Latent Trajectories in Looped Transformers: The Journey Over the Destination

C.Caccavella [1,2,3], D.Bonnet[2], E.Neftci[3,4]

[1] Zurich University of Applied Sciences, [2] ETH Zürich, [3] Forschungszentrum Jülich, Germany, [4] RWTH Aachen

Loop-based Transformer architectures have recently emerged as a promising approach to enhance reasoning capabilities reducing parameter count. By introducing recurrent computation over a fixed set of parameters, these models iteratively refine internal representations, resembling dynamical systems or energy minimization processes [1]. This paradigm improves performance on structured reasoning tasks, including visual domains such as mazes, Sudoku, and abstract reasoning benchmarks (e.g., ARC) [3]. Furthermore, the number of recurrent iterations ("loops") can serve as an adaptive computation mechanism, allocating more processing to complex inputs while remaining efficient on simpler ones [1]. However, the true benefits and underlying mechanisms of these models remain insufficiently understood, often mixing multiple sources of complexity and failing to disentangle model capacity and memorization from true reasoning [1, 2].

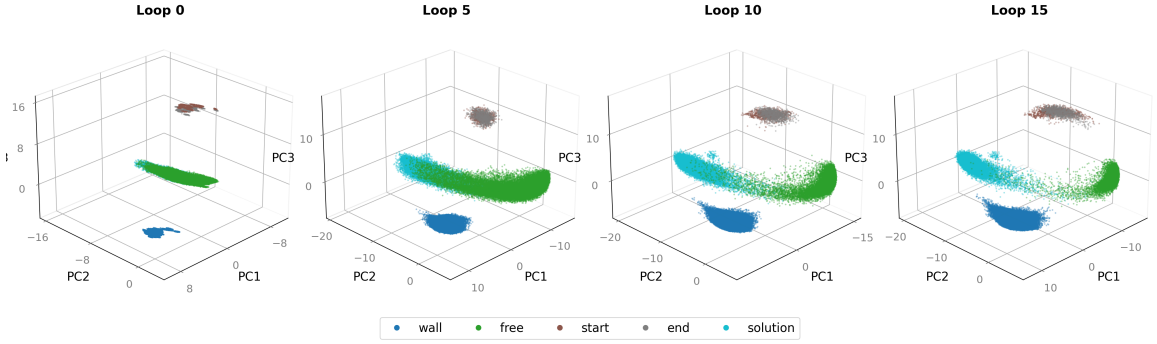

Figure 1: Evolution of hidden-state geometry across recurrent loop iterations.

In this work, we investigate loop-based Vision Transformers on maze tasks, focusing on the model's ability to refine predictions over recurrent latent steps. Early analysis shows that, over successive loops, latent representations progressively organize into well-separated clusters corresponding to meaningful maze components (walls, paths, start, end, and solution trajectory). To characterize these dynamics, we perform trajectory analysis using entropy-based measures and PCA probing. We observe that incorrectly classified solutions remain farther from the cluster centers associated with correct solutions and do not converge toward them as closely over iterations, providing a clear signal of solution quality. Despite these promising results supporting latent reasoning, we find that the model only partially relies on recurrent loops. In practice, performance is driven more by model capacity than by iterative refinement, resulting in poor generalization.We further investigate the mechanisms and advantages of loop-based models in this setting to better understand their limitations and how iterative computation supports reasoning. Building on this analysis, we explore approaches to improve generalization across variations in maze size and input structure, encouraging stronger reliance on latent recurrent refinement. Furthermore, we leverage token trajectory dynamics as a signal for early detection of incorrect solutions, enabling out-of-distribution detection and improved prediction accuracy at inference time. Preliminary experiments also suggest that incorporating full trajectory information can enhance model capability beyond final-step predictions.

[1] J. Geiping et al., arXiv:2502.05171, 2025.

[2] A. Jolicoeur-Martineau, arXiv:2510.04871, 2025.

[3] W. J. Shu et al., arXiv:2602.02156, 2026.