

Beyond Experience Retrieval: Learning to Generate Utility-Optimized Structured Experience for Frozen LLMs

Xuancheng Li^{1,2,†}, Haitao Li^{1,2,†}, Yujia Zhou^{1,2}, Yiqun Liu¹, Qingyao Ai^{2,1,*}

¹Department of Computer Science and Technology, Tsinghua University, Beijing, China

²Quancheng Laboratory

lixuancheng23@mails.tsinghua.edu.cn

liht22@mails.tsinghua.edu.cn, zhouyujia@mail.tsinghua.edu.cn

yiqunliu@tsinghua.edu.cn, aiqy@tsinghua.edu.cn

Abstract

Large language models (LLMs) are largely static and often redo reasoning or repeat mistakes. Prior experience reuse typically relies on external retrieval, which is similarity-based, can introduce noise, and adds latency. We introduce **SEAM** (Structured Experience Adapter Module), a lightweight, executor-specific plug-in that stores experience in its parameters and generates a structured, instance-tailored experience entry in a single forward pass to guide a frozen LLM executor. SEAM is trained for utility via executor rollouts and GRPO while keeping the executor frozen, and can be further improved with logged-success SFT after deployment. Experiments on mathematical reasoning benchmarks show consistent accuracy gains across executors with low overhead. Extensive ablation and analysis further elucidate the mechanisms underlying SEAM’s effectiveness and robustness.¹

1 Introduction

Large language models (LLMs) have achieved strong performance across diverse tasks, including machine translation, information retrieval, and mathematical reasoning (Ahn et al., 2024; Naveed et al., 2025; Shi et al., 2025). Yet, they remain largely *static*: when confronted with new problems, LLMs often reason from scratch, re-explore familiar solution paths, and repeat avoidable mistakes (Feng et al., 2025; Fan et al., 2025; Yu et al., 2025). In contrast, humans distill procedural experience from past problem solving and apply it to new situations, improving efficiency and avoiding prior failures (Efklides, 2009; Kolb, 2014; Usher, 1993). Ideally, an agent should also accumulate reusable experience for fast adaptation to novel problems.

Effectively leveraging prior experience requires answering two questions: how to extract useful experience from interactions and how to apply it to new problems. A common approach is to maintain an explicit experience library and consult it at inference time via retrieval-augmented generation (RAG) (Chhikara et al., 2025; Fang et al., 2025; Zhou et al., 2025). However, a key limitation is that retrieval is typically optimized for surface-level similarity rather than practical utility to the executor, and similarity does not imply effectiveness. Even when semantically close, retrieved entries may fail to yield meaningful improvements. They can miss decisive constraints or critical checks, or add noise that distracts reasoning and destabilizes solving. In addition, external libraries add noticeable inference overhead by requiring retrieval-time computation and often extra LLM calls for summarization or refinement, resulting in higher latency.

In this paper, we take a different approach: instead of maintaining a structured experience library as an external store, we encode experience in the parameters of a lightweight model and use it to guide a frozen downstream LLM executor. Accordingly, we introduce **SEAM** (Structured Experience Adapter Module), a compact, trainable, and executor-specific plug-in. We view each executor as a distinct problem solver with its own strengths, inductive biases, and failure modes; SEAM therefore stores executor-tailored experience in its parameters and, at inference time, generates a structured, instance-tailored experience entry in a single forward pass, steering the frozen executor without modifying its parameters. This parameterized design eliminates the need for complex external curation and organization, since these choices are learned implicitly during training. More importantly, unlike prior RAG-style experience retrieval that relies on similarity-based selection, SEAM is optimized for utility—it is trained to generate experience entries that actually improve

[†]Equal contribution.

^{*}Corresponding author.

¹We release our code at <https://github.com/Cocobalt/SEAM>.

the frozen executor’s task performance.

We train SEAM with a forward learning procedure consisting of three training steps: (i) forward exploration, where SEAM samples multiple schema-constrained experience entries per instance; (ii) rollout-based evaluation, where a frozen executor solves each instance conditioned on each candidate and we score candidates by task success; and (iii) parametric library evolution, where we compute group-relative advantages from candidates’ returns and update SEAM under the GRPO objective while keeping the executor frozen. At test time, SEAM synthesizes a task-relevant experience entry in a single forward pass to steer the executor. Optionally, SEAM can be further improved after deployment by logging successful entries and periodically applying SFT on accumulated successes.

We conduct experiments on diverse mathematical reasoning benchmarks to assess SEAM. Across multiple benchmarks and executors, SEAM improves accuracy with low additional overhead. We further conduct a series of analyses to investigate SEAM’s internal mechanisms and dynamics.

2 Related Work

Learning from Experience. Learning from experience enables LLM-based agents to reuse prior knowledge for new tasks, improving autonomy, efficiency, and performance in dynamic settings (Tan et al., 2025; Zheng et al., 2023; Wang et al., 2024b; Liu et al., 2025). Most prior work instills this capability by directly training the agent (e.g., reinforcement learning, imitation learning, and multi-agent learning) (Dong et al., 2025; Yang et al., 2024; Fang et al., 2025), but often suffers from limited use of negative examples, weak cross-task generalization, and catastrophic forgetting. As an alternative, RAG-based frameworks extract structured experience from past rollouts, store it in a memory bank, and retrieve relevant snippets to augment prompts for new tasks (Suzgun et al., 2025; Mishra et al., 2025; Chhikara et al., 2025). Despite their flexibility, these pipelines rely on surface-level similarity matching that can yield low-utility prompts (Du et al., 2025; Salama et al., 2025), introduce nontrivial latency and compute overhead (Wang et al., 2024a; Quinn et al., 2025), and depend on hand-crafted storage templates that require continual maintenance as tasks evolve (Xu et al., 2025).

LLM and Reinforcement Learning. Reinforcement learning (RL) is increasingly used to move

beyond supervised fine-tuning and enable LLMs to learn from interactive feedback. RLHF aligns outputs with human preferences (Ouyang et al., 2022), and recent work adopts GRPO to improve long-horizon reasoning and tool use under environment interaction signals. For example, Search-R1 trains models to issue web queries to maximize answer correctness (Jin et al., 2025), Tool-N1 learns when to invoke external tools (Zhang et al., 2025), and RAGEN extends GRPO to multi-turn self-evolution (Wang et al., 2025). In learning-from-experience settings, RL is typically applied either to directly optimize the executor (Lu et al., 2025) or to train a memory agent within RAG to extract and inject memories (Yan et al., 2025). In contrast, SEAM applies GRPO to train a lightweight, executor-specific guidance policy that internalizes experience in parameters and generates problem-conditioned structured experience guidance to steer a frozen executor.

3 Method

3.1 Problem Formulation

Once deployed, LLMs operate in a fundamentally stateless manner: let \mathcal{D} denote a distribution over task instances s , where each instance may be associated with a verifiable target $a^*(s)$. Given a new instance s sampled from \mathcal{D} , a frozen executor LLM \mathcal{E}_ϕ typically generates its output \hat{a} without conditioning on accumulated experience:

$$\hat{a} \sim \pi_\phi(\hat{a} | s). \quad (1)$$

This limitation arises because LLMs are bounded by a finite context window, which prevents them from retaining and reusing information across long interactions or evolving task sequences. Consequently, each new instance must be reasoned through from scratch, leading to redundant exploration of solution paths already discovered and frequent repetition of past errors.

This motivates us to equip LLMs with an experience module \mathcal{A}_θ that maintains experience and provides guidance to the frozen executor. Formally, for an instance s sampled from \mathcal{D} and a specific executor \mathcal{E}_ϕ , the experience module produces an auxiliary structured experience prompt z of limited length:

$$z \sim \pi_\theta(z | s, \mathcal{E}_\phi), \quad (2)$$

The executor \mathcal{E}_ϕ then solves the task conditioned on this experience:

$$\hat{a} \sim \pi_\phi(\hat{a} | s, z). \quad (3)$$

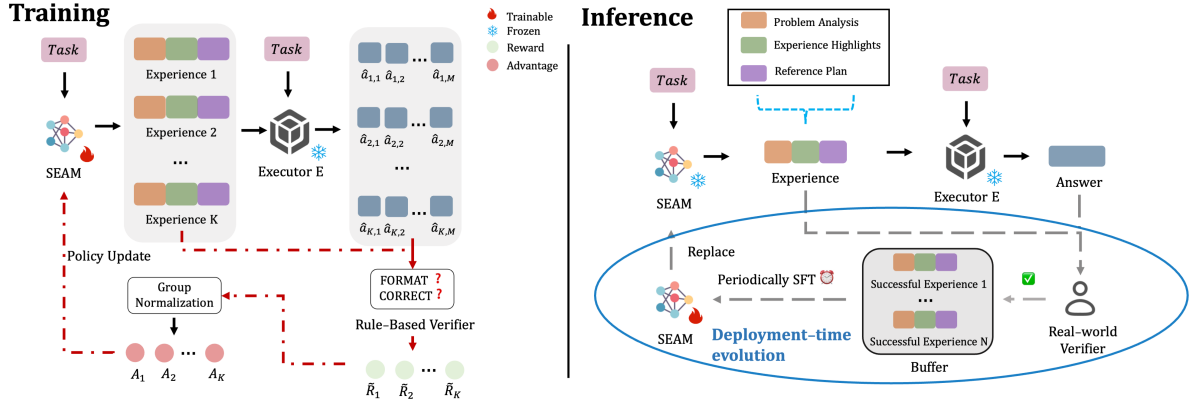


Figure 1: Illustration of SEAM training and inference pipelines.

The design of the experience module \mathcal{A}_θ can thus be formulated as the following constrained optimization problem:

$$\begin{aligned} \max_{\pi_\theta} \quad & \mathbb{E}_{s \sim \mathcal{D}, z \sim \pi_\theta(\cdot|s, \mathcal{E}_\phi), \hat{a} \sim \pi_\phi(\cdot|s, z)} [R(\hat{a}, a^*(s))] \\ \text{s.t.} \quad & \|z\| \leq L. \end{aligned} \quad (4)$$

where $R(\cdot)$ denotes a reward measuring task correctness, $a^*(s)$ is the ground-truth (or verifiable) target for instance s , and L is a predefined maximum length of the generated experience. In our main experiments, we instantiate R using correctness on verifiable reasoning tasks; the concrete reward definition is given in Section 3.3.

Most existing work instantiates \mathcal{A}_θ with the RAG pipeline: a LLM is first prompted to retrospectively summarize its trajectories into explicit experience entries, which are stored in an external bank. At inference time, the experience policy performs a search step that retrieves k candidates from this store, followed by a generation step that rewrites the selected items into the prompt z . However, two key limitations remain: (i) **Similarity \neq utility**—retrieval is optimized for surface relevance, so retrieved items can be noisy or miss decisive constraints/checks, yielding little improvement or even destabilizing the solver; (ii) **Inference overhead**—external libraries add noticeable latency due to retrieval-time computation and often extra LLM calls for refinement.

3.2 Architecture

To address these limitations, we adopt a fundamentally different paradigm: instead of maintaining an external experience store and retrieving entries via a RAG pipeline, we *parameterize* the structured experience library by encoding it in a lightweight generator, which can be accessed with a single forward pass. We introduce **SEAM** (Structured Experience Adapter Module), a compact and train-

able plug-in that implicitly maintains a structured experience library in its parameters. Given a new problem instance s for a downstream LLM executor \mathcal{E}_ϕ , SEAM synthesizes a schema-constrained experience entry z in one forward pass, and \mathcal{E}_ϕ solves the task conditioned on z .

This design shifts experience management away from explicit retrieval systems into a lightweight generative module and has several key characteristics. **(1) Efficient parametric experience access.** SEAM internalizes a structured experience library in its parameters and produces guidance via a single forward pass. This avoids the high cost of RAG-style pipelines, notably external storage, online retrieval, and retrospective summarization by large models, yielding low overhead and low latency at inference time. Moreover, this parameterized design eliminates the need for complex external curation and organization (e.g., indexing, clustering, deduplication, and retrieval policies), as these choices are absorbed into SEAM and learned implicitly during training. **(2) Utility-oriented guidance instead of surface-level retrieval.** At test time, SEAM does not retrieve entries based on embedding similarity. Instead, it *synthesizes* an instance-tailored experience entry under strict schema constraints and is trained using executor-conditioned utility signals, instantiated in our current setting by task success. As a result, SEAM learns to generate structured experience with actionable utility for solving the current instance, rather than returning superficially relevant but low-utility snippets that may dilute or destabilize the solver. **(3) Executor-specific experience libraries.** We adopt an *executor-specific* perspective: different downstream agents behave like different problem solvers, with distinct strengths, inductive biases, and failure modes. Accordingly, SEAM maintains a dedicated parametric experience library for a

particular frozen executor and generates *dedicated*, customized, and structured guidance for that executor, without modifying the executor’s parameters. **(4) Lightweight, pluggable modular adaptation.** We train SEAM while keeping the downstream executor \mathcal{E}_ϕ frozen, which reduces training and deployment costs, enables hot-swapping across domains, and preserves \mathcal{E}_ϕ ’s general competence and stability. Moreover, operational rollouts can be readily logged and recycled as feedback, enabling continual training and steady evolution of SEAM without modifying \mathcal{E}_ϕ .

As illustrated in Figure 1, SEAM provides the executor with a *structured experience entry* z as a textual prompt. Each entry follows a fixed schema with three complementary components:

- *Problem analysis* — a concise assessment of the instance s , highlighting intrinsic difficulty, salient substructures, and likely failure modes;
- *Experience highlights* — executor-aligned insights distilled from prior rollouts, specifying reusable strategies, heuristics, or cautionary checks that help \mathcal{E}_ϕ solve s ;
- *Reference plan* — an *illustrative* step-by-step decomposition procedure that demonstrates a reliable workflow, rather than revealing instance-specific shortcuts or the final answer.

This schema-bound design encourages SEAM to produce guidance that is simultaneously diagnostic (problem analysis), prescriptive (experience highlights), and procedural (reference plan), providing a stable and controllable interface for steering the frozen executor. We emphasize that the schema primarily serves as a controllable interface between SEAM and the frozen executor. Its role is to constrain generated guidance into a form that can be stably and effectively exploited by a fixed solver, rather than to provide a task-specific template that is inherently well matched to a particular benchmark. In this sense, the current schema is especially natural for procedural domains, while broader domains may benefit from alternative schema designs under the same training principle.

3.3 Training of SEAM

We train SEAM with a forward learning procedure that treats SEAM as an executor-conditioned guidance policy and improves it using observable outcomes from running a frozen executor. SEAM is not a surrogate reasoner; it provides procedure-level, schema-constrained guidance that biases the

executor’s exploration toward more reliable solution trajectories. Concretely, training iterates over three steps: (i) **forward exploration**, where SEAM samples a group of schema-constrained candidate experience entries for each instance; (ii) **rollout-based evaluation**, where we condition the frozen executor on each candidate and score its utility by downstream task success; and (iii) **policy evolution**, where we update SEAM (e.g., via GRPO) to increase the likelihood of higher-utility entries. Throughout training, the executor remains frozen and only SEAM is updated.

Step 1: Forward exploration (candidate generation). For each training instance $s \sim \mathcal{D}$, SEAM samples a group of K candidate experience entries to guide the specific executor \mathcal{E}_ϕ :

$$\{z_j\}_{j=1}^K \sim \pi_\theta(\cdot | s, \mathcal{E}_\phi). \quad (5)$$

Step 2: Rollout-based evaluation (guidance utility). To evaluate the guidance utility of each candidate for the target executor \mathcal{E}_ϕ , we condition the frozen executor on each z_j and perform M stochastic rollouts.

$$\hat{a}_{j,m} \sim \pi_\phi(\cdot | s, z_j), \quad m = 1, \dots, M. \quad (6)$$

We score each rollout using a binary reward, which jointly reflects task success and structural completeness:

$$\mathcal{R}(\hat{a}, z; s) = \begin{cases} 1, & \hat{a} \text{ correct, } z \text{ complete,} \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Equation (7) defines the concrete reward used in the present instantiation of SEAM. More generally, SEAM requires a utility signal for comparing multiple schema-constrained experience candidates for the same instance under a fixed executor, thereby identifying which forms of guidance are more effectively exploitable by that executor. Such a signal need not be a hard rule-based reward; depending on the setting, it may instead be instantiated using preferences, judge-based assessments, or other task-level evaluations. We aggregate rollout outcomes by averaging rewards across rollouts for the same candidate:

$$\tilde{R}_j = \frac{1}{M} \sum_{m=1}^M \mathcal{R}(\hat{a}_{j,m}, z_j; s). \quad (8)$$

Step 3: Parametric library evolution (GRPO update). We update SEAM using GRPO with the group rewards $\{\tilde{R}_j\}_{j=1}^K$. Let π_{old} denote the behavior policy used to sample $\{z_j\}$ (a snapshot of SEAM before the update), and π_{ref} be a fixed reference policy (a copy of the initial SEAM). We

compute group-relative advantages by normalizing rewards within the group:

$$A_j = \frac{\tilde{R}_j - \bar{R}}{\sqrt{\frac{1}{K} \sum_{i=1}^K (\tilde{R}_i - \bar{R})^2} + \delta}, \quad \bar{R} = \frac{1}{K} \sum_{i=1}^K \tilde{R}_i, \quad (9)$$

and perform a PPO-style clipped update over experience candidates:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = & -\frac{1}{K} \sum_{j=1}^K \min\left(\rho_j A_j, \text{clip}(\rho_j, 1 - \epsilon, 1 + \epsilon) A_j\right) \\ & + \beta \text{KL}\left(\pi_\theta(\cdot | s, \mathcal{E}_\phi) \parallel \pi_{\text{ref}}(\cdot | s, \mathcal{E}_\phi)\right), \\ \rho_j = & \frac{\pi_\theta(z_j | s, \mathcal{E}_\phi)}{\pi_{\text{old}}(z_j | s, \mathcal{E}_\phi)}. \end{aligned} \quad (10)$$

Crucially, gradients are backpropagated only through SEAM; the executor \mathcal{E}_ϕ remains frozen. From a high-level perspective, we optimize one model using the observable execution behavior of another: we treat \mathcal{E}_ϕ as a fixed solver that provides behavioral signals, and train SEAM to generate the structured guidance that \mathcal{E}_ϕ can best exploit. In this way, SEAM is not a surrogate reasoner; instead, it supplies procedure-level, schema-constrained cues that bias the executor’s exploration toward more reliable solution trajectories. This improves guidance and end-to-end effectiveness without backpropagating through \mathcal{E}_ϕ . Such a decoupled paradigm enables SEAM to continually improve from execution feedback without modifying \mathcal{E}_ϕ ’s parameters, thereby preserving the executor’s general competence while enhancing robustness and overall performance.

Optional: Deployment-time evolution via logged SFT. As an optional extension, SEAM can further improve during deployment by logging successful experience entries that actually guide the frozen executor to correct solutions. When enabled, we store these pairs (s, z^*) in a buffer \mathcal{B} and periodically fine-tune SEAM with teacher forcing:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(s, z^*) \sim \mathcal{B}} \left[\log \pi_\theta(z^* | s, \mathcal{E}_\phi) \right]. \quad (11)$$

This update again modifies only SEAM while keeping the executor frozen, progressively internalizing effective experience from concrete past episodes into the parametric library.

4 Experiment

4.1 Experiment Settings

Datasets. For training, we randomly sample 5k instances from the DAPO dataset, which provides diverse problem statements with annotated solutions. Evaluation is conducted on four widely used

mathematical reasoning benchmarks, including GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), AIME24 and AIME25. These benchmarks cover both routine arithmetic and competition-level mathematics, enabling us to assess generalization across varying difficulty. We use pass@1 as the primary metric and report the mean over 10 independent runs per dataset. In addition, for our cross-domain analysis on code and QA benchmarks including CodeContests (Li et al., 2022), MBPP (Austin et al., 2021), HotpotQA (Yang et al., 2018), and Natural Questions (NQ) (Kwiatkowski et al., 2019), we evaluate on a random subset of 300 instances from each benchmark.²

Baselines. In our main experiments, we compare SEAM with several representative baselines. Together, these baselines span three families: (i) direct use of the original model without experience memory; (ii) directly training the executor, which can be viewed as enabling the executor to internalize and apply experience; and (iii) RAG-based approaches for learning from experience.

- **Original Model.** The frozen executor \mathcal{E}_ϕ is directly applied to test problems without any memory augmentation or additional training. This serves as the minimal baseline.
- **Training-based method.** Directly train the executor \mathcal{E}_ϕ using GRPO, baking prior experience into \mathcal{E}_ϕ ’s parameters.
- **RAG-based methods.**
 - **MEM-0** (Chhikara et al., 2025): A scalable long-term memory framework that dynamically extracts–integrates–retrieves salient information to form a persistent external memory for relational reasoning.
 - **Dynamic-Cheatsheet (DC)** (Suzgun et al., 2025): Augments black-box LMs with a persistent, self-curated test-time memory of concise, transferable strategies. We adopt the DC-RS (retrieve-and-synthesize) variant.
 - **Memento** (Zhou et al., 2025): Equips black-box LLMs with an external episodic memory, retrieving and adapting prior trajectories for case-based planning without fine-tuning; supports learned Q-function retrieval with online write/read.

²Dataset details and access links are provided in § A.

Implementation Details. We adopt QWEN3-0.6B (Yang et al., 2025) as the base model for SEAM and, unless otherwise specified, report results with the GRPO-trained SEAM (without deployment-time SFT). We use QWEN3-4B (Yang et al., 2025) and DEEPSEEK-R1-DISTILLED-LLAMA-3.1-8B (Guo et al., 2025) as executors in the main experiments, and all baselines share the same frozen executor (i.e., their *executor* is identical to ours). We enable QWEN3 thinking mode throughout. Unless otherwise specified, we cap the maximum length of each generated experience entry to 4k tokens and the executor output to 8k tokens. For RAG-based baselines, we run them on top of the same frozen executor; for methods with an external experience bank, we pre-fill the bank using experiences constructed from the same training set, and then update it sequentially after each evaluation instance. All methods use greedy decoding for evaluation, and all experiments are conducted on 4 NVIDIA A100 (80GB) GPUs.³

4.2 Main Results

Table 1 shows across four math-reasoning benchmarks and two executors, SEAM achieves the best performance on all datasets, demonstrating the effectiveness and generality across different executors. Relative to directly training the executor, SEAM attains superior overall performance, indicating that internalizing the experience in a small parameterized generator that provides task-conditioned test-time guidance is effective and can even surpass directly optimizing the executor itself. This likely reflects an inductive-bias difference: directly training the executor updates a global behavior policy and aggregates credit across heterogeneous instances, whereas SEAM learns instance-conditioned guidance that injects procedural priors and reshapes the executor’s search space on a per-problem basis without perturbing the executor’s general reasoning capabilities.

Compared with RAG-based approaches, SEAM delivers substantially stronger performance, indicating that its mechanism of implicitly parameterizing experience and directly generating task-specific guidance addresses key limitations of RAG such as low-quality experience summarization and weak downstream guidance. Regarding task difficulty, the largest performance gains of SEAM emerge on the more challenging AIME problems,

whereas on the comparatively saturated MATH500 and GSM8K benchmarks, SEAM still provides modest improvements. This pattern suggests that experience-based guidance becomes increasingly critical and effective for the executor when tackling harder tasks. Collectively, these findings affirm the robustness and effectiveness of our design across diverse models and datasets.

4.3 Ablation Study

Training regimes. We compare four training regimes: (i) *SEAM without training*, which uses the same prompting strategy and experience format but applies no learning; (ii) *SEAM*, trained on the sampled DAPO dataset using GRPO; (iii) *SEAM with SFT-only*, trained only with online teacher forcing on logged successful data; and (iv) *SEAM with deployment-time evolution*, which after GRPO training applies the same logged successful data.

For each benchmark, we randomly split the test set into a 30% stream $\mathcal{D}_{\text{stream}}$ and a disjoint 70% evaluation set $\mathcal{D}_{\text{eval}}$. To collect data for online SFT, we keep the executor frozen and process instances from $\mathcal{D}_{\text{stream}}$. Whenever SEAM’s experience leads the executor to a correct solution, we log the successful pair (s, z^*) into a buffer. We iterate over $\mathcal{D}_{\text{stream}}$ for up to 10 rounds, stopping early once the buffer reaches 1k successful samples or when 10 rounds are completed.

Results and discussion. Table 2 shows that training is essential beyond prompting and experience formatting. SFT-only provides a modest improvement over no training, but remains less effective than GRPO, suggesting that directly maximizing the likelihood of logged successful experiences is insufficient to reliably optimize guidance utility under a frozen executor. In contrast, GRPO-trained SEAM learns to generate structured experiences that are better aligned with both the task and the frozen executor’s execution behavior, yielding more informative problem analyses and actionable step-wise strategies, and thereby providing consistent improvements. Finally, deployment-time evolution offers an additional refinement by internalizing experiences that demonstrably succeed in situ, helping SEAM adapt to deployment-specific patterns and distributional shifts. More principled deployment-time update schemes remain an interesting direction for future work.

³More details are provided in § B.

Table 1: Main results across four math-reasoning benchmarks, evaluated by pass@1.

Methods	AIME24	AIME25	MATH500	GSM8K
QWEN3-4B				
Original Model	0.407	0.380	0.784	0.932
GRPO	0.473	0.440	0.756	0.923
MEM-0	0.413	0.403	0.792	0.938
Dynamic-Cheatsheet	0.447	0.417	0.803	0.943
Memento	0.463	0.433	0.811	0.947
SEAM	0.504	0.460	0.818	0.958
DEEPSEEK-R1-DISTILLED-LLAMA-3.1-8B				
Original Model	0.357	0.243	0.734	0.773
GRPO	0.437	0.323	0.795	0.852
MEM-0	0.397	0.297	0.768	0.812
Dynamic-Cheatsheet	0.383	0.283	0.774	0.817
Memento	0.413	0.320	0.787	0.829
SEAM	0.442	0.352	0.794	0.856

Table 2: Ablation on training regimes with the QWEN3-4B executor (pass@1; mean over 10 random \mathcal{D}_{eval} splits).

Methods	AIME25	MATH500
SEAM	0.460	0.818
SEAM without training	0.374	0.790
SEAM with SFT only	0.425	0.809
SEAM with deployment-time evolution	0.482	0.836

5 Analysis

To better understand SEAM, we analyze its efficiency, robustness to experience size, and generalization across model pairings and domains.

5.1 Efficiency

Table 3 summarizes efficiency along three axes: training efficiency, per-instance inference latency, and time-to-correct⁴. During training, our method is substantially more efficient than directly fine-tuning a large model, because only the lightweight module is optimized while the executor remains frozen, which reduces compute requirements. At inference time, SEAM introduces a small, fixed, per-instance latency due to generating a structured experience that augments the executor. When measured by time-to-correct, SEAM exhibits clear advantages: the auxiliary experience it provides raises task success rates and shortens average problem-

⁴We run the system once per instance. Let N be the number of evaluation instances, t_i the end-to-end wall-clock latency for instance i , and $c_i = \mathbb{I}[\hat{a}_i \text{ is correct}]$ the correctness indicator. Then $TTC = \frac{\sum_{i=1}^N t_i}{\sum_{i=1}^N c_i} = \frac{\bar{t}}{\text{Acc}}$, where $\bar{t} = \frac{1}{N} \sum_{i=1}^N t_i$ and $\text{Acc} = \frac{1}{N} \sum_{i=1}^N c_i$.

Table 3: Efficiency comparison on DEEPSEEK-R1-DISTILLED-LLAMA-3.1-8B using $4 \times A100$ 80GB, evaluated on AIME25. GRPO (Full) denotes full-parameter GRPO training, while GRPO (LoRA) denotes GRPO training with LoRA adapters. Training efficiency is reported as GPU-hours to convergence. Inference efficiency is measured by the average end-to-end wall-clock latency per problem (*Per-inst.*) and by *Time-to-correct (TTC)*, defined as the total inference time over the evaluation set divided by the number of correctly solved problems.

Method	Training Efficiency	Inference Efficiency	
	GPU-hours to conv.	Per-inst. (s)	TTC (s)
Original Model	–	0.85	3.50
GRPO (Full)	104	0.81	2.51
GRPO (LoRA)	104	0.81	2.51
MEM-0	–	3.86	13.00
Dynamic-Cheatsheet	–	4.05	14.31
Memento	–	3.93	12.28
SEAM	53	0.95	2.70

solving time compared with RAG-based baselines. Overall, the design delivers notable gains in both training efficiency and solution speed while maintaining competitive inference latency.

5.2 Model Pairing

We systematically examine the pairings of SEAM’s different lightweight models with various executors in Table 4. On the side of smaller lightweight models, Qwen3-0.6B yields some improvements, but its limited generalization and generation quality lead to experience entries that lack stability and coverage in complex tasks, resulting in only modest gains. In contrast, Qwen3-1.7B exhibits

Table 4: AIME24 pass@1 for pairing SEAM scales with executors. All SEAM models are trained independently with rollouts from their respective paired executors; *w/o SEAM* reports the executor-only baseline without any extra training.

SEAM	Executor		
	Qwen3-4B	Qwen3-8B	Avg.
<i>w/o SEAM</i>	0.407	0.635	0.521
SEAM-Qwen3-0.6B	0.504	0.643	0.574
SEAM-Qwen3-1.7B	0.517	0.697	0.607

Table 5: Cross-executor transfer on AIME25: relative improvement (%) when applying a SEAM trained with QWEN3-4B as the executor to other executors, without any additional training. Relative improvement is computed as $[(\text{with SEAM} - \text{original}) / \text{original}] \times 100\%$.

Target Executor	Rel. imp.
DEEPSEEK-R1-DISTILLED-LLAMA-3.1-8B	+27.9%
GPT-4o	+22.7%
CLAUDE-3.5-SONNET	+25.3%

stronger generalization and more structured generation, thereby providing more stable and significant improvements for the executor on harder tasks. On the side of executors, stronger LLMs are already capable of solving a substantial portion of problems correctly on their own, which reduces the marginal benefit brought by external experience, making the improvements from SEAM less pronounced. Overall, there exists a complementary relationship between the scale of SEAM and the capability of the executor: given a fixed executor, enhancing the generative capacity of SEAM effectively amplifies the value of external experience; whereas when the executor itself is already sufficiently strong, the gains from external experience saturate, in which case further strengthening SEAM (e.g., with a more powerful base model or improved training) may be necessary to achieve additional benefits.

5.2.1 Cross-model Generalization

We further conducted a cross-executor transfer experiment to examine the portability of SEAM across different executors. Specifically, we directly migrated the SEAM trained in §4.2 with QWEN3-4B as the executor to other executors of different architecture or scale, without any additional training or adaptation. The results in Table 5 show that the transferred SEAM can still provide a certain degree of performance improvement on the new executor, indicating that the generated experience exhibits some level of generality. However, for DEEPSEEK-R1-DISTILLED-LLAMA-3.1-8B executor, the improvement is relatively limited and remains below

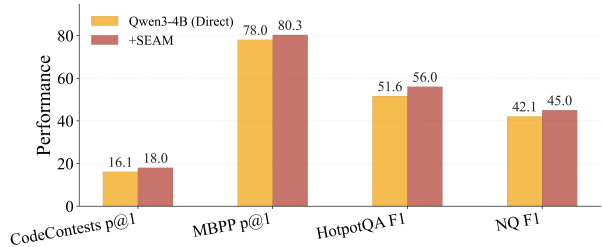


Figure 2: The cross-domain results for SEAM with Qwen3-4B as executor.

the performance of the original SEAM-executor pairing (44.8%⁵ vs. 27.9%). This limitation arises from SEAM being optimized for a specific executor; variations in executors’ capabilities and characteristics hinder portability and can degrade performance under direct transfer. To address this, future work could explore joint training across multiple executors.

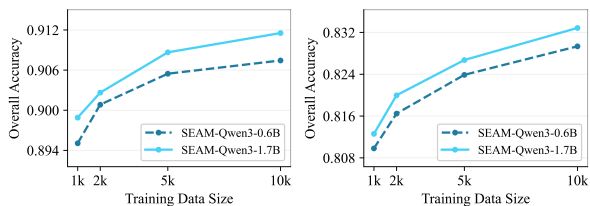
5.2.2 Cross-domain Generalization

We also investigated the cross-domain capability of SEAM. We directly use the SEAM trained in §4.2 with QWEN3-4B as the executor as well. As illustrated in Figure 2, although SEAM is trained solely on mathematical data, it still enhances performance in two markedly different domains. This suggests that SEAM acquires not just math-specific skills but a more general reasoning framework, including problem decomposition, explicit step planning, and the avoidance of common pitfalls. When the executor already possesses a baseline level of competence, these structured cues further improve its effectiveness. However, the cross-domain gains are relatively limited, because SEAM lacks experience for these unfamiliar domains. Enabling a single SEAM to efficiently switch across and adapt to different domains is a key direction for future.

5.3 Experience Scaling

We study the effect of experience size by changing the number of entries used to train SEAM while keeping the executor fixed, and evaluating overall accuracy aggregated over all instances from the four mathematical benchmarks (Figure 3). Performance improves monotonically as the experience size grows, indicating that SEAM benefits primarily from accumulating and reusing more high-quality experiences rather than from sampling luck. The gains show diminishing returns at larger scales, with improvements gradually tapering off, suggest-

⁵The result is computed from the data reported in Table 1.



(a) QWEN3-4B executor. (b) DEEPSEEK-R1 executor.

Figure 3: The results of scaling with experience size.

ing that further progress likely requires richer diversity and better coverage of experience. Larger SEAM variants consistently outperform smaller ones and gain more from additional experiences, revealing a complementary relationship between model capacity and experience scale. These trends are consistent across executors, demonstrating robust and transferable scaling behavior.

6 Conclusion

We presented **SEAM**, a lightweight, executor-specific module that parameterizes experience and produces structured, problem-conditioned guidance in a single forward pass to steer a frozen LLM executor. SEAM is trained with a three-step pipeline: (1) generate multiple schema-constrained experience candidates; (2) evaluate each candidate via executor rollouts and task success; and (3) update SEAM with GRPO using group-relative advantages to internalize the most effective experiences. SEAM can further evolve at deployment time by applying SFT on accumulated successes. Experiments and additional analyses demonstrate SEAM’s effectiveness and robustness across different frozen executors, and position it as a practical alternative to retrieval-based experience reuse. We hope this work motivates a broader paradigm of learning from experience beyond RAG.

Limitations

Despite the promising results, our current evidence remains subject to several important limitations.

Limited scope of current evidence. Our empirical evaluation is primarily concentrated on mathematical reasoning benchmarks and a small set of executors. While these results support the effectiveness of SEAM in this setting, they do not yet establish equally strong conclusions for broader task families, more diverse domains, or substantially different executor classes.

Executor-specificity as a design trade-off.

SEAM is deliberately trained in an executor- and domain-specific manner so that the generated guidance matches a fixed solver’s strengths, reasoning preferences, and failure modes. As a result, a single SEAM instance that adaptively transfers across heterogeneous models and domains is not a demonstrated property of the current system, but rather an open problem for future work.

Continual evolution remains underexplored.

Although SEAM can be further updated using logged feedback, we do not yet study its robustness under distribution shift, noisy or delayed supervision, and long-term continual adaptation. In particular, reliable experience evolution over time without degradation or catastrophic forgetting remains insufficiently characterized in the current work.

Ethical Considerations

Our work studies how to improve a frozen large-language executor by training a lightweight guidance model (SEAM) from execution feedback. We discuss potential ethical risks and mitigations below.

Intended use and misuse. SEAM is designed to generate structured, task-conditioned guidance that improves reliability and efficiency for benign problem-solving tasks (e.g., mathematical reasoning). Like other LLM-based systems, however, it could be repurposed to assist with harmful activities by providing step-by-step procedural guidance. We do not propose SEAM as a safety mechanism, and our method should be deployed with standard safeguards (e.g., content filtering, policy enforcement, and monitoring) when used in real-world applications.

Data and privacy. Our experiments use public datasets and do not involve personal data collection. In the optional deployment-time evolution setting, the system may log successful interactions for supervised fine-tuning. If applied beyond our experimental setup, such logging could capture user-provided sensitive information. Deployments should therefore minimize data retention, anonymize or redact sensitive fields, and obtain appropriate user consent; access controls and secure storage should be used for any retained logs.

Bias and fairness. SEAM internalizes patterns from the executor’s rollouts and the training data distribution. As a result, any biases present in the executor or datasets may be propagated or amplified in generated guidance. While our benchmarks are primarily technical, future applications in open-domain settings should include bias assessment and mitigation, such as evaluating guidance quality across user groups and enforcing constraints on unsafe or discriminatory content.

Robustness and over-reliance. Generated guidance may be incorrect, incomplete, or misleading, which could cause downstream errors if users or systems rely on it uncritically. Our design partially mitigates this by keeping the executor frozen and by training SEAM to optimize downstream task success, but this does not guarantee correctness on all inputs. For higher-stakes deployments, additional measures (e.g., verification, calibration, or human oversight) are necessary.

Environmental impact. Training SEAM requires additional computation beyond using a fixed executor alone. However, SEAM is lightweight compared to the executor and can reduce time-to-correct at inference by improving solution success rates. We report efficiency metrics to support transparent accounting of computational costs.

Responsible release. We release our code in the supplementary material and recommend accompanying it with clear documentation on intended use, limitations, and safety considerations, as well as default configurations that disable deployment-time logging unless explicitly enabled by users.

Acknowledgments

This work was supported by the Research Project of Quancheng Laboratory, China (Grant No. QCL20250105), and the Key R&D Program of Shandong Province (Grant No. SYS202201).

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. [Program synthesis with large language models](#). *Preprint*, arXiv:2108.07732.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. 2025. Tool-star: Empowering llm-brained multi-tool reasoner via reinforcement learning. *arXiv preprint arXiv:2505.16410*.

Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z Pan. 2025. Rethinking memory in ai: Taxonomy, operations, topics, and future directions. *arXiv preprint arXiv:2505.00675*.

Anastasia Efklides. 2009. The role of metacognitive experiences in the learning process. *Psicothema*, pages 76–82.

Siqi Fan, Xiusheng Huang, Yiqun Yao, Xuezhi Fang, Kang Liu, Peng Han, Shuo Shang, Aixin Sun, and Yequan Wang. 2025. If an llm were a character, would it know its own story? evaluating lifelong learning in llms. *arXiv preprint arXiv:2503.23514*.

Runnan Fang, Yuan Liang, Xiaobin Wang, Jialong Wu, Shuofei Qiao, Pengjun Xie, Fei Huang, Hua-jun Chen, and Ningyu Zhang. 2025. Memp: Exploring agent procedural memory. *arXiv preprint arXiv:2508.06433*.

Erhu Feng, Wenbo Zhou, Zibin Liu, Le Chen, Yunpeng Dong, Cheng Zhang, Yisheng Zhao, Dong Du, Zhichao Hua, Yubin Xia, and 1 others. 2025. Get experience from practice: Llm agents with record & replay. *arXiv preprint arXiv:2505.17716*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

- David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, and 7 others. 2022. [Competition-level code generation with alphacode](#). *Science*, 378(6624):1092–1097.
- Yitao Liu, Chenglei Si, Karthik Narasimhan, and Shunyu Yao. 2025. Contextual experience replay for self-improvement of language agents. *arXiv preprint arXiv:2506.06698*.
- Fanbin Lu, Zhisheng Zhong, Shu Liu, Chi-Wing Fu, and Jiaya Jia. 2025. Arpo: End-to-end policy optimization for gui agents with experience replay. *arXiv preprint arXiv:2505.16282*.
- Venkatesh Mishra, Amir Saeidi, Satyam Raj, Mutsumi Nakamura, Jayanth Srinivasa, Gaowen Liu, Ali Payani, and Chitta Baral. 2025. How can input reformulation improve tool usage accuracy in a complex dynamic environment? a study on τ -bench. *arXiv preprint arXiv:2508.20931*.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Derrick Quinn, Mohammad Nouri, Neel Patel, John Salihu, Alireza Salemi, Sukhan Lee, Hamed Zamani, and Mohammad Alian. 2025. Accelerating retrieval-augmented generation. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pages 15–32.
- Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. 2025. Meminsight: Autonomous memory augmentation for llm agents. *arXiv preprint arXiv:2503.21760*.
- Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, Qiujie Xie, Xinyu Guo, Qu Yang, Jiayi Wu, Jujia Zhao, Xiaqiang Tang, Xinbei Ma, Cunxiang Wang, Jiaxin Mao, and 7 others. 2025. [Deep research: A systematic survey](#). *Preprint*, arXiv:2512.02038.
- Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. 2025. Dynamic cheat-sheet: Test-time learning with adaptive memory. *arXiv preprint arXiv:2504.07952*.
- Xiaoyu Tan, Bin Li, Xihe Qiu, Chao Qu, Wei Chu, Yinghui Xu, and Yuan Qi. 2025. Meta-agent-workflow: Streamlining tool usage in llms through workflow construction, retrieval, and refinement. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 458–467.
- Robin Usher. 1993. Experiential learning or learning from experience: Does it make a difference. *Using experience for learning*, pages 169–180.
- Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, and 1 others. 2025. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*.
- Ziting Wang, Haitao Yuan, Wei Dong, Gao Cong, and Feifei Li. 2024a. Corag: A cost-constrained retrieval optimization system for retrieval-augmented generation. *arXiv preprint arXiv:2411.00744*.
- Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2024b. Agent workflow memory. *arXiv preprint arXiv:2409.07429*.
- Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.
- Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Hinrich Schütze, Volker Tresp, and Yunpu Ma. 2025. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lu-song Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. 2024. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26275–26285.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Lingfan Yu, Jinkun Lin, and Jinyang Li. 2025. Stateful large language model serving with pensieve. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 144–158.

Shaokun Zhang, Yi Dong, Jieyu Zhang, Jan Kautz, Bryan Catanzaro, Andrew Tao, Qingyun Wu, Zhiding Yu, and Guilin Liu. 2025. Nemotron-research-tool-n1: Tool-using language models with reinforced reasoning. *arXiv preprint arXiv:2505.00024*.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. 2023. Synapse: Trajectory-as-exemplar prompting with memory for computer control. *arXiv preprint arXiv:2306.07863*.

Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, and 1 others. 2025. Memento: Fine-tuning llm agents without fine-tuning llms. *Preprint*.

Appendix

A Dataset Details

Table 6 summarizes the datasets used in our experiments and provides the corresponding source links.

B Implementation Details

B.1 Detailed SEAM Implementations

B.1.1 Hyperparameters

Table 7 and Table 8 summarize the key hyperparameters used for GRPO training and deployment-time periodic SFT. Notably, for training efficiency, we decode the executor with zero temperature ($T = 0$) during training; accordingly, we set the number of executor rollouts per experience entry to $M = 1$.

B.1.2 Prompts

In Figure 4 and Figure 5, we present the input prompts for SEAM and the executor.

B.2 Detailed Baseline Implementations

B.2.1 General protocol and fairness controls

Shared frozen executor. All baselines use the same frozen executor \mathcal{E}_ϕ as the final answer generator (QWEN3-4B or DEEPSEEK-R1-DISTILLED-LLAMA-3.1-8B, matching the main experiments),

so performance differences come from the memory mechanism rather than the executor.

Memory budget and decoding. For all memory-augmented baselines, we cap the total memory text concatenated into the executor prompt to 4k tokens. If a method produces longer memory (e.g., iterative curation), we truncate by keeping higher-priority sections first (e.g., reusable strategies / formulas) and then cutting from the tail. All baseline sub-modules (retrieval, extraction/curation, and final answering) use greedy decoding with temperature = 0.0.

Online evaluation and run averaging.

DYNAMIC-CHEATSHEET, MEM-0, and MEMENTO can update memory online and may be order-sensitive. Thus, for each dataset we sample a fresh random permutation of the evaluation set per run, reset to the same prefilled bank using the same training set of SEAM, process instances sequentially, and report the mean over 10 independent runs.

B.2.2 MEM-0 (graph-based long-term memory)

Adaptation to math reasoning. MEM-0 is originally designed for incremental conversational memory. We adapt it by treating each processed training/evaluation instance as a single interaction unit: (user message = problem statement; assistant message = model solution). The external memory stores structured “facts” that can help future problems, such as reusable lemmas, problem templates, common transformations, and frequent error patterns.

Two-stage pipeline: extraction and update. We follow the standard MEM-0 pipeline consisting of: (i) *extraction*, which extracts a set of salient candidate memories from the latest interaction given (a) a global summary of the history and (b) a recency window of recent interactions; (ii) *update*, which retrieves semantically similar existing memories and decides one of four actions: ADD, UPDATE, DELETE, or NOOP for each candidate memory, to maintain consistency and avoid redundancy.

Hyperparameters. We use a recency window $m = 10$ interactions and retrieve $s = 10$ similar memories during the update phase. All similarity search uses dense embeddings with cosine similarity. After the memory is updated, we retrieve the

Table 6: Datasets used in our experiments.

Domain	Dataset	Link
Training (MATH eval)	DAPO-Math-17k-Processed	open-r1/DAPO-Math-17k-Processed
Math eval	GSM8K	openai/gsm8k
Math eval	MATH500 (subset of MATH)	HuggingFaceH4/MATH-500
Math eval	AIME24	HuggingFaceH4/aime_2024
Math eval	AIME25	math-ai/aime25
Code analysis	CodeContests	dshihk/code-contests
Code analysis	MBPP	google-research-datasets/mbpp
QA analysis	HotpotQA	hotpotqa/hotpot_qa
QA analysis	Natural Questions (NQ)	google-research-datasets/natural_questions

Table 7: Key hyperparameters for GRPO training of SEAM.

Hyperparameter	Value
Train batch size	128
Max prompt / response length for SEAM	2048/4096
Max prompt / response length for Executor	5120/8192
Experience candidates per instance	8
Executor solutions per instance	1
Executor temperature	0.0
Learning rate	1×10^{-6}
KL regularization coef	0.001
Training epochs	10

Table 8: Key hyperparameters for deployment-time periodic SFT of SEAM.

Hyperparameter	Value
Train steps	500
Micro-batch size	8
Learning rate	1×10^{-5}
Max prompt / response length	2048/4096

top- r memories most relevant to the current problem (default $r = 10$), serialize them into text, and prepend them to the executor prompt as additional context (while respecting the 4k-token memory budget).

B.2.3 DYNAMIC-CHEATSHEET (DC-RS)

Method overview. We adopt the DC-RS (retrieve-and-synthesize) variant, which maintains a persistent text cheatsheet (memory) and updates it *before* solving each new problem. At step i , DC-RS performs: (1) retrieve top- k most similar past inputs and their model-generated outputs, (2) update the cheatsheet by a *curator* conditioned on the previous cheatsheet and retrieved pairs, and (3) solve the current problem by a *generator* conditioned on the updated cheatsheet.

Retrieval and prompts. We set $k = 3$ for retrieval. The retriever ranks historical inputs by embedding cosine similarity and returns the top- k

input–output pairs. We follow DC’s prompt separation between a generator (problem solver) and a curator (memory editor): the generator prompt contains a `[[CHEATSHEET]]` placeholder and a `[[QUESTION]]` placeholder; the curator prompt takes the previous cheatsheet and retrieved notes and outputs an updated cheatsheet. We keep the cheatsheet as a structured document (e.g., reusable strategies / heuristics) and enforce the global 4k-token memory budget.

Online update without labels. Consistent with DC, the curator does not use ground-truth answers. After each problem is answered, we append the (problem, model output) pair to the retriever store; the next step may retrieve it to refine the cheatsheet.

B.2.4 MEMENTO (case-based reasoning with a case bank)

Adaptation to single-step math problems. MEMENTO is proposed as a planner–executor agent with a growing case bank. For math benchmarks without tool interaction, we instantiate the core case-based reasoning (CBR) component in a single-step setting: each problem is treated as a state s , the model solution as an action a , and the scalar reward r is computed from answer correctness.

Case bank (Write/Read). We use the *non-parametric* case memory variant. **Write** appends each completed case (s, a, r) to the case bank. **Read** retrieves top- K nearest cases by cosine similarity between a frozen text-encoder representation of the current state and stored past states.

Hyperparameters and prompting. We set $K = 3$ to control context length and match the retrieval scale of other baselines. Retrieved cases are formatted as short exemplars (problem statement + key steps + final answer, optionally with the reward tag), then concatenated to the executor prompt as

reference experiences. We again enforce the 4k-token memory budget for the concatenated exemplars.

Why we use non-parametric CBR. While MEMENTO also supports a parametric retrieval policy (online-updated Q-function), we use the non-parametric variant to avoid introducing an additional trainable module and to keep the comparison focused on retrieval-augmented reuse.

C Case Study

To examine the behavior of our method on real task instances, we selected a geometry problem from Math500 as a case study. The problem is: in triangle ABC , with $AB = 17$, $AC = 8$, and $BC = 15$, let D be the foot of the altitude from C to AB . Find the area of $\triangle ACD$. The correct answer is $\frac{3840}{289}$.

In this case, the analysis section of SEAM provided a concise and correct problem analysis: it first verified that $\triangle ABC$ is right-angled at C by the Pythagorean theorem, and then suggested placing the triangle in a coordinate system (e.g., $C = (0, 0)$, $A = (0, 8)$, $B = (15, 0)$), finding the coordinates of D , and finally computing the area of $\triangle ACD$ using the shoelace formula (or equivalently, base-height/2). The experience section listed multiple possible solution strategies (similar triangles, area decomposition, coordinate geometry, etc.) and explicitly recommended the coordinate method as the most direct, efficient, and reliable approach for this problem. The example section further provided the concrete steps of the coordinate method (deriving the equation of line AB , the perpendicular line through C , their intersection point D , and then applying the area formula), yielding a reusable “action checklist.”

The downstream executor successfully solved the problem and obtained the correct result $\frac{3840}{289}$. Notably, the executor was not additionally fine-tuned, and thus exhibited a degree of autonomy: it initially attempted two alternative approaches (e.g., area ratios/similar triangles and direct algebraic decomposition), but after cross-checking intermediate results, it eventually followed SEAM’s suggestion, switched to the simpler coordinate method, and completed the exact computation. This illustrates two points: (i) SEAM outputs provide a clear strategic prior that guides the executor toward convergence when multiple strategies compete, and (ii) even without specialized training, the executor may

perform exploratory reasoning, but under the structured guidance of SEAM, it ultimately converges to the most efficient and verifiable solution path.

This case study demonstrates that SEAM’s structured experience and task decomposition not only improve efficiency (by reducing unproductive exploration) but also enhance interpretability and reusability of reasoning steps. The executor’s limited autonomous exploration does not hinder final correctness; rather, in combination with SEAM, it forms a cooperative paradigm of “prior guidance + verification-based convergence.” This division of labor allows SEAM to focus on *experiential strategy generation*, while the executor specializes in *step execution and verification*. Without additional fine-tuning, the system can still produce stable and correct solutions on benchmark problems.

Case Study

Original Problem

Given triangle ABC with $AB = 17$, $AC = 8$, $BC = 15$. Let D be the foot of the altitude from C to AB . Find the area of triangle ACD .

SEAM Output

<analysis>

The problem asks for the area of triangle ACD . In triangle ABC , the side lengths are $AB = 17$, $AC = 8$, and $BC = 15$, and D is the foot of the altitude from C to AB .

First, I need to determine the type of triangle ABC . Check the Pythagorean relation: $8^2 + 15^2 = 64 + 225 = 289 = 17^2$, so triangle ABC is right-angled at C .

This lets me place the triangle on a coordinate plane for convenience. Set C at the origin $(0,0)$, A at $(0,8)$, and B at $(15,0)$. Then draw the perpendicular from C to AB and find the coordinates of the foot D .

With the coordinates of A , C , and D in hand, I can compute the area of triangle ACD using the shoelace formula or the “base \times height / 2” formula. This is more direct and efficient than using trigonometry or similar triangles for this problem.

</analysis>

<experience>

- Identify the triangle type: use the Pythagorean theorem to check if it is a right triangle. Given $AB = 17$, $AC = 8$, and $BC = 15$, verify $8^2 + 15^2 = 64 + 225 = 289 = 17^2$; therefore, triangle ABC is right-angled at C .

- Use coordinate geometry: place C at $(0,0)$, A at $(0,8)$, and B at $(15,0)$, then find the coordinates of D , the foot of the altitude

Input prompt for SEAM

You are an Experience Generator that produces structured guidance for a downstream large-language executor.

Here is the information you receive:

```
<problem>
{problem}
</problem>
```

Using this information, create one structured experience entry with three parts:

1. <analysis> - a brief, executor-aware diagnosis of what the problem is asking and what is likely to be tricky (e.g., key constraints, typical failure modes, and the most suitable strategy family). Keep it short.
2. <experience> - a bullet list of distilled tips designed for the downstream executor that will help it solve the current problem <problem> {problem} </problem> and similar tasks in the future. Draw on the executor's typical strengths and weaknesses to give algorithms, heuristics, and pitfalls that will most help it on future attempts.
3. <example> - 3-8 concise, high-level steps you judge to be the most effective solution path for the executor on the problem <problem> {problem} </problem>.

Output Format

```
<analysis>
... brief diagnosis and strategy
</analysis>
```

```
<experience>
... first distilled insight
... second distilled insight
... etc.
</experience>
```

```
<example>
1. Step 1 - ...
2. Step 2 - ...
3. ...
</example>
```

Figure 4: Input prompt for SEAM.

from C to AB .

- Use the triangle area formula: since ABC is a right triangle, its area is $\frac{1}{2} \times AC \times BC = \frac{1}{2} \times 8 \times 15 = 60$.
- Find the area of triangle ACD using coordinates: with A , C , and D , apply the shoelace formula or the "base \times height / 2" formula.
- Similar triangles or trigonometry could also be used, but in this case, coordinate geometry is more efficient.

```
</experience>
```

- ```
<example>
```
1. Step 1 - Use the Pythagorean theorem to determine whether triangle  $ABC$  is right-angled.
  2. Step 2 - Place the triangle on a coordinate plane with  $C$  at  $(0,0)$ ,  $A$  at  $(0,8)$ , and  $B$  at  $(15,0)$ .
  3. Step 3 - Find the coordinates of  $D$ , the foot of the altitude from  $C$  to  $AB$ .
  4. Step 4 - Using the coordinates of  $A$ ,  $C$ , and  $D$ , compute the area of triangle  $ACD$

with the shoelace formula.  
</example>

### Executor Output

```
<think>
```

We are asked to find the area of triangle  $ACD$  inside triangle  $ABC$ , where  $AB = 17$ ,  $AC = 8$ ,  $BC = 15$ , and  $D$  is the foot of the altitude from  $C$  to  $AB$ .

First, the user mentioned that the triangle is a right triangle. Let me confirm that. Using the Pythagorean theorem:  $AC^2 + BC^2$  should equal  $AB^2$ .  $8^2$  is 64,  $15^2$  is 225, adding them gives 289, which is  $17^2$ . So yes, triangle  $ABC$  is right-angled at  $C$ . That simplifies things because the area of  $ABC$  is  $(AC \times BC)/2 = (8 \times 15)/2 = 60$ .

Now, the altitude from  $C$  to  $AB$ . In a right triangle, the altitude from the right angle to the hypotenuse has a known length. The formula for the altitude ( $h$ ) is (product of the legs)/hypotenuse. So

### Input prompt for the executor Qwen3

The above is a Q&A dialogue between a user and a problem-solving guidance model. Treat the output of the guidance model as advisory context to solve the problem: prefer using its techniques when they fit, but you may use alternative correct methods if they are more efficient or clearer. If you diverge from the advisory context, briefly explain why. Be concise and accurate. Present your reasoning and answer in the following format:  
<think> Content of Thinking</think><answer>[Final result only]</answer>

(a) Qwen3 executor.

### Input prompt for the executor Deepseek-R1-distilled

The above is a Q&A dialogue between a user and a problem-solving guidance model. Treat the output of the guidance model as advisory context to solve the math problem: prefer using its techniques when they fit, but you may use alternative correct methods if they are more efficient or clearer. If you diverge from the advisory context, briefly explain why. Be concise and accurate. Please reason step by step, and put your final answer within `\boxed{\}`.

(b) Deepseek-R1-distilled executor.

Figure 5: Input prompts for different executors.

$h = (8 \times 15)/17 = 120/17$ . But wait, that's the length of the altitude from C to AB, which is CD.

But the question asks for the area of triangle ACD. Since D is the foot of the altitude, CD is the height of triangle ACD with base AD. So I need to find AD.

In a right triangle, the length of the segments into which the altitude divides the hypotenuse can be found using similar triangles. The formula for AD is  $(AC^2)/AB$ . So  $AD = (8^2)/17 = 64/17$ .

Then the area of ACD is  $(AD \times CD)/2$ . Plugging in the values:  $(64/17 \times 120/17)/2 = (7680/289)/2 = 3840/289$ . Let me check that calculation again.

Alternatively, since the area of ABC is 60, and D divides AB into AD and DB. The area of ACD would be proportional to  $AD/AB$  times the area of ABC. So  $60 \times (AD/AB) = 60 \times (64/17)/17 = 60 \times 64/(17^2) = 3840/289$ . That's the same result.

So the area of triangle ACD is  $3840/289$ . Let me compute that as a decimal to check if it makes sense. 289 goes into 3840 about 13.3 times. But since the problem likely expects an exact fraction,  $3840/289$  is the answer. Wait, but maybe there's a simpler way.

It is suggested that the coordinate method is simpler, so let me give it a try. Use coordinates: set  $C = (0,0)$ ,  $A = (0,8)$ ,  $B = (15,0)$ .

Line AB passes through  $(0,8)$  and  $(15,0)$ ,

so its slope is  $-\frac{8}{15}$  and its equation is

$$y = -\frac{8}{15}x + 8.$$

The altitude from C is perpendicular to AB, hence has slope  $\frac{15}{8}$  and equation

$$y = \frac{15}{8}x.$$

Intersecting these gives

$$\begin{aligned} \frac{15}{8}x &= -\frac{8}{15}x + 8 \implies 289x = 960, \\ x &= \frac{960}{289}, \quad y = \frac{15}{8}x = \frac{1800}{289}. \end{aligned}$$

Thus

$$D = \left( \frac{960}{289}, \frac{1800}{289} \right).$$

Compute the area of  $\triangle ACD$  using the determinant (shoelace) formula with  $A(0,8)$ ,  $C(0,0)$ ,  $D\left(\frac{960}{289}, \frac{1800}{289}\right)$ :

$$\begin{aligned} [ACD] &= \frac{1}{2} |x_A(y_C - y_D) + x_C(y_D - y_A) + x_D(y_A - y_C)| \\ &= \frac{1}{2} \cdot \frac{960}{289} \cdot 8 = \frac{3840}{289}. \end{aligned}$$

Therefore, the area is  $\boxed{\frac{3840}{289}}$ . </think>

<answer>  $\frac{3840}{289}$  </answer>