

Looped Transformers

Machine Learning Singapore #MSLG

■ Martin Andrews = Generalisation @ mdda.net

24-September-2026

About Me

- Machine Intelligence / Startups / Finance
 - Moved from NYC to Singapore in Sep-2013
- 2014 = 'fun' :
 - Machine Learning, Deep Learning, NLP
 - Robots, drones
- Since 2015 = 'serious' :: NLP + deep learning
 - Including Papers...
 - & GDE ML; ML-Singapore co-organisier...
 - & Red Dragon AI...

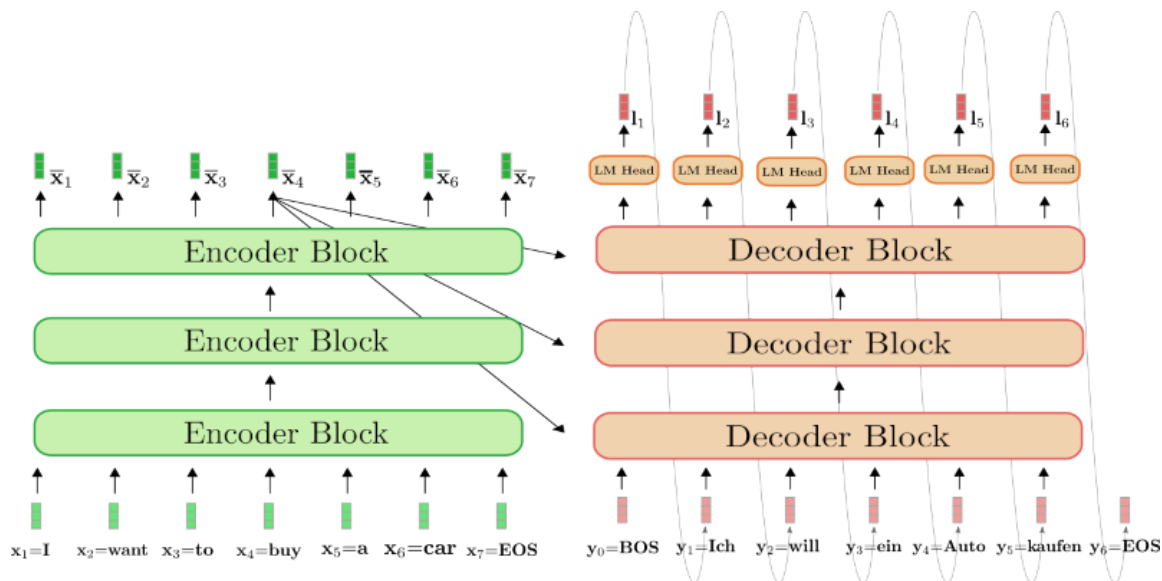
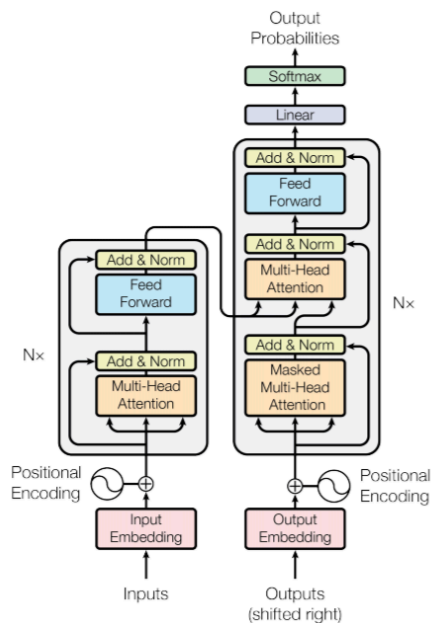


Outline

- Basic Transformers
- The Astra rumours
- Tokenwise Reasoning
- Vertical Expansion
 - Extensions
- Anxiety : Interpretability
- In the news...
- Wrap-up & QR-code

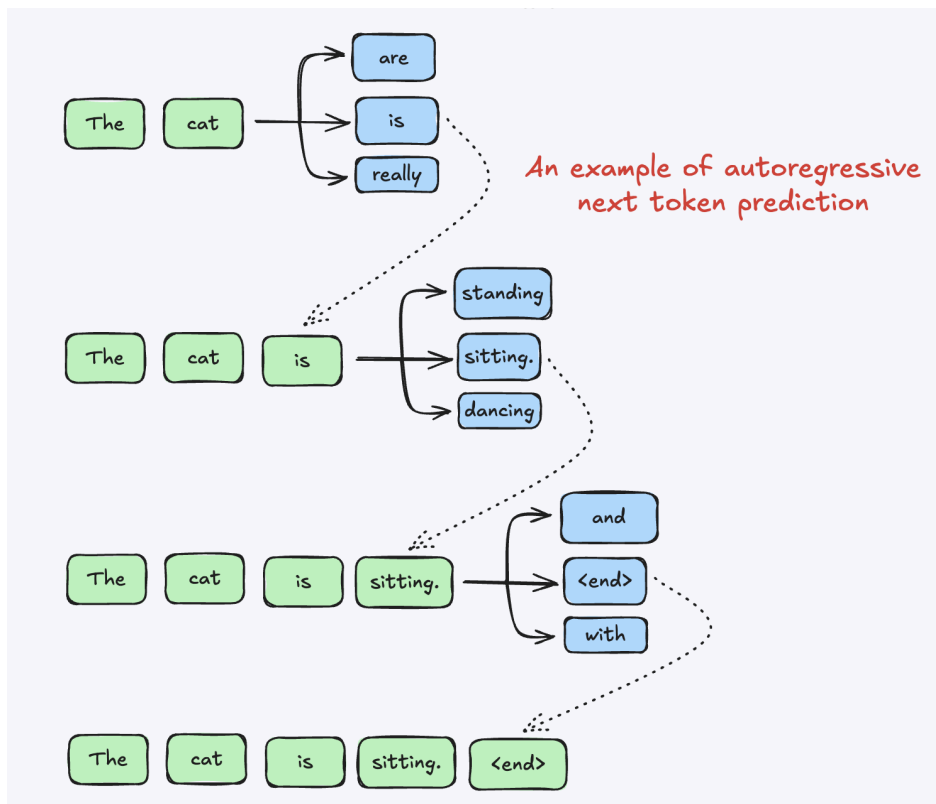
Basic Transformers

Attention Is All You Need (2017)



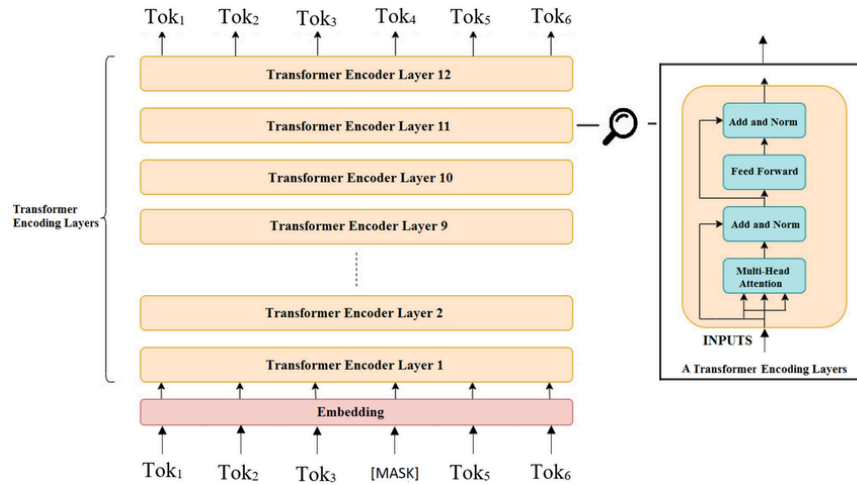
- Encoder side : BERT
- Decoder side : GPT 1, 2, 3, Chat, 4, 5

Basic Rollouts



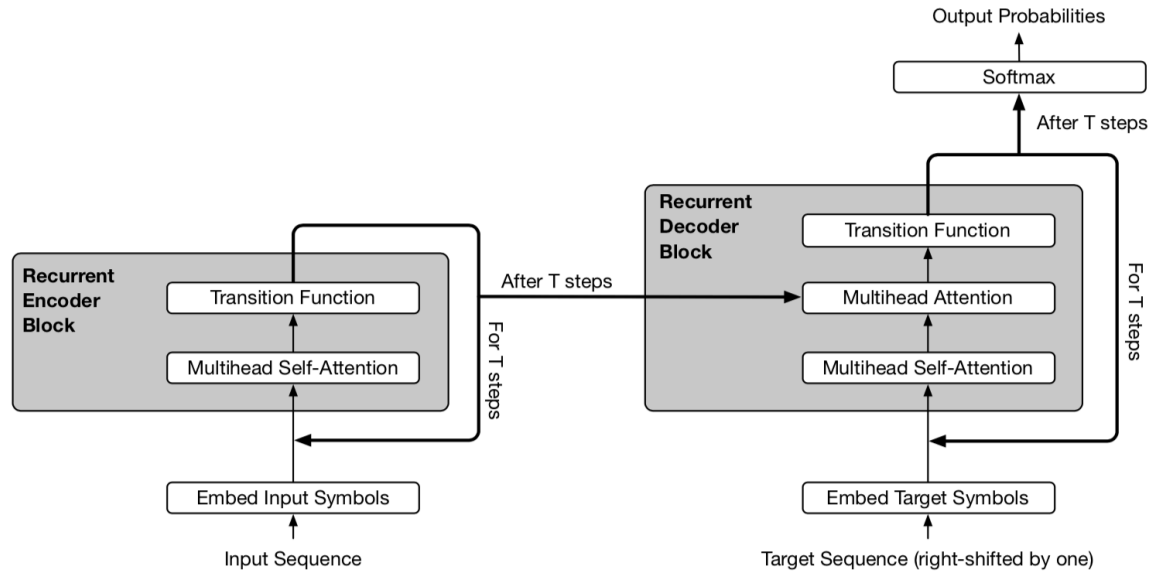
Early Ideas

ALBERT



- ALBERT: A Lite BERT for Self-supervised Learning of Language Representations - Lan *et al* (2019)
 - fixed # of layers
 - focus : Save parameters
 - downside : No time saving

Universal Transformers (2018)



- stopping criteria = Adaptive Computation Time (2016)

Rumours

Astra reportedly is a Looped Transformer

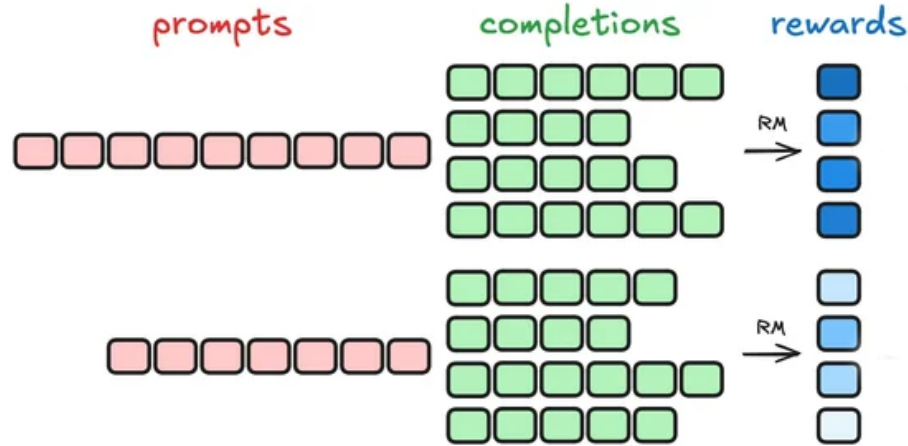
“ The new technique OpenAI is using, known as **recurrent depth or looped transformer**, allows an AI model to improve its answers by processing the same text multiple times.

Unlike commercially available state-of-the-art models, which show in writing how they are “thinking” about a task before completing it, the new technique works in a way that obscures some or all of the AI’s reasoning, otherwise known as its “chain of thought.” That means the steps that the model takes to accomplish a task can’t easily be read or understood by humans. ”

- Quote from The Information
- Axes of scaling:
 - More Data
 - More Parameters
 - "Thinking" (eg: o1 and DeepSeek)
 - i.e. token rollouts
 - Data scaling / improvements
- New Axes: "Test-time scaling"
 - Thinking in latent space
 - More processing per token

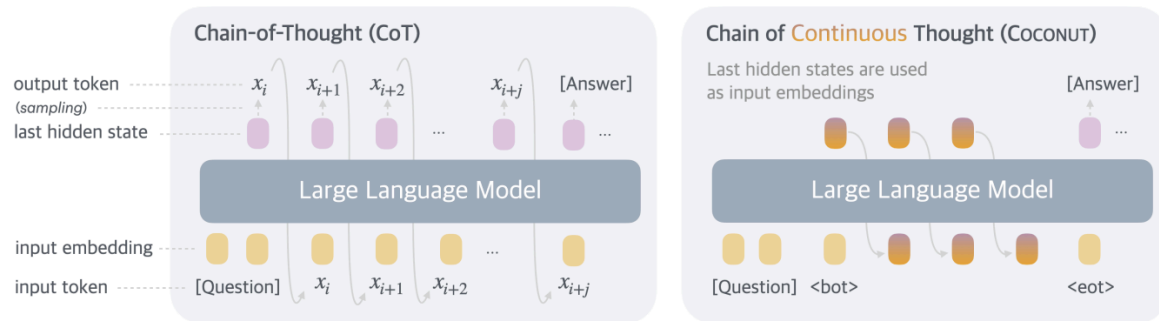
Tokenwise Reasoning

Thinking models



- DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning - Guo *et al* (2025)
 - = Reinforcement Learning from Verifiable Rewards ("RLVR")
 - made 2025 into "The Year of Reasoning"
 - and showed that Open Source models are not getting left behind

Rolling out Latents



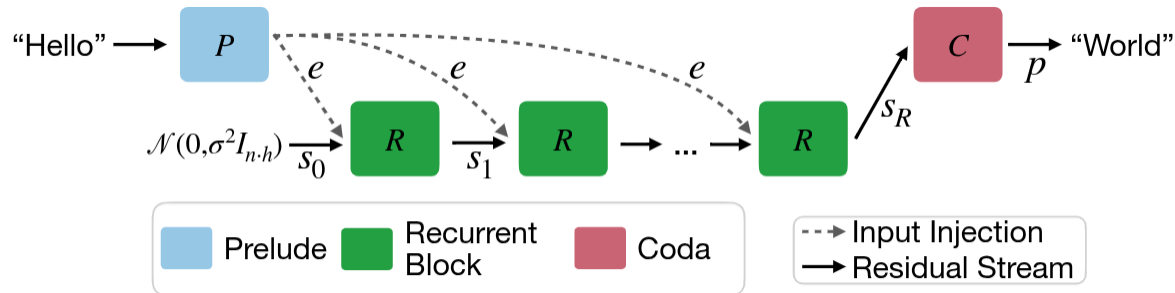
- COCONUT : Training Large Language Models to Reason in a Continuous Latent Space - Hao *et al.* (2024)
 - Facebook Code on GitHub (MIT)
 - previously discussed at MLSG in April-2025
- Instead of decoding the last hidden state into a token:
 - feed state directly as input to the decoder
 - ... as an embedding for the next step
 - ... in the autoregressive generation process

Vertical Expansion

Recurrent Depth

- Why didn't "Universal Transformers" win?
 - compute-vs-parameters trade-off wasn't understood then
 - in 2018-19, Scaling Laws for Neural Language Models hadn't even been published
 - timing was wrong
 - infrastructure wasn't ready
 - gains were real but not dramatic enough
 - scale they operated in was small
 - advances in tech
 - dense encoder-decoder architecture
- Xue Fuzhao (Frio) (pronounced "Full-Draw")
 - already spoke at the MLSG Singapore Spotlight in Oct-2023:
 - "Large Language Model Scaling, Conditional Computation and Adaptive Computation"
 - worked on Recurrent Depth - Sept-2026 tweet
 - Looped MoE transformer : Go Wider Instead of Deeper
 - Looped model has better data efficiency: To Repeat or Not To Repeat: Insights from Scaling LLM under Token-Crisis
 - now works "upstairs"

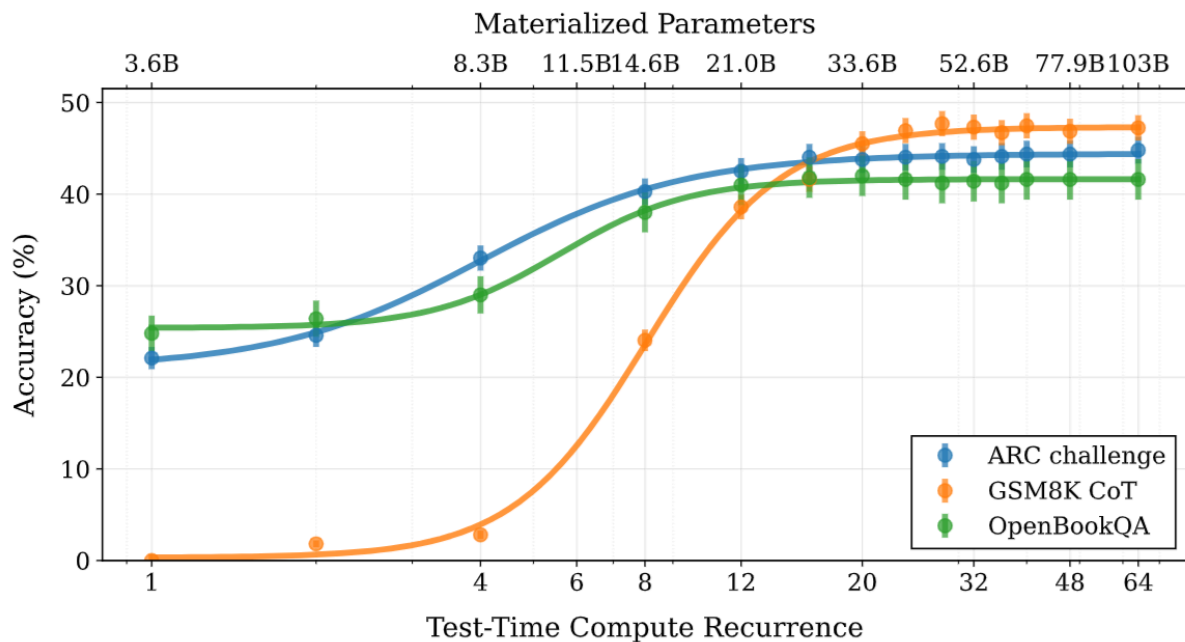
Approach from ICLR 2025



- Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach - Geiping et al. (2025)
 - "Hugging" = More modern take on which layers to repeat
 - prelude P , which embeds the input data into a latent space using multiple transformer layers
 - recurrent block R , which is the central unit of recurrent computation modifying states
 - coda C : un-embeds from latent space; and has the prediction head
 - "We scale a proof-of-concept model to 3.5 billion parameters and 800 billion tokens"
 - trained on *random* numbers of iterations
 - previously discussed at MLSG in April-2025

Effectiveness of Huginn

Scaling up Test-Time Compute with Recurrent Depth

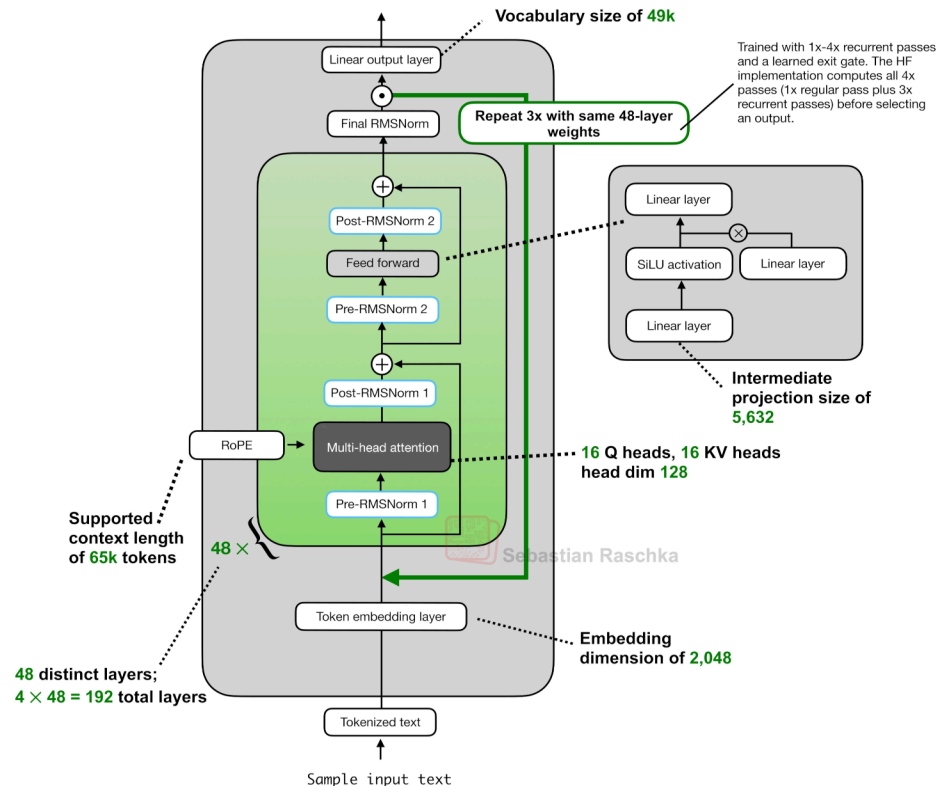


Ouro-Thinking

- Optional repetition of layers
- Scaling Latent Reasoning via Looped Language Models - Zhu *et al* (Oct-2025)
 - Ouro model from ByteDance
 - four passes and a learned exit gate
 - selects the first step where the cumulative exit probability crosses a chosen threshold
 - HF release has not implemented this exit gating...

- Image from Raschka, S. (2026). Looped depth sharing. LLM Architecture Gallery

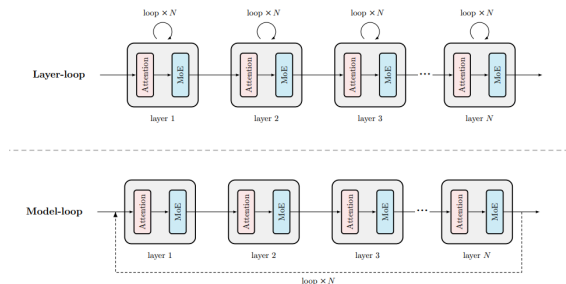
Ouro-Thinking (2.6B)



Something from this year

■ Nanbeige 4.2 Technical Report (2026)

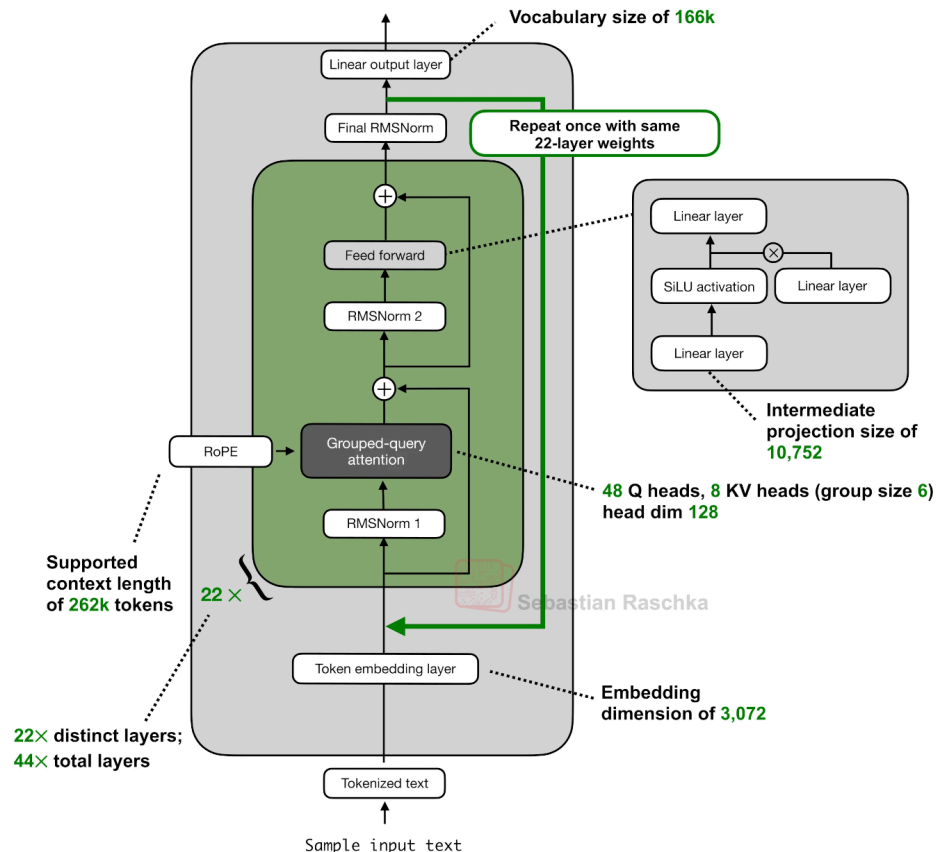
- 'simplifies the looped transformer idea'
- repeats the same 22-layer stack
 - without specific routing or stopping strategies
 - so this is 'just' weight sharing



■ Loop the Loopies! - Gao *et al* (July-2026)

- suggest Layer-Loop has:
 - has better empirical scaling
 - infrastructure-friendly execution
 - 'natural' parameter-sharing pattern

Nanbeige 4.2 (3B)



Latent Thought Flow (@SMU)

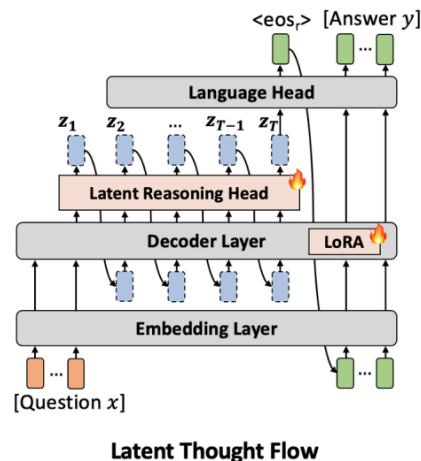
Latent Thought Flow: Efficient Latent Reasoning in Large Language Models

Xiandong Zou¹, Jing Huang², Jianshu Li², Pan Zhou^{1*}

¹Singapore Management University ²Ant Group

Abstract

Large Language Models (LLMs) increasingly rely on intermediate reasoning, yet explicit Chain-of-Thought (CoT) suffers from a linguistic space bottleneck: each thought must be decoded into tokens, causing high inference overhead. Latent reasoning moves deliberation into continuous space, but existing methods mostly learn deterministic or reward-maximizing paths, lacking a principled way to allocate



- Latent Thought Flow: Efficient Latent Reasoning in Large Language Models - Zou *et al* (2026)
 - latent flows are within frozen internal layers + trainable latent reasoning head + trainable LoRA
 - continuous GFlowNet perspective
 - improved accuracy by 9.5%
 - reduced reasoning length by 27.2% on average
 - compared with strong latent reasoning baselines

Anxiety : Interpretability

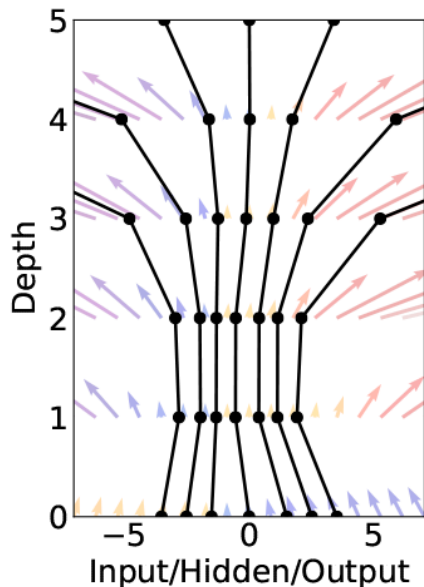
Interpretability of Latent Reasoning

- No directly readable trace
 - intermediate computation happens in hidden states
 - representations evolve and influence later computation
- Unclear step semantics
 - a latent update or loop need not correspond to a human reasoning step
- Decoding $\neq$ explanation
 - readable probes can reveal information
 - ... without showing how the model uses it
- Harder process oversight
 - detecting shortcuts / locating errors / verifying intermediate steps
 - all require internal access and causal tests
- NB: Textual chain-of-thought is not guaranteed faithful either!
 - Looped models can still emit CoT
 - but their internal iterations remain implicit

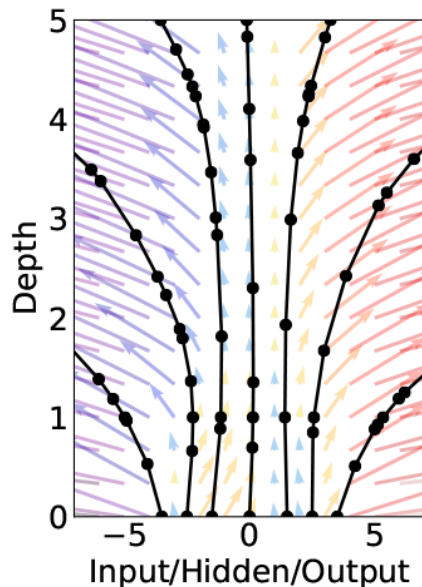
Extensions

Are Transformer layers 'small steps'?

Residual Network



ODE Network



- Perhaps : LLM community will notice that looped transformers are fixed point iteration
 - and will rediscover: Halpern algorithm, Resolvent Operators, and other animals...
 - ... and we will soon get Neural ODE back
 - Deep Implicit Layers - Neural ODEs, Deep Equilibrium Models, and Beyond = NeurIPS 2020 tutorial
 - under new names (something like "inverse residual looping")
 - comments suggested : Solve the Loop: Attractor Models for Language and Reasoning
- Fixed-Point Reasoners: Stable and Adaptive Deep Looped Transformers
 - Code applied to:
 - Sudoku, Maze, state-tracking, and ARC-AGI

In the news...

Actual Doom or IPO Salesmanship?

- (source)



State of the World Venn Diagram



Benchmark update

- BREAKING: Gemini has begun to hillclimb Felony Bench /s

EXCLUSIVE

Gemini Hacked Three Companies in First Known Breakout by Google's AI

The episode resembled similar hacks by other AI models, but Google said it didn't consider it an instance of model misalignment



Friends of MLSG Update

- @niceaunties' Auntieverse was the subject of a 22-page Cambridge University Press study



Figure 3. Artworks from the 'Auntiesocial' and 'Auntie City' chapters.

Wrap-Up

- Shows how a rumour can stir up interest in academia
- Being early is no guarantee
 - too early is cursed
 - many lessons have to be learned first
- Interesting work is still happening at the 3B level
 - and who knows where the Frontier will take us?



Link to Slides

- https://bit.ly/MLSG_2026-09

