PACE

Parity-Aware Compute Evaluation

Is the multi-agent gain coordination, or just more compute?

−0.038
mean drop when the baseline changes from greedy to sampled (5 studies)

74%
of 474 contrasts: multi-agent not ahead of a sampled or budget-matched baseline

8 of 9
study means at or below zero (the exception did not match compute)

1 Find your baseline's rung

D Frontier	accuracy against compute, swept	**+0.020 (2 contrasts, compute not matched)**
C Budget-matched	same token or iteration ceiling	**−0.012 (114 contrasts)**
B Sampled	vote over k independent samples	**−0.051 (360 contrasts)**
A Single pass	one greedy or chain-of-thought pass	**−0.032 (358 contrasts)**

Mean accuracy difference, multi-agent minus single-agent, 9 studies. Credit a gain only against a baseline given at least as much compute.

2 Record six resources for both systems

R_{reason}	thinking tokens	R_{token}	total tokens	R_{cost}	dollar cost
R_{lat}	latency	R_{sample}	independent samples	R_{sys}	tool and verifier calls

Mark each one: controlled, approximately matched, unmatched, or unreported.

3 Three rules

1. Before crediting communication, run self-consistency with k ≥ k_tok samples and report the gap.
   k_tok = KT + (T−1)K(K−1)r, r = completion share of one sample's tokens
2. Use several agents for task structure (split evidence, parallel tools, separate state), not for accuracy.
3. Give any verifier or tool to both arms, and report the tool calls.

Paper arxiv.org/abs/2611.17610
Try it huggingface.co/spaces/sajjadi/pace-studio
Code github.com/mahmoudsajjadi/pace