# MARS: Unleashing the Power of Speculative Decoding via Margin-Aware Verification

**Jingwei Song**[1*], **Xinyu Wang**[2,7*], **Hanbin Wang**[3], **Xiaoxuan Lei**[2,6],
**Bill Shi**[4†], **Shixin Han**[5], **Eric Yang**[4], **Xiao-Wen Chang**[2†], **Lynn Ai**[4]

[1]The University of Hong Kong, [2]McGill University, [3]Peking University,
[4]gradient, [5]Alibaba Cloud, [6]Mila, [7]SimpleWay

songjingwei@connect.hku.hk, xinyu.wang5@mail.mcgill.ca,
tianyu@gradient.network, chang@cs.mcgill.ca

## Abstract

Speculative Decoding (SD) accelerates autoregressive large language model (LLM) inference by decoupling generation and verification. While recent methods improve draft quality by tightly coupling the drafter with the target model, the verification mechanism itself remains largely unchanged, relying on strict token-level rejection sampling. In practice, modern LLMs frequently operate in low-margin regimes where the target model exhibits weak preference among top candidates. In such cases, rejecting plausible runner-up tokens yields negligible information gain while incurring substantial rollback cost, leading to a fundamental inefficiency in verification.

We propose **Margin-Aware Speculative Verification**, a training-free and domain-agnostic verification strategy that adapts to the target model's local decisiveness. Our method conditions verification on decision stability measured directly from the target logits and relaxes rejection only when strict verification provides minimal benefit. Importantly, the approach modifies only the verification rule and is fully compatible with existing target-coupled speculative decoding frameworks. Extensive experiments across model scales ranging from **8B to 235B** demonstrate that our method delivers consistent and significant inference speedups over state-of-the-art baselines while preserving generation quality across diverse benchmarks. **The code is available at** https://github.com/5SSjw/MARS.

## 1 Introduction

Autoregressive Large Language Models (LLMs) suffer from high inference latency due to memory-bandwidth constraints (Shazeer, 2019). Speculative Decoding (SD) addresses this by decoupling generation and verification: a lightweight *draft model* proposes a sequence of tokens, which are then verified in parallel by the target model (Leviathan et al., 2023). State-of-the-art methods like Medusa (Cai et al., 2024) and EAGLE 2 (Li et al., 2024) have significantly advanced this paradigm by integrating draft heads directly with target features, enabling high-fidelity drafting with minimal overhead. Despite these drafting improvements, the verification mechanism remains rigidly tied to exact-match rejection sampling. This standard approach implicitly assumes the target model holds a decisive preference at every step. In practice, however, LLMs frequently operate in *low-margin regimes*, where the likelihood difference between top candidates is statistically negligible. Strictly rejecting a plausible runner-up token in these cases yields negligible information gain while incurring substantial computational costs. While some recent works explore relaxed verification via learned semantic judges (Bachmann and Alistarh, 2025; Li et al., 2025a), they often introduce complexity through supervised training and auxiliary models, limiting their plug-and-play capability.

In this work, we focus on the essence of the verification bottleneck: the mismatch between verification strictness and the target model's intrinsic uncertainty. We propose a training-free paradigm that analyzes the target model's *logit margin*—a direct proxy for decision stability. We observe that wasteful rejections cluster precisely in low-margin regions. By conditioning verification on the stability of the target logits, we adaptively relax requirements when the model itself is indifferent.

We introduce <u>**Mar**</u>**gin-Aware** <u>**S**</u>**peculative Verification** (MARS), a simple strategy that performs stability-aware tie-breaking. Our method modifies *only* the verification rule, operating entirely at inference time without parameter updates. Extensive experiments demonstrate that this lightweight principle unlocks the full potential of high-quality drafters.

* Equal contributions.
†Corresponding authors.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 8652–8663
July 2-7, 2026 

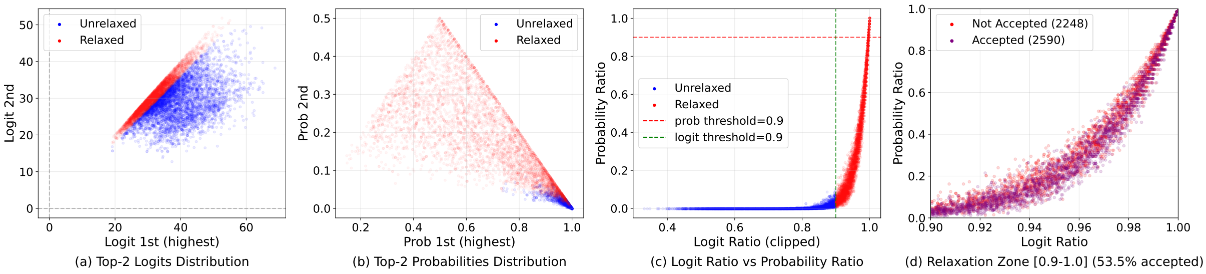


Figure 1: **Comparison of Logit Ratio vs. Probability Ratio for Adaptive Verification (Qwen3-8B).** Red points denote tokens accepted by our relaxation strategy (Logit Ratio $> 0.9$). **(a) Top-2 Logits:** Relaxed tokens cluster along the diagonal ($z_2 \approx z_1$), indicating our method captures candidates with similar raw scores regardless of scale. **(b) Top-2 Probabilities:** Unlike probability-based metrics, logit-based relaxation is not confined to high-entropy regions ($p_1 \approx p_2$); it also captures candidates with relatively lower probability ratios. **(c) Metric Decoupling:** High logit ratios (red) do not imply high probability ratios. The red points span the full range of probability ratios (y-axis), demonstrating that our method recovers valid candidates that are otherwise suppressed by the exponential sensitivity of the softmax function. **(d) Relaxation Zone Effectiveness:** A zoom-in on the low-margin regime (Logit Ratio $> 0.9$), where red points denote rejected tokens and purple points denote accepted ones.

**Contributions.**

- We identify *low-margin rejections* as a fundamental inefficiency in strict verification, showing that adaptive tie-breaking can significantly reduce rollback costs.

- We propose a **training-free** verification algorithm MARS that leverages the logit margin to dynamically adapt verification rigor, avoiding the overhead of auxiliary judges.

- We demonstrate the scalability and robustness of MARS across a wide spectrum of model sizes (from **8B to 235B**) and diverse task settings. Extensive experiments confirm that our method consistently outperforms state-of-the-art baselines across chat, coding, and reasoning benchmarks, delivering significant acceleration with **near-lossless** accuracy preservation.

## 2 Preliminaries

### 2.1 Autoregressive Generation

Autoregressive Large Language Models (LLMs) generate text token-by-token. Given a prefix sequence $x_{<t}$, the model computes the probability distribution for the next token $x_t$ as $P(x_t \mid x_{<t})$. This serial dependency prevents parallelization across time steps, making inference memory-bandwidth bound on modern hardware (e.g., GPUs) and resulting in high latency.

### 2.2 Speculative Decoding

Speculative Decoding (SD) (Leviathan et al., 2023; Chen et al., 2023) alleviates this bottleneck by employing a lightweight approximation model, referred to as the *draft model* $\mathcal{M}_s$. At each step, $\mathcal{M}_s$ tentatively generates a short chain of $K$ candidate tokens (the draft). These tokens are then verified in parallel by the *target model* $\mathcal{M}_t$ in a single forward pass, accepting those that match the target's distribution and discarding the rest.

Let $x_{<t}$ denote the current context. The draft model $\mathcal{M}_s$ autoregressively generates a sequence of speculative tokens $\hat{V} = [\hat{v}_1, \ldots, \hat{v}_K]$. Subsequently, the target model $\mathcal{M}_t$ performs a parallel forward pass to compute the conditional distributions $P(\cdot \mid x_{<t}, \hat{v}_{<i})$ for all positions $i \in \{1, \ldots, K\}$. The core challenge lies in efficiently verifying these candidates while ensuring the final output distribution remains faithful to the target model $\mathcal{M}_t$.

### 2.3 Speculative Verification

Speculative Decoding verifies draft tokens by comparing them against the target model's output distribution. In greedy decoding ($T$=0), a drafted token is accepted only if it matches the target's argmax prediction; the first mismatch truncates the draft chain and emits the target's top-1 token as correction.

In stochastic decoding ($T$>0), Leviathan et al. (2023) propose a rejection-sampling scheme that preserves the target distribution exactly. Each

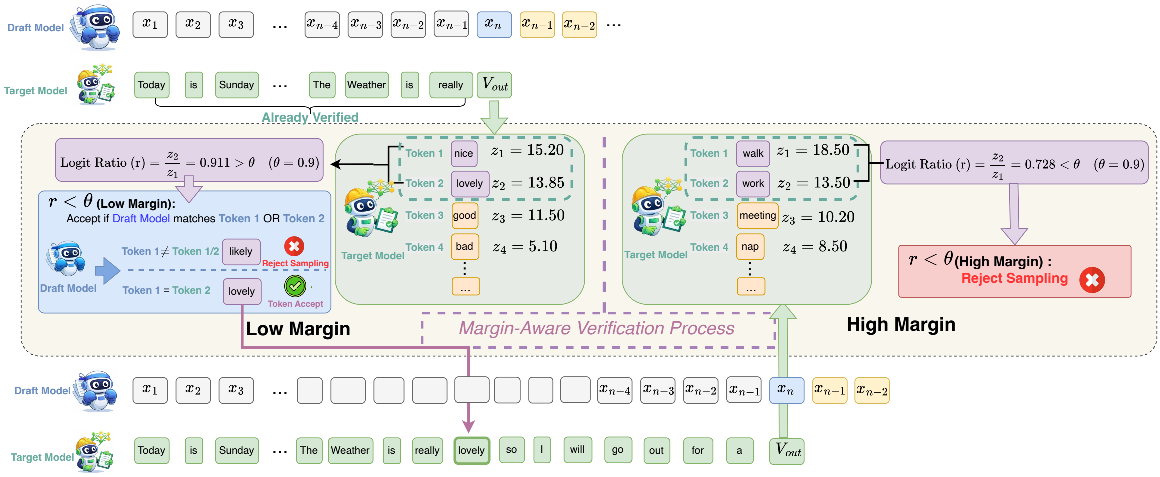


Figure 2: **Overview of the MARS Verification Workflow.** We illustrate the adaptive decision mechanism with a threshold $\theta = 0.9$. **(Left) Low-Margin Regime:** The target model exhibits weak preference between the top candidates "nice" and "lovely" (Logit Ratio $r = 0.911 > \theta$). Identifying this local indifference, MARS accepts the draft token "lovely" (Top-2) as a valid tie-breaker, avoiding unnecessary rollback. **(Right) High-Margin Regime:** The target model decisively prefers "walk" over the runner-up "work" (Logit Ratio $r = 0.728 < \theta$). MARS detects this stability and reverts to strict verification, rejecting the draft to preserve generation fidelity.

drafted token $\hat{v}_i$ is accepted with probability

$$\min\left(1,\ \frac{P_{\mathcal{M}_t}(\hat{v}_i \mid x_{<t}, \hat{v}_{<i})}{P_{\mathcal{M}_s}(\hat{v}_i \mid x_{<t}, \hat{v}_{<i})}\right), \tag{1}$$

where $P_{\mathcal{M}_s}$ and $P_{\mathcal{M}_t}$ denote the draft and target distributions. Upon rejection at position $i$, a corrected token is sampled from a residual distribution that compensates for the draft's approximation error. In both settings, **tree-based methods** (Miao et al., 2024; Li et al., 2024) extend this logic by organizing candidates into a token tree, allowing the target model to verify many competing prefixes in a single forward pass.

Our proposed method, MARS, modifies only the *accept/reject decision* during verification and is agnostic to the decoding strategy: it applies on top of both greedy and sampling-based verification, as well as chain- and tree-based draft structures.

# 3 Methodology

## 3.1 The Brittleness of Strict Verification

Speculative Decoding enforces strict token-level verification to guarantee exact output matching with the target model. This design implicitly treats the target model's argmax prediction as a decisive preference at every decoding step. In practice, however, large language models frequently encounter *low-margin regimes*, where the likelihood difference between the top candidate tokens is small. In such cases, the target model exhibits weak preference, and the choice of the argmax token becomes arbitrary. Rejecting a plausible draft token in these regimes provides negligible quality improvement, yet incurs a substantial computational cost due to rollback and resampling. This mismatch between verification cost and benefit motivates a verification strategy that adapts to the target model's local decisiveness, rather than enforcing rigid equality checks uniformly.

## 3.2 Output distribution formulation

Let $\mathcal{V}$ denote the vocabulary and $x_{<t}$ context at decoding step $t$. An autoregressive language model computes a hidden representation $\mathbf{h}_t \in \mathbb{R}^d$, which is projected to a logit vector

$$\mathbf{z}_t = \mathbf{W}_o \mathbf{h}_t + \mathbf{b}_o, \tag{2}$$

where $\mathbf{W}_o \in \mathbb{R}^{|\mathcal{V}| \times d}$ and $\mathbf{b}_o \in \mathbb{R}^{|\mathcal{V}|}$ are the output projection parameters. We denote the component of $\mathbf{z}_t$ corresponding to a token $v$ as $z_{t,v}$. The conditional probability of $v \in \mathcal{V}$ is obtained via the softmax transformation:

$$P(v \mid x_{<t}) = \frac{\exp(z_{t,v})}{\sum_{u \in \mathcal{V}} \exp(z_{t,u})}. \tag{3}$$

Here, the logits $\mathbf{z}_t$ represent unnormalized scores that encode the model's preference ranking over the vocabulary.

### 3.3 Logit Ratio as a Confidence Margin Indicator

To quantify the model's confidence margin at token level, we consider the ratio between the raw logits $\mathbf{z}_t$ of the top 2 candidate tokens at step $t$. Let $v^{(1)}$ and $v^{(2)}$ denote the top-1 and top-2 token, and their corresponding logits are $z_{t,v^{(1)}}$ and $z_{t,v^{(2)}}$ such that $z_{t,v^{(1)}} \geq z_{t,v^{(2)}}$. For brevity, we denote these sorted logits as $z_{t,(1)}$ and $z_{t,(2)}$.

We propose the **Logit Ratio** ($r_t$) to characterize the confidence margin, instead of using standard probability-based metrics. Specifically, we compute:

$$r_t = \frac{z_{t,(2)}}{z_{t,(1)}}, \tag{4}$$

A ratio approaching unity ($r_t \to 1$) implies that the gap between the best and second-best token is negligible, signaling *low decisiveness*. We empirically examined the logit distributions of our target models (Figure 1a) and observed that the top-ranked candidates consistently exhibit positive logit values. This *positive domain dominance* ensures that $r_t$ remains well-defined and strictly bounded in $(0, 1]$, where $r_t \to 1$ reliably signals low decisiveness.

We opt for logits because the softmax function introduces a non-linear distortion: it exponentially magnifies the difference between candidates based on their absolute magnitude (Figure 1b).

Specifically, the probability ratio $P(v^{(2)})/P(v^{(1)})$ is determined by the exponentiated difference of logits. Consequently, a fixed logit ratio ($z_{t,(2)}/z_{t,(1)}$) maps to vastly different probability ratios depending on the scale of $\mathbf{z}_t$ (Figure 1c). In regimes where logit values are large, the probability ratio becomes hypersensitive, often overestimating the model's actual confidence. The logit ratio avoids this exponential scaling, providing a metric that is invariant to the global magnitude of the output vector.

To ground the ratio metric in standard decision theory, we relate it to the logit margin, defined as $\Delta_t = z_{t,(1)} - z_{t,(2)}$. Substituting $z_{t,(2)} = z_{t,(1)} - \Delta_t$ into Eq. 4, the ratio can be rewritten as:

$$r_t = 1 - \frac{\Delta_t}{z_{t,(1)}}. \tag{5}$$

Consequently, enforcing the stability condition $r_t > \theta$ is equivalent to imposing an *adaptive margin constraint*:

$$\Delta_t < (1 - \theta) \cdot z_{t,(1)}. \tag{6}$$

Equation 6 reveals a critical property: the required margin for stability is not fixed, but scales linearly with the magnitude of the top logit. This counteracts the tendency of standard probability thresholds to overestimate confidence in high-logit regimes. By effectively demanding a larger raw margin $\Delta_t$ when the model projects larger values ($z_{t,(1)}$), the metric remains robust against calibration shifts that affect logit scale without altering the underlying preference ranking. Empirically, a substantial fraction of decoding steps fall into this low-margin region; Figure 1d visualizes the relaxation zone $r_t \in [0.9, 1.0]$, where the target model exhibits weak preference between the top candidates.

### 3.4 Margin-Aware Speculative Verification Algorithm

Building on the logit ratio analysis, we propose an *Margin-Aware Speculative Verification Algorithm* that dynamically adjusts the acceptance rigor. Unlike static verification, which enforces uniform exact matching, our method modulates the strictness of the check based on the target model's local decisiveness. This effectively balances the trade-off between *generation quality* and *decoding latency*, relaxing constraints only when the target model exhibits negligible preference differences.

Given a draft token $\hat{v}_t$ proposed by the draft model $\mathcal{M}_s$, the verification workflow is illustrated in Figure 2 and proceeds as follows:

- **Exact Match.** If $\hat{v}_t = v^{(1)}$, the draft token matches the target model's primary prediction. It is accepted immediately, consistent with standard speculative decoding.
- **Adaptive Relaxation.** If $\hat{v}_t = v^{(2)}$ and $r_t > \theta$, the target model exhibits weak preference between the top two candidates. Rejecting the draft in this regime would incur a high rollback cost for negligible quality improvement. We therefore accept the draft token $\hat{v}_t$, effectively treating the prediction as a tie.
- **Rejection and Correction.** Otherwise, the draft token represents a significant deviation from the target distribution. The draft is rejected, and the target model's top choice $v^{(1)}$ is used as the correct token (standard rollback).

Throughout this work, we use a fixed threshold $\theta = 0.9$, a choice justified by extensive ablation

**Algorithm 1** Margin-Aware Speculative Verification Algorithm

**Require:** Draft $\hat{V} = [\hat{v}_1, \ldots, \hat{v}_K]$, Target Model $\mathcal{M}_t$, Threshold $\theta$
**Ensure:** Verified sequence segment $V_{out}$

Initialize $V_{out} \leftarrow []$
Compute target logits $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_K]$ in parallel
**for** $i = 1$ to $K$ **do**
  Identify top-2 tokens $v^{(1)}, v^{(2)}$ from $\mathbf{z}_i$
  Compute ratio $r_i \leftarrow z_{i,(2)}/z_{i,(1)}$
  **if** $\hat{v}_i = v^{(1)}$ **then**
    $V_{out}$.append($\hat{v}_i$) ▷ Exact Match: Keep draft
  **else if** $\hat{v}_i = v^{(2)}$ **and** $r_i > \theta$ **then**
    $V_{out}$.append($\hat{v}_i$) ▷ Adaptive Relaxation: Keep draft
  **else**
    $V_{out}$.append($v^{(1)}$) ▷ Correction: Use Target Top-1
    **return** $V_{out}$ ▷ Stop and discard remaining draft
  **end if**
**end for**
**return** $V_{out}$ ▷ All draft tokens accepted

studies in Section 4.3. This policy effectively acts as a **cost-aware verification mechanism**: strict verification is relaxed only in regimes where the target model is indecisive and where the expected quality improvement from rejection is negligible relative to the computational cost. The complete procedure is summarized in Algorithm 1.

# 4 Experiments

## 4.1 Experimental Setup

In this section, we describe the models, datasets, evaluation metrics, baselines, and implementation details.

**Models.** We evaluate MARS on state-of-the-art open-source LLMs that cover both chat-oriented and reasoning-oriented settings, including Vicuna-13B v1.3 (Chiang et al., 2023), LLaMA-3.1-8B-Instruct, LLaMA-3.1-70B-Instruct (Grattafiori et al., 2024), Qwen3-8B, Qwen3-32B, and Qwen3-235B-A22B (Team, 2025). All models listed above are used as the **target models** to be accelerated. For each target model, we additionally employ the corresponding **EAGLE-3** draft model trained for that target to generate speculative drafts. MARS is applied *without* modifying the target model parameters. Due to hardware constraints, we do not include experiments on models larger than **235**B parameters.

**Datasets.** Following prior work on lossless LLM decoding acceleration (Li et al., 2024, 2025b), we evaluate on a diverse set of tasks that are commonly used for chat, code generation, reasoning, instruction following, translation, and summarization: MT-Bench (Zheng et al., 2023), HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021), GSM8K (Cobbe et al., 2021), AlpacaEval/Alpaca (Taori et al., 2023), WMT19 (Foundation, 2019), and CNN/DailyMail (Hermann et al., 2015).

**Evaluation Metrics.** Different from strictly lossless acceleration, our method is a *lossy* variant of speculative decoding (SD),which may deviate from the target model's exact token distribution. Therefore, we evaluate MARS along two axes: **generation quality** and **efficiency**. Specifically, we report:

- **Accuracy**: task-specific accuracy to quantify the quality impact introduced by lossy SD. For classification-style benchmarks, we report exact-match accuracy. For reasoning/QA benchmarks (e.g., GSM8K), we follow standard exact-match accuracy over final answers (Cobbe et al., 2021). For code-generation benchmarks (e.g., HumanEval), we report avg@4 as the primary accuracy-oriented metric (Chen et al., 2021).
- **Speedup Ratio**: end-to-end decoding speed relative to vanilla autoregressive decoding, measured under the same hardware, decoding parameters, and stopping criteria.
- **Average Acceptance Length** $\tau$: the average number of tokens committed per draft–verify cycle, capturing how many draft tokens are used per verification step (larger is generally more efficient).

**Comparison.** We use vanilla autoregressive decoding as the primary baseline (Speedup $= 1.00\times$). We compare MARS with representative lossless (or commonly used) decoding-acceleration methods, including standard speculative sampling (Leviathan et al., 2023; Chen et al., 2023), Prompt Lookup Decoding (Somasundaram et al., 2024),

| Model | Method | MT-bench | | HumanEval | | GSM8K | | Alpaca | | CNN/DM | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Speedup | $\tau$ | Speedup | $\tau$ | Speedup | $\tau$ | Speedup | $\tau$ | Speedup | $\tau$ | Speedup | $\tau$ |
| V 13B | SpS | 1.62× | 1.84 | 1.72× | 1.97 | 1.46× | 1.73 | 1.52× | 1.78 | 1.66× | 1.89 | 1.60× | 1.84 |
| | Lookahead | 1.45× | 1.39 | 1.41× | 1.25 | 1.61× | 1.60 | 1.36× | 1.36 | 1.26× | 1.20 | 1.42× | 1.36 |
| | PLD | 1.38× | 1.33 | 1.55× | 1.43 | 1.48× | 1.43 | 1.06× | 1.04 | 2.22× | 2.20 | 1.54× | 1.49 |
| | Medusa | 1.87× | 2.29 | 2.20× | 2.28 | 2.03× | 2.34 | 1.98× | 2.30 | 1.51× | 1.79 | 1.92× | 2.20 |
| | EAGLE | 2.32× | 3.20 | 2.65× | 3.63 | 2.57× | 3.60 | 2.45× | 3.57 | 2.23× | 3.26 | 2.44× | 3.45 |
| | EAGLE-2 | 2.90× | 4.38 | 3.01× | 4.90 | 2.97× | 4.30 | 2.89× | 4.35 | 2.46× | 3.94 | 2.85× | 4.37 |
| | EAGLE-3 | 3.24× | 5.46 | 3.60× | 6.10 | 3.26× | 5.65 | 3.20× | 5.41 | 2.69× | 5.57 | 3.12× | 5.64 |
| | MARS | **3.75×** | **7.11** | **4.20×** | **7.57** | **3.74×** | **7.19** | **3.74×** | **7.20** | **3.29×** | **6.94** | **3.74×** | **7.20** |
| L31 8B | EAGLE-3 | 3.63× | 5.67 | 2.90× | 4.10 | 3.41× | 4.87 | 3.51× | 5.09 | 2.77× | 4.36 | 3.24× | 4.82 |
| | MARS | **4.01×** | **6.78** | **4.09×** | **6.59** | **4.04×** | **6.68** | **4.26×** | **7.09** | **3.60×** | **5.91** | **4.00×** | **6.61** |
| L31 70B | EAGLE-3 | 4.20× | 5.41 | 4.72× | 5.95 | 4.62× | 5.94 | 4.81× | 6.04 | 3.63× | 4.96 | 4.40× | 5.66 |
| | MARS | **4.62×** | **6.24** | **5.12×** | **6.64** | **4.98×** | **6.66** | **5.09×** | **6.81** | **3.99×** | **5.61** | **4.76×** | **6.39** |
| Q3 8B | EAGLE-3 | 2.91× | 4.08 | 3.17× | 4.73 | 3.26× | 5.09 | 2.88× | 4.42 | 2.56× | 3.75 | 2.96× | 4.41 |
| | MARS | **3.14×** | **5.06** | **3.44×** | **5.32** | **3.69×** | **5.84** | **3.17×** | **5.47** | **2.70×** | **4.39** | **3.23×** | **5.22** |
| Q3 32B | EAGLE-3 | 2.62× | 3.63 | 3.30× | 4.78 | 3.53× | 5.12 | 2.79× | 3.56 | 2.82× | 3.97 | 3.01× | 4.21 |
| | MARS | **3.18×** | **4.79** | **3.73×** | **5.63** | **3.94×** | **6.04** | **3.34×** | **4.76** | **3.28×** | **4.93** | **3.49×** | **5.23** |
| Q3 235B | EAGLE-3 | 2.55× | 3.55 | 3.22× | 4.70 | 3.46× | 5.05 | 2.73× | 3.50 | 2.78× | 3.90 | 2.95× | 4.14 |
| | MARS | **3.12×** | **4.65** | **3.67×** | **5.55** | **3.88×** | **5.95** | **3.25×** | **4.70** | **3.20×** | **4.85** | **3.42×** | **5.14** |

Table 1: **Overall performance.** Speedup ratios and average acceptance lengths $\tau$ of different methods under non-greedy decoding (temperature= 1). V denotes Vicuna, L31 denotes LLaMA-Instruct 3.1, and Q3 denotes Qwen3. SpS denotes standard speculative sampling with Vicuna-68M as the draft model. Because of the characteristics of SD, once all drafts are accepted, the sequence will be appended with a bonus token generated by the target model. Therefore, the theoretical upper limit of $\tau$ is $K$+1, which is 8.

Medusa (Cai et al., 2024), Lookahead decoding (Fu et al., 2024), EAGLE-2 (Li et al., 2024), and EAGLE-3 (Li et al., 2025b). For all baselines, we follow the officially released implementations and default hyperparameters when available and otherwise match decoding settings (e.g., temperature, max length, and stopping criteria) to ensure fairness.

**Implementation Details.** For a fair comparison in our performance evaluation, we fix the draft length to 7 for all methods. In the draft process, we maintain Topk = 10 during the draft tree building, and no additional pruning is performed during the draft stage, keeping the same settings as Eagle3. Since our study primarily targets sampling scenarios with temperature $> 0$, we do not report results at temperature $= 0$. All experiments are conducted on NVIDIA H100 80GB GPUs. Models with $\leq 32$B parameters are benchmarked on a single GPU, while models with $> 32$B parameters are evaluated using 8 GPUs.

## 4.2 Overall Performance

Table 1 summarizes overall decoding efficiency at temperature= 1. Across all model families and scales, our method consistently improves over EAGLE-3, yielding higher speedup ratios and longer average accepted lengths $\tau$.

On **Vicuna-13B**, our method achieves a 3.74× mean speedup with $\tau = 7.20$, compared to 3.12× and $\tau = 5.64$ for EAGLE-3. The improvements remain consistent on **LLaMA-Instruct 3.1-8B** (4.00× vs. 3.24×) and **LLaMA-Instruct 3.1-70B** (4.76× vs. 4.40×), with $\tau$ increasing in all cases. We observe similar gains on **Qwen3-8B** (3.23× vs. 2.96×), **Qwen3-32B** (3.49× vs. 3.01×), and **Qwen3-235B-A22B** (3.42× vs. 2.95×).

Overall, the consistent increase in $\tau$ indicates better draft efficiency under stochastic decoding, translating into reliable end-to-end acceleration.

Although our experiments focus on sampling ($T$>0), MARS is equally applicable to greedy decoding ($T$=0). We provide $T$=0 results in Appendix B: MARS remains consistently faster than EAGLE-3 while preserving comparable task accuracy.

## 4.3 Ablation Studies

In this section, we quantify the contribution of key hyper-parameters in MARS and analyze the quality–efficiency trade-off under a unified evaluation setup. Unlike strictly lossless speculative decoding, our approach is a *lossy* SD variant, so we report both **task performance** and **decoding**

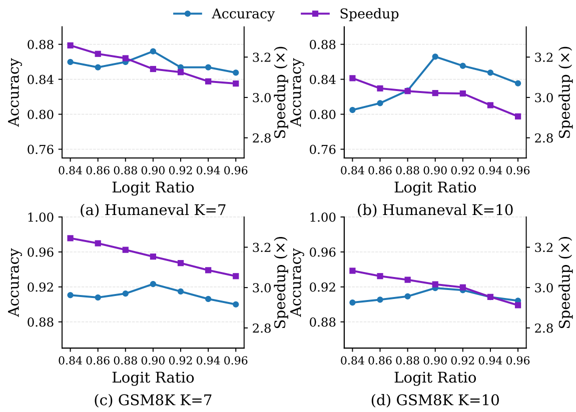


Figure 3: **Effect of the logit ratio threshold $\theta$ on the quality-efficiency trade-off.** We sweep over $\theta \in \{0.84, 0.86, 0.88, 0.90, 0.92, 0.94, 0.96\}$ and report accuracy (blue, left axis) and speedup (purple, right axis) on HumanEval and GSM8K with $K \in \{7, 10\}$ and $T$=1. Increasing $\theta$ consistently reduces speedup, while accuracy typically peaks around $\theta \approx 0.90$, indicating a balanced default choice.

**efficiency**. Concretely, we measure (i) **Accuracy** on each benchmark, (ii) **Speedup** (end-to-end latency reduction relative to vanilla autoregressive decoding), and (iii) **Precision**, which captures the agreement/consistency of the generated output with the vanilla autoregressive output under the same decoding configuration.

**Logit ratio threshold $\theta$.** The logit ratio threshold $\theta$ controls the degree of *adaptive relaxation* in MARS. Concretely, when the draft token $\hat{v}_i$ does not match the target model's top-1 token $v^{(1)}$ but matches the top-2 token $v^{(2)}$, MARS relaxes the verification rule and still commits $\hat{v}_i$ if the logit ratio $r_i = z_{i,(2)}/z_{i,(1)}$ exceeds $\theta$ (i.e., the top-1/top-2 margin is small). A smaller $\theta$ makes the policy more permissive, increasing speedup, but potentially introducing additional deviations; a larger $\theta$ yields more conservative verification with lower speedup.

As shown in Figure 3, speedup decreases monotonically as $\theta$ increases, while accuracy consistently peaks around $\theta \approx 0.90$ across both HumanEval and GSM8K with $K \in \{7, 10\}$ (all at $T$=1, Qwen3-8B). Notably, accuracy is *not* monotonically increasing with $\theta$ under sampling. This is because overly strict verification (high $\theta$) triggers frequent rollbacks, forcing the model to resample from the target's full distribution. Under $T$=1, this resampling can introduce stochastic tail noise that offsets the benefit of stricter verification.

The WMT19 results in Table 4 corroborate this pattern: on large-vocabulary models (Qwen3-8B/32B, ∼152K tokens), quality peaks near $\theta$=0.90 and fluctuates at higher values, whereas on Vicuna-13B (∼32K tokens) the trend is more monotonic, suggesting that the effect is amplified by vocabulary size. We therefore use $\theta = \mathbf{0.90}$ as the default for a favorable accuracy-speed trade-off.

**Temperature $t$.** Temperature $t$ reshapes the sampling distribution and thus affects both generation quality and SD behavior. As shown in Table 2, **Speedup** and $\boldsymbol{\tau}$ are largely stable across temperatures ($t \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$) on both GSM8K and HumanEval, while **accuracy consistently decreases as $t$ increases** (for both the AR baseline and our method). Overall, temperature do not affect the generation quality of our method within acceptable limits. Stable efficiency improvements are achieved under different temperature settings.

**Draft length $K$.** Table 2 shows that increasing $K$ leads to larger $\tau$ (more tokens committed per draft-verify cycle), but **speedup is not monotonic in $K$**. Moderate draft lengths (often $K$=9) achieve the best acceleration, whereas overly large $K$ can slightly reduce speedup, suggesting additional drafter/verification overhead. Accuracy varies with $K$ and task type (more sensitive on HumanEval than GSM8K), indicating a quality–efficiency trade-off. We treat these results as a *sensitivity analysis* and do not select hyperparameters based on this study.

### 4.4 Quality Preservation across Tasks and Granularities

Since MARS is a lossy variant of speculative decoding, we evaluate holistic generation quality on MT-Bench (GPT-5 as judge, $T$=1) across five models and eight categories; MARS preserves the baseline's average score within ±0.18 points on all models (see Appendix C for full results).

**Segment-Level Fidelity (CNN/DailyMail).** To measure whether relaxation alters the structural overlap between generated and reference texts, we report ROUGE-L on CNN/DailyMail (zero-shot, $K$=7, temperature= 1). As shown in Table 3, MARS scores are within 0.001–0.002 of the baseline across all five models (average: 0.1935 vs. 0.1943), well within the range of stochastic decoding variance. This indicates that the longest

| | | Temperature | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.2 | | | 0.4 | | | 0.6 | | | 0.8 | | | 1 | | |
| Dataset | $K$ | Speedup | $\tau$ | Acc | Speedup | $\tau$ | Acc | Speedup | $\tau$ | Acc | Speedup | $\tau$ | Acc | Speedup | $\tau$ | Acc |
| GSM8K | baseline | 1.0x | \ | 0.934 | 1.0x | \ | 0.932 | 1.0x | \ | 0.909 | 1.0x | \ | 0.901 | 1.0x | \ | 0.857 |
| | 6 | 3.96x | 5.79 | 0.929 | 3.89x | 5.75 | 0.927 | 3.97x | 5.76 | 0.894 | 3.95x | 5.76 | 0.896 | 3.97x | 5.74 | 0.851 |
| | 9 | 4.15x | 6.58 | 0.921 | 4.14x | 6.60 | 0.927 | 4.17x | 6.62 | 0.893 | 4.23x | 6.60 | 0.896 | 4.19x | 6.52 | 0.893 |
| | 12 | 4.09x | 6.87 | 0.921 | 4.07x | 6.85 | 0.924 | 4.13x | 6.83 | 0.904 | 4.05x | 6.80 | 0.904 | 4.07x | 6.76 | 0.916 |
| | 15 | 3.94x | 6.92 | 0.933 | 4.00x | 7.00 | 0.933 | 3.95x | 6.93 | 0.907 | 3.91x | 6.90 | 0.909 | 3.90x | 6.85 | 0.917 |
| Humaneval | baseline | 1.0x | \ | 0.896 | 1.0x | \ | 0.878 | 1.0x | \ | 0.878 | 1.0x | \ | 0.866 | 1.0x | \ | 0.854 |
| | 6 | 3.72x | 5.35 | 0.896 | 3.65x | 5.34 | 0.890 | 3.65x | 5.35 | 0.878 | 3.63x | 5.31 | 0.878 | 3.70x | 5.31 | 0.848 |
| | 9 | 3.84x | 5.98 | 0.884 | 3.86x | 5.94 | 0.860 | 3.84x | 5.93 | 0.860 | 3.83x | 5.93 | 0.842 | 3.90x | 6.03 | 0.866 |
| | 12 | 3.78x | 6.12 | 0.878 | 3.70x | 6.12 | 0.884 | 3.74x | 6.06 | 0.872 | 3.79x | 6.13 | 0.860 | 3.78x | 6.08 | 0.872 |
| | 15 | 3.58x | 6.18 | 0.878 | 3.59x | 6.15 | 0.878 | 3.61x | 6.08 | 0.854 | 3.63x | 6.14 | 0.860 | 3.59x | 6.14 | 0.848 |

Table 2: Ablation results on **Qwen3-32B** over **GSM8K** and **HumanEval**, studying the effects of **temperature** ($t$) and **draft length** ($K$) in MARS. For each dataset, we report **Speedup** (end-to-end decoding speed relative to vanilla autoregressive decoding), $\boldsymbol{\tau}$ (average committed tokens per draft–verify cycle), and **Acc** (task accuracy). "baseline" denotes vanilla autoregressive decoding under the same temperature.

| Method | L31 8B | L31 70B | Q3 8B | Q3 32B | V 13B |
|---|---|---|---|---|---|
| Baseline | 0.1699 | 0.1827 | 0.1972 | 0.1944 | 0.2272 |
| EAGLE-3 | 0.1707 | 0.1812 | 0.1985 | 0.1948 | 0.2260 |
| MARS | 0.1692 | 0.1812 | 0.1966 | 0.1945 | 0.2260 |

Table 3: Segment-level quality preservation on CNN/-DailyMail (ROUGE-L, zero-shot, $\theta$=0.9, $K$=7, $T$=1). For all models, the difference between the average MARS scores and the baseline is 0.0008.

common subsequence structure of the generated summaries is effectively unchanged.

**Token-Level Fidelity (WMT19).** We further evaluate on WMT19 Zh-En translation using BLEU and chrF($\beta$=2), both of which are sensitive to token and character-level deviations. Table 4 reports results across a range of thresholds $\theta \in \{0.84, \ldots, 0.98\}$. At the default $\theta$=0.9, MARS closely tracks the baseline on all models: for example, Qwen3-8B achieves 29.67 BLEU (baseline: 29.71) and 59.81 chrF (baseline: 60.25). Quality is well preserved at $\theta \geq 0.90$, while more aggressive relaxation ($\theta < 0.88$) can introduce measurable degradation, particularly on larger-vocabulary models.

### 4.5 Framework-Decoupled Verification

MARS modifies only the verification accept/reject rule and is therefore decoupled from the drafting architecture. To verify this, we integrate MARS into **Standard Speculative Decoding (SPD)**, keeping its draft→verify pipeline entirely unchanged and only replacing the verification criterion. We evaluate two draft→target pairs—Qwen3-0.6B→Qwen3-32B and LLaMA-3.1-8B→LLaMA-3.1-70B—on GSM8K, HumanEval, and WMT19 ($T$=1, $\gamma$=6).

| | Qwen3-8B | | Qwen3-32B | | Vicuna-13B | |
|---|---|---|---|---|---|---|
| Setting | BLEU | chrF | BLEU | chrF | BLEU | chrF |
| Baseline | 29.71 | 60.25 | 30.01 | 60.47 | 22.62 | 53.84 |
| EAGLE-3 | 29.99 | 60.40 | 30.02 | 60.35 | 22.26 | 53.66 |
| $\theta = 0.84$ | 28.70 | 58.88 | 26.28 | 58.61 | 20.33 | 52.32 |
| $\theta = 0.86$ | 29.12 | 58.97 | 26.98 | 59.01 | 20.40 | 52.56 |
| $\theta = 0.88$ | 29.34 | 59.12 | 29.08 | 59.53 | 21.21 | 52.42 |
| $\theta = 0.90$ | 29.67 | 59.81 | 29.76 | 60.10 | 21.50 | 53.19 |
| $\theta = 0.92$ | 29.59 | 59.46 | 29.62 | 60.10 | 21.58 | 53.14 |
| $\theta = 0.94$ | 29.10 | 59.61 | 29.57 | 60.18 | 21.95 | 53.35 |
| $\theta = 0.96$ | 29.39 | 59.88 | 28.29 | 60.07 | 21.90 | 53.09 |
| $\theta = 0.98$ | 29.22 | 59.78 | 28.57 | 60.24 | 21.97 | 53.73 |

Table 4: Token-level quality on WMT19 Zh-En (BLEU / chrF($\beta$=2), $K$=7, $T$=1) under varying logit ratio thresholds. Baseline and EAGLE-3 are shown for reference. Quality is well preserved at $\theta \geq 0.90$; aggressive relaxation ($\theta < 0.88$) degrades notably.

As shown in Table 5, MARS consistently increases the average accepted length $\tau$ over vanilla SPD on both model pairs, translating into higher end-to-end speedups, and task accuracy is preserved. These results confirm that MARS is not tied to EAGLE-3 and provides consistent gains as a plug-and-play verification strategy across different speculative decoding frameworks.

## 5 Related work

Speculative Decoding (SD) accelerates LLM inference by verifying cheap draft tokens in parallel, a paradigm established by (Leviathan et al., 2023; Chen et al., 2023). Current research enhances SD through three key dimensions: model alignment, verification strategies, and lossy trade-offs.

| Model | Method | GSM8K | | | HumanEval | | | WMT19 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Speedup | $\tau$ | Acc | Speedup | $\tau$ | Acc | Speedup | $\tau$ | BLEU |
| Qwen3 0.6B→32B | Baseline | 1× | – | 0.901 | 1× | – | 0.896 | 1× | – | 29.13 |
| | SPD | 2.15× | 4.23 | 0.893 | 2.23× | 4.33 | 0.896 | 1.51× | 2.92 | 29.16 |
| | SPD+MARS | 2.70× | 4.94 | 0.901 | 2.76× | 4.90 | 0.902 | 1.85× | 3.44 | 29.27 |
| LLaMA3 8B→70B | Baseline | 1× | – | 0.942 | 1× | – | 0.730 | 1× | – | 27.96 |
| | SPD | 3.05× | 5.07 | 0.934 | 3.15× | 5.73 | 0.712 | 2.20× | 4.16 | 27.84 |
| | SPD+MARS | 3.81× | 5.88 | 0.934 | 4.36× | 6.37 | 0.732 | 2.55× | 4.59 | 27.80 |

Table 5: Integration of MARS into Standard Speculative Decoding (SPD) ($T$=1, $\gamma$=6). MARS increases $\tau$ and speedup over vanilla SPD on both draft→target pairs while preserving or improving task accuracy, confirming its framework-decoupled nature.

### 5.1 Alignment of Draft and Target Models

The acceptance rate $\alpha$ relies heavily on the distributional alignment between models. To minimize divergence, **Knowledge Distillation** methods like DistillSpec (Zhou et al., 2024) and SSD (Spector and Re, 2023) fine-tune drafters to mimic the target, while (Liu et al., 2024) updates the drafter online using rejected tokens. Recently, **Architectural Alignment** has proven more effective than independent models. Medusa (Cai et al., 2024) and Hydra (Ankner et al., 2024) append decoding heads to the target backbone, whereas EAGLE (Li et al., 2025c, 2024) utilizes the target's feature history for context-aware drafting. Alternatively, REST (He et al., 2024) aligns generation with corpus statistics via retrieval-based drafting.

### 5.2 Better Verification Strategies

Standard verification often under-utilizes GPU parallelism. **Tree-based Verification** addresses this by evaluating token trees instead of chains. SpecInfer (Miao et al., 2024) and SpecTr (Sun et al., 2024) leverage tree attention masks to verify multiple branches in a single pass, with Sequoia (Chen et al., 2025) further optimizing tree structures dynamically. Distinctly, **Draft-Free** strategies operate without proxy models. Jacobi Decoding (Leviathan et al., 2023) employs fixed-point iteration for multi-token prediction. Similarly, Lookahead Decoding (Fu et al., 2024) utilizes parallel n-gram generation within the target model to break sequential dependencies efficiently.

### 5.3 Lossy Speculative Decoding

Recent research advances from lossless verification to *lossy* SD, trading negligible quality degradation for significant latency reduction. One primary direction relaxes verification criteria based on model confidence. Kim et al. (2023) proposed bypassing target model verification for high-confidence draft tokens, while Zhang et al. (2025) introduced adaptive thresholds tuned by context entropy to dynamically balance coherence and speed.

## 6 Conclusion

In this paper, we identify a fundamental inefficiency in speculative decoding: strict exact-match verification incurs substantial overhead in low-margin regimes where the target model exhibits weak decisiveness. To address this issue, we propose MARS, a training-free and plug-and-play verification strategy that adaptively adjusts acceptance rigor based on the target model's logit margin. By accepting plausible runner-up draft tokens in **low-margin regimes**, MARS significantly reduces unnecessary rollbacks while preserving generation quality. Extensive experiments across diverse model families (8B to 235B) show that our approach consistently outperforms state-of-the-art baselines, achieving a maximum speedup of **4.76**× on LLaMA-3.1-70B. Our results suggest that replacing rigid token matching with margin-aware adaptive verification is a critical step toward more efficient LLM inference.

## 7 Limitations

MARS focuses on adapting the verification rule based on local decision margins derived from the target model's logits. While we demonstrate that a single global threshold performs consistently across a wide range of models and tasks, more fine-grained or context-dependent adaptation strategies may further improve robustness. In addition, our method operates at the token level and is designed to be fully plug-and-play with existing speculative decoding frameworks; exploring margin-aware decisions at higher semantic or structural levels is beyond the scope of this work.

## References


Zachary Ankner, Rishab Parthasarathy, Aniruddha Nrusimha, Christopher Rinard, Jonathan Ragan-Kelley, and William Brandon. 2024. Hydra: Sequentially-dependent draft heads for medusa decoding. *arXiv preprint arXiv:2402.05109*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Gregor Bachmann and Dan Alistarh. 2025. Judge decoding: Faster speculative sampling requires going beyond model alignment. In *The Thirteenth International Conference on Learning Representations (ICLR)*. Oral Presentation.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code.

Zhuoming Chen, Avner May, Ruslan Svirschevski, Yuhsun Huang, Max Ryabinin, Zhihao Jia, and Beidi Chen. 2025. Sequoia: Scalable, robust, and hardware-aware speculative decoding. *Preprint*, arXiv:2402.12374.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Wikimedia Foundation. 2019. Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news.

Yichao Fu, Peter Bailis, Ion Stoica, and Hao Zhang. 2024. Break the sequential dependency of llm inference using lookahead decoding. *Preprint*, arXiv:2402.02057.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhenyu He, Zexuan Zhong, Tianle Cai, Jason D. Lee, and Di He. 2024. Rest: Retrieval-based speculative decoding. *Preprint*, arXiv:2311.08252.

Karl Moritz Hermann, Tomáš Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Sehoon Kim, Karttikeya Mangalam, Suhong Moon, Jitendra Malik, Michael W. Mahoney, Amir Gholami, and Kurt Keutzer. 2023. Speculative decoding with big little decoder. *Preprint*, arXiv:2302.07863.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. *Preprint*, arXiv:2211.17192.

Jinze Li, Yixing Xu, Guanchen Li, and 1 others. 2025a. Training-free loosely speculative decoding: Accepting semantically correct drafts beyond exact match. *arXiv preprint arXiv:2511.22972*.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024. Eagle-2: Faster inference of language models with dynamic draft trees. *Preprint*, arXiv:2406.16858.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2025b. Eagle-3: Scaling up inference acceleration of large language models via training-time test. *Preprint*, arXiv:2503.01840.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2025c. Eagle: Speculative sampling requires rethinking feature uncertainty. *Preprint*, arXiv:2401.15077.

Xiaoxuan Liu, Lanxiang Hu, Peter Bailis, Alvin Cheung, Zhijie Deng, Ion Stoica, and Hao Zhang. 2024. Online speculative decoding. *Preprint*, arXiv:2310.07177.

Xupeng Miao, Gabriele Oliaro, Zhihao He, Aaron Schulze, Cheng-Hao Parisot, Adam Paszke, and Zhihao Jia. 2024. Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

Shwetha Somasundaram, Anirudh Phukan, and Apoorv Saxena. 2024. Pld+: Accelerating llm inference by leveraging language model artifacts. *Preprint*, arXiv:2412.01447.

Benjamin Spector and Chris Re. 2023. Accelerating llm inference with staged speculative decoding. *Preprint*, arXiv:2308.04623.

Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix Yu. 2024. Spectr: Fast speculative decoding via optimal transport. *Preprint*, arXiv:2310.15141.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ziyin Zhang, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Rui Wang, and Zhaopeng Tu. 2025. Draft model knows when to stop: Self-verification speculative decoding for long-form generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16696–16708, Suzhou, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Yongchao Zhou, Kaifeng Lyu, Ankit Singh Rawat, Aditya Krishna Menon, Afshin Rostamizadeh, Sanjiv Kumar, Jean-François Kagy, and Rishabh Agarwal. 2024. Distillspec: Improving speculative decoding via knowledge distillation. *Preprint*, arXiv:2310.08461.

# A Additional Analysis

This appendix provides supplementary analyses for understanding the behavior of our relaxation criterion. Figure 4 reports the empirical distributions of (a) top-1 logit values, (b) logit ratios between the top-2 candidates ($z_2/z_1$), and (c) probability ratios between the top-2 candidates ($p_2/p_1$) on **Qwen3-8B**. We mark key reference points with red dashed lines: $x{=}0$ in (a), and a fixed threshold of 0.9 in (b–c).

# B Greedy Decoding Results

Although our experiments primarily target sampling settings ($T{>}0$), MARS is fully compatible with greedy decoding ($T{=}0$). We report results with $K{=}7$ in Table 6. Across all three model families, MARS consistently achieves higher speedup and longer average acceptance length $\tau$ than EAGLE-3 while maintaining comparable task accuracy. We note that under greedy decoding, both EAGLE-3 and MARS may exhibit minor accuracy deviations from the baseline due to numerical differences in tree attention and KV cache management; this is a known artifact of speculative decoding implementations rather than a property of the verification rule itself.

# C MT-Bench Fine-Grained Results

To validate generation quality beyond math and coding tasks, we evaluate on MT-Bench across eight diverse categories using GPT-5 as the judge (single-turn, $T{=}1$). As shown in Table 7, MARS preserves the baseline's generation quality across all categories and model scales, with average scores within $\pm 0.1$ of the baseline in most cases.

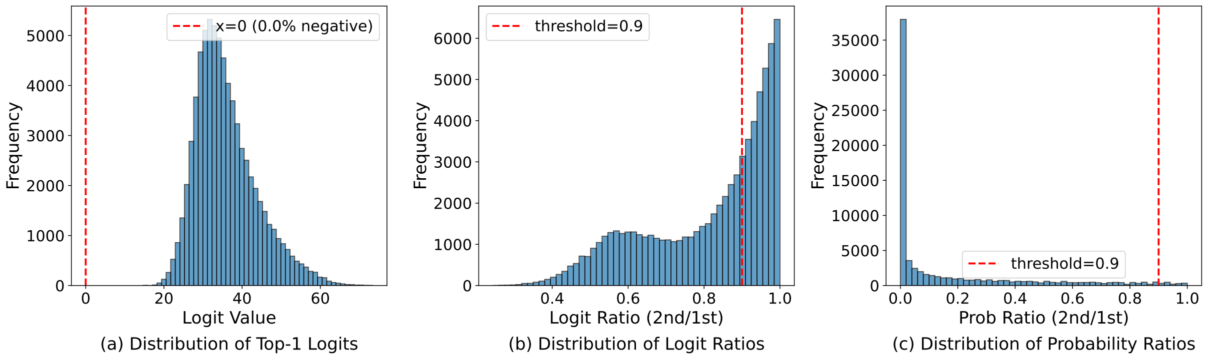


Figure 4: Histograms of top-2 statistics on **Qwen3-8B**. (a) Distribution of the top-1 logit values; the red dashed line marks $x$=0 (0.0% of top-1 logits are negative). (b) Distribution of logit ratios between the 2nd and 1st candidates ($z_2/z_1$); the red dashed line indicates the threshold (0.9). (c) Distribution of probability ratios between the 2nd and 1st candidates ($p_2/p_1$); the red dashed line indicates the same threshold (0.9).

| | | **GSM8K** | | | **HumanEval** | | | **Avg** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Method** | Speedup | $\tau$ | Acc | Speedup | $\tau$ | Acc | Speedup | $\tau$ | Acc |
| Qwen3-8B | Baseline | 1.00× | – | 0.967 | 1.00× | – | 0.835 | 1.00× | – | 0.901 |
| | EAGLE-3 | 2.83× | 5.20 | 0.959 | 2.53× | 4.87 | 0.848 | 2.68× | 5.03 | 0.903 |
| | MARS | 3.13× | 5.93 | 0.950 | 2.69× | 5.48 | 0.854 | 2.91× | 5.71 | 0.902 |
| Vicuna-13B | Baseline | 1.00× | – | 0.223 | 1.00× | – | 0.184 | 1.00× | – | 0.204 |
| | EAGLE-3 | 3.93× | 6.35 | 0.207 | 4.30× | 7.27 | 0.190 | 4.12× | 6.81 | 0.198 |
| | MARS | 4.13× | 7.18 | 0.215 | 4.61× | 7.73 | 0.183 | 4.37× | 7.46 | 0.199 |
| LLaMA-3.1-8B | Baseline | 1.00× | – | 0.893 | 1.00× | – | 0.677 | 1.00× | – | 0.785 |
| | EAGLE-3 | 3.57× | 5.86 | 0.901 | 4.10× | 6.42 | 0.677 | 3.93× | 6.14 | 0.789 |
| | MARS | 3.89× | 6.69 | 0.884 | 4.37× | 6.93 | 0.671 | 4.13× | 6.81 | 0.778 |

Table 6: Performance under greedy decoding ($T$=0, $K$=7). Speedup is relative to vanilla autoregressive decoding. $\tau$ denotes the average acceptance length per draft-verify cycle. Acc reports task-specific accuracy (exact-match for GSM8K, avg@4 for HumanEval).

| **Model** | **Method** | **Coding** | **Extraction** | **Humanities** | **Math** | **Reasoning** | **Roleplay** | **STEM** | **Writing** | **Avg** |
|---|---|---|---|---|---|---|---|---|---|---|
| Vicuna-13B | Baseline | 2.56 | 4.22 | 4.40 | 2.90 | 4.11 | 3.80 | 4.10 | 4.90 | 3.87 |
| | EAGLE-3 | 2.60 | 3.89 | 4.50 | 2.70 | 4.11 | 3.70 | 3.90 | 5.20 | 3.83 |
| | MARS | 2.60 | 4.22 | 4.50 | 2.80 | 4.11 | 3.80 | 3.90 | 5.20 | 3.89 |
| Qwen3-8B | Baseline | 6.22 | 7.78 | 6.20 | 10.00 | 7.67 | 6.30 | 7.78 | 7.30 | 7.41 |
| | EAGLE-3 | 6.11 | 7.78 | 6.50 | 10.00 | 7.78 | 6.50 | 7.67 | 7.10 | 7.43 |
| | MARS | 7.00 | 8.11 | 6.40 | 10.00 | 7.78 | 6.30 | 7.90 | 7.20 | 7.59 |
| Qwen3-32B | Baseline | 8.50 | 8.44 | 6.50 | 9.30 | 8.78 | 6.90 | 7.30 | 8.00 | 7.97 |
| | EAGLE-3 | 8.71 | 8.33 | 6.80 | 9.20 | 8.70 | 7.00 | 7.70 | 7.90 | 8.04 |
| | MARS | 9.29 | 8.44 | 6.70 | 9.30 | 8.56 | 6.90 | 7.30 | 8.10 | 8.07 |
| LLaMA-3.1-8B | Baseline | 4.12 | 6.33 | 4.50 | 6.80 | 6.56 | 5.00 | 4.70 | 6.10 | 5.51 |
| | EAGLE-3 | 4.71 | 6.22 | 4.20 | 6.80 | 6.67 | 5.10 | 4.70 | 6.40 | 5.60 |
| | MARS | 4.43 | 6.22 | 4.10 | 6.80 | 6.56 | 5.40 | 5.00 | 6.10 | 5.58 |
| LLaMA-3.1-70B | Baseline | 6.71 | 7.89 | 5.00 | 9.90 | 8.44 | 6.90 | 5.80 | 7.00 | 7.21 |
| | EAGLE-3 | 6.43 | 8.11 | 5.30 | 9.90 | 8.44 | 7.20 | 5.60 | 6.70 | 7.21 |
| | MARS | 6.25 | 8.00 | 5.10 | 9.90 | 8.33 | 6.90 | 5.60 | 6.90 | 7.12 |

Table 7: MT-Bench single-turn scores by category (GPT-5 judge, $T$=1). Higher is better. MARS shows no systematic quality degradation across diverse generation tasks.