Fig 1: NIAH task filled with repeative noise sentences

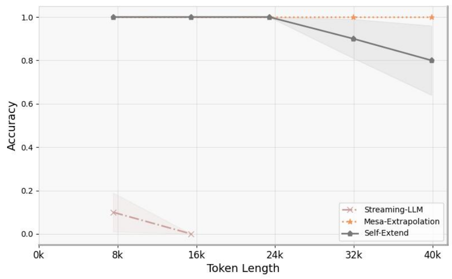


Fig 2: NIAH task filled with Paul Graham essays

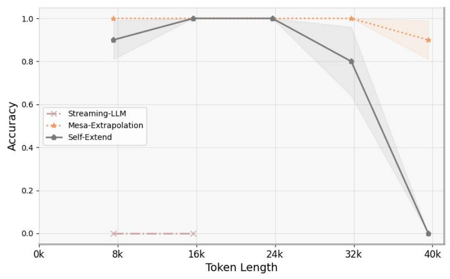


Fig 3: refer from Figure 2 with added Self-Extend

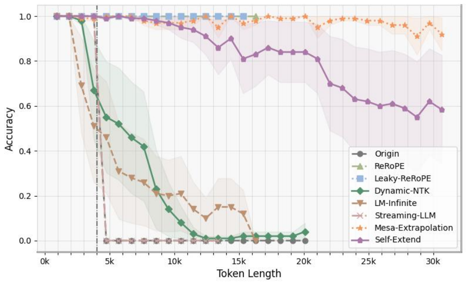


Fig 4: NIAH tasks with single-multi keys using Phi3-mini-128k for Origin & ours

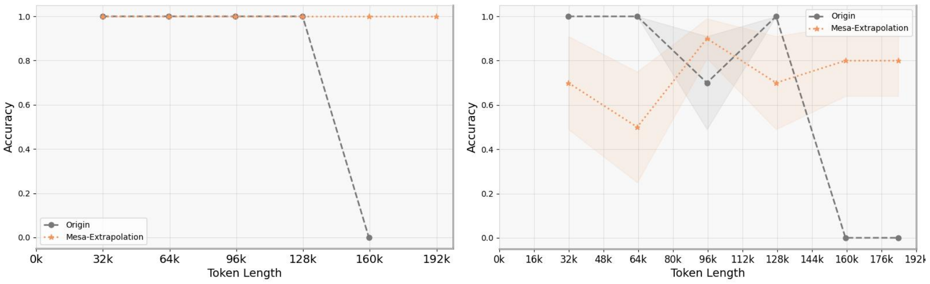


Fig 5: Memory Usage & Latency on Open-LLaMA-3B model

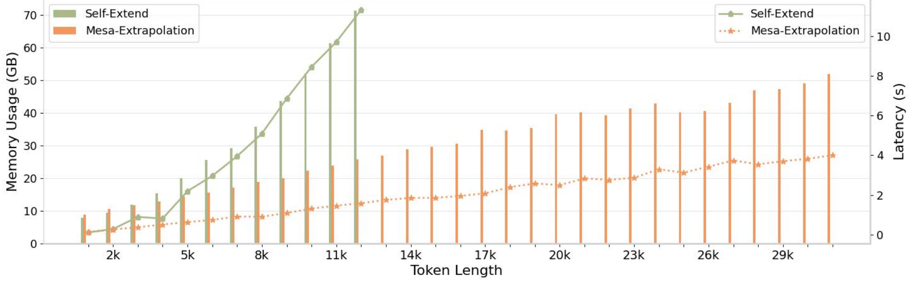
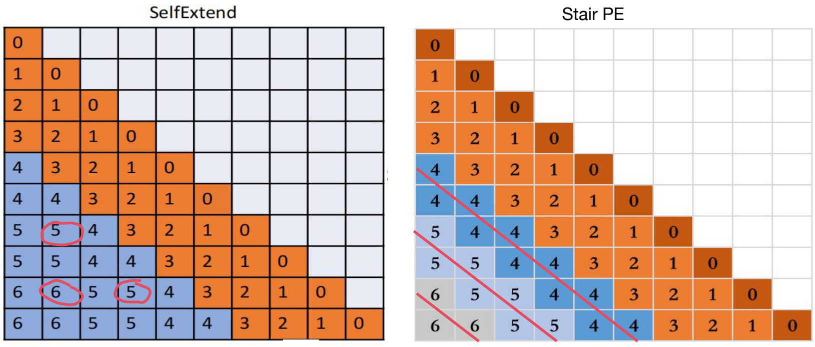


Fig 6: attention matrix (before Softmax operation) compared between SelfExtend and Stair PE



SelfExtend

0									
1	0								
2	1	0							
3	2	1	0						
4	3	2	1	0					
4	4	3	2	1	0				
5	5	4	3	2	1	0			
5	5	4	4	3	2	1	0		
6	6	5	5	4	3	2	1	0	
6	6	5	5	4	4	3	2	1	0

Stair PE

0									
1	0								
2	1	0							
3	2	1	0						
4	3	2	1	0					
4	4	3	2	1	0				
5	4	4	3	2	1	0			
5	5	4	4	3	2	1	0		
6	5	5	4	4	3	2	1	0	
6	6	5	5	4	4	3	2	1	0

Fig 7: Dim 14 and 19 are observed with “good” probe

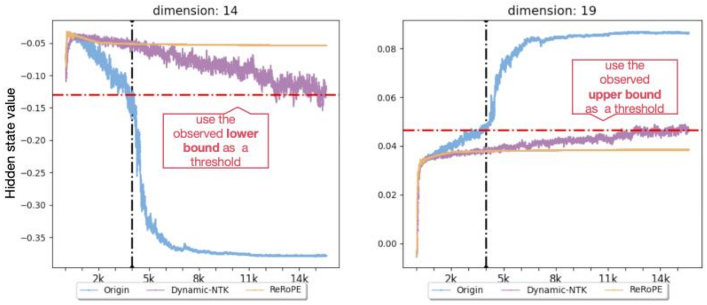
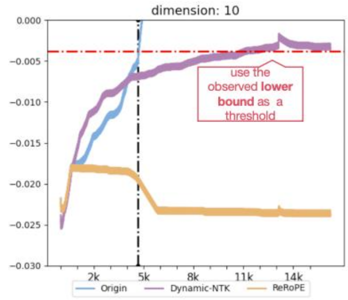


Fig 8: Dim 10 is observed with same passkey input.


dimension: 10
use the observed lower bound as a threshold
2k
5k
8k
11k
14k
Origin
Dynamic-NTK
ReRoPE