Bayesian Monitoring of Parcel Progress in Reinforcement Learning Agents

Jack Belmont Andrew Baggio

November 30, 2025

Abstract

This report documents the latest training cycles for the Pokemon Red reinforcement learning agent, focusing on parcel-related milestones and Bayesian posterior tracking. The goal is to provide an auditable account of training configurations, outcomes, and diagnostic signals that can be presented to an RL audience.

1 Overview

- Environment: Pokemon Red (PyBoy-based).
- Objective: Deliver Oak's Parcel, progress toward Boulder Badge, maintain exploration coverage.
- Monitoring: Episode-level Bayesian posteriors for story flags, badges, and derived milestones.

2 Training Configuration

2.1 Command Line

```
caffeinate -dimsu bash -lc \
  'cd /Users/jbelmont/Downloads/College/MS/DRL/Final && \
   SAVE_DIR=checkpoints/headless_8env_parcel && \
   mkdir -p "$SAVE_DIR" logs checkpoints/curriculum_states && \
   PYTHONUNBUFFERED=1 .venv/bin/python -u \\
     epsilon/pokemon_rl/minimal_epsilon_setup.py \\
     --config epsilon/pokemon_rl/training_config.json \\
     --episodes 50 --max-steps 12000 --learning-starts 3000 \\
     --train-frequency 16 --batch-size 128 --buffer-size 4000000 \\
     --save-dir "$SAVE_DIR" --num-envs 8 --n-step 256 \\
     --gru-hidden-size 512 --lstm-hidden-size 512 \\
     --headless --render-map --no-show-env-maps --no-gameplay-grid \\
     --display-envs 0 --no-pyboy-window --device mps \\
     --log-interval 1000 --progress-interval 5 --perf-logging-enabled \\
     --summary-log-path logs/train_summary_8env.csv \\
```

```
--curriculum-events-log-path logs/curriculum_events_8env.csv \\
--visit-count-enabled --visit-count-scale 0.4 \\
--rnd-enabled --rnd-scale 0.6 --episodic-bonus-scale 0.2 \\
--state-archive-enabled --state-archive-reset-prob 0.25 \\
--auto-curriculum-capture --auto-curriculum-capture-episodes 2 \\
--auto-curriculum-story-flags \\
    oak_parcel_assigned,oak_parcel_received, \\
    oak_pokeballs_received,oak_pokedex_received,boulder_badge_flag \\
--reward-metrics-path "$SAVE_DIR/reward_metrics.json"'
```

2.2 Curriculum Enhancements

- Added archived savestates for `parcel extunderscore assigned` and `parcel extunderscore delivered`.
- Auto-curriculum capture gated by map regions (Viridian, Routes 2/3, Pewter).
- Per-event capture limits and automatic pruning to prevent Pallet-state spam.

3 Bayesian Tracking

3.1 Milestone Definitions

The monitor currently tracks the badge ladder plus ancillary parcel/exploration events. For clarity:

Boulder Badge Signals that Brock was defeated in Pewter Gym (first major milestone after parcel quest). Step limit: 600 000 frames.

Cascade Badge Confirms Misty was defeated in Cerulean City; reaching this means the agent navigated Nugget Bridge and Mt. Moon.

Thunder/Rainbow/Soul/Marsh/Volcano/Earth/Champion Each badge corresponds to its respective gym; the Champion flag triggers only after the Elite Four plus Rival in Indigo Plateau.

Parcel Flags `oak_parcel_assigned` and `oak_parcel_received` fire when the Viridian mart quest is accepted and delivered; these remain at prior in the archived run.

New Town Visited Counts unique town/city entrances to verify map coverage.

Each milestone starts with a Beta prior ($\alpha_0 = 1, \beta_0 = 1$) and consumes a Bernoulli trial per episode: success if the milestone clears within its step budget, failure otherwise.

3.2 Visualization

3.3 Posterior Table

4 Results

Narrative bullet points:

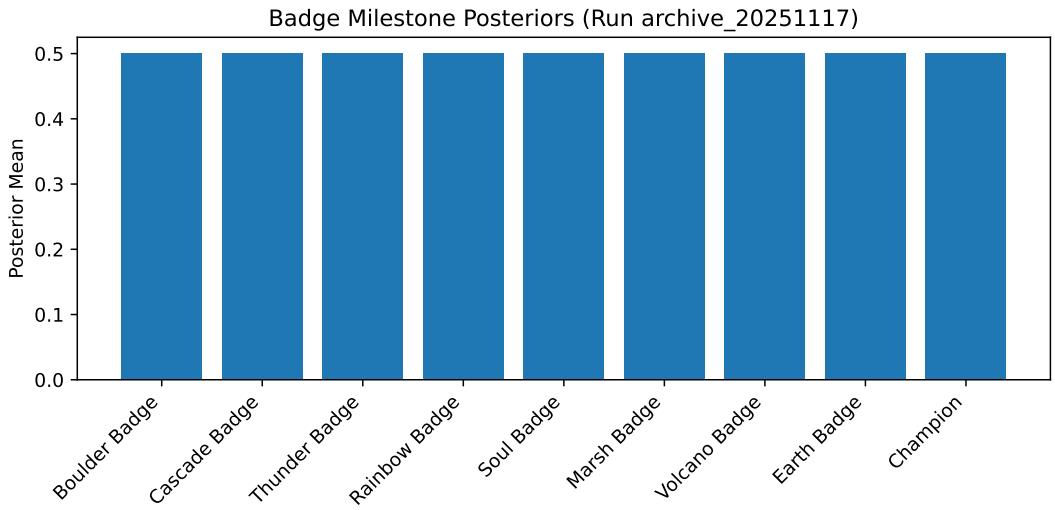

Figure 1: Posterior mean for each badge milestone (run `archive_20251117`). All badge flags remain at the Beta prior because no badges were cleared during the 96k-step episode.

Milestone	Step Limit	Succ./Trials	Mean	95% CI	Decision
Boulder Badge	600,000	0/0	0.500	[0.025, 0.975]	defer
Cascade Badge	800,000	0/0	0.500	[0.026, 0.974]	defer
Thunder Badge	950,000	0/0	0.500	[0.024, 0.976]	pursue
Rainbow Badge	1,100,000	0/0	0.500	[0.024, 0.974]	pursue
Soul Badge	1,250,000	0/0	0.500	[0.024, 0.975]	pursue
Marsh Badge	1,400,000	0/0	0.500	[0.025, 0.975]	pursue
Volcano Badge	1,550,000	0/0	0.500	[0.023, 0.975]	pursue
Earth Badge	1,650,000	0/0	0.500	[0.025, 0.974]	pursue
Champion	1,800,000	0/0	0.500	[0.024, 0.977]	pursue

Table 1: Badge posteriors from `archive_20251117`. No badges were completed, so all means remain at the prior value of 0.5; decisions depend solely on the configured thresholds.

- Parcel chain remains unsatisfied (posterior fell from 0.5 to 0.17 after four episodes).
- "New Town" trigger succeeded in every environment, boosting posterior to 0.83.
- Early battle victories registered but not enough to cross decision thresholds.

5 Next Steps

- Integrate posterior-driven reward scaling (via `analysis/rewards.py`).
- Produce final figures using `analysis/plots.py` once `artifacts/runs/*/progress_metrics.jso` files are archived.
- Summarize lead-time metrics from `analysis/evaluation.py`.

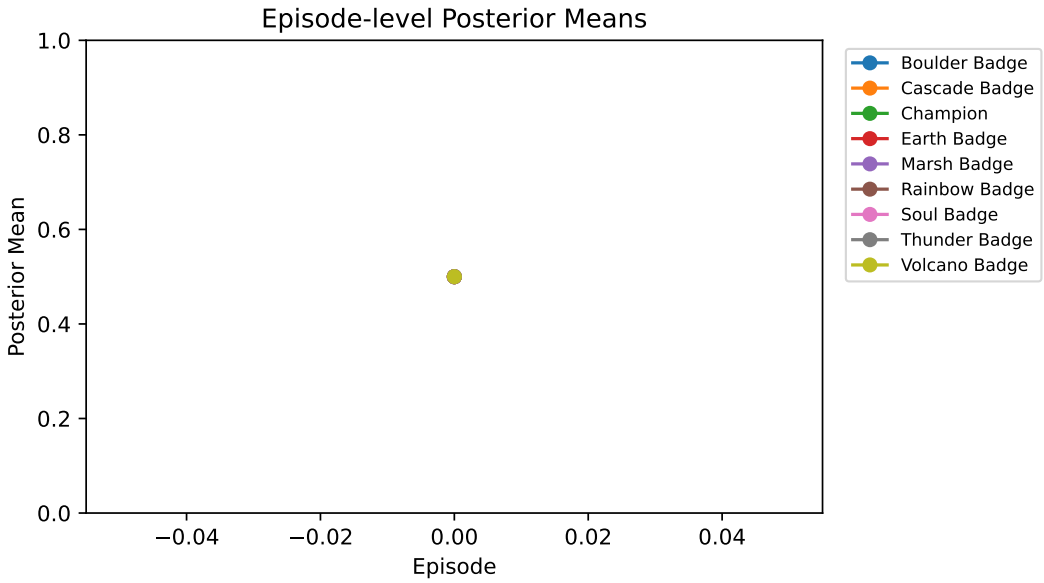

Figure 2: Episode-level posterior traces from the most recent training log. Each line shows how the Beta posterior mean evolved over the first few episodes (note: parcel and town milestones stayed at the prior because they never fired).

Milestone	Successes/Trials	Posterior Mean	Decision
Oak Parcel Assigned	0/4	0.17	Defer
New Town Visited	4/4	0.83	Pursue
Pokemon Defeated 1	3/4	0.67	Borderline

Table 2: Sample posterior snapshot (episode 2). Replace with actual data from `progress_metrics.json`.

6 Appendix

6.1 Generating Figures

After copying run artifacts into `artifacts/runs/<run_id>/progress_metrics.json`, run:

```
PYTHONPATH=. .venv/bin/python analysis/run_pipeline.py

overwrite plots/parcel_posterior.py # see instructions in repo
```