LOOPAHEAD: PARALLELIZING RECURRENT DEPTH IN LOOPED TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Looped Transformers increase computation per token by repeatedly applying a shared backbone, offering a way to improve language modeling and downstream task performance at a fixed backbone parameter count. However, each loop depends on the preceding loop, so increasing recurrent depth also increases decoding latency. Existing approaches include Parallel Loop Transformers, which reduce latency by overlapping loop computation across token positions, but yield rapidly diminishing marginal accuracy gains at larger depths. To address this challenge, we introduce **LoopAhead**, which processes hidden states from multiple loop depths in parallel for the same token prefix. A lightweight predictor estimates inputs to later loops so they can start before earlier loops finish. A corrector then refines the outputs of these later loops using newly available results from earlier loops. All components are trained jointly using only the final next-token prediction loss. In experiments with 1.4B models, deeper LoopAhead configurations achieve progressively higher average zero-shot accuracy across eight tasks, rising from 53.41% to 56.38%. Vanilla looped models and PLT show limited gains over the same depth range. Compared with a nine-loop vanilla model, LoopAhead uses five sequential stages and additional parallel computation to reduce decoding latency by 46%.

1 INTRODUCTION

Looped Transformers increase computation per token by repeatedly applying a shared backbone (Dehghani et al., 2019; Geiping et al., 2025a). This recurrent depth can improve language modeling and downstream task performance at a fixed backbone parameter count (Saunshi et al., 2025). However, each loop depends on the preceding loop's output, so greater depth requires more sequential backbone stages and increases the wall-clock time needed to generate each token.

Existing approaches reduce decoding latency by overlapping recurrent computation across token positions. Parallel Loop Transformers (PLT) shift hidden states by one token between loops, allowing different loop depths to run concurrently using states from neighboring tokens (Wu et al., 2025). Adding loops initially improves accuracy, but the positional mismatch persists at every loop and the marginal gains quickly diminish at larger depths (Yang et al., 2026). Recurrent-depth parallel samplers refine several token drafts in parallel, but reducing the number of recurrent refinement steps between draft updates trades accuracy for faster decoding (Geiping et al., 2025b). The challenge is to start later loops before their inputs are available while still incorporating the results of earlier loops.

We address this challenge with LoopAhead, which batches hidden states from different loop depths so that the shared backbone can process them in parallel. Specifically, a lightweight predictor estimates the inputs to later loops, allowing them to start before the preceding loops finish. After each batched evaluation, a corrector uses the newly available earlier-loop results to refine outputs computed from the estimated inputs. These corrected states feed the next stage, and the deepest corrected state predicts the next token.

With 1.4B models, LoopAhead achieves lower PPL and higher mean zero-shot accuracy as the target depth index increases (Figures 3 and 4). At target depth index $D = 9$, LoopAhead requires only five sequential stages rather than nine, reducing latency by 46%. It improves accuracy on all eight evaluated tasks, with Avg-8 increasing by 2.20 percentage points relative to the vanilla

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

looped model, at the cost of additional parallel computation. At matched training backbone-compute budgets of 1.68 and 2.16 ZFLOPs, and separately at comparable decoding latency, LoopAhead achieves lower PPL and higher mean zero-shot accuracy than vanilla looped models. LoopAhead thus improves the use of recurrent computation in two complementary ways: it can yield better quality at a fixed total backbone budget, and it can exploit available parallelism to perform additional recurrent computation without proportionally increasing decoding latency.

2 Related Work

Scaling latent computation. Recurrent-depth language models increase computation by repeatedly applying a shared transition, as in Universal Transformers (Dehghani et al., 2019). Such looping can approach the accuracy of deeper models with untied layers on arithmetic and question-answering tasks (Saunshi et al., 2025). PonderLM changes the input passed between loops, feeding a probability-weighted average of token embeddings into the next loop (Zeng et al., 2026). Geiping et al. (2025a) sample recurrent depths during training to support variable test-time computation. Ouro instead learns an input-dependent exit gate for adaptive recurrent depth (Zhu et al., 2025). Hidden Decoding takes a different route, expanding each token into multiple embedding streams that the backbone processes together, while retaining their intermediate KV states as context (Liu et al., 2026).

Parallel recurrent computation. To shorten the sequential path between loops, PLT and recurrent-depth parallel samplers overlap computation across token positions. PLT pipelines loop depths using states shifted across token positions (Wu et al., 2025). Recurrent-depth parallel samplers accelerate decoding from pretrained checkpoints by refining several token drafts concurrently, relying on per-loop input injection and decodable intermediate states (Geiping et al., 2025b). LoopAhead uses predicted inputs and output correction to run several backbone loops in parallel for the same token prefix.

Parallel correction methods. Classical Parareal uses a fine propagator for accurate parallel local solves and a cheaper coarse propagator for sequential correction, iterating toward the fine propagator's sequential solution (Lions et al., 2001). Neural variants replace or parameterize the coarse component with learned networks (Lee et al., 2021). Related methods parallelize network evaluation through fixed-point or Newton iterations (Song et al., 2021; Lim et al., 2024), with SNLP using structured Jacobian approximations for Transformer layers (Han et al., 2026). LoopAhead adapts parallel evaluation and correction to a shared recurrent backbone. Its predictor initializes later-loop inputs once, allowing several loops to run together, while its corrector incorporates newly available earlier-loop states into the parallel outputs. Unlike classical Parareal, which iteratively seeks to recover the sequential fine solution, LoopAhead learns a fixed-stage parallel computation optimized directly for next-token prediction. Its objective is therefore not to match the sequential recurrent trajectory, but to produce predictive states under a constrained number of sequential stages.

3 Method

LoopAhead supplies predicted inputs to later loops so that several recurrent depths can be evaluated together in a batch. For example, $1 + 3 + 3$ first computes the state at depth index 1 and predicts provisional states at indices 2 and 3. The next two sequential stages produce corrected states at indices $2, 3, 4$ and then $3, 4, 5$, so the final state has target depth index $D = 5$ (Figure 1).

Let $H_0 = \mathrm{Embed}(x_{1:L}) \in \mathbb{R}^{d\times L}$, with each column representing one token, and let $\mathcal{F}$ be the shared causal Transformer backbone. Its sequential states satisfy $H_j = \mathcal{F}(H_{j-1})$. For parallel width n and total stage count S, a schedule $1+n+\cdots+n$ uses one initial backbone evaluation followed by $S-1$ stages of n parallel evaluations. Each schedule entry counts backbone evaluations in one sequential stage. A schedule is specified by (n, S); $D = n+S-1$ is the target depth index of its final corrected state. Different schedules can share the same D. Let $Z_j^{(s)}$ denote the provisional state at depth index j after stage s.

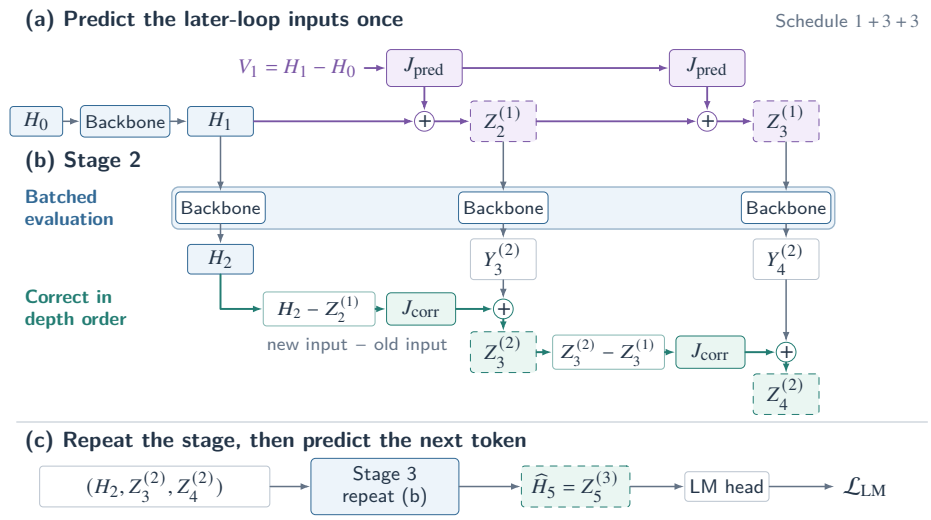


Figure 1: **LoopAhead with schedule** $1 + 3 + 3$**.** Three stages and width 3 reach target depth index $D = 5$ using seven backbone evaluations. (a) Predict deeper inputs from $H_1 - H_0$. (b) Evaluate three states together and correct using earlier-loop updates. (c) Repeat the stage and predict the next token.

3.1 Predicting Inputs for Later Loops

After computing $H_1 = \mathcal{F}(H_0)$, we use its state change $V_1 = H_1 - H_0$ as an initial increment along recurrent depth. A shared token-wise linear map J_{pred} repeatedly transforms this increment, and the accumulated changes initialize deeper states. At identity initialization, this is simply linear extrapolation of the first loop's state change:

$$Z_{1+\ell}^{(1)} = H_1 + \sum_{m=1}^{\ell} J_{\text{pred}}^m V_1, \qquad \ell = 1, \ldots, n-1. \tag{1}$$

Here, J_{pred}^m denotes m successive applications of the shared predictor map. For $J_{\text{pred}} = I$, the formula reduces to $Z_{1+\ell}^{(1)} = H_1 + \ell V_1$. Training adapts the increments to the language-modeling objective. Prediction runs only once to supply the $n-1$ states that join H_1 in the next batch.

3.2 Parallel Evaluation and Correction

At stage $s = 2, \ldots, S$, we batch one sequential state with the provisional states from the preceding stage and evaluate them together:

$$\begin{aligned} &\left(H_s, Y_{s+1}^{(s)}, \ldots, Y_{s+n-1}^{(s)}\right) \\ &\quad = \text{Split}\left[\mathcal{F}\left(\text{Concat}_{\text{batch}}\left(H_{s-1}, Z_s^{(s-1)}, \ldots, Z_{s+n-2}^{(s-1)}\right)\right)\right]. \end{aligned} \tag{2}$$

Here, $\text{Concat}_{\text{batch}}$ concatenates the n recurrent states along the batch dimension, $\mathcal{F}$ evaluates the n states in parallel, and Split separates the batched output back into its n lanes. All lanes use the same token positions and causal mask, while attention is computed independently within each lane. The first output advances the sequential recurrence to H_s; the remaining $Y_j^{(s)}$ are raw backbone outputs from the provisional inputs.

The backbone output $Y_{j+1}^{(s)}$ is stale because it was computed from the previous estimate of its input. We therefore correct the output according to how much that input estimate has changed. A shared token-wise linear map J_{corr} translates this input change into an output adjustment. Starting from $Z_s^{(s)} = H_s$, we update the remaining outputs in depth order:

$$Z_{j+1}^{(s)} = Y_{j+1}^{(s)} + J_{\text{corr}}\left(Z_j^{(s)} - Z_j^{(s-1)}\right), \qquad j = s, \ldots, s+n-2. \tag{3}$$

Table 1: **Backbone computation per token (any model size).** D: target depth index; S: sequential stages; n: parallel width after the first loop. Evaluations count all parallel lanes.

Model	Target depth index D	Backbone evaluations	Sequential stages S	Width n
Vanilla looped model	D	D	D	1
LoopAhead	$n+S-1$	$1+n(S-1)$	S	n

Because each corrected state is used to correct the next depth, newly available information propagates sequentially across the parallel outputs.

3.3 Training Objective and Computational Cost

After S stages, the deepest state $\widehat{H}_D = Z^{(S)}_{S+n-1}$, with target depth index $D = n+S-1$, produces the next-token logits and language-modeling loss:

$$\ell_t = W_{\mathrm{LM}}\,\mathrm{RMSNorm}(\widehat{H}_D)_t, \qquad \mathcal{L}_{\mathrm{LM}} = -\sum_{t=1}^{L-1} \log \mathrm{softmax}(\ell_t)_{x_{t+1}}. \tag{4}$$

All components are trained jointly using only this loss, learning intermediate states for next-token prediction. The predictor and corrector use separate parameter matrices, each written as $J = I+\Delta J$, with $\Delta J \in \mathbb{R}^{d\times d}$ initialized to zero. Each map is shared across depths and token positions; the corrector is also shared across stages. During generation, the prefix stays fixed throughout the stages, and the next token is selected from the final output.

Counting each parallel lane gives $1+n(S-1)$ backbone evaluations in S sequential stages (Table 1). The two maps add $2d^2$ parameters (8.39M at 1.4B) and $S(n-1)$ token-wise linear-map applications per schedule.

4 Experiments

We evaluate LoopAhead under three complementary comparison regimes. Matched-training-compute comparisons test whether LoopAhead can improve model quality at the same total backbone budget. Comparable-latency comparisons test whether additional backbone evaluations can be shifted onto the parallel dimension to improve quality without increasing the number of sequential stages. Finally, matched-target-depth comparisons characterize quality when LoopAhead reaches the same target depth index with fewer sequential stages. Because these regimes hold different quantities fixed, we report them separately rather than interpreting them interchangeably.

4.1 Experimental Setup

Models and training. We train 1.4B models from scratch on the Pile (Gao et al., 2020) using a LLaMA-style backbone (Touvron et al., 2023), a global batch size of 256, and a common 30B-token budget. Each schedule is trained separately with the same tokenizer, data order, and training recipe. The learning rate warms up to 2×10^{-4} over the first 1% of steps, then follows cosine decay to 2×10^{-5}. The ablations use 410M models trained on 25B tokens.

Baselines. A vanilla looped model applies the shared backbone sequentially D times. Our PLT-D baselines use cross-loop pipelining without KV sharing (Wu et al., 2025). Hidden Decoding uses 5, 7, or 9 embedding streams per token (Liu et al., 2026). We omit recurrent-depth parallel samplers because they require per-loop input injection that our backbone does not use (Geiping et al., 2025b).

Quality metrics. We report Pile validation perplexity (PPL) and zero-shot accuracy on LAMBADA-OpenAI (LAMBADA-OAI) (Paperno et al., 2016; Radford et al., 2019), ARC-Easy and ARC-Challenge (Clark et al., 2018), WinoGrande (Sakaguchi et al., 2020), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), SciQ (Welbl et al., 2017), and RACE (Lai et al., 2017).

Training computation. We report total training backbone computation in ZFLOPs. For the 1.4B models, one backbone evaluation per original token corresponds to 0.24 ZFLOPs over the train-

ing run. The evaluation count per token is $1 + n(S - 1)$ for LoopAhead, D for vanilla looped models and PLT, and the stream count for Hidden Decoding. This counts computation across all sequential stages, parallel lanes, and expanded token representations; the learned-map cost is given in Section 3.3.

Decoding latency. We measure wall-clock milliseconds per generated token with a 1,024-token prompt and a 1,024-token greedy continuation. The main comparisons use a decoding batch size of 1. Section 4.5 reports aggregate throughput for batches up to 16. All decoding benchmarks run on the same GPU in BF16 with tensor parallelism disabled, and timings cover the complete decoding schedule.

4.2 Parallel Width at Fixed Stage Count

Parallel width controls how many recurrent states the backbone evaluates in each batched evaluation. Figure 2 varies width while holding the number of sequential stages fixed at $S = 3, 4, 5$, testing whether additional parallel computation improves PPL within the same sequential-stage budget.

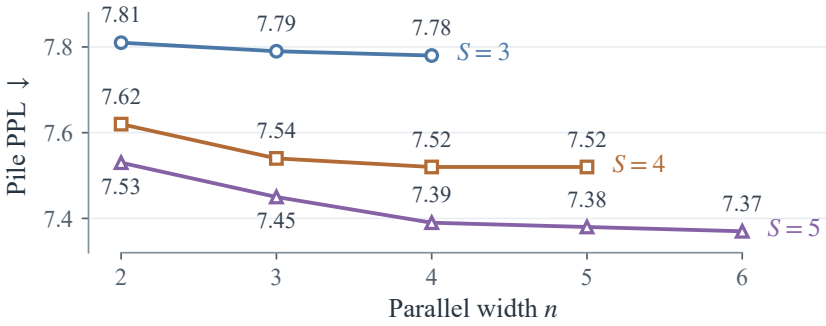


Figure 2: **Parallel width and Pile perplexity in LoopAhead (1.4B, 30B training tokens).** Each curve fixes sequential stages S and varies width n. Schedule $1 + n + \cdots + n$ reaches target depth index $D = n + S - 1$.

Increasing parallel width at a fixed stage count consistently improves PPL, showing that additional recurrent computation can improve model quality without extending the sequential path. However, the gain per added parallel lane decreases as width grows, indicating progressively lower returns from additional parallel computation.

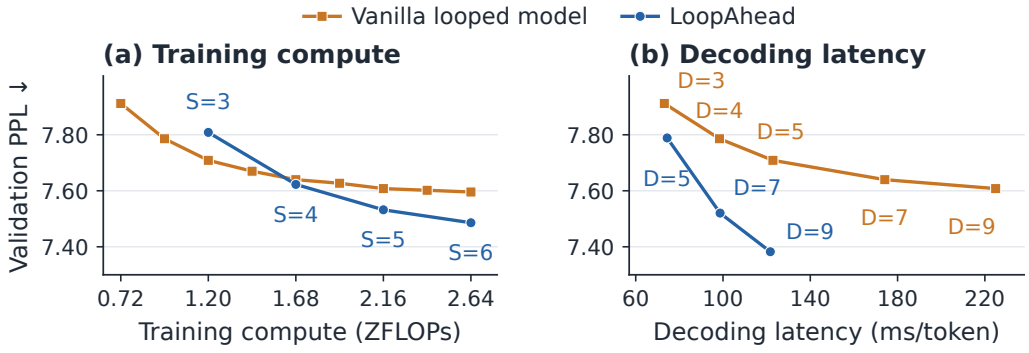


Figure 3: **Pile PPL versus training compute and decoding latency (1.4B, 30B tokens).** (a) LoopAhead: $n = 2$, $S = 3$–6, ending with $1 + 2 + 2 + 2 + 2 + 2$; vanilla looped: 3–11 loops. (b) LoopAhead uses $1 + 3 + 3$, $1 + 4 + 4 + 4$, and $1 + 5 + 5 + 5 + 5$ at $D = 5, 7, 9$. Timing: batch 1, 1,024-token prompts/continuations, greedy decoding, same GPU, BF16, tensor parallelism disabled.

Table 2: **Zero-shot accuracy at matched training backbone compute (1.4B, 30B Pile tokens).** Vanilla looped models and PLT use 5/7/9 loops; Hidden Decoding uses 5/7/9 streams per token. LoopAhead schedules $1+2+2$, $1+2+2+2$, and $1+2+2+2+2$ correspond to $D=4,5,6$. Accuracy is in percent; Avg-8 is the unweighted mean across eight tasks.

Model	Training (ZFLOPs)	LAMBADA OAI	ARC-E	ARC-C	Wino	PIQA	Hella	SciQ	RACE	Avg-8
Vanilla Looped	1.20	**59.65**	52.90	**28.24**	**57.14**	**69.86**	**50.55**	78.90	32.25	**53.69**
PLT		57.27	50.42	26.96	54.70	69.26	48.47	79.10	**33.11**	52.41
Hidden Decoding		54.41	48.23	26.62	52.09	68.99	43.66	75.70	30.33	50.00
LoopAhead ($n=2, S=3$)		57.60	**53.16**	28.16	55.33	69.59	49.57	**79.40**	**33.11**	53.24
Vanilla Looped	1.68	**60.14**	53.32	**28.75**	**57.77**	70.35	51.44	77.20	**34.64**	54.20
PLT		55.89	51.85	27.56	54.70	68.99	47.70	77.80	33.01	52.19
Hidden Decoding		52.92	48.32	25.60	52.49	68.88	43.44	74.70	30.62	49.62
LoopAhead ($n=2, S=4$)		58.08	**54.80**	**28.75**	57.14	**71.33**	**51.80**	**82.60**	33.21	**54.71**
Vanilla Looped	2.16	60.00	52.36	27.13	**57.85**	70.84	51.66	79.70	33.88	54.18
PLT		56.84	51.05	27.22	55.41	70.29	47.72	76.30	31.96	52.10
Hidden Decoding		53.19	49.20	26.54	51.78	68.23	43.65	73.40	30.62	49.57
LoopAhead ($n=2, S=5$)		**60.90**	**55.39**	**28.92**	56.75	**71.27**	**52.79**	**80.30**	**34.74**	**55.13**

Table 3: **Zero-shot accuracy and decoding latency (1.4B, 30B Pile tokens).** LoopAhead uses $1+3+3$, $1+4+4+4$, and $1+5+5+5+5$ at $D=5,7,9$, paired with 3/4/5-loop vanilla looped models. PLT/Hidden Decoding suffixes count loops/streams per token. Timing: batch 1, 1,024-token prompts/continuations, greedy decoding, same GPU, BF16, tensor parallelism disabled. Accuracy is in percent; Avg-8 is the unweighted mean across eight tasks.

Model	Latency (ms/token)	LAMBADA OAI	ARC-E	ARC-C	Wino	PIQA	Hella	SciQ	RACE	Avg-8
PLT-5	24.93	57.27	50.42	26.96	54.70	69.26	48.47	79.10	33.11	52.41
PLT-7	25.35	55.89	51.85	27.56	54.70	68.99	47.70	77.80	33.01	52.19
PLT-9	25.16	56.84	51.05	27.22	55.41	70.29	47.72	76.30	31.96	52.10
Hidden Decoding-5	23.63	54.41	48.23	26.62	52.09	68.99	43.66	75.70	30.33	50.00
Hidden Decoding-7	23.95	52.92	48.32	25.60	52.49	68.88	43.44	74.70	30.62	49.62
Hidden Decoding-9	23.27	53.19	49.20	26.54	51.78	68.23	43.65	73.40	30.62	49.57
Vanilla Looped $D=3$	73.11	56.36	50.21	27.65	**55.17**	69.86	48.40	75.20	33.01	51.98
LoopAhead $D=5$	74.41	**58.01**	**51.64**	**27.73**	54.78	**70.24**	**49.81**	**81.80**	**33.30**	**53.41**
Vanilla Looped $D=4$	98.40	58.24	52.57	28.16	54.85	70.02	50.39	77.90	34.35	53.31
LoopAhead $D=7$	98.66	**60.28**	**54.25**	**30.03**	**57.22**	**70.46**	**52.67**	**82.30**	**34.64**	**55.23**
Vanilla Looped $D=5$	122.88	59.65	52.90	28.24	57.14	69.86	50.55	78.90	32.25	53.69
LoopAhead $D=9$	121.68	**61.61**	**55.13**	**29.35**	**59.75**	**72.42**	**54.46**	**82.30**	**35.98**	**56.38**

4.3 Quality versus Training Compute and Decoding Latency

At matched training backbone compute, we fix LoopAhead's width at 2 and increase its stage count to compare models across the same training budgets (Figure 3a and Table 2). The vanilla looped model leads at 1.20 ZFLOPs, whereas LoopAhead achieves lower PPL and higher mean zero-shot accuracy at 1.68 and 2.16 ZFLOPs. Thus, the parallel schedule can improve prediction quality within the same training backbone budget while using fewer sequential stages. LoopAhead also exceeds PLT and Hidden Decoding in mean accuracy across all three budgets.

At comparable decoding latency, LoopAhead at target depth indices $D=5,7,9$ is compared with vanilla looped models using 3, 4, and 5 loops, respectively (Figure 3b and Table 3). Each pair uses the same number of sequential stages, allowing LoopAhead to perform additional recurrent computation in parallel within a similar time per token. This yields lower PPL and higher mean accuracy in all three pairs, with gains on all eight tasks for the two deeper schedules. PLT and Hidden Decoding decode faster but attain lower mean accuracy.

4.4 Quality Across Target Depth Indices

To compare how model quality evolves across deeper configurations, Figure 4 reports mean zero-shot accuracy at matched target depth indices $D = 5, 7, 9$. The corresponding LoopAhead schedules are $1+3+3$, $1+4+4+4$, and $1+5+5+5+5$, which jointly increase parallel width, sequential stage count, and total backbone computation.

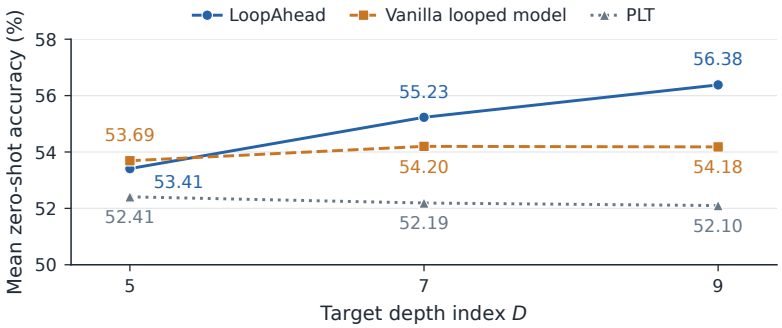


Figure 4: **Mean zero-shot accuracy across target depth indices (1.4B, 30B Pile tokens).** At $D = 5, 7, 9$, LoopAhead uses $1+3+3$, $1+4+4+4$, and $1+5+5+5+5$ in 3/4/5 stages; vanilla looped models and PLT use D loops. Accuracy is the unweighted mean (%) across eight tasks.

Across these configurations, LoopAhead's mean zero-shot accuracy increases from 53.41% at $D = 5$ to 56.38% at $D = 9$, whereas the vanilla looped and PLT results vary less over the same target depth indices. At $D = 9$, the evaluated LoopAhead configuration achieves 2.20 percentage points higher Avg-8 accuracy, lower PPL, and 46% lower decoding latency than the vanilla looped model, while using additional parallel backbone computation.

4.5 Throughput with Larger Decoding Batches

Because LoopAhead batches recurrent states, decoding more sequences also enlarges the batch processed by the backbone. Table 4 reports throughput across decoding batch sizes from 1 to 16. We compare LoopAhead $1+3+3$ ($D = 5$, three sequential stages) with a five-loop vanilla looped model.

Table 4: **Decoding throughput at $D = 5$ (1.4B, tokens/s).** LoopAhead uses $1+3+3$, with seven backbone evaluations in three stages, versus five loops in the vanilla looped model. B is batch size. Timing: 1,024-token prompts/continuations, greedy decoding, same GPU, BF16, tensor parallelism disabled. Speed ratio: LoopAhead/vanilla looped throughput at equal B; blanks indicate unavailable measurements.

Model	$B = 1$	$B = 2$	$B = 4$	$B = 8$	$B = 16$
Vanilla Looped $D = 5$	8.14	16.15	31.67	65.31	130.18
LoopAhead $D = 5$	13.44	26.24	52.26	105.35	209.12
Speed ratio to vanilla looped model	1.65×	1.62×	1.65×	1.61×	1.61×

At $B = 1$, LoopAhead achieves 1.65× the throughput of the five-loop vanilla looped model. As batch size increases, LoopAhead throughput grows approximately linearly, maintaining stable per-sequence decoding throughput across the evaluated batch sizes.

4.6 KV Sharing Across Recurrent Stages

Although LoopAhead decodes at latency close to a shallower vanilla looped model, storing an independent KV cache for every recurrent evaluation adds memory cost. We address this cost in the 1.4B $1+2+2+2$ model at target depth index $D = 5$, comparing the default *No KV sharing* policy with two reuse policies. *First-KV reuse* retains the KV tensors from the first evaluation at each loop

depth and reuses them in later stages, reducing the seven independent caches to five. To incorporate later-stage updates for recent tokens, *Hard overwrite-64* additionally refreshes the most recent 64 cached positions using newly computed sequential states. Each variant uses the same cache policy in training and inference.

Table 5: **KV sharing (1.4B, 30B Pile tokens, $1 + 2 + 2 + 2$, $D = 5$).** First-KV reuse retains each depth's first cache; Hard overwrite-64 also refreshes the latest 64 positions. Storage is relative to one full-backbone cache. Variants train with their inference cache policies. Timing: batch 1, 1,024-token prompts/continuations, greedy decoding, same GPU, BF16, tensor parallelism disabled.

Variant	Later-stage KV policy	KV storage	Latency (ms/token)	Pile PPL ↓
No KV sharing	independent cache	7×	**99.89**	**7.62**
First-KV reuse	reuse all positions	5×	103.21	7.68
Hard overwrite-64	overwrite most recent 64	5×	104.94	7.63

First-KV reuse reduces KV storage from $7\times$ to $5\times$, with slightly higher PPL and decoding latency. Hard overwrite-64 retains this memory saving and recovers most of the PPL loss by refreshing recent cached states. This refresh adds a small further latency overhead.

4.7 Contributions of Prediction and Correction

To identify how prediction and correction contribute to the quality gains, we use 410M models with the $1 + 4 + 4 + 4$ schedule. Component removal tests the value of each operation, identity maps test the benefit of learning its transformation, and state measurements examine how provisional states change across stages.

4.7.1 Component Removal

No prediction initializes every lane with the same H_1 while retaining correction. *No correction* keeps the learned initialization but passes raw backbone outputs directly to the next stage. All three variants retain the thirteen scheduled backbone evaluations and are trained separately.

Table 6: **Component ablations (410M, 25B Pile tokens, $1 + 4 + 4 + 4$, $D = 7$).** J_{pred} and J_{corr} denote prediction and correction. No prediction repeats H_1; No correction uses raw backbone outputs. Variants train separately; ΔPPL is the increase over the complete model.

Variant	Initial states	Output correction	Pile PPL ↓	ΔPPL
LoopAhead	J_{pred}	J_{corr}	**9.13**	–
No prediction	repeated H_1	J_{corr}	9.29	+0.16
No correction	J_{pred}	none	9.47	+0.34

Both components improve PPL under the same backbone budget. Removing correction has the larger penalty, highlighting the value of incorporating new earlier-loop results into the parallel states.

4.7.2 Learning the Prediction and Correction Maps

To test the value of learning each map, we hold one map fixed at identity throughout training while learning the backbone and the other map. *Identity prediction* therefore uses linear extrapolation, whereas *Identity correction* adds the input discrepancy directly to the output.

Both identity variants outperform removal of the corresponding component, showing that linear extrapolation and direct discrepancy addition already provide useful operations. Learning both maps further lowers PPL, so adapting these operations to the jointly trained backbone adds value beyond their identity forms.

Table 7: **Learned versus identity maps (410M, 25B Pile tokens, $1 + 4 + 4 + 4$, $D = 7$).** Each identity variant fixes prediction J_{pred} or correction J_{corr} at I throughout training, while learning the backbone and the other map.

Variant	J_{pred}	J_{corr}	Pile PPL ↓
LoopAhead	Learned	Learned	**9.13**
Identity prediction	I	Learned	9.21
Identity correction	Learned	I	9.20

4.7.3 Evolution of Learned States

To examine how the parallel states evolve, Table 8 revisits the three variants in Table 6. We compare their provisional states with H_2, H_3, H_4 from sequential rollouts of their own trained backbones on the same input. For each $Z_j^{(s)}$, we measure global cosine similarity and relative L_2 distance:

$$c_j^{(s)} = \frac{\langle Z_j^{(s)}, H_j \rangle_F}{\|Z_j^{(s)}\|_F \|H_j\|_F}, \qquad r_j^{(s)} = \frac{\|Z_j^{(s)} - H_j\|_F}{\|H_j\|_F}. \tag{5}$$

Table 8: **Provisional-state evolution (410M, 25B Pile tokens, $1 + 4 + 4 + 4$, $D = 7$).** Reference states come from each variant's own sequential backbone on the same input. S1 denotes initialization; arrows mark subsequent stages. Cosine uses full state tensors; relative L_2 divides the Frobenius error by the reference state's norm.

Variant	Metric	H_2 (S1)	H_3 (S1→S2)	H_4 (S1→S2→S3)
LoopAhead	Cosine ↑	0.922	0.774→0.811	0.636→0.653→0.823
	Relative L_2 ↓	0.394	0.642→0.588	0.809→0.792→0.573
No prediction	Cosine ↑	0.971	0.874→0.924	0.643→0.701→0.805
	Relative L_2 ↓	0.239	0.494→0.389	0.777→0.726→0.594
No correction	Cosine ↑	0.506	0.517→0.870	0.234→0.224→0.414
	Relative L_2 ↓	0.866	0.857→0.516	0.977→1.027→0.920

The complete model's state at depth index $j = 4$ moves closer to its sequential reference across stages, while removing correction leaves weaker alignment. This is consistent with correction carrying earlier-loop information into deeper provisional states.

Repeated-H_1 initialization starts closer to its own sequential states at depth indices 2 and 3, yet the complete model achieves better final PPL. This may indicate that minimizing instantaneous distance to the sequential trajectory is not necessary for good parallel refinement.

5 Conclusion

We introduced LoopAhead to parallelize recurrent computation through a jointly trained backbone, predictor, and corrector. Experiments show that this design can improve language modeling and zero-shot downstream performance with fewer sequential backbone stages, while the component analyses highlight the importance of correction. These results suggest that recurrent depth and sequential decoding depth need not coincide: additional recurrent computation can be shifted onto the parallel dimension when provisional states are subsequently corrected.

LoopAhead reduces sequential depth at the cost of additional backbone computation and thus higher FLOPs. It also increases KV-cache requirements; although KV sharing mitigates this overhead, our experiments show a modest trade-off in quality and/or latency. Reducing these computational and memory costs while preserving the benefits of parallel recurrent evaluation remains an important direction for future work.

AI USE STATEMENT

Generative AI tools were used for literature search, language editing, and figure preparation. The authors designed and conducted the experiments, and verified the citations, claims, and reported results against primary sources and experiment records. The authors take full responsibility for the final manuscript.

REFERENCES

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL https://arxiv.org/abs/2101.00027.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025a. URL https://arxiv.org/abs/2502.05171.

Jonas Geiping, Xinyu Yang, and Guinan Su. Efficient parallel samplers for recurrent-depth models and their connection to diffusion language models. *arXiv preprint arXiv:2510.14961*, 2025b. URL https://arxiv.org/abs/2510.14961.

Ligong Han, Kai Xu, Hao Wang, and Akash Srivastava. SNLP: Layer-parallel inference via structured newton corrections. *arXiv preprint arXiv:2605.17842*, 2026. URL https://arxiv.org/abs/2605.17842.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1082.

Chang-Ock Lee, Youngkyu Lee, and Jongho Park. Parareal neural networks emulating a parallel-in-time algorithm. *arXiv preprint arXiv:2103.08802*, 2021. URL https://arxiv.org/abs/2103.08802.

Yi Heng Lim, Qi Zhu, Joshua Selfridge, and Muhammad Firmansyah Kasim. Parallelizing non-linear sequential models over the sequence length. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=E34AlVLN0v.

Jacques-Louis Lions, Yvon Maday, and Gabriel Turinici. A "parareal" in time discretization of PDE's. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 332:661–668, 2001. doi: 10.1016/S0764-4442(00)01793-6.

Aiwei Liu, Cheng Shi, Chuhan Wu, Ci Lei, Di Lu, Donald He, Fan Zhang, Fanhao Kong, Feifei Zhang, Guan Wang, Haicheng Wang, Haoyu Liu, Houjin Yu, Jiachen Ding, Jiayi Feng, Jie Zhou, Jijun Chi, Jindi Shi, Jing Lei, Junjie Zhang, Laiyi Li, Le Tian, Linhao Zhang, Miao Fan, Sijun Zhang, Wei Jia, Weiwei Shi, Wenhan Li, Wentao Zhao, Wenteng Liang, Xiao Zhou, Xiaojin Zhou, Xihuai Wang, Xinyu Gao, Xuanliang Wang, Xuyang Ao, Yang Yu, Yangxiu You, Yinuo

Zhao, Yufei Kuang, Yufei Wang, Yuan Liu, Yuan Liu, Yuwen Chen, Zhencong Tian, Zhongyin Zhao, Zilin Yu, and Zitao Wang. Hidden decoding at scale: Latent computation scaling for large language models. *arXiv preprint arXiv:2607.08186*, 2026. URL https://arxiv.org/abs/2607.08186.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1525–1534. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1144.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=din0lGfZFd.

Yang Song, Chenlin Meng, Renjie Liao, and Stefano Ermon. Accelerating feedforward computation via parallel nonlinear equation solving. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9791–9800. PMLR, 2021. URL https://proceedings.mlr.press/v139/song21a.html.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-4413.

Bohong Wu, Mengzhao Chen, Xiang Luo, Shen Yan, Qifan Yu, Fan Xia, Tianqi Zhang, Hongrui Zhan, Zheng Zhong, Xun Zhou, Siyuan Qiao, and Xingyan Bin. Parallel loop transformer for efficient test-time computation scaling. *arXiv preprint arXiv:2510.24824*, 2025. URL https://arxiv.org/abs/2510.24824.

Jian Yang, Shawn Guo, Wei Zhang, Tianyu Zheng, Yaxin Du, Haau-Sing Li, Jiajun Wu, Yue Song, Yan Xing, Qingsong Cai, Zelong Huang, Chuan Hao, Ran Tao, Xianglong Liu, Wayne Xin Zhao, Mingjie Tang, Weifeng Lv, Ming Zhou, and Bryan Dai. LoopCoder-v2: Only loop once for efficient test-time computation scaling. *arXiv preprint arXiv:2606.18023*, 2026. URL https://arxiv.org/abs/2606.18023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472.

Boyi Zeng, Shixiang Song, Siyuan Huang, Yixuan Wang, He Li, Ziwei He, Xinbing Wang, Zhiyu Li, and Zhouhan Lin. PonderLM: Pretraining language models to ponder in continuous space. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UrM4MNRYZm.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, Lu Li, Jiajun Shi, Kaijing Ma, Shanda Li, Taylor Kergan, Andrew Smith, Xingwei Qu, Mude Hui, Bohong Wu, Qiyang Min, Hongzhi Huang, Xun Zhou, Wei Ye, Jiaheng

Liu, Jian Yang, Yunfeng Shi, Chenghua Lin, Enduo Zhao, Tianle Cai, Ge Zhang, Wenhao Huang, Yoshua Bengio, and Jason Eshraghian. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025. URL https://arxiv.org/abs/2510.25741.