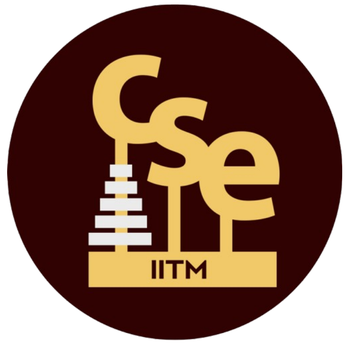
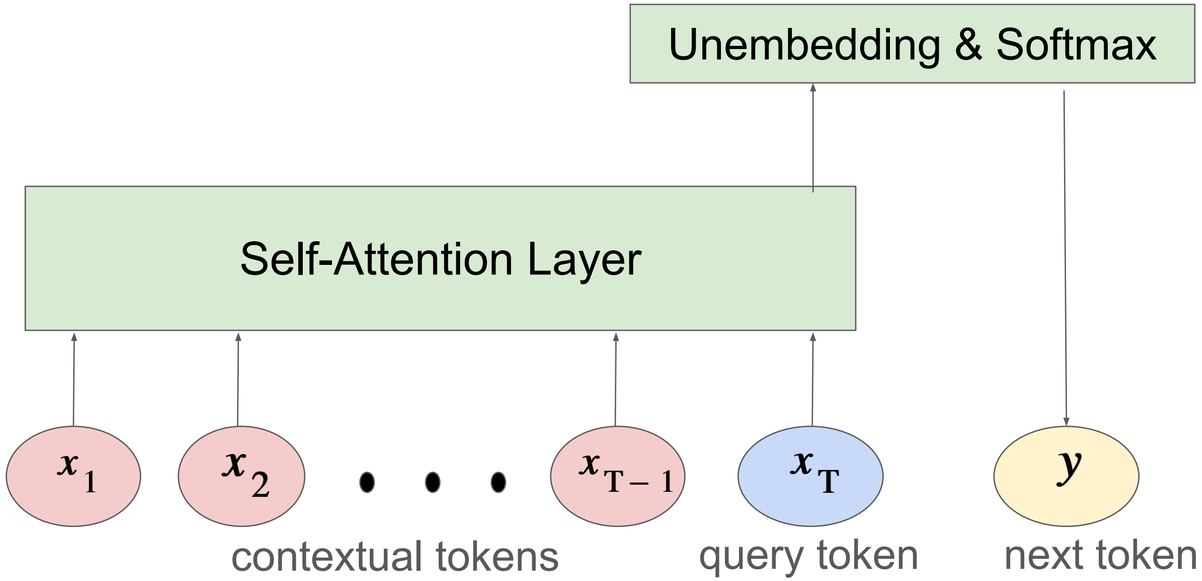


Faster Query-Key Learning Sharpens Attention in Self-Attention Models

Rahul Vashisht, Harish G. Ramaswamy

ICML 2026

Setup: Two interacting circuits in self-attention



QK Circuit

Controls how attention is assigned across tokens.

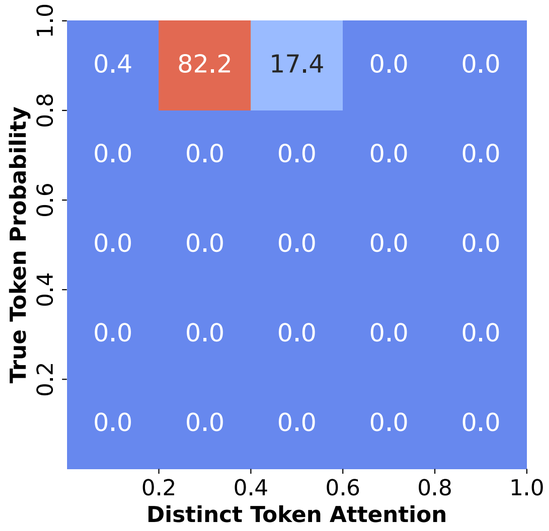
OV Circuit

Maps attended information to predictions.

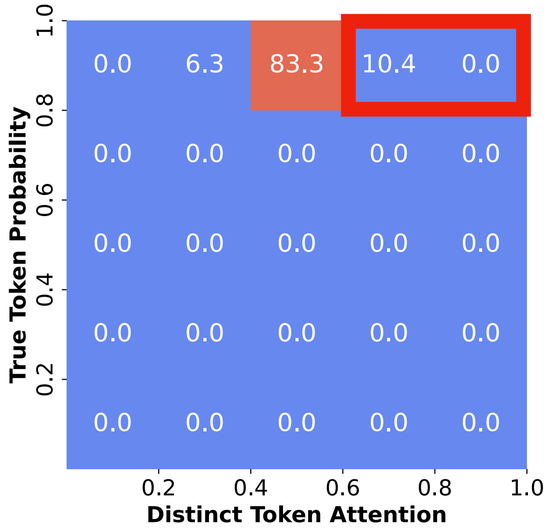
Our Focus: How does the relative learning speed of QK and OV circuits shape the final attention pattern?

Parameterization & Learning-rate Interventions

QK and OV circuits can be trained as factorized matrices ($W_Q W_K$, $W_O W_V$) or collapsed single matrices (W_{QK}, W_{OV}).

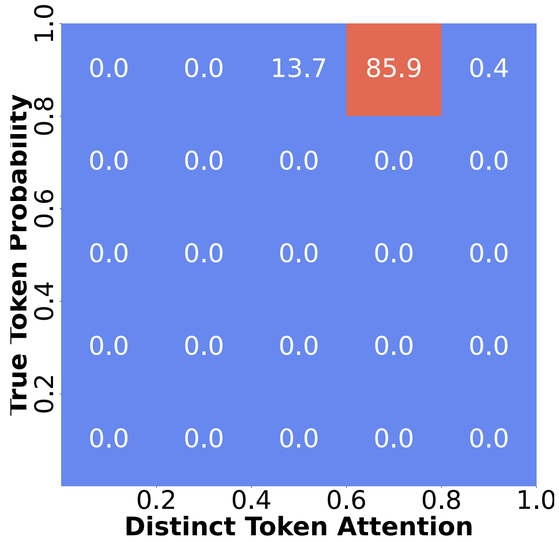


Factorized Attention Factorized Output (FAFO)

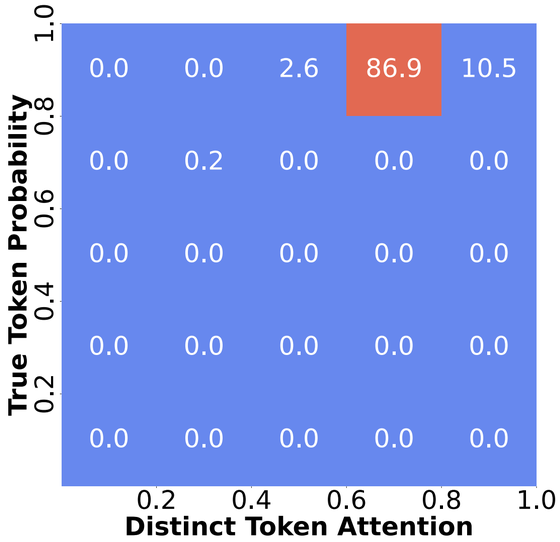


Collapsed Attention Collapsed Output (CACO)

Fraction of Instances with high relevant token attention and correct prediction



FAFO linearly increasing QK learning rate



CACO linearly increasing QK learning rate

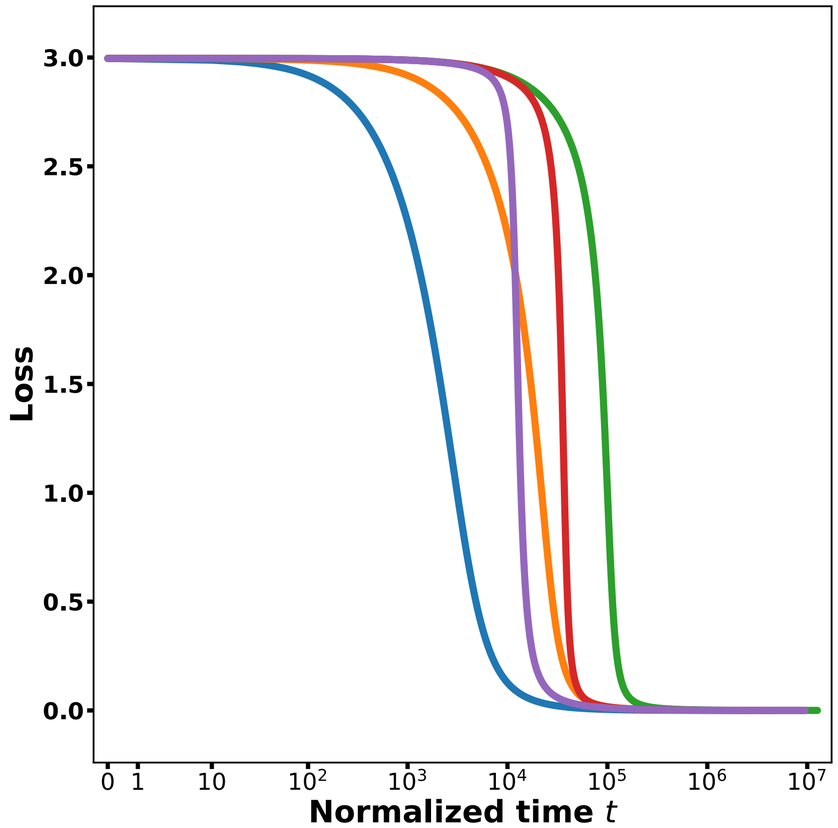
Parameterization Effect: different parameterizations lead to different attention patterns.

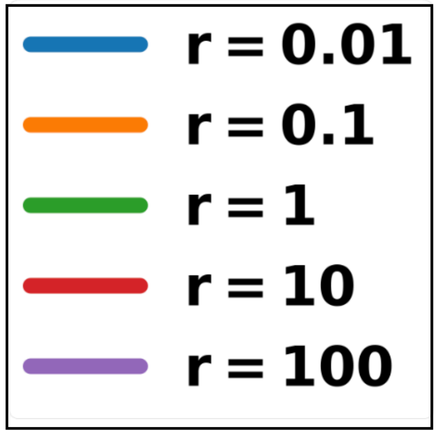
Learning-rate intervention: faster QK learning produces a similar attention-sharpening effect.

Informal Theorem: factorized training can be approximated by collapsed training with an increasing learning rate.

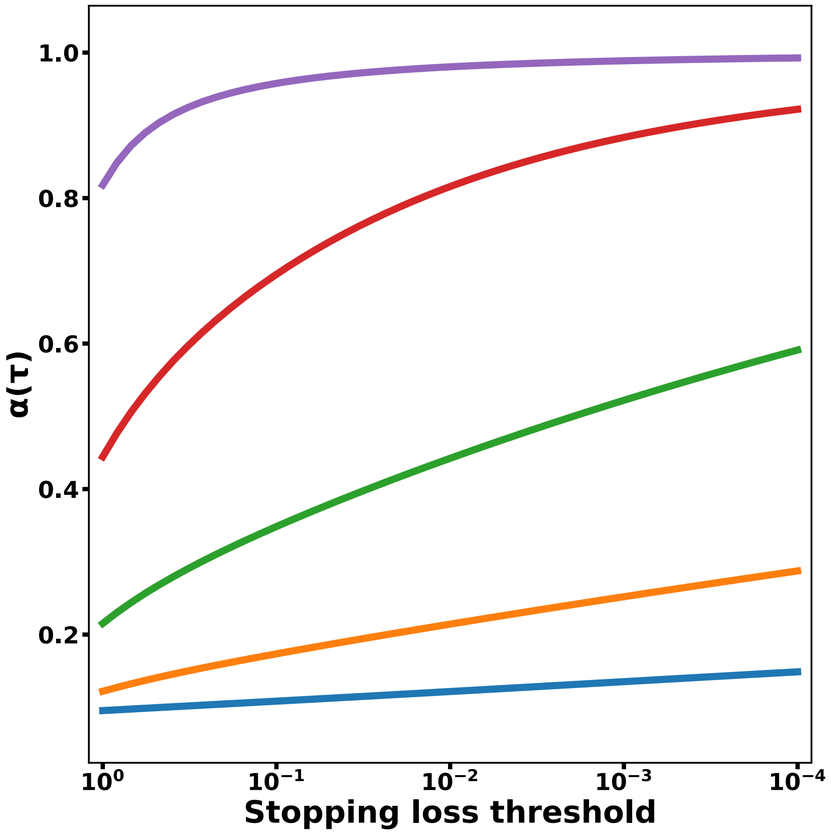
Theory & Learning Dynamics

Larger $r = \dfrac{\eta_{QK}}{\eta_{OV}}$ leads to a larger QK scale and higher attention mass on relevant tokens.





$\alpha(\tau)$: total distinct token attention



If OV circuit learns faster (r<<1): predictions can improve even before model attends to relevant tokens.

If QK learns faster (r>>1): the model compensates by concentrating attention on relevant tokens.

Real-Data Validation & Takeaway

Across SQuAD, SVA, and HateXplain: task performance remains comparable, while attention-based metrics improve.

Dataset	Setting	Test Perf.	AC	Comp.
SQuAD QA	Baseline Faster QK	56.11 56.86	12.65 **42.95**	0.38 **0.53**
SVA	Baseline Faster QK	92.43 93.41	41.67 **76.85**	0.24 **0.30**
HateXplain	Baseline Faster QK	55.74 56.90	0.00 **43.50**	0.40 **0.48**

Attention Metrics

Attention Confidence (AC): measures attention concentration on relevant tokens.

Comprehensiveness: measures the drop in prediction probability after removing important tokens.

Key Takeaway

Training choices can affect learned attention patterns without changing the model.

Thanks

Any Questions?

rahul@cse.iitm.ac.in

Paper

Code