SAGE: Mitigating Long-Horizon Reasoning Biases via Topological Guidance

Xinyue Zeng
CS Department
Virginia Tech

Jiawei Zhang
CS Department
University of Wisconsin Madison

Yujun Yan
CS Department
Dartmouth College

Dawei Zhou
CS Department
Virginia Tech

Abstract

Long-horizon reasoning remains a central challenge for large language models (LLMs) under sparse-reward regimes. We argue that this brittleness arises from two biases induced by complex reasoning spaces: an exploration bias, where models are drawn toward locally plausible but structurally unstable branches, and a compounding bias, where small local deviations accumulate across depth and suppress rare rewards. We introduce Symbolic Closure Analysis (SCA) as a theoretical lens characterizing how branching structures and sparse rewards induce these biases in long-horizon reasoning with local admissibility, and as a design principle for structural priors in less formal reasoning tasks. Motivated by this analysis, we propose SAGE (**S**tructural **A**dmissibility-**G**uided **E**xploration), a unified framework that injects structural guidance to alleviate exploration bias and compounding bias in long-horizon reasoning. SAGE combines two complementary structural guidance: algebraic sparsification, which projects locally admissible candidates onto operator-indexed algebraic subspaces to suppress spurious branching and mitigate exploration bias, and hyperbolic structural guidance, which embeds reasoning states into a negatively curved space to provide dense depth-wise signals and mitigate compounding bias. Across 12 benchmarks and 10 model families, SAGE outperforms competitive baselines. In particular, SAGE achieves up to an 8-fold improvement on the Andrews-Curtis problem, an open real-world long-horizon task. Code is available at: `https://github.com/Susan571/SAGE-NeurIPS2026`.

1 Introduction

Long-horizon reasoning remains a central challenge for large language models (LLMs), where success often requires discovering deep, structured reasoning trajectories across many steps. Post-training learning has become a dominant paradigm for improving LLM reasoning, with strong progress in mathematics and coding through outcome-based objectives (Lyu et al., 2025; Zhang et al., 2025b; Zeng et al., 2025; Lin et al., 2026a). However, as reasoning horizons grow, outcome-based post-training becomes brittle in sparse-reward regimes: limited intermediate rewards provide insufficient guidance for preserving long-term feasibility, leading to locally plausible but globally unstable patterns (Suo et al., 2025; Zeng et al., 2026).

This brittleness reflects two distinct but coupled long-horizon reasoning biases, illustrated in Figure 1 with the Andrews-Curtis (AC) trivialization task as a running example, where the task is to reach a trivial presentation through a long sequence of admissible transformations. The first is *exploration bias*, which arises from the structural complexity of the reasoning space: as the reasoning tree expands with depth, successful trajectories occupy a narrow feasible region, while much larger regions contain locally plausible but globally unproductive branches (Dziri et al., 2023), so outcome-based post-training tends to favor trajectories that are easy to sample rather than trajectories that remain extendable to success. The second is *compounding bias*, which arises from sparse-reward regimes with limited intermediate rewards: small local deviations accumulate across depth and

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

progressively move the trajectory away from long-term feasibility (Casper et al., 2023; Weaver & Tao, 2001). Together, these biases explain why LLMs may produce plausible intermediate steps while still failing at end-to-end long-horizon reasoning, and existing mitigation efforts address them along two corresponding axes. To counter exploration bias, one line of work introduces dense supervision that decomposes long-horizon reasoning into locally verifiable steps, ranging from interactive verifiers that enforce rigorous logical transitions (Yang et al., 2023; Hsiang et al., 2025) to learned verifiers or LLM-as-a-Judge that guide Monte Carlo Tree Search via step-wise critiques (Lightman et al., 2024); however, reliable process supervision remains structurally expensive to scale, ultimately shifting the bottleneck without resolving the fundamental difficulty of learning from sparse rewards. To counter compounding bias, a second line of work adapts outcome-based RL to sparse-reward regimes by augmenting the objective with auxiliary heuristics, including iterative bootstrapping methods like STaR or ReST (Zelikman et al., 2022; Gülçehre et al., 2023) and intrinsic motivation mechanisms based on entropy regularization or syntactic constraints (She et al., 2025; Zhang et al., 2026; Chen et al., 2025; Gai et al., 2025); yet these methods do not model the structure of the long-horizon reasoning space, leaving the geometry of feasible trajectories implicit.

This gap motivates two fundamental research questions: **Q1:** *Can we characterize the dominant factors that drive long-horizon reasoning failures under structural complexity and sparse-reward regimes?* **Q2:** *Can this characterization be instantiated as a unified framework that injects structural guidance to alleviate exploration bias and compounding bias?*

To address this gap, we first introduce *Symbolic Closure Analysis* (SCA), a theoretical framework for characterizing the dominant factors behind long-horizon reasoning failures under structural complexity and sparse-reward regimes. SCA models reasoning as a sequence of locally admissible transformations and studies how feasible support evolves in expanding reasoning spaces. Our analysis shows that structural complexity dilutes feasible support across high-volume but unproductive branches, producing exploration bias, while sparse-reward regimes provide limited intermediate rewards, allowing local deviations to accumulate with depth and produce compounding bias.

Motivated by SCA, we propose ***S**tructural **A**dmissibility-**G**uided **E**xploration* (SAGE), a unified framework for alleviating long-horizon reasoning biases. SAGE injects structural guidance during policy optimization so that the learned policy internalizes useful properties of structured reasoning space, through two complementary structural guidance: *algebraic sparsification*, which reduces spurious branching and alleviates exploration bias, and *hyperbolic structural guidance*, which provides dense depth-aware signals and alleviates compounding bias. Together, these components translate the SCA diagnosis into trainable guidance signals.

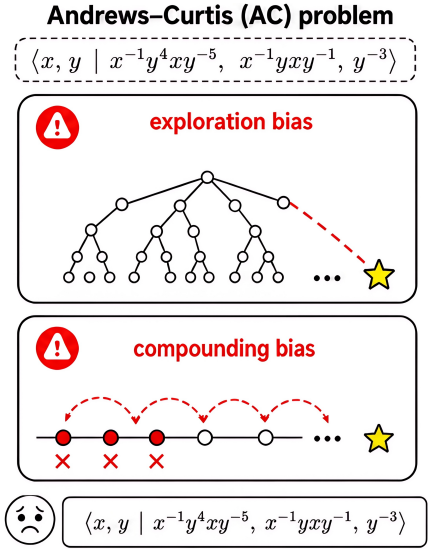


Figure 1: AC problem illustrates two long-horizon reasoning biases: exploration bias from many locally valid but low-promise branches, and compounding bias from early plausible deviations whose failures appear in later steps.

We evaluate SAGE across 10 model families and 12 benchmarks spanning closed-form mathematical reasoning, free-form natural reasoning, and open long-horizon reasoning task. SAGE always outperforms competitive baselines and particularly achieves up to 8-fold improvement on AC problem, an open real-world long-horizon reasoning task.

Our contributions are threefold:

- We introduce SCA as a theoretical lens to characterize the dominant factors behind long-horizon reasoning failures under structural complexity and sparse-reward regimes.
- We propose SAGE, a unified framework for alleviating long-horizon reasoning biases.
- Evaluation across 12 benchmarks and 10 models show that SAGE consistently outperforms competitive baseline. Code is open-sourced at: `https://github.com/Susan571/SAGE-NeurIPS2026`.

2 Preliminaries

2.1 Long-Horizon Reasoning

Following prior work (Yao et al., 2023; Lightman et al., 2024), we study long-horizon reasoning as a sequential decision process over discrete symbolic manipulations. Let $\mathfrak{S}$ denote the task-specific symbolic interface with local admissibility predicate $\mathrm{Adm}_{\mathfrak{S}}$. At each step $t \in \{0, \ldots, T-1\}$, the system selects $a_t \in \mathcal{A}(s_t)$ with $\mathrm{Adm}_{\mathfrak{S}}(s_t, a_t) = 1$ and transitions via $s_{t+1} = \Phi(s_t, a_t)$, yielding a trajectory $\tau = (s_0, a_0, \ldots, s_T)$ of maximum path length T. For analytical convenience, we cast this as a finite-horizon MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, T \rangle$ over reasoning contexts, symbolic manipulations, transition dynamics, and task-level outcome signal. The difficulty grows rapidly with T, shaped by two salient properties. The first is *structural complexity*: the reachable search space scales as $|\Omega| = O(\bar{A}^T)$ for average branching factor $\bar{A}$, while successful trajectories form only a small subset $\mathcal{T}^*$ (treated as an analytical object, not as supervision), occupying a vanishing fraction of the reachable manifold as T grows. The second is *sparse outcome feedback*: meaningful supervision is often available only at the trajectory end. Under terminal reward $r(\tau) = \mathbf{1}[\tau \in \mathcal{T}^*]$, informative feedback is inherently rare because successful trajectories themselves are rare, making credit assignment progressively harder and often yielding ineffective optimization in sparse-reward regimes (Uesato et al., 2022; Weaver & Tao, 2001).

2.2 Biases in Long-Horizon Reasoning

Structural complexity and sparse outcome feedback do not merely make long-horizon reasoning more difficult; they induce recurring distortions in how trajectories are explored and preserved (Zhou et al., 2025; Brantley et al., 2025; Liu et al., 2025; Jahin et al., 2025). Figure 1 illustrates these distortions in the AC problem, where an LLM policy transforms a group presentation toward the trivial presentation through legal symbolic moves such as inverting a relator, multiplying relators, or conjugating by a generator. Although many moves are locally valid, only a small subset continues to simplify the presentation over long horizons, giving rise to two coupled biases: *structural complexity* mainly induces an *exploration bias*, while *sparse outcome feedback* mainly induces a *compounding bias*, and the two interact as depth increases.

Exploration bias. As search volume grows exponentially, locally admissible trajectories that do not extend to success can overwhelmingly dominate the reachable set (Dziri et al., 2023): many moves are locally admissible, but only a few are structurally extendable. In the AC example, many legal moves branch from the same presentation, yet only a few continue to simplify the residual algebraic structure, so a policy sampling broadly from admissible moves spends most of its budget on locally plausible but structurally unstable paths.

Compounding bias. Sparse outcome feedback yields a different but equally persistent failure mode: with supervision only at the end of a long chain, small local deviations cannot be corrected early and instead accumulate. As shown in Figure 1, an early AC move such as replacing one relator by its product with another may look locally plausible, yet redirect the trajectory into a region where subsequent legal moves preserve or amplify complexity, with failure surfacing only several steps later. This is especially problematic in KL-regularized post-training: when terminal rewards are weak, the update is dominated by the KL term (Lyu et al., 2025; Schulman et al., 2017), and in the limit where the expected reward contribution vanishes, it degenerates toward preserving the reference policy, $\nabla J \approx -\beta \nabla D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}})$, allowing locally plausible deviations to persist rather than being corrected by task-level structure (Casper et al., 2023).

Together, we formalize the problem as follows:

Problem 2.1 (Alleviating Long-Horizon Reasoning Biases)**.**

Given: (i) a task space $\mathcal{Q}$ with complex reasoning structures whose search volume scales exponentially with path length T ($|\Omega| = O(\bar{A}^T)$); and (ii) a sparse-reward regime with a pretrained policy π_0 and sparse terminal reward $\mathcal{O}(\tau)$, where the initial success probability is negligible ($\mathbb{E}_{\tau \sim \pi_0}[\mathcal{O}(\tau)] \approx 0$).

Find: a reasoning policy π^* that increases the probability of successful trajectories by alleviating both biases, without access to dense process labels or ground-truth solution paths: $\pi^* = \mathrm{argmax}_\pi \, \mathbb{E}_{q \sim \mathcal{Q}} \left[\mathbb{P}(\tau \in \mathcal{T}^* \mid \pi, q) \right]$.

3 Theoretical Analysis

We introduce *Symbolic Closure Analysis* (SCA) as a theoretical lens for understanding the two biases identified in Section 2.2. SCA characterizes the reasoning manifold through a prefix-closed feasible region induced by local admissibility, providing a principled lens for diagnosing why standard outcome-based learning is structurally biased under sparse-reward regimes.

3.1 SCA: Symbolic Closure Analysis

To analyze long-horizon reasoning in sparse-reward regimes, we first revisit reference-regularized post-training (Ziegler et al., 2019; Ouyang et al., 2022), which stabilizes learning through a likelihood-based closure $\Omega_{\text{stat}} = \tau : \mathbb{P}\pi\text{ref}(\tau) \geq \epsilon$ but preserves trajectories that are easy to sample rather than those feasible over long horizons. We propose SCA, which instead defines feasibility through local admissibility. The resulting feasible region $\mathcal{F}$ is prefix-closed by construction, providing an analytical object against which exploration and compounding biases can be characterized.

Definition 3.1 (SCA Feasible Closure)**.** Let $\mathfrak{S}$ denote a domain-specific symbolic system specifying admissible operators and a local admissibility predicate $\mathrm{Adm}_{\mathfrak{S}}$. For a trajectory $\tau = (s_0, a_0, \ldots, a_{T-1}, s_T)$ with $a_t \in \mathcal{A}(s_t)$ and $s_{t+1} = \Phi(s_t, a_t)$, the SCA-feasible set $\mathcal{F}$ is $\tau \in \mathcal{F} \iff \{\forall t \in \{0, \ldots, T-1\},\ \mathrm{Adm}_{\mathfrak{S}}(s_t, a_t, s_{t+1}) = 1\}$.

Remark 3.2 (Local Admissibility). $\mathrm{Adm}_{\mathfrak{S}}$ checks only whether a transition is locally well-formed under $\mathfrak{S}$; it does not certify global correctness, provide ground-truth steps, reveal successful trajectories, or use outcome labels.

Remark 3.3 (Prefix Closure). $\mathrm{Adm}_{\mathcal{S}}$ induces a prefix-closed feasible region: if any prefix of τ violates local admissibility, all its continuations are infeasible. Equivalently, for any $\tau \in \mathcal{F}$, every prefix $\tau_{\leq t} \in \mathcal{F}$.

SCA admits three levels of instantiation. In explicit symbolic systems, $\mathrm{Adm}_{\mathfrak{S}}$ is given by the task interface and $\mathcal{F}$ is exact. In semi-structured domains such as closed-form mathematics, unresolved variables, equations, and answer-schema constraints serve as computable proxies for residual structure. In free-form natural reasoning, residuals and target anchors are estimated from prompt-conditioned semantic structure.

3.2 Theoretical Analysis of Biases under SCA

We now show how SCA explains the two biases of Section 2.2. Let $\mathcal{F}$ be the prefix-closed feasible set (Definition 3.1) and $\mathcal{G} = \Omega \setminus \mathcal{F}$ its inadmissible complement, with $\mathcal{T}^* \subseteq \mathcal{F}$ unobserved. We treat $\mathcal{F}$ as a tractable structural proxy: induced by local admissibility in symbolic domains, and approximating the long-horizon extendable subset elsewhere.

For a rollout policy π, let $\mu_\pi := \mathbb{P}_{\tau\sim\pi}(\tau \in \mathcal{G})$ and $p_\pi := \mathbb{P}_{\tau\sim\pi}(\tau \in \mathcal{T}^*)$. Under SCA, exploration bias appears as $\mu_{\pi_0} \approx 1$—rollouts concentrate outside $\mathcal{F}$—while compounding bias appears as the persistence of early inadmissible deviations: terminal-only rewards let locally unstable prefixes survive long enough to dominate the trajectory distribution.

Characterizing Exploration Bias via Feasible-Region Geometry. Structural complexity induces exploration bias because the reachable space grows exponentially with T, while admissible and successful trajectories occupy only a small fraction. Let $\tau_{\leq t}$ denote a prefix and define the local feasibility indicator $\phi(\tau_{\leq t}) = \mathbf{1}[\tau_{\leq t}$ is locally admissible under $\mathrm{Adm}_{\mathcal{S}}]$. By prefix closure, $\phi(\tau_{\leq t}) = 0 \Rightarrow \phi(\tau_{\leq t'}) = 0$ for all $t' \geq t$, hence $\mathcal{G} = \{\tau \in \Omega : \exists t,\ \phi(\tau_{\leq t}) = 0\}$.

Let B_t and $B_t^{\mathcal{F}}$ denote the effective and locally admissible branching factors at depth t. When $B_t^{\mathcal{F}} \ll B_t$ for many t, a crude volume comparison gives

$$\frac{|\{\tau_{\leq T} : \tau \in \mathcal{F}\}|}{|\{\tau_{\leq T} : \tau \in \Omega\}|} \lesssim \prod_{t=1}^{T} \frac{B_t^{\mathcal{F}}}{B_t}. \tag{1}$$

Thus, unless the base policy places exponentially increasing preference on $\mathcal{F}$, unconstrained rollouts concentrate outside the feasible region ($\mu_{\pi_0} \approx 1$). SCA identifies this volume mismatch as the structural origin of exploration bias and isolates feasible-support concentration as the property any successful mitigation must achieve and, as the next proposition shows, the structural quantity controlling gradient variance.

Proposition 3.4 (Variance Decomposition over Feasible and Inadmissible Regions)**.** *Let π be any rollout policy and $\hat{\rho}_\pi$ its empirical rollout distribution. For any score-function gradient term $g(\tau)$,*

$$\begin{aligned}\mathrm{Var}_{\tau\sim\hat{\rho}_\pi}[g(\tau)] &= \mathbb{P}(\tau\in\mathcal{F})\,\mathrm{Var}[g(\tau)\mid\tau\in\mathcal{F}] \\ &\quad+\mathbb{P}(\tau\in\mathcal{G})\,\mathrm{Var}[g(\tau)\mid\tau\in\mathcal{G}] \\ &\quad+\mathbb{P}(\tau\in\mathcal{F})\mathbb{P}(\tau\in\mathcal{G})\Big(\mathbb{E}[g(\tau)\mid\tau\in\mathcal{F}]-\mathbb{E}[g(\tau)\mid\tau\in\mathcal{G}]\Big)^2. \end{aligned} \tag{2}$$

Consequently, any policy with $\mathrm{supp}(\pi)\subseteq\mathcal{F}$ removes both the $\mathcal{G}$-conditioned variance and the between-region term.

Characterizing Compounding Bias under Sparse-reward Regimes. Under sparse rewards, local deviations go uncorrected until terminal evaluation, allowing trajectories to drift from feasible structure without intermediate signal. KL-regularized post-training sharpens this: when successful trajectories are rare under the reference policy, the reward signal is too weak to pull the learned policy away from reference-model behavior.

Consider the KL-regularized objective $\max_\pi\ \mathbb{E}_{\tau\sim\pi}[R(\tau)]-\lambda D_{\mathrm{KL}}(\pi\|\pi_{\mathrm{ref}})$, with terminal-only $R(\tau)\in[0,R_{\max}]$ and successful set $S=\{\tau: R(\tau)>0\}$. Then $\mathbb{E}_{\tau\sim\pi}[R(\tau)]\le R_{\max}\pi(S)$, while the KL penalty is dense over the full rollout space. The following theorem makes this mismatch precise: when S is rare under π_{ref}, the full-support KL-regularized optimizer cannot move meaningfully away from π_{ref}, so locally inadmissible prefixes under π_{ref} remain after optimization.

Theorem 3.5 (Reference anchoring under rare terminal rewards)**.** *For Section 3.2 with terminal-only $R(\tau)\in[0,R_{\max}]$, let $S=\{\tau: R(\tau)>0\}$ and $p=\mathbb{P}_{\tau\sim\pi_{\mathrm{ref}}}(S)$. The full-support optimizer $\pi^*_\Omega(\tau)=\pi_{\mathrm{ref}}(\tau)\exp(R(\tau)/\lambda)/\mathbb{E}_{\tau'\sim\pi_{\mathrm{ref}}}[\exp(R(\tau')/\lambda)]$ satisfies*

$$D_{\mathrm{TV}}(\pi^*_\Omega,\pi_{\mathrm{ref}})\le\big(e^{R_{\max}/\lambda}-1\big)p. \tag{3}$$

Hence if $(e^{R_{\max}/\lambda}-1)p\ll 1$, the optimizer remains close to π_{ref}. In the high-KL or weak-reward regime $R_{\max}\ll\lambda$, this becomes $D_{\mathrm{TV}}(\pi^*_\Omega,\pi_{\mathrm{ref}})\le(R_{\max}/\lambda)p+O(R_{\max}^2p/\lambda^2)$. Thus, when successful trajectories are rare under the reference, terminal-only KL-regularized optimization has limited leverage to move probability mass away from reference-likely prefixes.

4 SAGE: Structural Admissibility-Guided Exploration

4.1 From SCA to SAGE

SCA turns the two biases into computational requirements: exploration bias requires feasible-support concentration without observing $\mathcal{T}^*$, and compounding bias requires prefix-level correction without dense process labels. Motivated by Theorem 3.5, we address both through ***S**tructural **A**dmissibility-**G**uided **E**xploration* (SAGE), which injects two complementary structural potentials into post-training.

Algebraic sparsification $\Psi_{\mathcal{P}}$ targets exploration bias by biasing sampling toward operators that explain the current symbolic residual $r_t\in\mathbb{R}^d$ (Equation (1)). For a candidate operator L_j with associated subspace S_j:

$$\Psi_{\mathcal{P}}(r_t,S_j)=\frac{\|P_{S_j}r_t\|_2^2}{\|r_t\|_2^2+\varepsilon}, \tag{4}$$

where P_{S_j} projects onto S_j. Larger $\Psi_{\mathcal{P}}$ indicates the operator addresses more of the unresolved residual, providing a soft compatibility score (not a replacement for $\mathrm{Adm}_{\mathfrak{S}}$) that promotes feasible-support concentration without inference-time pruning.

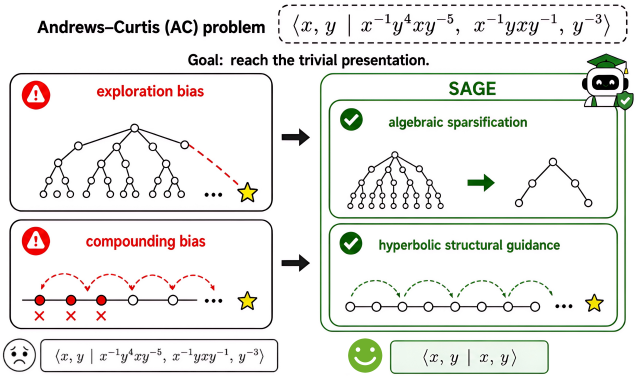


Figure 2: Overview of SAGE workflow. Outcome-only training induces exploration and compounding biases; SAGE injects algebraic sparsification and hyperbolic structural guidance during RL, yielding focused search, stable reasoning, and no additional inference-time cost.

Hyperbolic structural guidance $\Psi_{\mathcal{H}}$ targets compounding bias by supplying prefix-level feedback before terminal rewards arrive (Theorem 3.5), exploiting hyperbolic geometry's natural fit for hierarchical

reasoning structure. For state s_t, operator L_j, target structure g, and Poincaré embedding $\mathcal{E}(\cdot)$: $\Psi_{\mathcal{H}}(s_t, L_j, g) = \exp\big(-d_{\mathbb{D}}\big(\mathcal{E}(s_t \circ L_j), \mathcal{E}(g)\big)/\kappa\big)$, with Poincaré distance $d_{\mathbb{D}}$ and sharpness $\kappa > 0$. For Andrews–Curtis, s_t, r_t, g are the current presentation, unresolved structure, and trivial presentation; for closed-form and free-form reasoning, r_t and g are training-time structural priors (Appendix E.1). Both potentials are computable without $\mathcal{T}^*$ or process labels, and the resulting structural preferences are absorbed into the policy during post-training, so inference incurs no additional search or filtering.

Proposition 4.1 (Non-vanishing structural advantage under sparse rewards)**.** *If $R_{\text{term}}(\tau) = 0$ across a rollout group and Ψ_{SAGE} has nonzero within-group variance, the SAGE advantage $A_i = \big(\eta\bar{\Psi}_{\text{SAGE}}(\tau_i) - \eta\frac{1}{G}\sum_j \bar{\Psi}_{\text{SAGE}}(\tau_j)\big)/\big(\text{std}_j(\eta\bar{\Psi}_{\text{SAGE}}(\tau_j)) + \epsilon\big)$ remains nonzero, providing an update signal even under uninformative terminal rewards.*

4.2 SAGE

The two potentials combine into the SAGE mechanism: $\Psi_{SAGE}(s_t, a_t) = \alpha\,\Psi_{\mathcal{P}}(s_t, a_t) + \gamma\,\Psi_{\mathcal{H}}(s_t, a_t)$, with $\alpha, \gamma \geq 0$ controlling relative strength.

Structure-Guided Sampling. SAGE modulates the rollout distribution via $P_{\text{sample}}(a_t^{(k)} \mid s_t, \mathcal{C}_t) \propto \exp\Big(\bar{\ell}_{\theta_{\text{old}}}(a_t^{(k)} \mid s_t) + \lambda\Psi_{SAGE}(s_t, a_t^{(k)})\Big), \quad a_t^{(k)} \in \mathcal{C}_t$, where $\lambda \geq 0$ controls guidance strength and $\mathcal{C}_t$ is, for non-symbolic tasks, a finite candidate set sampled from $\pi_{\theta_{\text{old}}}$ (normalization is over $\mathcal{C}_t$, not the full vocabulary). This implements both SCA requirements as soft biases-feasible-support concentration via $\Psi_{\mathcal{P}}$, prefix-level signal via $\Psi_{\mathcal{H}}$-rather than projecting onto $\mathcal{F}$ through hard constraints.

Guidance-Augmented Advantage. For each τ_i, SAGE forms a reward combining terminal outcome with structural guidance: $\widetilde{R}(\tau_i) = R_{\text{term}}(\tau_i) + \eta \cdot \frac{1}{T_i}\sum_{t=1}^{T_i} \Psi_{SAGE}(s_t^i, a_t^i)$, with group-relative advantage $A_i = (\widetilde{R}(\tau_i) - \text{mean}_j\,\widetilde{R}(\tau_j))/(\text{std}_j\,\widetilde{R}(\tau_j) + \varepsilon)$. This makes the sparse terminal signal usable in low-resource regimes by supplementing it with dense structural feedback.

Policy Update. We optimize a step-level KL-regularized group-relative objective:

$$\mathcal{L}_{\text{SAGE}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\left[\min\{\rho_{i,t}(\theta)A_i, \text{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_i\} - \delta\log\frac{\pi_\theta(a_{i,t}|s_{i,t})}{\pi_{\text{ref}}(a_{i,t}|s_{i,t})}\right], \tag{5}$$

with $\rho_{i,t}(\theta) = \pi_\theta(a_{i,t} \mid s_{i,t})/\pi_{\theta_{\text{old}}}(a_{i,t} \mid s_{i,t})$. For concentration analysis, we use the trajectory-level distribution $P_{\text{EBM}}(\tau) = \pi_{\theta_{\text{old}}}(\tau)\exp(\lambda\Psi_{\text{SAGE}}(\tau))/Z_\lambda$, $\Psi_{\text{SAGE}}(\tau) = \sum_{t=1}^{T}\Psi_{\text{SAGE}}(s_t, a_t)$.

Proposition 4.2 (Guidance-Induced Feasible-Support Concentration)**.** *Let P_{EBM} be the trajectory-level distribution, $\Psi_{SAGE}(\tau) = \sum_{t=1}^{T}\Psi_{SAGE}(s_t, a_t)$, and $\mathcal{F}$, $\mathcal{G} = \Omega \setminus \mathcal{F}$ the SCA-feasible and inadmissible regions. If there exists $\Delta > 0$ with $\inf_{\tau\in\mathcal{F}}\Psi_{SAGE}(\tau) - \sup_{\tau\in\mathcal{G}}\Psi_{SAGE}(\tau) \geq \Delta$, then $P_{\text{EBM}}(\mathcal{G})/P_{\text{EBM}}(\mathcal{F}) \leq e^{-\lambda\Delta}\,\pi_{\theta_{\text{old}}}(\mathcal{G})/\pi_{\theta_{\text{old}}}(\mathcal{F})$. Thus, increasing λ exponentially suppresses locally inadmissible mass relative to feasible mass.*

5 Experiment

We comprehensively evaluate SAGE across 12 diverse benchmarks, 10 backbone models and 4 competitive baselines. Our experiments address the following three questions: **Q1:** *Does SAGE improve reasoning performance across diverse benchmarks and model scales?* **Q2:** *Does SAGE alleviate exploration and compounding biases on long-horizon reasoning tasks?* **Q3:** *Do algebraic sparsification and hyperbolic structural guidance each contribute to the gains predicted by SCA?*

Section 5.1 describes the experimental setup. Section 5.2 reports performance on mathematical and free-form natural reasoning benchmarks. Section 5.3 evaluates long-horizon symbolic reasoning as a stress test for exploration and compounding biases. Section 5.4 is the ablation study of the two components. We report implementation details and additional results in Appendix E.

5.1 Experimental Settings

Models. We evaluate SAGE across multiple model backbones, including Qwen3.5 (2B, 9B, 35B) (Team, 2026b), Qwen3.6-27B (Team, 2026b), Qwen3-32B (Team, 2025), DeepSeekMath-

7B (Shao et al., 2024), DeepSeek-Prover-V2-7B (Ren et al., 2025), Kimina-Prover (7B, Distill-8B) (Wang et al., 2025), Llama-3.3-70B-Instruct (Patterson et al., 2022), MiniMax M2.5 (MiniMax, 2026), GLM-5 (GLM, 2026) and Kimi K2.6 (Team, 2026a).

Benchmarks. We evaluate across three families of tasks: (i) *Closed-form Mathematical Reasoning:* MATH (Lightman et al., 2024), Minerva Math (Lewkowycz et al., 2022), AMC23 (Math AI, 2025), AIME 2024 (Hugging Face H4, 2025), OlympiadBench (He et al., 2024), GSM8K (Cobbe et al., 2021), and Putnam (Tsoukalas et al., 2024); (ii) *Free-form Natural Reasoning:* MMLU-Pro (Wang et al., 2024), GPQA (Rein et al., 2023), BBH-H (Suzgun et al., 2023), and ARC-C (Clark et al., 2018); and (iii) *Real-world Long-horizon Reasoning:* AC problem task, for which we follow Shehper et al. (2025) to construct 1190 AC presentations with $n \leq 7$ and $|w| \leq 7$.

Baselines. We compare with 4 representative baselines, including supervised fine-tuning (SFT), GRPO (Shao et al., 2024), EMPO (Zhang et al., 2025a) and GRPO-PRM (Sullivan & Koller, 2025). All methods use comparable rollout budgets, generation lengths, and decoding constraints.

Table 1: Accuracy (%) on mathematical reasoning benchmarks. Best in **bold**, second best underlined.

Model	MATH	Minerva	Olympiad	AIME24	AMC23	GSM8K	Putnam	Avg.
Flagship								
Llama-3.3-70B-Instruct	66.08	33.61	32.94	17.31	29.02	80.47	9.91	38.48
MiniMax M2.5	82.43	49.76	65.71	64.28	69.43	93.54	20.84	63.71
GLM-5	83.02	50.51	68.74	66.08	70.02	94.01	21.08	64.78
Kimi K2.6	83.45	50.82	67.85	65.22	70.28	94.07	21.91	64.80
2B models								
Qwen3.5	50.64	11.62	24.58	9.41	43.36	47.38	3.66	27.24
Qwen3.5 w/SFT	60.41	26.52	27.96	3.74	37.88	55.06	4.63	30.89
Qwen3.5 w/GRPO	73.39	33.27	33.86	**16.05**	50.94	63.12	6.79	39.63
Qwen3.5 w/EMPO	71.94	31.39	37.04	12.88	54.71	66.54	7.48	40.28
Qwen3.5 w/SAGE	**74.61**	**34.02**	**38.25**	15.72	**55.81**	**67.06**	**9.31**	**42.11**
9B models								
Qwen3.5	65.11	14.34	27.31	6.38	39.73	45.59	4.61	29.01
Qwen3.5 w/SFT	77.48	29.73	40.15	24.20	62.93	70.91	9.12	44.93
Qwen3.5 w/GRPO	75.96	40.41	38.82	19.51	56.96	62.77	7.66	43.16
Qwen3.5 w/EMPO	78.24	39.27	37.03	20.88	**64.88**	65.55	8.01	44.84
Qwen3.5 w/SAGE	**79.97**	**41.38**	**41.59**	**25.58**	63.91	**71.72**	**11.47**	**47.95**
35B models								
Qwen3.5	70.28	32.36	48.81	35.05	42.33	76.28	11.36	45.21
Qwen3.5 w/SFT	74.69	38.57	52.96	41.84	49.18	82.73	14.67	50.66
Qwen3.5 w/GRPO	82.18	48.07	65.31	57.94	68.92	91.51	19.48	61.92
Qwen3.5 w/EMPO	81.84	48.39	64.41	**62.31**	67.38	92.29	19.97	62.37
Qwen3.5 w/SAGE	**84.72**	**51.34**	**68.26**	62.04	**70.87**	**94.16**	**22.62**	**64.86**

5.2 Main Results

Table 1 shows SAGE consistently improves outcome-level accuracy across scales without gold reasoning traces or process labels. At the 2B, 9B, and 35B scales, average accuracy improves from 27.24% to 42.11%, 29.01% to 47.95%, and 45.21% to 64.86%, surpassing the strongest baseline by $+1.83$, $+3.02$, and $+2.49$ points respectively. Notably, the 35B SAGE variant (64.86%) substantially exceeds the Llama-3.3-70B-Instruct flagship (38.48%) at roughly half the parameters, with consistent gains on the hardest competition-style benchmarks (Olympiad, AIME24, Putnam).

Free-form Natural Reasoning. The benefits generalize beyond structured formal domains (Table 2). At 9B, SAGE raises MMLU-Pro average from 37.91% to 39.99%, BBH-H from 44.04% to 45.31%, and ARC-C from 39.73% to 42.04%; similar scaling holds at 27B (BBH-H 60.25%, ARC-C 58.11%) and 35B (BBH-H 69.07%, ARC-C 66.41%), all leading among question-only post-training methods. Five-seed mean-std results are in Appendix H.

5.3 Long-Horizon Reasoning

We evaluate the AC problem with two metrics (Yang et al., 2023; Hsiang et al., 2025): *AC Validity* (percentage of syntactically and logically valid steps, a proxy for local precision) and *Lean-Verified Proofs* (success rate of compiler-checked proofs, the gold standard for end-to-end rigor). As shown in Figure 3, SAGE consistently outperforms baselines on both: AC Validity gains of $+19.2\%$ to $+26.0\%$ across architectures, and Lean-Verified gains over $+13\%$ in every case and Qwen3 reaches a nearly 8-fold increase over base. These results are consistent with the SCA prediction that controlling exploration and compounding biases yields more stable long-horizon trajectories.

Table 2: Accuracy (%) on free-form natural reasoning benchmarks. The best is in **bold** with second best in underline.

Model	STEM	MMLU-Pro				GPQA	BBH-H	ARC-C
		Humanity	Social	Other	Avg.			
9B models								
Qwen3.5	12.71	8.02	14.95	10.21	11.06	10.91	21.58	18.49
Qwen3.5 w/SFT	20.18	11.36	28.97	19.22	19.85	12.31	32.44	29.56
Qwen3.5 w/GRPO	33.38	28.31	50.41	39.26	39.33	18.21	41.69	37.82
Qwen3.5 w/EMPO	32.57	27.32	48.73	37.69	37.91	**21.11**	44.04	39.73
Qwen3.5 w/SAGE	**34.11**	**29.08**	**51.17**	**39.71**	**39.99**	20.86	**45.31**	**42.04**
27B models								
Qwen3.6	30.29	24.34	46.55	35.31	35.40	16.09	38.70	34.78
Qwen3.6 w/SFT	34.18	28.51	41.52	37.28	35.77	22.67	45.27	41.12
Qwen3.6 w/GRPO	**57.74**	37.02	**65.16**	57.55	53.24	**34.29**	55.74	51.78
Qwen3.6 w/EMPO	53.96	35.58	60.04	52.18	49.27	29.52	57.19	53.26
Qwen3.6 w/SAGE	56.31	**37.91**	64.43	**58.25**	**53.53**	32.51	**60.25**	**57.11**
35B models								
Qwen3.5	45.08	36.31	52.29	44.28	44.29	31.05	47.52	45.37
Qwen3.5 w/SFT	49.84	38.26	54.47	48.62	47.12	28.97	52.96	49.18
Qwen3.5 w/GRPO	63.58	43.19	69.08	60.71	57.66	35.78	63.29	60.91
Qwen3.5 w/EMPO	61.96	42.02	68.61	59.54	56.72	35.43	65.02	62.97
Qwen3.5 w/SAGE	**65.27**	**44.51**	**71.22**	**62.25**	**59.33**	**38.58**	**69.07**	**66.41**

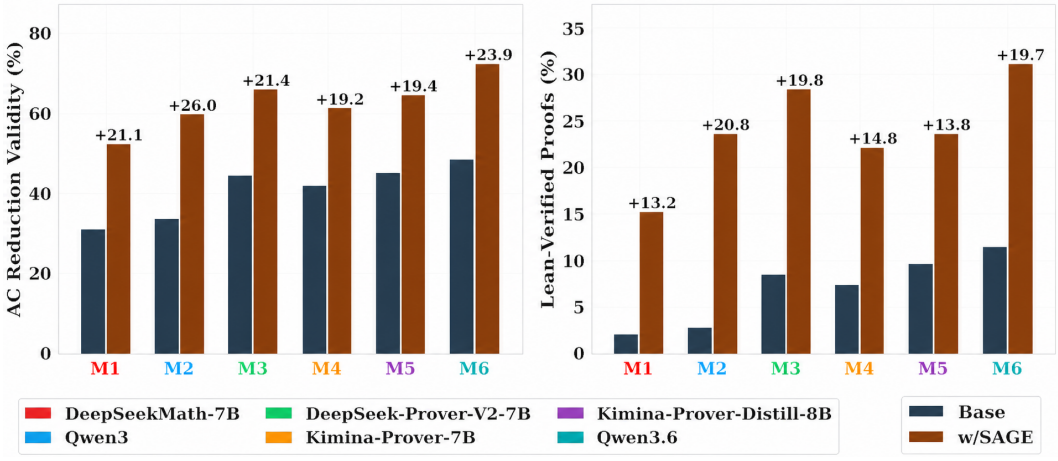


Figure 3: Comparative performance of models with SAGE versus base models across two primary metrics: AC Validity (Left) and Lean-Verified Proofs (Right). Numbers annotated above bars indicate the absolute percentage point improvement.

5.4 Component Ablation Analysis

We isolate the two guidance components: $\Psi_{\mathcal{P}}$ (algebraic sparsification, controlling exploration bias) and $\Psi_{\mathcal{H}}$ (hyperbolic structural guidance, mitigating compounding bias). On Qwen3.5-9B

mathematical and free-form reasoning, removing either degrades performance (Table 3), showing the two signals are complementary rather than redundant. Two controls rule out a generic dense-shaping explanation: replacing $\Psi_{\mathcal{H}}$ with Euclidean distance weakens performance and shuffling target anchors substantially reduces both accuracy and reward density.

Comparison with learned process rewards. Table 4a shows that GRPO-PRM improves over GRPO, confirming that dense process feedback helps under sparse rewards. However, SAGE remains stronger, especially on the AC problem, which demonstrates that the gains are thus not explained by generic dense reward shaping alone, but by target-aligned structural guidance.

Direct bias ablation on AC task. To directly test whether the two components mitigate the symbolic long-horizon failure modes predicted by SCA, we repeat the ablation on AC task. As shown in Table 4b, removing either $\Psi_{\mathcal{P}}$ or $\Psi_{\mathcal{H}}$ reduces AC Validity, AC Path Solving, and Lean-Verified success, while full SAGE achieves the strongest end-to-end verified performance-confirming that both signals are needed when structural validity and long-horizon extension are jointly required.

Table 3: Core ablations on Qwen3.5-9B. All variants share the same entropy-filtered training subset, rollout budget, decoding constraints, optimization steps, and KL coefficient.

Variant	**Olympiad**		**BBH-H**	
	Acc	**Rew./1k**	**Acc**	**Rew./1k**
GRPO	38.82	16.08	41.69	20.06
EMPO	37.03	15.03	44.04	19.02
SAGE w/o $\Psi_{\mathcal{H}}$	39.84	18.47	39.06	22.36
SAGE w/o $\Psi_{\mathcal{P}}$	39.12	17.68	38.85	21.18
SAGE w/ Euclidean	38.01	16.92	38.02	20.89
SAGE w/ Shuffled g	37.99	17.01	37.83	19.91
SAGE	**41.59**	**21.38**	**45.31**	**25.76**

Table 4: Left: comparison with a learned process-reward baseline (Qwen3.5-35B). Right: component ablation on AC task (Qwen3.5-35B).

(a) PRM baseline comparison.

Method	Olym.	BBH-H	Lean
GRPO	65.31	63.29	14.64
GRPO-PRM	63.27	59.12	17.36
SAGE	**68.26**	**69.07**	**23.69**

(b) AC component ablation.

Variant	AC Valid.	AC Path	Lean
GRPO	54.28	23.05	14.64
EMPO	53.18	21.52	13.78
SAGE w/o $\Psi_{\mathcal{H}}$	57.02	27.14	18.82
SAGE w/o $\Psi_{\mathcal{P}}$	56.31	25.98	18.07
SAGE	**59.83**	**31.76**	**23.69**

6 Related Work

Reinforcement Learning and Supervision for Reasoning. Outcome-based RL has become a dominant paradigm for improving LLM reasoning (Ouyang et al., 2022; Shao et al., 2024; Yu et al., 2025), yet its effectiveness degrades as reasoning horizon T grows and terminal feedback becomes sparse (Suo et al., 2025; Lan et al., 2025, 2026). Process-level supervision via verifiers, formal proofs, or step-wise reward models (Yang et al., 2023; Lightman et al., 2024; Yao et al., 2024) mitigates this issue but shifts the bottleneck to annotation cost and verifier coverage.

Exploration Strategies and Optimization Biases. A parallel line of work improves exploration under sparse rewards through entropy regularization, syntactic constraints, empirical consistency, and self-improvement objectives (Zhang et al., 2025a; She et al., 2025; Zhang et al., 2026; Chen et al., 2025; Gai et al., 2025; Lin et al., 2026b). Recent studies further reveal that KL-regularized RL can induce mode collapse through the optimization objective itself (GX-Chen et al., 2026), while latent Poincaré shaping leverages hyperbolic geometry to provide dense process rewards and guide search (Xia et al., 2026). Despite these advances, how logical constraints shape the admissible trajectory space in long-horizon reasoning, and how such structure can be exploited for principled exploration, remains largely underexplored (Wu et al., 2024).

7 Conclusion

In this work, we study long-horizon reasoning under sparse and delayed rewards, showing that its failures arise not only from task difficulty but from two structural biases in outcome-based post-training: exploration bias and compounding bias. We introduce SCA as a theoretical lens for characterizing these biases through prefix-closed feasible regions and for identifying two mitigation requirements: feasible-support concentration and prefix-level structural correction. Building on SCA, we propose SAGE, a topology-guided post-training framework that implements these requirements through algebraic sparsification and hyperbolic structural guidance. Across 12 benchmarks and 10 model backbones, SAGE consistently improves over strong post-training baselines. On the open real-world AC task, SAGE achieves a nearly 8-fold improvement.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This work is supported by the National Science Foundation under Award No. 2339989, No. 2449769, No. 2406439 and No. IIS 2543142; DARPA under contract No. HR00112490370 and No. HR001124S0013, U.S. Department of Homeland Security under Grant Award No. 17STCIN00001-08-00, Amazon-Virginia Tech Initiative for Efficient and Robust Machine Learning, Amazon AWS, Google, Cisco, 4-VA, Commonwealth Cyber Initiative, National Surface Transportation Safety Center for Excellence, and Virginia Tech. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

References

Brantley, K., Chen, M., Gao, Z., Lee, J. D., Sun, W., Zhan, W., and Zhang, X. Accelerating RL for LLM reasoning with optimal advantage regression. In Belgrave, D., Zhang, C., Montoya, L. N., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Ruíz, I. V. M., and Loaiza-Bonilla, A. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/de8e1aebdb1e8a96970a58a641eac3cc-Abstract-Conference.html.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Ségerie, C., Carroll, M., Peng, A., Christoffersen, P. J. K., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A. D., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=bx24KpJ4Eb.

Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., Song, S., and Huang, G. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? In Belgrave, D., Zhang, C., Montoya, L. N., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Ruíz, I. V. M., and Loaiza-Bonilla, A. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/537d5aa768c2d534016a4d06f87bc8fb-Abstract-Conference.html.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018. doi: 10.48550/ARXIV.1803.05457. URL https://doi.org/10.48550/arXiv.1803.05457.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. doi: 10.48550/ARXIV.2110.14168. URL https://doi.org/10.48550/arXiv.2110.14168.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Ren, X., Ettinger, A., Harchaoui, Z., and Choi, Y. Faith and fate: Limits of transformers on compositionality. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/deb3c28192f979302c157cb653c15e90-Abstract-Conference.html.

Gai, J., Zeng, G., Zhang, H., and Raghunathan, A. Differential smoothing mitigates sharpening and improves LLM reasoning. *CoRR*, abs/2511.19942, 2025. doi: 10.48550/ARXIV.2511.19942. URL https://doi.org/10.48550/arXiv.2511.19942.

GLM. GLM-5: from vibe coding to agentic engineering. *CoRR*, abs/2602.15763, 2026. doi: 10.48550/ARXIV.2602.15763. URL https://doi.org/10.48550/arXiv.2602.15763.

Gülçehre, Ç., Paine, T. L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., Siddhant, A., Ahern, A., Wang, M., Gu, C., Macherey, W., Doucet, A., Firat, O., and de Freitas, N. Reinforced self-training (rest) for language modeling. *CoRR*, abs/2308.08998, 2023. doi: 10.48550/ARXIV.2308.08998. URL https://doi.org/10.48550/arXiv.2308.08998.

GX-Chen, A., Prakash, J., Guo, J., Fergus, R., and Ranganath, R. Kl-regularized reinforcement learning for generative modelling is designed to mode collapse. In Vondrick, C., Hariharan, B., Raffel, C., Pinto, L., Yang, D., and Faust, A. (eds.), *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5306f6ce1082d29e4968acdacab84fd7-Abstract-Conference.html.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., and Sun, M. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.211. URL https://doi.org/10.18653/v1/2024.acl-long.211.

Hsiang, R., Adkisson, W., George, R. J., and Anandkumar, A. Leandojo-v2: A comprehensive library for ai-assisted theorem proving in lean. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*, 2025.

Hugging Face H4. Aime 2024 dataset. https://huggingface.co/datasets/HuggingFaceH4/aime_2024, 2025.

Jahin, A., Zidan, A. H., Zhang, W., Bao, Y., and Liu, T. Evaluating mathematical reasoning across large language models: A fine-grained approach, 2025. URL https://arxiv.org/abs/2503.10573.

Lan, G., Inan, H. A., Abdelnabi, S., Kulkarni, J., Wutschitz, L., Shokri, R., Brinton, C., and Sim, R. Contextual integrity in llms via reasoning and reinforcement learning. In Belgrave, D., Zhang, C., Montoya, L. N., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Ruíz, I. V. M., and Loaiza-Bonilla, A. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/968591afcfe4ba889eccffb6a4ca2d45-Abstract-Conference.html.

Lan, G., Xiong, L., Zhou, X., Cui, H., Zhang, Y., Li, M., Shi, Z., Fetahu, B., Li, L., and Li, X. Alternating reinforcement learning with contextual rubric rewards: Beyond the scalarization strategy. *arXiv preprint arXiv:2603.15646*, 2026.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V. V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. Solving quantitative reasoning problems with language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract-Conference.html.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Lin, J., Zeng, X., Zhu, J., Wang, S., Shun, J., Wu, J., and Zhou, D. Plan and budget: Effective and efficient test-time scaling on reasoning large language models. In Vondrick, C., Hariharan, B., Raffel, C., Pinto, L., Yang, D., and Faust, A. (eds.), *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ae8d4084f418bb51575c2ca6c658a05b-Abstract-Conference.html.

Lin, Z., Zhao, K., Zhang, S., Yu, P., and Xiao, C. Cec-zero: Zero-supervision character error correction with self-generated rewards. In Koenig, S., Jenkins, C., and Taylor, M. E. (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 23612–23620. AAAI Press, 2026b. doi: 10.1609/AAAI.V40I28.39534. URL https://doi.org/10.1609/aaai.v40i28.39534.

Liu, H., Ding, Y., Fu, Z., Zhang, C., Liu, X., and Zhang, Y. Evaluating the logical reasoning abilities of large reasoning models. *CoRR*, abs/2505.11854, 2025. doi: 10.48550/ARXIV.2505.11854. URL https://doi.org/10.48550/arXiv.2505.11854.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lyu, C., Gao, S., Gu, Y., Zhang, W., Gao, J., Liu, K., Wang, Z., Li, S., Zhao, Q., Huang, H., Cao, W., Liu, J., Liu, H., Liu, J., Zhang, S., Lin, D., and Chen, K. Exploring the limit of outcome reward for learning mathematical reasoning. *CoRR*, abs/2502.06781, 2025. doi: 10.48550/ARXIV.2502.06781. URL https://doi.org/10.48550/arXiv.2502.06781.

Math AI. Amc23 dataset. https://huggingface.co/datasets/math-ai/amc23, 2025.

MiniMax. The minimax-m2 series: Mini activations unleashing max real-world intelligence. *CoRR*, abs/2605.26494, 2026. doi: 10.48550/ARXIV.2605.26494. URL https://doi.org/10.48550/arXiv.2605.26494.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Patterson, D. A., Gonzalez, J., Hölzle, U., Le, Q. V., Liang, C., Munguia, L., Rothchild, D., So, D. R., Texier, M., and Dean, J. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022. doi: 10.1109/MC.2022.3148714. URL https://doi.org/10.1109/MC.2022.3148714.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48550/arXiv.2311.12022.

Ren, Z. Z., Shao, Z., Song, J., Xin, H., Wang, H., Zhao, W., Zhang, L., Fu, Z., Zhu, Q., Yang, D., Wu, Z. F., Gou, Z., Ma, S., Tang, H., Liu, Y., Gao, W., Guo, D., and Ruan, C. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *CoRR*, abs/2504.21801, 2025. doi: 10.48550/ARXIV.2504.21801. URL https://doi.org/10.48550/arXiv.2504.21801.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. doi: 10.48550/ARXIV.1707.06347. URL https://doi.org/10.48550/arXiv.1707.06347.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

She, S., Liu, J., Liu, Y., Chen, J., Huang, X., and Huang, S. R-PRM: reasoning-driven process reward modeling. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 13438–13451. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.679. URL https://doi.org/10.18653/v1/2025.emnlp-main.679.

Shehper, A., Medina-Mardones, A. M., Fagan, L., Lewandowski, B., Gruen, A., Qiu, Y., Kucharski, P., Wang, Z., and Gukov, S. What makes math problems hard for reinforcement learning: A case study. In Belgrave, D., Zhang, C., Montoya, L. N., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Ruíz, I. V. M., and Loaiza-Bonilla, A. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/ea8c9cd935dc62e0288bbf33a15e65e9-Abstract-Conference.html.

Sullivan, M. and Koller, A. GRPO is secretly a process reward model. *CoRR*, abs/2509.21154, 2025. doi: 10.48550/ARXIV.2509.21154. URL https://doi.org/10.48550/arXiv.2509.21154.

Suo, Y., Ma, F., Shen, K., Zhu, L., and Yang, Y. Long-horizon visual instruction generation with logic and attribute self-reflection. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=EdMb9TqqDY.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.824. URL https://doi.org/10.18653/v1/2023.findings-acl.824.

Team, K. Kimi K2.5: visual agentic intelligence. *CoRR*, abs/2602.02276, 2026a. doi: 10.48550/ARXIV.2602.02276. URL https://doi.org/10.48550/arXiv.2602.02276.

Team, Q. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Team, Q. Qwen3.5-omni technical report. *CoRR*, abs/2604.15804, 2026b. doi: 10.48550/ARXIV.2604.15804. URL https://doi.org/10.48550/arXiv.2604.15804.

Tropp, J. A. and Gilbert, A. C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory*, 53(12):4655–4666, 2007. doi: 10.1109/TIT.2007.909108. URL https://doi.org/10.1109/TIT.2007.909108.

Tsoukalas, G., Lee, J., Jennings, J., Xin, J., Ding, M., Jennings, M., Thakur, A., and Chaudhuri, S. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/1582eaf9e0cf349e1e5a6ee453100aa1-Abstract-Datasets_and_Benchmarks_Track.html.

Uesato, J., Kushman, N., Kumar, R., Song, H. F., Siegel, N. Y., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process- and outcome-based feedback. *CoRR*, abs/2211.14275, 2022. doi: 10.48550/ARXIV.2211.14275. URL https://doi.org/10.48550/arXiv.2211.14275.

Wang, H., Unsal, M., Lin, X., Baksys, M., Liu, J., Santos, M. D., Sung, F., Vinyes, M., Ying, Z., Zhu, Z., Lu, J., de Saxcé, H., Bailey, B., Song, C., Xiao, C., Zhang, D., Zhang, E., Pu, F., Zhu, H., Liu, J., Bayer, J., Michel, J., Yu, L., Dreyfus-Schmidt, L., Tunstall, L., Pagani, L., Machado, M., Bourigault, P., Wang, R., Polu, S., Barroyer, T., Li, W., Niu, Y., Fleureau, Y., Hu, Y., Yu, Z., Wang, Z., Yang, Z., Liu, Z., and Li, J. Kimina-prover preview: Towards large formal reasoning models with reinforcement learning. *CoRR*, abs/2504.11354, 2025. doi: 10.48550/ARXIV.2504.11354. URL https://doi.org/10.48550/arXiv.2504.11354.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

Weaver, L. and Tao, N. The optimal reward baseline for gradient-based reinforcement learning. In Breese, J. S. and Koller, D. (eds.), *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, pp. 538–545. Morgan Kaufmann, 2001. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=141&proceeding_id=17.

Wu, F., Zhang, R., Yi, Q., Gao, Y., Guo, J., Peng, S., Lan, S., Han, H., Pan, Y., Yuan, K., Jin, P., Chen, R., Chen, Y., and Li, L. OCEAN-MBRL: offline conservative exploration for model-based offline reinforcement learning. In Wooldridge, M. J., Dy, J. G., and Natarajan, S. (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 15897–15905. AAAI Press, 2024. doi: 10.1609/AAAI.V38I14.29520. URL https://doi.org/10.1609/aaai.v38i14.29520.

Xia, H., Chen, B., Zang, Z., Ge, Y., Zhao, G., and Zhu, S. Latent poincaré shaping for agentic reinforcement learning. *CoRR*, abs/2602.09375, 2026. doi: 10.48550/ARXIV.2602.09375. URL https://doi.org/10.48550/arXiv.2602.09375.

Yang, K., Swope, A. M., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., and Anandkumar, A. Leandojo: Theorem proving with retrieval-augmented language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/4441469427094f8873d0fecb0c4e1cee-Abstract-Datasets_and_Benchmarks.html.

Yao, J., Li, C., and Xiao, C. Swift sampler: Efficient learning of sampler by 10 parameters. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6c9dcffe0b9cc3b05d83bcdddb250690-Abstract-Conference.html.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, J., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, R., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W., Zhang, Y., Yan, L., Wu, Y., and Wang, M. DAPO: an open-source LLM reinforcement learning system at scale. In Belgrave, D., Zhang, C., Montoya, L. N., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Ruíz, I. V. M., and Loaiza-Bonilla, A. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. D. Star: Bootstrapping reasoning with reasoning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Zeng, X., Wang, H., Lin, J., Wu, J., Cody, T., and Zhou, D. Lensllm: Unveiling fine-tuning dynamics for LLM selection. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Forty-second International Conference on*

Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zeng25g.html.

Zeng, X., Lin, J., Yan, Y., Guo, F., Shi, L., Wu, J., and Zhou, D. Halluguard: Demystifying data-driven and reasoning-driven hallucinations in llms. In Vondrick, C., Hariharan, B., Raffel, C., Pinto, L., Yang, D., and Faust, A. (eds.), *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8c99af22c5becbbb4fa58ebf6e2dc2f6-Abstract-Conference.html.

Zhang, Q., Wu, H., Zhang, C., Zhao, P., and Bian, Y. Right question is already half the answer: Fully unsupervised LLM reasoning incentivization. In Belgrave, D., Zhang, C., Montoya, L. N., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Ruíz, I. V. M., and Loaiza-Bonilla, A. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025a. URL http://papers.nips.cc/paper_files/paper/2025/hash/614f8eba720cfc7ff00274bd64fb0a3f-Abstract-Conference.html.

Zhang, X., Li, R., Zhou, Z., Li, L., Qin, Y., Li, K., Sun, X., Tan, X., Qu, C., and Qi, Y. Count counts: Motivating exploration in LLM reasoning with count-based intrinsic rewards. In Vondrick, C., Hariharan, B., Raffel, C., Pinto, L., Yang, D., and Faust, A. (eds.), *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/af3bcaace529a7fe101217ec0ed42bd2-Abstract-Conference.html.

Zhang, Z., Xu, J., He, Z., Liang, T., Liu, Q., Li, Y., Song, L., Liang, Z., Zhang, Z., Wang, R., Tu, Z., Mi, H., and Yu, D. Deeptheorem: Advancing LLM reasoning for theorem proving through natural language and reinforcement learning. *CoRR*, abs/2505.23754, 2025b. doi: 10.48550/ARXIV.2505.23754. URL https://doi.org/10.48550/arXiv.2505.23754.

Zhou, Y., Ye, J., Ling, Z., Han, Y., Huang, Y., Zhuang, H., Liang, Z., Guo, K., Guo, T., Wang, X., and Zhang, X. Dissecting logical reasoning in llms: A fine-grained evaluation and supervision study. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 17075–17098. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.926. URL https://doi.org/10.18653/v1/2025.findings-emnlp.926.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P. F., and Irving, G. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. doi: 10.48550/ARXIV.1909.08593. URL https://doi.org/10.48550/arXiv.1909.08593.

A Formal Properties of Symbolic Closure Analysis

We restate the formal object used by Symbolic Closure Analysis (SCA). SCA is an analytical lens for characterizing feasible-support concentration under local admissibility. It is not itself an inference-time search procedure.

Let $\mathfrak{S}$ be a domain-specific symbolic interface that specifies admissible operators and a local admissibility predicate. A trajectory $\tau = \{(v_t, op_t)\}_{t=1}^{T}$ belongs to the SCA-feasible region $\mathcal{F}$ iff every local transition is admissible:

$$\tau \in \mathcal{F} \quad \Longleftrightarrow \quad \forall t,\ \mathrm{Adm}_{\mathfrak{S}}\big(\rho(v_t),\ op_t,\ v_t\big) = 1. \tag{6}$$

Here $\rho(v_t)$ denotes the predecessor context of v_t, and op_t is the operator used to obtain v_t.

Local admissibility is not process supervision. The predicate $\mathrm{Adm}_{\mathfrak{S}}$ checks only local well-formedness or rule consistency under the task interface $\mathcal{S}$. It does not provide ground-truth solution steps, does not reveal successful trajectories, and does not use terminal outcome labels. Thus, SCA separates local admissibility from global correctness.

Prefix closure. The feasible region $\mathcal{F}$ is prefix-closed by construction. If a prefix violates local admissibility, then any continuation of that prefix remains outside $\mathcal{F}$. Equivalently, if $\tau \in \mathcal{F}$, then every prefix $\tau_{\le t}$ also lies in $\mathcal{F}$.

Lemma A.1 (Prefix closure under local admissibility)**.** *Let $\tau = \{(v_t, op_t)\}_{t=1}^{T}$ and suppose that for some step t,*

$$\mathrm{Adm}_{\mathfrak{S}}\big(\rho(v_t), op_t, v_t\big) = 0.$$

Then every trajectory τ' that contains the same prefix up to step t satisfies $\tau' \notin \mathcal{F}$.

Proof. By Definition (6), membership in $\mathcal{F}$ requires every transition to satisfy local admissibility. If the step-t transition violates $\mathrm{Adm}_{\mathfrak{S}}$, then the conjunction in (6) fails. Any continuation that preserves this invalid prefix also contains the same failed transition, and therefore cannot belong to $\mathcal{F}$. □

Feasible but unsuccessful trajectories. Local admissibility does not imply terminal success. Let $\mathcal{T}^* \subseteq \Omega$ denote the set of globally successful trajectories, such as trajectories that produce a correct final answer or a compiler-checked proof. We assume $\mathcal{T}^* \subseteq \mathcal{F}$, but $\mathcal{F}$ may contain many feasible-yet-unsuccessful trajectories. Define

$$\mathcal{B} := \mathcal{F} \setminus \mathcal{T}^*.$$

For a rollout distribution π supported primarily on $\mathcal{F}$, define the success density inside the feasible region as

$$p_{\mathcal{F}} := \Pr_{\tau\sim\pi}\big(\tau \in \mathcal{T}^* \mid \tau \in \mathcal{F}\big),$$

and the feasible-but-failing mass as

$$\beta_{\mathcal{F}} := \Pr_{\tau\sim\pi}\big(\tau \in \mathcal{B} \mid \tau \in \mathcal{F}\big) = 1 - p_{\mathcal{F}}.$$

In intrinsically complex long-horizon tasks, $\beta_{\mathcal{F}}$ can remain large even when local admissibility is high. This explains why improving step-level validity alone is insufficient: SAGE must also provide depth-wise structural guidance that helps feasible prefixes extend toward successful trajectories.

Relation to SAGE. SAGE implements the requirements identified by SCA through soft training-time guidance. Rather than imposing a hard projection onto $\mathcal{F}$, SAGE uses structural potentials to bias rollout sampling and reward shaping:

$$\Psi_{\mathrm{SAGE}}(s_t, a_t) = \alpha\Psi_P(s_t, a_t) + \gamma\Psi_H(s_t, a_t).$$

This produces a training-time rollout distribution that increases probability mass on structurally informative trajectories while preserving direct inference with the trained policy. In domains where an exact local checker is available, hard validity masking can be viewed as a limiting or diagnostic variant, but it is not required by the general SAGE framework and is not used at inference time.

B Proof of Theorem 3.5

We prove the total-variation bound for the KL-regularized optimizer under sparse-reward regimes.

Theorem 3.5 (Reference anchoring under rare terminal rewards)**.** *For Section 3.2 with terminal-only $R(\tau) \in [0, R_{\max}]$, let $S = \{\tau : R(\tau) > 0\}$ and $p = \mathbb{P}_{\tau\sim\pi_{\mathrm{ref}}}(S)$. The full-support optimizer $\pi^*_\Omega(\tau) = \pi_{\mathrm{ref}}(\tau)\exp(R(\tau)/\lambda)/\mathbb{E}_{\tau'\sim\pi_{\mathrm{ref}}}[\exp(R(\tau')/\lambda)]$ satisfies*

$$D_{\mathrm{TV}}(\pi^*_\Omega, \pi_{\mathrm{ref}}) \le \big(e^{R_{\max}/\lambda} - 1\big)p. \tag{3}$$

Hence if $(e^{R_{\max}/\lambda} - 1)p \ll 1$, the optimizer remains close to π_{ref}. In the high-KL or weak-reward regime $R_{\max} \ll \lambda$, this becomes $D_{\mathrm{TV}}(\pi^*_\Omega, \pi_{\mathrm{ref}}) \le (R_{\max}/\lambda)p + O(R^2_{\max}p/\lambda^2)$. Thus, when successful trajectories are rare under the reference, terminal-only KL-regularized optimization has limited leverage to move probability mass away from reference-likely prefixes.

Proof. The optimizer of the KL-regularized objective is the exponentially tilted distribution

$$\pi^*_\Omega(\tau) = \frac{\pi_{\mathrm{ref}}(\tau)\exp(R(\tau)/\lambda)}{Z}, \qquad Z = \mathbb{E}_{\tau\sim\pi_{\mathrm{ref}}}\left[\exp(R(\tau)/\lambda)\right].$$

Define

$$h(\tau) = \exp(R(\tau)/\lambda) - 1.$$

Since $R(\tau) = 0$ for $\tau \notin S$ and $R(\tau) \in [0, R_{\max}]$, we have

$$0 \le h(\tau) \le e^{R_{\max}/\lambda} - 1, \qquad h(\tau) = 0 \text{ for } \tau \notin S.$$

Let

$$H = \mathbb{E}_{\pi_{\mathrm{ref}}}[h(\tau)].$$

Then

$$0 \le H \le \big(e^{R_{\max}/\lambda} - 1\big)p, \qquad Z = 1 + H.$$

Therefore,

$$\pi^*_\Omega(\tau) - \pi_{\mathrm{ref}}(\tau) = \pi_{\mathrm{ref}}(\tau)\left(\frac{1 + h(\tau)}{1 + H} - 1\right) = \pi_{\mathrm{ref}}(\tau)\frac{h(\tau) - H}{1 + H}.$$

Hence

$$D_{\mathrm{TV}}(\pi^*_\Omega, \pi_{\mathrm{ref}}) = \frac{1}{2}\mathbb{E}_{\pi_{\mathrm{ref}}}\left[\frac{|h(\tau) - H|}{1 + H}\right] \le \frac{1}{2(1 + H)}\left(\mathbb{E}_{\pi_{\mathrm{ref}}}[h(\tau)] + H\right) = \frac{H}{1 + H} \le H.$$

Using the bound on H gives

$$D_{\mathrm{TV}}(\pi^*_\Omega, \pi_{\mathrm{ref}}) \le \big(e^{R_{\max}/\lambda} - 1\big)p.$$

Finally, when $R_{\max} \ll \lambda$,

$$e^{R_{\max}/\lambda} - 1 = \frac{R_{\max}}{\lambda} + O\left(\frac{R^2_{\max}}{\lambda^2}\right),$$

which gives the stated asymptotic scaling. □

C Recovery Guarantee for the Greedy Sparse Locator

This appendix provides a standard recovery guarantee for the greedy sparse locator used to instantiate algebraic sparsification in tangent-space coordinates. The result supports the intuition that, when the residual admits a sparse decomposition over operator-indexed directions, greedy projection can identify structurally relevant operators under standard coherence conditions.

C.1 Setup

Let $D = [d_1, \dots, d_N] \in \mathbb{R}^{d \times N}$ be a dictionary with normalized atoms $\|d_j\|_2 = 1$. Assume that a residual vector $r \in \mathbb{R}^d$ has a k-sparse representation

$$r = D_I \alpha_I,$$

where $I \subseteq [N]$ is the support with $|I| = k$. Define the mutual coherence

$$\mu := \max_{i \neq j} |\langle d_i, d_j \rangle|.$$

The greedy sparse locator selects atoms by maximum correlation with the current residual and then orthogonally projects, matching the classical Orthogonal Matching Pursuit (OMP) procedure.

Theorem C.1 (OMP support recovery under coherence (Tropp & Gilbert, 2007))**.** *Let $D = [d_1, \dots, d_N] \in \mathbb{R}^{m \times N}$ be a dictionary with unit-norm columns, and let*

$$\mu(D) = \max_{p \neq q} |\langle d_p, d_q \rangle|$$

denote its mutual coherence. Suppose $y = D\alpha$ is noiseless and α is k-sparse with support I. If

$$\mu(D) < \frac{1}{2k-1},$$

then OMP run for k iterations recovers the exact support I.

Proof. We use the standard exact recovery condition (ERC) for OMP. For a fixed support I, OMP exactly recovers every signal supported on I in the noiseless setting if

$$\max_{j \notin I} \|D_I^\dagger d_j\|_1 < 1,$$

where D_I is the subdictionary indexed by I and $D_I^\dagger = (D_I^\top D_I)^{-1} D_I^\top$.

It remains to show that the mutual coherence condition implies this ERC. Let

$$G_I = D_I^\top D_I.$$

Since the columns of D are normalized, G_I has diagonal entries equal to 1 and off-diagonal entries bounded in absolute value by $\mu(D)$. Hence

$$\|I - G_I\|_1 \leq (k-1)\mu(D).$$

Under $\mu(D) < 1/(2k-1)$, we have $(k-1)\mu(D) < 1$, so G_I is invertible, and the Neumann-series bound gives

$$\|G_I^{-1}\|_1 \leq \frac{1}{1-(k-1)\mu(D)}.$$

For any $j \notin I$,

$$\|D_I^\top d_j\|_1 \leq k\mu(D).$$

Therefore,

$$\|D_I^\dagger d_j\|_1 = \|(D_I^\top D_I)^{-1} D_I^\top d_j\|_1 \leq \frac{k\mu(D)}{1-(k-1)\mu(D)}.$$

The condition $\mu(D) < 1/(2k-1)$ is equivalent to

$$\frac{k\mu(D)}{1-(k-1)\mu(D)} < 1.$$

Thus the ERC holds. By the standard OMP exact recovery theorem (Tropp & Gilbert, 2007), OMP selects atoms from the true support at every iteration and recovers I after k iterations. □

The theorem is not a guarantee that SAGE globally solves the reasoning task. It only justifies the algebraic sparsification step under a standard sparse-residual model: when the unresolved residual is concentrated on a small number of operator-aligned directions, greedy projection can identify the relevant structural directions. This supports using Ψ_P as a soft compatibility score for feasible-support concentration.

D Soft Feasible-Support Concentration

This section analyzes an idealized trajectory-level reweighting induced by SAGE. The result complements Proposition 4.2 in the main paper and formalizes how a structural potential suppresses locally inadmissible rollout mass when it separates feasible and infeasible trajectories. This is a conditional concentration result, not a universal guarantee.

Trajectory-level reweighting. Let $\mathcal{T}$ denote the discrete set of trajectories and let $\pi_0(\tau \mid q)$ be a base rollout distribution. We consider the trajectory-level reweighted distribution

$$\pi_\lambda(\tau \mid q) = \frac{\pi_0(\tau \mid q)\exp\big(\lambda\Psi(\tau,q)\big)}{Z_\lambda(q)}, \qquad Z_\lambda(q) = \sum_{\tau\in\mathcal{T}} \pi_0(\tau \mid q)\exp\big(\lambda\Psi(\tau,q)\big), \tag{7}$$

where $\Psi(\tau, q)$ denotes the trajectory-level SAGE potential and $\lambda \geq 0$ controls guidance strength.

Feasible and infeasible sets. Let

$$\mathcal{T}_{\mathrm{bad}}(q) := \{\tau \in \mathcal{T} : \tau \notin \mathcal{F}(q)\}, \qquad \mathcal{T}_{\mathrm{good}}(q) := \mathcal{T} \setminus \mathcal{T}_{\mathrm{bad}}(q).$$

Define the base invalid mass as

$$p_0(q) := \Pr_{\tau\sim\pi_0}\big(\tau \in \mathcal{T}_{\mathrm{bad}}(q)\big).$$

Relative margin condition. Assume that the structural potential separates feasible and infeasible trajectories by a relative gap:

$$\inf_{\tau\in\mathcal{T}_{\mathrm{good}}(q)} \Psi(\tau,q) - \sup_{\tau\in\mathcal{T}_{\mathrm{bad}}(q)} \Psi(\tau,q) \geq m. \tag{8}$$

Proposition D.1 (Soft suppression under a relative structural margin)**.** *Under Equations* (7) *and* (8)*,*

$$\Pr_{\tau\sim\pi_\lambda}\big(\tau \in \mathcal{T}_{\mathrm{bad}}(q)\big) \leq \frac{p_0(q)e^{-\lambda m}}{1 - p_0(q) + p_0(q)e^{-\lambda m}}.$$

Proof. Let

$$b(q) = \sup_{\tau\in\mathcal{T}_{\mathrm{bad}}(q)} \Psi(\tau,q), \qquad \widetilde{\Psi}(\tau,q) = \Psi(\tau,q) - b(q).$$

This additive shift does not change π_λ, because the factor $\exp(-\lambda b(q))$ cancels between the numerator and the normalizing constant. By the relative margin assumption,

$$\widetilde{\Psi}(\tau,q) \leq 0 \quad \text{for } \tau \in \mathcal{T}_{\mathrm{bad}}(q),$$

and

$$\widetilde{\Psi}(\tau,q) \geq m \quad \text{for } \tau \in \mathcal{T}_{\mathrm{good}}(q).$$

Therefore,

$$\sum_{\tau\in\mathcal{T}_{\mathrm{bad}}(q)} \pi_0(\tau \mid q)\exp(\lambda\widetilde{\Psi}(\tau,q)) \leq p_0(q),$$

while

$$\sum_{\tau\in\mathcal{T}_{\mathrm{good}}(q)} \pi_0(\tau \mid q)\exp(\lambda\widetilde{\Psi}(\tau,q)) \geq (1-p_0(q))e^{\lambda m}.$$

Hence

$$\Pr_{\pi_\lambda}\big(\mathcal{T}_{\mathrm{bad}}(q) \mid q\big) \leq \frac{p_0(q)}{p_0(q) + (1-p_0(q))e^{\lambda m}}.$$

Multiplying the numerator and denominator by $e^{-\lambda m}$ gives

$$\Pr_{\pi_\lambda}\big(\mathcal{T}_{\mathrm{bad}}(q) \mid q\big) \leq \frac{p_0(q)e^{-\lambda m}}{1 - p_0(q) + p_0(q)e^{-\lambda m}}.$$

□

The bound shows that the infeasible mass is suppressed exponentially in the guidance strength λ when the structural potential separates feasible and infeasible trajectories by a relative margin. The result does not require bad trajectories to receive negative potential values; it is invariant to additive shifts of Ψ.

E Experimental Details

This appendix provides additional implementation details for SAGE. We describe the task-specific instantiation of the structural potentials, the rollout sampling and reward construction used during post-training, and the auxiliary probes used to construct structural signals in non-symbolic tasks. All structural modules described below are used only during post-training. At inference time, we use the trained policy directly without evaluating Ψ_P, Ψ_H, residual probes, semantic clusters, or local checkers.

E.1 Task-Specific Instantiation

SAGE instantiates the two structural requirements identified by SCA: feasible-support concentration for mitigating exploration bias and prefix-level structural correction for mitigating compounding bias. The concrete implementation depends on the task interface.

AC Task. The AC task setting provides an explicit local admissibility interface. The state s_t is the current group presentation after the generated prefix, and the action a_t is an AC move. The residual r_t represents unresolved algebraic structure in the current presentation, including generator counts, relator lengths, and unresolved relator components. The target anchor g is the task-specified trivial presentation. In this setting, algebraic sparsification scores the compatibility between a candidate move and the unresolved algebraic residual, while hyperbolic structural guidance measures progress toward the target presentation in a geometry suited to tree-like long-horizon transformations. This is the domain where SCA gives an exact symbolic instantiation through the local admissibility interface.

Closed-form mathematical reasoning. For mathematical reasoning, we do not claim an exact SCA guarantee. Instead, SAGE uses the SCA requirements as design principles for learned training-time structural priors. The state is the current solution prefix concatenated with the original problem. The residual r_t is estimated from unresolved symbolic and semantic constraints extracted from the problem statement and the current prefix. These constraints include problem entities, variables, equation structure, operation category, and answer-schema status. The target anchor g is prompt-conditioned and derived from the expected constraint structure of the task, not from gold solution traces, gold rationales, test labels, or terminal correctness labels.

Free-form natural reasoning. For free-form reasoning, the state is the partial rationale or response prefix concatenated with the original prompt. The residual r_t captures unresolved prompt requirements and prefix-level semantic structure. The target anchor g is a prompt-conditioned semantic anchor estimated from training-rollout representations. It is not derived from test labels, gold answers, gold rationales, or ground-truth solution traces. We instantiate SCA's structural principles via learned proxies; their fidelity to the symbolic guarantees is empirical.

For non-symbolic reasoning tasks, SAGE does not normalize over the full token vocabulary or the full space of textual continuations. During post-training, each action is a candidate reasoning step proposed by the old policy. We sample a finite candidate set and reweight candidates using length-normalized policy log-probability and structural potentials. This finite-candidate procedure is used only for rollout generation during training. At inference time, all models decode directly from the trained policy without structural scoring or candidate reweighting.

To isolate the effect of structural guidance from prompt-selection effects, all ablations and dense-reward comparisons use a fixed entropy-filtered training subset constructed once from reference-policy rollouts. GRPO, EMPO, PRM-GRPO, and SAGE are trained on the same retained prompts under the same rollout budgets and optimization schedule. All test results are computed on the complete benchmark test sets.

E.2 Training-Time Rollout Sampling and Reward Construction

For symbolic Andrews-Curtis reasoning, an action a_t is a discrete Andrews-Curtis move provided by the task interface. For closed-form mathematical reasoning and free-form natural reasoning, an action a_t denotes a coarse candidate reasoning step rather than a single token. In practice, a candidate step is generated by $\pi_{\theta_{\text{old}}}$ until a step delimiter, sentence boundary, final-answer marker, end-of-sequence token, or a maximum step length L_{step} is reached.

At each state s_t, we first sample a finite candidate set

$$\mathcal{C}_t = \{a_t^{(1)}, \ldots, a_t^{(K)}\}$$

from $\pi_{\theta_{\text{old}}}(\cdot \mid s_t)$. We then evaluate the structural potentials only on this finite candidate set and sample the next step according to

$$P_{\text{sample}}(a_t^{(k)} \mid s_t, \mathcal{C}_t) = \frac{\exp\Big(\bar{\ell}_{\theta_{\text{old}}}(a_t^{(k)} \mid s_t) + \lambda \Psi_{SAGE}(s_t, a_t^{(k)})\Big)}{\sum_{k'=1}^{K} \exp\Big(\bar{\ell}_{\theta_{\text{old}}}(a_t^{(k')} \mid s_t) + \lambda \Psi_{SAGE}(s_t, a_t^{(k')})\Big)},$$

where

$$\bar{\ell}_{\theta_{\text{old}}}(a_t^{(k)} \mid s_t) = \frac{1}{|a_t^{(k)}|} \sum_{u=1}^{|a_t^{(k)}|} \log \pi_{\theta_{\text{old}}}\big(a_{t,u}^{(k)} \mid s_t, a_{t,<u}^{(k)}\big)$$

is the length-normalized log-probability of the candidate step. Length normalization prevents the finite-candidate reweighting rule from favoring shorter steps solely because of token-product probability effects.

The SAGE potential is

$$\Psi_{SAGE}(s_t, a_t) = \alpha \Psi_P(s_t, a_t) + \gamma \Psi_H(s_t, a_t).$$

This finite-candidate sampler is used only during post-training rollout generation. It is not an exact normalization over the full space of textual actions or the full token vocabulary. We use the AdamW optimizer (Loshchilov & Hutter, 2019) if applicable. At inference time, we use the trained policy directly without candidate-step reweighting or structural scoring.

Algorithm 1 SAGE-Guided Policy Optimization

Require: Training prompts $\mathcal{D}_{\text{filt}}$, old policy $\pi_{\theta_{\text{old}}}$, reference policy π_{ref}, frozen structural potentials Ψ_P, Ψ_H, terminal reward R_{term}, weights $\alpha, \gamma, \lambda, \eta$

for each training minibatch $\mathcal{B} \subseteq \mathcal{D}_{\text{filt}}$ **do**
  for each problem $q \in \mathcal{B}$ **do**
    for $i = 1, \ldots, G$ **do**
      Initialize state s_0 from q
      for $t = 0, \ldots, T-1$ until termination **do**
        Sample K candidate steps $C_t \sim \pi_{\theta_{\text{old}}}(\cdot \mid s_t)$
        for each $a \in C_t$ **do**
          Compute $\Psi_P(s_t, a)$ and $\Psi_H(s_t, a)$
          $\Psi_{\text{SAGE}}(s_t, a) \leftarrow \alpha \Psi_P(s_t, a) + \gamma \Psi_H(s_t, a)$
          $u_t(a) \leftarrow \bar{\ell}_{\theta_{\text{old}}}(a \mid s_t) + \lambda \Psi_{\text{SAGE}}(s_t, a)$
        end for
        Sample a_t from C_t with probabilities proportional to $\exp(u_t(a))$
        $s_{t+1} \leftarrow s_t \circ a_t$
      end for
      Construct trajectory τ_i
      $\widetilde{R}(\tau_i) \leftarrow R_{\text{term}}(\tau_i) + \frac{\eta}{T_i} \sum_{t=0}^{T_i-1} \Psi_{\text{SAGE}}(s_t, a_t)$
    end for
    Compute group-relative advantages
    $A_i \leftarrow \frac{\widetilde{R}(\tau_i) - \text{mean}_j\, \widetilde{R}(\tau_j)}{\text{std}_j\, \widetilde{R}(\tau_j) + \epsilon}$ for $i = 1, \ldots, G$
  end for
  Update π_θ using the clipped group-relative objective with KL regularization in Section 4.2
end for

Training and Inference Distinction. Algorithm 1 summarizes the training-time integration of structural guidance. The two potentials are evaluated only on finite sets of candidate reasoning steps sampled during rollout generation. Their scores affect both candidate selection and group-relative advantages, allowing the policy to internalize structural preferences during optimization. At inference time, the trained policy generates responses directly without additional candidate reweighting, structural-potential computation, or external search.

E.3 State Encoder and Hyperbolic Embedding

For each intermediate state s_t, we first construct a canonical textual representation $x(s_t)$. In AC task, $x(s_t)$ is the canonicalized group presentation after applying the generated prefix up to step t. In mathematical and free-form reasoning tasks, $x(s_t)$ is the generated reasoning prefix concatenated with the original prompt.

Let $\mathcal{E}$ denote the frozen reference-model encoder used to featurize intermediate states. Unless otherwise stated, we use the hidden state from layer ℓ_{enc} at the final generated token:

$$h_t = \mathcal{E}_{\ell_{\text{enc}}}(x(s_t)) \in \mathbb{R}^{d_h}.$$

The vector h_t is projected into a lower-dimensional structural representation by a fixed linear map $W_E \in \mathbb{R}^{d_E \times d_h}$ fitted only on training-rollout states:

$$z_t = W_E h_t.$$

In our implementation, W_E is fitted without correctness supervision using training-rollout representations only. It is fixed before guided rollout generation and is not updated during policy optimization. We then map z_t into the Poincaré ball using radial projection:

$$E(s_t) = \frac{\tanh(\sqrt{c}\|z_t\|_2)}{\sqrt{c}\|z_t\|_2} z_t,$$

where $c > 0$ is the curvature parameter. The same encoder and projection are used for the task anchor g:

$$E(g) = \frac{\tanh(\sqrt{c}\|W_E h_g\|_2)}{\sqrt{c}\|W_E h_g\|_2} W_E h_g.$$

For AC task, g is the canonical trivial presentation. For mathematical and free-form tasks, g is a prompt-conditioned structural anchor derived from the input format and training-rollout representations, not from test labels, gold rationales, or terminal correctness labels.

The hyperbolic structural potential is

$$\Psi_H(s_t, a_t, g) = \exp\left(-\frac{d_{\mathbb{D}_c}(E(s_t \circ a_t), E(g))}{\kappa}\right),$$

where $d_{\mathbb{D}_c}$ is the Poincaré distance with curvature c, and $\kappa > 0$ controls the sharpness of the guidance signal.

E.4 Residual Representation and Algebraic Sparsification

For symbolic domains, the residual r_t is computed from the unresolved symbolic structure of the current state. In Andrews-Curtis-style tasks, r_t is derived from the canonicalized presentation after the generated prefix, including generator counts, relator lengths, and unresolved relator components. Each candidate move L_j is associated with a structural subspace S_j, and the algebraic sparsification potential is

$$\Psi_P(r_t, S_j) = \frac{\|P_{S_j} r_t\|_2^2}{\|r_t\|_2^2 + \epsilon}.$$

For non-symbolic tasks, we use a finite operator taxonomy over coarse reasoning-step types. For mathematical reasoning, the operator taxonomy includes simplification, substitution, numerical evaluation, equation formation, formula invocation, case split, constraint checking, and final-answer extraction. For free-form natural reasoning, the taxonomy includes factual retrieval, comparison, elimination, aggregation, inference, format normalization, and final-answer commitment.

Given a rollout prefix s_t, we compute a frozen encoder representation

$$h_t = \mathcal{E}_{\ell_{\text{enc}}}(x(s_t)).$$

The residual vector is predicted by a fixed probe

$$r_t = W_R h_t.$$

The probe is trained only on training-rollout pseudo-labels. Let y_t denote the rollout-local structural pseudo-label vector, containing operator type, unresolved constraint coverage, answer-schema status, and format-consistency indicators. We fit W_R by the regularized regression objective

$$W_R = \arg\min_W \sum_{(s_t, y_t) \in \mathcal{D}_{\text{probe}}} \|W h_t - y_t\|_2^2 + \xi \|W\|_F^2.$$

No gold answers, gold rationales, ground-truth solution traces, terminal rewards, test labels, or test-set information are used to train this probe.

For each operator type j, we construct a subspace S_j from training rollouts. Let

$$\mathcal{R}_j = \{r_t : \text{the transition from } s_t \text{ is assigned operator type } j\}$$

be the residual vectors associated with operator type j. We compute the top d_j principal directions of $\mathcal{R}_j$ and write them as

$$U_j \in \mathbb{R}^{d_R \times d_j}.$$

The projection matrix is then

$$P_{S_j} = U_j U_j^\top.$$

For a candidate reasoning step $a_t^{(k)}$, we assign a coarse operator type

$$j(a_t^{(k)}) = C_{\text{op}}(s_t, a_t^{(k)}),$$

where C_{op} is the same parser or cluster-based operator classifier used to construct the pseudo-labels. The algebraic sparsification score for the candidate step is

$$\Psi_P(s_t, a_t^{(k)}) = \frac{\|P_{S_{j(a_t^{(k)})}} r_t\|_2^2}{\|r_t\|_2^2 + \epsilon}.$$

Thus, Ψ_P measures whether the candidate step's coarse operator type acts on the unresolved structural residual predicted for the current prefix. For non-symbolic tasks, this score is a learned training-time structural prior, not a formal local-admissibility certificate.

E.5 Meaning Clustering and Fixed-Subset Entropy Filtering

Entropy filtering is used only to define a fixed training subset and is not used during test-time evaluation. To avoid confounding method performance with method-dependent prompt selection, we compute the entropy filter once using rollouts from the same reference policy π_{ref} before training any compared method.

For each training prompt q, we sample G reference rollouts and group the outputs into meaning clusters $\{c_1, \ldots, c_M\}$. In structured domains, clustering is implemented by extracting and canonicalizing the final answer followed by deterministic matching. In free-form domains, we use a binary semantic-equivalence function

$$V(q, o_a, o_b) \in \{0, 1\}$$

that returns whether two outputs express the same answer meaning. This function is applied only to training rollouts for constructing the training subset.

We compute semantic entropy as

$$H_{\text{ref}}(q) = -\sum_{j=1}^{M} p(c_j \mid q) \log p(c_j \mid q), \qquad p(c_j \mid q) = \frac{|c_j|}{G}.$$

The fixed filtered training subset is

$$\mathcal{D}_{\text{filt}} = \{q \in \mathcal{D}_{\text{train}} : \delta_{\text{low}} < H_{\text{ref}}(q) < \delta_{\text{high}}\}.$$

All compared methods in the controlled ablation study, including GRPO, EMPO, PRM-GRPO, and SAGE, are trained on the same $\mathcal{D}_{\text{filt}}$ with the same number of prompts, rollout groups, optimization steps, decoding constraints, and KL coefficient. Therefore, differences in performance cannot be attributed to method-specific prompt retention.

All reported evaluation results are computed on the full benchmark test sets without filtering test examples, without semantic clustering, and without evaluating Ψ_P or Ψ_H at inference time.

Training Data. We use separate training pools for mathematical and natural reasoning, with all 11 benchmark test sets reserved exclusively for evaluation. For mathematical reasoning, we randomly sample 20,000 prompts from NuminaMath-CoT, evaluating the post-trained models on MATH, Minerva Math, OlympiadBench, AIME24, AMC23, GSM8K, and Putnam. For natural reasoning, we use 18,000 prompts from Natural Reasoning after filtering examples whose prompt, reference answer, or reference-model response exceeds 4,096 tokens, evaluating on MMLU-Pro, GPQA, BBH-H, and ARC-C. Following entropy-based prompt filtering, all methods within each controlled comparison use the same fixed training subset and optimization settings (Appendix E.5). The 1,190 independently constructed AC presentations are used exclusively in the separate symbolic AC setting and are not mixed with either LLM training pool.

EMPO Baseline. EMPO is a fully unsupervised reinforcement learning baseline that groups sampled responses according to semantic equivalence of their final answers and constructs entropy-based rewards from the empirical distribution over answer clusters. By encouraging low-entropy, self-consistent predictions, EMPO provides a learning signal without gold answers, verified reasoning trajectories, or external reward models. It additionally employs entropy thresholds and KL regularization for training stability. Unlike SAGE, which provides prefix-level guidance through algebraic sparsification and hyperbolic structural potentials, EMPO operates primarily on the final-response distribution and does not explicitly model the structural feasibility or depth-wise progress of intermediate reasoning trajectories.

AC Path Solving. We distinguish three evaluation metrics for the Andrews–Curtis (AC) task. ACValidity measures the percentage of syntactically and logically valid intermediate transitions, while Lean-Verified Proofs measures end-to-end success certified by the Lean proof assistant. These two metrics serve as the primary evaluation criteria in Section 5.3.

ACPathSolving is an additional path-level diagnostic reported in the component ablation and reward-side analyses. Following the long-horizon AC formulation, it measures the percentage of evaluated instances for which a generated trajectory identifies a valid sequence of Andrews–Curtis moves reaching the target trivial presentation within the allowed horizon T:

$$\mathrm{ACPathSolving} = \frac{100}{N_{\mathrm{AC}}} \sum_{i=1}^{N_{\mathrm{AC}}} \mathbf{1}\left[\exists\, t \le T : s_t^{(i)} = s_{\mathrm{triv}}^{(i)} \wedge \forall k < t,\ \mathrm{Adm}_{\mathcal{S}}(s_k^{(i)}, a_k^{(i)}, s_{k+1}^{(i)}) = 1\right], \tag{9}$$

where N_{AC} denotes the number of evaluated instances and $s_{\mathrm{triv}}^{(i)}$ is the target trivial presentation. Unlike ACValidity, which evaluates local transition validity, ACPathSolving evaluates whether a valid trajectory reaches the target within the prescribed horizon. Lean-Verified Proofs further requires independent formal verification.

F Hyperparameter Ablation

We evaluate the robustness of SAGE on the Olympiad dataset under three sources of variation: Pass@K scaling, decoding temperature, and the KL regularization coefficient β_{KL}. As shown in Figure 4, SAGE maintains a consistent performance lead over GRPO across sampling budgets. At Pass@64, SAGE reaches approximately 87%, compared with approximately 84% for GRPO.

Lower decoding temperatures generally favor more deterministic reasoning. However, SAGE remains competitive under higher stochasticity, indicating that structural post-training improves the quality of the sampled reasoning distribution rather than merely exploiting a narrow decoding regime. We also observe that SAGE is less sensitive to the KL coefficient than GRPO. While GRPO performance varies substantially across β_{KL}, SAGE with $\beta_{\mathrm{KL}} = 10^{-3}$ outperforms the strongest GRPO variants in this sweep.

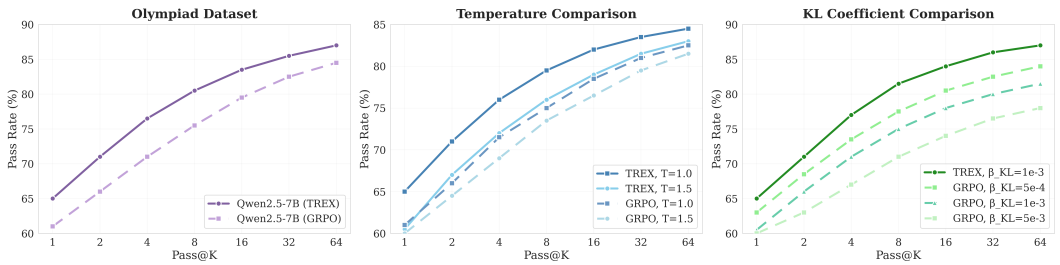


Figure 4: Hyperparameter ablation on the Olympiad dataset. Left: SAGE scales consistently across Pass@K budgets. Middle: SAGE remains robust under different decoding temperatures. Right: SAGE is less sensitive to the KL coefficient β_{KL} than GRPO.

G Process Reward Model Baseline

To further compare SAGE against dense process-level reward shaping, we include a process-reward-model baseline, denoted PRM-GRPO. This baseline addresses whether the improvements of SAGE can be explained simply by adding a learned dense reward signal during post-training.

PRM construction. PRM-GRPO trains a step-level reward model $R_\phi(s_t, a_t)$ on the same training rollouts used by the RL baselines. For mathematical and free-form reasoning tasks, process labels are derived from terminal-outcome-labeled rollouts through outcome-to-prefix credit assignment: prefixes from successful rollouts are treated as positive process examples, while prefixes from failed rollouts are treated as negative process examples. For AC task, we additionally use the task interface to derive local symbolic-validity targets for generated transitions. The PRM is trained only on training rollouts and does not use test-set labels or test-time feedback.

PRM-guided post-training. During post-training, PRM-GRPO uses the same rollout budget, decoding constraints, KL coefficient, and optimization schedule as GRPO and SAGE. The only difference from GRPO is that the sparse terminal reward is augmented with the learned process reward:

$$\widetilde{R}_{\mathrm{PRM}}(\tau) = R_{\mathrm{term}}(\tau) + \eta_{\mathrm{PRM}} \frac{1}{T} \sum_{t=1}^{T} R_\phi(s_t, a_t).$$

The resulting trajectory-level reward is then normalized within each rollout group and optimized with the same group-relative policy update. The PRM is used only during post-training reward shaping and is not used to filter test examples or guide inference.

H Additional Results

H.1 Generalization Beyond Symbolic Reasoning

Extension to Non-Tree Reasoning Spaces. SAGE does not require the underlying environment state graph to be a tree. For non-symbolic tasks, structural guidance operates on history-conditioned reasoning prefixes, where each state representation combines the original task input with the generated reasoning or action history. Consequently, trajectories reaching the same environment state through different histories can still receive distinct structural scores. This formulation allows SAGE to operate on reasoning problems involving cyclic dependencies, multi-hop evidence integration, and non-tree action spaces.

For MuSiQue, the residual representation captures unresolved question requirements, including entities, relations, and evidence dependencies, while the target anchor encodes prompt-conditioned structural information estimated from training rollouts. For ALFWorld, the residual captures unresolved task requirements and subgoals conditioned on the instruction, current observation, and action history. Both settings use the learned structural-prior construction described in Appendix E.

Multi-Hop Question Answering. Table 8 presents results on MuSiQue across three model scales. SAGE consistently outperforms both GRPO and EMPO in exact-match accuracy and token-level

Answer F1. Compared with GRPO, SAGE improves Answer EM by 5.1, 4.6, and 4.6 percentage points on the 9B, 27B, and 35B models, respectively. The corresponding Answer F1 gains are 4.8, 4.2, and 4.6 points. These results suggest that structural guidance remains beneficial when successful reasoning requires integrating evidence across multiple dependent hops rather than following a predefined symbolic transformation sequence.

Agentic Reasoning. Table 9 reports task success and action efficiency on ALFWorld. SAGE improves task success over GRPO by 7.5, 7.1, and 7.1 percentage points across the three model scales, while reducing the average number of environment actions among successful episodes by 3.3, 2.9, and 2.7 steps, respectively. The simultaneous improvements in success rate and action efficiency suggest that structural guidance helps policies maintain task-directed progress across extended interaction sequences.

Together, these results provide empirical evidence that SAGE can generalize beyond explicit symbolic reasoning to multi-hop question answering and interactive agentic tasks. However, the learned structural priors in these settings remain approximate, and the exact symbolic guarantees of SCA do not directly extend to arbitrary non-symbolic environments.

Table 5: Accuracy (%) on free-form natural reasoning benchmarks over 5 random seeds, reported as mean with standard deviation. The best mean is in **bold** with second best in underline.

Model	STEM	MMLU-Pro				GPQA	BBH-H	ARC-C
		Humanity	Social	Other	Avg.			
9B models								
Qwen3.5	12.71±0.21	8.02±0.18	14.95±0.25	10.21±0.20	11.06±0.16	10.91±0.31	21.58±0.38	18.49±0.35
Qwen3.5 w/SFT	20.18±0.34	11.36±0.27	28.97±0.43	19.22±0.36	19.85±0.31	12.31±0.42	32.44±0.55	29.56±0.51
Qwen3.5 w/GRPO	33.38±0.56	28.31±0.49	50.41±0.62	39.26±0.55	39.33±0.47	18.21±0.73	41.69±0.84	37.82±0.79
Qwen3.5 w/EMPO	32.57±0.53	27.32±0.51	48.73±0.66	37.69±0.58	37.91±0.50	**21.11±0.69**	44.04±0.82	39.73±0.76
Qwen3.5 w/SAGE	**34.11±0.48**	**29.08±0.44**	**51.17±0.59**	**39.71±0.52**	**39.99±0.43**	20.86±0.64	**45.31±0.71**	**42.04±0.68**
27B models								
Qwen3.6	30.29±0.25	24.34±0.22	46.55±0.34	35.31±0.29	35.40±0.24	16.09±0.38	38.70±0.49	34.78±0.46
Qwen3.6 w/SFT	34.18±0.31	28.51±0.27	41.52±0.39	37.28±0.33	35.77±0.29	22.67±0.47	45.27±0.57	41.12±0.53
Qwen3.6 w/GRPO	**57.74±0.49**	37.0±0.43	**65.16±0.54**	57.55±0.48	53.24±0.41	**34.29±0.62**	55.74±0.71	51.78±0.67
Qwen3.6 w/EMPO	53.96±0.51	35.58±0.45	60.04±0.58	52.18±0.50	49.27±0.44	29.52±0.66	57.19±0.69	53.26±0.63
Qwen3.6 w/SAGE	56.31±0.44	**37.91±0.39**	64.43±0.51	**58.25±0.43**	**53.53±0.37**	32.51±0.57	**60.25±0.62**	**57.11±0.58**
35B models								
Qwen3.5	45.08±0.18	36.31±0.17	52.29±0.25	44.28±0.21	44.29±0.18	31.05±0.34	47.52±0.41	45.37±0.39
Qwen3.5 w/SFT	49.84±0.23	38.26±0.20	54.47±0.31	48.62±0.26	47.12±0.22	28.97±0.39	52.96±0.46	49.18±0.43
Qwen3.5 w/GRPO	63.58±0.38	43.19±0.34	69.08±0.45	60.71±0.39	57.66±0.33	35.78±0.51	63.29±0.57	60.91±0.55
Qwen3.5 w/EMPO	61.96±0.41	42.02±0.36	68.61±0.47	59.54±0.42	56.72±0.35	35.43±0.54	65.02±0.55	62.97±0.51
Qwen3.5 w/SAGE	**65.27±0.33**	**44.51±0.30**	**71.22±0.40**	**62.25±0.35**	**59.33±0.29**	**38.58±0.46**	**69.07±0.49**	**66.41±0.47**

H.2 Mechanism Validation Under Sparse Rewards

Non-Vanishing Structural Advantages. Table 11 evaluates whether SAGE maintains informative optimization signals when terminal rewards are sparse. Among rollout groups in which all terminal rewards are zero, SAGE retains non-zero within-group advantages in 56.8% of cases, compared with 0.0% for GRPO and 14.2% for EMPO. This observation is consistent with Proposition 4.1, which establishes that variation in structural potentials can produce non-vanishing group-relative advantages even when terminal feedback is uninformative.

Importantly, stronger optimization signals are accompanied by improvements in independent correctness-related metrics. Relative to GRPO, SAGE increases Lean-verified intermediate steps from 18.6 to 31.9 per 1,000 tokens, a 71% improvement, while improving AC Path Solving from 23.05% to 31.76% and Lean-Verified success from 14.64% to 23.69%. These findings support the interpretation that structural guidance provides useful learning signals in sparse-reward regimes rather than merely increasing auxiliary reward values.

Table 6: Comparison on Math Reasoning Benchmarks.

Backbone	Baseline	MATH	Minerva	Olympiad	AIME24	AMC23	GSM8K	Putnam	Avg.
DeepSeekMath-7B		67.20	34.50	39.80	25.70	50.30	78.40	10.80	43.81
	w/SFT	72.80	39.40	46.10	31.80	56.70	83.20	13.60	49.09
	w/GRPO	79.30	45.70	55.80	45.20	63.90	88.10	17.40	56.49
	w/EMPO	78.80	46.10	**59.00**	**49.20**	64.80	88.90	17.90	57.81
	w/SAGE	**81.40**	**48.60**	58.70	48.70	**66.70**	**90.00**	**19.80**	**59.13**
DeepSeek-Prover-V2-7B		70.40	37.80	49.60	34.20	54.80	80.10	14.70	48.80
	w/SFT	74.90	41.60	53.80	38.50	59.10	83.70	16.30	52.56
	w/GRPO	80.80	47.90	61.40	50.60	65.20	**91.00**	20.40	59.61
	w/EMPO	80.20	48.40	62.10	**54.50**	65.80	89.10	20.80	60.13
	w/SAGE	**82.60**	**50.70**	**64.90**	53.80	**67.40**	90.70	**22.60**	**61.81**
Kimina-Prover-7B		71.60	38.90	51.20	35.80	56.10	80.80	15.60	50.00
	w/SFT	75.80	42.70	55.60	40.40	60.30	84.10	17.10	53.71
	w/GRPO	**83.40**	48.60	62.90	52.10	66.40	89.00	21.10	60.50
	w/EMPO	80.70	49.10	63.40	54.80	67.00	89.40	21.60	60.86
	w/SAGE	83.10	**51.30**	**66.10**	**55.20**	**68.60**	**90.70**	**23.40**	**62.63**
Kimina-Prover-Distill-8B		73.80	41.20	54.60	39.30	59.20	82.40	17.30	52.54
	w/SFT	77.40	45.10	58.70	43.70	63.10	85.60	18.90	56.07
	w/GRPO	82.60	50.50	65.10	55.30	68.20	**92.00**	22.80	62.36
	w/EMPO	82.10	**53.40**	65.80	**59.00**	68.90	90.50	23.20	63.27
	w/SAGE	**84.20**	53.20	**68.00**	58.10	**70.40**	91.70	**25.00**	**64.37**
Llama-3.3-70B-Instruct		66.08	33.61	32.94	17.31	29.02	80.47	9.91	38.48
	w/SFT	72.40	40.20	44.60	30.80	48.90	85.30	14.10	48.04
	w/GRPO	80.10	47.80	59.40	50.60	64.70	90.50	19.00	58.87
	w/EMPO	79.60	**51.30**	60.10	**55.00**	65.40	91.00	19.60	60.29
	w/SAGE	**82.50**	51.00	**63.40**	54.00	**67.10**	**92.30**	**21.70**	**61.71**

Representative Trajectory Analysis. Table 12 provides a trajectory-level illustration using the median-length SAGE trajectory among instances successfully solved by SAGE. All methods are evaluated on the same AC instance, with the first 12 transitions compared under a matched horizon.

SAGE maintains 12 valid transitions out of 12 and reduces residual complexity from 18 to 5. In comparison, GRPO achieves 6 valid transitions and increases residual complexity from 18 to 19, while EMPO achieves 7 valid transitions and reduces residual complexity only to 16. SAGE subsequently completes the AC path and passes independent Lean verification, whereas neither baseline succeeds on this instance.

This case illustrates how maintaining a valid prefix and reducing residual complexity can jointly support long-horizon completion. As a selected representative example, it complements rather than replaces the aggregate evaluation.

Structural Reward and Terminal Correctness. Structural reward density and terminal correctness measure different aspects of reasoning quality. Higher structural rewards do not necessarily imply greater end-to-end accuracy, particularly in heterogeneous free-form reasoning settings such as BBH-H, where locally plausible reasoning steps may still lead to incorrect final answers.

To examine this distinction, Table 13 evaluates the correspondence between individual structural potentials and terminal correctness on held-out BBH-H rollouts. Algebraic sparsification achieves a Spearman correlation of 0.17 and success-ranking AUROC of 0.58, while hyperbolic guidance achieves 0.20 and 0.60, respectively. In contrast, the combined SAGE signal achieves substantially stronger alignment, with $\rho = 0.37$ and AUROC $= 0.69$.

These results suggest that the two potentials capture complementary aspects of reasoning behavior: algebraic sparsification measures structural exploration compatibility, whereas hyperbolic guidance captures sustained depth-wise progress. Their combination provides a more informative structural signal than either component alone, helping explain why higher reward density from an isolated component does not necessarily translate into improved terminal accuracy.

Overall, the AC diagnostics, trajectory-level case study, and BBH-H correctness-alignment analysis provide complementary evidence for SAGE's mechanism: structural guidance sustains optimization signals under sparse rewards, promotes long-horizon progress, and exhibits stronger correspondence with correctness when both components are combined. These findings are consistent with the

proposed mechanism, without establishing that structural rewards guarantee correctness for every trajectory.

Table 7: Comparison on Natural Reasoning Benchmarks.

Backbone	Baseline	MMLU-Pro					GPQA	BBH-H	ARC-C
		STEM	Humanity	Social	Other	Avg.			
DeepSeekMath-7B		26.40	21.80	35.70	29.20	28.27	19.60	31.80	30.70
	w/SFT	31.20	24.60	39.40	33.10	32.08	21.80	35.20	34.40
	w/GRPO	**41.70**	31.40	50.20	42.10	41.35	26.70	44.80	42.60
	w/EMPO	38.90	32.10	49.50	41.70	40.55	**27.40**	46.00	43.10
	w/SAGE	41.50	**33.70**	**52.10**	**44.00**	**42.83**	27.00	**47.90**	**45.30**
DeepSeek-Prover-V2-7B		23.80	19.70	32.40	26.10	25.50	18.40	29.10	28.50
	w/SFT	28.60	22.30	36.80	30.20	29.48	20.60	33.00	31.70
	w/GRPO	36.70	28.80	46.30	38.50	37.58	**25.80**	41.20	39.40
	w/EMPO	35.90	**31.40**	45.80	38.10	37.80	25.60	42.00	40.20
	w/SAGE	**38.40**	31.00	**48.60**	**40.70**	**39.67**	25.20	**44.30**	**42.10**
Kimina-Prover-7B		25.10	20.80	34.10	27.30	26.83	18.90	30.40	29.60
	w/SFT	29.70	23.50	37.90	31.50	30.65	21.10	34.20	32.80
	w/GRPO	37.80	29.90	**50.10**	39.80	39.40	25.70	42.70	40.60
	w/EMPO	37.10	30.50	47.00	39.40	38.50	26.00	**46.10**	41.10
	w/SAGE	**39.60**	**32.20**	49.80	**42.00**	**40.90**	**26.50**	45.80	**43.40**
Kimina-Prover-Distill-8B		28.40	23.10	37.20	30.60	29.83	20.80	33.60	32.20
	w/SFT	32.70	25.90	41.10	34.40	33.52	22.90	36.90	35.10
	w/GRPO	40.50	32.40	50.60	42.80	41.58	27.90	**48.70**	43.20
	w/EMPO	39.80	**35.00**	50.10	42.40	41.83	**28.50**	46.10	43.80
	w/SAGE	**42.30**	34.60	**52.90**	**44.90**	**43.67**	28.10	48.40	**46.00**
Llama-3.3-70B-Instruct		52.80	43.60	61.90	54.10	53.10	38.20	59.40	57.80
	w/SFT	55.10	45.30	64.20	56.60	55.30	39.70	61.80	59.90
	w/GRPO	61.80	50.40	**72.60**	63.20	62.00	43.60	66.70	64.10
	w/EMPO	61.00	51.10	69.80	62.70	61.15	**44.20**	68.10	64.80
	w/SAGE	**63.40**	**52.80**	72.10	**65.00**	**63.32**	43.90	**69.80**	**67.10**

Table 8: Multi-hop QA results on MuSiQue. We report exact-match accuracy (EM) and token-level Answer F1.

Model	Method	Answer EM (%) ↑	Answer F1 (%) ↑
Qwen3.5-9B	GRPO	49.8	62.4
	EMPO	50.6	63.1
	SAGE	**54.9**	**67.2**
Qwen3.6-27B	GRPO	55.7	67.9
	EMPO	56.4	68.5
	SAGE	**60.3**	**72.1**
Qwen3.5-35B	GRPO	59.1	70.8
	EMPO	59.8	71.5
	SAGE	**63.7**	**75.4**

I Computation Resources

All experiments were conducted on an internal GPU cluster with 4 NVIDIA A100 GPUs and 2 NVIDIA H200 GPUs. We did not train any foundation model from scratch; compute was mainly used for post-training rollout generation, structural-potential computation, policy optimization, ablations, and benchmark evaluation. SAGE adds training-time overhead for computing algebraic sparsification and hyperbolic structural guidance, but incurs no extra inference-time cost because the guidance is absorbed into the trained policy.

Table 9: Agentic reasoning results on ALFWorld. Task Success measures completed goals, while Avg. Steps reports the number of environment actions among successful episodes.

Model	Method	Task Success (%) ↑	Avg. Steps (Success) ↓
Qwen3.5-9B	GRPO	57.2	19.6
	EMPO	58.4	18.9
	SAGE	**64.7**	**16.3**
Qwen3.6-27B	GRPO	63.5	17.8
	EMPO	64.1	17.2
	SAGE	**70.6**	**14.9**
Qwen3.5-35B	GRPO	67.8	16.5
	EMPO	68.4	16.0
	SAGE	**74.9**	**13.8**

Table 10: Training cost on Qwen3.5-35B. Reported costs include auxiliary reward acquisition and one-time preprocessing.

Method	Train Wall-clock ↓	GPU-hours ↓	Peak GPU Memory ↓	Lean-Verified ↑
GRPO	1.00×	1.00×	1.00×	14.64
GRPO-PRM	1.30×	1.28×	1.15×	17.36
SAGE	**1.16**×	**1.15**×	**1.08**×	**23.69**

Table 11: Reward-side diagnostics on AC. Zero-Terminal Groups w/ Non-zero Advantage denotes rollout groups where all terminal rewards are zero but the within-group advantage remains non-zero.

Method	Terminal Rew./1k ↑	Zero-Terminal Groups w/ Non-zero Advantage (%) ↑	Lean-Verified Mid-Steps/1k ↑	AC Path Solving ↑	Lean-Verified ↑
GRPO	1.8	0.0	18.6	23.05	14.64
EMPO	1.6	14.2	17.9	21.52	13.78
SAGE	**3.7**	**56.8**	**31.9**	**31.76**	**23.69**

Table 12: Representative AC case study. We compare all methods on the same instance over the first 12 transitions, using the median-length SAGE trajectory among instances solved by SAGE.

Method	Lean-Valid Steps ↑	Longest Valid Prefix ↑	Residual Complexity ↓	Final AC Path Solved	Final Lean-Verified
GRPO	6/12	6	$18 \to 19$	×	×
EMPO	7/12	7	$18 \to 16$	×	×
SAGE	**12/12**	**12**	$\mathbf{18 \to 5}$	✓	✓

Table 13: Reward–correctness alignment on held-out BBH-H rollouts. We evaluate how individual structural potentials and their combined SAGE signal correlate with terminal correctness.

Structural Signal	Spearman ρ ↑	Success-ranking AUROC ↑
Algebraic Sparsification (Ψ_P)	0.17	0.58
Hyperbolic Guidance (Ψ_H)	0.20	0.60
Full SAGE (Ψ_{SAGE})	**0.37**	**0.69**

J Limitations

SAGE relies on training-time structural priors whose quality depends on the task interface. In explicit symbolic domains, local admissibility can be defined exactly, whereas mathematical and free-form reasoning rely on approximate learned residuals, anchors, and operator subspaces. Consequently, the theoretical concentration guarantees apply directly to symbolic settings and conditionally to settings where the learned potentials separate productive and unproductive prefixes.

Moreover, SAGE's structural reward augmentation is not strict potential-difference shaping and therefore does not guarantee policy invariance. Although length averaging mitigates incentives for artificially extending trajectories, imperfect structural priors may introduce reward misspecification or favor structurally plausible but incorrect reasoning paths. KL regularization and clipped updates

constrain policy drift but do not eliminate this risk. We therefore evaluate SAGE using independent correctness metrics, including terminal task success and formal Lean verification, rather than relying solely on structural rewards.

Finally, SAGE incurs additional training-time overhead from candidate-step scoring, residual probing, and hyperbolic distance computation, although inference requires no additional structural scoring. Future work should investigate more robust structural priors, stronger correctness alignment, and tighter guarantees for non-symbolic reasoning.

K Broader Impact

This work aims to improve long-horizon reasoning under sparse-reward regimes. Its potential positive impact is to make LLM reasoning more reliable in domains requiring extended symbolic or semi-symbolic reasoning, such as mathematics, formal verification, and scientific reasoning, while reducing dependence on dense human-written process supervision.

The main risk is that stronger long-horizon reasoning may also improve dual-use capabilities when integrated into external tools or autonomous systems. SAGE improves training-time reasoning stability, but it does not guarantee factual correctness, harmlessness, fairness, or robustness under distribution shift. Deployments in consequential settings should therefore include domain-specific validation, uncertainty estimation, human oversight, and task-level safety checks.

L Safeguards

This work does not release a new pretrained foundation model, user-facing agent, scraped dataset, or deployment system. The released artifacts are limited to code, training and evaluation scripts, and reproducible reasoning resources. The experiments use public benchmarks and synthetic or symbolic reasoning tasks, without private user data or human-subject data collection.

SAGE is intended as a research framework, not a deployment-ready safety mechanism. Future extensions involving external tools, web access, code execution, or autonomous planning should add safeguards such as sandboxing, rate limits, misuse monitoring, restricted access for high-risk capabilities, and domain-specific safety evaluation before release.

M LLM usage

Large language models were used only for language polishing and wording refinement. Specifically, LLM assistance was limited to improving grammar, clarity, conciseness, and presentation of author-written text. All scientific ideas, problem formulation, theoretical results, proofs, algorithms, experimental design, implementation, data processing, numerical results, analysis, and conclusions were produced and verified by the authors. LLMs were not used to generate experimental results, fabricate data, perform reviewer simulation for reported claims, or make autonomous scientific decisions. The authors take full responsibility for the final content of the paper.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main theoretical, methodological, and empirical claims, including SCA as a theoretical lens, SAGE as the proposed framework, and evaluation across 12 benchmarks and 10 model backbones. The claims are scoped to long-horizon reasoning under sparse-reward regimes and are supported by the theoretical analysis and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations paragraph discussing dependence on training-time structural priors, the distinction between exact symbolic admissibility and approximate learned proxies in less structured domains, and additional training-time overhead in Appendix J.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions for the theoretical results in the relevant propositions and theorem, and provides complete proofs in the appendix. The theoretical claims are explicitly scoped to symbolic settings and conditionally to settings where learned potentials separate productive and unproductive prefixes in Appendix A-Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the evaluated models, benchmarks, baselines, metrics, rollout settings, ablations, and implementation details, and provides an anonymized code repository for reproducing the main results in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides an anonymized code repository. The experiments use public benchmarks and a reproducible construction protocol for the Andrews–Curtis presentations; scripts and instructions are provided with the released code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main text describes the evaluated model families, datasets, baselines, and metrics, while the appendix E provides training-time details, filtering procedures, hyperparameter settings, and ablation protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports five-seed mean-standard deviation results in the appendix for the main benchmark tables in Appendix H. The reported variability corresponds to independent training/evaluation runs under the same experimental conditions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the compute resources required for the experiments in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The work uses public benchmarks and synthetic/symbolic reasoning tasks, does not involve human-subject data collection, and preserves anonymity in the released materials.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential positive impacts, including improving reliable long-horizon reasoning and reducing dependence on dense process supervision, as well as potential risks from stronger reasoning models, such as misuse in automated generation or decision-support contexts in Appendix K.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper provides a safeguards section in Appendix L.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original sources for the datasets, model backbones, and baseline methods used in the experiments. The released code and documentation include the licenses or terms of use for existing assets where available.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces and releases code and experimental resources for SAGE, including scripts for constructing and evaluating the Andrews–Curtis reasoning instances. These assets are documented in the anonymized repository with instructions for reproducing the reported experiments.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, human-subject experiments, or participant compensation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject research, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper discloses LLM usage in Appnedix M. LLMs were used only for wording refinement, including grammar, clarity, conciseness, and presentation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.