REINFORCE Memory Experiment Report (v1)

February 13, 2026

1 Objective

This experiment evaluates whether a network-based policy gradient method (REINFORCE) can learn maze navigation when the observation explicitly includes memory of visited locations.

2 Method

Policy model. Two-layer MLP policy trained with REINFORCE.

State encoding.

$$x_t = [\text{onehot}(s_t)\,;\,\text{visited\_table}_t],$$

where `visited_table` is flattened binary memory indicating visited cells in the current episode.

Environment/reward. Step reward $= -0.01$, terminal goal reward $= +1.0$, deterministic transition (no slip), max steps $= 350$.

Hyperparameters. Episodes $= 800$, runs $= 8$, hidden dim $= 64$, learning rate $= 0.002$, $\gamma = 0.99$, entropy coefficient $= 0.001$.

3 Results

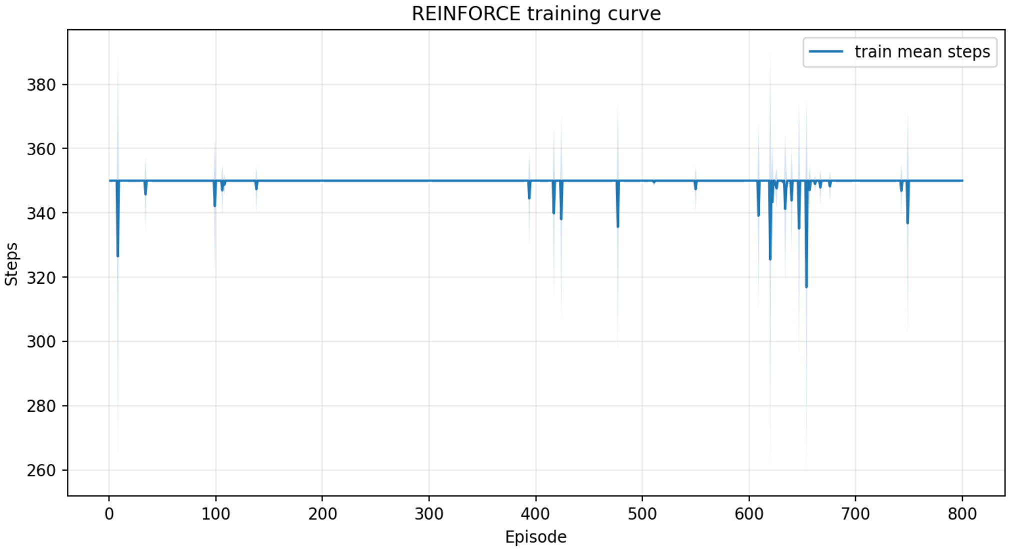

Figure 1: Training curve: mean steps to goal over episodes.

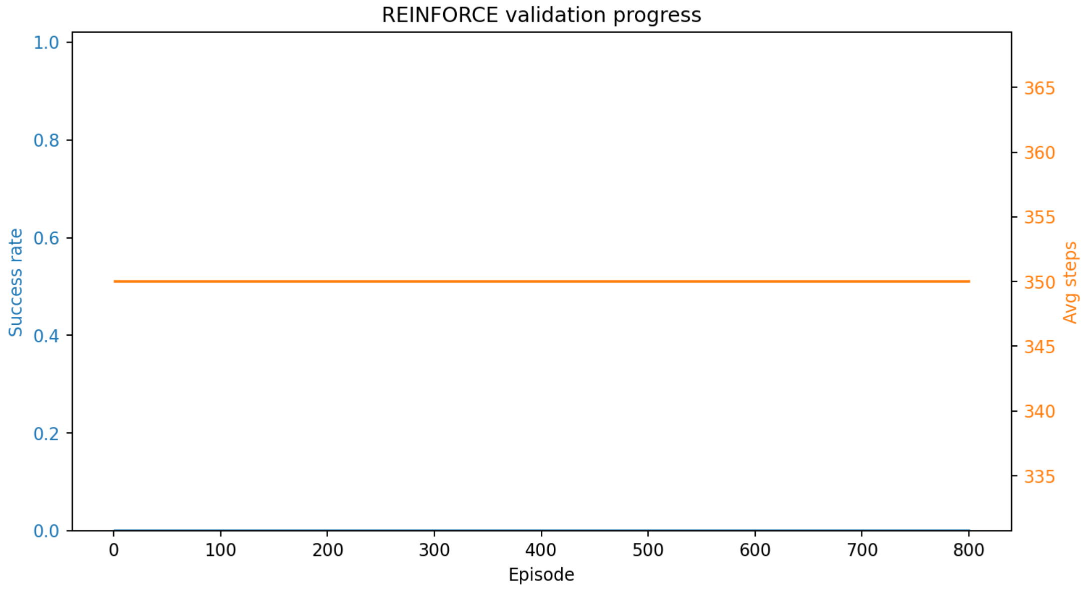

Figure 2: Validation success rate and validation average steps.

Metric	Value
Final train mean steps (ep 800)	350.0
Final validation success rate (ep 800)	0.0
Final validation mean steps (ep 800)	350.0
Best validation success rate (all checkpoints)	0.0

Table 1: Summary of REINFORCE memory experiment outcome.

4 Final Metrics

5 Interpretation

Under the tested configuration, the agent did not learn a successful policy. This run is recorded as a negative result and indicates that the current reward scale / optimization setting is insufficient despite memory-augmented state encoding.

6 Artifacts

- Output directory: `../outputs/reinforce_memory_v1`
- Curves: `learning_curve.csv/png`, `validation_progress.csv/png`
- Run config: `run_summary.json`
- Rollout GIFs: `gifs/policy_rollout_ep_*.gif`