

EAGLE: Speculative Sampling Requires Rethinking Feature Uncertainty

Yuhui Li¹, Fangyun Wei, Chao Zhang, Hongyang Zhang

Peking University, Microsoft Research, University of Waterloo 4Vector Institute

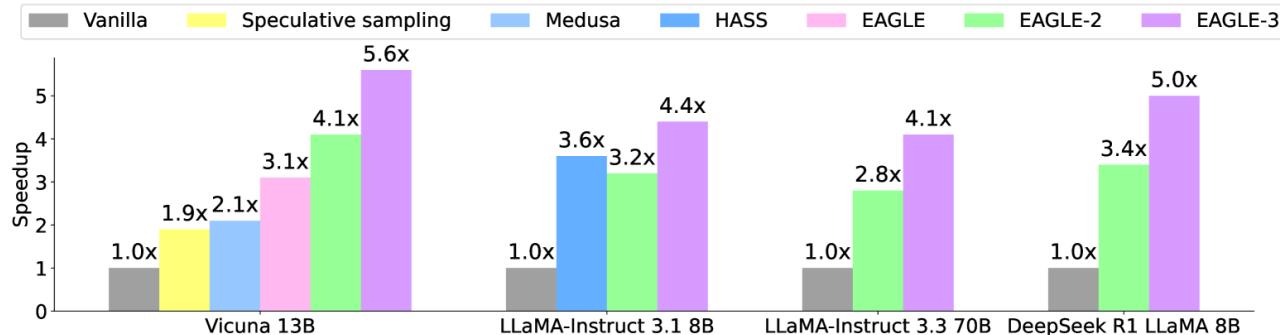
NeurIPS 2025

2025/10/20

Jun-Chen, Hung

Introduction

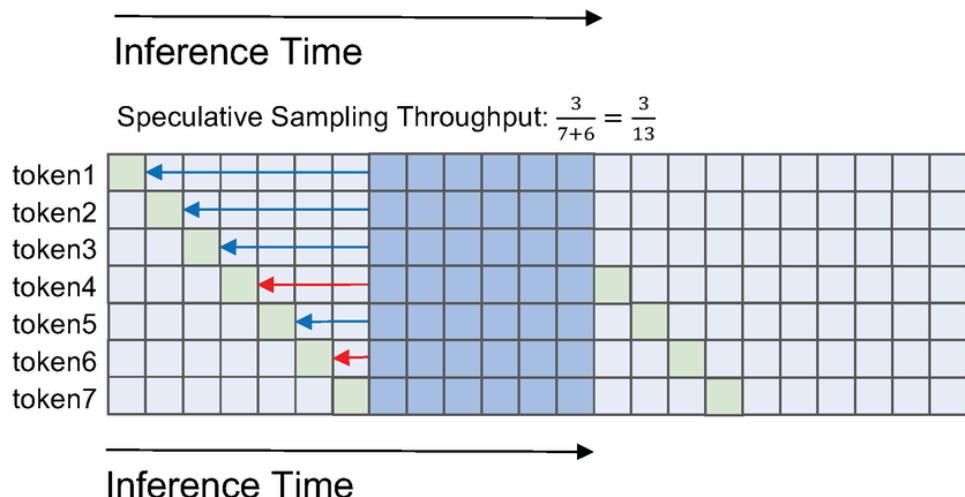
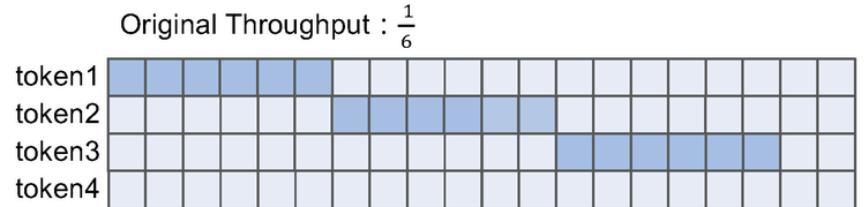
- The remarkable capabilities of modern Large Language Models (LLMs) are often offset by a significant operational challenge: their **autoregressive, token-by-token generation process** makes inference slow and computationally expensive.
- Speculative sampling accelerates this process using a "**draft and verify**" approach, where a smaller, faster model generates a sequence of draft tokens that are then validated in parallel by the larger target model, significantly reducing latency.
- EAGLE-3 stands as the culmination of the EAGLE series, a state-of-the-art speculative sampling method that achieves unprecedented speedups. It introduces a novel "Training-Time Test" (TTT) architecture and multi-layer feature fusion to overcome the performance plateaus of its predecessors.



Related Work - Speculative Sampling

Google Deepmind

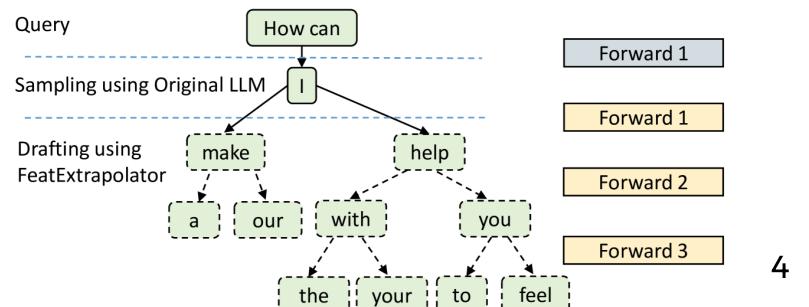
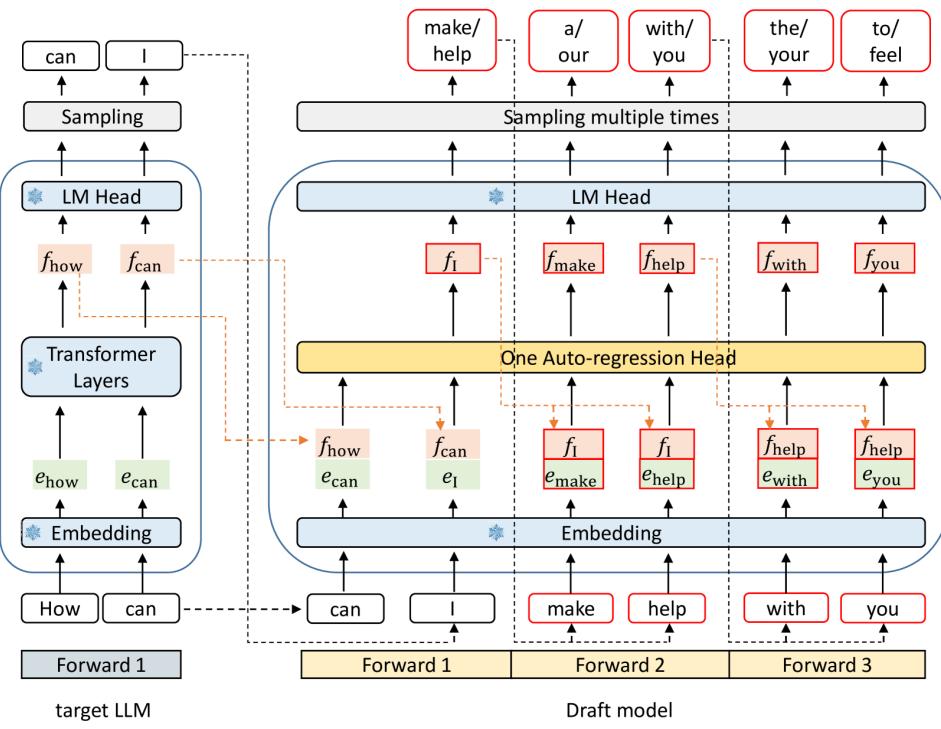
- Involves using a smaller, faster draft model to generate a sequence of tokens. These tokens are then passed to the larger target model for parallel verification, accepting a prefix of the draft that matches the target model's predictions.
- Its primary limitation is that the **high overhead** and **lower accuracy** of the draft model can constrain acceleration gains, as the target model frequently rejects inaccurate drafts.



Related Work - Eagle

ICML 2024

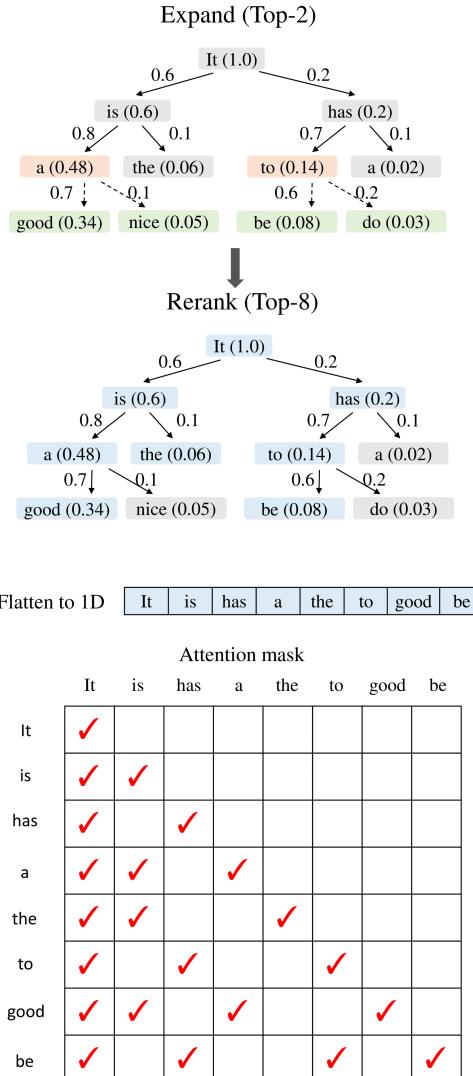
- Instead of token prediction, EAGLE use **feature prediction**, which proved to be a simpler and more effective prediction target.
- To fix feature uncertainty, EAGLE predict feature with token to make it stable
- Use **tree-structured** predict chain instead of chain-structured => increase accept length τ



Related Work - Eagle2

EMNLP 2024

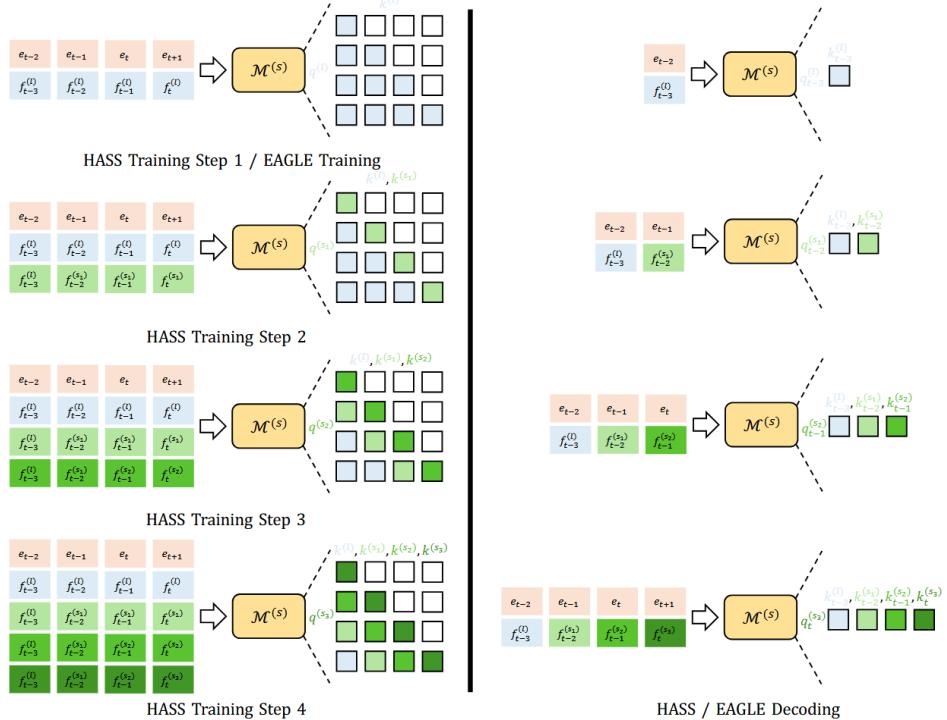
- Identified the limitation of EAGLE's **static draft tree**, which implicitly assumes that the acceptance rate of draft tokens depends only on their position in the tree structure.
- EAGLE-2's solution was a **context-aware dynamic draft tree**. Use **Expand** and **Rerank** to dynamically adjust the shape of the draft tree based on the immediate context, allocating resources more efficiently.



Related Work - HASS

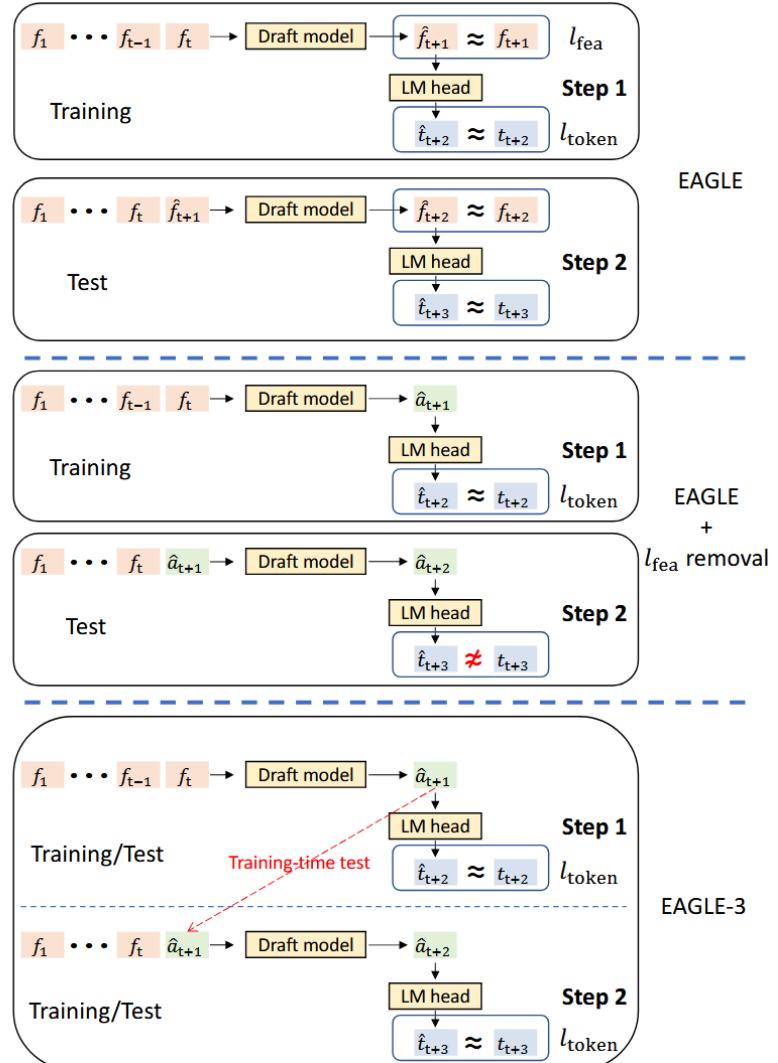
ICLR 2025

- EAGLE2 uses target model feature as training data, but use self predicted feature as inference input, causes **Exposure Bias** (暴露偏差)
- Previous training focus on target model vocabulary, but real target is predict **expect token** by target model
- By modify training process and loss function, HASS gain **8% ~ 20%** speed up compare to EAGLE2



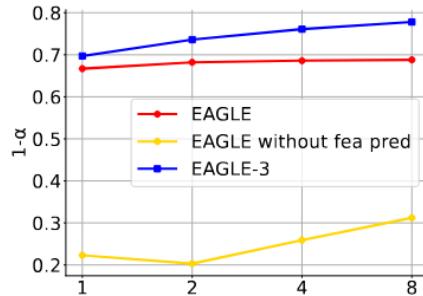
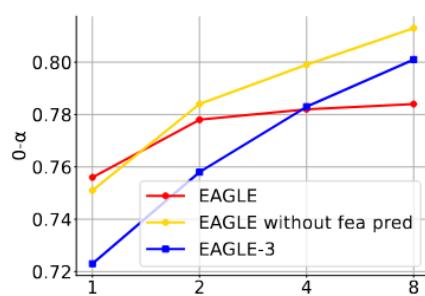
Problem

- Recent LLMs increase training tokens, but not for EAGLE draft model
 - Old EAGLE use feature loss L_{fea} and token loss L_{token} to train the model
 - Two losses cause additional constraint



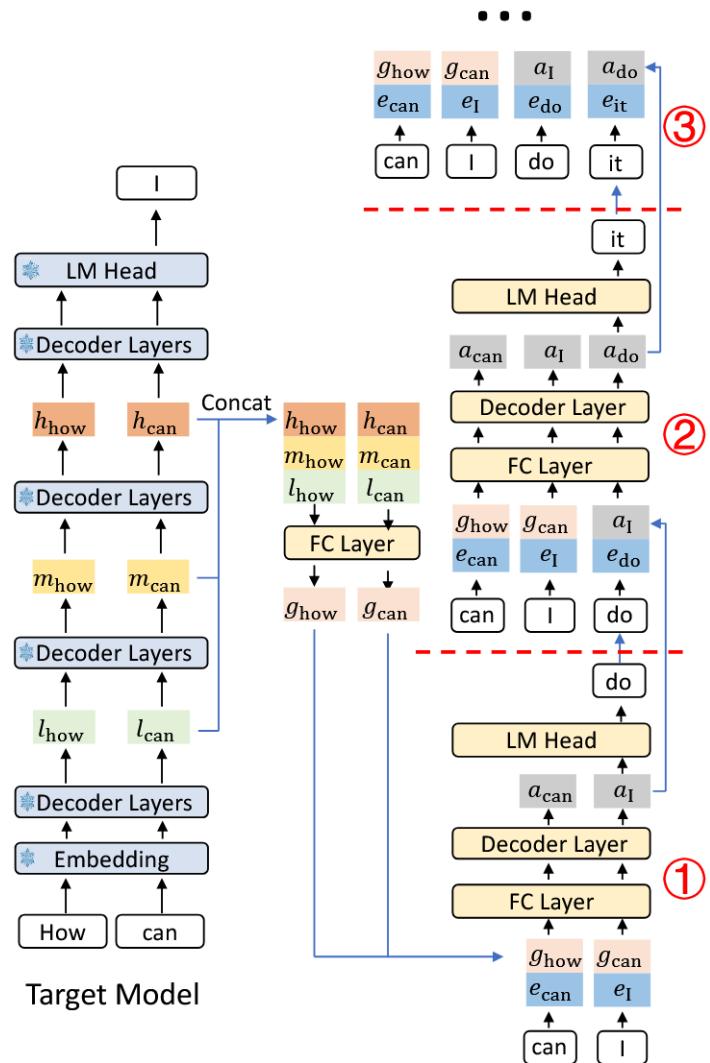
Problem

- These top-layer features correspond to the logits of the next token.
 - Relying solely on these limited features makes predicting the **next-next token** a significant challenge for the draft model
- HASS mitigate the error accumulation caused by inaccurate feature predictions in EAGLE
 - But still use **feature prediction**



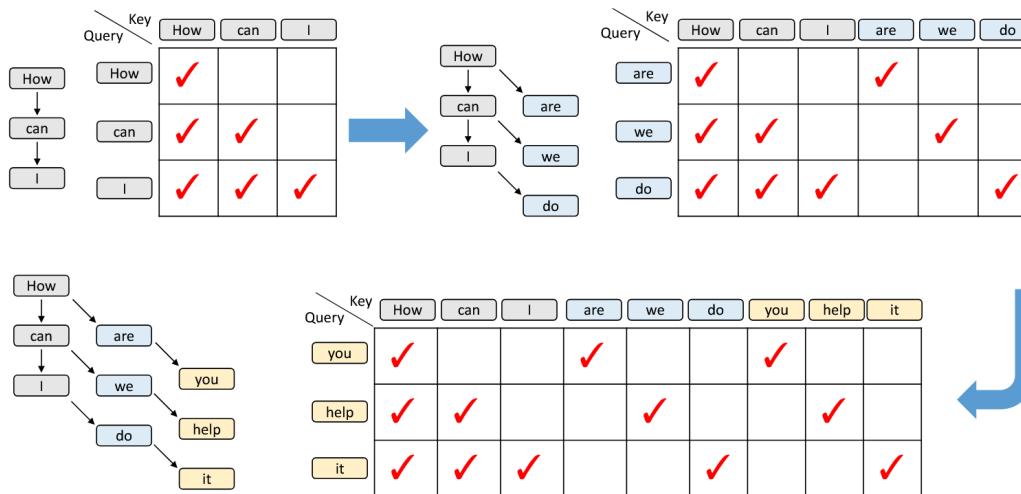
Solution

- Multi-Layer Feature Fusion
 - Concatenating **low**, **middle**, and **high**-level feature sequences from the target model.
 - Use an FC Layer reduce $3 \times k$ back to k dimension
 - When inference, use previous generated token to predict next token



Solution

- Training-Time Test (TTT):
 - Removes the feature prediction constraint.
 - Simulates multi-step generation during the training phase by feeding the draft model's own outputs back into it.



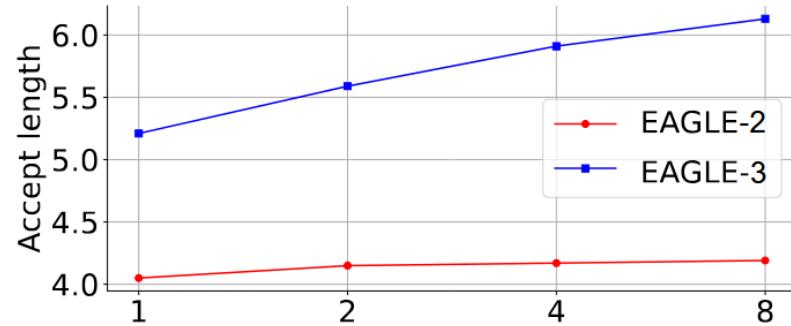
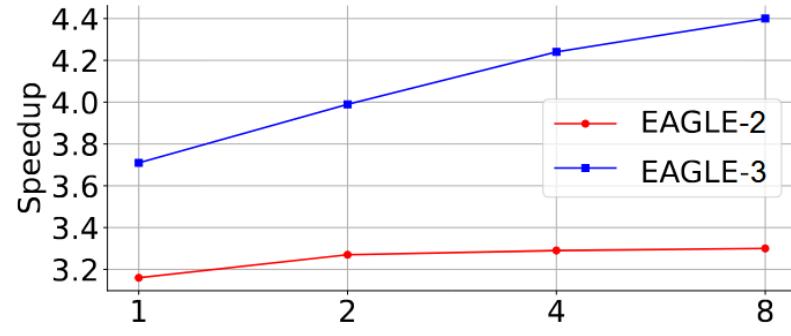
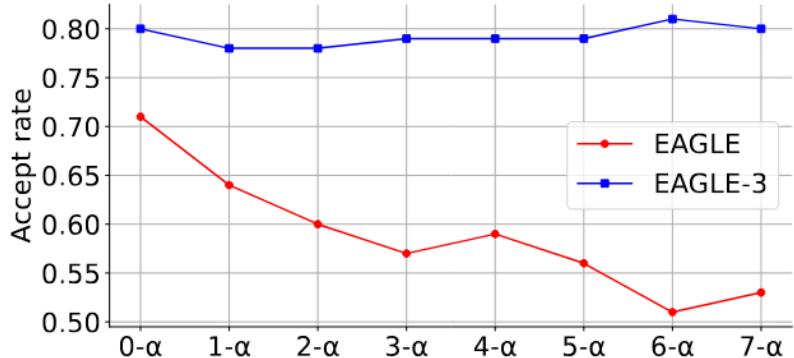
Experiment

- Models: Vicuna 13B, LLaMA-Instruct 3.18B, LLaMA-Instruct 3.3 70B, DeepSeek-R1-Distill-LLaMA 8B.
- Task: MT-bench(multi-turn conversation), HumanEval(code generation), GSM8K(mathematical reasoning), Alpaca(instruction following), CNN/DM(summarization),
- Metrics:
 - Speedup ratio
 - Average Accept Length τ
 - Acceptance Rate $n - \alpha$
- Training dataset: ShareGPT & Ultra-Chat200k

Experiment

Model	Method	MT-bench		HumanEval		GSM8K		Alpaca		CNN/DM		Mean	
		Speedup	τ										
Temperature=0													
V 13B	SpS	1.93x	2.27	2.23x	2.57	1.77x	2.01	1.76x	2.03	1.93x	2.33	1.92x	2.24
	PLD	1.58x	1.63	1.85x	1.93	1.68x	1.73	1.16x	1.19	2.42x	2.50	1.74x	1.80
	Medusa	2.07x	2.59	2.50x	2.78	2.23x	2.64	2.08x	2.45	1.71x	2.09	2.12x	2.51
	Lookahead	1.65x	1.69	1.71x	1.75	1.81x	1.90	1.46x	1.51	1.46x	1.50	1.62x	1.67
	Hydra	2.88x	3.65	3.28x	3.87	2.93x	3.66	2.86x	3.53	2.05x	2.81	2.80x	3.50
	EAGLE	3.07x	3.98	3.58x	4.39	3.08x	3.97	3.03x	3.95	2.49x	3.52	3.05x	3.96
	EAGLE-2	4.26x	4.83	4.96x	5.41	4.22x	4.79	4.25x	4.89	3.40x	4.21	4.22x	4.83
	EAGLE-3	5.58x	6.65	6.47x	7.54	5.32x	6.29	5.16x	6.17	5.01x	6.47	5.51x	6.62
L31 8B	EAGLE-2	3.16x	4.05	3.66x	4.71	3.39x	4.24	3.28x	4.12	2.65x	3.45	3.23x	4.11
	EAGLE-3	4.40x	6.13	4.85x	6.74	4.48x	6.23	4.82x	6.70	3.65x	5.34	4.44x	6.23
L33 70B	EAGLE-2	2.83x	3.67	3.12x	4.09	2.83x	3.69	3.03x	3.92	2.44x	3.55	2.85x	3.78
	EAGLE-3	4.11x	5.63	4.79x	6.52	4.34x	6.15	4.30x	6.09	3.27x	5.02	4.12x	5.88
DSL 8B	EAGLE-2	2.92x	3.80	3.42x	4.29	3.40x	4.40	3.01x	3.80	3.53x	3.33	3.26x	3.92
	EAGLE-3	4.05x	5.58	4.59x	6.38	5.01x	6.93	3.65x	5.37	3.52x	4.92	4.16x	5.84
Temperature=1													
V 13B	SpS	1.62x	1.84	1.72x	1.97	1.46x	1.73	1.52x	1.78	1.66x	1.89	1.60x	1.84
	EAGLE	2.32x	3.20	2.65x	3.63	2.57x	3.60	2.45x	3.57	2.23x	3.26	2.44x	3.45
	EAGLE-2	3.80x	4.40	4.22x	4.89	3.77x	4.41	3.78x	4.37	3.25x	3.97	3.76x	4.41
	EAGLE-3	4.57x	5.42	5.15x	6.22	4.71x	5.58	4.49x	5.39	4.33x	5.72	4.65x	5.67
L31 8B	EAGLE-2	2.44x	3.16	3.39x	4.39	2.86x	3.74	2.83x	3.65	2.44x	3.14	2.80x	3.62
	EAGLE-3	3.07x	4.24	4.13x	5.82	3.32x	4.59	3.90x	5.56	2.99x	4.39	3.45x	4.92
L33 70B	EAGLE-2	2.73x	3.51	2.89x	3.81	2.52x	3.36	2.77x	3.73	2.32x	3.27	2.65x	3.54
	EAGLE-3	3.96x	5.45	4.36x	6.16	4.17x	5.95	4.14x	5.87	3.11x	4.88	3.95x	5.66
DSL 8B	EAGLE-2	2.69x	3.41	3.01x	3.82	3.16x	4.05	2.64x	3.29	2.35x	3.13	2.77x	3.54
	EAGLE-3	3.20x	4.49	3.77x	5.28	4.38x	6.10	3.16x	4.30	3.08x	4.27	3.52x	4.89

Experiment



Ablation Study

Method	MT-bench		GSM8K	
	Speedup	τ	Speedup	τ
EAGLE-2	3.16x	4.05	3.39x	4.24
+ remove fea con	3.82x	5.37	3.77x	5.22
+ fused features (ours)	4.40x	6.13	4.48x	6.23

Conclusion

- **Problem:** Earlier speculative sampling methods like EAGLE-2 were fundamentally limited by a feature prediction constraint that prevented performance from improving with more training data.
- **Solution:** EAGLE-3 overcomes this by introducing a Training-Time Test (TTT) architecture for direct token prediction and using multi-layer feature fusion for richer contextual input.
- **Experiment:** Comprehensive benchmarks confirmed state-of-the-art performance, achieving up to a 6.5x speedup and demonstrating a unique scaling law where more data yields better acceleration.

EAGLE-3 continues to benefit from the augmentation of training data, achieving a maximum speedup of 6.5x.

Pros and Cons

Pros

- SOTA in speculative decoding
- Ready to use model and can be reproduced
- Framework supported

Cons

- Long context failure
- No in-domain training dataset analyst