The Illusion of Automatically Orchestrated Multi-Agent Advantage

Prathyusha Jwalapuram*[1], Hehai Lin*[2], Chuyuan Li[3], Fangkai Jiao[4], Sudong Wang[2], Yifei Ming[1], Zixuan Ke[1], Chengwei Qin[2,5], Giuseppe Carenini[3], Shafiq Joty[1,4]

[1]Salesforce AI Research
[2]HKUST (Guangzhou)
[3]University of British Columbia
[4]Nanyang Technological University
[5]HKUST
*Equal contribution



NeurIPS 2026

Scan for paper, dataset + code

salesforce

Automated MAS: More Cost, No Gain

① INEFFICIENT

salesforce

6 Automated MAS

INFERENCE-TIME

DyLAN | MAS-Zero

OPTIMIZED (SEARCH / RL)

ADAS | AFlow | MaAS | Orch.

4 Benchmarks

GPQA-Diamond | HLE-Math

SWE-Bench Lite | BrowseComp+

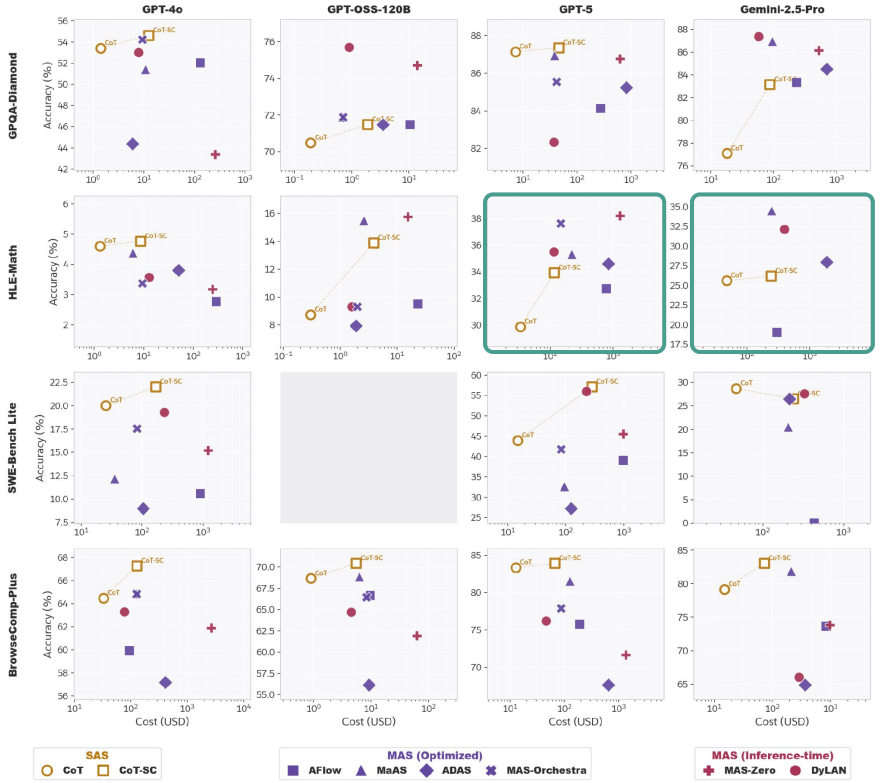
4 Backbone LLMs

GPT-4o | GPT-OSS-120B

GPT-5 | Gemini-2.5-Pro

Baseline: CoT-SC · k=3 majority vote

cost incl. search + validation · 3 runs



Single agent wins

CoT-SC matches or beats MAS at a fraction of the cost

HLE-Math uplift

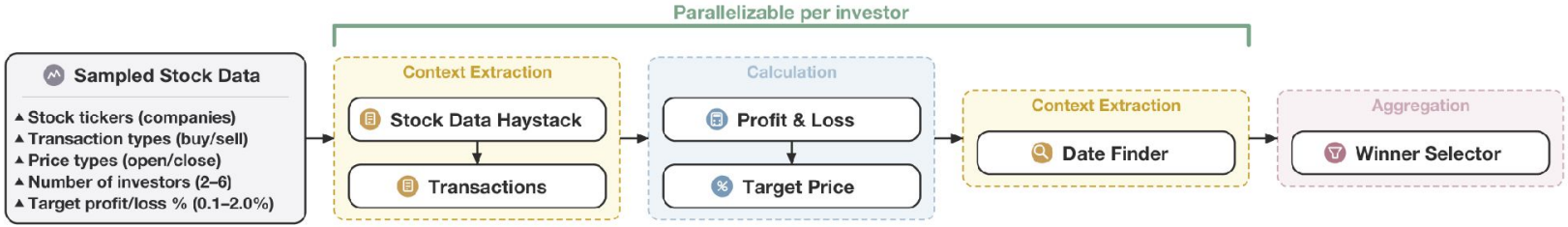
Only real MAS gain, but at 11× cost

No capability bridging

GPT-5 CoT-SC beats the best GPT-4o MAS at <½ the tokens

Was the Task the Problem? Enter SMFR!

② TASK ≠ PROBLEM



Decomposition | Context separation | Parallelism per investor (N = 2–6)

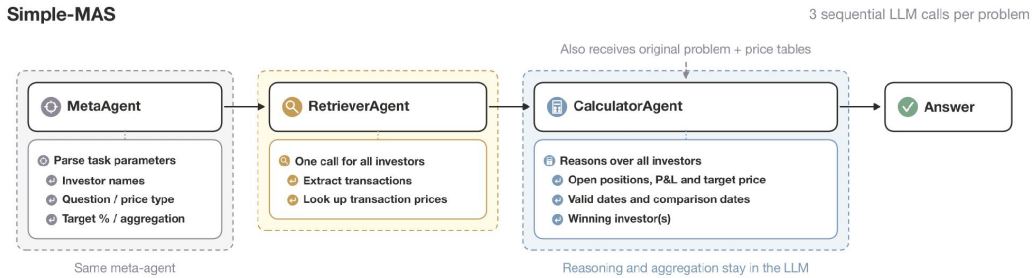
- **Prior work:** common MAS benchmarks test **static reasoning**, not what MAS are for Kapoor+ '24 · Kim+ '25
- **S**ynthetic **M**ulti-hop **F**inancial **R**easoning: find when each investor hits a P&L target in a 30-day price haystack, then pick the winner built per Anthropic's MAS guidelines
- **588** test (+16 val) · real US equity prices · procedurally generated (contamination-free) · ~**4.3K**-token prompts

Engineered MAS: Simple → Expert

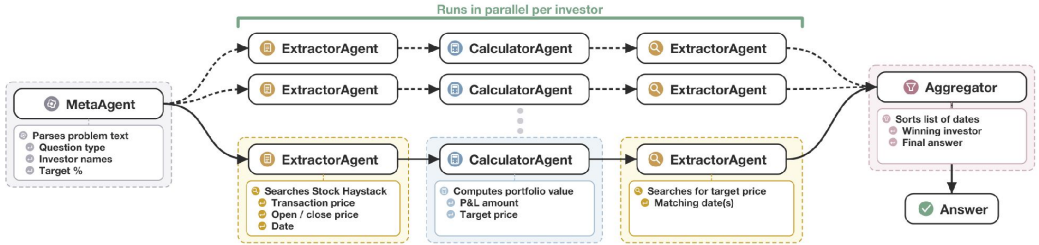
② TASK ≠ PROBLEM

salesforce

TIER 2 · CONTEXT-SEPARATION ANCHOR **Simple-MAS**: retrieval ⊥ reasoning · 3 calls



TIER 3 · TOPOLOGY CEILING **Expert-MAS**: parallel per investor · winner chosen in code



Context separation drives accuracy;
code scaffolding + parallelism cut latency

Model	System	Acc	Cost $	Lat (tok)
GPT-4o	Simple-MAS	15.8	86.0	3,634
GPT-4o	Expert-MAS	22.6	245.1	902
GPT-5	Simple-MAS	94.8	201.0	16,659
GPT-5	Expert-MAS	96.5	554.8	10,114

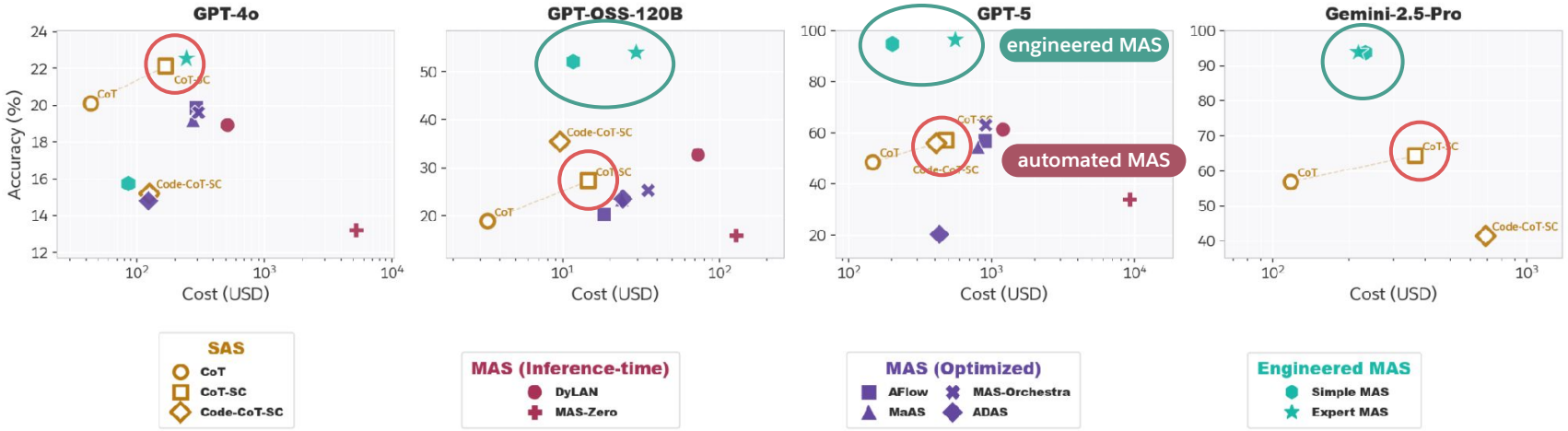
Latency = bottleneck output tokens per problem.
Expert-MAS is O(1) in N; Simple-MAS O(N).

4×
lower latency · GPT-4o

1.65×
lower latency · GPT-5

Automated MAS Architecture Was the Problem

② TASK ≠ PROBLEM

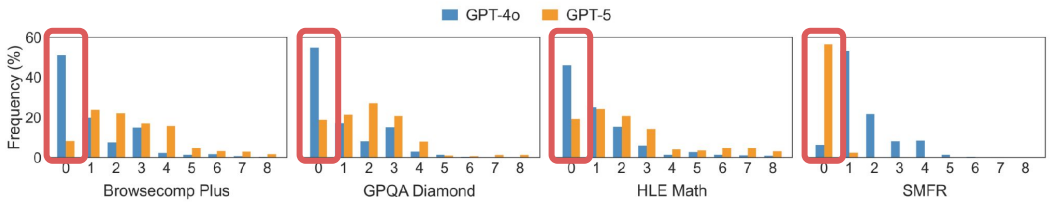


Automated MAS: ≤ +6.6 pp, strong backbones only, at 1.9–5× cost

Context separation alone (Simple-MAS) delivers nearly all the gain

Why Automated MAS Fail

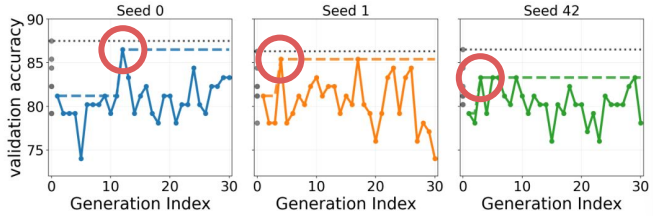
③ WHY IT FAILS



1 · ROLES COLLAPSE MAS-Zero · DyLAN

- Verifier picks **1st candidate** >45%
- DyLAN: instant consensus **70–90%**

→ **extra agents = expensive witnesses**



AFlow, GPT-5 · GPQA

```
for i in range(3):
  custom(problem)
vote(solutions)
```

= CoT-SC, k=3

2 · SEARCH REDISCOVERS CoT-SC ADAS · AFlow

- ADAS **peaks early**, then regresses
- **7/14** AFlow flows = prompt ×3 + vote

→ **~10× cost buys a relabelled CoT-SC**

MAS-Orchestra: agent selection (%)

Agent Type	GPQA Diamond	HLE-MATH	SWE-Bench Lite	BrowseComp-Plus	SMFR
CoT	0.0	0.0	0.3	0.0	0.0
CoT-SC	0.0	0.0	0.0	0.0	0.0
Reflexion	15.1	20.8	**56.0**	38.7	**71.1**
Debate	**84.9**	**79.2**	43.7	**61.3**	28.9

3 · ROUTERS IGNORE THE TASK MaAS · Orchestra

- MaAS: **74.2%** trivial I/O calls
- Orchestra: more **Debate** on easier tasks

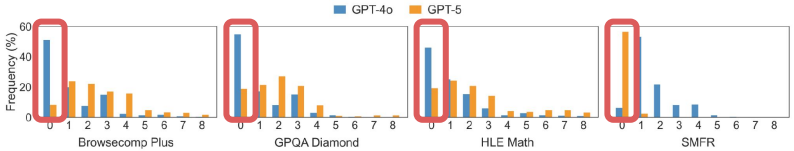
→ **static, greedy policies, no decomposition**

Why Automated MAS Fail

③ WHY IT FAILS

INFERENCE-TIME DyLAN · MAS-Zero

MAS-Zero: which candidate the verifier picks



0 = first base block · 4–8 = later search rounds

>45%
GPT-4o verifier picks the 1st block

<15%
of picks come from later rounds

DyLAN: does agent diversity show up?

70% / 90%+
instant consensus, 4o / GPT-5

54.4 ≥ 53.4
generic roles ≥ task experts (%)

→ extra agents are expensive witnesses

SEARCH-OPTIMIZED ADAS · AFlow

ADAS: motifs in discovered architectures (GPQA)

Self-consistency
+1.34 pp
only positive signal (82.19%)

Specialized coordination
≈ 0
negligible gain

Search is non-monotonic: peaks early, then regresses ("lucky" iterations)

AFlow: the "optimized" workflow (GPT-5 · GPQA)

```
for i in range(3):
    custom(problem)
vote(solutions)
```

= CoT-SC, k = 3

7 / 14
final workflows; 4 of 7 below CoT-SC

→ search rediscovers CoT-SC

Use SMFR to Diagnose Your Automated MAS

Tiered, cost- and topology-efficient anchors: optimize **task decomposition** without blowing up cost

WHAT WE FOUND	THE SMFR ANCHOR THAT EXPOSES IT
1 Inefficient — rarely beat CoT-SC, at up to 10× cost	→ TIER 1 · COST-EFFICIENCY FLOOR: **CoT-SC** 57.0% · $478 — ✓ Beat it, or the extra compute isn't justified
2 The design, not the task — engineered MAS hit 96.5% where automated MAS stall	→ TIER 2 · CONTEXT-SEPARATION ANCHOR: **Simple-MAS** 94.8% · $201 — ✓ Find at least a retrieval / reasoning split
3 Bloated architectures — roles collapse · search → CoT-SC · routers ignore the task	→ TIER 3 · TOPOLOGY CEILING: **Expert-MAS** 96.5% · $555 — ✓ Decompose + parallelize: 1.65× lower latency

Benchmark
your automated MAS

dataset · code · anchors
CoT-SC → Simple → Expert

Today's best automated MAS (GPT-5): 63.0% at $910 – above the floor, far below Tier 2 at 4.5× its cost

The Illusion of Automatically Orchestrated Multi-Agent Advantage · NeurIPS 2026