3-Model Speculative Decoding

Sanghyun Byun, Mohanad Odema, Jung Guack,
Baisub Lee, Jacob Song, Woo Seong Chung

LG Electronics USA

Abstract

Speculative decoding (SD) accelerates inference in large language models (LLMs) by using a smaller draft model to propose tokens, which are then verified by a larger target model. However, the throughput gains of SD are fundamentally limited by a trade-off between draft model size and token acceptance: smaller draft models generate tokens more quickly but exhibit greater divergence from the target model, resulting in lower acceptance rates and reduced speedups. We introduce *Pyramid Speculative Decoding (PyramidSD)*, an extension of SD that inserts an intermediate qualifier model between the draft and target to bridge the distributional gap in output predictions, allowing smaller model to be used for drafting. This hierarchical decoding strategy improves alignment across models, enabling higher acceptance rates and allowing the use of significantly smaller draft models without sacrificing overall performance. PyramidSD builds on fuzzy acceptance criteria to support relaxed divergence thresholds at each stage, improving throughput. In experiments, PyramidSD achieves up to 1.91× generation speed over standard SD, reaching 124 tokens per second on a consumer GPU (RTX 4090). In small-memory settings with a 1B-parameter draft model and an 8B target model, PyramidSD minimally trades target model quality for improved throughput. Overall, PyramidSD offers a practical approach to enhancing speculative decoding efficiency and can be readily applied to existing inference pipelines.

1 Introduction

Large language models (LLMs) have achieved remarkable performance across a wide range of natural language processing tasks. However, their rapidly increasing parameter count has led to prohibitively high inference costs, making fast decoding a critical challenge for real-world deployment. Generating each token with a large model result in high latency (e.g., approximately 100ms for Llama-3.2-70B with IQ2_S on an RTX 4090), limiting interactivity and scalability.

To address these limitations, *speculative decoding* (SD) [7, 2] has emerged, leveraging a smaller, faster draft model to propose multiple tokens, which are then verified by a larger target model. If the target agrees with the draft's predictions, the tokens are accepted; otherwise, decoding falls back to the target. The overall throughput improves when the acceptance rate is high. However, as the size gap between the draft and target models increases, their output distributions diverge as well, leading to lower acceptance rates and diminishing the speedups achievable with SD.

We propose *Pyramid Speculative Decoding (PyramidSD)*, a new decoding framework that introduces an intermediate qualifier model between the draft and target models. This three-model hierarchy performs incremental verification across two speculative stages, bridging the distributional gap and improving acceptance rates without compromising correctness. PyramidSD builds on fuzzy speculative decoding by applying relaxed divergence thresholds, enabling higher throughput with smaller draft models. PyramidSD requires no additional training and works seamlessly with off-the-shelf families of models. Our key contributions are as follows:

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: SPIGM.

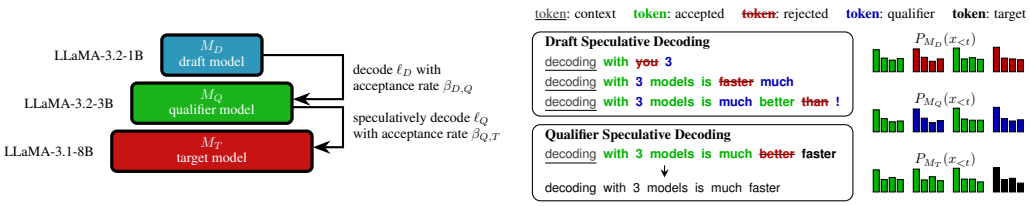

Figure 1: *PyramidSD* leverages the common practice of training family of models sharing a common tokenizer to effectively accelerate decoding speed. ℓ_D tokens are drafted by M_D and verified by M_Q until ℓ_Q tokens are generated, which are passed to M_T for final verification. As fuzzy threshold compares the logits, $\tau = (\tau_Q, \tau_T)$ can be adjusted to trade quality and acceptance rate β. P_{M_D} is replaced with P_{M_Q} in target comparison for closer distribution to target model. Example shown has been implemented on instruction-tuned LLaMA 3.2 and 3.1 family of models [4].

- We introduce a novel three-model decoding framework that accelerates inference with an intermediate qualifier model by layering two speculative steps.
- We derive two extensions of the proposed speculative decoding criterion, allowing fine-grained trade-off between performance, stability, and speed.
- We experimentally demonstrate that PyramidSD achieves higher decoding speeds than standard SD while maintaining relative performance.
- We provide a quantitative analysis of PyramidSD's acceleration behavior, showing how speculative lengths and fuzzy thresholds interact to determine efficiency.

2 Related Works

Several approaches proposed to improve speculative decoding efficiency. *Medusa* [1] and *EAGLE* [8] enhance draft quality through specialized architectures and uncertainty modeling. *SpecInfer* [9] introduces token tree verification to explore multiple decoding paths simultaneously, while *SpecTr* [11] applies optimal transport theory for improved token matching. *Lookahead decoding* [5] further accelerates inference by breaking sequential dependencies through parallel speculation windows.

Most relevant to our work is *Fuzzy Speculative Decoding* [6], which relaxes the strict equality constraint in standard SD by allowing tokens to be accepted when their output distributions are sufficiently similar under a divergence threshold. This approach improves acceptance rates but still faces significant limitations when the draft-target gap is large.

Recent work has also explored multi-stage speculative decoding. *Cascade speculative drafting* [3] chains multiple draft models, and *staged speculative decoding* [10] introduces progressive verification. However, these methods often require specialized training or complex coordination mechanisms, which limit their practicality. Similarly, *Ouroboros* [14] reuses recurring phrases to propose longer drafts with fewer forward passes, offering sizable acceleration without fine-tuning.

This motivates the need for approaches that preserve SD's acceleration benefits while maintaining high acceptance rates even across large performance gaps. Notably, many LLMs are released as families of models (e.g., Llama 3.2/3.1: 1B/3B/8B [4]) that share tokenizers and vocabularies. This architectural compatibility presents an underexplored opportunity: leveraging intermediate-sized models to bridge the distributional gap between small draft models and large target models.

3 3-Model Speculative Decoding

To push the performance limits of speculative decoding (SD), we exploit the fact that most model families come in at least three sizes with a shared tokenizer. We call this three-model variant *Pyramid Speculative Decoding (PyramidSD)* and describe its components below.

3.1 Motivation

Similar to SD, PyramidSD's speedup depends on the acceptance rate of tokens proposed by the draft model M_D and ultimately accepted by the target model M_T. As the performance gap between M_T

and M_D increases, their output distributions diverge, leading to lower acceptance rates:

$$\beta = P\left(\text{Div}(P_{M_T}(x_t), P_{M_D}(x_t)) \leq \tau\right),$$

where P_M denotes the logit distribution over the next token x_t, Div is a divergence measure, and τ is a divergence threshold. In standard SD, the Div formula between the target and the draft model is given by the difference in their logits while having $\tau = 0$, meaning the draft's top prediction must exactly match the target's.

PyramidSD addresses the low-acceptance rate problem by introducing an intermediate qualifier model M_Q between M_D and M_T, where parameter size $\|theta\|$ follows $\|theta_T\| > \|theta_Q\| > \|theta_D\|$. This is enabled by the prevalence of modern LLM families that share tokenizers and vocabularies, ensuring architectural compatibility across sizes. Empirically, we observe a systematic entropy gradient across model scales (Figure 3): larger models produce low-entropy, high-confidence distributions, whereas smaller models have higher uncertainty. The qualifier M_Q sits naturally between these extremes, making it well-suited to refine the draft's predictions by filtering out low-confidence tokens before they reach the target. This synergistically offers speedups to the costly inference process.

3.2 PyramidSD

If we directly extend the standard SD token acceptance criterion to three models, we encounter a key limitation: the qualifier M_Q contributes no benefit. Under strict equality matching (assisted decoding), a token x_t would be accepted only if the following holds:

$$P_{M_T}(x_t) \geq P_{M_Q}(x_t) \geq P_{M_D}(x_t), \quad x_t = \arg\max_{x \in \nu} P_{M_D}(x),$$

where ν is the vocabulary. Since the draft must already agree with the target, M_Q becomes redundant, nullifying its role. To overcome this, PyramidSD adopts Fuzzy Speculative Decoding [6] (FSD), relaxing acceptance criteria based on divergence thresholds instead of strict equality:

$$\text{Div}(P_{M_Q}(x_t), P_{M_D}(x_t)) \leq \tau_Q \quad \cap \quad \text{Div}(P_{M_T}(x_t), P_{M_Q}(x_t)) \leq \tau_T.$$

This two-stage criterion allows the qualifier to approve plausible tokens from the draft, increasing the effective acceptance rate β, while the target verifies correctness with more relaxed thresholds. Crucially, PyramidSD assumes P_{M_Q} lies closer to P_{M_T} than P_{M_D}, enabling efficient bridging between distributions.

3.3 Acceleration Analysis

The throughput (tokens/sec) of speculative decoding $\tilde{V}_{SD}$ with speculative length ℓ_D depends on both the acceptance rate $\beta_{T,D}$ and model speeds $\tilde{V}_D, \tilde{V}_T$. Following Leviathan et al. [7], the throughput under SD is:

$$\tilde{V}_{SD} = \beta_{T,D}(\ell_D + 1) \Bigg/ \left(\frac{\ell_D}{\tilde{V}_D} + \frac{1}{\tilde{V}_T}\right).$$

This extends naturally to FSD, where $\beta_{T,D}$ reflects divergence-threshold-based acceptance rather than exact token matching.

For PyramidSD, acceleration arises from two nested speculative stages: draft $\rightarrow$ qualifier and qualifier $\rightarrow$ target. We first compute the effective speed of fuzzy speculation between M_D and M_Q:

$$\tilde{V}_{FSD_{Q,D}} = \beta_{Q,D}(\ell_D + 1) \Bigg/ \left(\frac{\ell_D}{\tilde{V}_D} + \frac{1}{\tilde{V}_Q}\right).$$

Then the overall PyramidSD throughput becomes:

$$\tilde{V}_{PSD} = \beta_{T,Q}(\ell_Q + 1) \Bigg/ \left(\frac{\ell_Q}{\tilde{V}_{FSD_{Q,D}}} + \frac{1}{\tilde{V}_T}\right).$$

This decomposition highlights two key design levers: speculative lengths (ℓ_D, ℓ_Q) and acceptance thresholds (τ_Q, τ_T). Larger speculative lengths improve parallelism but may also reduce acceptance due to error compounding, while tighter thresholds stabilize acceptance at the cost of speed. Achieving optimal acceleration thus requires balancing these competing effects.

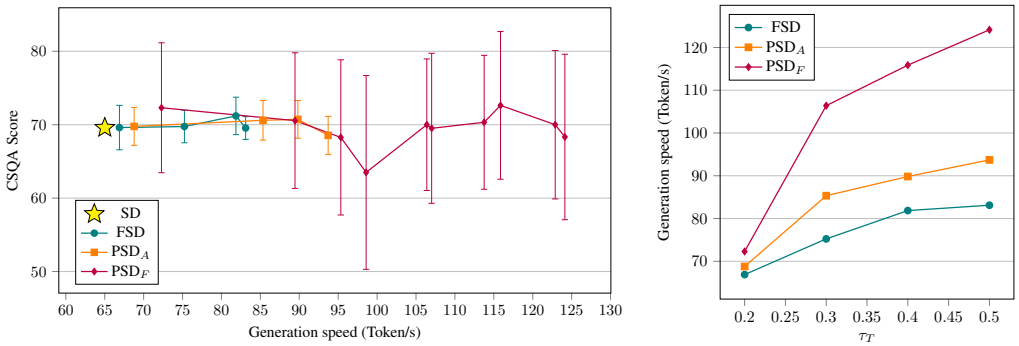

Figure 2: (Right) Evaluation result comparison on CSQA [12] dataset with Llama-3.1-8B, Llama-3.2-3B, and Llama-3.2-1B [4] Generation speeds for FSD, PSD$_A$, and PSD$_F$ are majorly controlled with τ_T (FSD, PSD) and τ_Q (PSD) adjustments. We report the fastest generation speed for each data point. Error bars represent standard deviation for CSQA score. (Left) Generation speed comparison with varying τ_T. Highest speeds are reported.

3.4 Assisted Decoding Variant

We introduce an assisted variant of PyramidSD (PSD$_A$) which improves stability by leveraging assisted decoding in place of τ_Q. When draft tokens are rejected by M_Q, instead of using stnadard SD probability rule, PSD$_A$ samples directly from the qualifier's distribution. This provides a guaranteed quality floor matching M_Q while still accelerating decoding via M_D, assuming no significant acceptance rate drop exists.

This design is motivated by the trade-off between speed and quality. High τ_Q values increase throughput by allowing more tokens from M_D to pass, but at the risk of forwarding low-quality predictions to M_T, potentially degrading output quality. PSD$_A$ resolves this tension by using M_Q as a fallback generator, combining the draft's acceleration with the qualifier's stability at the cost of reduced acceleration.

4 Experiments

4.1 Experimental Setup

We evaluate PyramidSD on CSQA [12] evaluation set using the Llama 3.2 and 3.1 model family [4]. All experiments are conducted on RTX 4090 GPUs, each with 24GB of memory. We build on the Hugging Face Transformers [13] implementation and use instruction-tuned variants: Llama 3.2 1B (draft), Llama 3.2 3B (qualifier), and Llama 3.1 8B (target) [4]. Unless otherwise noted, we set the maximum generation length to 2048 tokens and use a sampling temperature of 0.7. Following prior work [6], we evaluate each prompt using five pre-determined QA samples and report the mean and standard deviation of results.

We compare four decoding strategies: (1) Standard Speculative Decoding (SD) [7], (2) Fuzzy Speculative Decoding (FSD) [6], (3) PyramidSD with Assisted Decoding (PSD$_A$), and (4) PyramidSD with Fuzzy Acceptance (PSD$_F$). For hyperparameter selection, we perform a grid search over divergence thresholds $\tau_Q, \tau_T \in \{0.2, 0.3, 0.4, 0.5\}$ and speculative lengths $\ell_D \in \{2, 4, 6\}$, $\ell_Q \in \{1, 2, 3, 4, 5, 7, 10, 15, 20, 25\}$. This setup allows us to isolate the impact of different parameter configurations on both throughput and stability.

4.2 PyramidSD Performance

Figure 2 demonstrates the superior performance of PyramidSD variants compared to baseline methods across varying divergence thresholds. Both PSD$_A$ and PSD$_F$ achieve substantial speedups over standard SD and FSD, confirming the benefits of introducing an intermediate qualifier model. Among the two, PSD$_A$ consistently provides up to 1.44× improvement over SD while maintaining low variance across runs, making it particularly well-suited for production scenarios where predictable performance is critical.

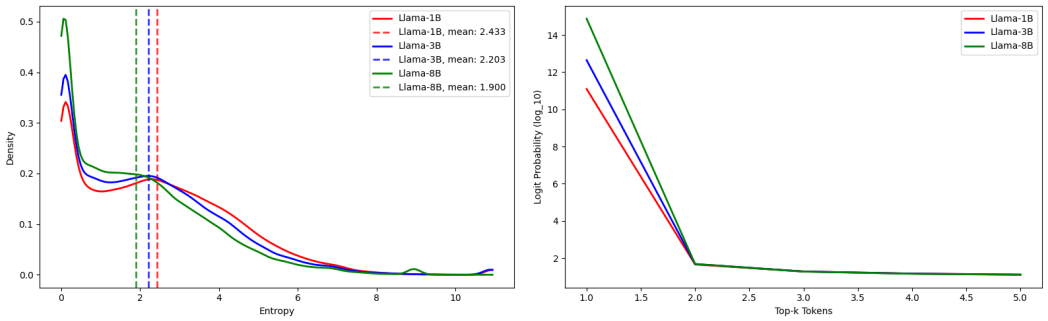

Figure 3: Behavior analysis on LLaMA 3.2/3.1 family with 1B, 3B, and 8B models on CSQA dataset. Both entropy and confidence scales with model size, revealing characteristics that can be capitalized for inference speed-up.

In contrast, PSD$_F$ delivers even greater acceleration—up to 1.91$\times$ faster than SD—but exhibits a larger standard deviation due to the compound effect of fuzzy acceptance across two speculative stages. The increased variability arises from the multiplicative nature of acceptance rates: errors at the draft-to-qualifier stage can propagate downstream, amplifying output fluctuations. However, when properly tuned, PSD$_F$ achieves the highest peak performance, as relaxed acceptance thresholds at both stages allow significantly larger speculative jumps.

In PyramidSD, performance is best maintained when setting $\tau_Q \leq \tau_T$: the qualifier first filters out clear mismatches while passing plausible candidates forward, allowing the target model to perform final verification efficiently. This hierarchical filtering helps PyramidSD balance speed and quality, achieving further acceleration while mostly preserving target-level correctness.

The benefits of PyramidSD become clearer when examining the performance of the two proposed variants, PSD$_A$ and PSD$_F$. PSD$_A$, which incorporates an assisted fallback, demonstrates robust stability across a wide range of threshold combinations. Its performance closely tracks FSD while achieving higher aggregate speedups and exhibits similarly robust sensitivity to thresholds as FSD. The fallback mechanism allows PyramidSD to reject low-confidence draft outputs early, preventing quality degradation. In contrast, PSD$_F$, which applies fuzzy relaxation at both the draft-to-qualifier and qualifier-to-target stages, achieves higher average throughput but displays noticeably larger variance across runs. This instability likely arises from compounding effects in the two fuzzy stages, where errors introduced early by the draft stage are amplified. These results underscore the importance of carefully tuning the pair (τ_Q, τ_T): a permissive τ_T combined with a moderately strict τ_Q can maximize PyramidSD's benefits, but extreme configurations in either direction diminish its advantage.

4.3 Model Behavior Analysis

Figure 3 illustrates this progression clearly. The 1B draft model produces relatively uniform entropy distributions, suggesting widespread uncertainty across its token predictions. The 3B qualifier model reduces this uncertainty, producing sharper distributions while still retaining flexibility, and the 8B target model shows strongly right-skewed entropy distributions, indicative of highly confident and selective predictions. This smooth gradient supports PyramidSD's core design: the qualifier operates naturally in an intermediate regime, capable of validating high-quality draft tokens while rejecting low-confidence outputs before they reach the target stage. As a result, the qualifier enables higher acceptance rates.

We further analyze token-level confidence by measuring the maximum predicted probability at each decoding step. The results align with the entropy findings: the 1B draft frequently exhibits low peak probabilities, reflecting its uncertainty and tendency toward diverse token sampling. The 3B qualifier achieves intermediate confidence, producing predictions that are sharper and better aligned with the target distribution. Finally, the 8B target produces consistently high-confidence predictions, selecting fewer but more reliable candidate tokens. This hierarchy of token confidence enables PyramidSD's cascaded filtering mechanism to function effectively. Together, the entropy gradient and confidence hierarchy demonstrate that the natural scaling properties validate the foundation of the proposed decoding scheme.

Table 1: Evaluation result comparison on CSQA [12] dataset with LLaMA-3.2-1B (draft for PSD), LLaMA-3.2-3B (qualifier for PSD / draft), and LlaMA-3.1-8B (target) [4] under different τ_Q/τ_T configurations, compared to vanilla and fuzzy speculative decoding. PSD_A and PSD_F denote that assisted and fuzzy speculative decoding (FSD) was used for qualifier model, respectively.

	SD [7]	FSD [6]				PSD_A			
τ_T	-	0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5
CSQA	69.58±2.20	69.61±3.03	69.75±2.21	71.20±2.55	69.55±1.55	69.77±2.58	70.60±2.70	70.73±2.57	68.55±2.59

	PSD_F									
τ_T	0.2	0.3		0.4			0.5			
τ_Q	0.2	0.2	0.3	0.2	0.3	0.4	0.2	0.3	0.4	0.5
CSQA	72.31±8.85	70.56±9.24	70.00±8.96	68.27±10.57	70.33±9.12	72.63±10.06	63.50±13.20	69.51±10.22	68.33±11.27	70.00±10.11

4.4 Ablation Studies

Table 1 presents detailed results for different divergence threshold combinations across four decoding strategies. Evaluations are performed on he CommonsenseQA (CSQA) [12] becnhmark. We report the evaluation scores with standard deviation across multiple ℓ_D/ℓ_Q pairs. As expected, standard SD maintains a stable CSQS score of 69.58 ± 2.20, since it relies on strict equality matching during verification. In contrast, Fuzzy SD exhibits a mild sensitivity to the divergence threshold τ_T. Performance does not scale directly with τ_T, achieving an average of 70.03 on CSQA; however, we observe that overly permissive thresholds ($\tau_T > 0.5$) degrade output quality without providing proportional speedups. This reflects a core limitation of single-stage fuzzy relaxation: while relaxing the matching criterion can improve acceptance rates, pushing it too far allows misaligned draft tokens to propagate, reducing overall correctness.

We conduct comprehensive ablation studies on divergence thresholds and speculative length configurations, with full results presented in Tables 2 and 3. We report the mean tokens per second for each configuration. These experiments provide deeper insights into the mechanisms driving PyramidSD's efficiency gains and reveal how the interplay between thresholds, speculative lengths, and verification stages shapes overall throughput.

Larger qualifier lengths (e.g., $\ell_Q = 25$) generally yield higher throughput, as they allow PyramidSD to propose longer speculative windows per iteration. However, simply increasing ℓ_D does not uniformly translate into speedups: beyond a certain point, the acceptance rate drops precipitously due to prediction errors. We find that setting ℓ_D to approximately half of ℓ_Q consistently achieves the best trade-off between speculation depth and acceptance stability, with optimal ratios typically falling between $1:2$ and $1:3$.

Finally, a striking observation from these studies is the presence of non-monotonic performance patterns across both thresholds and speculative lengths. Contrary to intuition, more aggressive speculation—whether through higher thresholds, longer draft sequences, or deeper relaxation—does not always yield higher throughput or lower performance. In several cases, medium-threshold, moderate-length configurations outperform more extreme setups. This suggests that PyramidSD's efficiency is governed by a subtle trade-off: aggressive speculation accelerates token generation but also increases the likelihood of rejection cascades, which ultimately reduces overall throughput. These findings highlight the need for careful joint tuning of speculative lengths and divergence thresholds to fully utilize the potential of three-model speculation.

5 Discussion

PyramidSD introduces additional complexity in its hyperparameter space. The interaction between divergence thresholds (τ_Q, τ_T) and speculative lengths (ℓ_D, ℓ_Q) strongly influences performance, and the optimal configuration varies across tasks, datasets, and model families. This variability further diminishes the plug-and-play capability of speculative decoding. Also, the multiplicative nature of acceptance rates across two speculative stages can lead to performance variability, especially for the fuzzy variant (PSD_F), where relaxed thresholds amplify fluctuations in throughput. Finally, PyramidSD currently relies on families of compatible models that share tokenizers and vocabularies, which may limit adoption for architectures without well-aligned models.

Despite these limitations, PyramidSD provides a new perspective on accelerating LLM inference. Our results suggest that hierarchical decoding strategies can exploit intrinsic scaling properties of model families to achieve better trade-offs between speed and accuracy than flat, single-stage approaches. Understanding and formalizing these dynamics opens opportunities for designing more principled speculative decoding algorithms.

6 Conclusion and Future Work

We introduced *Pyramid Speculative Decoding (PyramidSD)*, a hierarchical framework that accelerates large language model inference up to 1.91x over vanilla speculative decoding by introducing an intermediate qualifier model between draft and target models. By exploiting the natural entropy gradient across model scales and incorporating fuzzy acceptance criteria, PyramidSD achieves significant speedups while maintaining generation quality. Looking forward, several directions emerge for extending PyramidSD. One promising avenue is the development of adaptive controllers that dynamically adjust divergence thresholds and speculative lengths based on real-time decoding context, combining PSD_F's peak performance with PSD_A's stability. Another is generalization beyond homogeneous model families: enabling cross-family and mixed-attention setups would broaden PyramidSD's applicability, especially as emerging architectures diverge in tokenizers and scaling behaviors. Finally, understanding how model size ratios and entropy distributions influence acceptance rates could guide principled model selection and hyperparameter tuning. Overall, PyramidSD demonstrates that hierarchical speculative decoding is a powerful paradigm for efficient inference. As model families continue to grow and diversify, we believe this approach provides a foundation for next-generation decoding strategies that balance throughput, quality, and scalability.

References

[1] Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. In International Conference on Machine Learning, 2024.

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

[3] Ziyi Chen, Xiaocheng Yang, Yiwei Lin, Chenkai Zhang, Jianwei Zhu, and Yuxian He. Cascade speculative drafting for even faster llm inference. arXiv preprint arXiv:2312.05024, 2024.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

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

[6] Maximilian Holsman, Yukun Huang, and Bhuwan Dhingra. Fuzzy speculative decoding for a tunable accuracy-runtime tradeoff. arXiv preprint arXiv:2502.20704, 2025.

[7] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In International Conference on Machine Learning, pages 19274–19286. PMLR, 2023.

[8] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. In International Conference on Machine Learning, 2024.

[9] Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification. In ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2024.

[10] Benjamin Spector and Christopher Re. Staged speculative decoding for large language model acceleration. arXiv preprint arXiv:2308.04623, 2023.

[11] Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix Yu. Spectr: Fast speculative decoding via optimal transport. Advances in Neural Information Processing Systems, 36, 2023.

[12] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[14] W. Zhao and colleagues. Ouroboros: Generating longer drafts phrases by phrase reuse in speculative decoding. In EMNLP, 2024.

A Full Ablation Results

Table 2: Complete generation speed ablation of PSD with fuzzy threshold and draft length exploration. We evaluate the speed on LLaMA-3.2-1B (draft for PSD), LLaMA-3.2-3B (qualifier for PSD / draft), and LlaMA-3.1-8B (target) under different τ_T/τ_Q/ℓ_Q/ℓ_D configurations, compared to vanilla and fuzzy speculative decoding. PSD$_A$ and PSD$_F$ denote that assisted and fuzzy speculative decoding (FSD) was used for qualifier model, respectively. Due to space limitation, evaluation for PSD$_F$ continues in Table 3.

SD	ℓ_D	1	2	3	4	5	7	10	15	20	25										
	tok/s	50.89	58.38	61.83	63.65	65.01	63.42	56.65	60.96	50.13	42.03										
FSD	τ_T	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	ℓ_D	1	2	3	4	5	7	10	15	20	25	1	2	3	4	5	7	10	15	20	25
	tok/s	51.6	59.14	61.95	66.13	66.89	66.51	63.99	61.01	55.83	57.22	52.14	60.81	66.09	68.58	71.16	74.48	75.24	71.74	67.93	76.01
	τ_T	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	ℓ_D	1	2	3	4	5	7	10	15	20	25	1	2	3	4	5	7	10	15	20	25
	tok/s	52.42	61.06	66.36	69.81	71.9	76.81	77.98	81.86	80.23	75.75	52.36	61.34	67.06	70.77	72.09	76.1	79.56	83.11	82.26	80.22
PSD$_A$	τ_T	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	ℓ_Q	2	2	3	3	3	4	4	4	4	5	5	5	5	5	7	7	7	7	7	7
	ℓ_D	1	2	1	2	3	1	2	3	4	1	2	3	4	5	1	2	3	4	5	7
	tok/s	45.03	45.32	48.03	54.4	52.37	53.36	53.87	56.46	54.84	55.96	55.63	59.36	65.01	65.29	60.03	63.47	65.78	67.83	62.66	60.94
	τ_T	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2					
	ℓ_Q	10	10	10	10	10	10	10	15	15	15	15	15	15	15	15					
	ℓ_D	1	2	3	4	5	7	10	1	2	3	4	5	7	10	15					
	tok/s	59.86	62.63	65.02	68.78	65.5	67.89	55.44	60.83	64.5	68.17	64.52	56.92	77.45	55.75	53.4					
	τ_T	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
	ℓ_Q	20	20	20	20	20	20	20	20	20	25	25	25	25	25	25	25	25	25	25	
	ℓ_D	1	2	3	4	5	7	10	15	20	1	2	3	4	5	7	10	15	20	25	
	tok/s	48.56	57.38	64.47	55.41	69.79	51.95	44.98	60.32	45.52	50.33	44.75	59.07	55.23	64.08	72.49	52.48	55.88	34.19	30.89	
	τ_T	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	ℓ_Q	2	2	3	3	3	4	4	4	4	5	5	5	5	5	7	7	7	7	7	7
	ℓ_D	1	2	1	2	3	1	2	3	4	1	2	3	4	5	1	2	3	4	5	7
	tok/s	45.52	45.86	51.47	54.08	52.94	54.88	57.72	59.23	58.38	58.18	61.03	62	62.96	65.02	61.32	67.63	69.11	65.45	67	73.6
	τ_T	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3					
	ℓ_Q	10	10	10	10	10	10	10	15	15	15	15	15	15	15	15					
	ℓ_D	1	2	3	4	5	7	10	1	2	3	4	5	7	10	15					
	tok/s	69.84	75.56	74.92	76.46	74.72	73.31	81.87	70.39	84.95	86.36	73.7	80.92	70.77	67.81	82.29					
	τ_T	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
	ℓ_Q	20	20	20	20	20	20	20	20	20	25	25	25	25	25	25	25	25	25	25	
	ℓ_D	1	2	3	4	5	7	10	15	20	1	2	3	4	5	7	10	15	20	25	
	tok/s	68.97	71.16	72.74	85.34	81.01	81.4	78.35	64.89	55.97	65.91	78.05	81.77	79.67	83.86	70.19	74.13	73.26	59.18	60.8	
	τ_T	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
	ℓ_Q	2	2	3	3	3	4	4	4	4	5	5	5	5	5	7	7	7	7	7	7
	ℓ_D	1	2	1	2	3	1	2	3	4	1	2	3	4	5	1	2	3	4	5	7
	tok/s	45.97	45.77	51.78	54.69	54.39	56.58	58.82	59.79	63.59	58.85	61.12	63.49	67.18	65.36	62.86	69.15	72.88	71.22	72.42	72.47
	τ_T	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4					
	ℓ_Q	10	10	10	10	10	10	10	15	15	15	15	15	15	15	15					
	ℓ_D	1	2	3	4	5	7	10	1	2	3	4	5	7	10	15					
	tok/s	69.69	72.31	76.73	76.92	78.36	74.11	80.2	72.37	79.24	83.63	91.16	85.21	89.82	68.08	66.21					
	τ_T	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
	ℓ_Q	20	20	20	20	20	20	20	20	20	25	25	25	25	25	25	25	25	25	25	
	ℓ_D	1	2	3	4	5	7	10	15	20	1	2	3	4	5	7	10	15	20	25	
	tok/s	74.86	81.72	78.07	88.65	84.69	86.72	73.47	71.63	67.01	75.54	82.08	82.19	89.47	85.11	81.87	76.76	66.51	65.16	72.05	
	τ_T	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	ℓ_Q	2	2	3	3	3	4	4	4	4	5	5	5	5	5	7	7	7	7	7	7
	ℓ_D	1	2	1	2	3	1	2	3	4	1	2	3	4	5	1	2	3	4	5	7
	tok/s	46.12	46.05	51.75	54.23	54.5	56.81	59.09	60.74	59.2	59.76	62.99	65.08	65.89	64.73	65.38	69.16	72.51	71.04	71.88	71.18
	τ_T	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5					
	ℓ_Q	10	10	10	10	10	10	10	15	15	15	15	15	15	15	15					
	ℓ_D	1	2	3	4	5	7	10	1	2	3	4	5	7	10	15					
	tok/s	69.79	78.67	78.63	77.57	79.83	77.84	72.88	75.62	83.24	84.45	87.11	83.44	77.61	82.43	80.14					
	τ_T	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
	ℓ_Q	20	20	20	20	20	20	20	20	20	25	25	25	25	25	25	25	25	25	25	
	ℓ_D	1	2	3	4	5	7	10	15	20	1	2	3	4	5	7	10	15	20	25	
	tok/s	74.71	85.64	87.7	90.31	93.71	80.53	79.17	71.89	78.64	80.51	88.78	93.54	88.06	90.58	87.73	84.59	83.3	72.4	67.29	

Table 3: Continuation of Table 2 for PSD_F.

PSD_F																					
τ_I	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
ℓ_Q	2	2	3	3	3	4	4	4	4	5	5	5	5	5	7	7	7	7	7	7	7
ℓ_D	1	2	1	2	3	1	2	3	4	1	2	3	4	5	1	2	3	4	5	7	
tok/s	49.1	46.94	49.08	54.53	56.68	54.5	61.32	59.92	55.62	56.63	56.88	61.19	70.91	65.18	60.74	61.91	61.06	66.39	72.29	65.82	
τ_I	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
ℓ_Q	10	10	10	10	10	10	10	15	15	15	15	15	15	15	15	20	20	20	20	20	
ℓ_D	1	2	3	4	5	7	10	1	2	3	4	5	7	10	15	1	2	3	4	5	
tok/s	63.93	69.46	61.23	67.61	67.64	66.32	66.81	57.91	64.78	62.06	69.34	66.13	67.78	70.57	60.66	61.23	54.63	57.44	63.99	63.51	
τ_I	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
ℓ_Q	20	20	20	20	25	25	25	25	25	25	25	25	25	25	2	2	3	3	3	4	
ℓ_D	7	10	15	20	1	2	3	4	5	7	10	15	20	25	1	2	1	2	3	1	
tok/s	59.41	69.92	52.12	57.36	51.13	64.49	65.13	52.78	56.81	56.58	55.31	47.89	46.31	35.75	45.31	51.47	54.3	55.63	53.26	56.66	
τ_I	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
ℓ_Q	4	4	4	5	5	5	5	7	7	7	7	7	7	10	10	10	10	10	10		
ℓ_D	2	3	4	1	2	3	4	5	1	2	3	4	5	7	1	2	3	4	5	7	
tok/s	58.77	62.19	63.26	63.5	62.79	66.03	64.88	63.25	62.96	70.79	69.33	71.77	70.35	72.32	67.53	73.34	71.14	75.13	76.92	79.05	
τ_I	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
ℓ_Q	10	15	15	15	15	15	15	15	20	20	20	20	20	20	20	20	20	25	25		
ℓ_D	10	1	2	3	4	5	7	10	15	1	2	3	4	5	7	10	15	20	1	2	
tok/s	68.6	74.75	85.81	86.65	89.45	80.74	81.49	84.81	72.47	66.76	78.99	84.07	85.06	87.73	88.5	80.17	73.18	63.51	68.87	82.13	
τ_I	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
ℓ_Q	25	25	25	25	25	25	25	25	2	2	3	3	3	4	4	4	4	5	5	5	
ℓ_D	3	4	5	7	10	15	20	25	1	2	1	2	3	1	2	3	4	1	2	3	
tok/s	78.8	86.99	72.14	80.7	78.01	63.87	62.82	62.85	45.86	46.13	51.83	54.13	57.35	59.48	64.59	64.88	63.06	65.54	64.89	71.54	
τ_I	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
τ_Q	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
ℓ_Q	5	5	7	7	7	7	7	7	10	10	10	10	10	10	10	15	15	15	15	15	
ℓ_D	4	5	1	2	3	4	5	7	1	2	3	4	5	7	10	1	2	3	4	5	
tok/s	68.86	70.19	65.55	77.61	78.74	70.35	81.23	79.58	76.1	73.29	79.22	81.19	85.96	84.67	87.5	73.69	77.66	90.57	95.96	82.15	
τ_I	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
τ_Q	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
ℓ_Q	15	15	15	20	20	20	20	20	20	20	20	20	25	25	25	25	25	25	25	25	
ℓ_D	7	10	15	1	2	3	4	5	7	10	15	20	1	2	3	4	5	7	10	15	
tok/s	93.03	90.93	96.76	73.73	77.17	79.74	87.22	93.01	98.51	86.59	88.77	92.88	74.68	82.4	87.68	92.22	104.16	88.78	89.21	91.54	
τ_I	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
τ_Q	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
ℓ_Q	25	25	2	2	3	3	3	4	4	4	4	5	5	5	5	5	7	7	7	7	
ℓ_D	20	25	1	2	1	2	3	1	2	3	4	1	2	3	4	5	1	2	3	4	
tok/s	106.39	86.19	49.73	49.53	57.58	54.25	54.84	60.62	59.79	61.68	62.21	64.29	63.27	71.1	66.06	68.36	67.32	70.33	76.53	76.28	
τ_I	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
ℓ_Q	7	7	10	10	10	10	10	10	15	15	15	15	15	15	15	20	20	20			
ℓ_D	5	7	1	2	3	4	5	7	10	1	2	3	4	5	7	10	15	1	2	3	
tok/s	73.83	72.8	69.12	80.49	78.68	82.28	82.66	79.58	85.47	75.74	82.75	79.57	91.36	88.84	83.44	76.27	79.16	79.67	80.07	85.26	
τ_I	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	
ℓ_Q	20	20	20	20	20	20	25	25	25	25	25	25	25	25	25	25	2	2	3	3	
ℓ_D	4	5	7	10	15	20	1	2	3	4	5	7	10	15	20	25	1	2	1	2	
tok/s	92.57	95.34	84.63	79.86	75.25	78.68	78	88.34	75.36	85.9	51.69	88.78	93.54	80.67	91.63	59.64	46.24	46.32	52.4	55.49	
τ_I	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
τ_Q	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
ℓ_Q	3	4	4	4	4	5	5	5	5	5	7	7	7	7	7	10	10	10	10		
ℓ_D	3	1	2	3	4	1	2	3	4	5	1	2	3	4	5	7	1	2	3	4	
tok/s	60.52	56.54	59.3	64.78	63.68	66.88	62.86	73.5	75	77.25	66.55	74.94	75.85	72.77	77.7	82.72	74.39	78.87	80.21	92.29	
τ_I	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
τ_Q	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
ℓ_Q	10	10	10	15	15	15	15	15	15	15	15	20	20	20	20	20	20	20	20	20	
ℓ_D	5	7	10	1	2	3	4	5	7	10	15	1	2	3	4	5	7	10	15	20	
tok/s	86.63	89.17	83.47	81.65	84.65	92.92	102.27	95.22	93.81	95.81	91.77	79.58	87.52	96.32	95.93	92.51	103.94	108.96	88.84	98.46	
τ_I	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
τ_Q	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
ℓ_Q	25	25	25	25	2	2	25	25	25	25	2	2	3	3	3	4	4	4	4	5	
ℓ_D	1	2	3	4	5	7	10	15	20	25	1	2	1	2	3	1	2	3	4	1	
tok/s	80.9	82.12	90.26	92.05	92.92	104.38	113.76	101.7	89.91	76.38	53.91	53.08	51.54	60.76	56.24	57.28	61.34	66.83	64.99	66	
τ_I	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
τ_Q	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
ℓ_Q	5	5	5	5	7	7	7	7	7	7	10	10	10	10	10	15	15	15			
ℓ_D	2	3	4	5	1	2	3	4	5	7	1	2	3	4	5	7	10	1	2	3	
tok/s	68	70.05	71.99	77.29	65.64	74.67	76.17	78.36	83.27	83.62	79.64	82.72	83.74	90.42	86.91	88.67	93.53	72.11	89.37	94.27	
τ_I	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
τ_Q	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
ℓ_Q	15	15	15	15	15	20	20	20	20	20	20	20	20	25	25	25	25	25	25	25	
ℓ_D	4	5	7	10	15	1	2	3	4	5	7	10	15	20	1	2	3	4	5	7	
tok/s	98.83	98.33	103.27	102.76	106.11	86.38	92.36	98.49	102.03	100.6	101.76	108.66	103.33	115.86	80.47	88.49	97.12	104.43	105.93	108.63	
τ_I	0.4	0.4	0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.4	0.4	0.4	0.4	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
ℓ_Q	25	25	25	25	2	2	3	3	3	4	4	4	4	5	5	5	5	5	7	7	
ℓ_D	10	15	20	25	1	2	1	2	3	1	2	3	4	1	2	3	4	5	1	2	
tok/s	114.54	108.55	101.49	107.33	45.95	53.95	53.23	54.11	58.33	58.03	62.3	61.62	62.87	59.42	62.59	64.9	76.44	73.29	69.7	71	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
ℓ_Q	7	7	7	7	10	10	10	10	10	10	10	15	15	15	15	15	15	15	15	20	
ℓ_D	3	4	5	7	1	2	3	4	5	7	10	1	2	3	4	5	7	10	15	1	
tok/s	73.93	71.83	76.01	79.34	74.9	74.48	79.4	83.83	80.56	81.32	80.11	75.26	84.76	85.23	89.45	88.51	86.76	87.77	86.24	80.45	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	
ℓ_Q	20	20	20	20	20	20	20	20	25	25	25	25	25	25	25	25	25	25	2	2	
ℓ_D	2	3	4	5	7	10	15	20	1	2	3	4	5	7	10	15	20	25	1	2	
tok/s	85.94	91.14	88.04	92.29	89.14	85.25	77.51	82.54	78.42	90.02	88.14	98.59	95.35	93.08	80.16	81.26	71.89	71.5	46.39	53.77	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
ℓ_Q	3	3	3	4	4	4	4	5	5	5	5	5	7	7	7	7	7	7	10	10	
ℓ_D	1	2	3	1	2	3	4	1	2	3	4	5	1	2	3	4	5	7	1	2	
tok/s	54.8	55.77	56	56.98	60.28	65.38	68.66	60.93	63.81	70.55	77.17	70.81	67.61	76.86	73.53	76.58	81.88	88.2	73.2	85.19	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
ℓ_Q	10	10	10	10	10	15	15	15	15	15	15	15	15	20	20	20	20	20	20	20	
ℓ_D	3	4	5	7	10	1	2	3	4	5	7	10	15	1	2	3	4	5	7	10	
tok/s	83.12	86.55	88.01	89.26	84.91	82.95	88.22	95.04	93.62	101.24	98.89	95.22	104.95	79.37	84.72	95.01	97.71	100.82	104.32	94.23	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
ℓ_Q	20	20	25	25	25	25	25	25	25	25	25	25	2	2	3	3	3	4	4	4	
ℓ_D	15	20	1	2	3	4	5	7	10	15	20	25	1	2	1	2	3	1	2	3	
tok/s	101.51	95.64	82.27	92.03	96.27	102.56	101.99	102.15	105.05	103.59	88.27	107.01	51.07	46.56	52.87	62.19	56.26	62.62	64.38	67.31	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
ℓ_Q	4	5	5	5	5	5	7	7	7	7	7	7	10	10	10	10	10	10	10	15	
ℓ_D	4	1	2	3	4	5	1	2	3	4	5	7	1	2	3	4	5	7	10	1	
tok/s	61.78	64.41	64.76	66.33	72.45	71.6	66.71	78.66	82.5	76.88	85.69	90.27	74.97	84.22	81.91	89.47	90.18	94.45	100.13	77.43	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	
ℓ_Q	15	15	15	15	15	15	15	20	20	20	20	20	20	20	20	20	25	25	25	25	
ℓ_D	2	3	4	5	7	10	15	1	2	3	4	5	7	10	15	20	1	2	3	4	
tok/s	87.43	93.89	98.38	100.92	105.59	101.77	109.85	78.51	88.36	105.02	101.6	108.39	109.36	112.73	107.47	101.49	80.31	95.01	101.95	108.33	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.4	0.4	0.4	0.4	0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
ℓ_Q	25	25	25	25	25	25	2	2	3	3	3	4	4	4	4	5	5	5	5	5	
ℓ_D	5	7	10	15	20	25	1	2	1	2	3	1	2	3	4	1	2	3	4	5	
tok/s	108.07	111.38	110.31	124.13	112.2	104.16	54.7	49.68	52.38	58.75	56.61	61.11	60.83	65.1	65.22	62.72	73.05	74.43	82.13	73.12	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
ℓ_Q	7	7	7	7	7	7	10	10	10	10	10	10	10	15	15	15	15	15	15	15	
ℓ_D	1	2	3	4	5	7	1	2	3	4	5	7	10	1	2	3	4	5	7	10	
tok/s	66.84	77.12	75.9	81.95	85.05	85.6	71.8	86.15	83.73	97.95	90.62	89.85	102.75	77.72	88.49	100.63	104.01	96.89	106.09	105.44	
τ_I	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
τ_Q	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	
ℓ_Q	15	20	20	20	20	20	20	20	20	20	25	25	25	25	25	25	25	25	25	25	
ℓ_D	15	1	2	3	4	5	7	10	15	20	1	2	3	4	5	7	10	15	20	25	
tok/s	115.62	82.3	97.91	100.83	104.66	106.45	112.75	114.15	118.58	112.41	85.75	96.7	102	110.19	112.85	115.03	116.68	117.93	122.89	121.66	