Maze Shaping Report: Manhattan PBRS + Exploration Bonus (v1)

February 13, 2026

1 Experiment Setup

This run extends Manhattan PBRS by adding an exploration bonus for first-time state visits within each episode.

Reward used in training:

$$r_t = r_{\text{env}}(s_t, a_t, s_{t+1}) + \gamma \Phi_\kappa(s_{t+1}) - \Phi_\kappa(s_t) + r_{\text{explore}}(s_{t+1}),$$

where

$$\Phi(s) = -(|r - r_g| + |c - c_g|), \quad \Phi_\kappa(s) = \kappa \Phi(s), \ \kappa \in \{0, 0.5, 1.0\},$$

and

$$r_{\text{explore}}(s_{t+1}) = \begin{cases} 0.05, & \text{if } s_{t+1} \text{ is first-visit in current episode} \\ 0, & \text{otherwise.} \end{cases}$$

Hyperparameters: $\gamma = 0.99$, $\alpha = 0.02$, $\epsilon = 0.10$, step reward $= 0.0$, goal reward $= 1.0$, episodes $= 500$, runs $= 12$, max steps $= 350$.

2 Learning and Validation Curves

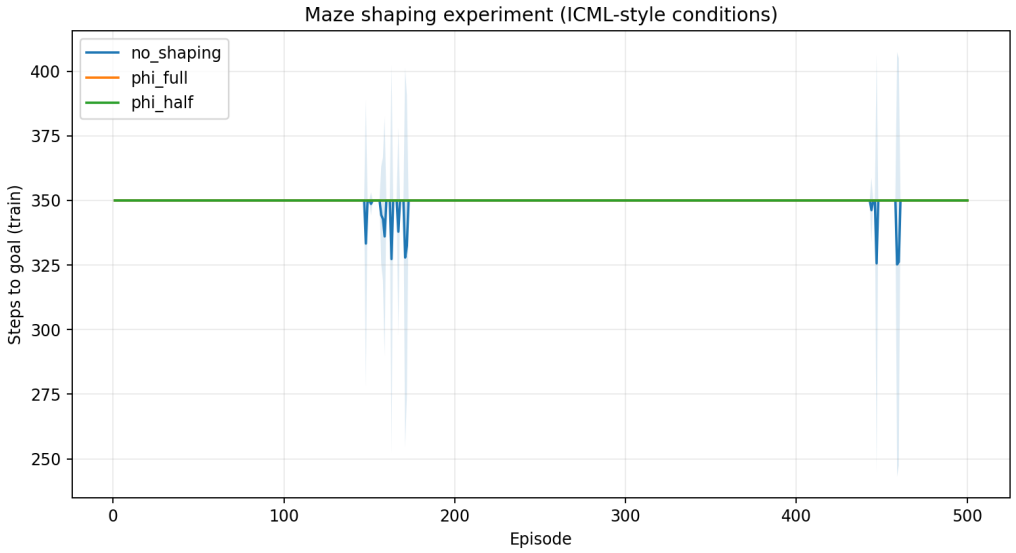

Figure 1: Training curve (mean steps to goal).

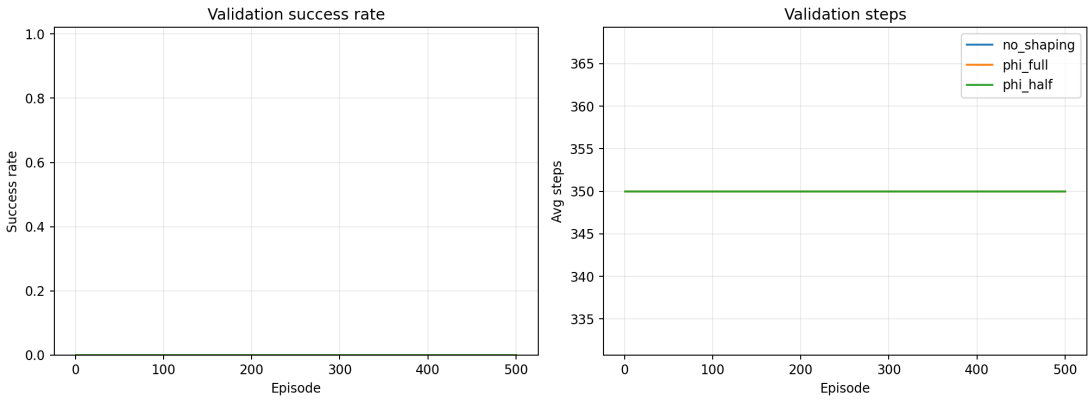

Figure 2: Validation success and validation average steps.

3 Final Summary (Episode 500)

Condition	Validation success rate	Validation mean steps
`no_shaping`	0.0	350.0
`phi_half`	0.0	350.0
`phi_full`	0.0	350.0

Table 1: All conditions remained unsuccessful at this budget.

4 Interpretation

Under this configuration, adding per-episode first-visit exploration bonus (0.05) did not recover successful greedy policies. This is recorded as a negative result for the current reward shaping design and parameter scale.

5 Artifacts

- Output folder: `../../outputs/maze_shaping_icml_style_pbrs_manhattan_explore_v1`
- CSVs: `learning_curve.csv`, `validation_progress.csv`
- Config dump: `run_summary.json`
- GIF rollouts: `gifs/policy_rollout_ep_*.gif`