DeepSeek-V3 Multi-Token Prediction vs. Next-Token Baselines in Low-Resource Code Completion

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the performance of DeepSeek-V3's multi-token prediction objective compare to standard next-token prediction on code completion accuracy in low-resource programming languages using the. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Efficient Training-Free Multi-Token Prediction via Embedding-Space Probing. Research question: How does the performance of DeepSeek-V3's multi-token prediction objective compare to standard next-token prediction on code completion accuracy in low-resource programming languages using the MBPP-XL benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

10 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
ESP achieves up to 12% higher τ than LADE on LLaMA3 models.	×	0.07
ESP achieves 13–18% higher τ than STAND on Qwen3 models.	×	0.04
ESP achieves up to 42% fewer model invocations at BC = 60.	×	0.02
ESP achieves the best speedup ratios across all models and budgets, outperforming the next-best baseline by up to $\sim$12% o	×	0.05
ESP does not rely on any auxiliary N-gram cache.	×	0.03
ESP uses single mask token for BC=10,30 and two mask tokens for BC=60.	×	0.10
Mask token design in ESP is based on mean of given prompt's embedding (soft initialization) with dynamic updates based o	×	0.08
Models are run with a maximum generation length of 100 tokens on a single NVIDIA A100 GPU.	×	0.04
Temperature=0.0, 1.0 with temperature=1.0 results provided in the Appendix Section G.3.	×	0.01
Tasks include summarization, translation, writing, coding, retrieval and math tasks (from GSM8K Cobbe et al. (2021)).	×	0.03
Baselines include Prompt Lookup Decoding (PLD) (Saxena, 2023), Stochastic Adaptive N-gram Drafting (STAND) (Song et al.,	×	0.02
Block Complexity (BC) values used are 10, 30, 60.	×	0.03
ESP consistently achieves the highest average accepted tokens across most tasks and BC settings for LLaMA3.1-8B-Instruct	×	0.04
ESP consistently achieves the highest average accepted tokens across most tasks and BC settings for Qwen3-32B.	×	0.04
Average acceptance length (τ) for BC=30 and BC=60 for LLaMA3.2-3B/LLaMA3.1-8B-Instruct is reported in Table 2.	×	0.04
Mask tokens are dynamically updated based on the last token generated following Equation (5), with λ = 0.1.	×	0.07

References

- http://arxiv.org/abs/2506.00404v1
- http://arxiv.org/abs/2603.17942v2
- http://arxiv.org/abs/1002.1144v1