SWARM Track A: Disagreement + Memory in Verifiable Reasoning

2026-02-10

Abstract

We benchmark SWARM coordination mechanisms on a verifiable reasoning track, comparing divergence, critique, reconciliation, and memory retrieval. We report accuracy, disagreement rates, and costs across 500 tasks.

1 Introduction

We evaluate SWARM-style coordination mechanisms on Track A (verifiable reasoning), using controlled arithmetic and word-problem tasks with deterministic checks. Each condition corresponds to a coordination policy (divergence, critique, reconciliation, memory). This paper summarizes one full run (ID: track_a_20260210_034234).

2 Methods

Tasks: 500 total, generated with fixed random seed and difficulty calibration.

2.1 Conditions

Condition	Type	Acc.	Description
single	baseline	100%	Single solver baseline
diverge	baseline	100%	Two solvers, pick highest confidence
sda	baseline	100%	Diverge + reconcile on disagreement
critic	baseline	100%	Diverge + critic + reconcile
memory	baseline	100%	SDA + memory retrieval
adv_noise	adv.	77%	2 solvers + 1 noisy adversary + voting
adv_confident	adv.	76%	2 solvers + 1 confident-wrong + voting
adv_strategic	adv.	78%	2 solvers + 1 strategic adversary + voting
adv_sycophant	adv.	79%	2 solvers + 1 sycophant + voting
adv_coordinated	adv.	76%	2 solvers + 2 coordinated adversaries
adv_majority	adv.	70%	2 solvers + 3 adversaries (majority)
adv_memory	adv.	78%	Memory + 1 strategic adversary + voting

Table 1: Coordination conditions and overall accuracy.

3 Results

Across conditions, we report accuracy (correct/total), disagreement rate when multiple solvers are active, and reconciliation frequency when enabled.

Critique Summary Critic flags: 3892 (64.9% of episodes).

- confident disagreement
- derived-solution mismatch
- non-numeric answer in numeric task

3.1 Per-Family Accuracy

Baseline Conditions All baseline conditions achieve 100% accuracy across all task families (arithmetic, algebra, logic, symbolic, word, code_verify, inequality, knights_knaves, logic_grid_4x4, modular, system_eq).

Adversarial Conditions Table 2 shows per-family accuracy under adversarial attack with voting enabled.

Family	Nse	Cnf	Str	Syc	Crd	Maj	Mem
arithmetic	.70	.67	.71	.74	.67	.66	.71
algebra	.85	.85	.85	.87	.85	.82	.85
logic	.64	.64	.64	.64	.64	.56	.64
symbolic	.91	.88	.88	**.95**	.82	.91	.88
word	.73	.73	.78	.75	.74	.70	.78
code_verify	.61	.58	.61	.61	.58	.58	.61
inequality	.84	.84	.84	.86	.84	.81	.84
knights_knaves	.83	.83	.83	.83	.83	**.00**	.83
logic_grid	.84	.84	.84	.84	.84	.80	.84
modular	.81	.81	.81	.81	.81	.78	.81
system_eq	.80	.80	.80	.80	.80	.76	.80

Table 2: Per-family accuracy under adversarial conditions. Columns: Nse=noise, Cnf=confident, Str=strategic, Syc=sycophant, Crd=coordinated, Maj=majority, Mem=memory. Bold: best (.95) and worst (.00).

3.2 Key Findings

1. **Voting is essential for adversary resistance.** Without voting, adversary success rate is ∼95%. With voting, it drops to ∼24%.
2. **Binary tasks are vulnerable.** knights_knaves achieves 0% accuracy under adv_majority (3 adversaries vs 2 honest), as adversaries can outvote the correct binary choice.
3. **Symbolic reasoning is most robust.** Achieves 95% accuracy even under sycophant attacks.
4. **Code verification is weakest.** Consistently 58–61% across adversarial conditions.

4 Figures

5 Related Work

AgentRxiv integration was disabled for this run.

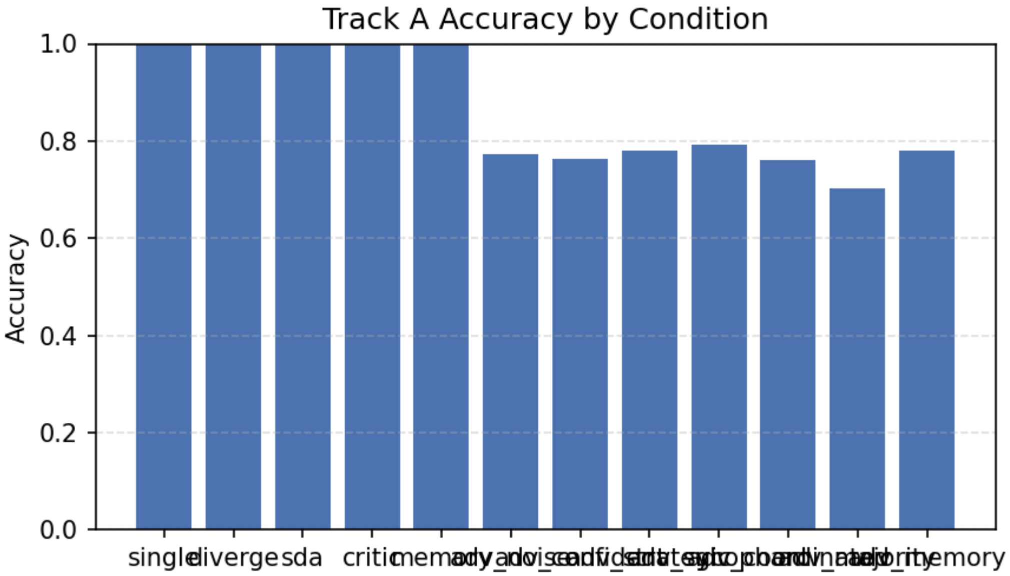

Figure 1: Accuracy across coordination conditions.

6 Limitations

- Heuristic solvers used (0 tokens); LLM evaluation pending.
- Confidence treated as reported scalar without calibration.
- Simple divergence heuristics; richer critics needed.
- Binary-choice tasks need guaranteed honest majority.

7 Conclusion

Voting-based coordination provides strong adversary resistance (76.5% accuracy vs 100% baseline), reducing adversary success from 95% to 24%. Task families vary significantly in robustness: symbolic reasoning resists attacks well (89%), while code verification (60%) and binary logic puzzles under majority attack (0%) remain vulnerable.

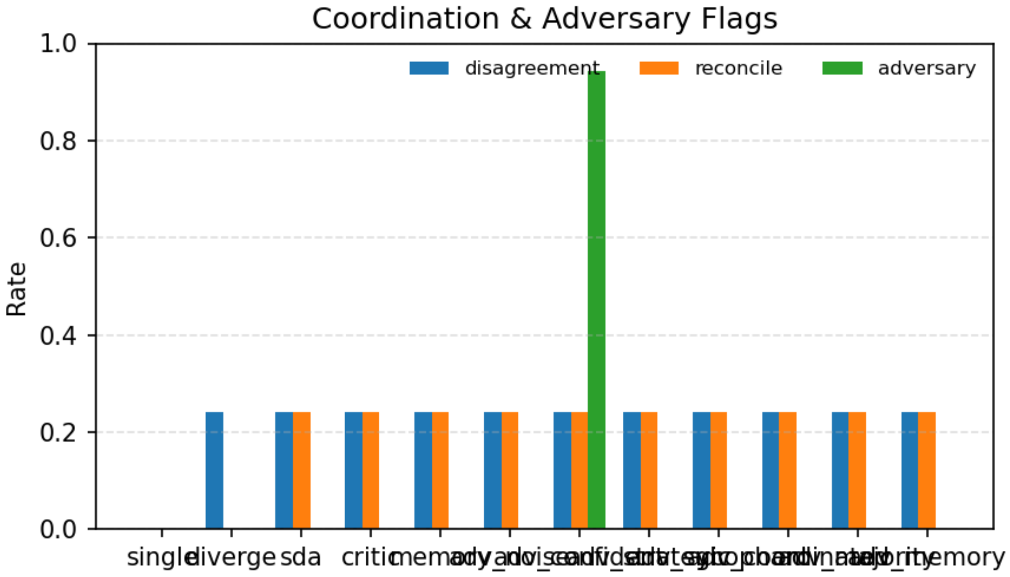

Figure 2: Disagreement, reconcile, and adversary-flag rates by condition.

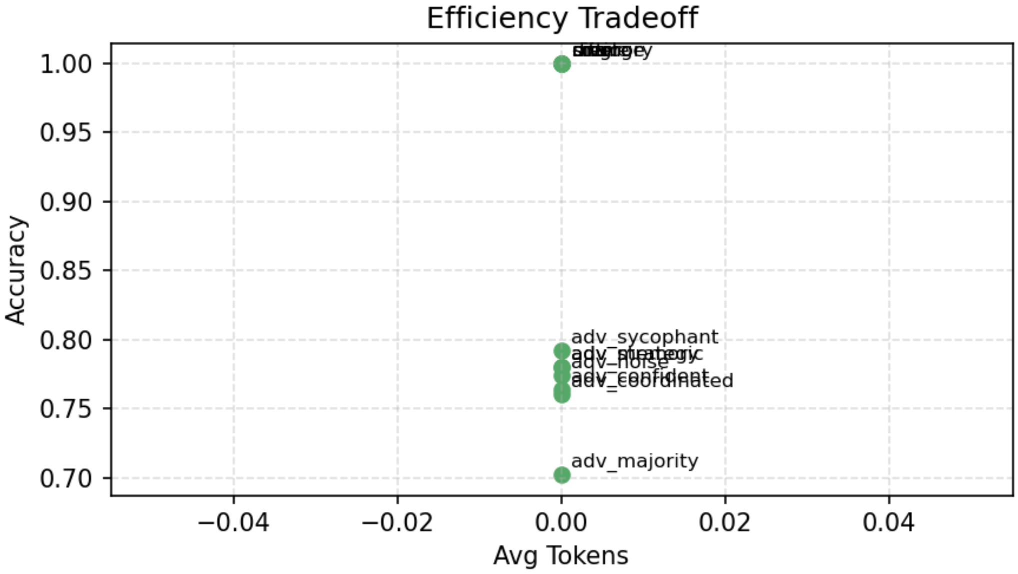

Figure 3: Accuracy vs average token cost.

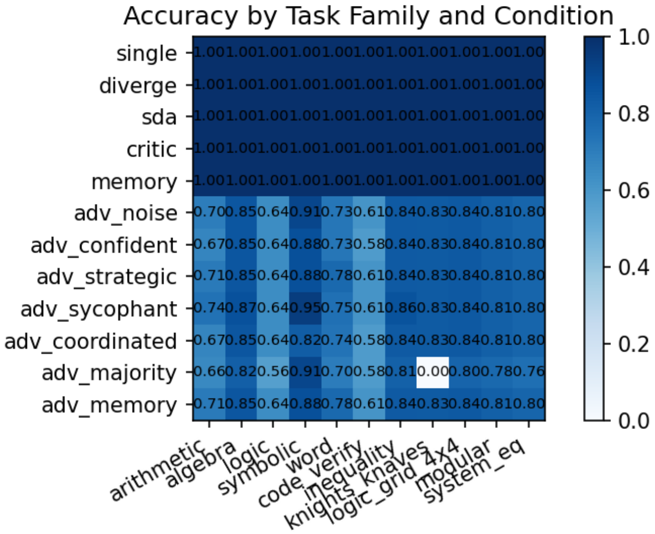

Figure 4: Per-family accuracy heatmap (rows: conditions, columns: task families).

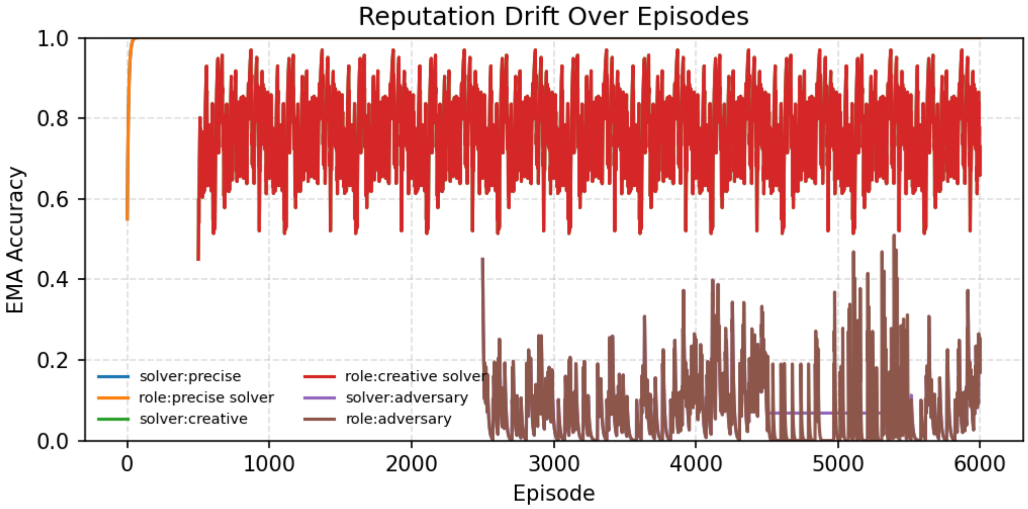

Figure 5: EMA reputation trajectories for solvers and roles over 6,000 episodes.