Non-stationary Causal Bandits

Reda Alami
Technology Innovation Institute
9639 Masdar City
Abu Dhabi, United Arab Emirates
reda.alami@tii.ae

Abstract

The causal bandit problem is an extension of the conventional multi-armed bandit problem in which the arms available are not independent of each other, but rather are correlated within themselves in a Bayesian graph. This extension is more natural, since day-to-day cases of bandits often have a causal relation between their actions and hence are better represented as a causal bandit problem. Moreover, the class of conventional multi-armed bandits lies within that of causal bandits, since any instance of the former can be modeled in the latter setting by using a Bayesian graph with all independent variables. However, it is generally assumed that the probabilistic distributions in the Bayesian graph are stationary. In this paper, we design non-stationary causal bandit algorithms by equipping the actual state of the art (mainly CAUSAL UCB, CAUSAL THOMPSON SAMPLING, CAUSAL KL UCB and ONLINE CAUSAL TS) with the restarted Bayesian online change-point detector [2]. Experimental results show the minimization of the regret when using optimal change-point detection.

1 Introduction

The multi-armed bandit is a stochastic scheduling problem wherein a user can choose amongst an independent set of arms (or actions), and the reward that they will get is stochastically dependent on the arm pulled. This problem covers many real life instances (for eg. which advertisement will have a higher click rate) and derives its name from the one-arm slot machines in casinos. Albeit its simple nature and relaxing assumptions, the multi-armed bandit can model a lot of problems making it widely prevalent.

However, there are several scenarios in which multi-armed bandits cannot be applied due to contradiction(s) with one or more of its fundamental assumptions. A very common example, especially in the context of our country, is related to the cropping season. Consider a young farmer who just started farming this season, naturally with the objective of procuring maximum yield. He is aware of various factors affecting the same, say quality of seeds, organic manure, pesticides and fertilizers, but can afford to spend on only one due to a fixed budget; other factors are out of his reach and obtained (randomly) via government schemes. Only at the end of the season can he estimate the quantity of each 'out-of-reach' resource consumed and the resulting yield, and plan his approach/expenditure for the next season. Another thing to be noted here is that the decisions of the farmer on spending his budget on a resource are not independent - if the farmer is already getting good quality of seeds (which are, say, robust to diseases) and decides not to spend on them, it is intuitive that he won't have to spend on anti-disease fungicides/pesticides either.

Clearly, a multi-armed bandit is not the best formulation for this problem since the actions are not independent. Instead, such instances are better modeled by causal bandits in which we assume that a Bayesian graph governs the dependencies amongst various inputs and other hidden factors, with the final 'reward' also being a node in the graph. Various algorithms have already been proposed

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

for causal bandits that try to model the Bayesian graph structure and distribution, and consequently find the best regret-minimizing action [4]. However, the algorithms like CAUSAL UCB [6], CAUSAL THOMPSON SAMPLING [6] and CAUSAL KL UCB [5] assume that the distributions carried out over the graph are stationary.

Our objective in this paper is to design new algorithms to handle non-stationary environments. We propose to equip the algorithms named CAUSAL UCB, CAUSAL THOMPSON SAMPLING, CAUSAL KL UCB and ONLINE CAUSAL THOMPSON SAMPLING with the restarted Bayesian online change-point detector in order to handle abrupt changes of the probabilistic distributions. By doing that, the resulting strategies named RESTARTED CAUSAL UCB, RESTARTED CAUSAL THOMPSON SAMPLING, RESTARTED CAUSAL KL UCB and RESTARTED ONLINE CAUSAL THOMPSON SAMPLING are able to adapt to the changes of the environment.

2 Problem Statement

A causal bandit instance can be described as follows. We are given a set of random variables $\mathcal{X} = \{X_1, X_2 \ldots X_N\}$, a reward variable Y and a set of allowed actions $\mathcal{A}$. The dependencies between $\mathcal{X}$ and Y are represented using a causal graph $\mathcal{G}$, and the amount of information we have beforehand on $\mathcal{G}$ can be varied. Each action $a_t \in \mathcal{A}$ is an intervention of the form $do\left(\mathbf{X}_t = \mathbf{x}_t\ \right)$ (where $\mathbf{X}_t = \left\{X_{j_1}, X_{j_2} \ldots X_{j_m}\right\}, X_{j_i} \in \mathcal{X} \forall i \in [m]$) that involves assigning X_{j_i} the corresponding value in $\mathbf{x}_t$ and removing all incoming edges for X_{j_i} in $\mathcal{G}$ to obtain a mutilated graph $\mathcal{G}_{a_t}$ (the empty intervention, $do()$, is also permitted). After the intervention has been performed, values of the non-intervened variables $\mathcal{X}_c = \mathcal{X} \backslash \mathbf{X}_t$ and the reward y_t are sampled from the distribution of $\mathcal{G}_{a_t}$. This process is repeated up to a fixed horizon T, and the objective is to choose $\{a_t\}_{t=1}^{T}$ such that the cumulative (or simple) regret is minimised.

For the purpose of this paper, we restrict ourselves to those causal bandit instances where the structure of $\mathcal{G}$ is known beforehand but the joint probability distribution (i.e. $\mathbb{P}(\mathcal{X} \cup \{Y\}\ )$ is unknown. Moreover, each $X \in \mathcal{X}$ and Y is a $\{0,1\}$-valued random variable, and the set of allowed actions $\mathcal{A}$ is the set of all possible size-1 interventions (i.e. $|\mathbf{X}_t| = |\mathbf{x}_t| = 1$, and hence $|\mathcal{A}| = 2N$). We shall minimise cumulative regret in our algorithms, which is defined as $R_T = \sum_{t=1}^{T} \mu_t^\star - \mathbb{E}\left[\sum_{t=1}^{T} y_t\right]$, where $\mu_t^\star = \mathbb{E}\left[Y = 1 \mid a_t^\star\right]$ and $a_t^\star$ is the optimal action at time t.

Piece-wise stationary processes We assume that the probabilistic distributions that are carried out over the Bayesian graph follow a piece-wise stationary process Indeed, we assume a global switching model that allows synchronous changes to happen, i.e. arm switches occur at the same time. We denote the sequence of break-points up to the horizon T by: $(\tau_1 = 1, \tau_2, ..., \tau_{K_T+1} = T+1)$.

3 Related Work

The causal bandit problem was first formulated in [4], where they assumed the causal graph $\mathcal{G}$ to be completely known beforehand. They suggested algorithms to minimise simple regret, which is defined as $R = \mu^* - \max_{a \in \mathcal{A}} \mu_a$ where $\mu_a = \mathbb{E}[Y \mid a]$, in two different scenarios - (1) when $\mathcal{G}$ was parallel i.e. all non-reward variables were independent of each other, and (2) when $\mathcal{G}$ was arbitrary. [4] used the graph structure by looking at parents of Y and optimising a distribution over $\mathcal{A}$, with the distribution itself changing to minimise the time taken to reach the best action.

[7] extend Thompson Sampling to the causal bandit setting of [4] by breaking the procedure into two parts - given an intervention $a = do(\mathbf{X} = \mathbf{x})$, they first try to estimate a distribution over the possible assignments of parent variables of Y given a, and then they estimate the reward distribution given a particular parent configuration; both distributions are updated after each trial as new samples are observed.

For many real-life causal bandit instances, it makes sense to minimise cumulative regret for a fixed horizon which is not explored in detail in these papers. A comprehensive analysis and comparison of the proposed algorithms with classical multi-armed bandit algorithms is also lacking. Moreover, in most practical scenarios, the structure of the causal graph is known beforehand - we know how variables affect each other as well as the reward it is the exact probability distribution which is unknown. We believe that these algorithms can be modified to exploit this additional information, which is one of the few things we plan to explore in this paper.

4 Agents Implemented - Restarted Causal bandits

4.1 The restarted Bayesian online change-point detector

The authors in [2] have designed a variant of the original Bayesian online change-point detector introduced by [1]. The resulting strategy is named restarted Bayesian online change-point detector RBOCPD. It is a pruning version of the original algorithm reinterpreted from the standpoint of forecasters aggregation and expressed as a restart procedure pruning the useless forecasters.

More formally, for a binary sequence $\big(y_r, ..., y_n\big) \in \{0, 1\}$, the final formulation of the RBOCPD strategy takes the following form:

$$\texttt{RBOCPD\_Restart}(y_r, ..., y_t) = \mathbb{I}\big\{\exists s \in (r, t] : \vartheta_{r,s,t} > \vartheta_{r,r,t}\big\} \tag{1}$$

where the weight of the forecasters $\vartheta_{r,s,t}$ are computed in a recursive way as follows (assuming an initial weight $\vartheta_{r,1,1} = 1$):

$$\vartheta_{r,s,t} = \begin{cases} \frac{\eta_{r,s,t}}{\eta_{r,s,t-1}} \exp\left(-l_{s,t}\right) \vartheta_{r,s,t-1} & \forall s < t, \\ \eta_{r,t,t} \times \mathcal{V}_{r:t} & s = t\,. \end{cases} \tag{2}$$

such that the initial weight of the forecaster takes the form of $\mathcal{V}_{r:t} := \exp\left(-\sum_{s'=r}^{t-1} l_{s',t-1}\right)$ and the instantaneous loss $l_{s,t} := -\log \texttt{Lp}\left(y_t | y_s...y_{t-1}\right)$ is computed based on the Laplace predictor $\texttt{Lp}\left(y_t | y_s...y_{t-1}\right) := \begin{cases} \frac{\sum_{i=s}^{t-1} y_i + 1}{t-s+2} & \text{if } y_t = 1 \\ \frac{\sum_{i=s}^{t-1} (1-y_i) + 1}{t-s+2} & \text{if } y_t = 0 \end{cases}$. The hyper-parameter $\eta_{r,s,t}$ is tuned as a decreasing function in t and depends also on the probability of false alarm δ.

The RBOCPD algorithm is chosen among all the sequential change-point detector algorithms in the state of the art for three main reasons.

- *Well adaptability to unknown priors.* Indeed, the RBOCPD algorithm has been designed to solve the problem of sequential change-point detection in a setting where both the change-points and the distributions before and after the change are assumed to be unknown. This setting corresponds exactly to the situation of an agent facing a multi armed bandit whose distributions are unknown and may change abruptly at some unknown instants.
- *Minimum detection delay.* This corresponds to the first criteria assessing the performance of a sequential change-point detector. The detection delay is defined as the number of samples needed to detect a change. In [2], the authors have shown that the detection delay of the RBOCPD strategy is asymptotically optimal in the sense that it reaches the existing lower bound stated in Theorem 3.1 in [3].
- *Well controlled false alarm rate.* The false alarm rate corresponds to the probability of detecting a change at some instant where there is no change. Again, in [2], the authors have demonstrated that $\forall \delta \in (0, 1)$ RBOCPD doesn't make any false alarm with a probability at least $1 - \delta$.

4.2 Restarted causal Bandit algorithms

In order to resolve a piece-wise stationary causal multi-armed bandit, we propose to equip the classical algorithms (CAUSAL UCB, CAUSAL THOMPSON SAMPLING and CAUSAL KL UCB) with the restarted Bayesian online change-point detector running on each arm $a \in \mathcal{A}$. The resulting strategies are called RESTARTED CAUSAL UCB, RESTARTED CAUSAL THOMPSON SAMPLING, RESTARTED CAUSAL KL UCB and RESTARTED CAUSAL OC-TS. At some time t, the designed strategies re-initialize the parameters related to arm a when the RBOCPD associated to arm a has raised an alarm.

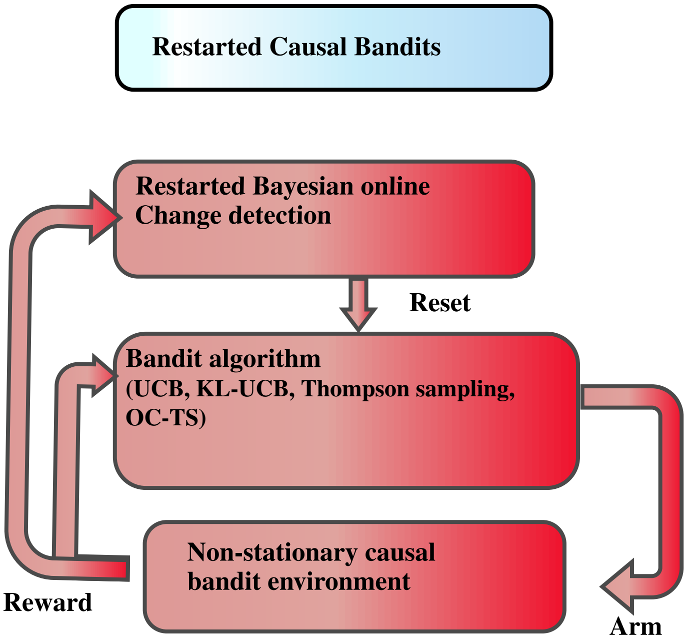
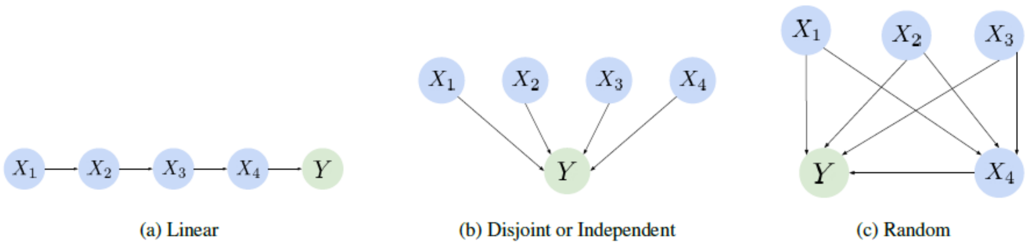

Restarted Causal Bandits
Restarted Bayesian online Change detection
Reset
Bandit algorithm (UCB, KL-UCB, Thompson sampling, OC-TS)
Non-stationary causal bandit environment
Reward
Arm

Figure 1: Caption

5 Experiments

5.1 Graph topologies

We have considered three types of Bayesian graph models (described below) with different topologies representing the causal dependencies amongst Bernoulli variables to show that our proposed modifications/algorithms perform better than existing ones. Figure 1 shows the representative model from each topology chosen for our experiments. - Linear. Every variable excluding one root variable has exactly one parent. The idea here is that the agent must learn that the best action is the one in which the closest possible ancestor is intervened (to either a value of 0 or 1 depending on the distribution). - Disjoint or Independent. Here all variables (except the reward variable) are pairwise independent and are parents of the reward variable. The best action in this case is the one in which the marginalised probability over non-intervened variables is maximum. - Random. It refers to a completely random Bayesian graph obtained by iteratively adding variables and randomly choosing whether each previously added variable is its parent or not.



(a) Linear

(b) Disjoint or Independent

(c) Random

Figure 2: Representative models from each topology.

5.2 Experimental results

We present the results of our experiments here. For each topology, we compare the performance of three designed agents (Restarted causal UCB, restarted causal Thompson Sampling, restarted causal KL UCB and restarted causal OC-TS) by plotting the regret $\sum_{t=1}^{T} \mu_t^\star -$

$\mathbb{E}\left[\sum_{t=1}^{T} y_t\right]$ as a function of horizon. These plots are shown in figures 3, 4 and 5. Each plot is obtained by averaging results over 20 random seeds; we also shade the regions of uncertainty corresponding to a deviation of $\pm\sigma$.

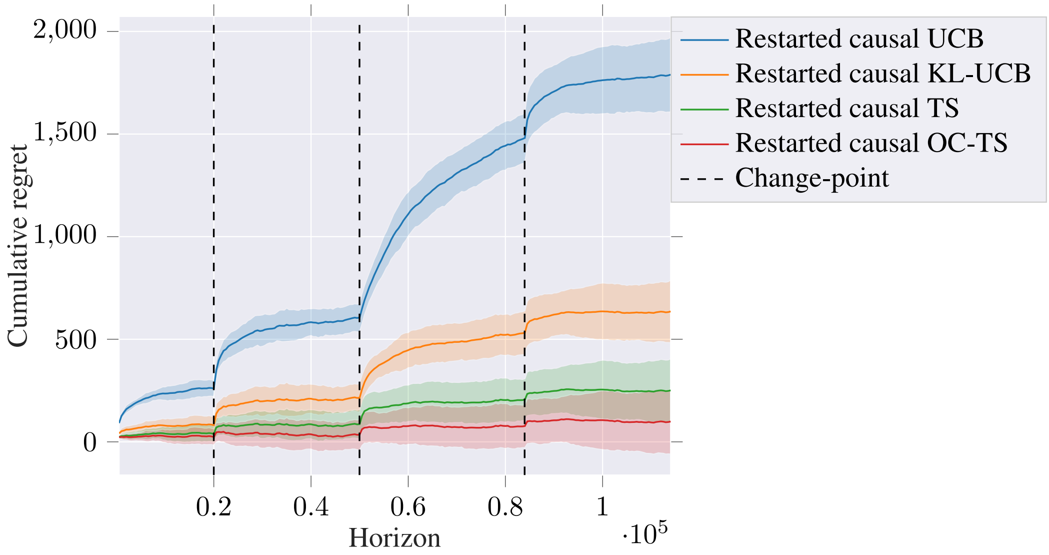


Figure 3: Regrets versus horizon for a linear graph

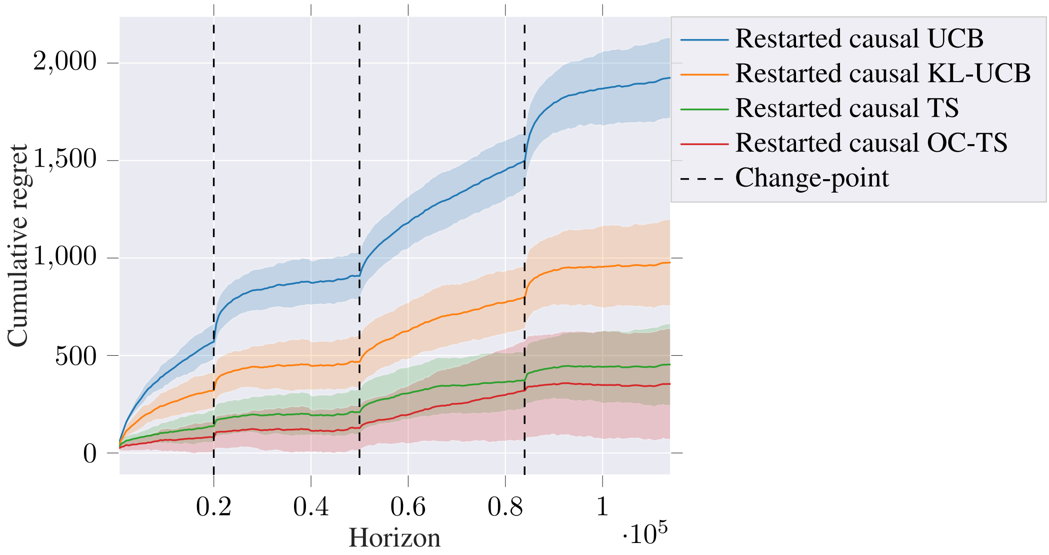


Figure 4: Regrets versus horizon for a disjoint graph

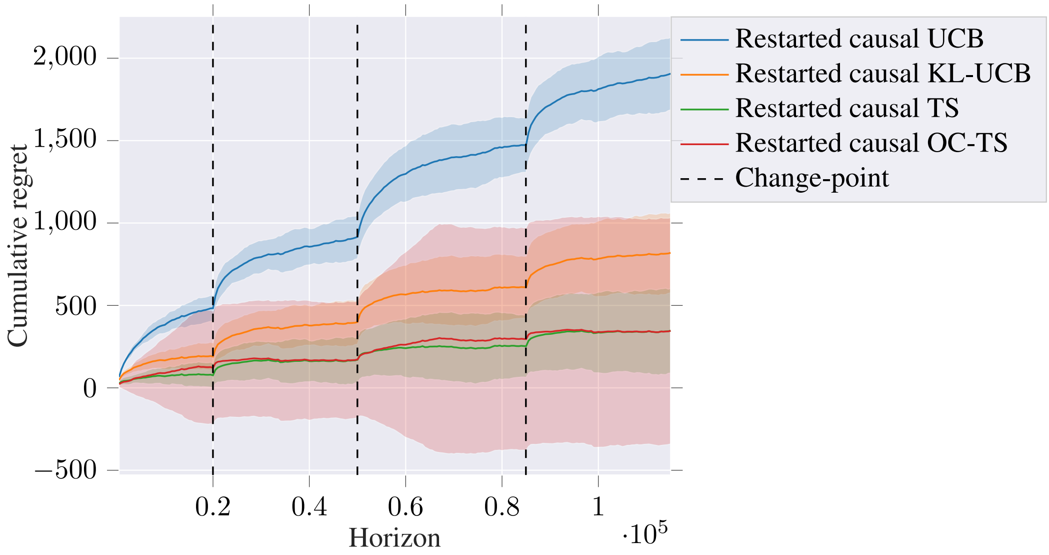


Figure 5: Regrets versus horizon for a random graph

From the figures 3, 4 and 5, the designed algorithms (RESTARTED CAUSAL UCB, RESTARTED CAUSAL THOMPSON SAMPLING, RESTARTED CAUSAL KL UCB and RESTARTED CAUSAL OC-TS) are able to adapt to the changes of the environment. Indeed, the restarted Bayesian online change-point detector is able to quickly detect the change-point thanks to it optimal detection delay and false alarm rate [2].

6 Conclusion

In this paper, we have designed non-stationary causal bandit algorithms for piece-wise stationary environments. The designed algorithms are resulting from a combination between the causal bandit algorithm in the state of the art with the restarted Bayesian online change-point detector. From the experiments, we show that the designed algorithms are able to quickly adapt to the change of the environment. As future work, we are planning to mathematically analyze the designed strategies in order to compute the regret upper bound.

References

[1] Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[2] Reda Alami, Odalric Maillard, and Raphael Feraud. Restarted Bayesian online change-point detector achieves optimal detection delay. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 211–221. PMLR, 13–18 Jul 2020.

[3] Tze Leung Lai and Haipeng Xing. Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential analysis*, 29(2):162–175, 2010.

[4] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*, 29, 2016.

[5] Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? *Advances in Neural Information Processing Systems*, 31, 2018.

[6] Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pages 141–150. PMLR, 2020.

[7] Vin Sachidananda and Emma Brunskill. Online learning for causal bandits. *Advances in Neural Information Processing Systems*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]