Enhancing Best-of-N Decoding by Speculative Rejection and Self-Certainty

José Lamir Gouvêa Junior[1,2]*, **Luan Fonseca Garcia**[1,2], **Ewerton de Oliveira**[3], **Thomas Paula**[3]

[1]Núcleo Avançado de Inteligência Artificial (NAIA)

[2]Escola Politécnica - Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

[3]Brazil R&D - HP Inc.
Porto Alegre – RS – Brazil

Abstract. *Controllable text generation techniques such as fine-tuning, reinforcement learning, and prompt engineering have significant potential to enhance reasoning, alignment, and efficiency in Large Language Models. However, these methods often struggle with memory management, generalization across diverse language tasks, and score function design. In contrast, enhancing the decoding process has proven to be an effective way to control generation without requiring additional training or external tools. This work proposes an improved parallel decoding strategy that not only alleviates resource requirements but also effectively leverages its guiding reward function.*

1. General Information

The latency reduction and alignment of Large Language Models (LLMs) output generation to human preferences at test-time has emerged as a promising area of research to maximize effectiveness while avoiding the resource and data intensive nature of the pretraining process [Snell et al. 2024]. Test-time scaling (TTS) proposes methods such as Prompt Engineering, tailoring contextual instruction prompts to elicit finer reasoning or behavior, and Latent Space Manipulation, modifying the inner latent space representation by altering the model internal structure at inference-time [Wang and Shu 2025]. Nevertheless, these methods are usually task and domain specific or lack explainability and transparency.

To overcome the limitations of TTS methods, attention has been directed to the decoding process of Transformer models [Turner 2024], where the text generation process can be guided dynamically by direct manipulation of the sampling method. Among efficient sampling techniques, the Best-of-N paradigm is a straightforward method of response selection that chooses the arg max of a defined score, or reward, function from a set of N candidate responses [Beirami et al. 2025]. A wide array of Best-of-N techniques, however, still rely on supporting-verifier models [Leviathan et al. 2023, Sun et al. 2024] to model reward functions. These reward models struggle with computational costs and narrow the potential for generalization.

This work proposes an improvement to a Best-of-N algorithm that both avoids fully generating responses and obviates the need for reward models altogether, potentially reducing GPU memory overhead. Our proposal combines Speculative Rejection

*Corresponding author: j.lamir@edu.pucrs.br

[Sun et al. 2024], as the method for selecting among the generated responses, with Self-Certainty [Kang et al. 2025], as the metric used for the Speculative Rejection's scoring system.

The remainder of this paper is structured as follows. In Section 2, we present Speculative Rejection and Self-Certainty. In Section 3, we present related works. In Section 4, we present our proposal and outline the approach we intend to use for its validation.

2. Background

Speculative Rejection. The number of candidate responses Best-of-N methods require to compete with other selection strategies is computationally infeasible. To avoid this issue, in [Sun et al. 2024] authors grounded on the observation that low-quality utterances can be distinguished at an early stage of generation, proposed a speculative approach for inference-time alignment where the number of candidates is gradually reduced to prevent memory exhaustion while ensuring that only promising responses are generated.

The process begins with a batch size N, large enough to avoid exhausting GPU memory. At each iteration, a reward model scores partially generated responses (or utterances), and the generation of those in the lowest-ranked α-th quantile is halted. The early rejection of candidate responses relies on the correlation between partial and final rewards, which does not generally hold, as generation may proceed in an unexpected way. Reward models are also usually trained to evaluate full responses, further degrading the quality of rewards for partial utterances.

Nonetheless, as Speculative Rejection obviates the need to modify a model's pre-trained weights, it offers a more practical deployment compared to post-training alignment methods and, since the memory required for KV caching grows linearly with the generation length, the immediate generation of a large pool of candidate responses also enables more efficacious resource utilization than conventional Best-of-N selection, especially at the early stages.

Self-Certainty and Best-of-N selection. Best-of-N improves LLMs' reasoning and alignment by generating N candidate responses and selecting the highest-scoring one using external reward models [Snell et al. 2024]. These models, however, raise significant computational and practical challenges: they are task-specific, sensitive to the base model, often require as many parameters as the LLM, and are predisposed to distributional shifts and reward hacking [Kang et al. 2025].

Self-Certainty [Kang et al. 2025] is proposed as a scalable metric that takes advantage of the LLMs' own probability distribution. It is defined as the average Kullback-Leibler(KL) divergence from a uniform distribution U over the vocabulary V:

$$\textbf{Self-Certainty} = \frac{1}{n}\mathrm{KL}(U \parallel p(\cdot|x, y_{<i})) = -\frac{1}{n|V|}\sum_{i=1}^{n}\sum_{j=1}^{|V|}\log\left(|V| \cdot p(j|x, y_{<i})\right) \quad (1)$$

Where $p(\cdot|x, y_{<i}) \in [0, 1]^V$ is the model's probability distribution for generating the i-th token conditioned on input x and the previously generated sequence of tokens $y_{<i}$.

The rationale is that a distribution that diverges from uniform indicates a more peaked, and thus more certain, prediction of the LLM output.

3. Related Works

Self-Truncation Best-of-N. Self-Truncation Best-of-N (ST-BoN) [Wang et al. 2025] truncates suboptimal responses via identification of early inconsistencies between samples while maintaining the generation of the sample with greater hidden-state consistency. ST-BoN is also a Best-of-N selection method that exchanges reward models for post-hoc strategies, notably, the proposed scoring function is heavily inspired by majority-voting Self-Consistency [Wang et al. 2023]. Limitations arise since Self-Consistency, unlike Self-Certainty, lacks a direct quality score that does not require exact string matching. [Wang et al. 2025] attempt to circumvent this limitation by operating at the level of latent representations. By comparison, the definition of Equation (1) allows Self-Certainty to be directly applied during inference.

4. Proposed Method

We propose a modification to the Speculative Rejection algorithm that replaces reward models with Self-Certainty, enabling faster scoring and widening the range of supported hyperparameters while still preserving the qualities of Speculative Rejection.

The performance of Speculative Rejection is highly dependent on the choice of the rejection rate α. A lower value maintains a larger pool of responses at any given point; however, it also increases latency due to both the larger batch size and the computational overhead required by the reward model [Sun et al. 2024]. The proposed modification has the potential to mitigate the latter issue, enabling a more efficient and broader range of viable values for α.

Using the average as an aggregation score in Equation 1 means that early-stage mistakes reduce the overall certainty score, making Self-Certainty particularly well-suited for the early-stop loop of Speculative Rejection. Moreover, due to its robustness to generation length compared to other measures such as Entropy and Perplexity [Kang et al. 2025], it is hypothesized that Self-Certainty can accurately score partial responses during the rejection step.

Moreover, by eliminating the reward model, additional memory resources are freed. These resources could be utilized to increase the initial batch size in Speculative Rejection: generating a larger number of candidate responses increases the likelihood of obtaining promising outputs. As demonstrated by [Kang et al. 2025], across the metrics considered, Self-Certainty is the only scoring metric that monotonically increases with N.

Experimental Setup. For evaluating the effect of the modification on both performance and resource efficiency, we propose conducting experiments under conditions similar to those used by [Sun et al. 2024]. The original work does not elaborate on the data profile besides mentioning the sele1ction of the Alpaca Farm Dataset [Dubois et al. 2024]. Nevertheless, metrics such as relative GPU compute or speedup factor should remain unaffected by the (instance) sampling method.

Further investigation could be conducted by analyzing the behavior of partial Self-Certainty, in order to determine whether the correlation between partial and full scores observed by [Sun et al. 2024] holds. A comparison with the results presented by [Kang et al. 2025] could also provide insights into the effect of early stopping on Self-Certainty.

Acknowledgements

This paper was achieved in a project supported by the Brazilian Informatics Law (Law nº 8.248 of 1991) and was developed over Agreement 001/2015 between Pontifícia Universidade Católica do Rio Grande do Sul and HP Brasil Indústria e Comércio de Equipamentos Eletrônicos Ltda.

References

Beirami, A., Agarwal, A., Berant, J., D'Amour, A., Eisenstein, J., Nagpal, C., and Suresh, A. T. (2025). Theoretical guarantees on the best-of-n alignment policy.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. (2024). Alpacafarm: A simulation framework for methods that learn from human feedback.

Kang, Z., Zhao, X., and Song, D. (2025). Scalable best-of-n selection for large language models via self-certainty.

Leviathan, Y., Kalman, M., and Matias, Y. (2023). Fast inference from transformers via speculative decoding.

Snell, C., Lee, J., Xu, K., and Kumar, A. (2024). Scaling llm test-time compute optimally can be more effective than scaling model parameters.

Sun, H., Haider, M., Zhang, R., Yang, H., Qiu, J., Yin, M., Wang, M., Bartlett, P., and Zanette, A. (2024). Fast best-of-n decoding via speculative rejection.

Turner, R. E. (2024). An introduction to transformers.

Wang, H. and Shu, K. (2025). Make every token count: A systematic survey on decoding methods for foundation models.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models.

Wang, Y., Zhang, P., Huang, S., Yang, B., Zhang, Z., Huang, F., and Wang, R. (2025). Sampling-efficient test-time scaling: Self-estimating the best-of-n sampling in early decoding.