Maze Reward Shaping Experiment Report (ICML 1999 Style Conditions)

February 13, 2026

1 Objective

This report summarizes a maze-navigation reinforcement learning experiment inspired by Ng, Harada, and Russell (ICML 1999), comparing three reward settings:

- `no_shaping`
- `phi_half` (potential scale 0.5)
- `phi_full` (potential scale 1.0)

Let the base reward be $R(s, a, s') = -1$ per step and let $\Phi_0(s)$ be the distance-based potential

$$\Phi_0(s) = -d(s, g),$$

where $d(s, g)$ is the shortest-path distance from state s to goal g on the maze graph (computed by BFS).

For a scale factor $\kappa \in \{0, 0.5, 1.0\}$, define

$$\Phi_\kappa(s) = \kappa\, \Phi_0(s),$$

and the shaped reward

$$R'_\kappa(s, a, s') = R(s, a, s') + \gamma \Phi_\kappa(s') - \Phi_\kappa(s).$$

In this experiment, $\gamma = 1.0$, so each condition becomes:

$$\texttt{no\_shaping}: \ \kappa = 0, \quad R'_0(s, a, s') = R(s, a, s') = -1,$$

$$\texttt{phi\_half}: \ \kappa = 0.5, \quad R'_{0.5}(s, a, s') = -1 + \Phi_{0.5}(s') - \Phi_{0.5}(s),$$

$$\texttt{phi\_full}: \ \kappa = 1.0, \quad R'_{1.0}(s, a, s') = -1 + \Phi_{1.0}(s') - \Phi_{1.0}(s).$$

2 Setup

- Environment: generated maze grid `../../outputs/maze_samples_v1/grids/maze_00.npy`
- Agent: tabular SARSA with ϵ-greedy behavior policy
- Transition stochasticity: slip probability 0.2
- Base reward: -1 per step
- Potential shaping: $R'(s, a, s') = R(s, a, s') + \gamma\Phi(s') - \Phi(s)$

- Episodes: 500, Runs per condition: 12
- Learning rate $\alpha = 0.02$, exploration $\epsilon = 0.10$, discount $\gamma = 1.0$
- Max steps per episode: 350
- Validation: every 25 episodes, 30 greedy rollout episodes

3 Used Maze Instance

The experiment used one fixed maze instance (`maze_00`) from the generated sample set.

Field	Value
Maze ID	`maze_00`
Seed	0
Cell size	10×10
Grid size	21×21
Start / Goal	$(1, 1)$ / $(19, 19)$
Shortest path length (BFS)	44
Wall count / ratio	242 / 0.5488
Dead-end count	13

Table 1: Metadata of the maze used in this run.

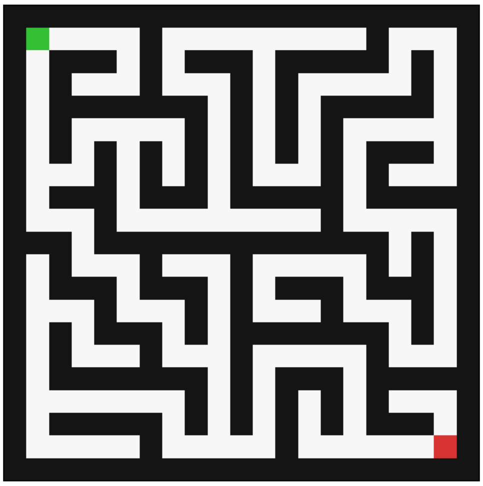

Figure 1: Maze instance `maze_00`: start and goal layout used in training and validation.

4 Quantitative Results

4.1 Training Curve

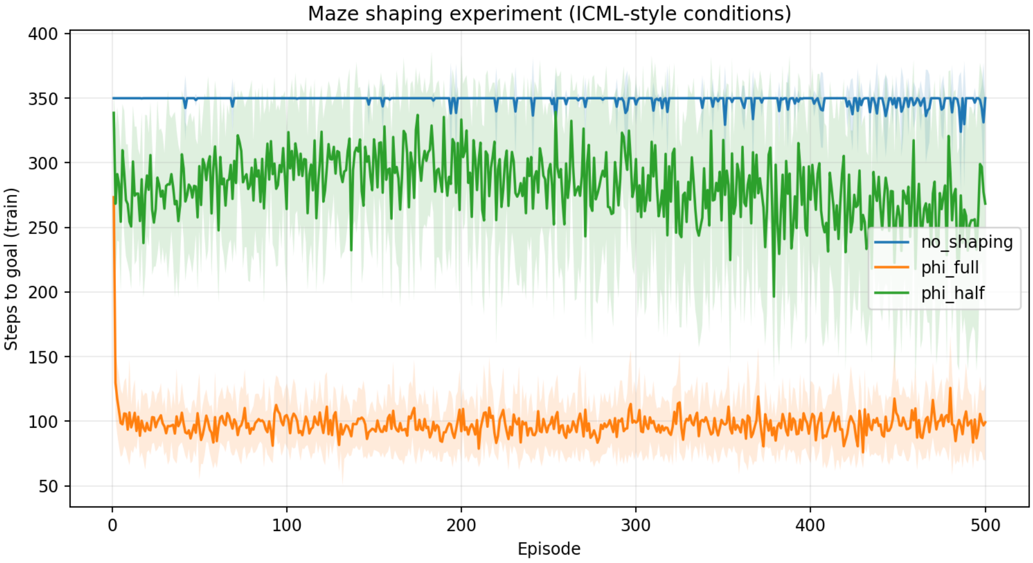

Figure 2: Training steps-to-goal over episodes (mean with std band).

4.2 Validation Progress

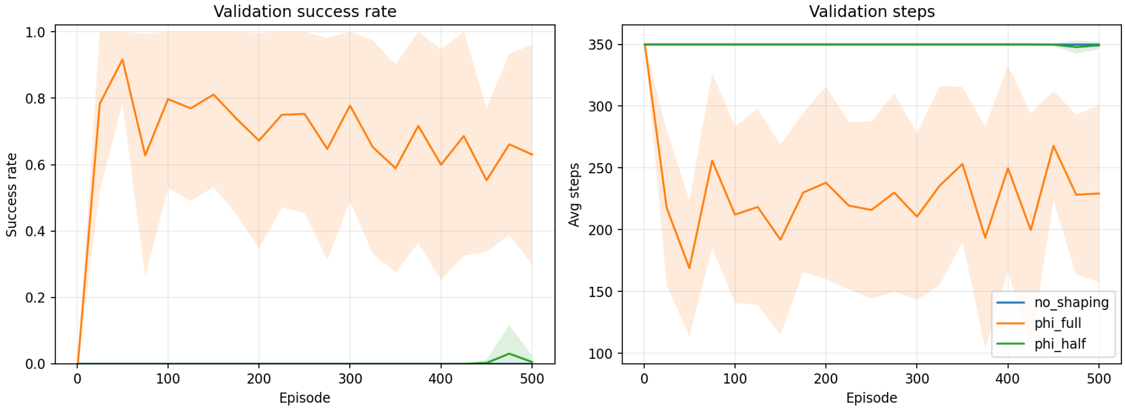

Figure 3: Validation success rate and validation average steps over training progress.

4.3 Final Episode Summary (Episode 500)

Condition	Train mean steps	Validation success rate	Validation mean steps
`no_shaping`	350.00	0.0000	350.00
`phi_half`	268.42	0.0056	349.12
`phi_full`	99.25	0.6306	229.26

Table 2: Performance comparison at the final episode.

5 Qualitative Policy Snapshots (GIF)

Five rollout GIFs were exported at policy checkpoints:

- `../outputs/maze_shaping_icml_style_v1/gifs/policy_rollout_ep_0000.gif`
- `../outputs/maze_shaping_icml_style_v1/gifs/policy_rollout_ep_0125.gif`
- `../outputs/maze_shaping_icml_style_v1/gifs/policy_rollout_ep_0250.gif`
- `../outputs/maze_shaping_icml_style_v1/gifs/policy_rollout_ep_0375.gif`
- `../outputs/maze_shaping_icml_style_v1/gifs/policy_rollout_ep_0500.gif`

6 Discussion

- `phi_full` consistently outperformed the other settings in this maze instance.
- `no_shaping` did not learn a successful strategy under the current budget and stochasticity.
- `phi_half` improved training steps somewhat but did not yield strong greedy validation success in this run.

7 Reproducibility Artifacts

Primary output files:

- `../outputs/maze_shaping_icml_style_v1/learning_curve.csv`
- `../outputs/maze_shaping_icml_style_v1/validation_progress.csv`
- `../outputs/maze_shaping_icml_style_v1/run_summary.json`