Understanding Dynamic Compute Allocation in Recurrent Transformers

Ibraheem Muhammad Moosa[1], Suhas Lohit[2], Ye Wang[2], Moitreya Chatterjee[2], Wenpeng Yin[1]
[2]Mitsubishi Electric Research Laboratories (MERL), [1]The Pennsylvania State University

Code

ANIRA

Adaptive Neural Iterative Reasoning Architecture

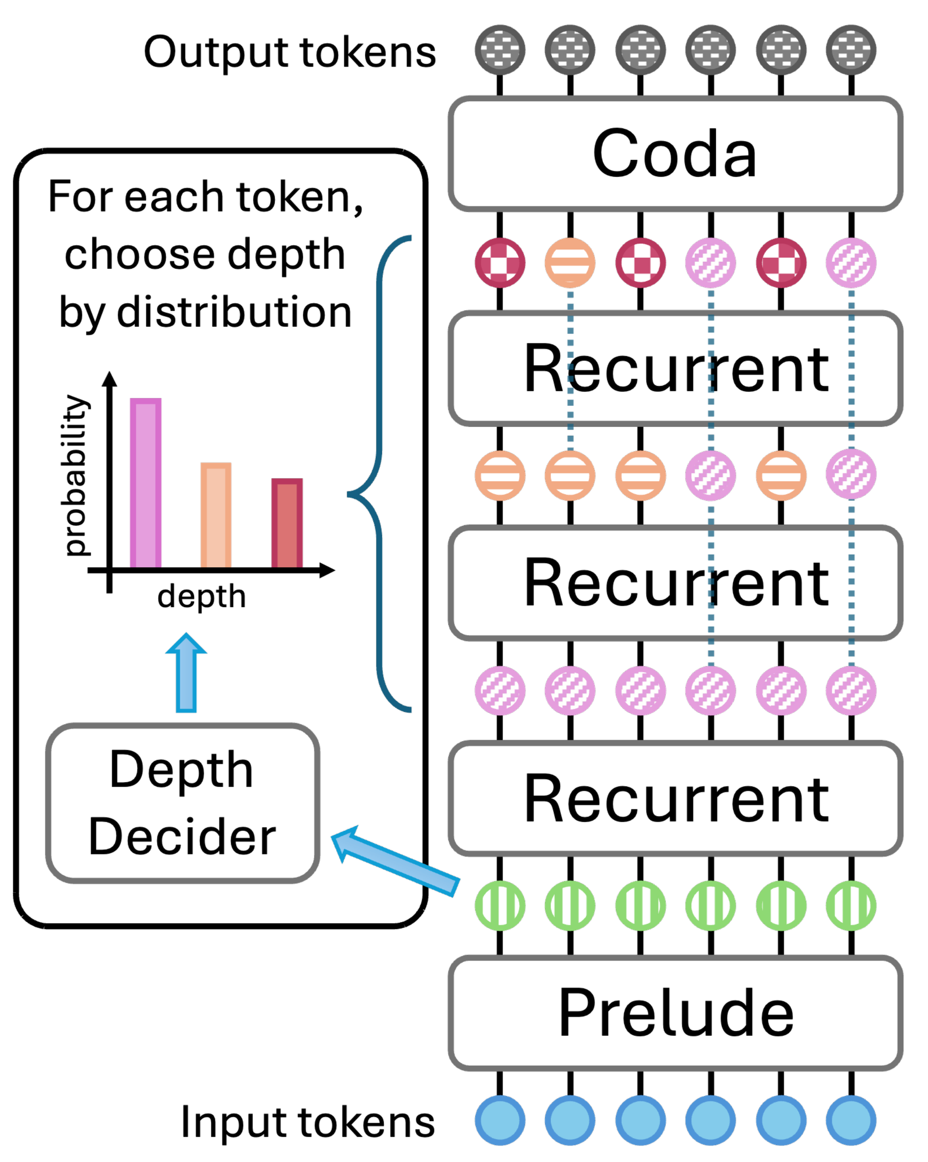
- We design ANIRA, a latent reasoning architecture using recurrent-depth/looped Transformers [1], **with the added capability of exiting early when possible**
- Loss function that encourages exiting early and improving next token prediction at the same time

$$L = L_{CE} + \gamma L_C$$
$$L_C = KL(P_{\text{ANIRA}}, P_{\text{prior}})$$

- ANIRA enforces early-exit semantics
- As natural language token generation complexity cannot be measured, we use synthetic tasks to test ANIRA's capabilities
  - Physics of Language Models [2]–MANO, BREVO, LANO, DEPO
  - Multi-task learning using CLRS-text [3] (see paper)

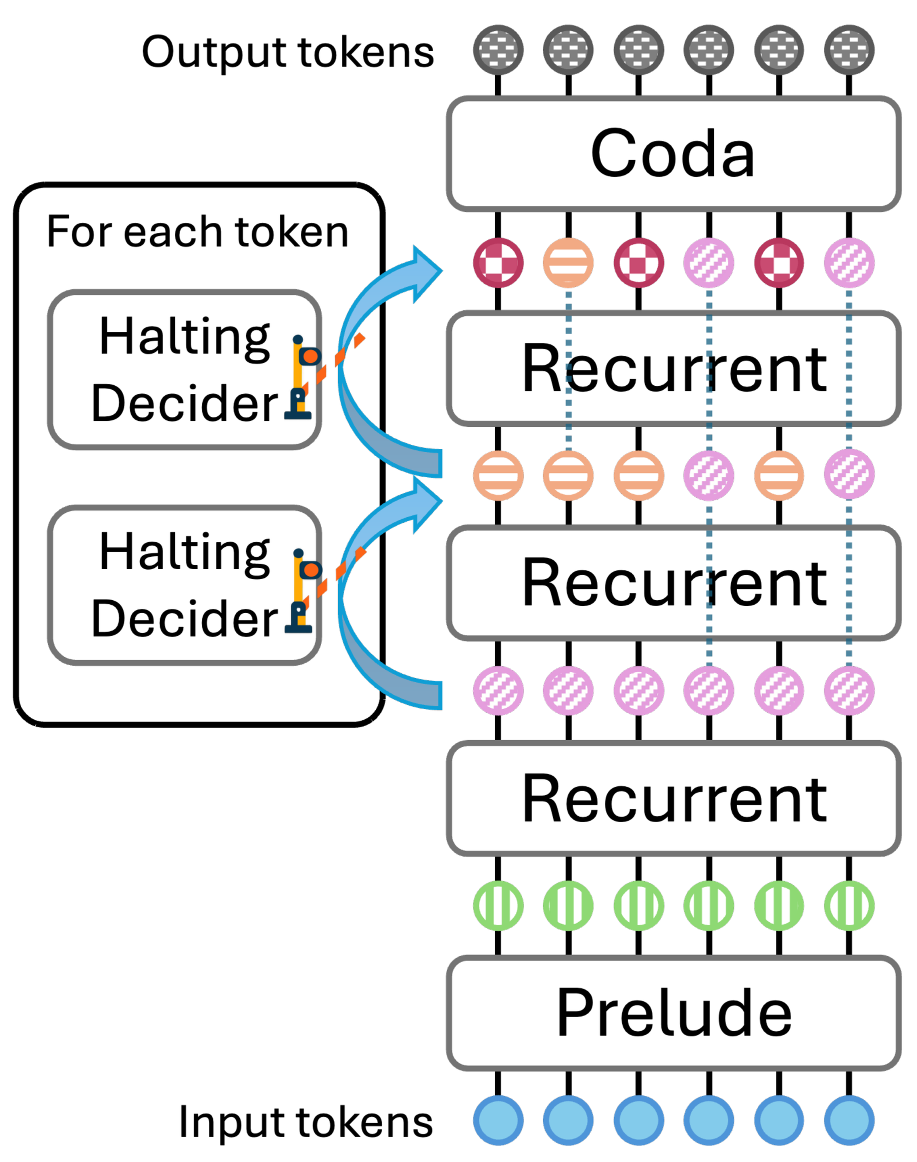
ANIRA-E

Make decision Early



ANIRA-O

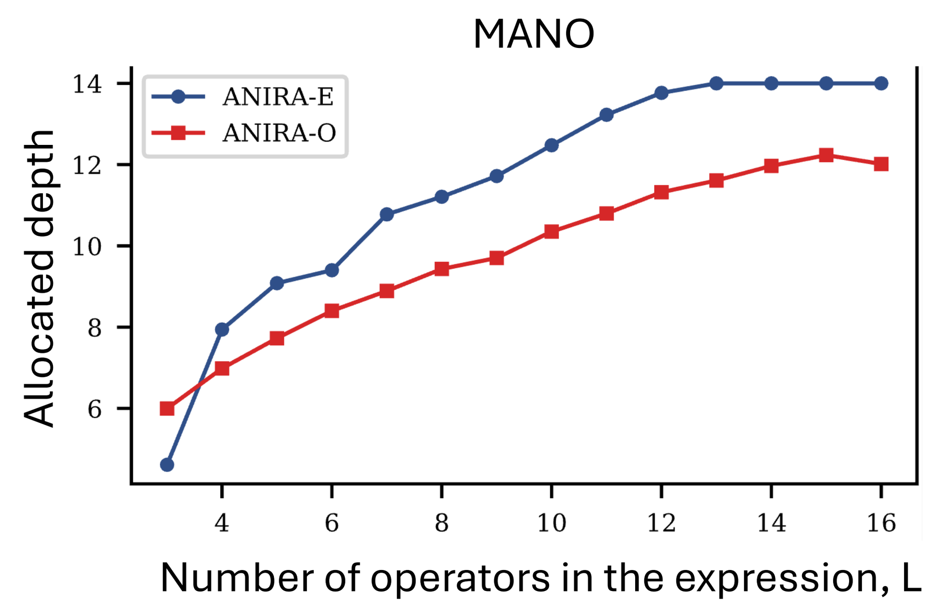
Make decision Online

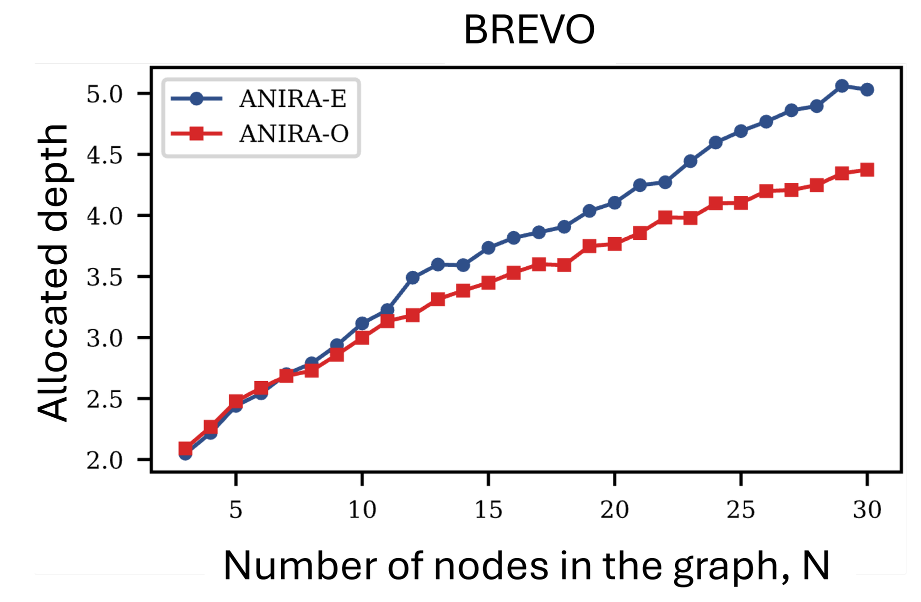


[1] Geiping, Jonas, et al. "Scaling up test-time compute with latent reasoning: A recurrent depth approach.", NeurIPS 2025
[2] Allen-Zhu, Zeyuan, "Physics of Language Models: Part 4.1, Architecture Design and the Magic of Canon Layers.", NeurIPS 2025
[3] Markeeva, Larisa, et al. "The clrs-text algorithmic reasoning language benchmark.", DMLR 2026

Allocated depth tracks task complexity, and ANIRA-O uses fewer steps

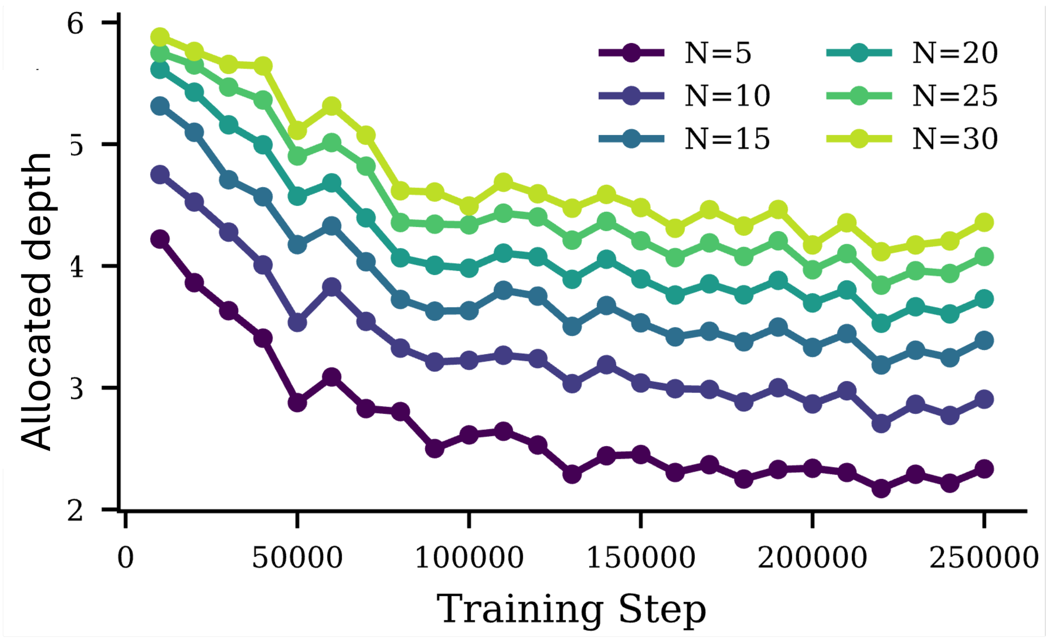
- BREVO: Topological sorting in DAGs
- MANO: Arithmetic using prefix notation
- ANIRA-O and ANIRA-E learn very different policies (see paper)

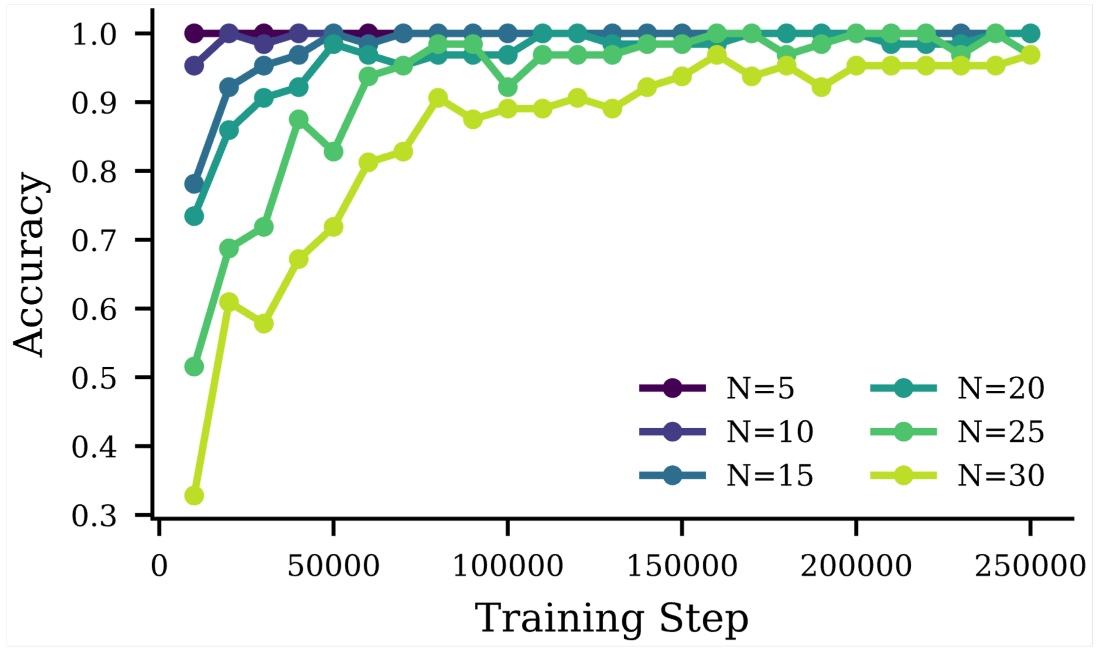




ANIRA learns first in easy-to-hard order, and reduces compute later

- BREVO training dynamics with ANIRA-O
- See paper for others





ANIRA can be more efficient for natural language mathematical reasoning, but allocation is not correlated with task complexity

- Adding Depth Decider (ANIRA-E) or Halting Decider (ANIRA-O) to pretrained baseline Recurrent-Llama-3.2-train-recurrence-8 model from [4]
- Finetune on Nemotron-CC-Math-v1-4plus [5] for 5B tokens, and test on GSM-Symbolic [6]

	Baseline	ANIRA-E		ANIRA-O	
Split	Acc. (%)	Acc. (%)	$\bar{d}$	Acc. (%)	$\bar{d}$
Main	46.26	45.70	6.060	43.68	4.108
P1	23.54	24.06	6.004	23.02	4.105
P2	5.80	5.80	5.982	5.08	4.001

[4] McLeish, Sean, et al. "Teaching Pretrained Language Models to Think Deeper with Retrofitted Recurrence.", arXiv (2025)
[5] Mahabadi, Rabeeh Karimi, et al. "Nemotron-CC-math: A 133 billion-token-scale high quality math pretraining dataset.", ICLR 2026
[6] Mirzadeh, Iman, et al. "GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models.", ICLR 2025