CS 690: Human-Centric Machine Learning

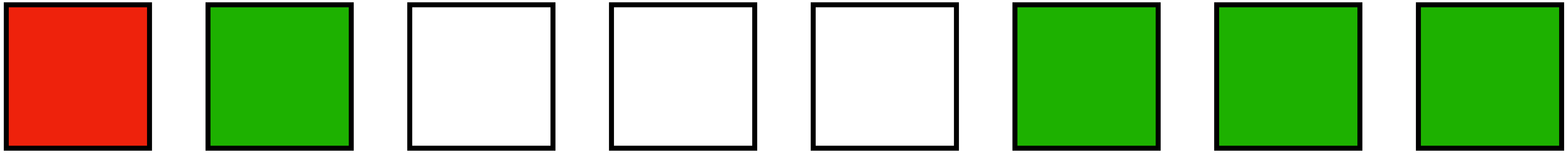
Prof. Scott Niekum

Improving human modeling assumptions

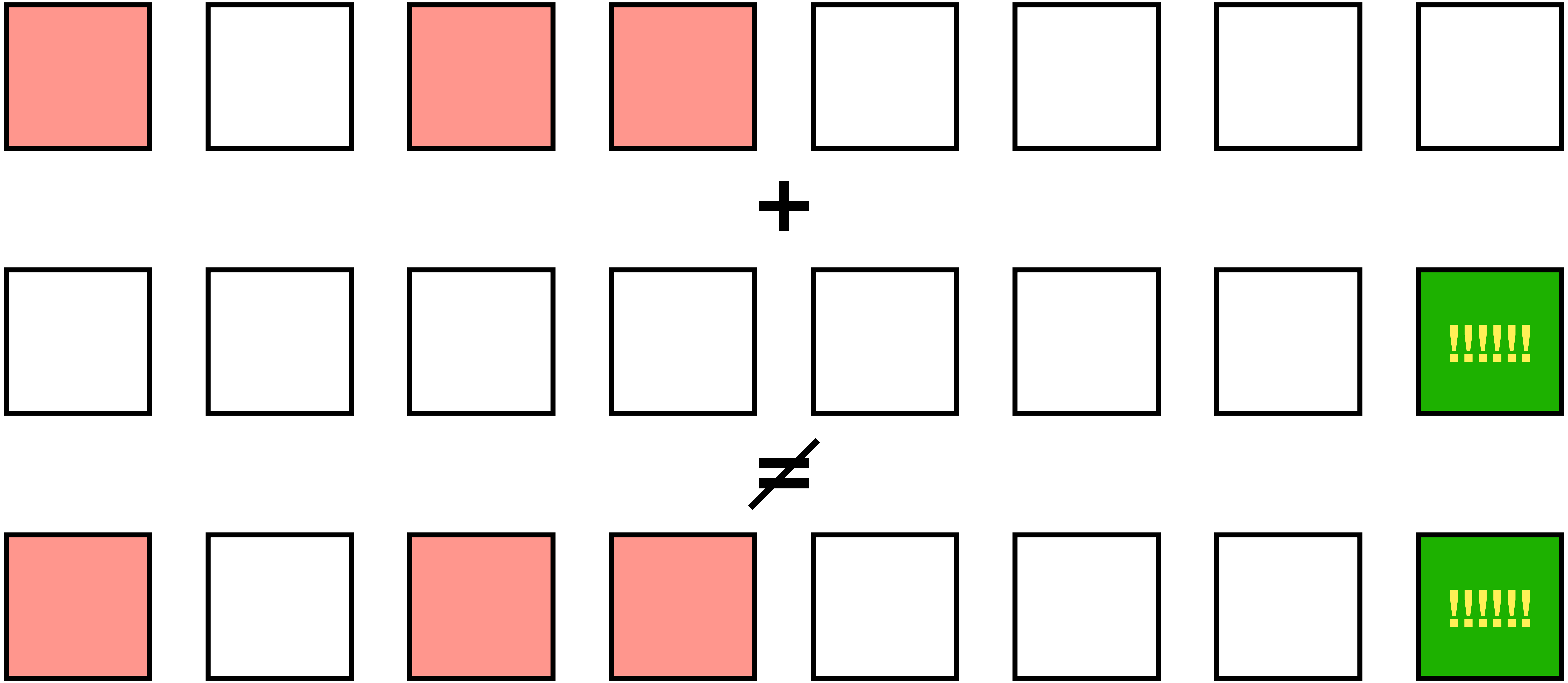
Equally-weighted markovian rewards?

Kim, Changyeon, et al. "Preference transformer: Modeling human preferences using transformers for rl." *arXiv preprint arXiv:2303.00957* (2023).

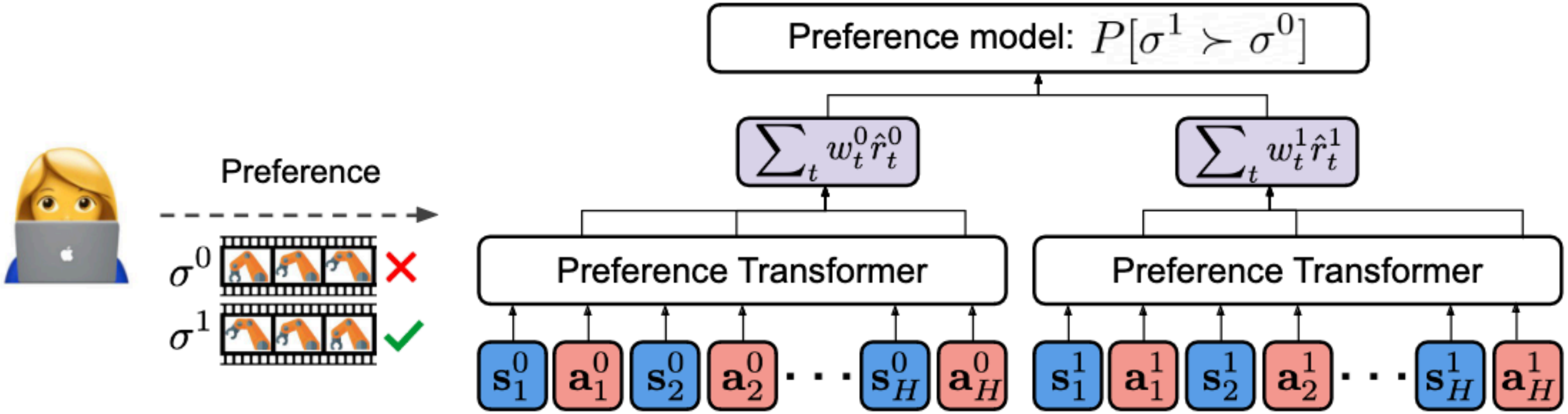
Equally-weighted markovian rewards?



Equally-weighted markovian rewards?



Preference transformer

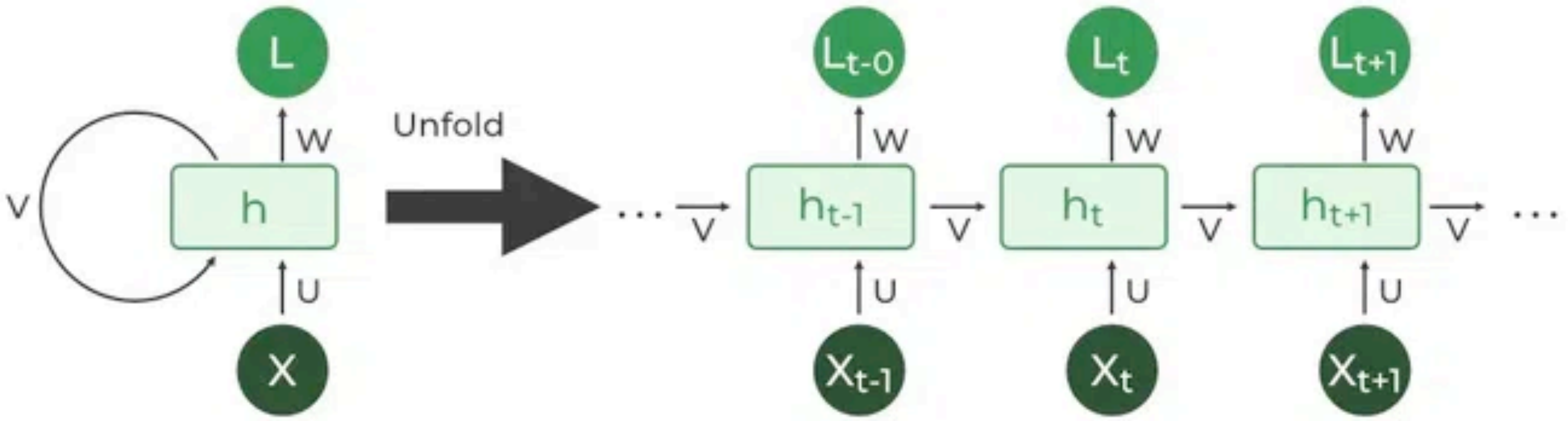


Kim, Changyeon, et al. "Preference transformer: Modeling human preferences using transformers for rl." *arXiv preprint arXiv:2303.00957* (2023).

Preference transformer

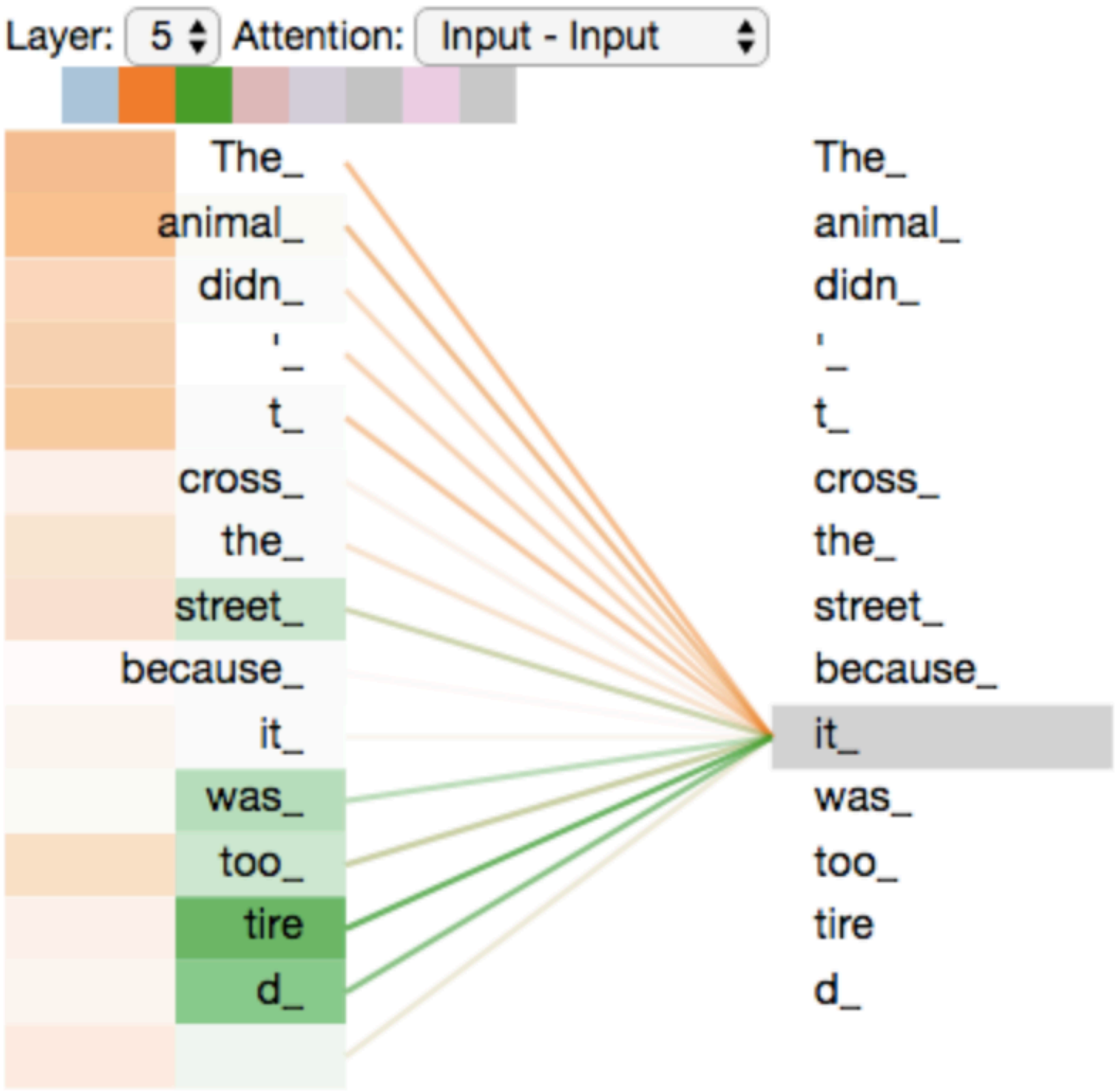
$$P[\sigma^1 \succ \sigma^0; \psi] = \frac{\exp\left(\sum_t w\left(\{(\mathbf{s}_i^1, \mathbf{a}_i^1)\}_{i=1}^H; \psi\right)_t \cdot \hat{r}\left(\{(\mathbf{s}_i^1, \mathbf{a}_i^1)\}_{i=1}^t; \psi\right)\right)}{\sum_{j\in\{0,1\}} \exp\left(\sum_t w\left(\{(\mathbf{s}_i^j, \mathbf{a}_i^j)\}_{i=1}^H; \psi\right)_t \cdot \hat{r}\left(\{(\mathbf{s}_i^j, \mathbf{a}_i^j)\}_{i=1}^t; \psi\right)\right)}$$

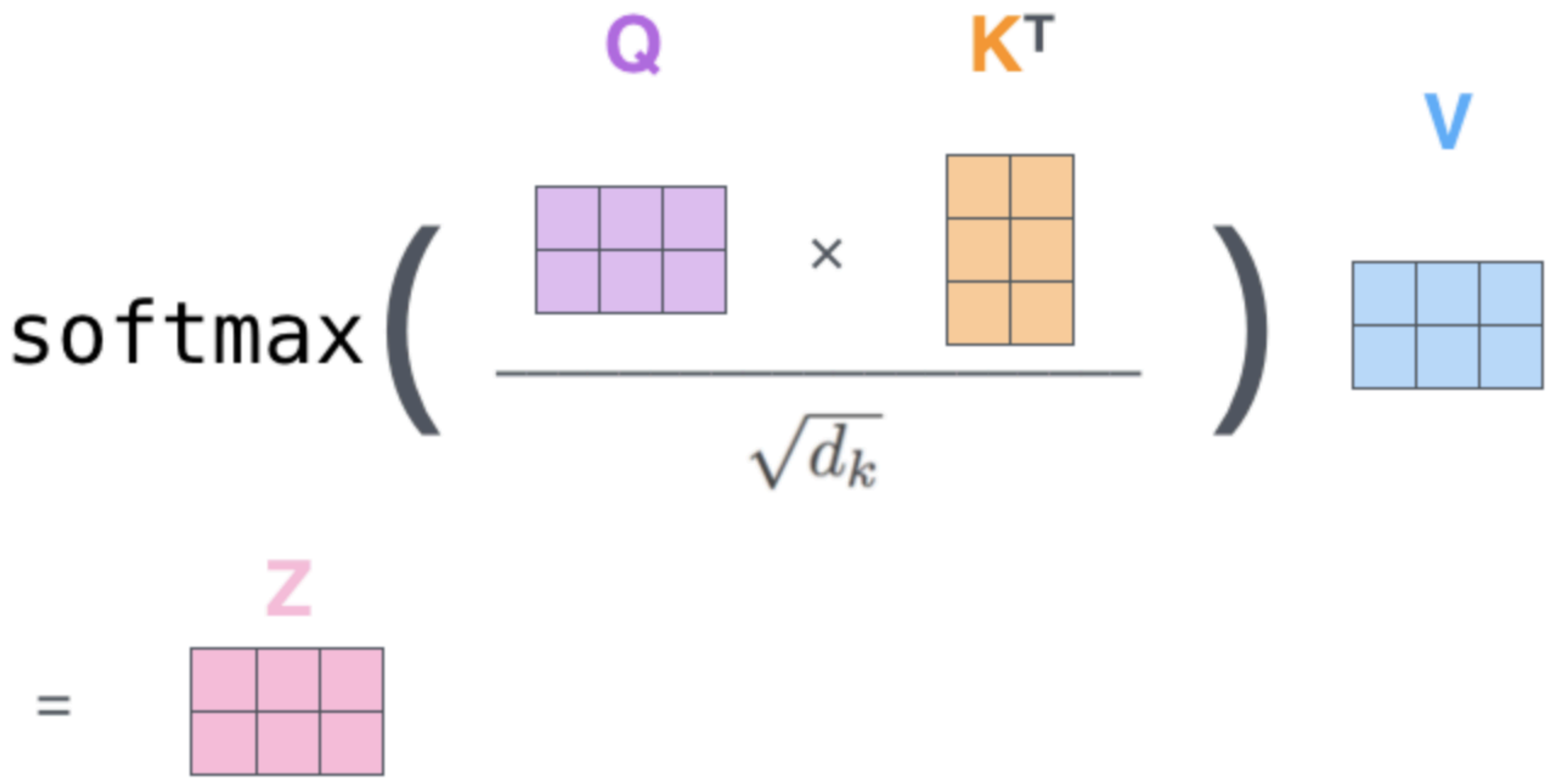
Some history: RNNs



https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/

Transformers





https://jalammar.github.io/illustrated-transformer/

Preference transformer

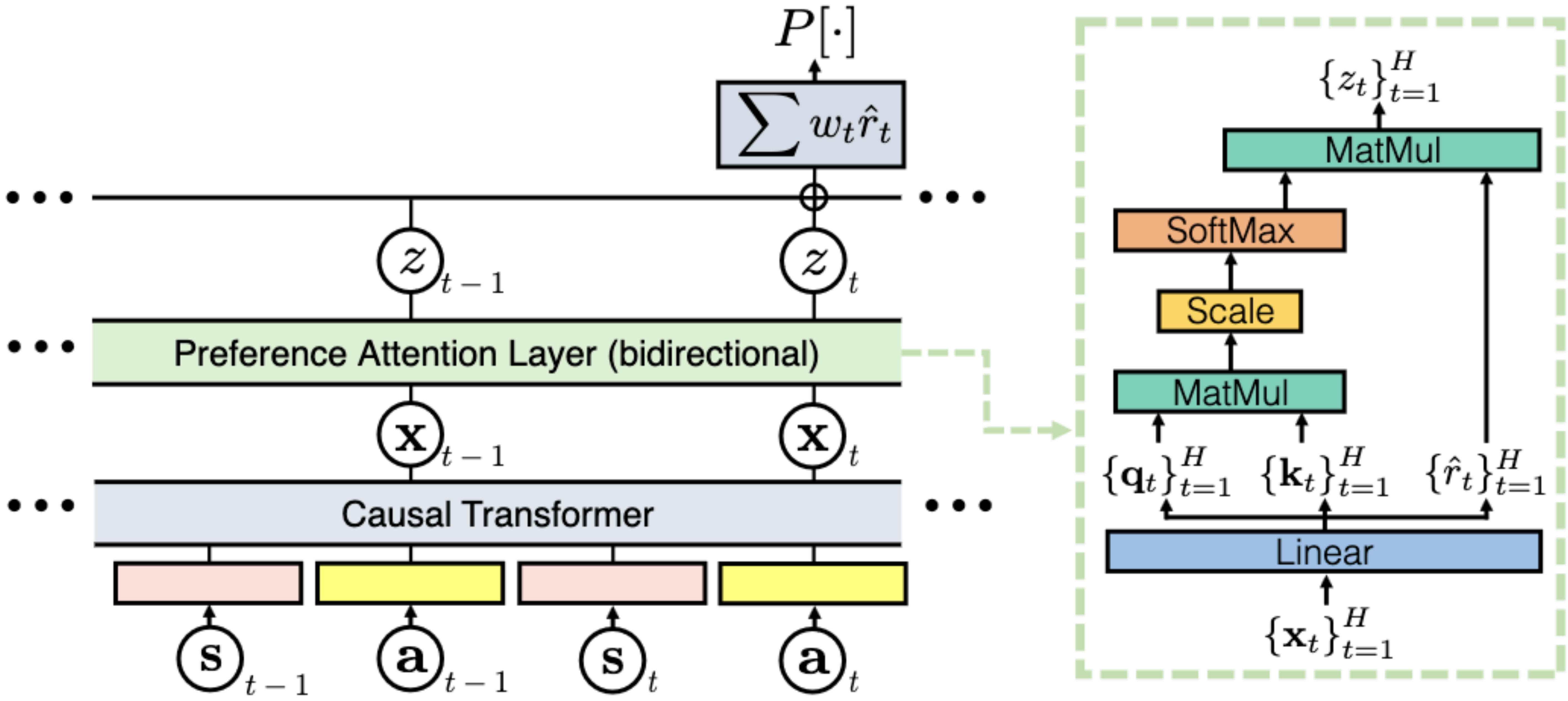
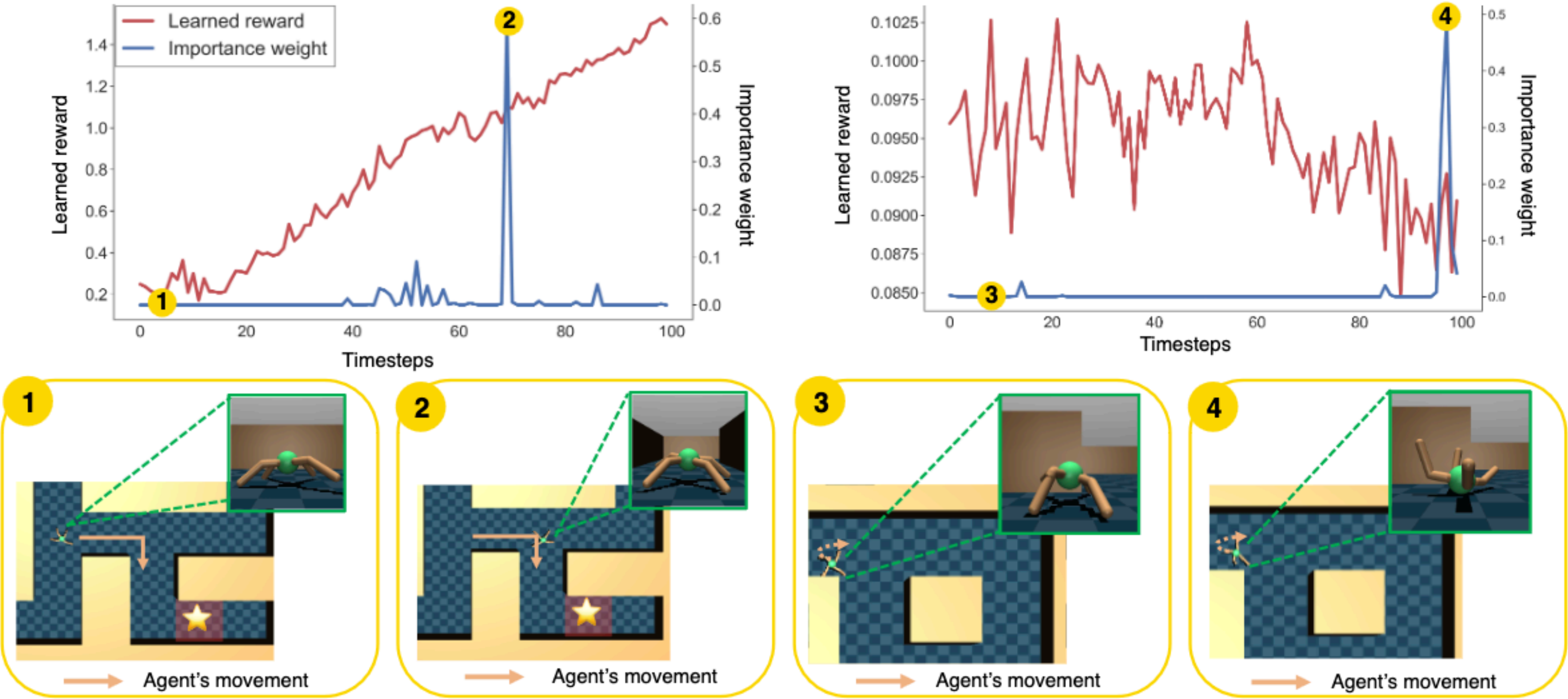


Figure 2: Overview of Preference Transformer. We first construct hidden embeddings $\{\mathbf{x}_t\}$ through the causal transformer, where each represents the context information from the initial timestep to timestep t. The preference attention layer with a bidirectional self-attention computes the non-Markovian rewards $\{\hat{r}_t\}$ and their convex combinations $\{z_t\}$ from those hidden embeddings, then we aggregate $\{z_t\}$ for modeling the weighted sum of non-Markovian rewards $\sum_t w_t \hat{r}_t$.

Preference transformer: results

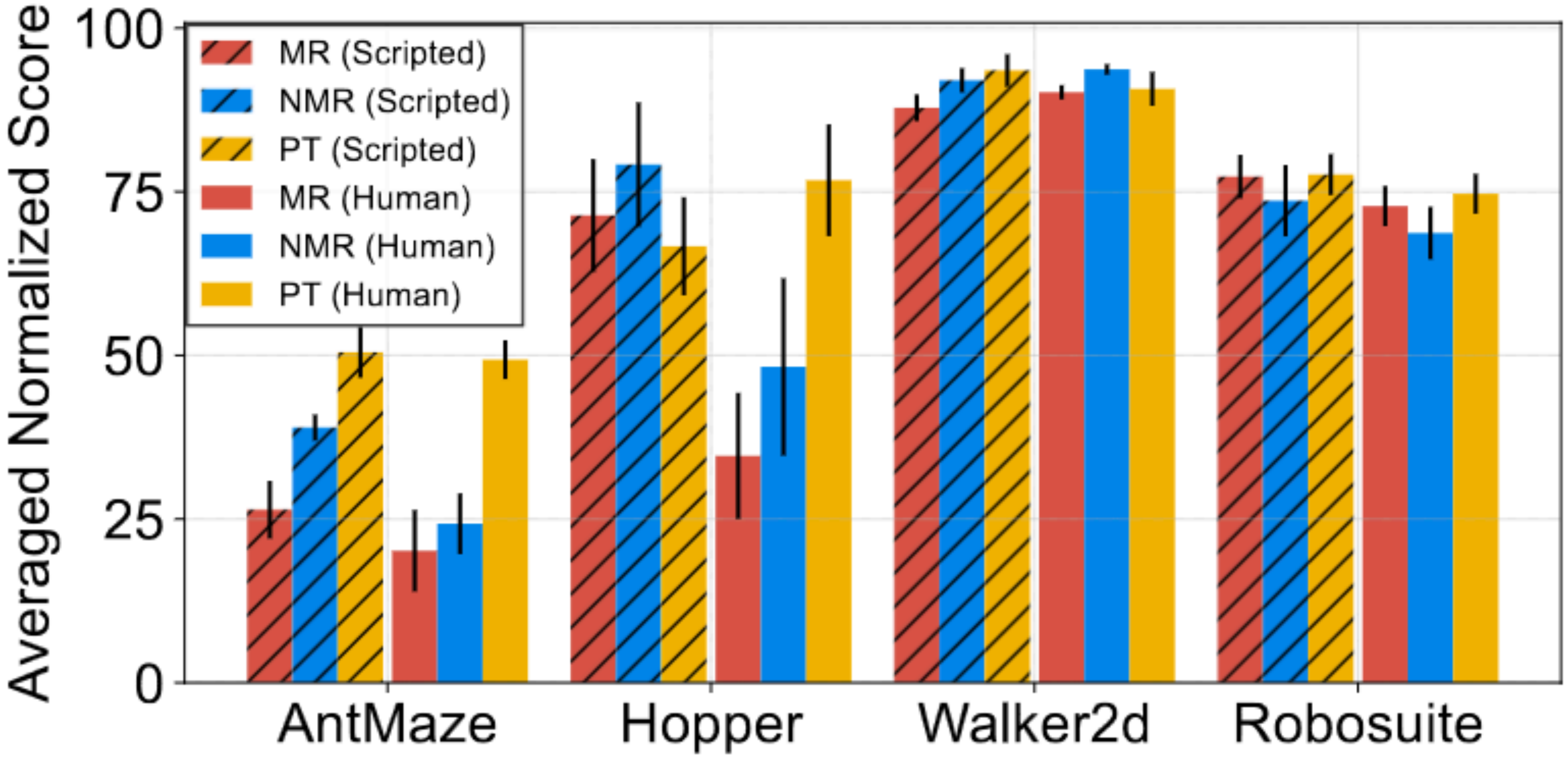
ent data collection schemes. For reward learning, we select queries (pairs of trajectory segments) uniformly at random from offline datasets and collect preferences from real human trainers (the authors).[2] Then, using the collected datasets of human preferences, we learn a reward function and

Preference transformer: results



(a) Successful trajectory (b) Failure trajectory

Preference transformer: results



How bad is irrationality?

Chan, Lawrence, Andrew Critch, and Anca Dragan. "Human irrationality: both bad and good for reward inference." *arXiv preprint arXiv:2111.06956* (2021).

A general model for irrationality

Assume human is a planner with irrationalities being deviations from the Bellman Equation:

Myopic Discounting
Myopic Horizon
Hyperbolic Discounting

Boltzmann-Rationality

Prospect Bias

$$V_{i+1}(s) = \max_a \sum_{s' \in S} P_{s,a}(s') \left(r_\theta(s, a, s') + \gamma V_i(s') \right)$$

Optimism / Pessimism
Illusion of Control

Extremal Bias

Types of irrationality

Modify max operator: Boltzmann rationality

$$V_{i+1}(s) = \mathrm{Boltz}_a^\beta \sum_{s' \in S} P_{s,a}(s') \left(r_\theta(s,a,s') + \gamma V_i(s')\right)$$

$$\text{where } \mathrm{Boltz}^\beta(\mathbf{x}) \; = \; \textstyle\sum_i x_i e^{\beta x_i} / \sum_i e^{\beta x_i}$$

Types of irrationality

Modify transition dynamics: Illusion of control

$$V_{i+1}(s) = \max_a \sum_{s' \in S} P^n_{s,a}(s') \left(r_\theta(s,a,s') + \gamma V_i(s')\right)$$

where $P^n_{s,a}(s') = \left(P_{s,a}(s')\right)^n / \sum_{s'' \in S} \left(P_{s,a}(s'')\right)^n$

Types of irrationality

Modify transition dynamics: Optimism / pessimism

$$V_{i+1}(s) = \max_a \sum_{s' \in S} P^{\omega}_{s,a}(s') \left(r_\theta(s, a, s') + \gamma V_i(s') \right)$$

$$\text{where} \;\; P^{\omega}_{s,a}(s') \;\; \propto \;\; P_{s,a}(s') e^{\omega \left(r_\theta(s,a,s') + \gamma V_i(s') \right)}$$

Types of irrationality

Modify reward: Prospect bias

$$V_{i+1}(s) = \max_a \sum_{s' \in S} P_{s,a}(s') \left(f(r_\theta(s, a, s')) + \gamma V_i(s') \right)$$

$$f_c(r) = \begin{cases} \log(1 + |r|) & r > 0 \\ 0 & r = 0 \\ -c \log(1 + |r|) & r < 0 \end{cases}$$

Types of irrationality

Modify relation between reward+future value: Extremal

$$V_{i+1}(s) = \max_{a} \sum_{s' \in S} P_{s,a}(s') \max \begin{cases} r_\theta(s, a, s') \\ (1-\alpha) r_\theta(s, a, s') + \alpha V_i(s') \end{cases}$$

Types of irrationality

Modify discounting

- Myopic discount (standard discounting with gamma)
- Myopic value iteration (only H steps performed)
- Hyperbolic discounting:

$$V_{i+1}(s) = \max_a \sum_{s' \in S} P_{s,a}(s') \frac{r_\theta(s,a,s') + V_i(s')}{1 + kV_i(s')}$$

Effects of irrationality on Bayesian inference

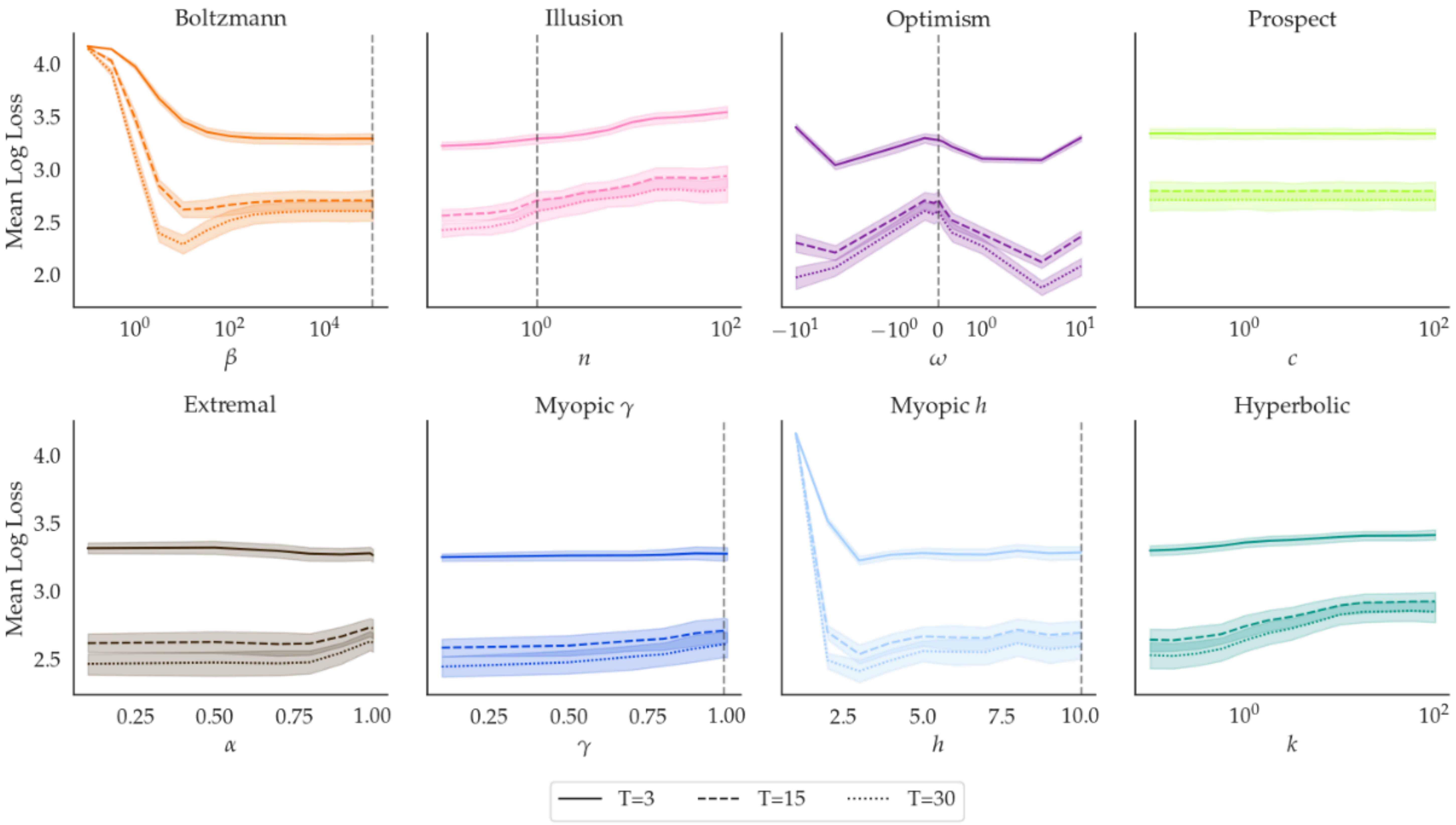
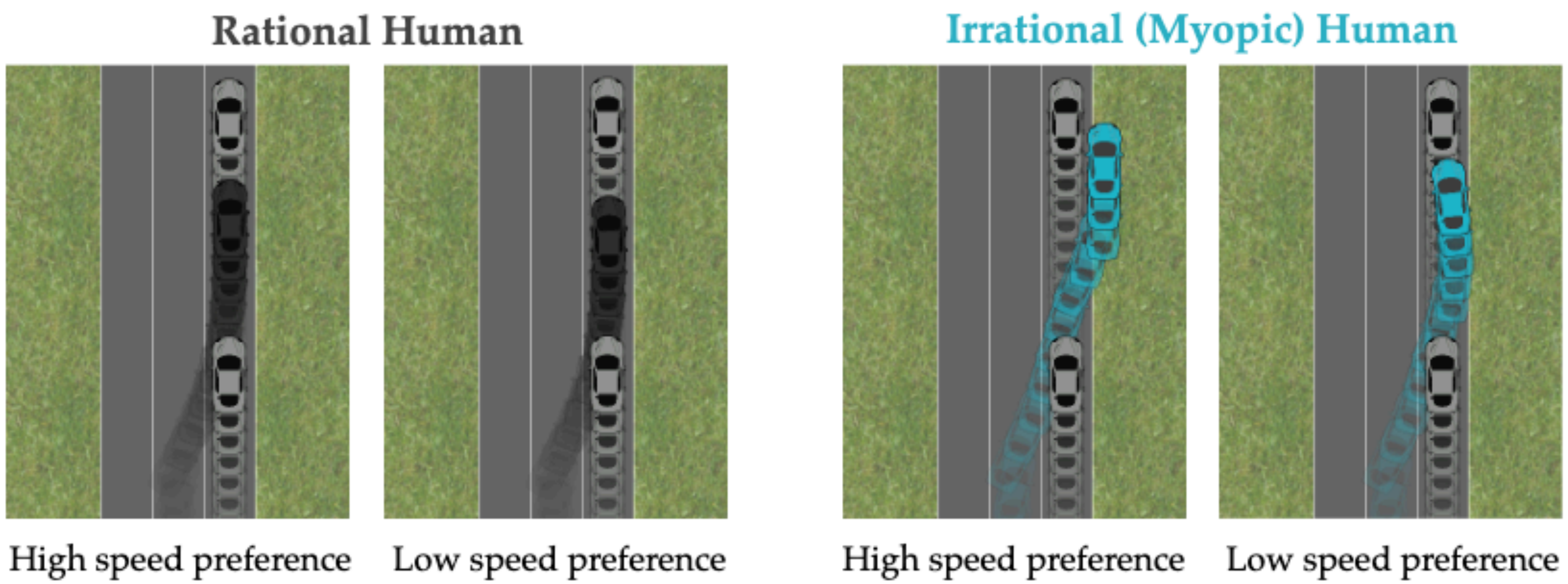


Fig. 3: The log loss (lower = better) of the posterior as a function of the parameter we vary for each irrationality type, on the random MDP environments. For the irrationalities that interpolate to the rational planner, we denote the value that is closest to rational using a dashed vertical line. Every irrationality except Prospect Bias all have parameter settings that outperform the rational planner. The error bars show the standard error of the mean, calculated by 1000 bootstraps across environments.

Irrationality can be good (if correctly modeled)!



Driving Scenario: Merging Our experiments were performed in a simple merging environment (Fig. 1). In it, the human wants to merge into the right lane while trying to maintain its 1.2 forward speed. In addition to the human car, the right lane contains two constant velocity cars, traveling at 0.8 speed. The features of this environment are composed of a squared penalty for deviating from 1.2 forward speed, features for the squared distances to the medians of each of the lanes, a feature for the minimum squared distance to any of the medians of the lanes, and a smooth collision feature.

Unmodeled irrationality is very bad

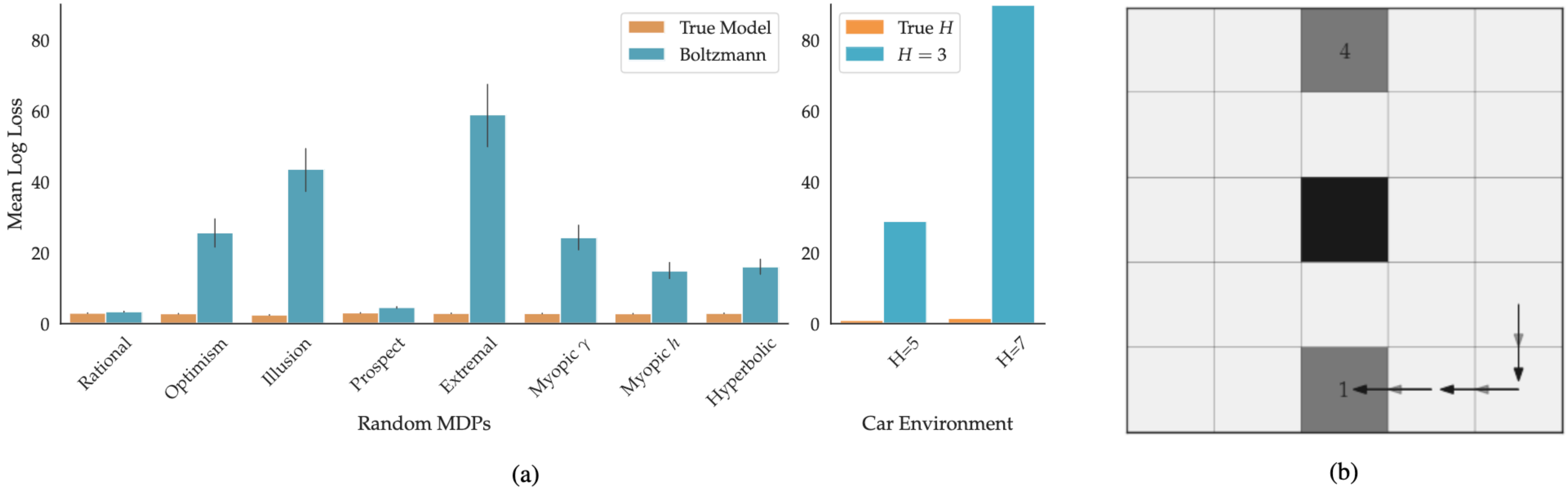


Fig. 6: (a) A comparison of reward inference using a correct model of the irrationality type, versus always using a Boltzmann-rational model ($\beta = 10$), on the random MDPs (left) and the car environment (right). The impairment due to model misspecification greatly outweighs the variation in inference performance caused by various irrationalities. The error bars show the standard error of the mean, calculated by the bootstrap across environments. (b) An example of why assuming Boltzmann is bad when the ground truth human is Myopic in the gridworld environment - the Boltzmann rational agent would take the trajectory depicted only if the reward at the bottom was not much less than the reward at the top. A myopic human with $n \leq 4$, however, only "sees" the reward at the bottom. Consequently, inferring the preferences of the myopic agent as if it were Boltzmann leads to poor performance in this case.

Approximate irrationality models might be enough

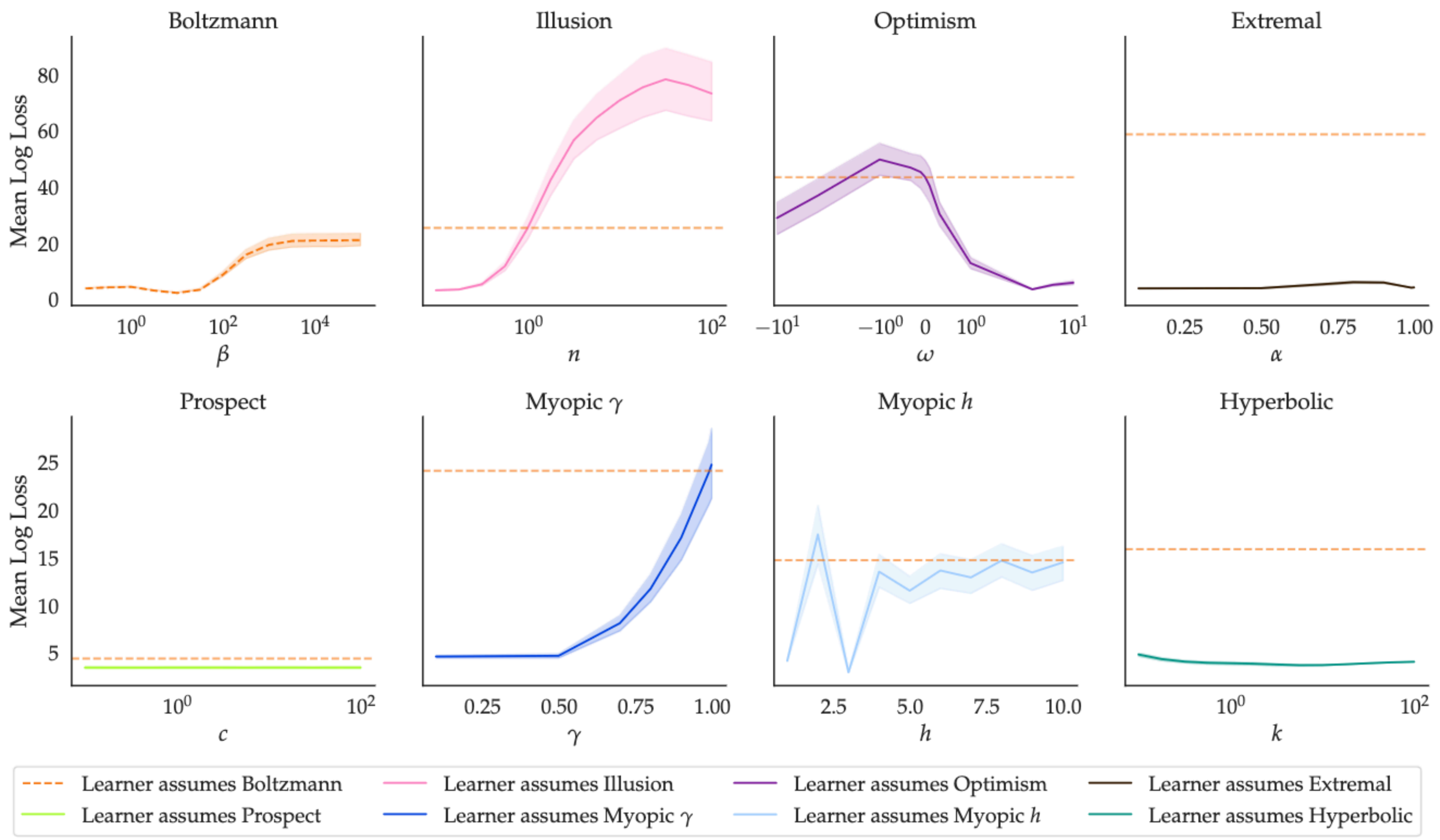


Fig. 7: The log loss (lower = better) of various models under parameter misspecification. Each x-axis shows the parameter that the robot assumes. The orange line represents the performance when the robot makes the faulty assumption that the human is Boltzmann-rational. In many cases, the robot perform better than by assuming Boltzmann-rational just by getting the type of the planner correct, even if they don't get the exact parameter correct. The error bars show the standard error of the mean, calculated by the bootstrap across environments.

Approximate irrationality models might be enough

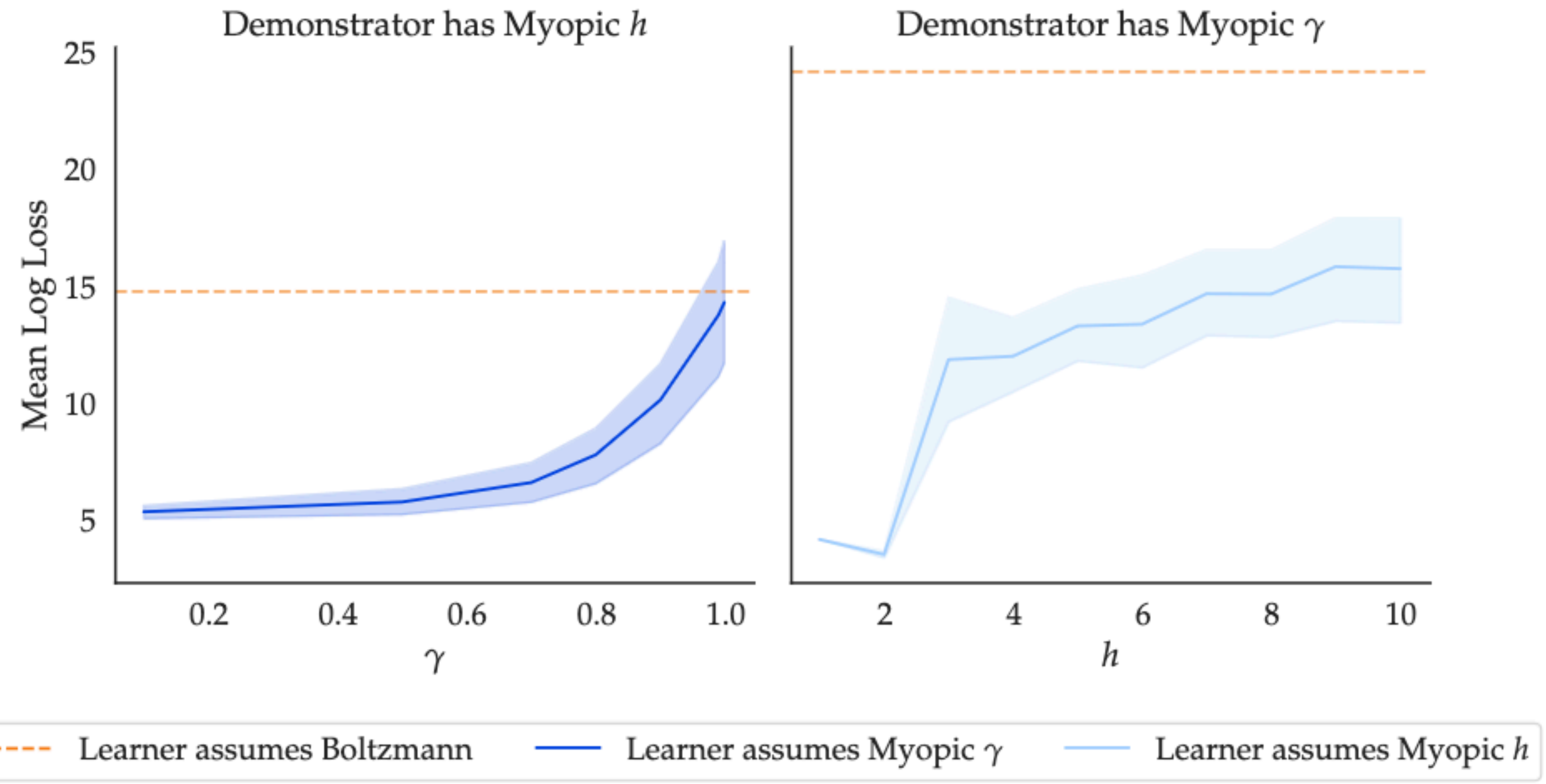


Fig. 8: The log loss (lower = better) of two myopic humans under type misspecification. On the left, the human performs myopic value iteration (Myopic h), but the robot assumes the human has a myopic discount rate γ (Myopic γ). On the right, the human has a myopic discount rate γ but the robot assumes myopic value iteration. However, in both cases, this leads to better inference than assuming Boltzmann-rationality.