AgentArk: Distilling Multi-Agent Intelligence into a Single LLM Agent

Yinyi Luo[1], Yiqiao Jin[2], Weichen Yu[1], Mengqi Zhang[3], Srijan Kumar[2], Xiaoxiao Li[4], Weijie Xu[5], Xin Chen[5], Jindong Wang[3]
[1]Carnegie Mellon University, [2]Georgia Institute of Technology, [3]William & Mary, [4]University of British Columbia, [5]Amazon

amazon Research Day

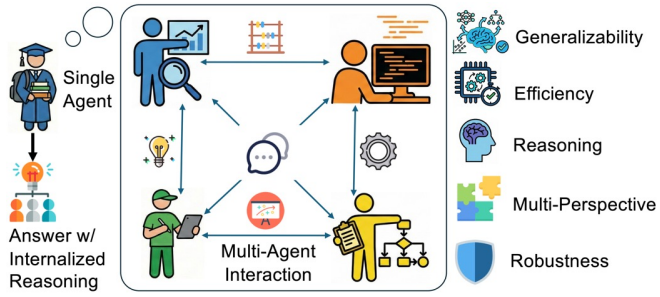
Scan for Our Paper

Multi-agent Systems (MASs)

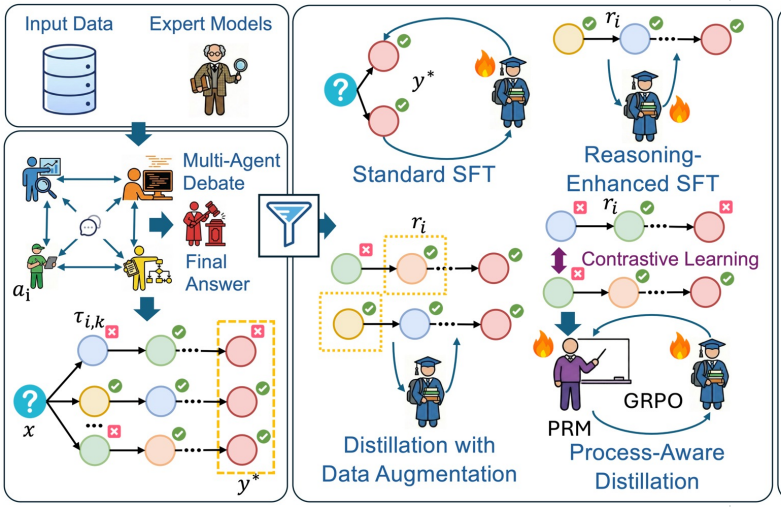
Agents interact via debate, critique, and consensus, but …

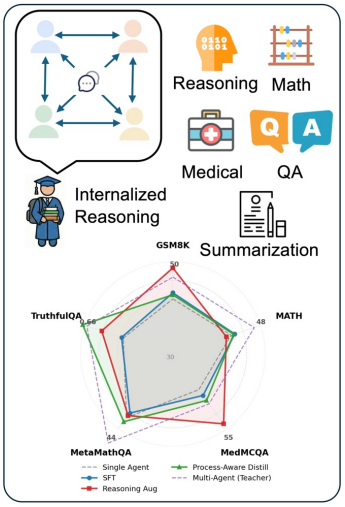
1. Multi-role and multi-turn interactions can be expensive and amplify errors ❌
2. Biases & hallucinations can spread across agents, reducing robustness and safety.

Can a single model internalize multi-agent reasoning without the inference-time cost and vulnerabilities ⚠️?



AgentArk distills the reasoning capability of multi-agent systems into one single agent.





- **Multi-Agent Debate** generates diverse reasoning trajectories
- **Knowledge Extraction** filters high-quality corrective traces
- **Distillation** uses RSFT, DA, and PAD (PRM + GRPO)
- **Student Model** learns efficient, generalizable reasoning

Promising results across 3 model families
- Qwen3-8B / 1.7B / 0.6B
- Gemma3-27B / 7B
- Llama3-8B-Instruct

Key Findings

- **Single-agent gains** AgentArk gives a single model multi-agent reasoning ability; all three methods help, and combining them helps more.
- **PRM matters most** PRM capacity matters more than student size, while student capacity limits the gains from larger teacher ensembles.
- **Quality > quantity** More trajectories alone do not help; PAD's high-signal process supervision yields more stable gains.
- **Better reasoning, not just accuracy** PAD improves step decomposition, self-checking, and error correction beyond RSFT and DA.
- **Stronger transfer** AgentArk improves generalization and robustness on unseen and robustness benchmarks.

Reasoning-enhanced Supervised Fine-Tuning (RSFT).
Use both the final answers and reasoning traces as supervision.

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,r,y^*)\sim\mathcal{D}}\mathcal{L}_{\text{res}} + \mathcal{L}_{\text{ans}}$$

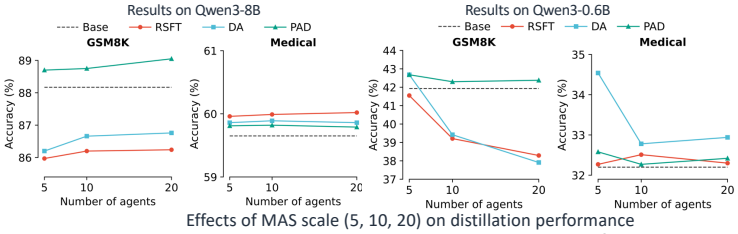
$$\mathcal{L}_{\text{res}} = \sum_{t=1}^{|r|} \log p_\theta(r_t \mid r_{<t}, x) \quad \mathcal{L}_{\text{ans}} = \log p_\theta(y^* \mid r, x)$$

Data Augmentation via Diverse Extraction (DA)
Filter the agents A into a subset of successful contributors $A_{correct} \subseteq A$.
Utilize a high-capacity teacher LLM as a "distiller" to parse the raw debate logs of $A_{correct}$

$$\mathcal{L}_{\text{Aug}}(\theta) = -\frac{1}{k}\sum_{i=1}^{k}\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, r_i, x)$$

Process-Aware Distillation using PRM and GRPO.
Given an input $x \sim D$, sample a group of G reasoning outputs $o_1, \ldots, o_G$ from a fixed behavior policy π_{old}

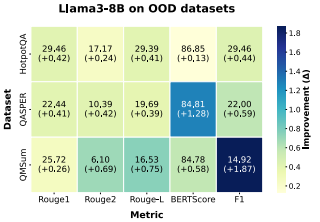
$$\mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{o_i\}\sim\pi_{\text{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\mathcal{L}_i(\theta) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\right]$$

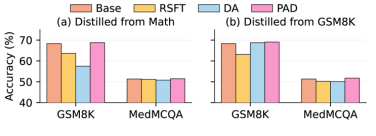


Effects of MAS scale (5, 10, 20) on distillation performance

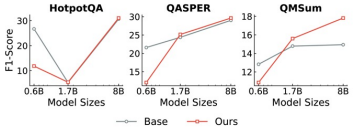
Metric	Single	RSFT	DA	PAD
Avg NLL (↓)	0.6529	0.4092	0.4449	0.5876
Perplexity (↓)	1.9211	1.6388	1.5603	1.7996
Step Decomposition (↑)	2.75	3.13	3.38	3.23
Intermediate Verification (↑)	2.41	3.48	4.04	4.07
Error Localization (↑)	1.97	2.19	2.91	2.78
Reasoning Coherence (↑)	1.88	2.25	3.07	3.96

Reasoning Quality





Multimodal distillation (GSM8K & MedMCQA)



AgentArk vs. base models (OOD datasets)

Contact: yinyil@andrew.cmu.edu, yjin328@gatech.edu
Corresponding authors: weijiexu@amazon.com, jdw@wm.edu