Paper Deep Dive / NeurIPS 2025 Best Paper, Oral

Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free

Apr 11, 2026

Dongwon Lee

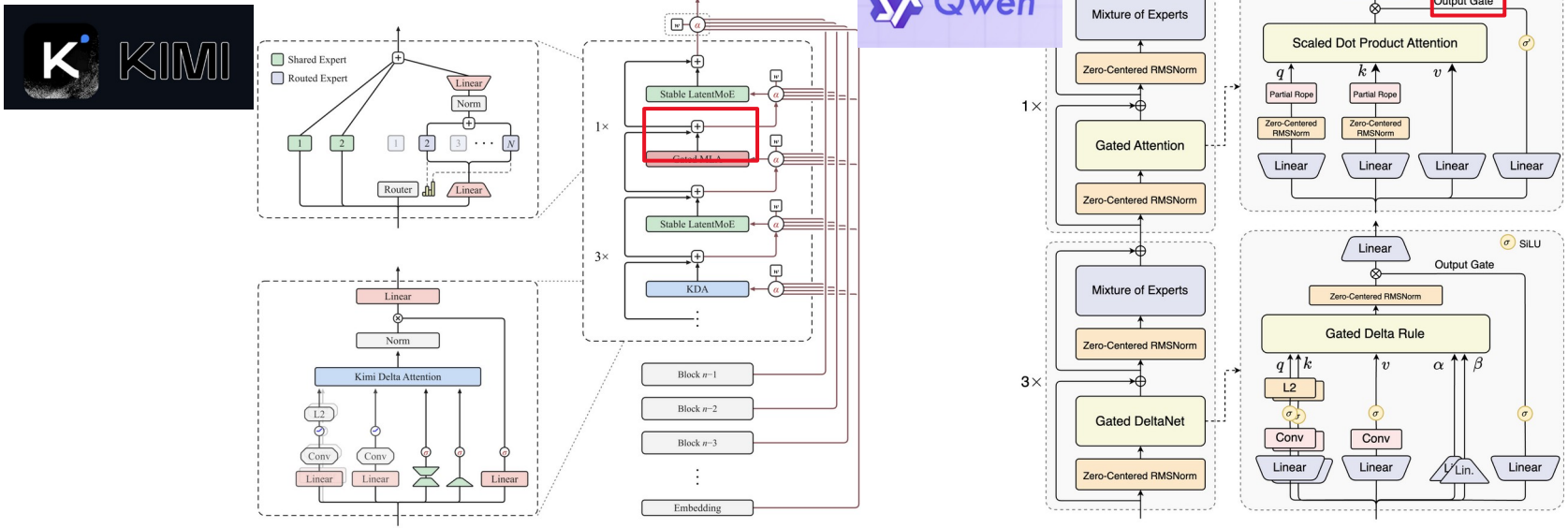
Agenda

- Preliminaries
  - Attention
  - Gating Mechanisms
- Experiment
- Analysis – Why gated attention works?
- Conclusion
- Limitations

Preliminary

Motivation

-
  - LLM(Qwen 3.5/Qwen-Next, Kimi K3) Gated Attention
  - Attention Sink
    - softmax attention Attention
    - Quantizable Transformer



Preliminary

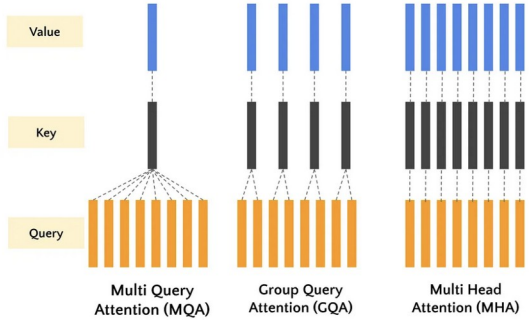
What is Gated Attention?

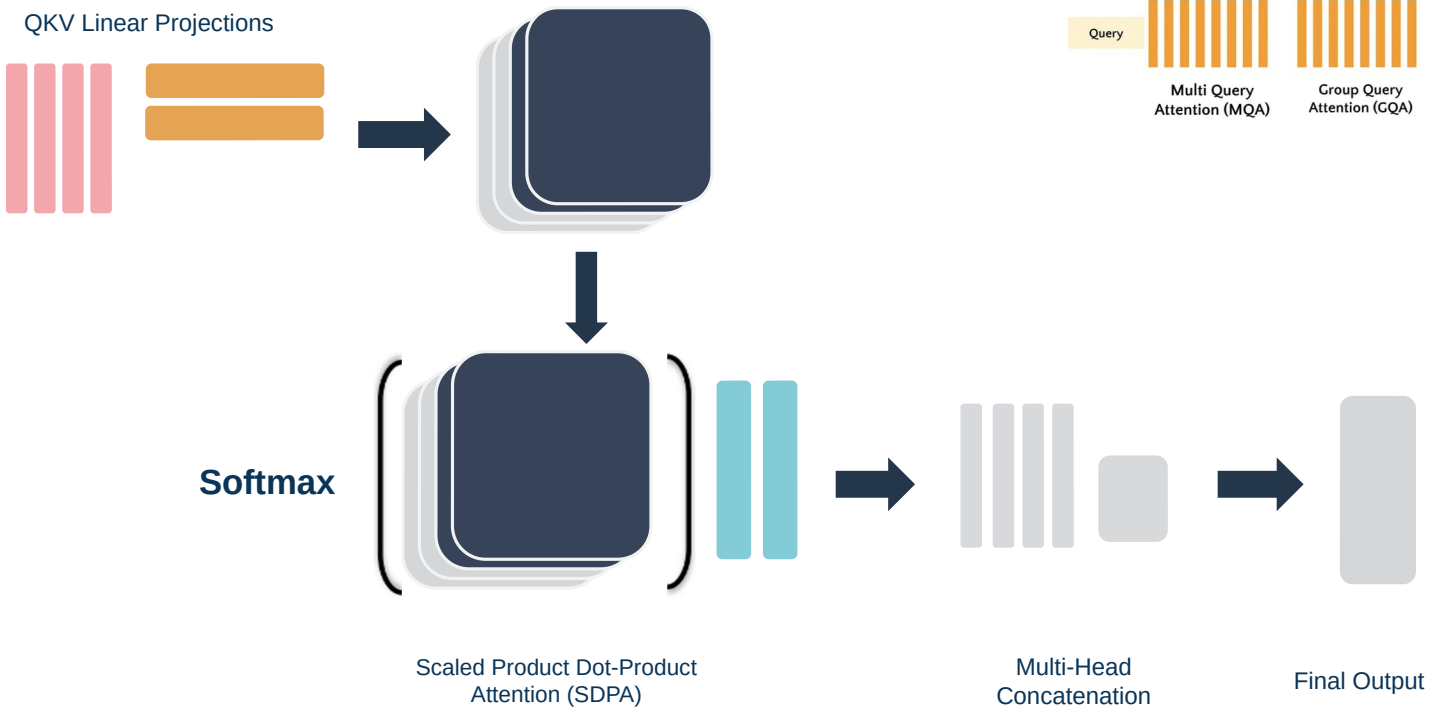
Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free

Zihan Qiu[*1], **Zekun Wang**[*1], **Bo Zheng**[*1], **Zeyu Huang**[*2],
Kaiyue Wen[3], **Songlin Yang**[4], **Rui Men**[1], **Le Yu**[1], **Fei Huang**[1], **Suozhi Huang**[5],
Dayiheng Liu[✉1], **Jingren Zhou**[1], **Junyang Lin**[✉1]
[1]Qwen Team, Alibaba Group [2]University of Edinburgh [3]Stanford University
[4]MIT [5]Tsinghua University

Preliminary

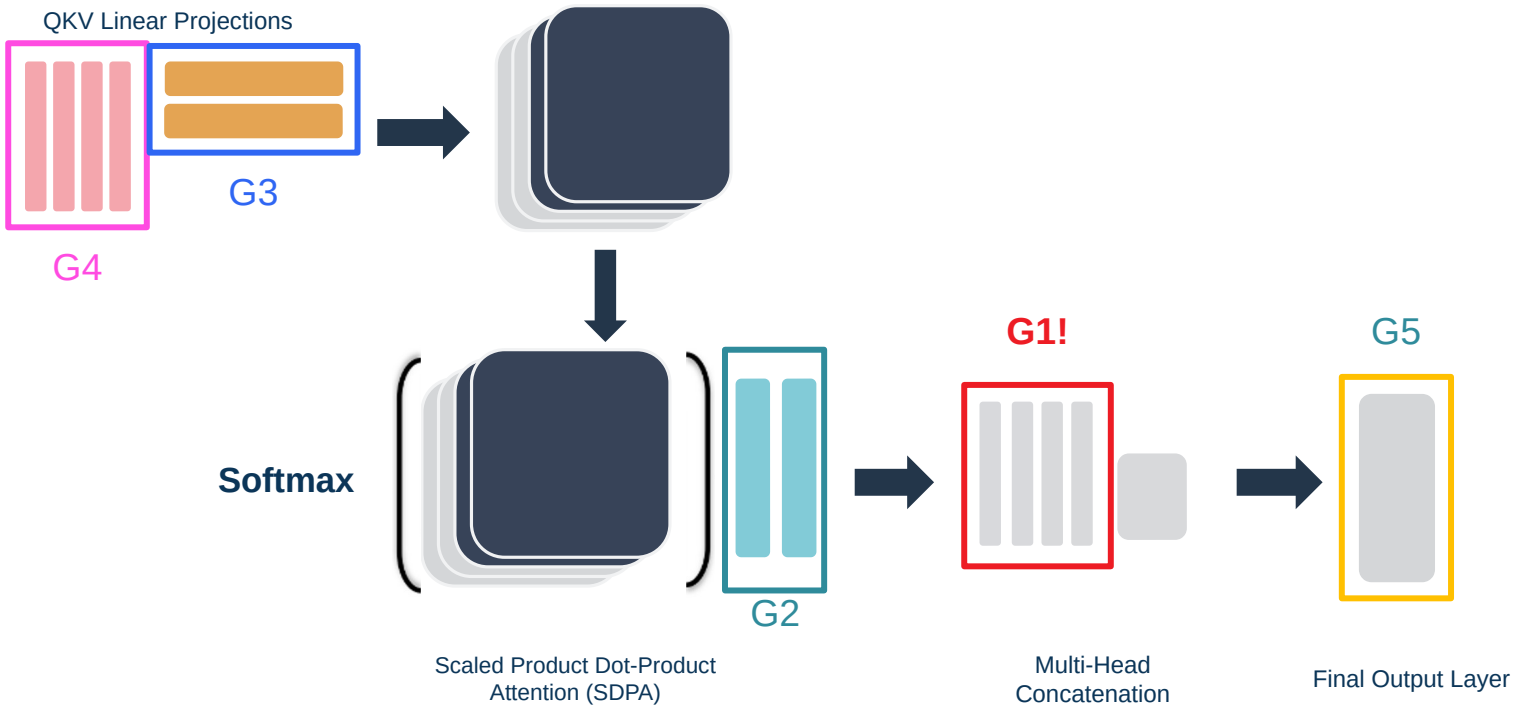
Softmax Attention + GQA(group query attention)





Preliminary

Softmax Attention + GQA(group query attention)



Preliminary

Gating Mechanisms

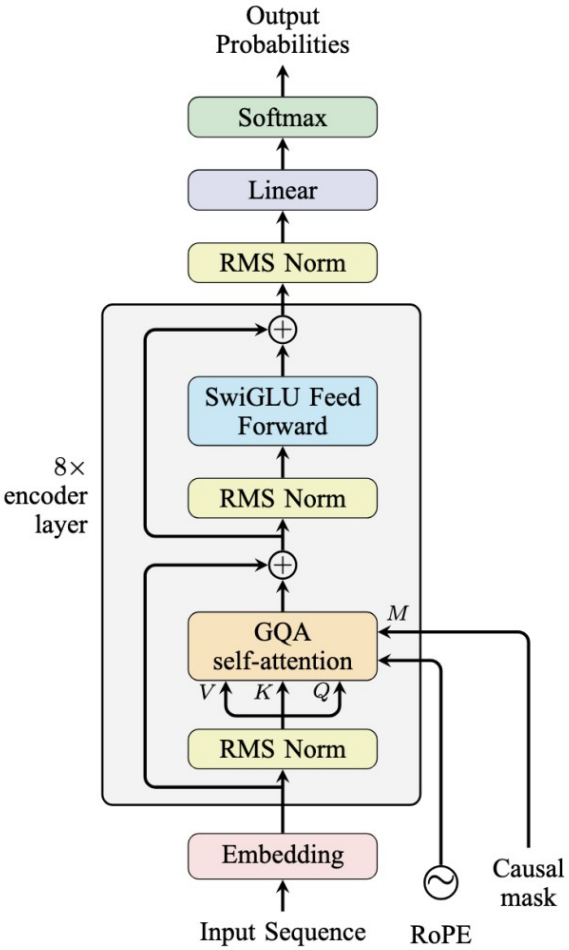
$$Y' = g(Y, X, W\theta, \sigma) = Y \odot \sigma(XW\theta)$$

Preliminary

Gating Mechanisms

$$Y' = g(Y, X, W\theta, \sigma) = Y \odot \sigma(XW\theta)$$

- **Widely Used in FFN (e.g., SwiGLU)**
- **The gating score, $\sigma(X\ W\theta)$ acts as a dynamic filter, controlling the information flow from Y**
- **$W\theta$ is learnable**



Problem Definition

Prior work never isolates the gate

- **Gating is widely used**
  - SwiGLU in every modern LLM's FFN
  - Not **inside the attention core**
- Prior works add gating — **with other changes**
  - **Switch Head** — sigmoid routing over top-k head experts
  - **NSA** — gating + sparse attention design

- Performance improves, but the gate's contribution stays unseparated

→ How much does gating alone contribute?

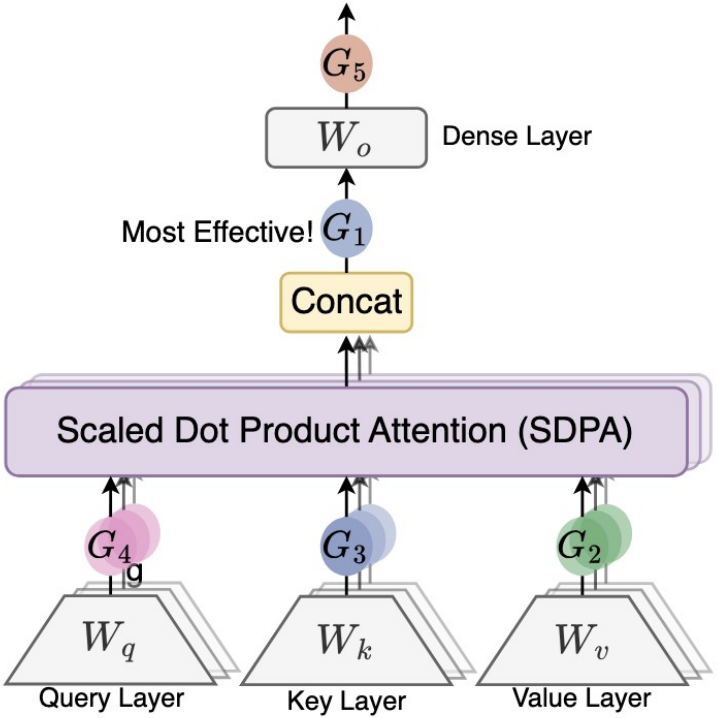
Main Contribution

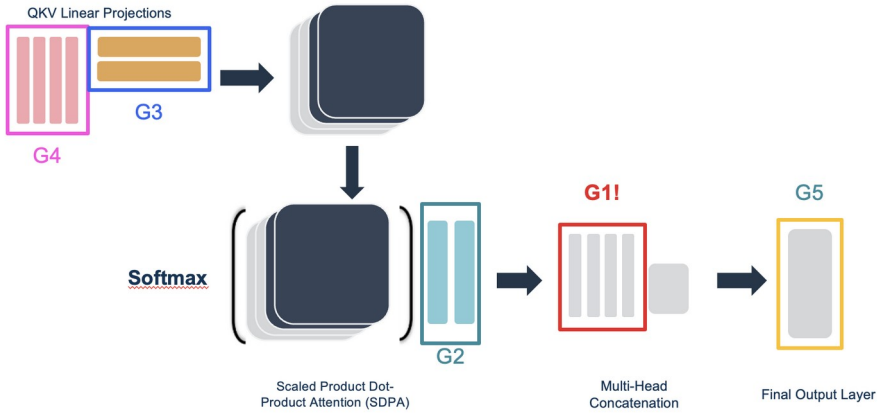
Ablation - Gating Design Space

- **Gating Design Space**
  - Positions
  - Granularity
  - Head Specific or Shared
  - Multiplicative or additive
  - Activation Function

Gating Design Space

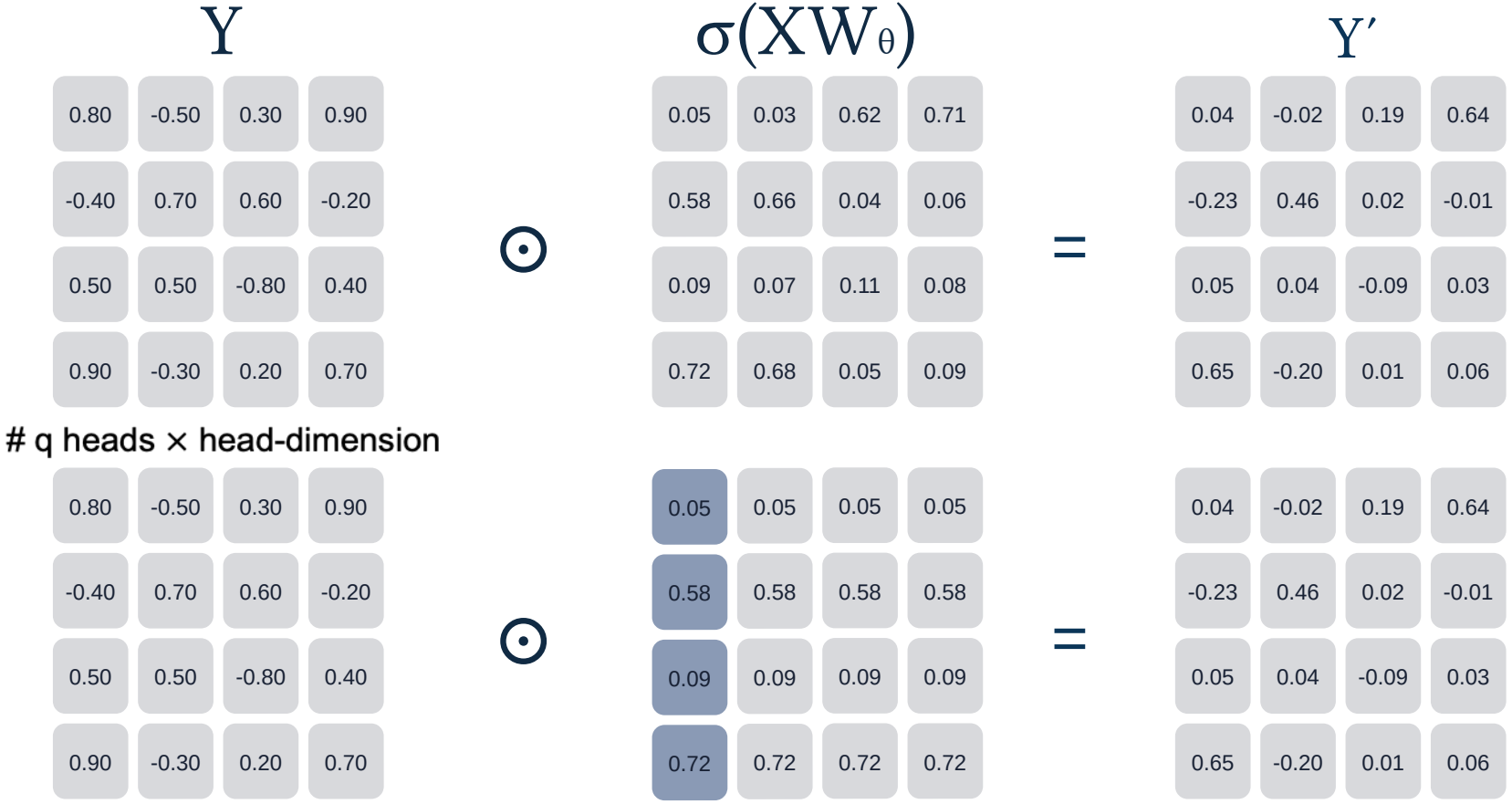
Position (1/5)





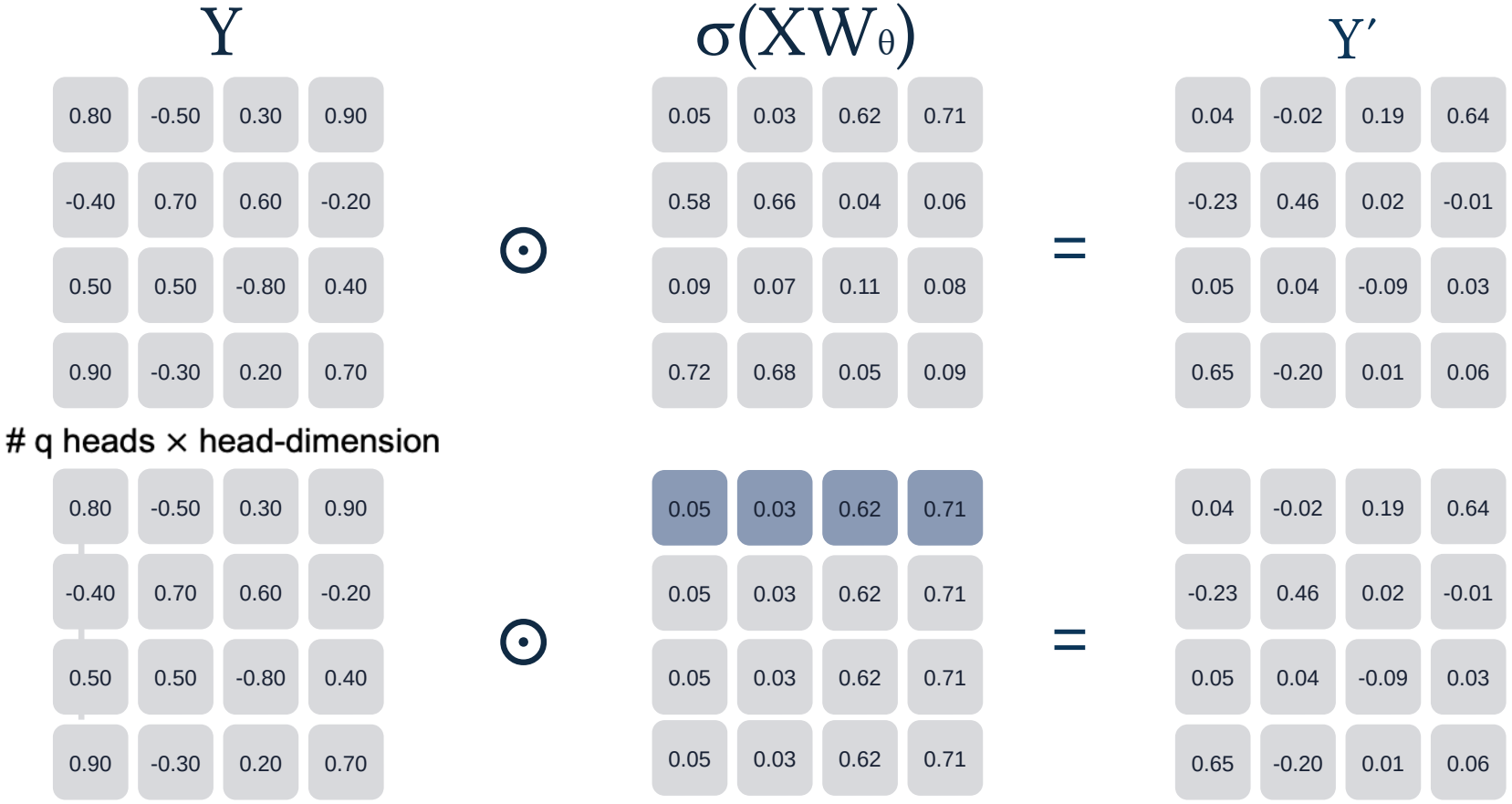
Gating Design Space

Granularity–Elementwise vs. Headwise(2/5)



Gating Design Space

Head Specific or Shared (3/5)



Gating Design Space

Multiplicative or Additive (4/5)

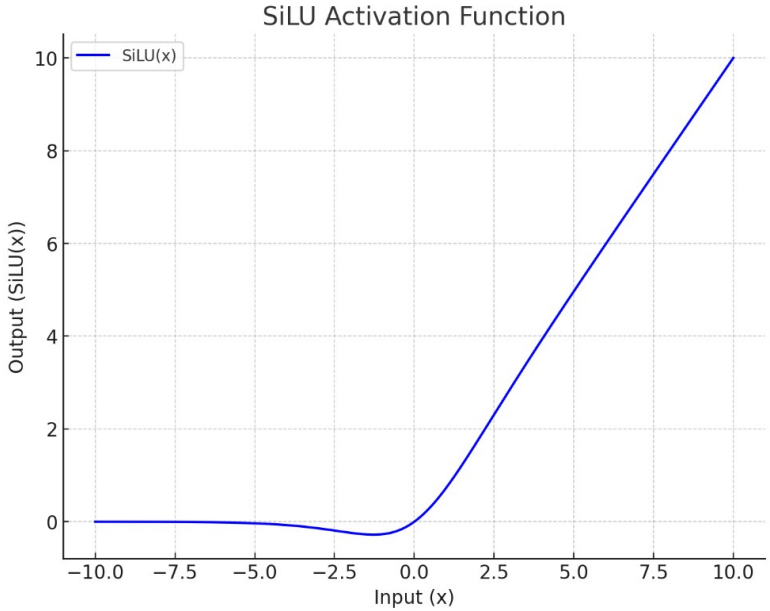
$$Additive: Y' = Y + \sigma(XW\theta)$$
$$Multiplicative: Y' = Y \cdot \sigma(XW\theta)$$

- Additive: SiLU
  - sigmoid bounded ([0, 1]) ->
  - Additive unbounded .

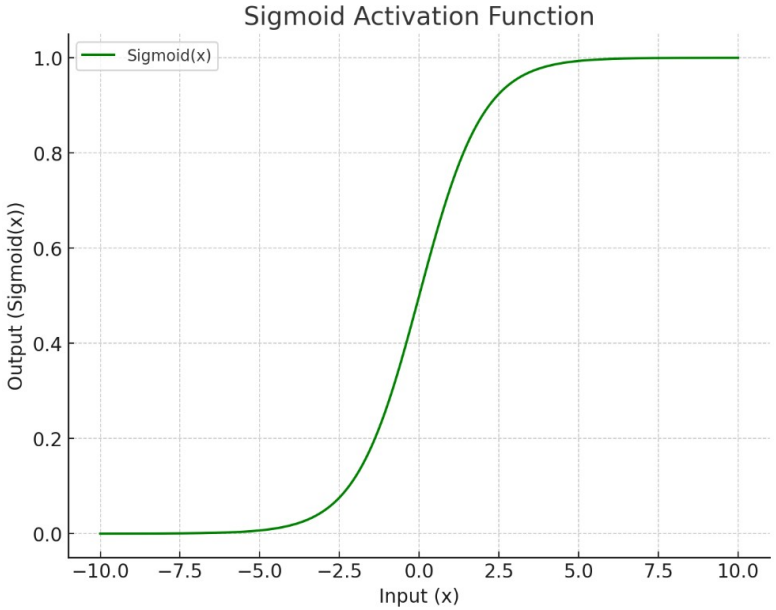
Gating Design Space

Activation Function (5/5)

$$\mathrm{SiLU}(x) = x \cdot \sigma(x) = x / (1 + e^{-x})$$



$$\sigma(x) = 1 / (1 + e^{-x})$$



Main Results

MoE models

Increasing #KV heads

Increasing #Q heads

Method	Act Func	Score Shape	Added Param	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
Reference Baselines (Baseline uses $q = 32, k = 4$. All methods use $d_k = 128$.)								
(1) Baseline	-	-	0	6.026	73.07	58.79	52.92	60.26
(2) $k = 8$	-	-	50	5.979	73.51	59.78	52.16	62.26
(3) $q = 48$	-	-	201	5.953	73.59	58.45	53.30	59.67
(4) Add 4 Experts	-	-	400	5.964	73.19	58.84	52.54	**63.19**
Gating Position Variants								
(5) SDPA Elementwise G_1	sigmoid	$n \times q \times d_k$	201	**5.761**	74.64	**60.82**	**55.27**	62.20
(6) v Elementwise G_2	sigmoid	$n \times k \times d_k$	25	5.820	74.38	59.17	53.97	61.00
(7) k Elementwise G_3	sigmoid	$n \times k \times d_k$	25	6.016	72.88	59.18	50.49	61.74
(8) q Elementwise G_4	sigmoid	$n \times q \times d_k$	201	5.981	73.01	58.74	53.97	62.14
(9) Dense Output G_5	sigmoid	$n \times d_{\text{model}}$	100	6.017	73.32	59.41	50.87	59.43
Gating Granularity Variants								
(10) SDPA Headwise G_1	sigmoid	$n \times q$	1.6	5.792	74.50	60.05	54.44	62.61
(11) v Headwise G_2	sigmoid	$n \times q$	0.2	5.808	74.38	59.32	53.53	62.61
Head-Specific v.s. Head-Shared Gating								
(12) SDPA Head-Shared G_1	sigmoid	$n \times d_k$	201	5.801	74.34	60.06	53.15	61.01
(13) v Head-Shared G_2	sigmoid	$n \times d_k$	25	5.867	74.10	59.02	53.03	60.61
Multiplicative v.s. Additive								
(14) SDPA Additive G_1	SiLU	$n \times q \times d_k$	201	5.821	**74.81**	60.06	53.30	60.98
Activation Variants								
(15) SDPA Elementwise G_1	SiLU	$n \times q \times d_k$	201	5.822	74.22	60.49	54.59	62.34

Main Results

MoE models– Position (1/5)

Increasing #KV heads

Increasing #Q heads

Method	Act Func	Score Shape	Added Param	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
Reference Baselines (Baseline uses $q = 32, k = 4$. All methods use $d_k = 128$.)								
(1) Baseline	-	-	0	6.026	73.07	58.79	52.92	60.26
(2) $k = 8$	-	-	50	5.979	73.51	59.78	52.16	62.26
(3) $q = 48$	-	-	201	5.953	73.59	58.45	53.30	59.67
(4) Add 4 Experts	-	-	400	5.964	73.19	58.84	52.54	**63.19**
Gating Position Variants								
(5) SDPA Elementwise G_1	sigmoid	$n \times q \times d_k$	201	**5.761**	74.64	**60.82**	**55.27**	62.20
(6) v Elementwise G_2	sigmoid	$n \times k \times d_k$	25	5.820	74.38	59.17	53.97	61.00
(7) k Elementwise G_3	sigmoid	$n \times k \times d_k$	25	6.016	72.88	59.18	50.49	61.74
(8) q Elementwise G_4	sigmoid	$n \times q \times d_k$	201	5.981	73.01	58.74	53.97	62.14
(9) Dense Output G_5	sigmoid	$n \times d_{\text{model}}$	100	6.017	73.32	59.41	50.87	59.43
Gating Granularity Variants								
(10) SDPA Headwise G_1	sigmoid	$n \times q$	1.6	5.792	74.50	60.05	54.44	62.61
(11) v Headwise G_2	sigmoid	$n \times q$	0.2	5.808	74.38	59.32	53.53	62.61
Head-Specific v.s. Head-Shared Gating								
(12) SDPA Head-Shared G_1	sigmoid	$n \times d_k$	201	5.801	74.34	60.06	53.15	61.01
(13) v Head-Shared G_2	sigmoid	$n \times d_k$	25	5.867	74.10	59.02	53.03	60.61
Multiplicative v.s. Additive								
(14) SDPA Additive G_1	SiLU	$n \times q \times d_k$	201	5.821	**74.81**	60.06	53.30	60.98
Activation Variants								
(15) SDPA Elementwise G_1	SiLU	$n \times q \times d_k$	201	5.822	74.22	60.49	54.59	62.34

Main Results

MoE models– Granularity (2/5)

Method	Act Func	Score Shape	Added Param	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
Reference Baselines (Baseline uses $q = 32, k = 4$. All methods use $d_k = 128$.)								
(1) Baseline	-	-	0	6.026	73.07	58.79	52.92	60.26
(2) $k = 8$	-	-	50	5.979	73.51	59.78	52.16	62.26
(3) $q = 48$	-	-	201	5.953	73.59	58.45	53.30	59.67
(4) Add 4 Experts	-	-	400	5.964	73.19	58.84	52.54	**63.19**
Gating Position Variants								
(5) SDPA Elementwise G_1	sigmoid	$n \times q \times d_k$	201	**5.761**	74.64	**60.82**	**55.27**	62.20
(6) v Elementwise G_2	sigmoid	$n \times k \times d_k$	25	5.820	74.38	59.17	53.97	61.00
(7) k Elementwise G_3	sigmoid	$n \times k \times d_k$	25	6.016	72.88	59.18	50.49	61.74
(8) q Elementwise G_4	sigmoid	$n \times q \times d_k$	201	5.981	73.01	58.74	53.97	62.14
(9) Dense Output G_5	sigmoid	$n \times d_{\text{model}}$	100	6.017	73.32	59.41	50.87	59.43
Gating Granularity Variants								
(10) SDPA Headwise G_1	sigmoid	$n \times q$	1.6	5.792	74.50	60.05	54.44	62.61
(11) v Headwise G_2	sigmoid	$n \times q$	0.2	5.808	74.38	59.32	53.53	62.61
Head-Specific v.s. Head-Shared Gating								
(12) SDPA Head-Shared G_1	sigmoid	$n \times d_k$	201	5.801	74.34	60.06	53.15	61.01
(13) v Head-Shared G_2	sigmoid	$n \times d_k$	25	5.867	74.10	59.02	53.03	60.61
Multiplicative v.s. Additive								
(14) SDPA Additive G_1	SiLU	$n \times q \times d_k$	201	5.821	**74.81**	60.06	53.30	60.98
Activation Variants								
(15) SDPA Elementwise G_1	SiLU	$n \times q \times d_k$	201	5.822	74.22	60.49	54.59	62.34

Main Results

MoE models – Head Specific or shared (3/5)

Method	Act Func	Score Shape	Added Param	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
Reference Baselines (Baseline uses $q = 32, k = 4$. All methods use $d_k = 128$.)								
(1) Baseline	-	-	0	6.026	73.07	58.79	52.92	60.26
(2) $k = 8$	-	-	50	5.979	73.51	59.78	52.16	62.26
(3) $q = 48$	-	-	201	5.953	73.59	58.45	53.30	59.67
(4) Add 4 Experts	-	-	400	5.964	73.19	58.84	52.54	**63.19**
Gating Position Variants								
(5) SDPA Elementwise G_1	sigmoid	$n \times q \times d_k$	201	**5.761**	74.64	**60.82**	**55.27**	62.20
(6) v Elementwise G_2	sigmoid	$n \times k \times d_k$	25	5.820	74.38	59.17	53.97	61.00
(7) k Elementwise G_3	sigmoid	$n \times k \times d_k$	25	6.016	72.88	59.18	50.49	61.74
(8) q Elementwise G_4	sigmoid	$n \times q \times d_k$	201	5.981	73.01	58.74	53.97	62.14
(9) Dense Output G_5	sigmoid	$n \times d_{\text{model}}$	100	6.017	73.32	59.41	50.87	59.43
Gating Granularity Variants								
(10) SDPA Headwise G_1	sigmoid	$n \times q$	1.6	5.792	74.50	60.05	54.44	62.61
(11) v Headwise G_2	sigmoid	$n \times q$	0.2	5.808	74.38	59.32	53.53	62.61
Head-Specific v.s. Head-Shared Gating								
(12) SDPA Head-Shared G_1	sigmoid	$n \times d_k$	201	5.801	74.34	60.06	53.15	61.01
(13) v Head-Shared G_2	sigmoid	$n \times d_k$	25	5.867	74.10	59.02	53.03	60.61
Multiplicative v.s. Additive								
(14) SDPA Additive G_1	SiLU	$n \times q \times d_k$	201	5.821	**74.81**	60.06	53.30	60.98
Activation Variants								
(15) SDPA Elementwise G_1	SiLU	$n \times q \times d_k$	201	5.822	74.22	60.49	54.59	62.34

Main Results

MoE models– Multiplicative or Additive (4/5)

Method	Act Func	Score Shape	Added Param	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
Reference Baselines (Baseline uses $q = 32, k = 4$. All methods use $d_k = 128$.)								
(1) Baseline	-	-	0	6.026	73.07	58.79	52.92	60.26
(2) $k = 8$	-	-	50	5.979	73.51	59.78	52.16	62.26
(3) $q = 48$	-	-	201	5.953	73.59	58.45	53.30	59.67
(4) Add 4 Experts	-	-	400	5.964	73.19	58.84	52.54	**63.19**
Gating Position Variants								
(5) SDPA Elementwise G_1	sigmoid	$n \times q \times d_k$	201	5.761	74.64	**60.82**	**55.27**	62.20
(6) v Elementwise G_2	sigmoid	$n \times k \times d_k$	25	5.820	74.38	59.17	53.97	61.00
(7) k Elementwise G_3	sigmoid	$n \times k \times d_k$	25	6.016	72.88	59.18	50.49	61.74
(8) q Elementwise G_4	sigmoid	$n \times q \times d_k$	201	5.981	73.01	58.74	53.97	62.14
(9) Dense Output G_5	sigmoid	$n \times d_{\text{model}}$	100	6.017	73.32	59.41	50.87	59.43
Gating Granularity Variants								
(10) SDPA Headwise G_1	sigmoid	$n \times q$	1.6	5.792	74.50	60.05	54.44	62.61
(11) v Headwise G_2	sigmoid	$n \times q$	0.2	5.808	74.38	59.32	53.53	62.61
Head-Specific v.s. Head-Shared Gating								
(12) SDPA Head-Shared G_1	sigmoid	$n \times d_k$	201	5.801	74.34	60.06	53.15	61.01
(13) v Head-Shared G_2	sigmoid	$n \times d_k$	25	5.867	74.10	59.02	53.03	60.61
Multiplicative v.s. Additive								
(14) SDPA Additive G_1	SiLU	$n \times q \times d_k$	201	5.821	**74.81**	60.06	53.30	60.98
Activation Variants								
(15) SDPA Elementwise G_1	SiLU	$n \times q \times d_k$	201	5.822	74.22	60.49	54.59	62.34

Main Results

MoE models– Activation Function (5/5)

Method	Act Func	Score Shape	Added Param	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
Reference Baselines (Baseline uses $q = 32, k = 4$. All methods use $d_k = 128$.)								
(1) Baseline	-	-	0	6.026	73.07	58.79	52.92	60.26
(2) $k = 8$	-	-	50	5.979	73.51	59.78	52.16	62.26
(3) $q = 48$	-	-	201	5.953	73.59	58.45	53.30	59.67
(4) Add 4 Experts	-	-	400	5.964	73.19	58.84	52.54	**63.19**
Gating Position Variants								
(5) SDPA Elementwise G_1	sigmoid	$n \times q \times d_k$	201	5.761	74.64	**60.82**	**55.27**	62.20
(6) v Elementwise G_2	sigmoid	$n \times k \times d_k$	25	5.820	74.38	59.17	53.97	61.00
(7) k Elementwise G_3	sigmoid	$n \times k \times d_k$	25	6.016	72.88	59.18	50.49	61.74
(8) q Elementwise G_4	sigmoid	$n \times q \times d_k$	201	5.981	73.01	58.74	53.97	62.14
(9) Dense Output G_5	sigmoid	$n \times d_{\text{model}}$	100	6.017	73.32	59.41	50.87	59.43
Gating Granularity Variants								
(10) SDPA Headwise G_1	sigmoid	$n \times q$	1.6	5.792	74.50	60.05	54.44	62.61
(11) v Headwise G_2	sigmoid	$n \times q$	0.2	5.808	74.38	59.32	53.53	62.61
Head-Specific v.s. Head-Shared Gating								
(12) SDPA Head-Shared G_1	sigmoid	$n \times d_k$	201	5.801	74.34	60.06	53.15	61.01
(13) v Head-Shared G_2	sigmoid	$n \times d_k$	25	5.867	74.10	59.02	53.03	60.61
Multiplicative v.s. Additive								
(14) SDPA Additive G_1	SiLU	$n \times q \times d_k$	201	5.821	**74.81**	60.06	53.30	60.98
Activation Variants								
(15) SDPA Elementwise G_1	SiLU	$n \times q \times d_k$	201	5.822	74.22	60.49	54.59	62.34

Main Results

Dense Models

reduce the width of FFN to maintain the parameter size.

Method	Max LR	Avg PPL	HumanEval	MMLU	GSM8k	Hellaswag	C-eval	CMMLU
28 Layer, 1.7B Parameters, **400B Tokens**, Batch Size=1024								
(1) Baseline	4.0×10^{-3}	7.499	28.66	50.21	27.82	64.94	49.15	49.52
(2) SDPA Elementwise	4.0×10^{-3}	**7.404**	**29.27**	**51.15**	**28.28**	**65.48**	**50.72**	**50.72**
28 Layer, 1.7B Parameters, **3.5T Tokens**, Batch Size=2048								
(3) Baseline	4.5×10^{-3}	6.180	34.15	59.10	69.07	68.02	68.19	64.95
(4) SDPA Elementwise	4.5×10^{-3}	**6.130**	**37.80**	**59.61**	**70.20**	**68.84**	**68.52**	**65.76**
48 Layer, 1.7B Parameters, **400B Tokens**, Batch Size=1024								
(5) Baseline	4.0×10^{-3}	7.421	28.05	52.04	32.98	65.96	51.11	51.86
(6) Baseline	8.0×10^{-3}	9.195	21.34	44.28	15.24	57.00	43.11	42.63
(7) Baseline+Sandwich Norm	8.0×10^{-3}	7.407	30.49	52.07	32.90	66.00	52.04	51.72
(8) SDPA Elementwise	4.0×10^{-3}	**7.288**	**31.71**	52.44	32.37	66.28	52.06	52.29
(9) SDPA Headwise	4.0×10^{-3}	7.370	31.10	53.83	34.12	65.59	**55.07**	52.38
(10) SDPA Elementwise	8.0×10^{-3}	7.325	31.10	**54.47**	**36.62**	**66.40**	53.91	**53.80**
48 Layer, 1.7B Parameters, **1T Tokens**, Batch Size=4096								
(11) Baseline	5.3×10^{-3}	7.363	29.88	54.44	32.22	65.43	53.72	53.37
(12) Baseline	8.0×10^{-3}	-	-	-	-	-	-	-
(13) SDPA Elementwise	5.3×10^{-3}	7.101	**34.15**	55.70	36.69	67.17	54.51	54.68
(14) SDPA Elementwise	8.0×10^{-3}	**7.078**	31.71	**56.47**	**39.73**	**67.38**	**55.52**	**55.77**

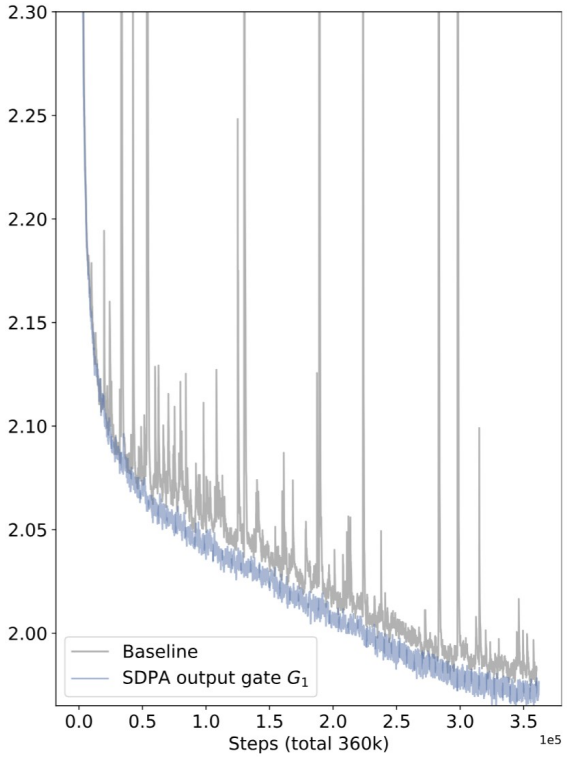
Main Results

Dense Models – Training Stability & Scaling (1/2)

Method	Max LR	Avg PPL	HumanEval	MMLU	GSM8k	Hellaswag	C-eval	CMMLU
28 Layer, 1.7B Parameters, **400B Tokens**, Batch Size=1024								
(1) Baseline	4.0×10^{-3}	7.499	28.66	50.21	27.82	64.94	49.15	49.52
(2) SDPA Elementwise	4.0×10^{-3}	**7.404**	**29.27**	**51.15**	**28.28**	**65.48**	**50.72**	**50.72**
28 Layer, 1.7B Parameters, **3.5T Tokens**, Batch Size=2048								
(3) Baseline	4.5×10^{-3}	6.180	34.15	59.10	69.07	68.02	68.19	64.95
(4) SDPA Elementwise	4.5×10^{-3}	**6.130**	**37.80**	**59.61**	**70.20**	**68.84**	**68.52**	**65.76**
48 Layer, 1.7B Parameters, **400B Tokens**, Batch Size=1024								
(5) Baseline	4.0×10^{-3}	7.421	28.05	52.04	32.98	65.96	51.11	51.86
(6) Baseline	8.0×10^{-3}	9.195	21.34	44.28	15.24	57.00	43.11	42.63
(7) Baseline+Sandwich Norm	8.0×10^{-3}	7.407	30.49	52.07	32.90	66.00	52.04	51.72
(8) SDPA Elementwise	4.0×10^{-3}	**7.288**	**31.71**	52.44	32.37	66.28	52.06	52.29
(9) SDPA Headwise	4.0×10^{-3}	7.370	31.10	53.83	34.12	65.59	**55.07**	52.38
(10) SDPA Elementwise	8.0×10^{-3}	7.325	31.10	**54.47**	**36.62**	**66.40**	53.91	**53.80**
48 Layer, 1.7B Parameters, **1T Tokens**, Batch Size=4096								
(11) Baseline	5.3×10^{-3}	7.363	29.88	54.44	32.22	65.43	53.72	53.37
(12) Baseline	8.0×10^{-3}	-	-	-	-	-	-	-
(13) SDPA Elementwise	5.3×10^{-3}	7.101	**34.15**	55.70	36.69	67.17	54.51	54.68
(14) SDPA Elementwise	8.0×10^{-3}	**7.078**	31.71	**56.47**	**39.73**	**67.38**	**55.52**	**55.77**

Main Results

Dense Models – Training Stability & Scaling (2/2)



Main Results

Dense Models – Training Stability & Scaling (2/2)

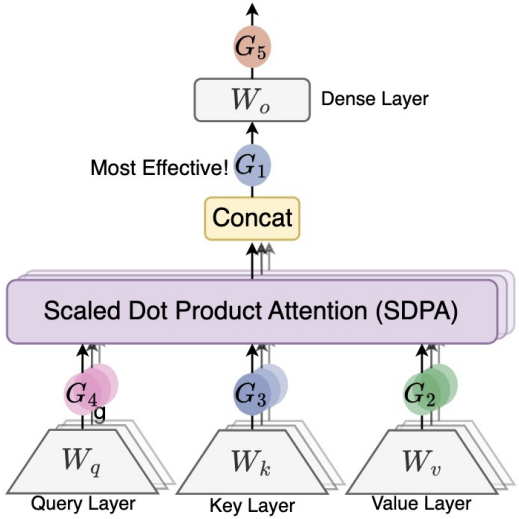
Method	Max LR	Avg PPL	HumanEval	MMLU	GSM8k	Hellaswag	C-eval	CMMLU
28 Layer, 1.7B Parameters, **400B Tokens**, Batch Size=1024								
(1) Baseline	4.0×10^{-3}	7.499	28.66	50.21	27.82	64.94	49.15	49.52
(2) SDPA Elementwise	4.0×10^{-3}	**7.404**	**29.27**	**51.15**	**28.28**	**65.48**	**50.72**	**50.72**
28 Layer, 1.7B Parameters, **3.5T Tokens**, Batch Size=2048								
(3) Baseline	4.5×10^{-3}	6.180	34.15	59.10	69.07	68.02	68.19	64.95
(4) SDPA Elementwise	4.5×10^{-3}	**6.130**	**37.80**	**59.61**	**70.20**	**68.84**	**68.52**	**65.76**
48 Layer, 1.7B Parameters, **400B Tokens**, Batch Size=1024								
(5) Baseline	4.0×10^{-3}	7.421	28.05	52.04	32.98	65.96	51.11	51.86
(6) Baseline	8.0×10^{-3}	9.195	21.34	44.28	15.24	57.00	43.11	42.63
(7) Baseline+Sandwich Norm	8.0×10^{-3}	7.407	30.49	52.07	32.90	66.00	52.04	51.72
(8) SDPA Elementwise	4.0×10^{-3}	**7.288**	**31.71**	52.44	32.37	66.28	52.06	52.29
(9) SDPA Headwise	4.0×10^{-3}	7.370	31.10	53.83	34.12	65.59	**55.07**	52.38
(10) SDPA Elementwise	8.0×10^{-3}	7.325	31.10	**54.47**	**36.62**	**66.40**	53.91	**53.80**
48 Layer, 1.7B Parameters, **1T Tokens**, Batch Size=4096								
(11) Baseline	5.3×10^{-3}	7.363	29.88	54.44	32.22	65.43	53.72	53.37
(12) Baseline	8.0×10^{-3}	-	-	-	-	-	-	-
(13) SDPA Elementwise	5.3×10^{-3}	7.101	**34.15**	55.70	36.69	67.17	54.51	54.68
(14) SDPA Elementwise	8.0×10^{-3}	7.078	31.71	**56.47**	**39.73**	**67.38**	**55.52**	**55.77**

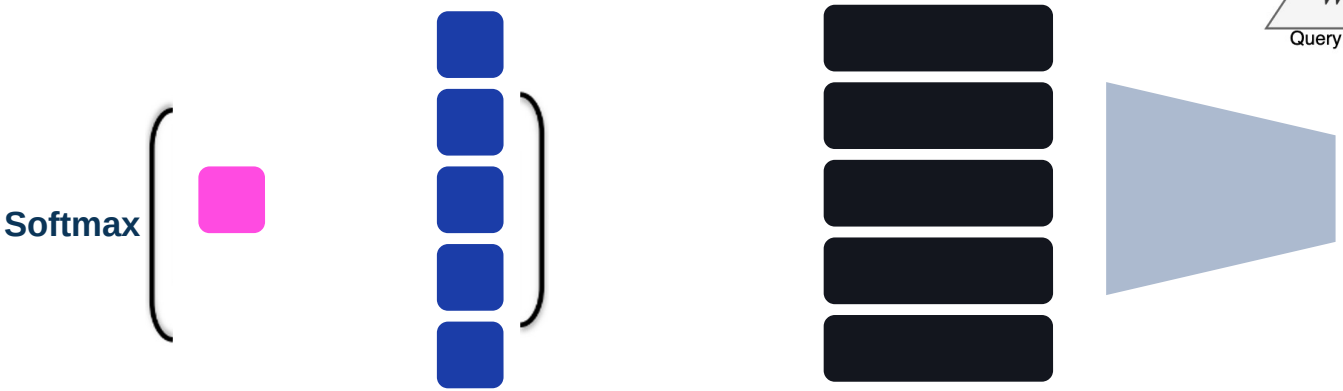
Why It Works?

Non-Linearity

Scaled Dot Product Attention(SDPA)

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$





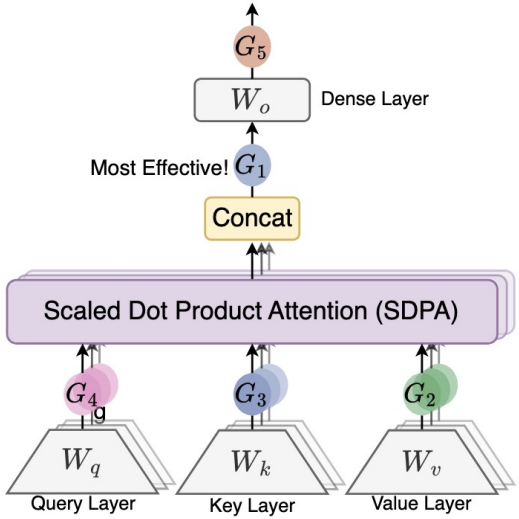
$$o_i^k = (\sum_{j=0}^{i} S_{ij}^k \cdot X_j W_V^k) W_O^k = \sum_{j=0}^{i} S_{ij}^k \cdot X_j (W_V^k W_O^k),$$

Why It Works?

Non-Linearity

Scaled Dot Product Attention(SDPA)

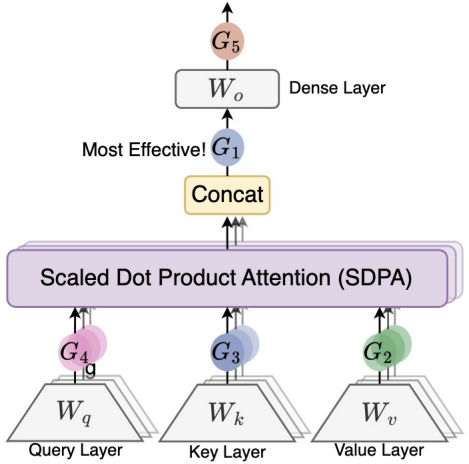
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

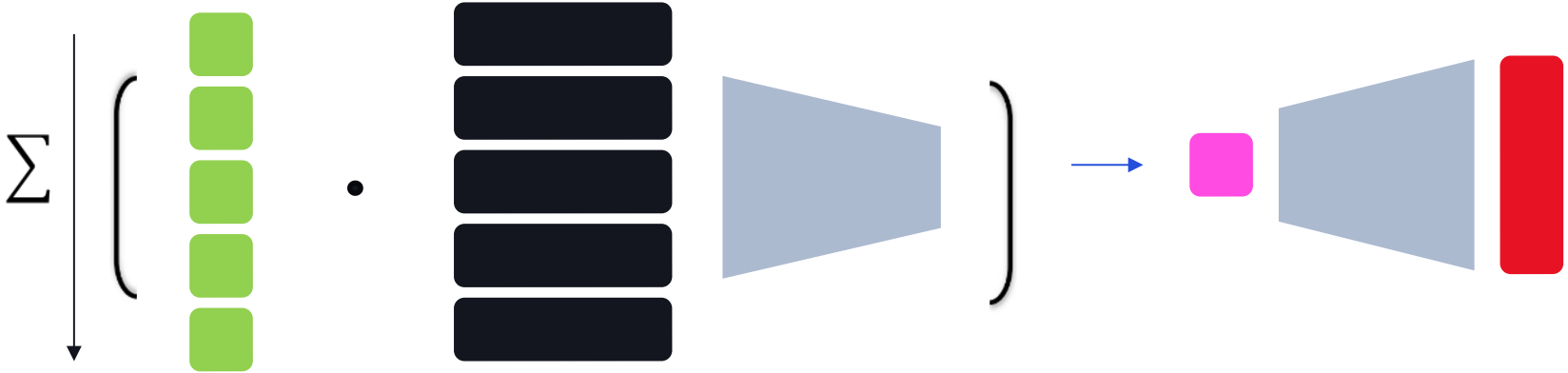


$$o_i^k = (\sum_{j=0}^{i} S_{ij}^k \cdot X_j W_V^k) W_O^k = \sum_{j=0}^{i} S_{ij}^k \cdot X_j (W_V^k W_O^k),$$

Why It Works?

Non-Linearity





$$o_i^k = (\sum_{j=0}^{i} S_{ij}^k \cdot X_j W_V^k) W_O^k = \sum_{j=0}^{i} S_{ij}^k \cdot X_j (W_V^k W_O^k),$$

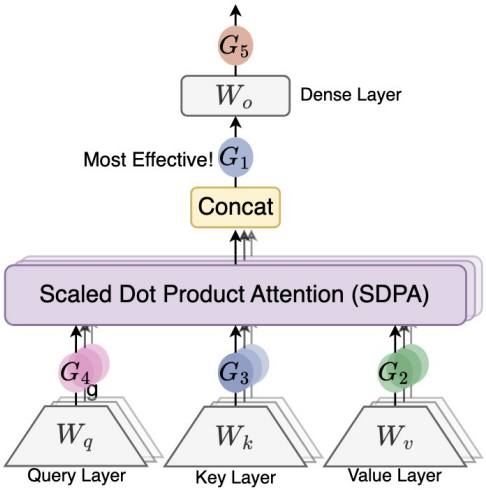
Why It Works?

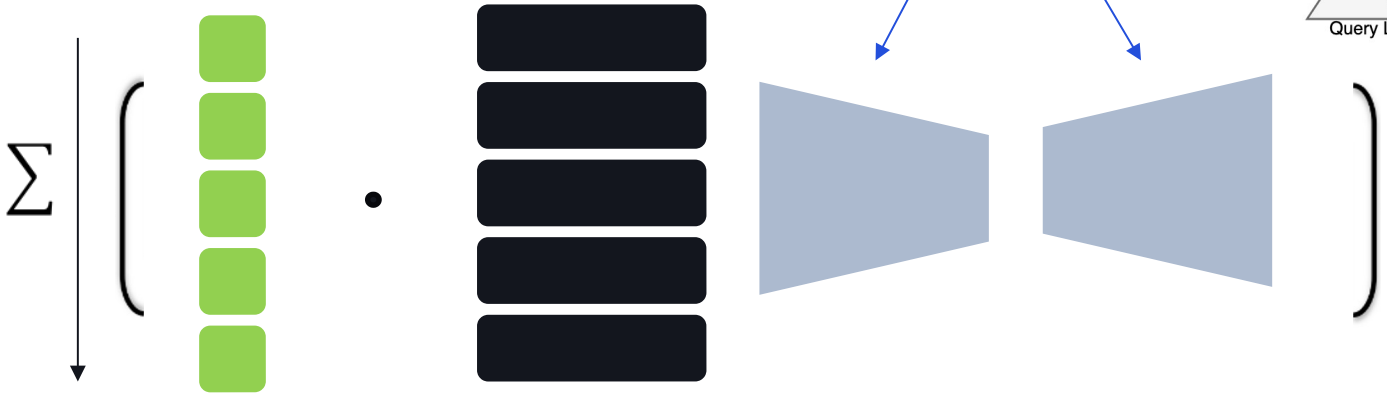
Non-Linearity

- **Low-Rank Mapping in ttention**

D_model X
D_model
low-rank

Linear!



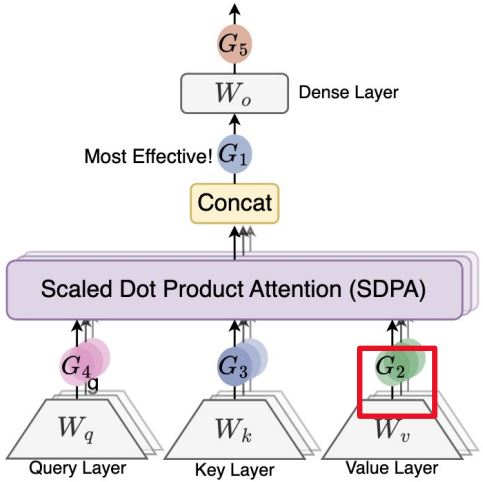


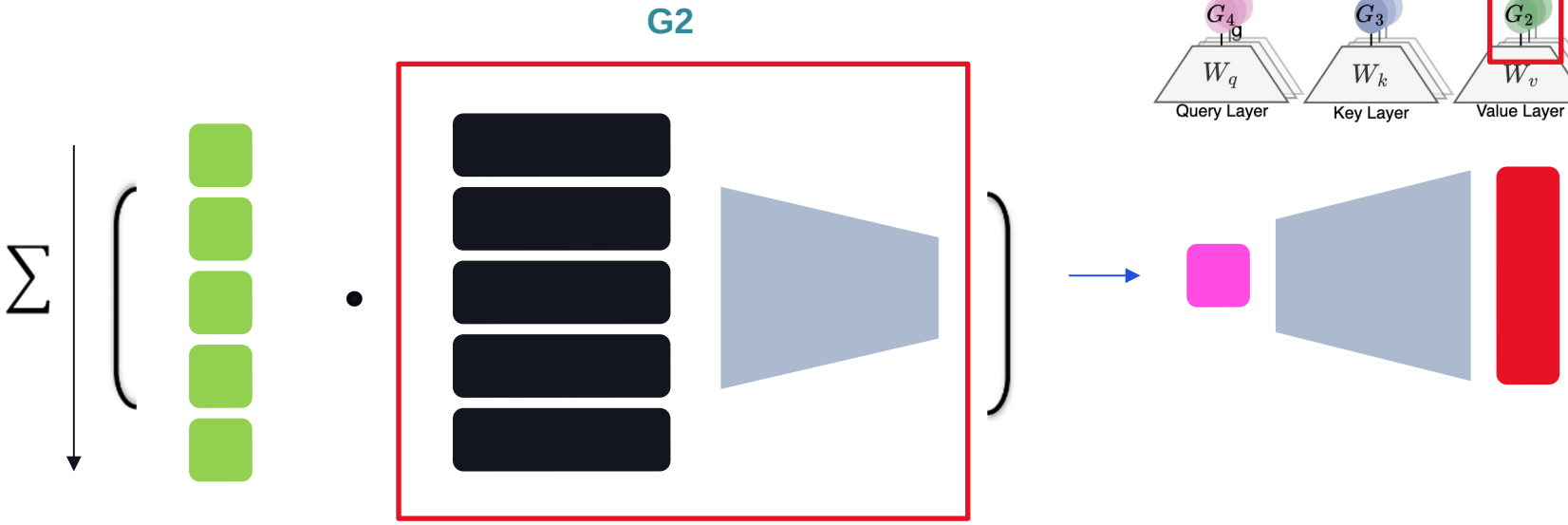
$$o_i^k = (\sum_{j=0}^{i} S_{ij}^k \cdot X_j W_V^k) W_O^k = \sum_{j=0}^{i} S_{ij}^k \cdot X_j (W_V^k W_O^k),$$

Why It Works?

Non-Linearity

- **Non-linearity Improves the Expressiveness of Low-Rank Mapping in ttention**



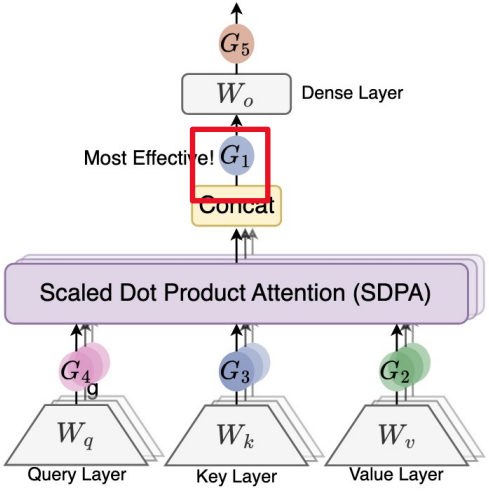


$$o_i^k = \left(\sum_{j=0}^{i} S_{ij}^k \cdot \text{Non-Linearity-Map}(X_j W_V^k) \right) W_O^k$$

Why It Works?

Non-Linearity

- **Non-linearity Improves the Expressiveness of Low-Rank Mapping in ttention**



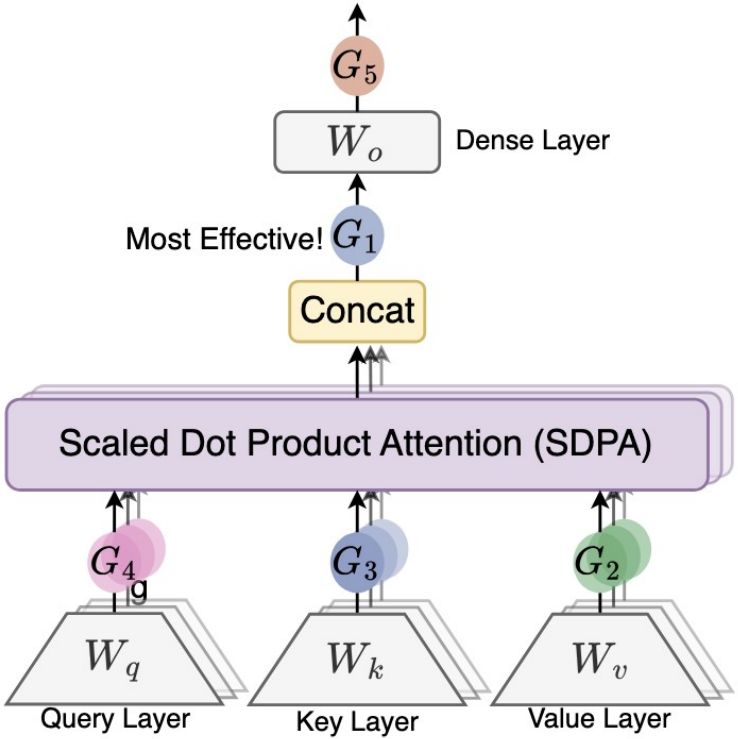
G1 – Most Effective!

$$o_i^k = \text{Non-Linearity-Map}\left(\sum_{j=0}^{i} S_{ij}^k \cdot X_j W_V^k\right) W_O^k$$

Why It Works?

Non-Linearity

- **Non-linearity Improves the Expressiveness of Low-Rank Mapping in ttention**



$$o_i^k = (\sum_{j=0}^{i} S_{ij}^k \cdot X_j W_V^k) W_O^k = \sum_{j=0}^{i} S_{ij}^k \cdot X_j (W_V^k W_O^k),$$

G5 G3 G4

Why It Works?

Non-Linearity

- **Question:** ? SiLU ?

$$Y = X \odot \sigma(XW_\theta)$$

$$Y = X + \mathrm{SiLU}(G(X))$$

Method	Activation Function	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
(1) Baseline	-	6.026	73.07	58.79	52.92	60.26
(2) SDPA Elementwise Gate	Sigmoid	**5.761**	74.64	**60.82**	**55.27**	**62.20**
(3) v Elementwise Gate	Sigmoid	5.820	74.38	59.17	53.97	61.00
(4) SDPA Additive Gate	SiLU	5.821	**74.81**	60.06	53.30	60.98
(5) SDPA GroupNorm	RMSNorm	5.847	74.10	60.15	53.75	61.14
(6) SDPA SiLU	SiLU	5.975	73.34	59.55	53.19	60.90
(7) SDPA Additive Gate	Identity	5.882	74.17	59.20	52.77	59.86

Why It Works?

Non-Linearity

- **SiLU , Gate bad**

$$Y = X + \mathrm{SiLU}(G(X)) \qquad Y = X + \mathrm{SiLU}(X)$$

Method	Activation Function	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
(1) Baseline	-	6.026	73.07	58.79	52.92	60.26
(2) SDPA Elementwise Gate	Sigmoid	**5.761**	74.64	**60.82**	**55.27**	**62.20**
(3) v Elementwise Gate	Sigmoid	5.820	74.38	59.17	53.97	61.00
(4) SDPA Additive Gate	SiLU	5.821	**74.81**	60.06	53.30	60.98
(5) SDPA GroupNorm	RMSNorm	5.847	74.10	60.15	53.75	61.14
(6) SDPA SiLU	SiLU	5.975	73.34	59.55	53.19	60.90
(7) SDPA Additive Gate	Identity	5.882	74.17	59.20	52.77	59.86

Why It Works?

Non-Linearity

- **Gate , SiLU bad**

$$Y = X + \mathrm{SiLU}(G(X)) \qquad Y = X + G(X)$$

Method	Activation Function	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
(1) Baseline	-	6.026	73.07	58.79	52.92	60.26
(2) SDPA Elementwise Gate	Sigmoid	**5.761**	74.64	**60.82**	**55.27**	**62.20**
(3) v Elementwise Gate	Sigmoid	5.820	74.38	59.17	53.97	61.00
(4) SDPA Additive Gate	SiLU	5.821	**74.81**	60.06	53.30	60.98
(5) SDPA GroupNorm	RMSNorm	5.847	74.10	60.15	53.75	61.14
(6) SDPA SiLU	SiLU	5.975	73.34	59.55	53.19	60.90
(7) SDPA Additive Gate	Identity	5.882	74.17	59.20	52.77	59.86

Why It Works?

Non-Linearity

- **SiLU , SiLU ✦ gating !**

$$Y = X + \mathrm{SiLU}(G(X)) \qquad Y = X + G(X) \qquad Y = X + \mathrm{SiLU}(X)$$

Method	Activation Function	Avg PPL	Hellaswag	MMLU	GSM8k	C-eval
(1) Baseline	-	6.026	73.07	58.79	52.92	60.26
(2) SDPA Elementwise Gate	Sigmoid	**5.761**	74.64	**60.82**	**55.27**	**62.20**
(3) v Elementwise Gate	Sigmoid	5.820	74.38	59.17	53.97	61.00
(4) SDPA Additive Gate	SiLU	5.821	**74.81**	60.06	53.30	60.98
(5) SDPA GroupNorm	RMSNorm	5.847	74.10	60.15	53.75	61.14
(6) SDPA SiLU	SiLU	5.975	73.34	59.55	53.19	60.90
(7) SDPA Additive Gate	Identity	5.882	74.17	59.20	52.77	59.86

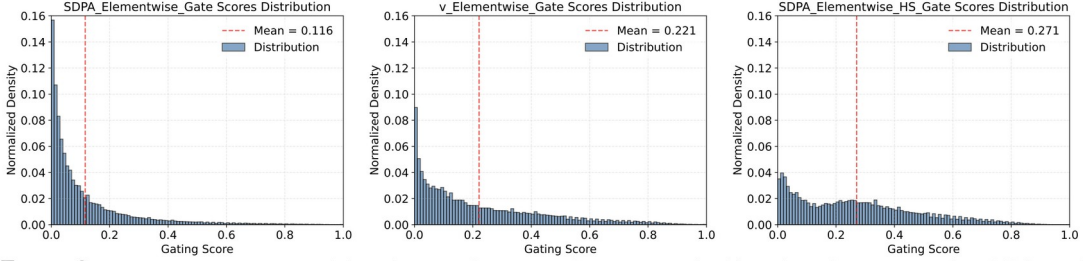
Why It Works?

Gating Introduces Input-Dependent Sparsity (1/3)

- **Effective Gating Scores are Sparse (1/3)**
  - SDPA output gating score distribution shows high sparsity, with superior performance
  - SDPA output gating exhibit the lowest gating scores

- **Head-Specific Sparsity Matters.**

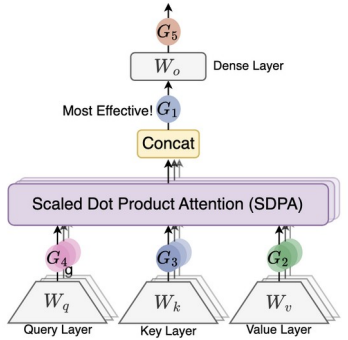
Method	Act-Func	Gate Score	M-Act	F-Attn	PPL	Hellaswag	MMLU	GSM8k
(1) Baseline	-	-	1053	0.467	6.026	73.07	58.79	52.92
(2) SDPA Elementwise Gate	Sigmoid	0.116	94	0.048	**5.761**	**74.64**	**60.82**	**55.27**
(3) SDPA Headwise Gate	Sigmoid	0.172	98	0.073	5.792	74.50	60.05	54.44
(4) SDPA Elementwise Head-shared Gate	Sigmoid	0.271	286	0.301	5.801	74.34	60.06	53.15
(5) v Elementwise Gate	Sigmoid	0.221	125	0.297	5.820	74.38	59.17	51.33
(6) SDPA Input Independent Gate	Sigmoid	0.335	471	0.364	5.917	73.64	59.02	52.40
(7) SDPA Elementwise Gate	NS-sigmoid	0.653	892	0.451	5.900	74.05	60.05	52.75



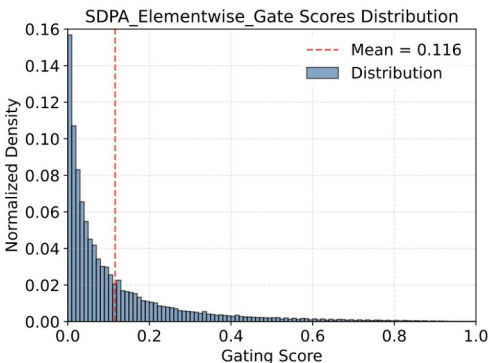
Why It Works?

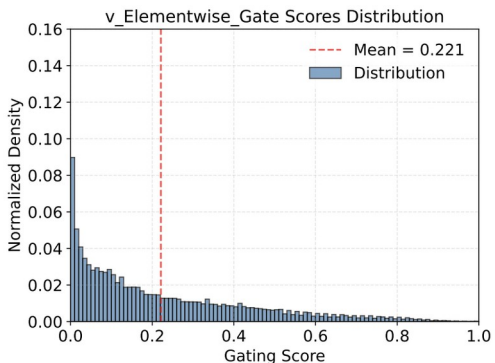
Gating Introduces Input-Dependent Sparsity (2/3)

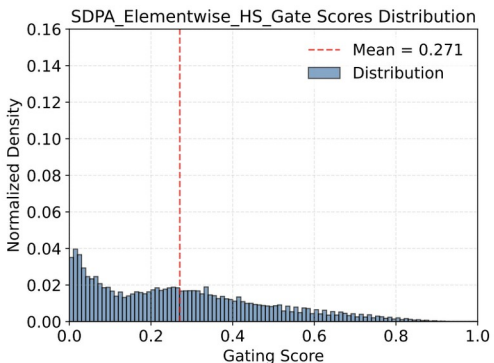
- **Why G1(SDP output gating) is better than G2(value gating)?**



Method	Act-Func	Gate Score	M-Act	F-Attn	PPL	Hellaswag	MMLU	GSM8k
(1) Baseline	-	-	1053	0.467	6.026	73.07	58.79	52.92
(2) SDPA Elementwise Gate	Sigmoid	0.116	94	0.048	**5.761**	**74.64**	**60.82**	**55.27**
(3) SDPA Headwise Gate	Sigmoid	0.172	98	0.073	5.792	74.50	60.05	54.44
(4) SDPA Elementwise Head-shared Gate	Sigmoid	0.271	286	0.301	5.801	74.34	60.06	53.15
(5) v Elementwise Gate	Sigmoid	0.221	125	0.297	5.820	74.38	59.17	51.33
(6) SDPA Input Independent Gate	Sigmoid	0.335	471	0.364	5.917	73.64	59.02	52.40
(7) SDPA Elementwise Gate	NS-sigmoid	0.653	892	0.451	5.900	74.05	60.05	52.75







Why It Works?

Gating Introduces Input-Dependent Sparsity (2/3)

- **Query-Dependency Matters (2/3)**

G1 $$o_i^k = \text{Non-Linearity-Map}\left(\sum_{j=0}^{i} S_{ij}^k \cdot X_j W_V^k\right) W_O^k$$

Depends on Xi

G2 $$o_i^k = \left(\sum_{j=0}^{i} S_{ij}^k \cdot \text{Non-Linearity-Map}(X_j W_V^k)\right) W_O^k$$

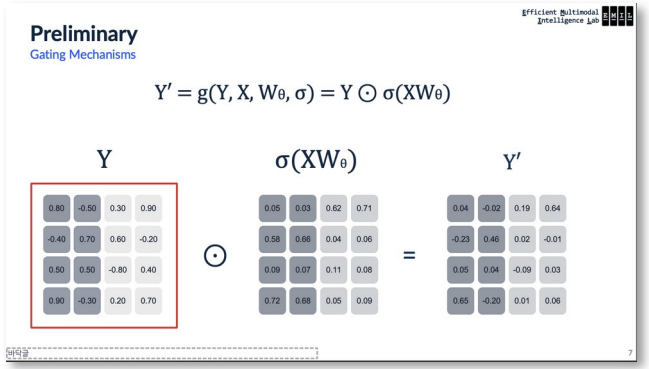
Depends on Xj

Why It Works?

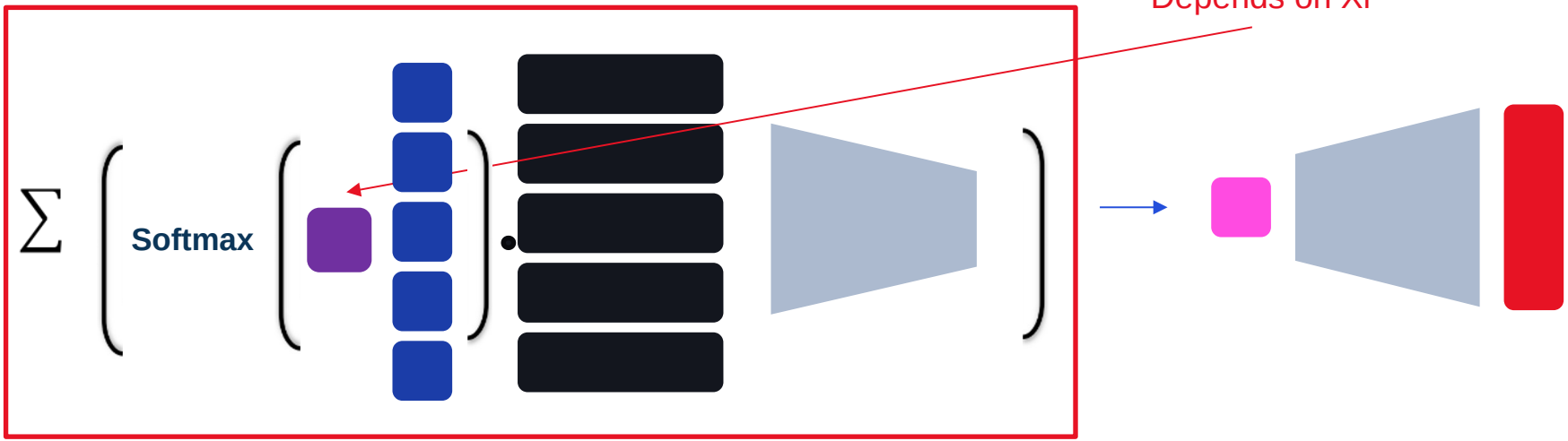
Gating Introduces Input-Dependent Sparsity (2/3)

- **Query-Dependency Matters (2/3)**

G1 $$o_i^k = \text{Non-Linearity-Map}\left(\sum_{j=0}^{i} S_{ij}^k \cdot X_j W_V^k\right) W_O^k$$



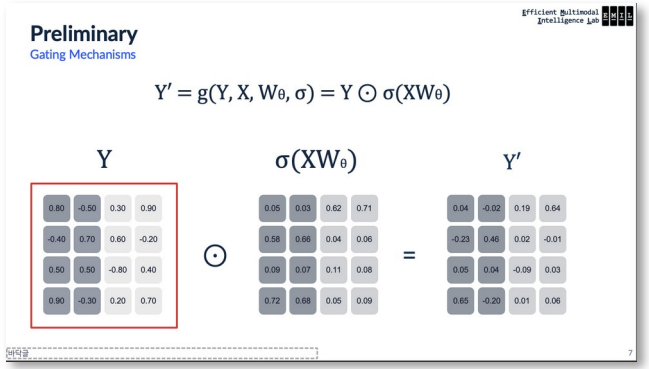
Depends on Xi



Why It Works?

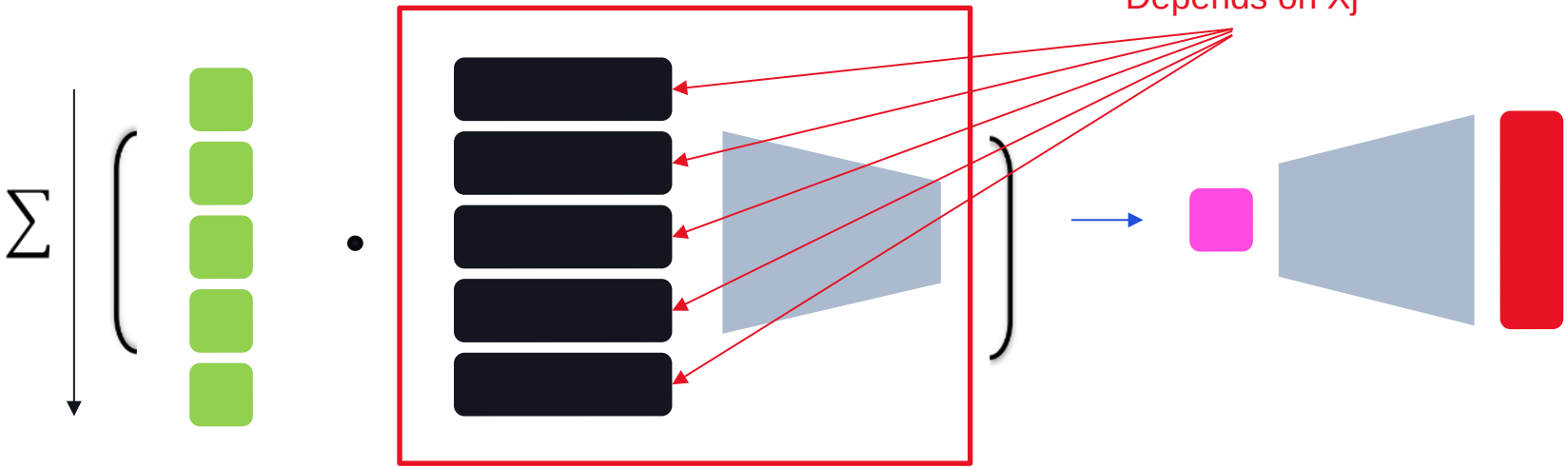
Gating Introduces Input-Dependent Sparsity (2/3)

- **Query-Dependency Matters (2/3)**



G2 $o_i^k = \left(\sum_{j=0}^{i} S_{ij}^k \cdot \text{Non-Linearity-Map}(X_j W_V^k) \right) W_O^k$

Depends on Xj



Why It Works?

Gating Introduces Input-Dependent Sparsity (2/3)

- **Query-Dependency Matters (2/3)**

G1 $$o_i^k = \text{Non-Linearity-Map}\left(\sum_{j=0}^{i} S_{ij}^k \cdot X_j W_V^k\right) W_O^k$$

G2 $$o_i^k = \left(\sum_{j=0}^{i} S_{ij}^k \cdot \text{Non-Linearity-Map}(X_j W_V^k)\right) W_O^k$$

Depends on Xi(Query Dependent)

Depends on Xj(Key, Value dependent)

- ***gating score sparsity may filter out irrelevant contextual information for the query!***

$Y = X \odot \sigma(XW_\theta)$ -> $Y = X \odot \sigma(P)$, P is zero-initializing learnable parameters

Method	Act-Func	Gate Score	M-Act	F-Attn	PPL	Hellaswag	MMLU	GSM8k
(1) Baseline	-	-	1053	0.467	6.026	73.07	58.79	52.92
(2) SDPA Elementwise Gate	Sigmoid	0.116	94	0.048	5.761	**74.64**	**60.82**	**55.27**
(3) SDPA Headwise Gate	Sigmoid	0.172	98	0.073	5.792	74.50	60.05	54.44
(4) SDPA Elementwise Head-shared Gate	Sigmoid	0.271	286	0.301	5.801	74.34	60.06	53.15
(5) v Elementwise Gate	Sigmoid	0.221	125	0.297	5.820	74.38	59.17	51.33
(6) SDPA Input Independent Gate	Sigmoid	0.335	471	0.364	5.917	73.64	59.02	52.40
(7) SDPA Elementwise Gate	NS-sigmoid	0.653	892	0.451	5.900	74.05	60.05	52.75

Why It Works?

Gating Introduces Input-Dependent Sparsity (3/3)

- **Less Sparse Gating is Worse (3/3)**
  - NS-sigmoid(x) = 0.5 + 0.5 · sigmoid(x) -> score range: [0.5, 1.0]

Method	Act-Func	Gate Score	M-Act	F-Attn	PPL	Hellaswag	MMLU	GSM8k
(1) Baseline	-	-	1053	0.467	6.026	73.07	58.79	52.92
(2) SDPA Elementwise Gate	Sigmoid	0.116	94	0.048	**5.761**	**74.64**	**60.82**	**55.27**
(3) SDPA Headwise Gate	Sigmoid	0.172	98	0.073	5.792	74.50	60.05	54.44
(4) SDPA Elementwise Head-shared Gate	Sigmoid	0.271	286	0.301	5.801	74.34	60.06	53.15
(5) v Elementwise Gate	Sigmoid	0.221	125	0.297	5.820	74.38	59.17	51.33
(6) SDPA Input Independent Gate	Sigmoid	0.335	471	0.364	5.917	73.64	59.02	52.40
(7) SDPA Elementwise Gate	NS-sigmoid	0.653	892	0.451	5.900	74.05	60.05	52.75

Why It Works?

SDPA Output Gating Reduces Attention-Sink

- **ttention Sink**
  - ( ) attention

$$S_{ij} = \frac{exp(q_i \cdot \kappa_j / \sqrt{a_k})}{\sum_l exp(q_i \cdot k_l / \sqrt{d_k})}$$

$$\sum_j S_{ij} = 1$$

1

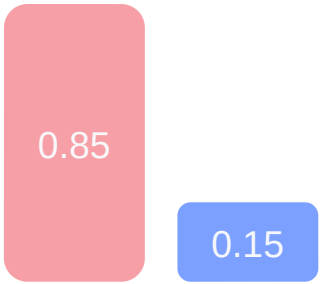
Why It Works?

SDPA Output Gating Reduces Attention-Sink

- **ttention Sink**
  - ( ) attention

$$S_{ij} = \frac{exp(q_i \cdot \kappa_j / \sqrt{a_k})}{\sum_l exp(q_i \cdot k_l / \sqrt{d_k})}$$

$$\sum \quad = 1$$



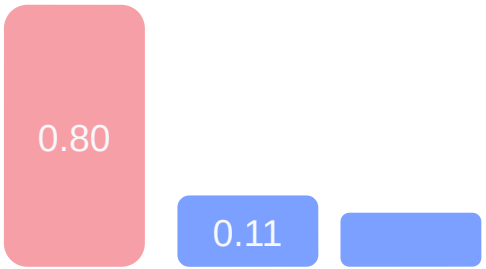
Why It Works?

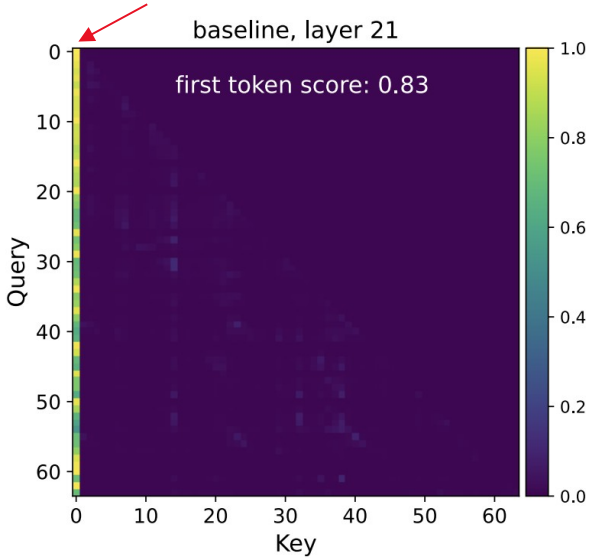
SDPA Output Gating Reduces Attention-Sink

- **ttention Sink**
  - (BOS ,) attention

$$S_{ij} = \frac{exp(q_i \cdot \kappa_j / \sqrt{a_k})}{\sum_l exp(q_i \cdot k_l / \sqrt{d_k})}$$

$$\sum_j S_{ij} = 1$$





Why It Works?

Attention-Sink-Free

Insight from sparsity: ***gating score sparsity may filter out irrelevant contextual information for the query!***

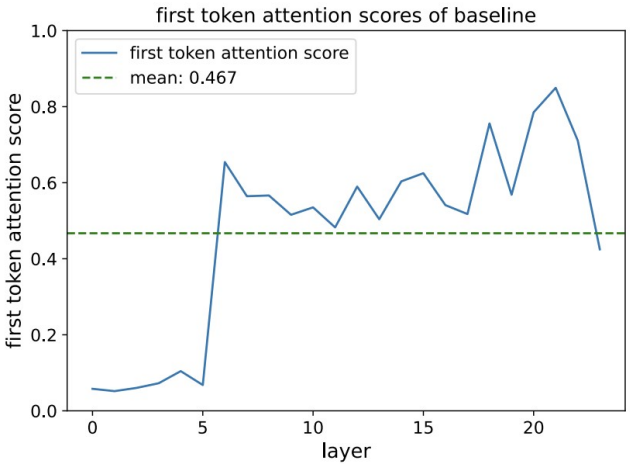
Hypothesis: ***Thereby mitigating the attention sink***

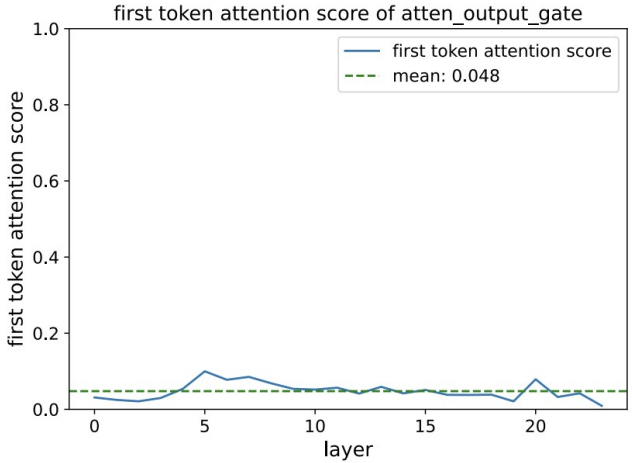
Previous Discussion: massive activation in hidden states and attention sinks (Sun et al., 2024)

Why It Works?

Attention-Sink-Free

- **Observation (1/3)**
- Gate -> attention score allocated to the first token reduced and decreases massive activations





Why It Works?

Attention-Sink-Free

• Observation (1/3)

• Gate -> attention score allocated to the first token reduced and decreases massive activations

Method	Act-Func	Gate Score	M-Act	F-Attn	PPL	Hellaswag	MMLU	GSM8k
(1) Baseline	-	-	1053	0.467	6.026	73.07	58.79	52.92
(2) SDPA Elementwise Gate	Sigmoid	0.116	94	0.048	**5.761**	**74.64**	**60.82**	**55.27**
(3) SDPA Headwise Gate	Sigmoid	0.172	98	0.073	5.792	74.50	60.05	54.44
(4) SDPA Elementwise Head-shared Gate	Sigmoid	0.271	286	0.301	5.801	74.34	60.06	53.15
(5) v Elementwise Gate	Sigmoid	0.221	125	0.297	5.820	74.38	59.17	51.33
(6) SDPA Input Independent Gate	Sigmoid	0.335	471	0.364	5.917	73.64	59.02	52.40
(7) SDPA Elementwise Gate	NS-sigmoid	0.653	892	0.451	5.900	74.05	60.05	52.75

Why It Works?

Attention-Sink-Free

- **Observation (2/3)**
- Head-shared Gate or the value gate (G2) -> massive activations ↓, but does not reduce attention scores to the first token

Method	Act-Func	Gate Score	M-Act	F-Attn	PPL	Hellaswag	MMLU	GSM8k
(1) Baseline	-	-	1053	0.467	6.026	73.07	58.79	52.92
(2) SDPA Elementwise Gate	Sigmoid	0.116	94	0.048	**5.761**	**74.64**	**60.82**	**55.27**
(3) SDPA Headwise Gate	Sigmoid	0.172	98	0.073	5.792	74.50	60.05	54.44
(4) SDPA Elementwise Head-shared Gate	Sigmoid	0.271	286	0.301	5.801	74.34	60.06	53.15
(5) v Elementwise Gate	Sigmoid	0.221	125	0.297	5.820	74.38	59.17	51.33
(6) SDPA Input Independent Gate	Sigmoid	0.335	471	0.364	5.917	73.64	59.02	52.40
(7) SDPA Elementwise Gate	NS-sigmoid	0.653	892	0.451	5.900	74.05	60.05	52.75

Insight: massive activations are not a prerequisite for attention sinks.

Why It Works?

Attention-Sink-Free

- **Observation (3/3)**
- Reducing the input-dependence of gating or using NS-sigmoid to reduce sparsity intensifies both massive activations and attention sink.

Method	Act-Func	Gate Score	M-Act	F-Attn	PPL	Hellaswag	MMLU	GSM8k
(1) Baseline	-	-	1053	0.467	6.026	73.07	58.79	52.92
(2) SDPA Elementwise Gate	Sigmoid	0.116	94	0.048	**5.761**	**74.64**	**60.82**	**55.27**
(3) SDPA Headwise Gate	Sigmoid	0.172	98	0.073	5.792	74.50	60.05	54.44
(4) SDPA Elementwise Head-shared Gate	Sigmoid	0.271	286	0.301	5.801	74.34	60.06	53.15
(5) v Elementwise Gate	Sigmoid	0.221	125	0.297	5.820	74.38	59.17	51.33
(6) SDPA Input Independent Gate	Sigmoid	0.335	471	0.364	5.917	73.64	59.02	52.40
(7) SDPA Elementwise Gate	NS-sigmoid	0.653	892	0.451	5.900	74.05	60.05	52.75

Why It Works?

Attention-Sink-Free

- **Observation**
- 1. Gate -> massive activation ↓, attention sink ↓
- 2. Head-Shared Gate, G2 gate -> massive activation ↓, not (attention sink ↓)
  massive activations are not a prerequisite for attention sinks.
- 3. Input-independent, NS-sigmoid(not sparse) -> bad(massive activation ↑↑↑, attention sink ↑↑↑)

- **Insight**
- 1. validate hypothesis: input-dependent, head-specific gating -> significant sparsity ->attention sink ↓
- 2. sparsity ↑ -> massive activations ↓ -> training stability ↑ by reducing massive activations, the model is less susceptible to numerical errors during BF16 training

Limitations

- (non-linearity) attention (dynamics) (training process)
- ttention sink - , long context
  - .
  .
- .

Thank You!