ICLR

Scaling Up, Speeding Up: A Benchmark of Speculative Decoding for Efficient LLM Test-Time Scaling

Shengyin Sun[1*], Yiming Li[2*], Xing Li[2], Yingzhao Lian[2], Weizhe Lin[2], Hui-Ling Zhen[2],

Zhiyuan Yang[2], Chen Chen[2], Xianzhi Yu[2], Mingxuan Yuan[2], Chen Ma[1†]

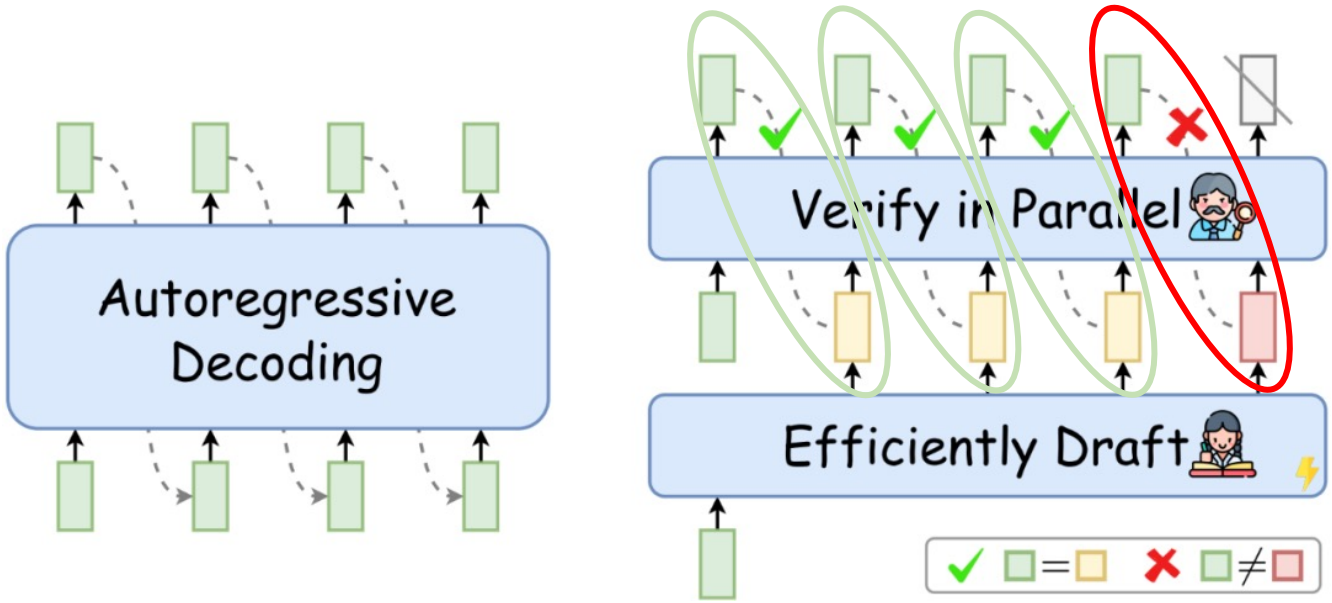
[1]City University of Hong Kong, [2]Huawei Technologies

香港城市大學
City University of Hong Kong

HUAWEI

Background [1/2] — Speculative Decoding

- Benefits of Speculative Decoding
  - ➢ Make full use of the parallel verification capabilities of LLMs.
  - ➢ Make full use of computing resources (LLM inference is largely not compute-bound).
  - ➢ Lossless!

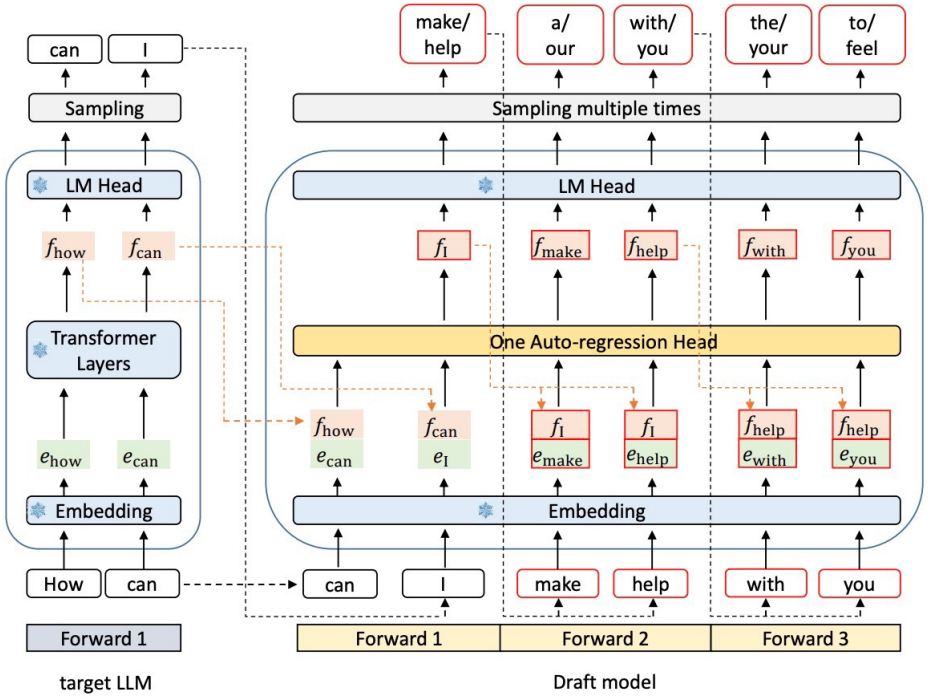


- ☐ The draft model's forward pass takes **much less** time than the target model's.
- ☐ The target model generates **4 tokens per forward pass**.

[1] Figure adapted from Figure 1 in Xia et al., 2024, Spec-Bench: A Comprehensive Benchmark and Unified Evaluation Platform for Speculative Decoding.

Background [2/2] — Related Work

- Training-based Speculative Decoding
  - ➢ Eagle-1 [1], Eagle-2 [2], Eagle-3 [3].



TL;DR:

- ☐ Train one auto-regression head.
  - ☐ Input: Token embedding.
  - ☐ Output: Hidden state of previous token.

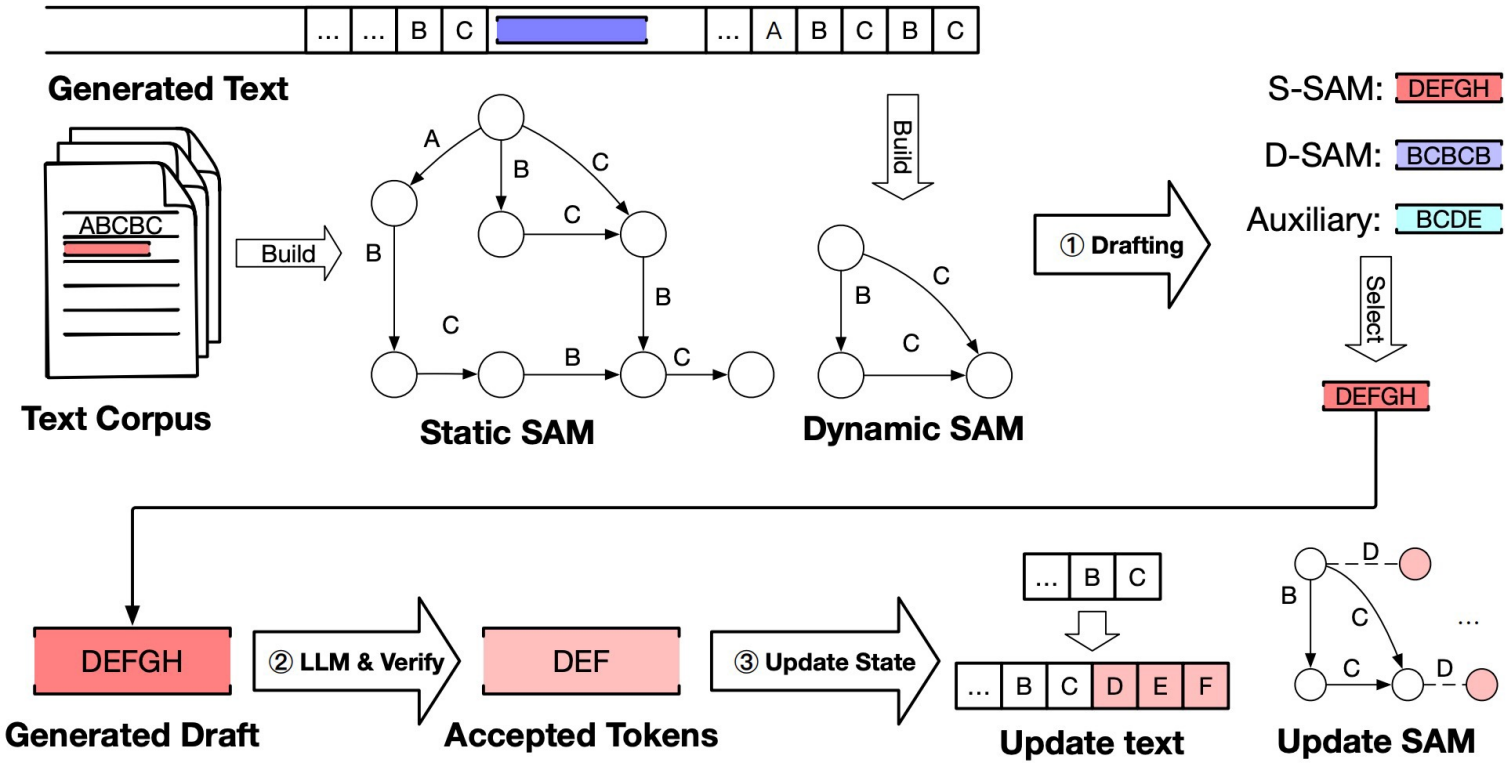
[1] Li et al., EAGLE: Speculative Sampling Requires Rethinking Feature Uncertainty (2024).

[2] Li et al., EAGLE-2: Faster Inference of Language Models with Dynamic Draft Trees (2024).

[3] Li et al., EAGLE-3: Scaling up Inference Acceleration of Large Language Models via Training-Time Test (2025).

Background [2/2] — Related Work

- Ngram-based Speculative Decoding
  - SAM [1].



TL;DR:

- Reuse historical information.
  - statistical regularities often indicate **fixed patterns: e.g., "Let me think step by step".**

[1] Hu et al., SAM Decoding: Speculative Decoding via Suffix Automaton (2024).

Motivation [1/1] — Redundancy

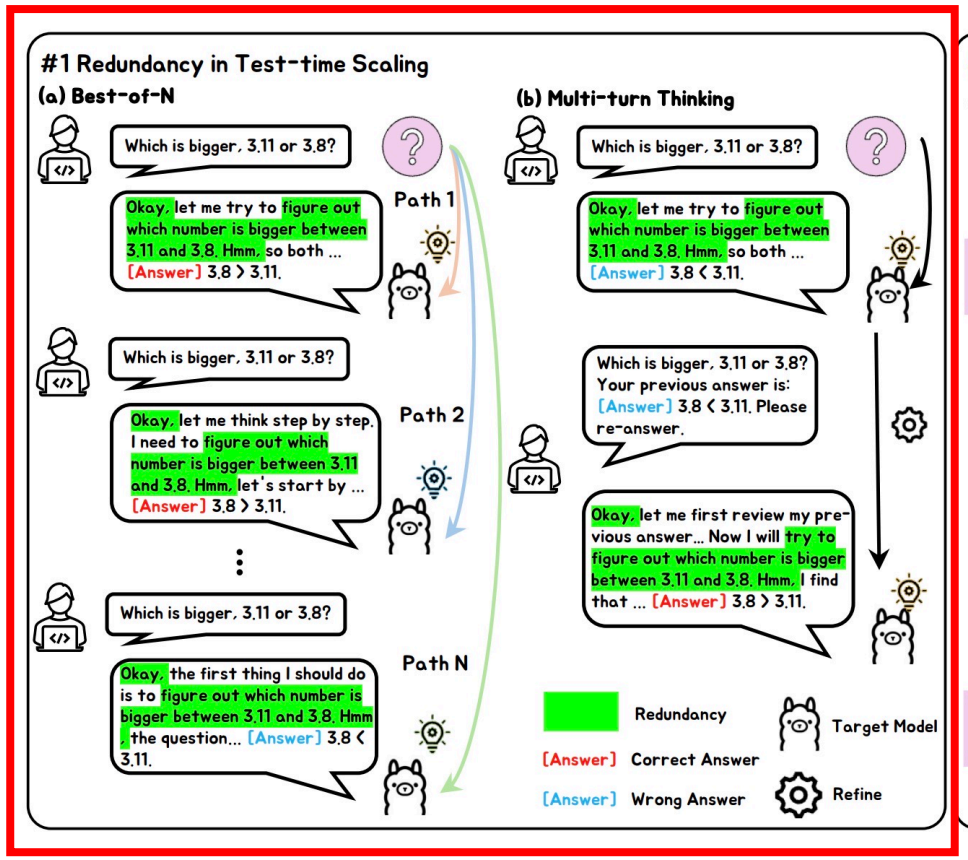
- Redundancy under the Test-time Scaling Paradigm
  - Most existing work focuses on evaluating short generations (e.g., 1024 tokens).
  - How does speculative decoding perform in reasoning models?

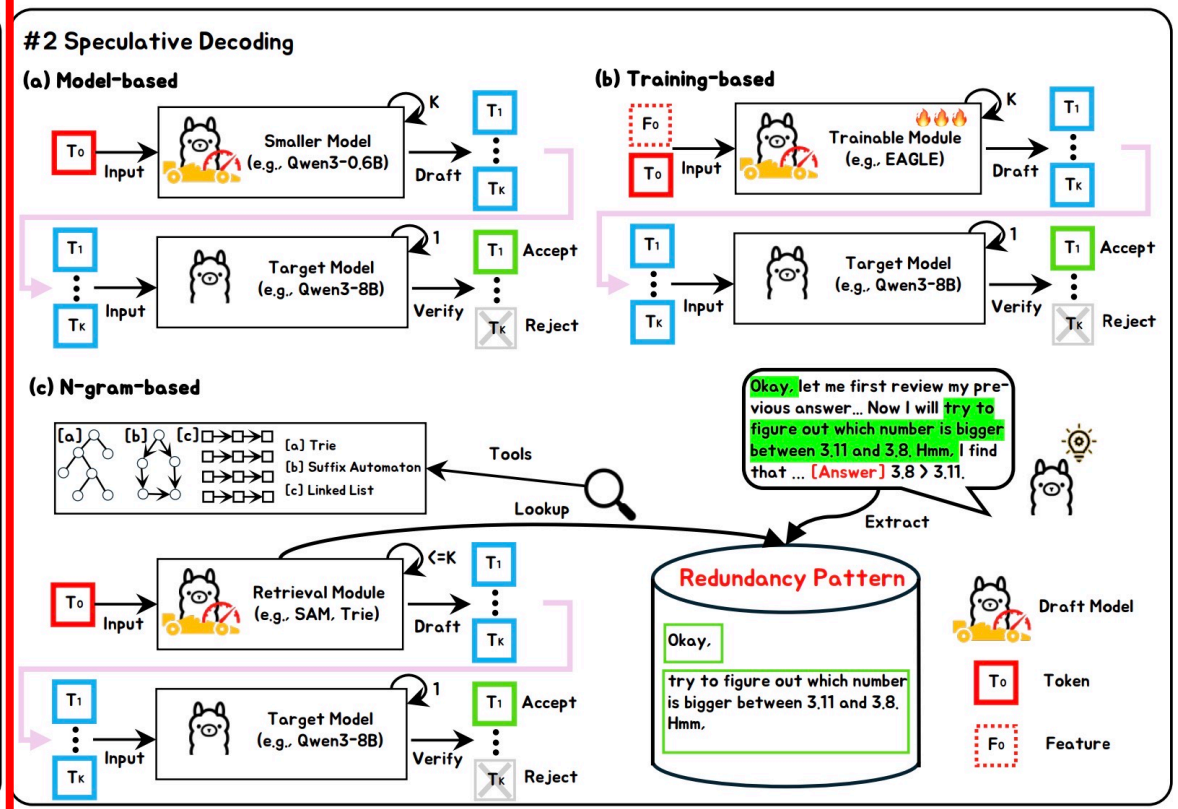
Question	Let p be the least prime number for which there exists a positive integer n such that $n^4 + 1$ is divisible by p^2. Find the least positive integer m such that $m^4 + 1$ is divisible by p^2.
Intra-turn Redundancy	**Turn 1:** Compute $8^3 = 512$. Then 512 mod 17: since $17 \times 30 = 510$, we have $512 \equiv 2 \pmod{17}$. Therefore, $4 \times 8^3 \equiv 4 \times 2 = 8 \pmod{17}$. Thus, $4 \times 8^3 \times 17k \equiv 8 \times 17k \pmod{289}$. So, $n^4 \equiv 50 + 8 \times 17k \pmod{289}$. Hence, $n^4 + 1 \equiv 51 + 136k \pmod{289}$. We need this to be 0 (mod 289), so $51 + 136k \equiv 0 \pmod{289}$. Dividing both sides by 17, we get $3 + 8k \equiv 0 \pmod{17}$, i.e., $\cdots$ Wait, earlier I said that $4 \times 8^3 \times 17k \equiv 8 \times 17k$ mod 289. Let me check that again. 4×8^3. Compute $8^3 = 512$. Then 512 mod 17: since $17 \times 30 = 510$, we have $512 \equiv 2 \pmod{17}$.
Inter-turn Redundancy	**Turn 1:** <think>\n Okay, so I need to find the least prime number p such that there's a positive integer n where p^2 divides $n^4 + 1$. Then, once I find that p, I have to find the smallest positive integer m such that p^2 divides $m^4 + 1$. Alright, let me start by understanding the problem $\cdots$ **Turn 2:** <think>\n Okay, so I need to find the least prime p such that there's a positive integer n with p^2 dividing $n^4 + 1$. Then, once I find that p, I need to find the smallest m such that $m^4 + 1$ is divisible by p^2. \n\nFirst, let me recall that$\cdots$

- A large amount of repetition!
- Which is superior: **statistically driven** methods or **data-driven** methods?

Benchmark [1/1] — Framework

- Accelerating the Reasoning Model under Test-time Scaling
  - 2 mainstream test-time scaling frameworks (BoN and Multi-round Thinking).
  - 2 categories of 4 reasoning models (Deepseek-R1-Distill-Llama-8B, Qwen3-4B/8B/14B).

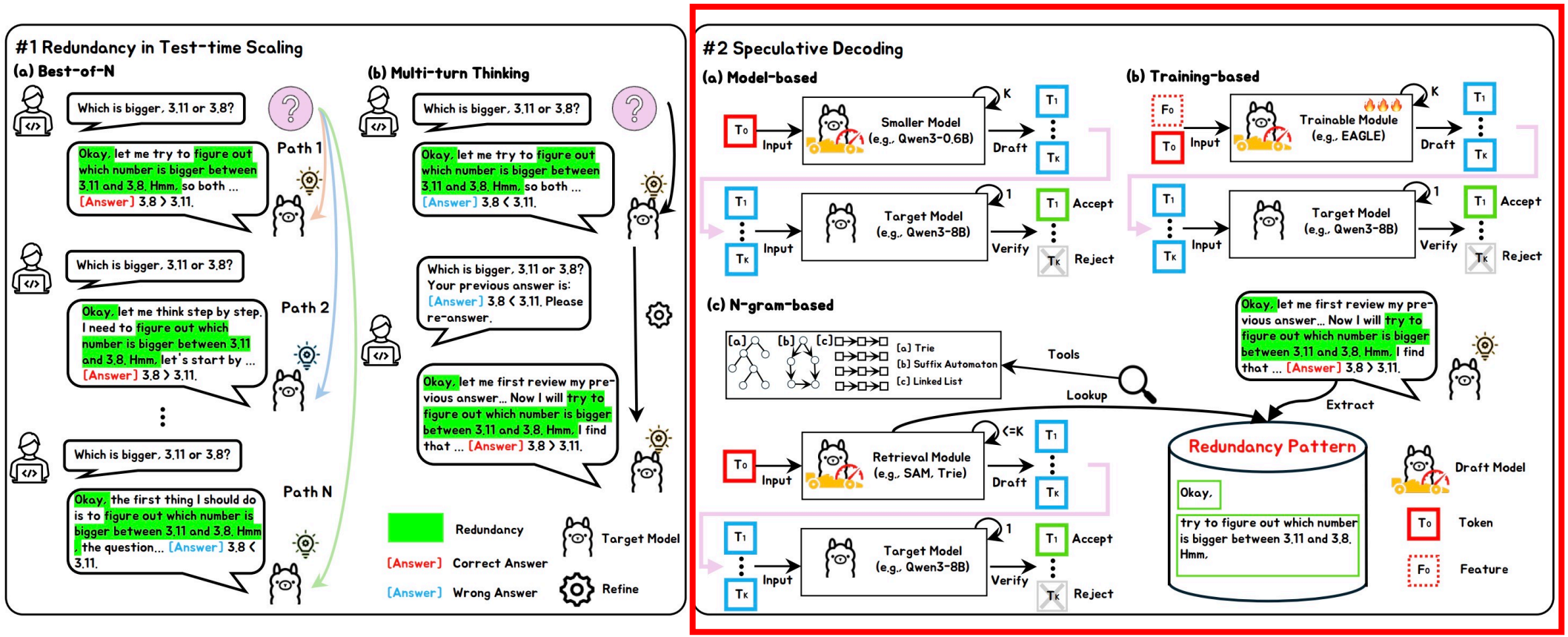




Selective Optimization & Iterative Optimization

Benchmark [1/1] — Framework

- Accelerating the Reasoning Model under Test-time Scaling
  - 4 Categories and 9 Types of Methods: (a) Training-based: Eagle-3; (b) Model-based: SpS; (c) N-gram-based: PLD, REST, Lookahead, PIA, SAM, Recycling; (d) **Hybrid: Eagle-3 [SAM].**



Experiments [1/3] — Analysis

- Lightweight Ngram-based Methods Perform Satisfactorily

Table 3: Performance comparison of speculative decoding methods for reasoning models under multi-round thinking framework with different temperature T (**Best** , Second Best).

Model	Bench	AIME24		AIME25		MATH500		GPQA		Overall	
	Method	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed
DSL-8B ($T=0$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	2.26	1.56×	2.24	1.53×	2.63	2.86×	2.46	1.61×	2.35	1.93×
	SAM[EAGLE-3]	**3.91**	**3.09**×	**4.50**	**3.85**×	**4.94**	**4.11**×	**7.00**	**4.70**×	**4.72**	**3.97**×
	SAM	2.64	2.41×	3.05	2.96×	2.60	2.14×	3.57	3.20×	2.93	2.66×
	Recycling	2.98	2.08×	2.98	2.05×	2.97	2.18×	3.02	2.08×	2.99	2.10×
	PLD	2.14	1.76×	2.38	1.89×	2.25	1.71×	2.65	2.02×	2.33	1.84×
	REST	1.33	1.00×	1.35	0.95×	1.37	1.08×	1.31	1.00×	1.34	1.01×
	Lookahead	2.21	1.59×	2.26	1.53×	2.28	1.55×	2.44	1.62×	2.28	1.57×
	PIA	1.84	1.51×	1.96	1.63×	1.91	1.48×	2.15	1.63×	1.95	1.56×
DSL-8B ($T=0.6$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	2.21	1.31×	2.17	1.50×	3.21	3.02×	2.45	1.61×	2.33	1.91×
	SAM[EAGLE-3]	2.54	1.82×	2.48	1.89×	**3.28**	**3.13**×	**2.86**	**2.15**×	2.66	**2.29**×
	SAM	1.84	1.65×	1.84	1.65×	1.93	1.71×	1.93	1.73×	1.87	1.69×
	Recycling	**2.84**	**1.91**×	**2.83**	**1.93**×	2.88	2.03×	2.81	1.98×	**2.84**	1.96×
	REST	1.35	1.00×	1.37	0.97×	1.38	1.03×	1.35	1.00×	1.36	1.00×
	PIA	1.58	1.26×	1.58	1.27×	1.63	1.39×	1.68	1.36×	1.61	1.33×
QW3-8B ($T=0$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	4.31	2.80×	4.37	2.87×	4.50	3.04×	4.39	2.91×	4.38	2.91×
	SAM[EAGLE-3]	4.40	**3.08**×	4.47	**3.18**×	4.82	**3.47**×	5.90	**4.17**×	4.76	**3.49**×
	SAM	2.15	1.97×	2.27	2.07×	2.18	1.95×	3.19	3.11×	2.37	2.28×
	Recycling	3.02	2.08×	2.98	2.13×	3.00	2.20×	3.07	2.17×	3.01	2.15×
	PLD	1.95	1.61×	1.97	1.66×	1.91	1.63×	2.49	2.04×	2.05	1.74×
	SpS	**7.40**	0.93×	**7.50**	0.96×	**6.69**	0.77×	**6.33**	0.83×	**7.07**	0.87×
	REST	1.39	1.12×	1.39	1.14×	1.42	1.16×	1.33	1.09×	1.38	1.13×
	Lookahead	2.14	1.54×	2.12	1.54×	2.10	1.58×	2.28	1.69×	2.15	1.59×
	PIA	1.83	1.48×	1.96	1.68×	1.89	1.63×	2.08	1.72×	1.93	1.63×
QW3-8B ($T=0.6$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	4.11	2.61×	4.21	2.71×	4.32	2.87×	4.03	2.68×	4.16	**2.73**×
	SAM[EAGLE-3]	3.92	**2.68**×	3.93	**2.72**×	4.19	**2.93**×	3.98	**2.82**×	3.97	**2.79**×
	SAM	1.91	1.73×	1.95	1.78×	1.93	1.74×	2.09	1.86×	1.96	1.78×
	Recycling	2.88	2.00×	2.92	2.04×	2.90	2.11×	2.92	2.08×	2.90	2.06×
	SpS	**6.22**	0.91×	**6.54**	0.95×	**6.17**	0.80×	**6.32**	0.86×	**6.34**	0.88×
	REST	1.41	1.08×	1.41	1.10×	1.42	1.12×	1.36	1.03×	1.40	1.08×
	PIA	1.69	1.40×	1.70	1.45×	1.71	1.45×	1.75	1.48×	1.71	1.44×

Takeaway 1: Training-free n-gram methods can outperform training-based approaches!

- The new hybrid solution SAM[EAGLE-3] offers **the best speedup** in all scenarios. For example, the speedup ratio on Deepseek-R1 (T=0) is 3.97x, achieving a 49.2% improvement compared to the second-best method.
- Training-based methods are sensitive to the training process For example, EAGLE-3 for DeepSeek R1 performs poorly (for generating long sequences).

Experiments [1/3] — Analysis

- Lightweight Ngram-based Methods Perform Satisfactorily

Table 3: Performance comparison of speculative decoding methods for reasoning models under multi-round thinking framework with different temperature T (**Best** , Second Best).

Model	Bench	AIME24		AIME25		MATH500		GPQA		Overall	
	Method	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed
DSL-8B ($T=0$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	2.26	1.56×	2.24	1.53×	2.63	2.86×	2.46	1.61×	2.35	1.93×
	SAM[EAGLE-3]	**3.91**	**3.09**×	**4.50**	**3.85**×	**4.94**	**4.11**×	**7.00**	**4.70**×	**4.72**	**3.97**×
	SAM	2.64	2.41×	3.05	2.96×	2.60	2.14×	3.57	3.20×	2.93	2.66×
	Recycling	2.98	2.08×	2.98	2.05×	2.97	2.18×	3.02	2.08×	2.99	2.10×
	PLD	2.14	1.76×	2.38	1.89×	2.25	1.71×	2.65	2.02×	2.33	1.84×
	REST	1.33	1.00×	1.35	0.95×	1.37	1.08×	1.31	1.00×	1.34	1.01×
	Lookahead	2.21	1.59×	2.26	1.53×	2.28	1.55×	2.44	1.62×	2.28	1.57×
	PIA	1.84	1.51×	1.96	1.63×	1.91	1.48×	2.15	1.63×	1.95	1.56×
DSL-8B ($T=0.6$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	2.21	1.31×	2.17	1.50×	3.21	3.02×	2.45	1.61×	2.33	1.91×
	SAM[EAGLE-3]	2.54	1.82×	2.48	1.89×	**3.28**	**3.13**×	**2.86**	**2.15**×	2.66	**2.29**×
	SAM	1.84	1.65×	1.84	1.65×	1.93	1.71×	1.93	1.73×	1.87	1.69×
	Recycling	**2.84**	**1.91**×	**2.83**	**1.93**×	2.88	2.03×	2.81	1.98×	**2.84**	1.96×
	REST	1.35	1.00×	1.37	0.97×	1.38	1.03×	1.35	1.00×	1.36	1.00×
	PIA	1.58	1.26×	1.58	1.27×	1.63	1.39×	1.68	1.36×	1.61	1.33×
QW3-8B ($T=0$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	4.31	2.80×	4.37	2.87×	4.50	3.04×	4.39	2.91×	4.38	2.91×
	SAM[EAGLE-3]	4.40	**3.08**×	4.47	**3.18**×	4.82	**3.47**×	5.90	**4.17**×	4.76	**3.49**×
	SAM	2.15	1.97×	2.27	2.07×	2.18	1.95×	3.19	3.11×	2.37	2.28×
	Recycling	3.02	2.08×	2.98	2.13×	3.00	2.20×	3.07	2.17×	3.01	2.15×
	PLD	1.95	1.61×	1.97	1.66×	1.91	1.63×	2.49	2.04×	2.05	1.74×
	SpS	**7.40**	0.93×	**7.50**	0.96×	**6.69**	0.77×	**6.33**	0.83×	**7.07**	0.87×
	REST	1.39	1.12×	1.39	1.14×	1.42	1.16×	1.33	1.09×	1.38	1.13×
	Lookahead	2.14	1.54×	2.12	1.54×	2.10	1.58×	2.28	1.69×	2.15	1.59×
	PIA	1.83	1.48×	1.96	1.68×	1.89	1.63×	2.08	1.72×	1.93	1.63×
QW3-8B ($T=0.6$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	4.11	2.61×	4.21	2.71×	4.32	2.87×	4.03	2.68×	4.16	**2.73**×
	SAM[EAGLE-3]	3.92	**2.68**×	3.93	**2.72**×	4.19	**2.93**×	3.98	**2.82**×	3.97	**2.79**×
	SAM	1.91	1.73×	1.95	1.78×	1.93	1.74×	2.09	1.86×	1.96	1.78×
	Recycling	2.88	2.00×	2.92	2.04×	2.90	2.11×	2.92	2.08×	2.90	2.06×
	SpS	**6.22**	0.91×	**6.54**	0.95×	**6.17**	0.80×	**6.32**	0.86×	**6.34**	0.88×
	REST	1.41	1.08×	1.41	1.10×	1.42	1.12×	1.36	1.03×	1.40	1.08×
	PIA	1.69	1.40×	1.70	1.45×	1.71	1.45×	1.75	1.48×	1.71	1.44×

- **Takeaway 1: Training-free n-gram methods can outperform training-based approaches!**
- **Takeaway 2: This hybrid approach demonstrates a clear 1 + 1 > 2 effect !**
- Training-based methods are sensitive to the training process For example, EAGLE-3 for DeepSeek R1 performs poorly (for generating long sequences).

Experiments [1/3] — Analysis

- Lightweight Ngram-based Methods Perform Satisfactorily

Table 3: Performance comparison of speculative decoding methods for reasoning models under multi-round thinking framework with different temperature T (**Best** , Second Best).

Model	Bench	AIME24		AIME25		MATH500		GPQA		Overall	
	Method	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed
DSL-8B ($T=0$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	2.26	1.56×	2.24	1.53×	2.63	2.86×	2.46	1.61×	2.35	1.93×
	SAM[EAGLE-3]	**3.91**	**3.09**×	**4.50**	**3.85**×	**4.94**	**4.11**×	**7.00**	**4.70**×	**4.72**	**3.97**×
	SAM	2.64	2.41×	3.05	2.96×	2.60	2.14×	3.57	3.20×	2.93	2.66×
	Recycling	2.98	2.08×	2.98	2.05×	2.97	2.18×	3.02	2.08×	2.99	2.10×
	PLD	2.14	1.76×	2.38	1.89×	2.25	1.71×	2.65	2.02×	2.33	1.84×
	REST	1.33	1.00×	1.35	0.95×	1.37	1.08×	1.31	1.00×	1.34	1.01×
	Lookahead	2.21	1.59×	2.26	1.53×	2.28	1.55×	2.44	1.62×	2.28	1.57×
	PIA	1.84	1.51×	1.96	1.63×	1.91	1.48×	2.15	1.63×	1.95	1.56×
DSL-8B ($T=0.6$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	2.21	1.31×	2.17	1.50×	3.21	3.02×	2.45	1.61×	2.33	1.91×
	SAM[EAGLE-3]	2.54	1.82×	2.48	1.89×	**3.28**	**3.13**×	**2.86**	**2.15**×	2.66	**2.29**×
	SAM	1.84	1.65×	1.84	1.65×	1.93	1.71×	1.93	1.73×	1.87	1.69×
	Recycling	**2.84**	**1.91**×	**2.83**	**1.93**×	2.88	2.03×	2.81	1.98×	**2.84**	1.96×
	REST	1.35	1.00×	1.37	0.97×	1.38	1.03×	1.35	1.00×	1.36	1.00×
	PIA	1.58	1.26×	1.58	1.27×	1.63	1.39×	1.68	1.36×	1.61	1.33×
QW3-8B ($T=0$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	4.31	2.80×	4.37	2.87×	4.50	3.04×	4.39	2.91×	4.38	2.91×
	SAM[EAGLE-3]	4.40	**3.08**×	4.47	**3.18**×	4.82	**3.47**×	5.90	**4.17**×	4.76	**3.49**×
	SAM	2.15	1.97×	2.27	2.07×	2.18	1.95×	3.19	3.11×	2.37	2.28×
	Recycling	3.02	2.08×	2.98	2.13×	3.00	2.20×	3.07	2.17×	3.01	2.15×
	PLD	1.95	1.61×	1.97	1.66×	1.91	1.63×	2.49	2.04×	2.05	1.74×
	SpS	**7.40**	0.93×	**7.50**	0.96×	**6.69**	0.77×	**6.33**	0.83×	**7.07**	0.87×
	REST	1.39	1.12×	1.39	1.14×	1.42	1.16×	1.33	1.09×	1.38	1.13×
	Lookahead	2.14	1.54×	2.12	1.54×	2.10	1.58×	2.28	1.69×	2.15	1.59×
	PIA	1.83	1.48×	1.96	1.68×	1.89	1.63×	2.08	1.72×	1.93	1.63×
QW3-8B ($T=0.6$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	4.11	2.61×	4.21	2.71×	4.32	2.87×	4.03	2.68×	4.16	**2.73**×
	SAM[EAGLE-3]	3.92	**2.68**×	3.93	**2.72**×	4.19	**2.93**×	3.98	**2.82**×	3.97	**2.79**×
	SAM	1.91	1.73×	1.95	1.78×	1.93	1.74×	2.09	1.86×	1.96	1.78×
	Recycling	2.88	2.00×	2.92	2.04×	2.90	2.11×	2.92	2.08×	2.90	2.06×
	SpS	**6.22**	0.91×	**6.54**	0.95×	**6.17**	0.80×	**6.32**	0.86×	**6.34**	0.88×
	REST	1.41	1.08×	1.41	1.10×	1.42	1.12×	1.36	1.03×	1.40	1.08×
	PIA	1.69	1.40×	1.70	1.45×	1.71	1.45×	1.75	1.48×	1.71	1.44×

- Takeaway 1: Training-free n-gram methods can outperform training-based approaches!
- Takeaway 2: This hybrid approach demonstrates a clear 1 + 1 > 2 effect !
- Takeaway 3: Training-based methods are sensitive to the training process !

Experiments [1/3] — Analysis

- Lightweight Ngram-based Methods Perform Satisfactorily

Table 3: Performance comparison of speculative decoding methods for reasoning models under multi-round thinking framework with different temperature T (**Best** , Second Best).

Model	Bench	AIME24		AIME25		MATH500		GPQA		Overall	
	Method	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed
DSL-8B ($T=0$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	2.26	1.56×	2.24	1.53×	2.63	2.86×	2.46	1.61×	2.35	1.93×
	SAM[EAGLE-3]	**3.91**	**3.09**×	**4.50**	**3.85**×	**4.94**	**4.11**×	**7.00**	**4.70**×	**4.72**	**3.97**×
	SAM	2.64	2.41×	3.05	2.96×	2.60	2.14×	3.57	3.20×	2.93	2.66×
	Recycling	2.98	2.08×	2.98	2.05×	2.97	2.18×	3.02	2.08×	2.99	2.10×
	PLD	2.14	1.76×	2.38	1.89×	2.25	1.71×	2.65	2.02×	2.33	1.84×
	REST	1.33	1.00×	1.35	0.95×	1.37	1.08×	1.31	1.00×	1.34	1.01×
	Lookahead	2.21	1.59×	2.26	1.53×	2.28	1.55×	2.44	1.62×	2.28	1.57×
	PIA	1.84	1.51×	1.96	1.63×	1.91	1.48×	2.15	1.63×	1.95	1.56×
DSL-8B ($T=0.6$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	2.21	1.31×	2.17	1.50×	3.21	3.02×	2.45	1.61×	2.33	1.91×
	SAM[EAGLE-3]	2.54	1.82×	2.48	1.89×	**3.28**	**3.13**×	**2.86**	**2.15**×	2.66	**2.29**×
	SAM	1.84	1.65×	1.84	1.65×	1.93	1.71×	1.93	1.73×	1.87	1.69×
	Recycling	**2.84**	**1.91**×	**2.83**	**1.93**×	2.88	2.03×	2.81	1.98×	**2.84**	1.96×
	REST	1.35	1.00×	1.37	0.97×	1.38	1.03×	1.35	1.00×	1.36	1.00×
	PIA	1.58	1.26×	1.58	1.27×	1.63	1.39×	1.68	1.36×	1.61	1.33×
QW3-8B ($T=0$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	4.31	2.80×	4.37	2.87×	4.50	3.04×	4.39	2.91×	4.38	2.91×
	SAM[EAGLE-3]	4.40	**3.08**×	4.47	**3.18**×	4.82	**3.47**×	5.90	**4.17**×	4.76	**3.49**×
	SAM	2.15	1.97×	2.27	2.07×	2.18	1.95×	3.19	3.11×	2.37	2.28×
	Recycling	3.02	2.08×	2.98	2.13×	3.00	2.20×	3.07	2.17×	3.01	2.15×
	PLD	1.95	1.61×	1.97	1.66×	1.91	1.63×	2.49	2.04×	2.05	1.74×
	SpS	**7.40**	0.93×	**7.50**	0.96×	**6.69**	0.77×	**6.33**	0.83×	**7.07**	0.87×
	REST	1.39	1.12×	1.39	1.14×	1.42	1.16×	1.33	1.09×	1.38	1.13×
	Lookahead	2.14	1.54×	2.12	1.54×	2.10	1.58×	2.28	1.69×	2.15	1.59×
	PIA	1.83	1.48×	1.96	1.68×	1.89	1.63×	2.08	1.72×	1.93	1.63×
QW3-8B ($T=0.6$)	AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×
	EAGLE-3	4.11	2.61×	4.21	2.71×	4.32	2.87×	4.03	2.68×	4.16	**2.73**×
	SAM[EAGLE-3]	3.92	**2.68**×	3.93	**2.72**×	4.19	**2.93**×	3.98	**2.82**×	3.97	**2.79**×
	SAM	1.91	1.73×	1.95	1.78×	1.93	1.74×	2.09	1.86×	1.96	1.78×
	Recycling	2.88	2.00×	2.92	2.04×	2.90	2.11×	2.92	2.08×	2.90	2.06×
	SpS	**6.22**	0.91×	**6.54**	0.95×	**6.17**	0.80×	**6.32**	0.86×	**6.34**	0.88×
	REST	1.41	1.08×	1.41	1.10×	1.42	1.12×	1.36	1.03×	1.40	1.08×
	PIA	1.69	1.40×	1.70	1.45×	1.71	1.45×	1.75	1.48×	1.71	1.44×

- Takeaway 1: Training-free n-gram methods can outperform training-based approaches!
- Takeaway 2: This hybrid approach demonstrates a clear 1 + 1 > 2 effect !
- Takeaway 3: Training-based methods are sensitive to the training process !

Experiments [1/3] — Analysis

- Lightweight Ngram-based Methods Perform Satisfactorily

Table 7: Performance comparison of speculative decoding methods for reasoning models under the multi-round thinking framework at different temperatures T on other domains (**Best** , Second Best).

(a) Experiments on LiveCodeBench.

Model	DSL-8B ($T=0$)		DSL-8B ($T=0.6$)		QW3-8B ($T=0$)		QW3-8B ($T=0.6$)	
Method	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed
EAGLE-3	2.12	1.64×	2.00	1.33×	4.37	2.85×	4.02	**2.58**×
SAM[EAGLE-3]	**4.17**	**3.34**×	2.30	**1.80**×	4.46	**3.27**×	3.70	2.54×
SAM	2.90	2.79×	1.72	1.44×	2.19	1.92×	1.77	1.58×
Recycling	2.91	2.02×	**2.69**	1.71×	2.90	1.98×	2.83	1.92×
PLD	2.24	1.81×	–	–	1.94	1.57×	–	–
SpS	–	–	–	–	**5.93**	0.83×	**5.16**	0.85×
REST	1.33	1.11×	1.37	1.02×	1.35	1.14×	1.37	1.09×
Lookahead	2.15	1.57×	–	–	2.08	1.53×	–	–
PIA	2.40	1.71×	1.63	1.04×	2.04	1.46×	1.72	1.32×
AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×

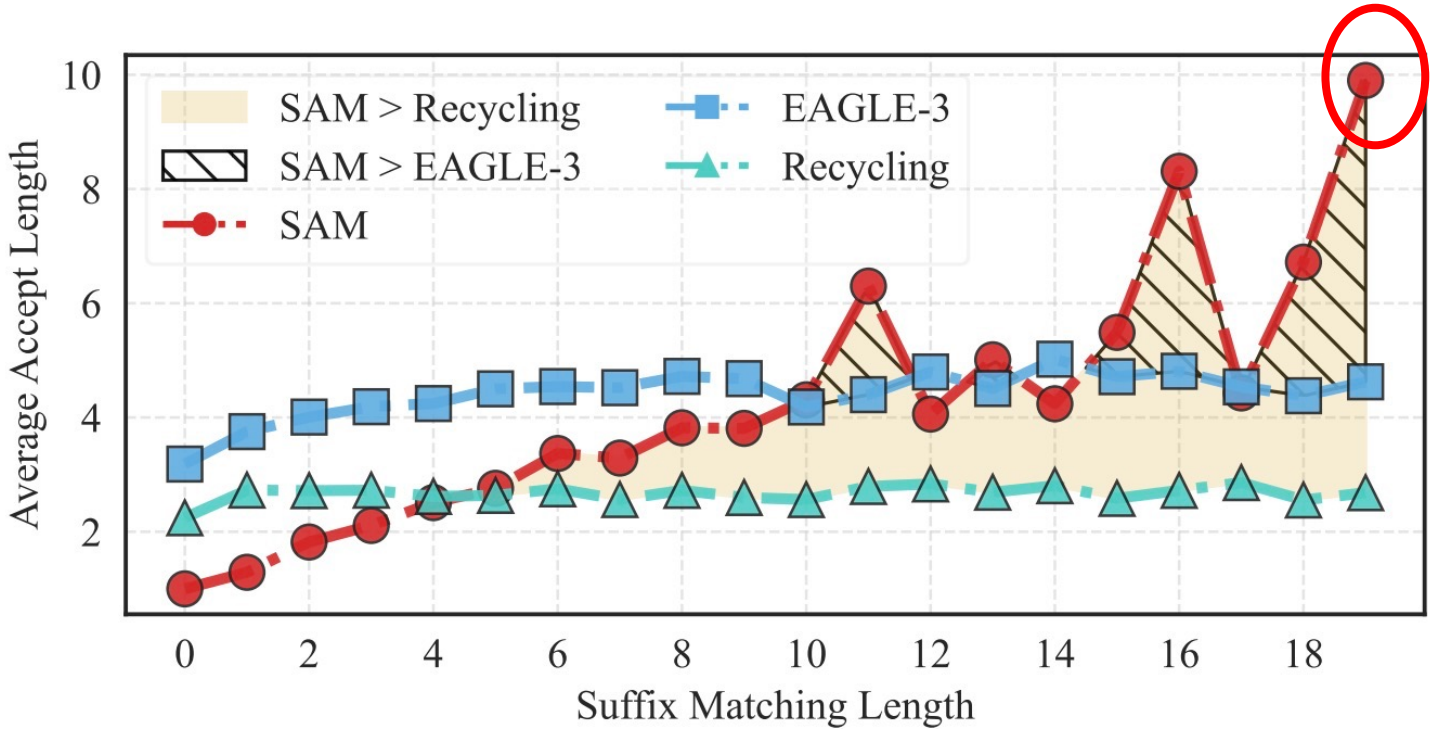
(b) Experiments on CNN/Daily Mail.

Model	DSL-8B ($T=0$)		DSL-8B ($T=0.6$)		QW3-8B ($T=0$)		QW3-8B ($T=0.6$)	
Method	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed
EAGLE-3	4.76	2.83×	4.60	2.65×	4.11	2.63×	4.01	2.53×
SAM[EAGLE-3]	**4.92**	**3.05**×	**4.70**	**2.79**×	**4.31**	**2.77**×	**4.17**	**2.61**×
SAM	1.53	1.45×	1.39	1.33×	1.55	1.45×	1.52	1.42×
Recycling	2.57	1.81×	2.50	1.71×	2.73	1.95×	2.70	1.86×
PLD	1.47	1.32×	–	–	1.46	1.30×	–	–
SpS	–	–	–	–	3.57	0.51×	3.44	0.54×
REST	1.28	1.04×	1.29	1.00×	1.28	1.05×	1.28	1.01×
Lookahead	1.59	1.29×	–	–	1.61	1.30×	–	–
PIA	1.61	1.42×	1.48	1.30×	1.58	1.39×	1.55	1.36×
AR	1.00	1.00×	1.00	1.00×	1.00	1.00×	1.00	1.00×

- Takeaway 1: Training-free n-gram methods can outperform training-based approaches!
- Takeaway 2: This hybrid approach demonstrates a clear 1 + 1 > 2 effect !
- Takeaway 3: Training-based methods are sensitive to the training process !

Experiments [2/3] — Analysis

- Hybrid Speculative Decoding is Promising



~ 2.5x Eagle-3

- We need more fine-grained hybrid strategies.
- The potential of hybrid strategies has not yet been fully realized.

Experiments [3/3] — Analysis

- Batch Inference Testing

Table 9: Performance comparison with varying batch sizes on DSL-8B (**Best** , Second Best).

(a) Batch Size = 4

Dataset	AIME24		AIME25		MATH500		GPQA	
Method	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed
EAGLE-3	**3.30**	**1.52**×	**3.34**	1.56×	3.10	1.46×	**3.08**	1.41×
SAM-Single	1.81	1.33×	1.87	1.44×	2.91	2.72×	2.01	1.58×
SAM-Cross	1.86	1.34×	2.11	**1.62**×	**3.19**	**2.89**×	2.21	**1.73**×

(b) Batch Size = 8

Dataset	AIME24		AIME25		MATH500		GPQA	
Method	MAT	Speed	MAT	Speed	MAT	Speed	MAT	Speed
EAGLE-3	**3.34**	**1.17**×	**3.37**	1.18×	**3.13**	1.11×	**3.06**	1.03×
SAM-Single	1.75	1.06×	1.81	1.18×	2.52	2.24×	1.92	**1.36**×
SAM-Cross	1.82	1.12×	1.92	**1.20**×	3.12	**2.61**×	1.99	1.34×

- The n-gram approach outperforms training-based methods because N-gram **draft generation is extremely lightweight**.
- We should arrange the N dimensions of the BON sequentially, while assigning different queries in parallel. This **maximizes overlap**, which is particularly useful in RL rollout scenarios.

Thank you for watching!

Project page: https://github.com/sunshy-1/SpecTTS-Bench

Correspondence: shengysun4-c@my.cityu.edu.hk