Speculative Decoding Efficiency Degradation in Qwen2.5 Across Single-Task and Multitask Scenarios on the LMSYS Chatbot Benchmark

Assignee Research

June 11, 2026

Abstract

Speculative decoding accelerates autoregressive inference by drafting candidate tokens with a fast model and verifying them in parallel with the target. Self-speculative methods avoid the need for an external drafter but have been studied exclusively in homogeneous Transformer architectures. We introduce component-aware self-speculative decoding, the first method to exploit the internal architectural heterogeneity of hybrid language models, isolating the SSM/linear-attention subgraph as a zero-cost internal draft. We evaluate this on two architecturally distinct hybrid families: Falcon-H1 (par

1 Introduction

This paper examines: Component-Aware Self-Speculative Decoding in Hybrid Language Models. Research question: How does speculative decoding efficiency degrade in Qwen2.5 when transitioning from single-task to multitask scenarios, as measured by tokens-per-second throughput on the LMSYS Chatbot Benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

13 papers retrieved. 15 claims extracted; 11 independently verified. Quality review score: 7.2/10.

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
The total variation distance between draft and full model output distributions was averaged over 100 prompts.	✓	0.28
Falcon-H1-0.5B-Base and Falcon-H1-3B-Base are parallel hybrid models where every layer contains both a Mamba-2 SSM branc	✓	0.32
The Falcon-H1-0.5B model has 36 layers.	✓	0.20
The Falcon-H1-3B model has 32 layers.	✓	0.19
Qwen3.5-0.8B-Base is a sequential hybrid model interleaving 18 Gated DeltaNet linear attention layers with 6 softmax att	✓	0.25
The ratio of linear attention layers to softmax attention layers in Qwen3.5-0.8B-Base is 3:1.	✓	0.18
Qwen2.5-0.5B is a pure Transformer model with 24 standard attention layers.	✓	0.20
LayerSkip applied to Qwen2.5-0.5B skips 33% of layers.	×	0.11
For Qwen3.5, LayerSkip skips 33% of layers selected uniformly.	×	0.14
For Qwen3.5, the early-exit strategy uses only the first 50% of layers.	×	0.12
All experiments use the WikiText-2 validation split as the primary evaluation corpus.	✓	0.16
Qwen3.5-0.8B-Base has 752M parameters.	✓	0.16
Falcon-H1-0.5B-Base has 521M parameters.	✓	0.18
Qwen2.5-0.5B has 494M parameters.	×	0.10
Falcon-H1-3B-Base has 3.15B parameters.	✓	0.17

References

- http://arxiv.org/abs/2001.06902v5

- http://arxiv.org/abs/2605.01106v1
- http://arxiv.org/abs/2201.11957v1