RAC: Request-adaptive Configuration for Efficient Speculative Decoding

Extended Abstract

Jinhao Sheng
School of Intelligent Medicine,
China Medical University
Shenyang, China
jhsheng@cmu.edu.cn

Xiao Huang
School of Artificial Intelligence,
Shanghai Jiao Tong University
Shanghai, China
sjtu1792369@sjtu.edu.cn

Hongxin Wei
Department of Statistics and Data Science,
Southern University of Science and Technology
Shenzhen, China
weihx@sustech.edu.cn

Feng Zhou*
Center for Applied Statistics and School of Statistics,
Renmin University of China
Beijing, China
feng.zhou@ruc.edu.cn

ABSTRACT

Speculative decoding enhances the inference efficiency of large language models by employing a lightweight draft model to generate candidate tokens, which are then verified in parallel by the target model. However, existing approaches typically use fixed speculative configurations—such as the draft model and speculative length—across similar requests, neglecting semantic and structural differences. This limits acceleration potential and reduces adaptability to diverse, dynamic real-world scenarios. To address this, we propose a reinforcement learning-based method called Request-Adaptive Configuration selection (RAC). By formulating speculative configuration selection as a Markov decision process, RAC dynamically determines the optimal draft model and speculative length for each incoming request. It integrates static request features with historical execution feedback to enable fine-grained, request-level inference optimization. Experiments on various text generation benchmarks demonstrate that RAC achieves maximum speedups of **2.02**× and **1.37**× over autoregressive decoding and vanilla speculative decoding, respectively.

KEYWORDS

Large Language Model, Speculative Decoding, Reinforcement Learning

ACM Reference Format:
Jinhao Sheng, Xiao Huang, Hongxin Wei, and Feng Zhou. 2026. RAC: Request-adaptive Configuration for Efficient Speculative Decoding: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/SGXL4097

*Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/SGXL4097

1 INTRODUCTION

In recent years, large language models (LLMs) such as GPT-4 [1], LLaMA [17], and Claude [2] have achieved impressive performance across a wide range of applications. However, as model sizes and parameter counts continue to grow, inference faces increasing challenges, including high latency and low throughput. To address this, speculative decoding (SD) [14] has emerged as an effective solution. SD employs a lightweight draft model to rapidly generate multiple candidate tokens, which are then verified in parallel by the target model. This approach reduces inference latency and improves computational efficiency.

However, most existing SD methods [8, 15, 20] adopt fixed configurations—such as predefined draft models and speculative lengths regardless of request variability. In real-world applications, different requests often require different optimal configurations, highlighting the need for more adaptive inference strategies. Several adaptive SD approaches have been proposed, including SpecDec++ [12], MetaSD [13], and SWIFT [18]. While these methods introduce a degree of adaptability, their mechanisms are inherently *static*, relying solely on features of the current request and neglecting the temporal correlation across user requests. In practice, user requests often exhibit historical dependencies, and leveraging execution feedback from past interactions can significantly enhance SD efficiency.

Motivated by this, we propose a *dynamic* request-adaptive configuration selection method (RAC) based on reinforcement learning (RL) [3]. RAC jointly considers the characteristics of the current request and historical execution feedback to enable fine-grained, request-level control over speculative configurations, including draft model and speculative length. We evaluate RAC on diverse text generation benchmarks, where it consistently outperforms existing baselines in inference speedup.

2 METHODOLOGY

We formulate the adaptive speculative configuration problem as a Markov decision process (MDP), as selecting the optimal configuration for each request in SD is a dynamic, state-dependent process with temporal feedback. The MDP framework offers a structured approach by defining states, actions, and rewards, and optimizing

Table 1: Experimental results on the HumanEval, GSM8K, and Hybrid datasets with Vicuna series models. The best and second-best results are marked in bold and underlined, respectively. The symbol "-" denote that the methods do not support current model configuration.

Model	Method	HumanEval		GSM8K		Hybrid		Avg.
		Token/s	Speedup	Token/s	Speedup	Token/s	Speedup	
Vicuna Series	Autoregressive	29.369	1.000×	29.938	1.000×	29.522	1.000×	1.000×
	Speculative Decoding	45.287	1.542×	42.977	1.435×	40.274	1.364×	1.432×
	Lookahead Decoding	40.487	1.378×	45.527	1.521×	36.991	1.252×	1.384×
	PLD	42.058	1.432×	44.036	1.470×	40.676	1.377×	1.426×
	SAM-Decoding	<u>46.382</u>	<u>1.579×</u>	<u>46.612</u>	<u>1.556×</u>	<u>43.721</u>	<u>1.481×</u>	<u>1.539×</u>
	Assisted Generation	45.205	1.539×	45.843	1.531×	37.972	1.286×	1.452×
	SWIFT	30.648	1.044×	25.874	0.864×	23.392	0.792×	0.9×
	RAC (ours)	**59.343**	**2.021×**	**54.472**	**1.819×**	**50.030**	**1.695×**	**1.845×**

a policy to maximize long-term performance. Our method is built on the following key components.

State. The state s_t at t-th step consists of the current request feature f_t and historical feedback h_t:

$$s_t = [f_t, h_t], \tag{1}$$

where the current request feature $f_t = [e_t, y_t, l_t, p_t]$ includes the semantic embedding e_t, task type y_t, input length l_t, and perplexity p_t under different draft models. These features reflect the complexity of the current request. The historical feedback h_t summarizes the configurations and performance statistics of the past k requests.

Action. In each decision step, the agent must select the optimal configuration for the current request, i.e., the draft model and the speculative length. Each action a_t in the action space $\mathcal{A}$ is defined as a combination of two components:

$$a_t = \left[M_t^q, w_t \mid M_t^q \in \mathcal{M}^q, w_t \in \mathcal{W}\right], \tag{2}$$

where $\mathcal{M}^q$ represents the set of draft models and $\mathcal{W}$ represents the set of speculative lengths.

Reward. The reward function plays a critical role in RL, as it directly influences both the convergence behavior and the final performance of the learned policy. In our setting, the primary objective is to improve inference speed. To this end, we define the reward as the normalized token generation speed relative to the autoregressive (AR) decoding. For each request, the reward under a given speculative configuration is computed as:

$$r = \frac{S_{\text{RAC}}}{S_{\text{AR}}}, \tag{3}$$

where S_{RAC} denotes the token generation speed under RAC and S_{AR} denotes the average speed achieved by AR.

To model the complex relationship between request characteristics and optimal configurations, RAC employs a diffusion-based policy integrated with a soft actor-critic (SAC) [10] framework. This design enables stable learning and expressive policy modeling.

3 EXPERIMENTS

We evaluate RAC on diverse benchmarks [4, 6, 7, 19], and compare it against representative baselines such as Speculative Decoding [5, 14], Lookahead Decoding [8], PLD [16], SAM-Decoding [11], Assisted Generation [9], and SWIFT [18].

The main results are shown in Table 1, we evaluate diverse datasets using Vicuna-33B-v1.3 as target model. For adaptive draft model selection, RAC utilizes a model pool consisting of Vicuna-160M and Vicuna-68M for evaluation. For speculative length selection, RAC adaptively chooses from candidate lengths ranging from 3 to 10. In Vicuna-based evaluation, RAC consistently outperforms Speculative Decoding, Lookahead, PLD, SAM-Decoding, Assisted Generation, and SWIFT across all datasets, achieving speedups of **2.021**×, **1.819**×, and **1.695**×, respectively. These results highlight the limitations of fixed speculative configurations and demonstrate the effectiveness of our adaptive approach.

4 CONCLUSIONS

In this paper, we proposed an RL-based request adaptive speculative configuration method RAC. The algorithm used the static characteristics of the request and historical execution feedback to dynamically adjust the speculation configuration, thereby improving the inference efficiency of LLM. RAC could also select draft models and adjust the speculative length. It can achieve refined optimization at the requested granularity and effectively overcome the performance bottleneck brought by fixed configuration. Many experimental results showed that RAC significantly outperforms existing baseline methods in multiple text generation benchmark tasks, showing good versatility and performance advantages.

ACKNOWLEDGMENTS

This work was supported by the NSFC Project (No.62576346), the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001), the fundamental research funds for the central universities, and the research funds of Renmin University of China (24XNKJ13), and Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www.anthropic.com

[3] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* 34, 6 (2017), 26–38.

[4] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*. Association for Computational Linguistics, 131–198.

[5] Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318* (2023).

[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[7] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[8] Yichao Fu, Peter Bailis, Ion Stoica, and Hao Zhang. 2024. Break the Sequential Dependency of LLM Inference Using Lookahead Decoding. In *International Conference on Machine Learning*. PMLR, 14060–14079.

[9] Joao Gante. 2023. Assisted generation: a new direction toward low-latency text generation.

[10] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. Pmlr, 1861–1870.

[11] Yuxuan Hu, Ke Wang, Xiaokang Zhang, Fanjin Zhang, Cuiping Li, Hong Chen, and Jing Zhang. 2024. SAM Decoding: Speculative Decoding via Suffix Automaton. *arXiv preprint arXiv:2411.10666* (2024).

[12] Kaixuan Huang, Xudong Guo, and Mengdi Wang. 2024. SpecDec++: Boosting Speculative Decoding via Adaptive Candidate Lengths. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*.

[13] Taehyeon Kim, Hojung Jung, and Se-Young Yun. 2024. A unified framework for speculative decoding with multiple drafters as a bandit. (2024).

[14] Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*. PMLR, 19274–19286.

[15] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024. EAGLE: Speculative Sampling Requires Rethinking Feature Uncertainty. In *International Conference on Machine Learning*. PMLR, 28935–28948.

[16] Apoorv Saxena. 2023. Prompt Lookup Decoding. https://github.com/apoorvumang/prompt-lookup-decoding/

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[18] Heming Xia, Yongqi Li, Jun Zhang, Cunxiao Du, and Wenjie Li. 2025. SWIFT: On-the-Fly Self-Speculative Decoding for LLM Inference Acceleration. In *ICLR*.

[19] Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. 2024. Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding. In *ACL (Findings)*.

[20] Weilin Zhao, Yuxiang Huang, Xu Han, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2024. Ouroboros: Speculative decoding with large model enhanced drafting. *arXiv e-prints* (2024), arXiv–2402.