Bayesian Agents Learning from Expert Outcomes

Daniel Jarne Ornia[*] University of Oxford

Joel Dyer University of Oxford

Nick Bishop University of Oxford

Ani Calinescu University of Oxford

Michael Wooldridge University of Oxford

Abstract

Complex learning agents are increasingly deployed alongside existing experts, such as human operators or previously trained agents. However, it remains unclear how learners should optimally incorporate certain forms of expert data, or how to quantify the potential benefit from doing so. We study this problem in the context of Bayesian multi-armed bandits, considering *offline settings*, where a learner receives a dataset of outcomes from an expert before interaction, and *online settings*, where posterior updates from different data sources may have different computational costs, so the learner must decide whether a given update should use its own experience or an outcome generated by an expert. We formalize how expert data influences the learner's posterior and quantify how pretraining on expert outcomes tightens information-theoretic regret bounds. We propose an information-directed rule for allocating a limited update budget across data sources, and we study strategies for how the learner can infer when to trust the expert, safeguarding for compromised experts. By disentangling and quantifying the value of expert data, our framework builds towards a practical, information-theoretic understanding of how agents should learn from others.

1 Introduction

Many learning systems are deployed *next to other learners*: an agent learning online may co-exist with other agents that already know how to act well in the same environment. Examples of this include clinical decision support (learning beside clinicians), robotics (learning beside a safe supervisor) and general AI agents (small agents learning beside a powerful, well-tuned large model). Furthermore, in many such settings agents may not have full observability of other agents' variables: Very often only states (or functions of these) are observable, but not actions or decisions. In many interactive robotic problems, one can observe the task outcome (an object placed correctly, success/failure, trajectory points...), but the agents' action representation may not be meaningful or even observable for nearby learners. Motivated by this observation, we study *online Bayesian learning next to an expert* in multi-armed bandits. In this setting, the learner interacts with a multi-armed bandit (with a prior belief over the structure of the bandit) alongside an expert who knows the true structure, and reveals the outcomes they experience, *but not their actions*. We investigate the following fundamental questions. **How should the learner incorporate expert outcomes to improve decision-making?** Intuitively, observing expert outcomes should help infer information about the optimal action distribution: The learner could prune "worlds" that cannot induce that optimal action distribution. We formalize this intuition and show how to use this expert dataset to warm-start the learner's prior, and how this allows the learner to perfectly learn the environment in some instances. **When learning online, which source should the learner pay attention to?** We consider budgeted online settings where the learner can observe its own action-outcome pair and an expert outcome, but performs only one

[*]daniel.jarneornia@cs.ox.ac.uk

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

posterior update per round. This models situations where evaluating a likelihood, querying an external trajectory, storing data, or marginalizing over latent expert actions has non-negligible cost. The learner's problem is therefore not whether additional data can ever help, but which source gives the largest expected reduction in uncertainty per unit of learning effort.

Contributions Our contributions with respect to the problem of Bayesian online learning observing expert outcomes are as follows. (i) We analyse how to incorporate expert data through a consistent Bayesian update (Proposition 3.1), (ii) We quantify how expert data tightens Bayesian regret bounds, with offline (Theorem 3.3) and online (Theorem 4.1), (iii) We extend the analysis to the general case where the expert is imperfect or incorrect. (iv) Finally we demonstrate how our method yields regret improvements in a wide range of bandit experimental scenarios, including a realistic model predictive control task. In summary, we show expert information is most valuable when it moves probability mass between optimal actions; its value can be quantified exactly by the reduction in uncertainty about the optimal action.

1.1 Related Work

Bandits and Beliefs Thompson Sampling [Thompson, 1933] has been a prevalent Bayesian algorithm for online learning for decades [Agrawal and Goyal, 2012, Chapelle and Li, 2011, Russo et al., 2018]. Russo and Van Roy [2016] made the explicit connection between the regret bounds of Thompson Sampling and information theoretic quantities. There are also many examples of multi-agent bandit problems [Brânzei and Peres, 2021, Chang and Lu, 2025] where the question of agent information is introduced. To the best of our knowledge, these works do not consider how to incorporate expert samples in a Bayesian update and how this affects Thompson Sampling regrets. Additionally, our work traces back to works such as Geanakoplos and Polemarchakis [1982], Moses and Nachum [1990], discussing the implications of agents with different belief structures sharing information to learn. Existing work on *opponent modeling* [Carmel and Markovitch, 1995, Yu et al., 2022b, Nashed and Zilberstein, 2022] deals with multi-agent systems where agents model each-other's behaviour, which resonates with our ideas on learning to trust the expert.

Learning from Experts and Demonstrations Our work connects to the broad literature on *bandits with expert advice*, showing tight regret bounds in adversarial or stochastic settings [McMahan and Streeter, 2009, Kale, 2014, Cesa-Bianchi et al., 2025]. Recent work has also extended this to stochastic experts with empirical data [Sharma et al., 2024]. However, these frameworks generally assume the learner observes expert *actions* or advice vectors to minimize regret against the best expert in hindsight. Our setting is fundamentally different: we never observe the expert's actions, but rather the *outcomes* generated by their policy. This distinguishes our work from imitation learning [Abbeel and Ng, 2004, Ross et al., 2011], RL with expert information [Gimelfarb et al., 2018], or offline best-action selection [Agrawal et al., 2023, Yang et al., 2025, Cheung and Lyu, 2024]. We focus on the optimal integration of outcome-only data into the learner's posterior.

Active Learning and Source Reliability Our results on deciding to learn from an expert echo Bayesian experimental design [Lindley, 1956] and Information-Directed Sampling (IDS) [Russo and Van Roy, 2014]. However, where standard IDS chooses an action to pull, our agent makes a meta-decision about which data stream to process. This relates to recent work on modeling source reliability and trust. For instance, Bayesian updates have been employed to track expert trustworthiness or maintain multiple reliability hypotheses in robotics and human-machine fusion [Muesing et al., 2021, Love et al., 2023, Wakayama and Ahmed, 2024]. While these works focus on calibrating trust to fuse inputs or for deferred decision-making, our simultaneous setting frames trust as an *active information acquisition* problem, connecting to Arumugam and Van Roy [2021] on rate distortion for sample selection and recent work on asking experts for help [Plaut et al., 2025a,b].

2 Single-Agent Bandit Problem

Preliminaries We define our problem on a probability space $(\Omega, \mathcal{F}, P)$ with all quantities including the true parameter of the bandit, and the agent's sampled parameters, actions, and outcomes, being random variables on this space. We use lower case $x \in \mathcal{X}$ to indicate items in a set, and upper case X for random variables. $\mathbb{E}[X]$ is the expected value of X, and the entropy of X with probability

mass function $p(X)$ is $\mathbb{H}(X) := -\sum_{x\in\mathcal{X}} p(x)\log p(x) = \mathbb{E}[-\log p(X)]$. The conditional entropy of X given another random variable Y is $\mathbb{H}(X|Y) := \mathbb{E}[-\log p(X|Y)]$, representing the remaining uncertainty in X once Y is known. The mutual information between X and Y is defined as $\mathbb{I}(X;Y) := \mathbb{H}(X) - \mathbb{H}(X|Y)$. It quantifies the reduction in uncertainty about X resulting from observing Y. Throughout the paper, we use a subscript t to denote conditioning on the history of variables up to time t, $\mathcal{H}_t = \{A_s, Y_s\}_{s<t}$. Furthermore, we use $P(X)$ to refer to the probability distribution of X, and $P(X = x)$ to refer to the probability of X taking value x. For instance, the posterior probability of X conditioned by history $\mathcal{H}_t$ is $P_t(X) := P(X \mid \mathcal{H}_t)$. Similarly, the conditional entropy of a random variable X given the history is $\mathbb{H}_t(X) := \mathbb{H}(X \mid \mathcal{H}_t)$, and the conditional mutual information is $\mathbb{I}_t(X;Y) := \mathbb{I}(X;Y \mid \mathcal{H}_t)$. For brevity and readability, we will shorten (conditional) densities as $P(A = a) \equiv P(a)$, and $P(\cdot \mid A = a) \equiv P(\cdot \mid a)$.

Single Agent Bandit An agent chooses actions $a \in \mathcal{A}$ at every time-step $t \in \mathbb{N}$, with $\mathcal{A}$ being a finite set of actions. Each action produces a (possibly random) outcome $Y_{t,a} \in \mathcal{Y}$, and the agent obtains a reward $R(Y_{t,a})$, with $R : \mathcal{Y} \to \mathbb{R}$. The outcomes are drawn from distributions p_a, of which the agents do not have knowledge of. We assume the outcome distribution $p_\theta := (p_{\theta,a})_{a\in\mathcal{A}}$ to be parameterised by some $\theta \in \Theta$ such that for any action, the (mean) reward is a function of θ, $\mu(a,\theta) := \mathbb{E}_{Y\sim p_{\theta,a}}[R(Y)]$. Furthermore, there is a *true* parameter θ^*, possibly sampled from some distribution, that defines the true bandit the agent is in. For some parameter θ, the *optimal action* $a^* \in \mathcal{A}$ is then the action that satisfies $a^*(\theta) \in \arg\max_{a\in\mathcal{A}} \mu(a,\theta)$ (and assume the maximiser is unique). The objective of such agent is to maximize the expected cumulative reward (or equivalently, minimize the expected regret relative to the best action). The regret is defined as $Reg(T) = \sum_{t=1}^{T} \big(R(Y_t^*) - R(Y_t)\big)$, where $Y_t^* \sim p_{\theta,a^*}$, and we use $Y_t \equiv Y_{t,A_t}$. We will use $p^* \equiv p_{\theta^*}$, and $p^*_{a^*} \equiv p_{\theta^*,a^*(\theta^*)}$, and $\mathbb{E}[Reg(T)]$ for the expected cumulative regret. Let $A^* := a^*(\theta^*)$ denote the optimal action under the true parameter. We define the (conditional) instantaneous expected regret at time t as $\Delta_t := \mu(A^*,\theta^*) - \mu(A_t,\theta^*)$. Then, conditional on the action sequence and the true parameter, $\mathbb{E}[Reg(T) \mid A_{1:T}, \theta^*] = \sum_{t=1}^{T} \Delta_t$, and therefore $\mathbb{E}[Reg(T)] = \mathbb{E}\Big[\sum_{t=1}^{T} \Delta_t\Big]$.

Thompson Sampling Thompson sampling is a Bayesian algorithm for bandit problems that works by sampling actions according to the (posterior) probability that they are the optimal action. Let $\mathcal{H}_t := \{A_s, Y_s\}_{s<t}$ be the history of the actions taken and outcomes observed up to time t. We write $\mathbb{E}_t[\cdot]$ for expectation conditional on $\mathcal{H}_t$. Thompson Sampling works by assuming the agent samples actions from a posterior distribution (or prior before any new observations) $P(\theta \mid \mathcal{H}_t)$ (abbreviated as $P_t(\theta)$) conditioned on $\mathcal{H}_t$ such that $P_t(A^* = a) = P_t(A_t = a)$[2]. Then, the agent samples a parameter $\hat{\theta}_t \sim P_t(\theta)$, and selects the action that maximises expected rewards *under the model* $\hat{\theta}_t$: $A_t \in \arg\max_{a\in\mathcal{A}} \mu(a,\hat{\theta}_t)$. Then, a new outcome Y_t is observed (when choosing A_t), and the belief $P_t(\theta)$ is updated according to the history $\mathcal{H}_{t+1} = \mathcal{H}_t \cup \{A_t, Y_t\}$ via Bayes:

$$P_{t+1}(\theta) = P(\theta \mid \mathcal{H}_{t+1}) \propto p_{\theta,A_t}(Y_t) P_t(\theta).$$

We define finally a quantity that will be of use for some of the results in the paper. From Russo and Van Roy [2016], we define the *information ratio* in a bandit as $\Gamma_t := \frac{\big(\mathbb{E}_t[\Delta_t]\big)^2}{\mathbb{I}_t(A^*;(A_t,Y_t))}$. In other words, it is the ratio of the squared expected regret at time t given the past history against the mutual information between the optimal action distribution and the current observation.

Expert Taxonomy We define three kinds of experts to be used through the rest of the work. An expert is an agent that has a fixed policy $\pi_e \in \Delta\mathcal{A}$. A *true expert* samples actions from policy $\pi_e^*(a) = \mathbb{1}[a = a^*(\theta^*)]$ (assigns probability 1 to the true optimal action). A *boundedly rational expert* (i.e. soft expert) has $\pi_e^\tau(a) = \frac{e^{\mu(a,\theta^*)/\tau}}{\sum_{a'} e^{\mu(a',\theta^*)/\tau}}$ with temperature τ. A *mistaken expert* (i.e. misaligned expert) assigns $\pi'_e(a') = 1$ for arbitrary action $a' \neq a^*(\theta^*)$. Finally, a *misleading expert* is a contamination model that mostly plays the optimal action for a parameter whose optimal action is poor in the true environment. Let $\theta^{mis} \in \arg\min_{\theta\in\Theta} \mu(a^*(\theta),\theta^*)$. Then $\pi_e^{mis}(a^*(\theta^{mis})) = 1-\epsilon$ and $\pi_e^{mis}(a^*(\theta^*)) = \epsilon$.

[2]Note that we use $A^* = a$ to refer to the random event corresponding to a being the optimal action under measure P_t. This differs from $a^*(\theta)$, which refers to the optimal action in expectation for a fixed θ.

3 Learning from Offline Expert Data

Consider the case where a *learner* has to act optimally in an unknown environment, and is able to observe a *true expert*. This can manifest as (i) the learner receiving an initial dataset $D_N^* = \{Y_n^*\}_{1 \leq n \leq N}$ and (ii) the learner observing new samples Y_t^* online.

Infinite Information To start our analysis, assume first that $N \to \infty$ and we can construct an unbiased density estimator with no errors, or, in other words, the learner has access to the likelihood $p_{a^*}^*(Y)$. Treating this as an offline data scenario, we can interpret the knowledge of $p_{a^*}^*(Y)$ as an observation to be incorporated into the learner's knowledge via posterior inference. Intuitively, knowing $p_{a^*}^*(Y)$ should restrict the set of non-zero likelihood parameters in our posterior to those which satisfy $\tilde{\Theta} := \{\theta \in \Theta : p_{\theta,a^*(\theta)} = p_{a^*}^*\}$. Let $\mathbb{1}[p_{a^*}^* \mid \theta] = 1$ if $p_{\theta,a^*(\theta)} = p_{a^*}^*$. Then, for the posterior to be consistent with the observed data, we want it to satisfy

$$P_1(\theta \mid p_{a^*}^*) \propto P_0(\theta)\mathbb{1}[p_{a^*}^* \mid \theta]. \tag{1}$$

We use P_0 to refer to the initial prior the learner has over the parameters Θ, and P_1 as the (offline) posterior resulting from incorporating the expert data. This posterior in (1) will assign zero mass[3] to any parameter θ which induces an optimal action distribution that does not match $p_{a^*}^*$. From the set of parameters that induce such a distribution, we cannot distinguish (have equal likelihood), so the prior will dominate the posterior mass. We show that this posterior update is consistent in the upcoming section, by showing it can be derived as a result of an infinite data limit.

Finite Information Next, consider the case where $N < \infty$, and therefore the learner starts with a finite dataset $D_N^* = \{Y_n^*\}_{1 \leq n \leq N}$ of samples from the optimal action, but cannot identify (yet) what action these correspond to. Following the intuition in the case of infinite information, one would want to incorporate this offline information into the prior, to afterwards proceed normally with TS, hopefully with a prior that is better informed. Recall that, under parameter $\theta \in \Theta$, the likelihood of a given sample Y^* being sampled from the bandit θ is $p_{\theta,a^*(\theta)}(Y^*)$. Then, given a set of N samples $D_N^* = \{Y_n^*\}_{1 \leq n \leq N}$, we can infer a posterior under the likelihood that the data comes from the current model as $P_1(\theta \mid D_N^*) \propto P_0(\theta)p_{\theta,a^*(\theta)}(D_N^*)$. Since the expert samples are i.i.d., the right hand side:

$$P_0(\theta)p_{\theta,a^*(\theta)}(D_N^*) = P_0(\theta)e^{\sum_{i=1}^N \log p_{\theta,a^*(\theta)}(Y_i^*)}. \tag{2}$$

Proposition 3.1. *Assume a finite set Θ. As the number of samples increases $N \to \infty$, the posterior update in* (2) *converges to the infinite data update in* (1)*. In other words,*

$$\lim_{N \to \infty} P_1(\theta \mid D_N^*) = P_1(\theta \mid p_{a^*}^*) \quad \textit{a.s.} \tag{3}$$

Regret Bounds with Offline Expert Data To estimate the Bayesian regret improvement of the agents when having access to offline expert data, let us first define the following concepts. The probability $P_0(A^* = a)$ under measure P_0 is the probability of a being optimal under the prior distribution $P_0(\theta)$. Let $\Theta_a^* := \{\theta \in \Theta : a^*(\theta) = a\}$; in other words, Θ_a^* is the subset of parameters that yields a to be the optimal action. Observe we can then write $P_0(A^* = a) = \int_{\Theta_a^*} P_0(\theta)d\theta$. Then, define $\mathbb{H}_0(A^*)$ to be the entropy of the optimal action distribution under measure P_0:

$$\mathbb{H}_0(A^*) = -\sum_{a \in \mathcal{A}} P_0(A^* = a) \log P_0(A^* = a).$$

We can first show that the probability mass of $A^* = a$ under posterior $P_0(\theta \mid D_N^*)$ (denoted as $P_1(A^* = a)$) is equal to the posterior probability $P_0(A^* = a \mid D_N^*)$ under prior $P_0(\theta)$.

Proposition 3.2. *Let $P_1(\theta) = P_0(\theta \mid D_N^*)$ be the posterior defined by the expert-outcome likelihood. Then the induced distribution over the optimal action satisfies $P_1(A^* = a) = P_0(A^* = a \mid D_N^*)$.*

Russo and Van Roy [2016] established an information-theoretic regret bound for Thompson Sampling that scales with the entropy of the optimal action. We can now show that, under expert data, the entropy of the (offline) posterior $P_0(\theta \mid D_N^*)$ is guaranteed to decrease in expectation over the

[3] In general, this needs to be dealt with care for compact parameter spaces and continuous measures to avoid measure theoretic issues. For this argument it suffices to assume a finite parameter set Θ, and all the results extend to compact parameter sets under appropriate measure theoretic assumptions prevalent in other Bayesian bandit work.

observed data. The following regret bound follows from the analysis of Russo and Van Roy [2016]; Our contribution here is to define the correct Bayesian update for outcome-only expert data, and show that evaluating the standard bound under the expert-updated posterior yields an exact decomposition of the benefit of expert data in terms of $\mathbb{I}_0(A^*; D_N^*)$.

Theorem 3.3 (Regret Reduction from Offline Expert Data)**.** *Let D_N^* denote N i.i.d. expert outcomes generated by the expert's (optimal) action, and let the learner form the expert-updated posterior $P_1(\theta) := P_0(\theta \mid D_N^*)$. Consider TS run with prior P_1. Then,*

$$\mathbb{E}[Reg_{TS_1}(T)] \leq \sqrt{\bar{\Gamma} T \big(\underbrace{\mathbb{H}_0(A^*)}_{\textit{uncertainty}} - \underbrace{\mathbb{I}_0(A^*; D_N^*)}_{\textit{information gain}} \big)}.$$

where $\bar{\Gamma}$ is a bound for the information ratio in TS, $\mathbb{H}_0(A^*)$ is the prior entropy of the optimal action and $\mathbb{I}_0(A^*; D_N^*)$ is the mutual information between the optimal action and the expert data under P_0.

In particular, Theorem 3.3 provides an upper bound on the expected regret that is no larger than the corresponding upper bound for a TS agent who observes no expert data and assumes the same prior P_0 given by Russo and Van Roy [Proposition 1, 2016]. Furthermore, *the information-theoretic bound is never worse than the bound for TS without expert data*. Intuitively, if the mutual information between the expert data and the optimal action distribution is high (*i.e.* the expert samples allow the agent to reduce the set of possible parameters to a much smaller subset), then the resulting regret will be significantly lower.

Remark 3.4. From the chain rule for mutual information, the bound in Theorem 3.3 is non-increasing in N.

4 Learning from Online Expert Data

Consider now the problem where the learner has no expert data to incorporate into its prior, but as it learns it can observe both its own pair (A_t, Y_t) and the outcome Y_t^e generated by a true expert. We study a budgeted update model: at each round the learner performs one Bayesian update, using either its own labeled observation or the expert's outcome-only observation. If both updates can be processed at negligible cost, then processing both is the natural Bayesian baseline; our question is the source-selection problem under a fixed update budget. We assume the learner still receives its own reward $R(Y_t)$, and thus the expected instantaneous regret $\mathbb{E}[R(Y_t^e) - R(Y_t)]$ does not depend on the expert sample, or on the agent's choice on which information source to incorporate. From Russo and Van Roy [2016] and Russo and Van Roy [2014], the expected regret of (general) Bayesian online learners is bounded by $\sqrt{\overline{\Gamma}\mathbb{H}_t(A^*)T}$, where $\overline{\Gamma}$ is an upper bound for the information ratio. Given that the agent's choice over what information to incorporate does not change the immediate rewards, this choice needs to be driven by the information gain from each source. Let $D_t \in \{Y_t^e, (Y_t, A_t)\}$ be the random variable representing the data processed at time t, which can be either the expert outcome or the pair (outcome, action) from the learner themselves. Then, the choice of data to learn from can be expressed through the choice

$$\arg\min_{D_t} \ \mathbb{E}[\mathbb{H}_t(A^* \mid D_t)] = \arg\max_{D_t} \ \mathbb{I}_t(A^*, D_t). \tag{4}$$

In other words, the agent should choose to learn from the sample that maximises the mutual information with the optimal action distribution. This is effectively a Bayesian experimental design framework [Lindley, 1956], where the *experiments* (self-generated data vs. expert data) need to be selected to maximise information gain[4]. This allows us to propose an online learning-and-acting Algorithm 1. Theorem 4.1 should be read as a bound on the information-ratio regret guarantee obtained by replacing the self observation with the routed observation, not as a claim that routing dominates a policy that processes all available data. The resulting routing rule estimates $\mathbb{I}_t(A^*; Y_t^e)$ and $\mathbb{I}_t(A^*; (A_t, Y_t))$, updates from the larger-information source, and then applies the corresponding Bayesian posterior update; see Algorithm 1.

Theorem 4.1 (Information-ratio bound under budgeted routing)**.** *Let an agent decide, at time t, whether it updates its posterior using $D_t \in \{(A_t, Y_t), (Y_t^e)\}$ under prior P_t, following the MI*

[4]Bayesian experimental design is usually framed in terms of the information gain of model parameters θ. In our case, we care about the mutual information between (A^*, D).

maximising rule (Algorithm 1) $D_t \in \arg\max_{D \in \{(A_t,Y_t),(Y_t^e)\}} \mathbb{I}_t(A^*, D)$*, and selecting actions via TS. Assume the standard TS information-ratio bound* $\Gamma_t \leq \bar{\Gamma}$ *is uniform over priors. Assume for some* $G \subseteq \{1, 2, \ldots, T\}$ *the expert data improves the MI by a non-zero amount, i.e.*

$$\exists \epsilon \in (0,1) : \mathbb{I}_t(A^*, Y_t^e) \geq \frac{1}{1-\epsilon}\mathbb{I}_t(A^*, (A_t, Y_t)) \ \forall t \in G.$$

Then, the expected regret of the resulting agent is

$$\mathbb{E}[Reg(T)] \leq \sqrt{\bar{\Gamma}(T - \epsilon|G|)\mathbb{H}_1(A^*)},$$

where $\mathbb{H}_1(A^*) = \mathbb{H}_0(A^*)$ *if no offline data is used.*

4.1 Estimating Mutual Information

The agent's information at time t is condensed in the current prior $P_t(\theta)$. In Thompson Sampling, the probability of selecting action a at time t (the history dependent policy π_t) is equal to the probability of a being optimal,

$$P_t(A_t = a) = P_t(A^* = a) = \int_{\Theta_a^*} P_t(\theta) d\theta =: \pi_t(a).$$

At time t, the learner evaluates the two candidate information sources through the quantities $\mathbb{I}_t(A^*, Y_t^e)$ and $\mathbb{I}_t(A^*, (A_t, Y_t))$. These may be computed exactly when the model permits, or estimated approximately under the current posterior P_t. These two quantities can be computed efficiently in the form:

$$\mathbb{I}_t(A^*, (A_t, Y_t)) = \sum_{a_t \in \mathcal{A}} \pi_t(a_t) \sum_{a^* \in \mathcal{A}} \pi_t(a^*) D_{\mathrm{KL}}(P_t(Y_t \mid A_t = a_t, A^* = a^*) \| P_t(Y_t \mid a_t)), \quad (5)$$

$$\mathbb{I}_t(A^*, Y_t^e) = \sum_{a^* \in \mathcal{A}} \pi_t(a^*) D_{\mathrm{KL}}(P_t(Y_t^e \mid A^* = a^*) \| P_t(Y_t^e)). \quad (6)$$

For a detailed derivation, we refer the reader to Appendix A.1.

4.2 When does Expert Information help?

A natural question after the results presented in previous sections is: when does expert information help (and when does it not help). Recall the *expert predictives* $P_t(Y_t^e \mid A^* = a)$, $P_t(Y_t^e)$, and recall the KL mixture expression for the MI in (6). We can quickly establish the following result.

Proposition 4.2. *The mutual information* $\mathbb{I}_t(A^*; Y_t^e) = 0$ *if and only if* $P_t(Y_t^e \mid A^* = a) = P_t(Y_t^e)$ *for every* $a \in \mathcal{A}$ *i.e., if and only if* $P_t(Y_t^e \mid A^* = a)$ *is identical across optimal-action labels* a*.*

Definition 4.3 (Expert-outcome symmetry taxonomy)**.** Fix a posterior P_t and define the expert-outcome predictive measures $Q_{t,a}(\cdot) := P_t(Y_t^e \in \cdot \mid A^* = a)$ for each $a \in \mathcal{A}$ with $P_t(A^* = a) > 0$. We say the problem is *symmetric at* P_t if $Q_{t,a} = Q_{t,b}$ for all such $a, b \in \mathcal{A}$, *asymmetric at* P_t if $Q_{t,a} \neq Q_{t,b}$ for some $a, b \in \mathcal{A}$ but the relevant predictive measures have overlapping support or *strongly asymmetric at* P_t if the predictive measures $\{Q_{t,a} : P_t(A^* = a) > 0\}$ are pairwise mutually singular.

Corollary 4.4 (Zero information in symmetric worlds)**.** *If the optimal-action likelihood is the same in every world,* $p_{\theta,a^*(\theta)}(\cdot) \equiv q(\cdot)$ *for all* $\theta \in \Theta$ *and some* $q \in \Delta(\mathcal{Y})$*, then* $\mathbb{I}_t(A^*; Y_t^e) = 0$ *for any* P_t*. The expert data leaves the posterior unchanged.*

Thus, symmetry in Definition 4.3 is exactly the case where expert outcomes carry no information about the optimal-action label.

When can $\mathbb{I}_t(A^*; Y_t^e) > 0$**?** From Proposition 4.2, expert outcomes help exactly when the action-indexed measures $Q_{t,a} = P_t(Y_t^e \mid A^* = a)$ differ. In the asymmetric case of Definition 4.3, expert observations shift probability mass between optimal-action hypotheses but need not identify the optimal action from a single sample. In the strongly asymmetric case, the predictive supports are disjoint, so an expert outcome can identify the optimal-action label whenever it falls in a label-specific support. These cases lower $\mathbb{H}_t(A^*)$ and improve the regret bound in Theorem 3.3 and 4.1.

5 Untrustworthy Experts

Until now, the analysis has focused on the setting in which the learning agent *fully trusts the expert*; there is an implicit assumption that the expert is a true expert. A natural question that follows is how this can be affected by misaligned or misleading experts, which can introduce robustness failure modes in agent learning. To address this, we first discuss the impact of imperfect experts in the current framework, and propose afterwards a strategy to incorporate expert trust in an online Bayesian learning agent.

Collapse under Naive Expert Trust Let us define $q \in \Delta(\mathcal{Y})$ as the marginal likelihood of outcomes induced by π_e: $q(Y) := \sum_{a\in\mathcal{A}} \pi_e(a) p^*_a(Y)$. Take N samples from q, $\{Y^e_n \sim q\}_{1\leq N}$. Recall that, since the learner is naive, it still updates its posterior based on the data: $P^q_1(\theta) \propto P_0(\theta) \exp\left(\sum_{n=1}^N \log p_{\theta,a^*(\theta)}(Y^q_n)\right)$. This is a specific form of a misspecified Bayesian inference problem; the agent is trying to infer a posterior from likelihood $p_{\theta,a^*(\theta)}$, while the data is sampled from a different distribution [Nott et al., 2023]. For $N \to \infty$, from established misspecified Bayes results [Berk, 1966, Bochkina, 2019] and under mild assumptions (measurability, compact Θ^*_a, $P_0(\theta) > 0$) the posterior $P^q_1(a = A^*)$ will concentrate probability mass around the set $\Theta_q := \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}(q \| p_{\theta,a^*(\theta)})$; in other words, the set of parameters that result in an optimal action distribution that is as close as possible to q. We discuss possible scenarios regarding non-optimal experts.

Boundedly Rational or Misleading Experts For a boundedly rational expert with π^τ_e, if τ is small, θ^* will still be the minimiser $\theta^* \in \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}(q \| p_{\theta,a^*(\theta)})$. In this case, the posterior will still concentrate around θ^* asymptotically and the agent will learn in the limit, but at a slower rate. However, for a *misleading expert* with π^{mis}_e (i.e. Huber contamination model [Huber, 1992]), depending on the problem instance there can be thresholds after which naive trust concentrates on parameters inducing suboptimal actions. Let $q_\epsilon := (1-\epsilon)p^*_{a^*(\theta^{mis})} + \epsilon p^*_{a^*(\theta^*)}$. If $D_{\mathrm{KL}}(q_\epsilon \| p^*_{a^*(\theta^{mis})}) < D_{\mathrm{KL}}(q_\epsilon \| p^*_{a^*(\theta^*)})$, then the misspecified posterior may concentrate on parameters whose optimal-action outcome distribution is closer to the misleading mixture than the true optimal-action outcome distribution is. In this case, the agent can end up being *confidently wrong*. For empirical examples of these cases, see Appendix 6.

5.1 Modeling the Expert: Learning to Trust

The existence of imperfect experts motivates the last question to be answered in this work: How should an online Bayesian learner estimate the expertise (or trustworthiness) of supposed experts, and how to incorporate it into their learning algorithm. Methods based on opponent modeling [Southey et al., 2012, Yu et al., 2022a] provide approaches for online Bayesian learners to estimate the behavior of other agents based on observations and use that for their own inference loops. Therefore, we conjecture that a solution to the possibility of experts being imperfect is to *infer the expert's policy based on their outcomes*.

Beliefs over Expert Policies We assume that the expert's capabilities to corrupt samples Y^e_t are limited; they cannot fabricate outcomes and are constrained to sampling from some (true) outcome distribution p^*_a. We assume therefore that the expert samples $A^e_t \sim \pi_e$, and the learner observes $Y^e_t \sim p^*_{A^e_t}$. A solution to handle the uncertainty over the expert policy π_e is to keep a (history dependent) joint prior $P_t(\theta, \pi_e)$, and update it as $P_{t+1}(\theta, \pi_e) \propto P_t(Y^e_t \mid \theta, \pi_e) P_t(\theta) P_t(\pi_e)$ [5]. Under policy π_e, the marginalised likelihood of a sampled expert outcome Y^e_t is $P_t(Y^e_t \mid \theta, \pi_e) = \sum_a \pi_e(a) p_{\theta,a}(Y^e_t)$.

Estimating MI with Expert Uncertainty We will now rely on the assumption that we can compute $\mathbb{E}_{P_t(\pi_e)}[\pi_e(a)] = \bar{\pi}_e(a)$. First, recall the MI can be written in terms of KL divergences:

$$\mathbb{I}_t(A^*, Y^e_t) = \sum_{a^*\in\mathcal{A}} \pi_t(a^*) D_{\mathrm{KL}}(P_t(Y^e_t \mid A^* = a^*) \| P_t(Y^e_t)).$$

[5] Note that this update is in general intractable since it couples the beliefs over θ and π_e. However, it can be practically approximated through *e.g.* a particle method.

We need to estimate $P_t(Y_t^e \mid A^* = a^*)$, $P_t(Y_t^e)$ to compute the corresponding MI and compare it with the MI computed from self-collected data in (5). One can show

$$P_t(Y_t^e) = \sum_a \bar{\pi}_e(a) P_t(Y_t^e \mid A_t^e = a), \quad P_t(Y_t^e \mid A^* = a^*) = \sum_a \bar{\pi}_e(a) P_t(Y_t^e \mid A_t^e = a, A^* = a^*).$$

With these, the learner can estimate the MI from observing expert data $\mathbb{I}_t(A^*, Y_t^e)$, which *incorporates the uncertainty about* π_e, and compare it against $\mathbb{I}_t(A^*, (A_t, Y_t))$ to decide which information source to process. For an algorithm to solve this expert modeling problem, see Algorithm 2 in Appendix D.

On (Un)Identifiable Expert Policies This strategy seems to rely on the expert policy π_e being *identifiable from outcomes* conditioned on an environment. However, identifiability of π_e is not required for safe behavior under our *MI routing* rule. Our online decision depends only on whether expert outcomes reduce uncertainty about the optimal action: by Proposition 4.2, $\mathbb{I}_t(A^*; Y_t^e) = 0$ if and only if the conditional predictives $P_t(Y_t^e \mid A^* = a)$ are identical across $a \in \mathcal{A}$. In many permutation-symmetric bandit classes under permutation-invariant priors, outcome ambiguity about π_e emerges exactly in this way: policies induce (approximately) the same outcome mixture, so $P_t(Y_t^e \mid A^* = a)$ varies little with a, and the resulting $\mathbb{I}_t(A^*; Y_t^e)$ is small. In this regime, our routing rule $D_t^* \in \arg\max_{D \in \{(A_t, Y_t), Y_t^e\}} \mathbb{I}_t(A^*; D)$ would fall back to self-labeled data whenever it is more informative. Thus, when expert outcomes are uninformative about A^* (whether due to non-identifiability or symmetry), MI routing rejects them; when expert outcomes are informative, MI routing exploits them. As a demonstration, see the results in Figure 7 on Appendix D.

6 Experiments

We validate our framework first on discrete bandit instances, Neural-Gaussian bandits (see Appendix D) and a bandit-adapted Model Predictive Control task. First, for the finite bandit experiments, the goal is to verify the theory in a setting where posterior distributions, Bayesian updates and MI can be computed exactly. In these experiments, the outcome space is $\mathcal{Y} = \{-50, ..., 50\}$, the reward is $R(Y) = Y$, and action space $|\mathcal{A}| = 70$. For all experiments, we fix the parameter space size $M = 500$. We use a uniform prior $P_0(\theta)$ and average results over 50 random seeds. We evaluate three problem classes instantiating Definition 4.3: *Symmetric*, *Asymmetric*, and *Strongly Asymmetric*. These classes vary only in the distinguishability of the expert-outcome predictives $\{P_t(Y_t^e \mid A^* = a)\}_a$. For details on the different scenarios, see Appendix D.

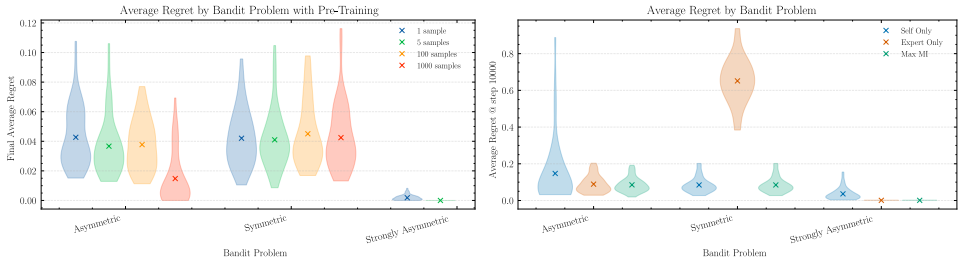


Figure 1: Regret in every problem instance for different pre-training budgets and information selection strategies.

Results As shown in Figure 1, offline pre-training yields no regret improvement for symmetric problems, confirming that expert data is uninformative when outcomes are independent of the action index. In the online setting, learning exclusively from the expert results in linear regret. However, our MI-estimating agent successfully identifies the expert as uninformative, consistently selecting its own experience and matching the performance of standard Thompson Sampling. Offline access to expert data significantly reduces initial uncertainty in asymmetric and strongly asymmetric problems. In the online setting dynamically selecting the information source via Algorithm 1 outperforms fixed strategies. By prioritizing the most informative source at each step, the agent achieves a learning rate improvement of approximately $8 - 12\%$ over standard Thompson Sampling at no additional sampling cost in the asymmetric instances, and matches performance of the best source in the other cases. Finally, the results show dramatic improvements for MI routing strategies in strongly asymmetric problems, around 99% reduction in regret after only a few steps as the agent effectively bypasses exploration by observing the expert.

Bayesian MPC with Outcome-Only Expert Trajectories Finally, we evaluate the online routing rule in control-derived contextual bandits using model-predictive control, on a task adapted from Tunyasuvunakool et al. [2020]. The learner observes state trajectories generated by an expert controller but not the expert's control inputs. The task is a point-mass control problem where an agent selects a sequence of discrete force vectors to move the mass close to an unknown goal. The goal position is part of θ and affects rewards but not transition dynamics, so expert trajectories reveal it through the direction of motion. We report regret per datapoint consumed, treating Both Updates as a higher-budget baseline that uses two observations per episode. For further details on the experiment results and implementation see Appendix D.6.

Table 1: Bayesian MPC point-mass results averaged across two scenarios: One where the expert is modelled correctly (softmax with $\beta = 8.0$), and one where the learner is misspecified and believes the expert is acting almost randomly (softmax with $\beta = 0.05$). Mean $\pm$ standard deviation.

Updates	MI Routing	Self-Only	Expert-Only	Both
2	2.26 ± 1.08	2.50 ± 1.16	$\mathbf{2.21 \pm 0.85}$	2.81 ± 1.30
10	$\mathbf{0.85 \pm 0.44}$	1.54 ± 0.84	1.25 ± 0.51	1.48 ± 0.88
20	$\mathbf{0.57 \pm 0.27}$	1.02 ± 0.63	1.02 ± 0.40	1.04 ± 0.64
50	$\mathbf{0.38 \pm 0.12}$	0.56 ± 0.25	0.85 ± 0.28	0.60 ± 0.31

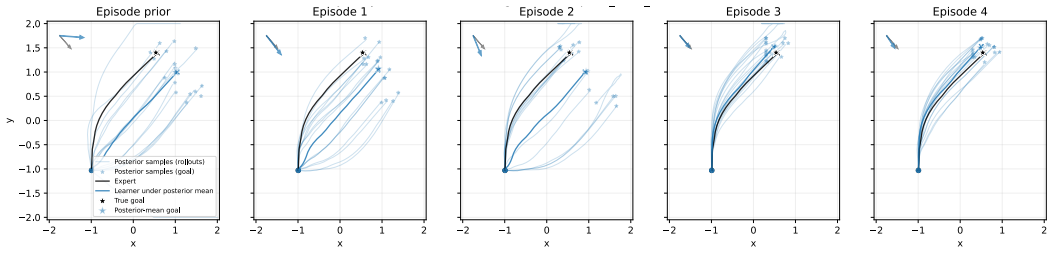


Figure 2: Trajectory evolution in the point-mass Bayesian MPC task for MI Routing with a modeled expert policy and $\beta = 8.0$. The expert trajectory is observed as outcome-only data; the learner updates its posterior and replans across episodes.

Results This class of tasks precisely motivates the results in the work. Expert trajectories may be very informative but are in practice *costly to learn from* because they marginalize over latent expert actions. Our MI routing agent makes use of this data only when it ensures net information gain over its self-collected data. The MI Routing agents essentially match (or improve) the learning trajectory from the best source in each instance. Furthermore, when the agent is misspecified with respect to the expert, the MI routing rule shields it and falls back to learning mostly from self-generated data.

7 Discussion

This paper studies a simple but relevant question: when an agent observes only an expert's outcomes but not the expert's actions, what can it learn (and when should it learn from it)? We answer these questions through information theoretic lenses. Expert outcomes help precisely when they **provide explicit evidence about which action is optimal** by helping learning agents pruning possible worlds in model-based Bayesian bandits. In the offline setting, this evidence reduces the entropy of A^* and tightens the corresponding Thompson Sampling regret bound.

In the online setting, the key issue is not whether expert data contains information, but whether it is more useful information than self-collected experiences. Expert trajectories may be highly informative, but *they can also be expensive* to incorporate. The learner must marginalize over latent expert actions, evaluate an internal expert model, or spend an external query budget. Our MI routing solution turns this trade-off into a Bayesian experimental-design problem: update from the expert **only if it provides more information** about A^*. The method also provides a robustness mechanism against misspecified agents. When the expert is uninformative, unidentifiable or poorly modeled, the same MI criterion induces the agent to reject expert outcomes and fall back to self-generated data.

Limitations and Future Work The framework inherits the usual limitations of Bayesian bandit and control models. Several theoretical results are stated for finite parameter and action spaces. Extending the guarantees to continuous models, function approximation, and adaptive experts would be a valuable direction. Furthermore, relevant followup work would be to study simultaneous-learning settings where the expert's policy is not static, but changes as the expert learns from its own experience, yielding a Bayesian game-theoretic problem.

Acknowledgements

The authors thank David Hyland and Francesco Fabiano for the useful conversations on this topic. Authors acknowledge funding from a UKRI AI World Leading Researcher Fellowship awarded to Wooldridge (grant EP/W002949/1). MW and AC also acknowledge funding from Trustworthy AI - Integrating Learning, Optimisation and Reasoning (TAILOR), a Horizon2020 research and innovation project (Grant Agreement 952215). DJO, NB and MW acknowledge funding from UK AISI *The Alignment Project*, grant AP-S2-100055.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via Inverse Reinforcement Learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shubhada Agrawal, Sandeep Juneja, Karthikeyan Shanmugam, and Arun Sai Suggala. Optimal best-arm identification in bandits with access to offline data. *arXiv preprint arXiv:2306.09048*, 2023.

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *International Conference on Machine Learning*, pages 373–382. PMLR, 2021.

Robert H Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

Natalia Bochkina. Bernstein–von mises theorem and misspecified models: A review. *Foundations of modern statistics*, pages 355–380, 2019.

Simina Brânzei and Yuval Peres. Multiplayer bandit learning, from competition to cooperation. In *Conference on Learning Theory*, pages 679–723. PMLR, 2021.

David Carmel and Shaul Markovitch. Opponent modeling in multi-agent systems. In *International joint conference on artificial intelligence*, pages 40–52. Springer, 1995.

Nicolò Cesa-Bianchi, Khaled Eldowa, Emmanuel Esposito, and Julia Olkhovskaya. Improved regret bounds for bandits with expert advice. *Journal of Artificial Intelligence Research*, 83, 2025.

William Chang and Yuanhao Lu. Multiplayer information asymmetric contextual bandits. *arXiv preprint arXiv:2503.08961*, 2025.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. *arXiv preprint arXiv:2405.02594*, 2024.

John D Geanakoplos and Heraklis M Polemarchakis. We can't disagree forever. *Journal of Economic theory*, 28(1):192–200, 1982.

Michael Gimelfarb, Scott Sanner, and Chi-Guhn Lee. Reinforcement learning with multiple experts: A bayesian model combination approach. *Advances in neural information processing systems*, 31, 2018.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

Satyen Kale. Multiarmed bandits with limited expert advice. In *Conference on learning theory*, pages 107–122. PMLR, 2014.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Tamlin Love, Ritesh Ajoodha, and Benjamin Rosman. Who should i trust? cautiously learning with unreliable experts. *Neural Computing and Applications*, 35(23):16865–16875, 2023.

H Brendan McMahan and Matthew J Streeter. Tighter bounds for multi-armed bandits with expert advice. In *COLT*, 2009.

Yoram Moses and Gal Nachum. Agreeing to disagree after all. In *Proceedings of the 3rd conference on Theoretical aspects of reasoning about knowledge*, pages 151–168, 1990.

Jeremy Muesing, Nisar Ahmed, Luke Burks, Michael Iuzzolino, and Danielle Albers Szafir. Fully bayesian human–machine data fusion for robust online dynamic target characterization. *Journal of Aerospace Information Systems*, 18(2):26–49, 2021.

Samer Nashed and Shlomo Zilberstein. A survey of opponent modeling in adversarial domains. *Journal of Artificial Intelligence Research*, 73:277–327, 2022.

David J Nott, Christopher Drovandi, and David T Frazier. Bayesian inference for misspecified generative models. *Annual Review of Statistics and Its Application*, 11, 2023.

Benjamin Plaut, Juan LiÃŠvano-Karim, and Stuart Russell. Asking for help enables safety guarantees without sacrificing effectiveness. *arXiv preprint arXiv:2502.14043*, 2025a.

Benjamin Plaut, Hanlin Zhu, and Stuart Russell. Avoiding catastrophe in online learning by asking for help. In *Forty-second International Conference on Machine Learning*, 2025b.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in neural information processing systems*, 27, 2014.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Nihal Sharma, Rajat Sen, Soumya Basu, Karthikeyan Shanmugam, and Sanjay Shakkottai. Bandits with stochastic experts: Constant regret, empirical experts and episodes. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 9(3):1–33, 2024.

Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180. PMLR, 2015.

Finnegan Southey, Michael P. Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: Opponent modelling in poker, 2012. URL `https://arxiv.org/abs/1207.1411`.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444. URL http://www.jstor.org/stable/2332286.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2020.100022. URL https://www.sciencedirect.com/science/article/pii/S2665963820300099.

Shohei Wakayama and Nisar Ahmed. Observation-augmented contextual multi-armed bandits for robotic search and exploration. *IEEE Robotics and Automation Letters*, 2024.

Le Yang, Vincent YF Tan, and Wang Chi Cheung. Best arm identification with possibly biased offline data. *arXiv preprint arXiv:2505.23165*, 2025.

XiaoPeng Yu, Jiechuan Jiang, Wanpeng Zhang, Haobin Jiang, and Zongqing Lu. Model-based opponent modeling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 28208–28221. Curran Associates, Inc., 2022a.

Xiaopeng Yu, Jiechuan Jiang, Wanpeng Zhang, Haobin Jiang, and Zongqing Lu. Model-based opponent modeling. *Advances in Neural Information Processing Systems*, 35:28208–28221, 2022b.

A Notation and Framework

The following table compiles notation and symbols.

Table 2: Notation used throughout the paper.

Symbol	Meaning
θ^*	True environment parameter
$P_t(\theta)$	Learner posterior over parameters after history $\mathcal{H}_t$
$\mathcal{A}, \mathcal{Y}$	Action and outcome spaces
$p_{\theta,a}$	Outcome distribution for action a under parameter θ
$\mu(a, \theta)$	Mean reward of action a under parameter θ
$a^*(\theta)$	Optimal action under parameter θ
A^*	Random optimal action induced by θ^* under the learner belief
Y_t, Y_t^e	Learner outcome and expert outcome at time t
D_t	Data source processed at time t
Δ_t	Conditional instantaneous expected regret
$\mathbb{H}_t, \mathbb{I}_t$	Entropy and mutual information conditional on $\mathcal{H}_t$
Γ_t	Thompson Sampling information ratio
π_e	Expert policy

A.1 Estimating Mutual Information

Self-Generated Data Predictives The (prior) outcome marginal predictive density is

$$P_t(Y_t \mid a_t) = \int_\Theta P_t(\theta) p_{\theta,a_t}(Y_t) d\theta;$$

this is the marginal likelihood of observing outcome Y_t having selected action $A_t = a_t$. For any $a^* \in \mathcal{A}$, the joint density of (A^*, Y_t) conditional on $A_t = a_t$ is

$$P_t(Y_t, A^* = a^* \mid A_t = a_t) = \int_{\Theta^*_{a^*}} P_t(\theta) p_{\theta,a_t}(Y_t) d\theta,$$

and observe that $P_t(Y_t \mid a_t) = \sum_{a^* \in \mathcal{A}} P_t(Y_t, A^* = a^* \mid A_t = a_t)$. This density indicates how much of the probability mass of any given outcome y comes from the worlds where $A^* = a^*$. Finally, the conditional density of Y_t under $A_t = a_t$ with the hypothesis that $A^* = a^*$ is $P_t(Y_t \mid A_t = a_t, A^* = a^*) = \frac{P_t(Y_t, A^* = a^* | A_t = a_t)}{\pi_t(a^*)}$.

By the chain rule for MI, $\mathbb{I}_t(A^*; (A_t, Y_t)) = \mathbb{I}_t(A^*; A_t) + \mathbb{I}_t(A^*; Y_t \mid A_t)$. Since we condition on history, $\mathbb{I}_t(A^*; A_t) = 0$ and the term $\mathbb{I}_t(A^*; Y_t \mid A_t)$ can be expanded as $\mathbb{I}_t(A^*; Y_t \mid A_t) = \sum_{a \in \mathcal{A}} \pi_t(a)\, \mathbb{I}_t(A^*; Y_t \mid A_t = a)$. Therefore, the MI resulting from observing self-collected data (A_t, Y_t) admits the representation

$$\mathbb{I}_t(A^*, (A_t, Y_t)) = \sum_{a_t \in \mathcal{A}} \pi_t(a_t) \sum_{a^* \in \mathcal{A}} \pi_t(a^*) D_{\mathrm{KL}}(P_t(Y_t \mid A_t = a_t, A^* = a^*) \| P_t(Y_t \mid a_t)). \quad (7)$$

Expert Data Predictives Similarly, we can now compute predictive densities for the case where the agent only observes the expert sample Y_t^e, knowing it comes from the optimal action. Analogous to previous definitions, we define the marginal and joint predictive for the expert output as:

$$P_t(Y_t^e) = \int_\Theta P_t(\theta) p_{\theta,a^*(\theta)}(Y_t^e) d\theta, \;\; P_t(Y_t^e, A^* = a^*) = \int_{\Theta^*_{a^*}} P_t(\theta) p_{\theta,a^*}(Y_t^e) d\theta.$$

The conditional expert predictive is $P_t(Y_t^e \mid A^* = a^*) = \frac{P_t(Y_t^e, A^* = a^*)}{P_t(A^* = a^*)}$, and $\mathbb{I}_t(A^*, Y_t^e)$ admits the representation

$$\mathbb{I}_t(A^*, Y_t^e) = \sum_{a^* \in \mathcal{A}} \pi_t(a^*) D_{\mathrm{KL}}(P_t(Y_t^e \mid A^* = a^*) \| P_t(Y_t^e)). \quad (8)$$

B Technical Results

Proposition 3.1. First, let us write

$$\sum_{i=1}^{N} \log p_{\theta,a^*(\theta)}(Y_i^*) = N\big(\frac{1}{N}\sum_{i=1}^{N} \log p_{\theta,a^*(\theta)}(Y_i^*)\big).$$

Now we have

$$\frac{1}{N}\sum_{i=1}^{N} \log p_{\theta,a^*(\theta)}(Y_i^*) = \mathbb{E}_{Y\sim p^*_{a^*}}[\log p_{\theta,a^*(\theta)}(Y)] + \delta_N(\theta),$$

and $\delta_N(\theta) := \frac{1}{N}\sum_{i=1}^{N} \log p_{\theta,a^*(\theta)}(Y_i^*) - \mathbb{E}_{Y\sim p^*_{a^*}}[\log p_{\theta,a^*(\theta)}(Y)]$, which goes to zero *almost surely* as $N \to \infty$ by the law of large numbers. Observe that $\mathbb{E}_{Y\sim p^*_{a^*}}[\log p_{\theta,a^*(\theta)}(Y)]$ is the cross entropy between $p^*_{a^*}$ and $p_{\theta,a^*(\theta)}$, and thus

$$\mathbb{E}_{Y\sim p^*_{a^*}}[\log p_{\theta,a^*(\theta)}(Y)] = -\mathbb{H}(p^*_{a^*}) - D_{\mathrm{KL}}(p^*_{a^*}||p_{\theta,a^*(\theta)}).$$

Then, substituting back in the posterior update,

$$\begin{aligned}
P_1(\theta \mid D_N^*) = \frac{\prod_{i=1}^N p_{\theta,a^*(\theta)}(Y_i^*)P_0(\theta)}{\sum_{\nu\in\Theta}\prod_{i=1}^N p_{\nu,a^*(\nu)}(Y_i^*)P_0(\nu)} = \\
\frac{\exp\big(\sum_{i=1}^N \log p_{\theta,a^*(\theta)}(Y_i^*)\big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\big(\sum_{i=1}^N \log p_{\nu,a^*(\nu)}(Y_i^*)\big)P_0(\nu)} = \\
= \frac{\exp\Big(N\Big(\mathbb{E}_{Y\sim p^*_{a^*}}[\log p_{\theta,a^*(\theta)}(Y)] + \delta_N(\theta)\Big)\Big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\Big(\mathbb{E}_{Y\sim p^*_{a^*}}[\log p_{\nu,a^*(\nu)}(Y)] + \delta_N(\nu)\Big)\Big)P_0(\nu)} = \\
= \frac{\exp\big(N\big(-\mathbb{H}(p^*_{a^*}) - D_{\mathrm{KL}}(p^*_{a^*}||p_{\theta,a^*(\theta)}) + \delta_N(\theta)\big)\big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\big(-\mathbb{H}(p^*_{a^*}) - D_{\mathrm{KL}}(p^*_{a^*}||p_{\nu,a^*(\nu)}) + \delta_N(\nu)\big)\Big)P_0(\nu)} = \\
= \frac{\exp\big(N\big(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\theta,a^*(\theta)}) + \delta_N(\theta)\big)\big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\big(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\nu,a^*(\nu)}) + \delta_N(\nu)\big)\Big)P_0(\nu)},
\end{aligned}$$

where the last step holds since $\exp(-\mathbb{H}(p^*_{a^*}))^N$ does not depend on θ and it cancels out with the normalization constant.

From the definition of almost sure convergence we have that, for any $\epsilon > 0$ and for almost every $\omega \in \Omega$, there exists a $0 < N'_{\omega,\epsilon} < \infty$ such that for all $N > N'_{\omega,\epsilon}$ we have $|\delta_N(\theta)| \le \epsilon$. Notice that in particular this means that for any $\epsilon_0 \in (0,1)$ and almost every $\omega \in \Omega$ there exists an $0 < N'_{\omega,\epsilon} < \infty$ with $\epsilon = \epsilon_0 \cdot \min_{\theta^*\in\Theta\setminus\tilde{\Theta}} D_{\mathrm{KL}}(p^*_{a^*}\|p_{a^*,\theta^*}) > 0$ such that $|\delta_N(\theta)| \le \epsilon$ for all $N > N'_{\omega,\epsilon}$. Finally, this implies that for almost any $\omega \in \Omega$ and for this ϵ, we have for $N > N'_{\omega,\epsilon}$

$$\begin{aligned}
0 &\le \exp\left(-D_{\mathrm{KL}}(p^*_{a^*}\|p_{a^*,\theta^*}) + \delta_N(\theta)\right) \\
&\le \exp\left(-D_{\mathrm{KL}}(p^*_{a^*}\|p_{a^*,\theta^*}) + |\delta_N(\theta)|\right) \\
&\le \exp\big(-[1-\epsilon_0]D_{\mathrm{KL}}(p^*_{a^*}\|p_{\theta,a^*(\theta)})\big) \\
&< 1.
\end{aligned} \tag{9}$$

Thus since $\lim_{N\to\infty}\left(\exp\left(-[1-\epsilon_0]D_{\mathrm{KL}}(p^*_{a^*}\|p_{\theta,a^*(\theta)})\right)\right)^N \to 0$ we also have almost surely that

$$\lim_{N\to\infty}\left(\exp\left(-D_{\mathrm{KL}}(p^*_{a^*}\|p_{a^*,\theta^*}) + \delta_N(\theta)\right)\right)^N \to 0.$$

Now, let us consider the subsets $\tilde{\Theta}$ and $\Theta \setminus \tilde{\Theta}$. First, take $\theta \in \tilde{\Theta}$. For any such theta, the posterior update $\forall\, \theta \in \tilde{\Theta}$ is

$$\begin{aligned}
&\lim_{N\to\infty} P_1(\theta \mid D_N^*) = \\
&= \lim_{N\to\infty} \frac{\exp\big(N\delta_N(\theta)\big)P_0(\theta)}{\sum_{\nu\in\Theta}\exp\Big(N\big(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\nu,a^*(\nu)}) + \delta_N(\nu)\big)\Big)P_0(\nu)}.
\end{aligned}$$

Dividing the numerator and denominator by $\exp\big(N\delta_N(\theta)\big)$,

$$\lim_{N\to\infty} P_1(\theta \mid D_N^*) =$$
$$= \lim_{N\to\infty} \frac{P_0(\theta)}{\sum_{\nu\in\Theta} \exp\Big(N\left(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\nu,a^*(\nu)}) + \delta_N(\nu) - \delta_N(\theta)\right)\Big)P_0(\nu)}.$$

First, any term in the denominator with $\nu \in \tilde{\Theta}$ has the same likelihood function for the optimal action. Therefore, $\delta_N(\nu) - \delta_N(\theta) = 0\ a.s.$ for any $\nu, \theta \in \tilde{\Theta}$. Second, by the same argument as (9), any term $\nu \notin \tilde{\Theta}$ goes to zero. Therefore,

$$\lim_{N\to\infty} \frac{P_0(\theta)}{\sum_{\nu\in\Theta} \exp\Big(N\left(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\nu,a^*(\nu)}) + \delta_N(\nu) - \delta_N(\theta)\right)\Big)P_0(\nu)} =$$
$$= \frac{1}{\sum_{\nu\in\tilde{\Theta}} P_0(\nu)} P_0(\theta) \quad \forall\, \theta \in \tilde{\Theta}.$$

Now consider $\theta \in \Theta \setminus \tilde{\Theta}$. Pick an arbitrary reference $\nu_0 \in \tilde{\Theta}$. We can bound the limit fraction as:

$$\lim_{N\to\infty} \frac{\exp\left(N\left(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\theta,a^*(\theta)}) + \delta_N(\theta)\right)\right)P_0(\theta)}{\sum_{\nu\in\Theta} \exp\Big(N\left(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\nu,a^*(\nu)}) + \delta_N(\nu)\right)\Big)P_0(\nu)} \leq$$
$$\leq \lim_{N\to\infty} \frac{\exp\left(N\left(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\theta,a^*(\theta)}) + \delta_N(\theta)\right)\right)P_0(\theta)}{\exp\Big(N\delta_N(\nu_0)\Big)P_0(\nu_0)} \quad \forall\, \theta \in \Theta \setminus \tilde{\Theta}.$$

Now, re-arranging terms,

$$\lim_{N\to\infty} \frac{\exp\left(N\left(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\theta,a^*(\theta)}) + \delta_N(\theta)\right)\right)P_0(\theta)}{\exp\Big(N\delta_N(\nu_0)\Big)P_0(\nu_0)} =$$
$$= \lim_{N\to\infty} \exp\left(N\left(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\theta,a^*(\theta)}) - \delta_N(\nu_0) + \delta_N(\theta)\right)\right)\frac{P_0(\theta)}{P_0(\nu_0)}.$$

By the same argument as (9), the exponent limit goes to zero, and thus $\forall\, \theta \in \Theta \setminus \tilde{\Theta}$:

$$\lim_{N\to\infty} \frac{\exp\left(N\left(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\theta,a^*(\theta)}) + \delta_N(\theta)\right)\right)P_0(\theta)}{\sum_{\nu\in\Theta} \exp\Big(N\left(-D_{\mathrm{KL}}(p^*_{a^*}||p_{\nu,a^*(\nu)}) + \delta_N(\nu)\right)\Big)P_0(\nu)} \leq 0.$$

This completes the proof, and we have

$$\lim_{N\to\infty} P_1(\theta \mid D_N^*) = \frac{\mathbb{I}[p^*_{a^*} \mid \theta]}{\sum_{\nu\in\tilde{\Theta}} P_0(\nu)} P_0(\theta).$$

□

Proposition 3.2. The event $A^* = a$ under the posterior density can be written as

$$P_1(A^* = a) = \int_{\Theta_a^*} P_0(\theta \mid D_N^*)d\theta =$$
$$\int_{\Theta_a^*} \frac{p_{\theta,a}(D_N^*)P_0(\theta)}{\int_\Theta p_{\nu,a^*(\nu)}(D_N^*)P_0(\nu)d\nu}d\theta = \frac{\int_{\Theta_a^*} p_{\theta,a}(D_N^*)P_0(\theta)d\theta}{\int_\Theta p_{\nu,a^*(\nu)}(D_N^*)P_0(\nu)d\nu}. \tag{10}$$

Now observe, by Bayes' Theorem,

$$P_0(A^* = a \mid D_N^*) = \frac{P(D_N^* \mid A^* = a)P_0(A^* = a)}{P_0(D_N^*)}. \tag{11}$$

The likelihood $P(D_N^* \mid A^* = a)$ represents the likelihood of the expert samples given that A is optimal. This can be computed by marginalising over all parameters that make A optimal,

$$P(D_N^* \mid A^* = a) = \int_\Theta p(D_N^* \mid \theta) P_0(\theta \mid A^* = a) d\theta = \\ = \int_{\Theta_a^*} p_{\theta,a}(D_N^*) \frac{P_0(\theta)}{P_0(A^* = a)} d\theta \implies$$

$$\implies P(D_N^* \mid A^* = a)\, P_0(A^* = a) = \int_{\Theta_a^*} p_{\nu,a^*(\nu)}(D_N^*) P_0(\theta) d\theta, \tag{12}$$

where the last equality is simply a conditional probability relation, and the last step holds since $P(A^* = a)$ does not depend on θ. Therefore, substituting (11) and (12) in (10):

$$P_1(A^* = a) = \frac{\int_{\Theta_a^*} p_{\theta,a}(D_N^*) P_0(\theta) d\theta}{\int_\Theta p_{\nu,a}(D_N^*) P_0(\nu) d\nu} = \\ = \frac{P(D_N^* \mid A^* = a) P_0(A^* = a)}{\int_\Theta p_{\nu,a}(D_N^*) P_0(\nu) d\nu} = P_0(A^* = a \mid D_N^*).$$

□

Theorem 3.3 (Regret Reduction from Offline Expert Data). The result follows directly from the information-theoretic analysis of Russo and Van Roy [2016], which bounds the Bayesian regret of a Thompson Sampling agent by the entropy of the optimal action under its current belief distribution. The agent has belief P_1. The regret, conditioned on a specific realization of D_N^*, is bounded by:

$$\mathbb{E}[Reg(T) \mid D_N^*] \leq \sqrt{\bar{\Gamma} T \cdot \mathbb{H}_1(A^*)}$$

To find the unconditional expected regret, we take the expectation over the expert data $D_N^* \sim p_{a^*}^*(\theta^*)$, where the uncertainty about θ^* is captured by the prior P_0:

$$\mathbb{E}[Reg(T)] = \mathbb{E}_{D_N^*}\left[\mathbb{E}[Reg(T) \mid D_N^*]\right] \leq \\ \leq \mathbb{E}_{D_N^*}\left[\sqrt{\bar{\Gamma} T \cdot \mathbb{H}_1(A^*)}\right].$$

Applying Jensen's inequality we have $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$, which gives:

$$\mathbb{E}[Reg(T)] \leq \sqrt{\bar{\Gamma} T \mathbb{E}_{D_N^*}[\mathbb{H}_1(A^*)]}.$$

The entropy $\mathbb{H}_1(A^*)$ is precisely the conditional entropy of the optimal action given the expert data, under the original measure P_0:

$$\mathbb{H}_1(A^*) = -\sum_{a \in \mathcal{A}} P_1(A^* = a) \log P_1(A^* = a) = \\ = -\sum_{a \in \mathcal{A}} P_0(A^* = a \mid D_N^*) \log P_0(A^* = a \mid D_N^*) =: \mathbb{H}_0(A^* \mid D_N^*).$$

Substituting this into the bound, we get:

$$\mathbb{E}[Reg(T)] \leq \sqrt{\bar{\Gamma} T \mathbb{E}_{D_N^*}[\mathbb{H}_0(A^* \mid D_N^*)]}.$$

Finally, from the definition of MI: $\mathbb{E}_{D_N^*}[\mathbb{H}_0(A^* \mid D_N^*)] = \mathbb{H}_0(A^*) - \mathbb{I}_0(A^*; D_N^*)$. This yields the main result:

$$\mathbb{E}[Reg(T)] \leq \sqrt{\bar{\Gamma} T \left(\mathbb{H}_0(A^*) - \mathbb{I}_0(A^*; D_N^*)\right)}.$$

This completes the proof. □

Proof. Define $\mathbb{I}_t(A^*, (A_t, Y_t)) = \mathbb{I}_t^{self}$, $\mathbb{I}_t(A^*, Y_t^e) = \mathbb{I}_t^e$. At every round, the immediate MI is $\mathbb{I}_t^* := \mathbb{I}_t(A^*, D_t) = \max\{\mathbb{I}_t^{self}, \mathbb{I}_t^e\} \geq \mathbb{I}_t^{self}$. Furthermore, in rounds $t \in G$, $\mathbb{I}_t^* \geq \frac{1}{1-\epsilon}\mathbb{I}_t^{self}$. By the prior-uniform TS information-ratio assumption, applied to the current posterior induced by the routed history, $\frac{\mathbb{E}_t[\Delta_t]^2}{\mathbb{I}_t^{self}} \leq \bar{\Gamma}$. For any t, $\frac{\mathbb{E}_t[\Delta_t]^2}{\mathbb{I}_t^*} \leq \frac{\mathbb{E}_t[\Delta_t]^2}{\mathbb{I}_t^{self}} \leq \bar{\Gamma}$, so the ratio bound holds regardless of the information source chosen. Furthermore, for $t \in G$, $\mathbb{E}_t[\Delta_t]^2 \leq (1-\epsilon)\bar{\Gamma}\mathbb{I}_t^*$. Defining $\alpha_t = 1 - \epsilon$ for $t \in G$ and $\alpha_t = 1$ for $t \notin G$, we derive the time bound $\mathbb{E}_t[\Delta_t]^2 \leq \alpha_t\bar{\Gamma}\mathbb{I}_t^*$. Taking square roots and summing over the horizon,

$$\mathbb{E}[Reg(T)] = \mathbb{E}[\sum_{t=1}^{T} \mathbb{E}_t[\Delta_t]] \leq \sqrt{\bar{\Gamma}}\mathbb{E}[\sum_{t=1}^{T} \sqrt{\alpha_t \mathbb{I}_t^*}].$$

Finally, $\mathbb{E}[\sum_{t=1}^{T} \sqrt{\alpha_t \mathbb{I}_t^*}] \leq \sqrt{\left(\sum_{t=1}^{T} \alpha_t\right) \mathbb{E}[\sum_{t=1}^{T} \mathbb{I}_t^*]}$. By definition, $\sum_{t=1}^{T} \alpha_t = T - \epsilon|G|$, and from Proposition 3.2 updating from expert data yields a well-specified Bayesian posterior, so $\mathbb{I}_t(A^*; D_t) = \mathbb{H}_t(A^*) - \mathbb{E}_t[\mathbb{H}_{t+1}(A^*)]$, and $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}_t^*\right] \leq \mathbb{H}_1(A^*)$. Substituting back in the regret bound,

$$\mathbb{E}[Reg(T)] \leq \sqrt{\bar{\Gamma}(T - \epsilon|G|)\mathbb{H}_1(A^*)}.$$

□

Proposition 4.2. By (6), $\mathbb{I}_t(A^*; Y_t^e) = 0$ if and only if each $D_{\mathrm{KL}}(P_t(Y_t^e \mid A^* = a) \,\|\, P_t(Y_t^e)) = 0$, which holds if and only if $P_t(Y_t^e \mid A^* = a) = P_t(Y_t^e)$ for all a. □

Corollary 4.4. Assume that for every $\theta \in \Theta$,

$$p_{\theta, a^*(\theta)}(y) \equiv q(y) \qquad \text{for all } y \in \mathcal{Y}, \tag{13}$$

for some fixed density function q on $\mathcal{Y}$. We now prove (i) A^* and Y_t^e are independent under P_t so $\mathbb{I}_t(A^*; Y_t^e) = 0$, and (ii) an expert sample leaves the posterior over θ unchanged.

First, for any action a and any measurable $B \subseteq \mathcal{Y}$,

$$P_t(Y_t^e \in B \mid A^* = a) = \int_{\Theta_a^*} \Big(\int_B p_{\theta,a}(y)\, dy \Big) P_t(\theta \mid A^* = a)\, d\theta = \int_B q(y)\, dy =: Q(B),$$

which does not depend on a. Note that the first equality comes from the definition of the conditional measure $P_t(Y_t^e \in B \mid A^* = a)$, and the second equality comes from assumption (13) and from the fact that Θ_a^* is the subset of Θ that includes all θ with $A^* = a$. Hence $P_t(Y_t^e \mid A^* = a) = Q$ for all a, and the (unconditional) predictive is also $P_t(Y_t^e) = \sum_a \pi_t(a) Q(Y_t^e) = Q(Y_t^e)$. From (6),

$$\mathbb{I}_t(A^*; Y_t^e) = \sum_{a \in \mathcal{A}} \pi_t(a)\, D_{\mathrm{KL}}\big(P_t(Y_t^e \mid A^* = a) \,\|\, P_t(Y_t^e)\big), \tag{14}$$

and each term equals $D_{\mathrm{KL}}(Q\|Q) = 0$, hence $\mathbb{I}_t(A^*; Y_t^e) = 0$.

Now let y^* be an observed expert outcome. Bayes' rule gives, for finite Θ (the general case follows by replacing sums with integrals),

$$P_{t+1}(\theta \mid Y_t^e = y^*) \;\propto\; p_{\theta, a^*(\theta)}(y^*)\, P_t(\theta) = q(y^*)\, P_t(\theta). \tag{15}$$

After normalizing by $\sum_\vartheta q(y^*) P_t(\vartheta) = q(y^*)$, we obtain $P_{t+1}(\theta \mid y^*) = P_t(\theta)$. This completes the proof. □

C Information Choice Algorithm

We describe here in detail the algorithm discussed in the main text.

D Experimental Implementation and Results

All experiments described here were run on a single 10 core Apple M4 chip.

Algorithm 1 Budgeted Information Routing

Input: Current posterior P_t.
Estimate (or compute) $\hat{I}_t^e = \mathbb{I}_t(A^*; Y_t^e)$, $\hat{I}_t^s = \mathbb{I}_t(A^*; (A_t, Y_t))$.
Select source $d_t \in \arg\max_{d \in \{e,s\}} \{\hat{I}_t^e, \hat{I}_t^s\}$.
Sample action $A_t \sim P_t(A^* \in \cdot)$.
if $d_t = s$ **then**
    Observe $D_t = (A_t, Y_t)$.
else
    Observe $D_t = Y_t^e$.
end if
Update posterior $P_{t+1}(\theta) = P_t(\theta \mid D_t)$.

D.1 Finite Bandit Experiment Construction

For the finite bandit scenarios (symmetric, asymmetric and strongly asymmetric) each arm's outcome probabilities are formed by starting from a uniform distribution over the discrete reward support, then adding a constant offset and Gaussian noise independently to each outcome probability. Concretely, for an arm with n_{out} outcomes, we set $p_i \leftarrow 1/(n_{\text{out}} + \alpha + \epsilon_i)$ with $\epsilon_i \sim \mathcal{N}(0, 1)$ i.i.d., α is a scalar concentration parameter (set to 10.0 by default). We then renormalise so $\sum_i p_i = 1$. Thus the only injected stochasticity is additive standard normal noise (variance 1) at the probability level before clipping/renormalization; there is no noise added directly to rewards beyond the discrete reward support itself (e.g., $-50, \ldots, 50$).

Table 3: Bandit Experiment Configuration

Parameter	Symmetric	Asymmetric	Strongly Asymmetric
Number of models	500	500	500
Number of arms	70	70	70
Horizon	20,000	4,000	4,000
Reward values	$\{-50, \ldots, 50\}$	$\{-50, \ldots, 50\}$	$\{-50, \ldots, 50\}$
Alpha	10.0	10.0	10.0
True model index	–	–	499

Table 4: Agent Pre-train Budgets and Policies

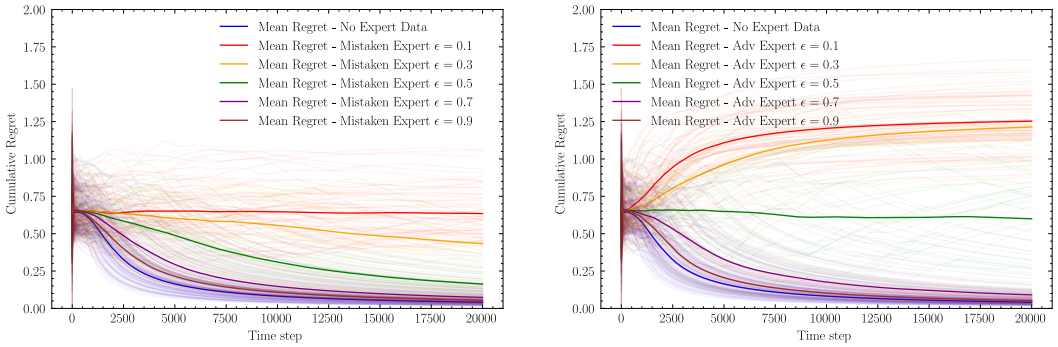
Bandit	Pretrain budgets	Online budget
Symmetric	(0,1,5,10,20,50,100,500,1000,2000)	horizon
Asymmetric	(0,1,5,10)	horizon
Strongly Asymmetric	(0,1,5,10)	horizon

D.2 Experiments on Misleading Experts

We include here empirical results on the misleading cases described in Section 5. We use the same asymmetric finite world setting as in Section 6. We compute results for the following:

- A scenario with a 'mistaken' expert, where the expert samples with probability ϵ a true optimal outcome and samples with probability $1 - \epsilon$ an outcome from a uniform action distribution over $\mathcal{A} \setminus A^*$.
- A scenario with a misleading expert, where the expert samples with probability ϵ a true optimal outcome and samples with probability $1 - \epsilon$ an outcome from an optimal action in a misleading world θ^{mis}.

Results for Misleading-Expert Experiments As discussed in Section 5, we can see how the mistaken expert, in the worst case, induces no improvement of regret, which is reasonable since it samples from all actions uniformly. As ϵ increases, the only minimiser in Θ_q becomes θ^* since this



(a) Regret when learning from mistaken expert data.

(b) Regret when learning from misleading expert data.

Figure 3: Regret obtained by TS agents with mistaken or misleading expert data.

is an asymmetric bandit class. Then, the cumulative regret still converges to zero, but at a much slower rate. For the misleading expert results, we can see how for low ϵ the regret actually increases away from the mean 'uninformed' initial value; the expert forces the agent to believe it lives in a completely different world θ^{mis}. Similarly, as ϵ increases, the set Θ_q becomes a singleton (θ^*) and the agent still manages to achieve zero regret.

D.3 Learning to Model the Expert

We include here a proposal for a particle-based algorithm that allows agents to decide whether they learn from the expert samples or their own, while modeling the expert with a general prior over expert policies.

Algorithm 2 Information Choice with Trust Inference

Initialize particles $\{(\theta^{(k)}, \pi_e^{(k)}, w_0^{(k)})\}_{k=1}^K$ from priors $P_0(\theta), P_0(\pi_e)$.
for $t = 1, ..., T$ **do**
  Compute agent's policy $\pi_t(a) = P_t(A^* = a)$ using particles.
  Estimate self-play MI, $I_s \leftarrow \mathbb{I}_t(A^*; (A_t, Y_t))$, using $\{\theta^{(k)}, w_t^{(k)}\}$.
  Estimate expert MI, $I_e \leftarrow \mathbb{I}_t(A^*; Y_t^e)$, using joint particles $\{(\theta^{(k)}, \pi_e^{(k)}, w_0^{(k)}\}_{k=1}^K$.
  Decision: $d_t \in \arg\max_{d \in \{s,e\}} \{I_s, I_e\}$
  if $d_t = s$ **then**
    Sample action $A_t \sim \pi_t$ and observe outcome Y_t.
    Update $P_{t+1}(\theta)$ using likelihood $p_{\theta, A_t}(Y_t)$.
  else
    Observe expert outcome $Y_t^e = y$.
    Compute $L^{(k)} = \sum_{a \in \mathcal{A}} \pi_e^{(k)}(a) p_{\theta^{(k)}, a}(y)$.
    Update posterior weights $w_{t+1}^{(k)} \propto w_t^{(k)} L^{(k)}$ and renormalise.
  end if
  Resample particles if necessary.
end for

D.4 True Experts

We include in this section additional plots and results for the experiments presented in Section 6.

Symmetric Worlds In *Symmetric* worlds, the induced expert-outcome predictives $P_t(Y_t^e \mid A^* = a)$ are identical across optimal-action labels under the initial prior. Then, observing an expert outcome provides no information about which action is optimal.

As shown in Figure 4, offline pre-training yields no regret improvement, confirming that expert data is uninformative when outcomes are independent of the action index. In the online budgeted

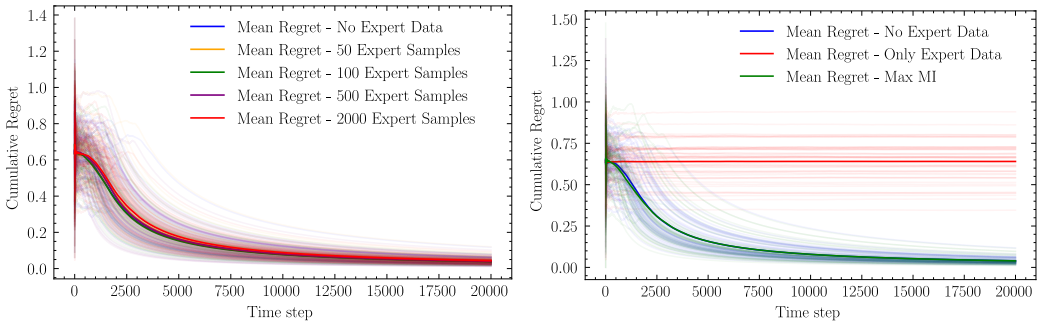


Figure 4: Regret in **Symmetric Bandits**. Left: Pretraining with expert samples. Right: Selecting information sources.

setting, learning exclusively from the expert results in linear regret. However, our MI-estimating agent successfully identifies the expert as uninformative, consistently selecting its own experience and matching the performance of standard Thompson Sampling.

Asymmetric Worlds In *Asymmetric* worlds, distributions $p_{\theta,a}$ are generated via Gaussian noise. While instances are numerically distinct, outcomes which uniquely identify the optimal action are still very numerically close to other outcomes. This represents the standard case where expert data is useful but not obviously better than self-collected data.

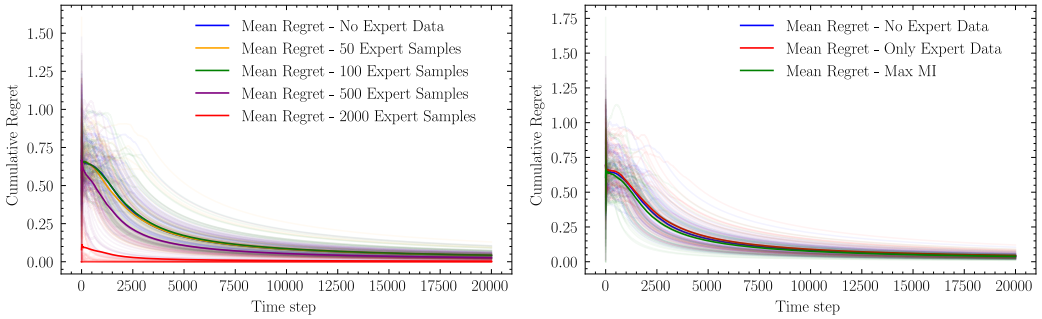


Figure 5: Regret in **Asymmetric Bandits**. Left: Pretraining with expert samples. Right: Selecting information sources.

Figure 5 demonstrates that offline access to expert data significantly reduces initial uncertainty, with 2000 samples yielding near-zero regret. In the online budgeted setting, dynamically selecting the information source via Algorithm 1 outperforms fixed strategies. By prioritizing the most informative source at each step, the agent achieves a learning rate improvement of approximately $8 - 12\%$ over standard Thompson Sampling at no additional sampling cost.

Strongly Asymmetric Worlds Finally, we consider *Strongly Asymmetric* worlds, constructed identically to the asymmetric case but with a key difference: the true parameter θ^* generates a deterministic, unique outcome y^* for the optimal action A^*. This makes the expert-outcome predictives approximately label-identifying, as in the strongly asymmetric case of Definition 4.3.

Results in Figure 6 show dramatic improvements. Because the expert outcome effectively identifies θ^*, even a single offline sample drastically warm-starts the posterior. In the online budgeted setting, the agent correctly estimates the high mutual information of the expert source. This leads to a 99% reduction in regret after 500 steps compared to standard Thompson Sampling, as the agent effectively bypasses exploration by observing the expert.

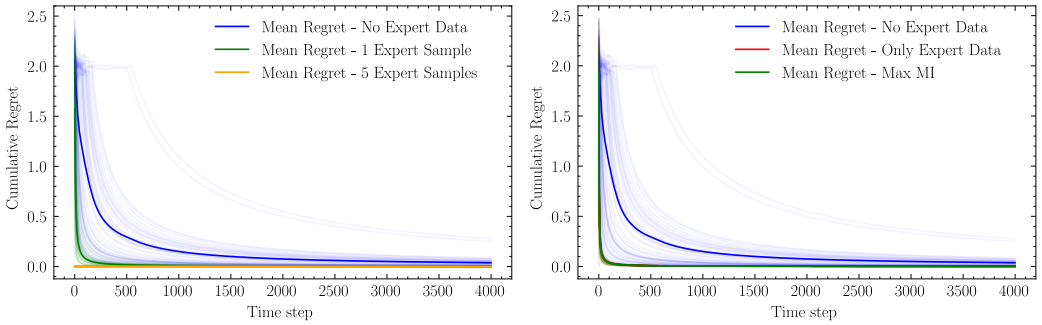


Figure 6: Regret in **Strongly Asymmetric Bandits**. Left: Pretraining with expert samples. Right: Selecting information sources.

D.5 Modeling the Expert: When to Trust

Finally, we simulate symmetric and strongly asymmetric scenarios with $M = 500$ models and $|\mathcal{A}| = 20$ actions to reduce the particle sampling requirements. We simulate three agents; one agent learns from its own data, a second agent assumes *the expert is optimal and truthful* and learns from the maximum MI source, and a third agent *models the expert* and uses Algorithm 2 to decide what source to learn from. Different to previous experiments, here the expert is *boundedly rational* with a softmax policy and temperature $\tau = 1$.

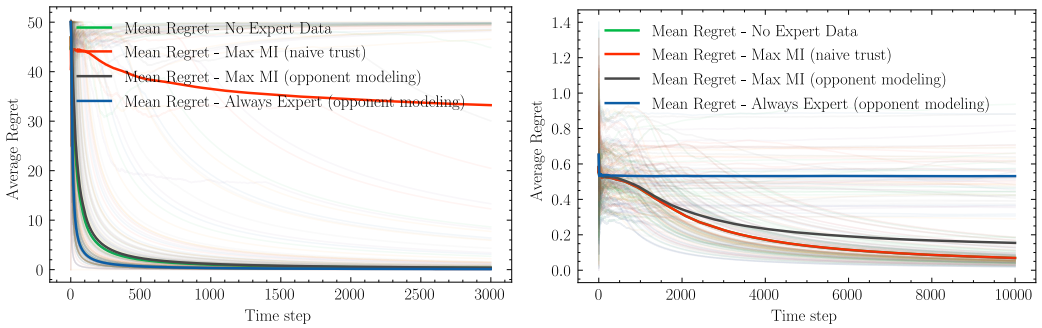


Figure 7: Regret obtained by MI estimating agents in symmetric (right) and asymmetric bandits (left).

Results for MI estimating Agents The results in Figure 7 allow us to distill two main conclusions. First, in symmetric bandits where expert information provides no gain, MI estimating agents are able to discard expert information since it provides no additional knowledge over the environment, regardless of the expert modeling misspecifications. Second, for the asymmetric problem, the naive trust agent incurs high sustained regret; it is attempting to update its posterior assuming the expert is optimal, but the expert is in fact sampling different actions, leading to model misspecifications. Interestingly, both opponent modeling agents learn to identify this, selecting their own samples as information source or making use of inferred expert policies for effective updates, confirming our hypotheses and motivations.

On Identifiability of Experts The symmetric case corresponds precisely to the unidentifiable case mentioned in Section 5.1. In the symmetric problem class, for a prior that places uniform mass on every possible parameter θ, the expert policy becomes unidentifiable. Without symmetry breaking, agents cannot distinguish what actions are generating the expert samples, as for any single θ there exists some policy that would generate the expert samples. As seen in Figure 7 (right), the agent modeling the expert and updating solely from expert data (using inferred policies $P_t(\pi_e)$) incurs linear regret. However, the MI-based agent (even under opponent modeling) learns to discriminate expert samples as uninformed (and furthermore, self-data steps will cause posterior updates that break symmetry, turning the expert data into an informative source). This confirms our conjecture at the end of Section 5.1; The MI selection rule shields agents from unidentifiable experts.

D.6 Bayesian MPC Experiment Details

We include here the implementation details for the Bayesian MPC experiments reported in Table 1. These experiments instantiate the outcome-only expert setting in episodic control problems: the learner observes trajectories generated by an expert controller, but does not observe the control inputs selected by that expert.

D.6.1 Control-Derived Contextual Bandits

Each episode is treated as one contextual bandit interaction. The context is the initial state s_0 of the controlled system. The learner selects a control policy or first control action using model-predictive control under a sampled parameter θ, executes for a fixed horizon, and observes its own trajectory and rewards. In parallel, an expert controller generates a trajectory from the same context. The expert's state trajectory is observable, but the sequence of expert actions is latent. Thus self data consists of labeled transitions and rewards, while expert data consists of outcome-only state transitions. In the reported comparison, routing determines which source is used for the posterior update; the learner still acts and receives reward in each episode.

Table 5: Bayesian MPC environment parameters.

Task	Episode length	State, actions, and unknown parameters
Point Mass	50	State (x, y, v_x, v_y); actions $(0,0), (\pm 1, 0), (0, \pm 1)$; unknown parameters act-over-mass, drag, wind direction, wind intensity, goal x, goal y

In the point-mass task, the state contains position and velocity in two dimensions, and the discrete actions are force vectors. The goal coordinates enter the reward but not the transition dynamics. Consequently, labeled state transitions generated by the learner are most informative about dynamics parameters, while the reward signal is the learner's direct source of goal information. Expert trajectories can reveal the goal more efficiently through the direction in which the expert moves.

D.6.2 Expert and Learner Models

The true expert is a stochastic Boltzmann planner with access to the true parameter θ^* and inverse temperature $\beta_{\text{true}} = 8.0$. At each state it estimates action values using MPPI and samples from the softmax policy proportional to $\exp(\beta_{\text{true}} Q(s, a))$. This stochasticity makes the outcome-only likelihood well-defined while keeping the expert close to greedy.

The learner maintains a particle posterior over θ with $N = 256$ particles. At the start of each episode it samples a particle as in Thompson Sampling and plans with MPPI. The reported runs use 50-step episodes, planning horizon $H = 5$, MPPI temperature 0.1, 256 learner rollouts per action, and 64 expert rollouts per action. Particle weights are updated by the log-likelihood of the selected data source and normalized. When the effective sample size falls below half the number of particles, systematic resampling with jitter is applied.

D.6.3 Posterior Updates

For self-generated data, the learner observes the action label and reward. For a transition (s, a, s'), the transition likelihood under particle θ is modeled as

$$p(s' \mid s, a, \theta) = \mathcal{N}\big(s'; f_\theta(s, a), \sigma_s^2 I\big)\,, \tag{16}$$

where f_θ is the deterministic dynamics model and σ_s is the transition observation scale. In point mass we also include a weak Gaussian reward likelihood, so self data can update reward-only goal parameters.

For expert data, the learner observes only the transition (s, s') from the expert trajectory. The action is marginalized using the learner's model of the expert policy:

$$p(s' \mid s, \theta) = \sum_{a \in \mathcal{A}} \pi_{\text{model}}(a \mid s, \theta) \mathcal{N}\big(s'; f_\theta(s, a), \sigma_s^2 I\big)\,. \tag{17}$$

The trajectory likelihood is the product of this marginal transition likelihood over the episode. The modeled expert policy is computed by solving the same planning problem under particle θ and applying a Boltzmann policy with inverse temperature β_{model}. The naive MI agent uses a fixed expert model. In the matched case, $\beta_{\text{model}} = \beta_{\text{true}} = 8.0$; in the misspecified case, $\beta_{\text{model}} = 0.05$ while $\beta_{\text{true}} = 8.0$. The modeled MI agent maintains a discrete belief over expert inverse temperatures and marginalizes the expert likelihood over the beta grid $\{0.05, 0.2, 1, 3, 8, 20\}$.

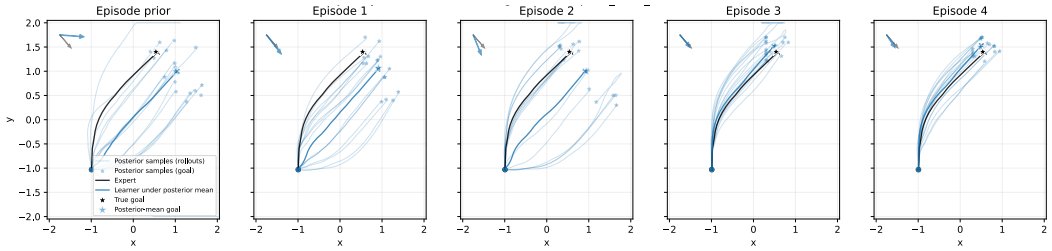


Figure 8: Trajectory evolution in the point-mass Bayesian MPC task for MI Routing with a modeled expert policy and $\beta_{\text{model}} = 8.0$. The expert trajectory is observed as outcome-only data; the learner updates its posterior and replans across episodes.

D.6.4 Routing and Baselines

The MI routing policy compares Monte Carlo estimates of the information in a self observation and an expert outcome-only observation at the current belief. It routes to the expert source when

$$\widehat{I}_{\text{expert}} > \widehat{I}_{\text{self}} + \epsilon,$$

and otherwise routes to self data. Routing is recomputed each episode. The reported results use $\epsilon = 0.004$ and 256 Monte Carlo samples per MI estimate.

We compare five update rules. **Self-Only** updates only from the learner's own labeled trajectory. **Expert-Only** updates only from the expert's outcome-only trajectory. **MI Naive** routes by mutual information while using a fixed expert model. **MI Modeled** routes by mutual information while inferring the expert's inverse-temperature parameter. **Both Updates** processes both self and expert observations in each episode. Since Both Updates consumes two datapoints per episode while the other methods consume one, we report both episode-indexed regret and regret at matched datapoint budgets.

D.6.5 Metrics and Results

Regret is measured relative to an oracle controller that uses the true parameter θ^* and a greedy planning policy. For each episode, the regret is the oracle episode reward minus the learner episode reward, averaged across seeds. The true parameter is fixed across seeds so that seed variation reflects initial states and inference randomness rather than changes in task difficulty. The point-mass summary in Table 1 uses 20 seeds.

D.7 Neural Gaussian Bandits

To demonstrate that our information-theoretic framework extends beyond finite, tabular bandits, we evaluate it on a *neural Gaussian bandit* model in which rewards are generated by a randomly initialised neural network, while the learner performs approximate Bayesian inference using a (possibly misspecified) linear model on top of learned features. The environment has rich nonlinear structure, but the learner maintains tractable Bayesian beliefs over a simpler parametric surrogate.

Environment We consider a K-armed bandit with action set $\mathcal{A} = \{1, \dots, K\}$. Each action $a \in \mathcal{A}$ has an embedding $e_a \in \mathbb{R}^{d_e}$ sampled i.i.d. from a normal distribution. The true mean reward for action a is produced by a two-layer ReLU network with random weights, $\mu^\star(a) = f_\star(e_a)$, and rewards are observed with Gaussian noise,

$$Y_t \mid A_t = a \sim \mathcal{N}\big(\mu^\star(a), \sigma^2\big). \tag{18}$$

Table 6: Bayesian MPC point-mass results under correct and misspecified expert models. Average regret is reported at matched datapoint budgets. MI Naive is an agent that believes the expert is a true expert. MI Modeled is an agent that infers the expert's β as part of the parameters.

Expert	Steps	MI (Naive)	MI (Modeled)	Self-Only	Expert-Only	Both
True	2	2.12 ± 0.93	2.26 ± 1.08	2.50 ± 1.16	1.80 ± 0.77	2.81 ± 1.30
	10	0.72 ± 0.35	0.85 ± 0.44	1.54 ± 0.84	0.54 ± 0.20	0.86 ± 0.38
	20	0.53 ± 0.23	0.57 ± 0.27	1.02 ± 0.63	0.37 ± 0.11	0.53 ± 0.19
	30	0.43 ± 0.16	0.47 ± 0.20	0.76 ± 0.42	0.31 ± 0.08	0.43 ± 0.15
	40	0.38 ± 0.11	0.41 ± 0.15	0.64 ± 0.31	0.29 ± 0.07	0.38 ± 0.11
	50	0.35 ± 0.09	0.38 ± 0.12	0.56 ± 0.25	0.27 ± 0.06	0.35 ± 0.10
Misspec.	2	2.41 ± 1.06	2.26 ± 1.08	2.50 ± 1.16	2.50 ± 0.86	2.81 ± 1.30
	10	1.56 ± 0.87	0.85 ± 0.44	1.54 ± 0.84	2.12 ± 0.84	2.10 ± 1.30
	20	0.99 ± 0.54	0.57 ± 0.27	1.02 ± 0.63	1.99 ± 0.78	1.61 ± 1.11
	30	0.74 ± 0.37	0.47 ± 0.20	0.76 ± 0.42	1.92 ± 0.72	1.27 ± 0.85
	40	0.61 ± 0.28	0.41 ± 0.15	0.64 ± 0.31	1.91 ± 0.69	1.04 ± 0.70
	50	0.54 ± 0.22	0.38 ± 0.12	0.56 ± 0.25	1.88 ± 0.65	0.89 ± 0.56

An *expert* samples an action A_t^e from a policy $\pi_e(\cdot)$ and produces an outcome-only observation $Y_t^e \sim \mathcal{N}(\mu^\star(A_t^e), \sigma^2)$. As in our main setting, the learner never observes A_t^e, only the scalar Y_t^e.

Learner Model and Approximate Bayesian Inference. The learner does not directly model the neural generator. Instead, it maintains a Bayesian linear surrogate on top of an action-dependent feature map $\phi : \mathcal{A} \to \mathbb{R}^d$,

$$\mu_w(a) \;=\; w^\top \phi(a), \qquad w \in \mathbb{R}^d, \tag{19}$$

with a Gaussian prior $w \sim \mathcal{N}(0, \lambda^{-1} I)$. We consider several feature families: (i) MLP (SMALL/GOOD): $\phi(a)$ is produced by a separate randomly initialised MLP applied to e_a (with different widths), and (ii) TRUE: $\phi(a)$ consists of the hidden activations of the true network $f_\star$ for action a (a realizable feature map). Posterior inference over w is approximated by sequential Monte Carlo (SMC) with N_p particles and resampling based on effective sample size.

Outcome-only Expert Updates When updating from *self data*, the learner observes (A_t, Y_t) and applies the labeled Gaussian likelihood

$$p(Y_t \mid w, A_t = a) \;=\; \mathcal{N}\big(Y_t;\, \mu_w(a),\, \sigma^2\big). \tag{20}$$

When updating from *expert data*, it observes only Y_t^e and must marginalize over the (unobserved) teacher action:

$$p(Y_t^e \mid w, \pi_e) \;=\; \sum_{a \in \mathcal{A}} \pi_e(a)\, \mathcal{N}\big(Y_t^e;\, \mu_w(a),\, \sigma^2\big). \tag{21}$$

We study both *naive* expert trust, which incorrectly treats Y_t^e as if generated by the learner's believed optimal arm, and *modeled* expert trust, which maintains uncertainty over π_e and uses the marginal likelihood (21). In particular, in our fully Bayesian expert modeling variant we place a prior $\pi_e \sim \mathrm{Dir}(\alpha_0)$ and approximate the joint posterior over (w, π_e) using joint particles.

Experimental Setting We evaluate a set of agents across feature-map classes and expert types. **Self-only**: update only from (A_t, Y_t). **Expert-only (naive)**: update only from Y_t^e under naive trust. update only from Y_t^e using a model of π_e. **MI routing (naive)**: choose between sources using MI, but process expert outcomes naively. **MI routing (modeled)**: choose between sources using MI with expert modeling.

Experts are either OPTIMAL (greedy), SOFT (Boltzmann with temperature τ), or BAD (play a random fixed action). We report average regret $\frac{1}{t}\sum_{s=1}^{t}(\max_a \mu^\star(a) - \mu^\star(A_s))$ over multiple random seeds, along with diagnostics such as the posterior entropy of the optimal-action distribution and the fraction of rounds in which each agent routes to the expert. These experiments test two claims: (i) MI-based routing continues to provide effective information selection under function approximation, and (ii) explicitly modeling expert reliability mitigates the failure modes of naive expert trust under misspecified expert policies.

D.7.1 Experiment Configuration

In the neural Gaussian experiments, each bandit instance is generated by sampling a latent "true" reward network and then defining Gaussian rewards around its per-action means. For each run we draw action embeddings $z_a \sim \mathcal{N}(0, I)$ in $\mathbb{R}^{d_{\text{embed}}}$ (default $d_{\text{embed}} = 8$), pass them through a 2-layer ReLU MLP with weights $W_1 \sim \mathcal{N}(0, 1/\sqrt{d_{\text{embed}}})$, $W_2 \sim \mathcal{N}(0, 1/\sqrt{d_{\text{hidden}}})$ and zero biases, and set the true mean reward to $\mu_a = 5 \cdot \text{MLP}(z_a)$ (default hidden width $d_{\text{hidden}} = 64$). Each observed reward is then $r \sim \mathcal{N}(\mu_a, \sigma^2)$ with $\sigma = 0.3$. The expert policy is derived from μ: optimal is deterministic argmax, soft uses a Boltzmann distribution $\propto \exp(\mu_a/\tau)$ with $\tau = 0.5$, and bad mixes a fixed "best" action 0 and fixed "worst" action 1 with probability 0.5 each. The agent's feature map is a separate randomly initialized mapping over actions: TRUE uses the hidden activations of the true network plus a bias term (dimension $d_{\text{hidden}}+1$, LINEAR uses the identity over actions, GOOD uses an independent random MLP feature map with hidden/output dimensions 64/64, and SMALL uses 16/16. All random draws are seeded per run. When updating the particle filter for estimating the posteriors $P_t(\theta, \pi_e)$, we compute the effective sample size as $1/\sum_i w_i^2$. If it is smaller than resample_threshold $\times N$, we resample and jitter particles. For resampling, we draw N particle indices with replacement using the current weights, we replace both the weight particles θ and the expert-policy particles π with the resampled copies and reset all weights to uniform $1/N$. For jittering, in the continuous weight particles θ we add Gaussian noise ($w \leftarrow w + \epsilon; \epsilon \sim \mathcal{N}(0, \text{resample_jitter_w}^2)$) elementwise. For the discrete policy particles π we compute $\alpha = \text{resample_jitter_pi_conc} \cdot \pi$ (row-wise normalized) and resample $\pi \sim \text{Dirichlet}(\alpha)$

Table 7: Neural Gaussian Experiment Parameters (Defaults)

Parameter	Default	Description
horizon	5000	Time steps per episode
n_actions	10	Number of actions
action_embed_dim	8	Action embedding dimension
true_hidden_dim	64	Hidden units in true reward network
reward_sigma	0.3	Std dev of reward noise
prior_precision	1.0	Prior precision for linear regression
dirichlet_alpha	1.0	Dirichlet prior concentration for expert policies
n_particles	1000	Posterior particles for belief tracking
mi_samples	50	Monte Carlo samples per MI estimate
mi_every	2	MI recompute interval
resample_threshold	0.3	Resample threshold
resample_jitter_w	0.05	Weight jitter after resampling
resample_jitter_pi_conc	100.0	Dirichlet refresh concentration
feature_type	good, small, true	Feature map families
teacher_type	optimal, soft, bad	Teacher policy types
teacher_tau	0.5	Soft teacher temperature
routing_method	Self, Expert-Only, MI-routing	Routing policy (see Table 8)

Table 8: Routing Methods (Default Sweep)

Method	Description (high level)
self_only	Use agent action only
expert_only_naive	Always follow expert; no modeling
expert_only_modeled	Always follow modeled expert
mi_routing_naive	Route via MI with naive expert
mi_routing_modeled	Route via MI with modeled expert

D.7.2 Results

To test whether our outcome-only expert framework scales beyond finite tabular models, we run experiments on *Neural Linear Gaussian Bandits* Snoek et al. [2015], Riquelme et al. [2018], where the true arm means are generated by a randomly initialised two-layer ReLU network applied to learned action embeddings, and observed through Gaussian noise. The learner maintains an approximate

Bayesian posterior over a set of last-layer weights w, $\mu_w(a) = w^\top \phi(a)$, using particles. We vary the feature map ϕ to control misspecification, including random MLP features of different capacities, and a realizable true map derived that matches the true generator[6]. We compare fixed baselines (self-only, expert-only) with our adaptive *information choice* rule (Algorithm 1). Across expert types (true, soft and mistaken) and feature-map regimes, these experiments show that MI-based routing remains effective under function approximation and that modeling the expert mitigates the brittleness of naive trust. Figure 9 presents a summary of the average regrets across different learner and expert profiles (Wrong expert, Soft-Optimal Expert) for the neural gaussian setting.

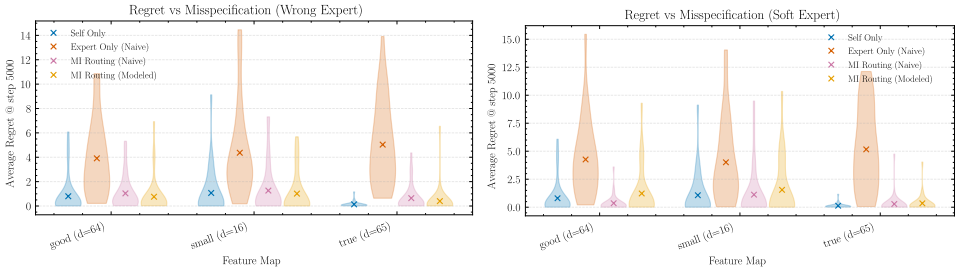


Figure 9: Neural Gaussian Bandits Regrets.

[6]Extended experimental details and results are in Appendix D.7.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the outcome-only expert setting, the offline Bayesian update and regret-bound results, the budgeted online routing rule, and the main assumptions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The discussion and appendices describe limitations including finite/exact-update assumptions, model misspecification, expert identifiability, and computational costs of expert likelihoods and MI estimation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The main theorem statements list their assumptions, and proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The experimental sections and appendices specify the synthetic bandit, neural Gaussian, and Bayesian MPC setups, including model classes, baselines, particle approximations, and evaluation metrics.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The current submission provides experimental details but does not yet include an anonymized public code release or exact reproduction commands, to be published with acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The appendices give the environment constructions, hyperparameters, posterior approximations, routing baselines, and evaluation protocols needed to understand the reported experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: Experimental results are averaged over multiple random seeds and reported with standard-deviation error bars in the relevant figures and tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The paper discusses computational overhead as well as hardware used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is theoretical and synthetic/benchmark-based and does not involve human subjects, private data, or deployment claims that would violate the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A].

Justification: The work is mostly foundational and theoretical. As any such work, there are possibilities for misuse, but none are immediate or evident from the work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release high-risk models, scraped datasets, or other assets requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper cites relevant prior work and frameworks and respects licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not introduce or release a new dataset, model checkpoint, or benchmark asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The research does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The research does not involve crowdsourcing or human subjects, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs are not used as an important, original, or non-standard component of the core method or experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.