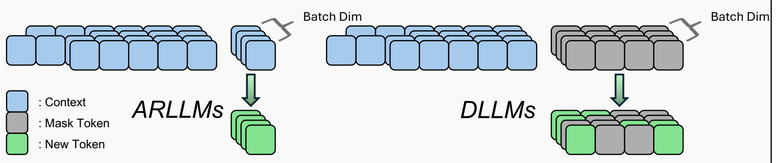
FOCUS: DLLMs Know How to Tame Their Compute Bound

Kaihua Liang[1], Xin Tan[2], An Zhong[1], Hong Xu[2], Marco Canini[1] | [1]*KAUST* [2]*CUHK*

Contact & Collaboration:
kaihua.liang@kaust.edu.sa
marco@kaust.edu.sa
Page:

Background & Motivation

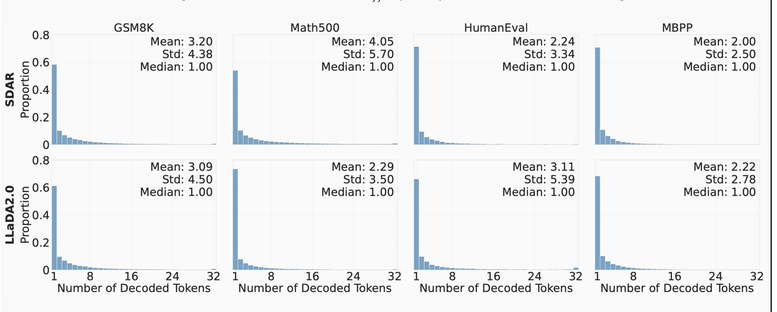


Diffusion LLMs offer a promising alternative paradigm to autoregressive LLMs. Yet, they represent a **computationally intensive workload**, even with block diffusion paradigm. Thus, their potential is limited by their **inference efficiency**.

FLOPs Formula forward of a layer of LLM:

$$\boldsymbol{Q} \cdot \left(8h^2 + 4d_{ff}h + 4hL\right)$$

Q: *Queries Length*, h: *Hidden Dimension*, d_{ff}: *Up − Projection Width*, L: *Context Length*

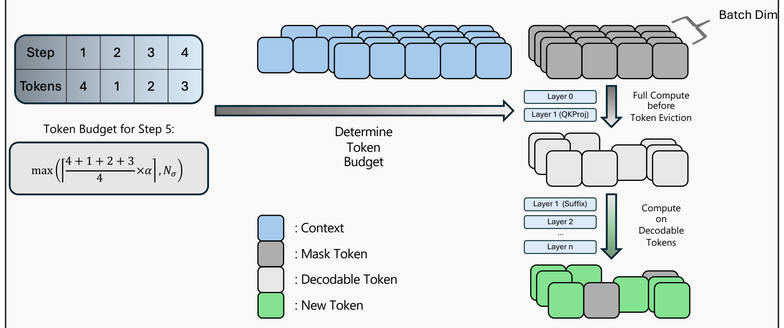


Decoding only a small portion per step **-> severe redundant computation**.

System Design

Comparison of DLLM inference paradigms. FOCUS further enhances the inference efficiency by predicting decodable tokens to eliminate computation redundancy.

Method	Parallel Decoding	KV Cache	Decodable Prediction	Inference Efficiency
LLaDA	✗	N/A	✗	★
Fast-dLLM	✓	Approx.	✗	★★
SDAR / LLaDA2.0	✓	Exact	✗	★★★
FOCUS (Ours)	✓	Exact	✓	★★★★



We propose the **FOCUS** system, being implemented on top of LMDeploy with customized Triton kernels, which is consists of two key components:

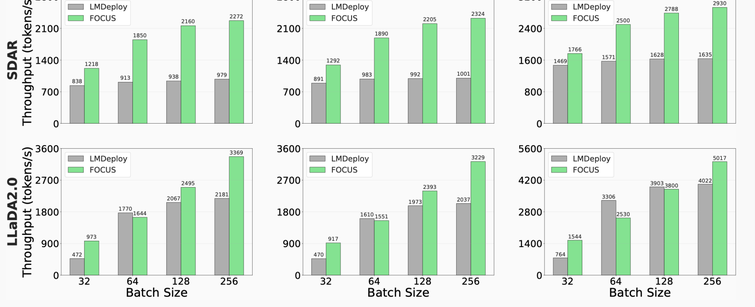
- **Importance-Based Token Eviction**: Preserve a dynamic budget of tokens with the highest importance delta in the first two layers.
- **Delayed Cache with Neighbor Awareness**: Inside processing block, a token's KV are cache in the forward after both itself and its right neighbor are decoded.

Evaluation

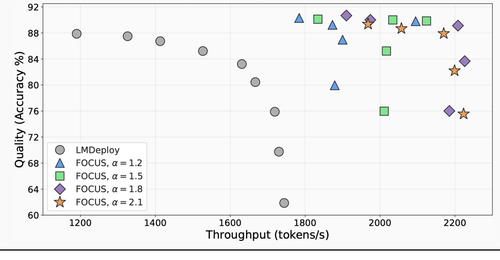
Improved generation quality under same confidence threshold.

Conf.	M	α	GSM8K	Math500	HumanEval	MBPP	IFEval
0.9	B	-	89.20	**64.70**	69.82	56.81	57.45
	F	1.2	89.84	64.60	**72.56**	58.17	**60.97**
	F	1.5	90.15	64.30	69.51	**61.09**	60.87
	F	1.8	**90.75**	63.80	71.34	58.95	60.11
0.8	B	-	87.57	59.30	65.85	50.78	58.69
	F	1.2	**90.33**	**64.10**	67.68	58.56	59.91
	F	1.5	89.73	62.20	69.21	**59.73**	60.97
	F	1.8	89.39	62.10	**69.82**	56.81	**61.14**
0.7	B	-	84.65	54.60	59.76	50.58	58.47
	F	1.2	89.24	62.00	**68.29**	**56.81**	61.91
	F	1.5	**89.31**	**62.20**	64.33	55.25	**62.26**
	F	1.8	88.03	59.80	64.33	53.89	60.08

Improved end-to-end throughput.



Improved quality-throughput tradeoff frontier.



Key Insight

Question: Is there a way to focus computation on decodable tokens?

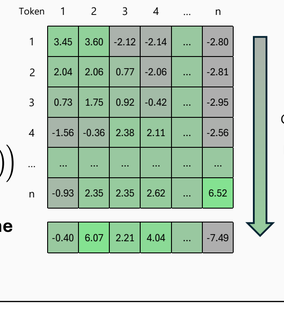
Token Importance Delta

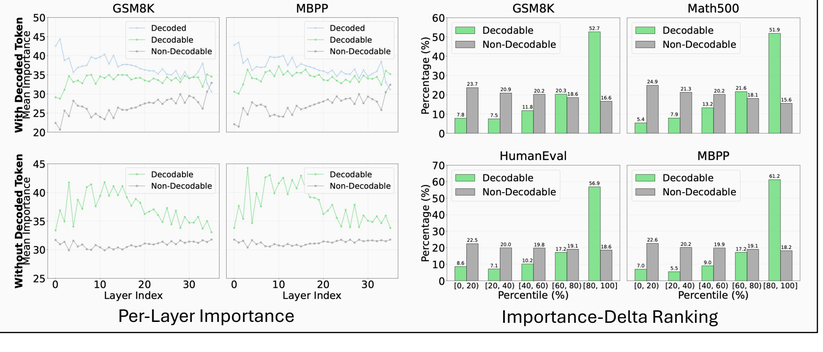
∝

Decoding Probability:

$$\Delta I_j = \Delta_{L1-L0} \sum_{i,h} Softmax\left(MaxPool1D\left(S_{i,j}^{(h)}\right)\right)$$

, where $S = \frac{\left(q_i^{(h)}\right)^\top k_j^{(h)}}{\sqrt{d_k}}$**,** $i, j \in \{1, \ldots, B\}$ **index the the intra-block** B **tokens.**





Conclusion

We propose FOCUS to revel that early-layer attention patterns serve as a predictor for token decodability in DLLMs, achieving up to 2.32× speed up and 3.52× on large blocks.

We hope to inspire future research with precise-compute, flexible-order parallel decoding.