The Price of Recurrence: Parameter Equivalence and Circuit Complexity in Token-Recurrent Models

Anonymous Authors

September 18, 2026

Abstract

Weight tying across depth has a measurable exchange rate: under scaling-law fits, an additional pass through a shared Transformer block contributes a fraction of the capacity of a fresh layer. We ask whether recurrence across tokens admits an analogous parameter-equivalence coefficient, ϕ_t, measuring the ordinary capacity contributed by each recurrent parameter. We find that it does, but only below a boundary set by circuit complexity. We study prefix products over three groups of order 60: the abelian $\mathbb{Z}_{60}$, the solvable D_{30}, and the non-solvable A_5, a standard NC^1-complete benchmark. On $\mathbb{Z}_{60}$ and shallow-depth D_{30}, recurrent and non-recurrent models obey a common capacity law with $\phi_t \approx 0.3$. On A_5, this finite exchange rate breaks down: non-recurrent Transformers require depth that grows with sequence length, while a single nonlinear recurrent state replaces an increasing number of feed-forward layers. At the measured crossover, recurrence is equivalent to 11 ± 2 additional layers, and the equivalence increases with sequence length. The effect persists from 0.5M to 340M parameters and when the group operation is embedded in natural-language text. The distinction is specifically tied to nonlinear recurrence. A diagonal state-space baseline fails on A_5 in all 72 runs, whereas nonlinear recurrence succeeds in 34/72; scratchpad decoding also succeeds, but uses roughly $60\times$ more inference compute. Optimization introduces a separate phenomenon: recurrent training is less reliable on D_{30} than on A_5 because models first discover a shallow solution on the rotation subgroup. A two-stage curriculum over the reflection coset raises D_{30} success from 22% to 97%. Across all 216 base runs, held-out loss is bimodal and a width-dependent phase-transition model predicts success rates to within four percentage points. Finally, recurrent A_5 solvers generalize to sequences $64\times$ longer than those seen in training in all 26/26 tested runs. These results suggest that token recurrence behaves as a conventional capacity multiplier on problems with shallow circuits, but becomes a capability switch once sequential computation is required.

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.