FastDraft: How to Train Your Draft

Ofir Zafrir* **Igor Margulis*** **Dorin Shteyman*** **Guy Boudoukh**
Intel Labs
{ofir.zafrir, igor.margulis, dorin.shteyman, guy.boudoukh}@intel.com

Abstract

Speculative Decoding has gained popularity as an effective technique for accelerating the auto-regressive inference process of Large Language Models (LLMs). However, Speculative Decoding entirely relies on the availability of efficient draft models, which are often lacking for many existing language models due to a stringent constraint of vocabulary incompatibility. In this work we introduce FastDraft, a novel and efficient approach for pre-training and aligning a draft model to any large language model by incorporating efficient pre-training, followed by fine-tuning over synthetic datasets generated by the target model. We demonstrate FastDraft by training two highly parameter efficient drafts for the popular Phi-3-mini and Llama-3.1-8B models. Using FastDraft, we were able to produce a draft with approximately 10 billion tokens on a single server with 8 accelerators in under 24 hours. Our results show that the draft model achieves impressive results in key metrics of acceptance rate, block efficiency and up to 3x memory bound speed up when evaluated on code completion and up to 2x in summarization, text completion and instruction tasks. Due to its high quality, FastDraft unlocks large language models inference on AI-PC and other edge-devices.

1 Introduction

The advent of Transformer architectures has fundamentally reshaped the field of natural language processing (NLP). In recent years, Transformer-based models have achieved remarkable success across a broad spectrum of natural language understanding and generation tasks [Achiam et al., 2023, Team et al., 2023, Abdin et al., 2024]. Their exceptional performance, particularly in large language models (LLMs), has made them highly desirable for deployment in numerous applications, ranging from conversational systems to content generation and beyond. Despite their outstanding performance, LLMs suffer from slow inference speeds due to substantial memory bandwidth requirements and the sequential nature of auto-regressive generation (ARG). The introduction of Speculative Decoding (SD) [Leviathan et al., 2023] offers a promising solution for accelerating ARG without sacrificing generation quality, making it a compelling approach for improving LLM inference efficiency. SD utilizes a draft language model (LM) to generate a sequence of tokens auto-regressively, while the target model validates the batched tokens in parallel. In certain applications, SD can achieve a 2-3x speedup in LLM inference without compromising the generation quality of the target model. Achieving significant speedup with SD requires a high-quality draft model that is both efficient and well-aligned with the target. To date, such draft models remain scarce, even for widely used open-source LLMs [Dubey et al., 2024, Abdin et al., 2024] due to vocabulary incompatibility. To address this limitation, we propose FastDraft, a method for producing hardware-efficient draft models that are orders of magnitude smaller than their corresponding target models.

While extensive research has focused on training and data generation for high-quality LLMs [Kaplan et al., 2020, Longpre et al., 2023], these frameworks are not necessarily applicable to training draft models for SD. LLMs are typically trained to generate helpful, high-quality responses, whereas

*Equal contribution

4th NeurIPS Efficient Natural Language and Speech Processing Workshop (ENLSP-IV 2024).

draft models should be trained to generate sequences that are likely to be accepted by the target model. We explore previously unexamined aspects of draft model training for SD. Our contributions include: (1) introducing a method for producing quality and highly efficient draft models with low resource requirements for any given target LLM and demonstrate it by training and benchmarking a draft for Phi-3-mini, (2) conducting extensive ablation studies on pre-training data size, pre-training for both code and natural language and target-draft alignment via knowledge distillation (KD) and (3) demonstrating the scalability of FastDraft by training a draft model for Llama-3.1-8B-Instruct, achieving performance improvements comparable to those attained for Phi-3-mini. We demonstrate significant improvements in key metrics using FastDraft, leading to theoretical speedups of up to 3x as measured by the Memory Bound Speedup (MBSU) metric. According to our findings, the small size of the draft model and the limited amount of data required to produce a high-quality draft enabled us to successfully train and align a draft model to Phi-3-mini, end-to-end, in under 24 hours using a single node with 8 accelerators.

2 Related work

The widespread adoption of LLMs in cloud and edge devices has driven a significant body of research focused on developing alternative strategies for ARG to address the slow performance of LLM inference [Santilli et al., 2023, Ghazvininejad et al., 2019, Stern et al., 2018]. However, many of these approaches compromise generation quality or require additional training data and architectural modifications. The introduction of speculative decoding as a lossless solution for accelerating LLM inference [Leviathan et al., 2023] has inspired a new wave of follow-up research. Some studies propose using plug-in prediction heads as a drafting mechanism [Zhang et al., 2024, Cai et al., 2024], while others focus on improving the serving latency of stand-alone draft models [Sun et al., 2024, Chen et al., 2023, Miao et al., 2023]. In contrast, our work focuses on directly enhancing stand-alone draft model's capabilities through pre-training and fine-tuning. Other works apply KD to draft models to improve alignment with the target model. These studies explore various divergence functions for the KD algorithm, rather than relying solely on the commonly used Kullback–Leibler Divergence (KLD). For instance, Zhou et al. [2023] proposed Total Variation Distance (TVD) based on Corollary 3.6 in [Leviathan et al., 2023], which posits that minimizing TVD maximizes the token-level acceptance rate. Goel et al. [2024] further developed this approach with TVD++, fine-tuning their model, pre-trained on 600 billion tokens. Another study, Yan et al. [2024], focuses on enhancing the hardware efficiency of draft models by extensively analyzing the trade-off between time latency and acceptance rates, rather than relying solely on the latter. While their work emphasizes optimizing the draft architecture, our approach fixes the architecture and introduces efficient approach for pre-training and aligning a draft model. Li et al. [2024] proposes training a compact draft model based on the target model, using a relatively limited dataset. The primary aim of the paper is to develop a draft architecture that effectively utilizes the target model's hidden representations and weights. In contrast, our paper centers on training and aligning any draft architecture that shares only the vocabulary with the target model.

3 Speculative decoding

SD is a lossless decoding paradigm introduced by Leviathan et al. [2023] for accelerating ARG with LLMs. It is inspired by speculative execution [Hennessy and Patterson, 2017] and aims to mitigate the inherent latency bottleneck caused by the sequential nature of ARG [Pope et al., 2023]. SD employs a draft LM to generate a block of γ candidate tokens. The LLM, referred to as the target model, then processes these candidate tokens in parallel. The algorithm examines each token's probability distribution, calculated by both the target and draft models, to determine whether the token should be accepted or rejected. As a result, any LM can function as a draft model, provided it shares the same vocabulary as the target model. However, since an LM's vocabulary is fixed during pre-training, leveraging existing models as draft models is only feasible if they were pre-trained on the same vocabulary. A key metric for benchmarking ARG inference performance is time per output token (TPOT), which represents the expected latency for generating a single output token, excluding pre-filling latency.[1] In SD, TPOT is a function of the draft model latency l_D, the speculation block

[1] Pre-filling refers to the action of populating the key-value cache with the input tokens' information.

size γ, the corresponding expected block efficiency τ^{γ}[2], and the target model latency l_T^{γ}. The equation is expressed as:

$$\mathrm{TPOT}_{\mathrm{SD}} = \frac{l_D \times \gamma + l_T^{\gamma}}{\tau^{\gamma}} \tag{1}$$

In comparison, for traditional ARG, TPOT is simply $\mathrm{TPOT}_{\mathrm{AR}} = l_T^1$. The expected speedup is calculated as follows:

$$\frac{\mathrm{TPOT}_{\mathrm{AR}}}{\mathrm{TPOT}_{\mathrm{SD}}} = \left(\frac{l_D}{l_T^1} \times \gamma + \frac{l_T^{\gamma}}{l_T^1}\right)^{-1} \times \tau^{\gamma} \tag{2}$$

From Equation 2, two key requirements emerge for SD to yield meaningful speedups. First, the latency for generating the speculated tokens block must be negligible compared to the target model's latency. Second, the increase in target model latency for block size γ should be insignificant compared to the latency for block size 1. The latter condition generally holds for most popular LLMs with sufficiently small γ, while the former depends on the availability of draft models that meet the vocabulary constraint.

4 FastDraft: Build your own draft model

4.1 Draft architecture & pre-training

The draft architecture imposes only one strict requirement, it must produce a probability distribution over the target's vocabulary. Beyond this, the design is flexible. Nevertheless, certain factors should be considered when selecting the draft's architecture, with latency being the primary concern, as discussed in Section 3. Then, the chosen draft is trained over a pre-training dataset of natural language for language modeling with the objective of predicting the next token. A common application of LLMs is code completion. While the pre-training dataset may include some code, unless the dataset is specifically focused on code, the draft's performance on tasks requiring an understanding of code is often suboptimal. To address this, Aryabumi et al. [2024] proposed continued pre-training (CP), where training begins with a pre-trained model and is extended using a combination of code and natural language data. FastDraft adopts this approach for producing drafts for code completion tasks.

4.2 Target-draft alignment

One of the primary objectives when designing a draft model is to maximize the acceptance rate with the target model, as a higher acceptance rate directly leads to greater speedup in SD. To closely mimic the target model's behavior in real-world scenarios, we investigate two knowledge distillation strategies that expose the draft model to data samples that closely reflect those generated by the target model.

Strategy 1: sequence-level knowledge distillation Kim and Rush [2016] proposed performing knowledge distillation (KD) by training the student model on sequences generated by the teacher model. This approach is widely used in large language model (LLM) training to improve model quality [Taori et al., 2023, Abdin et al., 2024]. To align the pre-trained draft model with the teacher model, we employ sequence-level KD by fine-tuning the draft model on a synthetic dataset generated by the target. In contrast to the sequence-level KD approach proposed by Kim and Rush [2016], we differentiate between training on teacher-generated data and training with knowledge distillation using the teacher's logits. Accordingly, in this work, sequence-level KD refers specifically to training on generated tokens with cross-entropy loss, without incorporating the teacher's soft targets (logits).

Strategy 2: token-level knowledge distillation This strategy aligns with the traditional KD method presented in [Hinton, 2015], which involves calculating the divergence of the token level probability distribution over the vocabulary between the teacher and the draft. Since computing the KD loss depends on the teacher's logits, the teacher is required to perform inference at every training step of

[2] The expected number of tokens accepted within a speculation block.

the draft, making this process both computationally intensive and time-consuming. An alternative approach is to precompute the teacher's logits beforehand and embed them into the dataset. However, the memory demands of such a dataset can quickly exceed several terabytes, as the number of logits per token equals the vocabulary size[3] posing a substantial memory overhead. To mitigate this issue, only a small subset of the most significant logits per token is extracted, significantly reducing the memory requirements. This optimization results in a 6x-9x reduction in fine-tuning time without noticeable degradation in model quality.

4.3 Draft evaluation

To scale both our experiments and evaluations, we created the FastDraft evaluation framework, specifically designed for assessing draft models. The metrics and benchmarks we implemented in FastDraft evaluation are detailed in Sections 4.3.1 and 4.3.2. Optimizing these key metrics on the proposed benchmarks is our key objective in this work.

4.3.1 Metrics

Acceptance rate A key metric which reflects the rate at which the target model accepts the draft model's speculated tokens. In this work, we calculate the acceptance rate (AR) α^γ by determining the expected number of tokens accepted per block normalized by the block's size. The formula is expressed as:

$$\alpha^\gamma = \frac{1}{N} \sum_{n=0}^{N} \frac{\#(\text{accepted tokens})}{\gamma} \tag{3}$$

where N is the number of blocks speculated by the draft model during evaluation, and γ is the block size.

Block efficiency The common use of SD with fixed-size blocks motivates the introduction of a more relevant efficiency metric: block efficiency, τ^γ. This is defined as the expected number of accepted tokens per block. For a given block size γ, block efficiency τ^γ serves as a more accurate measure of performance, reflecting the average rate at which tokens are accepted within each block.

$$\tau^\gamma = 1 + \alpha^\gamma \times \gamma \tag{4}$$

Wall-clock time and Memory-Bound Speedup Given the block efficiency τ^γ, the expected speedup achieved by applying Speculative Decoding (SD) is expressed as $\frac{\tau^\gamma}{c\gamma+1}$, where c represents the ratio of latencies between the draft and target models (see Equation 2 or [Leviathan et al., 2023]). Since this metric is hardware-dependent, it is preferable to use a hardware-agnostic measure. Considering that the expected speedup occurs in a memory-bound regime, we utilize the Memory-Bound Speedup (MBSU). We define $\hat{c}$ as the ratio of parameter counts between the draft and target models.

$$\text{MBSU} = \frac{\tau^\gamma}{\hat{c}\gamma + 1} \tag{5}$$

4.3.2 Benchmarks

The most commonly used benchmarks for assessing the quality of draft models, typically open-ended generation and summarization tasks, include the XSum [Narayan et al., 2018], CNN/DailyMail [Hermann et al., 2015], and HumanEval [Chen et al., 2021] datasets. In addition to CNN/DailyMail and HumanEval, we include in our evaluation the TinyStories [Eldan et al., 2023] and Dolly [Conover et al., 2023] datasets, providing a more comprehensive assessment of model performance across diverse tasks. When utilizing the TinyStories dataset, we generate text by starting from a random position within each sample, using the preceding tokens as the input context.

[3] The vocabulary size usually exceeds 30,000 tokens

Table 1: Pre-training data size effect on acceptance rate and perplexity. Results for block size $\gamma = 3$ and multinomial sampling with temperature $T = 0.6$

Draft size	Data size	Perplexity	CNN-DailyMail	TinyStories	Dolly
50M	2BT	297.4	**0.323**	0.264	0.241
	5BT	256.6	0.311	0.277	**0.245**
	10BT	**240.9**	0.312	**0.283**	0.234
120M	2BT	199.6	0.362	0.297	**0.284**
	5BT	167.7	**0.366**	0.327	0.281
	10BT	**147.4**	0.351	**0.331**	0.251

5 Experiments

In this section we report descriptions, setup and results of our experiments. In the following experiments, we report acceptance rate (Section 4.3.1) results using two sampling methods, greedy and multinomial sampling decoding with temperature $T = 0.6$ with block sizes $\gamma = 3$ and $\gamma = 5$ unless stated otherwise. The hyper-parameters used in our draft pre-training experiments are detailed Appendix A.3 unless stated otherwise.

5.1 Experimental setup

We demonstrate FastDraft by training a draft model for Phi-3-mini-4k-Instruct Abdin et al. [2024] target model. Phi-3-mini, a 3.8 billion parameters LLM, was selected as our case study due to its outstanding performance across multiple open-source LLM benchmarks, while also being capable of running locally on edge devices, such as personal computers and smartphones.

Draft architecture Exploring different architectures for draft models lies beyond the scope of this paper. Therefore, our experiments focus on drafts with the Phi-3 architecture, modified to smaller dimensions. Specifically, we reduced the dimensions of Phi-3 to create drafts in the size of 50M and 120M parameters which are approximately 76x and 32x smaller than Phi-3-mini. Full details of drafts configurations can be found in Table 5 in the Appendix.

Pre-training datasets For our natural language dataset, we use a 10 billion tokens (BT) sample from the FineWeb dataset [Penedo et al., 2024]. FineWeb is a widely used open-source, de-duplicated, and quality-filtered dataset, comprising 15 trillion tokens derived from 96 Common Crawl snapshots [Crawl]. It has been demonstrated to yield better-performing LLMs compared to other open-source pre-training datasets such as C4, RedPajama, and The Pile (Section 3.7 in Penedo et al. [2024]). For our code dataset, we use a 10BT sample from The Stack v2 smol dataset [Lozhkov et al., 2024]. This variant of The Stack v2 consists of 17 commonly used programming languages, as well as a substantial collection of documentation languages, configuration languages, and configuration files. An overview of the 10BT dataset composition is provided in Appendix A.2.

Synthetic alignment dataset To perform sequence-level KD as discussed in Section 4.2, we produce an instruction fine-tuning collection using the seed instructions from several open instruction datasets: Alpaca [Taori et al., 2023], OIG-small-chip2 [Huu et al., 2023] and Evol-Instruct [Luo et al., 2023]. We collect response sequences generated by Phi-3-mini with greedy sampling along with multinomial sampling with temperature in $\{0.6, 0.8, 1.0\}$ to improve the diversity of the generated sequences. Additionally, we adopt the approach proposed by Xu et al. [2024], directly soliciting instructions from the target model. These instructions are then used to generate the corresponding responses in the same manner as previously described.

5.2 Pre-training dataset size

We study the impact of the pre-training dataset size on the performance of the draft model, aiming to optimize runtime and resource usage and prevent dimishing returns. Consequently, we uniformly sample subsets of sizes $\{0.1, 0.5, 1, 2, 5, 10\}$BT and pre-train the 50M and 120M draft configurations on these subsets. We report the AR results for the $\{2, 5, 10\}$BT subsets with block size $\gamma = 3$ and

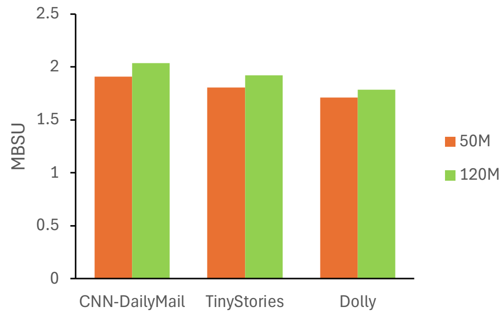

Figure 1: MBSU of 50M and 120M pre-trained drafts on 5BT FineWeb sample.

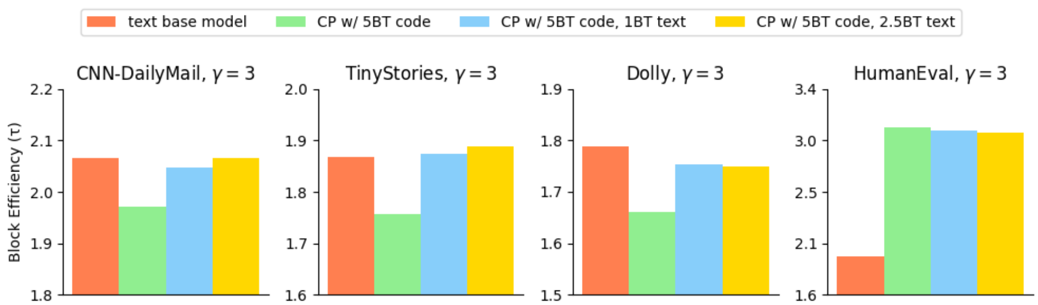

Figure 2: Block efficiency results for continued pre-training with code on tasks: CNN-DailyMail, TinyStories, Dolly and HumanEval using greedy decoding with block size $\gamma = 3$

multinomial sampling in addition to perplexity results measured on Wikitext2 [Salesforce] in Table 1. The full results are reported in Table 7 in the Appendix.

As anticipated, perplexity values decrease as the models are exposed to more training data, though the rate of improvement slows down. However, when looking at the AR results, it is not the case. In CNN-DailyMail and Dolly AR either plateaus or decreases as the data size grows. In the case of TinyStories, AR increases with the amount of data which can be expected due to the nature of the benchmark of text completion versus instruction following and summarization. Overall, we obtain strong results across all dataset sizes, with the models trained on the 5BT dataset emerging as a promising middle-ground option. When comparing the 50M draft to the 120M draft, the 120M draft demonstrates superior performance in terms of AR. Additionally, while the 120M draft also outperforms the 50M model in MBSU, the margin of improvement is relatively small, see Figure 1. However, it is important to note that training the 120M draft requires twice the time compared to the 50M draft.

For the subsequent experiments, we selected the 50M draft model trained on 5BT tokens as our pre-trained draft. Unless otherwise specified, references to a pre-trained draft in this section pertain to this model.

5.3 Continued pre-training for code

We investigate the continued pre-training (CP) method introduced by [Aryabumi et al., 2024] to refine our draft model on code-related tasks. In this approach, we extend the training of our pre-trained draft model using three distinct combinations of code and natural language datasets. Each combination incorporates 5BT of code data, along with varying amounts of natural language data: specifically, $0, 1, 2.5$BT. We then evaluate the resulting models on our benchmarks, including the HumanEval benchmark, and present the block efficiency results, τ^{γ}, in Figure 2. Our findings indicate that mixed CP datasets lead to significantly higher block efficiency in natural language tasks compared to using CP with code alone. Additionally, block efficiency improves across natural language benchmarks, surpassing the performance of the base draft model that was trained solely on natural language, with

Table 2: AR results for fine-tuning with original data vs synthetic data

Sampling method	Data source	CNN-DailyMail		TinyStories		Dolly	
		$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$
Greedy	None	0.350	0.246	0.295	0.196	0.259	0.174
	Original	0.357	0.251	0.282	0.189	0.328	0.228
	Target	**0.378**	**0.270**	**0.296**	**0.201**	**0.370**	**0.262**
Multinomial sampling	None	0.311	0.221	0.227	0.184	0.245	0.163
	Original	0.328	0.227	0.268	0.181	0.311	0.215
	Target	**0.339**	**0.242**	**0.286**	**0.193**	**0.352**	**0.234**

Table 3: Knowledge distillation with KL-Divergence and TVD

Sampling method	Loss	CNN-DailyMail		TinyStories		Dolly	
		$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$
Greedy	$\mathcal{L}_{CE}$	0.378	0.270	0.296	0.201	0.370	0.262
	$0.5\mathcal{L}_{CE} + 0.5\mathcal{L}_{KL}$	0.376	0.269	0.295	0.199	0.368	0.261
	$0.5\mathcal{L}_{CE} + 0.5\mathcal{L}_{TVD}$	0.377	0.270	0.297	0.202	0.370	**0.263**
	$\mathcal{L}_{KL}$	0.374	0.266	0.294	0.198	**0.371**	0.261
	$\mathcal{L}_{TVD}$	**0.384**	**0.274**	**0.301**	**0.204**	0.362	0.254
Multinomial sampling	$\mathcal{L}_{CE}$	0.339	0.242	0.286	0.193	**0.352**	0.234
	$0.5\mathcal{L}_{CE} + 0.5\mathcal{L}_{KL}$	0.350	0.248	0.289	0.190	0.346	0.235
	$0.5\mathcal{L}_{CE} + 0.5\mathcal{L}_{TVD}$	**0.356**	0.243	**0.297**	0.194	0.347	**0.244**
	$\mathcal{L}_{KL}$	0.355	0.235	0.280	0.188	0.343	0.237
	$\mathcal{L}_{TVD}$	0.347	**0.250**	**0.297**	**0.197**	0.344	0.237

only a slight decline observed in the Dolly benchmark. Notably, the base draft shows a substantial improvement on the HumanEval benchmark. The enhancement in the draft's performance on natural language tasks with CP was anticipated, based on the findings of [Aryabumi et al., 2024]. However, unlike the conclusions presented in that study, which advocate for a balanced dataset of natural language and code, our results suggest that CP is the preferred approach for integrating code and natural language datasets when training a draft model.

Given the structured nature of code, it serves as a highly suitable domain for generation with SD. To investigate how to optimize pre-training for code drafts, we conducted a comprehensive ablation study. This study involved pre-training from scratch on a mixture of code and pre-training with CP approach, where the base draft model was pre-trained on code, and CP was applied to adapt the model for natural language instead of code. Detailed results and analysis of this ablation study are provided in Appendix A.5. Overall, our findings indicate that initializing continued pre-training from a text-based model using a mixed CP dataset offers the most effective pre-training strategy for draft models, compared to other configurations tested.

5.4 Fine tuning with synthetic data vs original data

Since our objective is to achieve the best draft alignment with the target model rather than the best answers, we investigate the impact of using the target's responses instead of the original dataset responses, as described in Section 4.2. We measure the draft's performance across several tasks. In this experiment, we utilize the OIG-small-chip dataset for fine-tuning.We generate the target's responses with multinomial sampling at temperature of 0.6. Token-level KD was applied in this experiment. The results are presented in Table 2. An analysis of the data reveals that using the target's answers leads to better draft alignment compared to using the original dataset answers across all tasks, with one exception. The draft's quality decreases on the TinyStories task, which is expected, as text completion quality commonly declines following instruction fine-tuning. Nevertheless, it is noteworthy that the draft fine-tuned on the target's answers maintained higher quality than the draft fine-tuned on the original answers.

Table 4: AR results for FastDraft stages, performed in subsequent order: Pre-training (PT), Continued Pre-training (CP) and Fine-tuning (FT) using multinomial sampling with temperature $T = 0.6$

Model	Type	CNN-DailyMail		TinyStories		Dolly		HumanEval	
		$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$
Phi3-mini 50M	PT	0.311	0.221	0.277	0.184	0.245	0.163	0.229	0.151
	$PT \rightarrow CP$	0.304	0.211	0.287	0.192	0.226	0.149	0.561	0.450
	$PT \rightarrow CP \rightarrow FT$	**0.369**	**0.267**	**0.306**	**0.208**	**0.370**	**0.265**	**0.562**	**0.472**
Llama3.1 150M	PT	0.280	0.186	0.227	0.147	0.247	0.158	0.248	0.168
	$PT \rightarrow CP$	0.280	0.192	0.235	0.155	0.273	0.176	0.606	0.480
	$PT \rightarrow CP \rightarrow FT$	**0.307**	**0.214**	**0.266**	**0.178**	**0.334**	**0.239**	**0.649**	**0.525**

5.5 Fine tuning with knowledge distillation

To further enhance the alignment of our base draft model, we incorporate knowledge distillation alongside training on target-generated data, as discussed in Section 4.2. We experimented with various combinations of Cross Entropy (CE) loss, denoted as $\mathcal{L}_{CE}$, applied to the ground truth labels, and knowledge distillation losses, using the sparse logits collected from the target model: $\mathcal{L}_{KL}$ and $\mathcal{L}_{TVD}$. We utilize the same dataset we generated in Section 5.4 for this experiment. The results of these experiments are presented in Table 3. Our findings indicate that knowledge distillation in this setup does not provide a significant advantage over CE loss on the target-generated data. Although Total Variation Distance (TVD) has been shown to be negatively correlated with acceptance rates, in our results, it offers only slight benefit on some benchmarks over CE and Kullback–Leibler Divergence (KLD).

6 Results & reproducibility

Combining the findings from the ablation studies presented in section 5, we outline our comprehensive pipeline for draft pre-training and fine-tuning to optimize performance on key metrics for speculative decoding and demonstrate it by producing a draft for Phi-3-mini. To further illustrate reproducibility, we employ the same pipeline to produce a draft for Llama-3.1-8B-Instruct Dubey et al. [2024]. We exhibit a 50M draft for the Phi-3-mini and a 150M draft for Llama-3.1-8B. Drafts' architecture details are presented in Appendix A.1. Considering both performance and resource efficiency, our findings from sections 5.2, 5.3 suggest pre-training over 5BT of FineWeb text data and continued pre-training on a mixture of 5BT code data from The Stack v2 and 2.5BT FineWeb text data as a best practice. For fine-tuning, we conclude from sections 5.4, 5.5 that sequence-level KD yields significant improvements while token-level KD benefits are not definitive, therefore, for FastDraft we only utilize sequence-level KD. We construct an alignment dataset for both Phi-3 and Llama-3.1 drafts by combining a number of synthetic datasets we generated with the appropriate target model as described in Section 5.1. Table 4 presents the AR improvements achieved through each stage of FastDraft generated with multinomial sampling. Results for greedy sampling are presented in Table 14 in the appendix. Using both sampling methods, we observe for datasets CNN-DailyMail, TinyStories, and Dolly, a substantial AR increase following fine tuning, with instruction-following dataset Dolly exhibiting a ~10% AR increase. For HumanEval the primary performance gains stem from pre-training on code domain knowledge during continued pre-training. These models achieve a MBSU of ~2x for natural language tasks and ~3x for code completion tasks.

7 Conclusion & future work

In this paper, we introduced FastDraft, a novel approach for training and evaluating draft models for speculative decoding. Our results demonstrate that FastDraft facilitates the rapid training of high-quality, efficient draft models that are well-aligned with target models. We conducted a comprehensive ablation study, addressing various aspects of draft training, including pre-training data composition and draft-target alignment. Using our method, we successfully trained a highly efficient 50M-parameter draft model for Phi-3-mini, achieving up to a 67% acceptance rate and up to a 3x memory-bound speedup. We successfully demonstrated the scalability of FastDraft by training a draft for Llama-3.1-8B-Instruct, showing its effectiveness for this model as well. We hope that the findings of this work will inspire further research into efficient draft training, particularly focusing on the development of resource-efficient draft architectures and hardware-aware designs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

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

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023.

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

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

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

Aonan Zhang, Chong Wang, Yi Wang, Xuanyu Zhang, and Yunfei Cheng. Recurrent drafter for fast speculative decoding in large language models. *arXiv preprint arXiv:2403.09919*, 2024.

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

Hanshi Sun, Zhuoming Chen, Xinyu Yang, Yuandong Tian, and Beidi Chen. Triforce: Lossless acceleration of long sequence generation with hierarchical speculative decoding. *arXiv preprint arXiv:2404.11912*, 2024.

Ziyi Chen, Xiaocong Yang, Jiacheng Lin, Chenkai Sun, Jie Huang, and Kevin Chen-Chuan Chang. Cascade speculative drafting for even faster llm inference. *arXiv preprint arXiv:2312.11462*, 2023.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, et al. Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification. *arXiv preprint arXiv:2305.09781*, 2023.

Yongchao Zhou, Kaifeng Lyu, Ankit Singh Rawat, Aditya Krishna Menon, Afshin Rostamizadeh, Sanjiv Kumar, Jean-François Kagy, and Rishabh Agarwal. Distillspec: Improving speculative decoding via knowledge distillation. *arXiv preprint arXiv:2310.08461*, 2023.

Raghavv Goel, Mukul Gagrani, Wonseok Jeon, Junyoung Park, Mingu Lee, and Christopher Lott. Direct alignment of draft model for speculative decoding with chat-fine-tuned llms. *arXiv preprint arXiv:2403.00858*, 2024.

Minghao Yan, Saurabh Agarwal, and Shivaram Venkataraman. Decoding speculative decoding. *arXiv preprint arXiv:2402.01528*, 2024.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*, 2024.

John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Morgan kaufmann, 2017.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5, 2023.

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? exploring impact of code in pre-training. *arXiv preprint arXiv:2408.10914*, 2024.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, 2016.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Xsum: A new dataset for abstractive summarization of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 701–711, 2018. doi: 10.18653/v1/D18-1050. URL `https://aclanthology.org/D18-1050`.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1693–1701, 2015. URL `https://arxiv.org/abs/1506.03340`.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Ronen Eldan, Elad Liebman, Colin Cherry, and Yinhan Liu. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023. URL `https://arxiv.org/abs/2305.07759`.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL `https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm`.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

Common Crawl. Common crawl. `https://commoncrawl.org/`.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024.

Nguyen Huu, Suri Sameer, Ken Tsui, Shahules786, Together.xyz team, and Christoph Schuhmann. Oig-small-chip2. `https://huggingface.co/datasets/0-hero/OIG-small-chip2`, 2023.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024.

Salesforce. Salesforce wikitext dataset. `https://huggingface.co/datasets/Salesforce/wikitext`.

A Pre-training details and results

A.1 Draft model architecture

We work with drafts built on the architecture of "Phi-3-mini-4k-instruct" and "Llama-3.1-8B-Instruct". We use float16 precision for model pre-training. Table 5 below provides a detailed view of the structures of the drafts.

Table 5: Phi-3-mini and Llama-3.1-8B-Instruct drafts configurations

Draft name	Phi3-mini 50M	Phi3-mini 120M	Llama3.1 150M
Hidden size	512	768	512
Intermediate size	1408	2048	1792
# Layers	6	12	6
#Attention heads	8	12	8
# Key-value heads	8	12	8
Vocabulary size	32064	32064	128256

A.2 Details of pre-training code dataset

In Table 6, we summarize the data composition of the code dataset we employ for pre-training.

A.3 Hyper-parameter configuration for datasets size in pre-training experiments

We train all draft variants, for both pre-training and continued pre-training (CP) in float16 precision for 1 epoch with a batch size of 128. We use Adam with learning rate of 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay of 0.01, learning rate warmup over 5% of the total training steps, and linear decay of the learning rate.

A.4 Pre-training model size ablation

We provide the full results of Table 1, pre-training on varying model sizes of 50 and 120 million parameters in Table 7. For a detailed view of the drafts' structure, see Table 5

A.5 Pre-training on code ablation

Pre-training on a mixed dataset Consistent with current best practices for pre-training large language models Dubey et al. [2024], Abdin et al. [2024], we experiment with pre-training the draft model on a mixture of code and natural language datasets with varying ratios of the domains. Similar to Aryabumi et al. [2024], we evaluate a balanced mixture of 5 billion token code and 5 billion token text datasets. Since our datasets are relatively small to begin with, emphasizing on text data can be beneficial for general language understanding across tasks. Accordingly, we also experiment with mixed datasets featuring 4 billion tokens of text and 6 billion tokens of code, as well as 7 billion tokens of text and 3 billion tokens of code.

Our experimental results suggest the following conclusions:

Continued pre-training is better than pre-training on a mixed dataset: Continued pre-training variants with CP mixed dataset of code and text were able to obtain 1-4% higher acceptance rate on CNNT-DailyMail, Tinystories and Dolly datasets, over greedy and multinomial sampling methods (see Table 9, 11) compared to CP datasets of only code or only text. These models have been trained on 1-2.5 billion more tokens compared to pre-trained models on mixed data, but as section 5.2 suggests, the performance gain likely stems from the choice of pre-training strategy rather than additional data. Surprisingly, mixed datasets with higher proportion of text yield better results on HumanEval, a code evaluation dataset, with little to no improvement on natural language tasks 12, 13. This strengths our findings in section 5.2 that additional text data doesn't contribute to our draft performance beyond a certain point.

Table 6: Overview of the data composition of the 10BT sample of the-stack-v2-train-smol

Language	Token count (B)	Sample Count	Avg. sample length (tokens)
Python	1.18	963923	1227
C	0.83	309496	2679
C#	0.78	783399	994
C++	1.35	676943	1994
Go	0.18	140831	1285
Java	1.02	1003702	1015
JavaScript	1.27	1109269	1145
Kotlin	0.09	142275	601
Lua	0.1	32335	3212
PHP	1.07	731103	1462
R	0.13	79108	1704
Ruby	0.13	285785	469
Rust	0.07	36811	1926
SQL	0.26	70937	3633
Shell	0.11	175985	613
Swift	0.1	118558	877
TypeScript	0.26	386644	674
Documentation languages	0.96	807360	1194
Configuration languages	0.09	140370	652
Configuration files	0.02	43507	554
Total	10.0	8038341	

Table 7: Pre-training data size effect on acceptance rate of natural language tasks. Results for block size $\gamma = 3$ on both greedy and multinomial sampling temperature $T = 0.6$

Draft size	Data size	Perplexity	CNN-DailyMail		TinyStories		Dolly	
			Greedy	Multinomial	Greedy	Multinomial	Greedy	Multinomial
50M	0.1BT	1594.5	0.168	0.152	0.158	0.143	0.173	0.151
	0.5BT	476.3	0.304	0.280	0.245	0.226	0.253	0.233
	1BT	379.0	0.315	0.295	0.267	0.247	0.255	0.236
	2BT	297.4	0.347	**0.323**	0.267	0.247	0.255	0.236
	5BT	256.6	**0.350**	0.311	0.295	0.277	**0.258**	**0.245**
	10BT	**240.9**	0.325	0.312	**0.300**	**0.283**	0.242	0.234
120M	0.1BT	1241.0	0.194	0.180	0.183	0.171	0.199	0.173
	0.5BT	343.2	0.322	0.302	0.279	0.260	0.286	0.266
	1BT	253.8	0.376	0.348	0.302	0.272	**0.314**	**0.294**
	2BT	199.6	0.391	0.362	0.321	0.297	0.295	0.284
	5BT	167.7	**0.393**	**0.366**	0.343	0.327	0.293	0.281
	10BT	**147.4**	0.381	0.351	**0.357**	**0.331**	0.272	0.251

Text Base dataset is better than code Base dataset for continued pre-training Variants of continued pre-training settings with Base dataset of Fineweb 5 billion token sample achieve higher acceptance rate on HumanEval (Table 8) compared to their code Base dataset counterparts of TheStack-v2 5 billion token sample. This is likely because these models acquire the majority of their code domain knowledge during the initial pre-training stage, and the later introduction of text domain knowledge can cause gradient shifts that are sub-optimal for code. While including portions of code in the CP dataset helps mitigate this issue, the performance degrades significantly by up to 50% when the CP dataset lacks code.

Table 8: Continued pre-training effect on acceptance rate of code using greedy decoding

Base dataset	CP mix dataset code	CP mix dataset natural lang.	HumanEval $\gamma = 3$	HumanEval $\gamma = 5$
Fineweb 5BT sample	5B	-	**0.688**	**0.578**
	5B	1B	0.678	0.560
	5B	2.5B	0.672	0.561
The stack-v2 5BT sample	-	5B	0.320	0.238
	1B	5B	0.628	0.519
	2.5B	5B	0.667	0.553

Table 9: Continued pre-training effect on acceptance rate of natural language tasks using greedy decoding

CP Base dataset	CP dataset code	CP dataset natural lang.	CNN-DailyMail		Tinystories		Dolly	
			$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$
Fineweb 5BT sample	5B	-	0.314	0.216	0.249	0.164	0.210	0.139
	5B	1B	0.343	0.241	0.297	0.200	**0.245**	**0.163**
	5B	2.5B	**0.349**	**0.244**	**0.304**	**0.204**	0.244	**0.163**
The stack-v2 5BT sample	-	5B	0.340	0.238	0.276	0.182	0.220	0.144
	1B	5B	0.339	0.239	0.283	0.190	0.222	0.148
	2.5B	5B	0.344	0.243	0.285	0.191	0.227	0.151

B FastDraft additional results

Table 4 summerizes results of each stage of the FastDraft scheme using greedy sampling.

Table 10: Continued pre-training effect on acceptance rate of code using multinomial sampling with temperature $T = 0.6$

Base dataset	CP mix dataset code	CP mix dataset natural lang.	HumanEval $\gamma = 3$	HumanEval $\gamma = 5$
Fineweb 5BT sample	5B	-	**0.578**	**0.462**
	5B	1B	0.560	0.451
	5B	2.5B	0.561	0.450
The stack-v2 5BT sample	-	5B	0.238	0.161
	1B	5B	0.519	0.400
	2.5B	5B	0.553	0.416

Table 11: Continued pre-training effect on acceptance rate of natural language tasks using multinomial sampling with temperature $T = 0.6$

CP Base dataset	CP mix dataset code	CP mix dataset natural lang.	CNN-DailyMail		Tinystories		Dolly	
			$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$
Fineweb 5BT sample	5B	-	0.271	0.170	0.232	0.153	0.191	0.126
	5B	1B	0.296	0.210	0.279	0.187	**0.233**	0.148
	5B	2.5B	0.304	0.211	**0.287**	**0.192**	0.226	**0.149**
The stack-v2 5BT sample	-	5B	0.297	0.205	0.254	0.173	0.201	0.135
	1B	5B	0.307	0.201	0.268	0.172	0.205	0.141
	2.5B	5B	**0.312**	**0.213**	0.264	0.176	0.215	0.143

Table 12: Impact of pre-training on mixed datasets on acceptance rate of natural language and code tasks. Evaluated across window sizes 3 and 5 using greedy decoding

Mix dataset natural lang.	Mix dataset code	CNN-DailyMail		Tinystories		Dolly		HumanEval	
		$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$
5B	5B	0.354	0.249	0.284	0.189	0.205	0.136	0.639	0.532
6B	4B	0.355	0.248	0.284	0.190	**0.220**	**0.144**	0.656	**0.551**
7B	3B	0.354	0.248	0.285	0.191	0.216	**0.144**	**0.657**	0.550

Table 13: Impact of pre-training on mixed datasets on acceptance rate of natural language and code tasks. Evaluated across window sizes 3 and 5 using multinomial sampling with temperature $T = 0.6$

Mix data natural lang.	Mix data code	CNN-DailyMail		Tinystories		Dolly		HumanEval	
		$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$
5B	5B	0.311	0.206	**0.265**	0.175	0.194	0.124	0.506	0.389
6B	4B	0.317	**0.218**	0.261	**0.178**	0.201	0.137	0.530	0.411
7B	3B	**0.326**	0.211	0.258	0.176	**0.205**	**0.137**	**0.533**	**0.424**

Table 14: AR results for FastDraft stages, performed in subsequent order: Pre-training (PT), Continued Pre-training (CP) and Fine-tuning (FT) using greedy decoding

Model	Type	CNN-DailyMail		TinyStories		Dolly		HumanEval	
		$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 3$	$\gamma = 5$
Phi3-mini 50M	PT	0.350	0.246	0.295	0.196	0.258	0.174	0.312	0.221
	$PT \rightarrow CP$	0.349	0.244	0.304	0.204	0.244	0.163	**0.672**	**0.563**
	$PT \rightarrow CP \rightarrow FT$	**0.399**	**0.289**	**0.321**	**0.217**	**0.390**	**0.279**	0.663	0.553
Llama3.1 150M	PT	0.298	0.204	0.243	0.162	0.257	0.171	0.282	0.198
	$PT \rightarrow CP$	0.300	0.203	0.250	0.166	0.284	0.193	0.658	0.546
	$PT \rightarrow CP \rightarrow FT$	**0.327**	**0.228**	**0.271**	**0.181**	**0.350**	**0.247**	**0.700**	**0.593**