



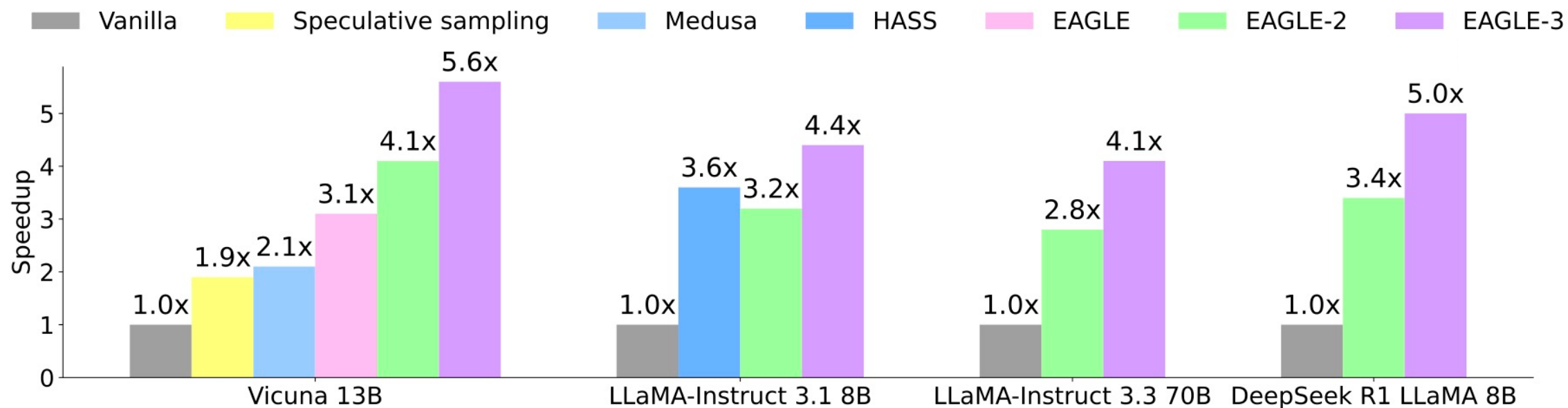
EAGLE-3: Scaling up Inference Acceleration of Large Language Models via Training-Time Test

Hongyang Zhang

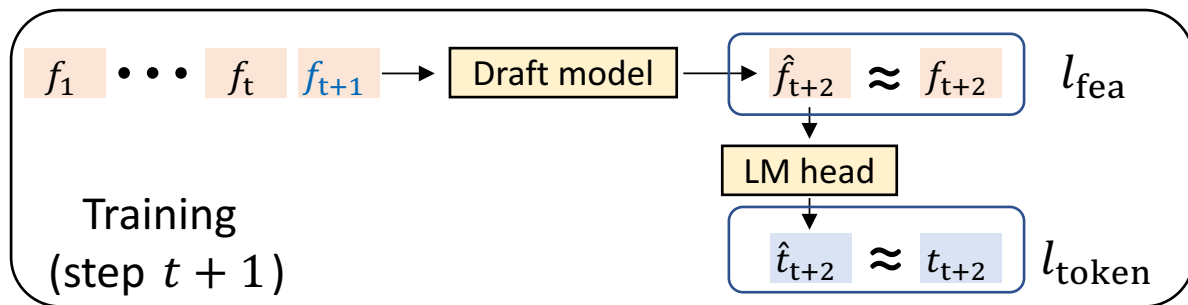
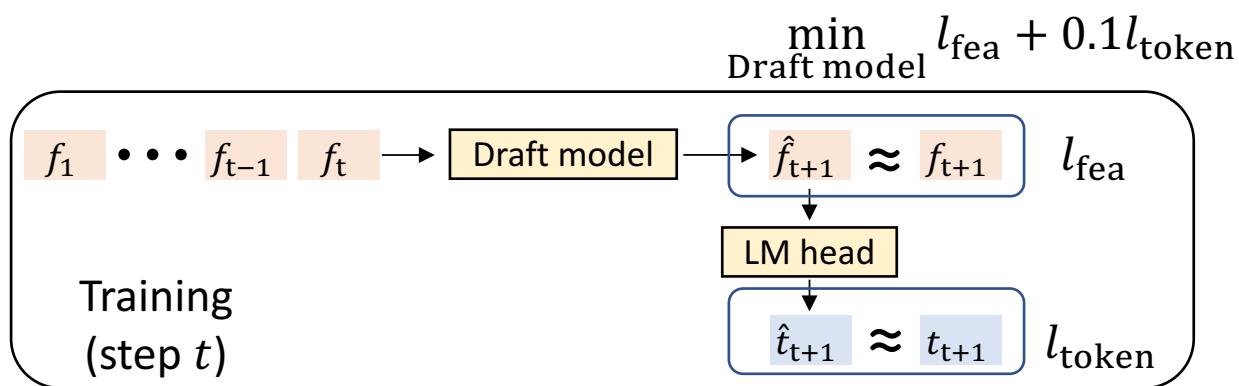
University of Waterloo & Vector Institute & EAGLE AI CORP

Joint work with Yuhui Li, Fangyun Wei, Chao Zhang

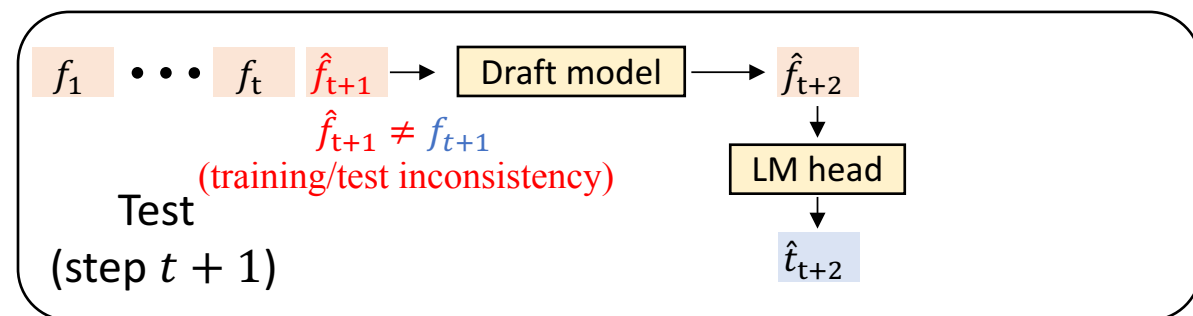
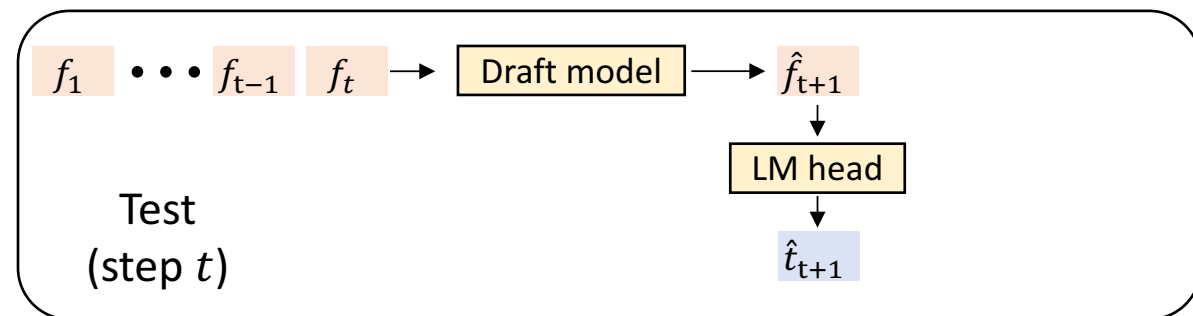
Benchmarking on MT-Bench



Inconsistency in EAGLE-1 Training/Test

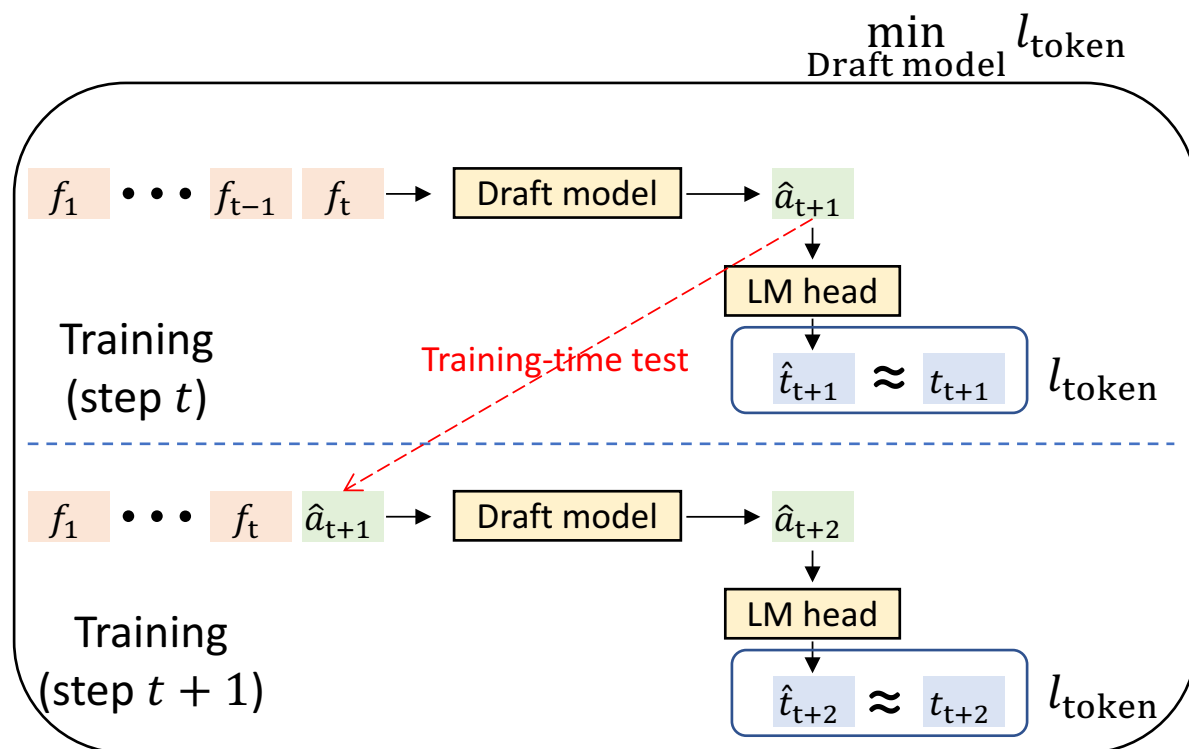


EAGLE-1 training

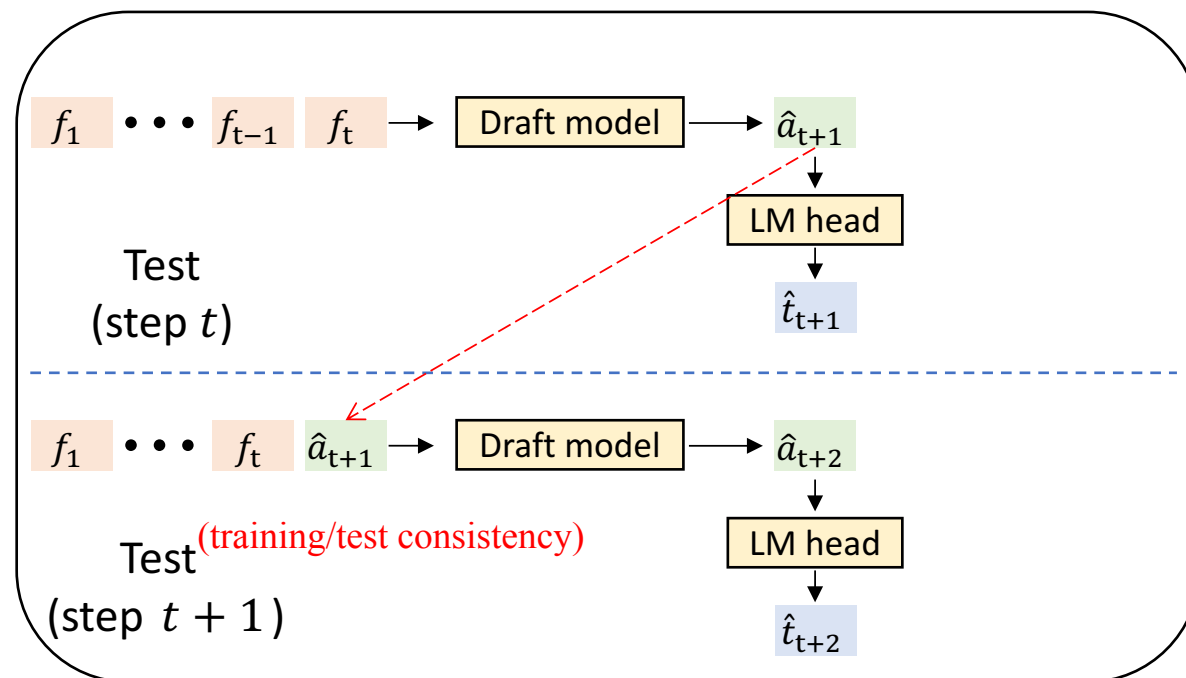


EAGLE-1 test

EAGLE-3 Training-Time Test

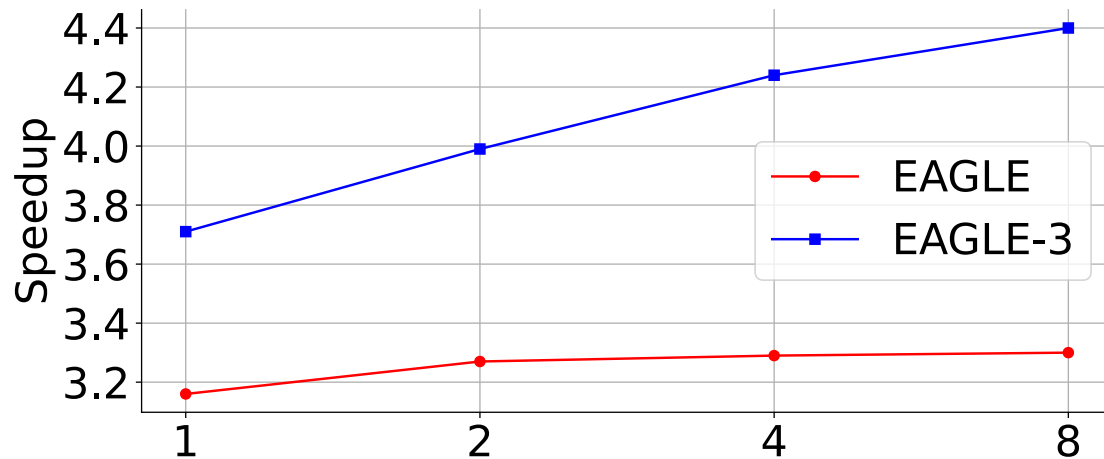


EAGLE-3 training

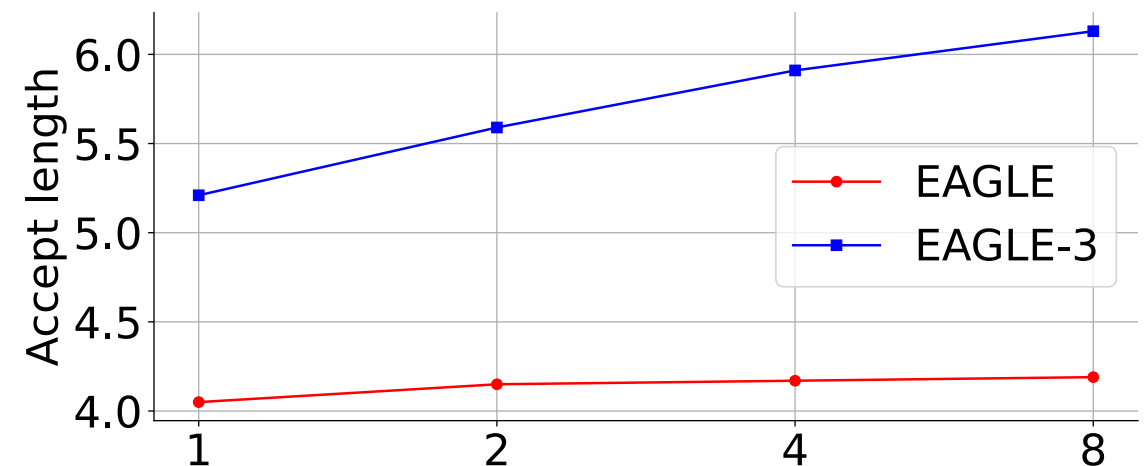


EAGLE-3 test

New Scaling Law for Inference Acceleration



Amount of training data (relative to ShareGPT)



Amount of training data (relative to ShareGPT)

Trained on UltraChat + ShareGPT

Performance (bs=1)

		MT-bench		HumanEval		GSM8K		Alpaca		CNN/DM		Mean	
Model	Method	Speedup	τ	Speedup	τ	Speedup	τ	Speedup	τ	Speedup	τ	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

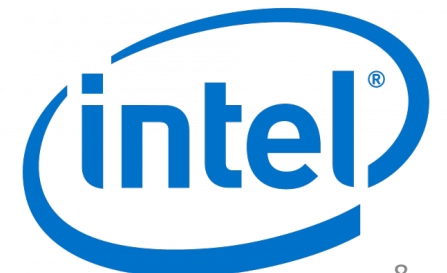
Large Batch (1x = SGLang w/o EAGLE)

Tested by the SGLang team on MT-Bench and LLama 3.1 8B

Batch size	2	4	8	16	24	32	48	56	64
EAGLE	1.40x	1.38x	1.23x	1.02x	0.93x	0.94x	0.88x	0.99x	0.99x
EAGLE-3	1.81x	1.82x	1.62x	1.48x	1.39x	1.32x	1.38x	1.34x	1.38x

EAGLE-3 in the community

- [SGLang, SpecForge](#)
- [vLLM](#)
- [NVIDIA TensorRT-LLM](#)
- [AMD ROCm](#)
- [DeepSeek V3 \(MTP\)](#)
- [AWS NeuronX Distributed Core](#)
- [Intel® Extension for Transformers](#)
- [Intel® LLM Library for PyTorch](#)
- [MLC-LLM](#)



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)
```

Thanks!