Why limit the Residual Stream to Layers and not Tokens? Persistent Memory for Continuous Latent Reasoning

Mujtaba Farhan[1] • Maheep Chaudary[2] • Sean Wu[3] • Ashwinee Panda[4]

[1]Algoverse AI Research · [2]Independent · [3]University of Oxford · [4]Princeton University

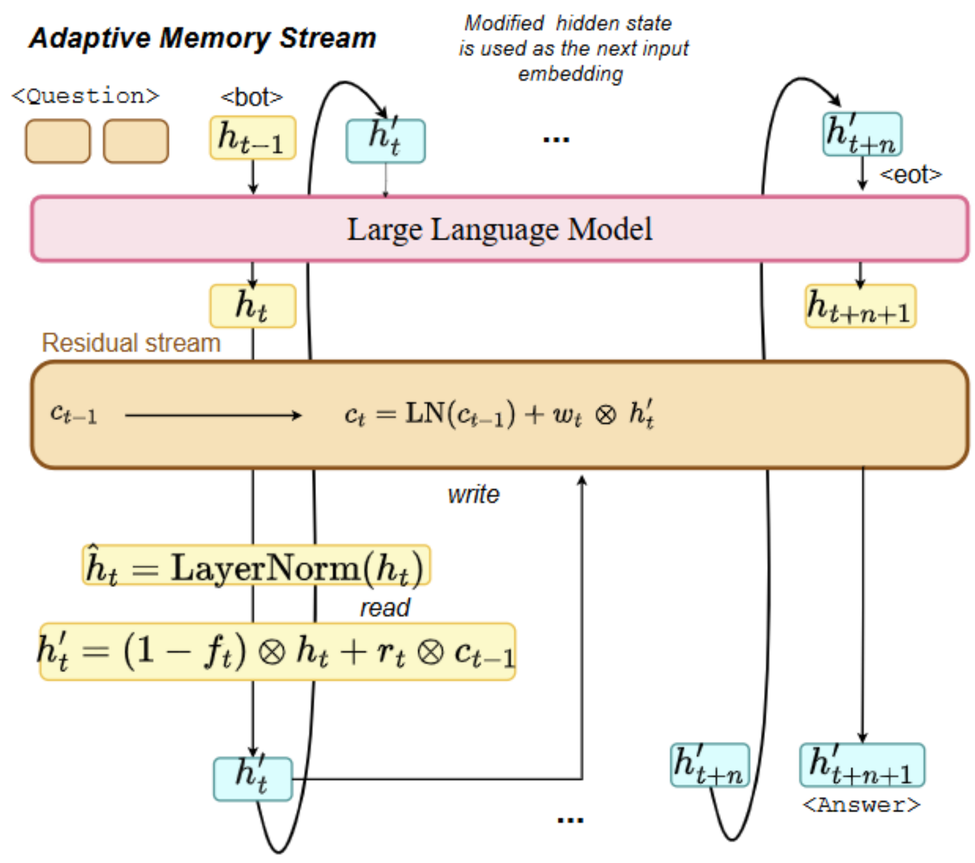
Background

- **CoCoNuT** (Chain of Continuous Thought) extends **Chain-of-Thought** beyond single-pass **language-space** reasoning by feeding each hidden state back as the next input, enabling **latent-space** reasoning with **implicit breadth-first search**.
- We identify a **concept bottleneck**: each reasoning pass **overwrites** the previous intermediate hidden states, so critical facts are lost as **reasoning depth** increases, reaching **87% information loss** by pass 6.
- Empirically, vanilla CoCoNuT (**10.4% EM**) still fails to beat CoT (**11.0% EM**) on **HotpotQA**, and it **degrades with curriculum depth** on GSM8K.

Objective / Aims

- **Research question:** Can a **persistent residual memory** across reasoning passes, extending the **residual stream** to **tokens** rather than only layers, resolve the **concept bottleneck** in continuous latent reasoning?
- **Aim 1:** Empirically **demonstrate the concept bottleneck** in vanilla CoCoNuT across **arithmetic** (GSM8K), **multi-hop QA** (HotpotQA), and **planning** (ProsQA).
- **Aim 2:** Design and evaluate **AGCLR**, a **gated concept stream** using learned **read, forget, and write gates** that preserves intermediate facts across passes with only **1.41% additional parameters**.

Methods

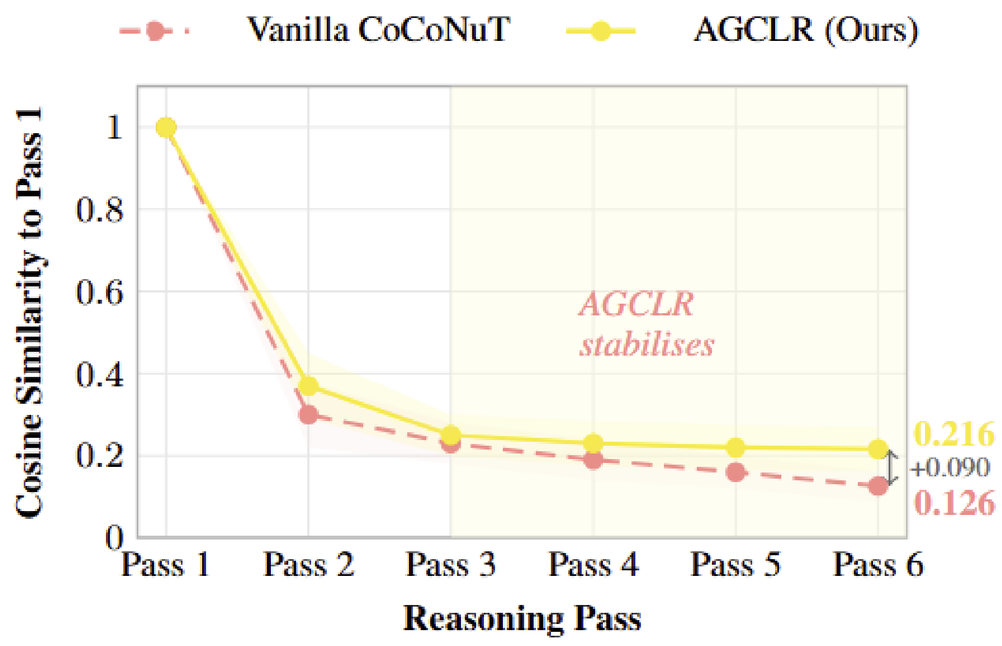


Limitations

- **Single-seed evaluation:** all results come from **one training run** per benchmark, with no **variance estimates** across seeds.
- **Model scale:** experiments use only **GPT-2 (124M)**; scalability to **1B+ models** remains unvalidated, though the architecture is **parameter-agnostic**.
- **Generalization claim:** evaluation covers three benchmarks; broader datasets (**MuSiQue**, **2WikiMultihopQA**, **StrategyQA**) and comparisons to **retrieval-augmented systems** are needed.

Results

Method	GSM8K	HotpotQA		ProsQA
	Acc. (%)	EM (%)	F1 (%)	Acc. (%)
CoT (Wei et al., 2022)	40.6	11.0	15.5	55.0
No-CoT (Hao et al., 2024)	16.5	4.0	7.6	76.7
iCoT (Deng et al., 2024)†	30.0	6.6	9.4	98.2
Pause Token (Goyal et al., 2023)†	16.4	10.6	14.6	75.9
Vanilla CoCoNuT (Hao et al., 2024)	31.4	10.4	15.2	92.0
AGCLR (Ours)	**34.0**$_{+2.6}$	**14.0**$_{+3.6}$	**19.4**$_{+4.2}$	**96.0**$_{+4.0}$



Conclusion

- **AGCLR** outperforms vanilla CoCoNuT on all three datasets: **+2.6%** (GSM8K), **+3.6% EM / +4.2% F1** (HotpotQA), **+4.0%** (ProsQA), with gains **compounding at deeper curriculum stages**.
- It resolves the **concept bottleneck**: AGCLR retains **71% more information** at final generation and reaches **96% on ProsQA at stage 6**, where vanilla **degrades to 92%**.
- Ablations confirm gains stem from **persistent memory, not parameters**: freezing the **write gate** after pass 2 costs only **0.8% EM**, so **early capture suffices**.

Future Directions

- **Scale** AGCLR to **larger base models** (1B+ parameters) and validate with **multi-seed evaluation** for variance estimates.
- **Broaden benchmarks** to additional multi-hop and reasoning datasets, including **MuSiQue**, **2WikiMultihopQA**, and **StrategyQA**, to further test generalization.
- Compare against **retrieval-augmented systems** and apply **fine-grained feature attribution** to deepen understanding of what the **concept stream** stores

References

linkedin

paper site