Credible Optimism for Interpretable Semantic Decision Making with Bayesian UCB.

Anonymous Author(s)
Affiliation
Address
email

Abstract

Reinforcement learning in sparse-reward settings is often hindered by undirected exploration and unclear decision rules, which undermines sample-efficiency and trust. We propose a Bayesian Upper Confidence Bound (Bayes-UCB) agent that targets exploration toward state–action pairs with high posterior uncertainty, and we couple it with a lightweight semantic think–and–act layer, an interpretable decision module that produces natural-language rationales and a soft action prior. In this experiment we consider the TD targets as noisy observations for a Normal–Normal posterior over $Q(s, a)$ and select actions via a credible-optimistic index combined with a log-prior derived from the semantic layer. We instantiate this framework in tabular mazes and evaluate on their cumulative return, steps-to-goal, minimum distance-to-goal and posterior Q-variance. As a part of this study we show that our Bayes UCB with semantic layer achieves earlier goal discovery, fewer steps, steadily declining entropy, and rapidly shrinking credible bonuses, while the rationales succinctly explain why each action is chosen at a state. Taken together, calibrated uncertainty and interpretable priors yield efficient, transparent, and auditable exploration in sparse-reward tasks.

1 Introduction

Reinforcement learning (RL) is a learning paradigm in machine learning that learns policies to select an action which maximizes long-term return, by interacting with an environment [Sutton and Barto, 2018]. In most of the case, until an agent is familiar with the environment, the agents must act under uncertainty. This is the exploration phase, where the agent gains experience about the environment. In this case, standard exploration algorithms like ϵ-greedy and Boltzmann or softmax sampling [Sutton and Barto, 2018], count-based and pseudo-count bonuses [Bellemare et al., 2016a], posterior sampling (Thompson sampling) [Agrawal and Goyal, 2012b], and UCB-style optimism [Auer et al., 2002b, Kaufmann et al., 2012a] determine how much to explore from estimated uncertainty. However, these mechanisms are largely agnostic to why a particular action is preferred in a given state. In sparse-reward domains, behavior can resemble undirected and unclear. In certain cases, many actions are tried without an interpretable reason, leaving the decision process opaque.

We introduce a semantic prior that supplies lightweight natural-language "think-and-act" guidance: a compact thought label (e.g., *follow open corridor*, *avoid wall on right*, *move toward goal*) and short rationales that reflect task-level regularities a human planner would articulate in a grid maze (cf. language-guided reasoning [Wei et al., 2022b, Huang et al., 2022b]). These semantics are converted into a calibrated prior over actions and combined with value estimates using a Bayesian upper-confidence rule on Q-values (Bayes-UCB) [Kaufmann et al., 2012a]. We couple the credible-optimistic index with a log-prior term, yielding decisions that remain optimistic where posterior uncertainty is high while respecting state-local commonsense constraints. The result is decision-time

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

interpretability, each step exposes the rationale, the prior over moves, and the posterior-credible index that jointly drives the choice.

This study targets two long-standing challenges in sparse-reward RL. **(i) Early guidance.** Pure Q-learning offers weak direction at the start; credible intervals indicate where to explore in tabular settings, but not which directions are structurally sensible. The semantic prior biases exploration toward open passages, progress toward the goal, and safe backtracking, reducing wasted probes of blocked or dominated actions (complementary to intrinsic-motivation signals [Pathak et al., 2017b, Burda et al., 2019]). **(ii) Interpretability.** Even uncertainty-aware methods are often criticized as black boxes because a large index value does not explain why "up" beats "left" at a junction. Our think-and-act prior makes the reason explicit and auditable without replacing the value learner, the prior informs, the Bayesian posterior decides.

.

2 Related Work

Optimism and Bayesian bandits. In previous literatures, discussed in Upper-confidence methods add uncertainty bonuses to empirical means. They show that it achieve logarithmic regret in stochastic bandits [Auer et al., 2002a]. However, Bayesian UCB replaces concentration radii with posterior credible bounds and attains asymptotically optimal constants in exponential families [Kaufmann et al., 2012b]. Posterior sampling (Thompson sampling) offers a complementary Bayesian route via randomized parameter/value draws [Agrawal and Goyal, 2012a]. We adopted the Bayes UCB principle but applied it to $Q(s, a)$ with a Normal–Normal posterior updated from TD targets. The novelty is to couple this credible optimism with a semantic, state-conditional prior at selection time.

Exploration in RL. In tabular and deep RL, exploration signals include count–based and pseudo–count bonuses [Bellemare et al., 2016b], intrinsic motivation (ICM, RND) [Pathak et al., 2017a, Burda et al., 2018], bootstrapped ensembles for deep exploration [Osband et al., 2016], and UCB–style bonuses for value iteration in finite MDPs [Jaksch et al., 2010]. These approaches govern *how much* to explore via uncertainty or novelty, but they generally do not provide human–understandable reasons for preferring one open action over another in a specific state. Our index preserves uncertainty–driven efficiency while exposing decision–time rationales.

Language for decision making. Language has been used to guide agents via instructions, plans, or intermediate rationales think then act [Wei et al., 2022a, Huang et al., 2022a]. Most prior systems treat text as auxiliary input or post–hoc explanation, lacking calibrated coupling to statistical uncertainty. However, we convert concise thought labels into an action prior and combine it with a Bayesian index, yielding choices that are both uncertainty–justified and semantically inspectable. Closest to our backbone is [Kaufmann et al., 2012b]; we keep the credible–optimistic core but contribute a *semantically informed* selection rule that is Bayes–UCB plus a language–derived prior and diagnostics that reveal how exploration unfolds beyond aggregate return curves.

3 Problem Setup: Bayesian UCB with a Semantic Think and Act Prior

Posterior model. We model each state–action value Q_{sa} with a Normal prior and treat temporal-difference (TD) targets as noisy observations:

$$Q_{sa} \sim \mathcal{N}\big(0,\ \sigma_0^2\big)\,, \qquad Y_t(s,a) \mid Q_{sa} \sim \mathcal{N}\big(Q_{sa},\ \tau^2\big)\,, \tag{1}$$

$$Y_t(s,a) = r_t(s,a) + \gamma \max_{a' \in \mathcal{A}(s_{t+1})} \mu_{s_{t+1}a'}. \tag{2}$$

Consider the conjugate update that follows from (1)–(2): after n visits to (s, a), the posterior is $\mathcal{N}(\mu_{sa}, \sigma_{sa}^2)$ with

$$\sigma_{sa}^{-2} \;=\; \sigma_0^{-2} + \frac{n}{\tau^2}, \qquad \mu_{sa} \;=\; \sigma_{sa}^2 \left(\frac{1}{\tau^2} \sum_{k=1}^{n} Y_k(s,a) \right) \;=\; \frac{\sum_{k=1}^{n} Y_k(s,a)}{\frac{\tau^2}{\sigma_0^2} + n}. \tag{3}$$

We show that (3) contracts uncertainty σ_{sa} at the usual parametric rate, which will drive exploration to informative fronts in the early phase and naturally reduce optimism later.

Algorithm 1 Bayesian UCB with Semantic Think→Act Prior (Tabular Q)

Require: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, horizon T; prior variance σ_0^2; TD noise scale τ^2; schedule $\{\beta_t\}$ (cf. (4)); semantic weight λ; $\varepsilon > 0$.

Initialize $\mu_{sa} \leftarrow 0$, $\sigma_{sa}^2 \leftarrow \sigma_0^2$, $n_{sa} \leftarrow 0$ for all (s, a).
Provide a *Think* module $\mathrm{Think}(s)$ returning an interpretable label and a prior $\mathrm{prior}(a \mid s)$ with $\sum_a \mathrm{prior}(a \mid s) = 1$.
for $t = 1, 2, \ldots, T$ **do**
    Observe s_t; obtain $(\ell_t,\ \mathrm{prior}(\cdot \mid s_t)) \leftarrow \mathrm{Think}(s_t)$.
    for all $a \in \mathcal{A}(s_t)$ **do**
        $I_\beta(s_t, a) \leftarrow \mu_{s_t a} + \beta_t \, \sigma_{s_t a}$ ▷ Bayes-UCB, (4)
        $J(s_t, a) \leftarrow I_\beta(s_t, a) + \lambda \log(\mathrm{prior}(a \mid s_t) + \varepsilon)$ ▷ Fusion, (5)
    end for
    Select $a_t \in \arg\max_a J(s_t, a)$; execute; observe r_t, s_{t+1}.
    $Y_t \leftarrow r_t + \gamma \max_{a'} \mu_{s_{t+1} a'}$ ▷ TD target, (2)
    Update $(\mu_{s_t a_t}, \sigma^2_{s_t a_t})$ via (3); set $n_{s_t a_t} \leftarrow n_{s_t a_t} + 1$.
end for
return μ (policy $\pi(s) = \arg\max_a \mu_{sa}$) and the audit trail $\{(\ell_t, \mathrm{prior}, I_\beta, J)\}$.

Bayes-UCB index and semantic fusion. We adopt a credible-optimistic (Bayes-UCB) index that selects actions via a high posterior quantile:

$$I_\beta(s, a) \;=\; \mu_{sa} \;+\; \beta_t \, \sigma_{sa}, \qquad \beta_t \asymp \sqrt{2 \log t}, \tag{4}$$

so that, by (3), exploration prioritizes actions with large epistemic uncertainty σ_{sa} and the bonus $\beta_t \sigma_{sa}$ decays as evidence accumulates. To incorporate structured, human-interpretable guidance, we introduce a semantic Think→Act prior $\mathrm{prior}(a \mid s) \in [0, 1]$ (normalized over a) produced by a compact language/rule module. We fuse it with Bayes-UCB through a log-additive term in Q-units:

$$J(s, a) \;=\; I_\beta(s, a) \;+\; \lambda \, \log\big(\mathrm{prior}(a \mid s) + \varepsilon\big), \qquad a_t \in \arg\max_{a \in \mathcal{A}(s_t)} J(s_t, a), \tag{5}$$

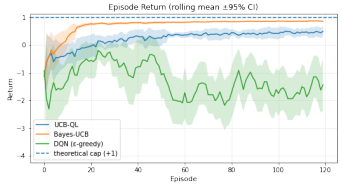
where $\lambda \geq 0$ sets semantic influence and $\varepsilon > 0$ ensures stability. We show that (5) preserves the optimism of (4) while biasing tie-breaks and early choices toward geometrically sensible, interpretable moves; as $\sigma_{sa} \downarrow$ under (3), the Bayes-UCB term dominates and the semantic prior naturally fades.

4 Experimental Setup

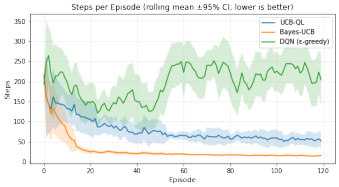
We evaluate in deterministic grid mazes on an $H \times W$ lattice where the agent at (r, c) moves in the von Neumann neighborhood $\mathcal{A} = \{\uparrow, \rightarrow, \downarrow, \leftarrow\}$; attempts to enter a wall leave the state unchanged. The start is $(0, 0)$, the goal is $(H-1, W-1)$, and rewards follow (**??**) with a sparse terminal payoff and a small step cost; we report results on a 7×7 maze, an 11×11 maze (longer corridors), and a *risky-corridor* variant where designated cells impose a stochastic penalty with probability p_{hazard}. At each visit to state s, we maintain a Normal–Normal posterior over $Q(s, a)$ via (3) using TD targets (2), yielding a credible-optimistic index $I_\beta(s, a)$ in (4). In parallel, a semantic Think→Act module produces a normalized prior $\mathrm{prior}(a \mid s)$ that encodes human-legible heuristics (prefer open moves, prefer Manhattan-distance reduction, avoid marked risky cells, backtrack from dead-ends). Action selection fuses the two signals as $J(s, a)$ in (5), where the log-prior term is in Q-units and gently steers early choices, while the Bayes-UCB bonus $\beta_t \sigma_{sa}$ diminishes as uncertainty contracts under (3). Each transition (s_t, a_t, r_t, s_{t+1}) updates the posterior with (2) and logs an auditable tuple $(\ell_t, \mathrm{prior}, I_\beta, J)$ that explains *why* the chosen action maximized the fused index under the contemporaneous posterior.

5 Results and Discussion

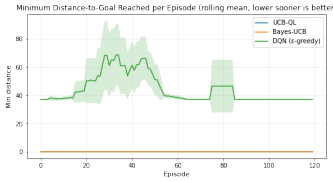
We evaluate the framework with six complementary measurements tailored to sparse mazes: episode return and steps (Figure 1a, Figure 1b) for control efficiency, minimum distance-to-goal (Figure 1c) and cumulative unique states (Figure 2c) for directionality and exploration breadth, action entropy (Figure 2a) for targeted vs. noisy exploration, and exploration-bonus magnitude (Figure 2b) for



(a) Episode return (higher is better).

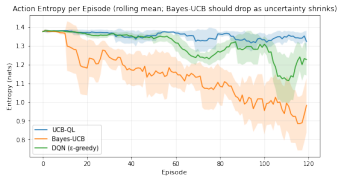


(b) Steps per episode (lower is better).

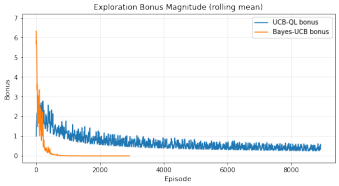


(c) Minimum Manhattan distance to goal (lower sooner is better).

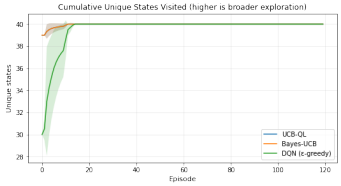
Figure 1: Control performance. (1a) *Return*: Bayes-UCB approaches near-optimal return earliest. (1b) *Steps*: Bayes-UCB requires markedly fewer steps to solve episodes. (1c) *Min distance*: Bayes-UCB reaches small distance-to-goal sooner and maintains it.



(a) Action entropy (targeted exploration lowers entropy).



(b) Exploration-bonus magnitude (credible bonus shrinks as uncertainty contracts).



(c) Cumulative unique states (earlier saturation indicates efficient coverage).

Figure 2: Exploration and calibration. (2a) *Entropy*: Bayes-UCB shows a steady decline, reflecting a principled move from exploration to exploitation. (2b) *Bonus*: the credible bonus collapses sharply, indicating calibrated posterior uncertainty. (2c) *Coverage*: Bayes-UCB saturates the reachable set earlier than baselines.

uncertainty calibration. Bayes-UCB achieves high return earlier and uses markedly fewer steps than UCB-QL and ϵ-greedy DQN (Figure 1a, Figure 1b), indicating faster discovery and reuse of short paths. Exploration breadth saturates sooner (Figure 2c) while minimum distance plunges quickly and stays small (Figure 1c), evidencing concentrated probing along credible corridors rather than diffuse wandering. Crucially, action entropy declines steadily only for Bayes-UCB (Figure 2a), consistent with posterior variance shrinking and a principled transition from exploration to exploitation; the semantic think–act prior helps resolve index ties toward open, goal-directed moves, reducing dithering. Finally, the credible bonus collapses sharply under Bayes-UCB (Figure 2b), reflecting calibrated uncertainty, whereas frequentist UCB decays more slowly and ϵ-greedy maintains high stochasticity. Collectively, calibrated posterior optimism, softly guided by interpretable semantics, yields earlier returns, fewer steps, and directed, auditable exploration.

6 Conclusion

We presented a Bayesian UCB agent augmented with a lightweight *think–act* semantic prior that guides optimism with interpretable hints. In sparse mazes, it reached high return earlier, required fewer steps, and showed calibrated exploration—declining entropy and rapidly shrinking credible bonuses—while exposing per-step rationales. The semantic head faithfully imitated the teacher and reproduced behavior, turning exploration into an auditable object. Future work: scale to deep function approximation and partial observability, incorporate multi-objective/risk-aware indices, and learn priors from human feedback.

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory (COLT)*, 2012a.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *COLT*, 2012b.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. In *Machine Learning*, volume 47, pages 235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. In *Machine Learning*, pages 235–256. Springer, 2002b.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *NeurIPS*, 2016a.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016b.

Yuri Burda, Harrison Edwards, Deepak Pathak, Amos Agarwal, Carlos Diuk, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Exploration by random network distillation. *arXiv:1810.12894*, 2018.

Yuri Burda, Harrison Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Exploration by random network distillation. In *ICLR*, 2019.

Wenlong Huang, Pieter Abbeel, Igor Mordatch, and Deepak Pathak. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning (CoRL)*, 2022a.

Wenlong Huang, Yifeng Miao, et al. Inner monologue: Embodied reasoning through planning with language models. In *CoRL*, 2022b.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. In *Journal of Machine Learning Research*, volume 11, pages 1563–1600, 2010.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *AISTATS*, 2012a.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Bayesian upper confidence bounds for bandit problems. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012b.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML Workshop on Abstraction in Reinforcement Learning*, 2017a.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML Workshop on RL*, 2017b.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv:2201.11903*, 2022b.