Paper

ACL 2026 SAN DIEGO JULY 2–7

Project Page

Guided by Gut: Efficient Test-Time Scaling with Reinforced Intrinsic Confidence

Amirhosein Ghasemabadi[1], **Keith G. Mills**[2], Baochun Li[3] and Di Niu[1]

[1]Dept. ECE, UAlberta [2]Division of CSE, LSU [3]Dept. ECE, UToronto

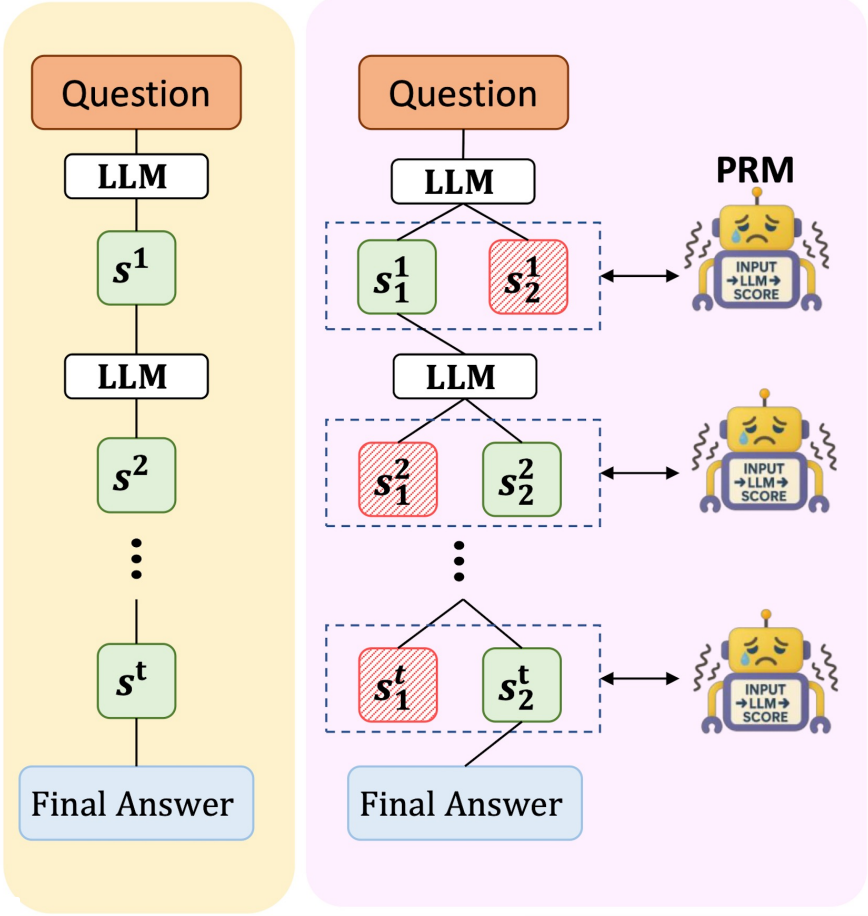
4TH OF July

LSU College of Engineering

ENGINEERING ALBERTA

The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

LLM Reasoning

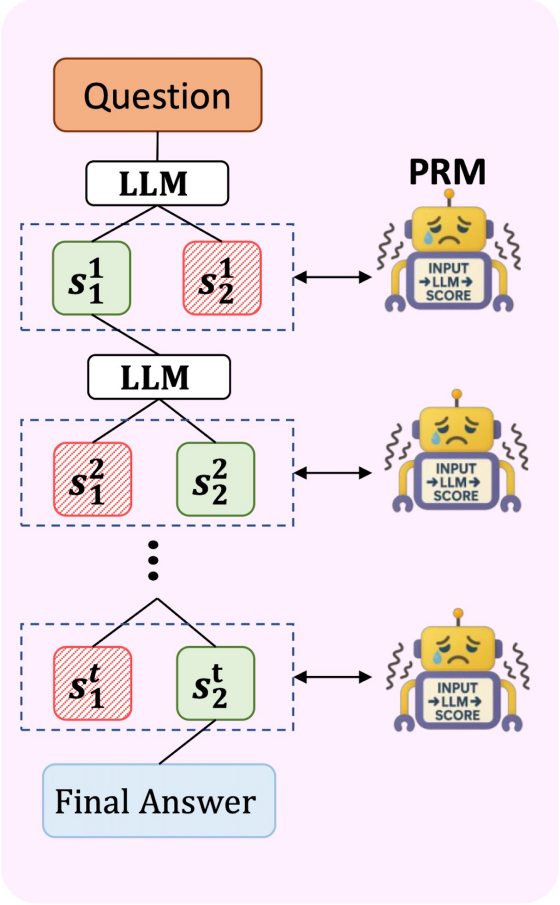


Chain-of-Thought (CoT) is good for larger models, like DeepSeek-R1-671B.

For smaller (<100B) LLMs, tree-based methods work:

- Best-of-N (BoN) sampling
- Search + Rewards

Tree-Based Reasoning



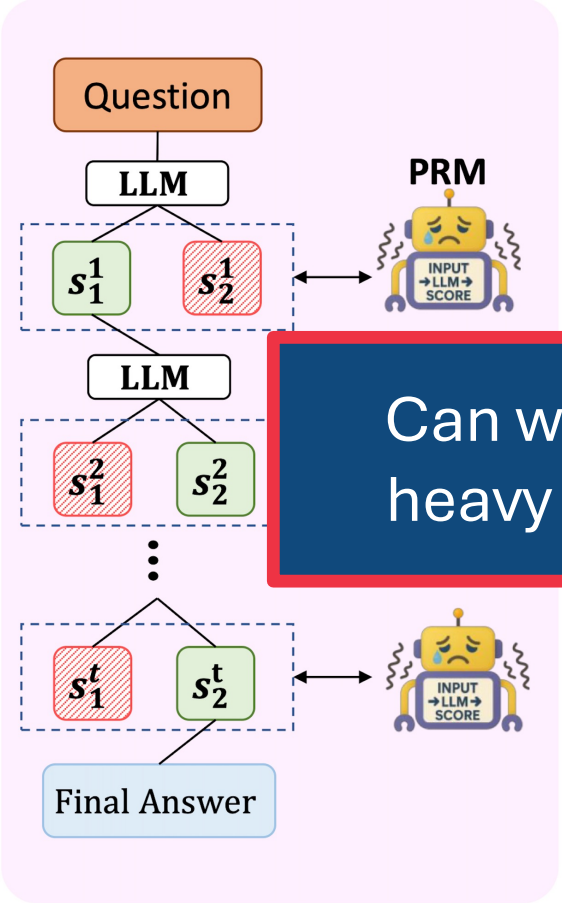
For **each** reasoning step, sample *N* possible answers.

Pick the *Best-of-N* answers.

Issue? How do we know which answer is the best? Process Reward Model (PRM)

- Usually, larger, supervisory LLM.
- Imposes memory & latency overhead.

Tree-Based Reasoning



For **each** reasoning step, sample *N* possible answers.

Can we get strong reasoning *without* heavy compute or external verifiers?

…answer is the best? Process Reward Model (PRM)

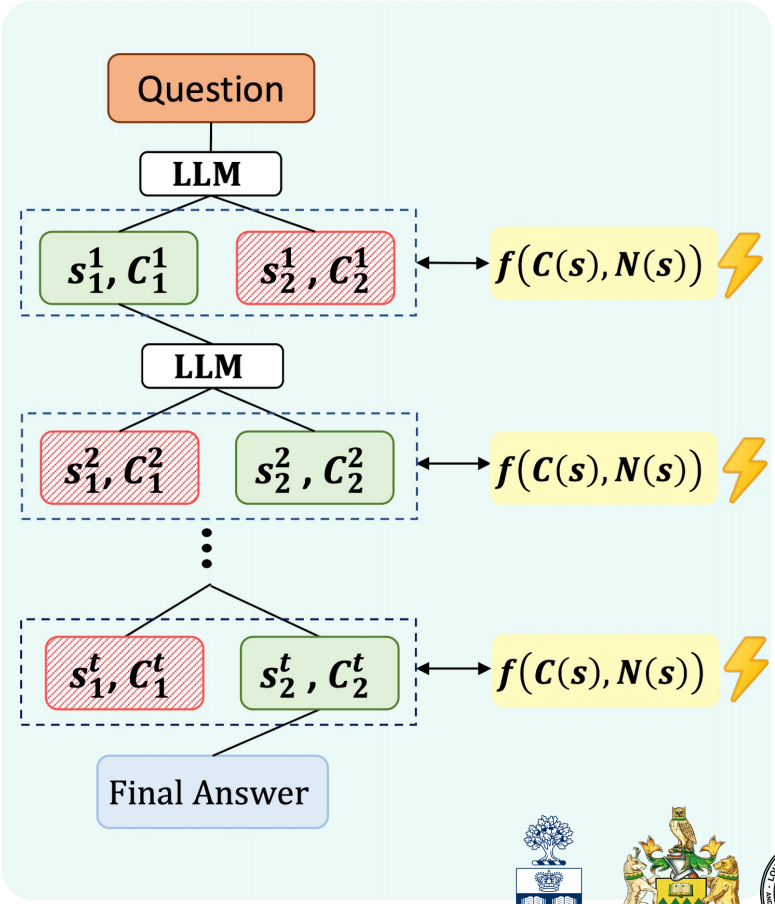
- Usually, larger, supervisory LLM.
- Imposes memory & latency overhead.

Guided by Gut: Intuition

ACL 2026 SAN DIEGO JULY 2–7

Ditch the PRM.
LLM already has an idea of how good a potential answer is.

$$p(t_i) = \prod p(t_i | t_{<i})$$



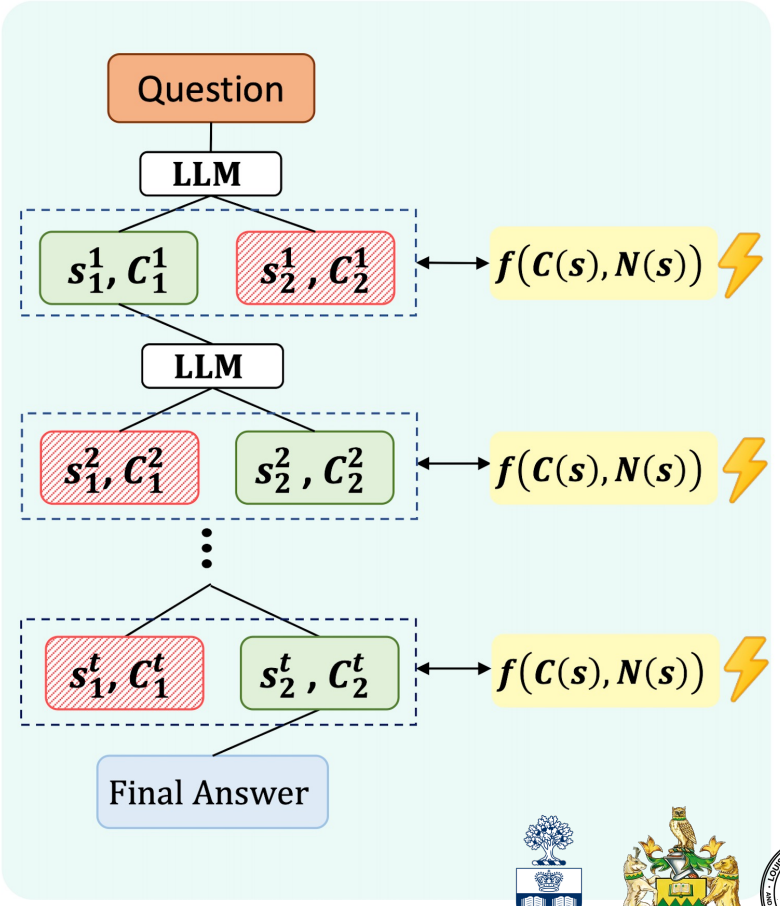
Guided by Gut: Method

Formulate reward r_t using internal statistics:

$$r_t = C(s^t) + \lambda_N N(s^t)$$

These terms?

- $C(s^t)$ - Confidence.
- $N(s^t)$ - Novelty.



Confidence and Novelty

$$r_t = C(s^t) + \lambda_N N(s^t)$$

Confidence $C(s^t)$ based on the internal **logits** for tokens selected to form reasoning step $s^t = [s_1^t, s_2^t, \ldots s_n^t]$

Novelty $N(s^t)$ proportion of **new tokens** considered relative to existing context.

Calibrate using **GRPO.** Confidence-driven r_t a reliable signal.

RLHF Calibration

Confidence != Correctness. But we can combine them.

Compute **weighted** confidence $\mathcal{C}(R_i)$ over k steps.

GRPO reward based on if the answer *is correct* or not:

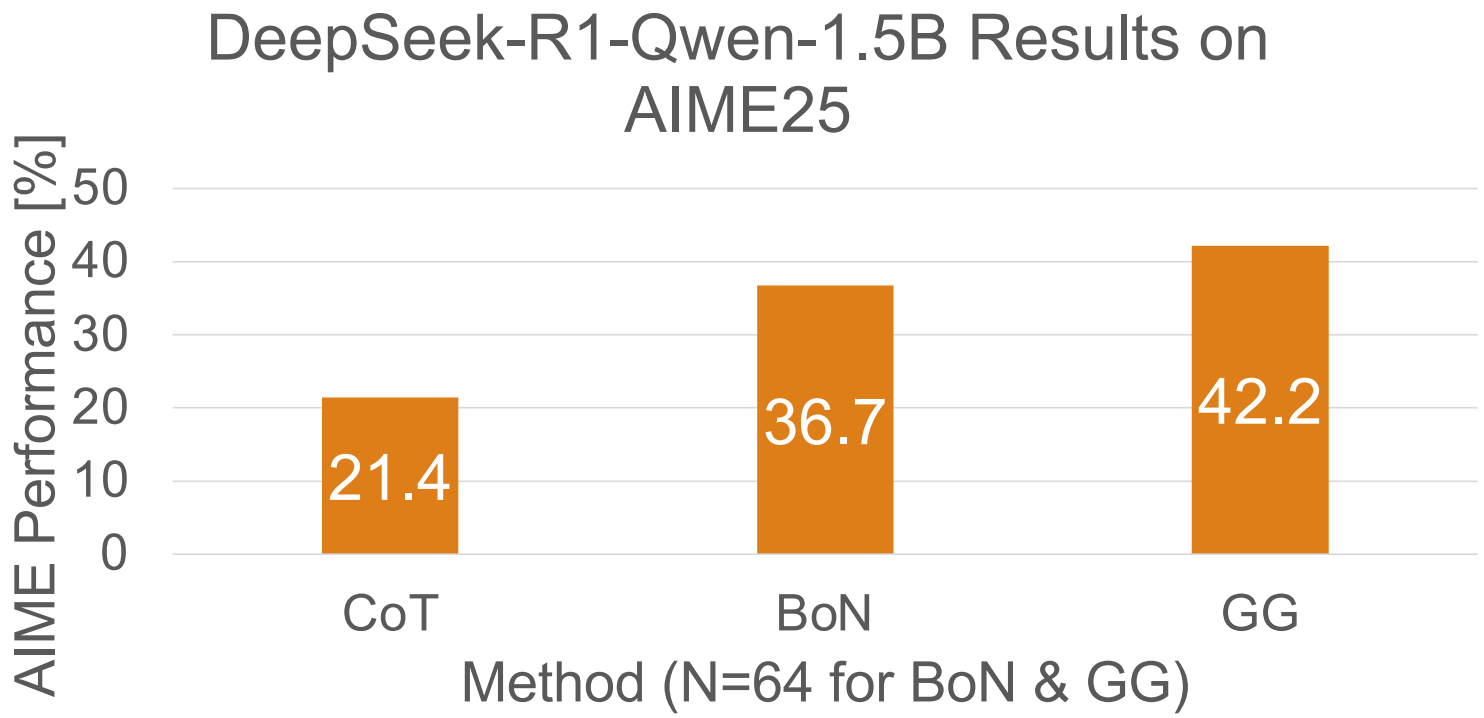
$$r_i = \begin{cases} 1 + \mathcal{C}(R_i)^4, & answer\ is\ correct \\ 1 - 10\mathcal{C}(R_i)^4, & answer\ is\ not\ correct \end{cases}$$

Experimental Baselines

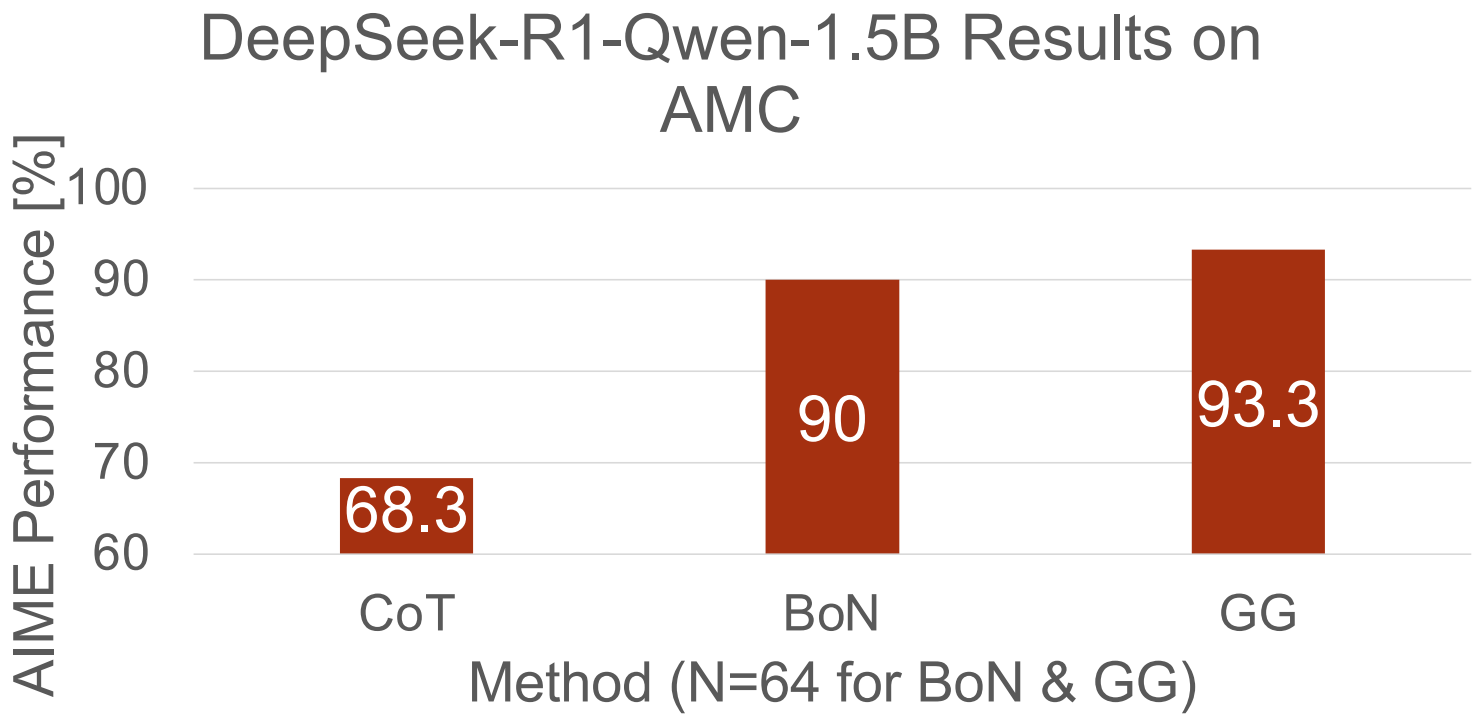
Method	Key Idea
Chain-of-Thought (CoT)	Single Reasoning Path
PRM Search	External Verifier
Guided by Gut	Internal Confidence-Guided Search

Benchmarks: AIME24/25, MATH 500, AMC

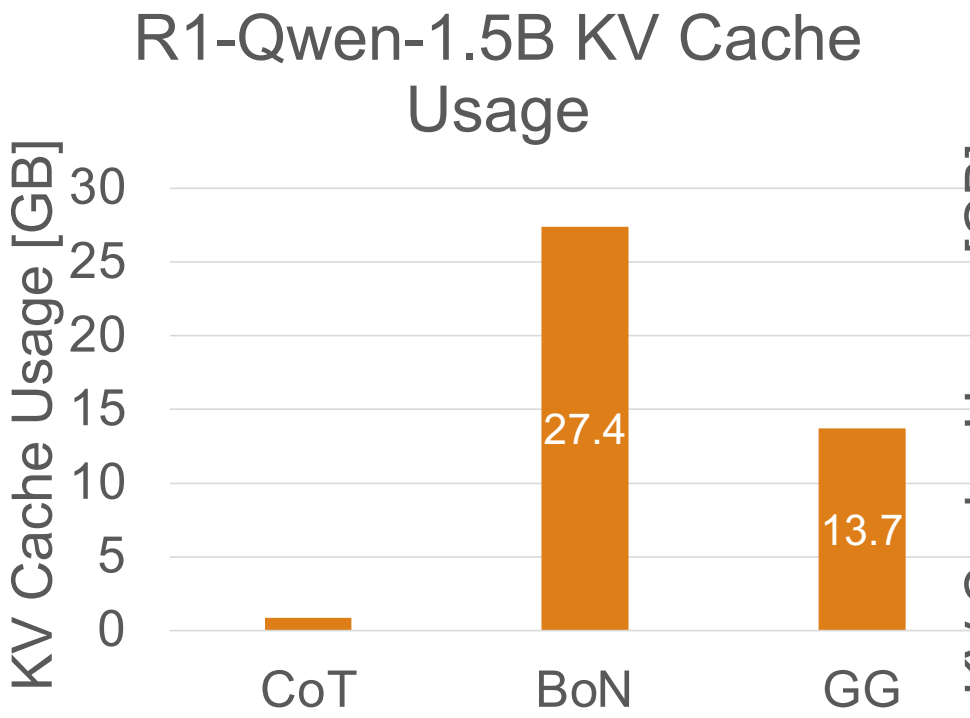
DeepSeek-Qwen-1.5B Results

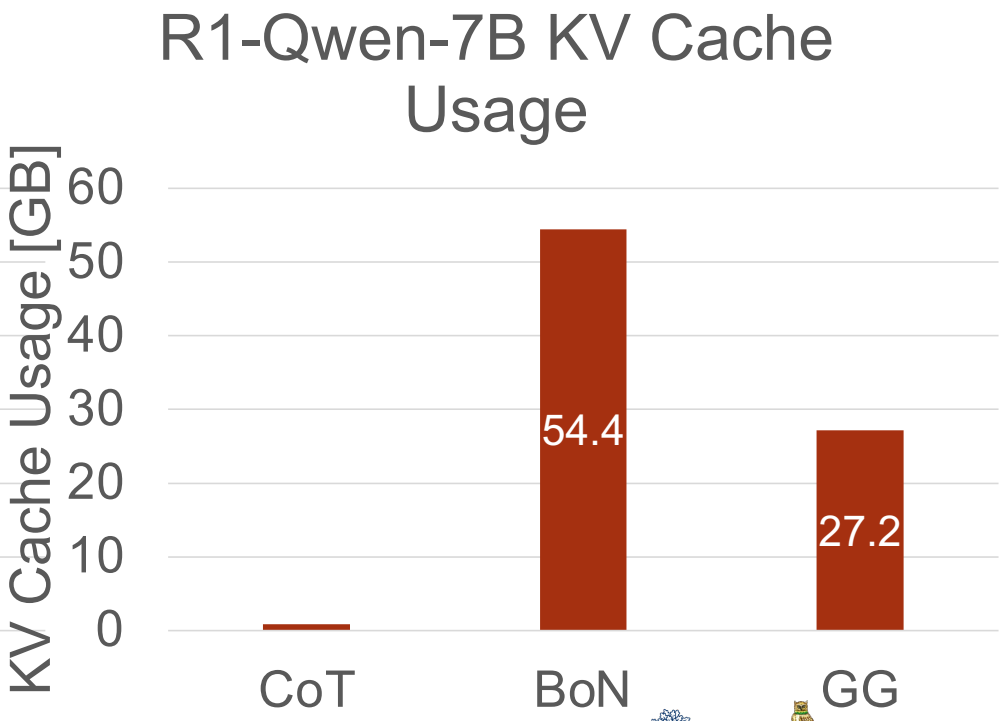


DeepSeek-Qwen-1.5B Results

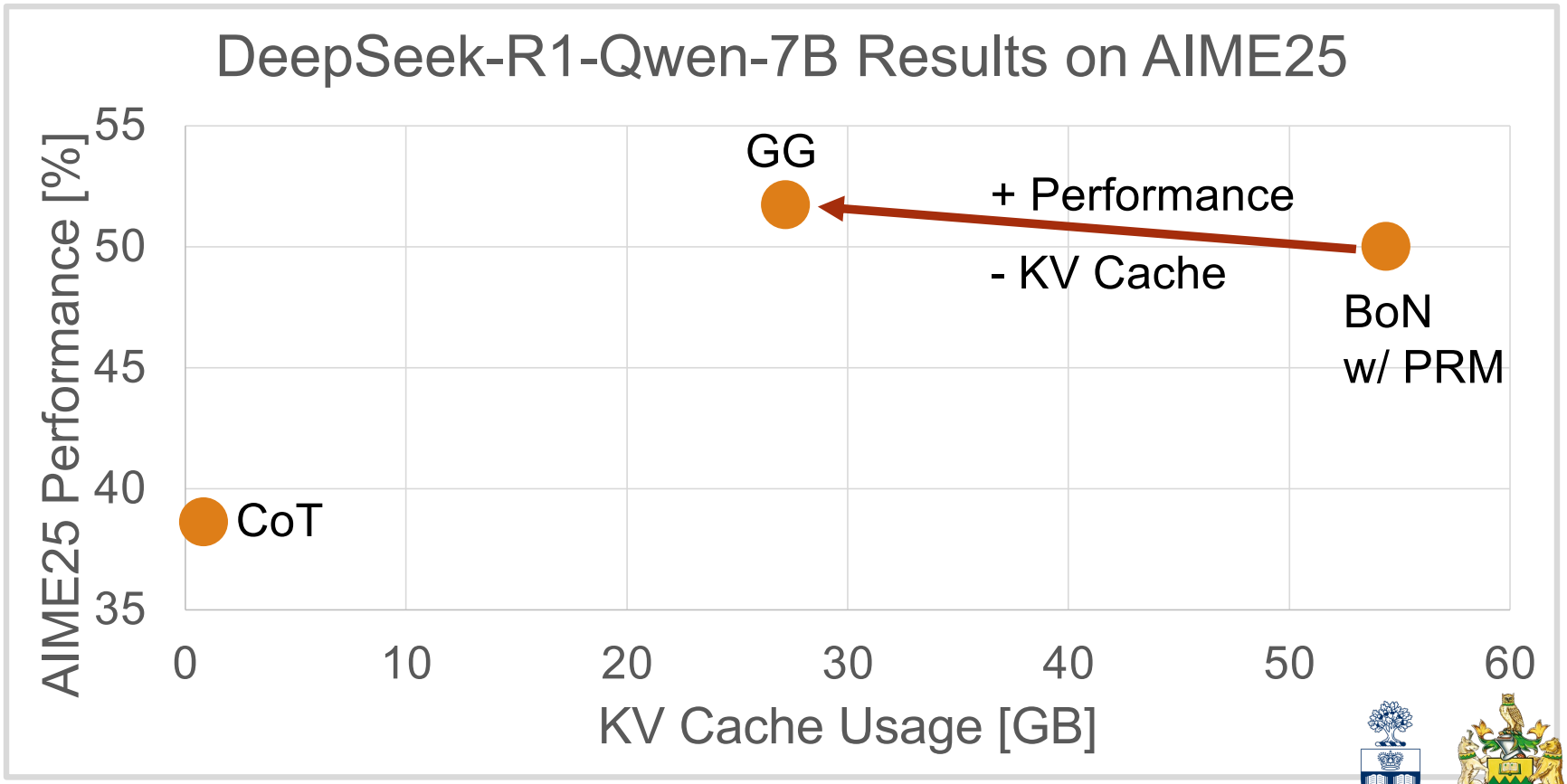


KV Cache Savings





Overall Look



Conclusion

ACL 2026 SAN DIEGO JULY 2–7

Don’t need larger models or heavy inference.

GG: Better reasoning from smarter search.

Paper

Project Page