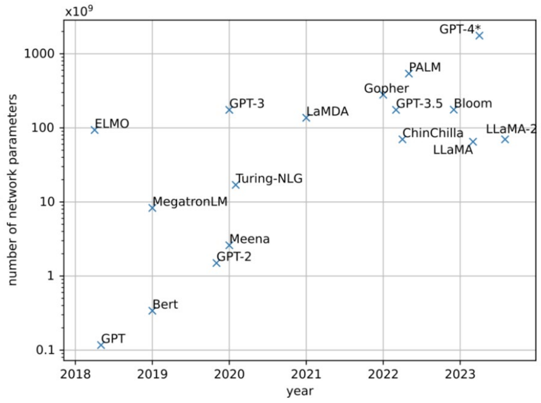
UCLA

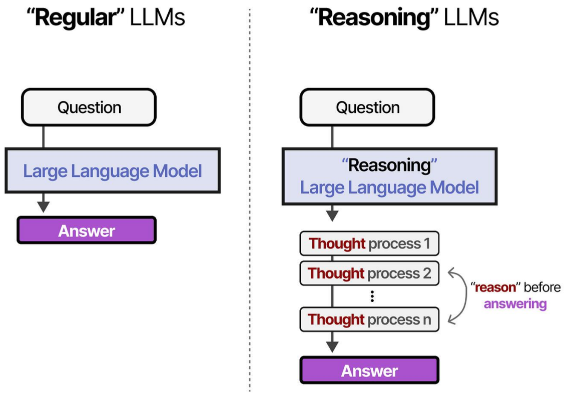
Optimized Multi-Token Joint Decoding with Auxiliary Model for LLM Inference

Zongyue Qin, Ziniu Hu, Zifan He, Neha Prakriya, Jason Cong, Yizhou Sun
University of California Los Angeles

LLM has become increasingly more popular

However, the time and energy consumption of LLM inference also increases significantly.

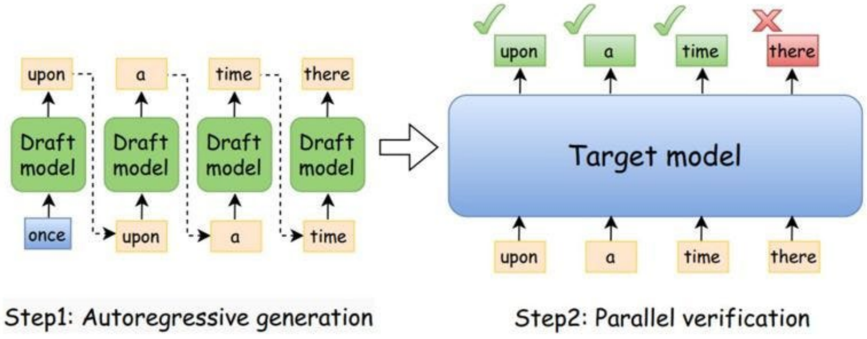




Left figure is from https://www.researchgate.net/figure/Number-of-parameters-of-LLM-over-the-past-five-years-Significant-advances-were-made-by_fig1_377469845
Right figure is from https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms

Speculative Decoding (SpD)

- Speculative decoding accelerates LLM inference speed by exploiting a small auxiliary LM.
- The small LM generates K draft tokens, then the large LM **verifies them in parallel**, and re-samples the first rejected token.



This figure is from https://medium.com/@genai.works/speed-up-llm-inference-with-speculative-decoding-1fc79701e9d6

SpD vs Our Goal

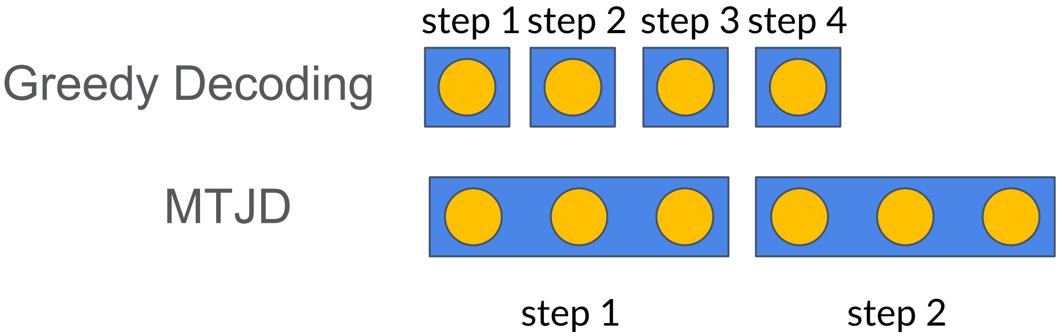
SPECULATIVE DECODING

- Energy Efficiency > Greedy Decoding
- Time Efficiency > Greedy Decoding
- Output Quality = Greedy Decoding

OUR GOAL

- Energy Efficiency > Greedy Decoding
- Time Efficiency > Greedy Decoding
- Output Quality > Greedy Decoding

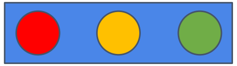
Multi-Token Joint Decoding (MTJD)



- MTJD selects multiple token based their joint conditional likelihood.
- Better effectiveness than greedy decoding.
- However, it is not efficient.
  - Solution: Speculative Decoding

Multi-Token Assisted Decoding (MTAD)

Small model generates K draft tokens via beam sampling

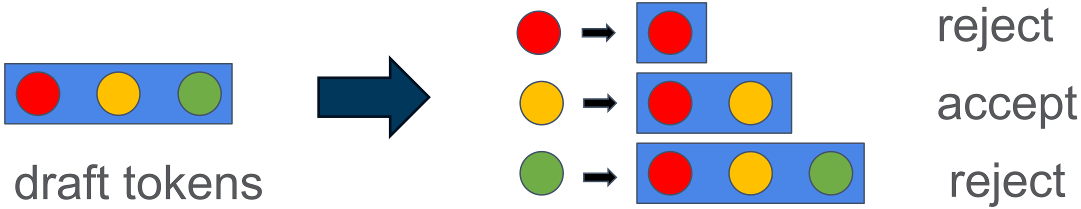


K draft tokens

Multi-Token Assisted Decoding (MTAD)

Accept each prefix sub-sequence iff $\frac{p(x_{1:k})}{q(x_{1:k})} \geq \tau$

p: target distribution, q: draft distribution, τ: pre-defined threshold



K draft tokens

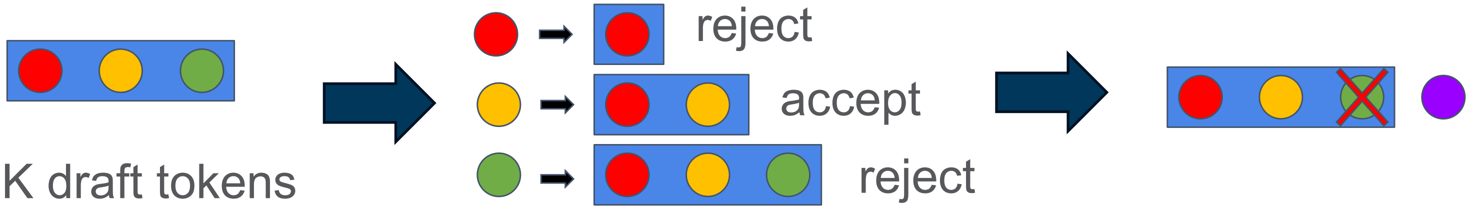
reject

accept

reject

Multi-Token Assisted Decoding (MTAD)

Select the longest accepted sequence and sample an additional token from target distribution p.



The perplexity ratio between approximate MTAD and exact MTJD can be bounded.

Experiment

Target and Draft Models

- Llama-3 (8B, 1B)
- Llama-3-Instruct (8B, 1B)

Datasets

- Spider (text-to-SQL)
- MTBench (various tasks)
- HumanEval (coding)

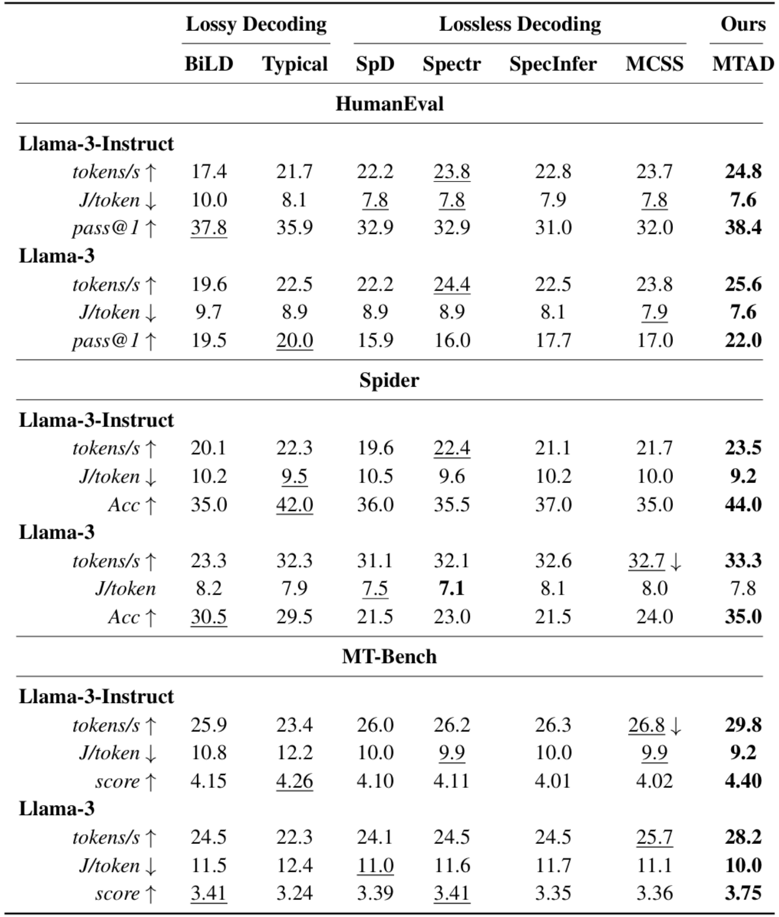
Metrics

- Speed (tokens/s)
- Energy (J/token)
- Output Quality
  - Spider: Execution accuracy
  - MT-Bench: LLM-evaluated score
  - HumanEval: Pass@1

Experiment Results

Compared to SpD, MTAD

- ✓ Improve output quality by **25%**
- ✓ Achieve **1.42x** speed-up
- ✓ Consume **1.54x** less energy



	Lossy Decoding		Lossless Decoding				Ours
	BiLD	**Typical**	**SpD**	**Spectr**	**SpecInfer**	**MCSS**	**MTAD**
				HumanEval			
Llama-3-Instruct							
tokens/s ↑	17.4	21.7	22.2	23.8	22.8	23.7	**24.8**
J/token ↓	10.0	8.1	7.8	7.8	7.9	7.8	**7.6**
pass@1 ↑	37.8	35.9	32.9	32.9	31.0	32.0	**38.4**
Llama-3							
tokens/s ↑	19.6	22.5	22.2	24.4	22.5	23.8	**25.6**
J/token ↓	9.7	8.9	8.9	8.9	8.1	7.9	**7.6**
pass@1 ↑	19.5	20.0	15.9	16.0	17.7	17.0	**22.0**
				Spider			
Llama-3-Instruct							
tokens/s ↑	20.1	22.3	19.6	22.4	21.1	21.7	**23.5**
J/token ↓	10.2	9.5	10.5	9.6	10.2	10.0	**9.2**
Acc ↑	35.0	42.0	36.0	35.5	37.0	35.0	**44.0**
Llama-3							
tokens/s ↑	23.3	32.3	31.1	32.1	32.6	32.7 ↓	**33.3**
J/token	8.2	7.9	7.5	**7.1**	8.1	8.0	7.8
Acc ↑	30.5	29.5	21.5	23.0	21.5	24.0	**35.0**
				MT-Bench			
Llama-3-Instruct							
tokens/s ↑	25.9	23.4	26.0	26.2	26.3	26.8 ↓	**29.8**
J/token ↓	10.8	12.2	10.0	9.9	10.0	9.9	**9.2**
score ↑	4.15	4.26	4.10	4.11	4.01	4.02	**4.40**
Llama-3							
tokens/s ↑	24.5	22.3	24.1	24.5	24.5	25.7	**28.2**
J/token ↓	11.5	12.4	11.0	11.6	11.7	11.1	**10.0**
score ↑	3.41	3.24	3.39	3.41	3.35	3.36	**3.75**

Thank You