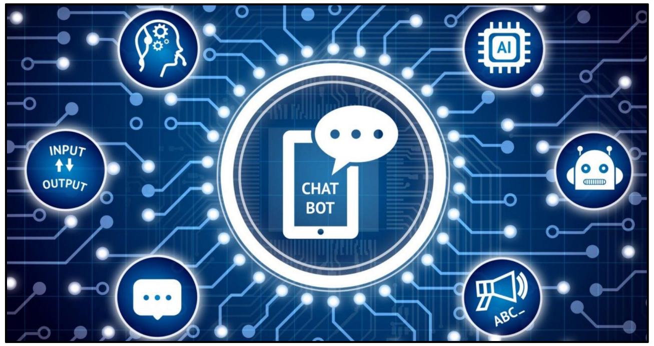
Mini-Sequence Transformer: Optimizing Intermediate Memory for Long Sequences Training

Cheng Luo[1], Jiawei Zhao[2], Zhuoming Chen[3], Beidi Chen[3], Anima Anandkumar [1]

California Institute of Technology[1], Meta Fair[2], Carnegie Mellon University[3]

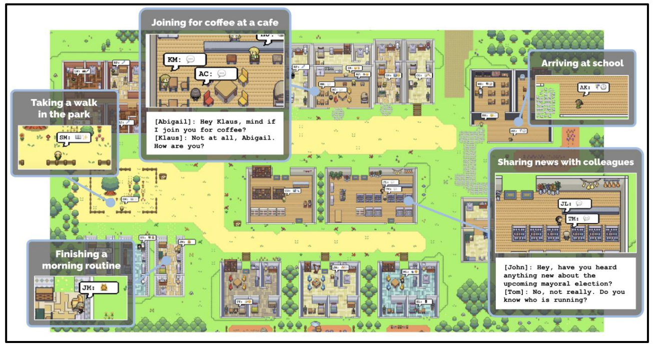
wdlctc@gmail.com
https://github.com/wdlctc/mini-s

Foundation Models

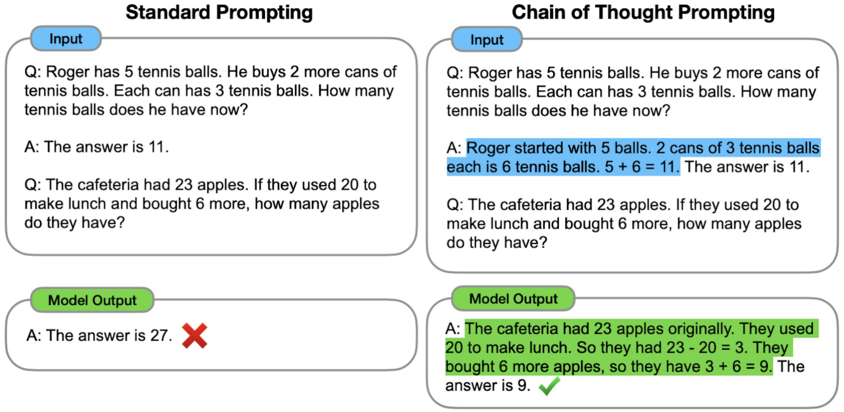


Conversational AI

Content Generation


Joining for coffee at a cafe
KM:
AC:
[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?
Taking a walk in the park
SM:
Arriving at school
AK:
Sharing news with colleagues
JL:
TM:
[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?
Finishing a morning routine
JM:

AI Agents



Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Reasoning

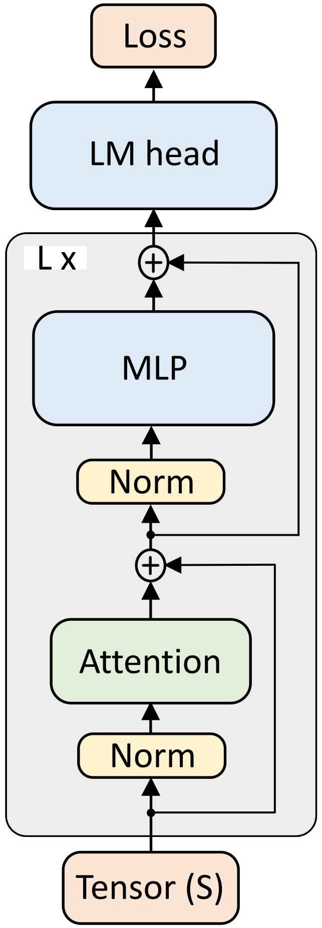
Planning

Evolution of Foundation Models
LLama 3 vs. LLaMa 2 - A Head-to-Head Comparison

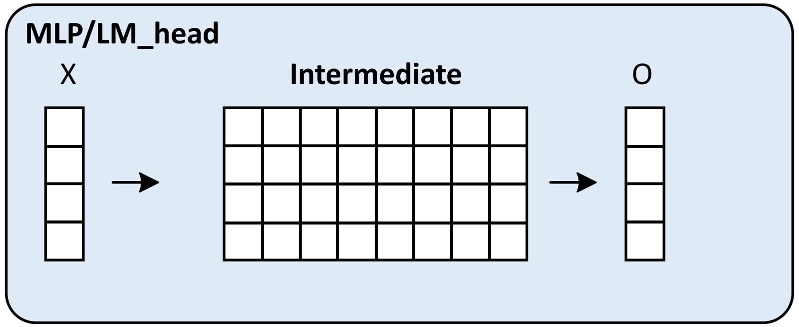
Feature	LLAMA2-7b	LLAMA3-8b
Hidden Length	4096	8192
Sequence Length	4096	**8192**
MLP Intermediate	11008	**14336**
Vocal Length	32000	**128256**

Evaluation of foundataion model leads to larger Vocal Length & MLP Intermediate

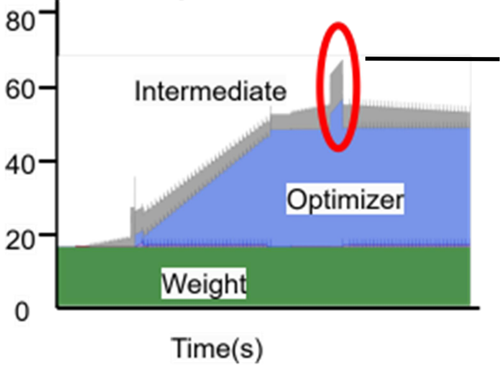
Mini-Sequence Transformer Yield Memory Saving & Long Sequence Enable



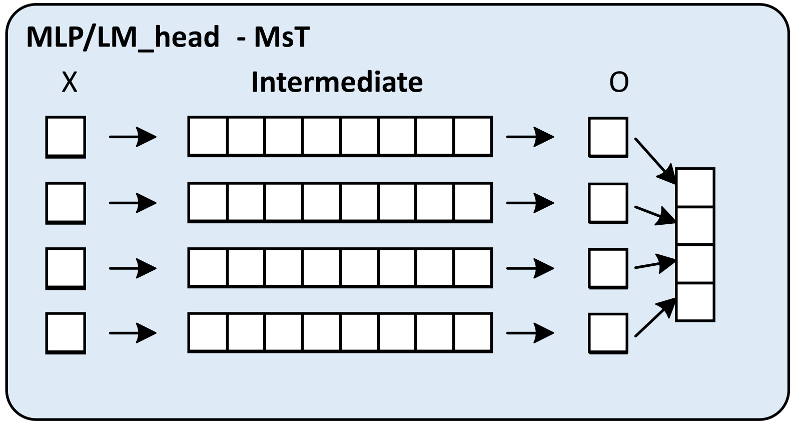
MLP/LM-head Computation Flow
Extremely Large Intermediate



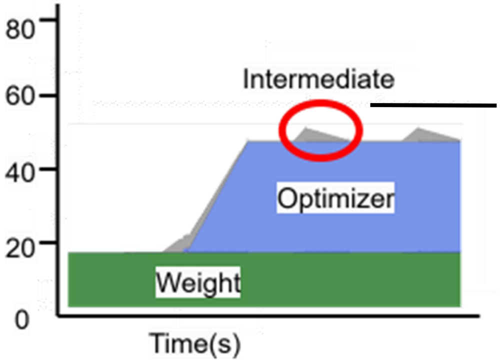
Llama-3 Sandard Implementation:
Peak Memory:67GB



Challenge:
Native implementation brought memory bottleneck



Llama-3 MST Implementation:
Peak Memory:47GB



Gain: Support 12x-24 longer Sequence Length

MsT: Splitting MLP/LM-Head Input
Concat output

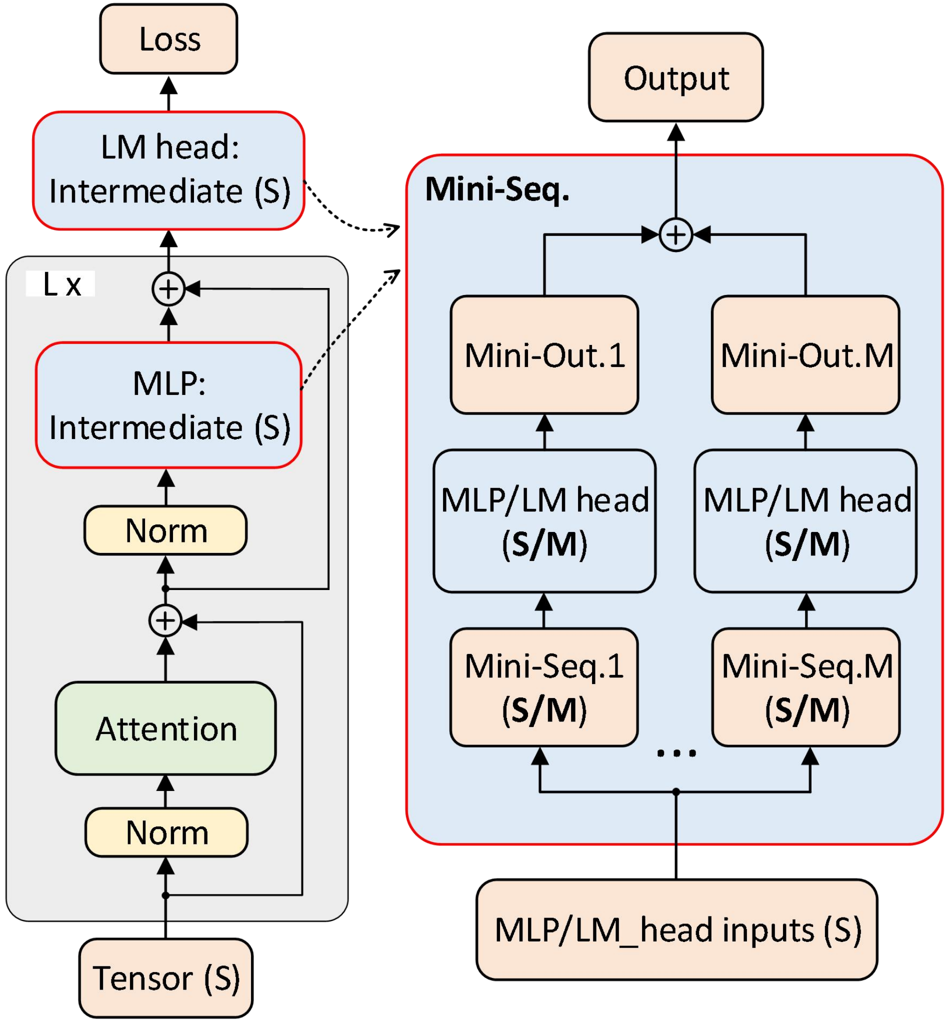
Mini-Sequence Transformer: Key Ideas

Key Ideas: Input Partition & Gradient Accumulation

Input Partition: Split the input sequence into Mini-sequence

Gradient Accumulation: Restore the full output/Gradients

Results: optimizing intermediate values, create more space for long sequence activation



Transformer -> Mini Sequence Transformer

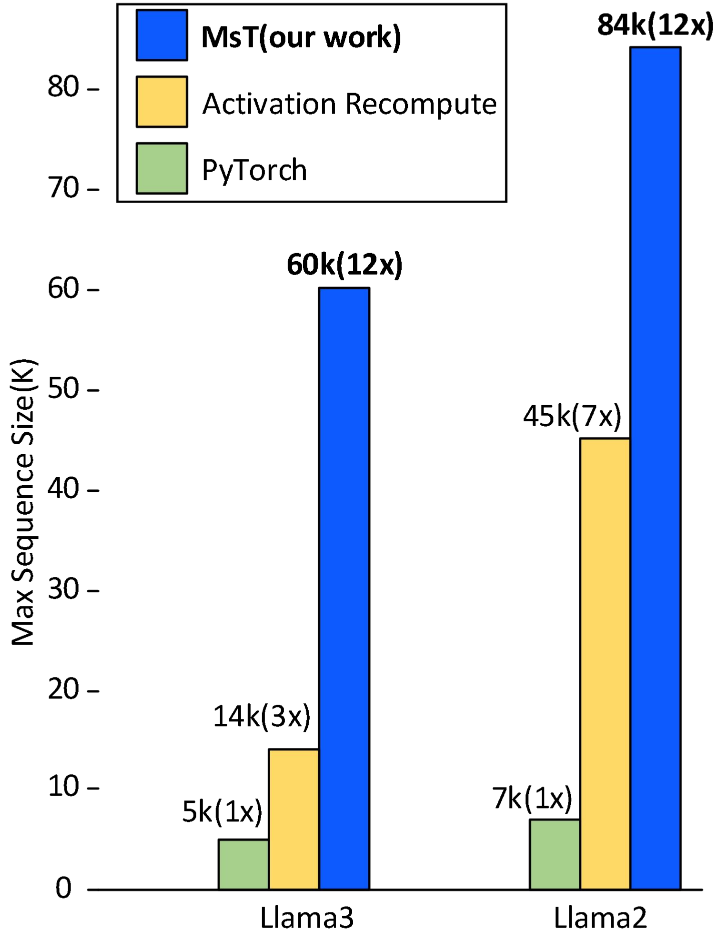
Mini-Sequence Transformer (MsT): 12-24x sequence & TFLOPS equivalence

12-24x longer sequence enable

Models	Maximum Sequence (K)		
	Vanilla	MST	Extend
LLAMA3-8b	5	**60**	**12x**
LLAMA2-78	7	**84**	**12x**
Qwen2-7B	4	**74**	**18x**
gemma-2-9b	1.5	**36**	**24x**

Equal TFLOPS for Llama3-8B

Models	TFLOPS		
	Vanilla	MST	Speedup
LLAMA3-8b	3271	**3386**	**1.02x**
LLAMA2-78	3290	**3656**	**1.11x**


MsT(our work)
Activation Recompute
PyTorch
Max Sequence Size(K)
5k(1x)
14k(3x)
60k(12x)
7k(1x)
45k(7x)
84k(12x)
Llama3
Llama2