Rethinking Optimal Verification Granularity for Compute-Efficient Test-Time Scaling

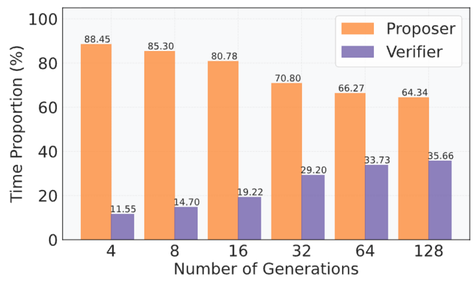
Hao (Mark) Chen[1] Guanxi Lu[1] Yasuyuki Okoshi[2] Zhiwen Mo[1] Masato Motomura[2] Hongxiang Fan[1]

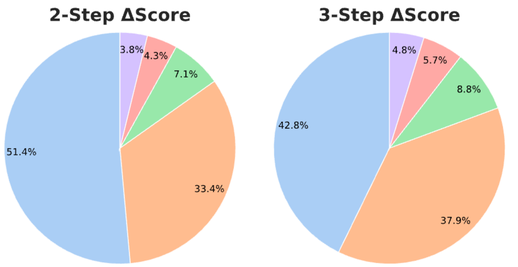
[1]Imperial College London, UK [2]Institute of Science Tokyo, Japan

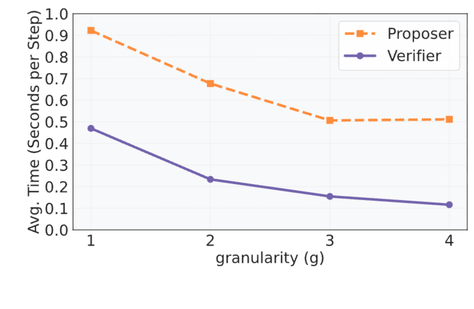
NEURAL INFORMATION PROCESSING SYSTEMS

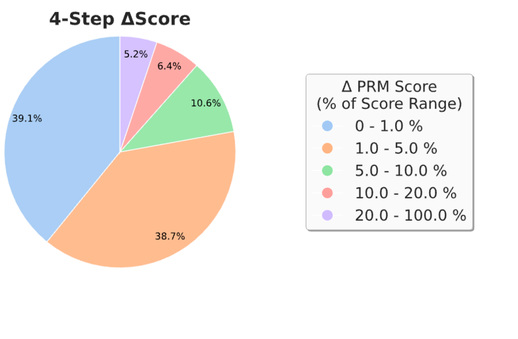
Motivation

1. Verification Cost in Test-Time Scaling is NOT Negligible
2. The Conventional Verification Granularity "\n\n" is Arbitrary

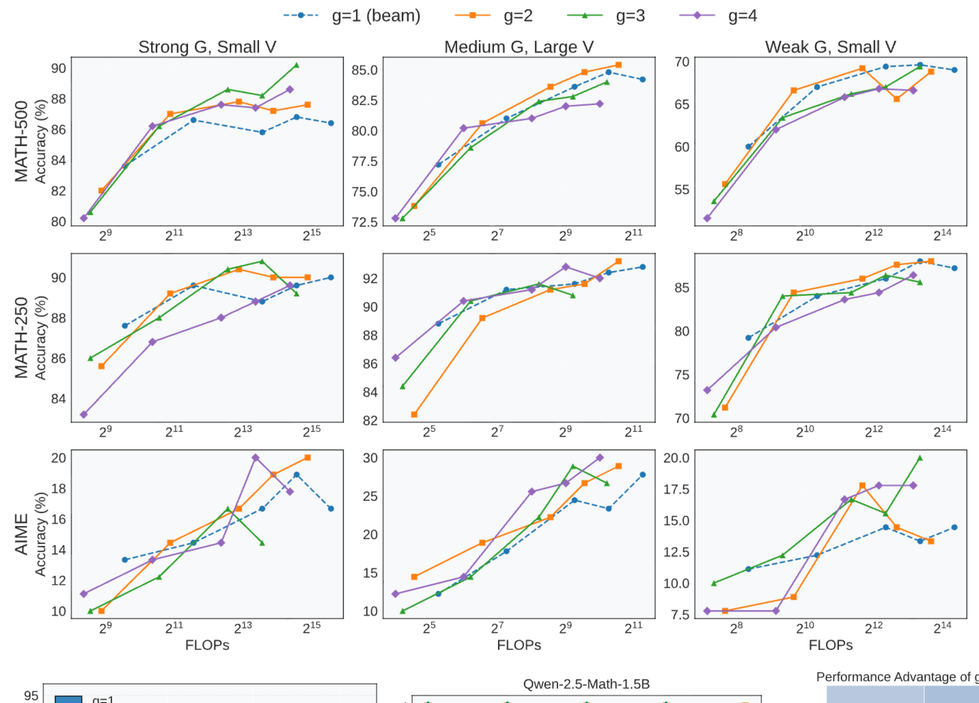




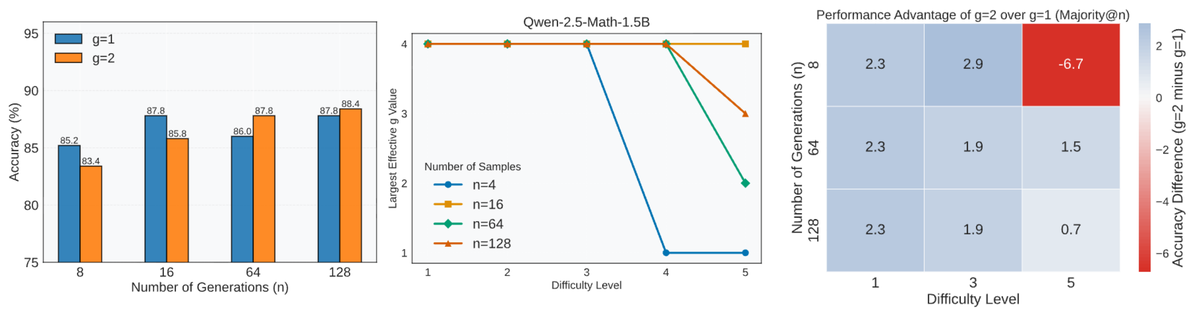




Understanding the Limits of the Verification Granularity Convention



- "\n\n" Is Not Always the Optimal Verification Granularity Boundary
- Strong Generators Prefer Sparse Verification
- Optimal Granularity Saves Computation Significantly
- Rethinking the Optimal "Thinking Step"

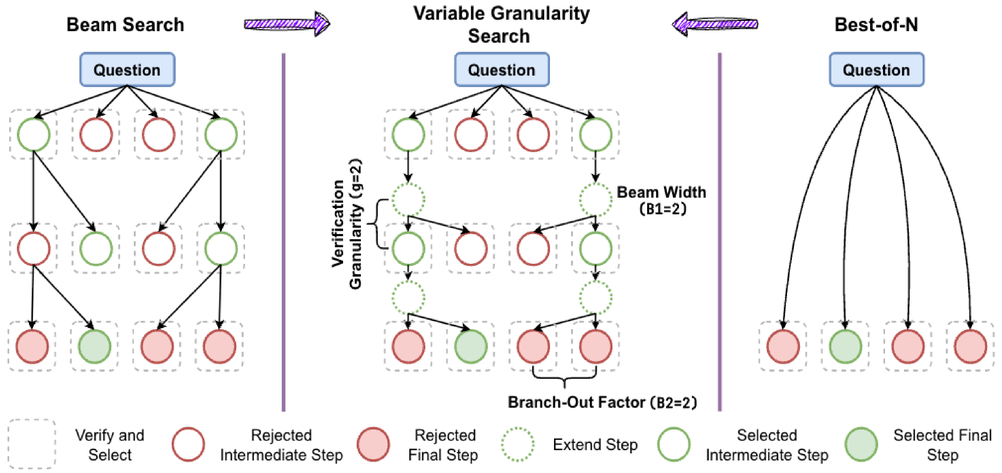


Optimal Granularity Varies with **Compute Budget** and **Problem Difficulty**

VG-Search

Unifying *Beam Search* and *Best-of-N* via Verification Granularity

Treat verification granularity as a tunable hyperparameter
Verify -> Select -> Extend -> Branch



Towards Adaptive Redefinition of the Verification Granularity

Search for Optimal Granularity based on **Compute Budget** and **Problem Difficulty** on a validation dataset

CM-g: finds the largest g that maintains accuracy

AM-g: finds the g that maximizes accuracy under a fixed compute budget

n	Metric	Adaptive *VG-Search*				Beam Search	DVTS	Best-of-N
		Oracle		Val				
		CM-g	*AM-g*	*CM-g*	*AM-g*			
64	Acc.	89.2	**90.1**	88.6	**90.1**	87.0	89.2	89.5
	FLOPs	**5393**	5844	5543	6145	12010	12010	11952
128	Acc.	90.5	**90.6**	89.9	90.1	89.3	89.7	89.3
	FLOPs	15899	16200	**11086**	12290	24021	24021	23904
256	Acc.	90.7	**90.8**	90.4	90.4	88.8	89.5	89.2
	FLOPs	**21572**	22775	28189	28189	48042	48042	47808