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_040154).

2 Methods

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

2.1 Conditions

Condition	Accuracy	Avg. 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

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: 430 (17.2

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

3.1 Per-Family Accuracy

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 Token Efficiency (Correct per 1k tokens)

Family	single	diverge	sda	critic	memory
arithmetic	0.00	0.00	0.00	0.00	0.00
algebra	0.00	0.00	0.00	0.00	0.00
logic	0.00	0.00	0.00	0.00	0.00
symbolic	0.00	0.00	0.00	0.00	0.00
word	0.00	0.00	0.00	0.00	0.00
code_verify	0.00	0.00	0.00	0.00	0.00
inequality	0.00	0.00	0.00	0.00	0.00
knights_knaves	0.00	0.00	0.00	0.00	0.00
logic_grid_4x4	0.00	0.00	0.00	0.00	0.00
modular	0.00	0.00	0.00	0.00	0.00
system_eq	0.00	0.00	0.00	0.00	0.00

4 Figures

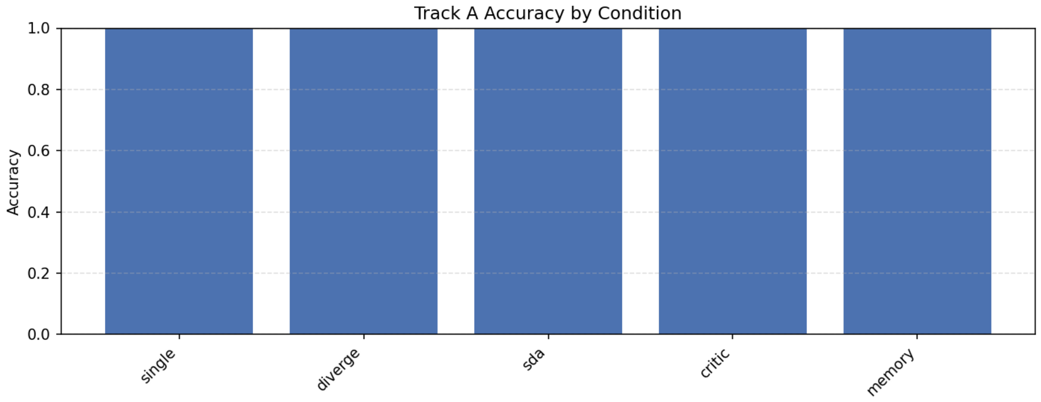

Figure 1: Accuracy across coordination conditions.

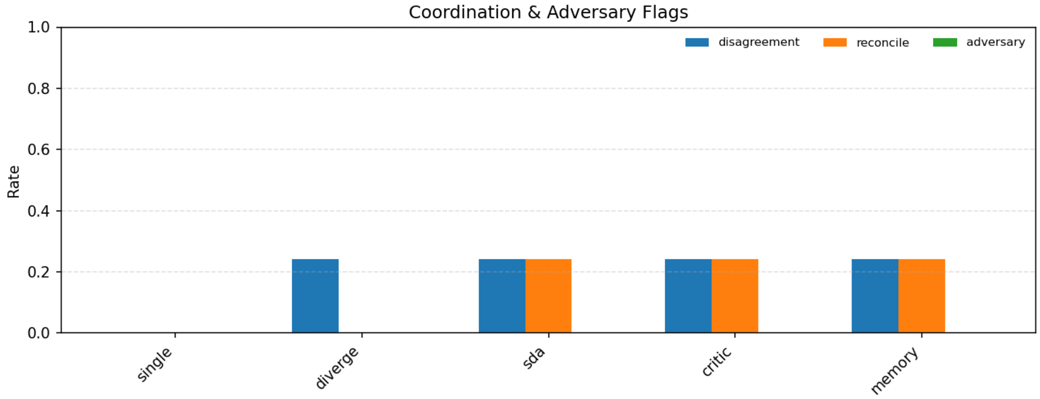

Figure 2: Disagreement, reconcile, and adversary-flag rates by condition.

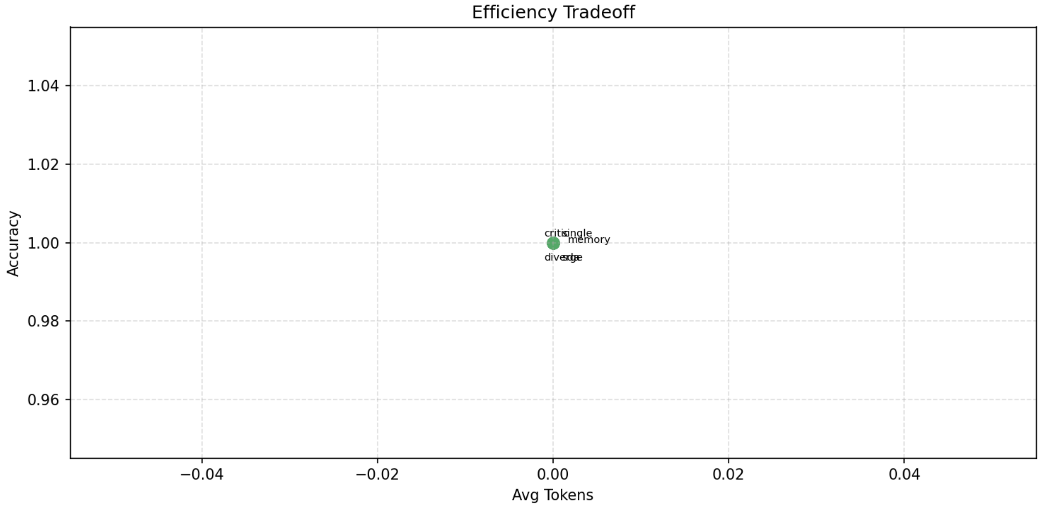

Figure 3: Accuracy vs average token cost.

5 Related Work (AgentRxiv)

- None.

6 Memory Artifacts

No memory artifacts were accepted in this run.

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.

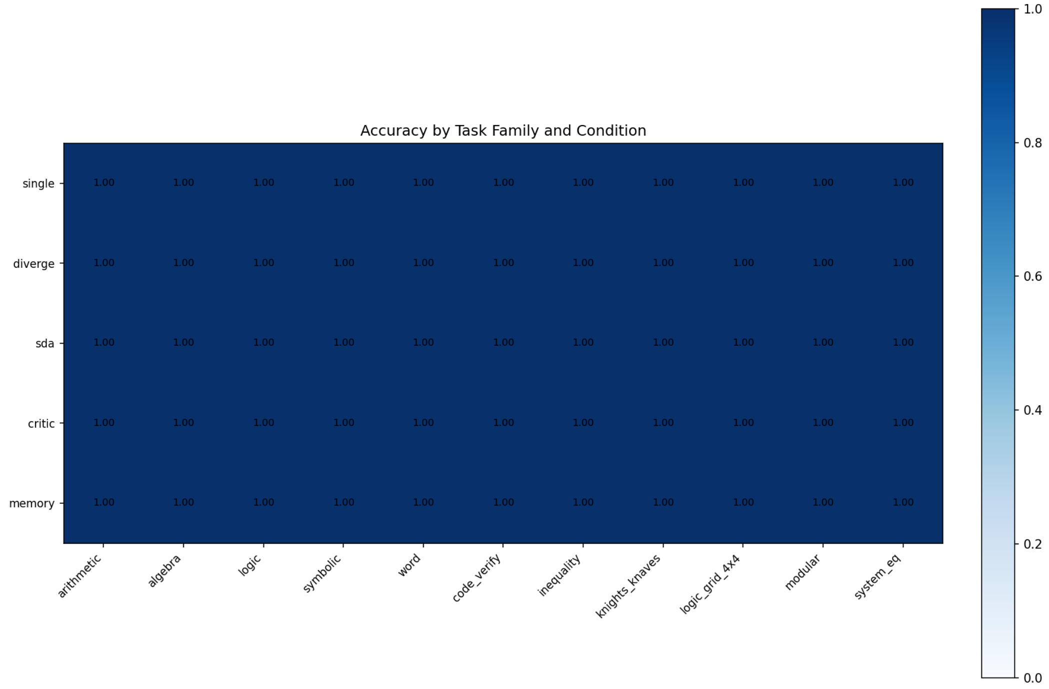

Figure 4: Per-family accuracy (arithmetic, algebra, logic, symbolic, word).

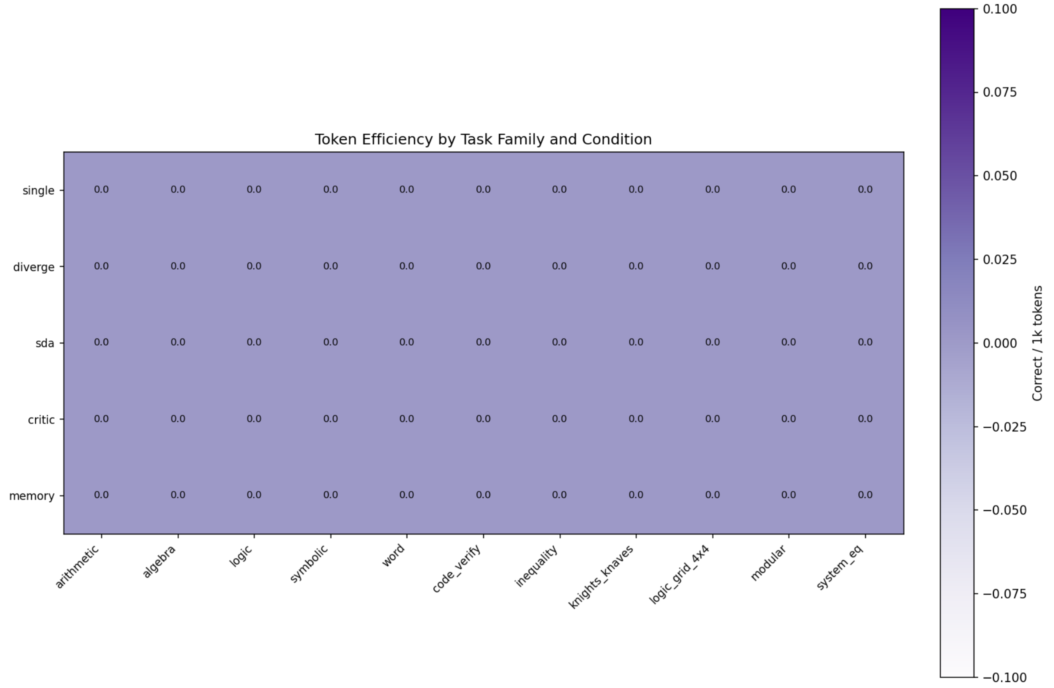

Figure 5: Per-family token efficiency (correct per 1k tokens).

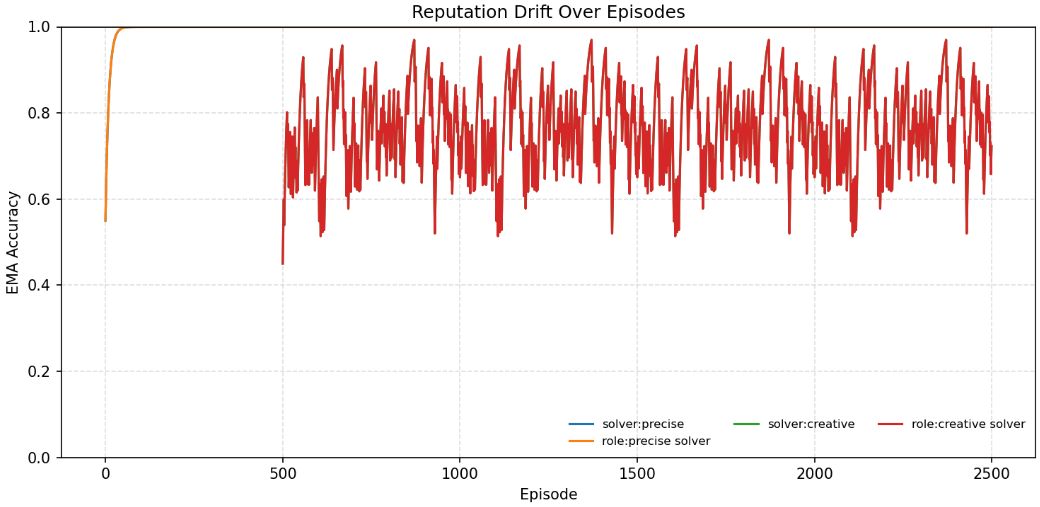

Figure 6: EMA reputation trajectories for key solvers/roles.