T²MLR: Transformers With Temporal Middle-Layer Recurrence

Ziyang Cai*, Xingyu Zhu*, Yihe Dong, Yinghui He, Sanjeev Arora
Princeton University

PRINCETON
Language + Intelligence

06/05/2026 @ FAI Seminar

Outline

- Related Works: Latent Reasoning
- Motivation for Middle-Layer Temporal Recurrence
- Separation on Synthetic Experiments: S5-Retrival
- Pre-training T²MLR
- Future Directions

Related Works: Latent Reasoning

- Information Bottleneck in Autoregressive Transformers:

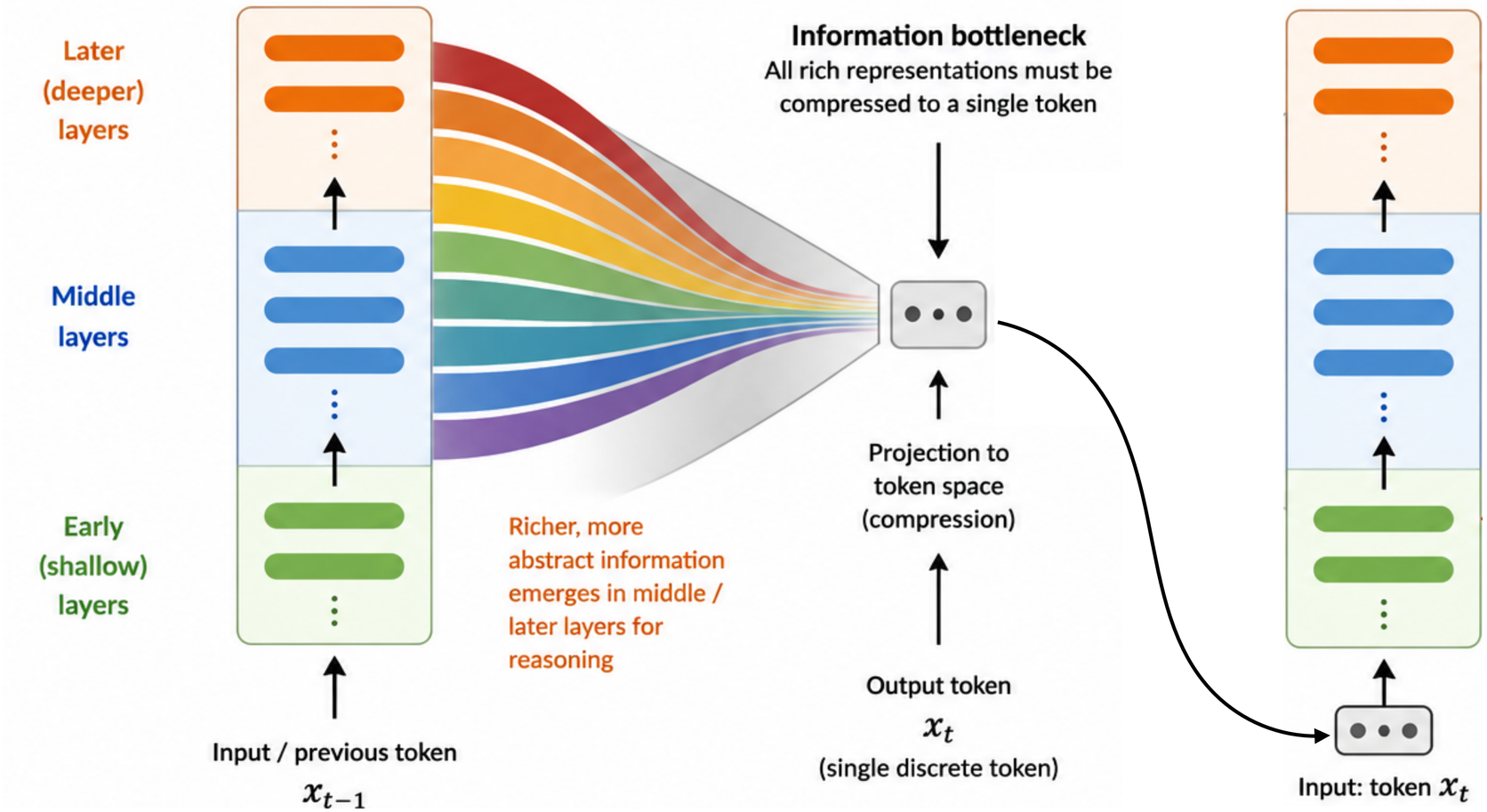
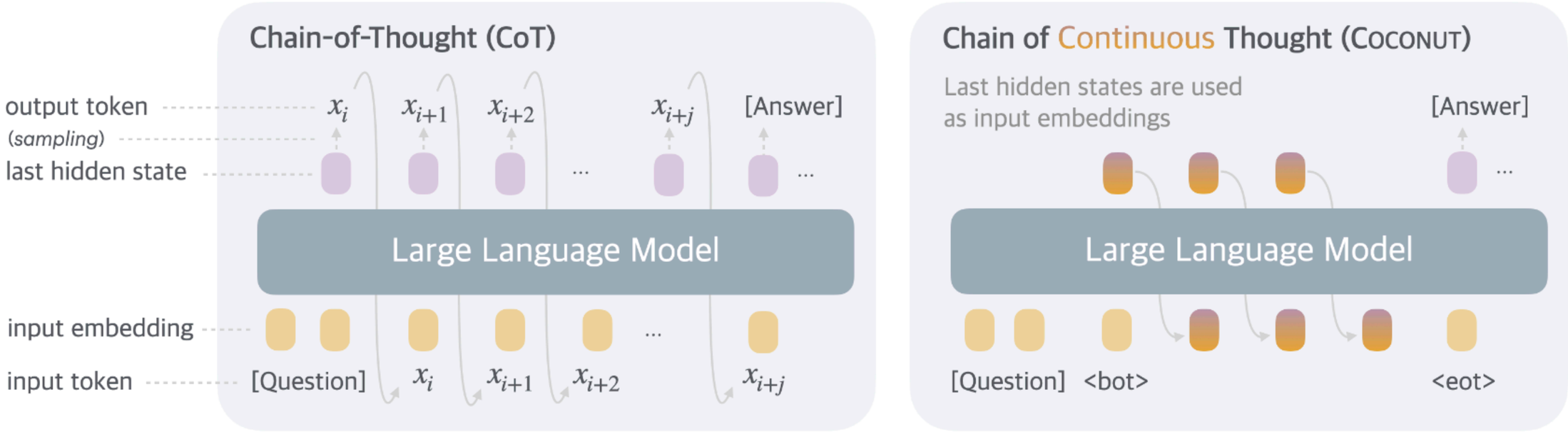


Figure Generated by GPT-Image-2

Background: Latent Reasoning

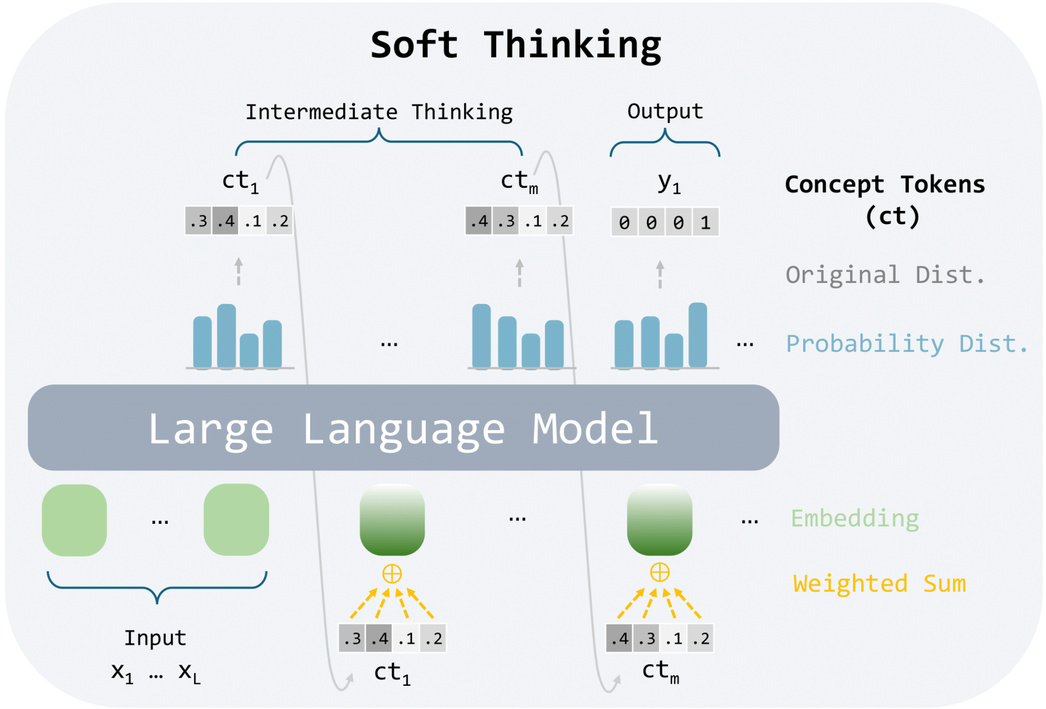
- Relaxation by embedding-level temporal recurrence: COCONUT / CODI

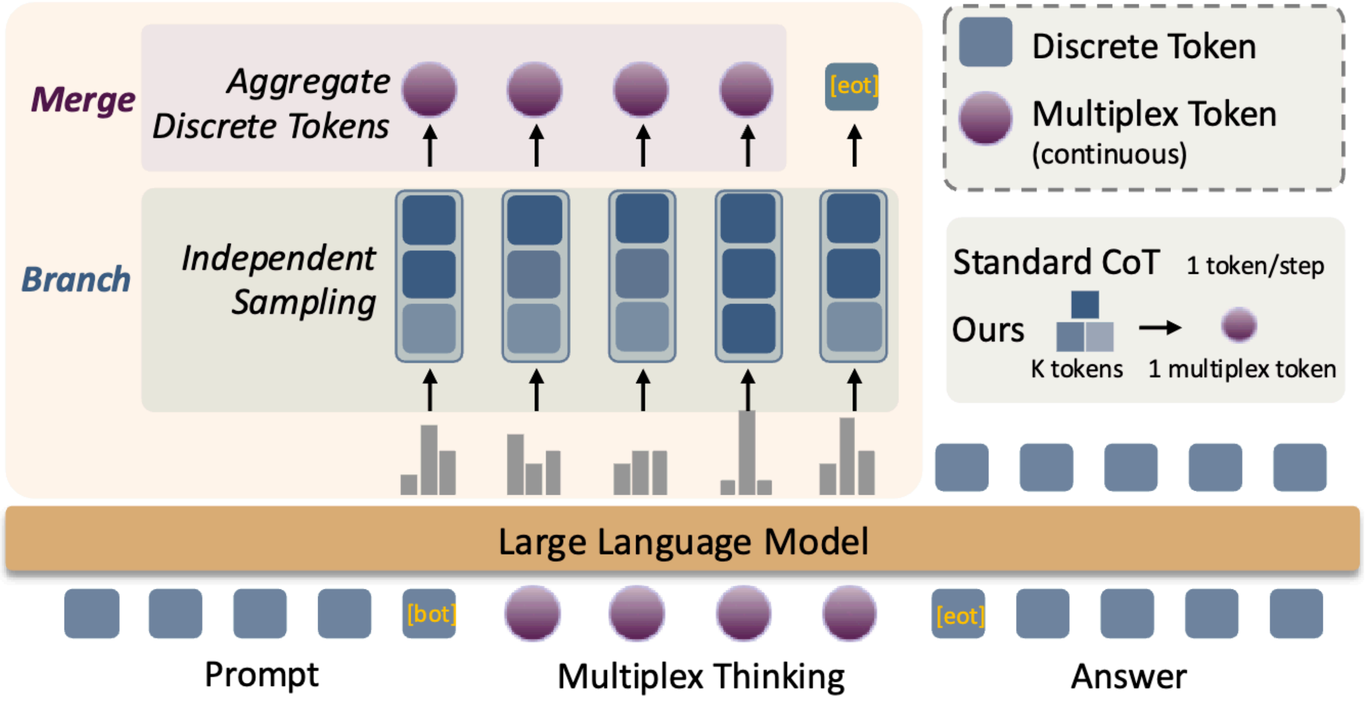


Hao, Shibo, et al. "Training large language models to reason in a continuous latent space." arXiv preprint arXiv:2412.06769 (2024).
Shen, Zhenyi, et al. "Codi: Compressing chain-of-thought into continuous space via self-distillation." Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025.
Zhu, Hanlin, et al. "Reasoning by superposition: A theoretical perspective on chain of continuous thought." Advances in Neural Information Processing Systems 38 (2026): 79931-79963.

Background: Latent Reasoning

- Relaxation by feeding back weighted combination of embeddings



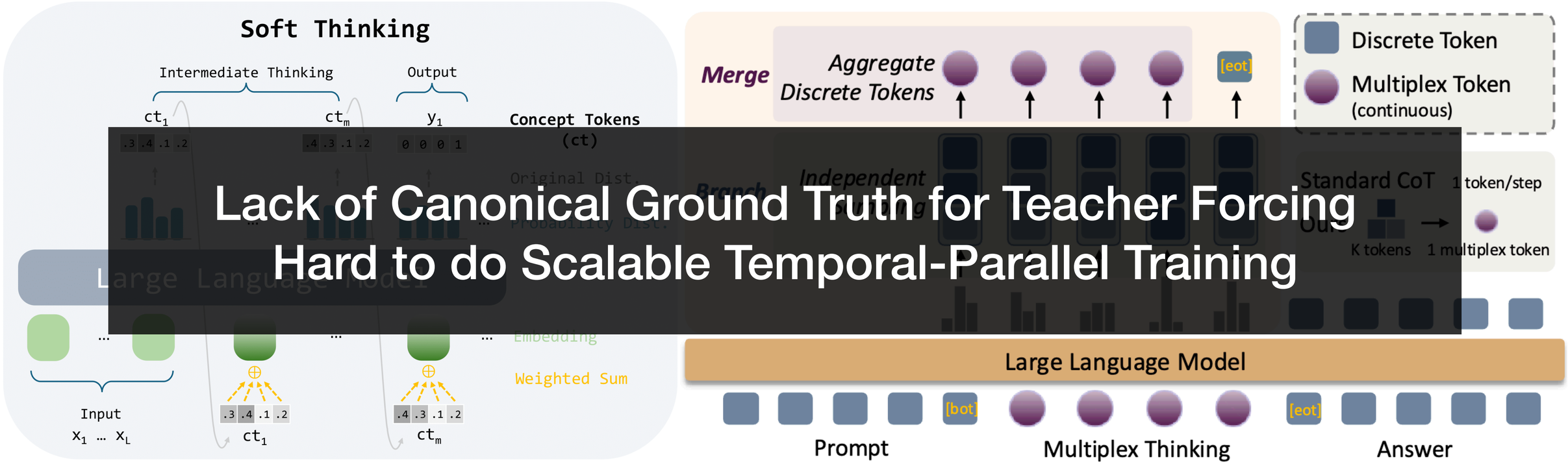


Zhang, Zhen, et al. "Soft thinking: Unlocking the reasoning potential of llms in continuous concept space." arXiv preprint arXiv:2505.15778 (2025).

Tang, Yao, et al. "Multiplex Thinking: Reasoning via Token-wise Branch-and-Merge." arXiv preprint arXiv:2601.08808 (2026).

Background: Latent Reasoning

- Relaxation by feeding back weighted combination of embeddings



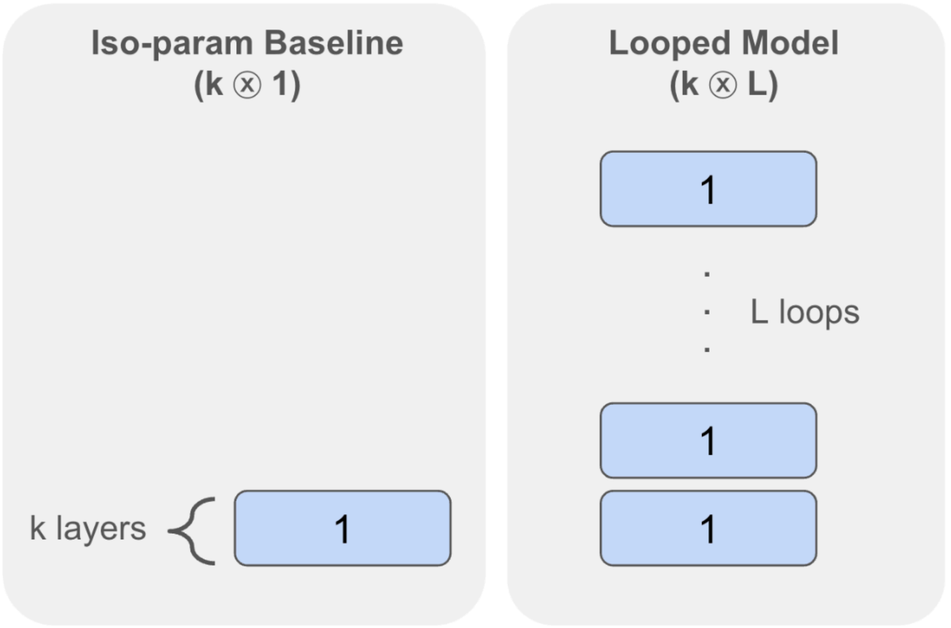
Lack of Canonical Ground Truth for Teacher Forcing
Hard to do Scalable Temporal-Parallel Training

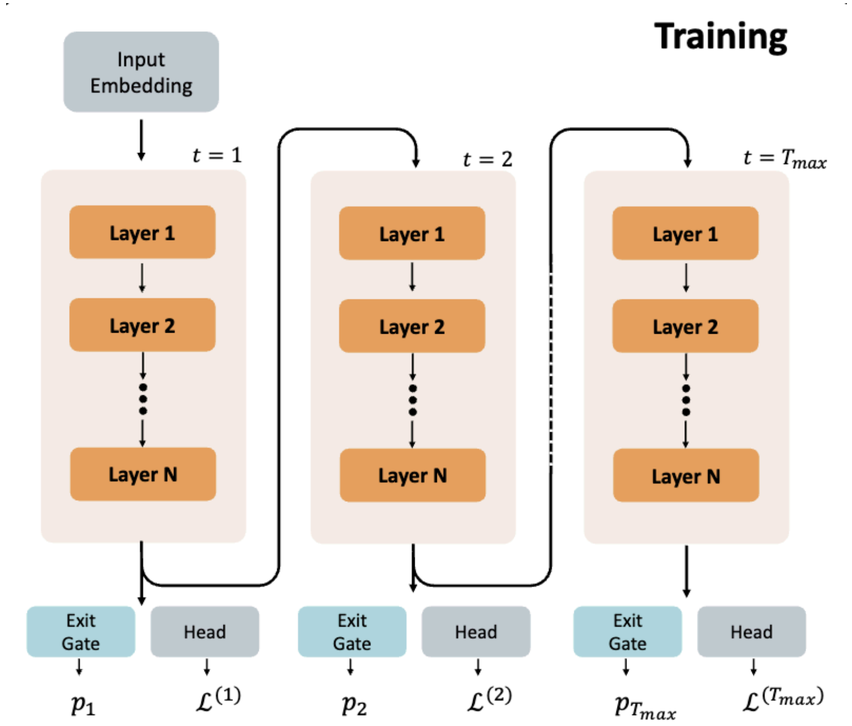
Zhang, Zhen, et al. "Soft thinking: Unlocking the reasoning potential of llms in continuous concept space." arXiv preprint arXiv:2505.15778 (2025).

Tang, Yao, et al. "Multiplex Thinking: Reasoning via Token-wise Branch-and-Merge." arXiv preprint arXiv:2601.08808 (2026).

Background: Looped Transformers

- Relaxation by feeding last-layer representation into first layer again.

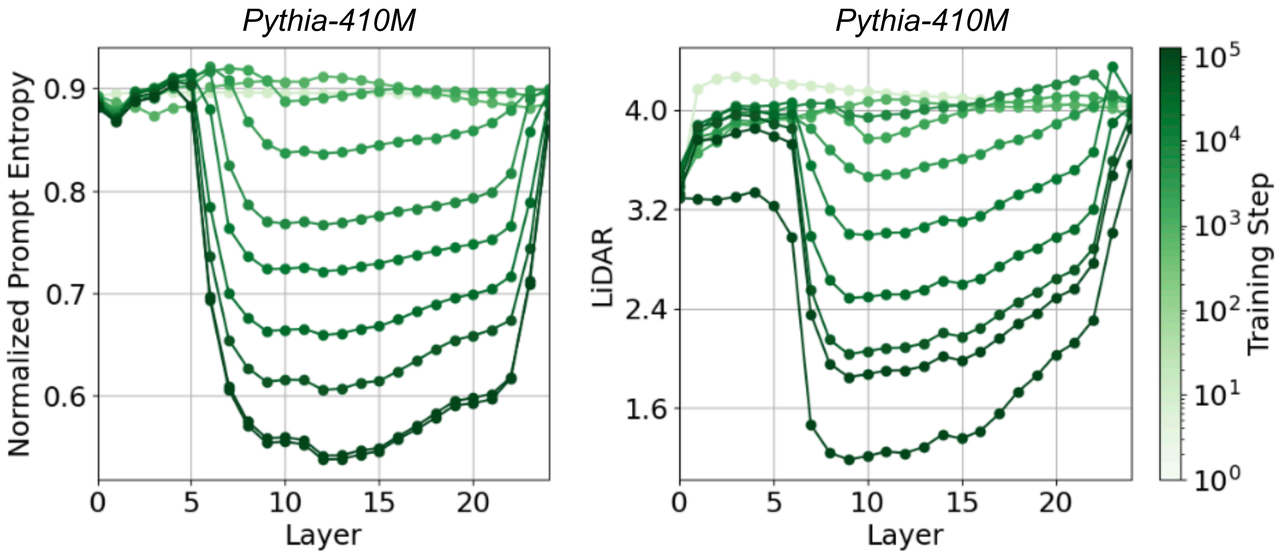


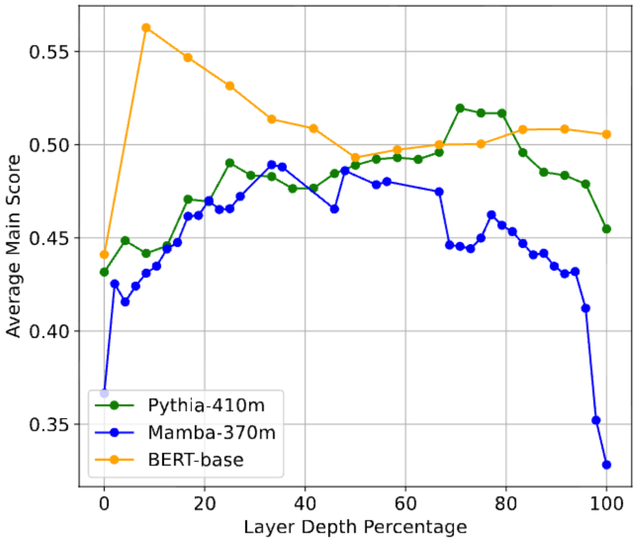


Saunshi, Nikunj, et al. "Reasoning with latent thoughts: On the power of looped transformers." International Conference on Learning Representations. Vol. 2025. 2025.
Zhu, Rui-Jie, et al. "Scaling latent reasoning via looped language models." arXiv preprint arXiv:2510.25741 (2025).

Background: Properties of Middle Layers

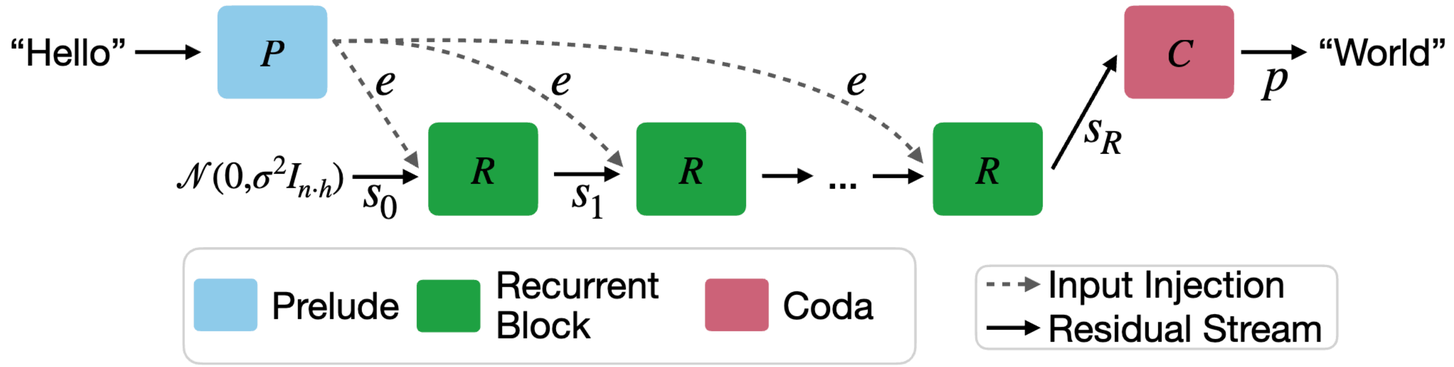
- Left: Middle layers in autoregressive Transformers encode more compressed / abstract information.
- Right: Middle-layer representations yield better performance on downstreams

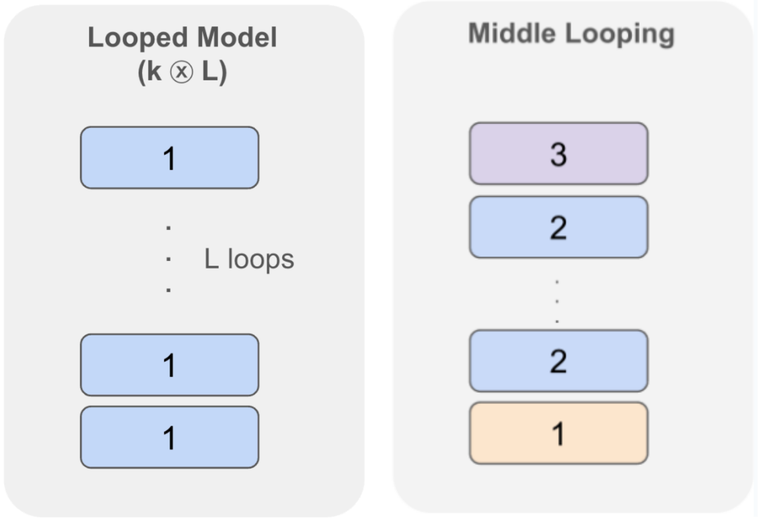




Skean, Oscar, et al. "Layer by layer: Uncovering hidden representations in language models." arXiv preprint arXiv:2502.02013 (2025).

Background: Middle-looping

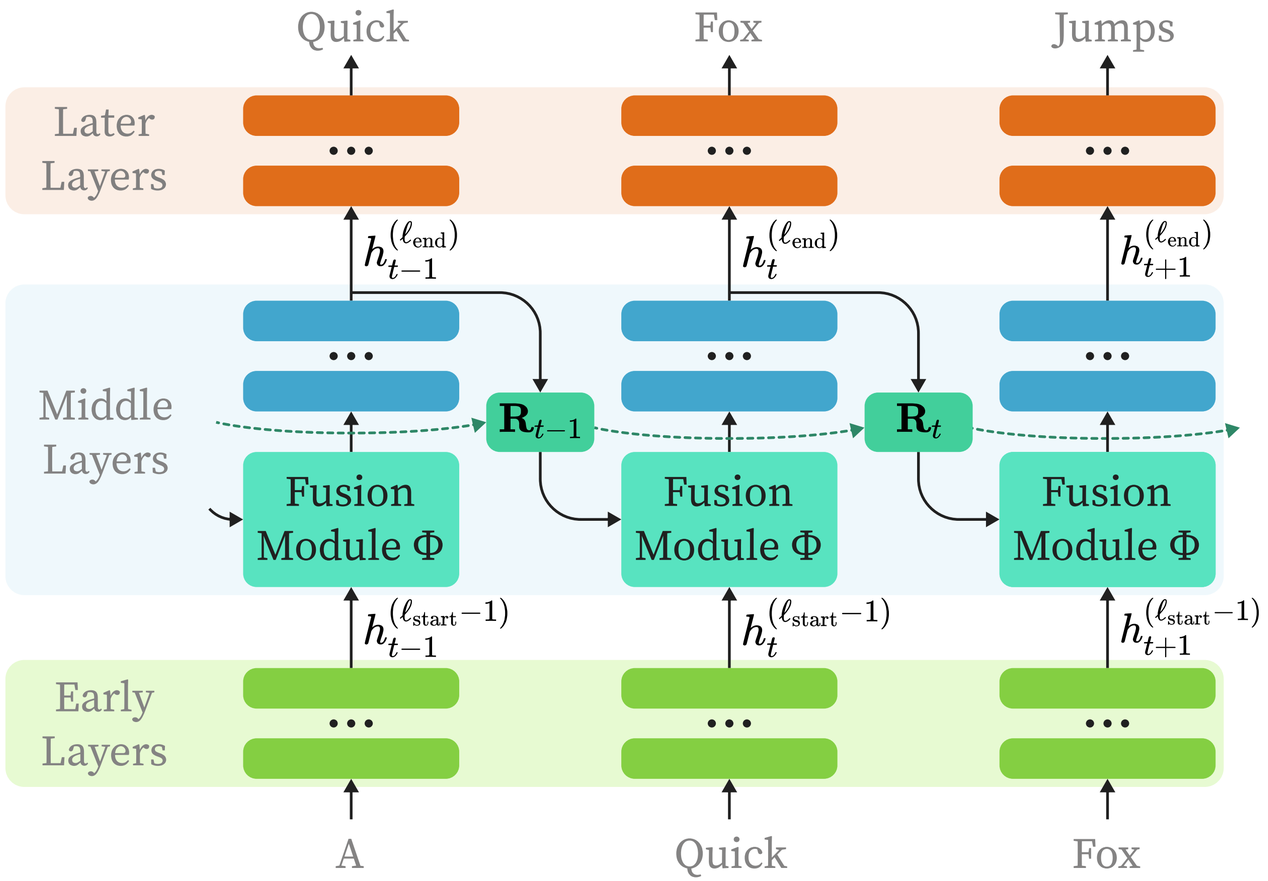




Geiping, Jonas, et al. "Scaling up test-time compute with latent reasoning: A recurrent depth approach." Advances in Neural Information Processing Systems 38 (2026): 41340-41391.

Saunshi, Nikunj, et al. "Reasoning with latent thoughts: On the power of looped transformers." International Conference on Learning Representations. Vol. 2025. 2025.

T²MLR: Transformers with Temporal Middle-Layer Recurrence



Standard Transformer:

$$\left(\boldsymbol{h}_t^{(\ell)}, \mathbf{KV}_{1:t}^{(\ell)}\right) = \mathcal{F}_\ell\left(\boldsymbol{h}_t^{(\ell-1)}, \mathbf{KV}_{1:t-1}^{(\ell)}\right)$$

$$\ell = 1, 2, \ldots, L.$$

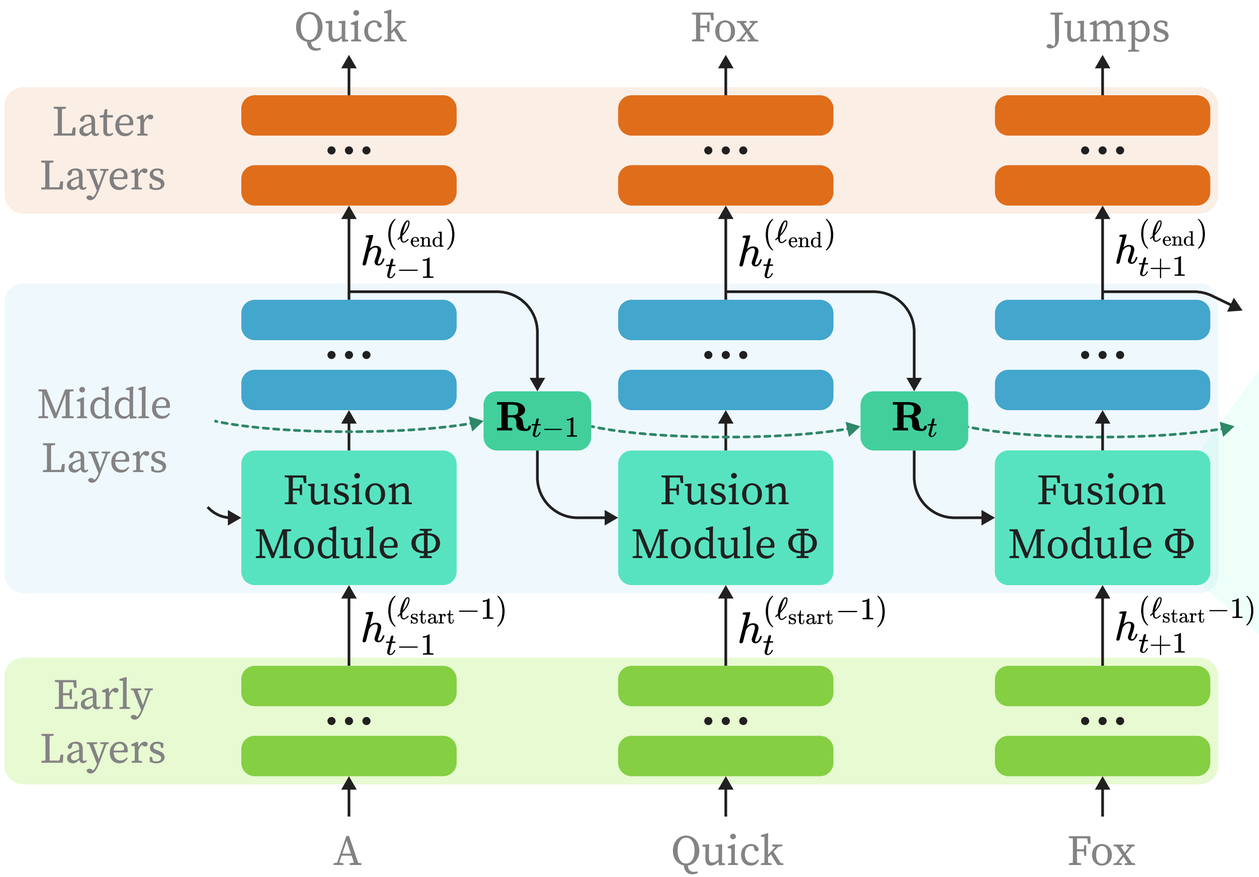
T²MLR:

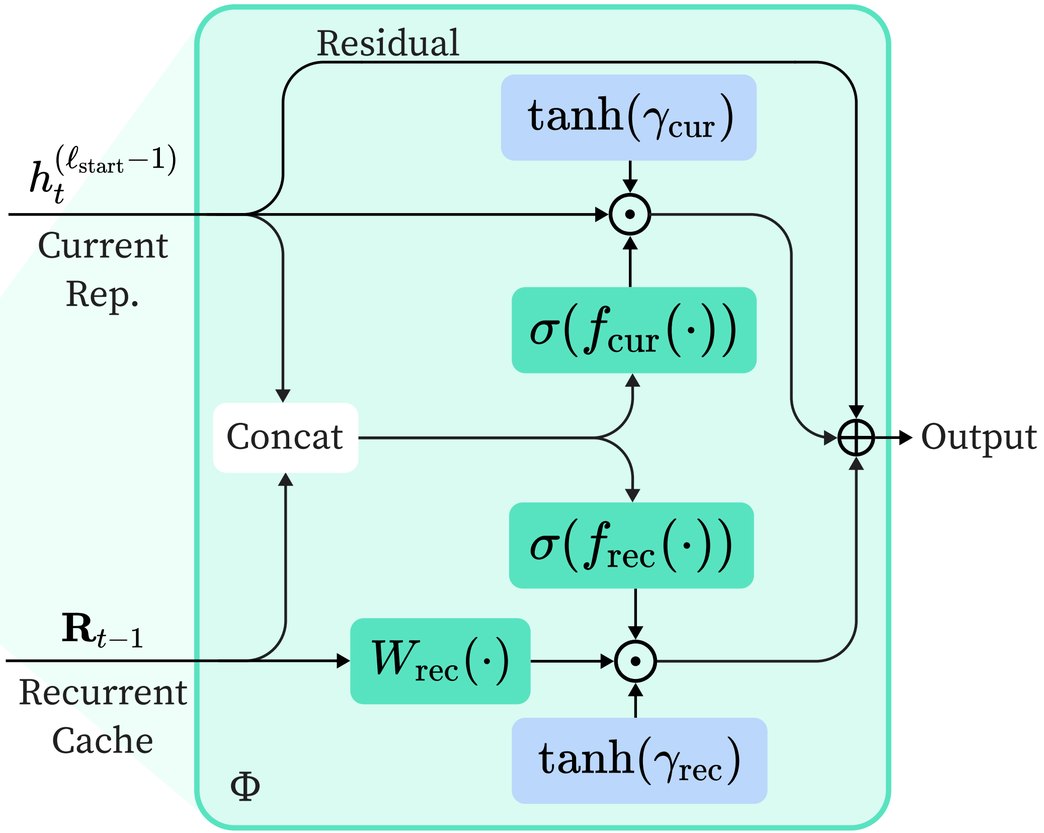
$$\mathbf{R}_t = \text{RMSNorm}\left(\boldsymbol{h}_t^{(\ell_{\text{end}})} + \mathbf{R}_{t-1}\right)$$

$$\left(\boldsymbol{h}_t^{(\ell_{\text{start}})}, \mathbf{KV}_{1:t}^{(\ell_{\text{start}})}\right) =$$

$$\mathcal{F}_{\ell_{\text{start}}}\left(\Phi\left(\boldsymbol{h}_t^{(\ell_{\text{start}}-1)}, \mathbf{R}_{t-1}\right), \mathbf{KV}_{1:t-1}^{(\ell_{\text{start}})}\right)$$

Design of the Fusion Gate



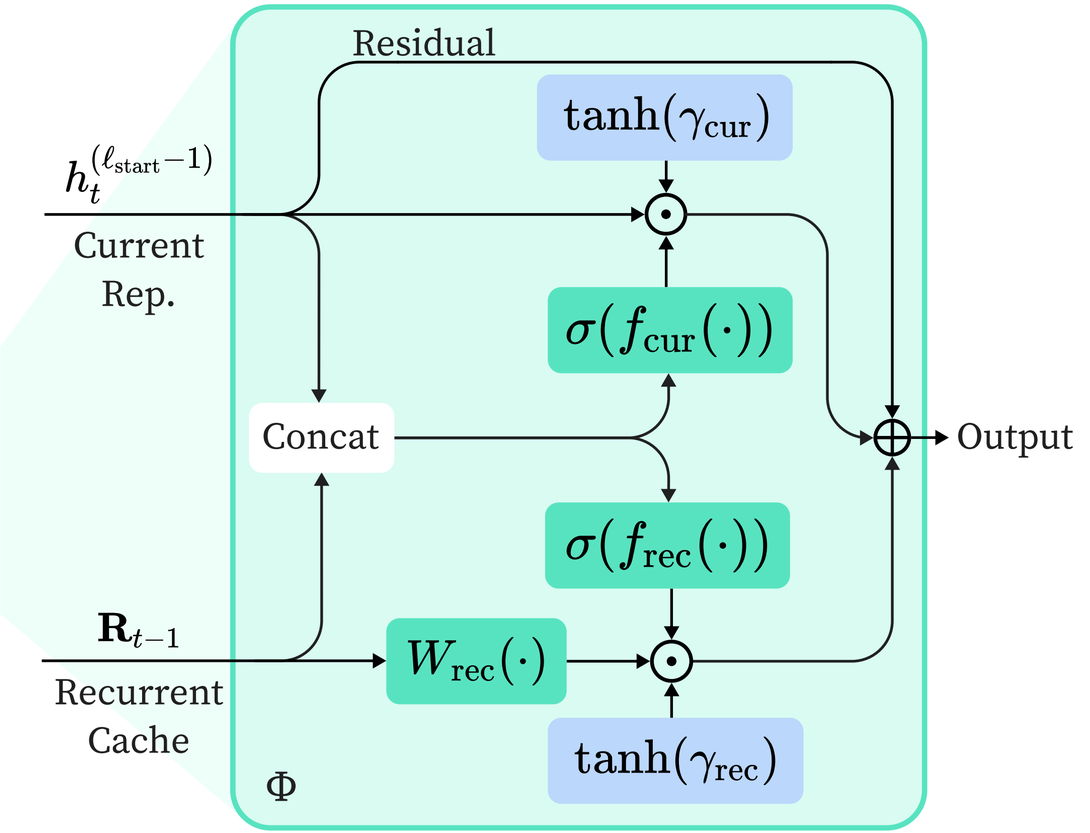


Design of the Fusion Gate

$$\Phi\left(\boldsymbol{h}_t^{(\ell_{\text{start}}-1)}, \mathbf{R}_{t-1}\right) = \boldsymbol{h}_t^{(\ell_{\text{start}}-1)} + \tanh\left(\gamma_{\text{rec}}\right)\, \sigma\left(f_{\text{rec}}\left(\left[\boldsymbol{h}_t^{(\ell_{\text{start}}-1)}, \mathbf{R}_{t-1}\right]\right)\right) \odot \boldsymbol{W}_{\text{rec}}\mathbf{R}_{t-1} + \tanh\left(\gamma_{\text{cur}}\right)\, \sigma\left(f_{\text{cur}}\left(\left[\boldsymbol{h}_t^{(\ell_{\text{start}}-1)}, \mathbf{R}_{t-1}\right]\right)\right) \odot \boldsymbol{h}_t^{(\ell_{\text{start}}-1)}$$

γ_{rec} and γ_{cur} : absolute scale (zero init.)

f_{rec} and f_{cur} : input-aware element-wise gating



Temporal Parallel Training of T²MLR

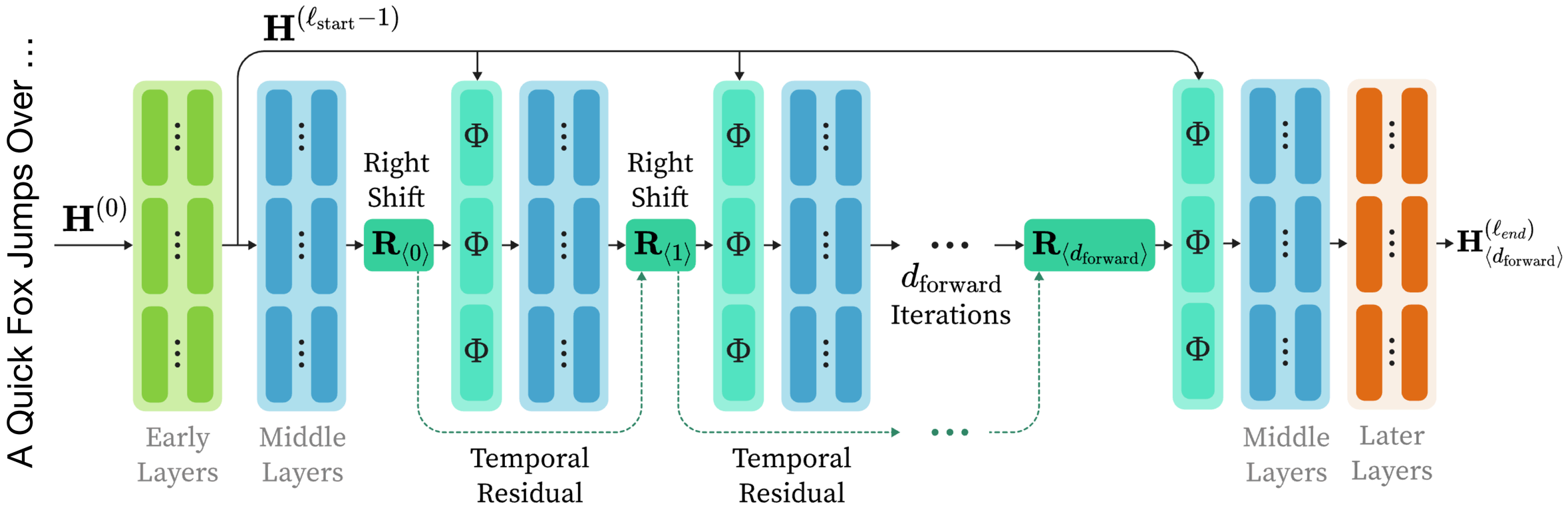
- Deep-to-shallow recurrence breaks temporal parallel training:
  - Coconut / CODI: sequential rollout + curriculum
  - Multiplex thinking / HRPO: RL-only
- Solution: **approximate** the recurrent state for all token positions in parallel
  - Optional truncated BPTT

Temporal Parallel Training of T²MLR

- Deep-to-shallow recurrence breaks temporal parallel training:
  - Coconut / CODI: sequential rollout
  - Multiplex thinking: RL-only
- Solution: approximate the recurrent state for all token positions in parallel
  - Optional truncated BPTT

Temporal Parallel Training of T²MLR

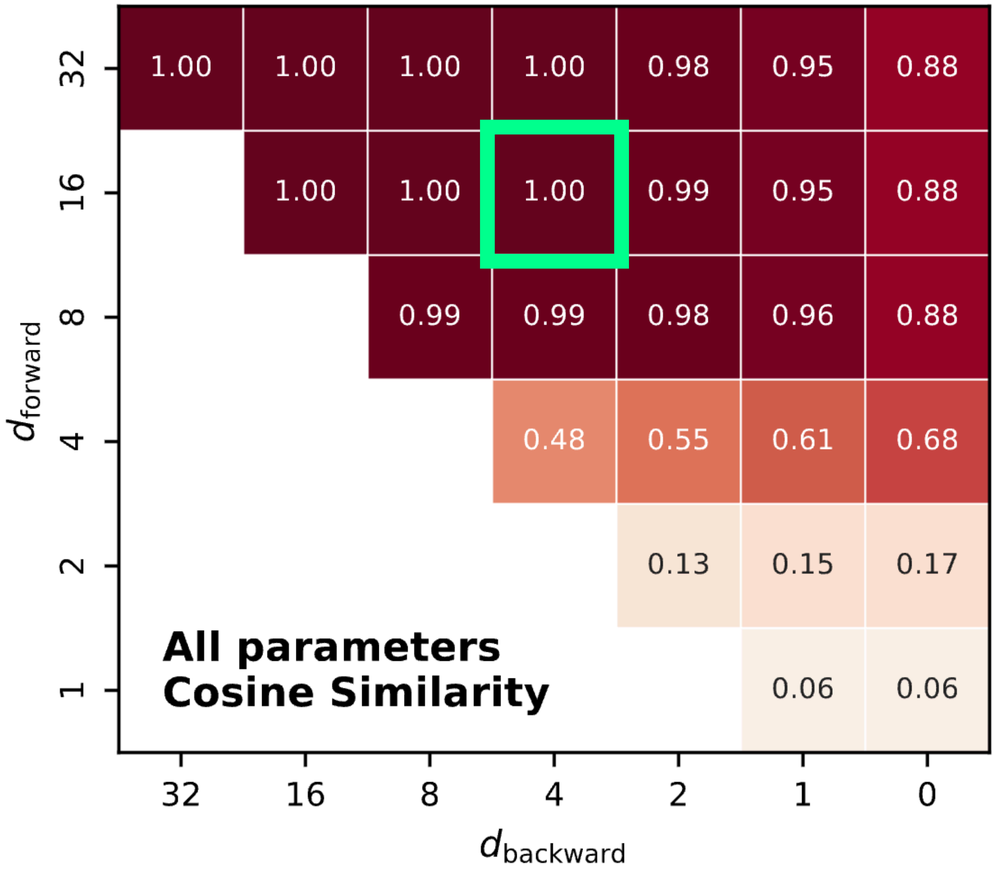
- Jacobi Approximation for Recurrent Cache

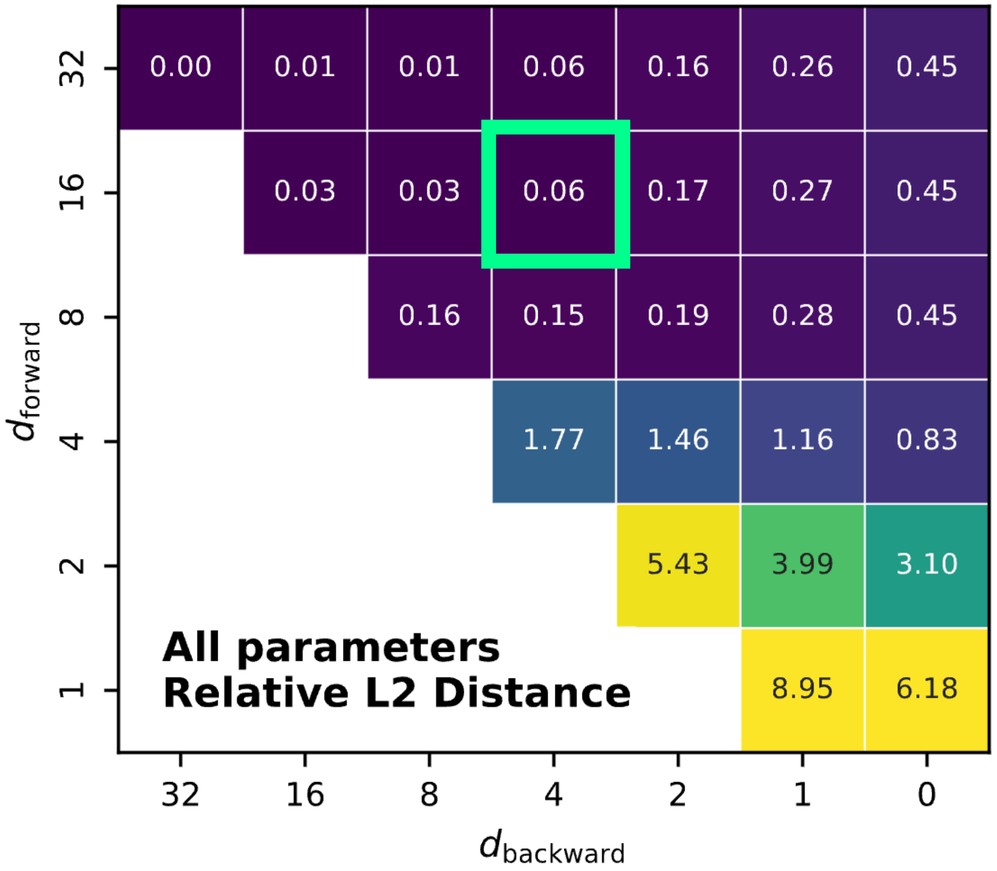


*For backprop: Truncated BPTT with fewer backward steps

Validity of Jacobi Approximation

- Fixing forward=32, backward=32 as “anchor”, sweep various configs.

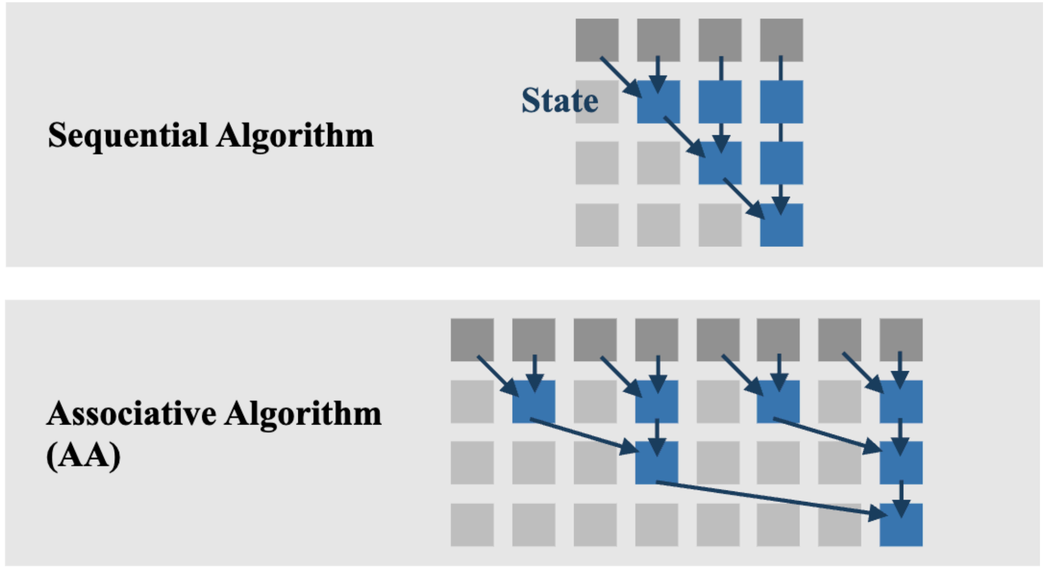




Inductive Bias of T²MLR

- Background: A state tracking task on the permutation group S_5.

$$(a_1, \quad a_2, \quad a_3, \quad \ldots, \quad a_N) \Rightarrow \left(a_1, \quad a_1 \circ a_2, \quad a_1 \circ a_2 \circ a_3, \ldots, \quad \Pi_{i=1}^{N} a_i\right)$$



Require N layers to solve for sequence of length N

Require log N layers to solve for sequence of length N

Theoretical LB: log N

Li, Belinda Z., Zifan Carl Guo, and Jacob Andreas. "(How) Do Language Models Track State?." arXiv preprint arXiv:2503.02854 (2025).

Merrill, William, Jackson Petty, and Ashish Sabharwal. "The illusion of state in state-space models." arXiv preprint arXiv:2404.08819 (2024).

Inductive Bias of T²MLR: S_5-Retrival:

- In addition to tracking composition, the model also needs to do in-context retrieval based on the tracked state.

$$(a_1, \quad a_2, \quad a_3, \quad \ldots, \quad a_N) \Rightarrow \left(a_1, \quad \mathcal{D}(a_1), \quad a_1 \circ a_2, \quad \mathcal{D}(a_1 \circ a_2), \quad \ldots, \quad \Pi_{i=1}^{N} a_i, \quad \mathcal{D}\left(\Pi_{i=1}^{N} a_i\right)\right)$$

- Example:

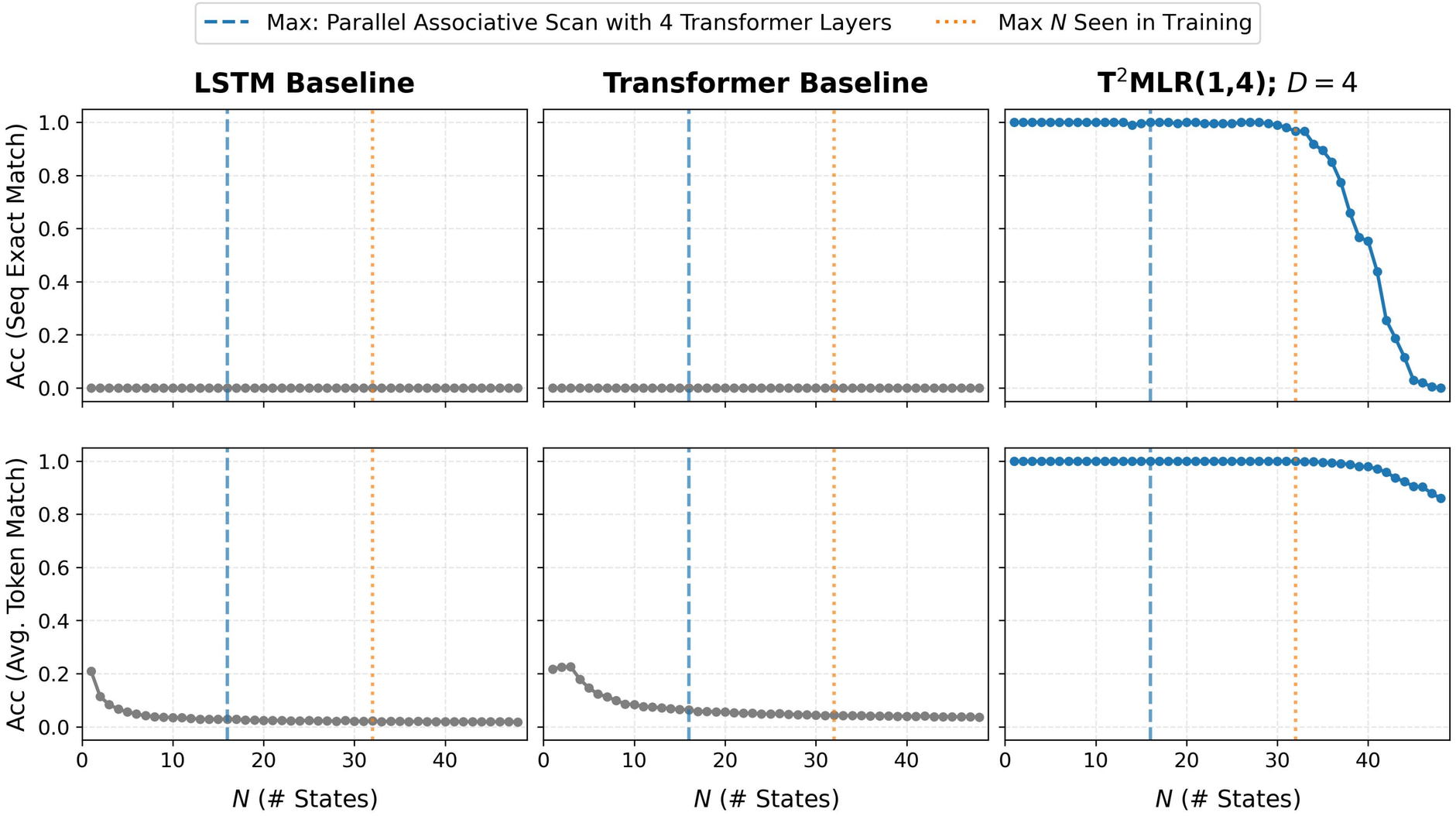
Input:

<A_32514>F4Nd <A_12543>908W <A_54213>Jccq ...
| <A_54213> <A_43125> <A_52314>

Target:

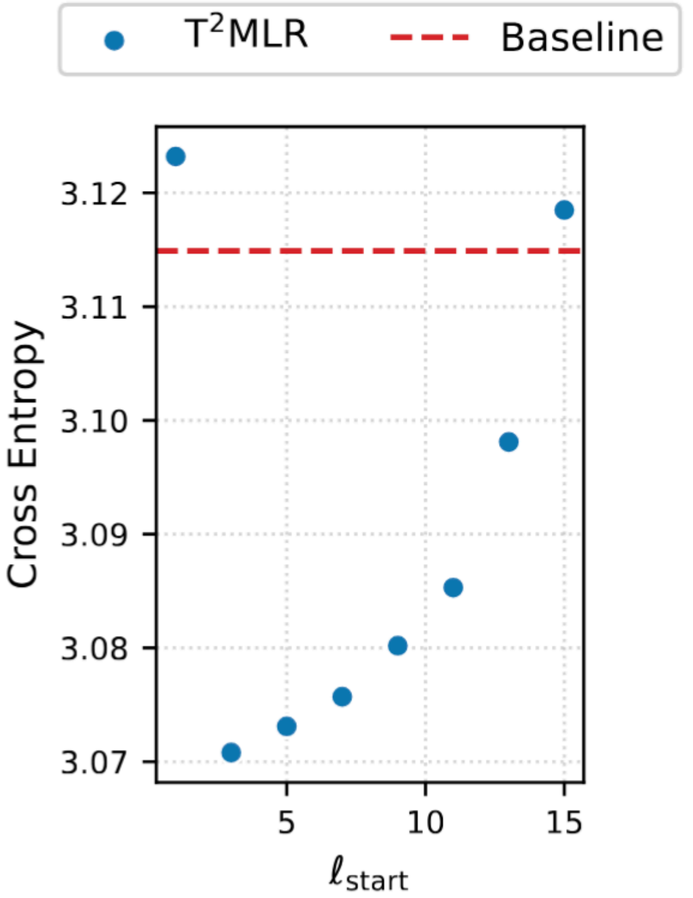
<A_32514>F4Nd <A_12543>908W <A_54213>Jccq ...
| <A_54213>Jccq <A_12543>908W <A_32514>F4Nd

Inductive Bias of T²MLR: S_5-Retrival:



Pre-training T²MLR

- Backbone: 135M (based on SmolLM2)
- Data: Fineweb-Edu (10B tokens)
- Context Length: 2048, batchsize 256
- d_{forward}=16, d_{backward}=4
- #Layers: 30
- #Recurrent Layers: 30, 26, …, 10, 6, 2



Penedo, Guilherme, et al. "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale." arXiv, 25 June 2024, arXiv:2406.17557.

Ben Allal, Loubna, et al. "SmolLM2: When Smol Goes Big -- Data-Centric Training of a Small Language Model." arXiv, 4 Feb. 2025, arXiv:2502.02737.

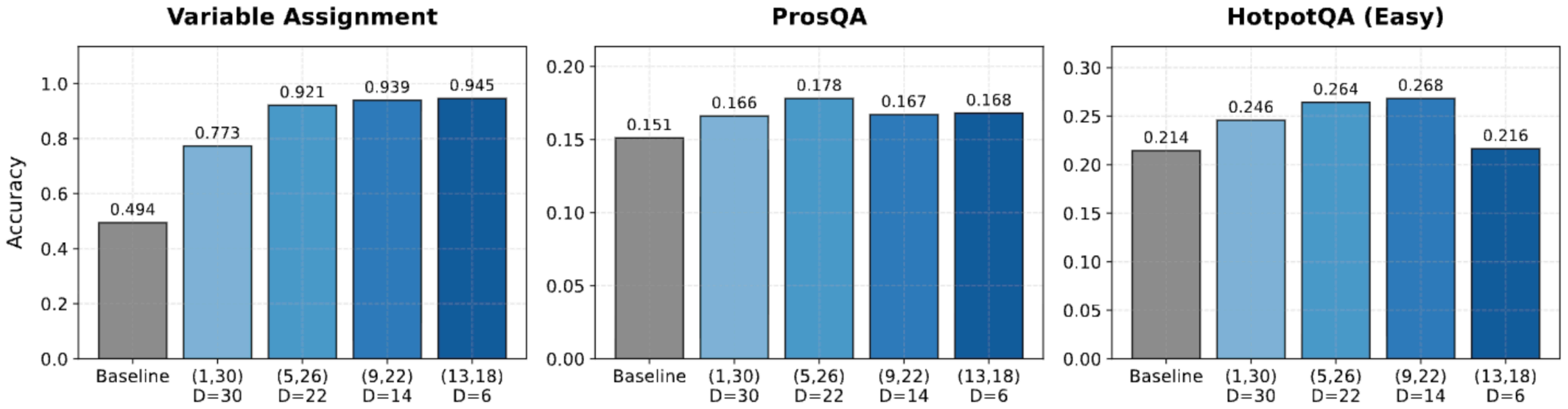
Pre-training T²MLR (135M, 10B Tokens)

Model/Config Metric	ARC-C acc_n ↑	ARC-E acc_n ↑	HS acc_n ↑	OBQA acc_n ↑	PIQA acc_n ↑	SciQ acc_n ↑	WG acc ↑	Average - ↑
Baseline	**24.74**	44.28	29.81	30.20	61.53	60.80	48.46	42.83
T²MLR (1,30); D=30	24.49	43.48	29.98	30.00	60.72	62.60	**52.25**	43.36
T²MLR (5,26); D=22	23.98	**46.13**	**30.75**	29.80	60.77	62.80	51.85	43.73
T²MLR (9,22); D=14	24.15	45.24	30.40	29.20	61.15	**66.40**	51.54	44.01
T²MLR (13,18); D=6	24.23	45.50	29.95	**31.20**	**61.70**	64.20	52.17	**44.14**
T²MLR (15,16); D=2	24.40	45.83	30.11	29.20	59.63	60.20	50.83	42.89

Best average performance when looping only 20% of the layers!

Fine-tuning T²MLR on Reasoning Tasks

- Variable Assignment: Multi-hop Reasoning
- ProsQA (Hard): Graph Connectivity
- HotpotQA (Easy): Natural Language Multi-hop Reasoning (w/ CoT)



ProsQA-Hard (Example)

Input: Every zolufibus is a conarus. Every conarus is a tanirus. Every tanirus is a gedatus. Every cemapus is a lovaperus. Every lovaperus is a zucirarus. Every zucirarus is a pilevus. Every sasarus is a madusurus. Every madusurus is a cidodivus. Every cidodivus is a bumobus. Jordan is a sasarus. Is Jordan a gedatus or a pilevus?

Steps: Jordan is a sasarus.
Every sasarus is a madusurus.
Every madusurus is a cidodivus.
Every cidodivus is a bumobus.
Every bumobus is a nuronus.
Every nuronus is a metuvelus.
Every metuvelus is a resamus.
Every resamus is a gedatus.

Output: Jordan is a gedatus.

Variable-Assignment (Hard)

Input: Variable assignment. Follow the assignments and fill in the blank.
a=5
b=2
c=8
x=c
y=x
z=y
w=b
v=w
z=___
Answer: |

Outputs: 8

Hotpot QA (Easy)

Input: Which magazine was founded first, *The Economist* or *The Atlantic*?

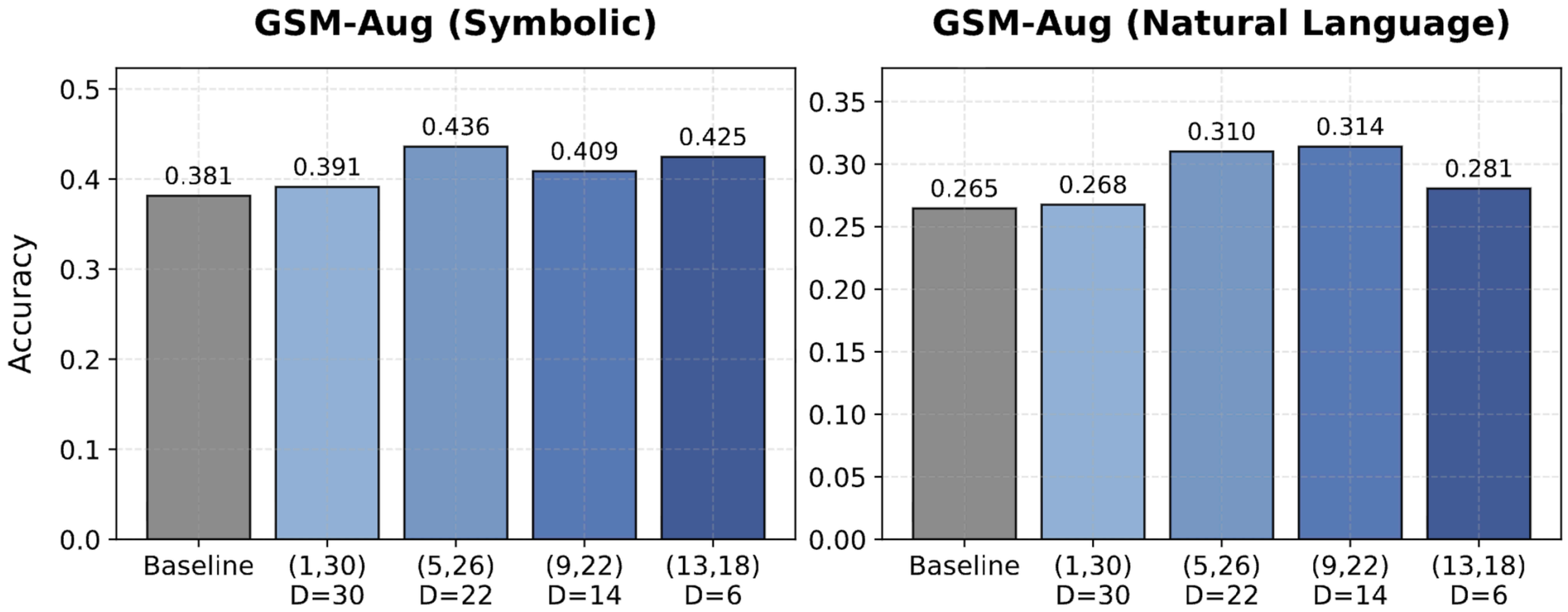
Context:
Document 1: *The Economist* is a weekly newspaper founded in 1843.
Document 2: *The Atlantic* is a magazine founded in 1857.

Steps:
The Economist was founded in 1843.
The Atlantic was founded in 1857.
1843 is earlier than 1857.

Answer: The Economist

Fine-tuning T²MLR on Grade School Math Tasks

- Middle-layer recurrence providing nontrivial boost, while full looping does not.



Future Directions

- Larger scale pre-training.
- RL? (Remove the need for approximation recurrent cache)
- Retrofitting to existing models.
- More flexible recurrent information flow (e.g. attention based)

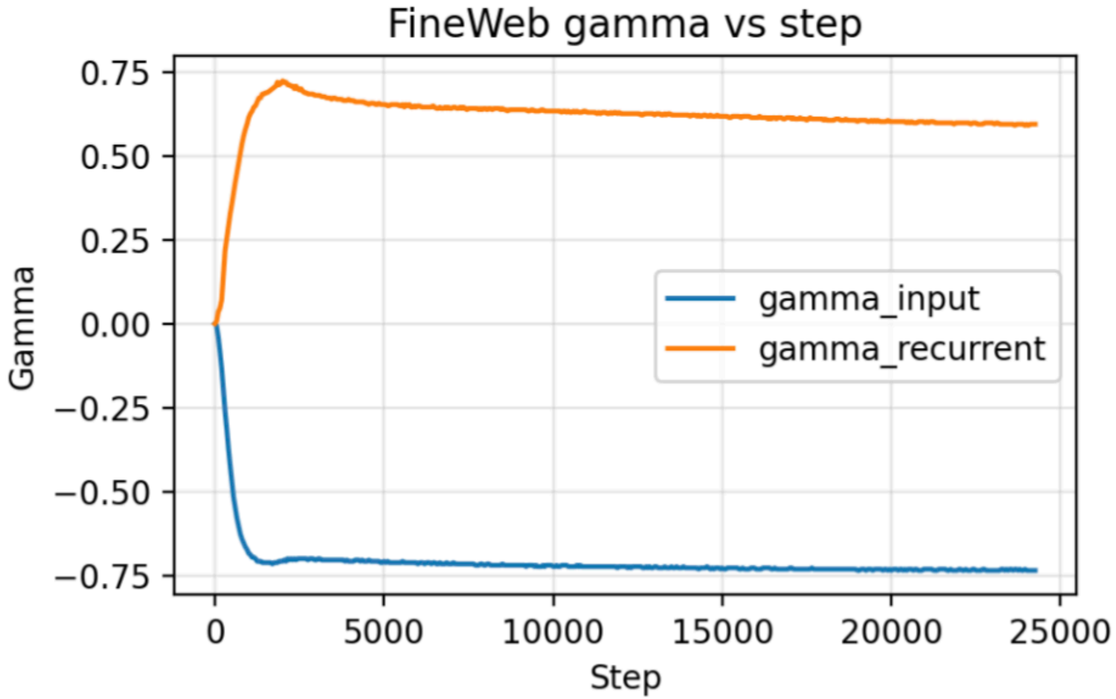
Thanks!

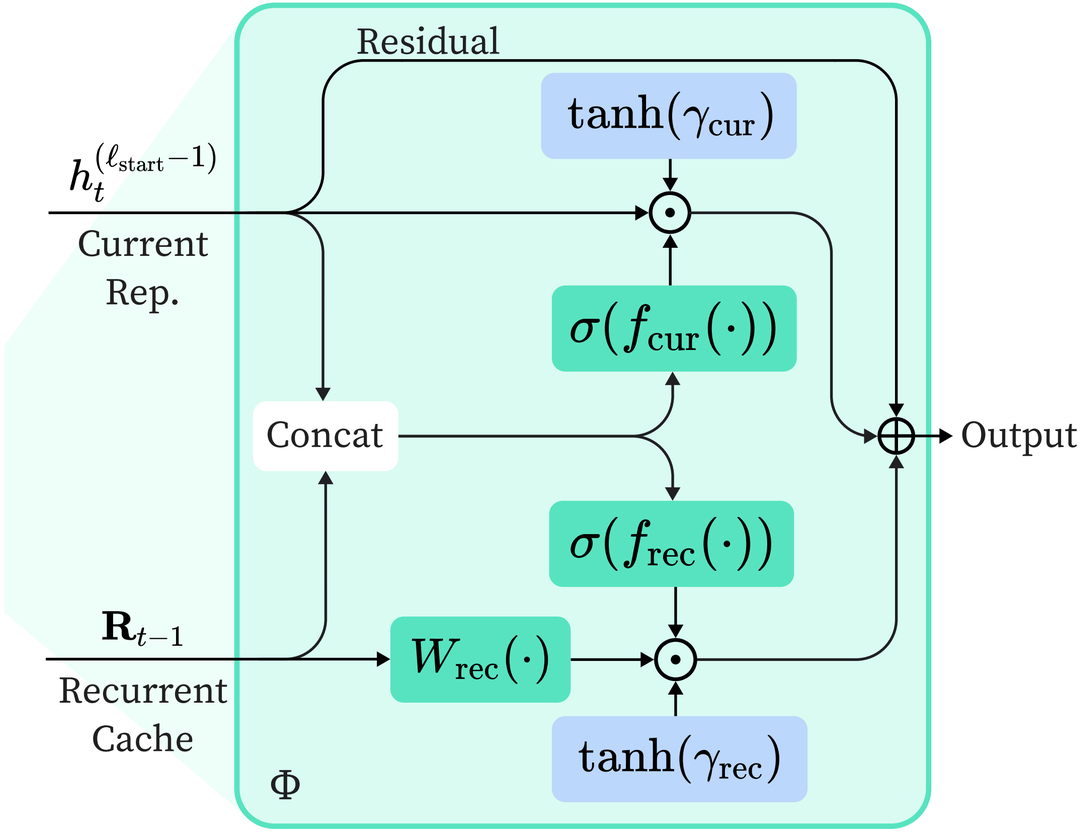
Larger Scale Experiment (370M / 1B) @ 10B Tokens

Model/Config Metric	ARC-C acc_n ↑	ARC-E acc_n ↑	HS acc_n ↑	OBQA acc_n ↑	PIQA acc_n ↑	SciQ acc_n ↑	WG acc ↑	Average - ↑
∼370M parameters								
Baseline (367.8M)	27.13	**52.65**	37.18	**34.00**	65.51	71.00	51.85	48.48
T²MLR (361.8M; gated-f16b4-L8)	**28.84**	52.48	**38.66**	33.40	**66.00**	**73.10**	**52.09**	**49.22**
∼1B parameters								
Baseline (996.9M)	30.97	**55.85**	42.91	33.40	67.74	**77.10**	51.70	51.38
T²MLR (981.5M; gated-f16b4-L8)	**32.00**	55.30	**45.11**	**35.40**	**68.99**	75.60	**54.22**	**52.37**

Model/Config Metric	HotpotQA-Easy acc ↑	GSM-Aug-NL acc ↑	GSM-Aug-Sym acc ↑
∼370M parameters			
Baseline (367.8M)	24.43	31.08	39.35
T²MLR (361.8M; gated-f16b4-L8)	**28.28** (+15.8%)	**34.12** (+9.8%)	**42.46** (+7.9%)
∼1B parameters			
Baseline (996.9M)	23.28	32.37	43.97
T²MLR (981.5M; gated-f16b4-L8)	**26.52** (+14.0%)	**36.69** (+13.3%)	**44.96** (+2.3%)

Gate Dynamics During Training





Longer Pre-training (135M / 370M on 50B tokens)

Model/Config Metric	ARC-C acc_n ↑	ARC-E acc_n ↑	HS acc_n ↑	OBQA acc_n ↑	PIQA acc_n ↑	SciQ acc_n ↑	WG acc ↑	Average - ↑
~135M parameters								
Baseline	**27.82**	**52.53**	36.82	33.60	65.67	72.40	49.64	48.35
T²MLR (T²MLR(9,22))	27.56	51.30	**38.74**	**34.00**	**66.10**	**75.20**	**53.04**	**49.42**
~370M parameters								
Baseline	31.23	**59.64**	46.28	34.60	**70.08**	76.00	52.01	52.83
T²MLR (T²MLR(9,24))	**33.28**	58.88	**48.57**	**36.80**	69.42	**80.30**	**56.20**	**54.78**

Middle Segment vs Early / Late Segment?

Model/Config Metric	ARC-C acc_n ↑	ARC-E acc_n ↑	HS acc_n ↑	OBQA acc_n ↑	PIQA acc_n ↑	SciQ acc_n ↑	WG acc ↑	Average - ↑
Baseline	24.74	44.28	29.81	30.20	61.53	60.80	48.46	42.83
6-Layer Recurrence								
13–18 (Middle)	24.23	**45.50**	**29.95**	**31.20**	**61.70**	**64.20**	**52.17**	**44.14**
1–6 (Early)	24.66	45.08	29.83	30.00	60.07	61.00	48.54	42.74
25–30 (Late)	**25.09**	45.24	29.65	28.20	60.12	61.20	50.59	42.87
14-Layer Recurrence								
9–22 (Middle)	**24.15**	**45.24**	**30.40**	29.20	**61.15**	**66.40**	**51.54**	**44.01**
1–14 (Early)	23.38	44.70	28.59	29.00	59.25	59.50	51.07	42.21
17–30 (Late)	23.98	44.02	29.97	**31.40**	60.28	61.90	50.43	43.14