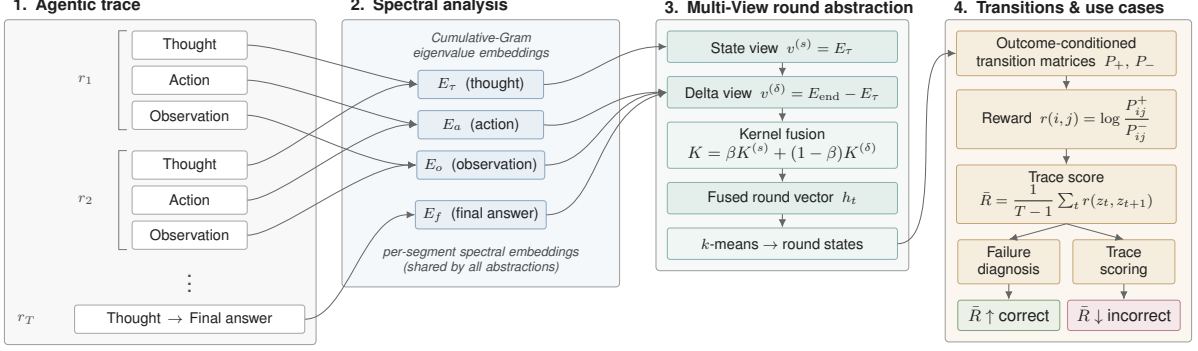


Figure 1: Overview of MULTI-VIEW-RA, which abstracts REACT trajectories into round-level latent states for transition-based failure diagnosis and trajectory correctness scoring.

the component of the endpoint not in state. Empirically, the two views exhibit limited global similarity across most evaluated conditions (Appendix F.3).

Preprocessing. Spectral eigenvalues span several orders of magnitude, so both views are first compressed dimension-wise via $\log(1 + |\cdot|)$, then each dimension is standardized independently to zero mean and unit variance. Independent standardization is essential: joint standardization would allow the larger-magnitude state view to dominate the fused representation.

Kernel fusion. We construct a linear kernel for each view:

$$K^{(s)} = \hat{\mathbf{V}}^{(s)} (\hat{\mathbf{V}}^{(s)})^\top / k_{\text{eig}},$$

$$K^{(\delta)} = \hat{\mathbf{V}}^{(\delta)} (\hat{\mathbf{V}}^{(\delta)})^\top / k_{\text{eig}}.$$

where $\hat{\mathbf{V}}^{(s)}$ and $\hat{\mathbf{V}}^{(\delta)}$ are the standardized state and delta matrices stacked over all rounds. The two kernels are fused as:

$$K_{\text{fused}} = \beta K^{(s)} + (1 - \beta) K^{(\delta)},$$

where $\beta \in [0, 1]$ balances the two views. We set $\beta = 0.5$ to weight them equally, as both views are independently standardized and thus on comparable scales. This is equivalent to concatenating the two standardized views and applying kernel PCA, with each view contributing equally.

Fused representation and clustering. We eigen-decompose $K_{\text{fused}} = V \Lambda V^\top$ and form a low-dimensional embedding by retaining the top m eigenvectors scaled by the square root of their eigenvalues:

$$Z[i, :] = [\sqrt{\lambda_1} V[i, 1], \dots, \sqrt{\lambda_m} V[i, m]].$$

k -means is run on Z to assign each round r_t a latent state $z_t \in \{1, \dots, k_{\text{clu}}\}$. A trace is thus abstracted into a round-level state sequence $z = (z_1, \dots, z_T)$.

4.2 Use Cases

Outcome-conditioned transition matrices. Given a training corpus of traces with binary outcome labels $\ell \in \{+, -\}$, we estimate separate transition matrices P_+ and $P_- \in \mathbb{R}^{k_{\text{clu}} \times k_{\text{clu}}}$ from correct and incorrect traces respectively:

$$P_{i,j}^\pm = P(z_{t+1} = j \mid z_t = i, \ell = \pm)$$

$$= \frac{C_{i,j}^\pm + \alpha}{\sum_{j'} C_{i,j'}^\pm + k_{\text{clu}} \alpha},$$

with $C_{i,j}^\pm$ the observed transition counts under each outcome. We use $\alpha = 10^{-3}$ throughout the main experiments; sensitivity to this choice is reported in Appendix F.5.

Transition reward and trace scoring. Our method provides a transition-level correctness signal that is a candidate source of process supervision. We assign each transition ($i \rightarrow j$) a reward equal to the log-likelihood ratio between its correct- and incorrect-trajectory probabilities:

$$r(i, j) = \log \frac{P_{i,j}^+}{P_{i,j}^-}.$$

A held-out trajectory is scored by its average transition reward:

$$\bar{R} = \frac{1}{T-1} \sum_{t=1}^{T-1} r(z_t, z_{t+1}).$$

This scalar serves as a process-level correctness signal. For *failure diagnosis*, the transitions with the largest positive and negative rewards identify the most likely successful and failed rounds.

5 Experimental Setup

Datasets and models We evaluate on three datasets spanning distinct reasoning regimes: MATH (symbolic/numerical reasoning), HotpotQA (multi-hop QA, no-context setting), and FEVER (claim verification). For each dataset we generate traces with five instruction-tuned models: Qwen2.5-32B, Qwen2.5-7B, Llama-3.1-8B, Llama-3.2-3B, and Gemma-2-2B. For each dataset-model condition we shuffle with seed 0 and generate traces for the first 1000 examples. Per-condition trace statistics are in Table 2. Our primary evaluation uses multiple Qwen and Llama models; Gemma-2-2B is reported separately as a diagnostic condition because its extreme outcome imbalance prevents stable split-level estimation.

Trace generation A single generic REACT-style agent is shared across tasks. At each step the model emits exactly one JSON object containing a tool action or a final answer. The agent has two tools—a calculator (one short numeric Python expression) and `wiki_search` (a short Wikipedia summary)—and continues until a `final_answer` or six REACT steps. We decode deterministically (temperature 0.0, top- p 1.0), limiting each turn to 192 new tokens. Full prompts are in Appendix A.

Protocol Following the explainable-CoT setting (Yu et al., 2025), we extract final-layer hidden states, project them to 128 dimensions, and first compute cumulative-Gram spectral embeddings for each typed segment. FLAT clusters these per-segment embeddings directly, MERGED-AO clusters action–observation merged units, RAW ROUND-CONCAT directly concatenates the thought, action, and observation embeddings within each round before clustering, and MULTI-VIEW-RA first converts the segment embeddings within each round into state and delta views, fuses them, and then clusters the resulting round-level representations into $k=5$ process states. Unless otherwise stated, all primary results use a fully inductive evaluation: on five stratified 70/30 splits over sample IDs usable under all four abstractions, all learned abstraction components are fit on training traces only and applied to held-out traces. *Clustering analysis* measures outcome-conditioned transition separation via transition LLR AUC and Cohen’s d , with silhouette as a geometric-compactness check (Table 3). *Correctness-scoring evaluation* estimates smoothed transition matrices

from training traces, scores held-out trajectories by average transition reward, and reports trajectory AUC, Cohen’s d , and reward gap (Table 4). We report the corresponding transductive comparison in Appendix F.1. FLAT and MERGED-AO are step-level structural baselines, while RAW ROUND-CONCAT serves as a direct round-level concatenation control.

Cluster labeling. For qualitative interpretation only, we label each cluster with a Large Language Model (LLM) judge (gpt-5.4-mini); no label enters the transition-reward estimation or any quantitative metric (details in Appendix D.1).

6 Results

We establish that cluster quality does not track outcome relevance. We further evaluate MULTI-VIEW-RA against two step-level explainability baselines FLAT and MERGED-AO and the round-level RAW ROUND-CONCAT control on two use cases: scoring trace correctness from process dynamics as a candidate source of process supervision, and diagnosing failures via the process states that separate correct from incorrect trajectories.

6.1 Clustering Quality Does Not Track Outcome Relevance

Cluster geometry and outcome separation are not aligned. Table 3 (Appendix B.2) reports complete clustering diagnostics, including silhouette together with transition LLR AUC and Cohen’s d , across all 12 dataset-model conditions. On MATH/Qwen2.5-7B, FLAT and MERGED-AO obtain the highest silhouette scores (0.497 each), compared with 0.410 for RAW ROUND-CONCAT and 0.449 for MULTI-VIEW-RA. In contrast, MULTI-VIEW-RA produces the strongest outcome-conditioned transition separation, reaching an LLR-AUC of 0.751 and Cohen’s d of 0.876, compared with 0.698/0.529 for FLAT, 0.705/0.569 for MERGED-AO, and 0.643/0.481 for RAW ROUND-CONCAT. This pattern is consistent across the 12 primary dataset-model conditions: MULTI-VIEW-RA achieves the highest LLR-AUC and Cohen’s d in every condition, even when another abstraction has higher silhouette. Thus, selecting an abstraction by cluster geometry would discard the structure that carries the correctness signal. This finding motivates our use of outcome-conditioned transition separation, rather than cluster geometry, as the primary criterion for evaluating the abstractions.

Dataset	Model	Flat	M-AO	Raw	MV	Δ	MV Cohen's d	MV Reward Gap
MATH	Qwen2.5-7B	0.690	0.696	0.596	0.748	+0.052	0.857	0.791
	Qwen2.5-32B	0.614	0.632	0.594	0.776	+0.144	1.023	0.742
	Llama-3.1-8B	0.792	0.808	0.703	0.910	+0.102	1.827	1.402
	Llama-3.2-3B	0.635	0.645	0.663	0.741	+0.078	0.854	0.497
HotpotQA	Qwen2.5-7B	0.660	0.662	0.562	0.701	+0.039	0.779	0.306
	Qwen2.5-32B	0.661	0.680	0.610	0.757	+0.077	0.885	0.438
	Llama-3.1-8B	0.681	0.692	0.573	0.853	+0.161	1.588	0.971
	Llama-3.2-3B	0.525	0.553	0.496	0.669	+0.116	0.584	0.212
FEVER	Qwen2.5-7B	0.676	0.678	0.625	0.692	+0.014	0.863	0.405
	Qwen2.5-32B	0.636	0.638	0.604	0.682	+0.044	0.680	0.427
	Llama-3.1-8B	0.565	0.585	0.590	0.771	+0.181	0.985	0.603
	Llama-3.2-3B	0.535	0.529	0.523	0.614	+0.079	0.444	0.076

Table 1: Held-out trajectory correctness-scoring headline results ($k=5$, five stratified 70/30 splits). Flat, M-AO, Raw, and MV denote FLAT, MERGED-AO, RAW ROUND-CONCAT, and MULTI-VIEW-RA, respectively. Δ is MV AUC minus the strongest baseline AUC. Full results are reported in Appendix B.2.

6.2 Scoring Trace Correctness as a Candidate Source of Process Supervision

We estimate the transition reward from binary outcome labels on the training split and apply it to held-out trajectories as a correctness score for reward modeling. TAcross all 12 primary Qwen and Llama dataset-model conditions, MULTI-VIEW-RA achieves the highest held-out trajectory AUC among the four structural abstractions as in Table 1; full comparisons with FLAT, MERGED-AO, and RAW ROUND-CONCAT are reported in Appendix B.2 On MATH, MULTI-VIEW-RA reaches an AUC of 0.748 for Qwen2.5-7B, compared with 0.690, 0.696, and 0.596 for FLAT, MERGED-AO, and RAW ROUND-CONCAT, respectively. For Llama-3.1-8B, the corresponding values are 0.792/0.808/0.703/0.910. MULTI-VIEW-RA likewise leads on HotpotQA/Llama-3.1-8B (0.681/0.692/0.573/0.853) and FEVER/Llama-3.1-8B (0.565/0.585/0.590/0.771), for FLAT/MERGED-AO/RAW ROUND-CONCAT/MULTI-VIEW-RA, respectively. The gains over RAW ROUND-CONCAT show that round-level grouping alone is not sufficient: the structured state-delta representation captures correctness-relevant dynamics that are obscured by direct concatenation. Full split-level paired comparisons are reported in Appendix B.2. Figure 2 visualizes the learned reward over normalized trace position for MATH; additional robustness analyses are reported in Appendix F.5.

The gains over the step-level baselines are often higher on longer, tool-rich traces, which contain more reasoning–action–observation structure. The trajectory-level trends in Figure 2 explain this gap: FLAT and MERGED-AO rewards stay near zero

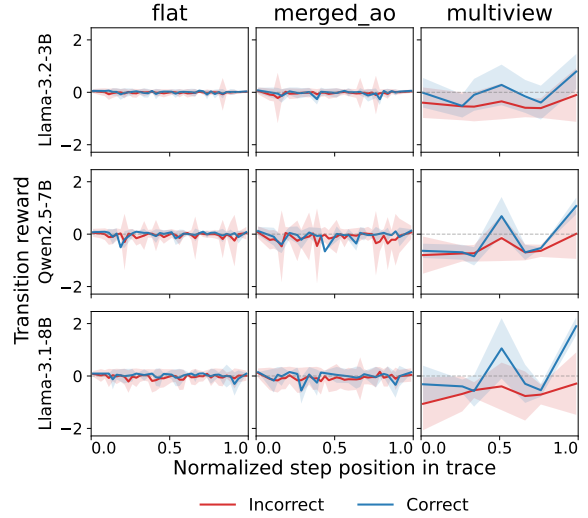


Figure 2: Transition-reward profiles on MATH for a representative held-out split ($k=5$, seed 1). “multiview” denotes MULTI-VIEW-RA. Compared with FLAT and MERGED-AO, MULTI-VIEW-RA yields clearer separation between correct and incorrect trajectories.

for both correct and incorrect traces, while MULTI-VIEW-RA separates the two outcomes across trace position, well before the final round and also in later regions, where correct trajectories receive higher reward. This suggests that the correctness signal is carried by round-level transition structure rather than by isolated step clusters. These comparisons are among structural abstractions that share the same clustering-and-transition pipeline; hand-crafted surface-feature classifiers over trace length, action counts, and observation statistics also carry substantial correctness signal and are complementary to MULTI-VIEW-RA (Appendix F.4).

A balanced outcome population is a precondition. The transition reward contrasts success- and failure-conditioned dynamics and therefore requires enough examples of both outcomes. Gemma-2-2B solves only 0.2%, 0.3%, and 0.1% of MATH, HotpotQA, and FEVER traces, respectively (Table 2), making stable fully inductive estimation infeasible. We therefore exclude Gemma-2-2B from the primary 12-condition comparison and report its results separately as diagnostic transductive analyses in Appendix F.2.

6.3 Diagnosing Failures Through Interpretable States

Because the reward is defined over transitions between latent cluster states, the transitions that separate correct from incorrect trajectories are directly



Figure 3: Outcome-conditioned transition differences on HotpotQA for Llama-3.1-8B across abstraction methods. Each cell shows $\Delta P = P_{\text{correct}} - P_{\text{incorrect}}$, the difference between transition probabilities estimated from correct and incorrect trajectories. “multiview” denotes MULTI-VIEW-RA.

interpretable. Figure 3 shows outcome-conditioned transition differences for a representative HotpotQA setting. Appendix D.2 provides additional transition-difference heatmaps and t-distributed stochastic neighbor embedding (t-SNE) projections for MATH with Qwen2.5-7B and Llama-3.1-8B, while Appendix D.1 reports LLM-judge labels for the recovered clusters. The MATH cluster-label tables (e.g., Table 7) show that the recovered states correspond to interpretable calculation, simplification, and tool-use patterns. The appendix transition heatmaps in Figures 5 and 6 show how those states differ between successful and failed trajectories.

To test whether the localized transitions correspond to recognizable failures, two annotators independently evaluated 54 HotpotQA/Qwen2.5-32B transitions: 40 selected by the most-negative MULTI-VIEW-RA reward, together with positive and random-reward controls. Annotators saw the local trace window and final outcome, but not the reward or gold answer. Agreement on whether a recognizable failure occurred at the marked locus was substantial (Cohen’s $\kappa = 0.71$), and each annotator identified 36/40 flagged transitions (90%) as failures. These were dominated by retrieval failures, whose categories map to concrete interventions such as query reformulation and retry policies (full protocol and category breakdown in Appendix C).

The same signal can also support deployment triage (Figure 4). At a 20% review budget, ranking traces by MULTI-VIEW-RA reward surfaces 33% of all failed trajectories, compared with 20% under random routing and close to the 33.1% oracle ceiling at that operating point. At low review budgets, simple trace-length heuristics perform similarly; as the budget increases, however, MULTI-VIEW-RA generally recovers more failed traces while routing

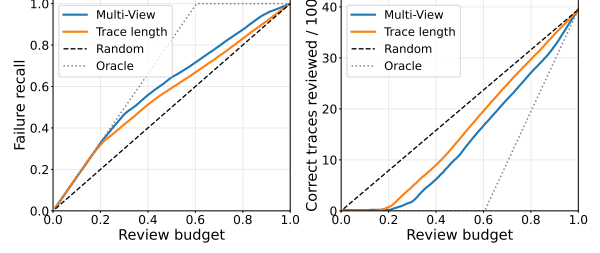


Figure 4: **Failure triage on HotpotQA/Qwen2.5-32B.** Held-out traces are ranked for review by MULTI-VIEW-RA reward, trace length, or random routing under the inductive protocol (five 70/30 splits; 60.5% failure base rate). **Left:** recall of failed traces as the review budget increases. **Right:** wasted review effort, measured as the number of correct traces queued per 100 traces. The oracle shows the achievable ceiling at each review budget. “Multi-View” denotes MULTI-VIEW-RA.

fewer correct traces for review. Thus, the transition reward provides a practical failure-ranking signal in addition to its role as a structural diagnostic.

7 Discussion and Conclusion

The unit of analysis determines how clearly agentic traces expose correctness-relevant structure. Representing each REACT round through accumulated-state and within-round-change views recovers outcome-relevant structure that step-level abstractions often obscure. Under an identical pipeline, and using only unsupervised clustering and binary outcome labels, MULTI-VIEW-RA outperforms FLAT and MERGED-AO on both use cases: it scores held-out trace correctness more accurately and yields interpretable process states that separate successful from failed trajectories.

The two use cases are complementary: failure diagnosis identifies where an agent’s process diverges, while correctness scoring converts the same transition structure into a process-level signal that requires no per-step annotation. The transition reward is a candidate process-supervision signal complementary to surface-feature trajectory classifiers, supporting analysis and diagnosis of agentic reasoning processes.

Acknowledgments

We thank the anonymous reviewers for the constructive feedback. Generative AI tools were used to assist with language editing and polishing, coding and debugging, and figure creation.

Limitations

Our approach requires access to model hidden states, from which we construct the spectral representations used for structural abstraction. These hidden states are already produced during generation, so the analysis adds no additional model invocations. Hence, this implementation is most directly applicable to open-weight models. Extending the framework to closed models remains an open direction for future work and may be possible by replacing hidden-state-derived spectral representations with embeddings from an external text encoder while retaining the subsequent abstraction and transition-analysis stages, but we do not explore or evaluate this setting in this work.

References

- Pierre-Luc Bacon, Jean Harb, and Doina Precup. 2017. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.
- Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. [Why Do Multi-Agent LLM Systems Fail?](#) In *Advances in Neural Information Processing Systems*.
- Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2019. [Diversity Is All You Need: Learning Skills Without a Reward Function](#). In *International Conference on Learning Representations*.
- Shangde Gao, Ke Liu, Yichao Fu, Hongxia Xu, and Jian Wu. 2025. Matrix factorization with dynamic multi-view clustering for recommender system. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Yuan Gao, Yiheng Zhu, Yuanbin Cao, Yinzhi Zhou, Zhen Wu, Yujie Chen, Shenglan Wu, Haoyuan Hu, and Xinyu Dai. 2024. [Dr3: Ask large language models not to give off-topic answers in open domain multi-hop question answering](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5350–5364, Torino, Italia. ELRA and ICCL.
- Jiawei Gu and Shangsong Liang. 2025. [Speculative reward model boosts decision making ability of LLMs cost-effectively](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 4–21, Vienna, Austria. Association for Computational Linguistics.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.
- Ming Li, Chenrui Fan, Yize Cheng, Soheil Feizi, and Tianyi Zhou. 2026. [Schoenfeld’s anatomy of mathematical reasoning by language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32773–32802, San Diego, California, United States. Association for Computational Linguistics.
- Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. Api-bank: A comprehensive benchmark for tool-augmented llms. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 3102–3116.
- Zhiwei Li, Yong Hu, and Wenqing Wang. 2025. [Encouraging good processes without the need for good answers: Reinforcement learning for LLM agent planning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1654–1666, Suzhou (China). Association for Computational Linguistics.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.
- Xiaoyu Liu, Di Liang, Hongyu Shan, Peiyang Liu, Yonghao Liu, Muling Wu, Yuntao Li, Xianjie Wu, Li Miao, Jiangrong Shen, and Minlong Peng. 2025. [Structural reward model: Enhancing interpretability, efficiency, and scalability in reward modeling](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 672–685, Suzhou (China). Association for Computational Linguistics.
- Yalan Qin, Xinpeng Zhang, Shui Yu, and Guorui Feng. 2025. A survey on representation learning for multi-view data. *Neural Networks*, 181:106842.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pages 9695–9717.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement

- learning. *Advances in neural information processing systems*, 36:8634–8652.
- Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. 2026. Llm reasoning as trajectories: Step-specific representation geometry and correctness signals. *arXiv preprint arXiv:2604.05655*.
- Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. 2024. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):9150–9162.
- Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.
- Junda Wu, Yuxin Xiong, Xintong Li, Sheldon Yu, Zhengmian Hu, Tong Yu, Rui Wang, Xiang Chen, Jingbo Shang, and Julian McAuley. 2025. Ctrls: Chain-of-thought reasoning via latent state-transition. *arXiv preprint arXiv:2507.08182*.
- Xihong Yang, Siwei Wang, Fangdi Wang, Jiaqi Jin, Suyuan Liu, Yue Liu, En Zhu, Xinwang Liu, and Yueming Jin. 2025. Automatically identify and rectify: Robust deep contrastive multi-view clustering in noisy scenarios. *arXiv preprint arXiv:2505.21387*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing Reasoning and Acting in Language Models](#). In *International Conference on Learning Representations*.
- Sheldon Yu, Yuxin Xiong, Junda Wu, Xintong Li, Tong Yu, Xiang Chen, Ritwik Sinha, Jingbo Shang, and Julian McAuley. 2025. Explainable chain-of-thought reasoning: An empirical analysis on state-aware reasoning dynamics. *arXiv preprint arXiv:2509.00190*.
- Jiaxin Zhang. 2025. [Confidence-aware reasoning: Optimizing self-guided thinking trajectories in large reasoning models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2081–2095, Suzhou (China). Association for Computational Linguistics.
- Shaowei Zhang and Deyi Xiong. 2025. [BackMATH: Towards backward reasoning for solving math problems step by step](#). In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 466–482, Abu Dhabi, UAE. Association for Computational Linguistics.
- Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Wang. 2025. [Soft thinking: Unlocking the reasoning potential of llms in continuous concept space](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 168990–169012. Curran Associates, Inc.
- Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, et al. 2025. Latent collaboration in multi-agent systems. *arXiv preprint arXiv:2511.20639*.

A Prompts

A.1 Agentic trace generation.

We generate traces with a single generic ReAct-style agent shared across all tasks. At each step, the model is prompted to emit exactly one JSON object containing either a tool action or a final answer.

System prompt.

You are a ReAct-style research agent.
 Use the available tools when needed and respond with JSON only.
 Use calculator only for a single concrete numeric Python expression.
 Do not send variables, assignments, equations, loops, comprehensions, or prose to calculator.
 Use ** for powers, not ^. If algebra is needed, do it in reasoning and calculate only numeric subexpressions.
 If a tool call fails, do not retry the same input; either use a simpler valid input or reason manually.
 Preferred response format A:

```
{ "thought": "<reasoning>", "action":
{ "tool_name": "<tool>", "tool_input":
"<input>" }}
```

 Also accepted response format A2:

```
{ "thought": "<reasoning>", "action":
{ "calculator": "<expression>" }}
```

 Also accepted response format A3:

```
{ "thought": "<reasoning>", "action":
{ "wiki_search": "<query>" }}
```

 Valid response format B:

```
{ "thought": "<reasoning>",
"final_answer": "<answer>" }
```

 Return exactly one JSON object per turn.
 Do not include Markdown fences or extra commentary.
 Available tools:
 - calculator: Evaluate one short numeric Python expression.
 - wiki_search: Search Wikipedia and return a short summary.

User prompt. For each example, we use the following user prompt:

Solve the problem by reasoning step by step.
 If external lookup or calculation is needed, use a tool.
 Give final_answer as only the final answer, not a full explanation.
 Question: {question}

For MATH, {question} is the original problem text. For HotpotQA no-context, {question} is the original question only. For FEVER, {question} is formatted as:

Claim: {claim}

Decide whether the claim is SUPPORTS, REFUTES, or NOT ENOUGH INFO.

A.2 Cluster-labeling for qualitative analysis

We label each learned cluster with an LLM judge. For each cluster, the judge receives the trace type, cluster ID, total number of clusters, and a sample of reasoning steps assigned to that cluster.

System prompt.

You are an expert in analyzing chain-of-thought (CoT) reasoning produced by large language models. Your task is to identify the semantic role of a group of reasoning steps that have been clustered together based on their latent representations.

You will receive:

1. The trace type: either “static_cot” (numbered step-by-step reasoning) or “react” (thought / action / observation / final_answer interleaved).
2. The cluster ID and total number of clusters K.
3. A sample of reasoning step texts assigned to this cluster.

Respond with a single JSON object (no markdown, no extra text) with exactly these keys:

“label”: short name for the cluster’s semantic role (2–5 words)
“description”: one sentence describing what steps in this cluster do
“rank_in_cot”: integer 1..K — your estimate of this cluster’s ordinal position in the CoT trajectory (1=earliest, K=latest), inferred from the semantic content of the steps alone
“is_coherent”: true if the sampled steps share a clear, consistent semantic role; false if the steps feel like a mixed bag with no single dominant function
“coherence_note”: one sentence explaining any incoherence, or null if the cluster is coherent

User prompt. For each cluster, we use the following user prompt:

Trace type: {trace_type}
Cluster ID: {cluster_id} (K = {k} total clusters)

Sample reasoning steps assigned to this cluster:

[1] {sample_text_1}
[2] {sample_text_2}
...
[n] {sample_text_n}

Respond with the JSON object only.

B Full Quantitative Results

B.1 Trace Statistics

Table 2 summarizes the generated ReAct trajectories across datasets and models, including success rate, average trace length, average number of actions, and the number of traces usable by all compared abstractions.

B.2 Full Results and Paired comparisons

We report the detailed results for both clustering diagnostics and held-out trajectory correctness scoring. Table 3 summarizes the clustering geometry and outcome-conditioned transition separation, while Table 4 reports held-out correctness-scoring performance for FLAT, MERGED-AO, RAW ROUND-CONCAT, and MULTI-VIEW-RA. We additionally report paired split-level comparisons between MULTI-VIEW-RA and the three baselines in Table 5 to assess the consistency.

C Human Evaluation of Failure Localization

C.1 Human Evaluation Protocol

We evaluate whether transitions assigned strongly negative MULTI-VIEW-RA reward correspond to failures that are recognizable to human reviewers. We used HotpotQA/Qwen2.5-32B and sampled 54 transitions: 40 selected by the most-negative MULTI-VIEW-RA reward, together with positive-reward and random-reward controls. Two annotators labelled all items independently.

For each item, annotators are shown the question, final outcome, one preceding round, and the marked round pair with its full thought, action, and observation text. Annotators do not see the gold answer, transition reward, or reason the transition was selected. They first judge whether a specific failure is recognizable at the marked locus (*yes*, *no*, or *unclear*). When a failure is recognizable, the root cause is assigned according to categories as in Table 6.

We adapt the failure categories from prior error analyses of ReAct-style trajectories on HotpotQA (Yao et al., 2023; Gao et al., 2024). Yao et al. (2023) distinguish retrieval, reasoning, repetition, and label ambiguity errors, while Gao et al. (2024) provide finer distinctions among retrieval, decomposition, and selection failures. The final protocol uses five categories: retrieval failure, repetition or looping, wrong answer type or question

Dataset	Model	Traces	Correct (%)	Avg. Steps	Avg. Actions	Common IDs
MATH	Qwen2.5-32B	1000	36.2	7.95	2.07	825
	Qwen2.5-7B	1000	30.9	7.83	2.03	832
	Llama-3.1-8B	1000	11.7	14.53	4.43	953
	Llama-3.2-3B	1000	13.0	10.23	3.04	732
	Gemma-2-2B	1000	0.2	15.86	4.99	845
HotpotQA	Qwen2.5-32B	1000	40.2	10.03	2.73	984
	Qwen2.5-7B	1000	33.0	8.25	2.13	833
	Llama-3.1-8B	1000	20.2	13.81	4.11	990
	Llama-3.2-3B	1000	19.1	10.01	2.75	940
	Gemma-2-2B	1000	0.3	18.49	5.83	978
FEVER	Qwen2.5-32B	1000	72.0	6.42	1.48	987
	Qwen2.5-7B	1000	62.1	5.41	1.16	654
	Llama-3.1-8B	1000	56.7	10.98	3.06	1,000
	Llama-3.2-3B	1000	40.8	10.79	2.99	978
	Gemma-2-2B	1000	0.1	18.95	5.98	999

Table 2: ReAct trace statistics. Common IDs = sample IDs usable under all four abstractions (≥ 2 transitions). **Caveat:** Gemma-2-2B solves almost no MATH/HotpotQA/FEVER traces (0.2%/0.3%/0.1%), so those rows are diagnostic, not balanced comparisons.

misread, selection error, and other. We report inter-annotator reliability using Cohen’s κ , following the inter-annotator validation practice of [Cemri et al. \(2025\)](#).

C.2 Human Evaluation Results

The annotators reach Cohen’s $\kappa = 0.71$ on the recognizable-failure judgment. Each annotator labels 36 of the 40 most-negative-reward transitions (90%) as recognizable failures. Retrieval failure is the dominant category among the flagged transitions; the remaining recognized failures fall into repetition or looping, wrong answer type or question misread, selection error, and other. Under the prespecified category-to-intervention mapping, the dominant retrieval failures correspond to concrete retrieval-side interventions such as query reformulation, retry policies, disambiguation, and reranking.

C.3 Example failure localization

The flagged transitions frequently localize retrieval failures. The following examples illustrate two forms covered by the retrieval-failure category.

Retrieval failure: On-topic but non-answering For a question asking when the English actress who starred in *London Road* was born, the agent identifies Olivia Colman and searches directly for her birth year:

Round 4, action: Olivia Colman birth year
 Observation: Olivia Colman: ... is an English actress noted for her versatility across both comedic and dramatic roles ...

[flagged] Round 5, action: Olivia Colman biography
 Observation: The Favourite (2018 film): ... stars Olivia Colman, Emma Stone, and Rachel Weisz ...

Both retrievals concern the correct subject, but neither provides the requested birth date. The agent subsequently reaches the step limit without an answer. This is a retrieval failure: the search returns relevant or related material without the information required to complete the task. A corresponding intervention is to detect non-answering retrievals and retry with a different query or retrieval strategy.

Retrieval failure: Wrong-entity For a question asking the height of the skyscraper hosting the WXNY-FM transmitter, the agent searches for a different transmitter location:

Round 4, action: WXNY-FM transmitter location
 Observation: WNYC-FM (93.9 MHz) is a non-commercial public radio station licensed to New York, New York ...

[flagged] Round 5, thought: The wiki_searches have not provided the specific information about the skyscraper where WXNY-FM has a transmitter. I need more specific information to proceed.
 Final: Insufficient information to determine the height of the skyscraper where WXNY-FM has a transmitter.

The retrieval substitutes WNYC-FM for WXNY-FM. The agent recognizes that the result does not answer the question and does not use it to produce a factual answer, but it is unable to recover before terminating. This is also a retrieval failure,

here caused by an entity mismatch. Relevant interventions would include entity-aware reranking or retrying when the retrieved entity does not match the query.

Repetition / looping For a question asking the birth year of the head coach of the 1992 Chicago Bears, the agent repeats the same search after the first result does not provide the needed information:

Round 3, action: 1992 Chicago Bears head coach

Observation: List of Chicago Bears head coaches: There have been 18 head coaches for the Chicago Bears ...

[flagged] Round 4, action: 1992 Chicago Bears head coach

Observation: List of Chicago Bears head coaches: There have been 18 head coaches for the Chicago Bears ...

The agent repeats the same action and receives the same non-answering result, without changing its search strategy. This is a repetition / looping failure: the repetition is in the agent’s own action, rather than only in the retriever’s output. A corresponding intervention is a stopping rule or duplicate-action check that prevents retrying the same query without modification.

D Qualitative Analysis of Learned States

D.1 LLM-as-a-judge Analysis

To make the learned states inspectable, we label each cluster using an LLM judge after clustering is complete. The judge receives representative segment texts from each cluster and returns a short functional label, a one-sentence description, an approximate trajectory rank, and the cluster’s average position. Tables 7 - 9 show these labels for Qwen2.5-7B on MATH, HotpotQA, and FEVER, respectively, under all three abstractions. The labels are used only for qualitative interpretation; they are not used when fitting transition rewards or computing any clustering or PRM metric.

Across datasets, the judge labels show that FLAT and MERGED-AO mostly recover local step functions, such as calculation attempts, evidence lookup, or answer extraction. In contrast, MULTI-VIEW-RA clusters summarize round-level behavior, grouping reasoning, action, and observation patterns into process states that can be connected back to the transition analyses in Appendix D.2.



Figure 5: Outcome-conditioned transition heatmaps for MATH / Qwen2.5-7B across FLAT, MERGED-AO, and MULTI-VIEW-RA. Each panel shows the difference between correct-trajectory and incorrect-trajectory transition probabilities, $\Delta P = P_{\text{correct}} - P_{\text{incorrect}}$. “multiview” denotes MULTI-VIEW-RA.



Figure 6: Outcome-conditioned transition heatmaps for MATH / Llama-3.1-8B across FLAT, MERGED-AO, and MULTI-VIEW-RA. Each panel shows the difference between correct-trajectory and incorrect-trajectory transition probabilities, $\Delta P = P_{\text{correct}} - P_{\text{incorrect}}$. “multiview” denotes MULTI-VIEW-RA.

D.2 Additional Transition and State Visualizations

We include additional visual diagnostics for MATH with Qwen2.5-7B and Llama-3.1-8B. Figures 5 and 6 show outcome-conditioned transition differences between correct and incorrect traces. Figures 7 and 8 show the geometry of the clustered state representations. These visualizations provide complementary qualitative evidence that MULTI-VIEW-RA recovers distinct round-level structure with clearer outcome-conditioned transition patterns than the step-level abstractions.

E Operational and Downstream Evaluation

E.1 Failure Triage Simulation

For each of the five inductive 70/30 splits, the abstraction and transition reward are fit using only the training portion, and triage is evaluated on the held-out trajectories. Each held-out trajectory is assigned its average transition reward, and trajectories are queued for review from lowest to highest reward. The trace-length baseline instead queues the longest trajectories first, while random routing

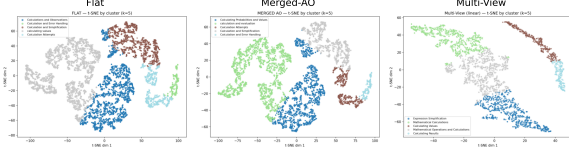


Figure 7: t-SNE projections for MATH / Qwen2.5-7B across FLAT, MERGED-AO, and MULTI-VIEW-RA. Points correspond to clustered states or rounds, colored by cluster assignment. “Multi-View” denotes MULTI-VIEW-RA.

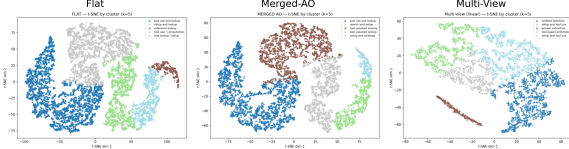


Figure 8: t-SNE projections for MATH / Llama-3.1-8B across FLAT, MERGED-AO, and MULTI-VIEW-RA. Points correspond to clustered states or rounds, colored by cluster assignment. “Multi-View” denotes MULTI-VIEW-RA.

uses a uniformly random ordering.

For a review budget b , we admit the first $b \cdot N$ held-out trajectories to the review queue. Failure recall is the fraction of all failed held-out trajectories contained in this queue. Wasted review effort is the number of correct trajectories in the queue per 100 held-out trajectories. Curves are averaged across the five splits. For rankers with tied scores, ties are broken randomly and the resulting curves are averaged over 20 tie-breakings.

The oracle ranking places all failed trajectories before all correct trajectories and therefore represents the best possible triage ordering given the observed failure base rate. If the failure rate is p , its failure recall at review budget b is $\min(b/p, 1)$. It incurs no wasted review effort until the review budget exceeds p , after which all remaining queued trajectories must be correct. The oracle is included as a ceiling on ranking performance, rather than as a deployable baseline.

E.2 Best-of- n Candidate Ranking

To connect the transition reward to downstream task performance beyond held-out trajectory ROC-AUC, we run a question-disjoint best-of- n re-ranking experiment on HotpotQA/Qwen2.5-32B. Unlike the main trace-generation experiments, which use deterministic decoding, candidate trajectories in this experiment are generated with stochastic decoding using temperature 0.7, top- p 0.95, and

sampling enabled. For 200 held-out questions, we generate eight candidate trajectories per question and rank them by average MULTI-VIEW-RA transition reward. These questions are excluded from all fitting, and all learned components are refit on the remaining 800 source questions.

We evaluate $n \in \{2, 4, 8\}$ against random selection, fixed-first, answer-level self-consistency, and supervised trajectory-level scorers. Results are reported in Table 10, with paired question-level bootstrap tests in Table 11.

At $n = 8$, MULTI-VIEW-RA improves top-1 accuracy over random selection by +0.027 (0.461 vs. 0.434), although the comparison does not survive Holm correction. At $n = 2$ and $n = 4$, the gains over random are +0.022 and +0.029, respectively, and both remain significant after correction. Answer-level self-consistency is the strongest deployable selector.

Overall, the transition reward provides modest same-question candidate-ranking utility. We include this result to make the connection to end-task performance concrete, not to claim that MULTI-VIEW-RA is the strongest re-ranking signal or that it improves the underlying agent policy.

E.3 Operational Cost

We measure post-hoc processing on a single Intel Xeon CPU core at 2.20 GHz with 13.6 GB RAM and no accelerator. The analysis requires no additional model invocation. Computing the cumulative-Gram spectral embedding from final-layer hidden states takes 88 ms per trace; this representation is shared by FLAT, MERGED-AO, RAW ROUND-CONCAT and MULTI-VIEW-RA. The additional MULTI-VIEW-RA cost is 0.55 ms per trace, consisting of 0.48 ms for round-state assignment and 0.07 ms for transition scoring. Fitting the abstraction is a one-time offline step taking 0.7 s with 7 MB peak memory over 700 training traces. Persisted state requires 2.8 KB per trace, approximately $1,580\times$ smaller than the 4.46 MB of raw hidden states. The raw hidden states are transient by-products of generation and need not be retained after post-hoc processing. These measurements do not include the substantially larger cost of generating the agent trajectories.

F Additional Evaluation, Ablation and Notes

F.1 Transductive Evaluation

Outcome labels never enter the unsupervised state abstraction: standardization, kernel construction, kernel PCA, and k -means clustering are fit without reference to correctness, and transition rewards are estimated only from training-split traces. In the primary evaluation (Tables 3 and 4), these components are fit on the training split only for each outer split.

For comparison, we additionally report a transductive variant in which the unlabeled state abstraction is fit once on the full corpus, while transition rewards are still estimated using training-split traces only. Thus, this setting is transductive with respect to the state abstraction, but never with respect to outcome labels. Tables 12 and 13 report the corresponding clustering diagnostics and held-out correctness-scoring results.

The qualitative conclusions are consistent with the primary inductive evaluation: MULTI-VIEW-RA retains stronger outcome-conditioned transition separation and higher correctness-scoring AUC than FLAT MERGED-AO and RAW ROUND-CONCAT across 12 dataset-model conditions. This suggests that fitting the unsupervised abstraction on the full corpus is not what drives the observed advantage.

F.2 Gemma-2-2B Results

Gemma-2-2B solves only 2/1000 MATH, 3/1000 HotpotQA, and 1/1000 FEVER traces (Table 2). With so few positive examples, success-conditioned transition statistics cannot be estimated reliably. We therefore treat the Gemma-2-2B results as diagnostic rather than as balanced comparisons among abstractions.

This class imbalance also prevents a stable fully inductive evaluation: after splitting the data, individual training sets may contain only zero or one correct trace. We therefore omit Gemma-2-2B from the main 12-condition inductive comparison and report it only in the transductive diagnostic tables (Tables 14 and 15). In this setting, MULTI-VIEW-RA is weaker than the step-level baselines on MATH, strongest on HotpotQA, and obtains a high FEVER AUC; however, the FEVER result is based on only one evaluable split and therefore has a degenerate confidence interval (Table 15). These results should not be interpreted as evidence for

or against comparative performance, but instead illustrate the need for sufficiently many examples of both outcomes to estimate the transition reward reliably.

F.3 State - Delta Complementarity

The state and delta views are related by construction, since $E_{\text{end}}(r_t) = \mathbf{v}_t^{(s)} + \mathbf{v}_t^{(\delta)}$. We measure their empirical similarity across 12 dataset - model conditions using two complementary diagnostics: the mean per-round cosine similarity and linear centered kernel alignment (CKA) between the pre-fusion view matrices.

As shown in Table 16, mean cosine similarity ranges from -0.553 to 0.289 . Linear CKA is at most 0.110 in 11 of the 12 conditions, with one higher-similarity condition (FEVER/Qwen2.5-32B, $\text{CKA} = 0.309$). These results indicate that the two representations are not orthogonal, but generally capture empirically distinct structure. The matched representation ablation in Appendix F.5 further shows that their predictive contributions are asymmetric: the delta view carries much of the correctness signal, while the state view provides additional information in some conditions.

F.4 Alternative Trajectory-Level Correctness Scorers

We compare MULTI-VIEW-RA with discriminative trajectory-level correctness scorers trained on alternative trace representations under the same fully inductive splits (Table 17). All supervised baselines use logistic regression (LR). Basic Pre-final Surface LR uses round and action counts, generated token counts, and observation-length statistics. Operational Pre-final LR additionally includes observable runtime diagnostics such as tool errors, invalid syntax, empty or no-result observations, repeated actions, and failure ratios. Full-trace Surface LR further includes final-answer length. We also train LR on normalized $k_{\text{clu}} \times k_{\text{clu}}$ MULTI-VIEW-RA transition counts, mean-pooled hidden-derived segment embeddings, and term frequency-inverse document frequency (TF-IDF) features over generated non-prompt trace text.

As Table 17 shows, global supervised classifiers can sometimes outperform the standalone MULTI-VIEW-RA transition score for final-correctness prediction. In particular, Full-trace Surface LR achieves higher held-out AUC than MULTI-VIEW-RA in 12 dataset-model conditions, with margins ranging from $+0.007$ on HotpotQA/Llama-3.1-8B

to +0.162 on MATH/Llama-3.2-3B. This is expected: MULTI-VIEW-RA is a single scalar derived from unsupervised process states, whereas these classifiers are directly trained on trajectory outcomes and, in the Full-trace variant, additionally exploit final-answer length. Unlike MULTI-VIEW-RA, they operate at the trajectory level and do not produce the per-transition attribution or interpretable process states used for failure diagnosis and triage. We therefore do not view MULTI-VIEW-RA as a replacement for optimized trajectory-level classifiers. Instead, we test whether its transition reward provides complementary information beyond readily observable trajectory statistics. Adding the scalar MULTI-VIEW-RA reward to Operational Pre-final LR improves mean AUC in all 12 conditions (Table 18), with larger gains on MATH and HotpotQA and only small gains on FEVER. Paired split-level comparisons are reported in Table 19.

F.5 Representation Ablations and Robustness Analyses

We further test the contribution of the proposed round representation and the sensitivity of MULTI-VIEW-RA advantage to implementation choices, terminal-answer information, trace complexity, or easily observable tool-execution failures. Unless otherwise stated, these analyses use the fully inductive evaluation on three representative conditions - MATH/Qwen2.5-7B, MATH/Llama-3.1-8B, and HotpotQA/Qwen2.5-32B. All learned abstraction components are fit on training traces only.

Matched Representation We compare alternative round representations while keeping the downstream pipeline fixed (Table 20). Delta is the strongest single-view representation. Adding state information has little effect on MATH/Qwen2.5-7B, but improves AUC on MATH/Llama-3.1-8B and HotpotQA/Qwen2.5-32B, indicating that the two views are complementary in some settings.

Number of clusters We vary the number of latent states over $k \in \{5, 10, 15\}$ (Table 21). MULTI-VIEW-RA achieves the highest held-out AUC in 8 of the 9 comparisons. The only exception is MATH/Llama-3.1-8B at $k = 15$, where FLAT slightly outperforms MULTI-VIEW-RA (0.800 vs. 0.792). Overall, the results are not specific to the default choice $k = 5$.

Transition smoothing. We vary the Laplace smoothing coefficient over

$\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ (Table 22). MULTI-VIEW-RA remains the strongest abstraction across all tested values, with its AUC varying by at most 0.006 across the three conditions. Thus, the results are not sensitive to the choice of smoothing coefficient.

Kernel and projection choices We vary the kernel family and spectral construction while keeping the state - delta representation fixed (Table 23). Alternative kernel choices yield inconsistent changes, and no variant improves uniformly across the three conditions. Direct clustering of the standardized state - delta concatenation also gives similar AUC, suggesting that the result is not sensitive to the specific kernel or projection choice.

Sensitivity to final answer information We remove final-answer segments before reclustering while retaining the preceding terminal thought (Table 24). MULTI-VIEW-RA AUC decreases in all three conditions, but remains the strongest abstraction for MATH/Llama-3.1-8B and HotpotQA/Qwen2.5-32B. This indicates that terminal-answer information contributes to the signal but does not fully account for the observed separation. On MATH/Qwen2.5-7B, the ablated MULTI-VIEW-RA AUC (0.662, 95% CI [0.636, 0.688]) overlaps substantially with FLAT (0.684, [0.675, 0.694]) and MERGED-AO (0.682, [0.664, 0.699]), so we do not claim an advantage over the step-level baselines on this condition after ablation.

Trace length We stratify held-out traces into length tertiles using thresholds derived from the training split (Table 25). MULTI-VIEW-RA achieves the highest AUC in all nine length strata, with the largest margins generally occurring on long traces. This indicates that its advantage is not explained solely by differences in trace length.

Action count We stratify held-out traces by pre-final action count using thresholds derived from the training split (Table 26). MULTI-VIEW-RA shows its largest gains in the high-action strata, reaching +0.259 AUC on MATH/Qwen2.5-7B and +0.358 on HotpotQA/Qwen2.5-32B. Some MATH strata are small or highly imbalanced; in particular, the high-action MATH/Llama-3.1-8B stratum contains no correct examples, so its AUC is undefined.

Tool-Error Analysis We separately evaluate traces with and without explicit pre-final tool errors (Table 27). MULTI-VIEW-RA remains strongest

on the no-error subsets for MATH/Llama-3.1-8B and HotpotQA/Qwen2.5-32B, while the step-level baselines are stronger for MATH/Qwen2.5-7B. This suggests that the overall MULTI-VIEW-RA advantage is not driven solely by obvious tool-execution failures.

F.6 Scope of the Trajectory-Level Correctness Scoring Claim

We use “reward modeling” in this paper to describe trajectory-level correctness scoring via the transition log-likelihood-ratio reward $\bar{R}$ (Sec. 6.2): the reward correlates with trajectory correctness and can rank or threshold held-out trajectories. We do not train a process reward model on top of $\bar{R}$, use it to guide search or policy improvement, or show that it transfers across tasks or models. Readers should treat MULTI-VIEW-RA’s reward-modeling contribution as a transition-level correctness signal that is a candidate source of process supervision, rather than as an evaluated process reward model; connecting $\bar{R}$ to policy improvement or transferable reward models is a direction for future work.

F.7 Dataset Regime Selection Rationale

MATH, HotpotQA (no-context), and FEVER were selected to span distinct reasoning regimes rather than to maximize surface-domain coverage, following the regime-based dataset selection principle used in prior explainability work on agentic traces (Yu et al., 2025): MATH exercises symbolic and numerical reasoning, HotpotQA (no-context) exercises multi-hop reasoning under retrieval, and FEVER exercises claim verification. We have not evaluated MULTI-VIEW-RA on additional domains beyond these three regimes in this work.

F.8 Protocol Requirements Beyond REACT

MULTI-VIEW-RA is evaluated on a single, widely instrumented protocol, the interleaved reasoning–action–observation loop of REACT (Yao et al., 2023; Sec. 3), so that the round-level abstraction could be characterized in depth rather than validated shallowly across many protocols. The abstraction itself does not depend on the specific JSON output format used in our prompts (Appendix A). What it requires is only that a trace decompose into rounds with an identifiable start and end (Sec. 3), so that the state view $\mathbf{v}_t^{(s)}$ and delta view $\mathbf{v}_t^{(\delta)}$ are well defined. Any protocol satisfying this condition is compatible in principle, including native tool-calling APIs and plan–execute–observe

loops; the output format only determines how the round boundary is parsed from the trace. We have not evaluated MULTI-VIEW-RA on protocols other than REACT in this work, and we flag this cross-protocol generalization as untested.

F.9 Algorithm for the End-to-End Pipeline

Algorithm 1 specifies the end-to-end pipeline, including the 128-dimensional projection procedure, k -means initialization and seeding, and Laplace smoothing of the transition matrices.

Algorithm 1 MULTI-VIEW-RA

Require: Training traces with segment hidden states $\{X_t\}$ and outcome labels $\ell \in \{+, -\}$; a held-out trace $r_1, \dots, r_T$

Require: Hyperparameters: projected dimension $d'=128$; retained eigenvalues $k_{\text{eig}}=64$; fusion weight $\beta=0.5$; retained spectral dimensions $m=32$; clusters $k_{\text{clu}}=5$; k -means++ with $n_{\text{init}}=10$; Laplace smoothing $\alpha=10^{-3}$; seed 42 for the random projection and k -means initialization; split seeds 1–5

Stage 1: Segment projection and spectral embedding

- 1: Draw a fixed Gaussian random projection $\Pi \in \mathbb{R}^{d \times d'}$ with seed 42; scale entries by $1/\sqrt{d'}$
 - 2: $G_0 \leftarrow \mathbf{0}$
 - 3: **for** each segment $t = 1, 2, \dots$ in trace order **do**
 - 4: $\tilde{X}_t \leftarrow X_t \Pi$ ▷ project to $d'=128$ dimensions
 - 5: $G_t \leftarrow G_{t-1} + \tilde{X}_t^\top \tilde{X}_t$
 - 6: $E_t \leftarrow$ top k_{eig} eigenvalues of G_t
 - 7: **end for**
 - Stage 2: State/delta views, kernel fusion, clustering (fit on training rounds)**
 - 8: **for** each round r_t **do**
 - 9: $v_t^{(s)} \leftarrow E_{\tau_t}$
 - 10: $E_{\text{end}}(r_t) \leftarrow E_{o_t}$ if r_t has an observation, E_{final} if r_t ends in a final answer, else E_{τ_t}
 - 11: $v_t^{(\delta)} \leftarrow E_{\text{end}}(r_t) - E_{\tau_t}$
 - 12: **end for**
 - 13: Apply elementwise $\log(1 + |\cdot|)$ compression to both views
 - 14: Standardize $v^{(s)}$ and $v^{(\delta)}$ independently per dimension (statistics fit on training rounds)
 - 15: Build linear kernels $K^{(s)}, K^{(\delta)}$; fuse $K_{\text{fused}} \leftarrow \beta K^{(s)} + (1 - \beta) K^{(\delta)}$
 - 16: Eigendecompose K_{fused} and form Z by retaining top m components scaled by $\sqrt{\lambda_i}$
 - 17: Fit k -means on Z with k -means++ initialization ($n_{\text{init}}=10$, seed 42); store centroids and assignments $z_t \in \{1, \dots, k_{\text{clu}}\}$
 - Stage 3: Outcome-conditioned transitions (fit on training traces only)**
 - 18: **for** $\ell \in \{+, -\}$ **do**
 - 19: $C_{ij}^\ell \leftarrow$ count of $z_t=i \rightarrow z_{t+1}=j$ transitions in traces with outcome ℓ
 - 20: $P_{ij}^\ell \leftarrow (C_{ij}^\ell + \alpha) / (\sum_{j'} C_{ij'}^\ell + k_{\text{clu}}\alpha)$ ▷ Laplace smoothing
 - 21: **end for**
 - 22: $r(i, j) \leftarrow \log(P_{ij}^+ / P_{ij}^-)$
 - Stage 4: Score / diagnose a held-out trace**
 - 23: Apply the training-fit projection, standardization, and fused-kernel eigenbasis to the held-out trace's rounds; assign each round to its nearest training-fit centroid to obtain $z_1, \dots, z_T$
 - 24: $\bar{R} \leftarrow \frac{1}{T-1} \sum_{t=1}^{T-1} r(z_t, z_{t+1})$ ▷ trajectory correctness score
 - 25: Rank transitions (z_t, z_{t+1}) by $r(z_t, z_{t+1})$ ▷ for failure diagnosis
 - 26: **return** $\bar{R}$ and the ranked transition list
-

Dataset	Model	Method	Silhouette	LLR-AUROC↑	Cohen's d ↑	Units
MATH	Qwen2.5-7B	FLAT	0.497	0.698	0.529	2,374
		MERGED-AO	0.497	0.705	0.569	1,757
		RAW ROUND-CONCAT	0.410	0.643	0.481	840
		MULTI-VIEW-RA	0.449	0.751	0.876	840
	Qwen2.5-32B	FLAT	0.479	0.573	0.396	2,433
		MERGED-AO	0.476	0.594	0.412	1,796
		RAW ROUND-CONCAT	0.402	0.568	0.408	858
		MULTI-VIEW-RA	0.436	0.780	1.068	858
	Llama-3.1-8B	FLAT	0.505	0.766	0.852	4,371
		MERGED-AO	0.508	0.773	0.879	3,038
		RAW ROUND-CONCAT	0.409	0.722	0.573	1,404
		MULTI-VIEW-RA	0.395	0.913	1.806	1,404
	Llama-3.2-3B	FLAT	0.514	0.630	0.377	3,050
		MERGED-AO	0.514	0.625	0.378	2,143
		RAW ROUND-CONCAT	0.435	0.628	0.312	936
		MULTI-VIEW-RA	0.338	0.764	0.906	936
HotpotQA	Qwen2.5-7B	FLAT	0.539	0.650	0.520	2,448
		MERGED-AO	0.545	0.653	0.376	1,820
		RAW ROUND-CONCAT	0.412	0.573	0.208	892
		MULTI-VIEW-RA	0.534	0.699	0.731	892
	Qwen2.5-32B	FLAT	0.521	0.604	0.414	2,971
		MERGED-AO	0.521	0.644	0.570	2,164
		RAW ROUND-CONCAT	0.411	0.604	0.185	1,057
		MULTI-VIEW-RA	0.546	0.765	1.006	1,057
	Llama-3.1-8B	FLAT	0.529	0.676	0.542	4,200
		MERGED-AO	0.533	0.687	0.595	2,948
		RAW ROUND-CONCAT	0.413	0.553	0.107	1,396
		MULTI-VIEW-RA	0.460	0.847	1.507	1,396
	Llama-3.2-3B	FLAT	0.532	0.513	0.210	2,978
		MERGED-AO	0.544	0.524	0.242	2,162
		RAW ROUND-CONCAT	0.409	0.495	-0.026	1,047
		MULTI-VIEW-RA	0.505	0.661	0.580	1,047
FEVER	Qwen2.5-7B	FLAT	0.580	0.667	0.777	1,613
		MERGED-AO	0.581	0.670	0.794	1,268
		RAW ROUND-CONCAT	0.412	0.564	0.192	622
		MULTI-VIEW-RA	0.623	0.697	0.995	622
	Qwen2.5-32B	FLAT	0.575	0.637	0.532	1,915
		MERGED-AO	0.538	0.632	0.597	1,475
		RAW ROUND-CONCAT	0.413	0.647	0.483	735
		MULTI-VIEW-RA	0.675	0.678	0.802	735
	Llama-3.1-8B	FLAT	0.512	0.573	0.177	3,288
		MERGED-AO	0.524	0.591	0.309	2,372
		RAW ROUND-CONCAT	0.420	0.576	0.204	1,156
		MULTI-VIEW-RA	0.574	0.773	1.280	1,156
	Llama-3.2-3B	FLAT	0.515	0.536	0.106	3,211
		MERGED-AO	0.527	0.534	0.139	2,321
		RAW ROUND-CONCAT	0.418	0.532	0.112	1,131
		MULTI-VIEW-RA	0.513	0.633	0.541	1,131

Table 3: Clustering diagnostics at $k=5$. Units = held-out clustered steps (FLAT and MERGED-AO) or aligned rounds (RAW ROUND-CONCAT and MULTI-VIEW-RA). LLR-AUROC is the area under the ROC curve obtained from transition log-likelihood-ratio (LLR) scores. Best per dataset-model block in **bold**; silhouette measures geometric compactness, while LLR-AUROC and Cohen's d measure outcome separation on held-out traces.

Dataset	Model	Method	AUC↑	95% CI	Cohen's d ↑	Reward Gap↑
MATH	Qwen2.5-7B	FLAT	0.690	[0.675, 0.704]	0.555	0.085
		MERGED-AO	0.696	[0.682, 0.711]	0.608	0.139
		RAW ROUND-CONCAT	0.596	[0.579, 0.614]	0.432	0.201
		MULTI-VIEW-RA	0.748	[0.727, 0.769]	0.857	0.791
	Qwen2.5-32B	FLAT	0.614	[0.591, 0.638]	0.418	0.064
		MERGED-AO	0.632	[0.604, 0.659]	0.392	0.118
		RAW ROUND-CONCAT	0.594	[0.565, 0.622]	0.343	0.104
		MULTI-VIEW-RA	0.776	[0.760, 0.791]	1.023	0.742
	Llama-3.1-8B	FLAT	0.792	[0.761, 0.822]	1.033	0.096
		MERGED-AO	0.808	[0.786, 0.831]	1.028	0.153
		RAW ROUND-CONCAT	0.703	[0.682, 0.723]	0.591	0.243
		MULTI-VIEW-RA	0.910	[0.894, 0.927]	1.827	1.402
	Llama-3.2-3B	FLAT	0.635	[0.615, 0.655]	0.391	0.021
		MERGED-AO	0.645	[0.625, 0.664]	0.397	0.035
		RAW ROUND-CONCAT	0.663	[0.612, 0.715]	0.435	0.079
		MULTI-VIEW-RA	0.741	[0.725, 0.757]	0.854	0.497
HotpotQA	Qwen2.5-7B	FLAT	0.660	[0.632, 0.689]	0.566	0.018
		MERGED-AO	0.662	[0.634, 0.690]	0.379	0.023
		RAW ROUND-CONCAT	0.562	[0.516, 0.608]	0.185	0.030
		MULTI-VIEW-RA	0.701	[0.674, 0.727]	0.779	0.306
	Qwen2.5-32B	FLAT	0.661	[0.635, 0.686]	0.428	0.028
		MERGED-AO	0.680	[0.663, 0.697]	0.636	0.049
		RAW ROUND-CONCAT	0.610	[0.590, 0.629]	0.212	0.039
		MULTI-VIEW-RA	0.757	[0.735, 0.778]	0.885	0.438
	Llama-3.1-8B	FLAT	0.681	[0.649, 0.714]	0.714	0.030
		MERGED-AO	0.692	[0.655, 0.729]	0.774	0.055
		RAW ROUND-CONCAT	0.573	[0.502, 0.645]	0.128	0.014
		MULTI-VIEW-RA	0.853	[0.837, 0.868]	1.588	0.971
	Llama-3.2-3B	FLAT	0.525	[0.503, 0.548]	0.205	0.005
		MERGED-AO	0.553	[0.534, 0.572]	0.288	0.008
		RAW ROUND-CONCAT	0.496	[0.476, 0.516]	-0.034	-0.003
		MULTI-VIEW-RA	0.669	[0.648, 0.689]	0.584	0.212
FEVER	Qwen2.5-7B	FLAT	0.676	[0.661, 0.691]	0.605	0.032
		MERGED-AO	0.678	[0.656, 0.699]	0.685	0.051
		RAW ROUND-CONCAT	0.625	[0.594, 0.655]	0.185	0.053
		MULTI-VIEW-RA	0.692	[0.668, 0.717]	0.863	0.405
	Qwen2.5-32B	FLAT	0.636	[0.596, 0.676]	0.574	0.040
		MERGED-AO	0.638	[0.618, 0.657]	0.584	0.070
		RAW ROUND-CONCAT	0.604	[0.600, 0.609]	0.484	0.236
		MULTI-VIEW-RA	0.682	[0.663, 0.701]	0.680	0.427
	Llama-3.1-8B	FLAT	0.565	[0.545, 0.586]	0.292	0.006
		MERGED-AO	0.585	[0.571, 0.598]	0.431	0.016
		RAW ROUND-CONCAT	0.590	[0.564, 0.616]	0.168	0.018
		MULTI-VIEW-RA	0.771	[0.756, 0.786]	0.985	0.603
	Llama-3.2-3B	FLAT	0.535	[0.516, 0.554]	0.053	0.000
		MERGED-AO	0.529	[0.502, 0.556]	0.098	0.003
		RAW ROUND-CONCAT	0.523	[0.481, 0.565]	0.025	-0.003
		MULTI-VIEW-RA	0.614	[0.592, 0.636]	0.444	0.076

Table 4: Held-out trajectory correctness-scoring results ($k=5$, five stratified 70/30 splits). Values are means over splits with 95% CIs. AUC is computed from the average per-trace transition reward; Cohen's d and Reward Gap measure correct and incorrect reward separation. Best per dataset-model block in **bold**.

Dataset	Model	MV-FLAT			MV-MERGED-AO			MV-RAW ROUND-CONCAT		
		Δ AUC [95% CI]	p_t/p_w	Wins	Δ AUC [95% CI]	p_t/p_w	Wins	Δ AUC [95% CI]	p_t/p_w	Wins
MATH	Qwen2.5-7B	+0.058 [+0.024, +0.092]	0.029/0.062	5/5	+0.052 [+0.019, +0.084]	0.035/0.062	5/5	+0.151 [+0.125, +0.177]	0.000/0.062	5/5
	Qwen2.5-32B	+0.161 [+0.134, +0.189]	0.000/0.062	5/5	+0.144 [+0.110, +0.178]	0.001/0.062	5/5	+0.182 [+0.152, +0.212]	0.000/0.062	5/5
	Llama-3.1-8B	+0.119 [+0.090, +0.147]	0.001/0.062	5/5	+0.102 [+0.081, +0.123]	0.001/0.062	5/5	+0.208 [+0.193, +0.222]	0.000/0.062	5/5
	Llama-3.2-3B	+0.106 [+0.094, +0.118]	0.000/0.062	5/5	+0.096 [+0.080, +0.112]	0.000/0.062	5/5	+0.078 [+0.036, +0.120]	0.022/0.062	5/5
HotpotQA	Qwen2.5-7B	+0.040 [+0.005, +0.075]	0.087/0.062	5/5	+0.039 [+0.009, +0.069]	0.061/0.062	5/5	+0.138 [+0.107, +0.170]	0.001/0.062	5/5
	Qwen2.5-32B	+0.096 [+0.089, +0.104]	0.000/0.062	5/5	+0.077 [+0.067, +0.088]	0.000/0.062	5/5	+0.147 [+0.139, +0.156]	0.000/0.062	5/5
	Llama-3.1-8B	+0.171 [+0.142, +0.200]	0.000/0.062	5/5	+0.160 [+0.129, +0.192]	0.001/0.062	5/5	+0.279 [+0.199, +0.359]	0.002/0.062	5/5
	Llama-3.2-3B	+0.143 [+0.113, +0.174]	0.001/0.062	5/5	+0.115 [+0.089, +0.142]	0.001/0.062	5/5	+0.173 [+0.134, +0.211]	0.001/0.062	5/5
FEVER	Qwen2.5-7B	+0.017 [+0.003, +0.030]	0.076/0.125	4/5	+0.015 [-0.016, +0.045]	0.395/0.438	4/5	+0.068 [+0.014, +0.122]	0.068/0.062	5/5
	Qwen2.5-32B	+0.046 [+0.019, +0.073]	0.029/0.062	5/5	+0.044 [+0.025, +0.064]	0.011/0.062	5/5	+0.078 [+0.060, +0.095]	0.001/0.062	5/5
	Llama-3.1-8B	+0.206 [+0.190, +0.221]	0.000/0.062	5/5	+0.187 [+0.167, +0.207]	0.000/0.062	5/5	+0.181 [+0.143, +0.220]	0.001/0.062	5/5
	Llama-3.2-3B	+0.079 [+0.057, +0.102]	0.002/0.062	5/5	+0.085 [+0.058, +0.112]	0.004/0.062	5/5	+0.091 [+0.048, +0.135]	0.015/0.062	5/5

Table 5: Paired split-level comparisons under the fully inductive evaluation. MV denotes MULTI-VIEW-RA. Δ is the mean paired AUC difference between MULTI-VIEW-RA and each baseline over the same five outer splits, with 95% confidence intervals. p_t/p_w denote paired t -test and two-sided exact Wilcoxon signed-rank p-values. Wins reports the number of outer splits in which MULTI-VIEW-RA achieves higher AUC than the corresponding baseline.

Category	Definition	Intervention
Retrieval failure	The retrieved result is about the wrong entity or passage, or provides no useful information for the query. This includes wrong-entity matches, off-topic passages, empty results, and generic or non-answering results returned across reformulated queries.	Query reformulation, retry, disambiguation, reranking, or result filtering.
Repetition / looping	The agent repeats a thought or action it has already attempted without making progress. Repeated or unhelpful observations do not count as repetition when the agent’s own actions differ across rounds; these are treated as retrieval failures.	Stopping rule or duplicate-action guard.
Wrong answer type / question misread	The agent answers a different question from the one asked or returns the wrong type of answer, such as a name when a year is requested.	Decomposition or planning prompt; answer-type check before termination.
Selection error	The relevant information is present in the shown context, but the agent selects the wrong candidate or conclusion.	Verification before finalization.
Other	A recognizable failure that does not fit one of the categories above.	Case-specific.

Table 6: Failure categories used for human evaluation. Actionability is derived from the prespecified mapping from failure category to intervention

Method	ID	Label	Description	Rank	Avg. Pos.
FLAT	3	calculating values	This cluster contains steps that involve performing calculations to find specific values or results based on mathematical expressions.	3	3.11
FLAT	0	Calculations and Observations	This cluster contains steps that involve performing calculations, making observations, and deriving results based on mathematical reasoning.	3	5.36
FLAT	2	Calculation and Simplification	This cluster contains steps focused on performing calculations, simplifying expressions, and deriving numerical results from mathematical operations.	3	9.74
FLAT	1	Calculation and Error Handling	This cluster contains steps that involve performing calculations, handling errors from those calculations, and attempting to derive results despite tool limitations.	3	15.02
FLAT	4	Calculation Attempts	This cluster contains steps that involve attempts to perform calculations, often resulting in errors or no final answers due to syntax issues or tool limitations.	4	12.15
MERGED-AO	1	calculation and evaluation	This cluster contains steps that involve performing calculations, evaluating expressions, and deriving numerical results based on mathematical operations.	3	2.89
MERGED-AO	3	Calculation and Simplification	This cluster contains steps focused on performing calculations, simplifying expressions, and deriving results from mathematical operations.	3	9.43
MERGED-AO	2	Calculation Attempts	This cluster contains steps that involve attempts to perform calculations or solve equations, often resulting in errors or no final answers.	3	12.08
MERGED-AO	4	Calculation and Error Handling	This cluster contains steps focused on performing calculations, handling errors from those calculations, and deriving values based on mathematical expressions.	3	14.75
MERGED-AO	0	Calculating Probabilities and Values	This cluster contains steps that involve calculating probabilities, values, and mathematical expressions, often leading to final numerical results.	4	5.10
MULTI-VIEW-RA	1	Mathematical Calculations	This cluster contains steps that involve performing various mathematical calculations and solving equations.	1	1.13
MULTI-VIEW-RA	0	Expression Simplification	The steps in this cluster focus on simplifying mathematical expressions and equations to derive values or relationships.	2	3.78
MULTI-VIEW-RA	3	Mathematical Operations and Calculations	This cluster contains steps that involve performing various mathematical operations, calculations, and simplifications to solve problems.	3	1.98
MULTI-VIEW-RA	2	Calculating Values	This cluster contains steps that involve calculating specific numerical values or results based on mathematical operations.	3	2.11
MULTI-VIEW-RA	4	Calculating Results	This cluster contains steps that involve calculating and presenting final numerical results based on various mathematical operations.	4	3.41

Table 7: LLM-judge cluster labels for MATH/Qwen2.5-7B at $k=5$. Rank denotes the judge-estimated stage of the reasoning trajectory, while Avg. Pos. is the empirical average position of items assigned to the cluster.

Method	ID	Label	Description	Rank	Avg. Pos.
FLAT	0	Information Extraction	This cluster consists of steps that extract specific information or facts from provided paragraphs.	1	3.27
FLAT	1	Information Retrieval	This cluster focuses on gathering specific information from various sources to answer questions or clarify details.	1	3.86
FLAT	2	Information Extraction	This cluster consists of steps that extract specific information or identify entities based on provided paragraphs.	2	1.83
FLAT	4	Identify and Confirm Information	This cluster consists of steps that focus on identifying, confirming, or deducing specific information based on provided data or external searches.	2	2.51
FLAT	3	Information Extraction	The steps in this cluster focus on extracting specific information from provided texts or paragraphs.	3	2.83
MERGED-AO	1	Information Extraction	This cluster consists of steps that extract specific information or facts from the provided paragraphs.	1	1.72
MERGED-AO	3	Identify First Magazine	The steps in this cluster focus on determining the chronological order of magazine launches based on their start dates.	1	2.34
MERGED-AO	0	Information Extraction	This cluster focuses on extracting specific information or facts from provided paragraphs or texts.	1	2.93
MERGED-AO	4	Information Extraction	The steps in this cluster extract specific pieces of information from provided texts or paragraphs.	2	2.70
MERGED-AO	2	Information Retrieval	This cluster focuses on gathering and confirming specific information from various sources to answer questions.	2	3.49
MULTI-VIEW-RA	0	Information Retrieval	The steps in this cluster focus on extracting specific information or details from provided texts or conducting searches to find relevant data.	1	1.40
MULTI-VIEW-RA	1	Information Extraction	The steps in this cluster extract specific information or facts from provided paragraphs to answer questions.	1	1.67
MULTI-VIEW-RA	2	Identifying Relationships	The steps in this cluster focus on identifying relationships, commonalities, or distinctions between various subjects based on provided information.	2	1.08
MULTI-VIEW-RA	4	Information Retrieval	This cluster consists of steps that focus on extracting specific information or facts from provided paragraphs or conducting searches to find relevant data.	2	1.09
MULTI-VIEW-RA	3	Identify Origins and Relationships	The steps in this cluster focus on identifying the origins, relationships, or comparisons between various entities based on provided information.	2	1.30

Table 8: LLM-judge cluster labels for HotpotQA/Qwen2.5-7B at $k=5$. Rank denotes the judge-estimated stage of the reasoning trajectory, while Avg. Pos. is the empirical average position of items assigned to the cluster.

Method	ID	Label	Description	Rank	Avg. Pos.
FLAT	1	Information Verification	This cluster contains steps focused on verifying claims by searching for factual information or context through external sources.	1	1.95
FLAT	4	Information Gathering	This cluster consists of steps focused on searching for specific information or clarifying details related to various topics.	2	4.57
FLAT	2	Information Gathering	The steps in this cluster involve searching for specific information and verifying details about various topics.	2	16.10
FLAT	0	Claim Evaluation	This cluster contains steps that evaluate claims based on gathered information, determining whether they are supported, refuted, or lack sufficient evidence.	3	9.03
FLAT	3	Information Gathering	This cluster consists of steps focused on searching for specific information and verifying claims through various sources.	3	12.74
MERGED-AO	0	Information Verification	This cluster contains steps focused on verifying claims through external searches for factual information.	1	1.86
MERGED-AO	1	Searching for Information	The steps in this cluster involve attempts to find specific information through searches and acknowledging the lack of results.	1	13.15
MERGED-AO	2	Information Retrieval Attempts	This cluster consists of steps focused on searching for specific information and acknowledging the lack of sufficient data from previous searches.	2	16.22
MERGED-AO	3	Information Gathering	This cluster consists of steps focused on searching for specific information to confirm or refute claims.	3	4.86
MERGED-AO	4	Information Gathering	This cluster consists of steps focused on searching for specific information and assessing the adequacy of the gathered data to support or refute various claims.	3	9.33
MULTI-VIEW-RA	1	Information Verification	This cluster contains steps that assess the validity of various claims based on available information, determining whether they are supported, refuted, or lack sufficient context.	2	1.00
MULTI-VIEW-RA	2	Claim Evaluation	This cluster contains steps that evaluate claims based on supporting or refuting evidence found in searches.	2	2.43
MULTI-VIEW-RA	3	Claim Verification Steps	This cluster contains steps focused on verifying specific claims by searching for factual information or evidence.	3	1.04
MULTI-VIEW-RA	4	Refuting False Claims	The steps in this cluster systematically refute various false claims by providing factual corrections.	3	1.12
MULTI-VIEW-RA	0	Information Retrieval Attempts	This cluster consists of steps where the model attempts to gather specific information through searches but often finds incomplete or irrelevant results.	3	3.37

Table 9: LLM-judge cluster labels for FEVER/Qwen2.5-7B at $k=5$. Rank denotes the judge-estimated stage of the reasoning trajectory, while Avg. Pos. is the empirical average position of items assigned to the cluster.

Method	Best-of-2	Best-of-4	Best-of-8
Random	0.434 [0.373, 0.494]	0.434 [0.373, 0.494]	0.434 [0.373, 0.494]
Fixed-first	0.432 [0.370, 0.493]	0.426 [0.362, 0.488]	0.425 [0.355, 0.495]
Self-consistency	0.463 [0.403, 0.523]	0.488 [0.423, 0.553]	0.495 [0.427, 0.562]
Basic Pre-final LR	0.454 [0.390, 0.517]	0.463 [0.397, 0.527]	0.460 [0.390, 0.530]
Operational Pre-final LR	0.453 [0.389, 0.516]	0.460 [0.394, 0.525]	0.465 [0.395, 0.535]
Transition-count LR	0.462 [0.398, 0.524]	0.472 [0.407, 0.535]	0.470 [0.403, 0.536]
MULTI-VIEW-RA Reward	0.457 [0.392, 0.520]	0.464 [0.397, 0.528]	0.461 [0.393, 0.527]
Operational + MULTI-VIEW-RA	0.461 [0.397, 0.524]	0.470 [0.403, 0.535]	0.470 [0.400, 0.540]
Oracle pass@ n	0.494	0.538	0.575

Table 10: Question-disjoint best-of- n candidate selection on 200 HotpotQA/Qwen2.5-32B questions with eight independently sampled candidates per question (temperature 0.7, top- p 0.95). LR denotes logistic regression. All learned components are fit on the remaining 800 source questions. Values are mean selection accuracy with question-level bootstrap 95% confidence intervals. Oracle pass@ n provides an upper bound.

n	Comparison	Δ Acc.	95% CI	p	Holm p	Favor A/B/Tie
2	MULTI-VIEW-RA – Random	+0.022	[+0.007, +0.036]	0.006	0.017	45/13/142
2	MULTI-VIEW-RA – Self-consistency	−0.006	[−0.036, +0.023]	0.708	1.000	41/31/128
2	Operational+MULTI-VIEW-RA – Operational LR	+0.008	[+0.003, +0.014]	0.004	0.014	22/10/168
2	Operational+MULTI-VIEW-RA – Self-consistency	−0.001	[−0.030, +0.026]	0.956	1.000	42/31/127
4	MULTI-VIEW-RA – Random	+0.029	[+0.011, +0.047]	0.002	0.006	42/16/142
4	MULTI-VIEW-RA – Self-consistency	−0.024	[−0.056, +0.005]	0.113	0.288	20/37/143
4	Operational+MULTI-VIEW-RA – Operational LR	+0.010	[−0.001, +0.022]	0.096	0.288	18/9/173
4	Operational+MULTI-VIEW-RA – Self-consistency	−0.018	[−0.050, +0.013]	0.247	0.288	25/30/145
8	MULTI-VIEW-RA – Random	+0.027	[+0.004, +0.050]	0.022	0.088	38/20/142
8	MULTI-VIEW-RA – Self-consistency	−0.034	[−0.070, +0.001]	0.061	0.183	7/18/175
8	Operational+MULTI-VIEW-RA – Operational LR	+0.005	[−0.020, +0.030]	0.869	0.869	4/3/193
8	Operational+MULTI-VIEW-RA – Self-consistency	−0.025	[−0.065, +0.015]	0.245	0.491	7/15/178

Table 11: Paired question-level bootstrap comparisons for best-of- n re-ranking. Holm correction is applied within each n across the four comparisons. Favor A/B/Tie gives the number of questions favoring the first method, the second method, or neither, respectively.

Dataset	Model	Method	Silhouette $\uparrow$	LLR-AUROC $\uparrow$	Cohen's d $\uparrow$	Units
MATH	Qwen2.5-32B	FLAT	0.484	0.564	0.366	7,952
		MERGED-AO	0.480	0.577	0.419	5,879
		RAW ROUND-CONCAT	0.399	0.582	0.399	2,806
		MULTI-VIEW-RA	0.435	0.774	1.115	2,806
	Qwen2.5-7B	FLAT	0.504	0.687	0.625	7,831
		MERGED-AO	0.507	0.692	0.658	5,802
		RAW ROUND-CONCAT	0.415	0.679	0.568	2,773
		MULTI-VIEW-RA	0.452	0.792	1.101	2,773
	Llama-3.1-8B	FLAT	0.506	0.738	0.854	14,530
		MERGED-AO	0.509	0.745	0.891	10,098
		RAW ROUND-CONCAT	0.413	0.724	0.695	4,666
		MULTI-VIEW-RA	0.397	0.905	1.814	4,666
	Llama-3.2-3B	FLAT	0.510	0.601	0.355	10,229
		MERGED-AO	0.510	0.602	0.371	7,191
		RAW ROUND-CONCAT	0.443	0.619	0.376	3,153
		MULTI-VIEW-RA	0.359	0.798	1.060	3,153
HotpotQA	Qwen2.5-32B	FLAT	0.523	0.621	0.583	10,033
		MERGED-AO	0.526	0.663	0.644	7,299
		RAW ROUND-CONCAT	0.415	0.604	0.354	3,565
		MULTI-VIEW-RA	0.546	0.767	1.062	3,565
	Qwen2.5-7B	FLAT	0.543	0.617	0.430	8,252
		MERGED-AO	0.548	0.627	0.469	6,126
		RAW ROUND-CONCAT	0.427	0.586	0.186	3,000
		MULTI-VIEW-RA	0.536	0.701	0.745	3,000
	Llama-3.1-8B	FLAT	0.527	0.671	0.567	13,811
		MERGED-AO	0.534	0.683	0.619	9,703
		RAW ROUND-CONCAT	0.416	0.579	0.246	4,595
		MULTI-VIEW-RA	0.463	0.854	1.538	4,595
	Llama-3.2-3B	FLAT	0.535	0.552	0.299	10,010
		MERGED-AO	0.544	0.546	0.289	7,259
		RAW ROUND-CONCAT	0.405	0.579	0.206	3,508
		MULTI-VIEW-RA	0.499	0.685	0.656	3,508
FEVER	Qwen2.5-32B	FLAT	0.581	0.660	0.754	6,415
		MERGED-AO	0.542	0.658	0.744	4,937
		RAW ROUND-CONCAT	0.420	0.656	0.517	2,459
		MULTI-VIEW-RA	0.676	0.696	0.977	2,459
	Qwen2.5-7B	FLAT	0.582	0.644	0.594	5,411
		MERGED-AO	0.581	0.647	0.586	4,249
		RAW ROUND-CONCAT	0.502	0.572	0.393	2,087
		MULTI-VIEW-RA	0.632	0.703	1.057	2,087
	Llama-3.1-8B	FLAT	0.514	0.582	0.302	10,984
		MERGED-AO	0.524	0.606	0.385	7,925
		RAW ROUND-CONCAT	0.422	0.608	0.285	3,866
		MULTI-VIEW-RA	0.577	0.773	1.299	3,866
	Llama-3.2-3B	FLAT	0.518	0.574	0.272	10,785
		MERGED-AO	0.530	0.573	0.268	7,792
		RAW ROUND-CONCAT	0.420	0.565	0.269	3,799
		MULTI-VIEW-RA	0.511	0.647	0.610	3,799

Table 12: Clustering diagnostics in the transductive setting at $k=5$ where the unlabeled structural abstraction is fit on the full corpus and outcome-conditioned transition statistics are estimated from training traces. Units = clustered steps (FLAT and MERGED-AO) or aligned rounds (RAW ROUND-CONCAT and MULTI-VIEW-RA). Best per dataset-model block in **bold**; note silhouette often favours flat/merged (more geometric compactness) while LLR-AUROC and Cohen's d favour multi-view (better outcome separation).

Dataset	Model	Method	AUC↑	95% CI	Cohen's d ↑	Reward Gap↑
MATH	Qwen2.5-32B	FLAT	0.581	[0.550, 0.611]	0.285	0.070
		MERGED-AO	0.592	[0.571, 0.614]	0.394	0.142
		RAW ROUND-CONCAT	0.589	[0.569, 0.608]	0.323	0.088
		MULTI-VIEW-RA	0.756	[0.735, 0.778]	0.886	0.745
	Qwen2.5-7B	FLAT	0.690	[0.651, 0.729]	0.508	0.088
		MERGED-AO	0.703	[0.664, 0.743]	0.604	0.143
		RAW ROUND-CONCAT	0.596	[0.558, 0.633]	0.431	0.168
		MULTI-VIEW-RA	0.778	[0.754, 0.801]	1.150	0.843
	Llama-3.1-8B	FLAT	0.768	[0.743, 0.793]	0.861	0.101
		MERGED-AO	0.782	[0.758, 0.807]	0.897	0.160
		RAW ROUND-CONCAT	0.661	[0.609, 0.713]	0.434	0.173
		MULTI-VIEW-RA	0.913	[0.907, 0.918]	1.842	1.403
	Llama-3.2-3B	FLAT	0.557	[0.522, 0.592]	0.223	0.008
		MERGED-AO	0.563	[0.532, 0.594]	0.248	0.019
		RAW ROUND-CONCAT	0.533	[0.477, 0.589]	0.084	0.013
		MULTI-VIEW-RA	0.775	[0.745, 0.804]	1.015	0.582
HotpotQA	Qwen2.5-32B	FLAT	0.634	[0.616, 0.652]	0.533	0.026
		MERGED-AO	0.670	[0.651, 0.688]	0.627	0.053
		RAW ROUND-CONCAT	0.621	[0.583, 0.660]	0.349	0.050
		MULTI-VIEW-RA	0.744	[0.721, 0.767]	0.810	0.422
	Qwen2.5-7B	FLAT	0.646	[0.635, 0.656]	0.517	0.018
		MERGED-AO	0.654	[0.636, 0.672]	0.482	0.030
		RAW ROUND-CONCAT	0.539	[0.497, 0.582]	0.064	0.007
		MULTI-VIEW-RA	0.699	[0.658, 0.740]	0.766	0.303
	Llama-3.1-8B	FLAT	0.661	[0.625, 0.698]	0.684	0.054
		MERGED-AO	0.689	[0.670, 0.707]	0.745	0.086
		RAW ROUND-CONCAT	0.569	[0.521, 0.617]	0.154	0.020
		MULTI-VIEW-RA	0.844	[0.830, 0.857]	1.530	0.969
	Llama-3.2-3B	FLAT	0.525	[0.503, 0.546]	0.251	0.005
		MERGED-AO	0.540	[0.519, 0.561]	0.189	0.006
		RAW ROUND-CONCAT	0.524	[0.485, 0.563]	-0.046	-0.008
		MULTI-VIEW-RA	0.673	[0.664, 0.682]	0.598	0.221
FEVER	Qwen2.5-32B	FLAT	0.662	[0.627, 0.696]	0.640	0.043
		MERGED-AO	0.659	[0.626, 0.692]	0.706	0.071
		RAW ROUND-CONCAT	0.600	[0.587, 0.612]	0.419	0.218
		MULTI-VIEW-RA	0.691	[0.660, 0.721]	0.825	0.422
	Qwen2.5-7B	FLAT	0.671	[0.647, 0.695]	0.595	0.035
		MERGED-AO	0.659	[0.639, 0.679]	0.608	0.048
		RAW ROUND-CONCAT	0.644	[0.628, 0.660]	0.254	0.079
		MULTI-VIEW-RA	0.690	[0.675, 0.706]	0.864	0.428
	Llama-3.1-8B	FLAT	0.578	[0.553, 0.604]	0.210	0.004
		MERGED-AO	0.587	[0.568, 0.606]	0.409	0.015
		RAW ROUND-CONCAT	0.589	[0.556, 0.621]	0.206	0.021
		MULTI-VIEW-RA	0.752	[0.732, 0.773]	0.900	0.564
	Llama-3.2-3B	FLAT	0.581	[0.534, 0.628]	0.248	0.004
		MERGED-AO	0.548	[0.512, 0.583]	0.163	0.006
		RAW ROUND-CONCAT	0.537	[0.494, 0.579]	0.030	0.003
		MULTI-VIEW-RA	0.627	[0.619, 0.635]	0.504	0.079

Table 13: Held-out trajectory correctness-scoring results in the transductive setting where the structural abstraction is fit on the full corpus ($k=5$, five stratified 70/30 splits). Values are means across splits with 95% confidence intervals. AUC is ROC-AUC of the average per-trace transition reward; best per dataset-model block in **bold**.

Dataset	Model	Method	Silhouette	LLR-AUROC↑	Cohen's d ↑	Units
MATH	Gemma-2-2B	FLAT	0.513	0.955	2.295	15,860
		MERGED-AO	0.514	0.957	2.439	10,867
		MULTI-VIEW-RA	0.324	0.783	1.336	4,874
HotpotQA	Gemma-2-2B	FLAT	0.387	0.817	0.734	18,494
		MERGED-AO	0.386	0.869	0.794	12,666
		MULTI-VIEW-RA	0.299	0.980	1.814	5,838
FEVER	Gemma-2-2B	FLAT	0.479	0.999	2.087	18,954
		MERGED-AO	0.486	0.996	2.113	12,971
		MULTI-VIEW-RA	0.378	1.000	2.870	5,988

Table 14: Clustering diagnostics for Gemma-2-2B in the transductive setting where the structural abstraction is fit on the full corpus at $k=5$. Best per dataset-model block in **bold**; Gemma-2-2B produces only one to three correct traces per dataset so these results are treated as diagnostic rather than as balanced comparisons among abstractions.

Dataset	Model	Method	AUC↑	95% CI	Cohen's d ↑	Reward Gap↑
MATH	Gemma-2-2B	FLAT	0.968	[0.949, 0.987]	4.058	1.171
		MERGED-AO	0.963	[0.944, 0.982]	4.238	1.326
		MULTI-VIEW-RA	0.618	[0.359, 0.877]	0.345	0.105
HotpotQA	Gemma-2-2B	FLAT	0.798	[0.621, 0.975]	0.949	0.313
		MERGED-AO	0.798	[0.606, 0.990]	0.991	0.334
		MULTI-VIEW-RA	0.963	[0.944, 0.983]	1.632	0.731
FEVER	Gemma-2-2B	FLAT	0.749	[0.749, 0.749]	1.286	0.070
		MERGED-AO	0.746	[0.746, 0.746]	0.797	0.054
		MULTI-VIEW-RA	0.993	[0.993, 0.993]	6.057	1.770

Table 15: Held-out correctness-scoring results for Gemma-2-2B in the transductive setting where the structural abstraction is fit on the full corpus ($k=5$, five stratified 70/30 splits where evaluable). Best per dataset-model block in **bold**; the FEVER confidence intervals are degenerate because only one split is evaluable.

Dataset	Model	# Rounds	Mean cosine	Linear CKA
MATH	Qwen2.5-7B	2773	-0.015	0.000
	Qwen2.5-32B	2806	-0.095	0.012
	Llama-3.1-8B	4666	0.289	0.035
	Llama-3.2-3B	3153	-0.045	0.007
HotpotQA	Qwen2.5-7B	3000	-0.133	0.054
	Qwen2.5-32B	3565	-0.149	0.072
	Llama-3.1-8B	4595	0.092	0.016
	Llama-3.2-3B	3508	-0.121	0.070
FEVER	Qwen2.5-7B	2087	-0.275	0.050
	Qwen2.5-32B	2459	-0.553	0.309
	Llama-3.1-8B	3866	-0.115	0.110
	Llama-3.2-3B	3799	-0.138	0.091

Table 16: Empirical similarity between the state and delta views across 12 conditions. Rounds are all aligned MULTI-VIEW-RA reasoning rounds. Mean cosine measures signed per-round alignment, while linear CKA measures global similarity between the state and delta view matrices after the same $\log(1 + |x|)$ compression and independent standardization used by MULTI-VIEW-RA.

Dataset	Model	Basic Pre-final	Operational	Full-trace	Transition-count	Pooled-hidden	Text	MULTI-VIEW-RA
MATH	Qwen2.5-7B	0.772 [0.756, 0.789]	0.788 [0.767, 0.809]	0.839 [0.832, 0.846]	0.770 [0.748, 0.793]	0.765 [0.748, 0.781]	0.766 [0.737, 0.796]	0.748 [0.727, 0.769]
	Qwen2.5-32B	0.753 [0.737, 0.768]	0.761 [0.742, 0.781]	0.832 [0.823, 0.841]	0.771 [0.755, 0.786]	0.757 [0.745, 0.768]	0.739 [0.702, 0.777]	0.776 [0.760, 0.791]
	Llama-3.1-8B	0.870 [0.852, 0.889]	0.878 [0.860, 0.896]	0.927 [0.906, 0.949]	0.898 [0.880, 0.916]	0.839 [0.818, 0.860]	0.801 [0.771, 0.831]	0.910 [0.894, 0.927]
	Llama-3.2-3B	0.757 [0.743, 0.770]	0.750 [0.735, 0.764]	0.903 [0.888, 0.919]	0.744 [0.739, 0.749]	0.722 [0.689, 0.755]	0.717 [0.668, 0.765]	0.741 [0.725, 0.757]
HotpotQA	Qwen2.5-7B	0.752 [0.733, 0.771]	0.753 [0.740, 0.766]	0.766 [0.748, 0.784]	0.688 [0.668, 0.709]	0.683 [0.666, 0.700]	0.656 [0.648, 0.664]	0.701 [0.674, 0.727]
	Qwen2.5-32B	0.730 [0.710, 0.749]	0.725 [0.709, 0.740]	0.789 [0.779, 0.799]	0.764 [0.739, 0.790]	0.703 [0.679, 0.727]	0.752 [0.731, 0.773]	0.757 [0.735, 0.778]
	Llama-3.1-8B	0.840 [0.824, 0.855]	0.840 [0.823, 0.857]	0.860 [0.842, 0.877]	0.847 [0.829, 0.864]	0.817 [0.801, 0.833]	0.646 [0.635, 0.658]	0.853 [0.837, 0.868]
	Llama-3.2-3B	0.675 [0.651, 0.700]	0.678 [0.662, 0.693]	0.708 [0.683, 0.732]	0.655 [0.638, 0.671]	0.612 [0.587, 0.636]	0.598 [0.571, 0.624]	0.669 [0.648, 0.689]
FEVER	Qwen2.5-7B	0.779 [0.754, 0.803]	0.770 [0.732, 0.808]	0.767 [0.741, 0.792]	0.693 [0.666, 0.720]	0.722 [0.700, 0.743]	0.800 [0.793, 0.807]	0.692 [0.668, 0.717]
	Qwen2.5-32B	0.749 [0.723, 0.776]	0.745 [0.720, 0.770]	0.744 [0.698, 0.789]	0.676 [0.653, 0.698]	0.714 [0.693, 0.735]	0.807 [0.791, 0.823]	0.682 [0.663, 0.701]
	Llama-3.1-8B	0.804 [0.786, 0.823]	0.804 [0.786, 0.823]	0.811 [0.796, 0.826]	0.771 [0.761, 0.781]	0.759 [0.737, 0.781]	0.687 [0.675, 0.698]	0.771 [0.756, 0.786]
	Llama-3.2-3B	0.626 [0.596, 0.656]	0.634 [0.607, 0.661]	0.640 [0.616, 0.664]	0.623 [0.594, 0.652]	0.561 [0.542, 0.580]	0.672 [0.657, 0.686]	0.614 [0.592, 0.636]

Table 17: Alternative trajectory-level correctness scorers under the fully inductive evaluation. Values are mean held-out ROC-AUC with 95% confidence intervals over five stratified 70/30 splits. All methods are evaluated on the same held-out trace IDs. Best AUC per dataset-model condition is shown in **bold**.

Dataset	Model	Operational Pre-final LR	+ MULTI-VIEW-RA Reward	Δ
MATH	Qwen2.5-7B	0.788 [0.767, 0.809]	0.805 [0.785, 0.825]	+0.016
	Qwen2.5-32B	0.761 [0.742, 0.781]	0.781 [0.764, 0.797]	+0.019
	Llama-3.1-8B	0.878 [0.860, 0.896]	0.894 [0.874, 0.914]	+0.016
	Llama-3.2-3B	0.750 [0.735, 0.764]	0.761 [0.745, 0.776]	+0.011
HotpotQA	Qwen2.5-7B	0.753 [0.740, 0.766]	0.755 [0.736, 0.774]	+0.002
	Qwen2.5-32B	0.725 [0.709, 0.740]	0.771 [0.755, 0.787]	+0.046
	Llama-3.1-8B	0.840 [0.823, 0.857]	0.847 [0.829, 0.866]	+0.007
	Llama-3.2-3B	0.678 [0.662, 0.693]	0.684 [0.671, 0.697]	+0.006
FEVER	Qwen2.5-7B	0.770 [0.732, 0.808]	0.771 [0.729, 0.812]	+0.001
	Qwen2.5-32B	0.745 [0.720, 0.770]	0.745 [0.719, 0.772]	+0.001
	Llama-3.1-8B	0.804 [0.786, 0.823]	0.806 [0.791, 0.820]	+0.001
	Llama-3.2-3B	0.634 [0.607, 0.661]	0.636 [0.608, 0.664]	+0.002

Table 18: Incremental correctness signal from adding the scalar MULTI-VIEW-RA transition reward to Operational Pre-final LR. Values are mean held-out trajectory ROC-AUC with 95% confidence intervals over five identical stratified splits. Training reward features are generated out-of-fold, while held-out rewards use transition matrices fit only on outer-training traces. Δ AUC is the difference relative to Operational Pre-final LR.

Dataset	Model	Δ AUC	p_t/p_w	Wins
MATH	Qwen2.5-7B	+0.016 [+0.008, +0.025]	0.022/0.062	5/5
	Qwen2.5-32B	+0.019 [+0.002, +0.037]	0.099/0.125	4/5
	Llama-3.1-8B	+0.016 [+0.013, +0.020]	0.000/0.062	5/5
	Llama-3.2-3B	+0.011 [+0.006, +0.016]	0.012/0.062	5/5
HotpotQA	Qwen2.5-7B	+0.002 [-0.006, +0.011]	0.640/0.625	4/5
	Qwen2.5-32B	+0.046 [+0.033, +0.059]	0.002/0.062	5/5
	Llama-3.1-8B	+0.007 [+0.004, +0.011]	0.014/0.062	5/5
	Llama-3.2-3B	+0.006 [+0.002, +0.010]	0.034/0.062	5/5
FEVER	Qwen2.5-7B	+0.001 [-0.004, +0.005]	0.787/0.625	3/5
	Qwen2.5-32B	+0.001 [-0.006, +0.008]	0.826/1.000	2/5
	Llama-3.1-8B	+0.001 [-0.003, +0.006]	0.595/0.625	3/5
	Llama-3.2-3B	+0.002 [-0.005, +0.009]	0.606/0.625	3/5

Table 19: Paired split-level comparisons for adding the scalar MULTI-VIEW-RA transition reward to Operational Pre-final LR. Δ AUC is the mean paired held-out AUC difference with a 95% confidence interval over identical outer splits. p_t and p_w denote paired t -test and two-sided exact Wilcoxon signed-rank p-values, respectively. Wins reports the number of outer splits in which adding the MULTI-VIEW-RA reward improves AUC.

Dataset	Model	Representation	Trajectory AUC $\uparrow$
MATH	Qwen2.5-7B	State	0.610 [0.577, 0.643]
		Delta	0.747 [0.733, 0.761]
		End	0.638 [0.610, 0.666]
		Mean	0.599 [0.580, 0.619]
		State+End	0.603 [0.585, 0.621]
		State+Delta (MULTI-VIEW-RA)	0.748 [0.727, 0.769]
	Llama-3.1-8B	State	0.708 [0.681, 0.735]
		Delta	0.854 [0.812, 0.896]
		End	0.762 [0.722, 0.801]
		Mean	0.700 [0.661, 0.738]
		State+End	0.688 [0.641, 0.736]
		State+Delta (MULTI-VIEW-RA)	0.910 [0.894, 0.927]
HotpotQA	Qwen2.5-32B	State	0.544 [0.478, 0.610]
		Delta	0.730 [0.709, 0.752]
		End	0.637 [0.606, 0.668]
		Mean	0.558 [0.548, 0.568]
		State+End	0.596 [0.568, 0.623]
		State+Delta (MULTI-VIEW-RA)	0.757 [0.735, 0.778]

Table 20: Matched representation ablation under the fully inductive evaluation. Values are mean held-out trajectory ROC-AUC with 95% confidence intervals over five stratified splits. All variants use the same aligned rounds and downstream clustering and scoring pipeline, differing only in the round representation. Best AUC per dataset-model condition is shown in **bold**.

Dataset	Model	Method	$k=5$	$k=10$	$k=15$
MATH	Qwen2.5-7B	FLAT	0.690 [0.675, 0.704]	0.688 [0.667, 0.708]	0.668 [0.647, 0.688]
		MERGED-AO	0.696 [0.682, 0.711]	0.706 [0.684, 0.728]	0.684 [0.651, 0.717]
		MULTI-VIEW-RA	0.748 [0.727, 0.769]	0.755 [0.733, 0.777]	0.745 [0.728, 0.763]
	Llama-3.1-8B	FLAT	0.792 [0.761, 0.822]	0.777 [0.734, 0.820]	0.800 [0.791, 0.810]
		MERGED-AO	0.808 [0.786, 0.831]	0.781 [0.755, 0.807]	0.792 [0.768, 0.816]
		MULTI-VIEW-RA	0.910 [0.894, 0.927]	0.843 [0.837, 0.849]	0.792 [0.764, 0.821]
HotpotQA	Qwen2.5-32B	FLAT	0.661 [0.635, 0.686]	0.656 [0.631, 0.680]	0.683 [0.661, 0.705]
		MERGED-AO	0.680 [0.663, 0.697]	0.659 [0.643, 0.675]	0.667 [0.648, 0.686]
		MULTI-VIEW-RA	0.757 [0.735, 0.778]	0.757 [0.728, 0.785]	0.749 [0.733, 0.765]

Table 21: Cluster-count robustness under the fully inductive evaluation. Values are mean held-out trajectory ROC-AUC with 95% confidence intervals over five stratified 70/30 splits. All settings use full traces and $\alpha = 10^{-3}$. Best AUC within each dataset-model and cluster-count setting is shown in **bold**.

Dataset	Model	Method	$\alpha = 10^{-4}$	$\alpha = 10^{-3}$	$\alpha = 10^{-2}$	$\alpha = 10^{-1}$	$\alpha = 1$
MATH	Qwen2.5-7B	FLAT	0.687 [0.672, 0.703]	0.690 [0.675, 0.704]	0.688 [0.670, 0.706]	0.691 [0.672, 0.710]	0.712 [0.692, 0.733]
		MERGED-AO	0.694 [0.681, 0.707]	0.696 [0.682, 0.711]	0.692 [0.679, 0.705]	0.699 [0.686, 0.711]	0.713 [0.700, 0.726]
		MULTI-VIEW-RA	0.748 [0.727, 0.769]	0.748 [0.727, 0.769]	0.748 [0.727, 0.769]	0.748 [0.727, 0.769]	0.753 [0.734, 0.773]
	Llama-3.1-8B	FLAT	0.795 [0.768, 0.822]	0.792 [0.761, 0.822]	0.801 [0.770, 0.831]	0.806 [0.783, 0.830]	0.820 [0.802, 0.838]
		MERGED-AO	0.806 [0.785, 0.827]	0.808 [0.786, 0.831]	0.806 [0.789, 0.824]	0.812 [0.790, 0.834]	0.815 [0.800, 0.830]
		MULTI-VIEW-RA	0.910 [0.894, 0.927]	0.910 [0.894, 0.927]	0.910 [0.894, 0.927]	0.909 [0.892, 0.926]	0.904 [0.886, 0.922]
HotpotQA	Qwen2.5-32B	FLAT	0.661 [0.635, 0.686]	0.661 [0.635, 0.686]	0.661 [0.635, 0.686]	0.661 [0.635, 0.687]	0.665 [0.638, 0.693]
		MERGED-AO	0.680 [0.663, 0.696]	0.680 [0.663, 0.697]	0.680 [0.663, 0.696]	0.680 [0.664, 0.697]	0.683 [0.666, 0.699]
		MULTI-VIEW-RA	0.757 [0.735, 0.778]	0.757 [0.735, 0.778]	0.757 [0.735, 0.778]	0.757 [0.735, 0.778]	0.758 [0.737, 0.780]

Table 22: Sensitivity to Laplace smoothing under the fully inductive evaluation. Values are mean held-out trajectory ROC-AUC with 95% confidence intervals over five stratified 70/30 splits. All settings use full traces and $k = 5$. Best AUC within each dataset-model and smoothing setting is shown in **bold**.

Dataset	Model	Linear/Linear	RBF/RBF	Linear/RBF	RBF/Linear	Concat RBF	Direct Concat
MATH	Qwen2.5-7B	0.748 [0.727, 0.769]	0.753 [0.729, 0.776]	0.743 [0.718, 0.768]	0.763 [0.740, 0.786]	0.762 [0.738, 0.787]	0.755 [0.737, 0.774]
	Llama-3.1-8B	0.910 [0.894, 0.927]	0.911 [0.892, 0.930]	0.835 [0.829, 0.842]	0.903 [0.880, 0.925]	0.908 [0.893, 0.924]	0.911 [0.895, 0.926]
HotpotQA	Qwen2.5-32B	0.757 [0.735, 0.778]	0.751 [0.723, 0.778]	0.762 [0.737, 0.787]	0.766 [0.749, 0.783]	0.755 [0.734, 0.776]	0.756 [0.735, 0.778]

Table 23: Sensitivity to kernel and projection choices under the fully inductive evaluation. Values are mean held-out trajectory ROC-AUC with 95% confidence intervals over five stratified splits. All settings use full traces, $k = 5$, and Laplace smoothing $\alpha = 10^{-3}$. RBF denotes the radial basis function kernel. RBF bandwidths are determined from training rounds using a median-distance heuristic. Direct Concat applies k -means directly to the standardized state-delta vectors without spectral projection. Best AUC per dataset-model condition is shown in **bold**.

Dataset	Model	Method	Full Trace	No Final Answer	Δ AUC
MATH	Qwen2.5-7B	FLAT	0.690 [0.675, 0.704]	0.684 [0.675, 0.694]	-0.005
		MERGED-AO	0.696 [0.682, 0.711]	0.682 [0.664, 0.699]	-0.015
		MULTI-VIEW-RA	0.748 [0.727, 0.769]	0.662 [0.636, 0.688]	-0.086
	Llama-3.1-8B	FLAT	0.792 [0.761, 0.822]	0.784 [0.748, 0.820]	-0.007
		MERGED-AO	0.808 [0.786, 0.831]	0.776 [0.734, 0.818]	-0.032
		MULTI-VIEW-RA	0.910 [0.894, 0.927]	0.846 [0.816, 0.876]	-0.064
HotpotQA	Qwen2.5-32B	FLAT	0.661 [0.635, 0.686]	0.620 [0.599, 0.641]	-0.040
		MERGED-AO	0.680 [0.663, 0.697]	0.612 [0.595, 0.629]	-0.068
		MULTI-VIEW-RA	0.757 [0.735, 0.778]	0.701 [0.677, 0.724]	-0.056

Table 24: Sensitivity to terminal-answer information under the fully inductive evaluation. Final-answer segments are removed before reclustering, while terminal thoughts are retained. Values are mean held-out trajectory ROC-AUC with 95% confidence intervals over five stratified 70/30 splits. Δ AUC denotes no-final-answer AUC minus full-trace AUC. Best AUC within each dataset-model and evaluation setting is shown in **bold**.

Dataset	Model	Method	All	Short	Medium	Long
MATH	Qwen2.5-7B	FLAT	0.690 [0.675, 0.704]	0.627 [0.608, 0.646]	0.571 [0.534, 0.608]	0.518 [0.486, 0.550]
		MERGED-AO	0.696 [0.682, 0.711]	0.631 [0.594, 0.668]	0.568 [0.542, 0.593]	0.528 [0.502, 0.553]
		MULTI-VIEW-RA	0.748 [0.727, 0.769]	0.645 [0.587, 0.703]	0.611 [0.573, 0.649]	0.786 [0.731, 0.841]
	Llama-3.1-8B	FLAT	0.792 [0.761, 0.822]	0.678 [0.634, 0.722]	0.523 [0.513, 0.533]	0.693 [0.472, 0.915]
		MERGED-AO	0.808 [0.786, 0.831]	0.696 [0.665, 0.727]	0.554 [0.518, 0.590]	0.689 [0.450, 0.928]
		MULTI-VIEW-RA	0.910 [0.894, 0.927]	0.777 [0.748, 0.807]	0.913 [0.879, 0.946]	0.981 [0.969, 0.993]
HotpotQA	Qwen2.5-32B	FLAT	0.661 [0.635, 0.686]	0.523 [0.497, 0.549]	0.577 [0.535, 0.619]	0.515 [0.483, 0.547]
		MERGED-AO	0.680 [0.663, 0.697]	0.578 [0.540, 0.616]	0.597 [0.575, 0.619]	0.510 [0.481, 0.539]
		MULTI-VIEW-RA	0.757 [0.735, 0.778]	0.634 [0.613, 0.656]	0.661 [0.621, 0.701]	0.872 [0.863, 0.881]

Table 25: Trace-length stratification under the fully inductive evaluation. Length-tertile thresholds are derived from training traces and applied unchanged to held-out traces. Values are mean held-out trajectory ROC-AUC with 95% confidence intervals over evaluable splits. All settings use full-trace abstractions, $k = 5$, and $\alpha = 10^{-3}$. Best AUC within each dataset-model and length stratum is shown in **bold**.

Dataset	Model	Action Stratum	Avg. N	Correct	Eval. Splits	FLAT	MERGED-AO	MULTI-VIEW-RA	Δ AUC
MATH	Qwen2.5-7B	Low	125.6	43.3%	5/5	0.655 [0.625, 0.685]	0.655 [0.634, 0.677]	0.632 [0.586, 0.677]	-0.024
		Medium	46.2	36.4%	5/5	0.806 [0.735, 0.877]	0.807 [0.733, 0.881]	0.763 [0.694, 0.832]	-0.044
		High	79.6	14.1%	5/5	0.501 [0.403, 0.600]	0.516 [0.434, 0.597]	0.774 [0.703, 0.846]	+0.259
	Llama-3.1-8B	Low	99.0	33.5%	5/5	0.618 [0.562, 0.673]	0.618 [0.575, 0.662]	0.690 [0.630, 0.750]	+0.072
		Medium	5.6	32.1%	5/5	0.608 [0.243, 0.973]	0.608 [0.243, 0.973]	0.306 [0.080, 0.531]	-0.303
		High	182.0	0.0%	0/5	—	—	—	—
HotpotQA	Qwen2.5-32B	Low	173.8	51.1%	5/5	0.580 [0.565, 0.596]	0.612 [0.594, 0.631]	0.650 [0.631, 0.668]	+0.037
		Medium	38.0	41.1%	5/5	0.536 [0.504, 0.567]	0.538 [0.514, 0.563]	0.677 [0.631, 0.724]	+0.139
		High	83.2	14.9%	5/5	0.522 [0.471, 0.572]	0.532 [0.483, 0.582]	0.891 [0.877, 0.904]	+0.358

Table 26: Action-count stratification under the fully inductive evaluation. Thresholds are derived from training traces and applied unchanged to held-out traces. Values report mean held-out stratum size, correctness rate, number of evaluable splits, and trajectory ROC-AUC with 95% confidence intervals. Δ AUC denotes MULTI-VIEW-RA AUC minus the stronger of FLAT and MERGED-AO. AUC is undefined when a stratum contains only one correctness class. Best AUC within each dataset-model and action stratum is shown in **bold**.

Dataset	Model	Method	All	No Tool Error	Tool Error
MATH	Qwen2.5-7B	FLAT	0.690 [0.675, 0.704]	0.703 [0.685, 0.720]	0.609 [0.595, 0.624]
		MERGED-AO	0.696 [0.682, 0.711]	0.706 [0.691, 0.721]	0.620 [0.600, 0.639]
		MULTI-VIEW-RA	0.748 [0.727, 0.769]	0.682 [0.657, 0.707]	0.857 [0.819, 0.895]
	Llama-3.1-8B	FLAT	0.792 [0.761, 0.822]	0.805 [0.773, 0.838]	0.803 [0.730, 0.876]
		MERGED-AO	0.808 [0.786, 0.831]	0.820 [0.802, 0.837]	0.812 [0.740, 0.883]
		MULTI-VIEW-RA	0.910 [0.894, 0.927]	0.908 [0.889, 0.927]	0.916 [0.897, 0.935]
HotpotQA	Qwen2.5-32B	FLAT	0.661 [0.635, 0.686]	0.657 [0.633, 0.682]	1.000 [1.000, 1.000]
		MERGED-AO	0.680 [0.663, 0.697]	0.678 [0.660, 0.695]	0.875 [0.630, 1.000]
		MULTI-VIEW-RA	0.757 [0.735, 0.778]	0.754 [0.734, 0.775]	1.000 [1.000, 1.000]

Table 27: Tool-error stratification under the fully inductive evaluation. Tool-error traces contain at least one pre-final action or observation with an explicit execution-error marker; empty or no-result retrievals are not counted as errors. Values are mean held-out trajectory ROC-AUC with 95% confidence intervals over evaluable splits. Best AUC within each dataset-model and error stratum is shown in **bold**. HotpotQA tool-error strata are very small and are therefore reported only descriptively.