Novelty Assessment Report

Paper: Training-Free Loosely Speculative Decoding: Accepting Semantically Correct Drafts Beyond Exact Match

PDF URL: https://openreview.net/pdf?id=JjoTg34YiU

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-30

Abstract

Large language models (LLMs) achieve strong performance across diverse tasks but suffer from high inference latency due to their autoregressive generation. Speculative Decoding (SPD) mitigates this issue by verifying candidate tokens from a smaller draft model in parallel, yet its strict exact-match verification discards many semantically valid continuations. We propose Training-Free Loosely Speculative Decoding (FLy), a novel method that loosens the rigid verification criterion by leveraging the target model's own corrective behavior to judge whether a draft–target mismatch remains semantically valid. FLy introduces a two-tier mechanism: an entropy-level gate that identifies whether the current token allows multiple plausible alternatives or is nearly deterministic, and a token-level deferred window that distinguishes genuine errors from differently worded yet semantically correct variants. To further reduce latency, we design a multi-level acceleration strategy that accelerates not only the target model but also the drafter itself. Owing to its training-free design, FLy composes seamlessly with arbitrary draft–target pairs and generalizes across models and domains without hyperparameter re-tuning. Experiments show that FLy preserves $\geq$99\% of the target model's accuracy while achieving an average 2.81$\times$ speedup on Llama-3.1-70B-Instruct and 5.07$\times$ speedup on the 405B variant. Notably, on out-of-domain datasets, our method remains highly effective and outperforms the training-based method EAGLE-3 by 1.62$\times$.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Accelerating Large Language Model Inference through Speculative Decoding**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Core Speculative Decoding Mechanisms and Theoretical Foundations**
- **Draft Model Design and Training Strategies**
- **Verification and Acceptance Mechanisms**
- **System-Level Optimization and Deployment Strategies**
- **Domain-Specific Applications and Extensions**
- **Advanced Decoding Strategies and Hybrid Methods**

Complete Taxonomy Tree

- Accelerating Large Language Model Inference through Speculative Decoding Survey Taxonomy
- Core Speculative Decoding Mechanisms and Theoretical Foundations
    - Foundational Algorithms and Sampling Methods (3 papers)
    - [6] Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding (Xia, 2024) View paper
    - [7] Fast inference from transformers via speculative decoding (Leviathan, 2023) View paper
    - [22] Accelerating Large Language Model Decoding with Speculative Sampling (Chen, 2023) View paper
    - Theoretical Analysis and Optimization Principles (3 papers)
    - [23] A theoretical perspective for speculative decoding algorithm (Yin Ming, 2024) View paper
    - [25] How Speculative Can Speculative Decoding Be? (Zhang, 2024) View paper
    - [49] Accelerated speculative sampling based on tree monte carlo (Z Hu, 2024) View paper
    - Survey and Tutorial Resources (5 papers)
    - [13] Beyond tokens: A survey on decoding methods for large language models and large vision-language models (Haoran Wang, 2025) View paper
    - [35] Beyond the speculative game: A survey of speculative execution in large language models (Zhang, 2024) View paper
    - [43] A comprehensive survey of accelerated generation techniques in large language models (Mahsa Khoshnoodi, 2024) View paper
    - [44] Speculative decoding and beyond: An in-depth survey of techniques (Hu Yunhai, 2025) View paper
    - [50] Tutorial Proposal: Speculative Decoding for Efficient LLM Inference (Xia, 2025) View paper
- Draft Model Design and Training Strategies
    - Self-Speculative and Model-Free Drafting (5 papers)
    - [3] Draft& verify: Lossless large language model acceleration via self-speculative decoding (Chen Ke, 2024) View paper
    - [27] Speculative streaming: Fast llm inference without auxiliary models (Nikhil Bhendawade, 2024) View paper
    - [32] An Adaptive Parallel Layer-Skipping Framework for Large Language Model Inference Speedup With Speculative Decoding (Zhe Wen, 2025) View paper
    - [33] Accelerated Test-Time Scaling with Model-Free Speculative Sampling (Woomin Song, 2025) View paper
    - [37] Speculative Decoding via Early-exiting for Faster LLM Inference with Thompson Sampling Control Mechanism (Liu Jiahao, 2024) View paper
    - Adaptive and Online Draft Model Optimization (3 papers)

    - [12] Online speculative decoding (Liu Xiao-xuan, 2023) View paper
    - [16] SDSAT: Accelerating LLM Inference through Speculative Decoding with Semantic Adaptive Tokens (Liu Chengbo, 2024) View paper
    - [24] A drop-in solution for on-the-fly adaptation of speculative decoding in large language models (Jiesong Liu, 2025) View paper
    - Specialized Draft Architectures and Training Methods (3 papers)
    - [8] Generation meets verification: Accelerating large language model inference with smart parallel auto-correct decoding (Yi, 2024) View paper
    - [39] KOALA: Enhancing Speculative Decoding for LLM via Multi-Layer Draft Heads with Adversarial Learning (Kaiqi Zhang, 2024) View paper
    - [46] Parallelspec: Parallel drafter for efficient speculative decoding (Zhang Hong-ming, 2024) View paper
- Verification and Acceptance Mechanisms
    - Tree-Based and Structured Verification (5 papers)
    - [4] Specinfer: Accelerating large language model serving with tree-based speculative inference and verification (Xupeng Miao, 2024) View paper
    - [11] Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification (Miao, 2023) View paper
    - [15] Accelerating llm inference with staged speculative decoding (Spector, 2023) View paper
    - [26] Graph-structured speculative decoding (Gong, 2024) View paper
    - [34] Recursive speculative decoding: Accelerating llm inference via sampling without replacement (Jeon, 2024) View paper
    - Semantic and Flexible Acceptance Criteria ★ (2 papers)
    - [0] Training-Free Loosely Speculative Decoding: Accepting Semantically Correct Drafts Beyond Exact Match (Anon et al., 2026) View paper
    - [1] Specee: Accelerating large language model inference with speculative early exiting (Jiaming Xu, 2025) View paper
    - Multi-Sample and Collaborative Verification (3 papers)
    - [19] Fast Large Language Model Collaborative Decoding via Speculation (Fu, 2025) View paper
    - [28] Speculate, then Collaborate: Fusing Knowledge of Language Models during Decoding (Wang Ziyao, 2025) View paper
    - [29] Speculative decoding for multi-sample inference (Li Yiwei, 2025) View paper
- System-Level Optimization and Deployment Strategies
    - Batch Processing and Throughput Optimization (3 papers)
    - [2] Spin: Accelerating large language model inference with heterogeneous speculative models (Fahao Chen, 2025) View paper
    - [40] TETRIS: Optimal Draft Token Selection for Batch Speculative Decoding (Wu Zhaoxuan, 2025) View paper
    - [45] SPIRe: Boosting LLM Inference Throughput with Speculative Decoding (Heinlein, 2025) View paper
    - Distributed and Pipeline Parallelism (3 papers)
    - [10] SpecPipe: Accelerating Pipeline Parallelism-based LLM Inference with Speculative Decoding (Yin, 2025) View paper
    - [17] Collaborative Large Language Model Inference via Resource-Aware Parallel Speculative Decoding (Yang Hyun Jong, 2025) View paper
    - [38] Communication-Efficient Collaborative LLM Inference via Distributed Speculative Decoding (Zheng Ce, 2025) View paper
    - Adaptive Scheduling and SLO-Aware Serving (2 papers)
    - [31] ALISE: Accelerating Large Language Model Serving with Speculative Scheduling (Youpeng Zhao, 2024) View paper
    - [41] SpecServe: Efficient and SLO-Aware Large Language Model Serving with Adaptive Speculative Decoding (Huang Kaiyu, 2025) View paper
    - Edge and On-Device Deployment (2 papers)
    - [42] Sled: A speculative llm decoding framework for efficient edge serving (Li Xiangchen, 2025) View paper
    - [47] Edgellm: Fast on-device llm inference with speculative decoding (Daliang Xu, 2024) View paper
- Domain-Specific Applications and Extensions
    - Vision-Language Model Acceleration (2 papers)
    - [5] ViSpec: Accelerating Vision-Language Models with Vision-Aware Speculative Decoding (Shu Han, 2025) View paper
    - [30] Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding (Huo, 2025) View paper
    - Reasoning and Test-Time Scaling (2 papers)
    - [14] Accelerating Large-Scale Reasoning Model Inference with Sparse Self-Speculative Decoding (Yilong Zhao, 2025) View paper
    - [36] Reward-guided speculative decoding for efficient llm reasoning (Liao, 2025) View paper
    - Recommendation Systems and Generative Tasks (2 papers)
    - [9] Efficient inference for large language model-based generative recommendation (Lin Xin-yu, 2024) View paper
    - [18] Efficiency unleashed: Inference acceleration for LLM-based recommender systems with speculative decoding (Yun-jia Xi, 2025) View paper
    - Heterogeneous Vocabularies and Cross-Model Inference (1 papers)
    - [21] Accelerating LLM Inference with Lossless Speculative Decoding Algorithms for Heterogeneous Vocabularies (Timor, 2025) View paper
- Advanced Decoding Strategies and Hybrid Methods (2 papers)
    - [20] Minions: Accelerating Large Language Model Inference with Aggregated Speculative Execution (Wang Siqi, 2024) View paper
    - [48] Dynamic-width speculative beam decoding for llm inference (Zongyue Qin, 2025) View paper

Narrative

Core task: Accelerating large language model inference through speculative decoding. The field has organized itself around several complementary research directions. At the foundation lie theoretical studies and core mechanisms that establish how draft models can propose tokens and target models verify them, exemplified by early works such as Speculative Sampling[22] and Fast Inference Speculative[7]. A dense branch focuses on draft model design and training strategies, exploring how to build efficient proposal generators through distillation, early-exiting, or self-speculation techniques like Sparse Self-Speculative[14]. Another major area addresses verification and acceptance mechanisms, where researchers investigate both strict token-level matching and more flexible semantic criteria. System-level optimization and deployment strategies examine batching, scheduling, and resource allocation across distributed settings, as seen in SpecInfer[4] and SpecServe[41]. Domain-specific applications extend speculative decoding to multimodal models, recommendation systems, and code generation, while advanced decoding strategies explore hybrid methods that combine speculation with beam search or Monte Carlo tree search.

Particularly active lines of work contrast strict versus relaxed acceptance policies and explore the trade-offs between draft quality and verification overhead. Loosely Speculative Decoding[0] sits within the semantic and flexible acceptance criteria cluster, emphasizing a more permissive verification strategy that tolerates minor deviations when draft tokens are semantically close to what the target model would produce. This approach contrasts with neighboring works like Specee[1], which may enforce tighter alignment constraints, and differs in philosophy from earlier strict token-matching schemes such as Draft Verify[3]. By relaxing acceptance rules, Loosely Speculative Decoding[0] aims to increase the average number of accepted tokens per verification step, potentially improving throughput when semantic equivalence suffices. Open questions in this area include how to define and measure semantic similarity efficiently, and whether such flexibility introduces quality risks in safety-critical applications.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Specee: Accelerating large language model inference with speculative early exiting

Authors: Jiaming Xu, Jianrong Xu, Jiayi Pan, Yongkang Zhou, Yongjun Zhou, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Early exiting has recently emerged as a promising technique for accelerating large language models (LLMs) by effectively reducing the hardware computation and memory access. In this paper, we identify that the LLM vocabulary serves as the runtime search space of the early exiting predictor and significantly influences the predictor workload (e.g., (\sim 20\%) overall inference latency with â□¼ 3 Ã□ 104 vocabulary size in Llama2). We propose a novel paradigm using speculative models to reduce...

Relationship Analysis

Both papers belong to the Semantic and Flexible Acceptance Criteria category, addressing verification mechanisms that go beyond exact token matching in speculative decoding. While the original paper (FLy) uses entropy-based gates and deferred windows to accept semantically correct drafts by monitoring the target model's self-corrective behavior, the candidate paper (SpecEE) focuses on early exiting within decoder layers using speculative tokens to reduce the predictor's search space, employing MLP-based predictors with probability shift features rather than flexible acceptance of draft tokens.

Contributions Analysis

Overall novelty summary. The paper proposes FLy, a training-free method that relaxes speculative decoding's strict exact-match verification by using the target model's corrective behavior to judge semantic validity. It resides in the 'Semantic and Flexible Acceptance Criteria' leaf, which contains only two papers total, indicating a relatively sparse research direction within the broader verification mechanisms branch. This positioning suggests the work addresses an emerging area where few prior methods have explored loosened acceptance beyond token-level matching.

The taxonomy reveals that FLy's parent branch, 'Verification and Acceptance Mechanisms,' contains three distinct approaches: tree-based verification with five papers, semantic acceptance with two papers, and multi-sample verification with three papers. Neighboring branches focus on draft model design and system-level optimization, which are orthogonal concerns. The scope note for FLy's leaf explicitly excludes strict token-level verification and tree-based methods, clarifying that this work diverges from the more populated tree-structured speculation approaches by prioritizing semantic correctness over structural exploration.

Among the three contributions analyzed, the core FLy framework examined ten candidates with zero refutations, while the two-tier verification mechanism examined only one candidate. The multi-level acceleration strategy, however, examined ten candidates and found one refutable match, suggesting this component has more substantial prior work. Given the limited search scope of twenty-one total candidates examined, these statistics indicate that the semantic acceptance approach appears relatively novel within the examined literature, though the acceleration strategy overlaps with existing techniques in a more crowded space.

Based on the top-21 semantic matches examined, FLy's core semantic acceptance mechanism appears to occupy a sparsely explored niche, while its acceleration component connects to more established optimization strategies. The analysis does not cover exhaustive citation networks or domain-specific applications beyond the taxonomy's scope, leaving open questions about how this work relates to broader semantic similarity research outside speculative decoding contexts.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Training-Free Loosely Speculative Decoding (FLy)

Description: FLy is a training-free speculative decoding method that relaxes the strict exact-match verification rule by accepting semantically correct draft tokens. It uses the target model's own behavior to distinguish genuine errors from differently worded yet semantically valid continuations, without requiring additional training or auxiliary models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. SDSAT: Accelerating LLM Inference through Speculative Decoding with Semantic Adaptive Tokens

URL: View paper

Brief Assessment

SDSAT[16] focuses on training semantic adaptive tokens for draft generation, whereas FLy is explicitly training-free and uses entropy-based gates with deferred windows for semantic verification without any model fine-tuning.

2. Grouped speculative decoding for autoregressive image generation

URL: View paper

Brief Assessment

Grouped Speculative[52] focuses on autoregressive image generation with grouped token verification, not on semantic verification for language models. The candidate addresses visual token redundancy in images, while the original addresses semantic correctness in text generation.

3. Speeding up Speculative Decoding via Sequential Approximate Verification

URL: View paper

Brief Assessment

Sequential Approximate Verification[55] focuses on using a trained verifier for sequential approximate verification in speculative decoding, whereas FLy is training-free and uses the target model's entropy and deferred windows for semantic verification without any auxiliary models.

4. Entropy-Aware Fusion Speculative Decoding for Reliable and Efficient Domain Text Generation

URL: View paper

Brief Assessment

Entropy-Aware Fusion[58] focuses on domain-specific text generation with entropy-aware fusion mechanisms for semantic alignment. The candidate does not address training-free semantic verification or self-corrective behavior detection in speculative decoding, which are core to FLy's novelty.

5. SpecVLM: Enhancing Speculative Decoding of Video LLMs via Verifier-Guided Token Pruning

URL: View paper

Brief Assessment

SpecVLM[54] focuses on video LLMs with visual token pruning for speculative decoding, not on semantic verification of text tokens. The candidate addresses video-specific challenges through spatial token reduction rather than relaxing exact-match verification rules for language generation.

6. Make every token count: A systematic survey on decoding methods for foundation models

URL: View paper

Brief Assessment

Make Every Token[53] is a survey paper that provides a systematic overview of decoding methods for foundation models. It does not present original research on speculative decoding with semantic verification, but rather reviews existing techniques in the field.

7. SelfJudge: Faster Speculative Decoding via Self-Supervised Judge Verification

URL: View paper

Brief Assessment

SelfJudge[56] focuses on training a judge verifier through self-supervision to accept semantically similar tokens, whereas FLy is training-free and uses entropy-level gates with deferred windows to distinguish semantic validity without any training or auxiliary models.

8. Alignment-Augmented Speculative Decoding with Alignment Sampling and Conditional Verification

URL: View paper

Brief Assessment

Alignment-Augmented Speculative[57] focuses on alignment sampling and conditional verification for retrieval-based speculative decoding in long-context generation scenarios. FLy addresses a different problem: relaxing exact-match verification through entropy-based gates and deferred windows to accept semantically correct tokens, without requiring retrieval or alignment sampling mechanisms.

9. Speculative Verification: Exploiting Information Gain to Refine Speculative Decoding

URL: View paper

Brief Assessment

Speculative Verification[59] focuses on dynamically adjusting verification length based on predicted token acceptance using a companion model, not on relaxing exact-match verification to accept semantically correct tokens as FLy does.

10. Beyond tokens: A survey on decoding methods for large language models and large vision-language models

URL: View paper

Brief Assessment

Beyond Tokens Survey[13] appears to be a survey paper that mentions speculative decoding methods in passing. The provided context only contains a brief fragment about 'specdec' and 'spec-drafter' without sufficient detail to assess whether it refutes FLy's novelty claims regarding semantic verification and training-free loosely speculative decoding.

Contribution 2: Two-tier verification mechanism with entropy-level gate and token-level deferred window

Description: The method introduces a two-tier verification scheme: an entropy-level gate determines if a mismatch position is ambiguous or deterministic, and a token-level deferred window monitors subsequent tokens to decide whether the mismatch is semantically valid or represents a genuine error requiring rejection.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. I Know What I Don't Know: Improving Model Cascades Through Confidence Tuning

URL: View paper

Brief Assessment

Confidence Tuning[51] focuses on calibrating confidence scores in model cascades for deferral decisions, not on speculative decoding verification mechanisms. The entropy usage is for routing between small/large models, not for accepting/rejecting draft tokens in speculative decoding.

Contribution 3: Multi-level acceleration strategy

Description: A multi-level acceleration mechanism is proposed that speeds up both the target model and the draft model itself. This prevents the drafting stage from becoming a bottleneck when longer draft sequences are accepted, thereby further reducing overall latency.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Fast inference from transformers via speculative decoding

URL: View paper

Brief Assessment

Fast Inference Speculative[7] focuses on parallel verification of draft tokens from a single approximation model, not on accelerating the draft model itself. The paper mentions hierarchical approaches only as future work: 'It could also be interesting to explore a hierarchical version of the algorithm, where the approximation model is itself accelerated by an even faster model.'

2. Draft& verify: Lossless large language model acceleration via self-speculative decoding

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

3. Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

4. Online speculative decoding

URL: View paper

Brief Assessment

Online Speculative Decoding[12] focuses on continuously updating draft models based on query distributions to improve token acceptance rates. While it mentions accelerating the draft model through knowledge distillation and online adaptation, it does not propose a multi-level acceleration mechanism that simultaneously speeds up both the target model and the draft model itself to prevent drafting bottlenecks when longer sequences are accepted.

5. Glide with a cape: A low-hassle method to accelerate speculative decoding

URL: View paper

Prior Art Analysis

Glide Cape[61] demonstrates prior work on multi-level acceleration in speculative decoding. The candidate paper explicitly proposes accelerating both the draft model and target model through a multi-level acceleration (MLA) strategy. The paper states that MLA 'applies speculative acceleration not only to the target model (as in standard spd) but also to the draft model' and that 'by reducing overhead at drafting stage, mla achieves greater end-to-end efficiency.' This directly addresses the same problem as the original paper's contribution: preventing the drafting stage from becoming a bottleneck when longer draft sequences are accepted.

Evidence

Evidence 1 - **Rationale**: Both papers recognize that improving acceptance rates increases the computational burden on the draft model, motivating the need for draft-side acceleration. - **Original**: by accepting semantically correct mismatches, the average number of accepted tokens (τ) rises markedly. thus, the drafter needs to propose a larger set of tokens per round, which raises the drafter's generation cost to the point where it becomes non-negligible compared to common spd methods - **Candidate**: as we can see from the figure, because the target model stores its computed keys and values corresponding to the tokens it has verified in the last round of verification, these kv cache entries are free for the draft model to use. by reusing these keys and values, the draft model is more likely to b...

6. Break the Sequential Dependency of LLM Inference Using Lookahead Decoding

URL: View paper

Brief Assessment

Lookahead Decoding[60] focuses on parallelizing the target model's own decoding process without using a separate draft model, whereas the original paper's multi-level acceleration specifically addresses speeding up both an external draft model and the target model in speculative decoding frameworks.

7. Closer look at efficient inference methods: A survey of speculative decoding

URL: View paper

Brief Assessment

Closer Look Survey[64] is a survey paper that categorizes and reviews existing speculative decoding methods. It does not propose novel acceleration mechanisms but rather taxonomizes existing work. The survey mentions multi-level approaches in passing but does not claim to introduce them as a novel contribution.

8. Sequoia: Scalable, Robust, and Hardware-aware Speculative Decoding

URL: View paper

Brief Assessment

Sequoia[62] focuses on optimizing tree structures for speculative decoding and hardware-aware optimization, not on accelerating both the draft and target models through a multi-level acceleration mechanism as described in the original paper.

9. Improving multi-candidate speculative decoding

URL: View paper

Brief Assessment

Improving Multi-candidate[65] focuses on accelerating the draft model itself within multi-candidate speculative decoding using target-initialized token trees and dynamic masking. The original paper's multi-level acceleration prevents the drafting stage from becoming a bottleneck when longer sequences are accepted, which is a different technical approach and context than the candidate's focus on multi-candidate generation optimization.

10. Multi-candidate speculative decoding

URL: View paper

Brief Assessment

Multi-candidate Speculative[63] focuses on sampling multiple candidate tokens at each position and organizing them for parallel verification, not on accelerating both the draft and target models simultaneously. The paper does not address the bottleneck of the drafting stage becoming dominant as acceptance length grows.

Appendix: Text Similarity Detection

Textual similarity detection checked 22 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Draft& verify: Lossless large language model acceleration via self-speculative decoding

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Training-Free Loosely Speculative Decoding: Accepting Semantically Correct Drafts Beyond Exact Match View paper
- [1] Specee: Accelerating large language model inference with speculative early exiting View paper
- [2] Spin: Accelerating large language model inference with heterogeneous speculative models View paper
- [3] Draft& verify: Lossless large language model acceleration via self-speculative decoding View paper
- [4] Specinfer: Accelerating large language model serving with tree-based speculative inference and verification View paper
- [5] ViSpec: Accelerating Vision-Language Models with Vision-Aware Speculative Decoding View paper
- [6] Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding View paper
- [7] Fast inference from transformers via speculative decoding View paper
- [8] Generation meets verification: Accelerating large language model inference with smart parallel auto-correct decoding View paper
- [9] Efficient inference for large language model-based generative recommendation View paper
- [10] SpecPipe: Accelerating Pipeline Parallelism-based LLM Inference with Speculative Decoding View paper
- [11] Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification View paper
- [12] Online speculative decoding View paper
- [13] Beyond tokens: A survey on decoding methods for large language models and large vision-language models View paper
- [14] Accelerating Large-Scale Reasoning Model Inference with Sparse Self-Speculative Decoding View paper
- [15] Accelerating llm inference with staged speculative decoding View paper
- [16] SDSAT: Accelerating LLM Inference through Speculative Decoding with Semantic Adaptive Tokens View paper
- [17] Collaborative Large Language Model Inference via Resource-Aware Parallel Speculative Decoding View paper
- [18] Efficiency unleashed: Inference acceleration for LLM-based recommender systems with speculative decoding View paper
- [19] Fast Large Language Model Collaborative Decoding via Speculation View paper
- [20] Minions: Accelerating Large Language Model Inference with Aggregated Speculative Execution View paper
- [21] Accelerating LLM Inference with Lossless Speculative Decoding Algorithms for Heterogeneous Vocabularies View paper
- [22] Accelerating Large Language Model Decoding with Speculative Sampling View paper
- [23] A theoretical perspective for speculative decoding algorithm View paper
- [24] A drop-in solution for on-the-fly adaptation of speculative decoding in large language models View paper
- [25] How Speculative Can Speculative Decoding Be? View paper
- [26] Graph-structured speculative decoding View paper
- [27] Speculative streaming: Fast llm inference without auxiliary models View paper
- [28] Speculate, then Collaborate: Fusing Knowledge of Language Models during Decoding View paper
- [29] Speculative decoding for multi-sample inference View paper
- [30] Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding View paper
- [31] ALISE: Accelerating Large Language Model Serving with Speculative Scheduling View paper
- [32] An Adaptive Parallel Layer-Skipping Framework for Large Language Model Inference Speedup With Speculative Decoding View paper
- [33] Accelerated Test-Time Scaling with Model-Free Speculative Sampling View paper
- [34] Recursive speculative decoding: Accelerating llm inference via sampling without replacement View paper
- [35] Beyond the speculative game: A survey of speculative execution in large language models View paper
- [36] Reward-guided speculative decoding for efficient llm reasoning View paper
- [37] Speculative Decoding via Early-exiting for Faster LLM Inference with Thompson Sampling Control Mechanism View paper
- [38] Communication-Efficient Collaborative LLM Inference via Distributed Speculative Decoding View paper
- [39] KOALA: Enhancing Speculative Decoding for LLM via Multi-Layer Draft Heads with Adversarial Learning View paper
- [40] TETRIS: Optimal Draft Token Selection for Batch Speculative Decoding View paper
- [41] SpecServe: Efficient and SLO-Aware Large Language Model Serving with Adaptive Speculative Decoding View paper
- [42] Sled: A speculative llm decoding framework for efficient edge serving View paper
- [43] A comprehensive survey of accelerated generation techniques in large language models View paper
- [44] Speculative decoding and beyond: An in-depth survey of techniques View paper
- [45] SPIRe: Boosting LLM Inference Throughput with Speculative Decoding View paper
- [46] Parallelspec: Parallel drafter for efficient speculative decoding View paper
- [47] Edgellm: Fast on-device llm inference with speculative decoding View paper
- [48] Dynamic-width speculative beam decoding for llm inference View paper
- [49] Accelerated speculative sampling based on tree monte carlo View paper
- [50] Tutorial Proposal: Speculative Decoding for Efficient LLM Inference View paper
- [51] I Know What I Don't Know: Improving Model Cascades Through Confidence Tuning View paper
- [52] Grouped speculative decoding for autoregressive image generation View paper
- [53] Make every token count: A systematic survey on decoding methods for foundation models View paper
- [54] SpecVLM: Enhancing Speculative Decoding of Video LLMs via Verifier-Guided Token Pruning View paper
- [55] Speeding up Speculative Decoding via Sequential Approximate Verification View paper
- [56] SelfJudge: Faster Speculative Decoding via Self-Supervised Judge Verification View paper
- [57] Alignment-Augmented Speculative Decoding with Alignment Sampling and Conditional Verification View paper
- [58] Entropy-Aware Fusion Speculative Decoding for Reliable and Efficient Domain Text Generation View paper
- [59] Speculative Verification: Exploiting Information Gain to Refine Speculative Decoding View paper
- [60] Break the Sequential Dependency of LLM Inference Using Lookahead Decoding View paper
- [61] Glide with a cape: A low-hassle method to accelerate speculative decoding View paper
- [62] Sequoia: Scalable, Robust, and Hardware-aware Speculative Decoding View paper
- [63] Multi-candidate speculative decoding View paper
- [64] Closer look at efficient inference methods: A survey of speculative decoding View paper
- [65] Improving multi-candidate speculative decoding View paper