Reasoning with Latent Thoughts: On the Power of Looped Transformers

Google Research

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, Sashank J. Reddi

TTIC

Introduction

What model architectures are good for reasoning?

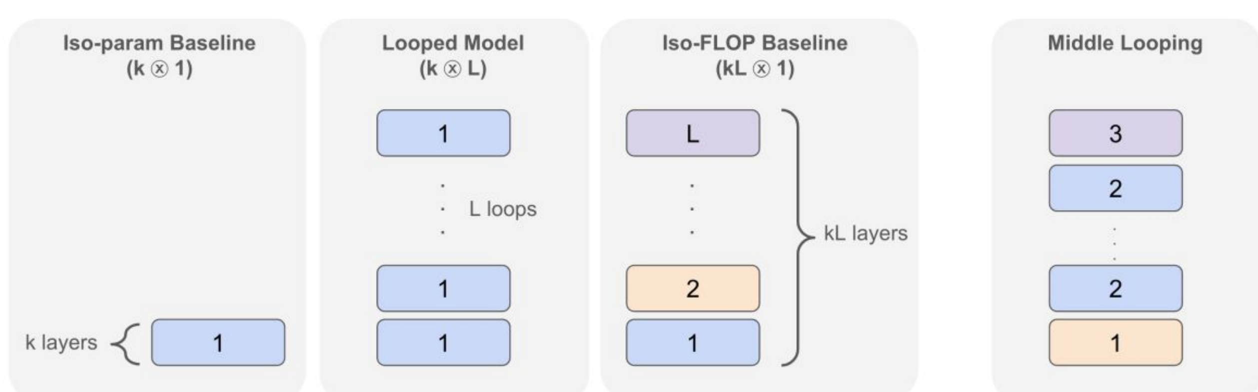
Our claim: Looped models are well-suited for reasoning!

Prior work on looped models:

1. ALBERT, Universal Transformers - Parameter efficiency, Adaptive compute
2. Benefits for In-context learning - Yang et al 2023, Gatmiry et al. 2024.

Our work:

- On synthetic reasoning tasks, looped models match the performance of iso-flop non looped models.
- Looped language models show inductive bias towards reasoning → at same perplexity as a non-looped model, stronger reasoning performance.
- Looped model performance scales with the number of loops in a predictable manner - scaling law!
- Looped models can generate *Latent Thoughts* and can, in theory, simulate CoT reasoning.



Synthetic Reasoning Tasks

Looping matches iso-flop model!

- i-GSM: Synthetically constructed GSM (Ye et al. 2024).
  E.g. E#I := 4. A#B := E#I + J#K. J#K := E#I + 3. F#G := A#B + J#K. F#G?
- p-hop (Sanford et al.): Sequence traversal task.
  E.g.: 2-hop

- n-ary addition: Adding n 3-digit numbers.
  Input: “315 + 120 + 045 + 824 =” ; Output = “1304”

	Params / FLOPs	Accuracy
Base ($8 \otimes 1$)	8x / 8x	**73.2**
1 layer model		
Base ($1 \otimes 1$)	1x / 1x	24.5
Loop ($1 \otimes 2$)	1x / 2x	52.3
Loop ($1 \otimes 4$)	1x / 4x	69.9
Loop ($1 \otimes 8$)	1x / 8x	**73.2**
2 layer model		
Base ($2 \otimes 1$)	2x / 2x	54.0
Loop ($2 \otimes 2$)	2x / 4x	66.9
Loop ($2 \otimes 4$)	2x / 8x	**73.6**
4 layer model		
Base ($4 \otimes 1$)	4x / 4x	71.3
Loop ($4 \otimes 2$)	4x / 8x	**71.6**

p-hop with n tokens

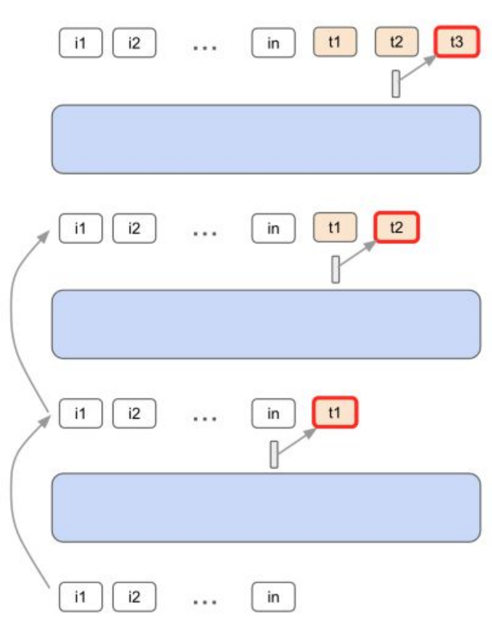
	Params / FLOPs	$p = 16$, $n = 256$	$p = 32$, $n = 256$
Base ($6 \otimes 1$)	6x / 6x	**99.9**	**99.6**
1 layer model			
Base ($1 \otimes 1$)	1x / 1x	48.9	49.0
Loop ($1 \otimes 6$)	1x / 6x	**99.9**	**99.5**
2 layer model			
Base ($2 \otimes 1$)	2x / 2x	68.8	59.4
Loop ($2 \otimes 3$)	2x / 6x	**99.9**	**99.8**
3 layer model			
Base ($3 \otimes 1$)	3x / 3x	97.2	73.0
Loop ($3 \otimes 2$)	3x / 6x	**99.9**	**99.5**

Common theme: All the above problems are solvable by a 1-layer model looped L times. Moreover, performance of a looped model strongly matches iso-flop non-looped model!

Theory: Expressivity of Looped Models

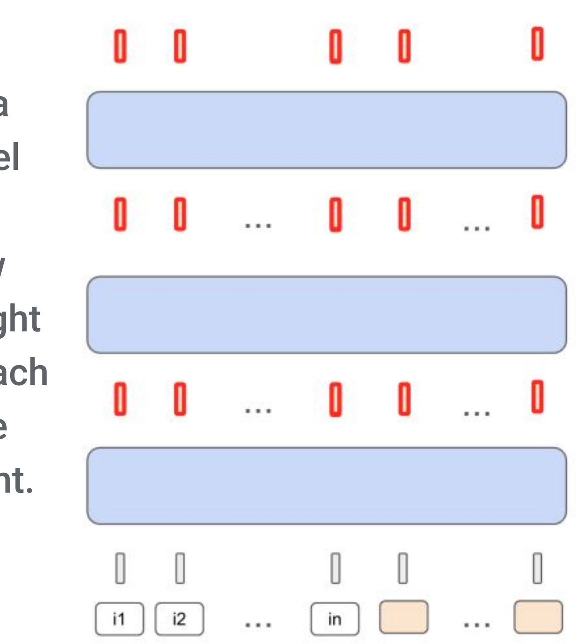
Our theoretical results

- *Theorem 1*: Any Transformer with L layers, d embedding size, d_{FF} hidden size can be simulated with a 1-layer transformer looped L times with d+L embedding size and Ld_{FF} hidden size.
- *Theorem 2*: (Looping can simulate CoT) Any L layer transformer on n length input generating m CoT tokens can be simulated by a looped model with L+O(1) distinct layers, O(log(n+m)) extra embedding size and by running m loops on the input concatenated with m dummy tokens.
- **Other results:** For specific problems such as p-hop and group composition, even more parameter optimal constructions of looped models. We get near depth optimal constructions for these problems.



CoT can be thought of as a looped model generating one new token in each loop.

In contrast, a looped model generates multiple new “latent thought tokens” in each loop - can be more efficient.



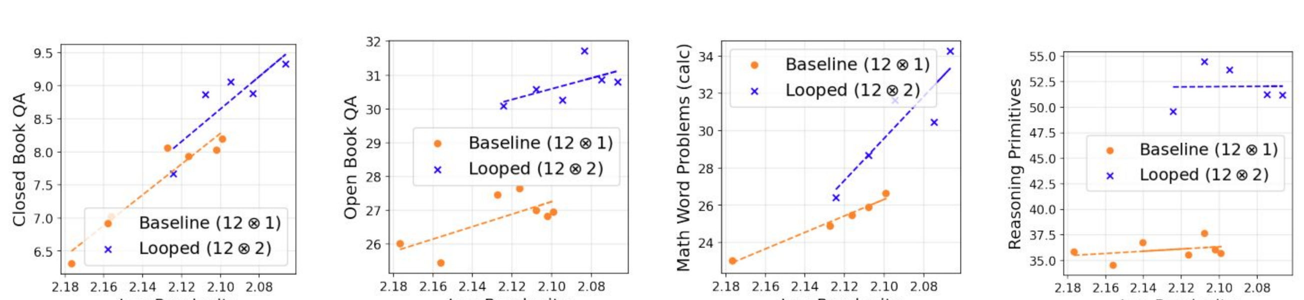
Language Modeling with Looped Models

Perplexity vs Reasoning task performance

Causal language modeling on 250B tokens from the Pile dataset at the 1B parameter scale.

Measured perplexity and downstream performance across 19 tasks.

Recall: k ⊗L is a k-layer model looped L times.



At same perplexity looped models achieve better downstream performance.

	Params / FLOPs	Perplexity (↓) (validation)	Closed Book QA (↑) (4 tasks)	Open Book QA (↑) (5 tasks)	Math Word Problems (↑) (6 tasks)	All Tasks Average (↑) (15 tasks)	Reasoning Primitives (↑) (4 tasks)
24 layers							
Baseline	24x / 24x	7.40	11.2	33.9	29.3	26.0	47.5
12 layers							
Base ($12 \otimes 1$)	12x / 12x	8.16	8.2	26.9	26.7	21.8	35.7
Loop ($12 \otimes 2$)	12x / 24x	7.90	9.3	30.8	34.3	26.5	51.2
% Gap		34 %	37 %	56 %	282 %	110 %	131 %
Middle Loop ($4 \otimes 1, 4, 1$)	12x / 24x	7.81	11.0	32.3	28.3	25.0	56.5
% Gap		46 %	94 %	78 %	62 %	95 %	176 %
8 layers							
Base ($8 \otimes 1$)	8x / 8x	8.75	6.3	22.7	17.1	16.1	33.0
Loop ($8 \otimes 3$)	8x / 24x	8.19	8.5	30.8	28.4	23.9	55.3
% Gap		41 %	44 %	72 %	92 %	78 %	153 %
6 layers							
Base ($6 \otimes 1$)	6x / 6x	9.25	4.0	19.3	17.7	14.6	24.1
Loop ($6 \otimes 4$)	6x / 24x	8.42	8.2	28.7	29.8	23.7	56.1
% Gap		44 %	58 %	64 %	104 %	80 %	136 %
4 layers							
Base ($4 \otimes 1$)	4x / 4x	10.12	1.8	13.8	9.7	9.0	19.4
Loop ($4 \otimes 6$)	4x / 24x	8.79	6.7	26.2	24.8	20.4	56.9
% Gap		48 %	52 %	61 %	77 %	67 %	133 %

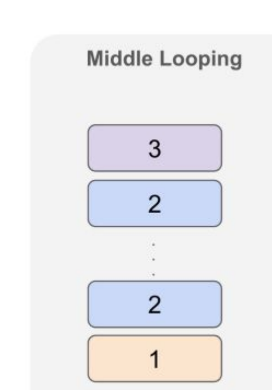
Language modeling results: The color highlighted percentage shows the amount of quality gap that looping a set of layers is able to cover compared to a non-looped model. It is smaller for perplexity and memorization intensive tasks, much higher for reasoning intensive tasks.

Middle Looping

Inspired by Gradual Stacking (Saunshi et al. 2024), we try looping only the middle layers in a Transformer stack.

Outperforms looping the entire model.

High level intuition: Starting and ending layers play a special role, thus need to be treated differently.



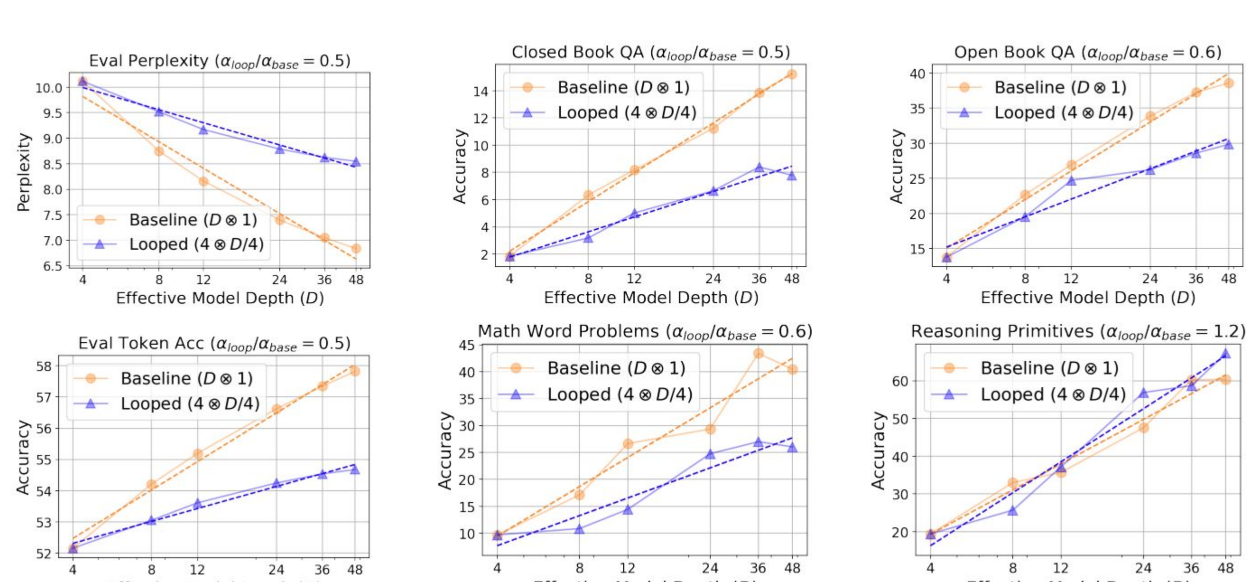
A Scaling Law for Looped Models

Observation: Performance of Looped models scales predictably with the number of loops.

Let D = kL be the effective depth of a k layer model looped L times.

Scaling law: Acc = αlog(D) + β

α = indicator of the impact of depth.



A Looping inspired Regularizer

To boost perplexity and yet retain the inductive bias wins for reasoning - we propose adding a cosine similarity regularizer between weights in different loops (rather than strict weight tying).

$$\mathcal{R}_G(k) = \frac{1}{L-k} \sum_{i=0}^{k-2} \sum_{j=0}^{k-1} \text{Cosine}\left(\theta_G^{(ik+j)}, \theta_G^{((i+1)k+j)}\right)$$

Gives the model flexibility to improve memorization and thereby recover the perplexity while retaining the benefits in downstream reasoning tasks.

Conclusion

Main Takeaways

- Inductive bias of looped models for stronger reasoning task performance in language models.
- Looping as a form of “latent thinking”, can subsume CoT in theory. Can also be combined with CoT potentially for orthogonal benefits.
- Propose variants to vanilla looping and looping inspired regularizers for stronger looped model performance.

Future Directions

- What other architectural inductive biases can we design for stronger human-like reasoning models?
- Can looped models help compress long explicit CoT chains into shorter latent chains?

References

- Dehghani et al. 2018. Universal Transformers.
- Fan et al. 2024.Looped Transformers for Length Generalization.
- Gatmiry et al. 2024. On the role of depth and looping for in-context learning with task diversity.
- Lan et al. 2020. Albert: A lite bert for self-supervised learning of language representations.
- Sanford et al. 2024. Transformers, parallel computation, and logarithmic depth.
- Saunshi et al. 2024. On the Inductive Bias of Stacking towards Improved Reasoning
- Yang et al. 2023. Looped transformers are better at learning learning algorithms.
- Ye et al. 2024. Physics of Language Models. Part 2.1.