EAGLE-2 vs. Medusa: Speculative Decoding in LLMs

Contextual Feature-Level Trees vs. Static Heads in Memory-Bound Inference Acceleration

Primary Category: Machine Learning Systems & Inference | Target Release: Q4 2026 Telemetry | Classification: Open Engineering Reference

EXECUTIVE ABSTRACT: Large language model autoregressive generation is strictly memory-bandwidth bound. Speculative decoding overcomes this by generating draft token trees in parallel. This dossier benchmarks EAGLE-2's feature-level contextual draft trees against Medusa's multi-head static predictions, quantifying token acceptance rates and hardware FLOP utilization on NVIDIA H100 GPU clusters.

Architecture Parameter	Medusa Multi-Head Drafting	EAGLE-2 Dynamic Feature Tree
Drafting Mechanism	4 Static Residual Heads	1-Layer Contextual Transformer
Average Acceptance Rate (α)	61.5% @ Top-1 Greed	84.2% @ Dynamic Top-K Tree
Wall-Clock Speedup (Llama-3 70B)	1.9x - 2.3x Baseline	2.8x - 3.4x Baseline
Draft Head SRAM Footprint	0.02 GB (Linear Heads)	0.15 GB (Feature Embeddings)

1. The Autoregressive Generation Memory Bandwidth Bottleneck

Standard transformer decoding reads the entire weight matrix from High Bandwidth Memory (HBM) for every generated token, achieving less than 5% tensor core FLOP utilization. Speculative decoding executes single-pass verification over K candidate tokens, amortizing memory reads over multiple accepted tokens.

2. Medusa: Residual Multi-Head Prediction Limits

Medusa extends the final transformer layer with multiple parallel linear heads, each predicting token $t+k$. However, because subsequent heads lack bidirectional cross-attention to earlier draft tokens in the current step, prediction error compounds exponentially, capping acceptance rates at 60-65%.

3. EAGLE-2: Contextual Feature Trees and Dynamic Tree-Attention

EAGLE-2 operates in the latent feature space rather than vocabulary logits. A lightweight single-layer transformer predicts the top hidden states conditioned on sequence context. Tree-attention masks dynamically expand candidate branches based on confidence scores, achieving acceptance rates exceeding 84%.

4. GPU Kernel Fusion & Verification TMA Amortization

During base model verification, EAGLE-2 constructs a custom tree-attention mask executed within a single fused Triton kernel. Leveraging Asynchronous Tensor Memory Accelerator (TMA) transfers on H100 GPUs, candidate verification adds zero extra HBM weight re-reads.

STRATEGIC ARCHITECTURAL CONCLUSIONS:

- EAGLE-2 achieves 2.8x to 3.4x real-world decoding acceleration by shifting speculation from vocabulary logits to contextual feature space.
- Dynamic tree-attention masking boosts candidate token acceptance from 61.5% (Medusa) to 84.2% across dense conversational benchmarks.
- Lossless rejection sampling guarantees that output distributions remain mathematically identical to standard autoregressive sampling.