EESP WORKSHOP · SC26 · LIGHTNING TALK

Sample Count Is Not Enough

Candidate-Generation Strategy Shapes the Energy and Performance of LLM Test-Time Scaling

Mobina Kashaniyan and Ali Jannesari

Department of Computer Science, Iowa State University



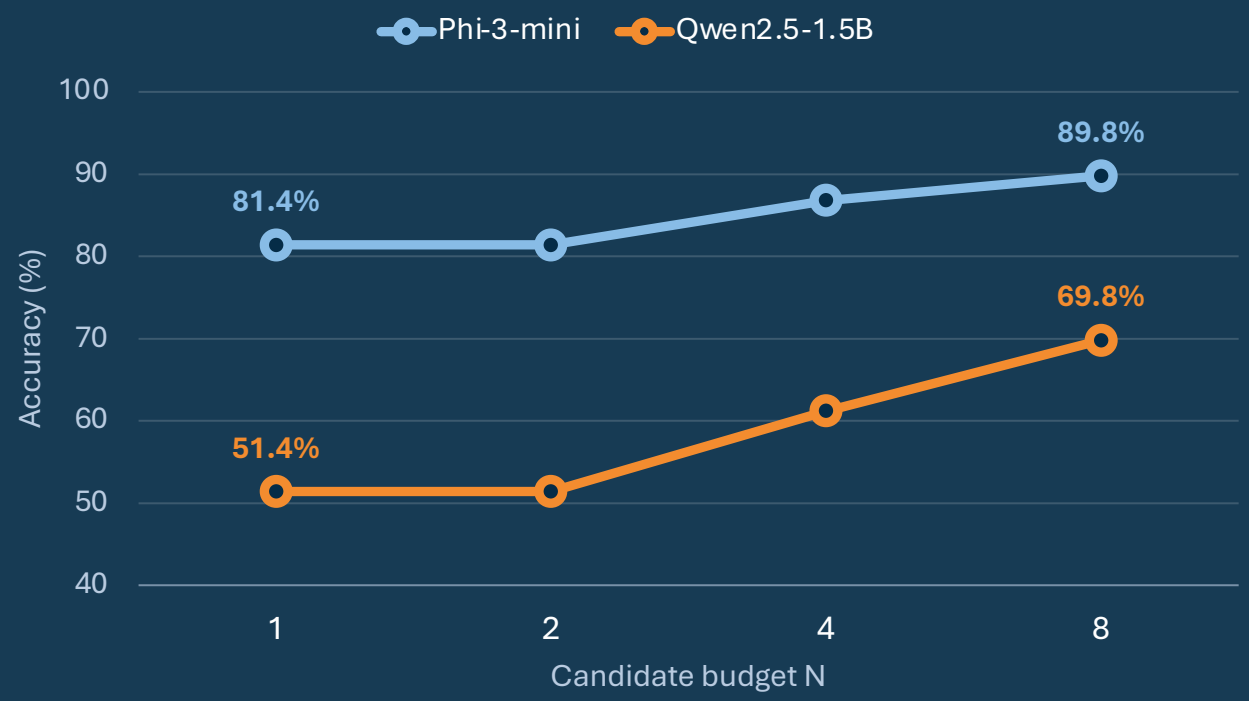
N = 8 candidates · same budget, different cost?

SC26 hpc unites | Chicago

More Candidates Improve Reasoning Accuracy

GSM8K accuracy vs. candidate budget N

Self-consistency on 500 prompts · majority vote over N sampled reasoning paths



+8.4 pp

Phi-3-mini · 81.4% → 89.8%

+18.4 pp

Qwen2.5-1.5B · 51.4% → 69.8%

Representative test-time-scaling papers don't report calls per query, candidates per call, or energy. They mostly report N.

Same N = 8, Four Ways to Execute It

Same prompt, decoding and majority vote · calls run back-to-back on one GPU · candidates within a call are batched

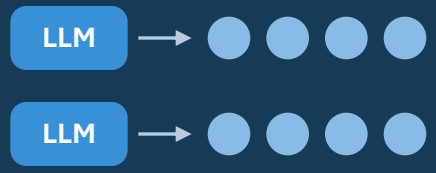
1 × 8

1 call → 8 candidates



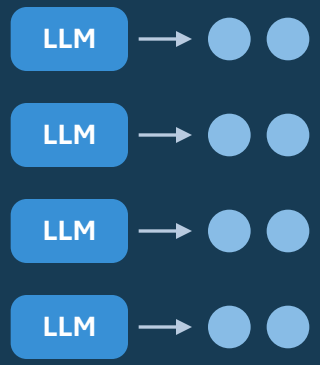
2 × 4

2 calls → 4 + 4



4 × 2

4 calls → 2 + 2 + 2 + 2



8 × 1

8 calls → 1 + … + 1



Phi-3-mini-4k-instruct
Qwen2.5-1.5B-Instruct

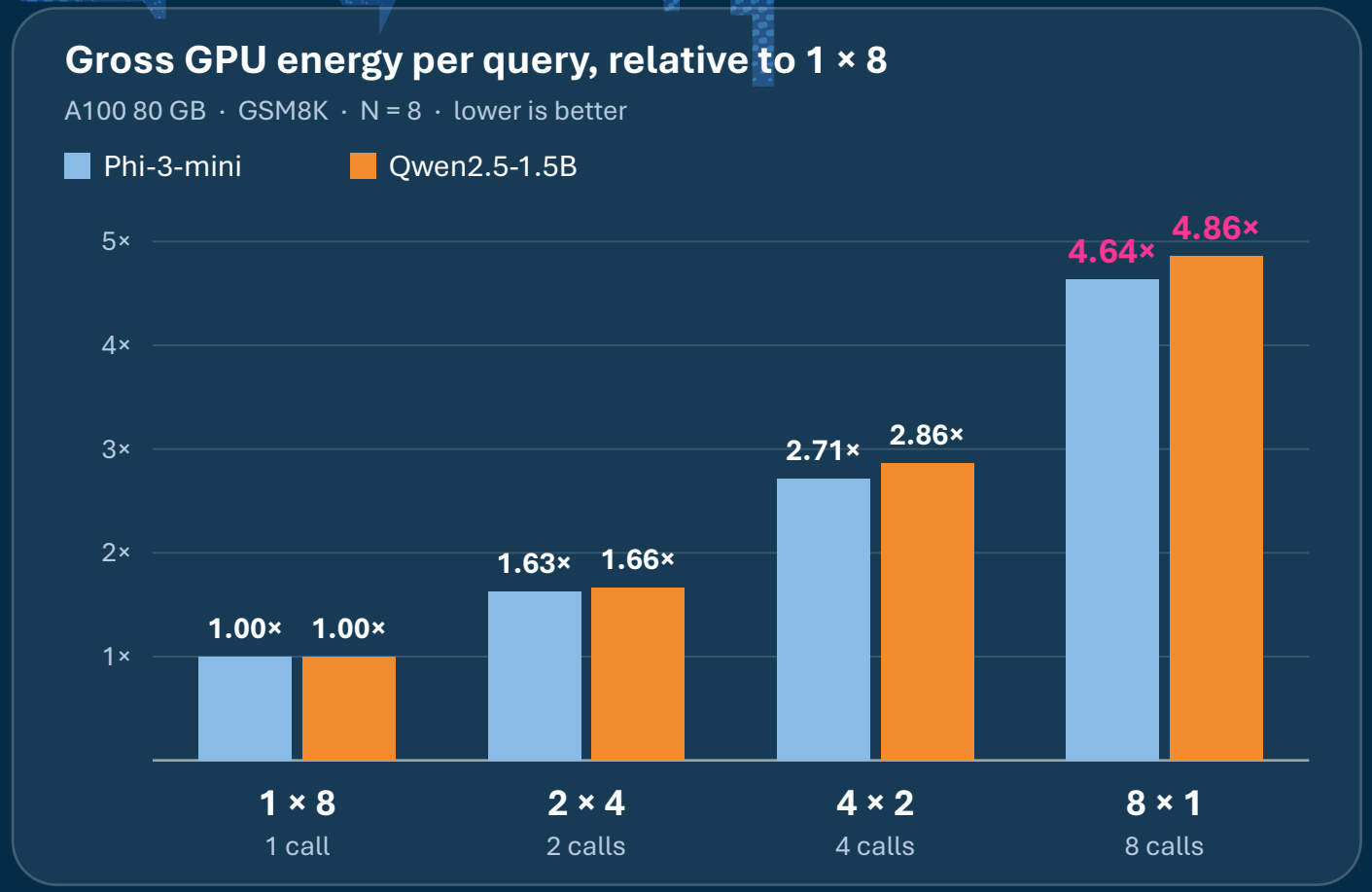
GSM8K · 100 prompts
× 3 repetitions

A100-SXM4 80 GB
500 W cap · fixed clock

NVML gross
GPU-device energy

Generated tokens
matched within ~1%

Same 8 Candidates, Up to 4.9× the Energy



5.8–6.1×

P95 latency, 8 × 1 vs. 1 × 8

17–18%

of batched throughput retained

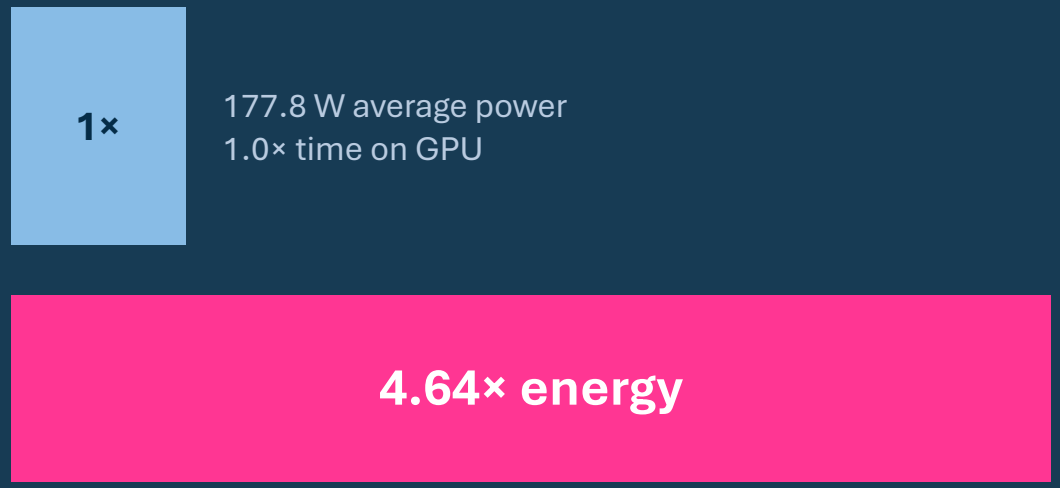
2.1 → 12.5

GPU-hours per 1,000 queries (Phi-3)

Lower Power ≠ Lower Energy, and the Effect Holds

Energy = Power × Time

Phi-3-mini on A100 · mean per query · box area = gross GPU energy



Average power	×	Time on GPU	=	Gross energy
−21%		5.97×		4.64×

The penalty holds up

✓ **3 independent A100 jobs**
4.4–4.9× energy, 8 × 1 vs. 1 × 8

✓ **All 4 response-length groups**
4.4–5.1× energy, 8 × 1 vs. 1 × 8

✓ **Short-output SciQ on V100**
2.6–3.3× energy, 8 × 1 vs. 1 × 8

Takeaways: Batch Candidates, Report the Schedule

1 For practitioners

Use as few calls as possible, with as many answers per call as GPU memory allows.

✓ **1 call × 8 answers**

✗ **8 calls × 1 answer**

when the answers are independent of each other

2 For researchers

Report more than N:

- calls per query
- candidates per call
- batching mode
- latency & throughput
- GPU-hours
- energy boundary

AT SCALE

1 million Phi-3 queries run as 8 × 1 instead of 1 × 8 would add:

+1.37 MWh

gross GPU-device energy

+10,401

GPU-hours

EESP WORKSHOP · SC26

Thank You! Questions?

Sample count is not enough: report the generation schedule, not just N.

Mobina Kashaniyan · **mobina@iastate.edu** **Ali Jannesari** · **jannesar@iastate.edu**

Department of Computer Science, Iowa State University

4.6–4.9× energy **5.8–6.1× P95 latency** **17–18% throughput kept**

SC26 hpc unites | Chicago