Polybasic Speculative Decoding Through a Theoretical Perspective

Ruilin Wang[1], Huixia Li[2], Yuexiao Ma[1 2], Xiawu Zheng[1 3 4], Fei Chao[1], Xuefeng Xiao[2], Rongrong Ji[1 3]

ICML International Conference On Machine Learning

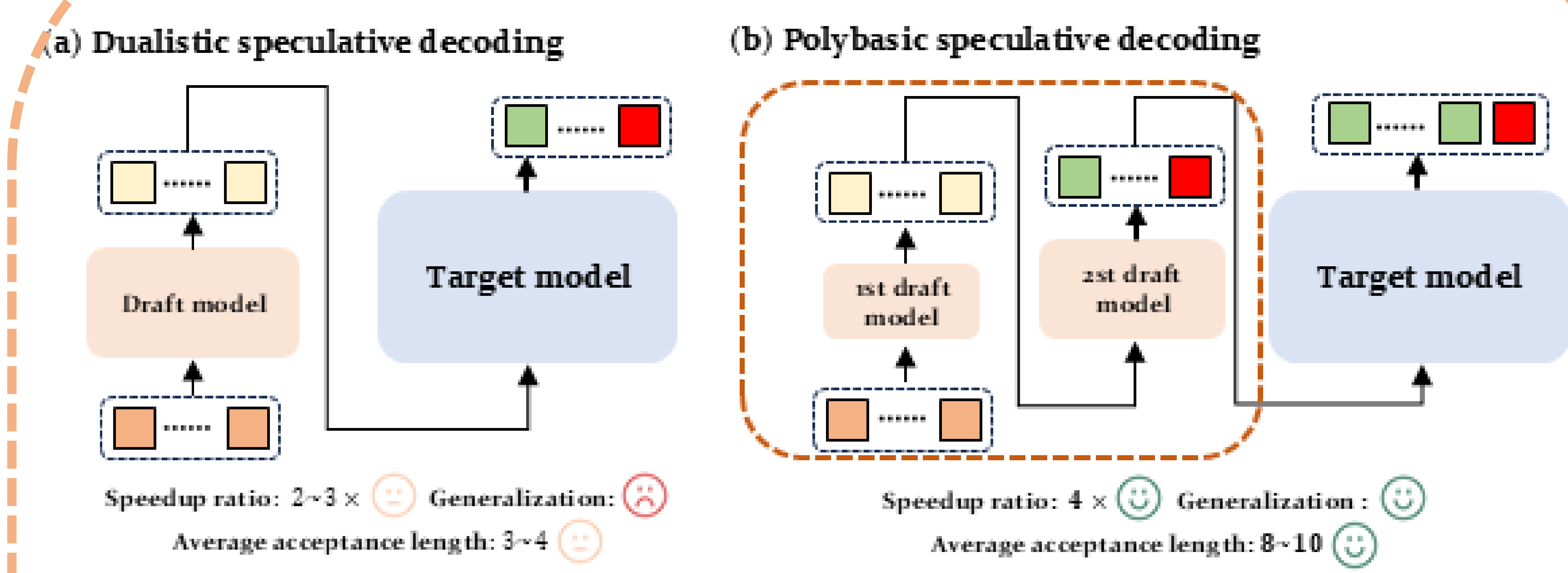
厦門大學 XIAMEN UNIVERSITY

Abstract

Inference latency stands as a critical bottleneck in the large-scale deployment of Large Language Models (LLMs). Speculative decoding methods have recently shown promise in accelerating inference without compromising the output distribution. However, existing work typically relies on a dualistic draft-verify framework and lacks rigorous theoretical grounding. In this paper, we introduce a novel polybasic speculative decoding framework, underpinned by a comprehensive theoretical analysis. Specifically, we prove a fundamental theorem that characterizes the optimal inference time for multi-model speculative decoding systems, shedding light on how to extend beyond the dualistic approach to a more general polybasic paradigm. Through our theoretical investigation of multi-model token generation, we expose and optimize the interplay between model capabilities, acceptance lengths, and overall computational cost. **Our framework supports both standalone implementation and integration with existing speculative techniques, leading to accelerated performance in practice.** Experimental results across multiple model families demonstrate that our approach yields speedup ratios ranging from 3.31× to 4.01× for LLaMA2-Chat 7B, up to 3.87× for LLaMA3-8B, up to 4.43× for Vicuna7B and up to 3.85× for Qwen2-7B—all while preserving the original output distribution. We release our theoretical proofs and implementation code to facilitate further investigation into polybasic speculative decoding.

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
[2]ByteDance Inc, [3]Institute of Artificial Intelligence, Xiamen University, [4]Peng Cheng Laboratory, Shenzhen, China

Framework



Theoretical Foundations

We establish fundamental properties of **polybasic speculative decoding** that govern how additional models impact computational cost and acceptance lengths. Our analysis focuses on two main aspects: (i) optimal inference time and (ii) stability of acceptance lengths.

Theorem 3.1 (Optimal Inference Time)

For an n-model polybasic system generating N tokens, the total inference time T is expressed as

$$T = \sum_{i=1}^{n-1} \frac{N}{L_i} \cdot T_i + \beta \cdot \frac{N}{L_{n-1}} T_n$$

Theorem 3.2 (Model Insertion Efficiency)

Adding M_{new} between M_i and M_{i+1} decreases total inference time if and only if it achieves a sufficiently large increase in acceptance lengths, balanced against its forward-pass cost T_{new} Concretely, if L_{new} is the acceptance length when verifying tokens from M_{new} against M_i, and L'_{I+1} is the acceptance length from M'_{I+1} 's perspective, then improvement occurs if:

$$\frac{T_{new}}{T_i} < \mathrm{L}_{new}\left(\frac{1}{L_i} - \frac{1}{L_{i-new}}\right) \; or \; \frac{T_{new}}{T_{i+1}} < \beta\left(\frac{L_{new-(i+1)}}{L_i} - 1\right)$$

Theorem 3.3 (Sampling Stability)

In the model chain using speculative sampling can ensure stable acceptance lengths.

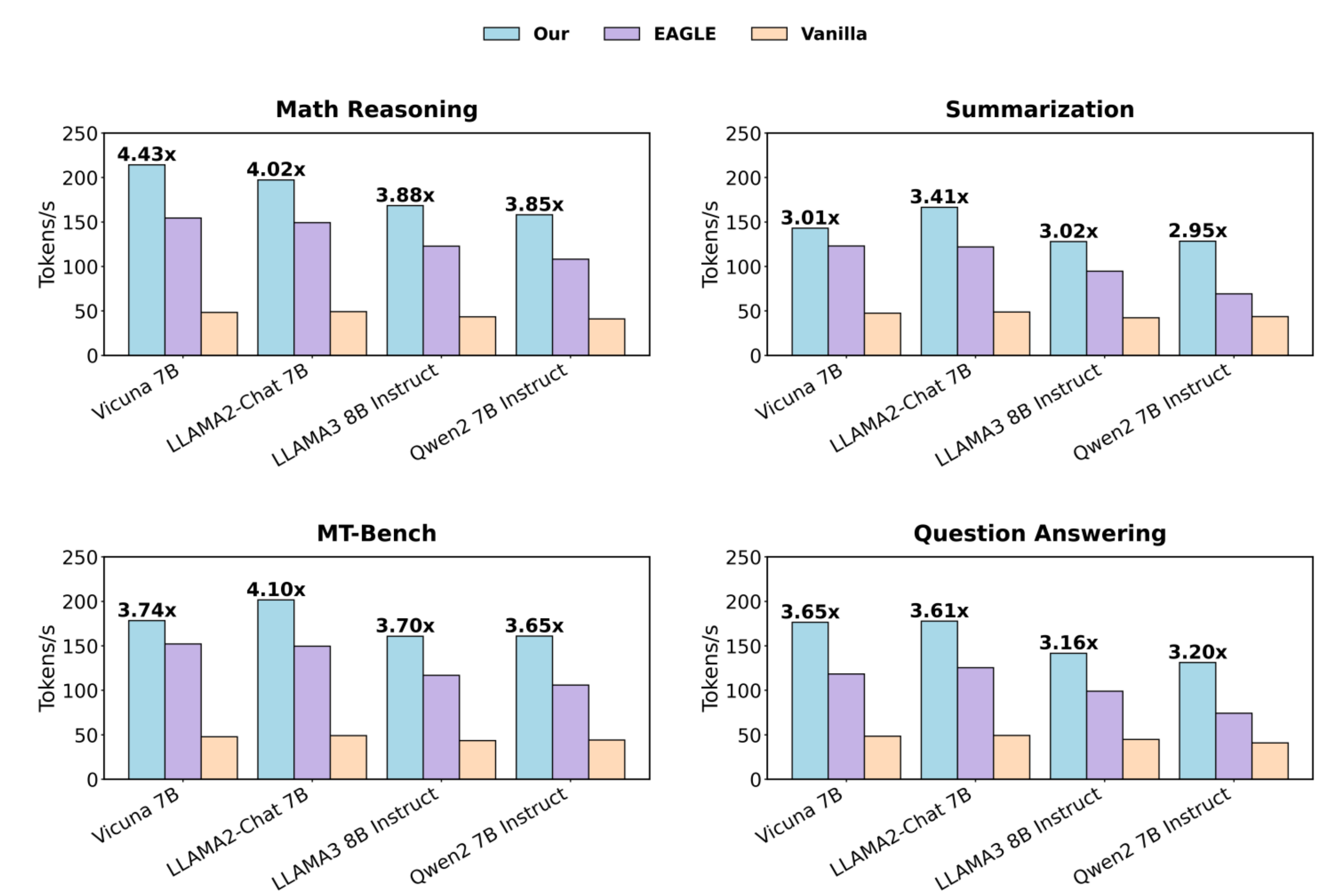
Experiment

Table 1. Theoretical Validation via Model Insertion

Case	T_i (ms)	$L_{i\text{-new}}$	T_{new} (ms)	L_{new}	T_{i+1} (ms)	L_i	Speedup
Non-compliant	22	3.83	17.61	3.77	4	4.34	2.61× → 1.08×
Compliant	22	6.26	7.00	4.67	4	4.34	2.61× → 3.48×
CS Drafting	47.52	3.50	19.16	3.02	12.42	2.28	3.19×→3.88×

Table 2. Average acceptance length (μ) and speedup ratio (c) on different tasks. V7B: Vicuna-7B, L3-8B: LLaMA3-8B-Instruct, L2-7B: LLaMA2-Chat-7B, Q2-7B: Qwen2-7B-Instruct.

		MT		Trans.		Sum.		QA		Math		RAG		Overall	
	Model	c	μ	c	μ	c	μ	c	μ	c	μ	c	μ	c	μ
Our	V7B	**3.77x**	**11.22**	**3.07x**	**7.76**	**3.01x**	**10.24**	**3.65x**	**9.53**	**4.43x**	**10.28**	**2.98x**	**10.30**	**3.48x**	**9.88**
	L3-8B	**3.70x**	**9.97**	**3.39x**	**8.86**	**3.02x**	**9.38**	**3.16x**	**9.08**	**3.87x**	**10.08**	**2.71x**	**9.24**	**3.31x**	**9.44**
	L2-7B	**4.10x**	**10.47**	**3.46x**	**9.15**	**3.41x**	**9.86**	**3.61x**	**9.49**	**4.02x**	**9.99**	**3.31x**	**10.08**	**3.66x**	**9.84**
	Q2-7B	**3.65x**	**9.85**	**3.15x**	**8.65**	**2.95x**	**9.15**	**3.25x**	**8.95**	**3.85x**	**9.95**	**2.85x**	**9.35**	**3.28x**	**9.32**
EAGLE2	V7B	3.19x	4.76	2.07x	3.22	2.59x	3.96	2.45x	3.71	3.19x	4.72	2.15x	3.95	2.61x	4.34
	L3-8B	2.69x	3.99	2.37x	3.53	2.23x	3.58	2.21x	3.42	2.83x	4.20	2.23x	3.95	2.44x	3.82
	L2-7B	3.04x	4.48	2.61x	3.96	2.50x	4.04	2.55x	4.05	3.04x	4.68	2.40x	4.19	2.70x	4.30
	Q2-7B	2.40x	3.74	1.45x	2.45	1.59x	3.06	1.81x	2.91	2.63x	4.26	1.72x	3.27	1.94x	3.51



Communicate us

Paper: https://openreview.net/forum?id=JrxJUMqqz4