Maze Reward Shaping Report (Manhattan Potential) Argumentative Rewrite

February 13, 2026

1 Research Objective

Why this section matters. A clear research question is necessary to interpret the same curves as either "theory-consistent" or "implementation-specific" behavior.

This report asks: *In a stochastic maze domain, how much practical learning benefit do potential-based shaping rewards provide under finite training, and how does that benefit change when the potential is based on Manhattan distance?*

We compare three conditions exactly as in the ICML-style setup:

- `no_shaping`
- `phi_half` ($\kappa = 0.5$)
- `phi_full` ($\kappa = 1.0$)

2 Theoretical Motivation (Ng et al., 1999)

Why this section matters. The experiment is only meaningful if we separate *policy invariance in theory* from *learning speed in practice.*

Ng et al. (1999) show that shaping of the form

$$F(s, a, s') = \gamma\Phi(s') - \Phi(s)$$

is policy-invariant (under their assumptions), meaning the set of optimal policies is unchanged after reward transformation. This motivates our design: we intentionally use potential-based shaping so that any difference across conditions should primarily reflect optimization dynamics (sample efficiency, exploration guidance), not a different task objective.

In this experiment, base reward is $R(s, a, s') = -1$ per step and shaped reward is

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

with

$$\Phi_\kappa(s) = \kappa\Phi_0(s), \quad \kappa \in \{0, 0.5, 1.0\}, \quad \gamma = 1.0.$$

For this Manhattan version,

$$\Phi_0(s) = -\left(|r - r_g| + |c - c_g|\right),$$

where (r_g, c_g) is the goal coordinate.

3 Experimental Design

Why this section matters. Design choices determine whether observed improvements answer the research question or are artifacts.

Design choices were tied directly to the question above:

- **Same environment, same algorithm, same hyperparameters across conditions:** isolates the effect of shaping magnitude κ.
- **Three shaping scales** $(0, 0.5, 1.0)$**:** tests whether stronger potential gradients produce stronger optimization bias.
- **Stochastic transitions (slip probability** 0.2**):** stresses robustness, making exploration quality measurable.
- **Repeated runs (12 seeds):** reduces single-seed variance and supports condition-level comparison.
- **Validation every 25 episodes (30 greedy rollouts):** tracks whether training improvements transfer to greedy policy quality.

Fixed settings: SARSA (tabular), $\alpha = 0.02$, $\epsilon = 0.10$, $\gamma = 1.0$, 500 episodes, max 350 steps per episode.

4 Used Maze Instance

Why this section matters. Maze geometry controls whether Manhattan potential is aligned or misaligned with true progress, directly affecting the research question.

The experiment used one fixed maze instance, `maze_00`:

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 Manhattan-potential run.

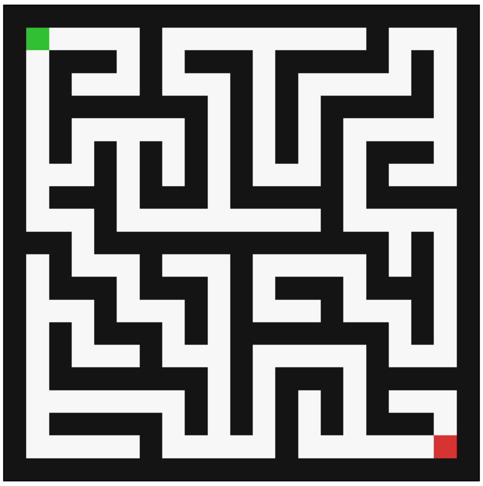

Figure 1: Maze instance `maze_00`.

5 Results

Why this section matters. This section quantifies whether potential scaling changes learning outcomes under finite data.

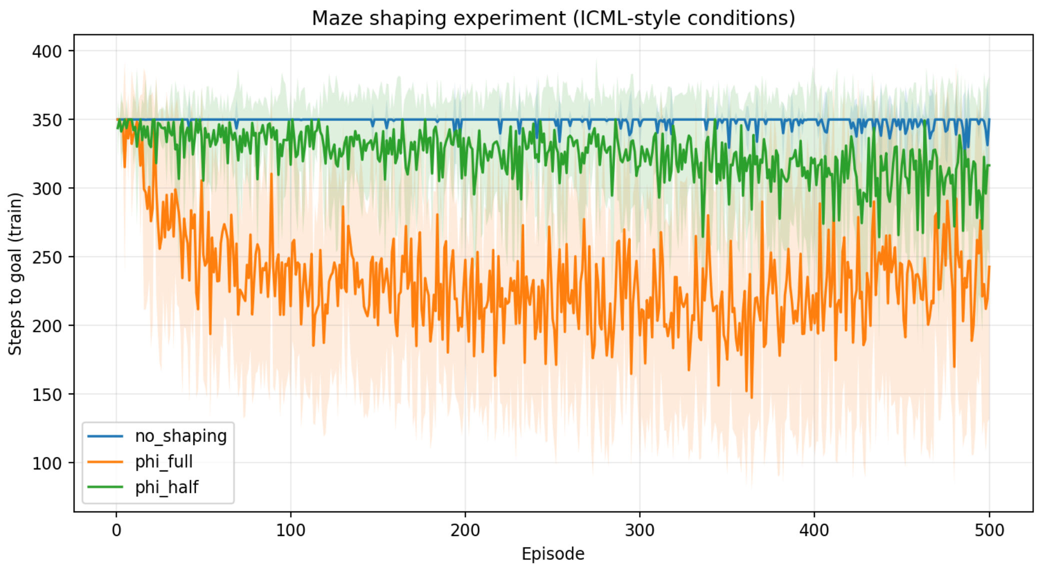

Figure 2: Training steps-to-goal (mean with std band).

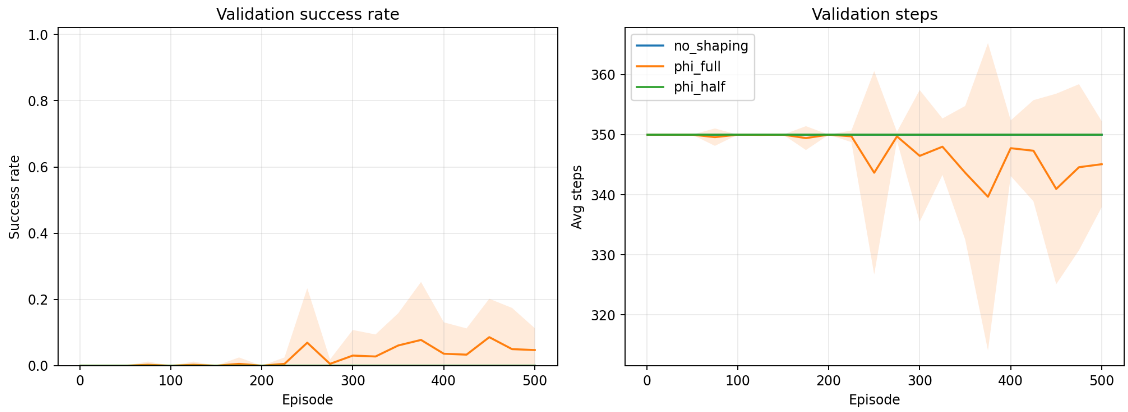

Figure 3: Validation success and validation average steps over training.

Condition	Train mean steps (ep 500)	Validation success (ep 500)	Validation mean steps (ep 500)
`no_shaping`	350.00	0.0000	350.00
`phi_half`	316.58	0.0000	350.00
`phi_full`	242.67	0.0472	345.08

Table 2: Final-episode summary for Manhattan potential.

Additional observation: peak validation success for `phi_full` was 0.0861, while `no_shaping` and `phi_half` stayed at 0.

6 Discussion: Mechanism-Level Interpretation

Why this section matters. Summary statistics alone do not explain *why* Manhattan shaping underperformed BFS shaping in the same maze.

Mechanism 1: gradient alignment with true geodesic progress. Manhattan distance ignores walls. In corridors and detours, moves that reduce Manhattan distance can be locally attractive but globally unhelpful. Therefore the shaping term can provide weak or misleading short-horizon guidance compared to BFS-based potential, which is aligned with maze topology.

Mechanism 2: magnitude vs direction trade-off. Comparing `phi_half` and `phi_full`, stronger shaping ($\kappa = 1.0$) produced better training metrics, suggesting that signal strength helps. However, the low validation success indicates that stronger but misaligned guidance does not reliably produce robust goal-reaching greedy policies.

Mechanism 3: finite-sample regime vs asymptotic invariance. Potential-based shaping is theoretically policy-invariant, but our experiment is finite-horizon training with stochastic transitions and nonzero exploration. In this regime, invariance does not guarantee equal sample efficiency. The observed gap is therefore compatible with Ng et al.: objective-equivalence can hold while optimization behavior differs substantially.

Mechanism 4: validation gap as a diagnostic. The large train/validation mismatch for Manhattan shaping suggests policies that partially exploit shaped reward structure without consistently solving the maze under greedy execution.

7 Qualitative Policy Snapshots (GIF)

Why this section matters. Rollout movies help verify whether numerical trends correspond to qualitatively better navigation.

Five checkpoint GIFs:

- `../../outputs/maze_shaping_icml_style_manhattan_v1/gifs/policy_rollout_ep_0000.gif`
- `../../outputs/maze_shaping_icml_style_manhattan_v1/gifs/policy_rollout_ep_0125.gif`
- `../../outputs/maze_shaping_icml_style_manhattan_v1/gifs/policy_rollout_ep_0250.gif`
- `../../outputs/maze_shaping_icml_style_manhattan_v1/gifs/policy_rollout_ep_0375.gif`
- `../../outputs/maze_shaping_icml_style_manhattan_v1/gifs/policy_rollout_ep_0500.gif`

8 Reproducibility Artifacts

Why this section matters. Explicit artifact paths enable exact reruns and auditing.

- Script: `../../experiments/maze_shaping_icml_style/run_maze_shaping_experiment.py`
- Learning CSV: `../../outputs/maze_shaping_icml_style_manhattan_v1/learning_curve.csv`
- Validation CSV: `../../outputs/maze_shaping_icml_style_manhattan_v1/validation_progress.csv`
- Run config summary: `../../outputs/maze_shaping_icml_style_manhattan_v1/run_summary.json`

9 Conclusion

Why this section matters. The conclusion maps evidence back to the original research question.

In this maze and training budget, Manhattan-based potential shaping improved training efficiency over no shaping, but the gain was limited and did not translate into strong validation success. This supports the view that potential-based shaping can preserve objective structure while still being highly sensitive to the *choice of potential* for practical learning speed.