

# EAGLE-2 Speculative Decoding Architecture

Feature-Level Speculative Drafting, Context-Aware Dynamic Trees, and Lossless Acceleration in Large Language Models

**EXECUTIVE ARCHITECTURAL SUMMARY:** Speculative decoding accelerates auto-regressive LLM inference by generating candidate token sequences with a lightweight draft mechanism and verifying them in parallel with the target model. Conventional draft models operate in raw token vocabulary space, suffering from low acceptance rates (~1.8 tokens/step). EAGLE-2 transitions speculative drafting into the top-layer hidden feature embedding space, combining contextual feature propagation with dynamic candidate tree construction (Tree Attention) to achieve 3.5 to 4.2 tokens per forward pass without loss of accuracy.

| Speculative Decoding Method        | Draft Overhead (FLOPs) | Acceptance Length (Tokens/Step) | Speedup on Llama-3-70B | Memory Overhead |
|------------------------------------|------------------------|---------------------------------|------------------------|-----------------|
| Vanilla Auto-Regressive (Baseline) | 0% (No Speculation)    | 1.00 token/step                 | 1.00x (Baseline)       | Zero            |
| Small Draft Model (Llama-3-8B)     | 12% - 18%              | 1.75 - 2.10 tokens/step         | 1.45x - 1.65x          | +16 GB VRAM     |
| Medusa (Multi-Head Heads)          | 4% - 6%                | 2.20 - 2.80 tokens/step         | 1.85x - 2.10x          | +1.2 GB VRAM    |
| EAGLE-2 (Dynamic Feature Trees)    | 1.5% - 2.5%            | 3.50 - 4.25 tokens/step         | 2.75x - 3.40x          | +0.4 GB VRAM    |

## 1. The Speculative Verification Paradigm

Auto-regressive language model generation is severely memory-bandwidth bound: generating one token requires reading billions of parameters from HBM to register SRAM.

- Speculative Execution: A fast draft generator proposes K candidate tokens. The large target LLM evaluates all K tokens in a single parallel forward pass using causal attention masking.
- Lossless Acceptance Criterion: Tokens are accepted or rejected based on speculative sampling:  $P_{\text{accept}} = \min(1, P_{\text{target}}(x) / P_{\text{draft}}(x))$ . Mathematical output distributions remain 100% identical to greedy or temperature-sampled generation from the base model.

## 2. The Collapse of Vocabulary-Level Draft Models

First-generation speculative decoding (Leviathan et al.) used small independent language models (e.g., Llama-3-8B drafting for Llama-3-70B).

- Semantic Misalignment: Small models frequently pick alternative valid synonyms that diverge from the large model's logits, causing speculative verification to reject valid tokens prematurely.
- Draft Compute Tax: Running an 8B model autoregressively consumes up to 20% of the GPU compute budget, wiping out speed gains when acceptance drops below 2.0 tokens/step.

### 3. EAGLE-2 Mechanics: Feature-Space Drafting & Dynamic Trees

EAGLE-2 bypasses vocabulary sampling by operating directly on the second-to-last transformer hidden layer embeddings (Feature-Level Drafting):

- Auto-Regressive Feature Head: A single lightweight transformer decoder layer takes the previous token's hidden state vector and predicts the next hidden feature representation.
- Context-Aware Dynamic Tree Construction: Unlike fixed tree masks (Medusa), EAGLE-2 evaluates the top-k entropy of candidate features. If the model is confident, it expands the draft tree deep (up to 8 steps); if uncertain, it widens the tree horizontally.
- Tree Attention Verification: The base LLM receives a non-linear 2D attention mask encoding all tree branches simultaneously, verifying dozens of candidate paths in a single GEMM operation.

### 4. Serving Infrastructure Integration (vLLM & TensorRT-LLM)

In production inference engines, EAGLE-2 draft heads are compiled into fused CUDA kernels that execute concurrently with PagedAttention KV-cache lookups.

- Zero Training Overhead on Base Model: The base target LLM weights remain 100% untouched. Only the single-layer EAGLE draft head is trained, requiring less than 48 hours on 8x A100 GPUs.

#### STRATEGIC ENGINEERING TAKEAWAYS & DEPLOYMENT RULES:

- EAGLE-2 generates speculative draft tokens in the hidden feature embedding space rather than token vocabulary logits.
- Dynamic tree attention adapts branch depth and width based on real-time prediction entropy, achieving 3.5+ tokens per step.
- Provides a 2.7x to 3.4x wall-clock inference speedup on 70B parameter models with zero loss in mathematical output fidelity.