SPEED-Bench: A Unified and Diverse Benchmark for Speculative Decoding

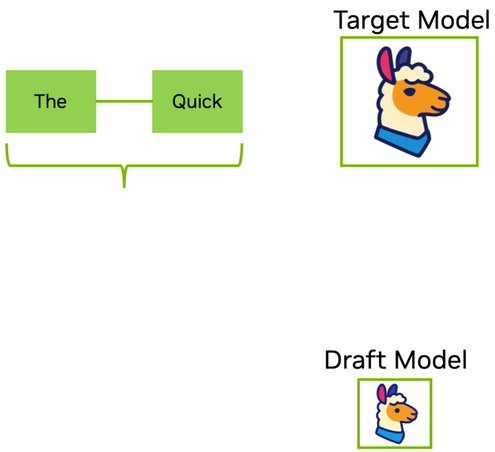
Talor Abramovich*, Maor Ashkenazi*, Izzy Putterman, Benjamin Chislett, Tiyasa Mitra, Bita Darvish Rouhani, Ran Zilberstein, Yonatan Geifman

ICML 2026

* Equal Contribution

Speculative Decoding

Background and Motivation: Why a Benchmark?



Speculative Decoding (SD) is a common technique for accelerating LLM inference

- By using a lightweight draft model to guess multiple future tokens, then verify in parallel using the target model

SD quality and inference speedups are inherently:

- **Data-dependent**
- **Serving-regime-dependent** (e.g., batch size)
- **System-dependent** (e.g., inference engine)

SD Benchmarking Today

Four Gaps in Existing Benchmarking Methodologies

1. **Fragmentation:** Papers evaluate on inconsistent datasets
2. **Diversity:** Existing benchmarks, such as *SpecBench*, rely on small prompt sets with limited semantic diversity
   - 10 samples per category
   - Input sequence length (ISL) < 100
3. **Environments:** Speedups need to be measured using production-like environments
   - Batch size > 1
   - Production inference engines such as vLLM, TensorRT-LLM and SGLang
4. **Hard to measure fixed-ISL high-batch loads:**
   - Standard throughput measurements use random tokens
   - SD speedups are highly data-dependent
   - Random tokens for throughput measurements can severely distort achieved speedups

SPEED-Bench introduces a benchmarking ecosystem for SD

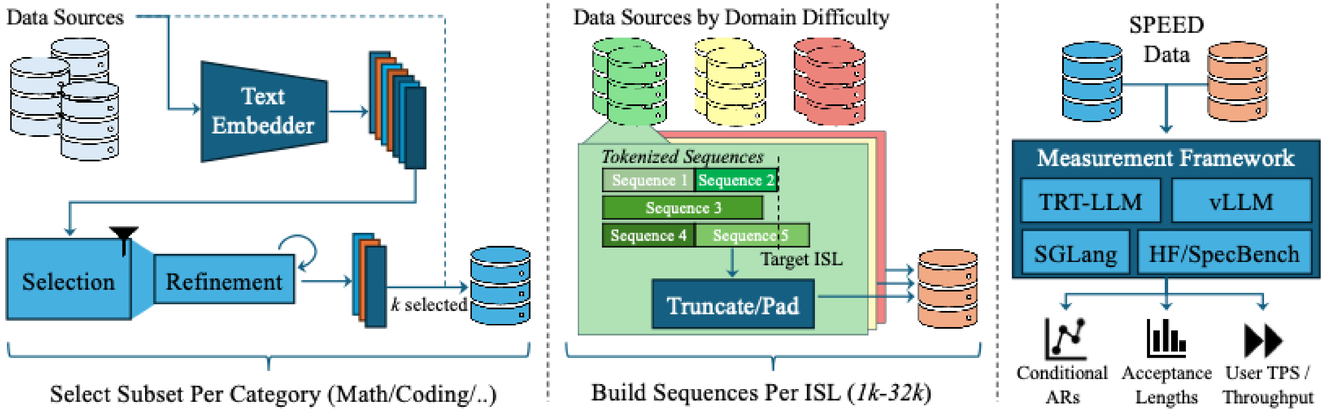
SPEED-Bench

One Ecosystem, Three Components

Qualitative Data Split: optimized for **semantic diversity** to measure **speculation quality** (draft accuracy) across domains

Throughput Data Split: constructed to evaluate **system-level speedups** across various **input sequence lengths** and **high-batch**

Measurement Framework: integrated with **production inference engines,** and standardizes evaluation across engines



The Qualitative Data Split

Diversity by Design

Goal: measure SD quality - **acceptance lengths (ALs)** - across diverse semantic domains

Sources: 18 publicly available benchmarks and datasets

Selection Algorithm: a greedy algorithm to **maximize semantic diversity** in the embedding space

Final split: **880 prompts**, uniformly distributed across **11 categories**

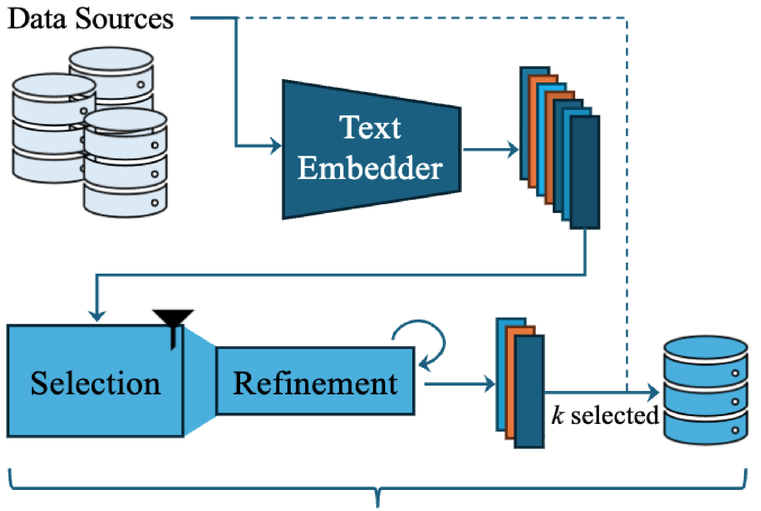


Figure 1: Qualitative Split Construction

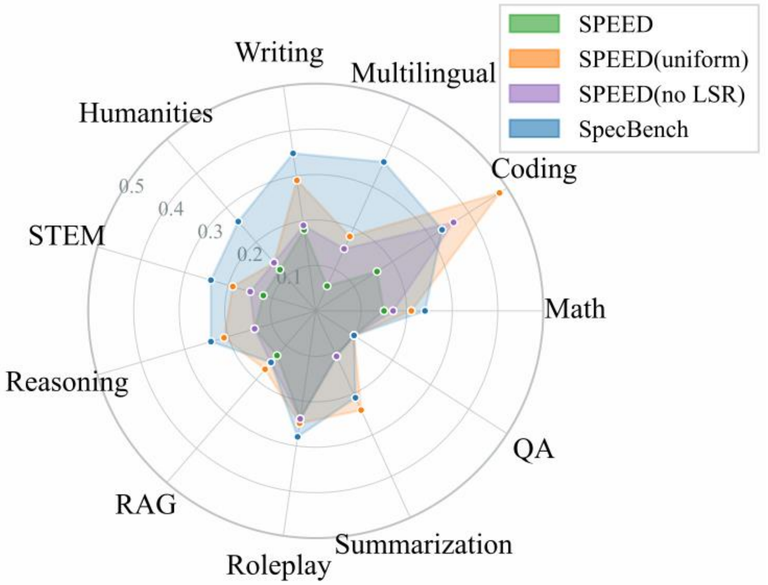


Figure 2: Comparison of average semantic similarity between samples (lower is better)

The Throughput Data Split

Realistic Serving Workloads

Goal: measure system-level speedups

- **Throughput**, output **tokens per second** (TPS)
- **Interactivity / User TPS**, per-request TPS

Sources: 9 publicly available benchmarks and datasets

Construction:

- 5 input sequence length (ISL) buckets, 1k, 2k, 8k, 16k, 32k
- Three coarse difficulty categories: low-, mixed- and high-entropy domains
- Batches up to 512 per category (1536 per-ISL bucket)

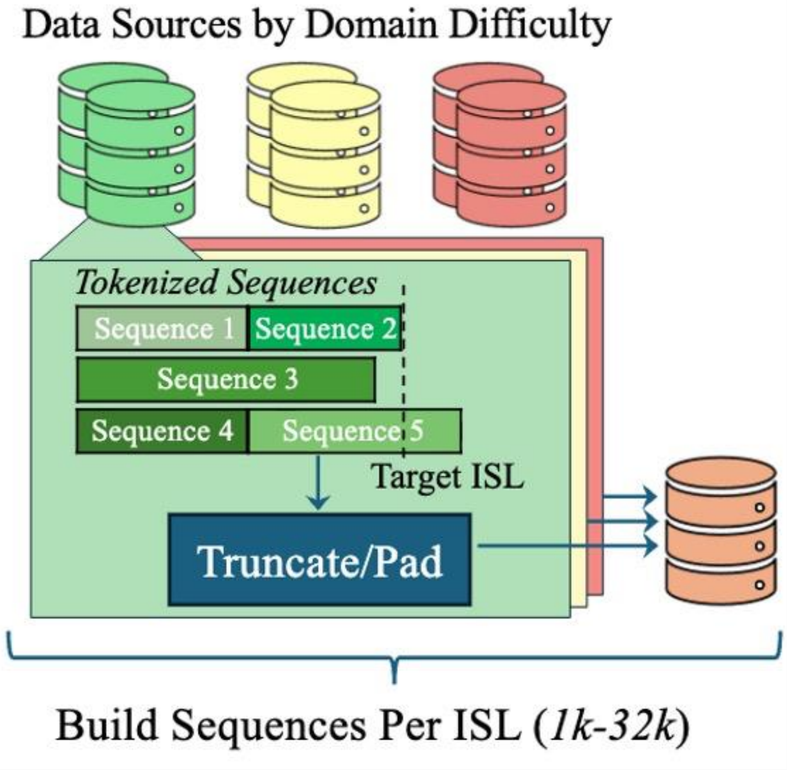


Figure 3: Throughput Split Construction

The Measurement Framework

Goal: measure using SPEED-Bench data across multiple inference engines

Problem: Different engines may apply different chat templates, handle BOS tokens differently or tokenize inputs inconsistently

Measurement framework: lightweight wrapper that handles external factors such as tokenization and prompt formatting, ensuring all systems process identical inputs

Metrics: acceptance length, acceptance rate, conditional acceptance rate, throughput and interactivity, and more fine-grained metrics

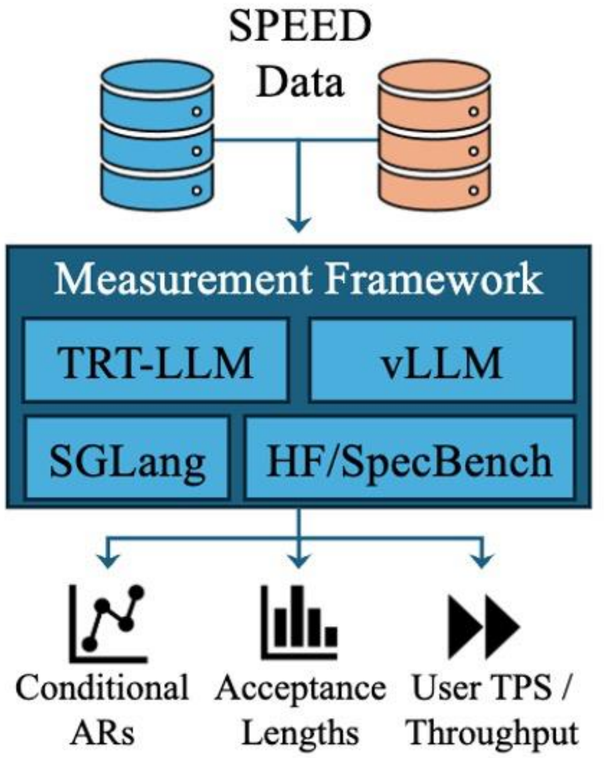


Figure 4: The Measurement Framework

SPEED-Bench Insights

Qualitative Data Split

SPEED-Bench semantic diversity helps to analyze and compare speculators quality better:

- In longer DLs and chain-drafting, we expect external draft to get high AL than EAGLE
- Unclear on SpecBench, SPEED-Bench reveals the gap

SPEED-Bench exposes side-effects of aggressive optimizations: EAGLE3 vocabulary pruning degrades AL unproportionally – highest drop on Multilingual, RAG and Summarization categories

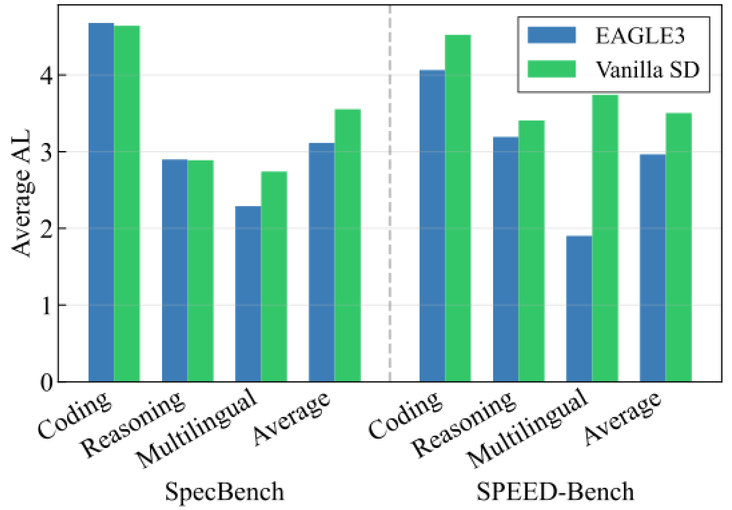


Figure 5: Average AL across selected categories in SpecBench vs SPEED-Bench. Target model is Llama 3.3 70B. DL=7

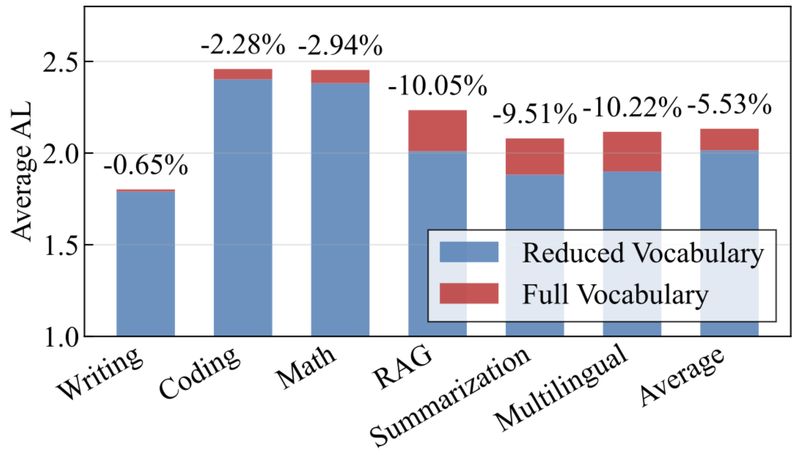


Figure 6: Average AL across selected categories using GPT-OSS-120B and EAGLE3 drafters (full vs. pruned vocabulary), DL=3

NVIDIA

SPEED-Bench Insights

Throughput Data Split

Random token trap: Random tokens often yield responses that has higher ALs, resulting in overestimation of throughput by an average of 23% for GPT-OSS

Optimal draft length: The higher the batch size, the lower the optimal draft length is

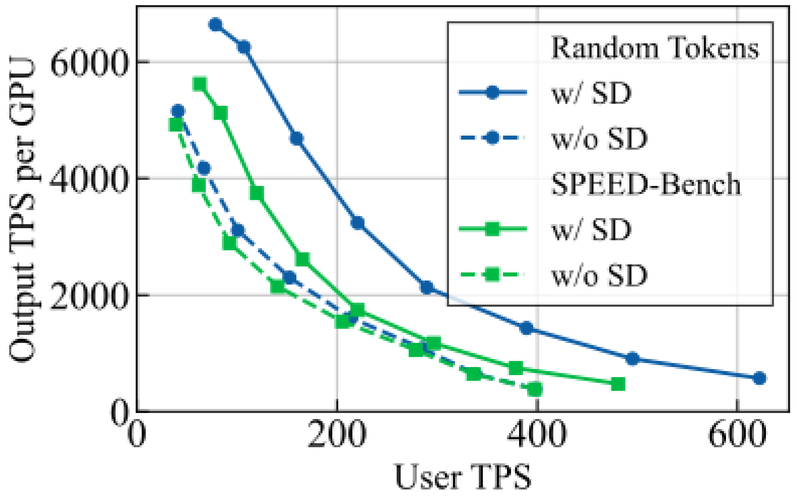


Figure 7: Throughput as a function of user TPS, comparing random input tokens to the Throughput Split (8k). Target is GPT-OSS 120B with EAGLE3 drafter, DL=3

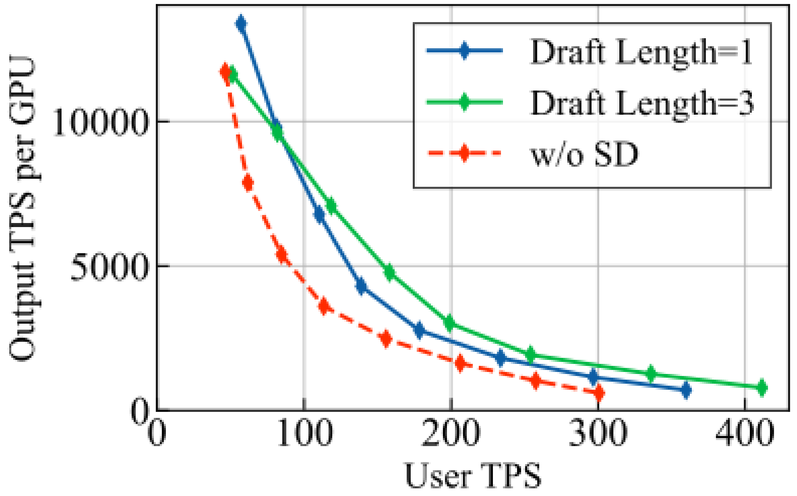


Figure 8: Throughput as a function of user TPS, comparing DL=1,3 on the Throughput Split (2k). Target is GPT-OSS 120B with EAGLE3

Try It Out!

SPEED-Bench

Data: huggingface.co/datasets/nvidia/SPEED-Bench

Paper: https://arxiv.org/abs/2604.09557


NVIDIA