

(SMEPO) Hide to Guide: Learning via Semantic Masking

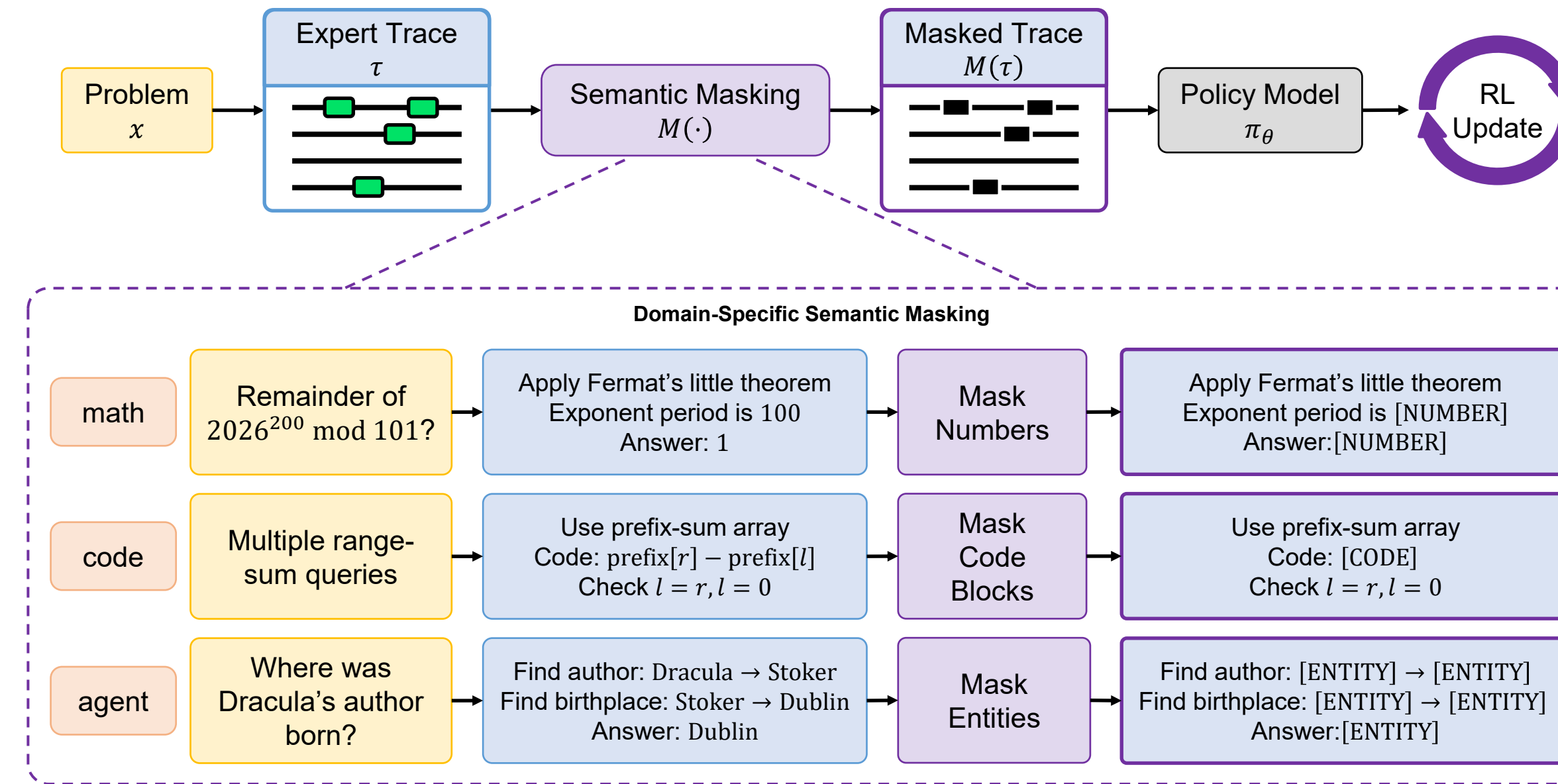
Ruitao Liu*, Qinghao Hu*, Alex Hu, Yecheng Wu, Shang Yang, Luke J. Huang, Zhuoyang Zhang, Han Cai, Song Han

Introduction

Exploration Bottleneck. RLVR improves mathematical reasoning, code generation, and agentic search, but often suffers from sparse rewards on hard problems, where the policy struggles to discover successful trajectories.

Expert-Guided RL. External expert traces provide valuable guidance for exploration, including solution plans, intermediate subgoals, and reasoning steps. However, exposing complete traces can also reveal reward-relevant content, allowing policies to obtain rewards by copying rather than learning the underlying problem-solving process.

Our Insight. Existing methods often rely on partial trajectories or truncated prefixes, controlling how much expert information is exposed rather than what should be hidden. **SMEPO preserves procedural guidance while masking reward-relevant semantic spans along the critical path.**

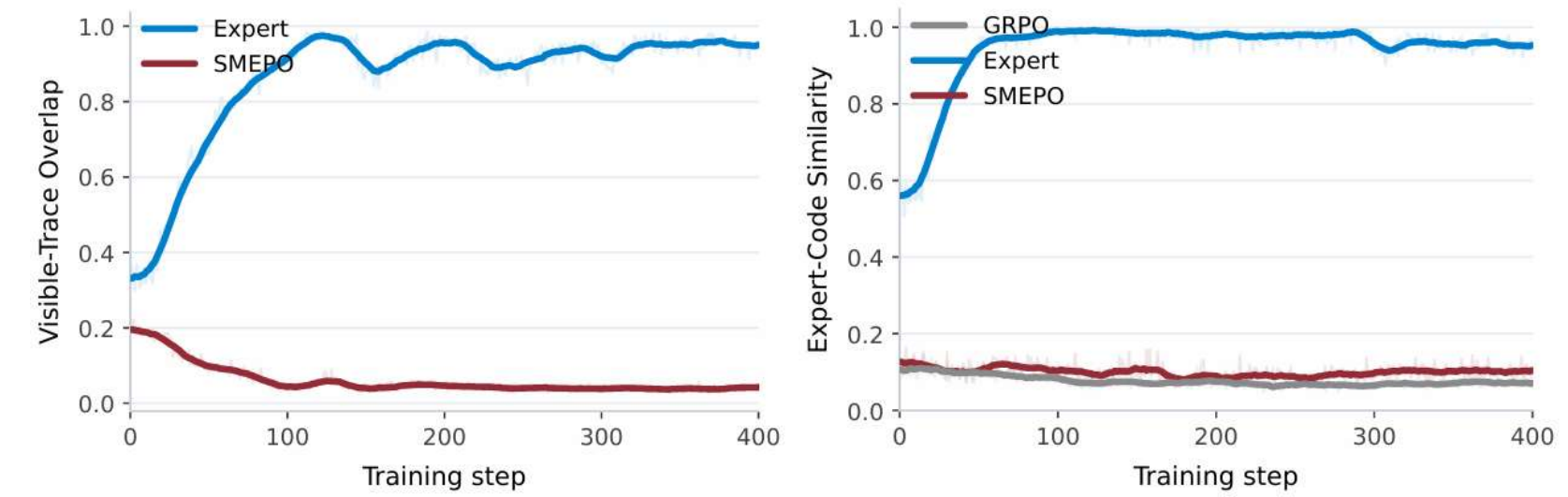


Training pipeline for SMEPO. We detect reward-relevant spans in the expert trace with domain-specific masking rules, replace those spans with mask placeholders, and then run the same GRPO loop conditioned on the original problem and the masked trace.

Diagnosis

Code-Copy

Diagnostics. Full expert traces encourage substantial copying and reuse of expert solutions, while SMEPO maintains near-zero visible-trace overlap and significantly



Code-copy diagnostics. Left: *Visible-trace overlap* captures copying from the expert content actually visible in the prompt. Right: *Expert-code similarity* captures reuse of the original expert program. The left figure does not include GRPO since it does not involve expert traces.

reduces expert-code similarity, providing evidence that semantic masking mitigates reward hacking through expert-solution copying.

$$\text{VisibleTraceOverlap} = \frac{\text{LCCS}(G_{\text{rollout}}, T_{\text{vis}})}{|G_{\text{rollout}}|}, \quad \text{ExpertCodeSimilarity} = \frac{\text{LCCS}(G_{\text{code}}, T_{\text{code}})}{|G_{\text{code}}|}.$$

Math Masking Ablations.

Prefix truncation yields non-monotonic gains, showing that reducing trace exposure is insufficient without hiding reward-relevant spans. Random masking performs poorly despite matched masking budgets. Reward-aware masking shows progressive improvements as more reward-relevant content is targeted, with SMEPO performing best by masking numerical content throughout the trace while preserving procedural guidance.

Cat.	Method	GSM8K	MATH500	AIME25	AIME26	AMC	Olympiad	Avg.
Full	Expert (100%)	76.4	69.2	11.4	8.6	52.5	27.7	41.0
	Expert (75%)	86.1	75.0	14.5	11.6	60.0	29.7	46.1
	Expert (50%)	85.2	72.6	13.5	10.3	57.5	28.8	44.7
	Expert (25%)	87.1	74.6	14.2	11.0	65.0	30.6	47.1
Random	Word	73.6	67.0	10.2	8.4	50.0	25.2	39.1
	Sentence	75.7	68.0	10.9	9.4	45.0	26.9	39.3
Targeted	Mask Final	82.9	70.4	11.7	9.1	57.5	27.9	43.2
	Mask Ans. Uses	89.9	75.2	14.4	13.2	65.0	30.4	48.0
	SMEPO	92.2	76.6	17.6	13.4	70.0	32.0	50.3

Math ablations with Qwen3-8B-Base. We compare full-expert-trace, prefix exposure, random masking, and targeted masking.

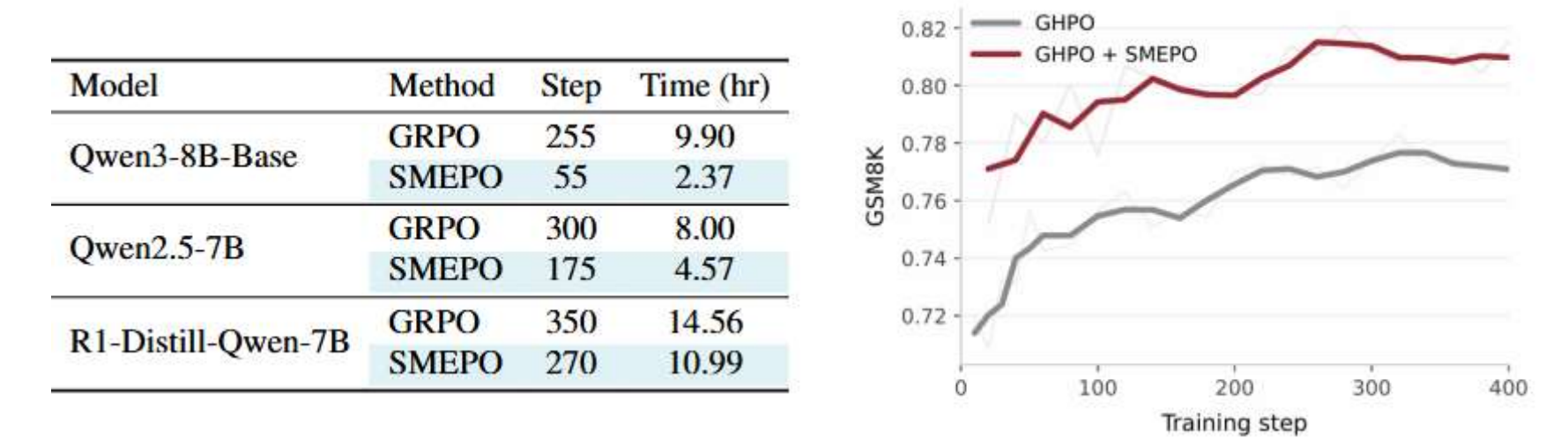
Results



Math							
Method	GSM8K	MATH500	AIME25	AIME26	AMC	Olympiad	Avg.
Qwen3-8B-Base							
GRPO	91.3	75.2	12.5	12.3	65.0	29.5	47.6
Full Expert Trace	76.4	69.2	11.4	8.6	52.5	27.7	41.0
SMEPO	92.2	76.6	17.6	13.4	70.0	32.0	50.3
Qwen2.5-7B							
GRPO	87.6	70.2	9.4	7.7	55.0	24.3	42.4
Full Expert Trace	85.3	63.8	5.6	6.4	47.5	22.8	38.6
SMEPO	89.0	72.4	10.8	9.7	60.0	25.1	44.5
DeepSeek-R1-Distill-Qwen-7B							
GRPO	86.4	84.2	34.0	38.8	90.0	37.2	61.8
Full Expert Trace	87.3	81.0	34.1	38.4	85.0	36.8	60.4
SMEPO	87.5	85.4	34.7	39.7	90.0	38.3	62.6

Code: Qwen3-8B-Base					Agent: Qwen3-8B-Base	
Method	HumanEval	HumanEval+	LCB	Avg.	Method	Bamboogle
GRPO	67.7	59.8	42.0	56.5	GRPO	45.6
Full Expert Trace	66.5	58.5	42.0	55.7	Full Expert Trace	23.2
SMEPO	69.5	62.2	45.3	59.0	SMEPO	48.8

Main results. Top: Math performance across three models. **Bottom:** Code and agentic search performance on Qwen3-8B-Base. Full expert traces often degrade performance, while SMEPO consistently outperforms vanilla GRPO and full-expert-trace conditioning.



Training efficiency and compatibility with GHPO. Left: Steps and wall-clock time to reach the GRPO peak mean@32 accuracy on AIME25 and AIME26. Right: GSM8K training curves showing performance improvements from adding SMEPO to GHPO.

Method	GSM8K	MATH500	AIME25	AIME26	AMC	Olympiad	Avg.
Qwen3-8B-Base	68.0	67.0	9.9	7.6	50.0	25.8	38.1
GHPO	78.3	72.8	13.4	10.2	60.0	29.5	44.0
GHPO + SMEPO	82.1	74.2	15.1	10.1	62.5	29.5	45.6

Method	HumanEval	HumanEval+	LCB	Avg.	Method	Bamboogle
Qwen3-8B-Base	63.4	55.5	35.0	51.3	Qwen3-8B-Base	20.8
GHPO	65.8	58.5	40.3	54.9	GHPO	42.4
GHPO + SMEPO	67.1	57.9	41.8	55.6	GHPO + SMEPO	44.0

Additional gains with GHPO. On math benchmarks with Qwen3-8B-Base, adding SMEPO to GHPO improves the six-benchmark average from 44.0 to 45.6, suggesting that masked expert traces can bring additional gains on top of prompt-refinement guidance. Similar improvements are observed in code (54.9 to 55.6) and agentic search (42.4 to 44.0), using adapted GHPO implementations.

Performance Gains. SMEPO consistently outperforms GRPO across math, code, and agentic search, with improvements of up to 3.2 points.

Efficiency. SMEPO reaches the GRPO accuracy target up to 4.2× faster.

Compatibility. Adding SMEPO to GHPO further improves performance across math, code, and agentic search, showing complementary benefits.

Method

Semantic Masking. SMEPO transforms expert traces into fill-in-the-blank guidance by masking reward-relevant semantic spans while preserving the expert's decomposition, intermediate subgoals, and procedural structure. The policy follows the expert's solution path but must reconstruct the missing information independently.

Domain-Specific Masking. In math, SMEPO masks numerical values, including intermediate results and final answers. In code generation, it masks executable code blocks while retaining algorithmic plans. In agentic search, it masks answer-relevant entities while preserving the multi-hop search trajectory.

Simple Integration. Masking is applied offline before rollout generation. The policy is conditioned on the original problem and the masked expert trace, while the verifier, reward function, and GRPO objective remain unchanged. Preprocessing takes only 37 seconds across all three training datasets.