EAGLE & EAGLE-2
Lossless Inference Acceleration for LLMs

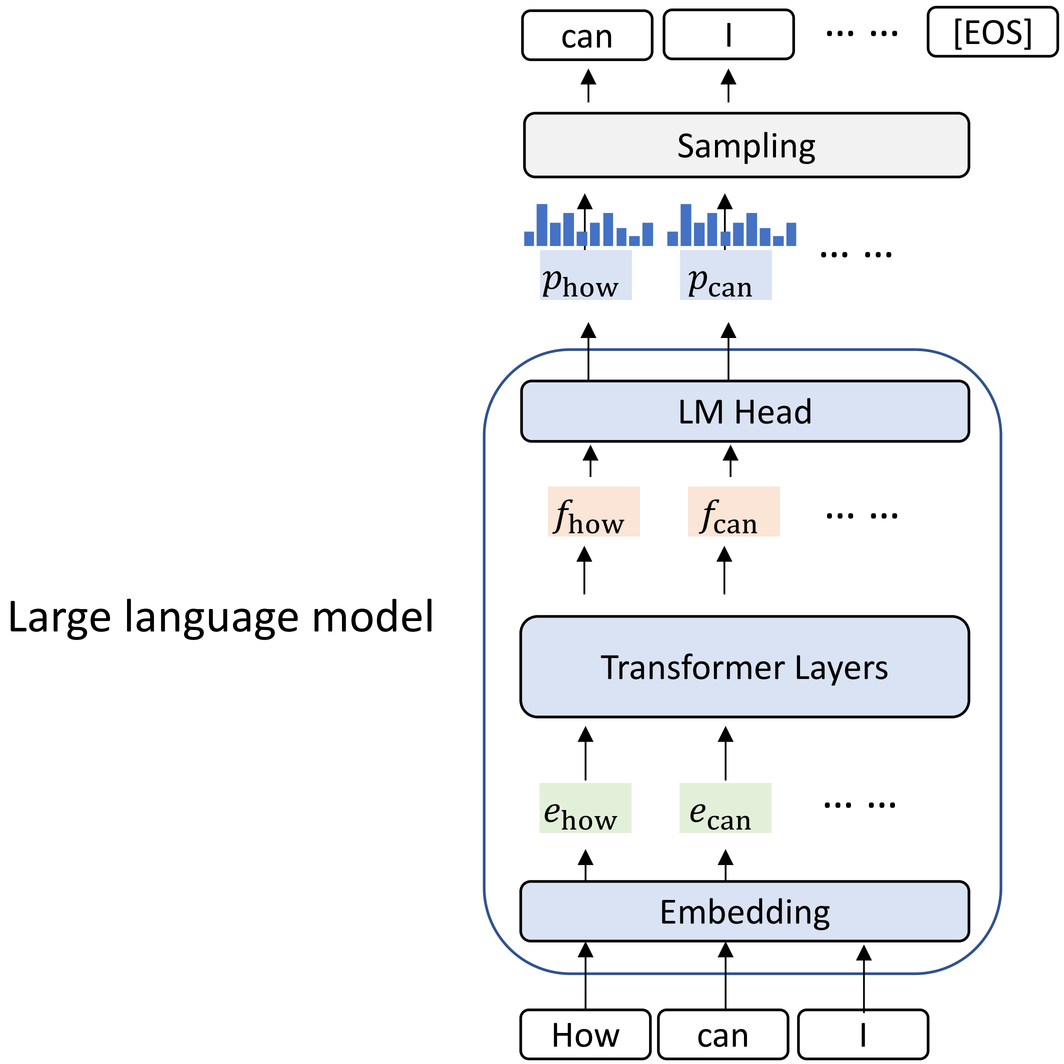
Hongyang Zhang

University of Waterloo & Vector Institute for AI

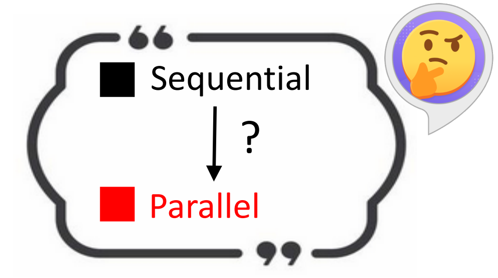
UNIVERSITY OF WATERLOO

VECTOR INSTITUTE

Vanilla autoregressive inference



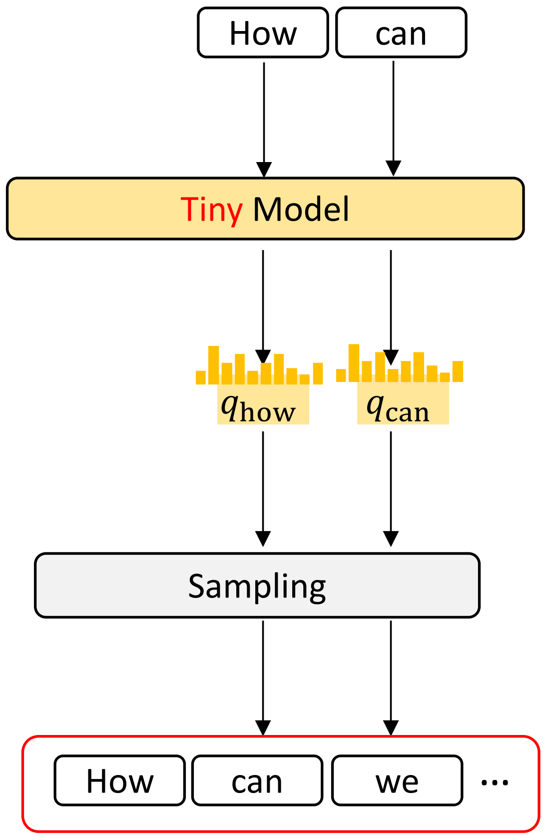
- Sequential inference



Speculative sampling framework (draft)

Good approx. of
original large model

Output
distribution



- Sequential inference
  (but fast due to tiny model)

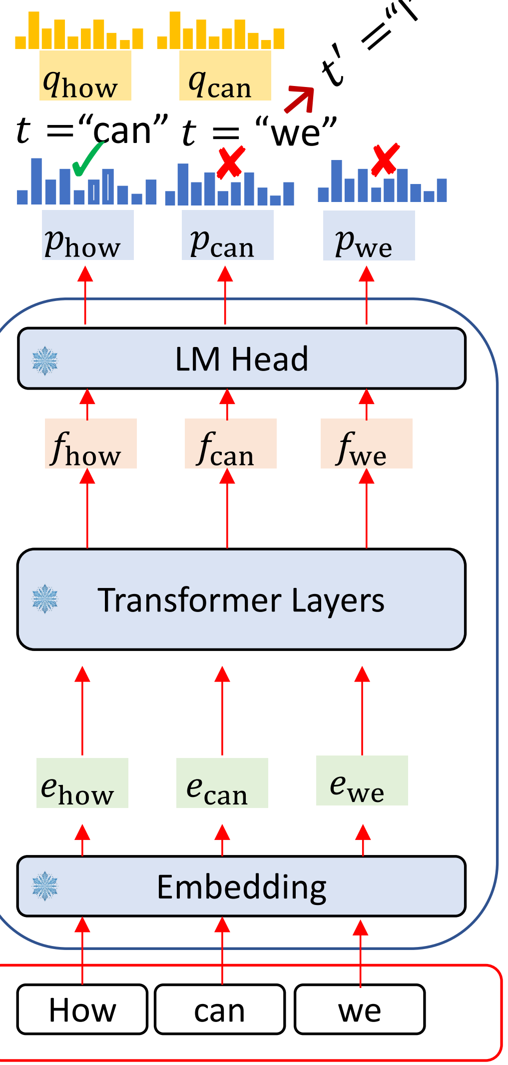
Leviathan, Yaniv, Matan Kalman, and Yossi Matias. "Fast inference from transformers via speculative decoding." *International Conference on Machine Learning*. PMLR, 2023.

Chen, Charlie, et al. "Accelerating large language model decoding with speculative sampling." arXiv preprint arXiv:2302.01318 (2023).

Speculative sampling framework (check)

Compare with tiny model

Original large model



Drafted by tiny model

1. $r \sim U(0,1)$, if $r < \min\left(1, \frac{p(t)}{q(t)}\right)$, next token = t

Accept rate

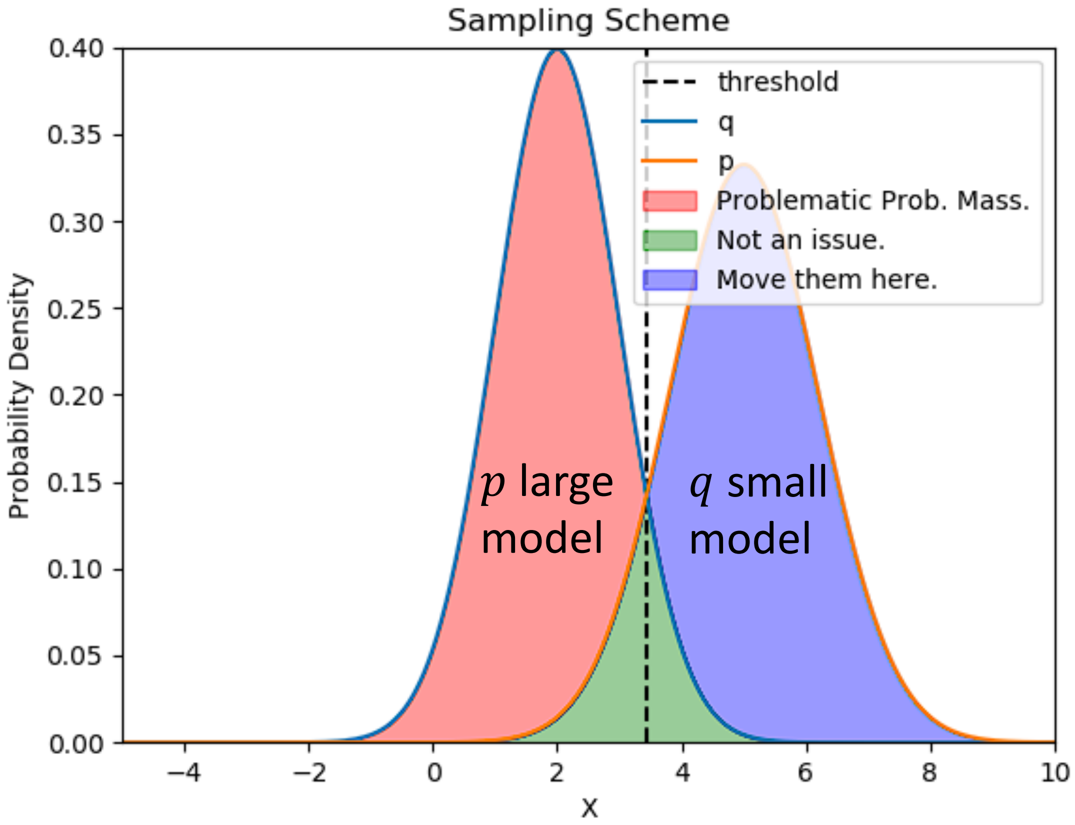
2. else: next token = $t' \sim \text{norm}(\max(0, p - q))$

- Check in parallel

Leviathan, Yaniv, Matan Kalman, and Yossi Matias. "Fast inference from transformers via speculative decoding." *International Conference on Machine Learning*. PMLR, 2023.

Chen, Charlie, et al. "Accelerating large language model decoding with speculative sampling." arXiv preprint arXiv:2302.01318 (2023).

Speculative sampling framework (check)



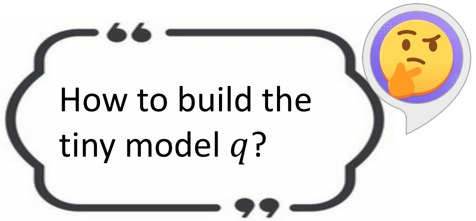
It is all about how to use sampling from (p, q) to mimic sampling from p.

1. $r \sim U(0,1)$, if $r < \min\left(1, \frac{p(t)}{q(t)}\right)$, next token = t

Accept rate

2. else: next token = $t' \sim \text{norm}(\max(0, p - q))$

Theorem:
The above (p, q) sampling procedure is equivalent to sampling from p.



Leviathan, Yaniv, Matan Kalman, and Yossi Matias. "Fast inference from transformers via speculative decoding." *International Conference on Machine Learning*. PMLR, 2023.

Chen, Charlie, et al. "Accelerating large language model decoding with speculative sampling." arXiv preprint arXiv:2302.01318 (2023).

How to build the tiny model q?

- Trade-off between accuracy and efficiency

Fast but inaccurate

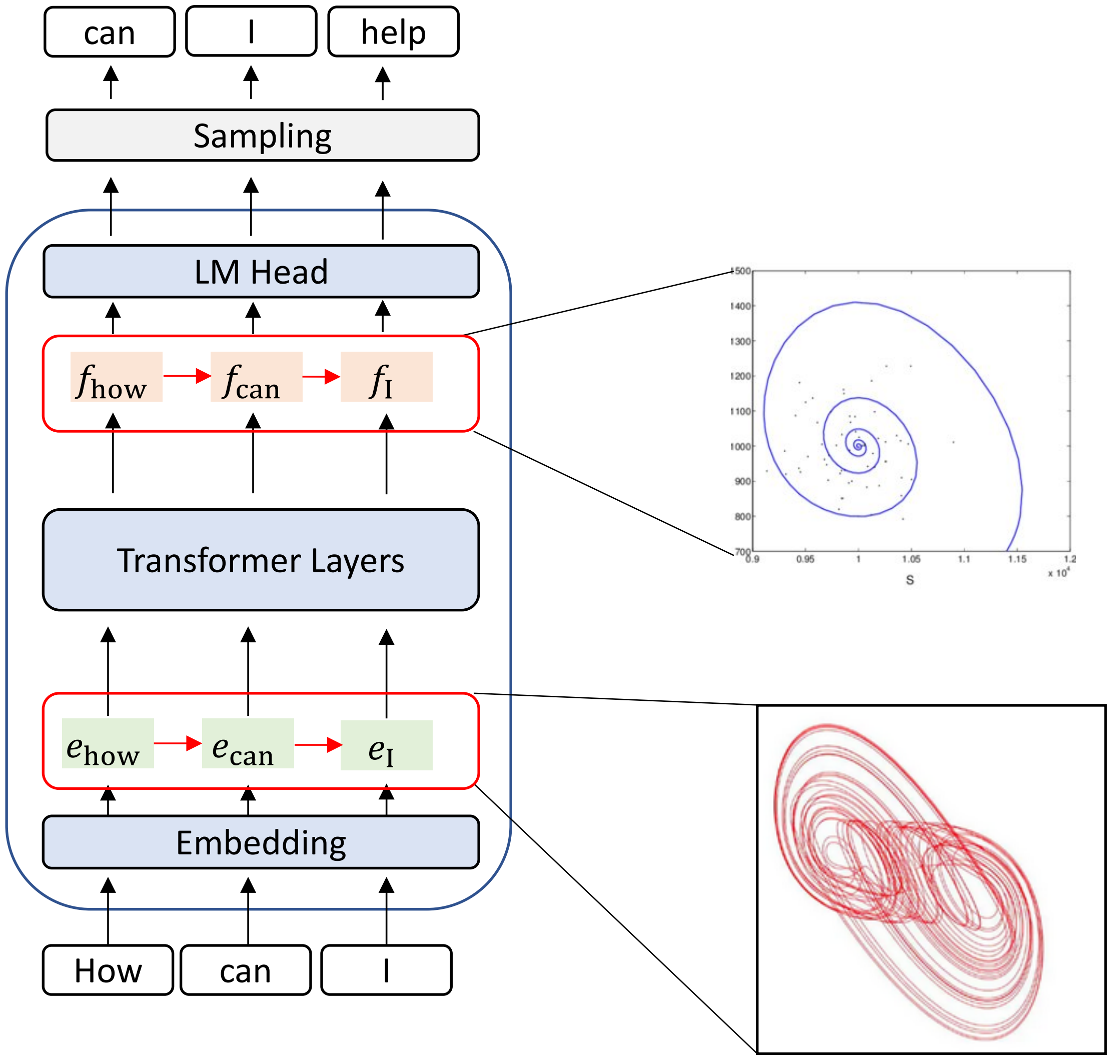
Accurate but slow

EAGLE: Speculative Sampling Requires Rethinking Feature Uncertainty

Yuhui Li, Fangyun Wei, Chao Zhang, Hongyang Zhang

(ICML 2024)

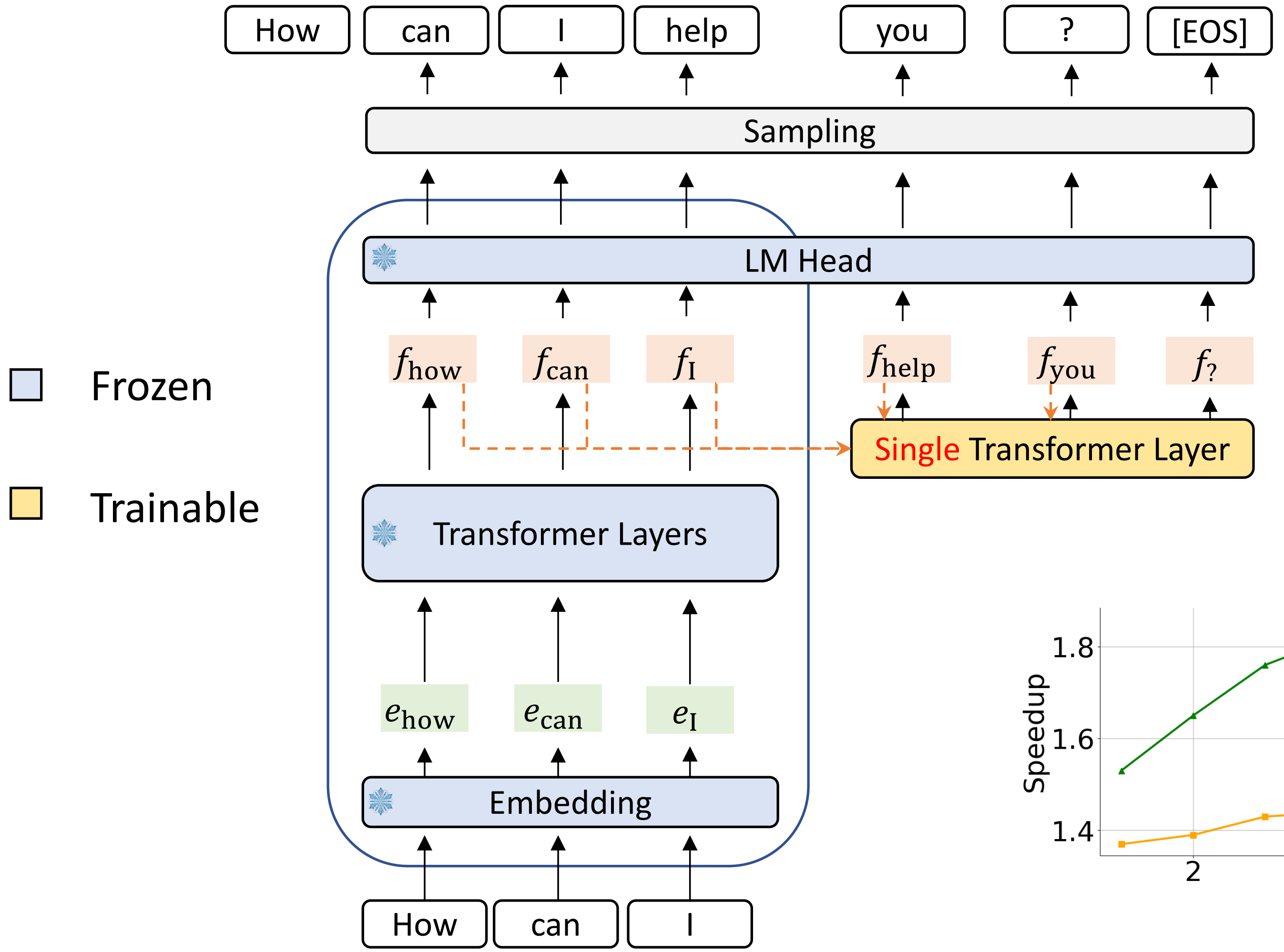
Observations

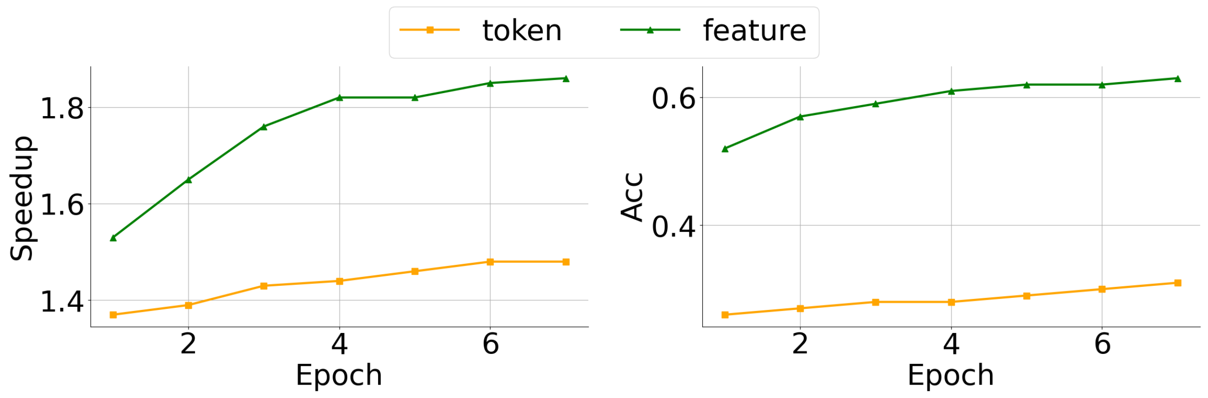


A simple dynamic
(Need a tiny model)

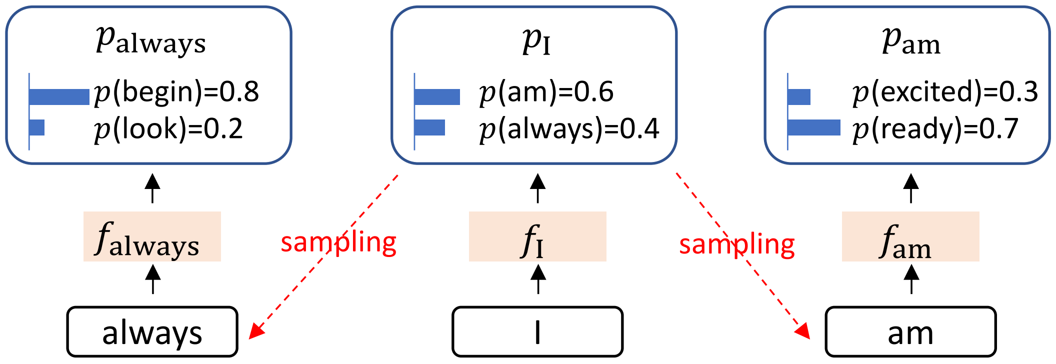
A complicated dynamic
(Need a large model)

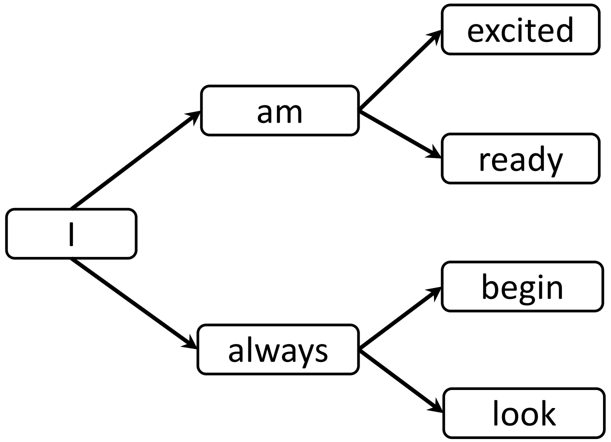
Our first trial: next-feature prediction

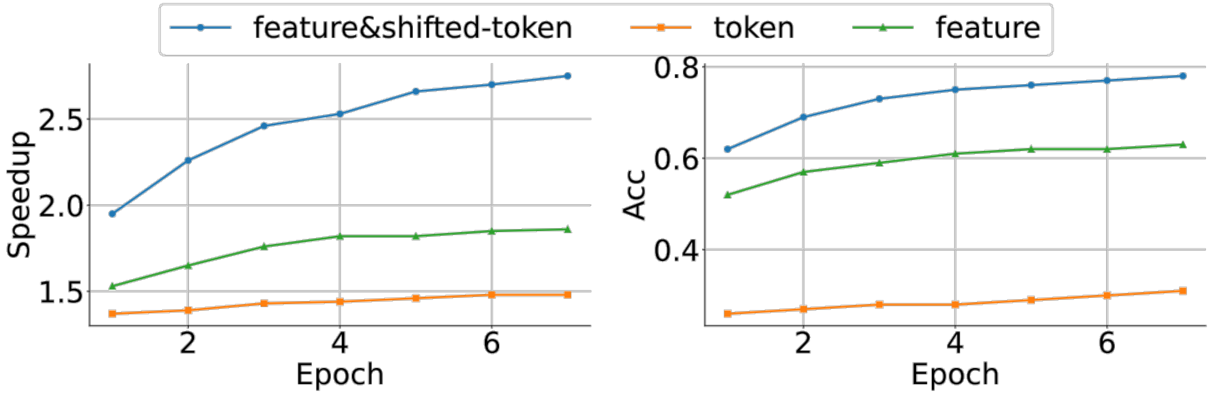




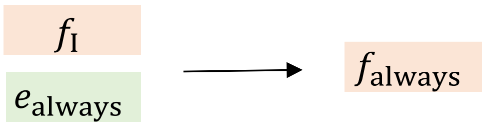
Feature uncertainty matters



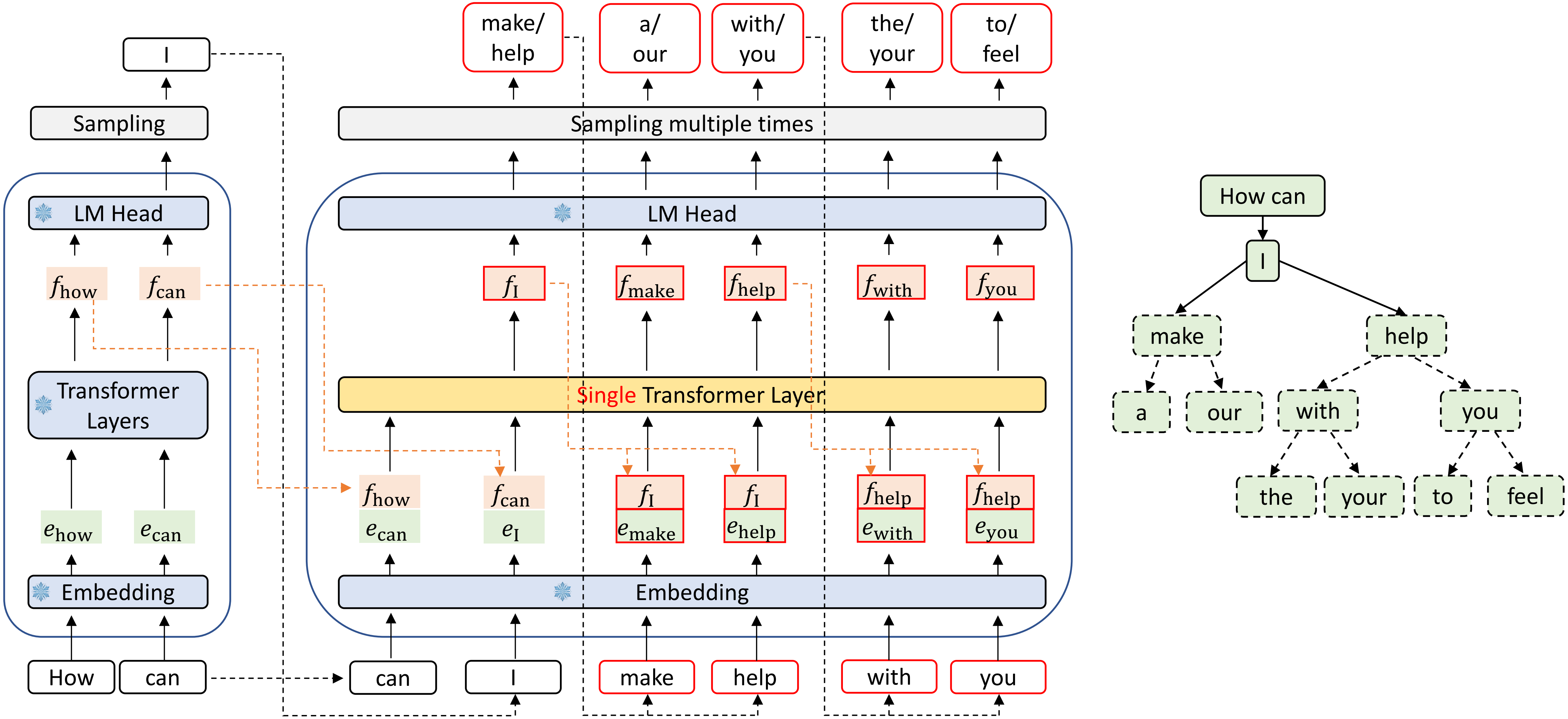




Idea: feature & shifted-token -> next feature



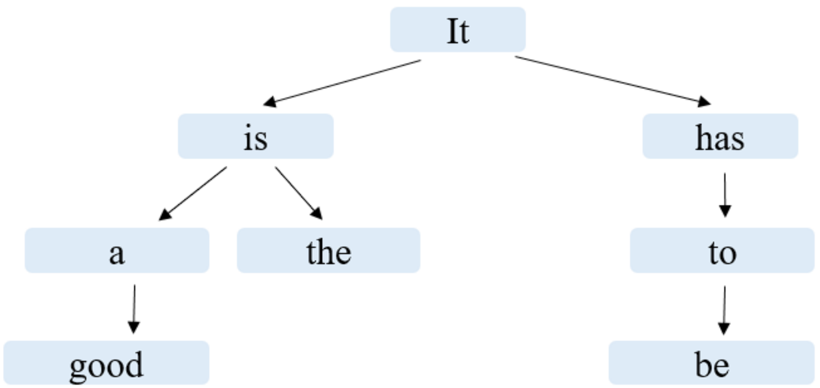
Our second trial: EAGLE



Original LLM

Draft model

Tree attention



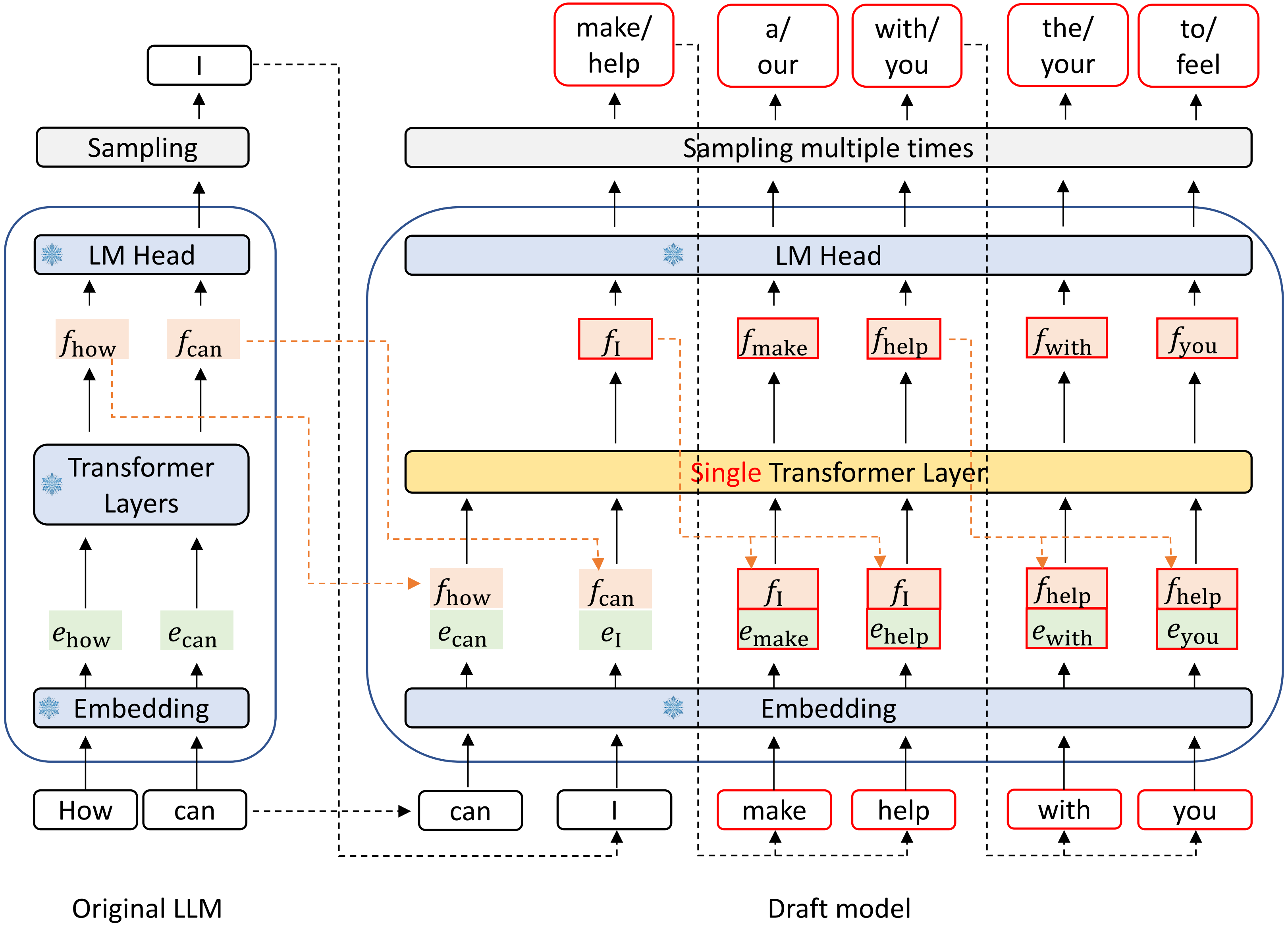
Flatten to 1D

It	is	has	a	the	to	good	be

Attention mask

	It	is	has	a	the	to	good	be
It	✓							
is	✓	✓						
has	✓		✓					
a	✓	✓		✓				
the	✓	✓			✓			
to	✓		✓			✓		
good	✓	✓		✓			✓	
be	✓		✓			✓		✓

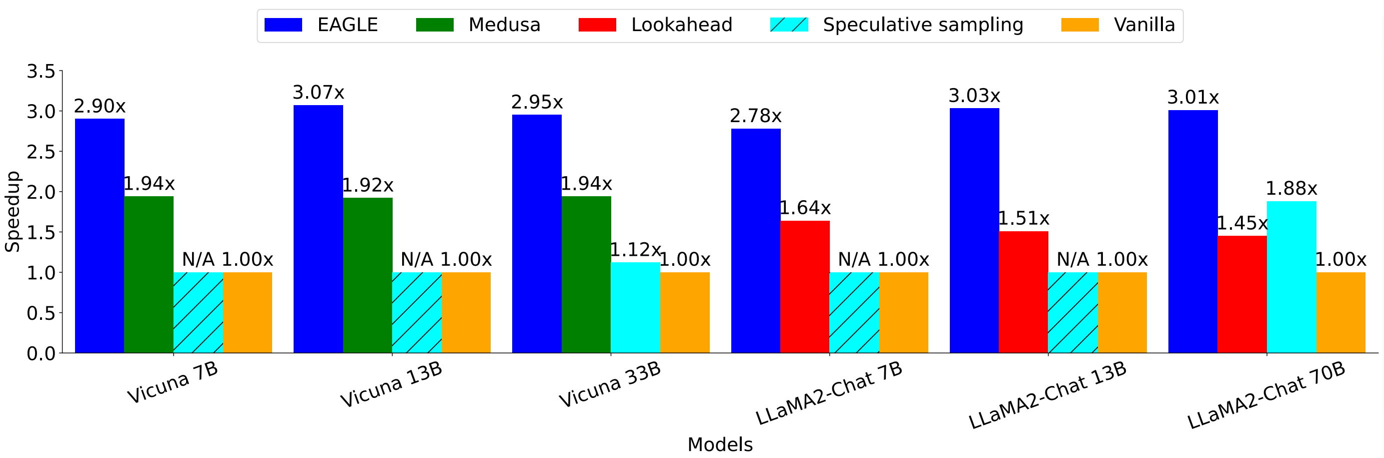
#Parameters of the drafted models

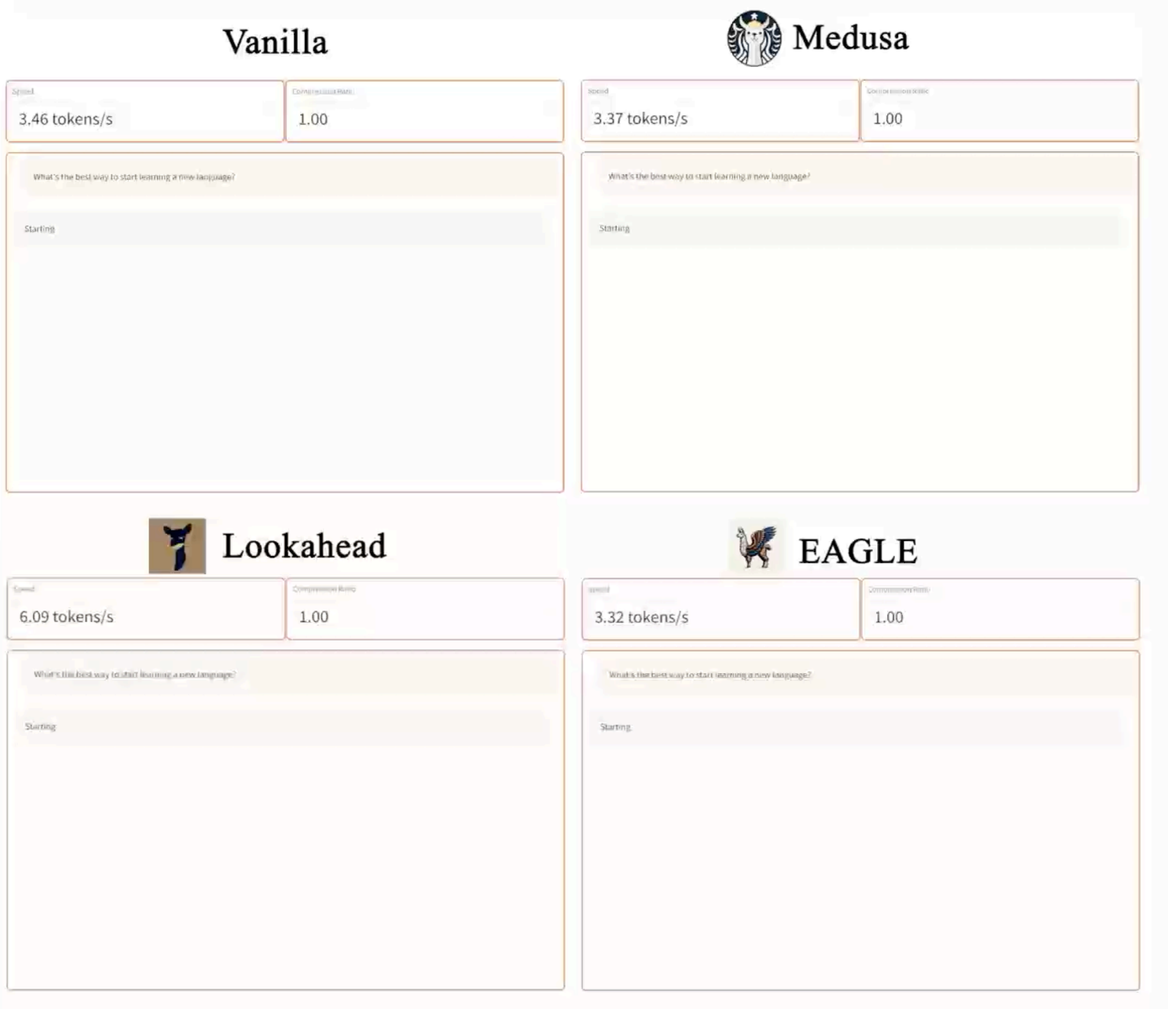


#Parameters (Original LLM)	#Parameters (Draft model)	Ratio
7B	0.24B	3.4%
13B	0.37B	2.8%
33B	0.56B	1.7%
70B	0.99B	1.4%

- Trained on RTX 3090 GPUs for 1 - 2 days

Performance on MT-bench





On MT-Bench, EAGLE is

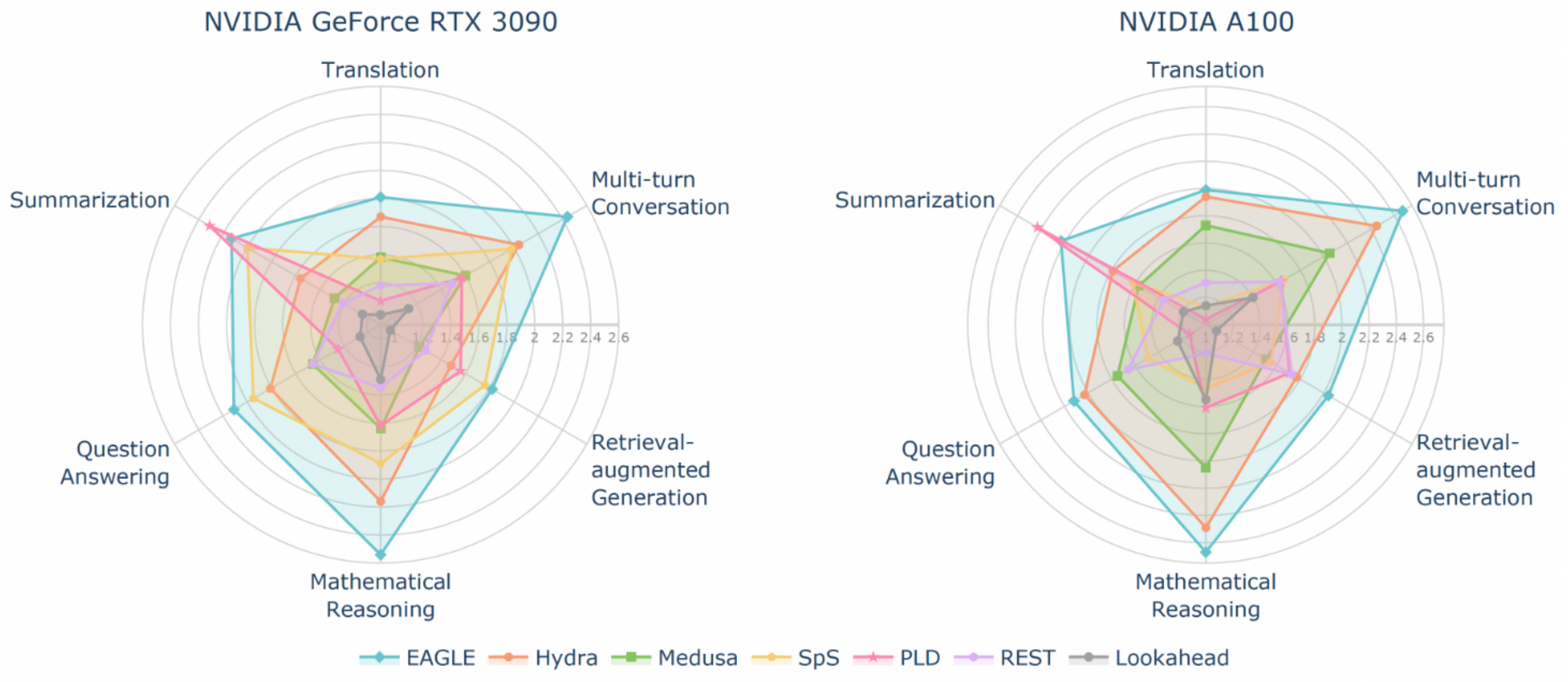
- 3x 🚀 than vanilla decoding
- 1.6x 🚀 than Medusa
- 2x 🚀 than Lookahead
- Provably maintaining text distribution

Third-party evaluations

Spec-bench

Unlocking Efficiency in Large Language Model Inference:
A Comprehensive Survey of Speculative Decoding

Heming Xia[1], Zhe Yang[2], Qingxiu Dong[2], Peiyi Wang[2],
Yongqi Li[1], Tao Ge[3], Tianyu Liu[4], Wenjie Li[1], Zhifang Sui[2]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Microsoft Research Asia [4]Alibaba Group



Speedup comparison of Speculative Decoding methods on Spec-Bench, evaluated by Vicuna-7B-v1.3.

Third-party evaluations

***Unlocking Efficiency in Large Language Model Inference:*
A Comprehensive Survey of Speculative Decoding**

**Heming Xia[1], Zhe Yang[2], Qingxiu Dong[2], Peiyi Wang[2],
Yongqi Li[1], Tao Ge[3], Tianyu Liu[4], Wenjie Li[1], Zhifang Sui[2]**
[1]Department of Computing, The Hong Kong Polytechnic University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Microsoft Research Asia [4]Alibaba Group

Spec-bench

RTX 3090, Vicuna 7B

Models	Multi-turn Conversation	Translation	Summa-rization	Question Answering	Mathematical Reasoning	Retrieval-aug. Generation	#Mean Accepted Tokens	Overall
EAGLE 🥇	**2.44x**	**1.81x**	2.13x	**2.11x**	**2.54x**	**1.82x**	**3.57**	**2.16x**
SpS 🥈	1.98x	1.37x	2.00x	1.95x	1.89x	1.76x	2.29	1.83x
Hydra 🥉	2.04x	1.67x	1.56x	1.81x	2.16x	1.48x	3.26	1.80x
PLD	1.57x	1.07x	**2.31x**	1.25x	1.62x	1.56x	1.74	1.55x
Medusa	1.60x	1.38x	1.28x	1.46x	1.64x	1.22x	2.32	1.44x
REST	1.49x	1.18x	1.21x	1.46x	1.35x	1.27x	1.63	1.32x
Lookahead	1.13x	0.97x	1.05x	1.07x	1.29x	0.98x	1.65	1.08x

Third-party evaluations

Unlocking Efficiency in Large Language Model Inference:
A Comprehensive Survey of Speculative Decoding

Heming Xia[1], Zhe Yang[2], Qingxiu Dong[2], Peiyi Wang[2], Yongqi Li[1], Tao Ge[3], Tianyu Liu[4], Wenjie Li[1], Zhifang Sui[2]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Microsoft Research Asia [4]Alibaba Group

Spec-bench

A100, Vicuna 7B

Models
EAGLE 🥇
SpS 🥈
Hydra 🥉
PLD
Medusa
REST
Lookahead

Models	Multi-turn Conversation	Translation	Summa-rization	Question Answering	Mathematical Reasoning	Retrieval-aug. Generation	#Mean Accepted Tokens	Overall
EAGLE 🥇	**2.67x**	**1.99x**	2.23x	**2.12x**	**2.67x**	**2.04x**	**3.61**	**2.29x**
Hydra 🥈	2.45x	1.94x	1.79x	2.03x	2.49x	1.77x	3.24	2.09x
Medusa 🥉	2.05x	1.73x	1.57x	1.75x	2.05x	1.51x	2.32	1.78x
PLD	1.64x	1.04x	**2.43x**	1.14x	1.61x	1.71x	1.73	1.59x
SpS	1.66x	1.13x	1.62x	1.49x	1.47x	1.55x	2.28	1.49x
REST	1.63x	1.31x	1.36x	1.66x	1.21x	1.73x	1.82	1.48x
Lookahead	1.40x	1.14x	1.19x	1.24x	1.55x	1.09x	1.66	1.27x

Third-party evaluations

Spec-bench

Unlocking Efficiency in Large Language Model Inference:
A Comprehensive Survey of Speculative Decoding

Heming Xia[1], Zhe Yang[2], Qingxiu Dong[2], Peiyi Wang[2],
Yongqi Li[1], Tao Ge[3], Tianyu Liu[4], Wenjie Li[1], Zhifang Sui[2]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Microsoft Research Asia [4]Alibaba Group

A100, Vicuna 13B

Models
EAGLE 🥇
SpS 🥈
Hydra 🥉
PLD
Medusa
REST
Lookahead

Models
EAGLE 🥇
Hydra 🥈
Medusa 🥉
PLD
SpS
REST
Lookahead

Models	Multi-turn Conversation	Translation	Summa-rization	Question Answering	Mathematical Reasoning	Retrieval-aug. Generation	#Mean Accepted Tokens	Overall
EAGLE 🥇	**2.68x**	**1.96x**	**2.44x**	**2.04x**	**2.70x**	**2.23x**	**3.64**	**2.34x**
Hydra 🥈	2.46x	1.90x	1.93x	1.96x	2.48x	1.92x	3.35	2.12x
Medusa 🥉	1.96x	1.66x	1.63x	1.63x	2.00x	1.58x	2.39	1.75x
SpS	1.60x	1.13x	1.68x	1.39x	1.53x	1.67x	2.18	1.49x
PLD	1.47x	1.02x	2.19x	1.03x	1.57x	1.71x	1.68	1.48x
REST	1.52x	1.17x	1.37x	1.53x	1.19x	1.55x	1.82	1.38x
Lookahead	1.30x	1.06x	1.20x	1.12x	1.48x	1.12x	1.63	1.22x

Third-party evaluations

Spec-bench

***Unlocking Efficiency in Large Language Model Inference:* A Comprehensive Survey of Speculative Decoding**

Heming Xia[1], Zhe Yang[2], Qingxiu Dong[2], Peiyi Wang[2], Yongqi Li[1], Tao Ge[3], Tianyu Liu[4], Wenjie Li[1], Zhifang Sui[2]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Microsoft Research Asia [4]Alibaba Group

A100, Vicuna 33B

Models
EAGLE 🥇
SpS 🥈
Hydra 🥉
PLD
Medusa
REST
Lookahead

Models
EAGLE 🥇
Hydra 🥈
Medusa 🥉
PLD
SpS
REST
Lookahead

Models
EAGLE 🥇
Hydra 🥈
Medusa 🥉
SpS
PLD
REST
Lookahead

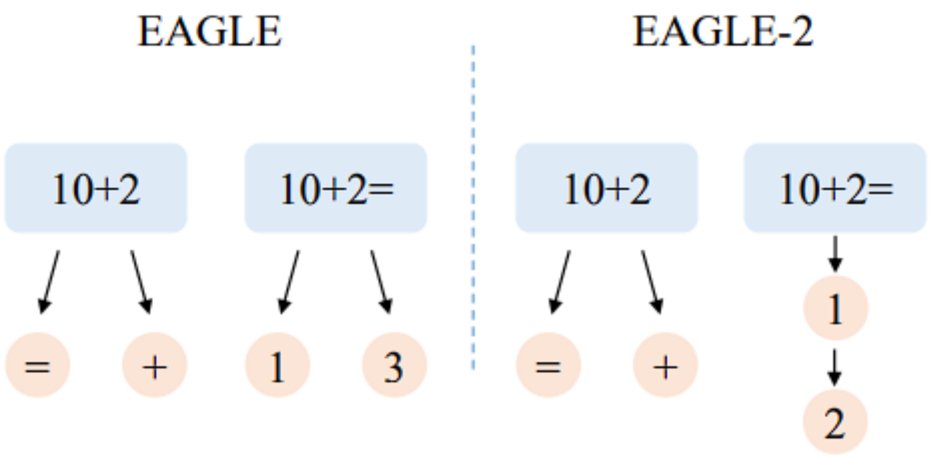
Models	Multi-turn Conversation	Translation	Summa-rization	Question Answering	Mathematical Reasoning	Retrieval-aug. Generation	#Mean Accepted Tokens	Overall
EAGLE 🥇	**2.79x**	**2.05x**	**2.51x**	**2.17x**	**2.99x**	**2.27x**	**3.39**	**2.47x**
Hydra 🥈	2.59x	2.01x	2.04x	2.11x	2.71x	2.06x	3.24	2.26x
Medusa 🥉	1.98x	1.73x	1.64x	1.66x	2.07x	1.62x	2.33	1.79x
SpS	1.75x	1.28x	1.76x	1.53x	1.69x	1.68x	2.01	1.61x
REST	1.63x	1.27x	1.45x	1.61x	1.30x	1.61x	1.80	1.48x
PLD	1.44x	1.06x	2.00x	1.07x	1.55x	1.45x	1.55	1.42x
Lookahead	1.32x	1.08x	1.20x	1.16x	1.54x	1.15x	1.61	1.24x

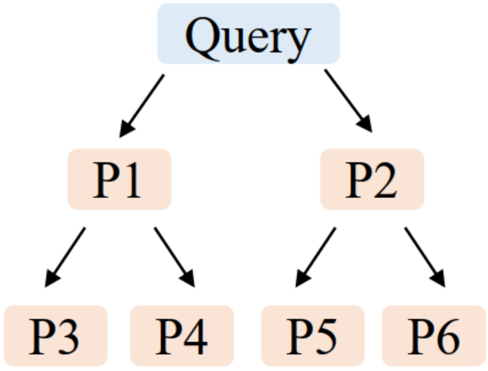
EAGLE-2: Faster Inference of Language Models with Dynamic Draft Trees

Yuhui Li, Fangyun Wei, Chao Zhang, Hongyang Zhang

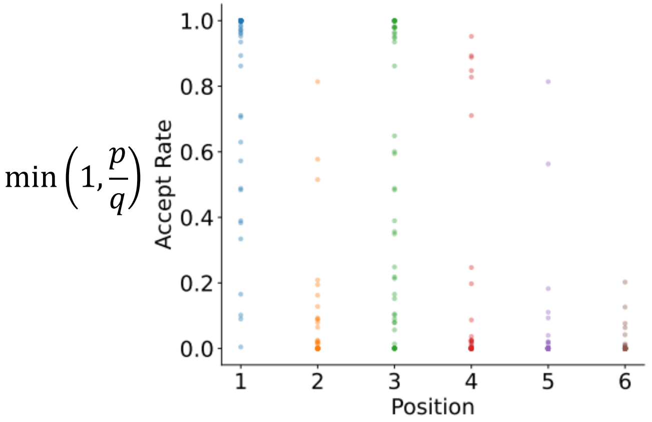
(EMNLP 2024)

Static tree structure?





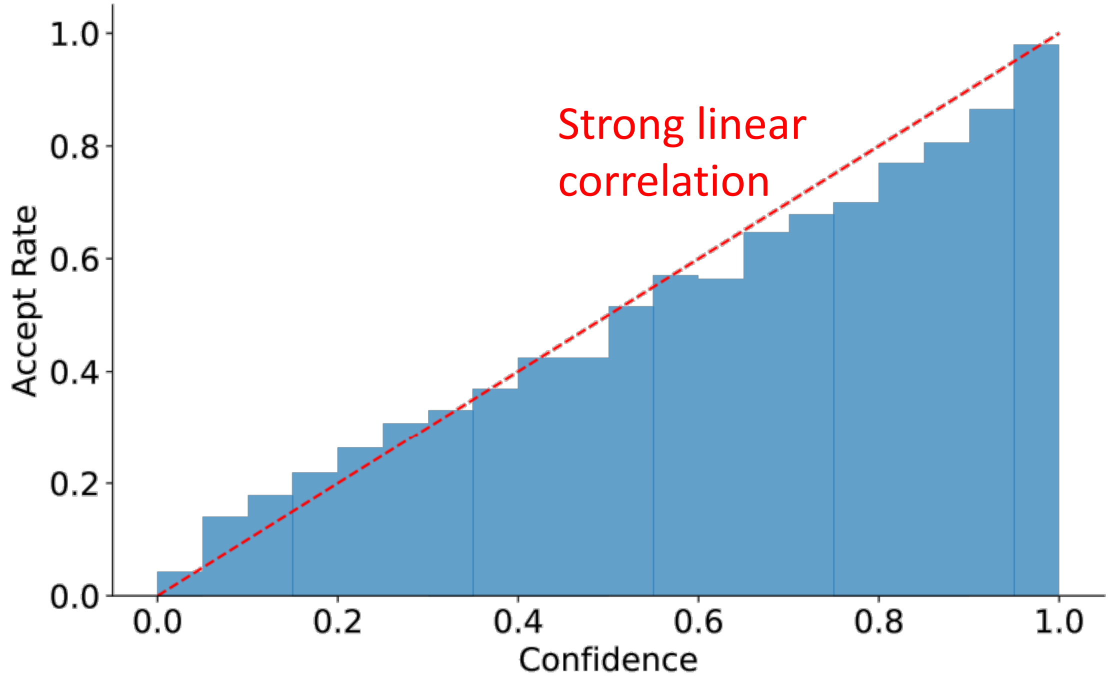
(a) Draft tree structure.



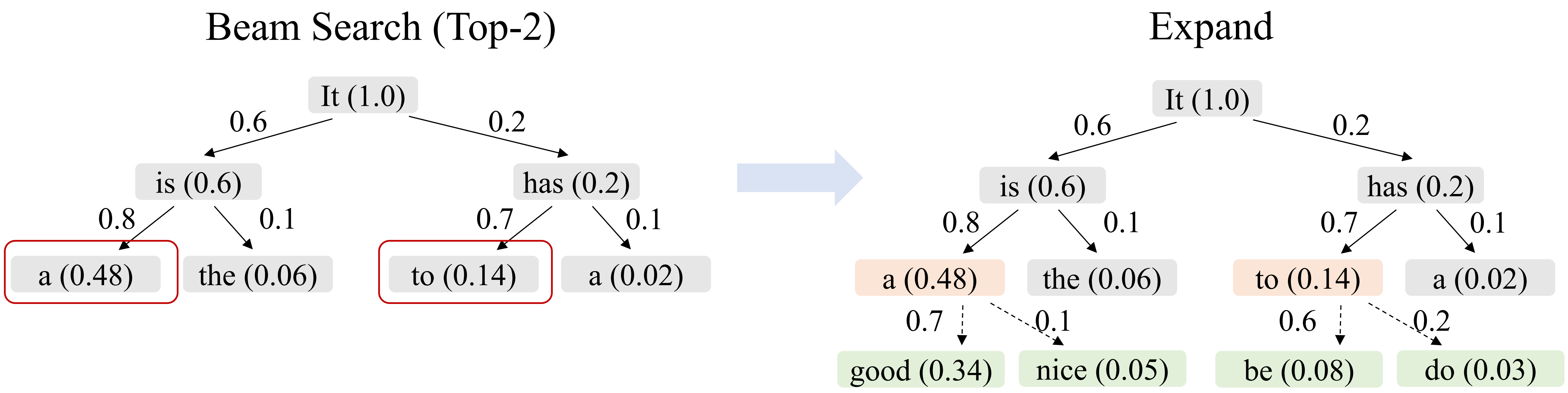
(b) Acceptance rates of tokens at different positions, with each point representing a query.

How to determine importance of each node?

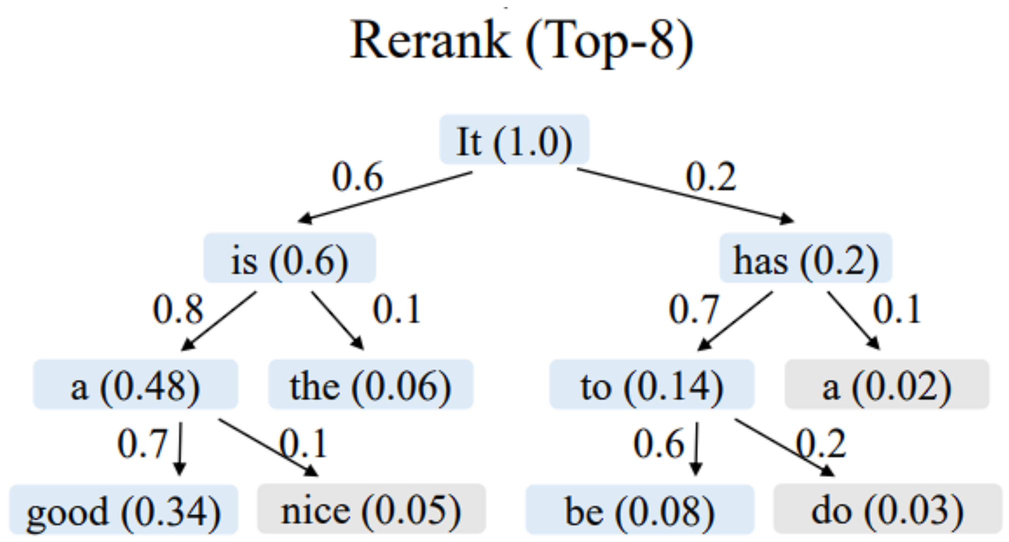
- Accept rate $\min\left(1, \frac{p(t)}{q(t)}\right)$
- However, it requires the computation from the original large model p
- The confidence of the draft model q is a good approximation



Context-aware dynamic draft tree



Context-aware dynamic draft tree



Flatten to 1D

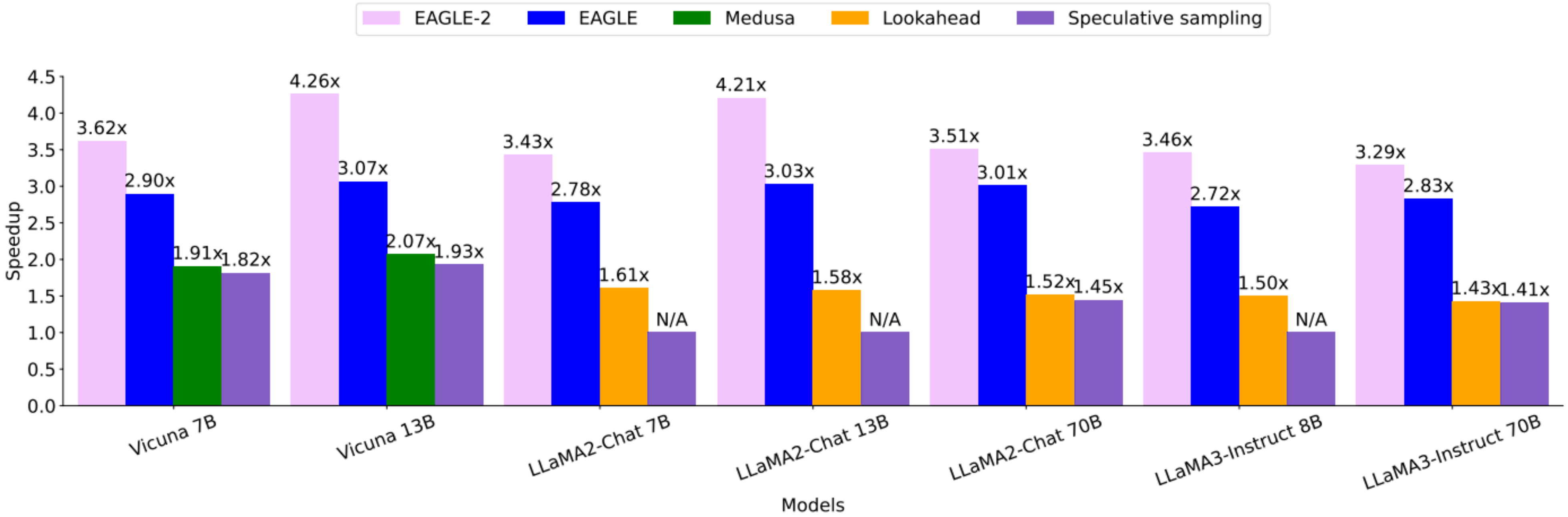
It	is	has	a	the	to	good	be

Attention mask

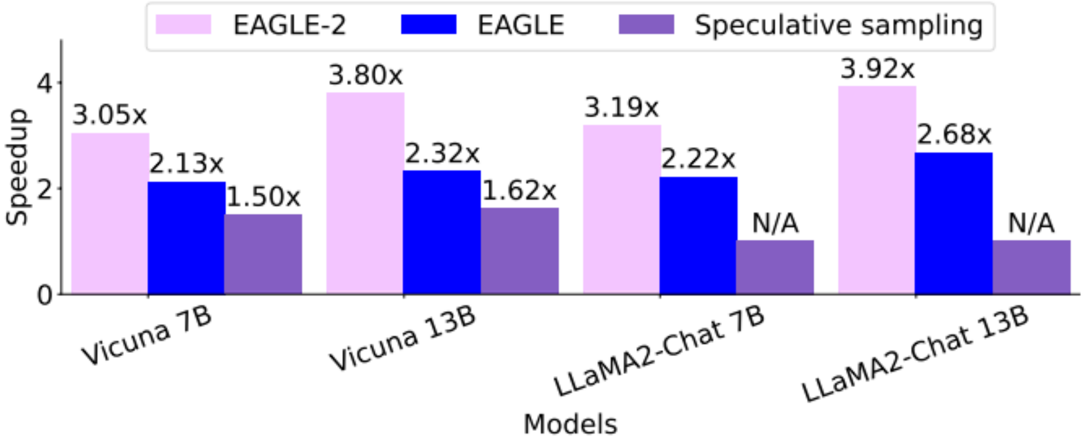
	It	is	has	a	the	to	good	be
It	✓							
is	✓	✓						
has	✓		✓					
a	✓	✓		✓				
the	✓	✓			✓			
to	✓		✓			✓		
good	✓	✓		✓			✓	
be	✓		✓			✓		✓

Performance on MT-bench

Greedy
T=0



Non-greedy
T=1



Performance (T=0, bs=1)

		MT-bench		HumanEval		GSM8K		Alpaca		CNN/DM		Natural Ques.		Mean	
Model	Method	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens
V 13B	SpS	1.93x	2.27	2.23x	2.57	1.77x	2.01	1.76x	2.03	1.93x	2.33	1.66x	1.88	1.88x	2.18
	PLD	1.58x	1.63	1.85x	1.93	1.68x	1.73	1.16x	1.19	2.42x	2.50	1.14x	1.17	1.64x	1.69
	Medusa	2.07x	2.59	2.50x	2.78	2.23x	2.64	2.08x	2.45	1.71x	2.09	1.81x	2.10	2.07x	2.44
	Lookahead	1.65x	1.69	1.71x	1.75	1.81x	1.90	1.46x	1.51	1.46x	1.50	1.36x	1.39	1.58x	1.62
	Hydra	2.88x	3.65	3.28x	3.87	2.93x	3.66	2.86x	3.53	2.05x	2.81	2.11x	2.88	2.69x	3.40
	EAGLE	3.07x	3.98	3.58x	4.39	3.08x	3.97	3.03x	3.95	2.49x	3.52	2.42x	3.11	2.95x	3.82
	EAGLE-2	**4.26x**	**4.83**	**4.96x**	**5.41**	**4.22x**	**4.79**	**4.25x**	**4.89**	**3.40x**	**4.21**	**3.13x**	**3.74**	**4.04x**	**4.65**
L2 13B	PLD	1.42x	1.46	1.63x	1.70	1.41x	1.44	1.16x	1.20	1.42x	1.45	1.12x	1.15	1.36x	1.40
	Lookahead	1.58x	1.64	1.80x	1.85	1.65x	1.69	1.47x	1.50	1.46x	1.53	1.42x	1.45	1.56x	1.61
	EAGLE	3.03x	3.90	3.76x	4.52	3.20x	4.03	3.01x	3.83	2.70x	3.59	2.83x	3.47	3.09x	3.89
	EAGLE-2	**4.21x**	**4.75**	**5.00x**	**5.52**	**4.31x**	**4.90**	**4.13x**	**4.61**	**3.45x**	**4.24**	**3.51x**	**4.04**	**4.10x**	**4.68**
V 7B	SpS	1.82x	2.36	1.99x	2.61	1.71x	2.26	1.65x	2.21	1.81x	2.44	1.60x	2.16	1.76x	2.34
	PLD	1.61x	1.68	1.82x	1.87	1.82x	1.99	1.21x	1.31	2.53x	2.72	1.23x	1.44	1.70x	1.84
	Medusa	1.91x	2.52	2.02x	2.67	1.89x	2.59	1.79x	2.48	1.42x	2.02	1.51x	2.09	1.76x	2.40
	Lookahead	1.63x	1.69	1.72x	1.77	1.84x	1.99	1.38x	1.57	1.44x	1.53	1.45x	1.60	1.58x	1.69
	Hydra	2.69x	3.60	2.98x	3.79	2.73x	3.66	2.66x	3.58	2.01x	2.70	2.25x	2.86	2.55x	3.37
	EAGLE	2.90x	3.94	3.33x	4.29	3.01x	4.00	2.79x	3.89	2.33x	3.42	2.31x	3.21	2.78x	3.79
	EAGLE-2	**3.62x**	**4.98**	**3.95x**	**5.33**	**3.63x**	**4.97**	**3.46x**	**4.86**	**2.94x**	**4.12**	**2.76x**	**3.82**	**3.39x**	**4.68**
L2 7B	PLD	1.38x	1.43	1.52x	1.59	1.32x	1.37	1.15x	1.19	1.48x	1.52	1.15x	1.20	1.33x	1.38
	Lookahead	1.61x	1.66	1.72x	1.77	1.58x	1.65	1.49x	1.52	1.49x	1.54	1.48x	1.53	1.56x	1.61
	EAGLE	2.78x	3.62	3.17x	4.24	2.91x	3.82	2.78x	3.71	2.43x	3.41	2.61x	3.44	2.78x	3.71
	EAGLE-2	**3.43x**	**4.70**	**4.03x**	**5.39**	**3.52x**	**4.77**	**3.45x**	**4.66**	**3.01x**	**4.12**	**3.15x**	**4.19**	**3.43x**	**4.64**

Performance (T=1, bs=1)

		MT-bench		HumanEval		GSM8K		Alpaca		CNN/DM		Natural Ques.		Mean	
Model	Method	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens	Speedup	#mean accepted tokens
V 13B	SpS	1.62x	1.84	1.72x	1.97	1.46x	1.73	1.52x	1.78	1.66x	1.89	1.43x	1.70	1.55x	1.82
	EAGLE	2.32x	3.20	2.65x	3.63	2.57x	3.60	2.45x	3.57	2.23x	3.26	2.14x	3.06	2.39x	3.39
	EAGLE-2	**3.80x**	**4.40**	**4.22x**	**4.89**	**3.77x**	**4.41**	**3.78x**	**4.37**	**3.25x**	**3.97**	**3.07x**	**3.54**	**3.65x**	**4.26**
L2 13B	EAGLE	2.68x	3.45	2.89x	3.78	2.82x	3.67	2.66x	3.55	2.41x	3.39	2.37x	3.31	2.64x	3.53
	EAGLE-2	**3.92x**	**4.51**	**4.58x**	**5.29**	**4.21x**	**4.80**	**3.85x**	**4.48**	**3.31x**	**4.08**	**3.43x**	**3.89**	**3.88x**	**4.51**
V 7B	SpS	1.50x	1.87	1.55x	1.95	1.53x	1.82	1.56x	1.85	1.63x	1.91	1.33x	1.72	1.52x	1.85
	EAGLE	2.13x	3.17	2.39x	3.43	2.34x	3.29	2.21x	3.30	2.08x	3.12	1.95x	2.86	2.18x	3.20
	EAGLE-2	**3.05x**	**4.28**	**3.33x**	**4.65**	**3.07x**	**4.49**	**3.08x**	**4.43**	**2.63x**	**3.76**	**2.48x**	**3.56**	**2.94x**	**4.20**
L2 7B	EAGLE	2.22x	3.30	2.61x	3.79	2.40x	3.52	2.29x	3.33	2.19x	3.15	2.22x	3.12	2.32x	3.37
	EAGLE-2	**3.19x**	**4.41**	**3.67x**	**5.06**	**3.35x**	**4.62**	**3.20x**	**4.48**	**2.73x**	**3.85**	**2.81x**	**4.01**	**3.15x**	**4.41**

Vanilla on A100 vs EAGLE-2 on RTX3060

Vanilla	EAGLE-2
A100 （$10000)	RTX 3060 （2×$300)

Vanilla

Speed: 26.06 tokens/s

Compression Ratio: 1.00

Introduce artificial intelligence to me.

I'm excited

EAGLE-2

Speed: 19.61 tokens/s

Compression Ratio: 5.00

Introduce artificial intelligence to me.

I'm excited to introduce

Third-party evaluations
(updated on Oct. 25, 2024)

⚡ ***Unlocking Efficiency in Large Language Model Inference:***
A Comprehensive Survey of Speculative Decoding

Heming Xia[1], Zhe Yang[2], Qingxiu Dong[2], Peiyi Wang[2], Yongqi Li[1], Tao Ge[3], Tianyu Liu[4], Wenjie Li[1], Zhifang Sui[2]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Microsoft Research Asia [4]Alibaba Group

Leaderboard on 3090

- Device: a single NVIDIA GeForce RTX 3090 GPU (24GB) with 12 CPU cores
- Testing environment: Pytorch 2.0.1, under CUDA 11.8
- Experimental Settings: Vicuna-7B-v1.3, greedy decoding, FP16 precision, batch size = 1

Models	Multi-turn Conversation	Translation	Summa-rization	Question Answering	Mathematical Reasoning	Retrieval-aug. Generation	#Mean Accepted Tokens	Overall
EAGLE2 🥇	**2.71x**	**1.82x**	2.19x	**2.11x**	**2.71x**	**1.91x**	**4.36**	**2.25x**
EAGLE 🥈	2.44x	1.81x	2.13x	2.11x	2.54x	1.82x	3.57	2.16x
SpS 🥉	1.98x	1.37x	2.00x	1.95x	1.89x	1.76x	2.29	1.83x
Hydra	2.04x	1.67x	1.56x	1.81x	2.16x	1.48x	3.26	1.80x
PLD	1.57x	1.07x	**2.31x**	1.25x	1.62x	1.56x	1.74	1.55x
Medusa	1.60x	1.38x	1.28x	1.46x	1.64x	1.22x	2.32	1.44x
REST	1.49x	1.18x	1.21x	1.46x	1.35x	1.27x	1.63	1.32x
Lookahead	1.13x	0.97x	1.05x	1.07x	1.29x	0.98x	1.65	1.08x

EAGLE in the community

- **Ongoing:** EAGLE is being merged with SGLang.
- **2024.8.23**: EAGLE is merged with vLLM.
- **2024.5.25**: EAGLE is merged with Intel® LLM Library for PyTorch.
- **2024.5.9**: EAGLE is merged with Intel® Extension for Transformers.

How to use?

- Code: https://github.com/SafeAILab/EAGLE

```
from eagle.model.ea_model import EaModel
model = EaModel.from_pretrained(base_model_path=base_model_path,
            ea_model_path=EAGLE_model_path,
            torch_dtype=torch.float16)
output_ids = model.eagenerate(input_ids,temperature=0.5,max_new_tokens=512)
```

- EAGLE-3 is on the way. Please stay tuned!
    - Focus on large batch size and throughput

Thanks!
Q&A

Code