Qeios

Peer Review

# Review of: "Draft Model Knows When to Stop: A Self-Verification Length Policy for Speculative Decoding"

**Hangbo Bao**[1]

1. Microsoft Research Asia (China), Beijing, China

Strengths:

1. The paper proposes SVIP, an improved dynamic draft length policy for speculative decoding, addressing the token difficulty variation and achieving substantial wall-time speedup.
2. SVIP is flexible, training-free, and can be integrated with existing speculative decoding frameworks, showing consistent improvements across multiple benchmarks.

Weaknesses:

1. The paper does not sufficiently describe the datasets used, such as SpecBench, lacking details on metrics, example tasks, and how these datasets evaluate performance.
2. The numerical values in tables 1-4 are not clearly explained, making it difficult for readers to understand the performance improvements.
3. The introduction to speculative decoding is brief, and the paper does not adequately compare SVIP to other existing methods or improvements in the field.

Suggestions for Improvement:

Include a comprehensive description of the datasets like SpecBench, detailing the metrics used for evaluation and providing examples of tasks or benchmarks.

Clarify the meaning and significance of the numbers in tables 1-4, perhaps by including a legend or additional explanations in the text.

Expand the introduction to speculative decoding to provide readers with a clear understanding of the concept and its relevance. Additionally, compare SVIP with other related improvements in speculative decoding to highlight its unique contributions and advantages.



doi.org/10.32388/W5GJ1C

## Declarations

**Potential competing interests:** No potential competing interests to declare.