From Proprietary to Open-Source: Bridging the Distribution Gap via Multi-Agent Protocol Distillation in Agentic Search

Junlin Liu[1*†], **Jiangwang Chen**[2*], **Zixin Song**[2*], **Shuaiyu Zhou**[3*], **Chunji Lv**[4], **Hank Wu**[5], **Kailin Jiang**[6], **Jinyang Wu**[2], **Bohan Yu**[1] **and Chenxi Zhou**[1]

[1]University of Chinese Academy of Sciences, [2]Tsinghua University, [3]Peking University, [4]Beijing Institute of Technology, [5]East China Normal University, [6]University of Science and Technology of China

[*]Equal contribution. [†]Corresponding author.

Agentic search enables large language models to solve knowledge-intensive tasks by interleaving multi-step reasoning with retrieval, yet optimizing this with outcome-based reinforcement learning (RL) provides only sparse supervision. Knowledge distillation can supply denser guidance, and advanced proprietary models with their strong reasoning capabilities are promising teachers. While distilling from proprietary models can densify this supervisory signal, conventional logit-matching is precluded by hidden logits and mismatched tokenizers, whereas raw natural language trajectory imitation transfers superficial stylistic artifacts rather than core reasoning competence. To address the heterogeneous distillation problem and bridge the distribution gap, we propose **M**ulti-**A**gent **P**rotocol **D**istillation (**MAPD**), a joint distillation and RL framework uses a structured, style-normalized protocol as an intermediate representation. An offline multi-agent system (MAS) decomposes each query, retrieves supporting evidence, repairs failed searches, and converts the resulting exploration trace into a JSON protocol containing the task type, reasoning plan, and extractive grounding facts. During training, the protocol is provided only to a privileged branch of the student policy, whose token distributions furnish a dense distillation signal alongside the sparse RL objective. Extensive evaluations across seven QA benchmarks demonstrate that MAPD consistently outperforms competitive distillation and RL, achieving average success rates of 39.4% on Qwen3-1.7B and 44.4% on Qwen3-4B. Crucially, the framework generalizes robustly across diverse proprietary teachers while effectively mitigating the student policy from style drift and verbosity degeneration. The code is open sourced in https://github.com/AaronLiu0702/MAPD

1. Introduction

Driven by the rapid advancement of large language models (LLMs), research has shifted from single-turn reasoning to multi-turn agentic tasks, with agentic search emerging as a crucial paradigm for tackling knowledge-intensive challenges (Jin et al., 2025; Zhang et al., 2025; Zheng et al., 2025; Luo et al., 2025; Liang et al., 2026). However, existing agentic search systems are often optimized with outcome-based reinforcement learning from verifiable rewards (RLVR), where final-answer correctness provides only sparse supervision for long interaction trajectories. Recent studies supplement RLVR with online policy distillation (OPD), using token-level distribution alignment to provide denser guidance for intermediate decisions (Dong et al., 2025b; Wang et al., 2026b; Lu et al., 2026).

Although combining RL with OPD provides a more reliable training signal, the overall efficacy of OPD strictly hinges on a high-quality teacher policy. To unlock the potential of open-source models in complex agentic search, leveraging state-of-the-art proprietary models as expert teachers is highly desirable, given their exceptional capabilities in dynamic task decomposition, multi-hop reasoning, and information synthesis. However, distilling knowledge from these proprietary teachers to open-source students presents two major bottlenecks. First, traditional OPD fundamentally relies on token-level KL divergence alignment. Yet, the discrepancy of heterogeneous tokenizers and the absence of accessible teacher logits render this mathematical

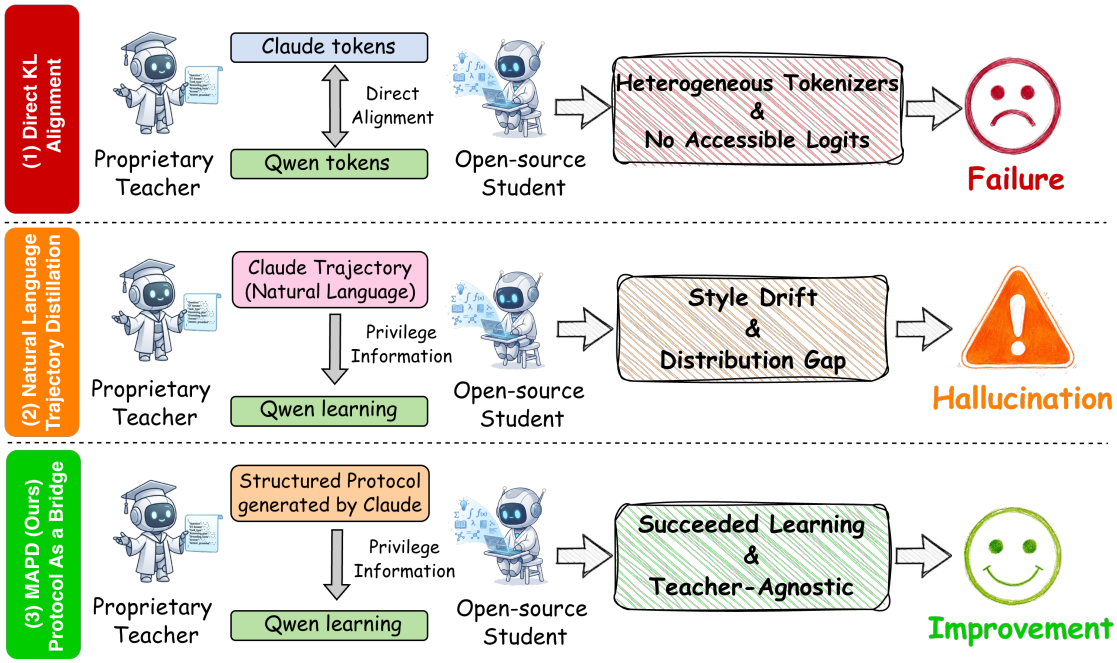

Figure 1 | Distilling from proprietary teacher to open-source student faces two bottlenecks: (1) heterogeneous tokenizers and inaccessible logits preclude direct alignment; (2) free-form natural-language trajectory imitation induces style drift and hallucination.

alignment undefined (Agarwal et al., 2024; Li et al., 2026). Second, as a common fallback, natural language trajectory distillation frequently fails in agentic scenarios. By forcing open-source student models to mimic the verbose and idiosyncratic reasoning styles of proprietary experts, this approach exacerbates the distribution gap. It induces severe style drift and hallucinations, failing to transfer the underlying cognitive strategies while degrading performance in agentic tasks (Gudibande et al., 2023; Zhao et al., 2026). Existing approaches therefore lack an intermediate representation that can convert black-box expert behavior into dense supervision without exposing the student to teacher-specific surface forms.

To address these bottlenecks, we propose **M**ulti-**A**gent **P**rotocol **D**istillation (**MAPD**), a joint distillation and RL framework built around a structured, style-normalized intermediate representation between proprietary experts and open-source students. MAPD separates offline protocol synthesis from online policy alignment. In the offline stage, a proprietary-model-powered MAS decomposes each query, retrieves supporting evidence, repairs failed searches, and converts the exploration trace into a JSON protocol. In the training stage, this protocol is reformatted as privileged information into a conditioned branch of the student policy. Since this privileged branch and the student branch share a same model, their distributions are aligned via a dense self-distillation objective that complements the sparse outcome-based RL objective. By distilling through structured protocols rather than free-form natural-language trajectories, MAPD reduces the student's exposure to teacher-specific verbosity and formatting artifacts. Crucially, the proprietary model and MAS are used only for offline protocol synthesis, adding zero overhead to inference. Our main contributions are summarized as follows:

- We propose MAPD, a joint on-policy self-distillation and outcome-based RL framework designed to bridge the distribution gap between proprietary and open-source. By aligning a protocol-conditioned branch with a student branch of the same policy, it provides dense guidance without requiring access to proprietary logits or cross-tokenizer alignment.
- We introduce a structured, style-normalized JSON protocol as the intermediate representation to decouple core strategies from the linguistic shell. To reliably synthesize protocols, we construct a robust MAS with specialized roles and rigorous automated checks.
- Experiments with three top-tier proprietary models (Claude-Opus-4.6, GPT-5.5, and Gemini-3.1-Pro) as expert teachers demonstrate that MAPD yields reliable gains and these improvements can transfer across different proprietary models without retuning the pipeline.

2. Problem Formulation & Preliminaries

2.1. Agentic Search

We formulate agentic search as a multi-turn interaction paradigm. Given a question x, a language model policy π_θ interacts with a retrieval-augmented environment for up to K turns. At each turn, the agent generates reasoning steps and optionally issues a retrieval query, to which the environment responds with relevant passages as the next observation. For notational simplicity, we flatten the entire generated trajectory across all turns into a single token sequence

$$y = (y_1, \ldots, y_T) \sim \pi_\theta(\cdot \mid x). \tag{1}$$

The interaction terminates either when the agent emits a designated final answer string or reaches the maximum turn limit K. Each completed episode receives a sparse, trajectory-level binary reward $R(x, y) \in \{0, 1\}$, which is determined by parsing the final predicted answer from y and evaluating it via strict exact-match (EM) against the ground-truth.

2.2. Group Relative Policy Optimization (GRPO)

For a given prompt x, GRPO (Shao et al., 2024) samples a group of G rollouts $\{y^{(i)}\}_{i=1}^{G}$ and computes advantages $\hat{A}^{(i)}$ derived from the terminal rewards. The policy is updated by optimizing a clipped surrogate objective $J_t^{(i)}$ alongside a per-token KL penalty against a reference model π_{ref}:

$$\begin{aligned} J_t^{(i)} &= \min\Big(\rho_t^{(i)} \hat{A}^{(i)},\ \text{clip}\big(\rho_t^{(i)}, 1-\epsilon, 1+\epsilon\big)\hat{A}^{(i)}\Big), \\ \mathcal{L}_{\text{GRPO}} &= -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\Big[J_t^{(i)} - \beta\mathbb{D}_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)\Big]. \end{aligned} \tag{2}$$

Where $\rho_t^{(i)}$ denotes the per-token importance weight. While GRPO effectively drives exploration, it strictly evaluates outcome success and lacks the granular supervision required to optimize complex intermediate reasoning steps.

2.3. Online Policy Self-Distillation (OPSD)

To supply denser token-level supervision and follow the context-conditioned self-distillation formulation of OPSD (Zhao et al., 2026), we obtain two next-token distributions from the same current policy parameters. At optimization step k, the student branch conditions on $s_t = (x, y_{<t})$, whereas the teacher branch additionally conditions on $s_t^+ = (x, p, y_{<t})$, where p denotes the privileged information (PI) available exclusively during training. By restricting gradient flow solely to the student path during backpropagation, OPSD minimizes the per-token reverse KL divergence:

$$\mathcal{L}_{\text{OPSD}}^{(t)} = \mathbb{D}_{\text{KL}}\big(\pi_{\theta_k}(\cdot \mid s_t) \,\|\, \pi_{\theta_k}(\cdot \mid s_t^+)\big). \tag{3}$$

This homologous formulation inherently circumvents the vocabulary mismatch issues encountered in proprietary model distillation. However, standard OPSD typically derives the PI p from the student's own correct rollouts, which fundamentally bounds the supervision quality by the student's existing reasoning limits.

3. Methodology

In this section, we first introduce the Structured JSON Protocol (Section 3.1) that normalizes privileged information and decouples cognitive strategies from proprietary model-specific linguistic patterns. We then detail the Multi-Agent System Generation Pipeline (Section 3.2), which automates the distillation of raw teacher trajectories into our compact protocol through a three-stage collaborative process with rigorous quality control. Finally, we present the Joint Training Objective (Section 3.3) that combines a token-level OPSD distillation signal with a sparse GRPO reinforcement learning signal to effectively align the student model.

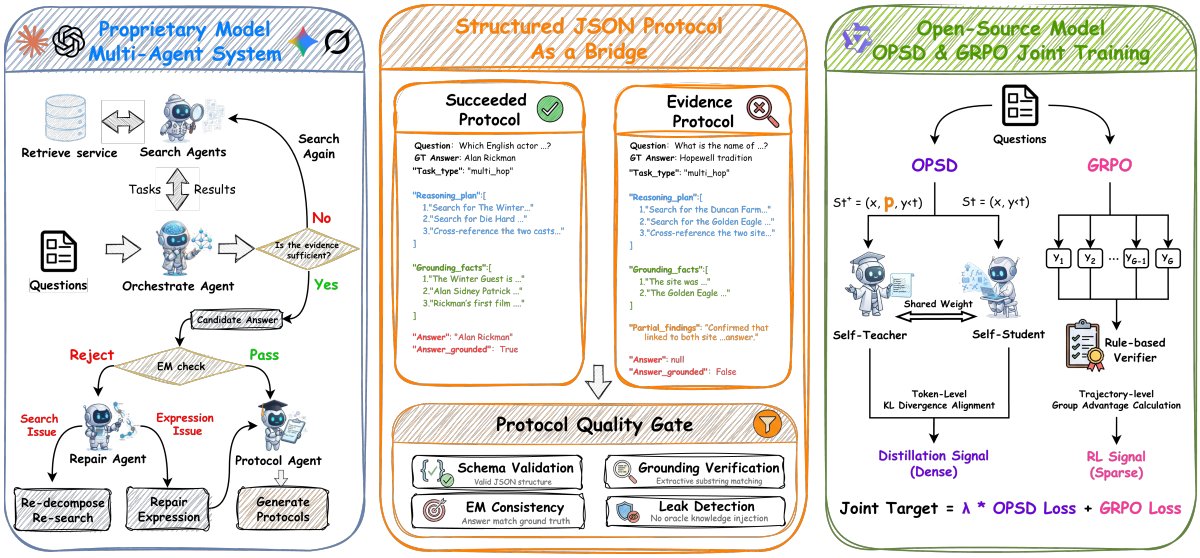

Figure 2 | Overview of the MAPD framework. (1) A proprietary-model-powered multi-agent system generates high-quality structured protocols. (2) The structured JSON protocol serves as a bridge between proprietary teacher and open-source student. (3) The student model internalizes the teacher's capability through joint training with protocols as privileged information.

3.1. Structured JSON Protocol

Directly utilizing free-form natural language trajectories as privileged information for OPSD can introduce a substantial representation mismatch or induce severe style drift and hallucinations, because they contain model-specific verbosity, formatting conventions, and linguistic patterns. To address this, we introduce a style-normalized Structured JSON Protocol as the fundamental unit of distillation, and this design reduces the student's direct exposure to teacher-specific language and provides traceable evidence for the privileged teacher branch. Specifically, rather than extracting raw text, this protocol establishes a purified, traceable format (e.g., Example 1). Formally, the protocol z encapsulates five essential dimensions of task:

- **Task Type:** Determined by the proprietary expert, this field categorizes the query into `single_hop`, `multi_hop`, `comparison`, `others`, thereby dictating the overarching cognitive strategy.
- **Reasoning Plan:** An ordered array of actionable sub-goals that strictly decomposes the complex agentic task into tractable search and inference steps.
- **Grounding Facts:** A curated set of evidence extracted from the environment, which enforces faithful reasoning and actively mitigates parametric hallucinations.
- **Partial Findings:** An optional field recording intermediate discoveries when multiple search rounds fail to yield the correct answer.
- **Answer Verification:** The final extracted answer paired with a boolean flag (`answer_grounded`), ensuring the conclusion is rigorously supported by the retrieved facts.

By formatting the protocol via a descriptive template $f(\cdot)$ such that the PI is defined as $p = f(z)$, we effectively decouple the core cognitive strategies from the proprietary model's surface linguistic patterns. This mechanism not only facilitates the transfer of high-quality reasoning capabilities into OPSD, but also effectively preserves the student's native token distribution.

3.2. Multi-Agent System Generation Pipeline

While the structured JSON protocol mitigates style drift, reliably extracting such knowledge from a proprietary teacher remains challenging. We automate this via a three-stage MAS pipeline that distills the teacher's raw problem-solving into our compact protocol. Crucially, to prevent data leakage, ground-truth answers are strictly confined to offline synthesis and the privileged teacher context, remaining completely invisible to the student.

Example 1: Representative Example of Structured Protocol — Evidence Protocol

Question: *What is the name of the culture cultivated in the area of Duncan Farm and Golden Eagle?*
Ground-Truth Answer: *Hopewell tradition* (*Search did not converge to this answer*)

"task_type": **"multi_hop"**

"reasoning_plan": [
1. "Search for the Duncan Farm archaeological site to determine its geographic location and cultural period."
2. "Search for the Golden Eagle archaeological site to determine the culture it belongs to."
3. "Cross-reference two sites to find the name of the jointly present in the Duncan Farm and Golden Eagle area."]

"grounding_facts": [
1. "The site was part of the Hopewell exchange system in Illinois and is the closest neighboring village site to the Golden Eagle regional transaction center, a major trade and social hub in the system."
2. "The Golden Eagle-Toppmeyer Site is a pre-Columbian archaeological site located near the confluence of the Illinois and Mississippi Rivers in Calhoun County, Illinois. The site is associated with the Havana Hopewell culture."]

"partial_findings": "Confirmed that Duncan Farm belongs to the Woodland period and is part of the Hopewell exchange system, and that the Golden Eagle-Toppmeyer Site is explicitly associated with the Havana Hopewell culture; the culture jointly linked to both sites is the Havana Hopewell culture, but the search did not resolve whether the precise sense of 'cultivated' points to a different answer."

"answer": **null**
"answer_grounded": **false**

Stage A: Multi-Agent Collaborative Search. Given a question x, the **Orchestrator agent** decomposes x into sub-tasks with explicit dependency annotations. Subsequently, these sub-tasks are dispatched in parallel to independent **Searcher agents**. In response, each Searcher issues up to K queries to a local Wikipedia corpus and compresses the retrieved passages into a concise finding. Once all findings within a topological level are collected, the Orchestrator synthesizes them; if the accumulated evidence remains insufficient, it formulates an additional round of sub-tasks (up to M rounds). After the initial exploration pass, an exact match (EM) check evaluates whether the candidate answer matches the ground truth answer. If the check fails, a **Repair agent** leverages the ground-truth answer as diagnostic guidance to trace back the reasoning chain (The GT answer is never propagated into the exploration log and retrieval queries). It diagnoses the failure as either an *expression* issue (correct evidence retrieved but answer poorly formatted) or a *search issue*. It then triggers targeted re-decomposition and additional search operations accordingly, for up to R repair iterations. Ultimately, the output of Stage A comprises an original question, a candidate answer, a binary success label, and a comprehensive exploration log documenting every sub-task, query, retrieved passage, and finding.

Stage B: Protocol Generation. To transform the exhaustive yet unstructured exploration log from Stage A into our structured JSON protocol target, a **Protocolizer agent** is deployed. Conditioned on the success label, the protocolizer compiles this raw information into one of two structured JSON variants. For succeeded trajectories, it generates a *Succeeded Protocol* that includes the fully verified answer. Conversely, if the search failed to converge, it generates an *Evidence Protocol*; here, a definitive answer is omitted and replaced with a neutral summary of partial findings (e.g., "Confirmed X..., but reliable evidence for Y... is missing"). To guarantee generation quality, two integrity constraints are strictly enforced: (i) *No Hindsight*: the reasoning_plan must be formulated step-by-step, strictly prohibiting any mention of the final outcome or subsequent search results to prevent data leakage; and (ii) *Extractive Grounding*: the grounding_facts must be strictly extractive, appearing as verbatim substrings within the retrieved passages.

Stage C: Quality Gate. Before a generated protocol is admitted into the final training phase, it must pass a rigorous quality gate comprising four automated checks: (i) *Schema Validation*, ensuring a well-formed and valid JSON structure; (ii) *EM Consistency* (for succeeded protocol), confirming the extracted answer strictly matches the ground truth; (iii) *Grounding Verification*, enforcing extractive substring matching for all provided facts; and (iv) *Leak Detection*, guaranteeing no oracle knowledge is injected into the reasoning steps. Protocols failing any of these criteria are discarded, and the corresponding training falls back to a self-rollout baseline. Finally, manual review confirms the robustness of this pipeline, demonstrating a 99.33% accuracy in yielding high-quality, strictly compliant structured JSON protocols (total $3,000$ instances across three representative models).

Model	Method	Single-hop QA			Multi-hop QA				
		NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Qwen3-1.7B	Vanilla	29.9±0.3	46.4±0.2	39.1±0.1	23.9±0.2	18.8±0.1	4.9±0.3	16.8±1.1	25.7±0.1
	OPSD	4.2±0.1	8.3±0.1	4.6±0.1	6.6±0.1	15.3±0.1	0.7±0.1	1.6±0.5	5.9±0.1
	GRPO	36.9±0.4	54.6±0.3	43.1±0.1	29.8±0.3	27.5±0.2	6.8±0.4	22.4±1.6	31.6±0.2
	GRPO+OPSD	37.0±0.3	54.6±0.2	41.4±0.1	29.9±0.3	23.3±0.2	6.5±0.3	20.8±1.6	30.5±0.2
	SDAR	43.4±0.4	58.1±0.3	47.6±0.2	37.0±0.3	34.3±0.3	10.6±0.5	32.0±1.6	37.6±0.3
	MAPD(Ours)	**45.1**±0.2	**58.6**±0.2	**48.6**±0.1	**39.0**±0.2	**36.2**±0.2	**11.2**±0.3	**36.8**±1.1	**39.4**±0.1
Improvement (%)		3.9%↑	0.9%↑	2.1%↑	5.4%↑	5.5%↑	5.7%↑	15.0%↑	4.8%↑
Qwen3-4B	Vanilla	29.3±0.3	48.0±0.2	34.9±0.1	26.5±0.2	30.9±0.2	6.1±0.3	25.6±1.6	28.8±0.2
	OPSD	16.2±0.2	36.7±0.1	23.1±0.1	20.5±0.2	24.6±0.1	5.7±0.2	19.2±1.1	20.9±0.1
	GRPO	40.4±0.3	61.0±0.2	43.5±0.1	36.7±0.3	37.3±0.2	10.3±0.4	32.8±1.6	37.4±0.2
	GRPO+OPSD	36.3±0.3	57.9±0.2	43.1±0.1	37.7±0.3	36.7±0.2	12.7±0.3	44.0±1.6	38.3±0.2
	SDAR	46.1±0.4	62.4±0.2	48.5±0.2	43.3±0.3	43.1±0.3	13.1±0.5	44.8±1.6	43.0±0.3
	MAPD(Ours)	**47.7**±0.2	**62.5**±0.2	**49.4**±0.1	**44.3**±0.2	**45.7**±0.2	**14.0**±0.3	**47.2**±1.1	**44.4**±0.1
Improvement (%)		3.5%↑	0.2%↑	1.9%↑	2.3%↑	6.0%↑	6.9%↑	5.4%↑	3.3%↑

Table 1 | Results are reported as success rate (%) (mean ± std across five independent evaluation seeds from a single training run). **Avg.** denotes the macro-averaged success rate. Improvement (%) is the relative gain of MAPD over the SDAR baseline. **Bold**: best performance; underline: second best.

3.3. Joint Training Objective

The overall training objective integrates a sparse, outcome-driven reinforcement learning signal with a dense distillation signal to achieve high-quality alignment:

$$\mathcal{L}(\theta) = \lambda_{OPSD} \cdot \mathcal{L}_{\text{OPSD}}(\theta) + \mathcal{L}_{\text{GRPO}}(\theta), \tag{4}$$

where $\mathcal{L}_{\text{OPSD}}$ provides a token-level distillation signal derived from the structured protocol, which facilitates efficient step-by-step credit assignment. Conversely, $\mathcal{L}_{\text{GRPO}}$ provides a sparse RL signal via clipped policy gradients, optimizing the model toward the final environment reward. The hyperparameter λ_{OPSD} dictates the relative strength of this distillation guidance (extensively ablated in Section 4.2).

4. Experiments and Analysis

4.1. Experimental Setup

Data Splits and Benchmarks. Following Search-R1 (Jin et al., 2025), our training uses only the NQ and HotpotQA train splits. Evaluation spans seven knowledge-intensive QA benchmarks, encompassing single-hop (NQ, TriviaQA, and PopQA) and multi-hop (HotpotQA, 2WikiMultihopQA, MuSiQue, and Bamboogle) reasoning tasks (Kwiatkowski et al., 2019; Min et al., 2019; Mallen et al., 2023; Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; Press et al., 2023). Five of the seven evaluation sets remain entirely absent from training, serving as out-of-distribution tests; for the two shared benchmarks, we ensure no question-level overlap between train and test splits. Performance is reported as success rate (%) via exact match against ground-truth answers.

Baselines. To rigorously assess the effectiveness of our approach, we compare against a spectrum of baselines, ranging from pure alignment strategies to advanced hybrid methods: (1) **Vanilla**: The pretrained base model evaluated zero-shot without any further alignment; (2) **OPSD** (Zhao et al., 2026): Pure online policy self-distillation, which utilizes the student's own successful rollouts as privileged information; (3) **GRPO** (Shao et al., 2024): Pure outcome-based reinforcement learning, relying solely on sparse environment rewards without any dense distillation guidance; (4) **GRPO+OPSD**: A naive hybrid approach combining RL with online policy self-distillation, using successful rollouts as privileged information; (5) **SDAR** (Lu et al., 2026): An advanced hybrid baseline combining RL with token-level gated distillation, utilizing skill-generated content as privileged information; (6) **MAPD (Ours)**: Our proposed Multi-Agent Protocol Distillation framework, integrating RL with dense, token-level distillation guided by the structured MAS protocol without gating mechanisms.

Method	PM	SP	MAS	Avg.	
				Qwen3-1.7B	Qwen3-4B
GRPO+OPSD	✗	✗	✗	30.5	38.3
SDAR	✗	✗	✗	37.6	43.0
Raw trajectory (single PM)	✓	✗	✗	30.1	37.3
Raw trajectory (MAS)	✓	✗	✓	29.2	36.1
Protocol (single PM)	✓	✓	✗	37.1	42.9
MAPD (Ours)	✓	✓	✓	**39.4**	**44.4**

Table 2 | Component ablation study of MAPD. PM = Proprietary Model; SP = Structured Protocol; MAS = Multi-Agent System.

Implementation Details. For the student policy, we train two open-source models, Qwen3-1.7B and Qwen3-4B (Yang et al., 2025), both initialized from their respective pretrained base checkpoints. During joint training, we utilize a group size of $n=8$ rollouts per prompt, a batch size of 128, a maximum prompt length of 4096, and a maximum response length of 512 tokens. The learning rate is set to 1×10^{-6} with a 10% linear warmup. Training runs for 200 steps across 8 GPUs. The agentic environment allows up to 4 interaction turns, utilizing a local Wikipedia corpus (wiki-18) as the retrieval backend, returning the top-3 passages per query.

MAS Pipeline Configuration. For protocol synthesis, the MAS agents (Orchestrator, Searcher, Repair, and Protocolizer) utilize Claude-Opus-4.6 as the default backbone LLM. We also explore GPT-5.5 and Gemini-3.1-Pro as alternative teacher sources (ablated in Section 4.2). The MAS parameters are constrained as follows: the Orchestrator decomposes each query into a maximum of 4 sub-questions over at most 2 rounds; each Searcher issues up to 3 retrieval queries; and the Repair agent is permitted up to 2 diagnostic rounds. To ensure training efficiency, all protocols are pre-synthesized offline and cached, entirely decoupling the teacher's generation latency from the student's RL loop.

4.2. Results and Analysis

Main Result. As shown in Table 1, our proposed MAPD framework consistently outperforms all baseline methods across the seven benchmarks, achieving average success rates of 39.4% on Qwen3-1.7B and 44.4% on Qwen3-4B. Compared to the strongest baseline SDAR, MAPD yields relative average improvements of 4.8% and 3.3%, respectively. Crucially, these performance gains are more pronounced on complex multi-hop QA tasks. For Qwen3-1.7B, the average relative gain on multi-hop benchmarks reaches 7.9%, substantially exceeding the 2.3% on single-hop benchmarks. A similar pattern holds for Qwen3-4B (5.1% vs. 1.8%). These larger relative gains on multi-hop benchmarks are consistent with the intended role of multi-agent decomposition and evidence aggregation, as the MAS pipeline in MAPD systematically breaks down complex queries, synthesizes evidence across multiple retrieval steps, and distills these strategies into the student policy via our structured protocol, thereby raising its upper bound for complex reasoning. Furthermore, we observe a catastrophic performance collapse when applying pure OPSD to the models under our agentic training configuration (dropping to 5.9% on Qwen3-1.7B and 20.9% on Qwen3-4B). This indicates that dense self-distillation is insufficient for agentic search and motivates the use of external PI. Lacking external information gain, self-rollouts merely drive the model toward pathologically long outputs——as evidenced by the length clip ratio surging from 5% to approximately 74% during training, a collapse that underscores the necessity of our MAS protocol.

Ablation Study. We systematically ablate the Proprietary Model (PM), Structured Protocol (SP), and Multi-Agent System (MAS) to assess the contribution of each component, results are reported in Table 2.

(1) Structured protocol is essential for cross-model distillation. The most striking finding is that directly distilling from a proprietary model's raw natural-language trajectories (*w/o SP&MAS*) performs worse than baselines using no proprietary model at all (*GRPO+OPSD*). By comparing *w/o SP&MAS* with *w/o MAS*, we isolate the impact of the representation format under a single-teacher setting: replacing raw text with our structured protocol drives an observed gain of 7.0 and 5.6 points, respectively. This is consistent with the

Proprietary Teacher	Model	λ	Single-hop QA			Multi-hop QA				
			NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Claude-Opus-4.6	Qwen3-1.7B	0.01	44.1	58.0	47.5	37.8	33.5	10.9	28.0	37.1
		0.05	**45.1**	**58.6**	48.6	**39.0**	36.2	**11.2**	**36.8**	**39.4**
		0.10	43.7	58.2	**49.1**	38.2	**36.6**	10.2	36.0	38.9
	Qwen3-4B	0.01	46.3	63.0	48.3	42.5	37.0	13.9	43.2	42.0
		0.05	**47.7**	62.5	49.4	**44.3**	**45.7**	**14.0**	**47.2**	**44.4**
		0.10	47.4	**63.7**	**50.1**	42.3	38.4	11.1	42.4	42.2
GPT-5.5	Qwen3-1.7B	0.01	**43.5**	58.6	**49.1**	38.4	35.5	**11.2**	**36.0**	38.9
		0.05	**43.5**	**59.7**	48.6	**38.7**	**38.0**	9.3	35.2	**39.0**
		0.10	43.3	58.4	48.6	36.4	31.9	9.2	32.8	37.2
	Qwen3-4B	0.01	47.4	62.2	48.7	43.5	**46.0**	14.3	46.4	44.1
		0.05	**48.3**	62.9	**50.6**	**44.2**	44.6	**14.6**	**47.2**	**44.6**
		0.10	47.4	**63.7**	50.3	43.4	43.1	12.6	44.0	43.5
Gemini-3.1-Pro	Qwen3-1.7B	0.01	42.8	58.2	**48.9**	37.4	32.4	9.9	33.6	37.6
		0.05	**43.2**	**58.9**	47.5	**37.7**	**36.3**	9.1	32.8	**37.9**
		0.10	42.5	58.3	47.9	**37.7**	32.6	**10.5**	**35.2**	37.8
	Qwen3-4B	0.01	48.8	**63.7**	47.9	**44.8**	**45.2**	13.0	43.2	43.8
		0.05	**48.9**	62.8	**50.8**	43.8	39.1	**14.0**	**48.8**	**44.0**
		0.10	47.4	62.5	49.3	43.0	42.3	12.8	43.2	42.9

Table 3 | Effect of three proprietary models as privileged information source and OPSD loss weight λ. Results are reported as Success Rate (%). **Bold**: best performance.

style drift and distribution gap bottlenecks discussed in Section 1: without structured decoupling, the student imitates the teacher's surface patterns (e.g., verbose reasoning chains, idiosyncratic phrasing), which conflicts with its own learned distribution and actively degrades performance.

(2) MAS cannot replace structured protocol. Adding the MAS pipeline without the structured protocol (*w/o SP*: 29.2%/36.1%) does not improve performance and is slightly worse, because the MAS produces richer but also noisier natural-language trajectories that exacerbate the style-transfer problem. Conversely, applying the structured protocol generated by single proprietary agent (*w/o MAS*: 37.1%/42.9%) recovers a relative gain, showing that the JSON protocol alone already alleviates style contamination. However, this single-agent protocol still underperforms SDAR, indicating that a single proprietary model call cannot match the retrieval depth and error recovery of the full MAS pipeline.

(3) MAS pipeline improves the quality of protocol generation. The full MAPD configuration performs best, suggesting that the structured representation and the MAS pipeline provide complementary benefits in the evaluated setting. Specifically, SP ensures that the distilled signal largely mitigates style drift and remains learnable, while MAS ensures that the signal is factually accurate, reasoning-complete and traceable. Compared with the single-agent protocol without MAS, this full protocol delivers a significantly higher-quality training signal that bridges the gap between proprietary and open-source models. Moreover, our results demonstrate that the primary bottleneck in distillation lies in PI quality rather than training mechanisms: with sufficiently clean, style-decoupled protocols, a simple token-level KL objective achieves performance comparable to gated approaches (SDAR).

Effect of the OPSD Loss Weight. Table 3 and Figure 3 summarizes the sensitivity of our framework to the OPSD loss weight $\lambda \in \{0.01, 0.05, 0.1\}$ in the joint objective. Across both model scales, $\lambda = 0.05$ consistently achieves the best among the tested values (39.4% on 1.7B and 44.4% on 4B), revealing a sweet-spot associated with two observed patterns at the extremes. Below we analyze these in detail.

(1) Weak distillation signal ($\lambda = 0.01$). A small OPSD weight provides only a weak distillation signal relative to the dominant GRPO policy-gradient objective. Training logs on Claude-Opus-4.6 reveal that the teacher-student gap collapses prematurely by step 100, indicating that the student quickly exhausts the available supervision. However, the KL divergence from the reference policy remains low (0.24 on 1.7B), suggesting that the model does not simply revert to the reference behavior. Despite the early disappearance of the distillation signal, this configuration still achieves a clear improvement over GRPO alone (37.1% vs. 31.6% on 1.7B), implying that even a short-lived OPSD term can provide beneficial guidance during the initial phase of training.

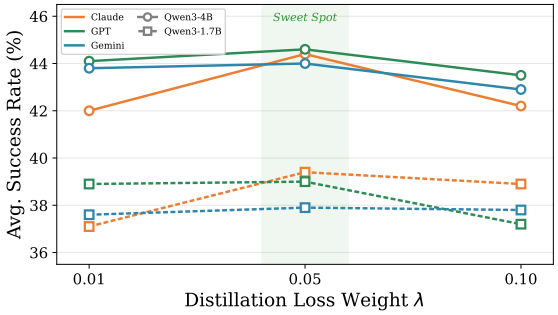

Figure 3 | The relationship between the OPSD loss weight λ and Avg success rate (%). $\lambda = 0.05$ balances distillation strength and policy stability, achieving the best Avg on both model scales.

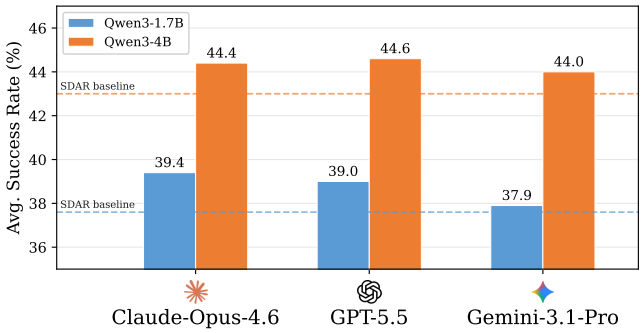

Figure 4 | Performance with three top-tier proprietary teacher backbones. All configurations use the same MAPD architecture and training settings without retuning the MAS pipeline.

(2) Excessive distillation pressure ($\lambda = 0.1$). A large OPSD weight pushes the student aggressively toward the teacher's distribution, which inflates the KL divergence from the reference policy (1.06 on 4B, nearly 4× that of $\lambda = 0.01$). More importantly, we observe a behavioral degeneration specific to agentic search: on 4B the mean response length collapses from 135 tokens to roughly 42 tokens while the number of tool calls saturates at the maximum of 3.0 per episode. This pattern suggests that excessive distillation pressure drives the model toward a "retrieve-without-reasoning" shortcut–issuing retrieval calls with minimal intermediate reasoning. As a result, single-hop accuracy slightly improves (e.g., PopQA rises from 49.4% to 50.1%, TriviaQA from 62.5% to 63.7% on 4B), but performance on multi-hop tasks drops sharply because coherent chain-of-thought across hops is essential (2WikiMultihopQA falls from 45.7% to 38.4%, MuSiQue from 14.0% to 11.1%). This trade-off reveals that over-regularization erodes the model's capacity for multi-step reasoning, ultimately harming the very capability that agentic search relies on.

Taken together, these results highlight a fundamental tension in agentic distillation: too little distillation provides insufficient guidance, while too much forces the student into shallow, retrieval-heavy behaviors. $\lambda = 0.05$ strikes a practical balance, maintaining a sustained distillation signal throughout training without destabilizing the policy, thereby fostering a robust balance between efficient retrieval and rigorous reasoning.

Transfer Across Proprietary Teacher Backbones. A key advantage of MAPD is transferability across different proprietary models. Because the structured JSON protocol successfully decouples the teacher's core reasoning strategies from its surface stylistic artifacts, any sufficiently capable model can serve as the teacher backbone. To validate this flexibility, we evaluate whether the framework can seamlessly utilize protocols generated by three top-tier commercial models without requiring any architectural modifications or teacher-specific hyperparameter tuning—Claude-Opus-4.6, GPT-5.5, and Gemini-3.1-Pro. Results are shown in Table 3 and Figure 4. All three configurations produce positive observed gains over SDAR on both student scales: Claude-Opus-4.6 (39.4%/44.4%), GPT-5.5 (39.0%/44.6%), and Gemini-3.1-Pro (37.9%/44.0%). This narrow variance across teachers (within 2 points) confirms that our protocol successfully abstracts away teacher-specific idiosyncrasies, isolating semantic reasoning from the source model. Consequently, practitioners can seamlessly swap APIs to optimize cost-latency tradeoffs without retuning the pipeline.

Cost Analysis. To quantify the overhead of protocol synthesis, we report the generation cost using Gemini-3.1-Pro as an illustrative example. The MAS pipeline processes 25,600 training instances, producing 25,584 valid protocols that pass the quality gate (99.94% yield). On average, the pipeline invokes the teacher LLM ~6.3 times per instance, consuming approximately 12.5K tokens per instance. The amortized cost is $0.057 per instance, yielding a total one-time generation cost of approximately $1,454 for the entire training set. Since all protocols are pre-synthesized and cached, this cost is incurred only once.

5. Related Work

LLM Reasoning and Agentic Search. The paradigm of LLM reasoning has recently shifted from static, single-turn tasks toward dynamic, multi-turn agentic tasks that require sustained interaction with external environments (Liu et al., 2026a,b; Dong et al., 2025a; Mai et al., 2025; Lv et al., 2026; Jiang et al., 2026c, 2025b, 2026b). As a critical extension of this reasoning paradigm, agentic search interleaves multi-step logical deduction with environment-augmented retrieval to tackle knowledge-intensive tasks. Recent frameworks have formalized this as an RL optimization problem. For instance, WebRL (Qi et al., 2025b) employs self-evolving RL with process-level rewards, Search-R1 (Jin et al., 2025) leverages classic RL algorithms for multi-turn retrieval, and RAGEN (Wang et al., 2025) optimizes retrieval-augmented reasoning chains via trajectory-based rewards. However, these policies remain predominantly driven by sparse, outcome-only signals (e.g., terminal answer accuracy), which fundamentally struggle to provide granular credit assignment for intermediate planning, searching, and diagnostic steps. To resolve this optimization bottleneck, our MAPD framework introduces a joint training paradigm that integrates outcome-driven RL signal with a dense distillation signal.

Knowledge Distillation from Proprietary Models. Transferring capabilities from proprietary LLMs to smaller open-source models has become a dominant paradigm. Classical knowledge distillation (Hinton et al., 2015) relies on matching output probability distributions, a mechanism fundamentally precluded by the opaque nature of black-box APIs and the mismatch between heterogeneous tokenizers. Consequently, contemporary approaches resort to text-level imitation, distilling the natural language reasoning traces or Chain-of-Thought (CoT) generated by proprietary teachers (Hsieh et al., 2023; Mukherjee et al., 2023; Xu et al., 2025). However, as highlighted by recent studies (Gudibande et al., 2023), directly mimicking unconstrained teacher trajectories forces the student to overfit to the teacher's verbose linguistic patterns and authoritative tone. This capacity mismatch induces severe style drift and exacerbates hallucinations. Our MAPD framework extracts the teacher's semantic strategies into a style-normalized structured JSON protocol to decouple the core reasoning logic from the proprietary linguistic shell, enabling safe online self-distillation.

Multi-Agent System for Complex Reasoning. LLM-based MAS have achieved remarkable success in complex reasoning and open-ended problem-solving (Wu et al., 2023; Hong et al., 2024; Lei et al., 2024). In retrieval-augmented scenarios, multi-agent collaboration can significantly enhance recall on multi-step questions through parallel sub-query orchestration and iterative debate (Du et al., 2023; Talebirad & Nadiri, 2023; Wang et al., 2026a). However, deploying these collaborative swarms at inference time incurs prohibitive computational overhead, severe latency, and massive API costs, rendering them highly impractical for real-time applications. Instead of relying on MAS during test-time interactions, MAPD utilizes the agent swarm exclusively as an offline training-time knowledge synthesizer, distilling the swarm's collaborative problem-solving strategies into a single student model via our structured protocol.

6. Conclusion

In this paper, we introduce Multi-Agent Protocol Distillation (MAPD), a novel joint distillation and RL framework designed to bridge the heterogeneous distribution gap between proprietary teachers and open-source students in agentic search. To overcome the severe style drift and verbosity collapse inherent in raw trajectory imitation, MAPD elegantly decouples core cognitive strategies from idiosyncratic linguistic shells via a style-normalized, structured JSON protocol. Synthesized exclusively offline by a robust multi-agent system, this protocol serves as privileged information to provide dense guidance that seamlessly complements sparse outcome-driven RL without incurring any inference overhead. Extensive evaluations across seven knowledge-intensive QA benchmarks demonstrate that MAPD not only consistently outperforms all evaluated baselines, but also generalizes seamlessly across different proprietary teachers without retuning. By isolating semantic reasoning from idiosyncratic stylistic artifacts, MAPD unlocks a safer and more efficient pathway for heterogenes models alignment. We expect that the structural decoupling pioneered by MAPD will serve as a catalyst for future research, inspiring the community to push beyond the reliance on proprietary models and toward the development of fully autonomous, self-improving agentic frameworks.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, et al. Agentic entropy-balanced policy optimization. *arXiv preprint arXiv:2510.14545*, 2025a.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*, 2025b.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Baochen Fu, Wenzhi Deng, Baihao Jin, Yang Li, Zihan Nie, Kailin Jiang, Yuntao Du, and Weiye Song. Can multimodal large language models understand oct? 2026a.

Baochen Fu, Yuntao Du, Cheng-Wei Chang, Baihao Jin, Wenzhi Deng, Muhao Xu, Hongmei Yan, Weiye Song, and Yi Wan. Mmku-bench: A multimodal update benchmark for diverse visual knowledge. *ArXiv*, abs/2603.15117, 2026b.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Yifan Jia, Yuntao Du, Kailin Jiang, Yuyang Liang, Qihan Ren, Yi Xin, Rui Yang, Fenze Feng, MingCai Chen, Hengyang Lu, et al. Benchmarking multimodal knowledge conflict for large multimodal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22283–22291, 2026.

Kailin Jiang, Zhi Gao, Chenrui Shi, Zilong Zheng, Siyuan Qi, Qing Li, et al. Mmke-bench: A multimodal editing benchmark for diverse visual knowledge. In *International Conference on Learning Representations*, volume 2025, pp. 526–555, 2025a.

Kailin Jiang, Hongbo Jiang, Ning Jiang, Zhi Gao, Jinhe Bi, Yuchen Ren, Bin Li, Yuntao Du, Lei Liu, and Qing Li. Kore: Enhancing knowledge injection for large multimodal models via knowledge-oriented augmentations and constraints. *arXiv preprint arXiv:2510.19316*, 2025b.

Kailin Jiang, Yuntao Du, Yukai Ding, Yuchen Ren, Ning Jiang, Zhi Gao, Zilong Zheng, Lei Liu, Bin Li, and Qing Li. When large multimodal models confront evolving knowledge: Challenges and explorations. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Kailin Jiang, Ning Jiang, Yuntao Du, Yuchen Ren, Yuchen Li, Yifan Gao, Jinhe Bi, Yunpu Ma, Bin Li, Lei Liu, et al. Mined: Probing and updating with multimodal time-sensitive knowledge for large multimodal models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13766–13795, 2026b.

Kailin Jiang, Lei Liu, Jian Xi, Hui Xu, Junlin Liu, Baochen Fu, Shaoqing Ren, Bin Li, Yu Lu, Haibo Shi, et al. Beyond relevance-centric retrieval: Rubric-oriented document set selection and ranking. *arXiv preprint arXiv:2607.19747*, 2026c.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. Macm: Utilizing a multi-agent system for condition mining in solving complex mathematical problems. *Advances in Neural Information Processing Systems*, 37: 53418–53437, 2024.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Zihan Liang, Yufei Ma, Ben Chen, Zhipeng Qian, Xuxin Zhang, Huangyu Dai, and Lingtao Mao. Search-e1: Self-distillation drives self-evolution in search-augmented reasoning. *arXiv preprint arXiv:2605.22511*, 2026.

Junlin Liu, Shengnan An, Shuang Zhou, Dan Ma, Yehao Lin, Xinxuan Lv, Xuanlin Wang, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, et al. Amo-bench: Large language models still struggle in high school math competitions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2120–2137, 2026a.

Junlin Liu, Shengnan An, Shuang Zhou, Dan Ma, Shixiong Luo, Ying Xie, Yuan Zhang, Wenling Yuan, Yifan Zhou, Xiaoyu Li, et al. General365: Benchmarking general reasoning in large language models across diverse and challenging tasks. *arXiv preprint arXiv:2604.11778*, 2026b.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, et al. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 2026.

Jinchang Luo, Mingquan Cheng, Fan Wan, Ni Li, Xiaoling Xia, Shuangshuang Tian, Tingcheng Bian, Haiwei Wang, Haohuan Fu, and Yan Tao. Globalrag: Enhancing global reasoning in multi-hop question answering via reinforcement learning. *arXiv preprint arXiv:2510.20548*, 2025.

Chunji Lv, Jiaxi Ye, Yuchen Jiang, Rexar Lin, and Changsheng Li. Physagent: Automating physics-based 4d synthesis via trajectory-grounded multi-agent feedback. *arXiv preprint arXiv:2606.08688*, 2026.

Xinji Mai, Haotian Xu, Zhong-Zhi Li, Weinong Wang, Jian Hu, Yingying Zhang, Wenqiang Zhang, et al. Agent rl scaling law: Agent rl with spontaneous code execution for mathematical problem solving. *arXiv preprint arXiv:2505.07773*, 2025.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. A discrete hard em approach for weakly supervised question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2851–2864, 2019.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Siyuan Qi, Bangcheng Yang, Kailin Jiang, Xiaobo Wang, Jiaqi Li, Yifan Zhong, Yaodong Yang, and Zilong Zheng. In-context editing: Learning knowledge from self-induced distributions. In *International Conference on Learning Representations*, volume 2025, pp. 77563–77585, 2025a.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, et al. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 79791–79821, 2025b.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. *URL https://arxiv. org/abs/2402.03300*, 2(3):5, 2024.

Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Jianze Wang, Ying Liu, Jinlong Chen, Xuchun Hu, Qilong Zhang, Yu Cao, Jun Wang, Hua Yang, Yong Xie, and Qianglong Chen. Mad-opd: Breaking the ceiling in on-policy distillation via multi-agent debate. *arXiv preprint arXiv:2605.01347*, 2026a.

Yinjie Wang, Xuyang Chen, Xiaolong Jin, Mengdi Wang, and Ling Yang. Openclaw-rl: Train any agent simply by talking. *arXiv preprint arXiv:2603.10165*, 2026b.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. In *International Conference on Learning Representations*, volume 2025, pp. 76346–76382, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, et al. From web search towards agentic deep research: Incentivizing search with reasoning agents. *arXiv preprint arXiv:2506.18959*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025.

A. Structured Protocol Example

Example 2, Example 3 and Example 4 show three complete structured protocols produced by our MAS pipeline with different teacher backbones (Claude-Opus-4.6, GPT-5.5 and Gemini-3.1-Pro, respectively).

Example 2: Structured Protocol Example of Claude-Opus-4.6

Question: Which English actor and director who was involved with the 1997 film *The Winter Guest* also appeared in "Die Hard"?
Ground-Truth Answer: Alan Rickman

"task_type": "multi_hop"

"reasoning_plan": [
1. "Search for *The Winter Guest* 1997 film to identify its director and cast."
2. "Search for *Die Hard* 1988 film to identify its cast."
3. "Cross-reference the two casts to find the English actor and director common to both."
]

"grounding_facts": [
1. "The Winter Guest is a 1997 film directed by Alan Rickman in his debut as a director."
2. "Alan Sidney Patrick Rickman (21 February 1946 – 14 January 2016) was an English actor and director."
3. "Rickman's first film role was as the German terrorist leader Hans Gruber in 'Die Hard' (1988)."
]

"answer": "Alan Rickman"
"answer_grounded": true

Example 3: Structured Protocol Example of GPT-5.5

Question: What year was the barber surgeon who served the last French monarch of the House of Valois born?
Ground-Truth Answer: 1510

"task_type": "multi_hop"

"reasoning_plan": [
1. "Search for the last French monarch of the House of Valois to identify the king's name."
2. "Search for that king's name + 'barber surgeon' to find the person who served him."
3. "Look up the barber surgeon's Wikipedia entry to verify the birth year."
]

"grounding_facts": [
1. "In 1589, at the death of Henry III of France, the House of Valois became extinct in the male line."
2. "Ambroise Paré (c. 1510 – 20 December 1590) was a French barber surgeon who served in that role for kings Henry II, Francis II, Charles IX and Henry III."
3. "Paré was born in 1510 in Bourg-Hersent in northwestern France."
]

"answer": "1510"
"answer_grounded": true

B. Training Dynamics

To provide a comprehensive overview of the optimization process, Figures 5 and 6 illustrate the detailed training dynamics of MAPD. These visualizations track the learning trajectories across different model scales, while systematically highlighting how varying the distillation weight, λ_{OPSD}, influences the overall convergence and performance evolution(Jiang et al., 2026a; Jia et al., 2026; Jiang et al., 2025a; Qi et al., 2025a; Fu et al., 2026a,b).

Example 4: Structured Protocol Example of Gemini-3.1-Pro

Question: *The band famous for the single Waterfalls, had a singer with the nickname Left Eye, who was born on what day?*
Ground-Truth Answer: May 27, 1971

```
"task_type": "multi_hop"
"reasoning_plan": [
    1. "Search for the band famous for the single Waterfalls to identify the musical group"
    2. "Search for the members of the identified band to find the singer who goes by the nickname Left Eye"
    3. "Search for the biographical details of the identified singer to find their exact date of birth"
]
"grounding_facts": [
    1. "'Waterfalls' is a song by American recording group TLC."
    2. "'Lisa Nicole Lopes (May 27, 1971 – April 25, 2002), better known by her stage name Left Eye, was an American hip hop singer, rapper, songwriter, and producer.'
    3. "'Lopes was best known as one-third of the R&B girl group TLC'
]
"answer": "May 27, 1971"
"answer_grounded": true
```

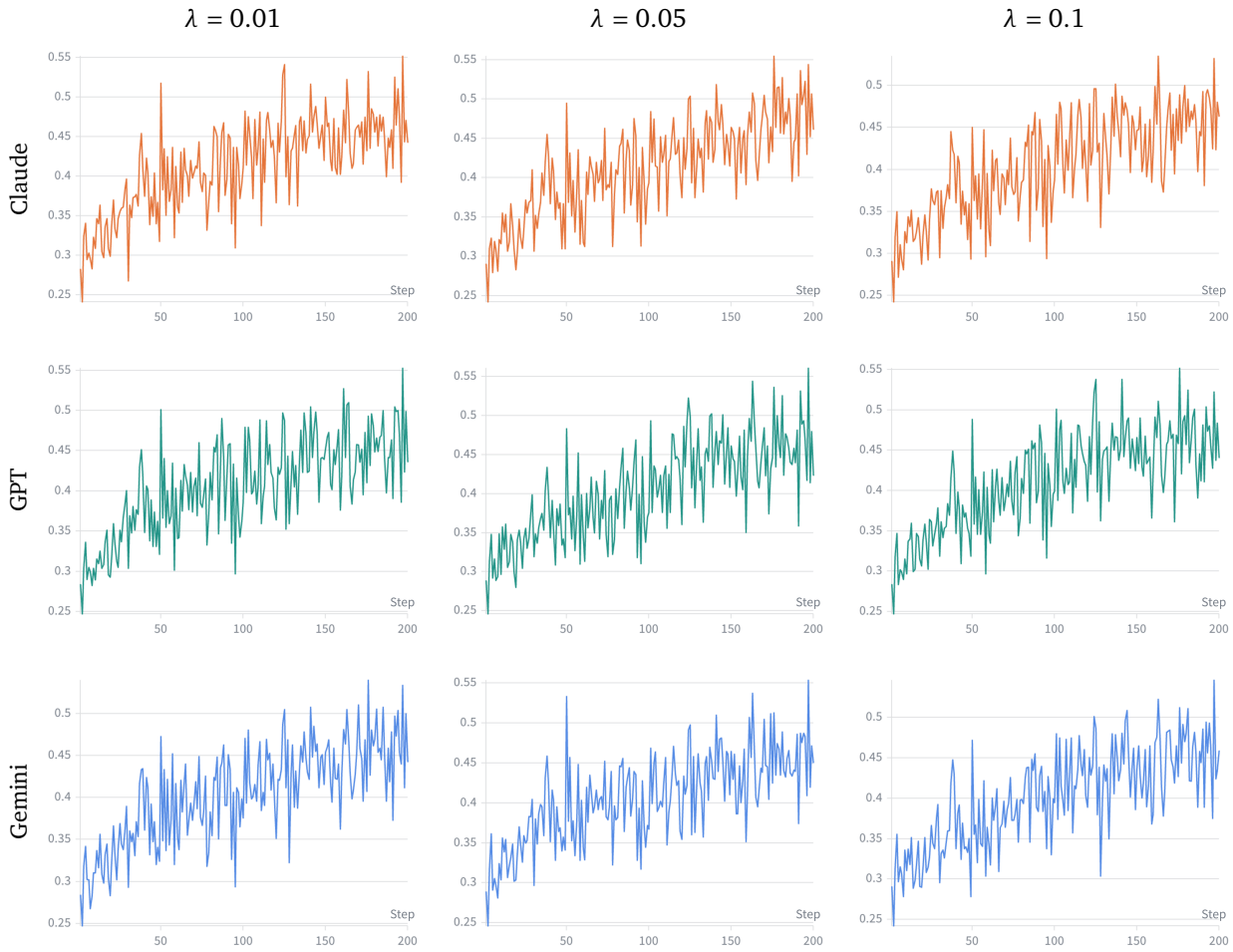

Figure 5 | **Training success rate** of **Qwen3-1.7B** under varying distillation weight λ_{OPSD} with different proprietary teacher models.

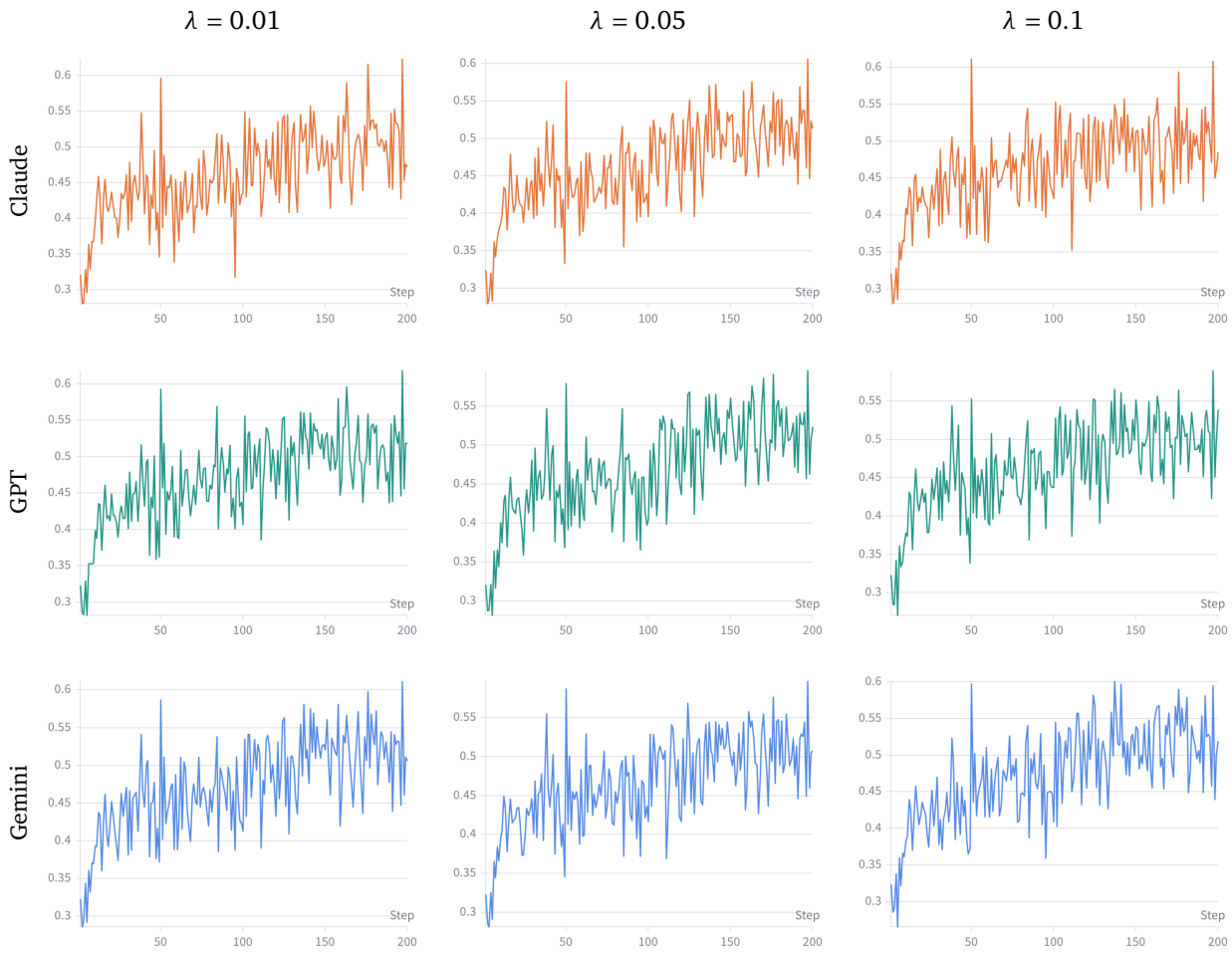

Figure 6 | **Training success rate** of **Qwen3-4B** under varying distillation weight λ_{OPSD} with different proprietary teacher models.