[OnlineSPEC] When Drafts Evolve: Speculative Decoding Meets Online Learning

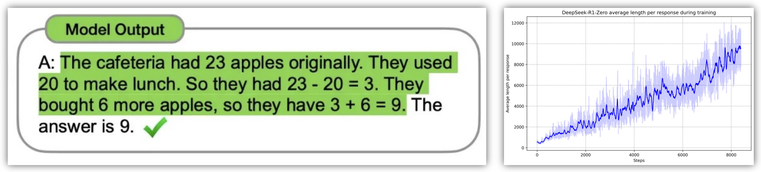
Yu-Yang Qian[1,2] **Hao-Cong Wu**[1,2] **Yichao Fu**[3] **Hao Zhang**[3] **Peng Zhao**[1,2*]

[1]National Key Laboratory for Novel Software Technology, Nanjing University
[2]School of Artificial Intelligence, Nanjing University [3]University of California, San Diego

ICML International Conference On Machine Learning
Seoul, South Korea

Background: LLM Inference

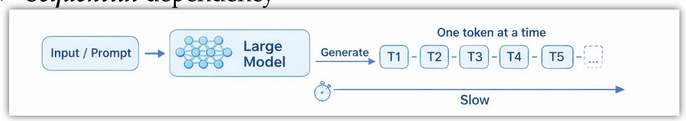
- Nowadays, many ***LLMs*** have been developed and used.
- Not only the model size, but also the ***inference cost***, is increasing.
  - Especially with ***Chain of Thought*** (CoT)
  - Recently emerged ***agentic LLMs*** further increase this burden



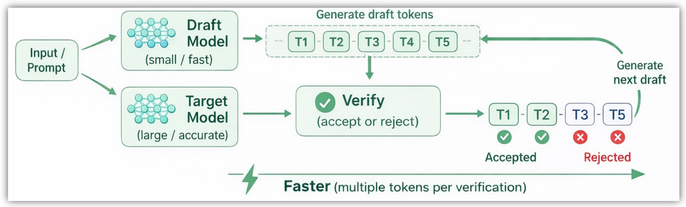
Inference cost and latency of LLMs lead to a decline in user experience.

Background: Speculative Decoding

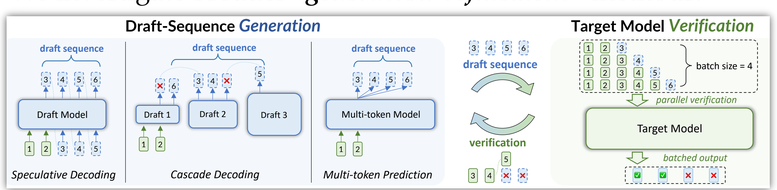
- Autoregressive (AR) models speedup bottleneck:
  - ***Sequential*** dependency



- LLM inference acceleration methods **Speculative Decoding**:
  - Use a small *"draft model"* to first quickly generate draft tokens;
  - Use a big *"target model"* to ***parallelly*** verify the drafts.

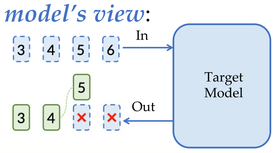


- We investigate broader "***generation-refinement***" methods:



- One round of speculative decoding in *draft model's view*:
  - Submits: k tokens
  - Gets: k labels denote accept or not

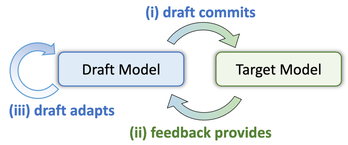
Goal: the more tokens accepted, the better



The "***acceptance rate***" of the *draft model* directly determines speedup.

Key Observation: Free Feedback can be Received

- ***Key observation:***
  - Draft model receives timely feedback ***at no cost!***
  - We can use this feedback to update draft model(s), exactly match **Online Learning.**



Naturally forms *"draft commits–feedback provides–draft adapts"* ***evolving loop.***

- Formulated as an ***online learning*** *problem:*

At each round $t = 1, \dots, T$:
(1) the learner submits draft model(s) $\mathbf{w}_t \in \mathcal{W}$; (draft model generating)
(2) target model reveals feedback $f_t(\cdot) : \mathcal{W} \mapsto \mathbb{R}$; (parallel verifying)
(3) the learner suffers loss $f_t(\mathbf{w}_t)$ and updates the draft model(s).

- Performance Metric: ***Dynamic Regret***

$$\mathbf{Reg}_T \triangleq \sum_{t=1}^{T} f_t(\mathbf{w}_t) - \sum_{t=1}^{T} f_t(\mathbf{w}_t^\star)$$

i.e., the ***cumulative gap*** between the algorithm and a sequence of time-varying comparators $\{\mathbf{w}_t^\star\}_{t=1}^T$.

By reducing this optimization problem to ***regret minimization***, we can leverage the rich toolkit from online learning.

Algorithm 1 OnlineSPEC Framework

Input: Initial draft model parameter $\mathbf{w}_0 \in \mathcal{W}$ with its predicted distribution $q_{\mathbf{w}_0}(\cdot \mid \mathbf{x})$, target model $\mathbf{v} \in \mathcal{V}$ with its predicted distribution $p_{\mathbf{v}}(\cdot \mid \mathbf{x})$, total steps T, candidate length k, input prompt $\mathbf{x}_0$.
Initialize: Input sequence $\mathbf{x} = \mathbf{x}_0$, draft model $\mathbf{w}_t = \mathbf{w}_0$.
for $t = 1, \dots, T$ **do**
  ▷ Generate the draft sequence:
  for $i = 1$ **to** k **do**
    $x_i \sim q_{\mathbf{w}_t}(x \mid \mathbf{x}_{<i})$, where $\mathbf{x}_{<i} = \{\mathbf{x}, x_1, \dots, x_{i-1}\}$.
  ▷ Verify by target model:
  Compute $\{p_{\mathbf{v}}(x_1|\mathbf{x}_{<1}), \dots, p_{\mathbf{v}}(x_k|\mathbf{x}_{<k})\}$ in parallel.
  ▷ Determine the accepted token length n_t:
  Sample $r_1 \sim U(0,1), \dots, r_k \sim U(0,1)$ uniformly;
  $n_t \leftarrow \min(\{j-1 \mid 1 \le j \le k, r_j > \frac{p_{\mathbf{v}}(x_j|\mathbf{x}_{<j})}{q_{\mathbf{w}_t}(x_j|\mathbf{x}_{<j})}\} \cup \{k\})$.
  ▷ Verify and update the draft sequence:
  if $n_t < k$ **then**
    $p'(x) \propto \max\left(0, p_{\mathbf{v}}(x \mid \mathbf{x}_{<n_t+1}) - q_{\mathbf{w}_t}(x \mid \mathbf{x}_{<n_t+1})\right)$;
  else
    $p'(x) \leftarrow p_{\mathbf{v}}(x \mid \mathbf{x}_{<k+1})$.
  $x_{n_t+1} \sim p'(x)$; $\mathbf{x} \leftarrow \mathbf{x} + [x_1, \dots, x_{n_t}, x_{n_t+1}]$.
  ▷ Receive interactive feedback: observe the loss function $f_t : \mathcal{W} \mapsto \mathbb{R}$ from the target model $\mathbf{v}$.
  ▷ Update draft model: $\mathbf{w}_{t+1} \leftarrow \mathrm{Update}(f_t; \mathbf{w}_t)$ as Sec. 3.
Output: Token sequence $\hat{\mathbf{x}} \triangleq [x_1, x_2, \dots]$.

A Unified Framework: *OnlineSPEC*

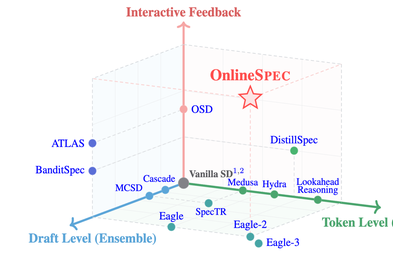
- ***Relationship*** between ***acceleration rate*** and ***regret***:

Theorem 1 (Acceleration Rate). Under Assumption 1, for OnlineSPEC with a total of T steps, the acceleration rate γ is bounded by

$$\frac{k+1}{(\alpha k+1)\left(1+(k+1)\sqrt{\mathrm{Reg}_T/(2T)}\right)} \le \gamma \le \frac{k+1}{\alpha k+1},$$

where $\alpha \triangleq \frac{a}{A} \ll 1$ is the ratio of the inference times between the draft model and the target model.

- ***OnlineSPEC*** framework can guide designs of new algorithms:
  - A ***unified framework*** to boost speculative decoding
  - Can leverage the rich toolkit from online learning



Can be ***seamlessly*** combined with existing methods to further enhance the acceleration rate.

- We demonstrate its generality through three ***instantiations***:

Base Method	+Feedback via	Instantiation	Acceleration rate
Lookahead Reasoning	Online gradient descent $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{w}_t - \eta \nabla f_t(\mathbf{w}_t)]$	*Online-LR*	$\gamma = \Omega\left(\frac{1}{\alpha k+1} \min\left\{k+1, \frac{T^{1/4}}{\sqrt{1+P_T}}\right\}\right)$
Hydra	Optimistic online learning $\mathbf{w}_t = \Pi_{\mathcal{W}}[\widehat{\mathbf{w}}_t - \eta \mathbf{h}_t]$; $\widehat{\mathbf{w}}_{t+1} = \Pi_{\mathcal{W}}[\widehat{\mathbf{w}}_t - \eta \nabla f_t(\mathbf{w}_t)]$	*Opt-Hydra*	$\gamma = \Omega\left(\frac{1}{\alpha k+1} \min\left\{k+1, \frac{\sqrt{T}}{(1+\delta_T)^{1/4}\sqrt{1+P_T}}\right\}\right)$
EAGLE/ EAGLE-3	Online ensemble learning $\mathbf{w}_t = \sum_{i=1}^N p_t^i \cdot \mathbf{w}_t^i$	*Ens-EAGLE*	$\gamma = \Omega\left(\frac{1}{\alpha k+1} \min\left\{k+1, \frac{T^{1/4}}{(1+P_T)^{1/4}}\right\}\right)$

Experimental Results: Speedup of our OnlineSPEC framework

		GSM8K		Spider		Code-Search		Alpaca-Finance	
		AvgLen ↑	SpeedUp ↑	AvgLen ↑	SpeedUp ↑	AvgLen ↑	SpeedUp ↑	AvgLen ↑	SpeedUp ↑
Target model: *lmsys / Vicuna-7B-v1.3*									
	Vanilla SD	1.25±0.01	1.00×	1.20±0.03	1.00×	1.22±0.02	1.00×	1.20±0.01	1.00×
	OSD	**1.52±0.03**	**1.23×**	**1.36±0.02**	**1.12×**	**1.37±0.02**	**1.09×**	**1.30±0.01**	**1.08×**
	Hydra	2.14±0.05	1.00×	2.65±0.31	1.00×	1.82±0.04	1.00×	1.78±0.01	1.00×
	OSD-Hydra	2.56±0.06	1.19×	3.11±0.31	1.11×	2.11±0.05	1.16×	2.19±0.03	1.25×
(OnlineSPEC)	**Opt-Hydra**	**2.69±0.08**	**1.26×**	**3.27±0.32**	**1.18×**	**2.37±0.07**	**1.31×**	**2.70±0.03**	**1.55×**
	EAGLE	1.48±0.03	1.00×	1.31±0.03	1.00×	1.41±0.03	1.00×	1.39±0.01	1.00×
	OSD-EAGLE	1.92±0.04	1.28×	1.53±0.03	1.21×	1.54±0.04	1.07×	1.51±0.02	1.09×
(OnlineSPEC)	**Ens-EAGLE**	**2.01±0.05**	**1.41×**	**1.60±0.03**	**1.32×**	**1.58±0.04**	**1.15×**	**1.61±0.04**	**1.14×**
	EAGLE-3	1.78±0.03	1.00×	1.85±0.07	1.00×	1.67±0.04	1.00×	1.62±0.01	1.00×
	OSD-EAGLE-3	2.07±0.04	1.20×	2.10±0.05	1.07×	1.97±0.04	1.13×	2.00±0.04	1.16×
(OnlineSPEC)	**Ens-EAGLE-3**	**2.16±0.03**	**1.26×**	**2.24±0.07**	**1.12×**	**2.01±0.07**	**1.18×**	**2.07±0.04**	**1.27×**
Target model: *meta-llama / Llama-2-7B-Chat*									
	Vanilla SD	1.25±0.01	1.00×	1.07±0.01	1.00×	1.09±0.01	1.00×	1.20±0.01	1.00×
	OSD	**1.40±0.03**	**1.06×**	**1.47±0.05**	**1.22×**	**1.34±0.02**	**1.26×**	**1.31±0.01**	**1.12×**
	Hydra	1.52±0.02	1.00×	1.48±0.08	1.00×	1.66±0.01	1.00×	1.64±0.01	1.00×
	OSD-Hydra	2.55±0.03	1.67×	3.10±0.18	1.91×	2.39±0.04	1.43×	2.25±0.03	1.36×
(OnlineSPEC)	**Opt-Hydra**	**2.94±0.03**	**1.90×**	**3.28±0.16**	**2.03×**	**2.71±0.05**	**1.61×**	**2.78±0.05**	**1.68×**
	EAGLE	1.19±0.01	1.00×	1.15±0.03	1.00×	1.17±0.01	1.00×	1.14±0.01	1.00×
	OSD-EAGLE	1.45±0.01	1.18×	1.72±0.06	1.46×	1.47±0.02	1.28×	1.43±0.01	1.20×
(OnlineSPEC)	**Ens-EAGLE**	**1.58±0.01**	**1.33×**	**1.82±0.08**	**1.61×**	**1.62±0.01**	**1.46×**	**1.56±0.02**	**1.41×**
	EAGLE-3	1.82±0.03	1.00×	1.61±0.03	1.00×	1.69±0.04	1.00×	1.70±0.02	1.00×
	OSD-EAGLE-3	2.25±0.02	1.23×	2.42±0.12	1.43×	2.09±0.10	1.27×	2.13±0.02	1.23×
(OnlineSPEC)	**Ens-EAGLE-3**	**2.33±0.02**	**1.27×**	**2.54±0.14**	**1.57×**	**2.18±0.10**	**1.33×**	**2.15±0.03**	**1.26×**

🚀 Up to **24%** ***speedup*** over previous SOTA methods.

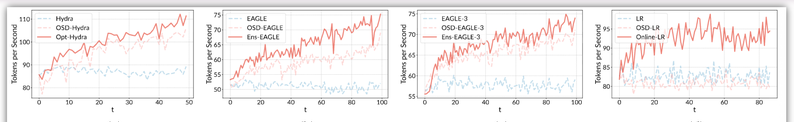


Figure 3. Evolution of tokens per second (TPS) on the *GSM8K* dataset: (a) *Opt-Hydra* with *lmsys/Vicuna-7B-v1.3*, (b) *Ens-EAGLE* with *lmsys/Vicuna-7B-v1.3*, (c) *Ens-EAGLE-3* with *lmsys/Vicuna-7B-v1.3*, and (d) *Online-LR* with *Qwen/Qwen3-8B*. This demonstrates consistent performance improvements via online learning, validating the effectiveness of our OnlineSPEC during deployment.

Continuously improving draft model and inference efficiency.

Model	Method	SpeedUp ↑	AvgLen ↑	Acc (%)
Vicuna-13B	Standard AR	1.00 (40.40)	1.00	–
	Vanilla SD	1.20 (48.52)	1.93	–
	OSD-EAGLE-3	1.37 (55.52)	2.24	–
	Ens-EAGLE-3	**1.46** (59.03)	**2.35**	–
Vicuna-13B	Standard AR	1.00 (40.40)	1.00	–
	Vanilla SD	1.50 (60.61)	2.22	–
	OSD-Hydra	1.69 (68.36)	2.50	–
	Opt-Hydra	**1.84** (74.37)	**2.73**	–
Qwen3-32B	Standard AR	1.00 (35.53)	1.00	96.08
	LR	1.11 (39.39)	3.66	95.76
	OSD-LR	1.09 (38.63)	3.18	95.44
	Online-LR	**1.31** (46.71)	**9.98**	95.68

- ***OnlineSPEC*** can be scaled to larger models.
- Still ***outperforms*** other methods in 13B/32B scale.

Scaling to ***larger*** target model.