

Falcon: Faster and Parallel Inference of Large Language Models Through Enhanced Semi-Autoregressive Drafting and Custom-Designed Decoding Tree

Xiangxiang Gao*, Weisheng Xie^{*†}, Yiwei Xiang, Feng Ji

Bestpay AI Lab, Shanghai, 200080 China
{gaoxiangxiang, xieweisheng, xiangyiwei, jifeng}@bestpay.com

Abstract

Striking an optimal balance between minimal drafting latency and high speculation accuracy to enhance the inference speed of Large Language Models remains a significant challenge in speculative decoding. In this paper, we introduce Falcon, an innovative semi-autoregressive speculative decoding framework fashioned to augment both the drafter’s parallelism and output quality. Falcon incorporates the Coupled Sequential Glancing Distillation technique, which fortifies inter-token dependencies within the same block, leading to increased speculation accuracy. We offer a comprehensive theoretical analysis to illuminate the underlying mechanisms. Additionally, we introduce a Custom-Designed Decoding Tree, which permits the drafter to generate multiple tokens in a single forward pass and accommodates multiple forward passes as needed, thereby boosting the number of drafted tokens and significantly improving the overall acceptance rate. Comprehensive evaluations on benchmark datasets such as MT-Bench, HumanEval, and GSM8K demonstrate Falcon’s superior acceleration capabilities. The framework achieves a lossless speedup ratio ranging from 2.91x to 3.51x when tested on the Vicuna and LLaMA2-Chat model series. These results outstrip existing speculative decoding methods for LLMs, including Eagle, Medusa, Lookahead, SPS, and PLD, while maintaining a compact drafter architecture equivalent to merely two Transformer layers. Our code is publicly available at <https://github.com/Bestpay-inc/Falcon>.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional performance across various benchmarks, reinforcing their practical significance. These models are primarily built on the Transformer architecture and utilize autoregressive (AR) decoding, effectively capturing complex dependencies and generating coherent sequences (Wan et al. 2024). However, due to their considerable model sizes, LLMs also face significant computational overhead and latency bottlenecks during inference. For example, the inference speed of the GLM 10B model is a mere 101 tokens per second when operated on a single Nvidia A100 GPU (Du et al. 2022). This

presents considerable challenges for the widespread deployment and application of LLMs (Zhu et al. 2024; Xiao et al. 2023; Ghazvininejad et al. 2019; Xu et al. 2021; Ding et al. 2021; Guo et al. 2021; Wang et al. 2019; Guo et al. 2019a).

Intriguingly, the inference speed of LLMs is considerably hampered by memory bandwidth, with the main latency bottleneck stemming from reading and writing model parameters to memory, rather than from arithmetic computations (Xia et al. 2024). To address this issue, researchers have proposed speculative decoding to expedite the inference speed of LLMs without sacrificing accuracy (Zhang et al. 2024; Guo et al. 2020). This strategy employs a drafter that efficiently generates k tokens as speculation of future decoding steps within the LLMs. Subsequently, the LLM verifies these tokens, with only successfully validated tokens being accepted as decoded tokens, thus maintaining the quality of the generated output (Xia et al. 2023). By directing computational resources towards validating pre-generated tokens, speculative decoding significantly diminishes the memory operations needed to access LLM parameters, consequently boosting overall inference efficiency (Xia et al. 2024).

While speculative decoding presents a potential solution to decrease the inference latency of LLMs, it also raises several important questions that need further exploration. Notably, existing speculative decoding methods primarily employ two drafting strategies: AR and semi-autoregressive (SAR) drafting. AR drafting generates tokens sequentially, each dependent on the preceding ones. This sequential dependency constrains the parallelism of the draft models, leading to substantial time overheads (Guo et al. 2019b; Sun et al. 2019; Geng, Feng, and Qin 2021; Graves et al. 2006). In contrast, SAR drafting generates multiple tokens simultaneously, enhancing the parallelization of the drafting process. However, a significant limitation of SAR drafting is its inability to fully capture the inter-dependencies between tokens within the same block, potentially resulting in a lower acceptance rate for the generated tokens (Bao et al. 2022; Ran et al. 2020). Consequently, balancing low drafting latency with high speculation accuracy poses a significant challenge in speculative decoding aimed at accelerating the inference speed of LLMs (Xia et al. 2024).

In this paper, we introduce Falcon, an advanced SAR speculative decoding framework engineered to boost both the drafter’s parallelism and the output quality, thereby am-

*These authors contributed equally.

[†]Corresponding Author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

plifying the inference efficiency of LLMs. Falcon integrates the Coupled Sequential Glancing Distillation method, which elevates the token acceptance rate of SAR drafting. We offer an in-depth theoretical explanation for this accuracy enhancement, ascribing it to the strengthened dependencies between tokens within the same block. Additionally, we have developed a specialized decoding tree to support SAR drafting, enabling the drafter to generate multiple tokens in a single forward pass and also accommodate multiple forward passes. This design leads to a higher acceptance rate of tokens by the LLMs, further hastening inference speed.

Our comprehensive evaluations of Falcon on various benchmarks showcase its outstanding acceleration capabilities. It achieves a lossless speedup ratio ranging from 2.91x to 3.51x on the Vicuna and LLaMA2 model series. This performance outshines top-tier speculative decoding methods that use either AR drafting (such as Eagle, SPS, and PLD) or SAR drafting (like Medusa and Lookahead). Additionally, unlike other methods that necessitate billions of parameters for drafters, Falcon can attain faster inference speeds using parameters equivalent to just two Transformer blocks. Therefore, Falcon is highly beneficial for applications requiring real-time responses while working under computationally constrained environments. Our key contributions to the community are summarized as:

- We introduce Falcon, an advanced SAR speculative decoding framework that enhances both the parallelism and output quality of the drafter. Falcon achieves a better equilibrium between reduced inference latency and elevated speculation accuracy.
- We develop Coupled Sequential Glancing Distillation, which markedly improves the accuracy of the SAR drafter. Additionally, a theoretical explanation is provided to clarify how this method enhances output quality.
- We design a specialized decoding tree to support SAR drafting, enabling the drafter to generate multiple tokens in a single forward and allowing multiple forward passes, resulting in an improved token acceptance rate and inference speed of the LLMs.
- Falcon outperforms existing speculative decoding methodologies for LLMs, attaining a speedup ratio ranging from 2.91x to 3.51x. Remarkably, it accomplishes this while preserving a compact drafter size, which is comparable to merely two Transformer layers.

2 Related Work

2.1 Autoregressive Drafting

Speculative decoding with AR drafting is a straightforward approach that employs a small language model as the drafter to generate tokens sequentially, each conditioned on its predecessors. The generated tokens are then validated by the LLMs to ensure alignment (Leviathan, Kalman, and Matias 2023; Spector and Re 2023; Chen et al. 2024; Zhao et al. 2024). SPS (Chen et al. 2023) is the pioneering work in this field, which generates draft tokens by invoking a 4B parameter AR drafter and validating these drafts using 70B parameter LLMs. PLD (Saxena 2023) replaces the draft model with

simple string matching in the prompt to generate candidate token sequences. BiLD (Kim et al. 2023) utilizes the T5-small model to generate tokens, which are then validated by a T5-XXL model. SpecInfer (Miao et al. 2024) accelerates LLM serving through a tree-based inference and verification mechanism, organizing the drafts into a token tree. The correctness of all candidate token sequences is verified against the LLM using a tree-based decoding mechanism.

While AR drafting considerably boosts inference speed, it also imposes extra GPU memory costs, particularly for draft models with billions of parameters. Moreover, it adds time overhead from the drafter that could potentially counterbalance the advantages of the improved inference speed. EAGLE (Li et al. 2024), the current state-of-the-art speculative decoding method with AR drafting, tackles the issue of high GPU memory usage by incorporating a token sequence advanced by one time step. However, due to its inherently sequential nature, EAGLE is still constrained by time overhead, which hampers its potential to further speed up the inference of LLMs.

2.2 Semi-autoregressive Drafting

SAR speculative decoding simultaneously generates multiple tokens, maintaining the AR feature globally while easing it locally (Wang, Zhang, and Chen 2018; Łukasz Kaiser et al. 2018; van den Oord et al. 2018, 2016). Santilli et al. (Santilli et al. 2023) propose that AR decoding can be restructured as parallel resolution of a non-linear equation through Jacobi and Gauss-Seidel fixed-point iterations. This technique directly appends multiple [PAD] tokens to the end of the input prompt, facilitating parallel generation and speeding up existing models. Lookahead (Zhao et al. 2024) utilizes this generation method to enable the LLMs to produce several separate n-grams concurrently within a single forward pass, thereby reducing the latency of LLM. PASS (Monea, Joulin, and Grave 2023) introduces multiple learnable tokens and fine-tunes these token embeddings to enhance parallel decoding performance. However, these methods deviate from the AR pre-training patterns, which can result in less optimal draft quality.

Medusa (Cai et al. 2024) represents the most significant advancements in SAR drafting, building upon the research of Stern et al. (Stern, Shazeer, and Uszkoreit 2018). It optimizes the process by freezing the backbone model and incorporating additional lightweight heads into it, enabling the concurrent generation of multiple tokens. Medusa effectively alleviates the computational cost typically associated with AR drafting, thus achieving remarkable speedups. However, its inference speed is constrained by the low accuracy of the drafter, which currently stands at 0.6. This drop in accuracy results from employing the parallel processing mechanism, whose predictions are solely based on input tokens without accounting for inter-token dependencies (Xia et al. 2023; Wertheimer et al. 2024). Therefore, the key to improving the output quality of SAR drafters lies in strengthening the inter-token dependencies within the same block. Such enhancements would enable an optimal balance between low drafting latency and high speculative accuracy.

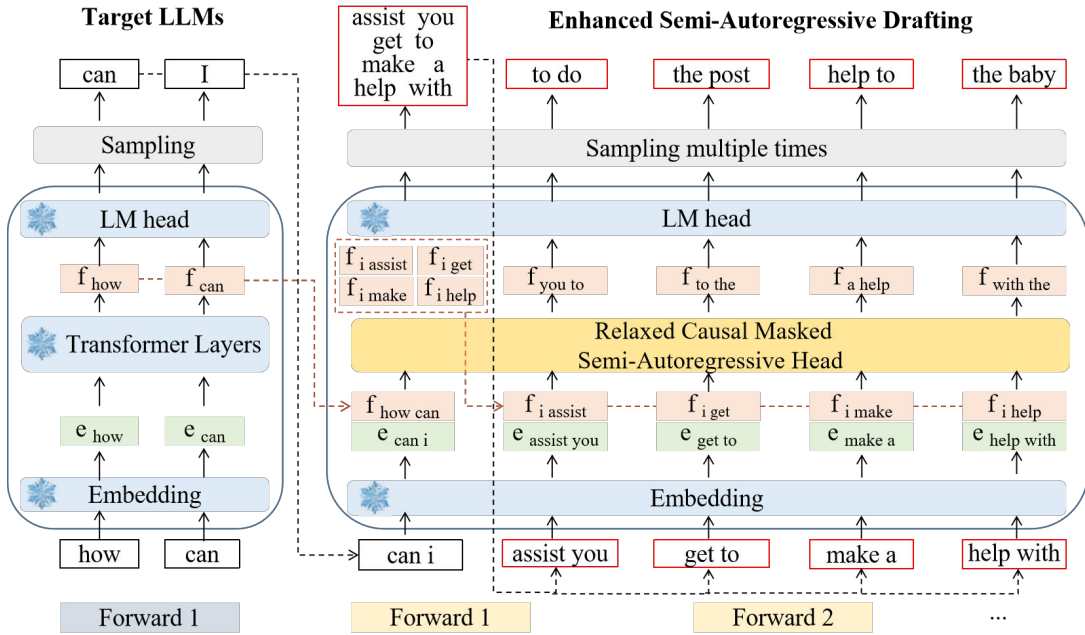


Figure 1: Framework of Falcon. It illustrates the computational process and displays the corresponding generation results of each forward pass for enhanced SAR drafting.

3 Falcon

The Falcon framework is an advanced SAR speculative decoding architecture that concurrently improves the drafter’s parallelism and token acceptance rate. A comprehensive overview and computational process of this framework are presented in Section 3.1. To improve the drafter’s accuracy, we introduce the Coupled Sequential Glancing Distillation (CSGD) method, which strengthens the inter-dependencies among tokens within each block. An in-depth description of the CSGD method is provided in Section 3.2, followed by a theoretical discussion in Section 3.3. In addition, a Custom-Designed Decoding Tree has been developed to support SAR drafting, which is described in Section 3.4.

3.1 Framework and Computational Process

The architecture and computational process of the Falcon are depicted in Figure 1. It comprises three main components: the Embedding Layer, the Language Model (LM) Head, and the Relaxed Causal SAR Decoding Head. The Embedding Layer transforms a sequence of tokens into a corresponding sequence of token embeddings. The LM Head computes the probability distribution based on the aggregated features, from which the next token is sampled. Both the Embedding Layer and the LM Head leverage the parameters of LLMs. Relaxed Causal SAR Head will be introduced in section 3.2.

EAGLE’s findings (Li et al. 2024) indicate that concatenating feature and token sequences from one time step ahead encapsulate richer semantic context. This enhancement allows the model to make more informed predictions by leveraging a broader scope of information. Consequently, we concatenate consecutive feature sequences and token se-

quences from one time step ahead to predict the next k tokens concurrently. For instance, when $k = 2$, Falcon predicts the feature sequence (f_3, f_4) using the initial feature sequence (f_1, f_2) and the token sequence (t_2, t_3) advanced by one time step. Subsequently, the predicted features (f_3, f_4) , along with the next token sequence (t_4, t_5) are concatenated to form the new input sequence. This is used to predict subsequent feature sequences (f_5, f_6) and token sequences (t_6, t_7) , facilitating the continuation of the drafting process.

3.2 Coupled Sequential Glancing Distillation

We have designed an SAR decoding method based on CSGD, aimed at enhancing the accuracy of the drafter. The training procedure is illustrated in Figure 2. Here, the feature and token sequences from one time step ahead are concatenated and fed into the drafter, resulting in a fused sequence of dimensions $(bs, seq_len, 2 * hidden_dim)$. The drafter is composed of a hybrid Transformer network, which includes two layers of LSTM (Hochreiter and Schmidhuber 1997), Relaxed Causal-Masked Multi-Head Self-Attention, and MLP networks. The LSTM network reduces the dimension of the fused sequence to $(bs, seq_len, hidden_dim)$ and retains information about past tokens, thereby improving the model’s accuracy. The Relaxed Causal-Masked Multi-Head Self-Attention mechanism enables the model to focus on relevant parts of the input sequence while preserving causality. The MLP layers further process this information to make the final predictions.

After the sequence passes through the drafter for the first time, an initial prediction of tokens, denoted as \hat{Y} is generated. We compare the Hamming Distance (Roman 1992) between the prediction from the drafter \hat{Y} and the predic-

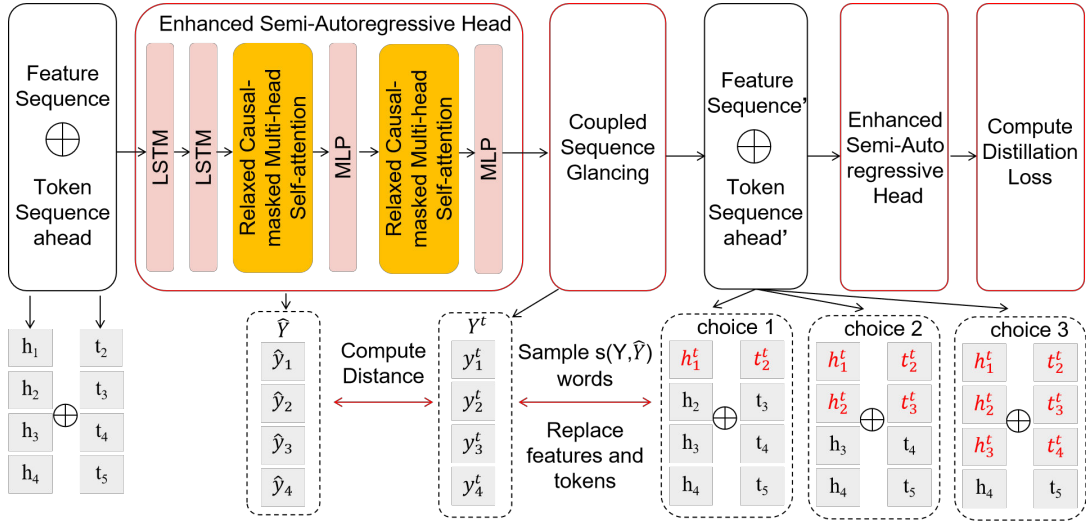


Figure 2: The training procedure of CSGD. \hat{Y} is the initial predicted feature representation sequence of the draft model, Y^t is the ground-truth feature calculated by LLMs, t_i^t is the target token generated by LLMs, h_i is the original feature sequence, and h_i^t is the target feature generated by LLMs.

tion from LLMs Y^t . Then, we replace a certain number of continuously predicted token sequence t_i and features sequence h_i with the correct ones t_i^t and h_i^t from LLMs. The number N is computed as $N = \lambda \cdot d(Y^t, \hat{Y})$, where $d(\cdot)$ is the hamming distance, $\lambda = \frac{0.4 * (ep_t - ep_c)}{ep_t}$, ep_t is the total epoch number and ep_c is the current epoch number. Note that our approach differs from the conventional glancing method (Qian et al. 2021), which only replaces tokens randomly. Instead, we concurrently replace continuous token and feature sequences preceding those to be predicted, which is illustrated in the dashed boxes labeled Choice 1, 2, 3 in Figure 2. This enhances the comprehension of inter-token relationships, as well as ensures the drafter can effectively utilize token sequences from ahead time step. This is particularly beneficial in SAR decoding, where preserving the sequence’s integrity is crucial in maintaining optimum performance. Subsequently, the revised token and feature sequences are re-input into the drafter to compute the training loss. The training loss consists of two components: the regression loss and the distillation loss. For regression loss, we utilize the Smooth L1 loss:

$$f_{i:i+k}^{draft} = Draft(t_{i:i+k}, f_{i-1:i+k-1}) \quad (1)$$

$$L_{reg} = \text{Smooth L1}(f_{i:i+k}, f_{i:i+k}^{draft}) \quad (2)$$

Correspondingly, we optimize the drafter by calculating the distillation loss between the LLMs and the drafter:

$$L_{soft} = KL_Div(p_{i:i+k}, p_{i:i+k}^{draft}) \quad (3)$$

$$L_{hard} = Cross_Entropy(t_{i:i+k}, t_{i:i+k}^{draft}) \quad (4)$$

$$L_{dist} = \alpha L_{soft} + (1 - \alpha) L_{hard} \quad (5)$$

where f and f^{draft} denote features, p and p^{draft} denote the probability distribution, t and t^{draft} represent the tokens

produced by the LLM and the drafter, respectively; and α , a constant coefficient set to 0.9. The losses L_{soft} and L_{hard} , represent the soft and hard label losses as described in (Hinton, Vinyals, and Dean 2015), are independently computed using the Kullback-Leibler divergence and Cross-Entropy, respectively.

Using the same weight matrix of the LM Head, the logits can be calculated as follows:

$$p_{i:i+k} = Softmax(LM(f_{i-1:i+k-1})) \quad (6)$$

$$p_{i:i+k}^{draft} = Softmax(LM(f_{i-1:i+k-1}^{draft})) \quad (7)$$

By integrating regression and distillation loss, we train the SAR Head with the combined loss function:

$$L = L_{reg} + \omega_{dist} L_{dist} \quad (8)$$

ω_{dist} is set to 0.1. Moreover, we employ data augmentation by adding random noise sampled from a uniform distribution $U(-0.1, 0.1)$ to avoid error accumulation in features.

3.3 Theoretical analysis of CSGD

We use a theory of information argument to illustrate the impact of CSGD. Take $k = 2$ as an example, let X represent the next token, Y the second-next token, and C the input context (Omitted from equations for simplicity). Traditional AR methods focus on $H(X)$, whereas SAR decoding with $k = 2$ targets $H(X) + H(Y)$, decomposed as (Gloeckle et al. 2024; Olsson et al. 2022):

$$H(X) = H(X|Y) + I(X; Y) \quad (9)$$

$$H(X) + H(Y) = H(Y|X) + 2I(X; Y) + H(X|Y) \quad (10)$$

In Equation (10), the left two terms represent the training loss of 2-token prediction models. The right terms decompose the loss into a local cross-entropy term for the pre-fix (C, X), denoted as $H(Y|X)$, a mutual information term

that captures the information about Y contained in X , denoted as $I(X; Y)$, and term $H(X|Y)$ describes the uncertainty about X given the prefix C and suffix Y . We can observe that SAR decoding increases the weight of $I(X; Y)$. However, conventional training methods for SAR decoding typically consider only the classical next-token entropy $H(Y|X)$, while the terms $I(X; Y)$ and $H(X|Y)$ are always overlooked. As a result, these methods do not effectively learn the dependencies between tokens within a block. However, CSGD addresses this issue. When predicting X , CSGD can leverage the features and tokens from one time step ahead in C . The feature sequences represent the prefix, and the token sequences represent the mutual information capturing the correlation between C and X . Therefore, the term $H(X|Y)$ should be changed to $H(X|C)$, and $I(X; Y)$ should be changed to $I(X; C)$. In addition, when predicting token Y , CSGD can see the information about C and X simultaneously. Therefore, the training loss of CSGD is changed to:

$$H(X) = H(X|C) + I(X; C) \quad (11)$$

$$H(X) + H(Y) = H(X|C) + I(X; C) + H(Y|X) + I(X; Y) \quad (12)$$

Equations (11) and (12) denote CSGD improves the dependency between tokens within a block, making the training loss of the SAR approach more similar to that of the AR approach. As a result, the probability distribution of multiple tokens predicted simultaneously by SAR decoding becomes more aligned with the distribution of tokens predicted sequentially by AR decoding. It further increases the probability that the tokens generated by SAR drafting will be accepted by the LLMs.

3.4 Custom-Designed Decoding Tree

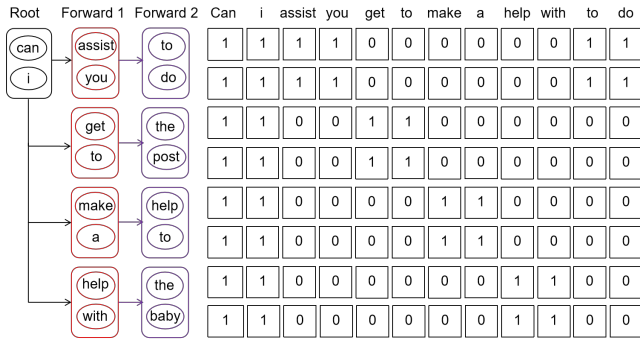


Figure 3: SAR decoding tree attention illustrated. This visualization demonstrates that SAR tree attention is utilized to process multiple candidates in parallel, and k is set to 2.

A Custom-Designed Decoding Tree has been introduced to accommodate SAR drafting. It enables the drafter to perform m forward passes and generate $n*k$ tokens for each forward pass, where n is the number of the tree nodes. These $m*n*k$ tokens are then organized in a tree structure. Meanwhile, the generated tokens and the corresponding features

are concatenated for the subsequent forward passes. In traditional AR decoding, a causal mask is used, structured as a lower triangular matrix. It ensures that the former tokens can not access the later information. In contrast, Falcon employs a Relaxed Causal Mask, which allows the model to access tokens within the same $k*k$ block and their predecessors in the tree, illustrated in Figure 3.

Subsequently, the drafts are organized as a token tree, whose nodes each represent a candidate token sequence. The correctness of the candidate token sequences are verified by the LLMs using a tree-based parallel decoding mechanism, which is consistent with SpecInfer (Miao et al. 2024). We record the accepted tokens and their corresponding features for the subsequent drafting phase.

Our custom-designed decoding tree, engineered specifically for SAR, is more efficient than those used for fully AR approaches. In SAR decoding, a tree-structure drafter generates k times as many tokens as an AR decoding drafter in a single forward pass. As a result, m forward passes can propose $k*m*n$ tokens, which is k times greater than an AR tree would generate. This enhancement significantly boosts the efficiency of token generation for drafters, allowing the LLMs to verify more tokens concurrently. Therefore, this improvement increases the likelihood of drafting tokens being accepted by the LLMs, thereby accelerating the overall inference speed of the LLMs.

4 Experiments and Analysis

4.1 Models and Tasks

We conducted experiments on Vicuna models (7B, 13B) and LLaMA2-Chat models (7B, 13B). We evaluated Falcon across multiple tasks, including multi-turn dialogue, code generation, and mathematical reasoning, employing the MT-bench (Zheng et al. 2024), HumanEval (Chen et al. 2021), and GSM8K (Cobbe et al. 2021) datasets, respectively. We conducted experiments with a batch size of 1, a temperature of 0 (greedy decoding), and adopted experimental settings consistent with other works like Eagle and Medusa.

4.2 Metrics

Like other speculative decoding methods, Falcon primarily focuses on latency rather than throughput. We assess acceleration effects using the following metrics:

- **Wall-time speedup ratio:** The actual test speedup ratio relative to vanilla AR decoding.
- **Acceptance rate (α):** The ratio of tokens generated by the drafter to all tokens gauges draft accuracy.
- **Average acceptance length (τ):** The number of accepted tokens in each draft phase.

4.3 Training

Falcon was trained on the ShareGPT dataset, utilizing 68,000 dialogue iterations. We employed the AdamW optimizer with beta values (β_1, β_2) set to (0.9, 0.95) with a learning rate set to $3e-5$. The settings are consistent with Eagle (Li et al. 2024). The semi-autoregressive head is trainable within two days on an H800 server.

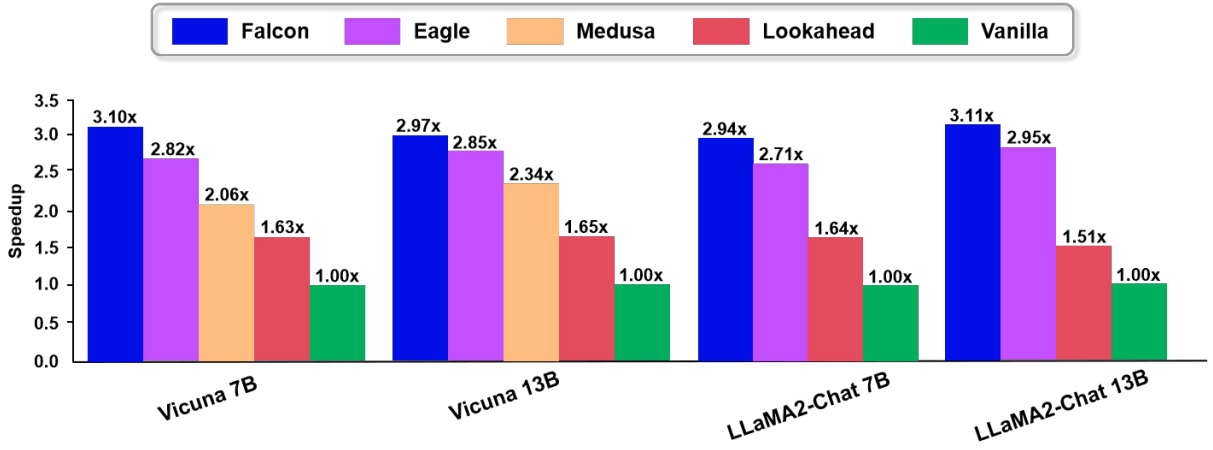


Figure 4: Speedup ratio of Vicuna and LLaMA2-Chat on MT-bench for greedy (temperature=0).

		MT-Bench	HumEval	GSM8K
Model	Method	speedup	speedup	speedup
V7B	SpS	1.82x	1.99x	1.71x
	PLD	1.61x	1.82x	1.82x
	Lookahead	1.63x	1.72x	1.84x
	Medusa	2.06x	2.41x	2.22x
	Eagle	2.82x	2.95x	2.72x
	Falcon	3.10x	3.21x	2.92x
V13B	SpS	1.93x	2.23x	1.77x
	PLD	1.58x	1.85x	1.68x
	Lookahead	1.65x	1.71x	1.81x
	Medusa	2.32x	2.44x	2.32x
	Eagle	2.85x	2.97x	2.52x
	Falcon	2.97x	3.12x	3.13x
LC7B	PLD	1.38x	1.52x	1.32x
	Lookahead	1.61x	1.72x	1.58x
	Eagle	2.71x	2.93x	2.89x
	Falcon	2.94x	2.95x	2.91x
LC13B	PLD	1.42x	1.63x	1.41x
	Lookahead	1.61x	1.72x	1.58x
	Eagle	2.95x	3.25x	2.93x
	Falcon	3.11x	3.51x	3.10x

Table 1: Comparison among Falcon and other speculative decoding methods in terms of speedup ratio.

4.4 Effectiveness

Figure 4 and Table 1 show the speedup comparison among Falcon and other state-of-the-art speculative decoding methods on MT-bench, HumanEval and GSM8K for greedy (temperature=0). Speedup ratios of Falcon, Eagle, and Medusa were fairly tested on an H800 server. The results of Speculative Sampling (SpS), Prompt lookup decoding (PLD), and Lookahead were copied from their technical reports. Falcon has achieved a speedup ratio ranging from 2.91x-3.51x. Compared to AR drafters, Lookahead and Eagle, Falcon is

		MT-Bench	HumEval	GSM8K
Model	Method	α	α	α
V7B	Medusa	60.53	63.26	61.09
	Eagle	74.59	76.18	75.47
	Falcon	77.64	79.92	78.03
V13B	Medusa	61.77	64.58	62.39
	Eagle	74.91	76.26	75.46
	Falcon	77.60	80.34	80.22
LC7B	Eagle	72.98	74.76	73.89
	Falcon	74.31	77.07	75.51
LC13B	Eagle	73.64	76.51	75.28
	Falcon	75.27	78.52	77.13

Table 2: Comparison among Falcon and other speculative decoding methods in terms of Acceptance rate (α).

faster by 1.79x-2.06x and 1.01x-1.24x, respectively. This improvement is because of Falcon’s SAR property; the draft model can generate multiple tokens each forward, which reduces the latency of each drafting phase and thus increases the total speedup ratio. Compared to Medusa, which is a SAR method as well, Falcon is faster by 1.24x-1.50x. It is attributed to the consideration of the dependency among tokens. Falcon adopts a CSGD method to maintain token dependency in the training stage. This enables the drafter to predict tokens more accurately than Medusa, effectively reducing the cost from the verification phase.

Table 2 illustrates the Acceptance rate (α) of Falcon, Medusa, and Eagle. The experiments were performed on the H800 server. Falcon outperforms Eagle by 3% to 5%, indicating that more tokens are generated by the draft model rather than the target LLM. Compared to the SAR method (Medusa), Falcon achieves a higher α by 15.76%-17.83%, which illustrates the importance of the token dependency in the draft phase.

Table 3 compares the average acceptance length (τ). AI-

though the Vicuna 7B model has a lower τ than Eagle, other sizes of models have shown a higher acceptance rate. It is demonstrated that the drafter of Falcon has a high prediction accuracy, even surpassing the AR models, and thus illustrates the effectiveness of CSGD. Compared to Medusa, the intrinsic design restricts it to conduct at most one forward each draft phase, which leads to a low τ . However, there is no such limitation to Falcon. In the experiment, Falcon performs four forward passes each drafting phase, achieving a higher τ than Medusa by 1.78-2.22.

		MT-Bench	HumEval	GSM8K
Model	Method	τ	τ	τ
V7B	Medusa	1.51	1.71	1.55
	Eagle	3.94	4.29	4.00
	Falcon	3.34	3.88	3.70
V13B	Medusa	1.59	1.81	1.64
	Eagle	2.91	3.13	2.97
	Falcon	3.37	3.97	3.86
LC7B	Eagle	2.43	2.91	2.78
	Falcon	2.82	3.30	3.02
LC13B	Eagle	2.73	3.21	2.98
	Falcon	2.95	3.60	3.47

Table 3: Comparison among Falcon, Eagle, and Medusa in terms of Average acceptance length (τ).

4.5 Ablation Study

Tree Attention Falcon, similar to Eagle and Medusa, employs tree attention in both drafting and verification phases, while methods like speculative decoding do not use it. To remove the effect of tree attention, we applied a chain to Falcon, whose length is equal to the tree height, in order to maintain the same forward times of the draft model. Table 4 shows the comparative results indicating the impact of tree attention. In Falcon, the implementation of tree attention mechanisms has been demonstrated to enhance the acceleration ratio by 1.12x, increase α by 1.22, and improve τ by 4.96%. This improvement is attributed to the fact that the tree attention increases the number of tokens validated and accepted, thereby augmenting the overall token throughput.

Coupled Sequential Glancing Distillation The SAR approach of Falcon enhances the token numbers of each forward at the cost of reduced accuracy. The CSGD training method was employed to mitigate the drop in precision. We conducted tests on Vicuna 7B under two conditions: training the SAR head with Eagle’s token-shift method (Li et al. 2024) alone and with the CSGD method. Table 4 illustrates the impact of CSGD. The implementation of CSGD resulted in a 1.17x increase in acceleration ratio, 0.56 in τ , and 3.26% improvement in α . The above results demonstrate that CSGD can significantly improve the drafter’s accuracy.

k factor The k factor determines the number of tokens the drafter generates in a single forward pass, which is important

Model	TA	CSGD	speedup	τ	α
V7B	×	×	2.06x	1.64	62.86
	×	✓	2.27x	2.13	68.68
	✓	×	2.65x	2.78	74.37
	✓	✓	3.10x	3.34	77.64

Table 4: The ablation study of Falcon of the impact of Tree Attention (TA) and CSGD on MT-Bench.

to SAR drafting. We tested conditions with $k = \{2, 3, 4\}$ on Vicuna-7B. Considering the accuracy descending of SAR, we adopted fewer forward pass times in the drafting phase for $k = \{3, 4\}$ to achieve speedup ratios as high as possible. The corresponding heights of the decoding trees are 6 and 8, respectively. The results are presented in Table ?? . With the increasing of the k factor, we see a drop in all three metrics, but the speedup ratios are still higher than Eagle (2.82x) and Medusa (2.06x). This is due to the reduced time it takes for an SAR drafter to pass forward with the same number of tokens generated by AR drafting, indicating the inherent advantages of enhanced SAR drafting.

Model	k	speedup	τ	α
V7B	2	3.10x	3.34	77.64
	3	2.96x	2.52	72.21
	4	2.86x	2.35	70.78

Table 5: The comparison of k factors on MT-Bench.

5 Conclusion and Future Work

In this paper, we propose Falcon, which utilizes the Coupled Sequential Glancing Distillation method to bolster the interdependencies among tokens, thereby enhancing the quality of the drafts. We also provide a theoretical analysis that clarifies the inner workings of our method. Additionally, we have developed a specialized decoding tree to support SAR drafting, which increases the token acceptance rate. Comprehensive evaluations indicate Falcon’s superior performance over the Vicuna and LLaMA2 model series. On the MT-bench, HumanEval, and GSM8K datasets, Falcon is 2.91x-3.51x faster than vanilla Transformer, while maintaining a compact size comparable to two Transformer layers. Falcon appeals to applications that demand real-time responses and have limited computing resources.

The main challenge in accelerating the LLMs through speculative decoding lies in improving the accuracy and efficiency of the drafter under resource-limited conditions. Therefore, our future research efforts will focus on developing drafters that attain a high token acceptance rate with minimal overhead. This goal will be met by advanced semi-autoregressive or potentially non-autoregressive decoding techniques that fortify the interdependence among tokens. In addition, the dynamic modification of the decoding tree is another avenue that warrants further exploration.

References

- Bao, Y.; Zhou, H.; Huang, S.; Wang, D.; Qian, L.; Dai, X.; Chen, J.; and Li, L. 2022. latent-GLAT: Glancing at Latent Variables for Parallel Text Generation. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8398–8409. Dublin, Ireland: Association for Computational Linguistics.
- Cai, T.; Li, Y.; Geng, Z.; Peng, H.; Lee, J. D.; Chen, D.; and Dao, T. 2024. Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads. *arXiv:2401.10774*.
- Chen, C.; Borgeaud, S.; Irving, G.; Lespiau, J.-B.; Sifre, L.; and Jumper, J. 2023. Accelerating Large Language Model Decoding with Speculative Sampling. *arXiv:2302.01318*.
- Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Chen, Z.; Yang, X.; Lin, J.; Sun, C.; Chang, K. C.-C.; and Huang, J. 2024. Cascade Speculative Drafting for Even Faster LLM Inference. *arXiv:2312.11462*.
- Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Ding, L.; Wang, L.; Liu, X.; Wong, D. F.; Tao, D.; and Tu, Z. 2021. Understanding and Improving Lexical Choice in Non-Autoregressive Translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. *arXiv:2103.10360*.
- Geng, X.; Feng, X.; and Qin, B. 2021. Learning to Rewrite for Non-Autoregressive Neural Machine Translation. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3297–3308. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Ghazvininejad, M.; Levy, O.; Liu, Y.; and Zettlemoyer, L. 2019. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6112–6121. Hong Kong, China: Association for Computational Linguistics.
- Gloeckle, F.; Idrissi, B. Y.; Rozière, B.; Lopez-Paz, D.; and Synnaeve, G. 2024. Better & Faster Large Language Models via Multi-token Prediction. *arXiv:2404.19737*.
- Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, 369–376. New York, NY, USA: Association for Computing Machinery. ISBN 1595933832.
- Guo, J.; Tan, X.; He, D.; Qin, T.; Xu, L.; and Liu, T.-Y. 2019a. Non-autoregressive neural machine translation with enhanced decoder input. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19*. AAAI Press. ISBN 978-1-57735-809-1.
- Guo, J.; Tan, X.; Xu, L.; Qin, T.; Chen, E.; and Liu, T.-Y. 2019b. Fine-Tuning by Curriculum Learning for Non-Autoregressive Neural Machine Translation. *arXiv preprint arXiv:1911.08717*.
- Guo, J.; Tan, X.; Xu, L.; Qin, T.; Chen, E.; and Liu, T.-Y. 2020. Fine-Tuning by Curriculum Learning for Non-Autoregressive Neural Machine Translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7839–7846.
- Guo, J.; Wang, M.; Wei, D.; Shang, H.; Wang, Y.; Li, Z.; Yu, Z.; Wu, Z.; Chen, Y.; Su, C.; Zhang, M.; Lei, L.; shimin tao; and Yang, H. 2021. Self-Distillation Mixup Training for Non-autoregressive Neural Machine Translation. *arXiv:2112.11640*.
- Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531*.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.
- Kim, S.; Mangalam, K.; Moon, S.; Malik, J.; Mahoney, M. W.; Gholami, A.; and Keutzer, K. 2023. Speculative Decoding with Big Little Decoder. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 39236–39256. Curran Associates, Inc.
- Leviathan, Y.; Kalman, M.; and Matias, Y. 2023. Fast Inference from Transformers via Speculative Decoding. *arXiv:2211.17192*.
- Li, Y.; Wei, F.; Zhang, C.; and Zhang, H. 2024. EAGLE: Speculative Sampling Requires Rethinking Feature Uncertainty. *arXiv:2401.15077*.
- Miao, X.; Oliaro, G.; Zhang, Z.; Cheng, X.; Wang, Z.; Zhang, Z.; Wong, R. Y. Y.; Zhu, A.; Yang, L.; Shi, X.; Shi, C.; Chen, Z.; Arfeen, D.; Abhyankar, R.; and Jia, Z. 2024. SpecInfer: Accelerating Large Language Model Serving with Tree-based Speculative Inference and Verification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, ASPLOS '24*, 932–949. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703867.
- Monea, G.; Joulin, A.; and Grave, E. 2023. PaSS: Parallel Speculative Sampling. *arXiv:2311.13581*.
- Olsson, C.; Elhage, N.; Nanda, N.; Joseph, N.; DasSarma, N.; Henighan, T.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Johnston, S.; Jones, A.; Kernion, J.; Lovitt, L.;

- Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2022. In-context Learning and Induction Heads. arXiv:2209.11895.
- Qian, L.; Zhou, H.; Bao, Y.; Wang, M.; Qiu, L.; Zhang, W.; Yu, Y.; and Li, L. 2021. Glancing Transformer for Non-Autoregressive Neural Machine Translation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1993–2003. Online: Association for Computational Linguistics.
- Ran, Q.; Lin, Y.; Li, P.; and Zhou, J. 2020. Learning to Recover from Multi-Modality Errors for Non-Autoregressive Neural Machine Translation. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3059–3069. Online: Association for Computational Linguistics.
- Roman, S. 1992. *Coding and information theory*, volume 134. Springer Science & Business Media.
- Santilli, A.; Severino, S.; Postolache, E.; Maiorca, V.; Mancusi, M.; Marin, R.; and Rodola, E. 2023. Accelerating Transformer Inference for Translation via Parallel Decoding. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12336–12355. Toronto, Canada: Association for Computational Linguistics.
- Saxena, A. 2023. Prompt Lookup Decoding.
- Spector, B.; and Re, C. 2023. Accelerating LLM Inference with Staged Speculative Decoding. arXiv:2308.04623.
- Stern, M.; Shazeer, N.; and Uszkoreit, J. 2018. Blockwise Parallel Decoding for Deep Autoregressive Models. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Sun, Z.; Li, Z.; Wang, H.; He, D.; Lin, Z.; and Deng, Z. 2019. Fast Structured Decoding for Sequence Models. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. arXiv:1609.03499.
- van den Oord, A.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; van den Driessche, G.; Lockhart, E.; Cobo, L.; Stimberg, F.; Casagrande, N.; Grewe, D.; Noury, S.; Dieleman, S.; Elsen, E.; Kalchbrenner, N.; Zen, H.; Graves, A.; King, H.; Walters, T.; Belov, D.; and Hassabis, D. 2018. Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 3918–3926. PMLR.
- Wan, Z.; Wang, X.; Liu, C.; Alam, S.; Zheng, Y.; Liu, J.; Qu, Z.; Yan, S.; Zhu, Y.; Zhang, Q.; Chowdhury, M.; and Zhang, M. 2024. Efficient Large Language Models: A Survey. arXiv:2312.03863.
- Wang, C.; Zhang, J.; and Chen, H. 2018. Semi-Autoregressive Neural Machine Translation. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 479–488. Brussels, Belgium: Association for Computational Linguistics.
- Wang, Y.; Tian, F.; He, D.; Qin, T.; Zhai, C.; and Liu, T.-Y. 2019. Non-Autoregressive Machine Translation with Auxiliary Regularization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 5377–5384.
- Wertheimer, D.; Rosenkranz, J.; Parnell, T.; Suneja, S.; Ranganathan, P.; Ganti, R.; and Srivatsa, M. 2024. Accelerating Production LLMs with Combined Token/Embedding Speculators. arXiv:2404.19124.
- Xia, H.; Ge, T.; Chen, S.-Q.; Wei, F.; and Sui, Z. 2023. Speculative Decoding: Lossless Speedup of Autoregressive Translation.
- Xia, H.; Yang, Z.; Dong, Q.; Wang, P.; Li, Y.; Ge, T.; Liu, T.; Li, W.; and Sui, Z. 2024. Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding. arXiv:2401.07851.
- Xiao, Y.; Wu, L.; Guo, J.; Li, J.; Zhang, M.; Qin, T.; and yan Liu, T. 2023. A Survey on Non-Autoregressive Generation for Neural Machine Translation and Beyond. arXiv:2204.09269.
- Xu, W.; Ma, S.; Zhang, D.; and Carpuat, M. 2021. How Does Distilled Data Complexity Impact the Quality and Confidence of Non-Autoregressive Machine Translation? In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4392–4400. Online: Association for Computational Linguistics.
- Zhang, J.; Wang, J.; Li, H.; Shou, L.; Chen, K.; Chen, G.; and Mehrotra, S. 2024. Draft & Verify: Lossless Large Language Model Acceleration via Self-Speculative Decoding. arXiv:2309.08168.
- Zhao, Y.; Xie, Z.; Liang, C.; Zhuang, C.; and Gu, J. 2024. Lookahead: An Inference Acceleration Framework for Large Language Model with Lossless Generation Accuracy. arXiv:2312.12728.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.
- Zhu, X.; Li, J.; Liu, Y.; Ma, C.; and Wang, W. 2024. A Survey on Model Compression for Large Language Models. arXiv:2308.07633.
- Łukasz Kaiser; Roy, A.; Vaswani, A.; Parmar, N.; Bengio, S.; Uszkoreit, J.; and Shazeer, N. 2018. Fast Decoding in Sequence Models using Discrete Latent Variables. arXiv:1803.03382.

A Tree Structure

A custom-designed decoding tree is introduced to accommodate SAR drafting. Figure 5 and figure 6 respectively display the tree structures utilized by the Vicuna models and the LLaMA models. It’s important to note that, apart from the root node, each node on the tree represents only one token. The number of tokens for the root node is determined by k (the number of tokens initially generated by LLM or produced by the draft model in each forward pass). In a greedy setting, we select the top m tokens with the highest probabilities as child nodes. In a non-greedy setting, we sample m tokens. The number of child nodes, m , can be inferred from Figure 5; for instance, $m = 20$ at the root node. For different nodes, the value of m can vary. For trees of the same depth and k , the draft model undergoes the same number of forward passes during the draft phase. For example, whether it’s Figure 5 or Figure 6 of these trees, when k equals 2, the draft model performs 4 forward passes.

Why do we use such a complex tree structure? The choices of tree structure are optimized based on intuition and findings during experiments: branches of higher-probability tokens should be deeper and wider. We also find that the optimal tree structure is likely both context-dependent and model-dependent. For instance, the tree structure that is effective for the Vicuna models may not necessarily yield the best acceleration results when applied to the LLaMA models, thus we design different tree structures for these models.

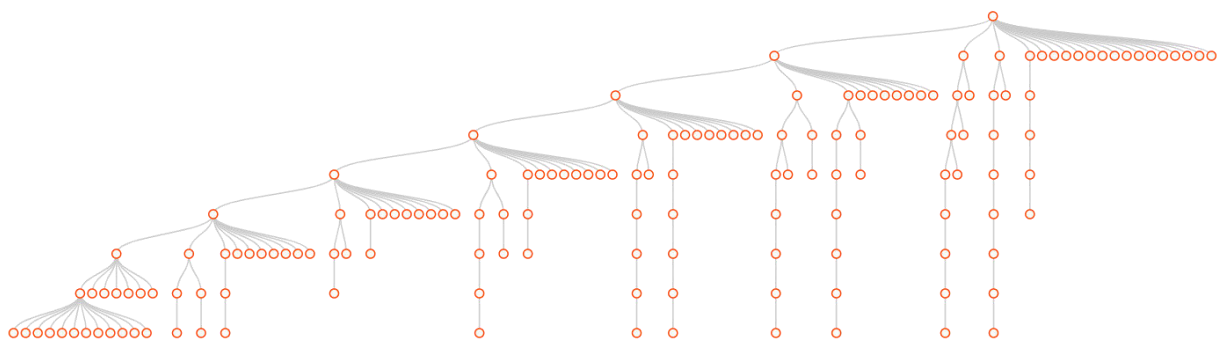


Figure 5: The tree structure used for Vicuna series.

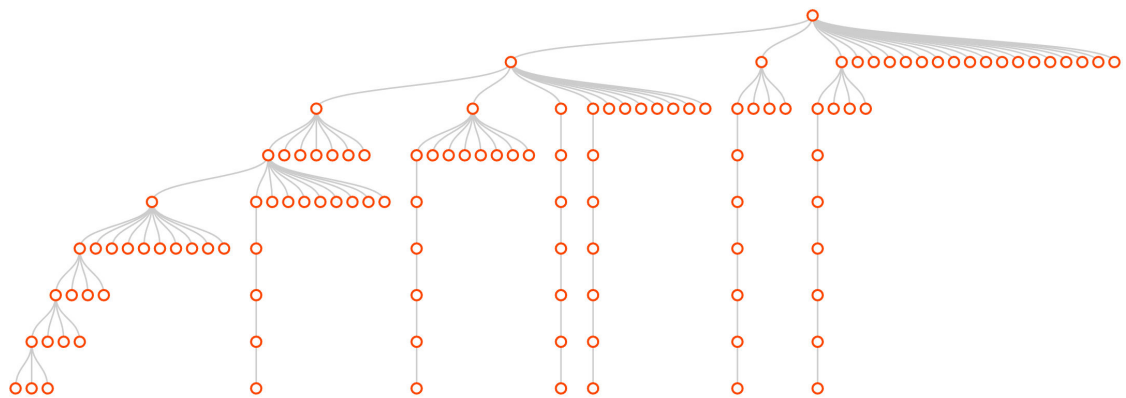


Figure 6: The tree structure used for LLaMA series.