Recirculation

Feedback across depth and token processing in a frozen Transformer

Paper by Michael C. Mozer, Shoaib Ahmed Siddiqui, Danny Sawyer, Sunny Sanyal, Rosanne Liu

Research group discussion 23 October 2026
arXiv:2608.17981v3

The paper tests two distinct interventions

Fixed Recirculation

Choose source layer, destination layer, and feedback strength; change the inference graph while keeping *all* weights frozen.

Adaptive Recirculation

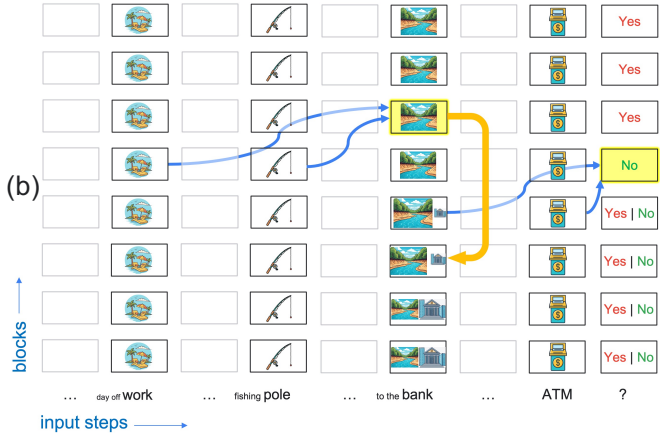
Freeze the foundation model, but train a small MLP to predict token-specific mixing vectors.

- Do not attribute the adaptive GSM8k gain to a training-free method.

Paper, Abstract, Secs. 2 and 4.6, PDF pp. 1, 4–5, 15–16.

Motivation: deep layers resolve context too late

- In a fishing story, *bank* means the river edge.
- Shallow layers may still mix river and financial senses.
- Prior targeted patching reuses the deep interpretation for a later ATM question.



Paper, Fig. 2(b), PDF p. 3. Cropped from the original vector figure. The cited targeted-patching result is prior work.

Recirculation gives a token a second pass

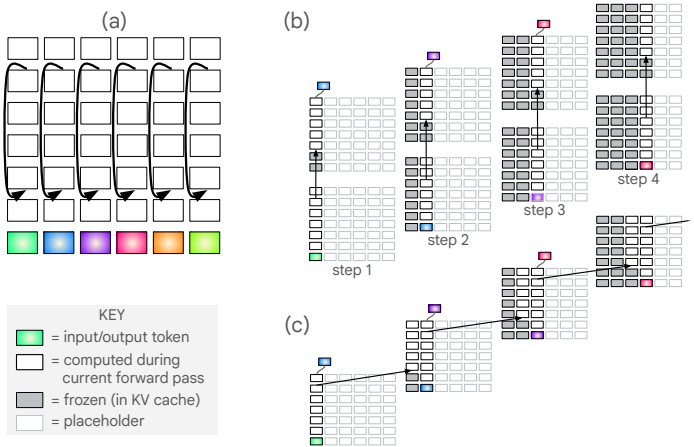
1. First pass of token t: compute a deep source state at layer s.
2. At the next processing step, feed a small amount back to a shallower layer $d < s$ of *the same token*.
3. Run that second pass alongside the first pass of token $t + 1$.
4. Later tokens can attend to the revised representation of token t.

Scope of experiments

Exactly one additional pass per token in the reported main experiments.

Paper, Sec. 2 and Fig. 3(c), PDF pp. 4–5; Appendix A, PDF p. 25.

The unrolled architecture pipelines two stacks



Paper, Fig. 3, PDF p. 5. (b) repeats layers within a token; (c) recirculates across processing steps.

Feedback is a small, norm-matched residual addition

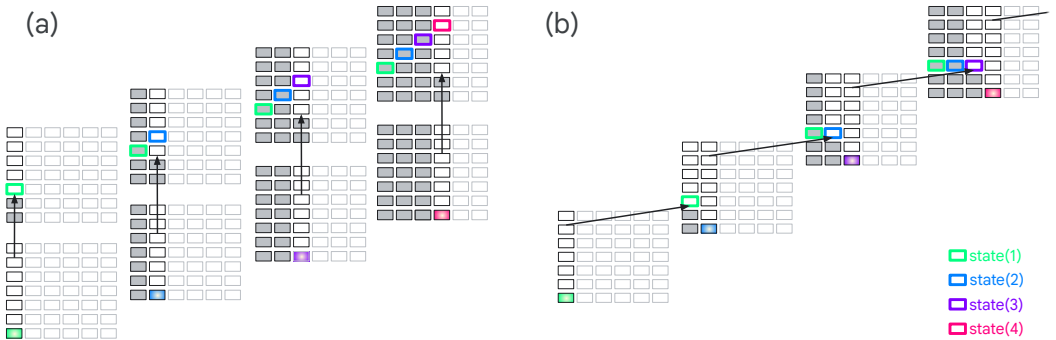
$$z_{t+1,t,d} = \alpha f(z_{t,t,s} \mid d, t) + \beta z_{t,t,d}, \qquad s > d.$$

$$f(z \mid d, t) = \frac{\|z_{t,t,d}\|_2}{\|z\|_2} z.$$

- $z_{i,j,l}$: residual output at processing step i, token j, layer l.
- Fixed experiments usually set $\beta = 1$; the update is *not* a convex average.
- Norm matching keeps the deeper activation from overwhelming the shallower one.

Paper, Eqs. (1)–(2), Sec. 2, PDF p. 5.

A state can persist at the same layer across tokens



Paper, Fig. 4, PDF p. 5. (a) Looped depth pushes state upward; (b) recirculation provides a same-layer path. This is an architectural argument, not a learned-state theorem.

The computational tradeoff sits in prefill

- **Generation:** the authors pipeline two stacks and argue that modern hardware can overlap them with little extra latency.
- **Prefill:** state updates depend on preceding tokens, so the full prompt must be processed sequentially.
- The paper reports no measured milliseconds, token/s, or extra-FLOP benchmark for this tradeoff.

Practical implication

A long prompt could dominate end-to-end latency even if per-token generation remains fast.

Paper, Secs. 2 and 5, PDF pp. 5, 16–18.

Evaluation uses separate tuning and scoring splits

- Search s, d, α using 1024-token windows from arXiv, C4, and PG-19 training splits.
- Score held-out perplexity on ten corpora; also report LongPPL for important context tokens.
- Evaluate instruction following, contextualization, eight zero-shot short-answer tasks, GSM8k, and HumanEval.
- Gemma3 1B/4B/12B are central; four other families receive qualitative layer sweeps.

Paper, Sec. 4 and Appendix B, PDF pp. 7–16, 26–28.

Useful source–destination pairs form a region

Gemma3 PT	Source s	Destination d	α
1B	10	4	0.07
4B	15	8	0.09
12B	22	13	0.08

- Three tuning datasets yield similar beneficial regions, mostly in middle layers.
- The wrong pair can harm: the paper reports about **+15% perplexity** at $s = 33, d = 14$.
- The optimal region changes with model size and instruction tuning.

Paper, Fig. 6 and Sec. 4.1, PDF pp. 7–8; Appendix Table B.2, PDF p. 27.

Standard PT perplexity gains are consistent but small

Gemma3 4B	PT + BOS	LongPPL	IT	PT without BOS
arXiv	0.25%	0.72%	9.22%	11.26%
PG-19	0.20%	1.18%	10.76%	15.72%
BillSum	0.09%	0.26%	18.09%	4.46%

- Cells are *relative perplexity reductions*; column conditions differ.
- Normal PT: all ten corpora improve by 0.09–0.62%; token-level effect size $d < 0.10$.
- Removing BOS deliberately raises baseline perplexity.

Paper, Table 1 and Sec. 4.2, PDF pp. 8–9.

Temperature tuning explains part of the IT result

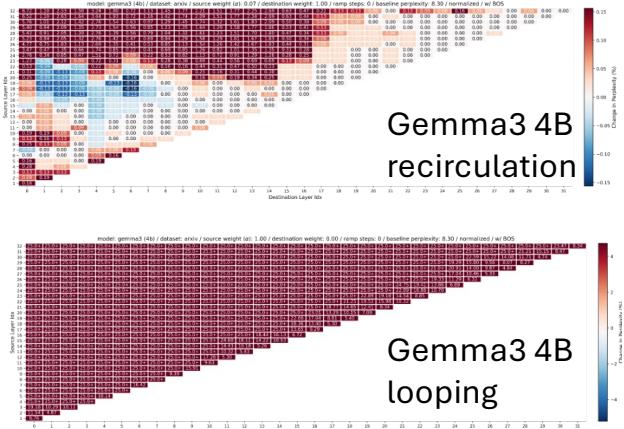
- On PG-19 with 4B IT, temperature tuning alone reduces perplexity by up to **19.5%**.
- Temperature plus recirculation reduces it by **20.2%**: an incremental 0.7 points.
- With ordinary BOS-present PT input, the incremental gain after temperature tuning is about 0.18 points.
- With BOS omitted, the extra gain is about 12 points.

Interpretation

Feedback does more than change output temperature, but calibration overlaps with its large IT perplexity effect.

Paper, Sec. 4.4 and Appendix C, PDF pp. 10–11, 33–34.

Naive training-free looping behaves differently



Blue indicates lower perplexity. The 1B and 12B panels show a similar broad contrast.

Paper, Fig. 8 and Sec. 4.4, PDF pp. 10–11. Gemma3 4B panels cropped from the original vector figure; top: recirculation, bottom: naive looping. This control does not assess trained looping.

Early and content-rich tokens help more

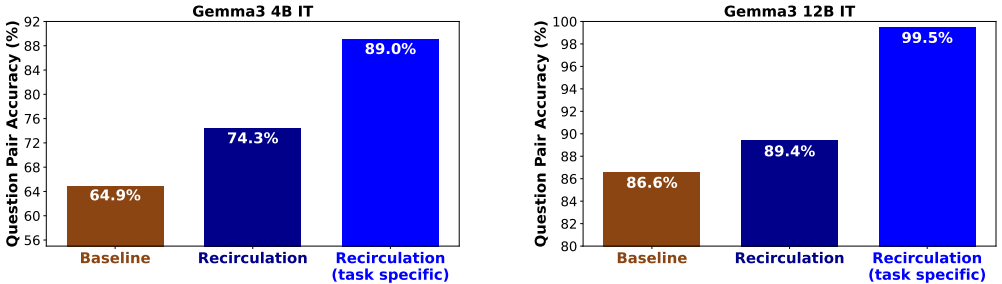
- A single recirculated token mostly helps nearby predictions, with a longer residual tail.
- Recirculating early tokens has a more persistent effect across a 1024-token window.
- Verbs and plural nouns tend to contribute more than numerals and determiners in the arXiv/Gemma3 4B analysis.

Why this matters

Uniformly applying feedback to every token may spend work where little useful state is propagated.

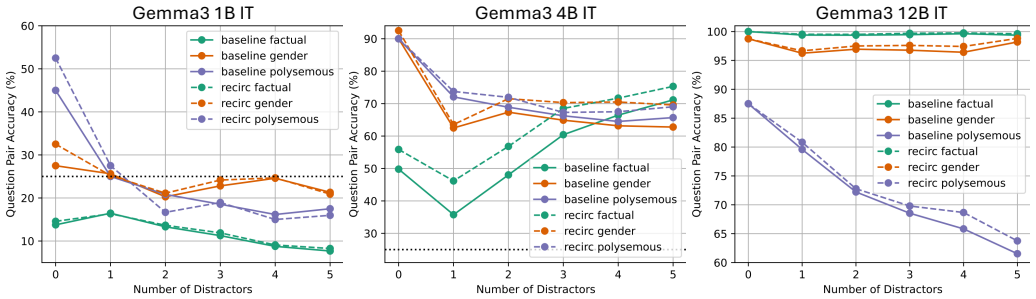
Paper, Fig. 9 and Sec. 4.4, PDF pp. 11–12; Appendix C, PDF pp. 34–35.

Instruction following improves with fixed feedback



Paper, Fig. 10 and Sec. 4.5, PDF p. 12. Preselected settings: 4B 64.9 to 74.3%, 12B 86.6 to 89.4%. Task-specific bars are tuned on this task and are upper bounds.

Contextualization gains vary by question type



Paper, Fig. 11, PDF p. 13; Appendix Table D.1, PDF p. 36. 12B factual/gender are near ceiling; with five distractors, polysemy error is 38.47 to 36.25%.

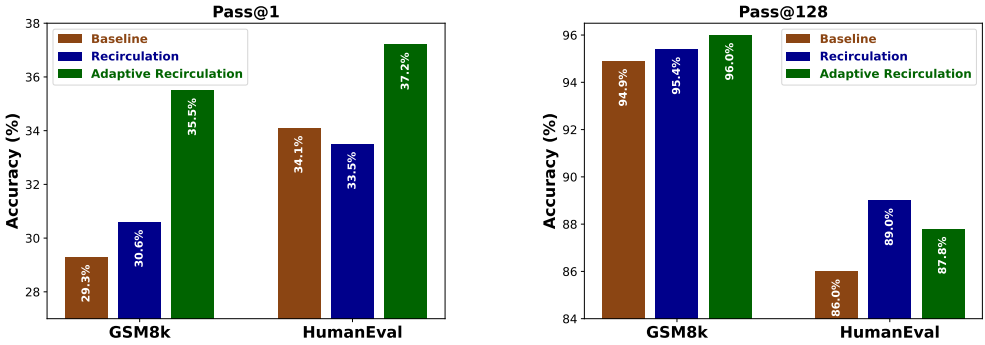
Single-token benchmarks move only slightly

Gemma3 4B PT, zero shot	Baseline	Fixed recirc.	Change (points)
MMLU	57.90	58.28	+0.38
ARC Easy	81.78	82.07	+0.29
PiQA	79.98	80.52	+0.54
HellaSwag	75.89	75.86	−0.03
WinoGrande	69.46	69.38	−0.08

- Six of eight tasks improve; eight-task mean is about **+0.23 points**.

Paper, Table 2 and Sec. 4.5, PDF p. 14. Layers/strength selected with MMLU development data.

The 21% GSM8k claim belongs to the adaptive variant



Paper, Fig. 12, PDF p. 15. GSM8k pass@1: 29.3 to 35.5% is +6.2 points, or +21% relative accuracy. The adaptive variant trains a task-related MLP.

Adaptive feedback learns per-token, per-dimension weights

$$z' = \boldsymbol{\alpha}(s, d) \odot f(s) + \boldsymbol{\beta}(s, d) \odot d.$$

- A two-hidden-layer MLP consumes source and destination residual vectors.
- It predicts $\boldsymbol{\alpha}, \boldsymbol{\beta} \in [0, 1]^D$ for each token.
- The foundation model is frozen; the MLP is trained with backpropagation through time.
- Language-model ablation uses 1B PT, 1024-token windows, 100 training steps, and many BOS-absent windows.

Paper, Sec. 4.6, PDF pp. 15–16; Appendix D.5, PDF p. 38.

Adaptive vectors beat scalar mixing in the ablation

1B PT variant	Mean relative PPL reduction, 9 corpora
Fixed coefficients	8.5%
Learned constant scalars	8.4%
Learned conditional scalars	11.7%
Learned constant vectors	20.6%
Learned conditional vectors	**23.0%**
Full model fine-tuning	21.6%

- These runs include BOS-absent windows; compare within this ablation.

Paper, Fig. 13 and Sec. 4.6, PDF pp. 15–16.

Adaptive task gains depend on the tuning set

- Training the MLP on part of the *MMLU test split* yields 59.83% on the full test split; training examples overlap evaluation.
- Using the separate MMLU auxiliary train split yields 58.63% (baseline 57.90%).
- Training on ARC Easy / ARC Challenge instead reduces MMLU to 56.05 / 55.02%.
- HumanEval pass@128: adaptive 87.8% is below fixed recirculation's 89.0%.

Paper, Table 2, Fig. 12 and Sec. 4.6, PDF pp. 14–16.

Take-home messages

1. Recirculation feeds a token's deep state into its later shallow pass and pipelines passes across tokens.
2. Fixed feedback is training-free: consistent but small normal-PT PPL gains, with larger gains on some state-sensitive tasks.
3. Adaptive feedback trains an MLP and brings larger gains under its training and evaluation conditions.
4. Long-prompt prefill cost, task-dependent tuning, and evaluation overlap matter before deployment claims.

Paper, Secs. 2, 4–5; Tables 1–2 and Figs. 10–13.

Backup: 1B and 12B do not improve every corpus

- With BOS omitted, the appendix reports broad PPL improvements for 1B, 4B, and 12B PT.
- LAMBADA is an exception: relative PPL "reduction" is -0.72% for 1B and -2.81% for 12B.
- The selected source/destination and feedback strength differ by model and input condition.

Paper, Appendix Tables B.2–B.3, PDF pp. 27–28.

Backup: why prefill remains serial

- Token t must produce its source representation before its second pass can update the representation available to later tokens.
- This dependence crosses token steps even when the full prompt is known.
- The paper proposes blockwise recirculation as future work, but does not evaluate the block-size/quality tradeoff.

Paper, Secs. 2 and 5, PDF pp. 4–5, 17–18.