Heuristic Transformer: Belief Augmented In-Context Reinforcement Learning

Extended Abstract

Oliver Dippel
University of Liverpool
Liverpool, United Kingdom
oliver.dippel@liverpool.ac.uk

Bei Peng
University of Sheffield
Sheffield, United Kingdom
bei.peng@sheffield.ac.uk

Alexei Lisitsa
University of Liverpool
Liverpool, United Kingdom
A.Lisitsa@liverpool.ac.uk

ABSTRACT

Transformers have recently shown that reinforcement learning can be reframed as an in-context prediction problem, allowing agents to adapt to new tasks without updating their parameters. However, existing in-context RL approaches rely solely on past trajectories as prompts, leaving the model to implicitly infer uncertainty about the reward structure from raw experience alone. We introduce the Heuristic Transformer (HT), an in-context RL method that explicitly augments the prompt with a learned belief over rewards. A low-dimensional stochastic latent variable captures the posterior distribution over rewards and is provided to the transformer alongside trajectories and query states. This enables the model to reason directly over uncertainty rather than inferring it indirectly from data. Across Darkroom, Miniworld, and MuJoCo benchmarks, HT consistently outperforms existing in-context RL baselines in both performance and generalization, particularly in stochastic settings. Our results show that combining belief-based representations with transformer policies is a powerful mechanism for improving in-context decision-making, and suggest a new direction for integrating probabilistic reasoning into transformer-based RL.

KEYWORDS

In-context reinforcement learning, meta-RL, Bayesian RL

ACM Reference Format:
Oliver Dippel, Bei Peng, and Alexei Lisitsa. 2026. Heuristic Transformer: Belief Augmented In-Context Reinforcement Learning: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/IWMS8291

1 INTRODUCTION

Transformers have recently enabled reinforcement learning to be framed as an in-context prediction problem, where an agent adapts to new tasks purely from past trajectories without updating its parameters. Existing in-context RL approaches, however, rely entirely on raw experience in the prompt and must implicitly infer uncertainty about the reward structure from these trajectories. This limits their ability to reason efficiently under task uncertainty, especially in stochastic or partially observable settings. In contrast, meta-RL methods based on Bayes-adaptive Markov decision processes explicitly maintain a belief over possible tasks, allowing agents to approximate Bayes-optimal behavior through belief augmentation. While effective, these approaches typically require specialized architectures and online interaction during training, and do not benefit from the scalability and flexibility of transformer-based in-context learning. We observe that these two paradigms are complementary: transformer policies excel at in-context adaptation from trajectories, while belief-based methods provide an explicit representation of task uncertainty. This motivates combining both ideas. We propose the Heuristic Transformer (HT), an in-context RL method that augments the transformer prompt with a learned belief over rewards. A variational auto-encoder infers a low-dimensional stochastic latent variable from past transitions that represents the posterior distribution over rewards. This belief is provided to the transformer alongside the trajectory context and query state, enabling the model to reason directly over task uncertainty rather than inferring it implicitly from data. HT is trained in two phases: first, learning the belief representation from offline data, and second, training the transformer policy to act conditioned on this belief and the in-context dataset. During evaluation, the agent adapts fully in context without parameter updates, using only recent experience to update its belief and prompt.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/IWMS8291

2 BACKGROUND

We consider reinforcement learning across a distribution of finite-horizon Markov decision processes where the reward function varies between tasks. When the underlying reward is unknown, principled decision-making requires maintaining a belief over possible rewards and updating this belief from experience. In Bayes-adaptive formulations, such belief representations enable agents to approximate Bayes-optimal behavior by explicitly reasoning about task uncertainty. In this work, we adopt a simplified Bayesian perspective and model only a posterior belief over rewards rather than over full environment dynamics. This belief is inferred from past transitions and used heuristically to guide decision-making under uncertainty. While this does not yield a fully Bayes-optimal policy, it provides a compact and informative representation of task uncertainty that can be leveraged by a policy model.

3 HEURISTIC TRANSFORMER

We propose Heuristic Transformer (HT), which is trained in the following two phases.

Phase 1. Learning a belief over rewards. From offline transition data collected across tasks, a variational auto-encoder learns a low-dimensional latent variable that represents the posterior distribution over rewards for a task. This latent belief is inferred from past transitions and provides a compact representation of task-specific reward structure. Rather than modeling full environment dynamics, we focus on learning a belief over rewards, which is sufficient to guide decision-making while remaining tractable.

Phase 2. Conditioning a transformer policy on belief and context. A transformer policy is then trained via supervised learning to predict optimal actions conditioned on (i) an in-context dataset of past transitions, (ii) the learned belief over rewards, and (iii) a query state. During training, tasks are sampled from a distribution, and the model learns to generalize action prediction across tasks given different beliefs and contexts. At evaluation time, the agent adapts entirely in context: recent experience updates the belief over rewards, which is fed into the transformer together with trajectories to produce actions, without any parameter updates.

4 RELATED WORK

Prior work has shown that transformers can model decision-making from offline trajectories by treating reinforcement learning as sequence prediction, but these approaches typically rely solely on trajectory context. In parallel, Bayesian and meta-RL methods approximate Bayes-optimal behavior by maintaining beliefs over tasks, but require specialized architectures and online interaction during training. Recent in-context RL methods combine transformers with meta-learning ideas, yet they either depend on implicit task cues in the prompt or require additional meta-knowledge to construct informative context. HT bridges these directions by explicitly learning a belief over rewards and using it as structured input to a transformer policy. This combines the scalability of transformer-based in-context learning with the principled uncertainty handling of belief-based methods, without requiring online interaction or task-specific prompt engineering.

Our work builds on recent efforts to treat reinforcement learning as sequence modeling with transformers, such as Decision Transformer [1] and subsequent in-context RL approaches including Algorithm Distillation [3], Decision-Pretrained Transformer [4], and Goal-Focused Transformer [2]. In parallel, Bayesian and meta-RL methods such as VariBAD [5] demonstrate the importance of belief representations for task generalization. HT connects these lines of work by explicitly integrating belief modeling into a transformer-based in-context RL framework, combining probabilistic reasoning with scalable sequence models.

5 EXPERIMENTS

We evaluate HT across three increasingly challenging settings to assess its in-context adaptation and generalization capabilities: (i) Darkroom, a sparse-reward gridworld, including a larger *Darkroom Hard* variant and stochastic versions with controlled action misdirection, (ii) Miniworld, a 3D visual navigation task with image-based observations, and (iii) four MuJoCo continuous-control benchmark tasks (Hopper, Walker2d, HalfCheetah, Swimmer).

We compare HT to closely related in-context RL baselines, primarily Decision-Pretrained Transformer (DPT) and, where applicable, Goal-Focused Transformer (GFT). We also report results relative to RL^2 as a soft upper bound, as it benefits from online interaction during training.

Darkroom and Darkroom Hard. HT demonstrates significantly faster online adaptation and higher returns during early episodes compared to DPT and GFT. The performance gap becomes particularly pronounced in Darkroom Hard, where the context size and environment complexity increase substantially. While baselines often fail to consistently locate the goal, HT steadily improves with increased pre-training, showing strong generalization to unseen tasks.

Transition Uncertainty. Under stochastic transition dynamics, HT shows only minor degradation in performance and continues to outperform DPT. This highlights HT's ability to reason effectively under uncertainty when the environment dynamics become noisy.

Miniworld. In the image-based Miniworld environment, HT consistently achieves higher returns and faster adaptation than DPT across all pre-training stages, demonstrating that belief-augmented prompting remains effective even with high-dimensional visual observations.

MuJoCo. Although MuJoCo tasks correspond to fixed MDPs with limited task variation, HT remains competitive and benefits from diverse offline training data. Training HT on a mixture of PPO and SAC rollouts (HT-SP) yields the strongest performance and consistently outperforms DPT across all tasks (e.g., Walker2d: 3565 ± 433 vs. 3099 ± 433; HalfCheetah: 1968 ± 61 vs. 1879 ± 61).

Overall, HT consistently achieves faster online adaptation and higher returns across environments, particularly in settings with task uncertainty or increased environmental complexity.

6 CONCLUSION

We presented the Heuristic Transformer (HT), an in-context reinforcement learning method that augments transformer policies with a learned belief over rewards. By conditioning action prediction on this belief together with trajectory context, HT enables agents to reason explicitly about task uncertainty while retaining the scalability and flexibility of transformer-based in-context learning.

Across gridworld, visual navigation, and continuous-control benchmarks, HT demonstrates faster online adaptation, stronger generalization, and improved robustness to stochastic dynamics compared to the in-context RL baselines. These results show that belief-augmented prompting is an effective mechanism for improving decision-making in transformer policies.

While HT currently relies on high-quality pre-training data and access to optimal actions during training, future work may relax these assumptions and explore broader cross-environment generalization. Overall, our findings suggest that integrating structured uncertainty representations into transformer-based RL is a promising direction for learning adaptive and generalizable policies.

REFERENCES

[1] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

[2] Oliver Dippel, Alexei Lisitsa, and Bei Peng. 2024. Contextual Transformers for Goal-Oriented Reinforcement Learning. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 207–220.

[3] Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, D Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. [n.d.]. In-context reinforcement learning with algorithm distillation, 2022. *URL https://arxiv. org/abs/2210.14215* ([n. d.]).

[4] Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. 2023. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 43057–43083.

[5] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. 2019. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348* (2019).