Kangaroo： Lossless Self-Speculative Decoding for Accelerating LLMs via Double Early Exiting

Fangcheng Liu[1], Yehui Tang[1], Zhenhua Liu[1], Yunsheng Ni[1], Duyu Tang[2], Kai Han[1]*, Yunhe Wang[1]*

[1] Huawei Noah's Ark Lab [2] Consumer Business Group, Huawei

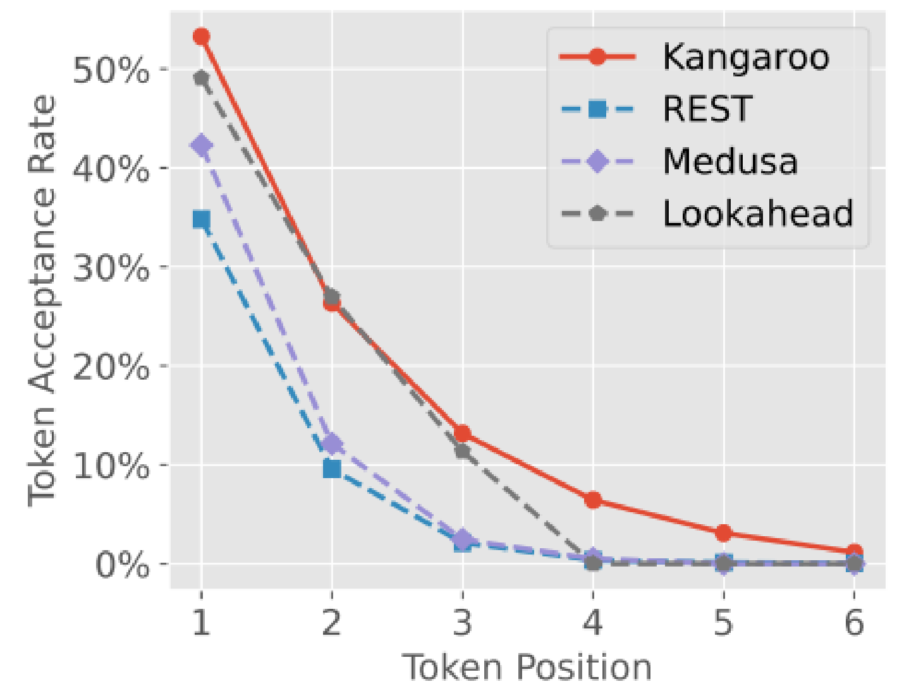
Code

NEURAL INFORMATION PROCESSING SYSTEMS

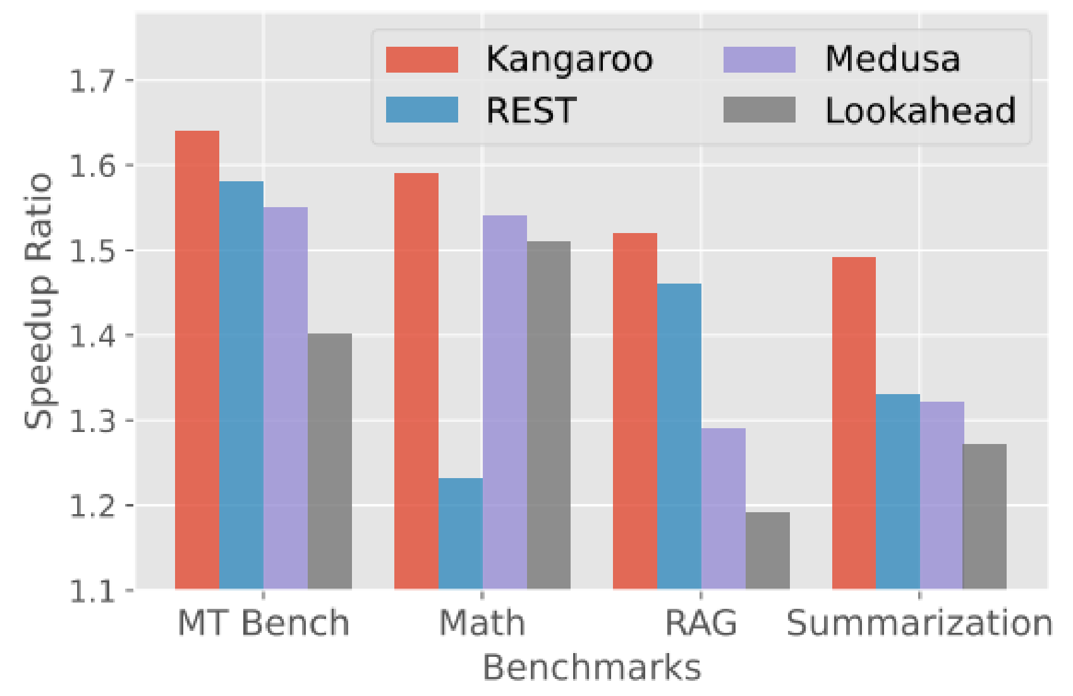
Background: Self-Speculative Decoding

There is a trade-off between token acceptance rate and drafting efficiency

- The conventional approach of training separate draft model to achieve a satisfactory token acceptance rate can be costly and impractical.
- To mitigate these costs, several studies have proposed ***self-drafting*** methods that do not rely on external drafter models.
- Although Medusa and REST could generate draft tokens efficiently, however, the token acceptance rate is not always satisfactory.
- Focusing exclusively on the token acceptance rate without considering the latency of generating draft tokens can lead to suboptimal speedup ratio.



(a) Token acceptance rate on the *mathematical reasoning* subtask. Token position "2" denotes the task to predict the next-next-token.



(b) End-to-end speedup ratio on four subtasks in Spec-Bench. "Math" and "RAG" denote mathematical reasoning and retrieval-augmented generation, respectively.

Kangaroo: Self-Drafting via Double Early Exiting

Early Exiting as Self-Drafting Model

Drawing inspiration from early exiting, we directly extract hidden states from a ***fixed shallow sub-network*** of the target LLM

$$f_t = \mathcal{M}_b[:l](x_t), \quad l \in \{1, 2, \cdots, L\}$$

Note that there is a representation gap between the shallow sub-network and full model. Therefore, we train a ***lightweight and efficient adapter*** to bridge this gap

$$f_t' = f_t + \texttt{MultiHead}(\texttt{LayerNorm}(f_t))$$

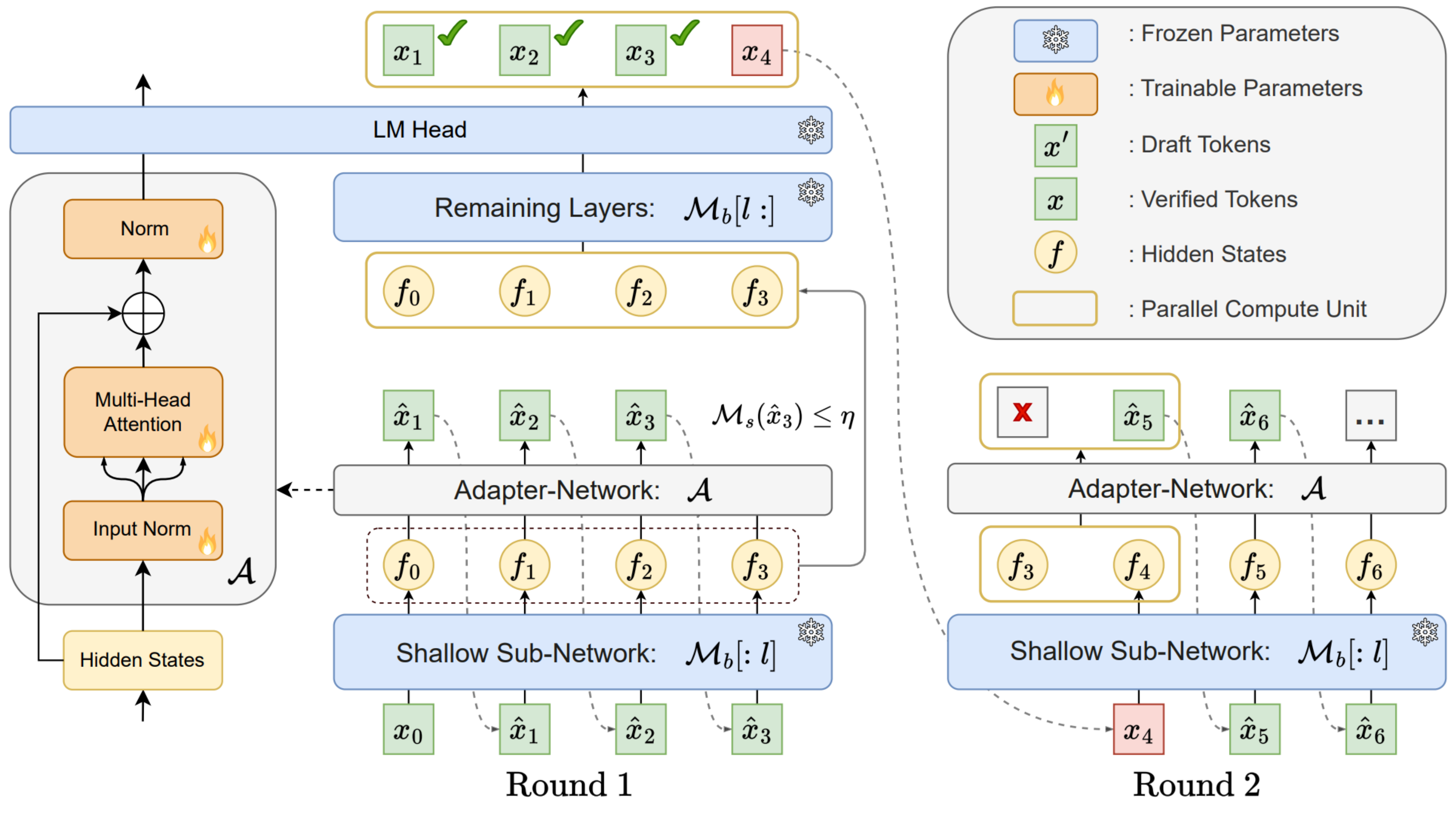
We keep the *residual connection and the* ***multi-head attention*** module but remove the FFN in a standard transformer block. Besides the shallow sub-network, Kangaroo also **reuses the *LM Head*** of the target model to get the final distribution

$$\mathcal{M}_s(x \mid x^t) = \texttt{Softmax}\left(\mathbf{W}^\top \texttt{LayerNorm}(f_t')\right)$$

Finally, we train the adapter with cross entropy loss

$$\mathcal{A}^* = \arg\min_{\mathcal{A}} \sum_t \sum_{x \in \mathcal{X}} -\mathcal{M}_b(x \mid x^t) \log \mathcal{M}_s(x \mid x^t)$$

Framework under Single-Sequence Verification


LM Head
Norm
Multi-Head Attention
Input Norm
$\mathcal{A}$
Hidden States
Remaining Layers: $\mathcal{M}_b[l:]$
f_0 f_1 f_2 f_3
$\hat{x}_1$ $\hat{x}_2$ $\hat{x}_3$
$\mathcal{M}_s(\hat{x}_3) \leq \eta$
Adapter-Network: $\mathcal{A}$
Shallow Sub-Network: $\mathcal{M}_b[:l]$
x_0 $\hat{x}_1$ $\hat{x}_2$ $\hat{x}_3$
x_1 x_2 x_3 x_4
Round 1
: Frozen Parameters
: Trainable Parameters
x' : Draft Tokens
x : Verified Tokens
f : Hidden States
: Parallel Compute Unit
$\hat{x}_5$ $\hat{x}_6$ ...
f_3 f_4 f_5 f_6
x_4 $\hat{x}_5$ $\hat{x}_6$
Round 2

The Second Early-Exit: Dynamic Drafting Steps

As shown in following figure, the difficulty of predicting the next token varies across different contextual scenarios, necessitating the design of a ***token-wise dynamic drafting strategy***. Fortunately, we observe a strong ***correlation*** between the small model's confidence level for the current sampled token and the likelihood that the token will be accepted by the big target model. Therefore, we stop drafting once the top-1 probability on the self-drafting model falls below a predefined threshold:

$$\max_{x \in \mathcal{X}} \mathcal{M}_s(x \mid x^t) \leq \eta$$

We also generalize the early stopping mechanism of Kangaroo in single-sequence decoding to tree decoding, where both the tree depth and node selection at each level of the tree are token-wise dynamic.

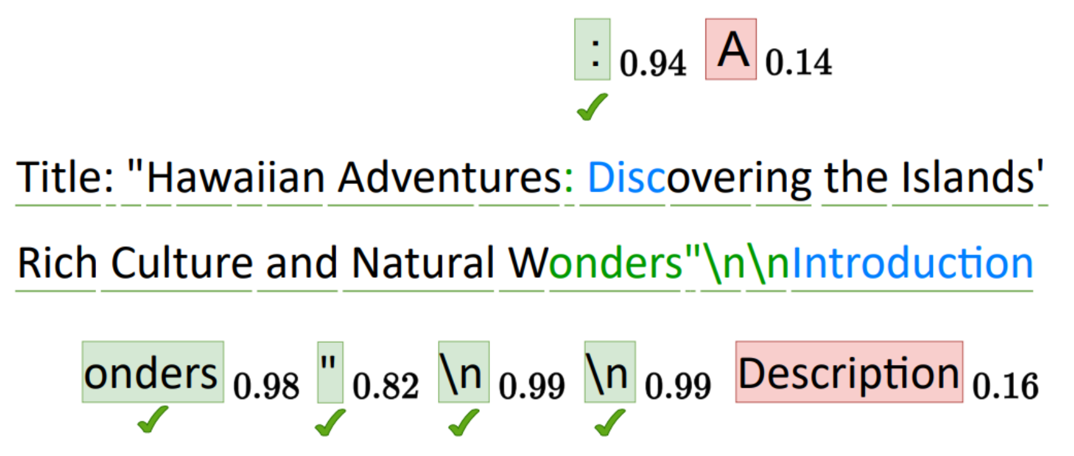

: 0.94
A 0.14
✓
Title: "Hawaiian Adventures: Discovering the Islands'
Rich Culture and Natural Wonders"\n\nIntroduction
onders 0.98
" 0.82
\n 0.99
\n 0.99
Description 0.16
✓ ✓ ✓ ✓

Figure 3: The difficulty of token acceptance varies across different contexts. The values on the right side of the rectangular blocks represent the top-1 probability of the tokens on the self-drafting model $\mathcal{M}_s$. Green boxes indicate accepted tokens, while red boxes represent rejected tokens. Blue tokens signify corrections made by the big target model $\mathcal{M}_b$.

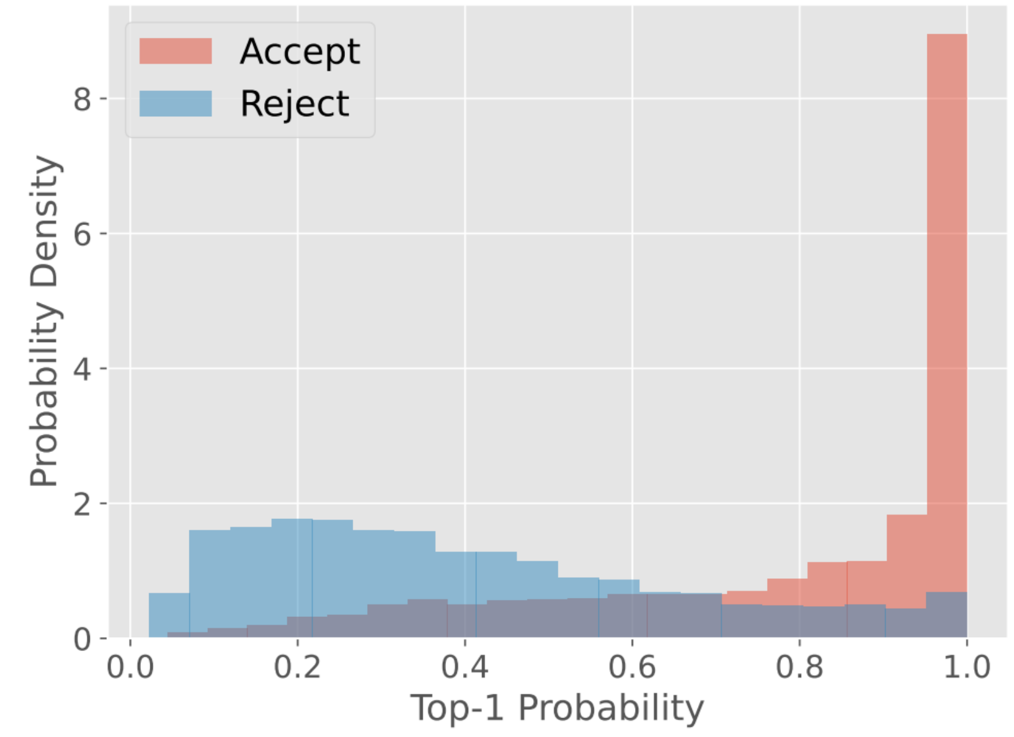
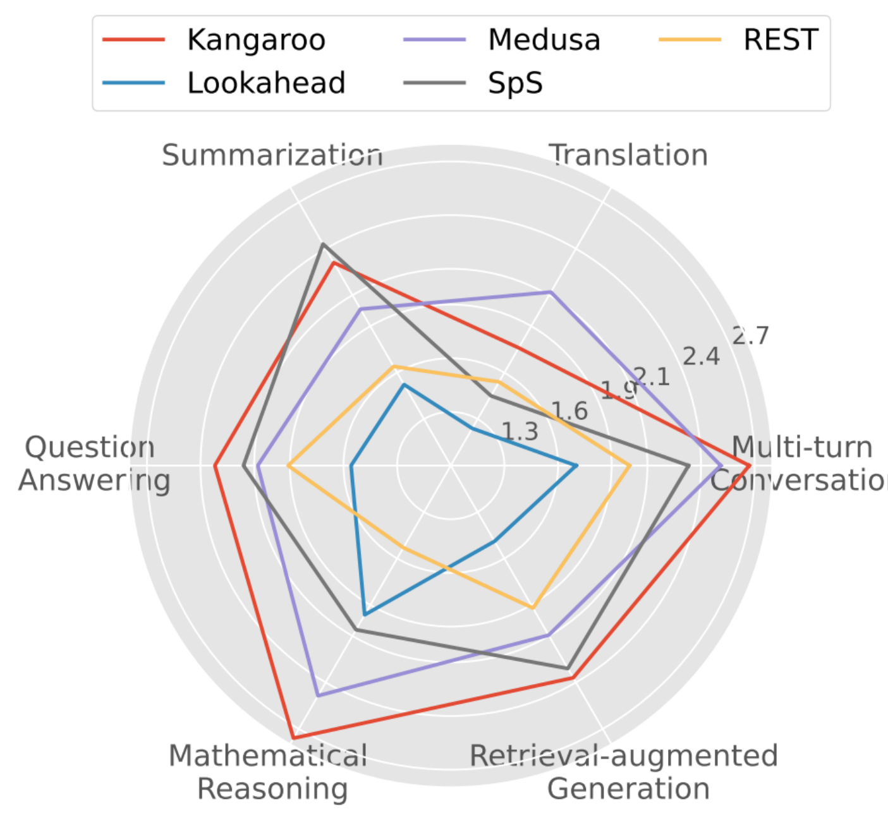
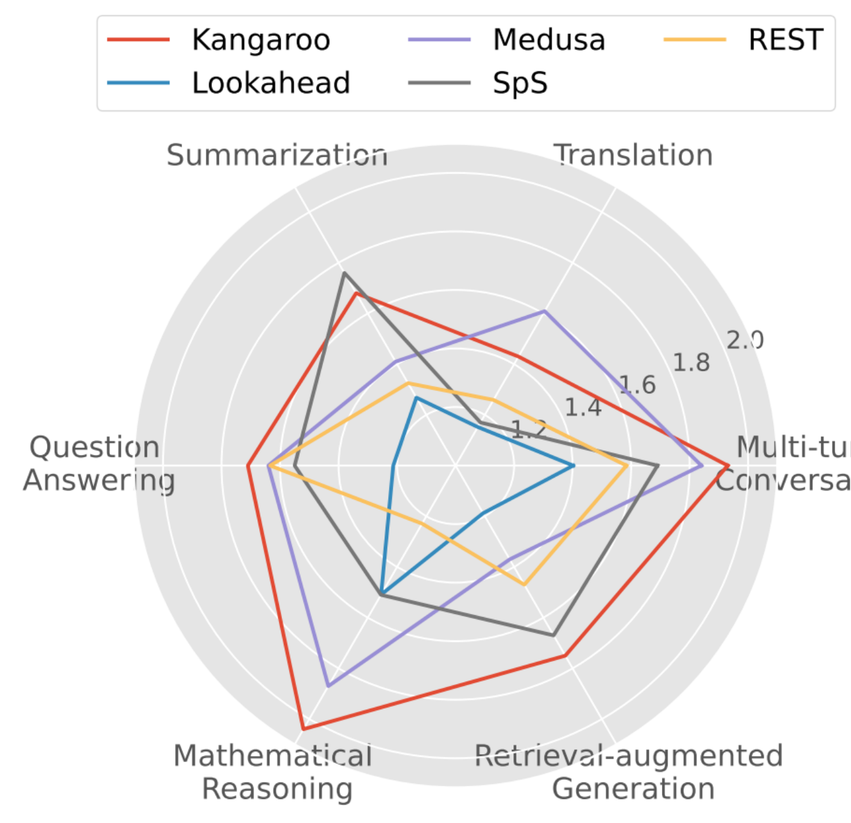


Figure 4: The density of top-1 conditional probability on the *mathematical reasoning* subtask. "Accept" denotes the top-1 confidence of accepted draft tokens while "Reject" denotes the corresponding confidence of rejected tokens.

Experiments



(a) Compression rate.



(b) End-to-end speedup ratio.

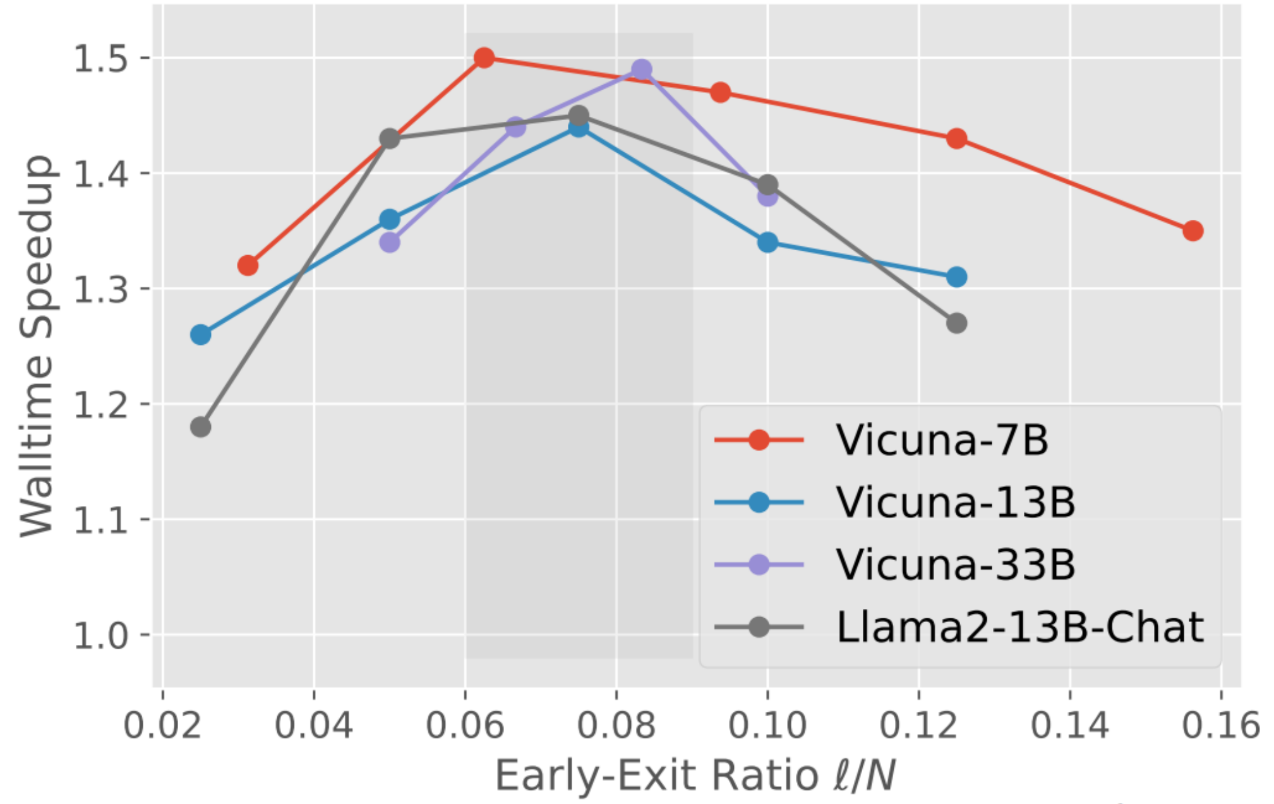


Figure 6: The optimal early-exit ratio $\frac{\ell}{N}$.

Table 1: Speedup comparison of various speculative decoding methods on Spec-Bench [22] for Vicuna [12]. Speedup is the walltime speedup ratio and CR denotes the compression rate.

Size	Method	Translation		QA		Summarization		Math		RAG		MT Bench		Avg.
		CR	Speedup	CR	Speedup	CR	Speedup	CR	Speedup	CR	Speedup	CR	Speedup	
7B	Lookahead [21]	1.24	1.15×	1.56	1.21×	1.53	1.27×	1.96	1.51×	1.49	1.19×	1.70	1.40×	1.29×
	Medusa *w/o Tree*	1.58	1.41×	1.50	1.34×	1.49	1.32×	1.73	1.54×	1.51	1.29×	1.76	1.55×	1.41×
	REST [19]	1.54	1.26×	1.91	1.63×	1.64	1.33×	1.53	1.23×	1.92	1.46×	2.00	1.58×	1.43×
	Kangaroo *w/o Tree*	1.41	1.24×	1.87	1.43×	1.87	1.50×	2.14	1.61×	2.05	1.52×	2.22	1.68×	1.50×
	SpS [13]	1.45	1.17×	2.16	1.55×	2.43	**1.76**×	2.06	1.51×	2.31	1.67×	2.33	1.69×	1.56×
	Medusa [20]	2.12	**1.60**×	2.08	1.64×	2.01	1.41×	2.48	1.87×	2.09	1.37×	2.51	1.84×	1.63×
	Kangaroo	1.76	1.43×	2.32	**1.71**×	2.31	1.68×	2.76	**2.04**×	2.37	**1.75**×	2.67	**1.93**×	**1.72**×
13B	Lookahead [21]	1.25	1.02×	1.39	0.99×	1.50	0.98×	1.94	1.24×	1.52	0.94×	1.68	1.08×	1.04×
	REST [19]	1.53	1.07×	1.92	1.41×	1.66	1.14×	1.55	1.06×	1.87	1.34×	1.98	1.36×	1.23×
	Medusa [20]	2.19	1.21×	2.11	1.17×	2.08	1.21×	2.59	1.41×	2.12	1.12×	2.58	1.38×	1.24×
	Medusa *w/o Tree*	1.61	1.33×	1.49	1.25×	1.53	1.25×	1.80	1.48×	1.53	1.23×	1.82	1.48×	1.34×
	Kangaroo *w/o Tree*	1.45	1.18×	1.79	1.34×	2.00	1.41×	2.42	1.63×	2.16	1.40×	2.44	1.66×	1.44×
	SpS [13]	1.44	1.21×	1.83	1.49×	2.32	**1.62**×	2.15	1.63×	2.43	1.61×	2.25	1.65×	1.54×
	Kangaroo	1.79	**1.39**×	2.25	**1.66**×	2.41	1.57×	2.82	**1.87**×	2.49	**1.68**×	2.71	**1.79**×	**1.65**×

Table 2: Ablation studies on the architecture of the adapter module $\mathcal{A}$ for Vicuna-7B. "Speedup“ denotes the average speedup ratio on Spec-Bench [22].

Architecture	Input LN	Attention	Post LN	FFN	*Linear*	Last LN	Head	# Parameters	Speedup
Medusa	✗	✗	✗	✗	× 4	✗	× 4	591M	1.41 ×
Kangaroo	✓	✓	✗	✗	✗	✓	✗	**67M**	**1.50** ×
Kangaroo + Head	✓	✓	✗	✗	✗	✓	✓	198M	1.44 ×
1-Layer Transformer	✓	✓	✓	✓	✗	✓	✗	202M	1.37 ×
MLP Only	✓	✗	✗	✗	× 2	✓	✓	165M	1.22 ×