SAS: Simulated Attention Score

Chuanyang Zheng, Jiankai Sun, Yihang Gao, Yuehao Wang, Peihao Wang, Jing Xiong, Liliang Ren, Hao Cheng, Janardhan Kulkarni, yelong shen, Zhangyang Wang, Mac Schwager, Anderson Schneider, Xiaodong Liu, Jianfeng Gao

NEURAL INFORMATION PROCESSING SYSTEMS

Simulated Attention Score

$$O_i = \sum_{c=1}^{H}\left(\sum_{j=1}^{i}\phi(\frac{x_iW_cx_j}{\sqrt{D}})x_jU_c\right)$$

1.The Transformer Performance with Hidden Size, Attention Head,

2.Attention Hidden Size Per head. Hidden Size Determine the maximum number of Attention Patterns.

3.Attention head Number Determine the Independence of Attention Pattern and Value Embedding.

4.Attention Hidden Size Per Head Determine the Independence of Attention Pattern and the matrix W_c

Example: x_i = [a b] x_j =[c d].--->**2-dimension**

Attention Pattern: 1) ac; 2)ad; 3) bc; 4) bd---**4 Attention Patterns**

W_1= [1 0]   W_2= [0 1] W_3= [0 0] W_4= [0 0]
[0 0]   [0 0]   [1 0]   [0 1]

With 4 attention head, the attention patterns are independently for the value embedding

If hidden size per head is 1: x_i = [w1*a+w2*b]
x_j =[w3*c+w4*d]
The Attention Patterns: w1*w3*ac; 2)w1*w4*ad; 3)w2*w3*bc; 4) w2*w4*bd
With Hidden Size Per Head as 1, appenratenly, the Attention Patterns are dependent

Method

Apply the convoltion operation on query, key and value embedding, mapping the shape from [B H T D] to [B H' T D].

Also apply Fully-Connect Operation to map key and value from [B H' T D] to [B H' T D']

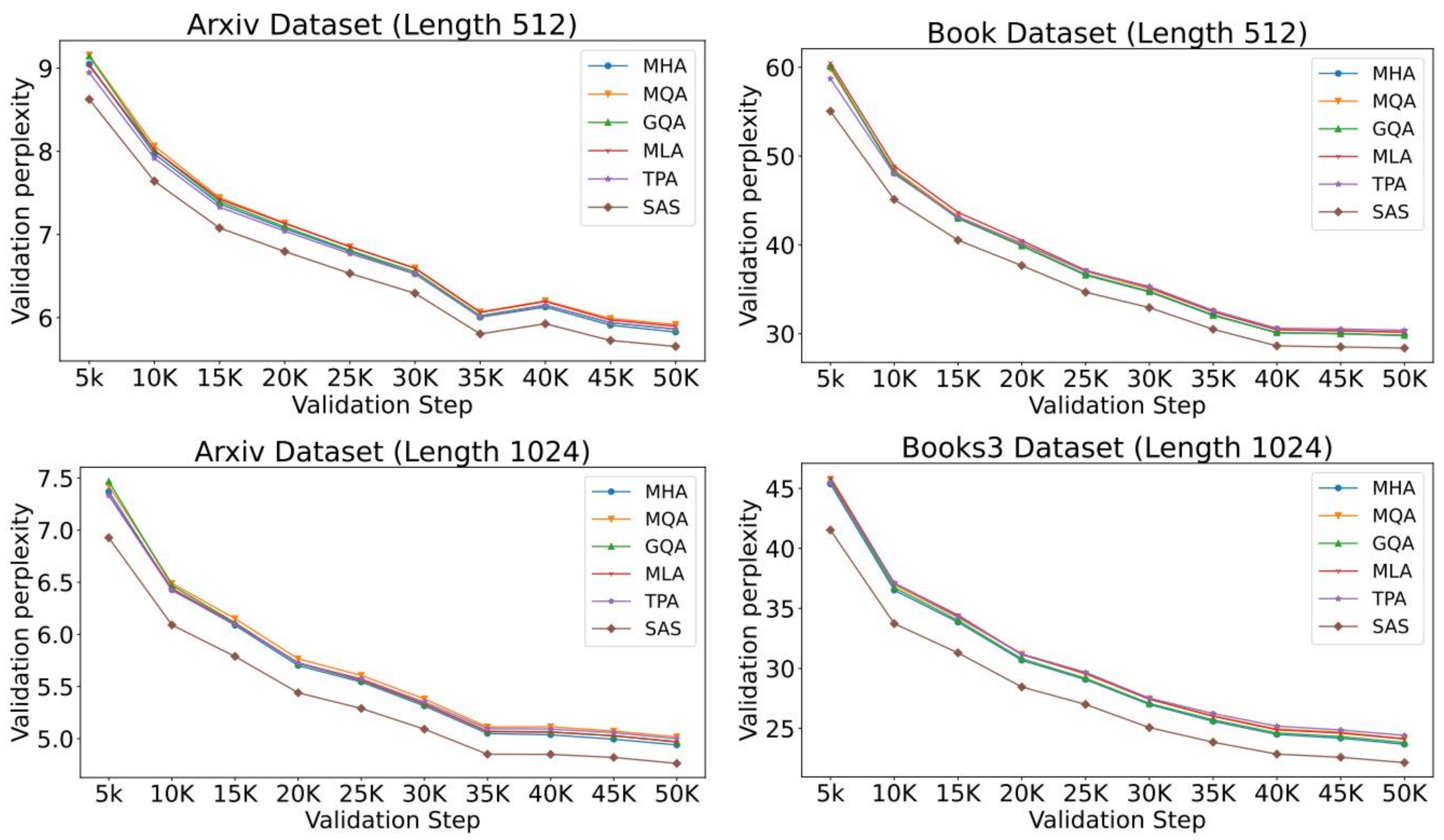
Finally, apply Parameter-Efficient Attention Aggregation

$$\begin{aligned}
\mathbf{Q}_0 &= \text{Reshape}(\mathbf{Q}, (B \cdot T \cdot H, D)),\\
\mathbf{Q}_1 &= \text{Linear}_1^{Q_f}(\mathbf{Q}_0),\\
\mathbf{Q}_2 &= \text{ReLU}(\mathbf{Q}_1),\\
\hat{\mathbf{Q}} &= \text{Linear}_2^{Q_f}(\mathbf{Q}_2) + \mathbf{Q}_1,
\end{aligned}$$

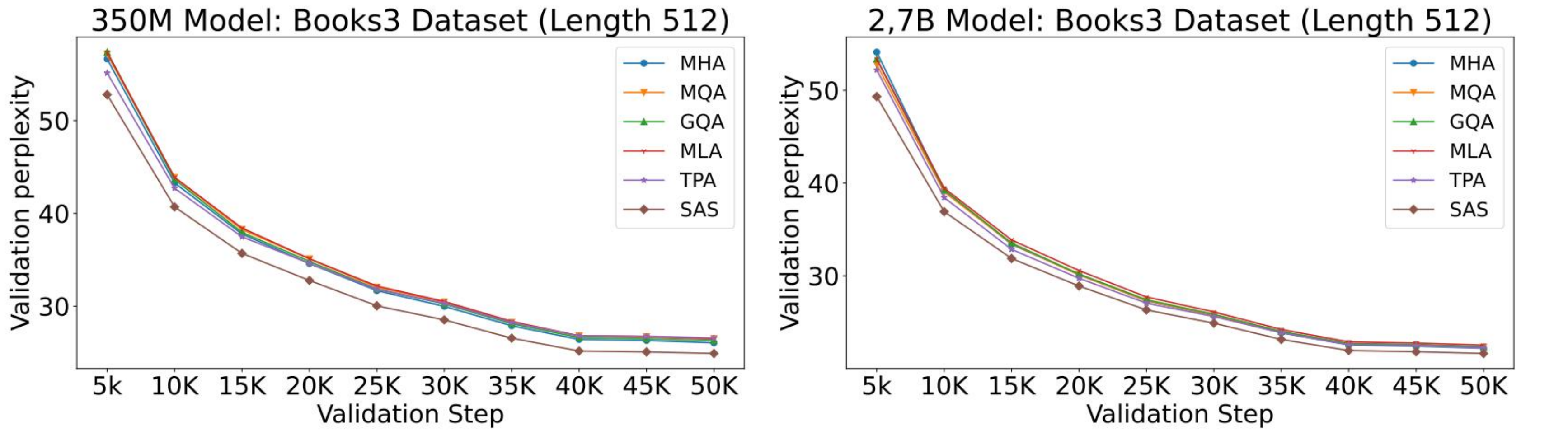
$$\begin{aligned}
\mathbf{Q}_1 &= \text{Conv}_1^{Q_h}(\mathbf{Q}_0),\\
\mathbf{Q}_2 &= \text{ReLU}(\mathbf{Q}_1),\\
\hat{\mathbf{Q}} &= \text{Conv}_2^{Q_h}(\mathbf{Q}_2) + \mathbf{Q}_1,
\end{aligned}$$

$$\textbf{Output} = \frac{1}{\hat{H}/H}\sum_{i=1}^{\hat{H}/H}\text{Concat}(\mathbf{h}_{(i-1)\times H+1}, \cdots, \mathbf{h}_{i\times H})\mathbf{W}^O,$$

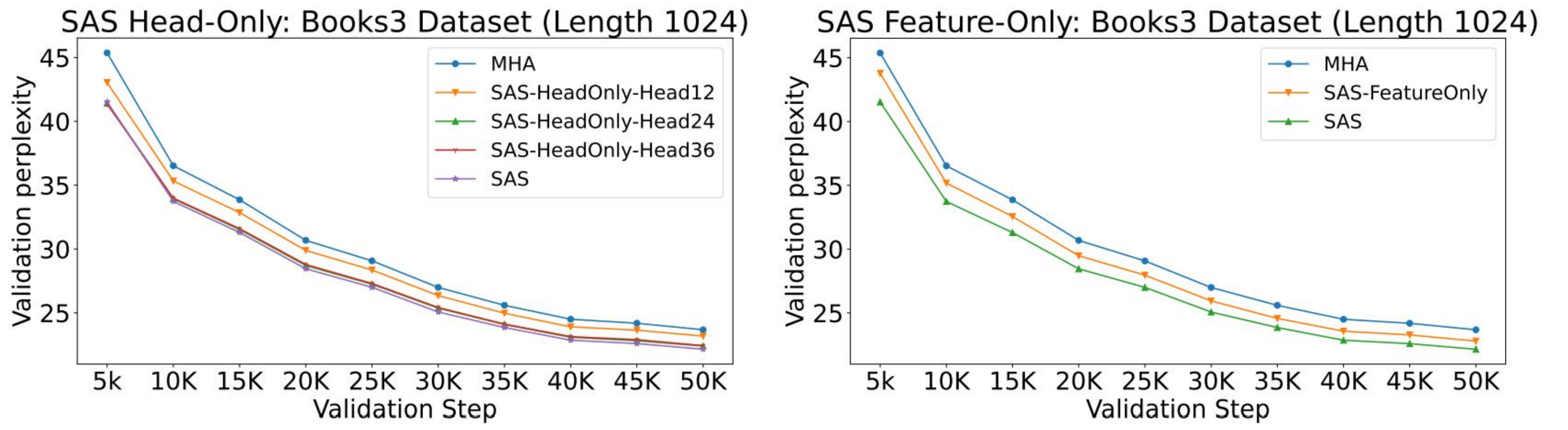
Comparisons with baselines



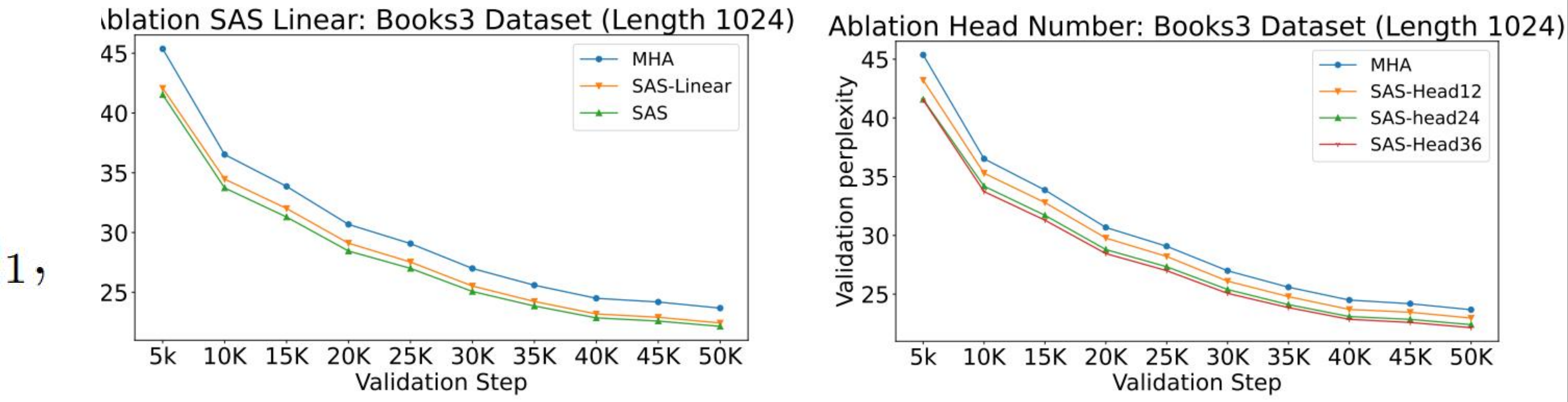
On Large Model Size



The Effect of Head and Feature Simulation



The Effect of Non-Linear and Head Number



Experiments on Downstream Task

Model Metric	ARC-E acc_n	ARC-C acc_n	Hellaswag acc_n	PIQA acc_n	ScIQ acc_n	SocialIQA acc	Winograde acc	Avg
350M params / 50B tokens								
MHA	56.57	29.01	45.45	69.10	76.40	**40.94**	52.96	52.92
MQA	58.12	30.55	45.71	69.10	78.70	39.25	51.46	53.27
GQA	57.62	30.38	46.04	68.99	76.20	39.56	53.28	53.15
MLA	57.15	28.92	44.77	67.95	75.90	38.89	53.67	52.46
TPA	59.09	**32.08**	46.51	69.53	76.20	39.82	53.12	53.76
SAS	**60.44**	31.66	**47.79**	**70.67**	**80.70**	40.07	**54.14**	**55.07**
350M params / 10B tokens								
MHA	54.29	27.99	39.94	66.38	74.00	37.67	53.59	50.55
MQA	54.45	28.75	40.51	66.21	73.10	37.92	51.30	50.32
GQA	53.03	27.73	40.34	66.59	74.20	38.95	51.22	50.29
MLA	53.28	27.65	39.51	66.10	**76.00**	38.08	52.09	50.39
TPA	52.44	27.99	41.27	67.57	72.60	38.33	53.35	50.51
SAS	**55.93**	**29.61**	**43.04**	**68.82**	75.90	**38.43**	**53.20**	**52.13**
760M params / 10B tokens								
MHA	56.57	29.01	45.45	67.25	77.20	38.54	52.96	52.43
MQA	55.85	29.95	43.63	67.85	76.20	39.30	53.35	52.32
GQA	54.21	29.35	44.40	68.34	77.70	38.38	52.17	52.08
MLA	57.15	29.10	44.77	67.95	75.90	38.89	53.67	52.49
TPA	58.12	30.89	44.71	69.37	77.90	39.30	52.80	53.30
SAS	**59.43**	**31.91**	**45.84**	**69.53**	**78.30**	**39.61**	**55.25**	**54.27**
1.5B params / 10B tokens								
MHA	57.87	31.40	45.51	68.82	75.90	38.18	52.33	52.86
MQA	55.85	31.74	46.40	69.53	77.30	38.13	54.70	53.38
GQA	57.62	30.20	46.20	69.48	76.30	38.95	53.59	53.19
MLA	57.24	29.95	44.90	68.50	75.20	39.76	53.43	52.71
TPA	59.81	31.23	46.84	68.99	75.60	39.46	**55.09**	53.86
SAS	**60.44**	**34.39**	**48.66**	**70.08**	**81.40**	**39.92**	54.93	**55.69**