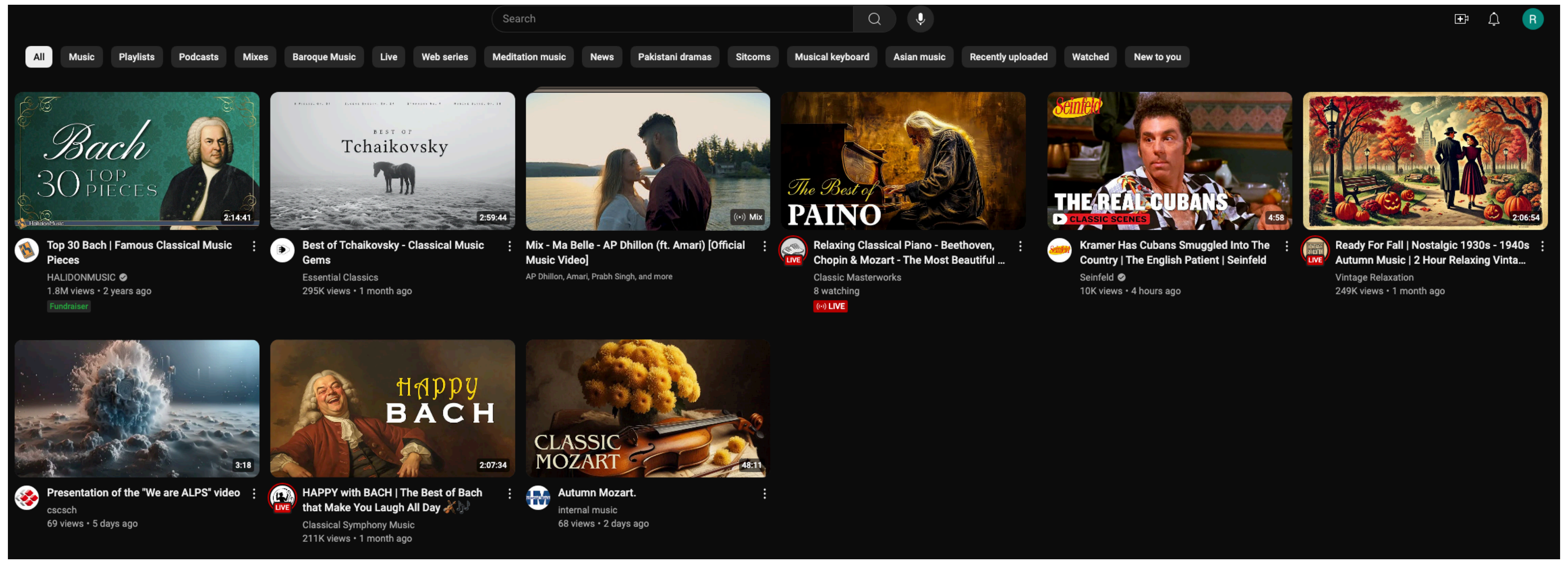
MultiScale Contextual Bandits for Long Term Objectives

Richa Rastogi, Yuta Saito, Thorsten Joachims
Cornell University

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- In many interactive AI systems, (recommender, conversational systems), there is abundant short term feedback (e.g., clicks, generated response quality)



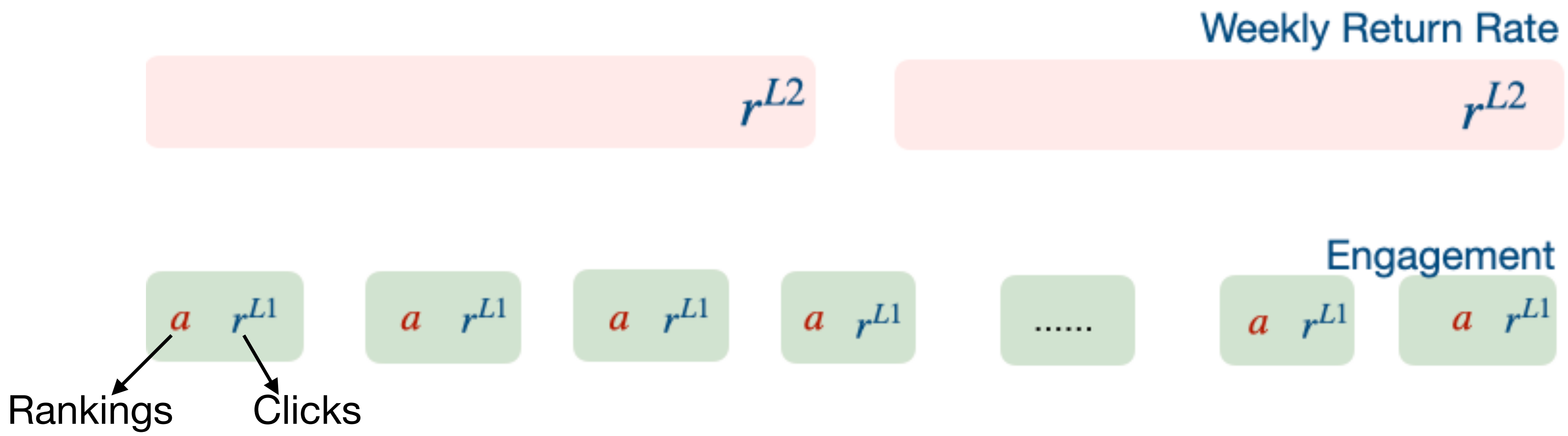
User Retention

Beneficial dialogue outcomes

- Prior work shows that optimizing for short term feedback does not necessarily achieve the desired long term objective (e.g., clickbait feeds do not lead to user retention)

Motivation

A key problem — long-term feedback is at a different timescale than the short-term interventions

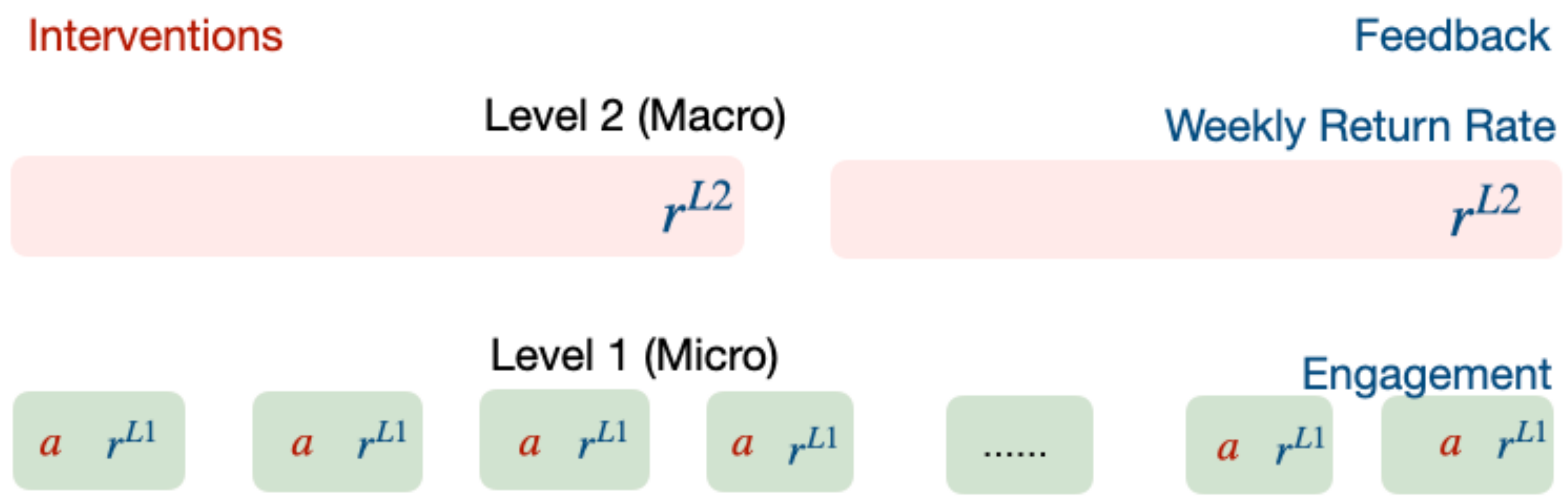


We address it by contextually reconciling this disconnect in timescales

MultiScale Policy Framework

Consider two levels

- A micro level that operates at faster timescale, e.g., clicks, response quality
- A macro level that operates at slower timescale, e.g., user retention



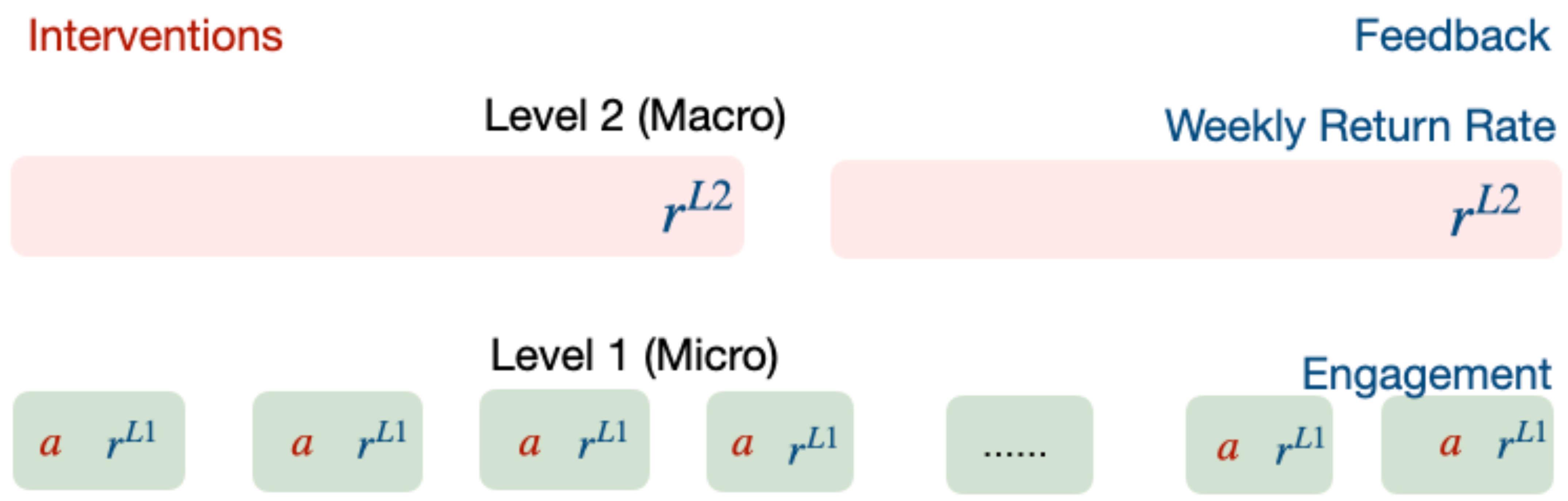
MultiScale Policy Framework

$$\pi^{L2*} \leftarrow \arg\max_{\pi \in \Pi} V^{L2}(\pi)$$

Hard

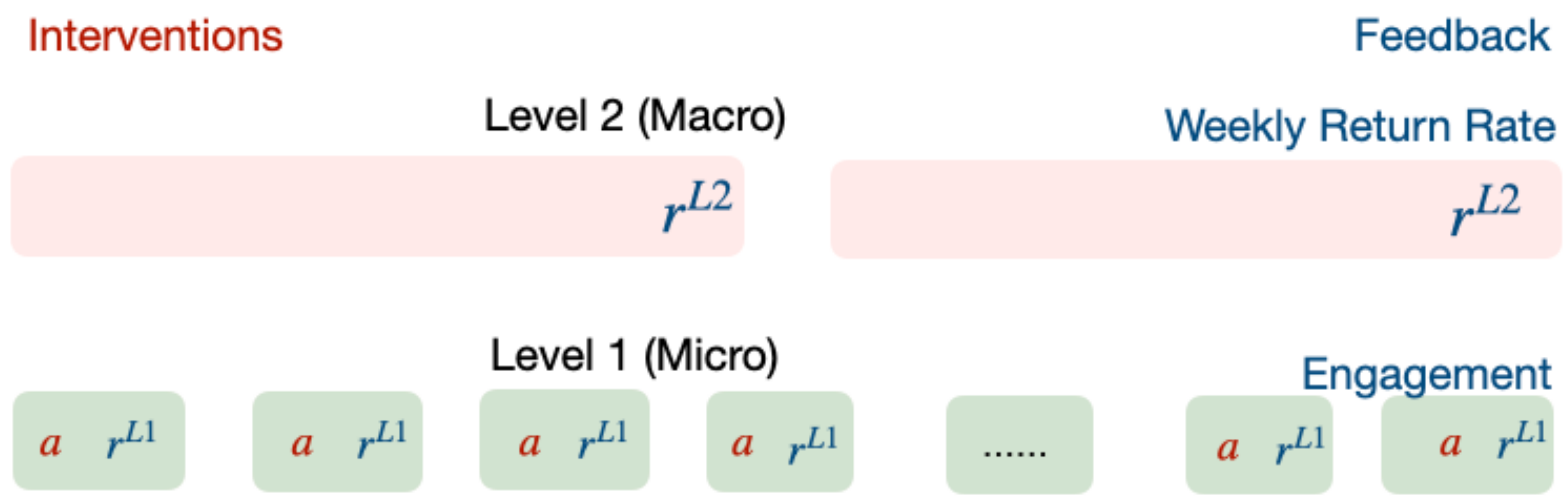
$$\pi^{L1*} \leftarrow \arg\max_{\pi \in \Pi} V^{L1}(\pi)$$

Easy



Even though $V^{L2}(\pi^{L1*}) < V^{L2}(\pi^{L2*})$, π^{L1*} is typically much better than a random policy from Π

MultiScale Policy Framework

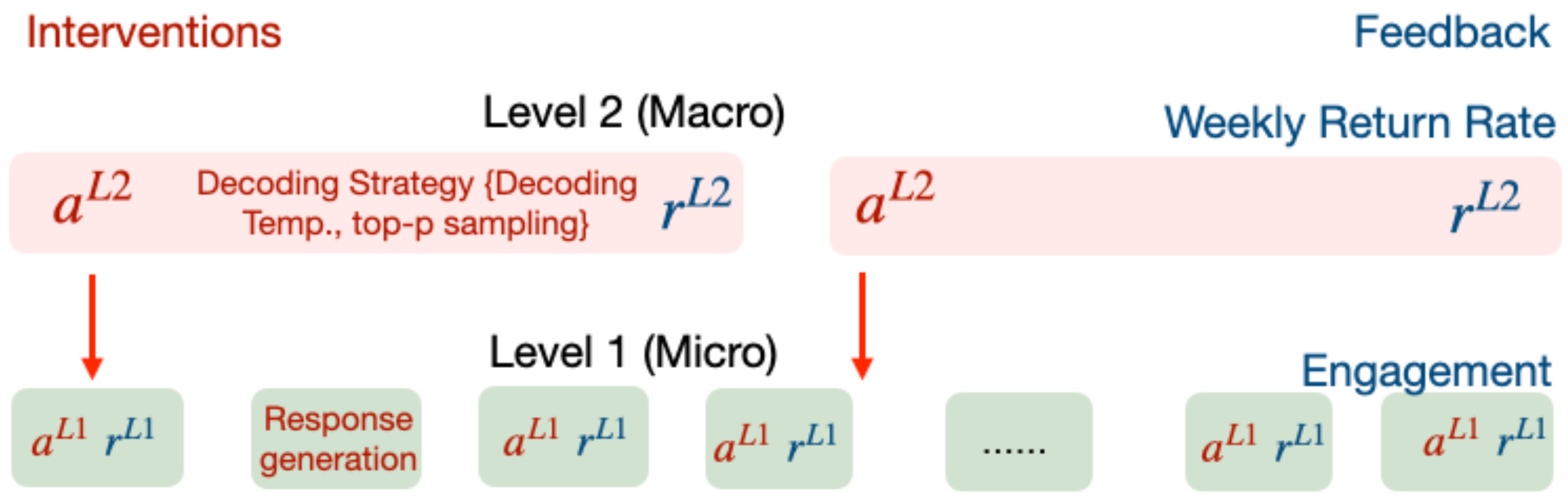


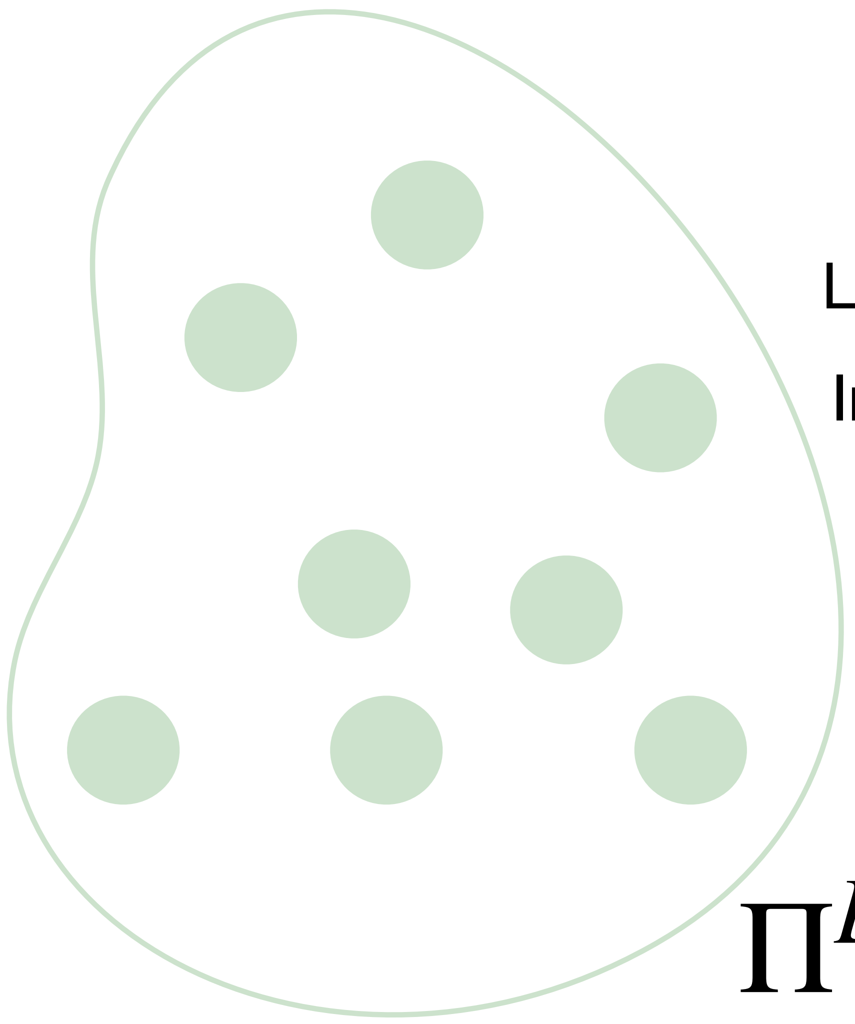
Can we exploit feedback at the micro level to learn the long term optimal policy?

MultiScale Policies

Factorization of policies

$$\Pi \triangleq \Pi^{L1} \cdot \Pi^{L2}$$

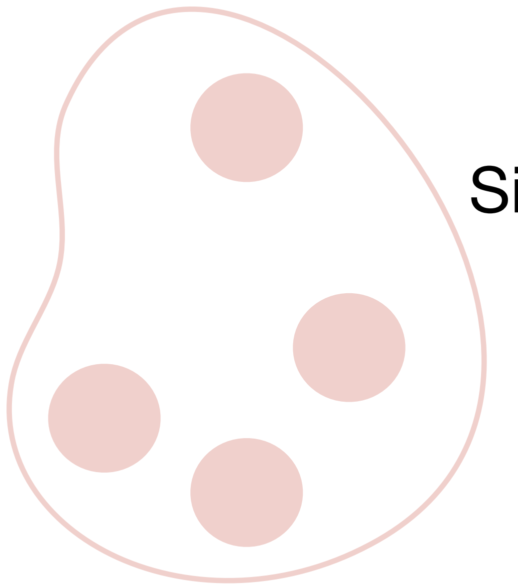




Learns a large part of the parameter space

Inductive bias for long term optimal policy

Π^{L1}

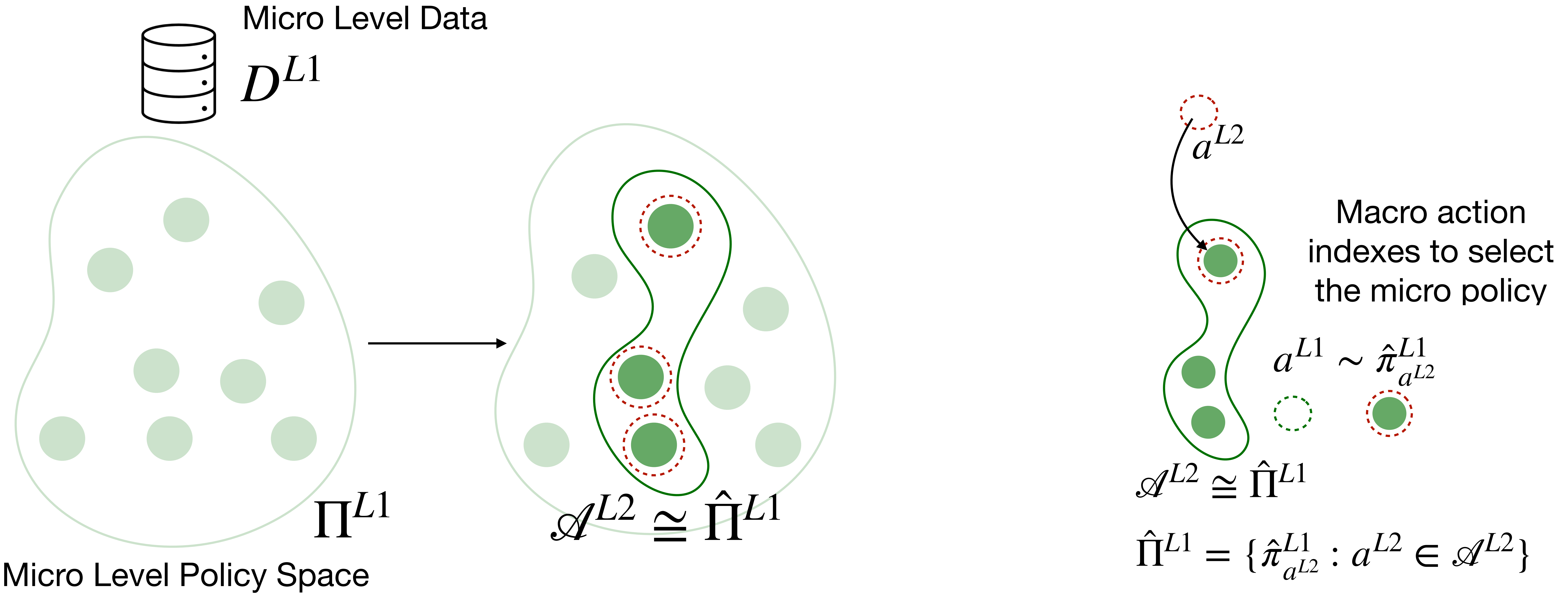


Simplified learning at macro level

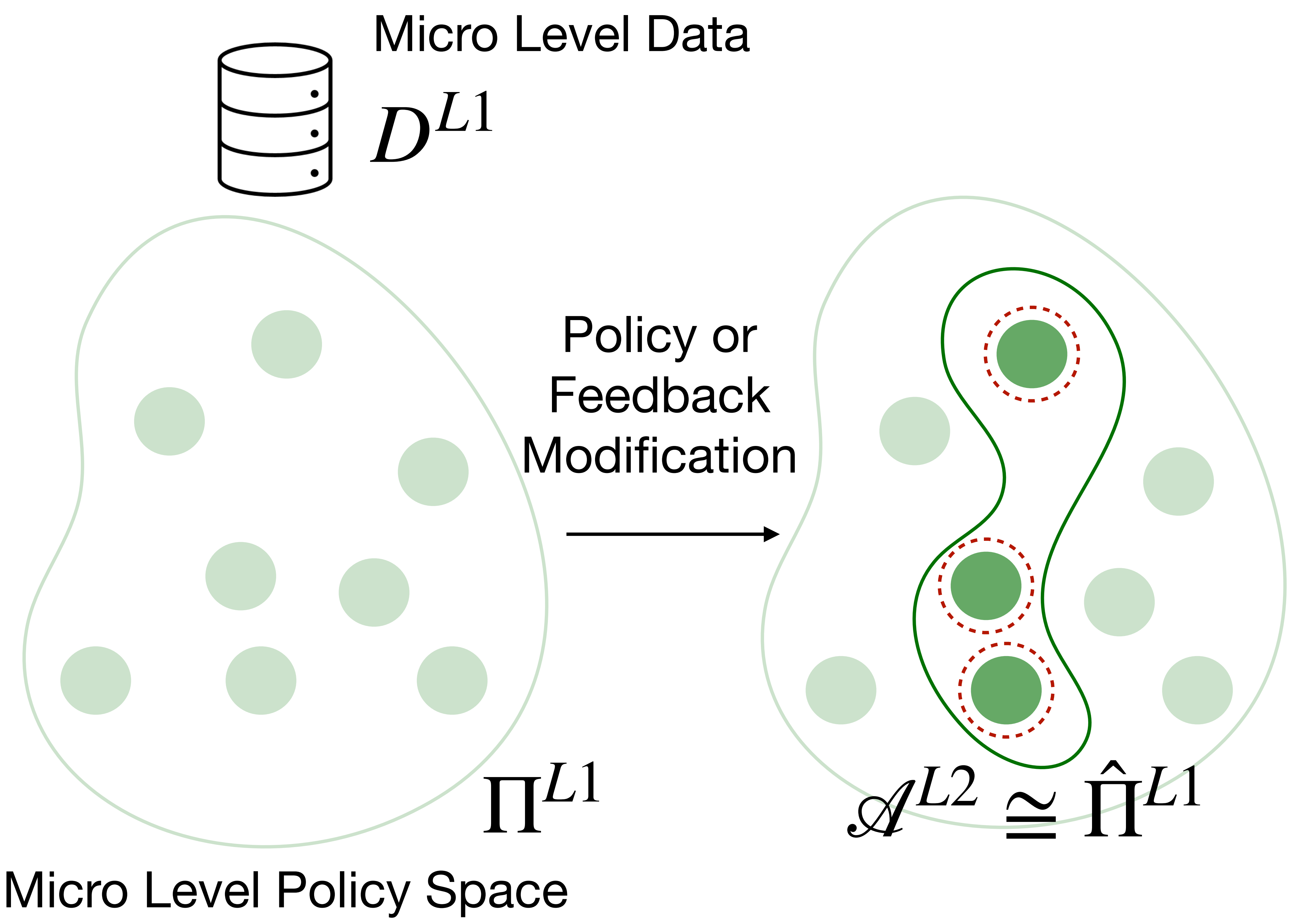
Small policy space

Π^{L2}

Policy Learning at micro level



Policy Learning at micro level



Example:

$$r^{L1} \triangleq \begin{bmatrix} \text{Clicks} \\ \text{Likes} \end{bmatrix}$$

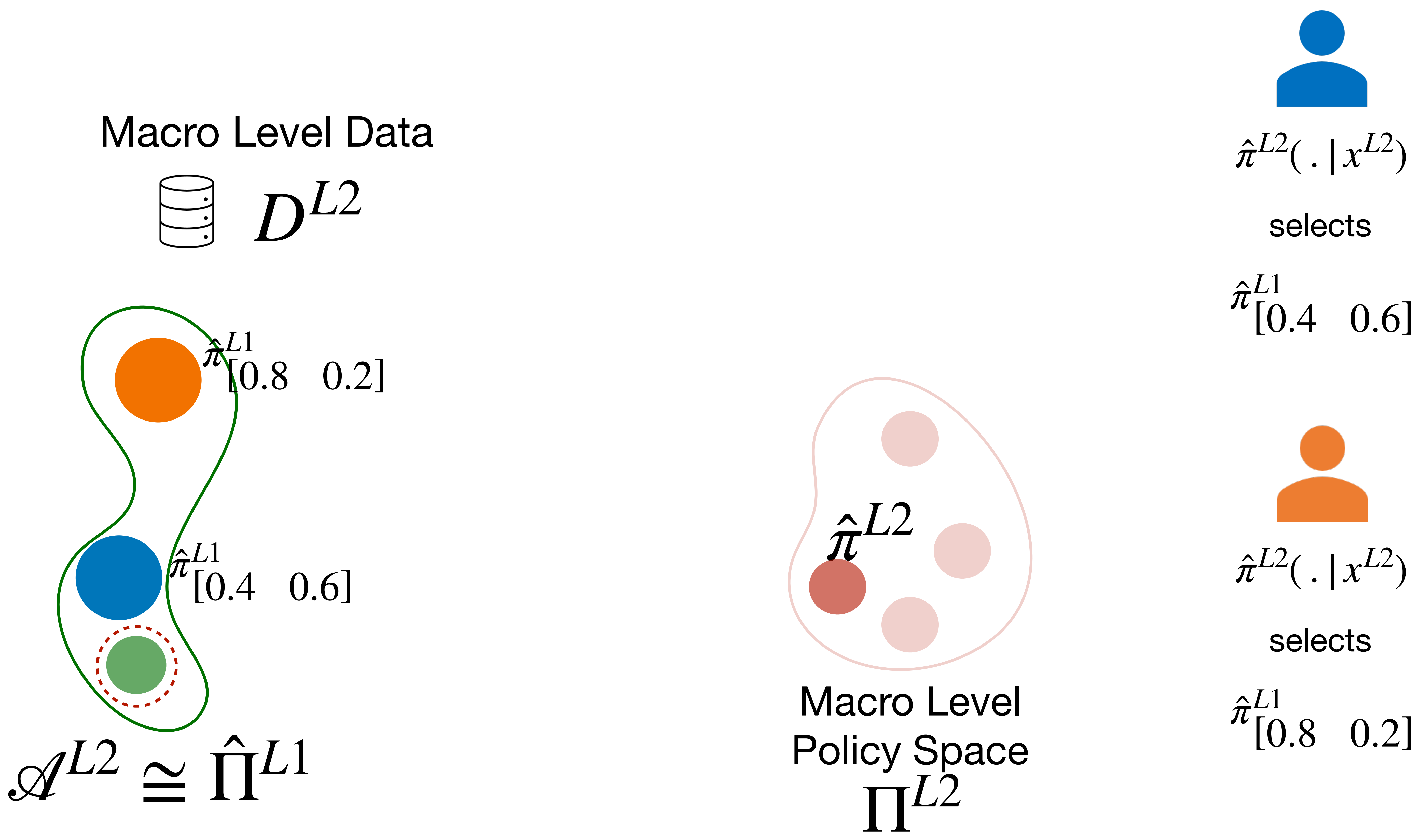
$$[0.8 \quad 0.2] \begin{bmatrix} \text{Clicks} \\ \text{Likes} \end{bmatrix}$$

$$[0.4 \quad 0.6] \begin{bmatrix} \text{Clicks} \\ \text{Likes} \end{bmatrix}$$

$$\cdots$$

$$\mathscr{A}^{L2} \cong \hat{\Pi}^{L1} = \{\hat{\pi}^{L1}_{[0.8 \quad 0.2]}, \hat{\pi}^{L1}_{[0.4 \quad 0.6]}, \cdots\}$$

Policy Learning at macro level



MultiScale Contextual Bandits Algorithm

Algorithm 1 MultiScale Training: Off-Policy Contextual Bandits

Procedure *PolicyLearning*(π_0^{L2}, π_0^{L1})

Collect Micro Logged dataset $D^{L1} := \{(x_i^{L1}, a_i^{L1}, r_i^{L1}, p_i^{L1})\}_{i=1}^{n^{L1}} \sim \pi_0^{L1}$

Learn Micro policies $\hat{\Pi}^{L1}$(Eq. ($\boxed{5}$) or ($\boxed{6}$) using D^{L1})

MultiScale Contextual Bandits Algorithm

Algorithm 1 MultiScale Training: Off-Policy Contextual Bandits

Procedure *PolicyLearning*(π_0^{L2}, π_0^{L1})

Collect Micro Logged dataset $D^{L1} := \{(x_i^{L1}, a_i^{L1}, r_i^{L1}, p_i^{L1})\}_{i=1}^{n^{L1}} \sim \pi_0^{L1}$

Learn Micro policies $\hat{\Pi}^{L1}$(Eq. ($\boxed{5}$) or ($\boxed{6}$) using D^{L1})

Collect Macro Logged dataset $D^{L2} := \{(x_j^{L2}, a_j^{L2}, r_j^{L2}, p_j^{L2})\}_{j=1}^{n^{L2}} \sim \pi_0^{L2}$

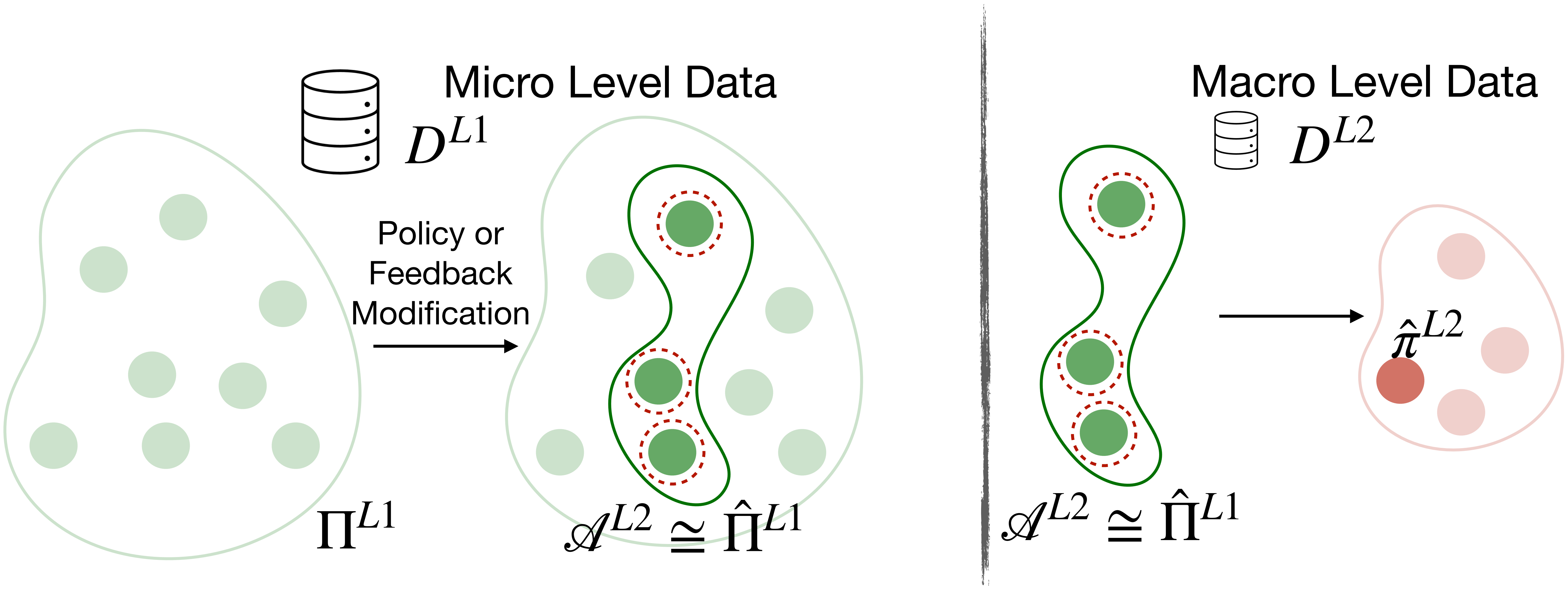
Learn Macro Policy $\hat{\pi}^{L2} \leftarrow \arg\max_{\pi^{L2}} \hat{V}^{L2}(\pi^{L2}; D^{L2})$ (Eq. ($\boxed{7}$))

return learned policies $\hat{\pi}^{L2}, \hat{\Pi}^{L1}$

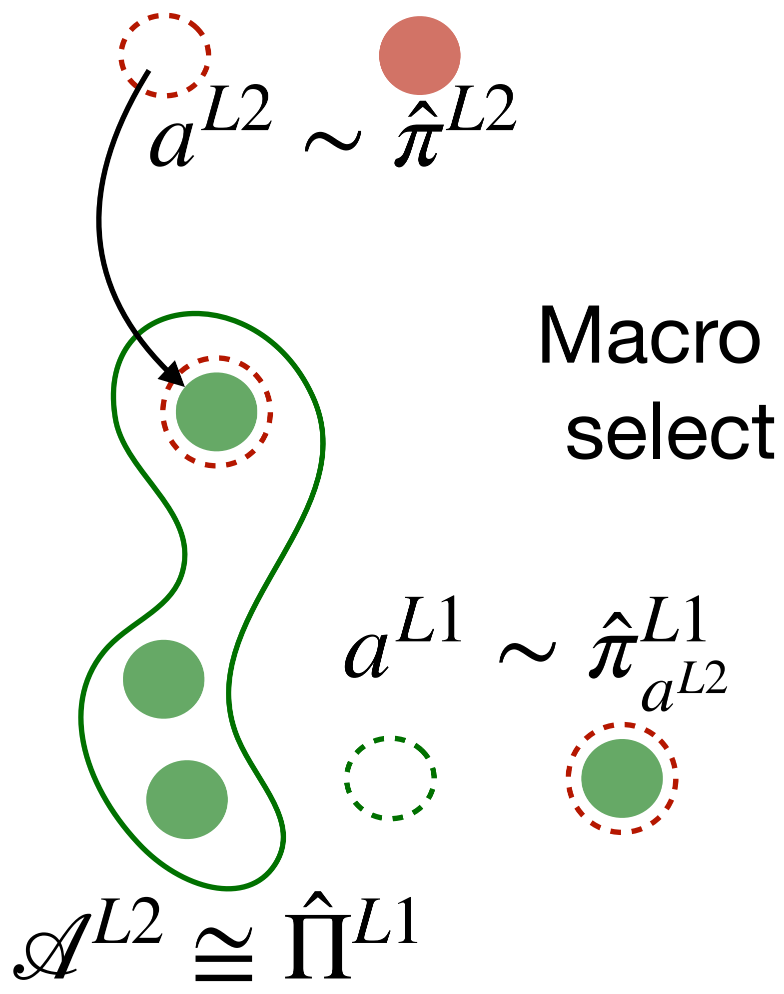
This procedure can be recursively called for extending to arbitrary number of levels

MultiScale Contextual Bandits

- Training
    - Bottom up



- Inference
    - Top down



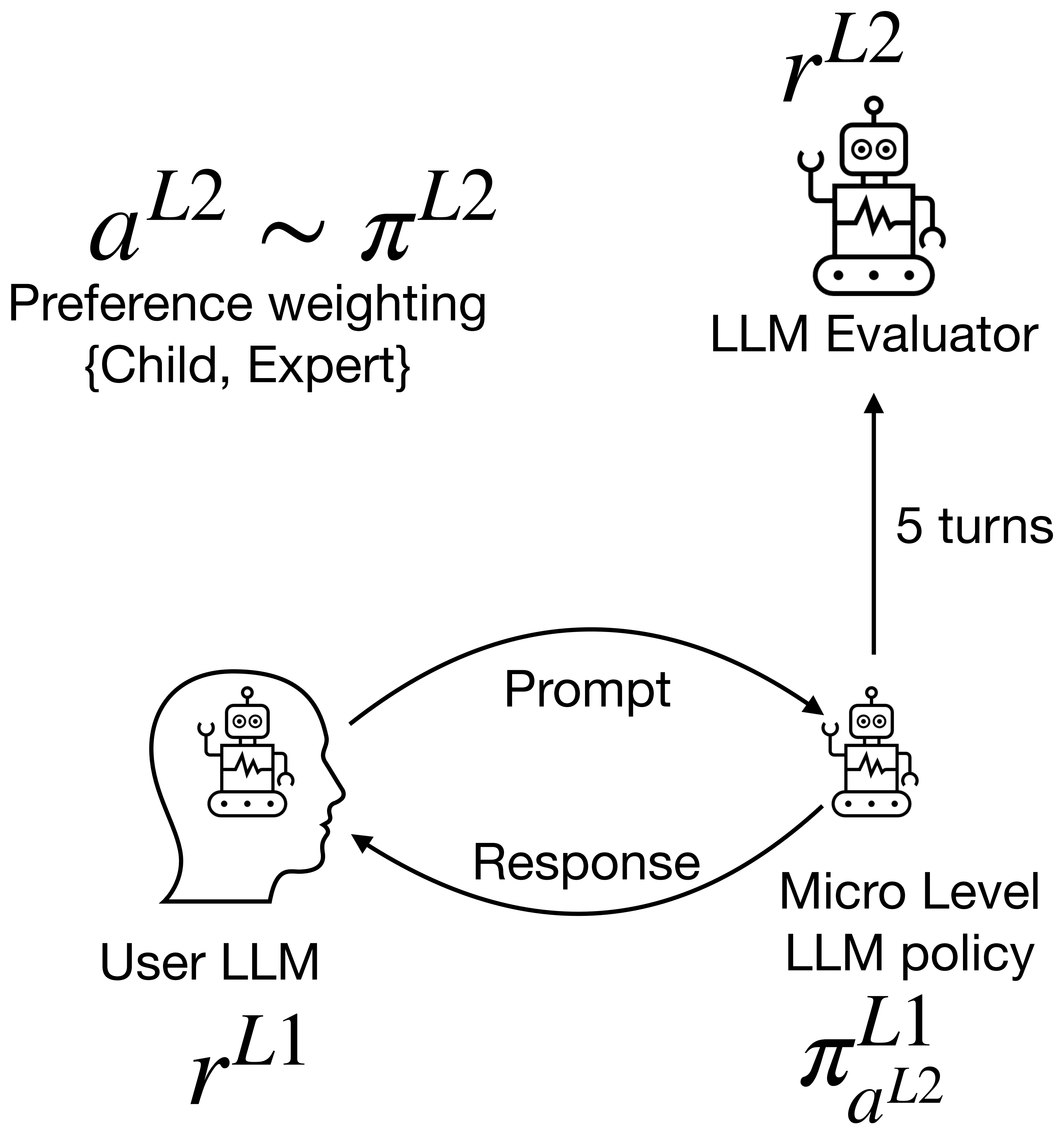
Macro action indexes to select the micro policy

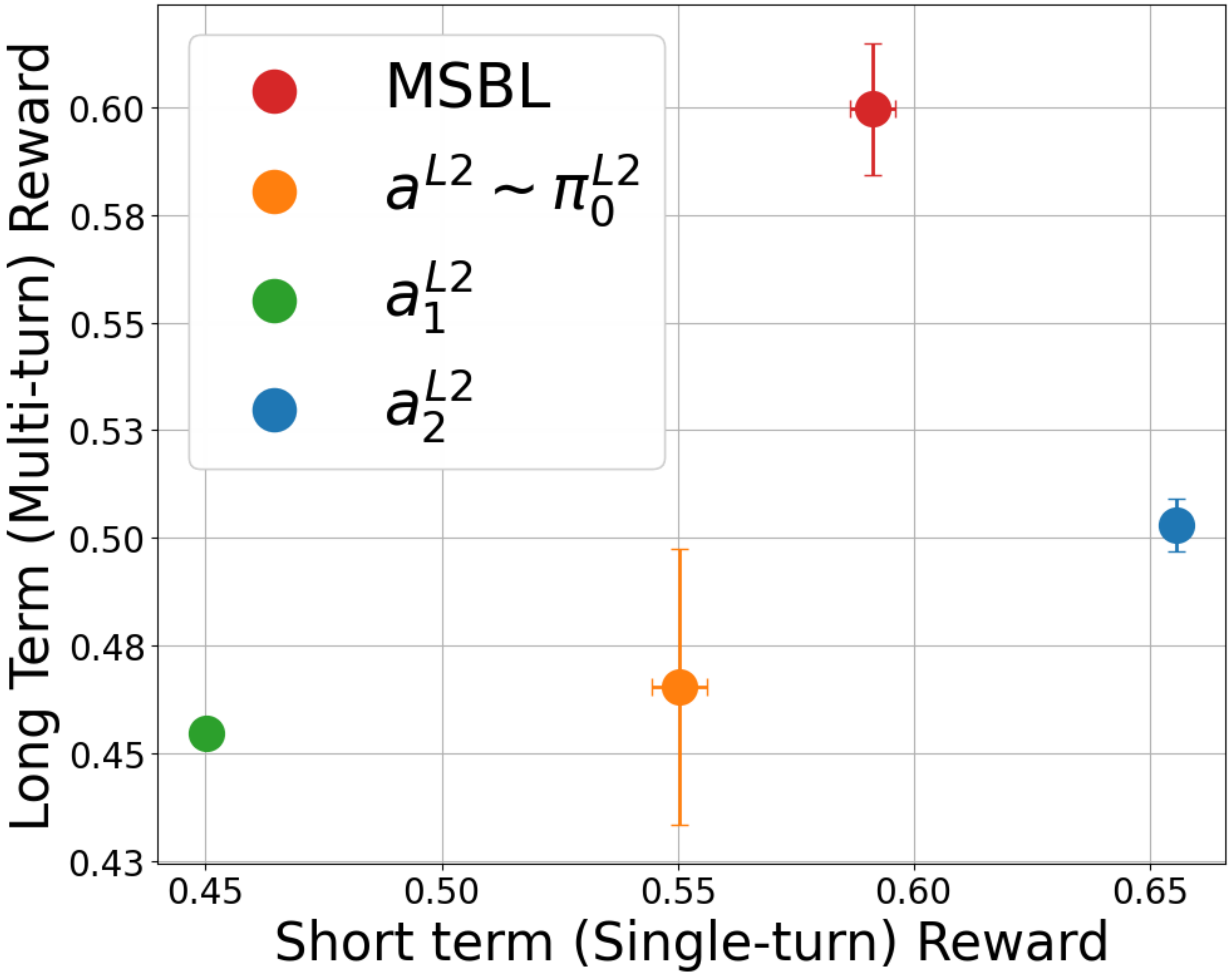
Experiments

- Multi turn Conversation
- Conversational recommender system
- Large Scale Recommender System

Experiments

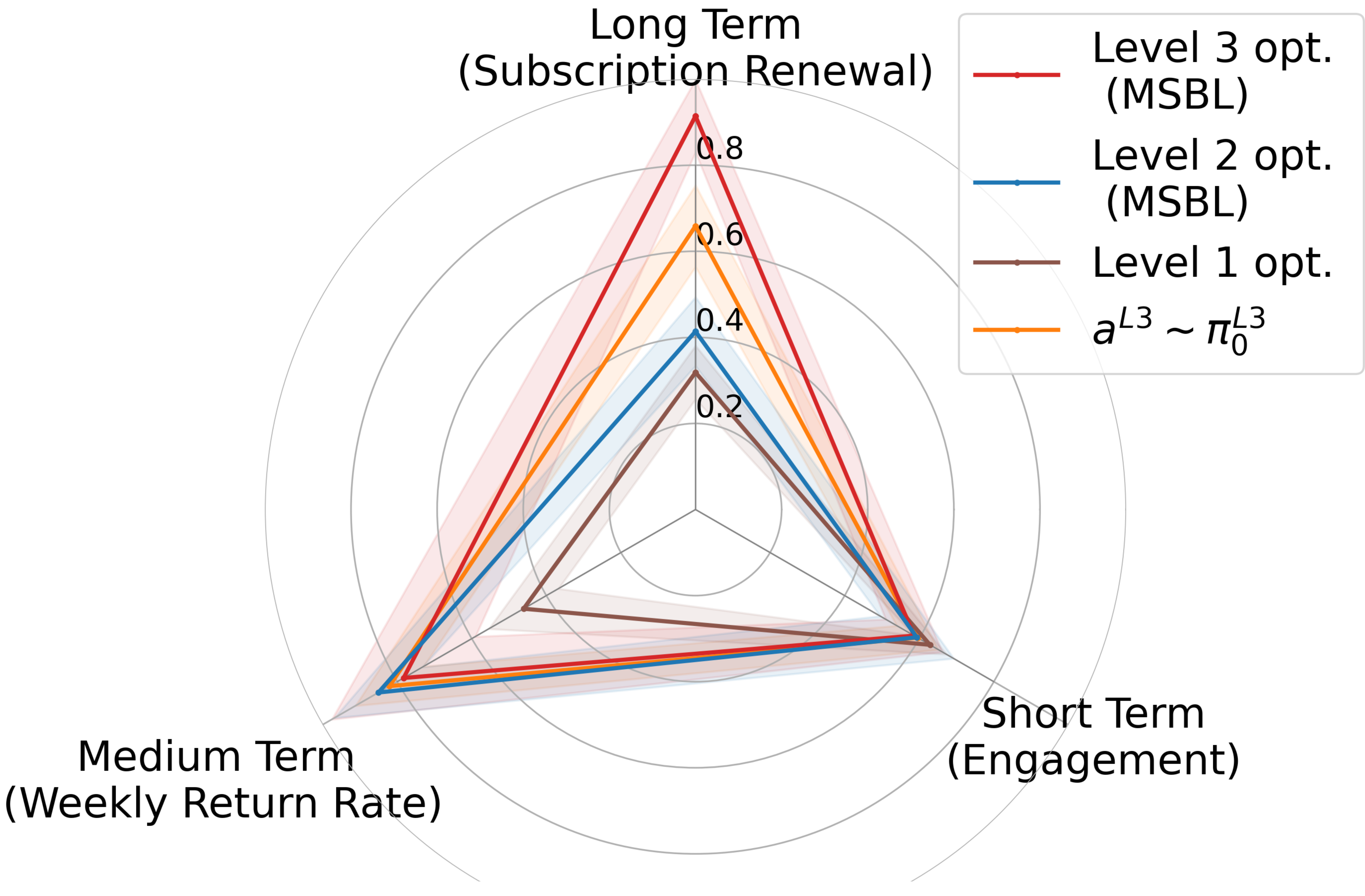
Multi turn Conversation





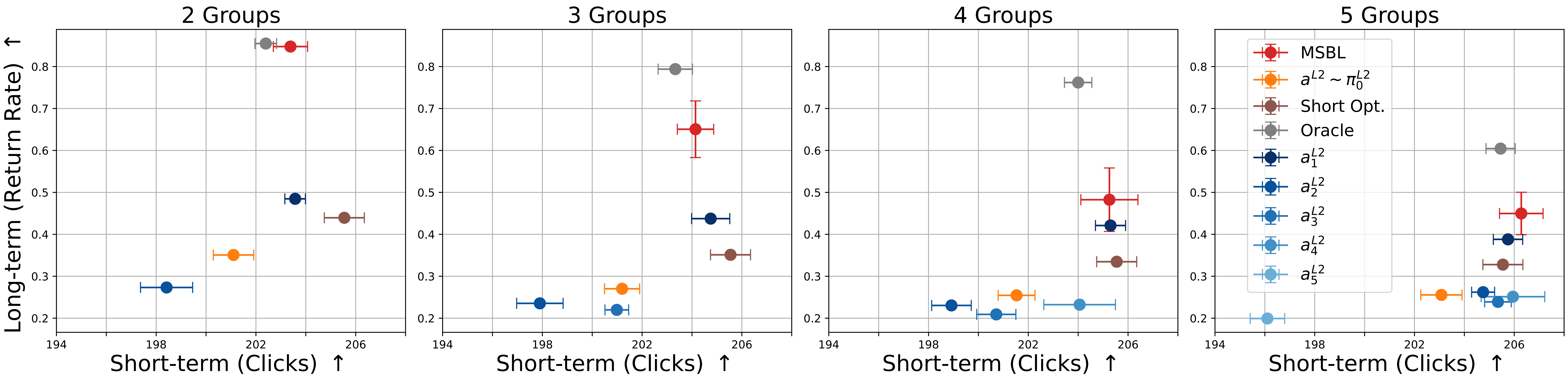
Experiments

Conversational Recommender System



Experiments

Large Scale Recommender System



Summary

- Introduce a principled framework to optimize for long term objectives.
- Motivated by using plentiful short-term data for faster learning with scarcer long term feedback
- We discuss two ways - policy and feedback modification to learn a family of policies at micro level.
- Propose a practical bandit algorithm for recursively learning policies at multiple interdependent levels.
- Checkout the paper for more results and analysis - PAC Bayesian motivation, updating micro and macro policies after deployment when new data is available, scaling action space and more !