Designing Draft Models for Speculative Decoding

Anonymous ACL submission

Abstract

Speculative Decoding is a widely used technique to speed up inference for Large Language Models (LLMs) without sacrificing quality. When performing inference, speculative decoding uses a smaller *draft* model to generate speculative tokens and then uses the *target* LLM to verify those draft tokens. The speedup provided by speculative decoding heavily depends on the choice of the draft model. In this work, we perform a detailed study comprising over 350 experiments with LLaMA-65B and OPT-66B using speculative decoding and delineate the factors that affect the performance gain provided by speculative decoding. Our experiments indicate that the performance of speculative decoding depends heavily on the latency of the draft model, and the draft model's capability in language modeling does not correlate strongly with its performance in speculative decoding. Based on these insights we explore a new design space for draft models and design hardware-efficient draft models for speculative decoding. Our newly designed draft model for LLaMA-65B can provide 60% higher throughput than existing draft models and can generalize further to the LLaMA-2 model family and supervised fine-tuned models.

1 Introduction

In recent years, Large Language Models (LLMs) have emerged as a cornerstone of modern computational linguistics, offering unprecedented capabilities in generating and interpreting human language. As the demand for faster and more efficient language processing grows, understanding and optimizing the inference throughput of these models becomes increasingly crucial. Decoder-only LLMs (Brown et al., 2020; Touvron et al., 2023a,b) use auto-regressive decoding to perform inference. Auto-regressive decoding is known to be hardware inefficient (Miao et al., 2023; Liu et al., 2023a), leading to poor resource utilization and low throughput during inference.

Several methods (Yu et al., 2022; Wang et al., 2020; Kwon et al., 2023; Dao et al., 2023; Hong et al., 2023) have been studied to optimize the serving of LLMs. One promising approach to improve the throughput for serving LLMs without accuracy loss is speculative decoding (Stern et al., 2018; Xia et al., 2023a; Leviathan et al., 2023). When using speculative decoding to serve an LLM (usually 100s of billion parameters), a draft model (a significantly smaller LLM) is used to generate speculative tokens. The target LLM model then verifies the output of the draft model and only outputs tokens that match its output. In the case of speculative decoding, the target LLM for inference acts as a *verifier* for the draft model. By leveraging faster inference of smaller draft models, speculative decoding turns autoregressive decoding on the target LLM into a more hardware-friendly batched operation (similar to "prefill"), thereby increasing throughput while preserving accuracy.

Given the promised benefits of speculative decoding, this paper first focuses on understanding the key factors that dictate the throughput improvements that can be obtained. We perform a comprehensive benchmarking study and profile speculative decoding to characterize bottlenecks. We perform over 350 experiments, using LLMs like LLaMA-65B, OPT-66B, and fine-tuned chat models such as Vicuna-33B (Chiang et al., 2023) as target models and LLaMA and OPT families as draft models, ranging from $\approx 5\times$ to $528\times$ fewer parameters than the target models. Our findings show that the key bottleneck in speculative decoding is the draft model's latency, highlighting the need to optimize draft model designs.

Next, we find that existing draft models, which are typically designed only for improving accuracy in a given parameter budget, are sub-optimal for maximizing the throughput with speculative decod-

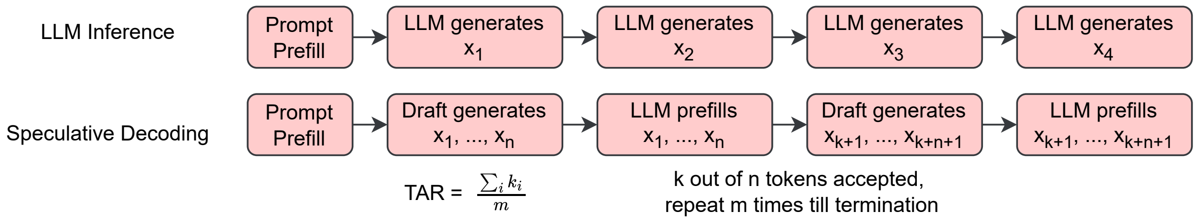


Figure 1: This figure shows the speculative decoding process. In vanilla LLM inference, after the prompt is processed into KV caches (Prefill phase), LLM generates the output token by token in an autoregressive manner (Autoregressive generation phase). In speculative decoding, a draft model is first used to generate n candidate tokens at each step (Draft token generation phase). The LLM verifies the candidate tokens and accepts k ($k \leq n$) tokens (LLM verification phase). Since LLM knows all n candidate tokens in advance, this step is identical to a prefill step of length n. In both cases, this process is repeated until either an end-of-sequence (EOS) token is generated or the maximum generation limit has been reached.

ing. From our experiments, we first observe that the draft model latency is bottlenecked by model depth, and higher model depth leads to increased latency (Section 3.2). Second, we also find that draft model accuracy on language modeling tasks does not correlate strongly with its performance in speculative decoding (Section 3.3), *i.e.*, a draft model with higher accuracy on language modeling task can have similar TAR to a model with lower accuracy. Based on these two insights, we propose designing new draft models that trade increased depth for width (thus retaining the same parameter count) and show that our new draft models can boost inference throughput using speculative decoding by over 60%. Finally, we show how pruning methods like Sheared-LLaMA (Xia et al., 2023b) can generate smaller draft models with favorable configurations.

Our Contributions:

- To the best of our knowledge, we are the first work to conduct comprehensive experiments on serving the open source LLaMA-65B and OPT-66B models utilizing speculative decoding, conducting more than 352 experiments to elucidate the factors one needs to consider while selecting and designing a draft model.

- We show the need for a systematic redesign of draft models used for speculative decoding. We demonstrate that using accuracy on language modeling tasks to choose the draft model for speculative decoding can lead to suboptimal choices, and our experiments highlight that redesigning draft models can improve the throughput of speculative decoding by up to 60%. Based on these insights, our pruned LLaMA-796M provides up to 60% higher throughput than Sheared-LLaMA-1.3B while using only 0.8% of tokens (0.42B vs 50.42B) used to train Sheared-LLaMA-1.3B. We also show that LLaMA-796M works well for other LLMs, such as LLaMA-2 families of models and supervised fine-tuned models (Vicuna-33B).

- Finally, we also study how improvements in models and hardware can further impact draft model design for future generations of LLMs. (Section 5.1).

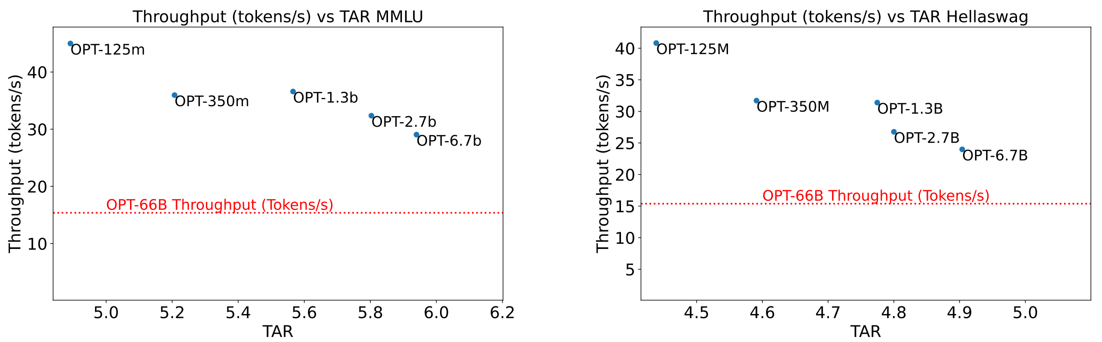
2 Background and Related Work

First, we provide a high-level overview of LLM inference and the use of speculative decoding.

2.1 Background

A decoder-only LLM performs inference in two phases: a prefill phase and an autoregressive-decoding phase. In the prefill phase, the LLM is initialized with a context or prompt, formulated as $C = \{c_1, c_2, ..., c_n\}$, where C represents the input context and n the length of the prefill. In the prefill phase, the model processes the whole input context in parallel and performs next-word prediction. During the autoregressive-decoding phase, the model generates new text sequentially, one token at a time, building upon the context provided in the prefill phase. Due to its sequential nature, the autoregressive decoding phase is widely known to be memory bandwidth bound on modern GPUs (Leviathan et al., 2023).

To improve hardware utilization and throughput, Schuster et al. (2022); Chen et al. (2023) proposed *speculative decoding*, where a significantly smaller *draft* model generates multiple tokens, and the *target* LLM performs verification on the generated tokens in parallel. The verification is akin to *prefill* stage in LLM inference. As long as more than



(a) OPT Series on MMLU

(b) OPT Series on Hellaswag

Figure 2: This figure shows the throughput of different draft models from the OPT series. As model size increases, throughput decreases due to higher inference latency despite consistent increases in TAR.

one token is accepted on average, speculative decoding can potentially provide speedups. Figure 1 shows how inference using speculative decoding differs from auto-regressive decoding. It is widely reported (Miao et al., 2023; Liu et al., 2023a) that the number of tokens accepted by the target model influences the speedup provided by speculative decoding.

In this work, we conduct a comprehensive empirical study to identify the performance bottleneck of speculative decoding and identify strategies to design the best draft model for a given LLM.

2.2 Related Work

LLM Inference There has been significant amount of work on improving LLM serving including work in Orca (Yu et al., 2022), LightSeq (Wang et al., 2020), DeepSpeed Inference (Aminabadi et al., 2022), PagedAttention (Kwon et al., 2023), FlashDecoding (Dao et al., 2023) and FlashDecoding++ (Hong et al., 2023). These works seek to improve LLM inference by better utilization of hardware. There are lines of work that have looked at pruning LLMs based on input context to speed up inference (Liu et al., 2023b) or using shallower and wider neural networks for machine translation (Kasai et al., 2020). However, in this work, we focus on speculative decoding (Leviathan et al., 2023; Chen et al., 2023; Santilli et al., 2023), which has been inspired by speculative execution in hardware (Hennessy and Patterson, 2011).

Speculative Decoding Several prior works have studied ways to improve speculative decoding. Liu et al. (2023a) seeks to continuously train the draft model on the output of the target model to improve the token acceptance rate. However, training on the same hardware during inference can be challenging, depending on the inference request rate and hardware utilization. Predictive Pipeline Decoding (PPD) (Yang et al., 2023) was one of the first methods to introduce the use of early exit (Schuster et al., 2022) from the target model to obtain draft tokens. Similar to PPD, Draft&Verify (Zhang et al., 2023) seeks to combine the use of early exit with speculative decoding, where the early exit (Schuster et al., 2022; Bae et al., 2023) from the target model acts as a draft token. A drawback of these methods is that the maximum benefit in latency is capped. For example, in speculative decoding, we can use draft models that are orders of magnitude (e.g., ≈100x-1000x) smaller than the target model, while early exit methods usually exit after performing inference over at least a fourth of the model (Schuster et al., 2022), thus, limiting the gain in throughput. Other lines of work, such as Medusa (Cai et al., 2024), propose fine-tuning multiple generation heads within the LLM that do not match the LLM output distribution exactly but maintain the generation quality.

In this work, we aim to understand how the choice of draft model affects the throughput provided by speculative decoding. We use insights from benchmarking to design draft models that maximize speculative decoding throughput.

3 Understanding Speculative Decoding

To study the effects of the choice of the draft model, we first perform a detailed study on serving OPT-65B and LLaMA-65B (two popular LLMs) using speculative decoding.

Setup. We implement speculative decoding in the Microsoft Deepspeed library (Microsoft, 2023). We use the same setup as SpecInfer (Miao et al., 2023), first using the draft model to generate draft

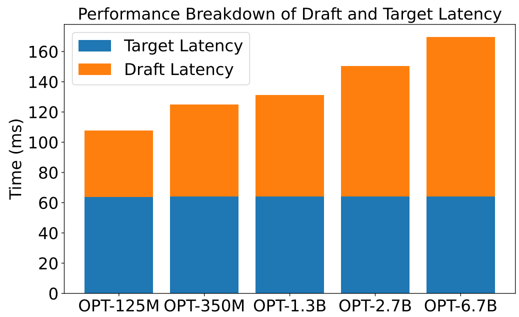
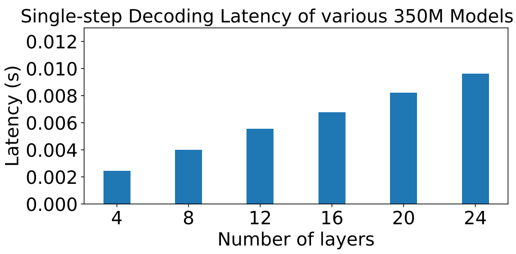
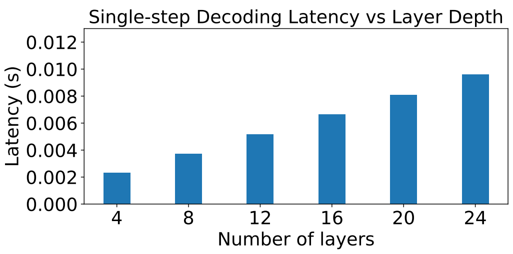


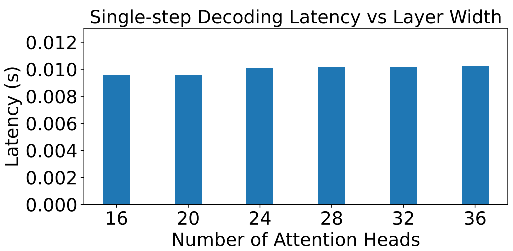
Figure 3: Performance Breakdown of OPT speculative decoding, lookahead length is set to be optimal for each draft model found empirically.



(a) In this figure, we fix model parameters to 350M and vary the number of layers and attention heads. As the number of layers decreases from 24 to 4, the number of attention heads increases from 16 to 56 (Table 6 in the Appendix).



(b) In this figure, we fix layer width and increase the number of layers. The number of parameters in the model increases from 79M to 350M.



(c) In this figure, we fix model depth and increase the number of attention heads in each layer. The number of parameters in the model increases from 350M to 1B.

Figure 4: This figure shows microbenchmarks on how model depth and width affect decoding latency.

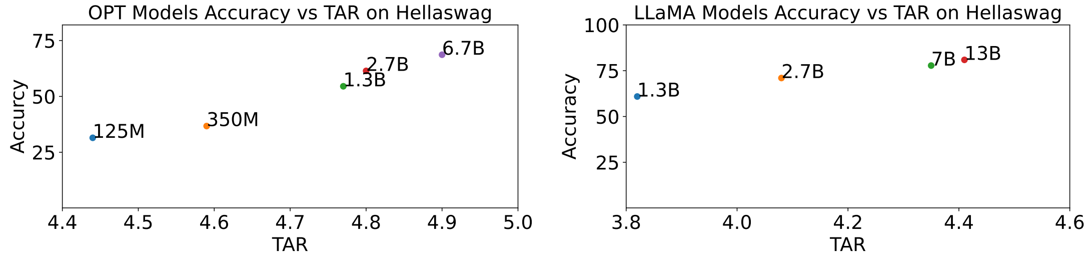
tokens and then using the target model to verify the output of the draft model. We set the batchsize to 1 and use greedy decoding. For all our experiments, we use 4 Nvidia 80GB A100 GPUs. We perform our experiment on the OPT and LLaMA base models (Zhang et al., 2022; Touvron et al., 2023a) on MMLU (Hendrycks et al., 2020), Hellaswag (Zellers et al., 2019), and Chatbot Arena datasets (Zheng et al., 2023). For MMLU, we use the standard 5-shot setup. The remaining datasets were evaluated in a zero-shot setting. We use OPT-66B and LLaMA-65B as the target LLM for OPT and LLaMA series and use OPT-125M, OPT-350M, OPT-1.3B, OPT-2.7B, and OPT-6.7B as draft models for OPT series, and LLaMA-7B, and LLaMA-13B as draft models for LLaMA series.

Metrics. To quantify the performance of different draft models when performing inference on a target model, we measure throughput (tokens generated per second) and TAR (Figure 2). We note that the primary goal of speculative decoding is to improve throughput.

3.1 Bottlenecks in Speculative Decoding

To understand the throughput of LLMs, we first plot a latency breakdown of speculative decoding in Figure 3. We show the latency breakdown between the draft token generation phase and the target model verification phase for serving OPT-66B model when using various variants of OPT as the draft model. A similar figure for LLaMA models (Figure 9) can be found in the Appendix.

In Figure 3, the time taken by the draft model for token generation increases with an increase in model sizes, going from 6.23 ms for OPT-125M to 18.56 ms for OPT-6.7B. However, even the smallest draft model, OPT-125M, still takes significant time in a speculative decoding iteration to perform draft model autoregressive decoding. Though the target LLM has a higher latency in each decoding iteration, it only has to perform one prefill operation on the entire candidate token sequence. In contrast, the draft model has to perform multi-step autoregressive decoding sequentially, creating a bottleneck. This highlights why draft model latency is one of the key bottlenecks in speculative decoding performance. We note that while Figure 3 uses lookahead values from 6 to 8, depending on the draft model, even if we scale lookahead length to hundreds of tokens, the target model verification time stays constant. The draft model latency remains the bottleneck due to the difference in efficiency between prefill and auto-regressive decoding. Next, we investigate how we can reduce draft model latency.



(a) Model accuracy vs TAR for OPT models

(b) Model accuracy vs TAR for LLaMA models

Figure 5: This figure shows the task accuracy versus TAR for OPT and LLaMA models on Hellaswag. The accuracy numbers are obtained from OpenLLM Leaderboard (HuggingFace, 2023).

3.2 Understanding Draft Model Latency

When studying the breakdown in latencies for speculative decoding in the previous section, we observed something intriguing in Figure 3. We see that OPT-350M has a similar draft-model latency as OPT-1.3B, a model almost four times its size. This indicates that OPT-350M is inefficient, and we can design better models.

We perform three microbenchmarks to validate our hypothesis and analyze decoding throughput: First, we fix the total model parameters at 350M and see how changing layer width and depth would affect decoding latency. Then, we fix either the layer width or depth to be the same as in OPT-350M and modify the other to see how latency scales with wider layers or shallower models.

Figure 4 shows the results of these three benchmarks. In the first benchmark (Figure 4a), we vary the number of attention heads, feed-forward dimension, and layers in a model to keep the model parameters at around 350M. The detailed configuration for each model can be found in Table 6 in the Appendix. The plot shows that the autoregressive decoding latency is linear in terms of layer depth despite each model having roughly the same parameter count.

The same is true for the second benchmark (Figure 4b). The original OPT-350M model has 24 layers. As we reduce the number of layers while keeping all other configurations the same, the autoregressive decoding latency decreases linearly. On the other hand, the third benchmark (Figure 4c) shows that as we scale the number of attention heads up from the original OPT-350's 16 heads to 36 heads, the decoding latency stays almost constant even if layer width has doubled.

These experiments indicate more latency-efficient model architectures with the same parameter budget exist. Changing the number of layers and attention heads not only changes the throughput but also affects the quality of predictions made by the model. We will next study how changes in model depth and width affect model accuracy and TAR and the correlation between them.

3.3 Understanding Draft Model TAR

In prior work (Leviathan et al., 2023), speculative decoding throughput is modeled by $\frac{1-\alpha^{\gamma+1}}{(1-\alpha)(\gamma c+1)}$, where $\frac{1-\alpha^{\gamma+1}}{1-\alpha}$ represents the improvement factor (expected number of tokens matched in each iteration) and $\gamma c+1$ represents the combined latency of draft and target models. Therefore, tokens matched per iteration (also known as TAR) have a linear effect on speculative decoding throughput.

In this section, we perform experiments to understand the correlation between the accuracy of a model on popular NLP tasks and its TAR. We plot the accuracy of a model against the TAR it achieves in Figure 5. Surprisingly, we find that TAR has little correlation to the model's accuracy on a task. We believe this lack of correlation is due to the majority of tokens in a sentence not being content words, which do not affect the accuracy of the model on a specific task.

For example, if a user asks a model: *What is the capital of Uruguay?* An LLM may correctly answer: *The capital of Uruguay is Montevideo.* But a draft model, without retaining this much knowledge, may respond incorrectly: *The capital of Uruguay is Paris.* For model accuracy evaluation, this would be a failure. However, this would be a good set of candidate tokens in speculative decoding, as the first five words are generated correctly. Therefore, as shown in Figure 5, TAR increases sub-linearly with an increase in model size, irrespective of its accuracy on the task. Results on more datasets can be found in the Appendix (Figure 10).

Combining insights from these experiments, we observe that current draft models are not designed to maximize speculative decoding throughput. Next, we will show how to design new draft models that outperform existing models.

4 Draft Model Design for Speculative Decoding

The above results indicate that to improve the throughput of speculative decoding, it is necessary to improve the latency of draft models, *i.e.*, can we design a model that provides a similar TAR at a lower inference cost? In the next section, we study the possibility of such a design based on the above insights.

4.1 Draft Model Design

In section 3.1, we show that model depth bottlenecks draft model latency, while in section 3.3, we show that a draft model's performance in speculative decoding is largely irrelevant to its accuracy on language modeling. These two insights prompted us to test if we can build a wider and shallower network and study how it affects latency and TAR.

Method: We leverage recent advances in structured LLM pruning, Sheared-LLaMA (Xia et al., 2023b), which provides a framework to prune larger models to a specified smaller configuration. Sheared-LLaMA (Xia et al., 2023b) learns layers, attention heads, and neurons to mask from the large model to prune it into the specified small model. The flexibility enables us to prune LLaMA-7B into desirable model configurations. In our experiments, we prune our models from LLaMA-7B using 0.4B tokens sampled from the RedPajama Dataset (Computer, 2023) (the same as in Sheared-LLaMA (Xia et al., 2023b)), but skipped the expensive fine-tuning step on 50B more tokens (and hence the name NoFT). We find that this is sufficient to achieve a significantly higher throughput.

Deep vs Wide Model Comparison: Our goal is to start with LLaMA-7B and produce a wider version of Sheared-LLaMA-1.3B while keeping the number of parameters the same as in Sheared-LLaMA-1.3B. We choose Sheared-LLaMA-1.3B since it achieves the highest throughput in our benchmark among existing models (Blue dots in Figure 6). We use two configurations: the first configuration was provided by the Sheared-LLaMA authors (NoFT-1.3B), and we designed the second configuration (NoFT-Wide-1.3B) to optimize for better speculative decoding throughput. Table 1 shows the detailed configuration of the two models. We slash the number of layers by half, from 24 to 12, and keep the total parameter count roughly the same by increasing the intermediate size from 5504 to 9280, the number of attention heads from 16 to 20, and the corresponding model dimension from 2048 to 2560.

Table 1: This table shows the model configuration of the two pruned models. Here l represents the number of layers, h represents the number of attention heads, d_{inter} represents intermediate size, and d_{model} represents model dimension.

Model	l	h	d_{inter}	d_{model}
NoFT-1.3B	24	16	5504	2048
NoFT-Wide-1.3B	12	20	9280	2560
NoFT-Wide-796M	5	32	11008	4096
NoFT-Wide-543M	3	32	11008	4096
NoFT-Wide-290M	1	32	11008	4096

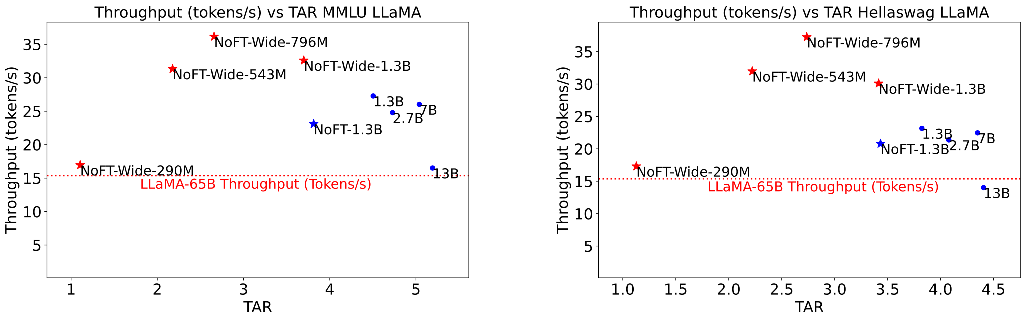
Figure 6a, 6b, and 8b show that we can achieve up to 30% higher speculative decoding throughput using only 0.8% of tokens used to train Sheared-LLaMA-1.3B.

Table 2 also shows the latency and TAR of the two sheared models on MMLU. The deep variant (NoFT-1.3B) can achieve 3% higher TAR, but the wide variant (NoFT-Wide-1.3B) reduces draft latency by 49%, improving overall throughput by 41%. We found results are very similar for other datasets, such as Chatbot Arena (Figure 8b in the Appendix) and Hellaswag (Figure 6b). This experiment shows a need to rethink the model design space for speculative decoding, where we should specifically design models for higher throughput.

Draft model scaling: To understand the limitation of draft model depth-width tradeoff in speculative decoding, we created three configurations, NoFT-Wide-796M, 543M, and 290M, that use the same number of attention heads, intermediate size, and model dimension as LLaMA-7B, but reduce the number of layers to 5, 3, and 1, respectively. This is the widest configuration possible using the Sheared-LLaMA pruning scheme.

Figure 6 shows that the NoFT-Wide-796M model provides another 20% improvement in throughput over NoFT-Wide-1.3B and demonstrates up to 60% throughput improvement over the existing Sheared-LLaMA-1.3B model. Though the smaller NoFT-Wide-543M provides up to 40% throughput improvements over Sheared-LLaMA-1.3B, it has a lower throughput than NoFT-Wide-796M.

Results in figure 6 show that reducing the layer count to less than 5 layers would cause the model's alignment capability to reduce dramatically. In addition, as we reduce models to 5 layers, target model latency takes more than 80% of the time



(a) LLAMA Series on MMLU

(b) LLAMA Series on Hellaswag

Figure 6: This figure shows the throughput scaling of different draft models from the LLAMA series on MMLU and Hellaswag. Asterisks represent models that are pruned but not fine-tuned. The red asterisks represent model configurations that we designed.

Table 2: This table shows the speculative decoding throughput and the latencies to generate 8 tokens using the two pruned draft models.

Draft Model	TAR	Latency (ms)	Tput (tokens/s)
NoFT-1.3B	**3.81**	105.1	23.10
NoFT-Wide-1.3B	3.70	**53.5**	**32.59**

in a decoding cycle. Therefore, further reducing the latency would only provide a marginal gain in overall decoding latency since the target model latency remains constant. In this case, the drop in TAR significantly outweighs the latency gain, causing decoding throughput to decrease.

4.2 Ablation Studies

In this section, we study if using a different or supervised fine-tuned target model would affect our draft model's performance.

Varying the target model: Prior experiments are performed with LLAMA-65B as the target model. As newer generations of models roll out, we would like to see if our conclusion holds on newer generations of models. In this ablation study, we evaluate our best NoFT-Wide-796M model against LLAMA-2-70B model. Table 3 shows that though our NoFT-Wide-796M is distilled from LLAMA-7B, it can achieve a similar token acceptance rate when the target model is from LLAMA-2 family. We believe the similarity in performance is due to similar training datasets. This shows that our distilled model can be applied to future families of models based on similar training recipes with little to no changes.

Supervised fine-tuned models: Prior experiments are performed on base models to study the scaling of draft models. In practice, supervised fine-tuned models are adopted for their better instruction-following capabilities. In this section,

Table 3: This table shows the tokens accepted per iteration when we use different target models. The draft model we use is NoFT-Wide-796M.

Target Model	MMLU	Hellaswag	Chatbot Arena
LLAMA-65B	2.66	2.74	2.61
LLAMA-2 70B	2.55	2.68	2.64

Table 4: This table shows the throughput of speculative decoding (tokens/s) with Vicuna 33B as the target model.

Draft Model	MMLU	Hellaswag	Chatbot Arena
Tiny-LLAMA-1.1B	20.78	18.25	18.73
NoFT-Wide-796M	**29.87**	**26.55**	**25.61**

we compare our best NoFT-Wide-796M model to Tiny-LLAMA-1.1B with Vicuna 33B as the target model. Note that our NoFT-Wide-796M is pruned from the base version of LLAMA-7B without fine-tuning. Table 4 shows that NoFT-Wide-796M outperforms Tiny-LLAMA-1.1B in all cases by up to 45%. While Tiny-LLAMA-1.1B has a TAR 35% and 32% higher than NoFT-Wide-796M on MMLU and Hellaswag, respectively, its latency is 4x higher due to having 22 layers in the model compared to NoFT-Wide-796M with merely 5 layers. This ablation study also demonstrates how speculative decoding is bottlenecked by draft model depth and that a draft model obtained from the non-fine-tuned base model, when appropriately designed, can still provide significant speedup over draft models fine-tuned for chatbot purposes.

5 Discussion

Next, we discuss how our insights can change if the models or the underlying hardware change.

5.1 Future Draft Model Design

To study how compute and performance changes can lead to different choices of draft models, we use a performance model. The original speculative de-

Table 5: This table shows the latency reduction needed for larger draft models to achieve parity throughput with OPT-125M on MMLU.

Model	Latency (ms)	Parity Latency	Reduction (%)
125M	43.7	43.7	0
350M	79.8	50.6	36.6
1.3B	87.1	58.7	32.6
2.7B	114.3	49.8	56.4
6.7B	139.5	68.2	51.1

coding (Leviathan et al., 2023) model $\frac{1-\alpha^{\gamma+1}}{(1-\alpha)(\gamma c+1)}$ can be simplified to the following to remove the unnecessary assumption of mutual independence between generated tokens in a sequence:

$$Tput = \begin{cases} \dfrac{TAR}{(t^d_{target} + t^d_{draft})} & \text{if } TAR > 1, \\ \dfrac{1}{(t^d_{target} + t^d_{draft})} & \text{if } TAR \leq 1. \end{cases}$$

In the Appendix (Figure 12), we show that this simplified formula almost perfectly captures the real speculative decoding throughput. Here, t^d represents the latency to generate d tokens autoregressively. In this section, with the aid of the performance model, we provide quantitative answers to several questions: First, we study the improvement in TAR a larger draft model needs to be provided to compensate for the additional inference cost. Next, we study how much improvement in latency is required to change the choice of the draft model.

Improvement in TAR needed to switch to a larger draft model model. In Figure 2, we observed that with existing datasets and models, we are better off with the smallest model as the draft model, *e.g.*, OPT-125M, than choosing a larger model. However, there is a possibility that the TAR difference will become greater for new datasets. In Figure 7, we plot the improvement in TAR (extra TAR), which larger models in the OPT model family should provide to match the throughput of the smallest model (OPT-125M) for MMLU. We find that if a 1.3B model can achieve a TAR advantage greater than 2 over OPT-125M for a new workload, we would choose the 1.3B model instead. Furthermore, given that the maximum TAR is capped at 8 in our scenario due to the length of draft token generation, it becomes unfeasible for OPT-2.7B and OPT-6.7B to surpass OPT-125M in performance. This is because the improvement needed in TAR for OPT-6.7B to match the throughput of OPT-125M would exceed this maximum limit.

Improvement in latency for switching to higher TAR model. As hardware evolves, latency scaling patterns may change with more computing

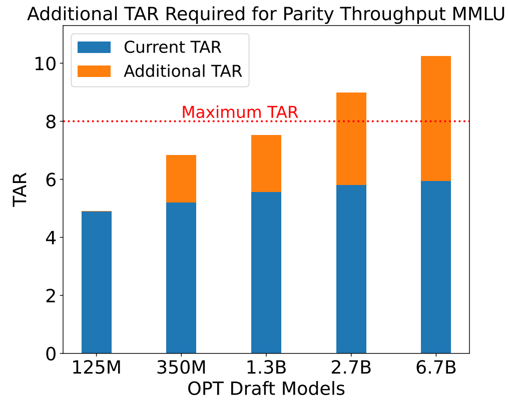


Figure 7: This figure shows the extra TAR needed for each model to achieve parity throughout with OPT-125M on MMLU.

power and memory bandwidth. Therefore, conclusions drawn on specific hardware (*e.g.*, A100) may not hold for newer or older hardware (*e.g.*, H100 or V100). To account for changing hardware, we study how much draft model latency improvement is needed to achieve throughput parity. To demonstrate this, we first compute the latency reduction needed for different members in OPT family to reach the same throughput as the smallest draft model in Table 5. We find that up to 56% of latency reduction is needed to achieve the same throughput. For instance, for OPT-1.3B to achieve parity throughput with OPT-125M, its latency needs to be reduced by 32.9%. This reinforces our finding that latency reduction provided by the smaller models has significantly more benefit than the extra TAR provided by a larger draft model.

6 Conclusion

In this work, we conduct a large-scale experimental study to understand how we can optimize the throughput of speculative decoding. Using our experiments, we outline the various factors that affect speculative decoding throughput. We observe that draft model accuracy on language modeling does not correlate strongly with its performance in speculative decoding. Further, we find that draft model latency is bottlenecked by model depth, and higher model depth increases latency. Based on these two insights, we propose new draft models pruned to align with the target model while trading model depth for width. Our proposed draft model can increase throughput by up to 60% over existing models. We find that the pruned models can be used for supervised fine-tuned target models without modification and discuss how future models may impact draft model selection.

7 Limitations

Our work aims to improve the inference efficiency of LLMs by designing better draft models for speculative decoding. Since speculative decoding preserves the output from the LLM, our work will not amplify existing biases in LLMs. However, limiting and reducing such biases are out of the scope of this work. Furthermore, since we are making LLM generation more efficient, we believe our work will not have a significant negative environmental impact.

References

Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. 2022. Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15. IEEE.

Sangmin Bae, Jongwoo Ko, Hwanjun Song, and Se-Young Yun. 2023. Fast and robust early-exiting framework for autoregressive language models with synchronized parallel decoding. *arXiv preprint arXiv:2310.05424*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Tri Dao, Daniel Haziza, Francisco Massa, and Grigory Sizov. 2023. Flashdecoding. `https://pytorch.org/blog/flash-decoding/`. Accessed: January 26, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

John L Hennessy and David A Patterson. 2011. *Computer architecture: a quantitative approach*. Elsevier.

Ke Hong, Guohao Dai, Jiaming Xu, Qiuli Mao, Xiuhong Li, Jun Liu, Kangdi Chen, Hanyu Dong, and Yu Wang. 2023. Flashdecoding++: Faster large language model inference on gpus. *arXiv preprint arXiv:2311.01282*.

HuggingFace. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`. Accessed: January 26, 2024.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A Smith. 2020. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. *arXiv preprint arXiv:2006.10369*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

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

Xiaoxuan Liu, Lanxiang Hu, Peter Bailis, Ion Stoica, Zhijie Deng, Alvin Cheung, and Hao Zhang. 2023a. Online speculative decoding. *arXiv preprint arXiv:2310.07177*.

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, et al. 2023b. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pages 22137–22176. PMLR.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Rae Ying Yee Wong, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. 2023. Specinfer: Accelerating generative llm serving with speculative inference and token tree verification. *arXiv preprint arXiv:2305.09781*.

Microsoft. 2023. Deepspeed. `https://github.com/microsoft/deepspeed`. Accessed: January 26, 2024.

Andrea Santilli, Silvio Severino, Emilian Postolache, Valentino Maiorca, Michele Mancusi, Riccardo Marin, and Emanuele Rodolà. 2023. Accelerating transformer inference for translation via parallel decoding. *arXiv preprint arXiv:2305.10427*.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. 2018. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiaohui Wang, Ying Xiong, Yang Wei, Mingxuan Wang, and Lei Li. 2020. Lightseq: A high performance inference library for transformers. *arXiv preprint arXiv:2010.13887*.

Heming Xia, Tao Ge, Peiyi Wang, Si-Qing Chen, Furu Wei, and Zhifang Sui. 2023a. Speculative decoding: Exploiting speculative execution for accelerating seq2seq generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3909–3925.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2023b. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*.

Seongjun Yang, Gibbeum Lee, Jaewoong Cho, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Predictive pipelined decoding: A compute-latency trade-off for exact llm decoding. *arXiv preprint arXiv:2307.05908*.

Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. 2022. Orca: A distributed serving system for {Transformer-Based} generative models. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 521–538.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2023. Draft & verify: Lossless large language model acceleration via self-speculative decoding. *arXiv preprint arXiv:2309.08168*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A More Experiment Results

In this section, we show more experimental analysis of speculative decoding. In Figure 8, we plot the throughput of OPT and LLaMA models against its TAR on Chatbot Arena. This figure shows that as model size increases, throughput generally decreases due to significantly higher inference latency despite consistent increases in TAR.

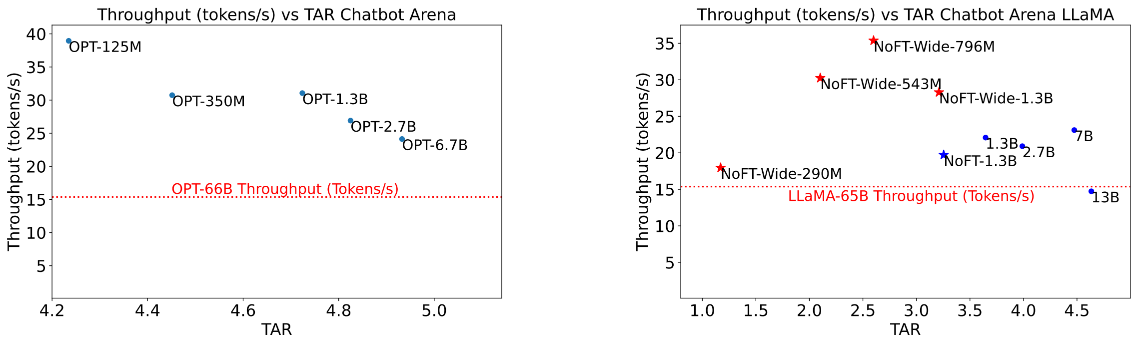
In Figure 9, we plot the throughput of OPT and LLaMA models against its TAR on Chatbot Arena. This figure shows that draft latency occupies a large chunk of time in a speculative decoding iteration, opening up new avenues for designing draft models optimal for speculative decoding.

In Figure 10, we plot the task accuracy versus TAR for OPT and LLaMA models on MMLU. The accuracy numbers are obtained from Open-LLM Leaderboard (HuggingFace, 2023). This figure shows that task accuracy is irrelevant to TAR.

Required TAR to match throughput. We can also use our analytical model to predict the TAR necessary for different models to achieve a target throughput. This can be useful in scenarios where developers deploy speculative decoding-based LLMs and must meet a throughput goal. In Figure 11, we plot the TAR needed by existing models to achieve a specific throughput. The figure shows that the TAR gap between draft models at each given throughput is much larger than we observed in Figure 2. When the throughput requirement is high, a large draft model, such as OPT-6.7B, can't achieve the desired throughput. This will allow model designers to quickly judge which draft and target model pair allows them to meet throughput requirements.

B OPT-350M Configurations

Table 6 shows the detailed model configurations of the OPT-350M variants we created. The goal is



(a) Throughput scaling with increasing TAR in OPT series on Chatbot Arena

(b) Throughput scaling with increasing TAR in LLaMA series on Chatbot Arena

Figure 8: This figure shows the throughput scaling against TAR for Chatbot Arena.

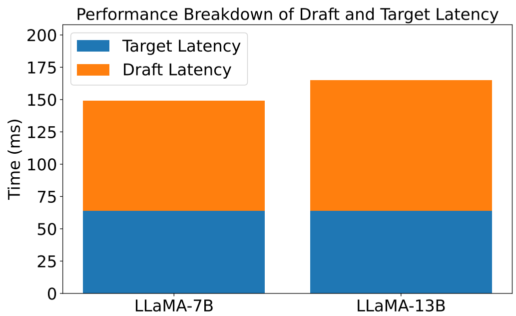


Figure 9: Performance Breakdown of LLaMA speculative decoding, lookahead length is set to be optimal lookahead length found empirically.

Table 6: This table shows model configuration of various OPT-350M models we created. The goal is to explore the tradeoff between model depth and width while keeping the total parameter count constant.

Num Layers	Attn. Heads	Hidden size	FFN Dim
24	16	1024	4096
20	20	1280	3448
16	22	1408	4096
12	28	1792	3448
8	36	2304	3448
4	56	3584	3448

to keep the total parameter count close to that of OPT-350M while adjusting model width and depth.

C Simplifying Analytical model

The original speculative decoding paper (Leviathan et al., 2023) proposed an analytical model $\frac{1-\alpha^{\gamma+1}}{(1-\alpha)(\gamma c+1)}$ to describe the speedup achieved by speculative decoding, where α denotes the expected token acceptance rate (in percentage) and γ denotes the lookahead length. However, this model is inaccurate since it assumes that the tokens generated in a sentence are mutually independent.

Table 7: This table shows the latency of each autoregressive generation step of the draft model.

Model	Latency (ms)
OPT-125M	6.23
OPT-350M	11.74
OPT-1.3B	12.64
OPT-2.7B	16.35
OPT-6.7B	18.56

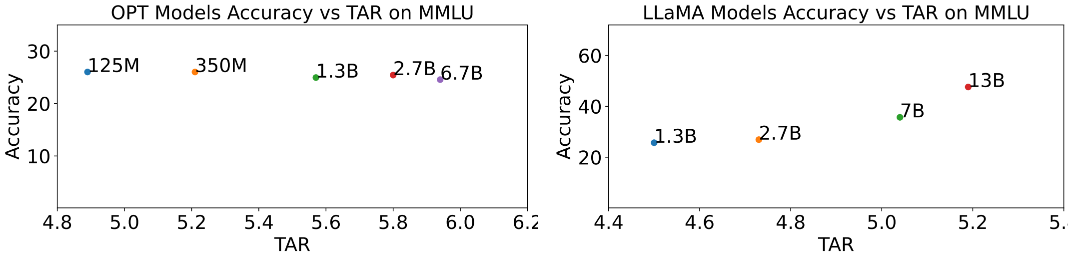
We simplify this cost model and use our updated analytical model in our experiments.

Assuming a setup similar to prior work (Leviathan et al., 2023; Chen et al., 2023; Miao et al., 2023) where speculative execution of the draft model and target model verification phases happen sequentially, the performance of speculative decoding can be decomposed into the following factors,

$$Tput = \begin{cases} \dfrac{TAR}{(t^d_{target} + t^d_{draft})} & \text{if } TAR > 1, \\ \dfrac{1}{(t^d_{target} + t^d_{draft})} & \text{if } TAR \leq 1. \end{cases}$$

Considering a case where, in each iteration, d tokens are generated by the draft model, t^d_{draft} depicts the time draft models take to generate d draft tokens, while t^d_{target} is the time taken by the target model for verifying those d draft tokens. TAR is used to denote the average number of tokens that were matched across a query or a dataset.

Verifying Analytical Model In Figure 12, we compare the throughput predicted by our model with throughput measured on real hardware for two model families: LLaMA (7B and 13B) and



(a) Model accuracy vs. TAR for OPT models on MMLU and Hellaswag

(b) Model accuracy vs TAR for LLAMA models on MMLU and Hellaswag

Figure 10: This figure shows the task accuracy versus TAR for OPT and LLAMA models on MMLU. The accuracy numbers are obtained from OpenLLM Leaderboard (HuggingFace, 2023).

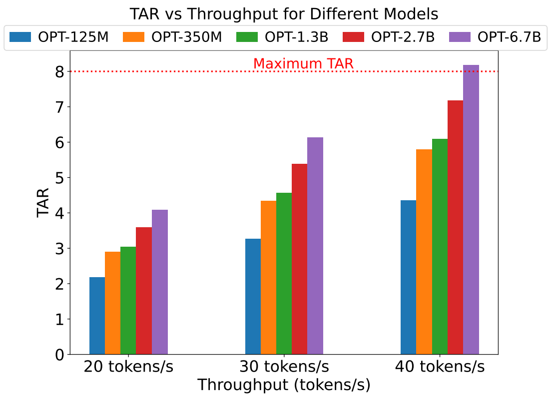


Figure 11: This figure shows the required TAR to achieve a given throughput.

OPT (125M, 350M, 1.3B, 2.7B, and 6.7B) to serve LLAMA-65B and OPT-66B on MMLU.

We run these experiments on 4 Nvidia 80GB A100 GPUs for 100 iterations on the real server, and the error bars in Figure 12 represent the standard deviation of the measurement. For the performance model, we collect t^d_{draft} and t^d_{target} on a real cluster with a single iteration. For TAR, we collect the average token acceptance rate from the MMLU dataset. The maximum deviation we observed between our proposed analytical model and the results obtained is 3.5%. The close correspondence between our performance model and real measurements shows that our performance model accurately predicts the throughput of speculative decoding.

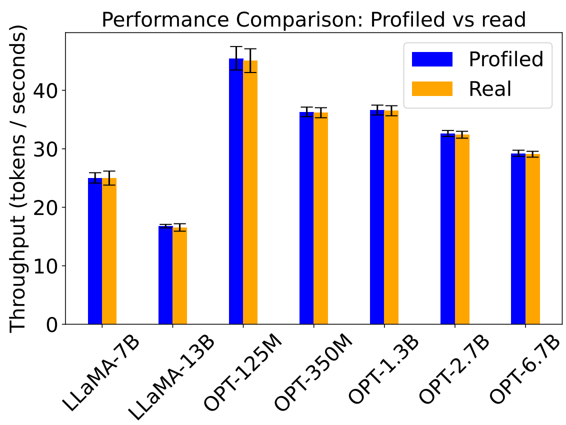


Figure 12: This figure shows that our performance model correctly captures the real performance of speculative decoding. We use LLAMA-65B and OPT-66B as the target model for each model family, respectively.