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_044024).

2 Methods

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

2.1 Conditions

Condition	Accuracy	Tokens	Notes
single	1.000	0.0	Single solver baseline
diverge	1.000	0.0	Two solvers, pick highest confidence
sda	1.000	0.0	Diverge + reconcile on disagreement
critic	1.000	0.0	Diverge + critic + reconcile
memory	1.000	0.0	SDA + memory retrieval
adv_noise	0.774	0.0	Two solvers + 1 noisy adversary + voting
adv_confident	0.764	0.0	Two solvers + 1 confident-wrong adversary + voting
adv_strategic	0.780	0.0	Two solvers + 1 strategic adversary + voting
adv_sycophant	0.792	0.0	Two solvers + 1 sycophant adversary + voting
adv_coordinated	0.760	0.0	Two solvers + 2 coordinated adversaries + voting
adv_majority	0.702	0.0	Two solvers + 3 adversaries (adversary majority) + voting
adv_memory	0.780	0.0	Memory condition + 1 strategic adversary + voting

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

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

3.1 Per-Family Accuracy (Baseline)

Family	single	diverge	sda	critic	memory
arithmetic	1.00	1.00	1.00	1.00	1.00
algebra	1.00	1.00	1.00	1.00	1.00
logic	1.00	1.00	1.00	1.00	1.00
symbolic	1.00	1.00	1.00	1.00	1.00
word	1.00	1.00	1.00	1.00	1.00
code_verify	1.00	1.00	1.00	1.00	1.00
inequality	1.00	1.00	1.00	1.00	1.00
knights_knaves	1.00	1.00	1.00	1.00	1.00
logic_grid_4x4	1.00	1.00	1.00	1.00	1.00
modular	1.00	1.00	1.00	1.00	1.00
system_eq	1.00	1.00	1.00	1.00	1.00

3.2 Per-Family Accuracy (Adversarial)

Family	nse	cnf	str	syc	crd	maj	mem
arithmetic	0.70	0.67	0.71	0.74	0.67	0.66	0.71
algebra	0.85	0.85	0.85	0.87	0.85	0.82	0.85
logic	0.64	0.64	0.64	0.64	0.64	0.56	0.64
symbolic	0.91	0.88	0.88	0.95	0.82	0.91	0.88
word	0.73	0.73	0.78	0.75	0.74	0.70	0.78
code_verify	0.61	0.58	0.61	0.61	0.58	0.58	0.61
inequality	0.84	0.84	0.84	0.86	0.84	0.81	0.84
knights_knaves	0.83	0.83	0.83	0.83	0.83	0.00	0.83
logic_grid_4x4	0.84	0.84	0.84	0.84	0.84	0.80	0.84
modular	0.81	0.81	0.81	0.81	0.81	0.78	0.81
system_eq	0.80	0.80	0.80	0.80	0.80	0.76	0.80

Legend: nse=noise, cnf=confident, str=strategic, syc=sycophant, crd=coordinated, maj=majority, mem=memory

4 Figures

5 Related Work (AgentRxiv)

- None.

6 Memory Artifacts

No memory artifacts were accepted in this run.

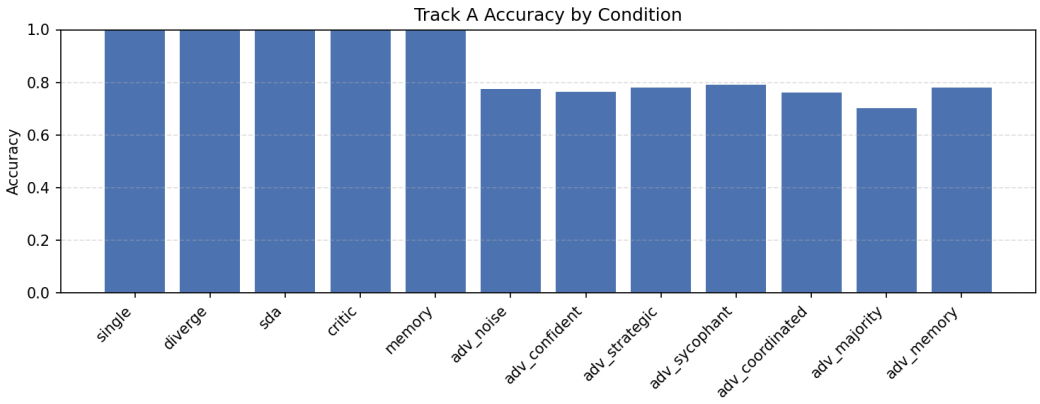

Figure 1: Accuracy across coordination conditions.

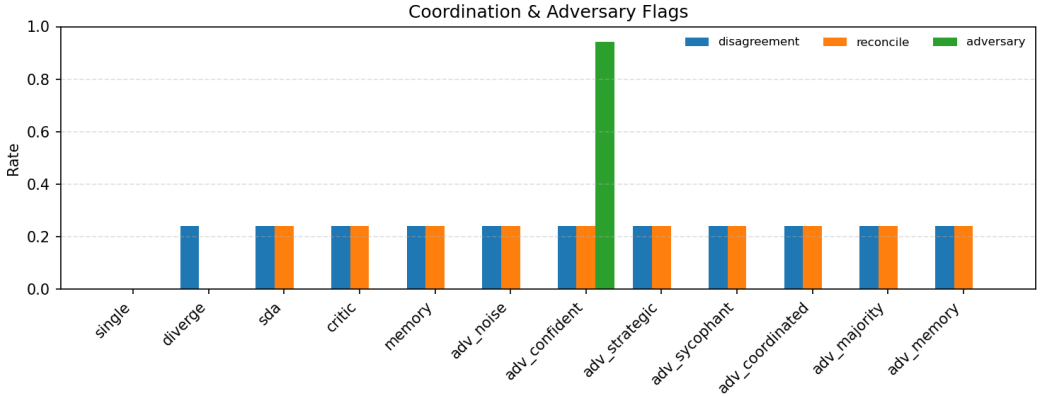

Figure 2: Disagreement, reconcile, and adversary-flag rates by condition.

7 Limitations

We treat confidence as a reported scalar and rely on simple divergence heuristics. Future runs should incorporate stronger validators and richer task suites.

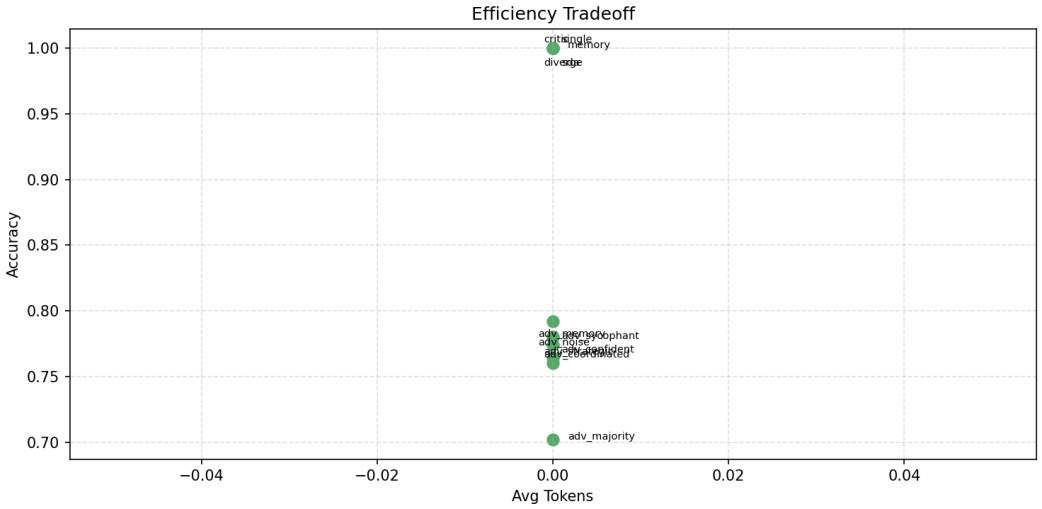

Figure 3: Accuracy vs average token cost.

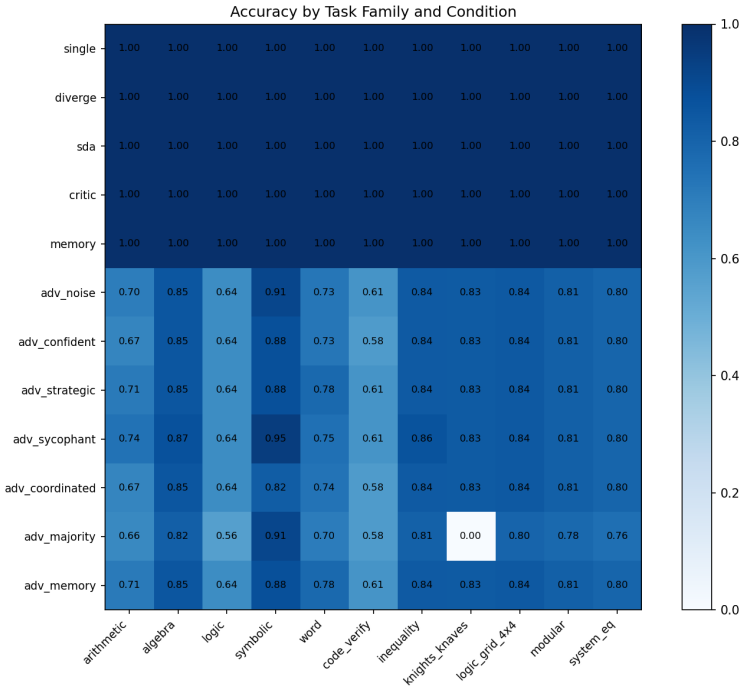

Figure 4: Per-family accuracy (arithmetic, algebra, logic, symbolic, word).

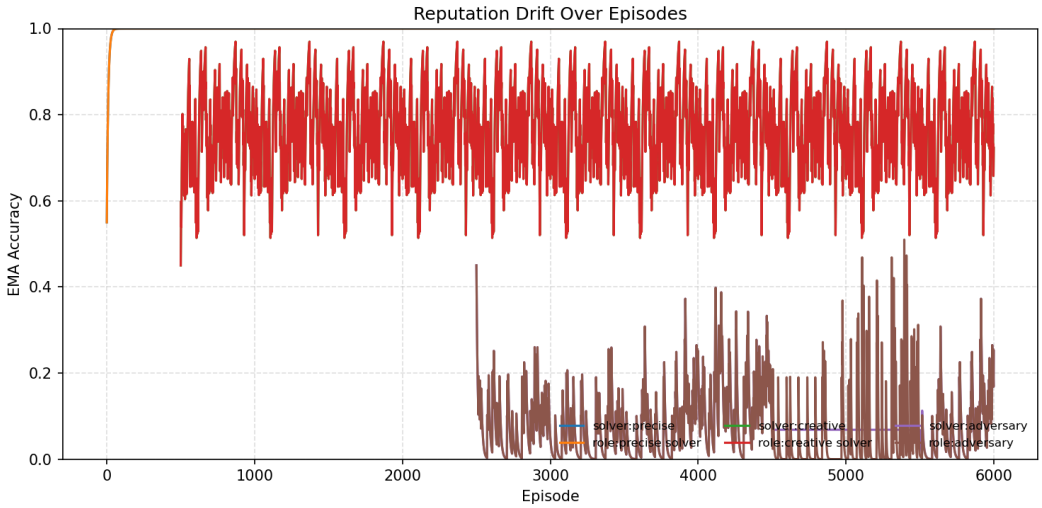

Figure 5: EMA reputation trajectories for key solvers/roles.