Reward-Guided Speculative Decoding vs. Standard Methods: Throughput and Quality on DS-1000

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 5 peer-reviewed papers addressing the following research question: How does Reward-Guided Speculative Decoding (RSD) compare to standard speculative decoding in terms of inference throughput and output quality on the DS-1000 code generation benchmark. Speculative decoding (SD) accelerates large language model inference by using a smaller draft model to propose draft tokens that are subsequently verified by a larger target model. However, the performance of standard SD is often limited by the strictly sequential execution of. 17 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MineDraft: A Framework for Batch Parallel Speculative Decoding. Research question: How does Reward-Guided Speculative Decoding (RSD) compare to standard speculative decoding in terms of inference throughput and output quality on the DS-1000 code generation benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.5/10.

3 Results

5 papers retrieved. 17 claims extracted; 3 independently verified. Quality review score: 5.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
MINEDRAFT improves throughput by up to 75% compared to standard speculative decoding.	×	0.15
MINEDRAFT reduces end-to-end latency by up to 39% compared to standard speculative decoding.	✓	0.20
MINEDRAFT is implemented as a plugin for vLLM.	×	0.12
MINEDRAFT uses a batch-parallel design to overlap drafting and verification processes.	✓	0.18
MINEDRAFT maintains two batches of requests to overlap drafting for one batch with verification for the other.	✓	0.21
MINEDRAFT is evaluated using throughput (tokens/second) and end-to-end latency (milliseconds).	×	0.12
MINEDRAFT can integrate existing drafting techniques to further improve speculative decoding.	×	0.08
Setting 1 uses Qwen3-32B as the target model and Qwen3-0.6B as the draft model.	×	0.06
Setting 2 uses Qwen3-32B as the target model and Qwen3-1.7B as the draft model.	×	0.06
Setting 3 uses Qwen3-32B as the target model and Qwen3-4B as the draft model.	×	0.06
Setting 4 uses Qwen3-32B as the target model and Qwen3-8B as the draft model.	×	0.06
Setting 5 uses Llama-3.3-70B-Instruct-AWQ-INT4 as the target model.	×	0.02
All settings use tensor parallelism set to four for the target model and a single GPU for the draft model.	×	0.04
Settings 1–4 are served on 5 L40 GPUs with m = 16.	×	0.00
MINEDRAFT adopts a modified PEARL parallelism to offset the limitation of strictly sequential execution.	×	0.08
MINEDRAFT re-drafts requests that failed verification before submitting the batch to the Verifier.	×	0.07
MINEDRAFT aims to optimize performance through a re-drafting mechanism for failed verification attempts.	×	0.06

References

- http://arxiv.org/abs/2203.16487v6
- http://arxiv.org/abs/2508.17739v2
- http://arxiv.org/abs/2603.18016v1