NeurIPS 2026 Poster

ValuSpec

Plug-and-Play Candidate Valuation Before Target Verification for Tree-Based Speculative Decoding

Liang He[1*] **SiYuan Ma**[2] **Qishi Zhan**[3] **Yongqi Fan**[4] **Mingyu Cao**[5] **Qizhen Lan**[6] **Zhaolu Kang**[7] **Xilu Wang**[5†]

[1]Tongji University [2]Nanyang Technological University [3]Marquette University [4]Tencent [5]University of Surrey
[6]University of Texas Health Science Center at Houston [7]Peking University [*]hel9919@163.com [†]Corresponding author, wangxilu@surrey.ac.uk

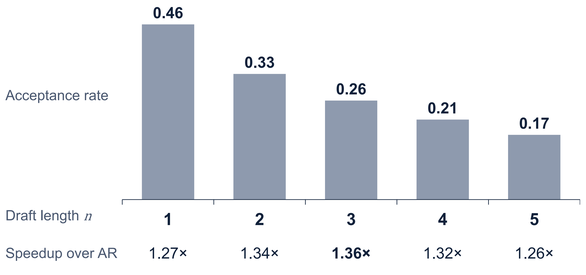
ValuSpec prunes low-value branches from a supplied candidate tree before the target model verifies it.
Up to 2.38× over autoregressive decoding on retrieval trees and 2.78× with EAGLE-3, with output identical to greedy decoding.

Bigger trees, more waste

Tree-based speculative decoding verifies many drafted branches in one target forward pass. Retrieval suppliers such as REST reuse historical drafts instead of regenerating them, but as the reused tree grows, many branches leave the target's greedy path early. They still cost verification, then get rejected.

More drafted tokens, lower acceptance

Standard SD on Dolly with Llama-2-7B on an A100. Bars show acceptance rate and the bottom row speedup over AR.



Where ValuSpec sits

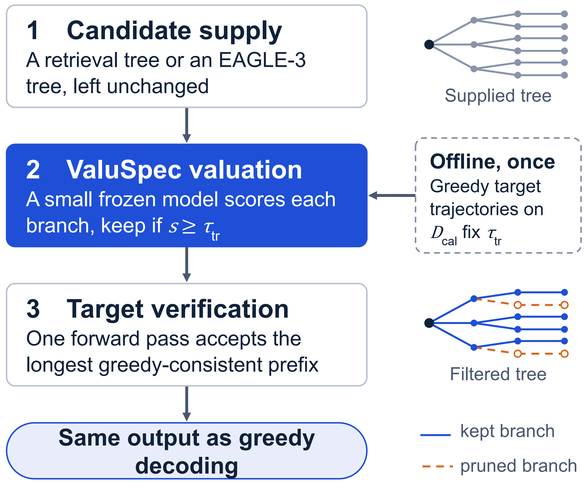
Conventional SD
Draft model writes a chain online → Target verifies the chain

Retrieval SD, e.g. REST
Retrieve a tree of past drafts → Target verifies every branch

ValuSpec
Supplied tree → **ValuSpec filters branches** → Target verifies the rest

Once candidates can be reused, the question is no longer how to generate them, but which ones are worth verifying.

Value before you verify



Transfer-based supervision

Entropy or top-1 confidence is not tied to what the target accepts. ValuSpec instead calibrates on the target model's own greedy trajectories, once and offline, with no parameter training ($\delta = 0.05$ throughout).

$$s_\phi(b \mid x) = (1/\ell) \Sigma_{i=1}^{\ell} \log q_\phi(y_i \mid x, y_{<i}) \quad (3)$$

$$\tau_{tr} = \text{Quantile}_\delta(\{ s_\phi(b^*(x) \mid x) \}_{x \in D_{cal}}) \quad (4)$$

$$\text{keep } b \in T_x \text{ if } s_\phi(b \mid x) \geq \tau_{tr} \quad (5)$$

Training-free
A frozen valuation model and one calibrated threshold

Plug-and-play
Works on any materialized tree, supplier untouched

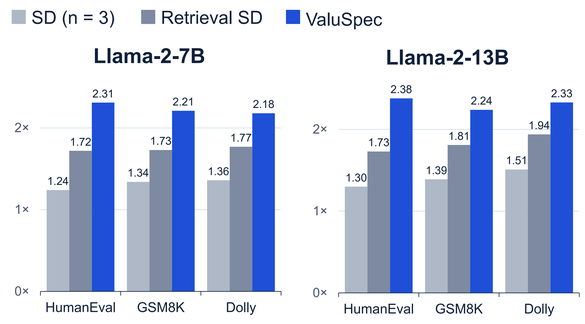
Output-identical
Greedy verification with AR fallback gives $Y = Y^*$

Small overhead
A 68M scorer in front of 7B and 13B targets

Up to 2.78× faster

Greedy decoding ($T = 0$), batch size 1, one A100. Each dataset is split 70% for calibration and 30% for evaluation with no overlap. Retrieval trees hold 64 draft tokens, and LLaMA-68M scores branches for Llama-2-7B and 13B targets.

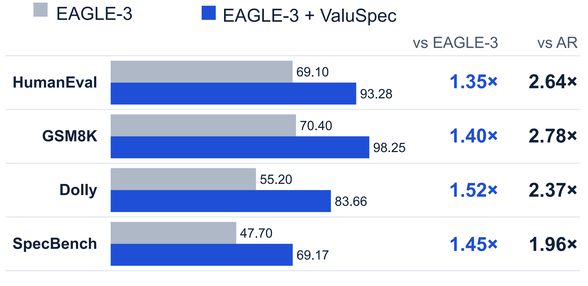
Retrieval trees, speedup over AR



ValuSpec reaches 2.18× to 2.38× over AR, while Retrieval SD on the same trees reaches 1.72× to 1.94×.

On top of EAGLE-3 dynamic trees

Llama-3.1-8B-Instruct in tokens/s. EAGLE-3 builds its tree as usual and ValuSpec only filters it; AR runs at 35.3.



Acceptance of retained candidates exceeds 86%, against 48.2% to 71.4% for EAGLE-3 alone.

Consumer GPU. On an RTX 4090 with Llama-2-7B, ValuSpec still reaches 1.67× to 1.84× over AR, ahead of SD and Retrieval SD.

Robust across settings

Thresholds transfer across datasets

τ_{tr} is calibrated on one dataset (rows) and applied to another (columns) without recalibration. Speedup over AR for Llama-2-7B and 13B; dashed cells are in-domain.

Calibrated on \ Evaluated on	HumanEval	GSM8K	Dolly
HumanEval	7B **2.31×** 13B **2.38×**	7B **2.17×** 13B **2.17×**	7B **2.05×** 13B **2.24×**
GSM8K	7B **2.29×** 13B **2.35×**	7B **2.21×** 13B **2.24×**	7B **2.08×** 13B **2.24×**
Dolly	7B **2.24×** 13B **2.30×**	7B **2.11×** 13B **2.07×**	7B **2.18×** 13B **2.33×**

Off-domain speedups stay between 2.05× and 2.35×.

Same module on Qwen3.5

Qwen3.5-9B target and Qwen3.5-0.8B valuation model. Ranges cover HumanEval, GSM8K and Dolly.

	Standard SD	Retrieval SD	EAGLE-3
Speedup over supplier	1.63–1.89×	1.97–2.14×	1.38–1.61×
Cross-dataset transfer	1.24–1.63×	1.78–2.01×	1.33–1.46×
Best fixed → ValuSpec	58.26 → 79.78	87.15 → 113.16	78.03 → 90.49
Branches kept	72.8–76.9%	79.1–82.6%	84.5–88.2%
False discards	1.35–1.91%	0.75–1.58%	0.13–0.44%
Batch 1/4/8	1.89/1.51/1.37×	2.01/1.73/1.47×	1.46/1.32/1.23×

The fixed row pits the best likelihood threshold from a 0.3 to 0.8 sweep against ValuSpec, in tokens/s on HumanEval. Empty trees stay below 1.4%.

Takeaways

- Valuation is its own stage between supply and verification, so it stacks on retrieval, EAGLE-3 and standard SD.
- A threshold calibrated once from target greedy runs transfers across datasets and still helps at batch sizes 4 and 8.
- **Limits.** Greedy decoding only for now; gains shrink when supplied trees are already selective or requests share little reusable structure.