# Neural Dueling Bandits: Preference-Based Optimization with Human Feedback

**Arun Verma[*1], Zhongxiang Dai[*2], Xiaoqiang Lin[3], Patrick Jaillet[4], Bryan Kian Hsiang Low[13]**

*[*]Equal contribution, [1]Singapore-MIT Alliance for Research and Technology, [2]The Chinese University of Hong Kong, Shenzhen, [3]National University of Singapore [4]MIT*

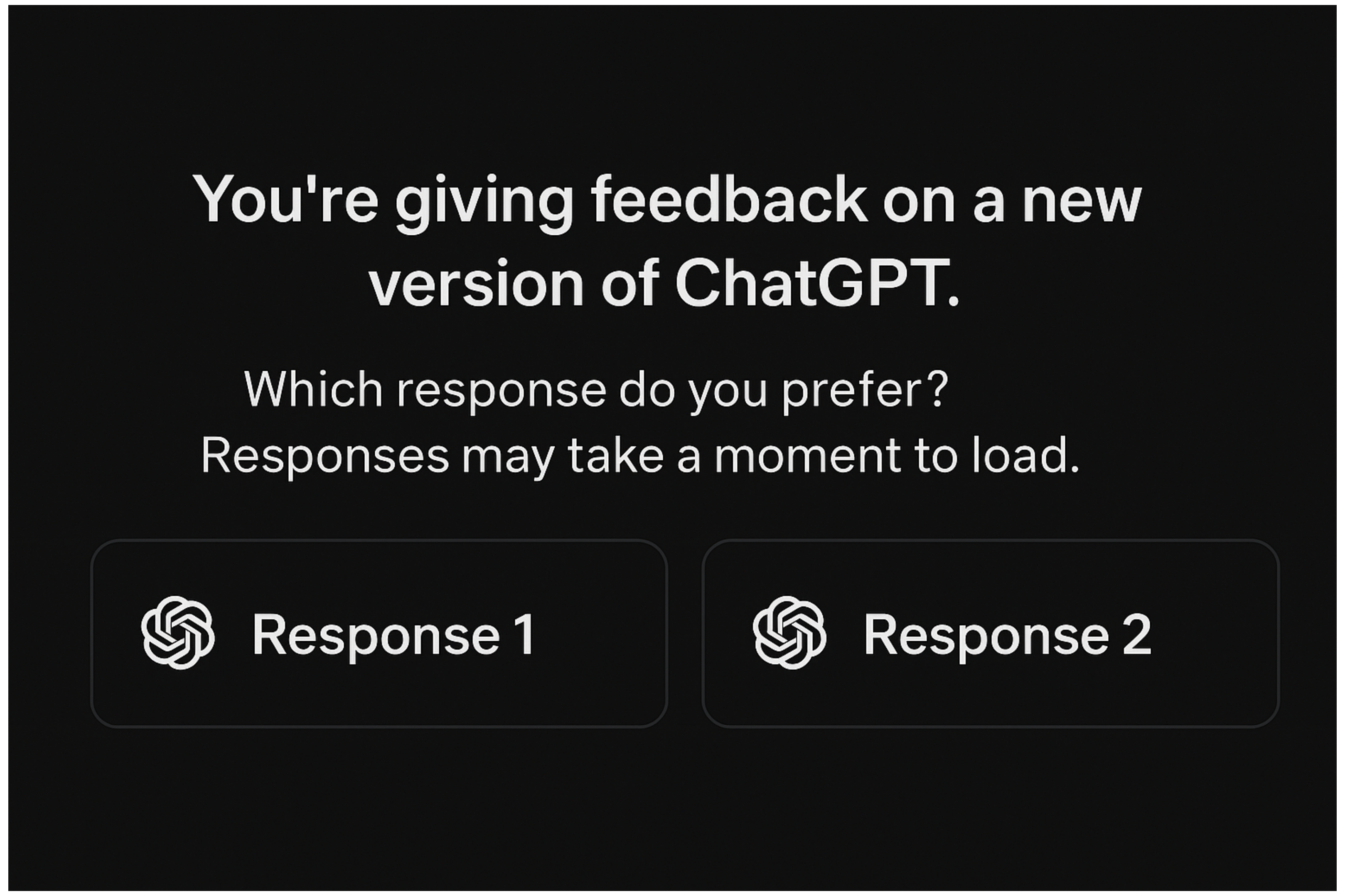
NUS National University of Singapore

SMART Singapore-MIT Alliance for Research and Technology

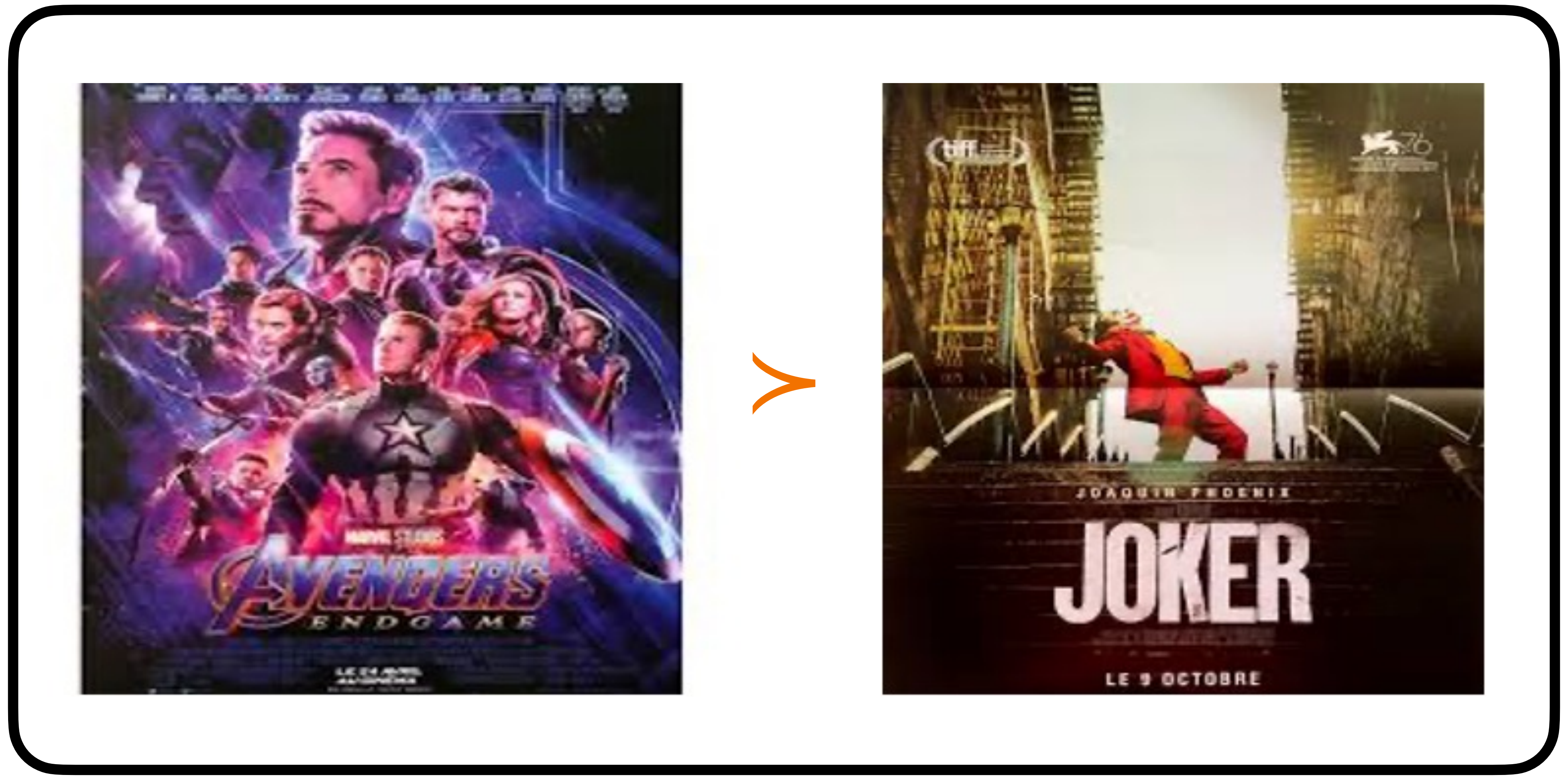
GLOW.AI

ICLR

# Motivation




*ChatGPT*



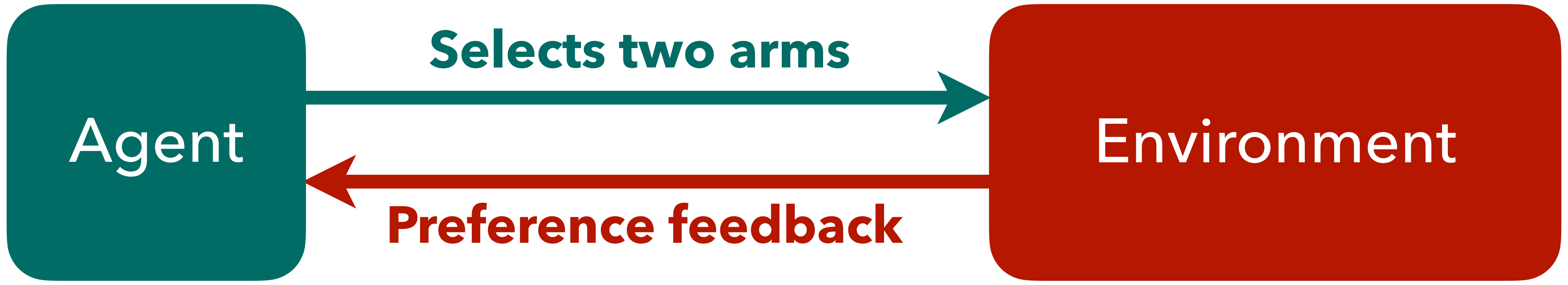

*Asking the user: Which movie they like?*

**Similar problems:** Online recommendation, ranking web search, rating two restaurants or movies, online LLM alignment, LLM response optimization, and many more.

*How to efficiently learn from human preference feedback in an online setting?*

# Contextual Dueling Bandits

- In each round, an agent (or decision-maker) first selects the two arms (actions) for a given context.




- Then, the environment returns a stochastic preference feedback (i.e., one arm is preferred over another).

- **Goal:** Find the best arm for a given context using observed preference feedback that minimizes the cumulative (average) regret for $T$ rounds is defined as:

$$\mathscr{R}_T = \sum_{t=1}^{T} \left( f(x_t^\star) - \frac{\big(f(x_{t,1}) + f(x_{t,2})\big)}{2} \right).$$

# Neural Dueling Bandits (NDB)

- **Assumption:** Preference feedback between two arms follows the Bradley-Terry-Luce (BTL) model.

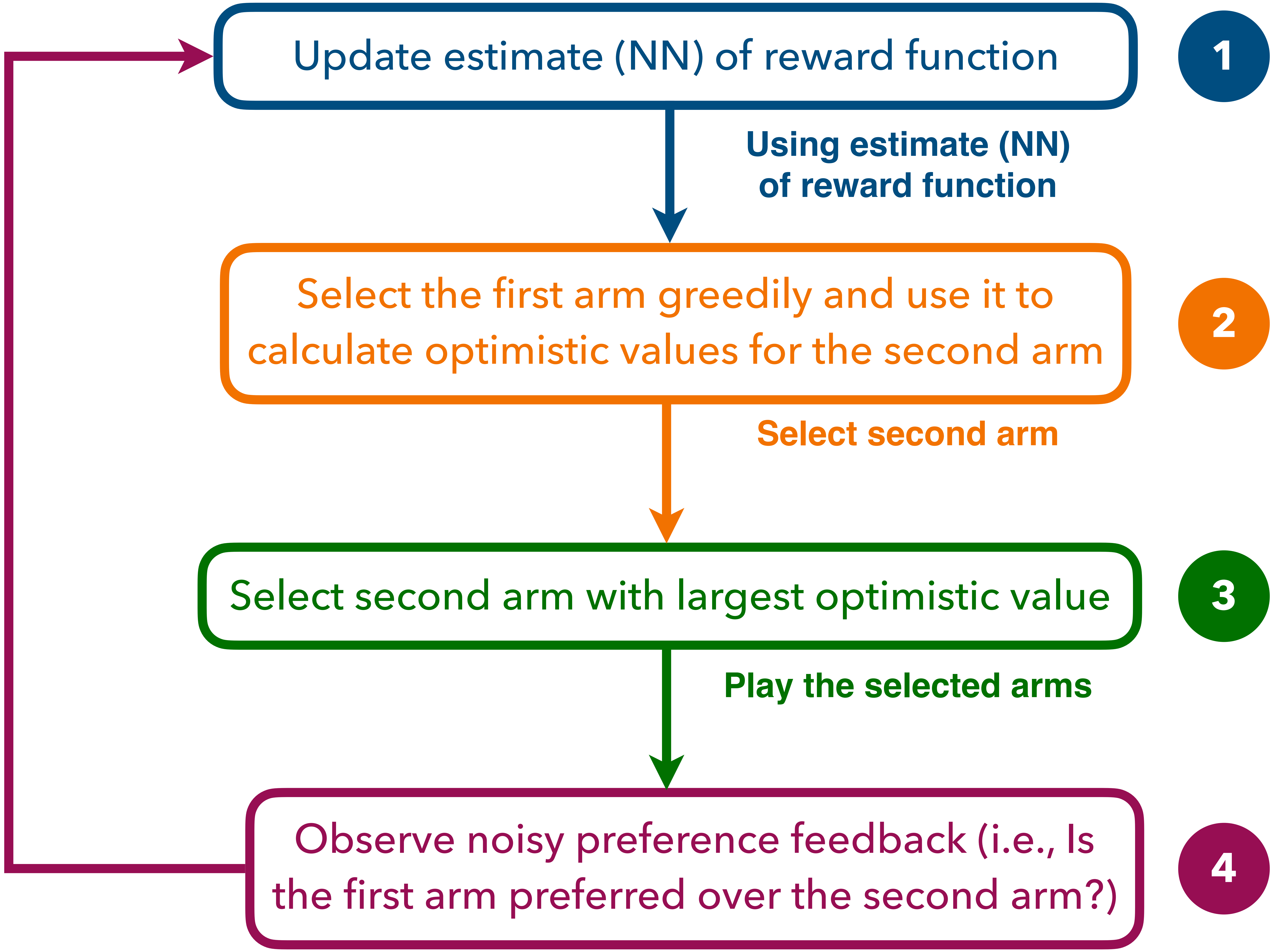
$$\mathbb{P}\{x_{t,1} \succ x_{t,2}\} = \frac{\exp\left(f(x_{t,1})\right)}{\exp\left(f(x_{t,1})\right) + \exp\left(f(x_{t,2})\right)},$$

  where $f$ is the latent non-linear reward function.

- Our algorithm uses a **neural network** (NN) to estimate the unknown latent reward function and then

- Selects the first arm greedily and chooses the second arm that maximizes the optimistic values (using trained NN and selected first arm with **upper confidence bound** (**UCB**) or **Thomson sampling** (**TS**)) to balance exploration and exploitation.

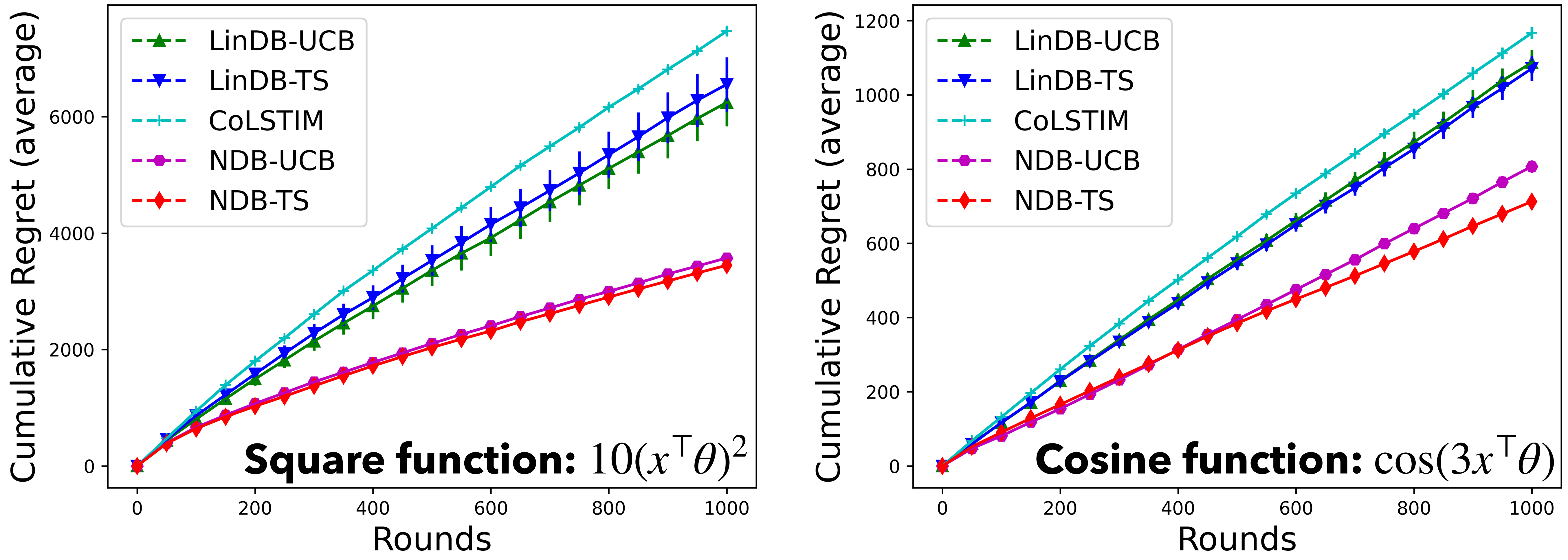
# Algorithm for NDB



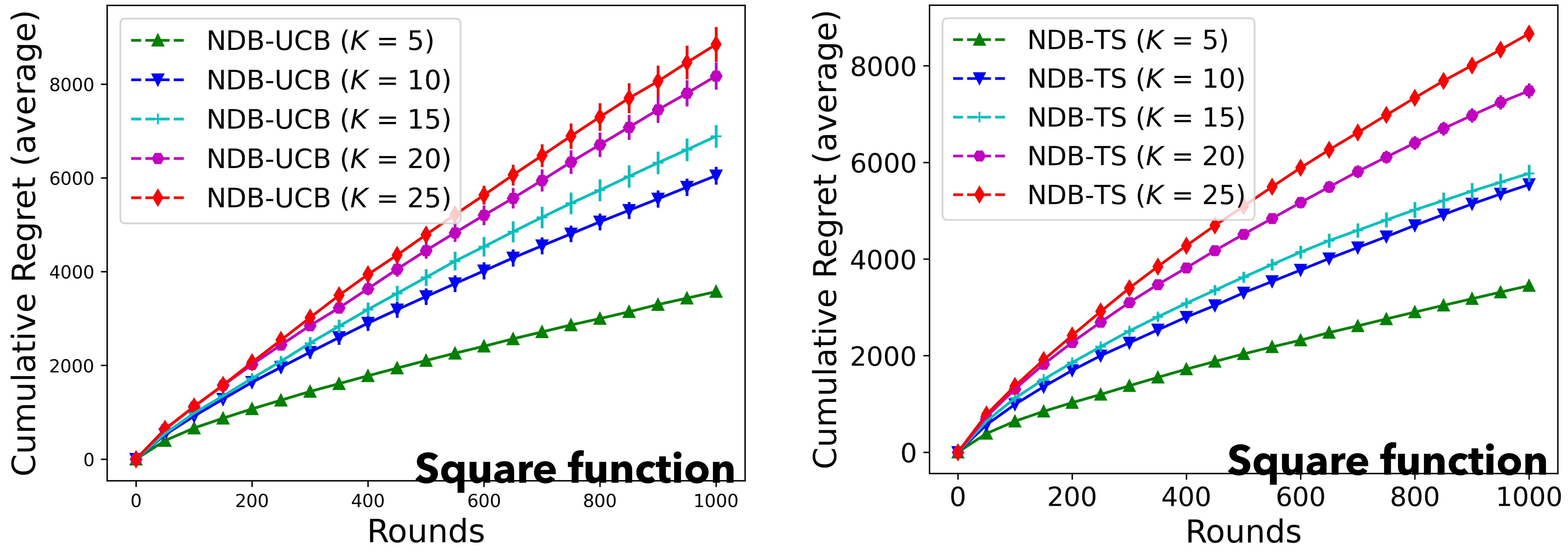
# Theoretical Results

- Novel confidence ellipsoid bounds tailored to NDB.
- Our UCB- and TS-based algorithms using NN have cumulative (average and weak) regret of $\tilde{O}\left(\tilde{d}\sqrt{T}\right)$ for $T$ rounds, where $\tilde{d}$ is the effective dimension.
- Also, drive sub-linear cumulative regret upper bounds for neural contextual bandit problems with binary feedback.

# Experimental Results

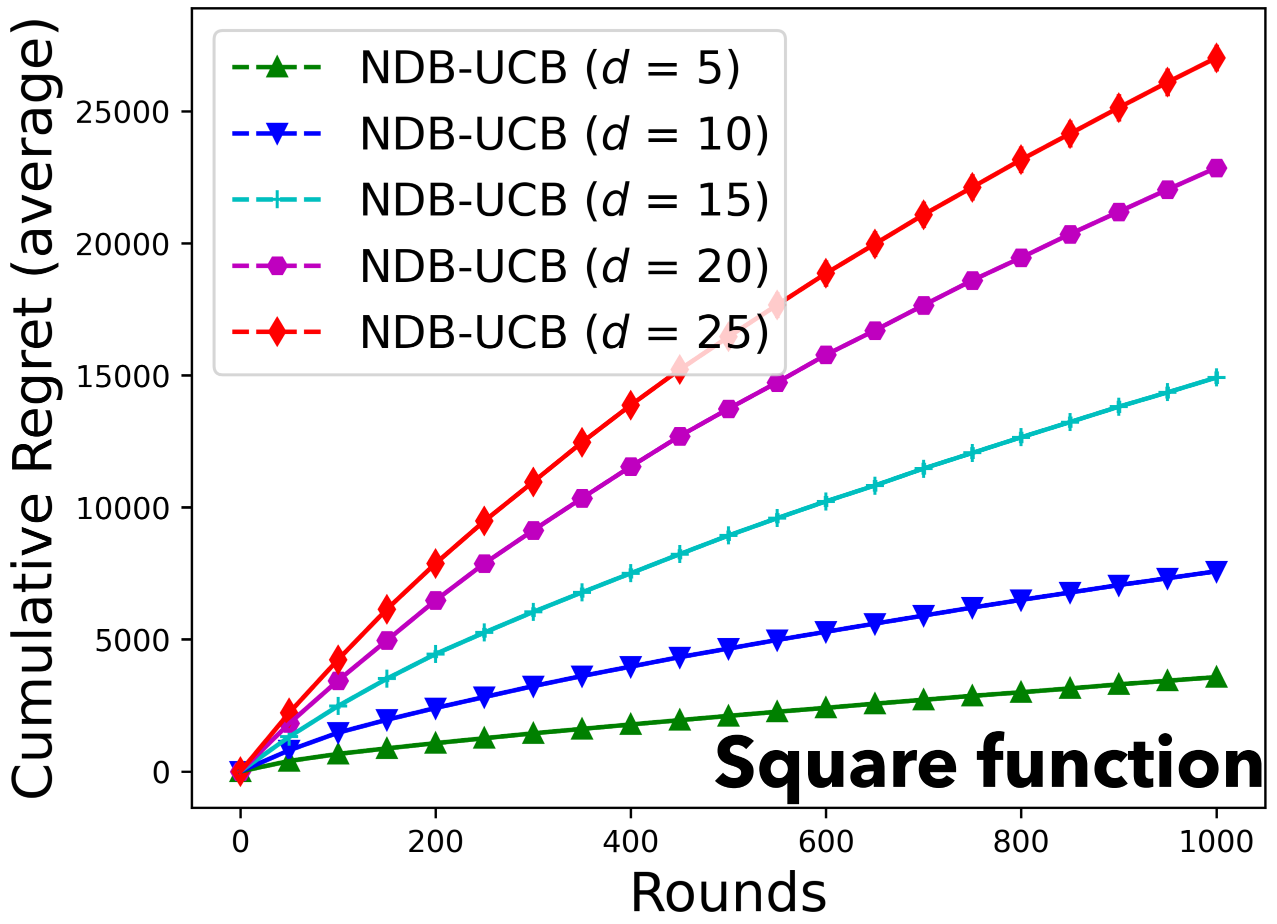


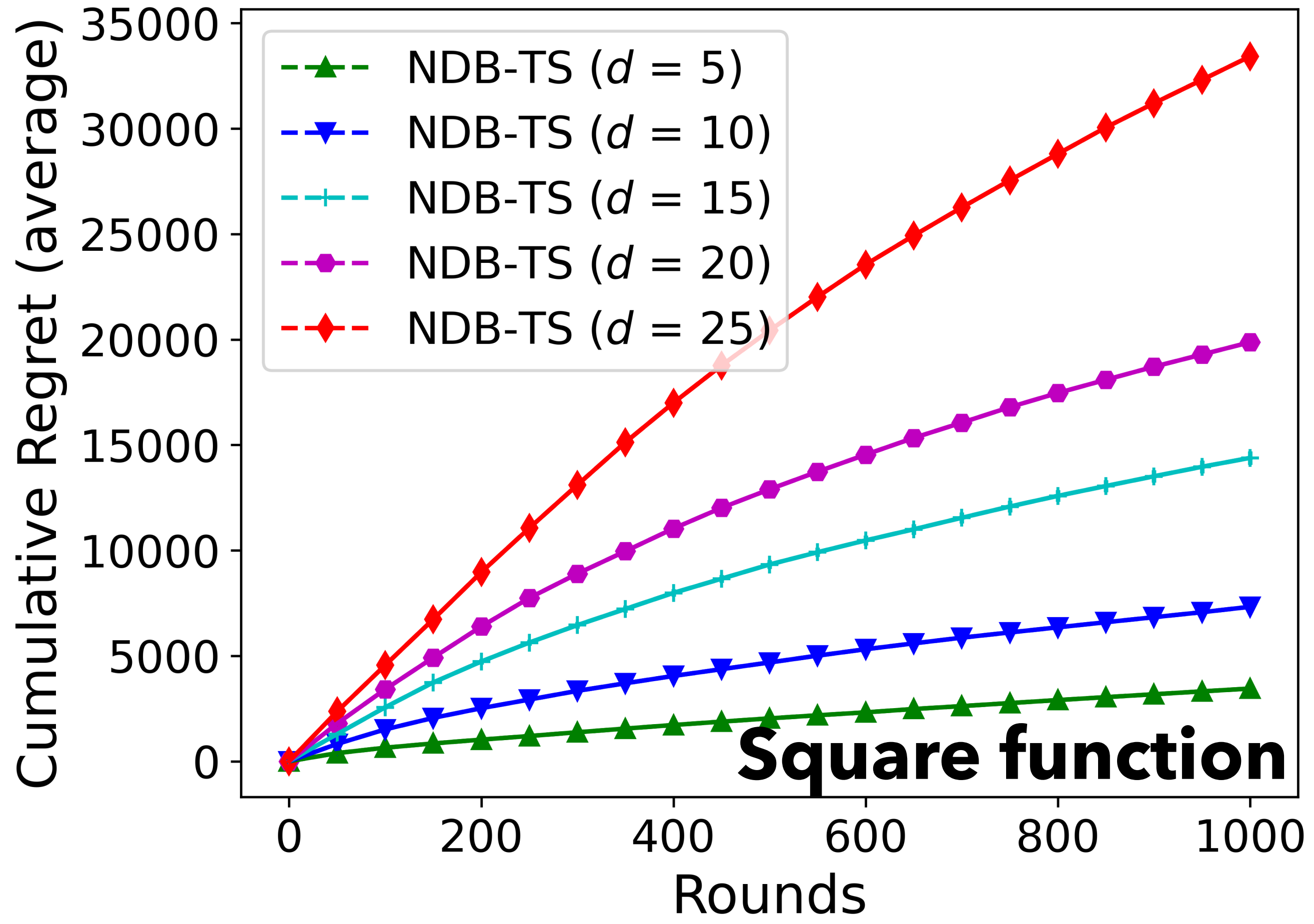

**Comparisons of cumulative (average) regret of dueling bandits algorithms.**



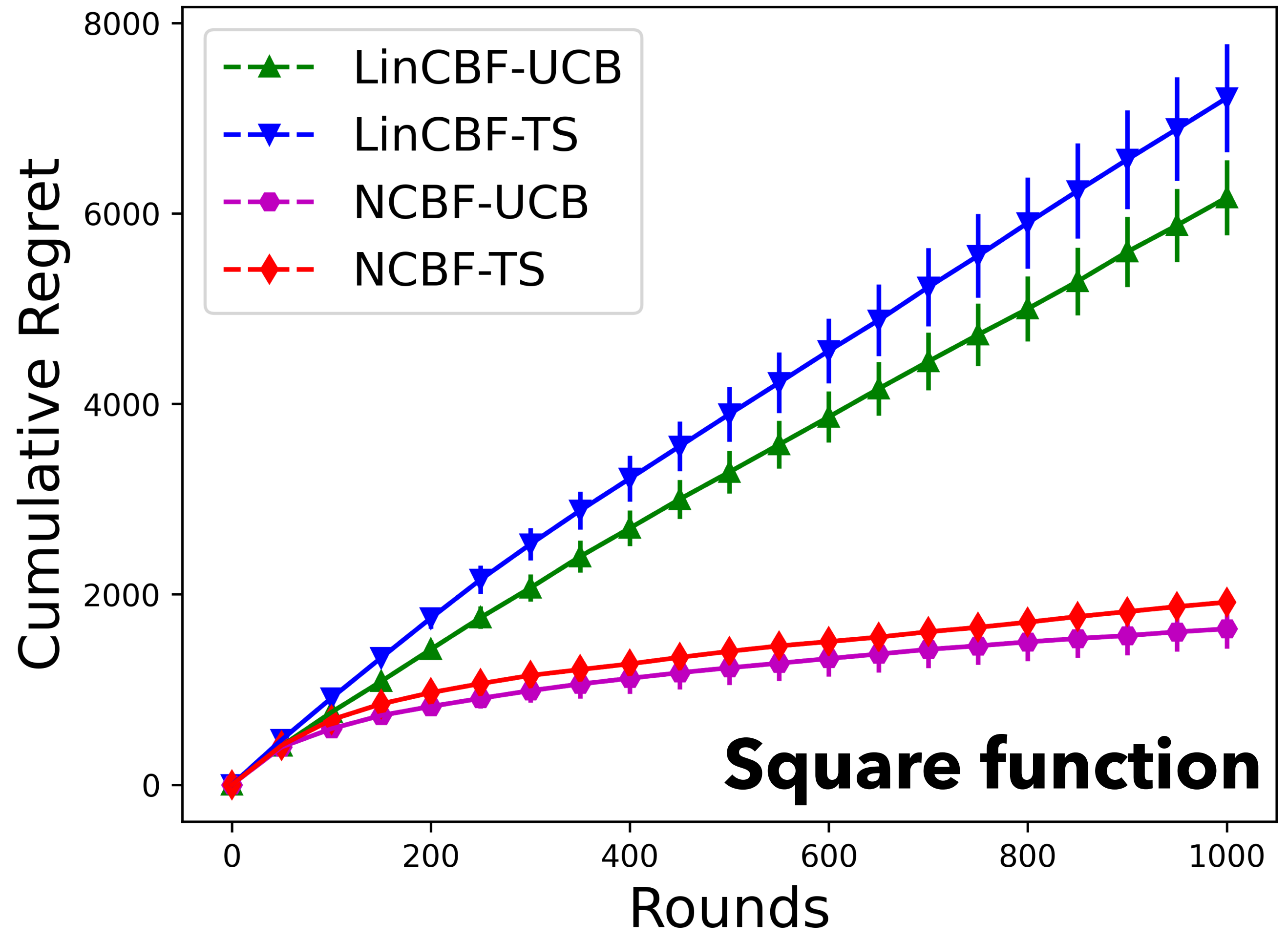

**Our proposed algorithms vs. different numbers of arms $K$.**

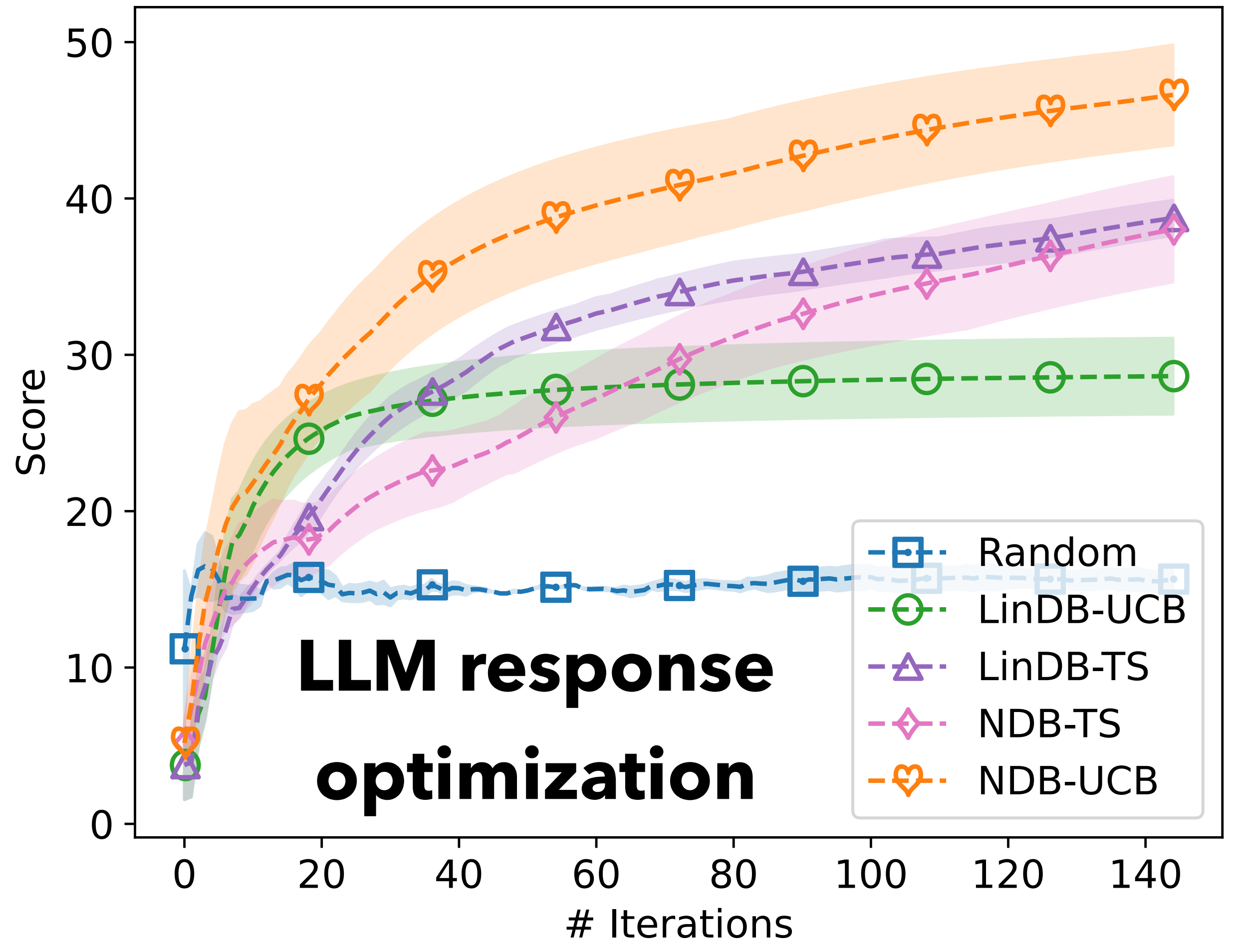
# Experimental Results




**Our proposed algorithms vs. dimension of the context-arm feature vector $d$.**




**Left: Comparisons of algorithms for contextual bandits with binary feedback.**

**Right: Scores of different algorithms for LLM response optimization.**