SAGE:

Mitigating Long-Horizon Reasoning Biases via Topological Guidance

Xinyue Zeng[1] · Jiawei Zhang[2] · Yujun Yan[3] · Dawei Zhou[1]

[1]Virginia Tech · [2]University of Wisconsin–Madison · [3]Dartmouth College

Paper

Code

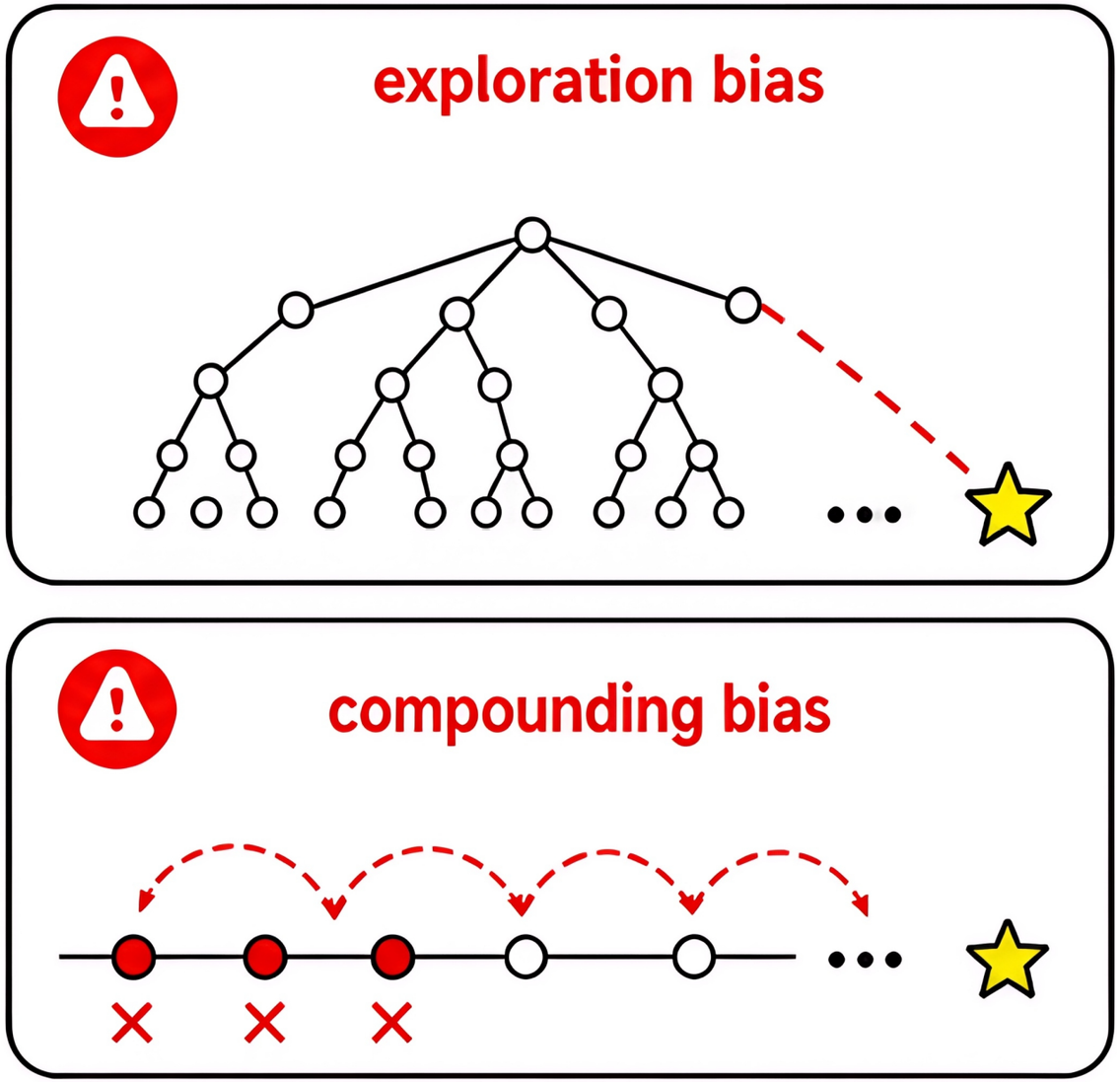


VIRGINIA TECH · WISCONSIN UNIVERSITY OF WISCONSIN–MADISON · DARTMOUTH

NEURAL INFORMATION PROCESSING SYSTEMS 2026

Outcome RL fails on long reasoning chains

PROBLEM



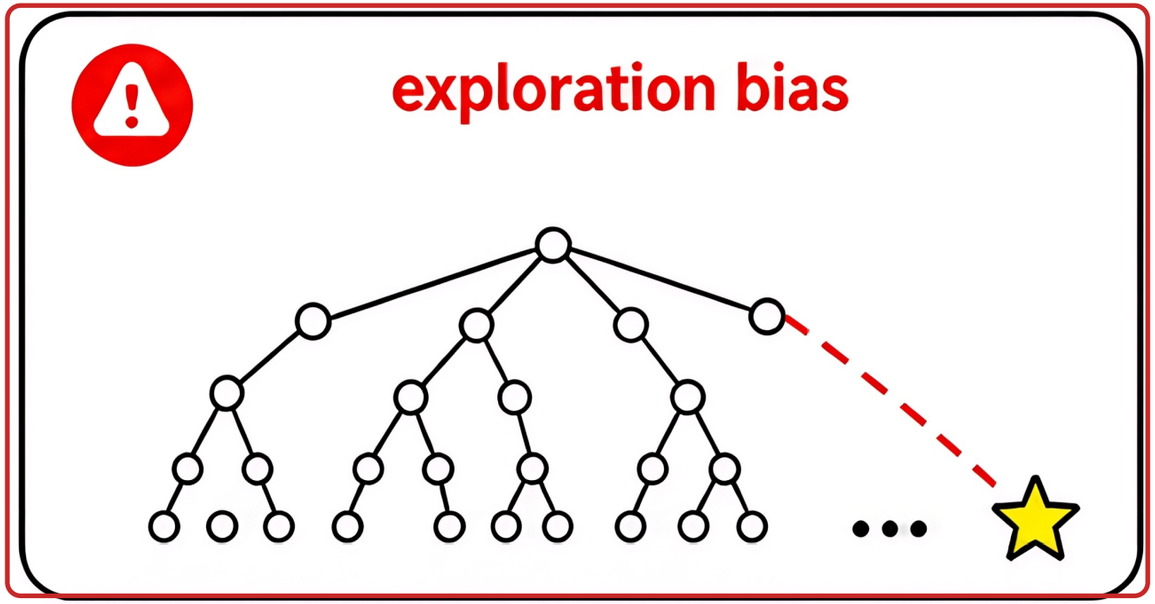
Running example: Andrews–Curtis

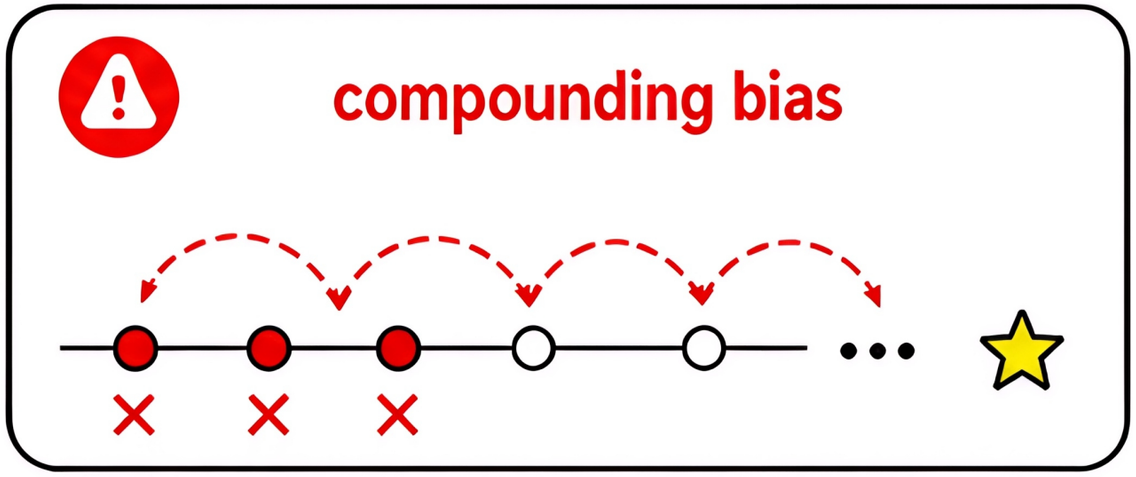
Simplify a group presentation to the trivial one, one move at a time.

$$\langle x, y \mid x^{-1}y^{4}xy^{-5},\ x^{-1}yxy^{-1},\ y^{-3} \rangle$$
$$\longrightarrow\ \langle x, y \mid x, y \rangle$$

Outcome RL fails on long reasoning chains

PROBLEM





Running example: Andrews–Curtis

Simplify a group presentation to the trivial one, one move at a time.

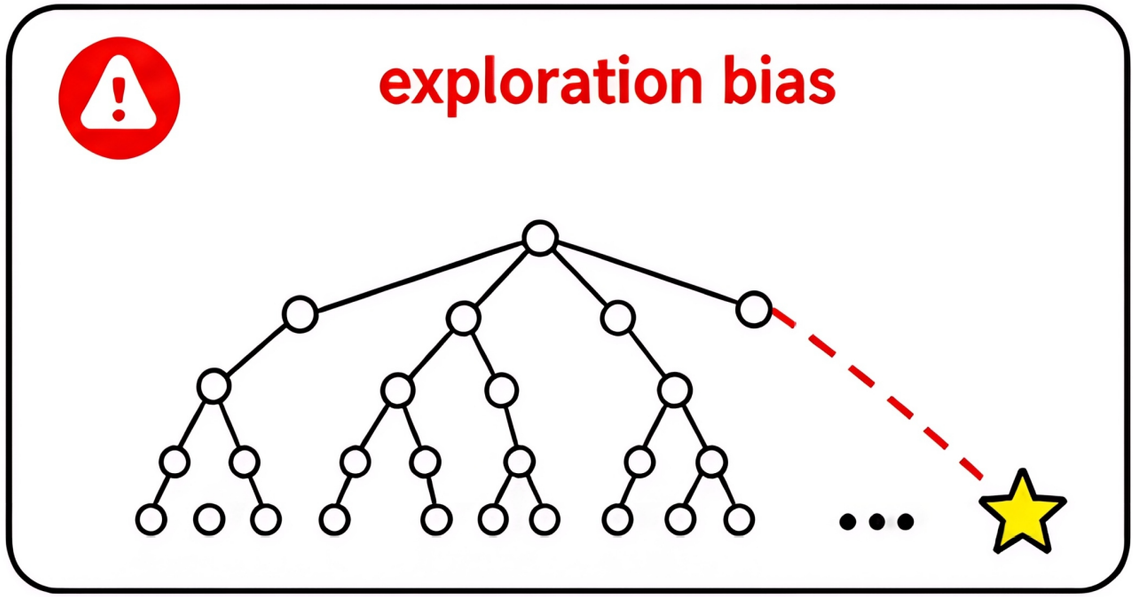
$$\langle x, y \mid x^{-1}y^{4}xy^{-5},\ x^{-1}yxy^{-1},\ y^{-3} \rangle$$
$$\longrightarrow\ \langle x, y \mid x, y \rangle$$

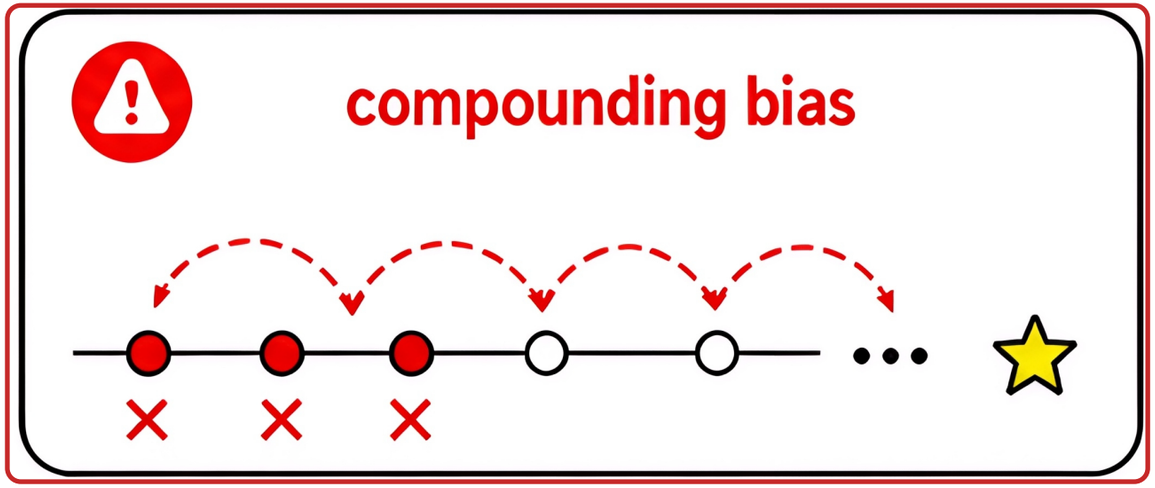
Exploration bias

Feasible paths are exponentially rare.

Outcome RL fails on long reasoning chains

PROBLEM





Running example: Andrews–Curtis

Simplify a group presentation to the trivial one, one move at a time.

$$\langle x, y \mid x^{-1}y^{4}xy^{-5},\ x^{-1}yxy^{-1},\ y^{-3} \rangle \longrightarrow \langle x, y \mid x, y \rangle$$

Exploration bias

Feasible paths are exponentially rare.

Compounding bias

Early mistakes pile up;
reward comes only at the end.

Both biases are provable, not anecdotal

Exploration bias

Feasible volume shrinks multiplicatively with depth.

$$\frac{|\{\tau \in \mathcal{F}\}|}{|\{\tau \in \Omega\}|} \lesssim \prod_{t=1}^{T} \frac{B_t^{\mathcal{F}}}{B_t}$$

Both biases are provable, not anecdotal

Exploration bias

Feasible volume shrinks multiplicatively with depth.

$$\frac{|\{\tau \in \mathcal{F}\}|}{|\{\tau \in \Omega\}|} \lesssim \prod_{t=1}^{T} \frac{B_t^{\mathcal{F}}}{B_t}$$

Compounding bias

Rare success keeps KL-regularized RL anchored to π_{ref}.

$$D_{\text{TV}}(\pi^*, \pi_{\text{ref}}) \leq (e^{R_{\max}/\lambda} - 1)\, p$$

Both biases are provable, not anecdotal

Exploration bias

Feasible volume shrinks multiplicatively with depth.

$$\frac{|\{\tau \in \mathcal{F}\}|}{|\{\tau \in \Omega\}|} \lesssim \prod_{t=1}^{T} \frac{B_t^{\mathcal{F}}}{B_t}$$

Compounding bias

Rare success keeps KL-regularized RL anchored to π_{ref}.

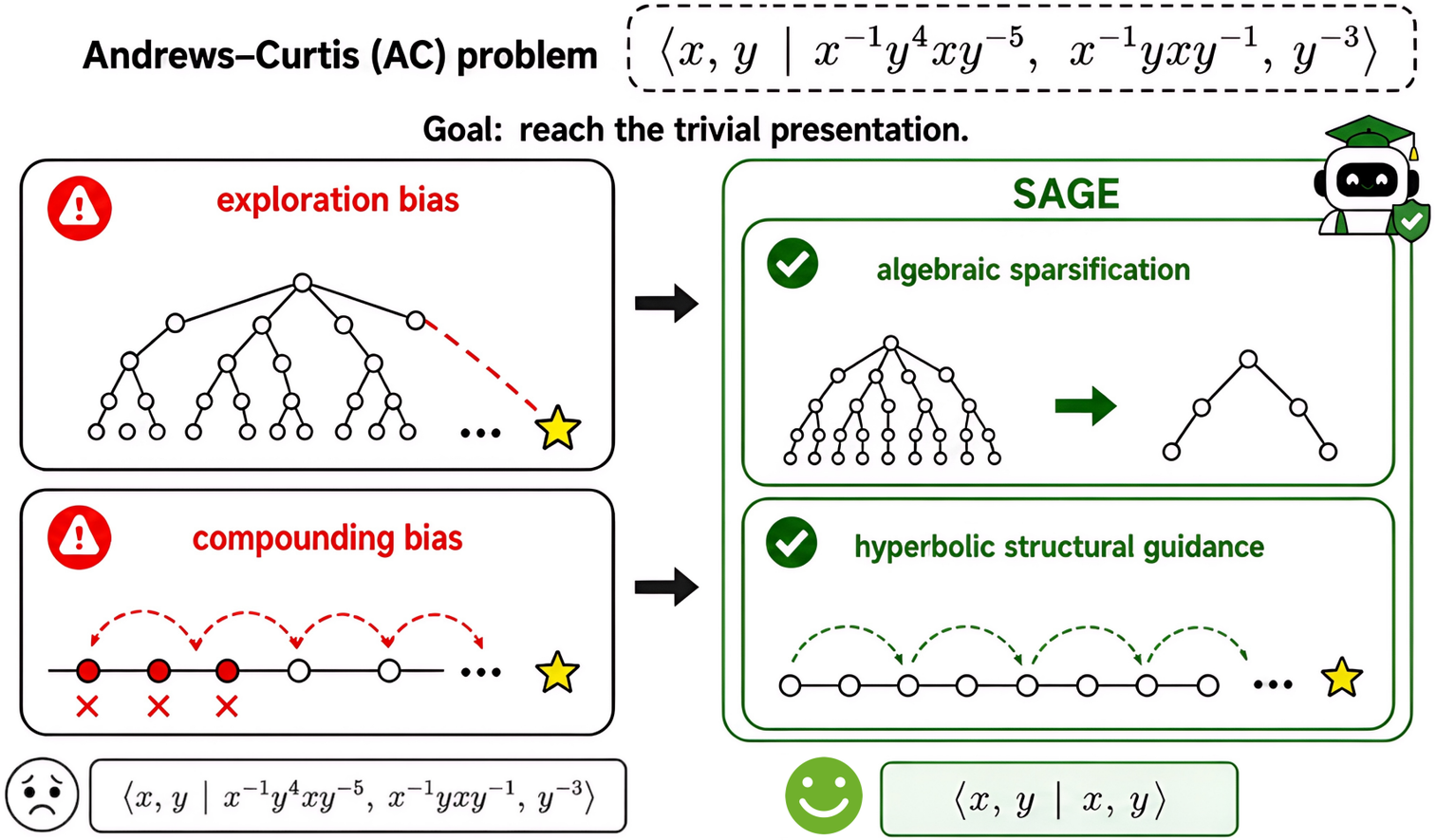
$$D_{\text{TV}}(\pi^*, \pi_{\text{ref}}) \leq (e^{R_{\max}/\lambda} - 1)\, p$$

A fix must

1. keep search inside the feasible region
2. give feedback at every step

SAGE: one structural potential per bias

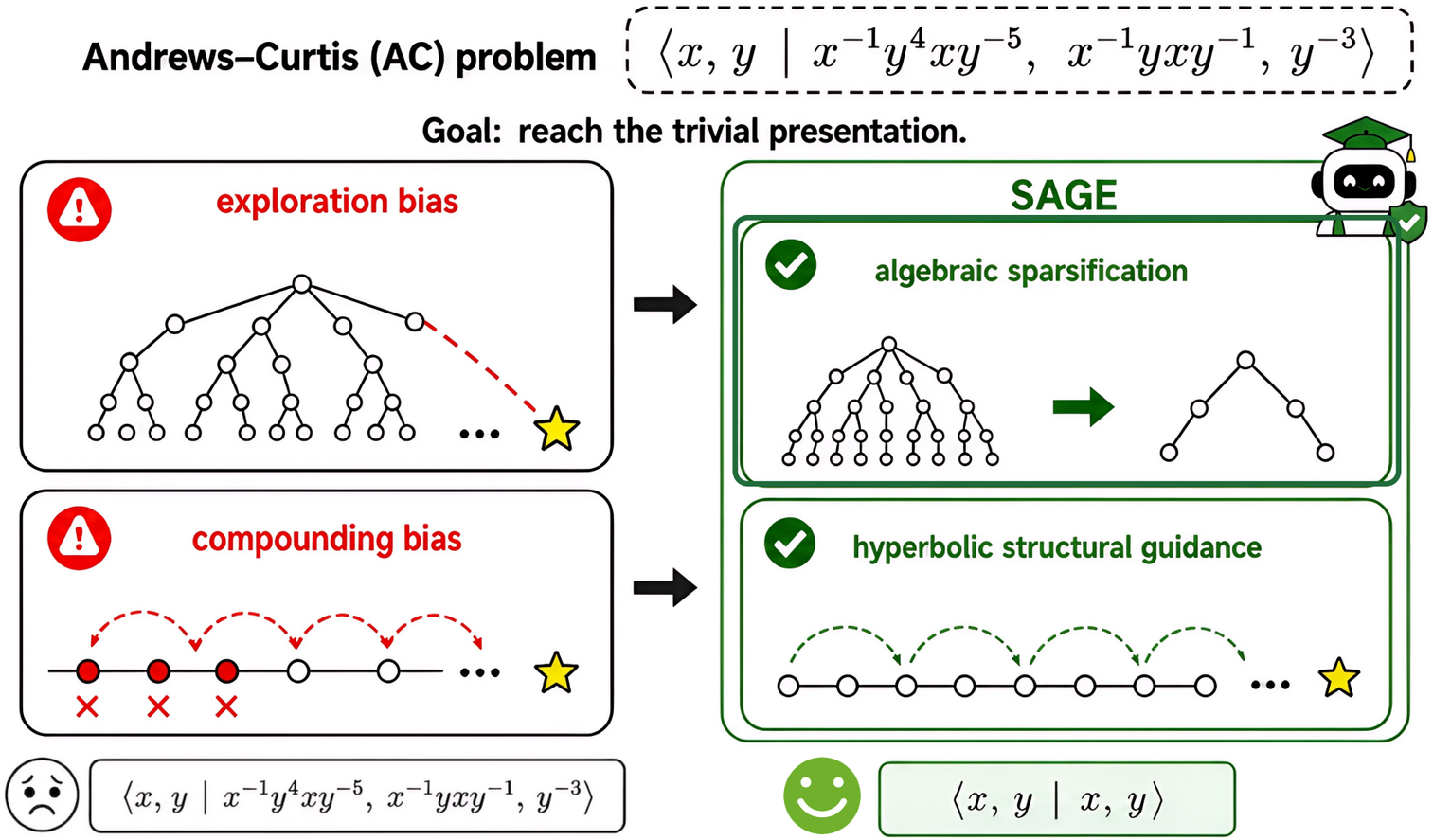
METHOD



Goal: reach the trivial presentation $\langle x, y \mid x, y\rangle$

SAGE: one structural potential per bias

METHOD



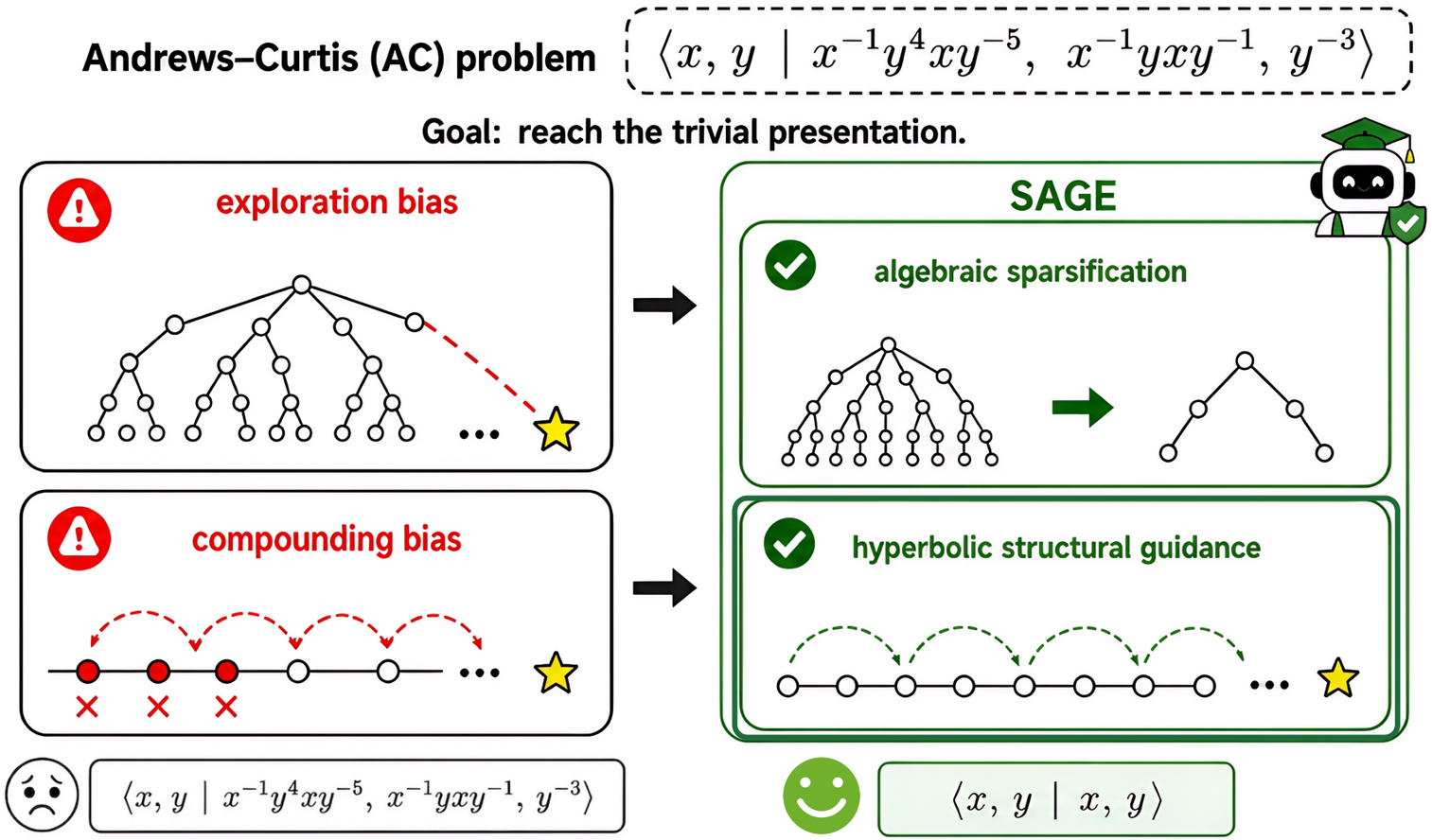
Algebraic sparsification $\Psi_{\mathcal{P}}$

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

Prunes spurious branches
→ fixes exploration bias

SAGE: one structural potential per bias

METHOD



Algebraic sparsification $\Psi_{\mathcal{P}}$

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

Prunes spurious branches
→ fixes exploration bias

Hyperbolic guidance $\Psi_{\mathcal{H}}$

$$\Psi_{\mathcal{H}} = \exp\left(-\frac{d_{\mathbb{D}}(E(s_t \circ L_j), E(g))}{\kappa}\right)$$

Dense, depth-aware signal
→ fixes compounding bias

Use the score twice: to explore and to learn

METHOD

$$\Psi_{\text{SAGE}} = \alpha\Psi_{\mathcal{P}} + \gamma\Psi_{\mathcal{H}}$$

Explore

Try high-score moves more

$$P(a_t) \propto \exp(\ell_\theta(a_t) + \lambda\Psi_{\text{SAGE}})$$

Learn

Every step earns credit in GRPO

$$\tilde{R} = R_{\text{term}} + \frac{\eta}{T}\sum_t \Psi_{\text{SAGE}}$$

Use the score twice: to explore and to learn

METHOD

$$\Psi_{\text{SAGE}} = \alpha\Psi_{\mathcal{P}} + \gamma\Psi_{\mathcal{H}}$$

Explore

Try high-score moves more

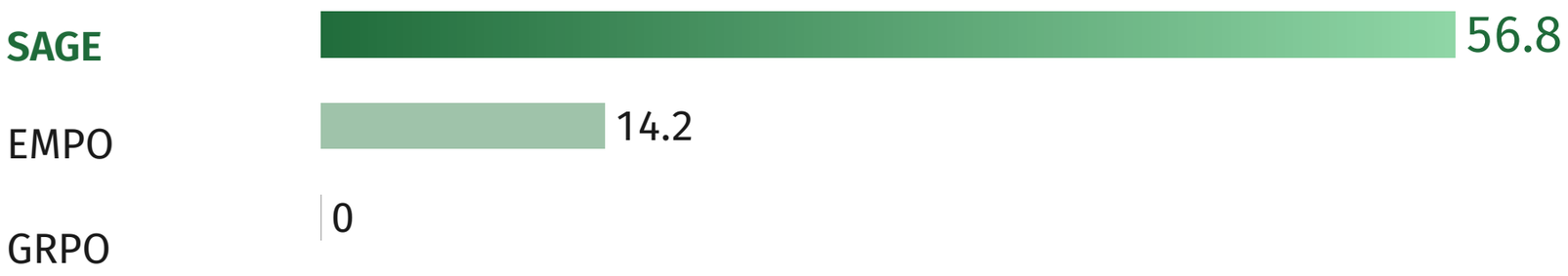
$$P(a_t) \propto \exp(\ell_\theta(a_t) + \lambda\Psi_{\text{SAGE}})$$

Learn

Every step earns credit in GRPO

$$\tilde{R} = R_{\text{term}} + \frac{\eta}{T}\sum_t \Psi_{\text{SAGE}}$$

Signal even when every rollout fails (%)



Use the score twice: to explore and to learn

METHOD

$$\Psi_{\text{SAGE}} = \alpha\Psi_{\mathcal{P}} + \gamma\Psi_{\mathcal{H}}$$

Explore

Try high-score moves more

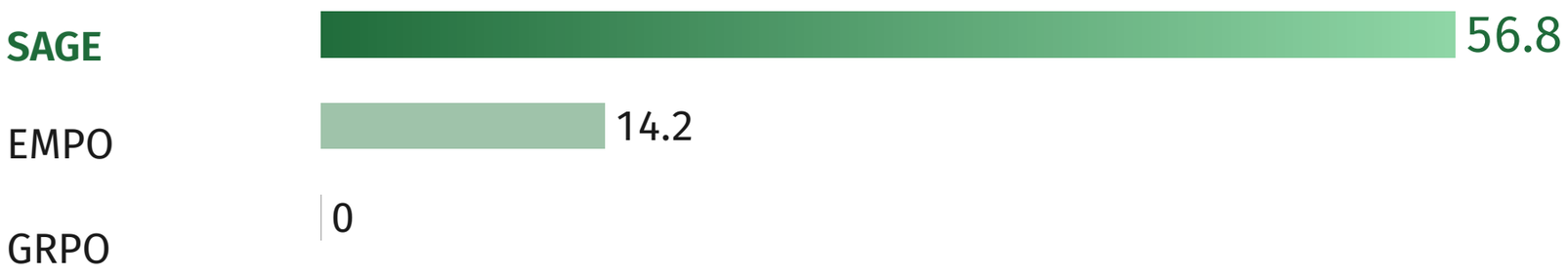
$$P(a_t) \propto \exp(\ell_\theta(a_t) + \lambda\Psi_{\text{SAGE}})$$

Learn

Every step earns credit in GRPO

$$\tilde{R} = R_{\text{term}} + \frac{\eta}{T}\sum_t \Psi_{\text{SAGE}}$$

Signal even when every rollout fails (%)

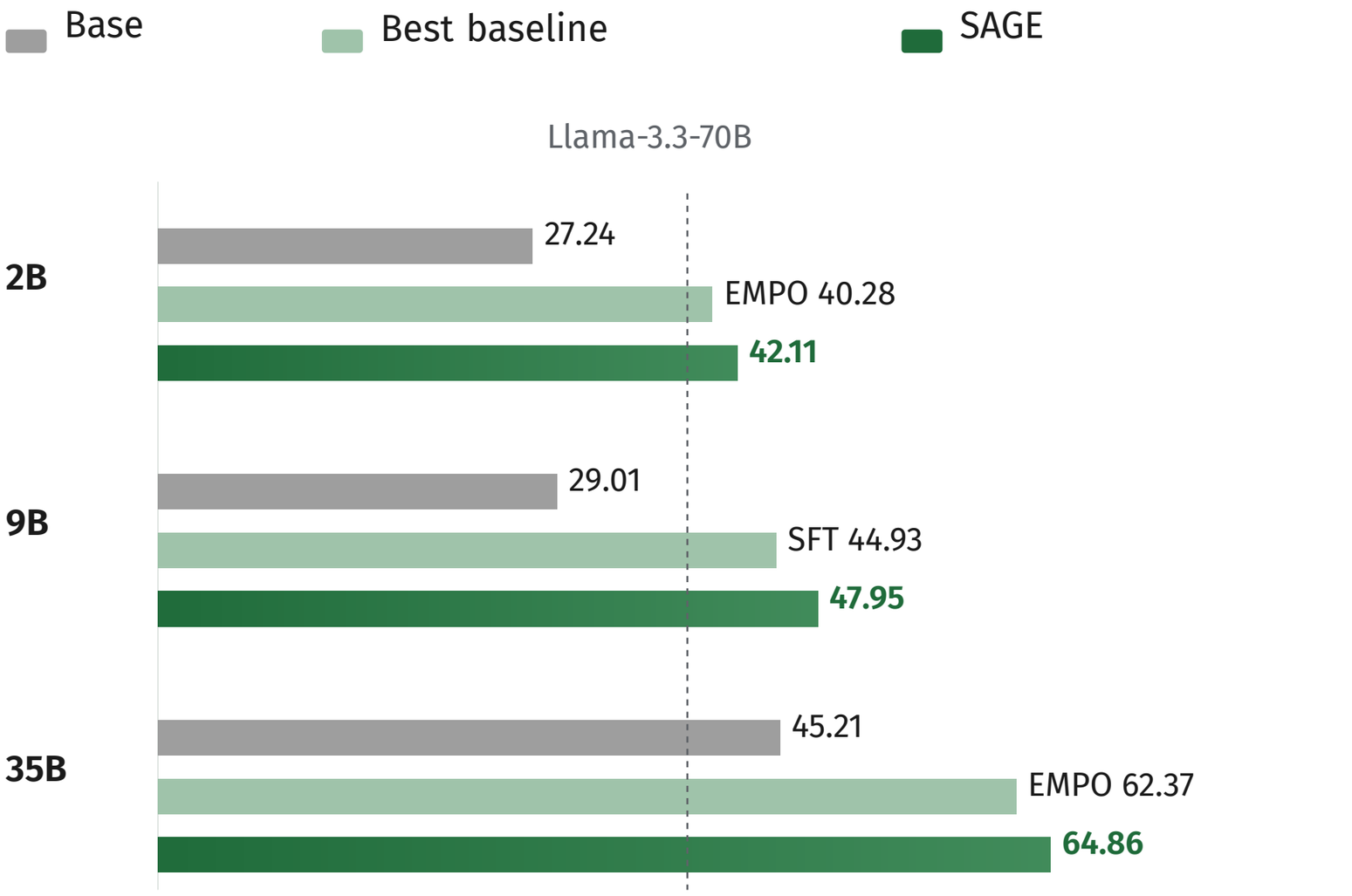


✓ **Training only: zero inference overhead**

Best math accuracy at every scale

RESULTS

Qwen3.5 · Average accuracy on 7 math benchmarks (%)



12 benchmarks

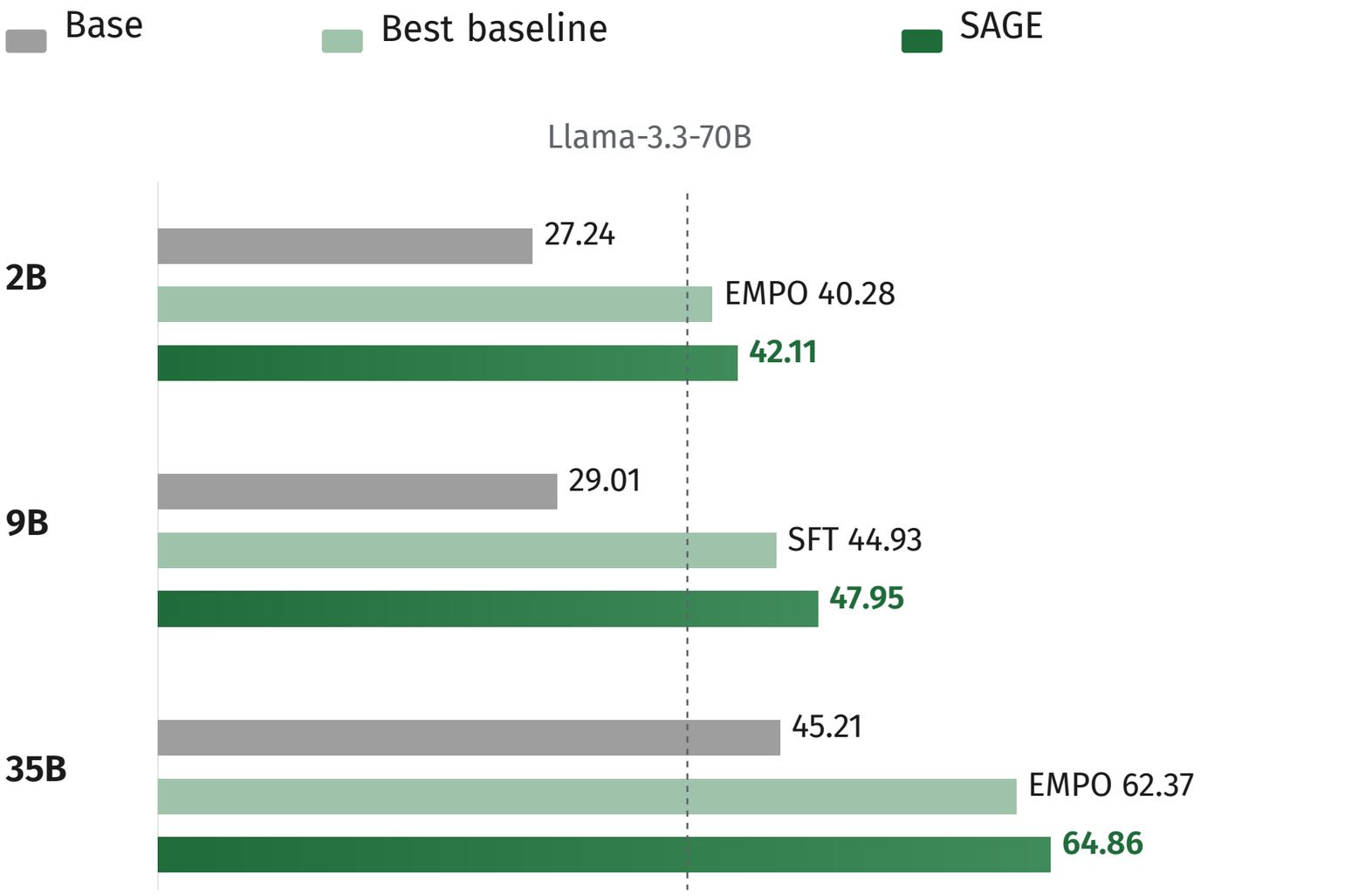
10 model families

4 baselines

Best math accuracy at every scale

RESULTS

Qwen3.5 · Average accuracy on 7 math benchmarks (%)



12 benchmarks

10 model families

4 baselines

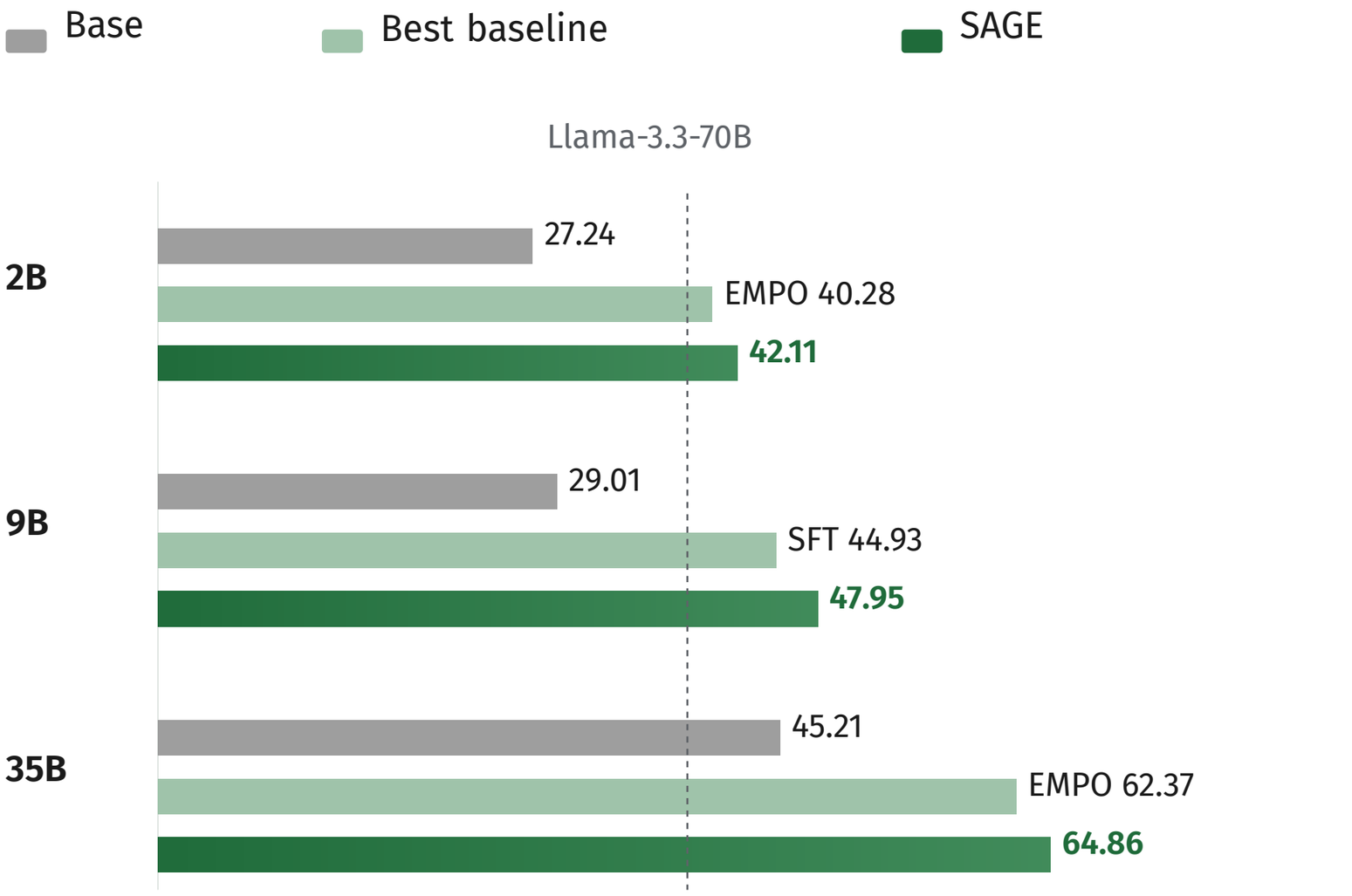
SAGE-2B beats Llama-3.3-70B

42.11 vs 38.48

Best math accuracy at every scale

RESULTS

Qwen3.5 · Average accuracy on 7 math benchmarks (%)



12 benchmarks

10 model families

4 baselines

SAGE-2B beats Llama-3.3-70B

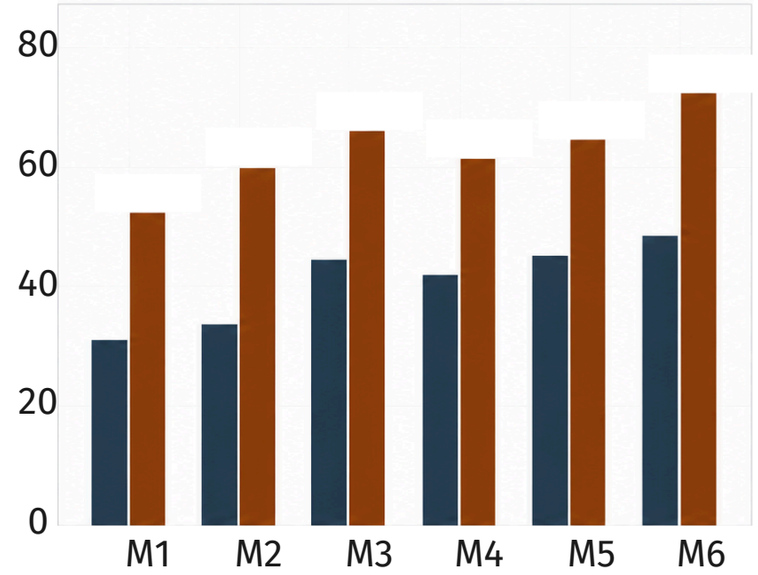
42.11 vs 38.48

Also best on free-form at 35B:
MMLU-Pro, GPQA,
BBH-H, ARC-C

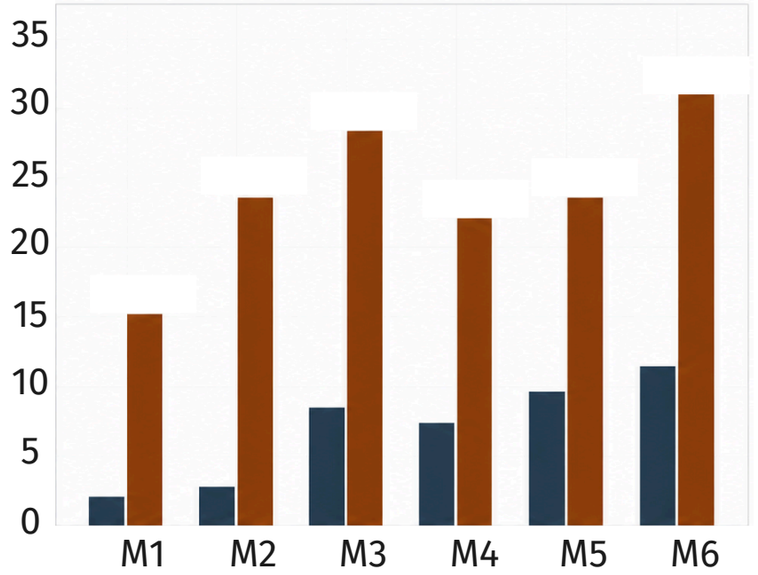
Up to 8× more Lean-verified proofs

RESULTS

AC validity: +19 to +26 points



Lean proofs: +13 to +21 points



Dark = base · Orange = + SAGE

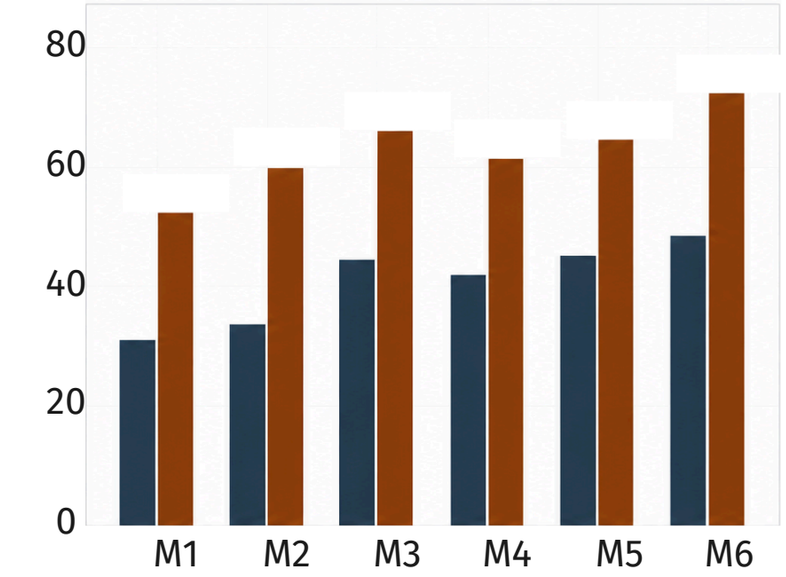
M1: DeepSeekMath-7B
M2: Qwen3
M3: DeepSeek-Prover-V2-7B
M4: Kimina-Prover-7B
M5: Kimina-Prover-8B
M6: Qwen3.6

8× Lean-verified proofs on Qwen3 vs base

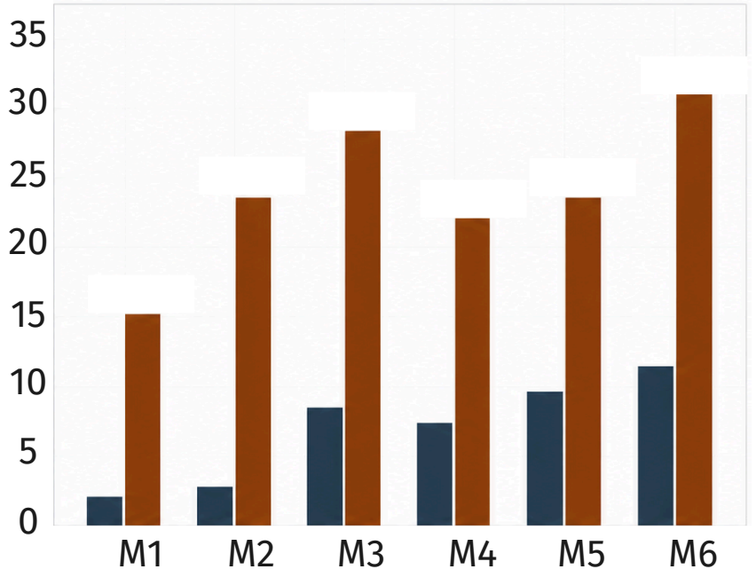
Up to 8× more Lean-verified proofs

RESULTS

AC validity: +19 to +26 points



Lean proofs: +13 to +21 points



Dark = base · Orange = + SAGE

M1: DeepSeekMath-7B
M2: Qwen3
M3: DeepSeek-Prover-V2-7B
M4: Kimina-Prover-7B
M5: Kimina-Prover-8B
M6: Qwen3.6

8× Lean-verified proofs on Qwen3 vs base

Both potentials matter · Structure beats a learned PRM

Qwen3.5-35B · Lean-verified (%)



Takeaway

1. Long-horizon RL suffers two provable biases
2. SAGE adds structure: denser signal, feasible exploration
3. Better math, free-form and proof search, no inference cost

Paper

Code

Thank you!

Xinyue Zeng[1] · Jiawei Zhang[2] · Yujun Yan[3] · Dawei Zhou[1]