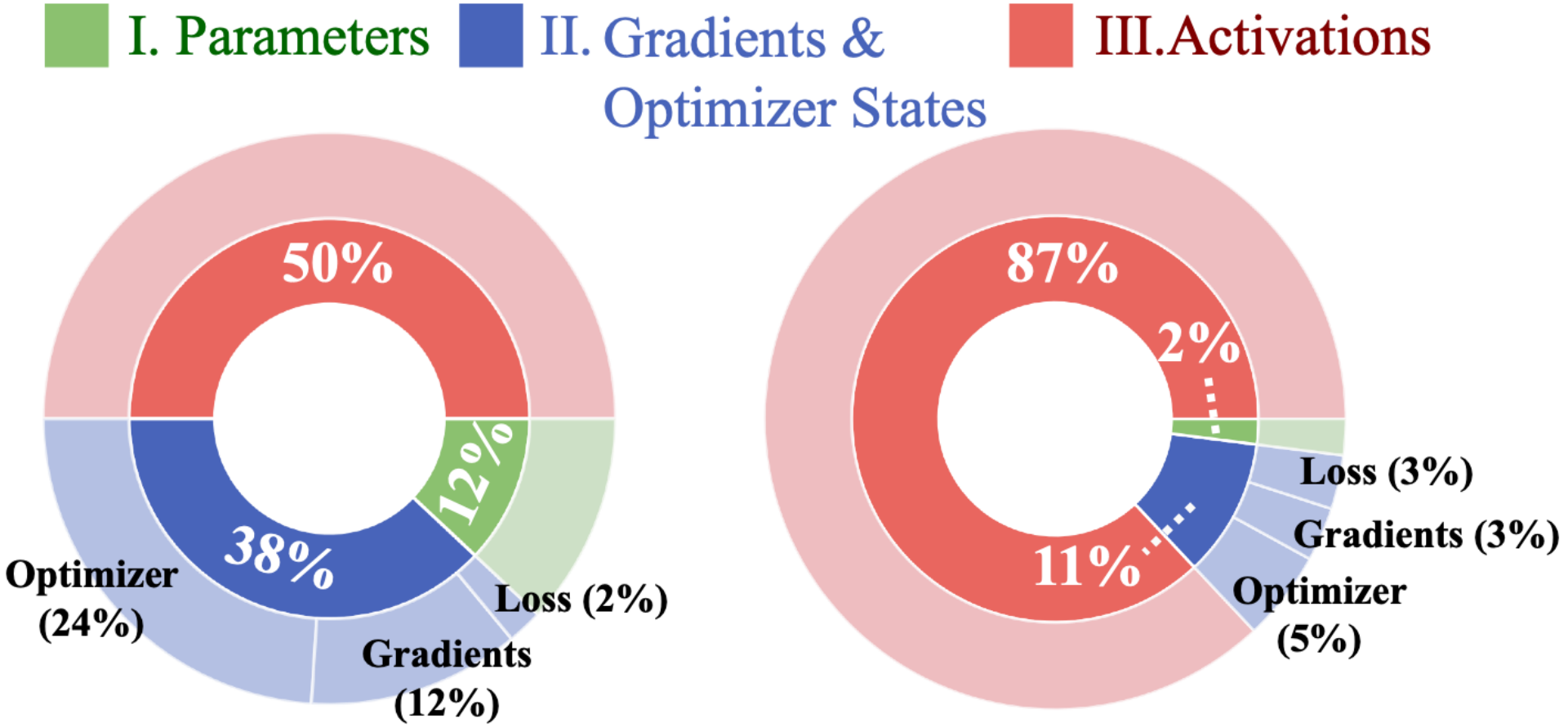
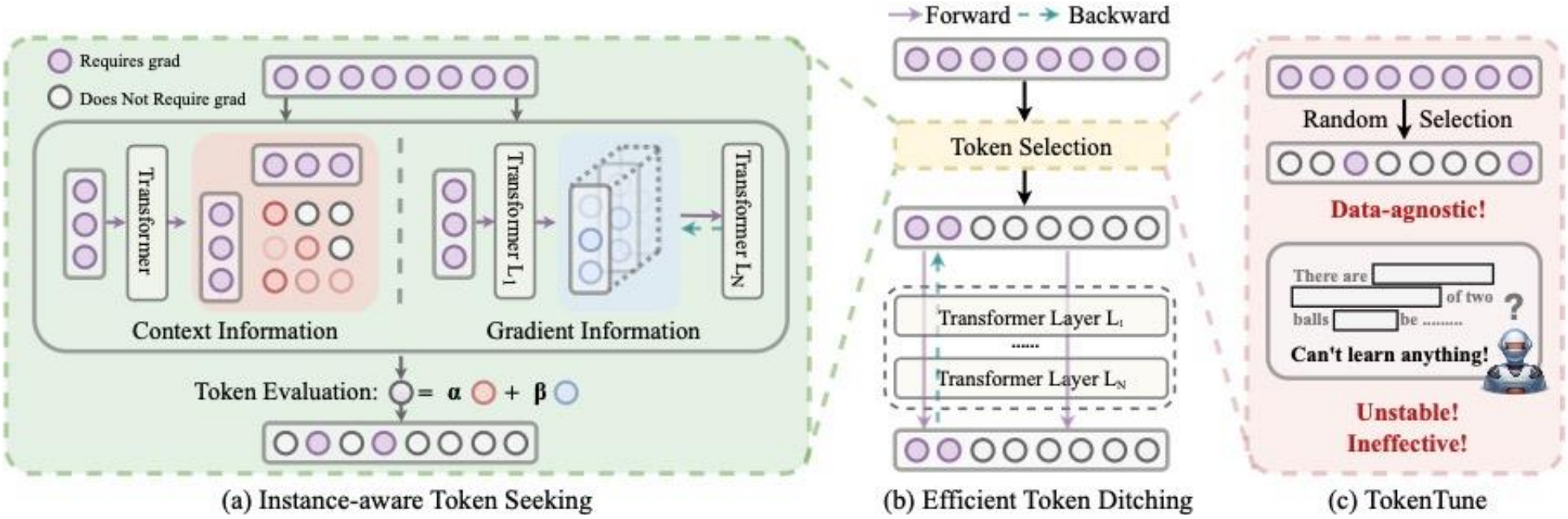
TokenSeek: Memory Efficient Fine Tuning via Instance-Aware Token Ditching

Runjia Zeng

RIT



(a) Training Memory Breakdown of Llama3 8B



(a) Instance-aware Token Seeking (b) Efficient Token Ditching (c) TokenTune

Instance-Aware Token Seeking

TokenSeek first leverages context and gradient information at the token level to evaluate and score individual tokens, selectively retaining more informative ones to mitigate performance degradation and fluctuation.

➤ **Context Information**:

Attention map in the last layers

$$\mathbf{A} = \mathrm{softmax}\Big(\mathrm{mask}\big(\mathbf{Q}\mathbf{K}^{\top}/\sqrt{d_k}\big)\Big).$$

The context importance of each token

$$I_1(t_j) = \sum_{i=1}^{n} \mathbf{A}_{ij}.$$

➤ **Gradient Information**:

Gradient matrix for the activations in the penultimate layer

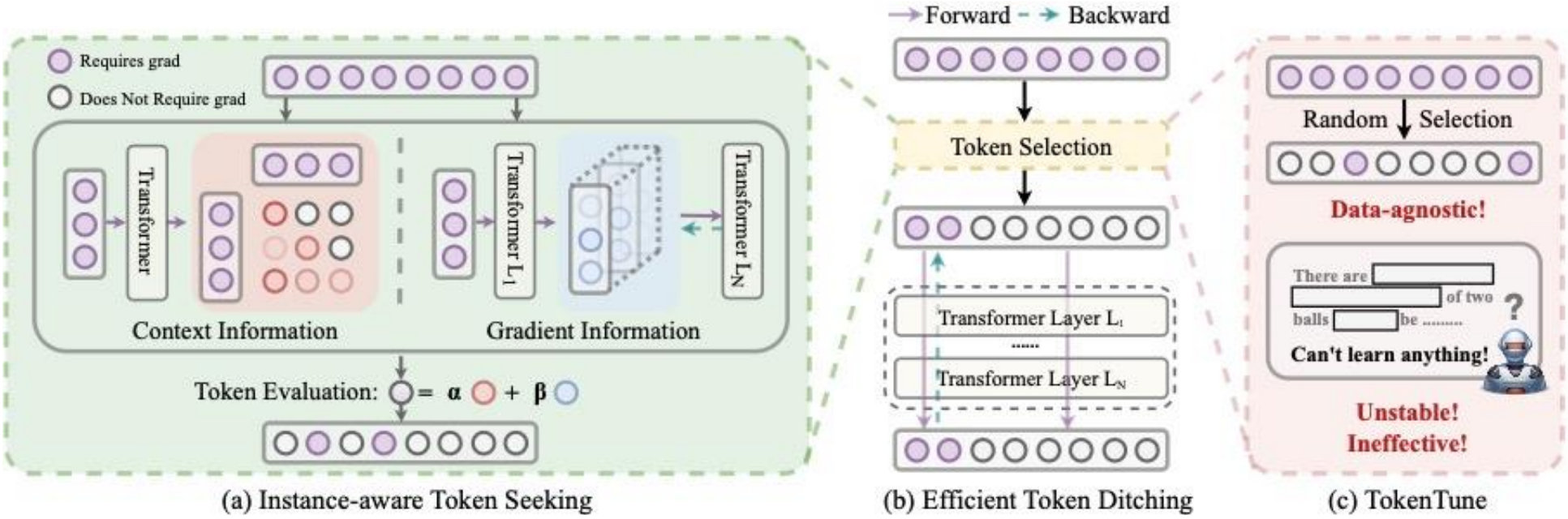
$$\mathbf{G} = \big[\partial \mathcal{L}/\partial z^{(L-1)}\big] \in \mathbb{R}^{n\times d}$$

The gradient-based importance of each token

$$I_2(t_j) = \mathrm{Accumulate}\Big[\frac{\partial \mathcal{L}}{\partial z^{(L-1)}}\Big], \quad \mathrm{Accumulate}[\cdot] = \sum_{k=1}^{d} \mathbf{G}_{jk}.$$

➤ **Token Evaluation**:

$$I(t_j) = \alpha \log\big[I_1(t_j)\big] + \beta\, \mathrm{Norm}\big[I_2(t_j)\big]$$



(a) Instance-aware Token Seeking (b) Efficient Token Ditching (c) TokenTune

Efficient Token Ditching

TokenSeek then significantly decreases the memory footprint for activations by updating model exclusively on selected tokens, thereby ditching the gradients of the others and thus eliminating these activations.

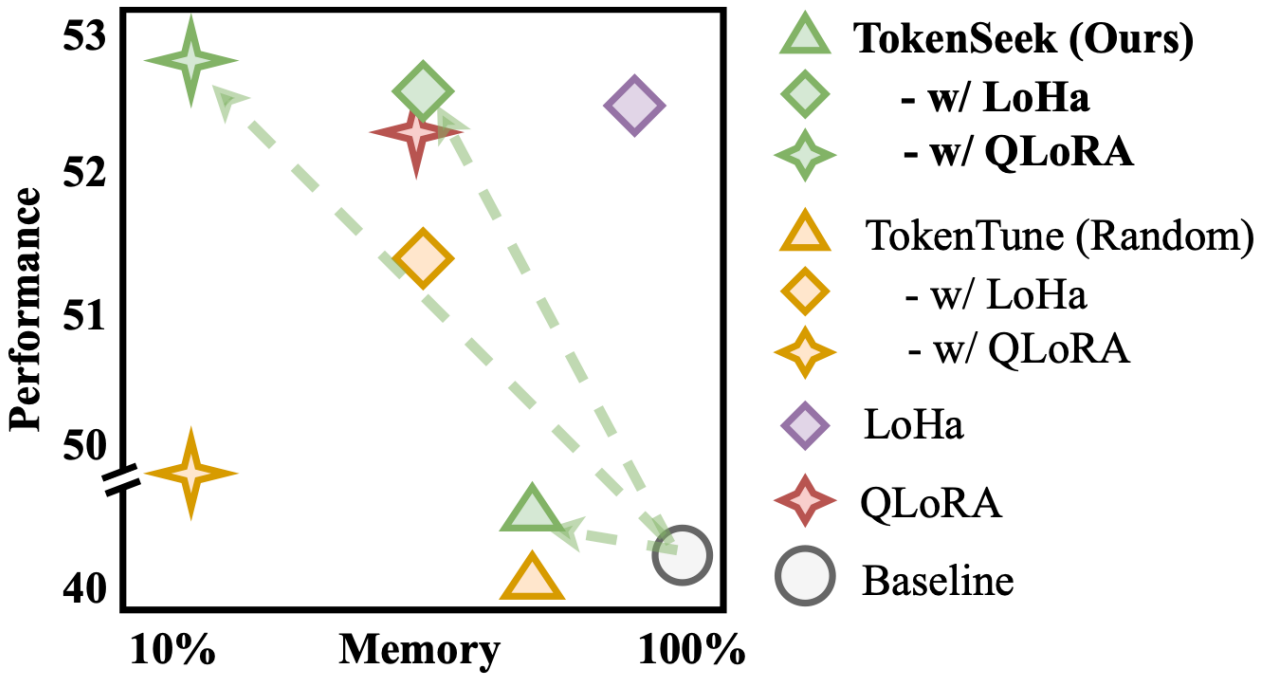
➤ **Originally:**

$$\frac{\partial \mathcal{L}}{\partial W^{(1)}} = \frac{\partial \mathcal{L}}{\partial z^{(L)}} \left(\prod_{\ell=2}^{L} \boxed{\frac{\partial z^{(\ell)}}{\partial z^{(\ell-1)}}} \right) \frac{\partial z^{(1)}}{\partial W^{(1)}}. \qquad \boxed{\frac{\partial z^{(l)}}{\partial z^{(l-1)}}} = \frac{\partial z^{(l)}}{\partial a^{(l)}} \frac{\partial a^{(l)}}{\partial z^{(l-1)}} = \sigma'\left(a^{(l)}\right) W^{(l)}.$$

➤ **TokenSeek:**

$$\boxed{\frac{\partial z^{(l)}}{\partial z^{(l-1)}}} = \left[\sigma'\left(a_t^{(l)}\right), \sigma'\left(a_{\tilde{t}}^{(l)}\right) \right] W^{(l)} = \left[\sigma'\left(a_t^{(l)}\right), \ 0 \right] W^{(l)}.$$

Method	Ave. Mem.	Max. Mem.	Average Score
Qwen2.5 (0.5B)			
Full Parameter/Token Tuning	100%	100%	48.43
- w/ TokenTune (Random)	48.3%	25.6%	39.88
- w/ **TokenSeek (Ours)**	48.3%	25.6%	41.94
IA3 (Liu et al., 2022)	84.3%	72.8%	48.01
LoRA (Hu et al., 2022a)	81.2%	71.8%	48.06
LoKr (Hyeon-Woo et al., 2021)	91.6%	79.3%	48.28
BOFT (Liu et al., 2024b)	145.1%	100.6%	48.10
Bone (Kang, 2024)	85.8%	76.2%	42.48
LoHa [1.33%] (Hyeon-Woo et al., 2021)	86.6%	76.9%	48.17
- w/ TokenTune (Random)	39.5%	22.5%	41.06
- w/ **TokenSeek (Ours)**	39.5%	22.5%	42.28
QLoRA [1.04%] (Dettmers et al., 2023)	51.7%	45.6%	47.66
- w/ TokenTune (Random)	19.2%	13.4%	46.34
- w/ **TokenSeek (Ours)**	**19.2%**	**13.4%**	**48.45**
Llama3.2 (1B)			
Full Parameter/Token Tuning	100%	100%	40.82
- w/ TokenTune (Random)	64.6%	34.3%	40.75
- w/ **TokenSeek (Ours)**	64.6%	34.3%	41.13
LoHa [0.63%]	92.3%	99.4%	52.28
- w/ TokenTune (Random)	45.9%	28.4%	51.37
- w/ **TokenSeek (Ours)**	45.9%	28.4%	52.58
QLoRA [0.52%]	45.6%	34.8%	52.13
- w/ TokenTune (Random)	14.8%	14.3%	49.22
- w/ **TokenSeek (Ours)**	**14.8%**	**14.3%**	**52.61**
Llama3.2 (3B)			
Full Parameter/Token Tuning	100%	100%	41.53
- w/ TokenTune (Random)	73.1%	39.3%	41.20
- w/ **TokenSeek (Ours)**	73.1%	39.3%	41.95
LoHa [0.47%]	90.7%	96.5%	60.01
- w/ TokenTune (Random)	49.6%	30.0%	59.13
- w/ **TokenSeek (Ours)**	49.6%	30.0%	**61.02**
QLoRA [0.42%]	33.6%	26.5%	60.39
- w/ TokenTune (Random)	13.3%	11.1%	59.31
- w/ **TokenSeek (Ours)**	**13.3%**	**11.1%**	60.42



(b) Performance-Memory Comparison

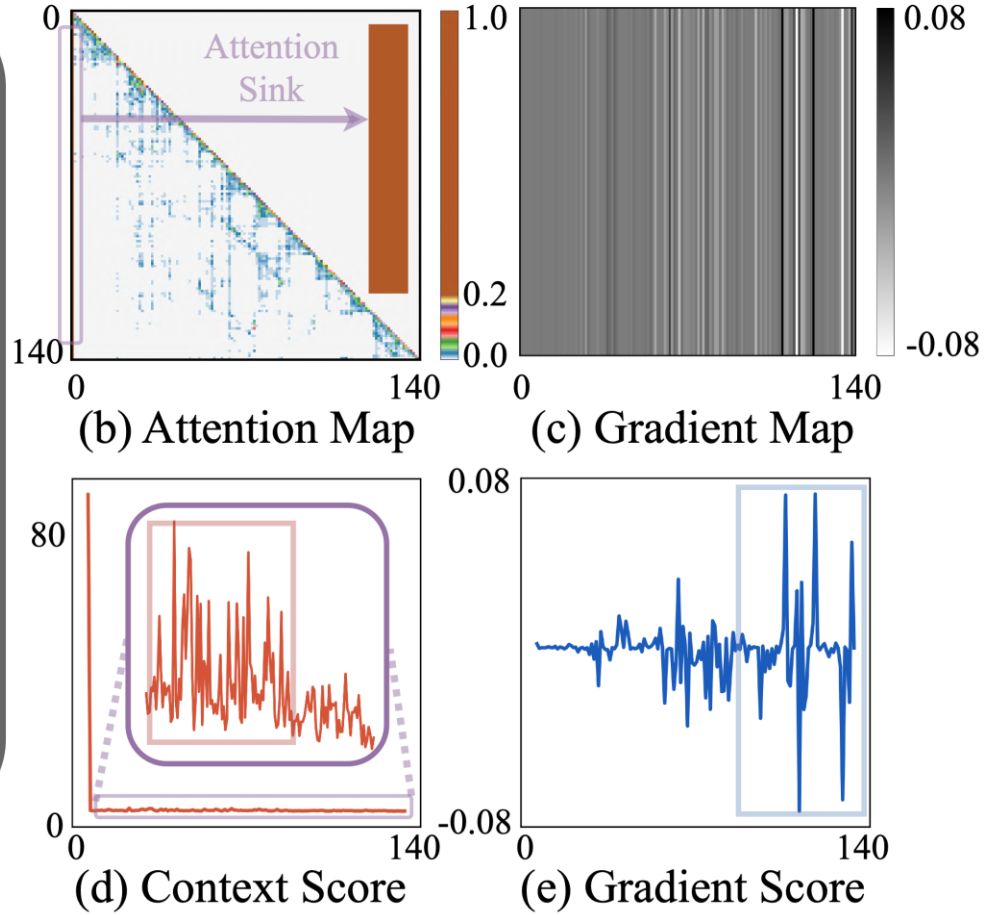
Selected Tokens by Context Score Selected Tokens by Gradient Score

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:
A box contains 5 white balls and 6 black balls. Two balls are drawn out of the box at random. What is the probability that they both are white?

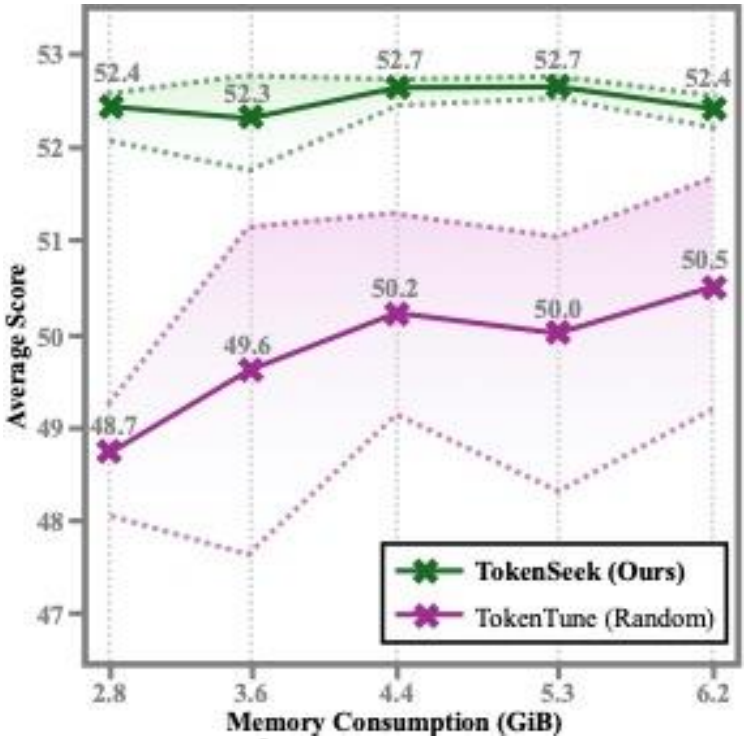
Response:
There are $\binom{11}{2} = 55$ combinations of two balls that can be drawn. There are $\binom{5}{2} = 10$ combinations of two white balls that can be drawn. So the probability that two balls pulled out are both white is $\dfrac{10}{55} = \dfrac{2}{11}$.

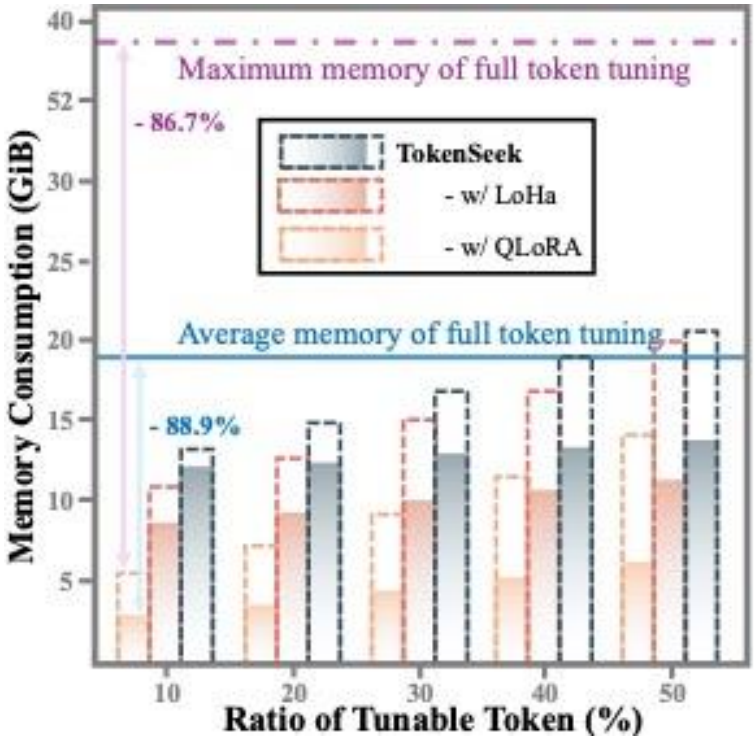
(a) Visualization of Token Seeking

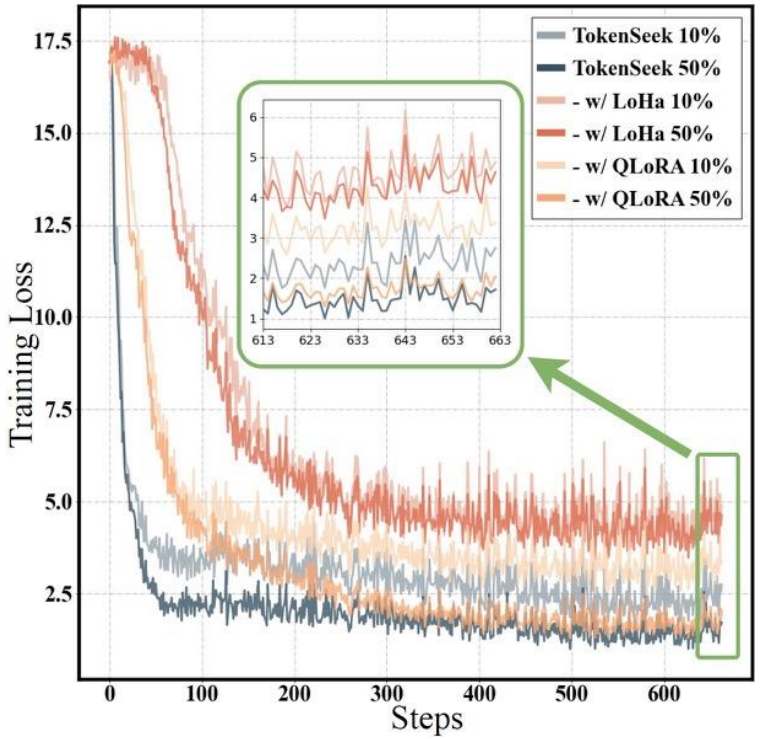


(b) Attention Map (c) Gradient Map

(d) Context Score (e) Gradient Score







✨*Effectiveness*

✨*Stability*

✨*MemoryControl*

✨*TokenSeek favors PEFT*

✨*Tokens contribute FT*

THANK YOU!

runjia.tech/iclr_tokenseek