# Compute-Aware Mixture-of-Agents

**Verifier-Gated Adaptive Aggregation under a Fixed Token Budget**

Shine Gupta[1] · S Akash[2] — [1]KIET Group of Institutions · [2]IIT Patna

SCALE · ICML 2026
Seoul

GSM8K · SIMILAR-CAPABILITY POOL · N=20 · AVG LAYERS RUN = 1.25

**85%**

**Spend deep compute only when it helps — CA-MoA ties top accuracy and Pareto-dominates fixed-depth MoA.**

- **+5.0 pt** over MoA-1, at 1.4× tokens
- **+20 pt** over Single (65→85%)
- **89–100%** of $\ell$ =1 stops correct
- **0–40%** of $\ell$ =3 runs correct

## 1 The problem

Standard **Mixture-of-Agents** spends the **same L · K** LLM calls on **every** query — yet most solve at layer 1.

Cast it as a **stochastic knapsack**:

$$\max_\pi \mathbb{E}_x[Q(\pi(x))] \quad \text{s.t.} \quad \mathbb{E}_x[C(\pi(x))] \le B$$

## 2 Method: CA-MoA

A verifier **V** scores each layer's aggregate and **stops** per-query once the score clears a budget-aware bar τ **(D1)**.

**ALGORITHM 1 · CA-MOA**

```
in: x; V; pool A; budget B;
    L_max; τ(·); width k
a ← x;  b ← B;  ℓ ← 0
while ℓ < L_max and b > 0:
  s ← V(x, a)          # score agg
  if s ≥ τ(b/B): break  # D1: GO
  for A_j in A:         # D2 route*
    û_j ← V(x,a|A_j) − s
  S ← top-k(û) s.t. cost ≤ b
  a ← aggregate( run(S, a) )
  b −= cost(S);  ℓ += 1
return a        # *D1-only eval
```

$$\tau(\varrho) = \tau_{min} + (\tau_{max}-\tau_{min})\cdot\varrho\ , \quad \varrho = b/B$$

**τ is a single Pareto knob** — monotone in τ; strictly beats fixed-depth MoA **in expected cost** when $E[L] < L_{max}$.

## Verifier V — three forms

- **V_SC** **Agreement**, zero cost. **Used in all main experiments.**
- **V_LLM** Judge LLM. *Ablated* — over-confident.
- **V_RM** Reward model. *Pre-registered.*

## How the gate works

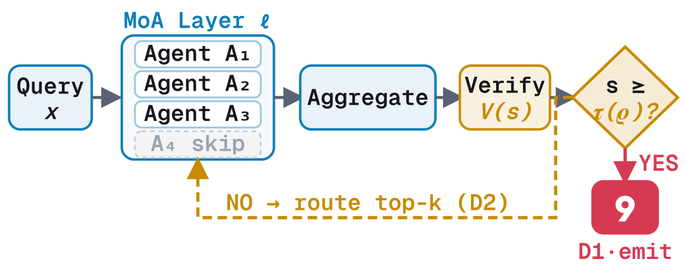


**Fig 1.** τ terminates (D1, red) or routes top-k onward (D2). Reported numbers: D1-only depth-adaptive policy.

## Pareto frontier — Exp C

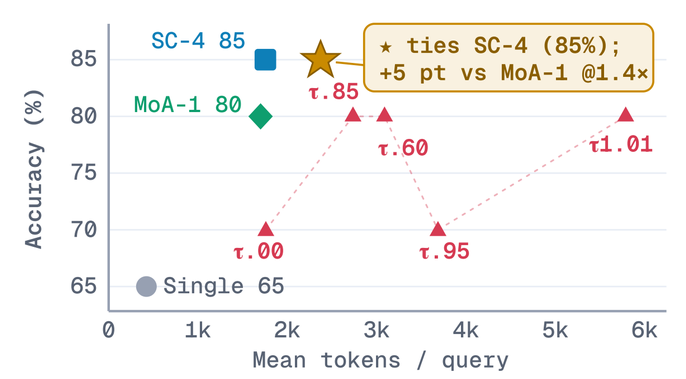


**Fig 2.** ★ CA-MoA headline (τmax=0.95) ties SC-4. Red ▲ = τ-sweep (labels = τmax); ◆ MoA-1, ■ SC-4, ● Single.

## Headline results — Exp C

| METHOD | ACC. | TOKENS | AVGL |
|---|---|---|---|
| Single (Qwen2.5-3B) | 65.0 | 423 | – |
| MoA-1 (Wang '24) | 80.0 | 1,700 | 1.00 |
| SC-4 (Wang '23) | 85.0 | 1,754 | – |
| **CA-MoA (ours)** | **85.0** | **2,372** | **1.25** |

Pool {Llama-3.2-3B, Qwen2.5-3B}, **n=20**. **AvgL 1.25** → most queries stop at ℓ1.

**Adaptation is real, measured**

Token CV runs **3–4× higher** than fixed-cost baselines:

- **.62–.96** CA-MoA token CV
- **.24–.34** fixed baselines

## Gating quality replicates

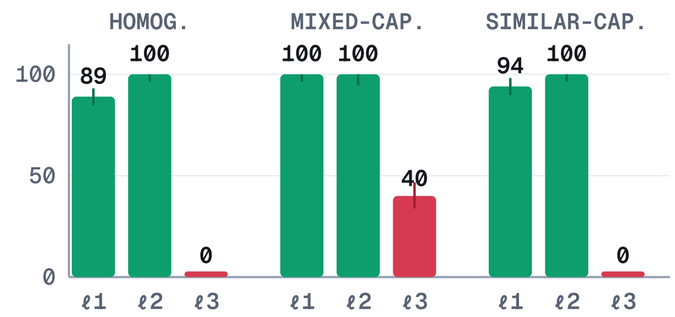


**Fig 4.** Accuracy by layers ℓ permitted. **ℓ=1 stops 89–100% correct; ℓ=3 runs 0–40%** across pools. n = 1–44 per bucket; small-n ℓ2/ℓ3 CIs wide.

## Where the win lives — 3 regimes

- **GO · C** **Balanced.** +5 vs MoA-1. **Clean win.**
- **WEIGH · B** **Broken.** +13.3, weak agent corrupts.
- **STOP · A** **Saturated.** −8: false-positive stops.

| METHOD | A·N50 | B·N15 | C·N20 |
|---|---|---|---|
| Single | 90.0 | 73.3 | 65.0 |
| SC-4 | 88.0 | 46.7 | 85.0 |
| MoA-1 | 92.0 | 46.7 | 80.0 |
| CA-MoA | 84.0 | 60.0 | 85.0 |

Accuracy (%) · **bold** = regime winner.

> *The signal is regime-independent; its payoff is not.*

## Honest limits & what's next

- Small **n** (50/15/20); no **p<0.05** — a single run swings **±15 pts** at n=20.
- Only **V_SC** validated; **D2** not yet exercised. Cost = tokens, not wall-clock.

**Pre-registered follow-up**

~7B × 3 families, **n=200**, MATH+GSM8K.
**Success:** match SC-4 at ≤1.5× tokens.

**Method:** verifier-gated, depth-adaptive MoA

**Correspondence:** shine.2226cse1184@kiet.edu