IMPLICIT THOMPSON SAMPLING VIA IMAGINED REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Thompson sampling (TS) is an efficient algorithm for addressing the exploration-exploitation trade-off in sequential decision-making. However, its application to modern auto-regressive predictive models (e.g., LLMs and TabPFN) is limited, since it requires sampling from the posterior of an *explicit* Bayesian model. To address this challenge, we propose Imagined Reward Thompson Sampling (IRTS) as a family of TS-style algorithms, based on the insight that in-context learning with autoregressive models provides predictive inference for an *implicitly defined* Bayesian model. IRTS uses autoregressive models to sample τ *imagined rewards* autoregressively per action, and acts greedily w.r.t. either their averaged rewards (IRTS-A), or the predictive mean reward conditioned on the observed and imagined rewards (IRTS-C). We provide a regret analysis and prove that IRTS-A and IRTS-C provide a "sandwich" of exploration behaviour over TS, where both algorithms approach TS when $\tau \to \infty$. With finite τ, IRTS-A over-explores due to over-estimation of posterior uncertainty, while IRTS-C under-explores by interpolating between greedy and TS. Experiments across multi-armed and contextual bandits, with analytic, TabPFN and LLM-based in-context learners, demonstrate that IRTS achieves efficient TS-style exploration with implicitly-defined Bayesian models.

1 INTRODUCTION

Sequential decision-making under uncertainty has long been studied through the lens of Bayesian inference, where the learner maintains a posterior distribution over a latent parameter and uses it to balance exploration and exploitation (Lattimore & Szepesvári, 2020). This framework underpins provably efficient algorithms, most notably Thompson Sampling (TS) (Thompson, 1933; Russo et al., 2018). However, this paradigm is challenging to reconcile with large modern machine learning architectures, where explicit posterior inference is often intractable (Riquelme et al., 2018). In fact, the remarkable expressivity of foundation models such as LLMs (Brown et al., 2020) and prior-fitted networks (Müller et al., 2021; Hollmann et al., 2025), together with their ability to learn *in-context* – that is, by conditioning on new data without retraining – makes them uniquely promising as flexible models for sequential decision-making. Indeed, prior work has considered in-context decision-making tasks, where simple instruction-based approaches fail (Krishnamurthy et al., 2024; Nie et al., 2024), again showing the need for efficient exploration in this setting. On the other hand, recent work argues that predictions from such in-context learners can be interpreted as arising from an *implicitly defined* Bayesian model, with their predictive distributions corresponding to a marginalisation over implicitly defined latent parameters (Xie et al., 2021; Chlon et al., 2025; Falck et al., 2024; Ye & Namkoong, 2024). To this end, we ask the following research question:

Can we exploit the implicit Bayesian model to develop a computationally tractable, TS-style efficient exploration algorithm for in-context decision-making?

We show an affirmative answer, by proposing Imagined Reward Thompson Sampling (IRTS), a framework that employs a finite *imagined-reward* sampling budget compared to TS. Here imagined rewards are sampled autoregressively conditioned on observed history. In the limit of infinitely many imagined rewards, this scheme recovers a policy identical to that derived from TS (Fong et al., 2023; Ng et al., 2025; Liu et al., 2023; Zhang et al., 2025). For computational tractability, we introduce two practical algorithms, IRTS-A and IRTS-C, each with its own distinct exploration-exploitation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

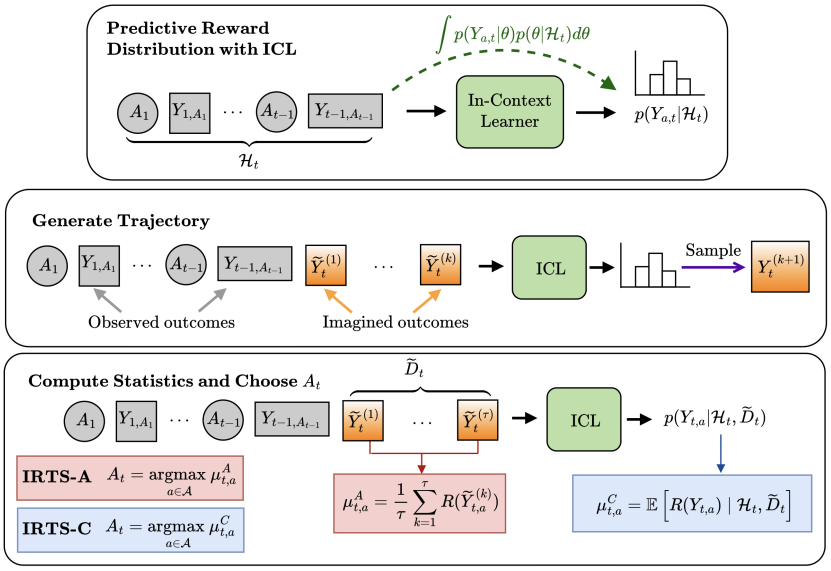

Figure 1: Overview of IRTS framework. (Top): An in-context learner predicts a reward distribution given an arm and the observed history $\mathcal{H}_t$. Under exchangeability, this corresponds to predictive inference in an implicit Bayesian model. (Middle): IRTS autoregressively generates one shared synthetic dataset $\widetilde{D}_t$, appending each imagined arm-outcome pair before drawing the next; the block contains τ imagined rewards for each arm. (Bottom): IRTS-A scores each arm by its empirical mean reward, while IRTS-C uses the conditional mean reward given the complete observed and synthetic history. The arm with the largest score is selected.

profile induced by the finiteness of the imagined reward trajectory samples: IRTS-A acts greedily with respect to the empirical *average reward* of the imagined outcomes, leading to *over-exploration* compared to TS, whereas IRTS-C acts greedily with respect to the model's conditional mean given the imagined and observed outcomes, leading to *under-exploration* relative to TS.

In summary, our contributions are:

1. Two finite-compute approximate TS algorithms, IRTS-A and IRTS-C, that operate purely through a predictive distribution and require no access to a parameter posterior;
2. A Bayesian regret analysis and variance characterisations for both algorithms showing the two methods bracket exact TS from above and below in exploration. Our results also unify and clarify the theoretical standing of related recent work (Sun et al., 2025; Liu et al., 2023);
3. Experiments across multi-armed and contextual bandits, with analytic, TabPFN and LLM-based in-context learners that demonstrate that IRTS-C extends efficient TS-style exploration to implicit Bayesian models including modern foundation models.

2 Background

Bayesian bandits. We consider a stochastic multi-armed bandit (MAB) with $K \geq 2$ arms, indexed by $[K] = \{1, \ldots, K\}$ (Lattimore & Szepesvári, 2020). The unknown environment parameter is drawn once from a prior, $\theta^\star \sim \pi$. Each round has an outcome vector

$$Y_t = (Y_{t,a})_{a\in[K]} \in \mathcal{Y}^K.$$

Conditional on $\theta^\star = \theta$, the sequence $(Y_t)_{t\geq 1}$ is i.i.d. with joint outcome density $p_Y(\cdot \mid \theta)$, and $p_{Y_a}(\cdot \mid \theta)$ denotes its arm-a marginal. The agent associates each outcome $y \in \mathcal{Y}$ with a reward through a fixed, known map $R : \mathcal{Y} \to \mathbb{R}$. The mean reward of arm a is

$$\mu_a(\theta) = \int_{\mathcal{Y}} R(y) p_{Y_a}(y \mid \theta)\, dy.$$

For the Bayesian regret bounds in Section 3, we assume $R(Y_{t,a}) \in [0,1]$ almost surely for every arm. The coordinates of Y_t, and hence the arm outcomes and means, may be dependent; only independence across rounds conditional on $\theta^\star$ is assumed. Let

$$\mu^\star(\theta) = \max_{a\in[K]} \mu_a(\theta), \qquad A^\star \in \arg\max_{a\in[K]} \mu_a(\theta^\star),$$

where a fixed rule is used to break ties. At round $t = 1, \dots, T$, the learner selects $A_t \in [K]$ and observes Y_{t,A_t}, receiving reward $R(Y_{t,A_t})$. The parameter posterior given history before round t is

$$\pi_t(\theta) = p(\theta^\star \mid \mathcal{H}_t), \quad \mathcal{H}_t = (A_1, Y_{1,A_1}, \dots, A_{t-1}, Y_{t-1,A_{t-1}}), \qquad \mathcal{H}_0 = \varnothing.$$

Following the information-theoretic convention of Russo & Van Roy (2016), the Bayesian regret through time T is

$$\mathcal{R}(T) := \mathbb{E}\left[\sum_{t=1}^{T} \left(R(Y_{t,A^\star}) - R(Y_{t,A_t})\right)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left(\mu^\star(\theta^\star) - \mu_{A_t}(\theta^\star)\right)\right]. \tag{1}$$

Contextual bandits. Contextual bandit augments the MAB task with state information. At each timestep t, the learner first observes a context $X_t \in \mathcal{X}$ sampled from a context distribution ρ, then chooses an action $A_t \in [K]$, observes Y_{t,A_t}, and receives reward $R(Y_{t,A_t})$. The expected reward of action a now depends on the context, with

$$\mu_a(x;\theta) = \mathbb{E}[R(Y_{t,a}) \mid X_t = x, \theta^\star = \theta].$$

Accordingly, the optimal value is $\mu^\star(x;\theta) = \max_{a\in[K]} \mu_a(x;\theta)$, and regret is defined relative to the best action for the observed context.

Thompson Sampling (TS). TS is a popular heuristic strategy for balancing exploration/exploitation, which acts greedily with respect to posterior samples. At each round t, the learner samples $\theta_t \sim \pi_t$ and acts greedily with respect to this draw, choosing an arm that maximises $\mu_a(\theta_t)$. This is summarised in Algorithm 1. TS is also known as 'probability-matching', as this induces a policy that picks an action with probability equal to the probability of it being the optimal action under the Bayesian model (Russo & Van Roy, 2016). A key theoretical property of TS is that it has provably optimal asymptotic regret, thus efficiently balancing exploration and exploitation (Agrawal & Goyal, 2017).

Implicit Bayesian Models and Sequential Sampling. Throughout this work, we are concerned with *in-context* learners. For our purposes, these are models that provide a predictive distribution $p(z_{n+1} \mid \mathcal{D})$ given observations $\mathcal{D} = \{z_i\}_{i=1}^n$ without further retraining. This kind of in-context predictive behaviour can be observed in some flexible foundation models, such as LLMs and prior-fitted networks. There is a growing body of work (Xie et al., 2021; Müller et al., 2021; Ye & Namkoong, 2024) that argues that, under certain assumptions, these models' predictive distributions can be interpreted as arising from some underlying, implicitly defined Bayesian model. In particular, if the predictions satisfy infinite exchangeability, that is $p(z_{\sigma(1)}, \dots, z_{\sigma(n)}) = p(z_1, \dots, z_n)$ for any permutation σ and any $n \in \mathbb{N}^+$, then by de Finetti's theorem (De Finetti, 1929) there exists an underlying latent parameter θ, with a prior $p(\theta)$ and likelihood $p(z_i \mid \theta)$ such that $p(z_{1:n}) = \int \prod_{i=1}^n p(z_i \mid \theta) p(\theta) d\theta$. We refer to models that admit such a factorisation as having an *implicit* Bayesian model even if the distribution over θ is never directly accessible or defined.

While the existence of such a model can be guaranteed under these assumptions, we do not have direct access to it from the in-context learner, as in this framework the latter only provides the marginalised predictive distribution over observations, i.e. $p_Y(y \mid \mathcal{D})$ rather than the full posterior distribution $p(\theta \mid \mathcal{D})$. For this reason, we do not assume access to posterior distribution samples or even a definition of the Bayesian model, but instead only assume access to the predictive distribution.

The following result establishes the equivalence of latent-variable and autoregressive sampling.

Proposition 2.1 (Equivalence of latent-variable and autoregressive sampling)**.** *Let $\theta \sim p(\theta)$, and assume $z_{1:n}$ are conditionally i.i.d. when conditioned on θ. Then the following methods for sampling $z_{1:n}$ from $p(z_{1:n}) = \int p(\theta) \prod_{i=1}^n p(z_i \mid \theta)\, d\theta$ are equivalent (in distribution):*

1. (Latent-variable sampling) Sample $\theta \sim p(\theta)$, then sample $z_i \overset{iid}{\sim} p(\cdot \mid \theta)$ for $i = 1, \dots, n$;
2. (Autoregressive sampling) Sample sequentially $z_i \sim p(\cdot \mid z_{1:i-1})$ for $i = 1, \dots, n$.

See Appendix A.1 for a proof. Autoregressive sampling can be implemented in practice in the in-context learning setting to enable us to obtain joint predictive posterior samples from implicit Bayesian models, while latent-variable sampling provides a convenient tool for our theoretical analysis.

Algorithm 1 Thompson Sampling

Require: Arms $[K]$, horizon T

for $t = 1, \dots, T$ **do**
  $\mathcal{H}_t \leftarrow (A_1, Y_{1,A_1}, \cdots, A_{t-1}, Y_{A_{t-1}})$
  Draw $\theta_t \sim \pi_t$
  for $a = 1, \cdots, K$ **do**
    $\mu_a(\theta_t) \leftarrow \mathbb{E}[R(Y_{t,a}) \mid \theta_t]$
  end for
  $A_t \in \arg\max_{a \in [K]} \mu_a(\theta_t)$
  Play A_t and observe Y_{t,A_t}
end for

Algorithm 2 IRTS-A/C

Require: Arms $[K]$, horizon T, budget τ, mode $M \in \{A, C\}$

for $t = 1, \dots, T$ **do**
  $\mathcal{H}_t \leftarrow (A_1, Y_{1,A_1}, \cdots, A_{t-1}, Y_{A_{t-1}})$, $\widetilde{D}_t \leftarrow \varnothing$
  for $i = 1, \dots, \tau$ **do**
    $\widetilde{Y}_t^{(i)} \leftarrow (\widetilde{Y}_{t,a}^{(i)})_{a \in [K]} \sim p_Y\left(\cdot \mid \mathcal{H}_t, \widetilde{D}_t\right)$
    $\widetilde{D}_t \leftarrow \widetilde{D}_t \cup \{\widetilde{Y}_t^{(i)}\}$
  end for
  for $a = 1, \dots, K$ **do**
    if $M = A$: $\widehat{\mu}_{t,a}^{A} \leftarrow \tau^{-1} \sum_{i=1}^{\tau} R(\widetilde{Y}_{t,a}^{(i)})$
    if $M = C$: $\widehat{\mu}_{t,a}^{C} \leftarrow \mathbb{E}[R(Y_{t,a}) \mid \mathcal{H}_t, \widetilde{D}_t]$
  end for
  $A_t \in \arg\max_{a \in [K]} \widehat{\mu}_{t,a}^{M}$
  Play A_t and observe Y_{t,A_t}
end for

3 Algorithms and Theoretical Results

We take inspiration from the fact that uncertainty about an implicit Bayesian model can be reduced by conditioning on a sufficiently long sequence of observations. At round t, IRTS generates a block of $K\tau$ imagined outcomes, containing τ outcomes for each arm. The block is sampled autoregressively: every newly imagined action-outcome vector pair is appended to the synthetic dataset $\widetilde{D}_t$ before the next outcome is drawn. By Proposition 2.1, this autoregressive generation is equivalent to first drawing one $\theta_t \sim \pi_t$ and then drawing τ independent "imagined" outcomes conditioned on θ_t,

$$\widetilde{Y}_t^{(1)}, \dots, \widetilde{Y}_t^{(\tau)} \sim p_Y(\cdot \mid \theta_t).$$

From the synthetic observations, it is then possible to infer an approximation to the usual TS score $\mathbb{E}[R(Y_t) \mid \theta_t]$ used to carry out action selection. We propose two IRTS variants that construct this approximation in two different ways:

- IRTS-A takes the empirical *average reward* of the imagined outcomes,

$$\widehat{\mu}_{t,a}^{A} = \frac{1}{\tau} \sum_{i=1}^{\tau} R(\widetilde{Y}_{t,a}^{(i)}). \tag{2}$$

- IRTS-C takes the *conditional mean reward* given the observed history and the complete synthetic dataset $\tilde{D}_t = \{\widetilde{Y}_t^{(1)}, \dots, \widetilde{Y}_t^{(\tau)}\}$,

$$\widehat{\mu}_{t,a}^{C} = \mathbb{E}\left[R(Y_{t,a}) \,\middle|\, \mathcal{H}_t, \widetilde{D}_t\right]. \tag{3}$$

In equation 3, the expectation is taken under the model's posterior-predictive continuation after treating $\widetilde{D}_t$ as additional data. The selected arm is $A_t \in \arg\max_{a \in [K]} \widehat{\mu}_{t,a}^{M}$, for method $M \in \{A, C\}$, using a fixed tie-breaking rule. Intuitively, with large enough τ, both variants of IRTS would recover the reward of the true posterior sample $\mathbb{E}[R(Y_{t,a}) \mid \theta_t]$ to a high degree of confidence, thus approximating TS without ever requiring an explicit posterior sample.

We remark that IRTS can be interpreted as a finite-compute adaptation of Algorithm 1 in Liu et al. (2023), which is shown to be equivalent to exact TS when $\tau = \infty$. In addition, IRTS-A with $\tau = 1$ corresponds to the recently proposed TS-LLM algorithm (Sun et al., 2025). Despite the increasing interest in predictive versions of TS (Cai et al., 2024; Zhang et al., 2025; Tan et al., 2026), to the best of our knowledge there is no analysis on the effect that finite synthetic data generation has on these algorithm. For the analogous contextual version of the IRTS-A and IRTS-C algorithms, see Appendix B.

3.1 Theoretical results on Bayesian regret

We first compare the distributions of the IRTS scores with the posterior mean samples used by exact TS. Proofs of the following results are in Appendix A.

Lemma 3.1 (IRTS-A variance)**.** *Let $\theta_t \sim \pi_t$, and define $\sigma_a^2(\theta) = \mathrm{Var}_{Y \sim p_{Y_a}(\cdot|\theta)}(R(Y))$. Then*

$$\mathbb{E}[\widehat{\mu}_{t,a}^A \mid \mathcal{H}_t] = \mathbb{E}[\mu_a(\theta_t) \mid \mathcal{H}_t], \quad \mathrm{Var}(\widehat{\mu}_{t,a}^A \mid \mathcal{H}_t) = \mathrm{Var}(\mu_a(\theta_t) \mid \mathcal{H}_t) + \frac{1}{\tau}\mathbb{E}[\sigma_a^2(\theta_t) \mid \mathcal{H}_t]. \quad (4)$$

Note that $\mathrm{Var}(\mu_a(\theta_t) \mid \mathcal{H}_t)$ is the variance of exact TS posterior mean samples. Thus IRTS-A uses scores with the same mean and at least as much conditional variance as exact TS, and the inequality is strict in general stochastic settings where $\mathbb{E}[\sigma_a^2(\theta_t) \mid \mathcal{H}_t] > 0$.

Lemma 3.2 (IRTS-C variance)**.** *Under the same notation,*

$$\mathbb{E}[\widehat{\mu}_{t,a}^C \mid \mathcal{H}_t] = \mathbb{E}[\mu_a(\theta_t) \mid \mathcal{H}_t], \quad \mathrm{Var}(\widehat{\mu}_{t,a}^C \mid \mathcal{H}_t) \leq \mathrm{Var}(\mu_a(\theta_t) \mid \mathcal{H}_t). \quad (5)$$

Moreover, the conditional variance of $\widehat{\mu}_{t,a}^C$ given $\mathcal{H}_t$ is nondecreasing in τ.

The variance analyses show that IRTS-A/C entails over-/under/exploration behaviour when compared with exact TS. We proceed to bound this excess regret in terms of τ. In the following, $I(X;Y)$ denotes the mutual information of random variables X and Y.

Theorem 3.3 (Bayesian regret of IRTS-A)**.** *The Bayesian regret of IRTS-A is upper bounded by*

$$\mathcal{R}(T) \leq T\sqrt{\frac{\log K}{2\tau}} + \sqrt{\frac{KT\log K}{2}}. \quad (6)$$

Theorem 3.4 (Bayesian regret of IRTS-C)**.** *The Bayesian regret of IRTS-C is upper bounded by*

$$\mathcal{R}(T) \leq T\sqrt{\frac{\log K}{2\tau}} + \sqrt{\frac{KT}{2}I(\theta^\star; \mathcal{H}_{T+1})}. \quad (7)$$

Theorem 3.5 (Bayesian regret of IRTS for contextual bandits.)**.** *The Bayesian regret of both IRTS-A and IRTS-C for contextual bandits with i.i.d. contexts sampled each round is upper bounded by*

$$\mathcal{R}(T) \leq T\sqrt{\frac{\log K}{2\tau}} + \sqrt{\frac{KT}{2}I(\theta^\star; \mathcal{H}_{T+1})}. \quad (8)$$

The bounds separate two cost terms. The first term is due to the approximation error of a finite imagined-reward block and vanishes as $\tau \to \infty$. The second term scales like the usual exact TS $\sqrt{T}$ regret (Russo & Van Roy, 2016; Neu et al., 2022). The difference in the $\sqrt{T}$ coefficient between IRTS-A and IRTS-C is due to the fact that the score for IRTS-C is not conditionally independent of the history given θ_t, thus requiring a slightly different proof strategy. The IRTS-C and contextual sublinear bound contributions are consistent with information-theoretic analyses, where the same $\sqrt{T}$ coefficient is recovered as in Neu et al. (2022). $I(\theta^*|\mathcal{H}_{T+1})$ can be bound by $H(A^\star) + H(\theta^\star \mid A^\star)$ if $\theta^\star$ is discrete with finite Shannon entropy. For continuous θ^*, suitable regularity and boundedness assumptions yield a useful $I(\theta^*; \mathcal{H}_{T+1}) = O(\log T)$ bound, as is discussed in detail in Neu et al. (2022).

Together with Lemmas 3.1 and 3.2, these results support the following interpretation:

- IRTS-A uses noisier scores than exact TS and can over-explore. At $\tau = 1$, its score includes both epistemic uncertainty and irreducible reward noise. This is consistent with the TS-LLM algorithm in (Sun et al., 2025) requiring heuristic sampling temperature varying throughout an episode to tune exploration, as it corresponds to IRTS-A with $\tau = 1$.
- IRTS-C interpolates between greedy ($\tau = 0$) and exact TS ($\tau = \infty$) policies, where for finite $\tau \geq 1$, we obtain a "greedier" version of TS. In practice TS can over-explore (Jin et al., 2023; Wang et al., 2005), so we expect IRTS-C to be more practically performant than IRTS-A.

4 Related work

Thompson sampling. Thompson sampling (TS) balances exploration and exploitation by choosing actions according to the probability that they are optimal given the observed data so far (Thompson, 1933; Lattimore & Szepesvári, 2020; Russo & Van Roy, 2016). This leads to asymptotically provably optimal regret under certain assumptions (Lai & Robbins, 1985; Agrawal & Goyal, 2017), making the TS heuristic a powerful principle that immediately generalises to the contextual setting (Neu et al.,

2022). Empirically, TS shows strong performance compared to other exploration heuristics, such as UCB and epsilon-greedy (Chapelle & Li, 2011).

A drawback of traditional Thompson sampling is its reliance on a manually specified Bayesian model and having access to posterior samples Liu & Li (2016). Due to its strong empirical performance, extending its applicability beyond simple Bayesian models has been a major line of work. Some of the contexts in which analysis of TS has been extended to have been TS with neural approximation (Zhang et al., 2021; Riquelme et al., 2018), approximate inference (Phan et al., 2019) and misspecified models (Jaiswal et al., 2023). In particular, Riquelme et al. (2018) evaluate a range of Bayesian deep learning methods on contextual bandit problems and highlight the necessity for robust online uncertainty estimation. While most existing approaches retain an explicit parametric form to enable inference, there is growing interest in Thompson Sampling-style methods that do not require direct access to model parameters (Wen et al., 2021). Such approaches have been explored in settings including non-stationary bandits (Liu et al., 2023) and meta-contextual bandits (Zhang et al., 2025). This line of work is particularly promising for integrating large black-box foundation models into online decision-making, and our approach builds on this paradigm.

In-context bandits. A recent line of work seeks to use foundation models' rich prior knowledge and in-context capabilities to drive online decision-making. In particular, there is increasing interest in empirical evaluation of LLMs as in-context decision-makers. One line of research involves structuring the prompt with instructions, asking the LLM to pick the next action (Song et al., 2025; Nie et al., 2024; Lim et al., 2025; Xia et al., 2025) The other uses LLMs to make predictions about outcomes, and decision-making is carried out on top of these predictions (Sun et al., 2025; Berdica et al., 2026). Beyond LLMs, in-context decision-making has been investigated with pre-trained architectures such as in Lee et al. (2023); Laskin et al. (2022), and Tan et al. (2026) approximate TS with PFNs using a Gaussian approximation to the mean-reward posterior based on observed data under predictive CLT assumptions.

Foundation models as Bayesian predictors. The necessity to specify a prior-likelihood pair and the computational difficulties in finding the corresponding posterior limit their scalability. A complementary view that has recently become popular is that of Bayesian predictive inference (BPI) (Fortini & Petrone, 2025). In this framework, the Bayesian model is not explicitly defined in terms of prior and likelihood, but rather inferred from a model's predictive distribution. In a foundation model context predictive distributions are readily available, for example with LLMs (Requeima et al., 2024) or prior fitted networks (Hollmann et al., 2025), whereas explicit parametric forms may be unavailable. BPI allows us to extend Bayesian reasoning and uncertainty quantification to these models (Fong et al., 2023; Jayasekera et al., 2025; Xie et al., 2021). We leverage the idea of using the predictive distribution as a proxy for the internal Bayesian model as foundation for our work.

5 Experiments

We empirically evaluate the exploration efficiency of IRTS-A/C on (1) analytic bandit cases, where we focus on comparing with exact TS, as well as in-context (contextual) bandits with (2) TabPFN and (3) LLM models. When exact TS is inapplicable, we consider alternative baselines such as TS with a linear model in TabPFN experiments, and an instruction baseline in LLM tasks following Krishnamurthy et al. (2024). Further experimental settings and results are available in Appendix C.

5.1 Analytic models

We consider a range of Bayesian bandit environments and models where both the posterior predictive and posterior distribution itself are available in closed-forms. This allows us to compare IRTS-A/C with exact TS and characterise empirically their exploration behaviour. Specifically, we simulate three MAB tasks , **Bernoulli**, **Gaussian**, **Mismatch**, and one contextual bandit task, **Contextual**, where they all have $K = 5$ arms, and their environment and model assumptions are listed below. In these experiments the outcomes are scalar and $R(y) = y$, so outcome and reward values coincide. We evaluate IRTS-A/C for $\tau \in \{1, 5, 10\}$, with greedy (w.r.t. the predictive mean reward conditioned on $\mathcal{H}_t$, equivalent to $\tau = 0$) and exact TS as baselines. Each algorithm is evaluated across 100 seeds, with each episode lasting one million steps. See Table 1 for a summary.

Table 1: Summary of analytic bandit environments and models

Task	Environment	Model
Bernoulli	$Y_{t,a} \sim \text{Bern}(\theta_a),\ \theta_a \sim \text{Beta}(1,1)$	No Model Mismatch
Gaussian	$Y_{t,a} \sim \mathcal{N}(\theta_a, 1),\ \theta_a \sim \mathcal{N}(0,1)$	No Model Mismatch
Mismatch	$Y_{t,a} \sim \text{Bern}(\theta_a),\ \theta_a \sim \text{Beta}(1,1)$	$Y_{t,a} \sim \mathcal{N}(\theta_a, 1),\ \theta_a \sim \mathcal{N}(0,1)$
Contextual	$\theta_a \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \theta_a^{(0)} \sim \mathcal{N}(0,1),\ X \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ $Y_{t,a} \mid X_t \sim \mathcal{N}(\theta_a^\top X_t + \theta_a^{(0)}, 1) \qquad (X_t \in \mathbb{R}^5)$	No Model Mismatch

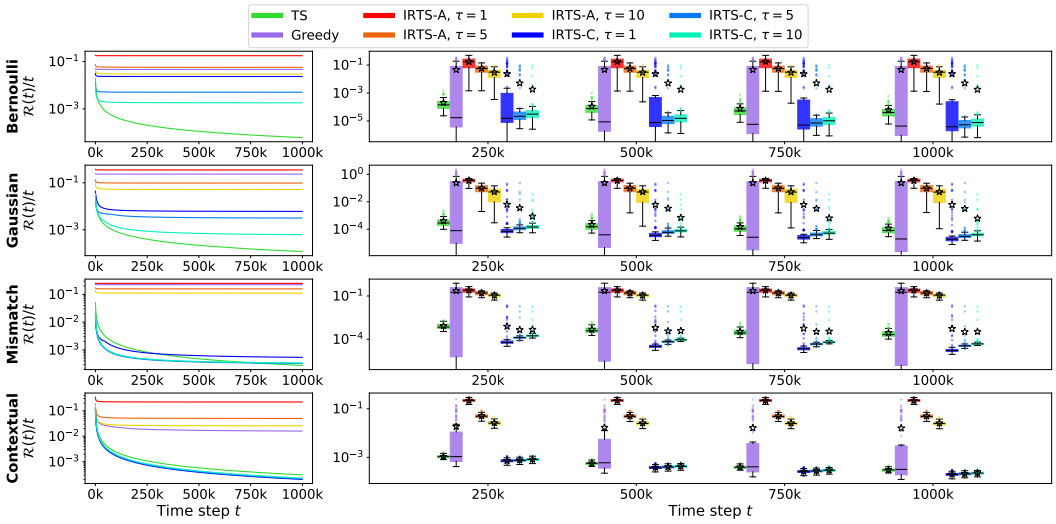

Figure 2: Average-regret-so-far $\mathcal{R}(t)/t$ (left column) and their corresponding box-plots (right column, with mean regrets displayed with a $\star$ symbol) for Bayesian bandits with analytic models (100 runs).

We show average-regret-so-far $\mathcal{R}(t)/t$ curves for each experiment in the left column of Figure 2 and summarise the final performance of each algorithm in Appendix C.1, Table 2. We observe the following behaviour:

- TS performs well across most tasks, achieving the best results in MAB tasks. For Bernoulli and Gaussian tasks, both IRTS-A and IRTS-C display linear regret, and their performance improves with increasing τ, agreeing with our theoretical results.
- The performance of IRTS-A is inferior to that of TS and IRTS-C for similar τ, as expected from excessive exploration.
- In the contextual bandit task, IRTS-C outperforms TS, which we conjecture is due to the additional exploration that arises with observing a new context each round, thus mitigating the under-exploration of IRTS-C compared to the MAB tasks.

To further understand the regret profiles, we show in the right column of Figure 2 box plots that show the distributions over regrets every 250k timesteps, including $10\%/25\%/75\%/90\%$ percentiles and upper-tails. Investigating the box-plots reveals a long-tail behaviour of IRTS-C results. While IRTS-C methods achieve linear average regrets, in most runs they obtain low-regret, often even lower than TS runs, and the few failure cases are catastrophic. This is consistent with the under-exploration behaviour of IRTS-C when compared with TS, so that a good trajectory of actions is not guaranteed. However, empirically even using a small $\tau = 1$ consistently outperforms a purely greedy strategy (equiv. to $\tau = 0$), significantly mitigating the severity of under-exploration in practice.

Motivated by the results in this section, going forward we report results in the main text for $\tau = 1$ as exemplar, due to its strong performance-computation tradeoff in the contextual bandit task in the case of IRTS-C, and its relevance to prior literature in the case of IRTS-A, which is equivalent to the TS-LLM method investigated in Sun et al. (2025) without the temperature scheduling.

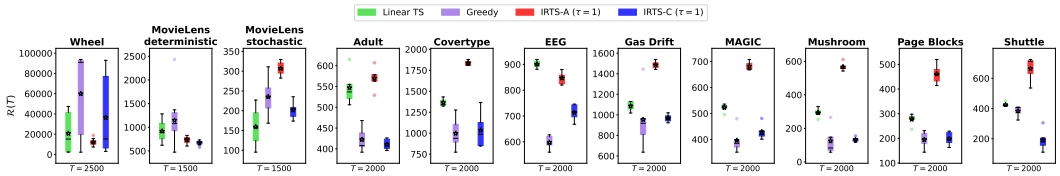

Figure 3: Box plot of cumulative regret $\mathcal{R}(T)$ across bandit tasks with horizons T. The box plots display the average across seeds with a $\star$ symbol. The Wheel and Movielens tasks are aggregated across 10 seeds and the OpenML tasks across 5.

5.2 In-context bandit with TabPFN

We present here results on a range of contextual bandit environments using a modern TabPFN Hollmann et al. (2025) in-context learning architecture.

- **Wheel bandit:** a bandit task proposed in Riquelme et al. (2018) specifically designed to test the exploration-exploitation capabilities of bandit algorithms.
- **Movielens recommendation:** a bandit task adapted from Nie et al. (2024) to emulate movie recommendations from on user features based on semi-synthetic data.
- **Open ML:** a bandit task derived from supervised learning datasets Bischl et al. (2025). We add noise to the observed labels to incentivise more thorough exploration.

We defer to Appendix C.2 for a precise description of each task and experimental details.

We compare to greedy (with respect to TabPFN predictions) and Linear TS (Agrawal & Goyal, 2013) baselines. Fig. 3 shows the final cumulative regrets per task. On the wheel bandit, IRTS-A outperform LinTS, IRTS-C has competitive median performance but a heavy upper tail, and Greedy is the weakest. The wheel bandit design tests the exploration capabilities while requiring a flexible, non-linear model for reward approximation. Greedy fails to adequately explore, and LinTS is not sufficiently expressive to capture the reward function, leading to high regret. TabPFN does not run into this limitation and IRTS-A ensures sufficient exploration. IRTS-C alleviates the frequency of the Greedy failure mode where most seeds get stuck with suboptimal regret, but does not resolve it entirely. As explained in Appendix C.2 and Nie et al. (2024), the Movielens data-generation process is well-aligned with the LinTS model, making it a strong baseline for this task. We observe that IRTS-C is competitive with LinTS, outperforming it in the Deterministic variant. While IRTS-C is competitive on Movielens variants tasks, IRTS-A does well on the Deterministic setting but badly, even worse than greedy, in the Stochastic environment. The OpenML tasks come from classification datasets, so do not as much exploration, and Greedy and IRTS-C approaches are more successful.

Overall, the optimality of a particular algorithm is dependent on the amount of exploration required by the task and, in the case of LinTS, whether the model is expressive enough. IRTS-C is the only method that ranks first or second across all non-Wheel tasks. Its weakness in the Wheel bandit is a mitigated version of the Greedy failure mode, where some seeds get stuck on suboptimal actions. This is consistent with the results observed in the previous section, showing that it suffers from the same fundamental weakness of Greedy but greatly mitigates the frequency of such failure cases. On the other hand, IRTS-A performs poorly in the OpenML tasks but is the strongest method on the Wheel bandit where exploration is crucial.

5.3 In-context bandit with LLMs

We evaluate the performance of various in-context methods for action-selection with LLMs. We choose the buttons bandit task (Krishnamurthy et al., 2024), where the agent selects between 5 buttons, each button yielding a Bernoulli reward. Four of them have Bernoulli parameter $p = 0.6$ whereas the optimal one, shuffled every seed, has a probability of $p = 0.8$. The experiments are run over 10 seeds and last 500 steps. In our comparison, we include two philosophies for in-context LLM action-selection. The first, common in the literature (Song et al., 2025; Nie et al., 2024; Lim et al., 2025; Xia et al., 2025), is to directly ask the LLM to choose which action it should pick next. We refer to this as the Instruction baseline. The second, which is more similar to how we implement IRTS with LLMs, involves using the LLM to predict what the next reward will be given a dataset in the prompt, and use these in-context predictions to carry out action-selection. Under this framework,

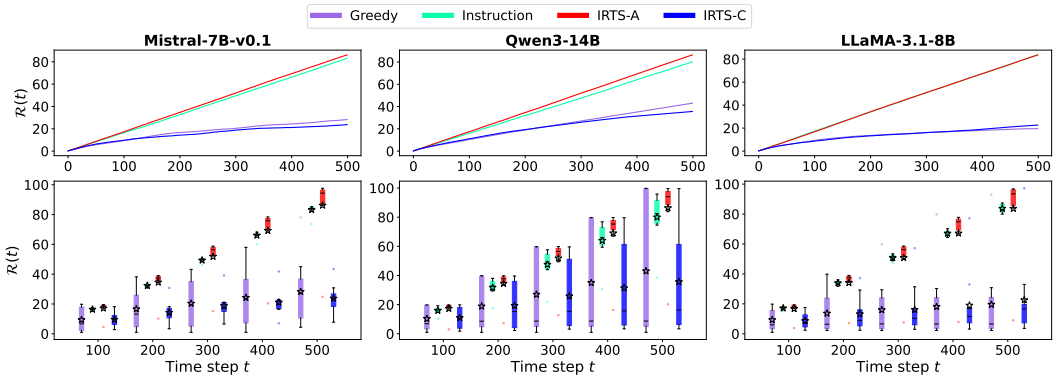

Figure 4: Average cumulative regret $\mathcal{R}(t)$ (top row) and corresponding box-plots (bottom row, with mean regrets displayed with a star $\star$ symbol) for the buttons experiments with LLMs (10 runs).

we evaluate greedy action selection, IRTS-A with $\tau = 1$ (which coincides with TS-LLM from (Sun et al., 2025) without temperature scheduling) and our IRTS-C with $\tau = 1$. As the experiments in this section strictly concern relative comparison of in-context action selection strategies with LLMs, we do not compare against traditional MAB baselines. We use the open source Mistral-7B (Jiang et al., 2023), Qwen3-14B (Yang et al., 2025) and LLaMA-3.1-8B (Grattafiori et al., 2024) models for our evaluations.

We observe that IRTS-C provides a competitive action-selection strategy. Our results highlight the importance of choosing an appropriate action-selection strategy. This is both true in terms of the method in which the LLM's predictions are leveraged, in terms of instructing the model to pick an action vs asking it for a reward prediction, as well as in terms of effectively balancing exploration and exploitation. The Instruction and IRTS-A methods exhibit slightly worse than random baseline, which would achieve an expected cumulative regret of 80 after 500 steps, suggesting that the LLM's noisy predictions already provide sufficient exploration. To avoid introducing systematic bias in the evaluation, we break ties at random. This is a departure from the theoretical analysis and increases the exploratory profile of IRTS-A. This further strengthens the empirical observation that, with complex models and noisy predictions, employing a greedier strategy, such as IRTS-C, is generally beneficial.

6 Discussion

Limitations. Our work has limitations. We build on the interpretation of ICL as a form of Bayesian inference, which is debatable in practice (Chlon et al., 2025; Falck et al., 2024). Despite this, we expect our work to become increasingly relevant as foundation models improve in quality and consistency of their in-context predictions, and IRTS still provides value regardless of whether this assumption is strictly true. Secondly, the quality of predictions with LLMs is affected by long context windows (Liu et al., 2024), although advances in ICL have led to improvements on this front (Baek et al., 2025; Hao et al., 2022). Furthermore, IRTS-C queries the predictive mean, which requires accessing the model's logits. If this is not provided, the predictive mean would need to be estimated via additional sampling of the model. Finally, the current implementation of our work is computationally sensitive to large action spaces, as it requires drawing τ imagined rewards for each of the K available arms, with computation that scales as $\mathcal{O}(K\tau)$. This problem is mitigated by the fact that, in practice, even a small $\tau = 1$ can still lead to improved performance with IRTS-C.

Summary and future work. We have introduced the IRTS framework, which enables approximate TS-like exploration for in-context learners, opening the door to leveraging the rich priors from foundation models for in-context decision-making in a principled and effective way. We proposed two practical instances (IRTS-A/C) and provided a theoretical analysis of their regret and variance profiles, where the theoretical results are verified on Bayesian bandits with analytic models. We conducted evaluations using TabPFN and LLM architectures, in which IRTS achieved competitive exploration-exploitation performance. Given that IRTS-A over-explores and IRTS-C under-explores, a hybrid approach combining both, or a new mean reward estimator based on imagined outcomes, can potentially improve the exploration efficiency further.

AI DISCLOSURE

In this work, we used generative AI tools to help formulate mathematical claims and to provide critical ingredients toward proving them, and to assist in writing and refining proofs, including proposing and refining candidate lemmas and proof strategies, and checking proofs and derivations for soundness and gaps, en route to the final theorems. All LLM-suggested mathematical claims, lemmas, and proof steps were independently verified by the authors.

We also used generative AI tools to provide feedback on our research methodology and experiments, for instance to review the overall manuscript.

We did not use generative AI tools to generate synthetic datasets, to help develop the theoretical models or conceptual frameworks underlying this work, to implement our methods (the core algorithms were written by hand by the authors), or to interpret our results. Assisting with translation and supporting qualitative/thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools (including code-completion and autocomplete tools) to help write auxiliary code (experiment scaffolding and code to run experiments and aggregate results or for instance to reformat OpenML data for model input), for figure plotting and table formatting when presenting results, and to proofread and review the manuscript. The core algorithmic code was written by the authors. All AI-assisted code was reviewed and tested by the authors for correctness, all AI-suggested proof steps, identified gaps, and methodology feedback were independently verified by the authors, and all AI-assisted writing/figures/tables were checked against the underlying results.

We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work enables principled Thompson Sampling-style exploration with modern auto-regressive models, with potential benefits in any domain requiring sequential decision-making under uncertainty, such as personalised recommendations, clinical trial design, and adaptive education (Guo et al., 2020). As with any method that improves exploration efficiency, deployment in sensitive settings should be accompanied by appropriate safeguards to mitigate risks arising from miscalibrated reward estimates or biased underlying models.

REPRODUCIBILITY STATEMENT

We have taken several steps to support the reproducibility of our results The anonymised codebase, including all model implementations, hyper-parameters, and scripts to reproduce the experiments presented in this paper, can be accessed at: `https://anonymous.4open.science/r/Implicit-Thompson-Sampling-Review-Version-CB18`.

For our theoretical results, we state all assumptions explicitly and provide complete proofs of our main claims in Section A.

REFERENCES

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for Thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

Jinheon Baek, Sun Jae Lee, Prakhar Gupta, Geunseob Oh, Siddharth Dalmia, and Prateek Kolhar. Revisiting in-context learning with long context language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26950–26966, 2025.

Uljad Berdica, Fernando Acero, Anton Ipsen, Parisa Zehtabi, Michael Cashmore, and Manuela Veloso. When do we need LLMs? a diagnostic for language-driven bandits. *Reinforcement Learning Journal*, 7, 2026.

Bernd Bischl, Giuseppe Casalicchio, Taniya Das, Matthias Feurer, Sebastian Fischer, Pieter Gijsbers, Subhaditya Mukherjee, Andreas C Müller, László Németh, Luis Oala, Lennart Purucker, Sahithya Ravi, Jan N van Rijn, Prabhant Singh, Joaquin Vanschoren, Jos van der Velde, and Marcel Wever. Openml: Insights from 10 years and more than a thousand papers. *Patterns*, 6(7):101317, 2025. doi: 10.1016/j.patter.2025.101317. URL https://www.cell.com/patterns/fulltext/S2666-3899(25)00165-5.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Tiffany Tianhui Cai, Hongseok Namkoong, Daniel Russo, and Kelly W Zhang. Active exploration via autoregressive generation of missing data. *arXiv preprint arXiv:2405.19466*, 2024.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf.

Leon Chlon, Zein Khamis, Maggie Chlon, Mahdi El Zein, and MarcAntonio M Awada. LLMs are Bayesian, in expectation, not in realization. *arXiv preprint arXiv:2507.11768*, 2025.

Bruno De Finetti. Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pp. 179–190, 1929.

Fabian Falck, Ziyu Wang, and Chris Holmes. Is in-context learning in large language models bayesian? a martingale perspective. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Edwin Fong, Chris Holmes, and Stephen G Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1357–1391, 2023.

Sandra Fortini and Sonia Petrone. Exchangeability, prediction and predictive modeling in Bayesian statistics. *Statistical Science*, 40(1):40–67, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu,

Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, and Zoe Papakipos. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dalin Guo, Sofia Ira Ktena, Pranay Kumar Myana, Ferenc Huszar, Wenzhe Shi, Alykhan Tejani, Michael Kneier, and Sourav Das. Deep bayesian bandits: Exploring in online personalized recommendations. In *Proceedings of the 14th ACM conference on recommender systems*, pp. 456–461, 2020.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. Structured prompting: Scaling in-context learning to 1,000 examples. *arXiv preprint arXiv:2212.06713*, 2022.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://doi.org/10.1038/s41586-024-08328-6.

Prateek Jaiswal, Debdeep Pati, Anirban Bhattacharya, and Bani K Mallick. Generalized regret analysis of Thompson sampling using fractional posteriors. *arXiv preprint arXiv:2309.06349*, 2023.

I Shavindra Jayasekera, Jacob Si, Filippo Valdettaro, Wenlong Chen, A Aldo Faisal, and Yingzhen Li. Variational uncertainty decomposition for in-context learning. *arXiv preprint arXiv:2509.02327*, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, and Pan Xu. Thompson sampling with less exploration is fast and optimal. In *International Conference on Machine Learning*, pp. 15239–15261. PMLR, 2023.

Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:43057–43083, 2023.

Jimin Lim, Arjun Damerla, Arthur Jiang, Nam Le, and Nikil Selladurai. Textbandit: Evaluating probabilistic reasoning in LLMs through language-only decision tasks. *RANLP'2025*, pp. 1, 2025.

Che-Yu Liu and Lihong Li. On the prior sensitivity of Thompson sampling. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles (eds.), *Algorithmic Learning Theory*, pp. 321–336, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46379-7.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Yueyang Liu, Benjamin Van Roy, and Kuang Xu. Nonstationary bandit learning via predictive sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 6215–6244. PMLR, 2023.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

Gergely Neu, Iuliia Olkhovskaia, Matteo Papini, and Ludovic Schwartz. Lifting the information ratio: An information-theoretic analysis of thompson sampling for contextual bandits. *Advances in Neural Information Processing Systems*, 35:9486–9498, 2022.

Kenyon Ng, Edwin Fong, David T Frazier, Jeremias Knoblauch, and Susan Wei. Tabmgp: Martingale posterior with tabpfn. *arXiv preprint arXiv:2510.25154*, 2025.

Allen Nie, Yi Su, Bo Chang, Jonathan N Lee, Ed H Chi, Quoc V Le, and Minmin Chen. Evolve: Evaluating and optimizing LLMs for in-context exploration. *arXiv preprint arXiv:2410.06238*, 2024.

My Phan, Yasin Abbasi Yadkori, and Justin Domke. Thompson sampling and approximate inference. *Advances in Neural Information Processing Systems*, 32, 2019.

James Requeima, John Bronskill, Dami Choi, Richard E Turner, and David Duvenaud. Llm processes: Numerical predictive distributions conditioned on natural language. *Advances in Neural Information Processing Systems*, 37:109609–109671, 2024.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. In *International Conference on Learning Representations*, 2018.

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, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–99, 2018.

Kefan Song, Amir Moeini, Peng Wang, Lei Gong, Rohan Chandra, Shangtong Zhang, and Yanjun Qi. Reward is enough: Llms are in-context reinforcement learners. *arXiv preprint arXiv:2506.06303*, 2025.

Jiahang Sun, Zhiyong Wang, Runhan Yang, Chenjun Xiao, John Lui, and Zhongxiang Dai. Large language model-enhanced multi-armed bandits. *arXiv preprint arXiv:2502.01118*, 2025.

Yan Shuo Tan, Kenyon Ng, Ruizhe Deng, Sumetha Loganathan, Qiong Zhang, and Bibhas Chakraborty. Pfn-ts: Thompson sampling for contextual bandits via prior-data fitted networks. *arXiv preprint arXiv:2605.10137*, 2026.

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.

Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans. Bayesian sparse sampling for on-line reward optimization. In *Proceedings of the 22nd international conference on Machine learning*, pp. 956–963, 2005.

Zheng Wen, Ian Osband, Chao Qin, Xiuyuan Lu, Morteza Ibrahimi, Vikranth Dwaracherla, Mohammad Asghari, and Benjamin Van Roy. From predictions to decisions: The importance of joint predictive distributions. *arXiv preprint arXiv:2107.09224*, 2021.

Fanzeng Xia, Hao Liu, Yisong Yue, and Tongxin Li. Beyond numeric rewards: In-context dueling bandits with llm agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9959–9988, 2025.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Naimeng Ye and Hongseok Namkoong. Exchangeable sequence models quantify uncertainty over latent concepts. *arXiv preprint arXiv:2408.03307*, 2024.

Kelly W Zhang, Tiffany Tianhui Cai, Hongseok Namkoong, and Daniel Russo. Contextual thompson sampling via generation of missing data. *arXiv preprint arXiv:2502.07064*, 2025.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural Thompson sampling. In *International Conference on Learning Representation (ICLR)*, 2021.

APPENDIX CONTENTS

The appendix is organised as follows.

- **Section A: Proofs of Theoretical Results.** Establishes the equivalence between sequentially generated imagined rewards and latent-conditioned samples, proves Lemmas 3.1 and 3.2 on the empirical-mean-reward and conditional-mean-reward variances, and gives the information-theoretic proofs of Theorems 3.3, 3.4 and 3.5.
- **Section B: Implicit Thompson Sampling for Contextual Bandits.** Extends IRTS-A and IRTS-C to the contextual setting, describing how imagined rewards are sampled conditional on the next context, and presents the corresponding pseudocode in Algorithm 3.
- **Section C: Additional Experimental Details.** Full specification of the experimental setup and supplementary results, organised by model class:
  - *Analytic experiments (Section C.1):* bandit instances, outline of methods, and additional analytic results.
  - *TabPFN experiments (Section C.2):* details on the experimental setup for the TabPFN setting. Description of each of the Wheel, Movielens and OpenML tasks. Numerical reporting fo the results presented in the main paper and individual per-task regret curves.
  - *LLM experiments (Section C.3.1):* buttons-bandit setup, the three open-weight base models used (Qwen3-14B, Llama-3.1-8B, Mistral-7B-v0.1), the predictive action-selection methods (IRTS-A, IRTS-C), the greedy and instruction baselines, the vLLM inference setup, and additional LLM results.
- **Section D: Code Availability and Computational Resources.** Link to the anonymised codebase and a breakdown of CPU/GPU resources used for the analytic, TabPFN, and LLM experiments.

A PROOFS FOR BAYESIAN REGRET

A.1 EQUIVALENCE OF AUTOREGRESSIVE AND LATENT-CONDITIONED SAMPLING

We first prove Proposition 2.1. Under the latent-variable sampling, conditional independence implies that

$$p(z_{1:n}) = \int p(\theta) \prod_{i=1}^{n} p(z_i \mid \theta)\, d\theta,$$

which is the distribution obtained by first sampling $\theta \sim p(\theta)$ and then sampling $z_{1:n}$ conditional on θ and then marginalising over θ.

On the other hand, by the chain rule for probability,

$$p(z_{1:n}) = \prod_{i=1}^{n} p(z_i \mid z_{1:i-1}),$$

which is exactly the joint distribution obtained by autoregressive sampling. Hence the two sampling procedures induce the same joint distribution over $z_{1:n}$. □

A.2 PROOFS OF THE VARIANCE LEMMAS

Proof of Lemma 3.1. By Proposition 2.1, there is a representation in which $\theta_t \sim \pi_t$ is drawn once and, conditional on θ_t, the variables $\widetilde{Y}_{t,a}^{(1)}, \ldots, \widetilde{Y}_{t,a}^{(\tau)}$ are the elements corresponding to action a from τ i.i.d. draws from $p_Y(\cdot \mid \theta_t)$. Their rewards $R(\widetilde{Y}_{t,a}^{(i)})$ have mean $\mu_a(\theta_t)$ and variance $\sigma_a^2(\theta_t)$. Hence

$$\mathbb{E}[\widehat{\mu}_{t,a}^{A} \mid \theta_t, \mathcal{H}_t] = \mu_a(\theta_t), \qquad \mathrm{Var}(\widehat{\mu}_{t,a}^{A} \mid \theta_t, \mathcal{H}_t) = \frac{\sigma_a^2(\theta_t)}{\tau}.$$

The tower property marginalising the left equation over θ_t then immediately proves the asserted expectation. On the other hand, the law of total variance gives

$$\begin{aligned}\mathrm{Var}(\widehat{\mu}_{t,a}^{A} \mid \mathcal{H}_t) &= \mathbb{E}\left[\mathrm{Var}(\widehat{\mu}_{t,a}^{A} \mid \theta_t, \mathcal{H}_t) \,\middle|\, \mathcal{H}_t\right] \\ &\quad + \mathrm{Var}\left(\mathbb{E}[\widehat{\mu}_{t,a}^{A} \mid \theta_t, \mathcal{H}_t] \,\middle|\, \mathcal{H}_t\right) \\ &= \frac{1}{\tau}\mathbb{E}[\sigma_a^2(\theta_t) \mid \mathcal{H}_t] + \mathrm{Var}(\mu_a(\theta_t) \mid \mathcal{H}_t).\end{aligned}$$

□

Proof of Lemma 3.2. Underthe same representation, let $Y_{t,a}^{+}$ be a fresh outcome from $p_{Y_a}(\cdot \mid \theta_t)$. Fix an arm a. Then,

$$\mathbb{E}\Big[R(Y_{t,a}^{+}) \mid \theta_t, \mathcal{H}_t, \widetilde{D}_t\Big] = \mathbb{E}\big[R(Y_{t,a}^{+}) \mid \theta_t\big] = \mu_a(\theta_t).$$

Hence, by the tower property, marginalising over $\theta_t | \mathcal{H}_t, \widetilde{D}_t$ gives

$$\widehat{\mu}_{t,a}^{C} = \mathbb{E}\Big[R(Y_{t,a}^{+}) \mid \mathcal{H}_t, \widetilde{D}_t\Big] = \mathbb{E}\Big[\mu_a(\theta_t) \mid \mathcal{H}_t, \widetilde{D}_t\Big].$$

Finally, marginalising over $\widetilde{D}_t | \mathcal{H}_t$, the tower property gives

$$\mathbb{E}\big[\widehat{\mu}_{t,a}^{C} \mid \mathcal{H}_t\big] = \mathbb{E}[\mu_a(\theta_t) \mid \mathcal{H}_t].$$

Applying the law of total variance,

$$\begin{aligned}\mathrm{Var}(\mu_a(\theta_t) \mid \mathcal{H}_t) &= \mathbb{E}\left[\mathrm{Var}(\mu_a(\theta_t) \mid \mathcal{H}_t, \widetilde{D}_t) \,\middle|\, \mathcal{H}_t\right] \\ &\quad + \mathrm{Var}\left(\mathbb{E}[\mu_a(\theta_t) \mid \mathcal{H}_t, \widetilde{D}_t] \,\middle|\, \mathcal{H}_t\right) \\ &\geq \mathrm{Var}(\widehat{\mu}_{t,a}^{C} \mid \mathcal{H}_t).\end{aligned}$$

Next, we address monotonicity in τ. Define $\widetilde{D}_t^{(\tau)} = (\widetilde{Y}_t^{(1)}, \ldots, \widetilde{Y}_t^{(\tau)})$, where the synthetic observations are generated autoregressively, so that $\widetilde{D}_t^{(\tau)}$ is a prefix of $\widetilde{D}_t^{(\tau+1)}$. Recall that

$$\widehat{\mu}_{t,a}^{C}(\tau) = \mathbb{E}\Big[\mu_a(\theta_t) \mid \mathcal{H}_t, \widetilde{D}_t^{(\tau)}\Big].$$

Since conditioning on $\mathcal{H}_t, \widetilde{D}_t^{(\tau+1)}$ reveals all the information in $\mathcal{H}_t, \widetilde{D}_t^{(\tau)}$, the tower property gives

$$\mathbb{E}\Big[\widehat{\mu}_{t,a}^{C}(\tau+1) \mid \mathcal{H}_t, \widetilde{D}_t^{(\tau)}\Big] = \widehat{\mu}_{t,a}^{C}(\tau).$$

Hence, by law of total variance conditional on $\mathcal{H}_t$,

$$\begin{aligned}\mathrm{Var}\big(\widehat{\mu}_{t,a}^{C}(\tau+1) \mid \mathcal{H}_t\big) &= \mathrm{Var}\big(\widehat{\mu}_{t,a}^{C}(\tau) \mid \mathcal{H}_t\big) \\ &\quad + \mathbb{E}\Big[\mathrm{Var}\Big(\widehat{\mu}_{t,a}^{C}(\tau+1) \mid \mathcal{H}_t, \widetilde{D}_t^{(\tau)}\Big)\Big|\mathcal{H}_t\Big] \\ &\geq \mathrm{Var}\big(\widehat{\mu}_{t,a}^{C}(\tau) \mid \mathcal{H}_t\big).\end{aligned}$$

Therefore, $\mathrm{Var}\big(\widehat{\mu}_{t,a}^{C}(\tau) \mid \mathcal{H}_t\big)$ is nondecreasing in τ. □

A.3 A FIXED-KERNEL POSTERIOR-MATCHING BOUND

We first establish a general bound for a family of algorithms that carry out action selection with a fixed-kernel randomised action rule, of which IRTS-A as a special case. Let $q(a \mid \theta)$ be any fixed randomised action rule:

$$q(a \mid \theta) \geq 0, \qquad \sum_{a=1}^{K} q(a \mid \theta) = 1.$$

Here "fixed" means that q does not depend on the round or realised history. Define its one-step approximation error by

$$\varepsilon(q) = \mathbb{E}\left[\sum_{a=1}^{K} q(a \mid \theta^\star)\,(\mu^\star(\theta^\star) - \mu_a(\theta^\star))\right]. \tag{9}$$

Consider the posterior-sampling policy that draws $\theta_t \sim \pi_t$ and then $A_t \sim q(\cdot \mid \theta_t)$. For reference, exact TS would correspond to q placing its probability mass entirely on the action with greatest mean reward given θ_t.

Proposition A.1 (Fixed-kernel posterior matching)**.** *Let B be an arm drawn once at the start of the episode, drawn conditionally on the true environment parameter which is itself drawn from the prior $\theta^\star \sim \pi(\theta^\star)$ at the start of the episode. Suppose B is then drawn according according to*

$$\mathbb{P}(B = a \mid \theta^\star) = q(a \mid \theta^\star),$$

using auxiliary randomness independent of the algorithm and reward process. Then

$$\mathcal{R}(T) \leq T\varepsilon(q) + \sqrt{\frac{KT\,H(B)}{2}} \leq T\varepsilon(q) + \sqrt{\frac{KT\log K}{2}}. \tag{10}$$

Proof. **Step 1: posterior matching for the auxiliary arm.** Because the auxiliary draw is not used to generate the history, B and $\mathcal{H}_t$ are conditionally independent given $\theta^\star$. For every $a \in [K]$,

$$\mathbb{P}(B = a \mid \mathcal{H}_t) = \int q(a \mid \theta^\star)\, p(\theta^\star|\mathcal{H}_t)\, d\theta^\star = \int q(a \mid \theta^\star)\, \pi_t(\theta^\star) d\theta^\star, \tag{11}$$

$$\mathbb{P}(A_t = a \mid \mathcal{H}_t) = \int q(a \mid \theta)\, \pi_t(\theta) d\theta. \tag{12}$$

Consequently,

$$\mathbb{P}(A_t = a \mid \mathcal{H}_t) = \mathbb{P}(B = a \mid \mathcal{H}_t). \tag{13}$$

Moreover, because A_t is generated from a fresh posterior draw and fresh rule randomisation,

$$A_t \perp\!\!\!\perp (\theta^\star, B) \mid \mathcal{H}_t. \tag{14}$$

Thus A_t is an independent draw from the posterior distribution of B. This is the probability-matching property used in Proposition 2 of Russo & Van Roy (2016, pp. 14–15), with B replacing their optimal arm $A^\star$. This is the key property underlying the machinery from the proof in Russo & Van Roy (2016), and probability matching on the reference arm B allows us to bound the regret relative to the policy that choses arms according to q. The rest of the proof involves applying the same steps in Russo & Van Roy (2016) to our setting, and accounting for the consequences of choosing arms that are suboptimal with respect to the θ_t samples.

Step 2: regret decomposition. For every t,

$$\mu^\star(\theta^\star) - \mu_{A_t}(\theta^\star) = (\mu^\star(\theta^\star) - \mu_B(\theta^\star)) + (\mu_B(\theta^\star) - \mu_{A_t}(\theta^\star)) .$$

Since

$$\mathbb{E}[\mu^\star(\theta^\star) - \mu_B(\theta^\star)] = \varepsilon(q),$$

taking expectations and summing gives

$$\mathcal{R}(T) = T\varepsilon(q) + \mathbb{E}\left[\sum_{t=1}^{T} (\mu_B(\theta^\star) - \mu_{A_t}(\theta^\star))\right]. \tag{15}$$

The arm B is drawn once and remains fixed over all rounds.

Step 3: one-round information bound. Fix a round t and, to avoid ambiguity in notation, a realised history $\mathcal{H}_t = h$. From equation 13 and equation 14,

$$\begin{aligned}&\mathbb{E}\left[\mu_B(\theta^\star) - \mu_{A_t}(\theta^\star) \mid \mathcal{H}_t = h\right] \\ &\quad = \sum_{a=1}^{K} \mathbb{P}(B = a \mid \mathcal{H}_t = h)\Big(\mathbb{E}[\mu_a(\theta^\star) \mid B = a, \mathcal{H}_t = h] - \mathbb{E}[\mu_a(\theta^\star) \mid \mathcal{H}_t = h]\Big).\end{aligned} \tag{16}$$

This is the analogue of the second identity in Proposition 2 of Russo & Van Roy (2016, p. 14).

Let $Y_{t,a}$ be a fresh outcome of arm a. We denote the law of a random variable with the symbol $\mathcal{L}$. Since $R(Y_{t,a}) \in [0, 1]$, Pinsker's inequality implies

$$\begin{aligned}&|\mathbb{E}[\mu_a(\theta^\star) \mid B = a, \mathcal{H}_t = h] - \mathbb{E}[\mu_a(\theta^\star) \mid \mathcal{H}_t = h]| \\ &\quad \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathcal{L}(Y_{t,a} \mid B = a, \mathcal{H}_t = h) \,\|\, \mathcal{L}(Y_{t,a} \mid \mathcal{H}_t = h)\right)}.\end{aligned} \tag{17}$$

This is Fact 9 of Russo & Van Roy (2016, p. 15).

The KL representation of mutual information gives

$$\begin{aligned}I(B; Y_{t,a} \mid \mathcal{H}_t = h) &= \sum_{b=1}^{K} \mathbb{P}(B = b \mid \mathcal{H}_t = h) \\ &\quad \cdot D_{\mathrm{KL}}\left(\mathcal{L}(Y_{t,a} \mid B = b, \mathcal{H}_t = h) \,\|\, \mathcal{L}(Y_{t,a} \mid \mathcal{H}_t = h)\right) \\ &\ge \mathbb{P}(B = a \mid \mathcal{H}_t = h) \\ &\quad \cdot D_{\mathrm{KL}}\left(\mathcal{L}(Y_{t,a} \mid B = a, \mathcal{H}_t = h) \,\|\, \mathcal{L}(Y_{t,a} \mid \mathcal{H}_t = h)\right).\end{aligned}$$

Therefore, arms having positive conditional probability satisfy

$$\begin{aligned}&|\mathbb{E}[\mu_a(\theta^\star) \mid B = a, \mathcal{H}_t = h] - \mathbb{E}[\mu_a(\theta^\star) \mid \mathcal{H}_t = h]| \\ &\quad \le \sqrt{\frac{I(B; Y_{t,a} \mid \mathcal{H}_t = h)}{2\mathbb{P}(B = a \mid \mathcal{H}_t = h)}}.\end{aligned} \tag{18}$$

Zero-probability arms contribute nothing to equation 16. Analogously to Russo & Van Roy (2016), substitution followed by the triangle inequality and Cauchy–Schwarz gives

$$\begin{aligned}&|\mathbb{E}\left[\mu_B(\theta^\star) - \mu_{A_t}(\theta^\star) \mid \mathcal{H}_t = h\right]| \\ &\quad \le \sqrt{\frac{K}{2} \sum_{a=1}^{K} \mathbb{P}(B = a \mid \mathcal{H}_t = h) I(B; Y_{t,a} \mid \mathcal{H}_t = h)}.\end{aligned} \tag{19}$$

Further, the mutual-information chain rule and the fact that A_t is conditionally independent of B given $\mathcal{H}_t$, in conjunction with probability matching, give

$$I(B;(A_t,Y_{t,A_t}) \mid \mathcal{H}_t = h) = I(B;A_t \mid \mathcal{H}_t = h) + I(B;Y_{t,A_t} \mid A_t, \mathcal{H}_t = h)$$
$$= \sum_{a=1}^{K} \mathbb{P}(B = a \mid \mathcal{H}_t = h) I(B;Y_{t,a} \mid \mathcal{H}_t = h). \tag{20}$$

Combining equation 19 and equation 20,

$$|\mathbb{E}\left[\mu_B(\theta^\star) - \mu_{A_t}(\theta^\star) \mid \mathcal{H}_t = h\right]| \le \sqrt{\frac{K}{2} I(B;(A_t,Y_{t,A_t}) \mid \mathcal{H}_t = h)}. \tag{21}$$

This is the counterpart of Proposition 3 of Russo & Van Roy (2016, p. 16); the calculation shows explicitly that probability matching, regardless of the optimality of the target, is the property used here.

Step 4: summing the information gains. Applying equation 21 for each realised history, followed by Jensen's inequality and Cauchy–Schwarz over time, yields

$$\mathbb{E}\left[\sum_{t=1}^{T} (\mu_B(\theta^\star) - \mu_{A_t}(\theta^\star))\right] \le \sqrt{\frac{KT}{2} \sum_{t=1}^{T} I(B;(A_t,Y_{t,A_t}) \mid \mathcal{H}_t)}. \tag{22}$$

Because $\mathcal{H}_t = (\mathcal{H}_t, A_t, Y_{t,A_t})$, the information chain rule gives

$$\sum_{t=1}^{T} I(B;(A_t,Y_{t,A_t}) \mid \mathcal{H}_t) = I(B;\mathcal{H}_{T+1}).$$

Finally,

$$I(B;\mathcal{H}_{T+1}) = H(B) - H(B \mid \mathcal{H}_{T+1}) \le H(B) \le \log K.$$

Substitution into equation 22, followed by equation 15, proves equation 10. □

A.4 Proof of Theorem 3.3

For a fixed environment θ, let the IRTS-A rule generate τ i.i.d. outcome vecotrs $\widetilde{Y}^{(1)}, \ldots, \widetilde{Y}^{(\tau)} \sim p_Y(\cdot \mid \theta)$ where $\widetilde{Y}_k^{(\tau)}$ corresponds to the kth draw's element corresponding to action a and define

$$\widehat{\mu}_a = \frac{1}{\tau} \sum_{i=1}^{\tau} R(\widetilde{Y}_a^{(i)}), \qquad \widehat{A} \in \arg\max_{a \in [K]} \widehat{\mu}_a.$$

In combination with any arbitrary fixed tie-breaking rule, this induces the special randomised action decision rule

$$q_\tau(a \mid \theta) = \mathbb{P}(\widehat{A} = a \mid \theta).$$

Therefore, the bound of IRTS-A can be analysed as a special case of the general fixed-kernel posterior matching bound. It remains to bound the approximation error term ε.

Lemma A.2 (IRTS-A approximation error)**.** *The approximation error of q_τ satisfies*

$$\varepsilon(q_\tau) \le \sqrt{\frac{\log K}{2\tau}}.$$

Proof. By definition,

$$\varepsilon(q_\tau) = \mathbb{E}\left[\mu^\star(\theta^\star) - \mu_{\widehat{A}}(\theta^\star)\right]. \tag{23}$$

Fix θ, let

$$a^\star(\theta) \in \arg\max_{a \in [K]} \mu_a(\theta),$$

and define $\xi_a = \widehat{\mu}_a - \mu_a(\theta)$. Conditional on this fixed θ, $a^\star(\theta)$ is nonrandom. Since $\widehat{A}$ maximises the empirical mean rewards,

$$\begin{aligned}\mu^\star(\theta) - \mu_{\widehat{A}}(\theta) &= \widehat{\mu}_{a^\star(\theta)} - \xi_{a^\star(\theta)} - \widehat{\mu}_{\widehat{A}} + \xi_{\widehat{A}} \\ &\le \xi_{\widehat{A}} - \xi_{a^\star(\theta)} \\ &\le \max_{a\in[K]} \xi_a - \xi_{a^\star(\theta)}. \end{aligned} \tag{24}$$

The empirical mean rewards are conditionally unbiased, and $a^\star(\theta)$ does not depend on the simulated outcomes. Hence

$$\mathbb{E}[\xi_{a^\star(\theta)} \mid \theta] = 0$$

and

$$\mathbb{E}[\mu^\star(\theta) - \mu_{\widehat{A}}(\theta) \mid \theta] \le \mathbb{E}\left[\max_{a\in[K]} \xi_a \,\middle|\, \theta\right]. \tag{25}$$

Hoeffding's lemma applied to $R(\widetilde{Y}_a^{(i)}) \in [0,1]$ (Hoeffding, 1963) implies that, for every $\lambda > 0$,

$$\mathbb{E}[e^{\lambda \xi_a} \mid \theta] \le \exp\left(\frac{\lambda^2}{8\tau}\right). \tag{26}$$

The log-sum-exp inequality and Jensen's inequality give

$$\begin{aligned}\mathbb{E}\left[\max_{a\in[K]} \xi_a \,\middle|\, \theta\right] &\le \frac{1}{\lambda} \log\left(\sum_{a=1}^{K} \mathbb{E}[e^{\lambda \xi_a} \mid \theta]\right) \\ &\le \frac{\log K}{\lambda} + \frac{\lambda}{8\tau}. \end{aligned}$$

This argument uses only linearity of expectation and does not require the ξ_a to be independent. Choosing $\lambda = \sqrt{8\tau \log K}$ yields

$$\mathbb{E}\left[\max_{a\in[K]} \xi_a \,\middle|\, \theta\right] \le \sqrt{\frac{\log K}{2\tau}}.$$

The bound holds for every θ. Combining it with equation 23 and equation 25, then integrating over the prior, proves the result. □

Proof of Theorem 3.3. Apply Proposition A.1 with $q = q_\tau$, which with Lemma A.2 then gives

$$\mathcal{R}(T) \le T\sqrt{\frac{\log K}{2\tau}} + \sqrt{\frac{KT \log K}{2}}.$$

□

A.5 Proof of Theorem 3.4

Unlike IRTS-A, the choice of arm is now no longer conditionally independent of history given θ_t, as the conditional mean used by IRTS-C has an explicit dependence on the history even after observing synthetic data. This means there is no fixed arm B that will enable us to carry out probability matching as is done in the proof IRTS-A. As such, we must adapt the proof to draw a different reference arm B_t each round.

Proof. **Step 1: posterior matching for a round-dependent auxiliary arm.** For each t, conditionally on $(\theta^\star, \mathcal{H}_t)$. Consider generating τ outcome vector samples from the ground-truth environment $\theta^\star$

$$\widetilde{Y}_t^{B,(1)}, \ldots, \widetilde{Y}_t^{B,(\tau)} \sim p_Y(\cdot \mid \theta^\star),$$

and let $\widetilde{D}_t^B$ denote this set of synthetic observations. All τ auxiliary outcome vectors are generated from the same $\theta^\star$, using fresh randomness independent of the algorithm and outcome process. Define

$$\widehat{\mu}_{t,a}^{C,B} = \mathbb{E}\left[R(Y_{t,a}) \,\middle|\, \mathcal{H}_t, \widetilde{D}_t^B\right], \qquad B_t \in \arg\max_{a\in[K]} \widehat{\mu}_{t,a}^{C,B}.$$

where B_t is chosen with some fixed arbitrary tie breaking rule if necessary. Equivalently, we have

$$B_t \sim q_{t,\tau}(\cdot \mid \theta^\star, \mathcal{H}_t),$$

for some induced randomised rule $q_{t,\tau}$, where the randomness is induced from the τ imagined reward samples. For every realised history h,

$$\mathbb{P}(B_t = a \mid \mathcal{H}_t = h) = \int q_{t,\tau}(a \mid \theta, h)\, \pi_t(\theta)\, d\theta, \tag{27}$$

$$\mathbb{P}(A_t = a \mid \mathcal{H}_t = h) = \int q_{t,\tau}(a \mid \theta, h)\, \pi_t(\theta)\, d\theta. \tag{28}$$

Consequently,

$$\mathbb{P}(A_t = a \mid \mathcal{H}_t) = \mathbb{P}(B_t = a \mid \mathcal{H}_t).$$

Fresh and mutually independent algorithmic and auxiliary randomness also gives

$$A_t \perp\!\!\!\perp (\theta^\star, B_t) \mid \mathcal{H}_t. \tag{29}$$

This posterior-matching property holds separately at every round. Unlike IRTS-A, a single auxiliary arm cannot be shared over time because $q_{t,\tau}$ changes with the current posterior.

Step 2: regret decomposition. For each t,

$$\mu^\star(\theta^\star) - \mu_{A_t}(\theta^\star) = (\mu^\star(\theta^\star) - \mu_{B_t}(\theta^\star)) + (\mu_{B_t}(\theta^\star) - \mu_{A_t}(\theta^\star))\,.$$

In step 5, we will prove the uniform approximation bound

$$\mathbb{E}\left[\mu^\star(\theta^\star) - \mu_{B_t}(\theta^\star) \,|\, \mathcal{H}_t\right] \le \sqrt{\frac{\log K}{2\tau}}. \tag{30}$$

It follows that

$$\mathcal{R}(T) \le T\sqrt{\frac{\log K}{2\tau}} + \mathbb{E}\left[\sum_{t=1}^{T} (\mu_{B_t}(\theta^\star) - \mu_{A_t}(\theta^\star))\right]. \tag{31}$$

Step 3: one-round information bound. Fix t and $\mathcal{H}_t = h$. By equation A.5 and equation 29,

$$\begin{aligned}
&\mathbb{E}\left[\mu_{B_t}(\theta^\star) - \mu_{A_t}(\theta^\star) \mid \mathcal{H}_t = h\right] \\
&\quad = \sum_{a=1}^{K} \mathbb{P}(B_t = a \mid \mathcal{H}_t = h)\Big(\mathbb{E}[\mu_a(\theta^\star) \mid B_t = a, \mathcal{H}_t = h] - \mathbb{E}[\mu_a(\theta^\star) \mid \mathcal{H}_t = h]\Big).
\end{aligned} \tag{32}$$

Let $Y_{t,a}$ again denote the potential outcome of arm a. As in the proof for Theorem 3.3, Pinsker's inequality, the KL representation of mutual information, and Cauchy–Schwarz give

$$\begin{aligned}
&\left|\mathbb{E}\left[\mu_{B_t}(\theta^\star) - \mu_{A_t}(\theta^\star) \mid \mathcal{H}_t = h\right]\right| \\
&\quad \le \sqrt{\frac{K}{2}\sum_{a=1}^{K} \mathbb{P}(B_t = a \mid \mathcal{H}_t = h) I(B_t; Y_{t,a} \mid \mathcal{H}_t = h)}.
\end{aligned} \tag{33}$$

The same chain-rule calculation as in equation 20, now with B_t, yields

$$I(B_t; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t = h) = \sum_{a=1}^{K} \mathbb{P}(B_t = a \mid \mathcal{H}_t = h) I(B_t; Y_{t,a} \mid \mathcal{H}_t = h). \tag{34}$$

Combining equation 33 and equation 34,

$$\left|\mathbb{E}\left[\mu_{B_t}(\theta^\star) - \mu_{A_t}(\theta^\star) \mid \mathcal{H}_t = h\right]\right| \le \sqrt{\frac{K}{2} I(B_t; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t = h)}. \tag{35}$$

Step 4: summing and telescoping the information gains. The tower property, Jensen's inequality, and Cauchy–Schwarz give

$$\mathbb{E}\left[\sum_{t=1}^{T}(\mu_{B_t}(\theta^\star)-\mu_{A_t}(\theta^\star))\right] \leq \sqrt{\frac{KT}{2}\sum_{t=1}^{T} I(B_t;(A_t,Y_{t,A_t})\mid \mathcal{H}_t)}. \tag{36}$$

The target B_t is freshly generated at each round, so unlike with IRTS-A its information gains do not themselves telescope. Conditional on $(\theta^\star,\mathcal{H}_t)$, however, the auxiliary randomisation generating B_t is independent of (A_t,Y_{t,A_t}). Thus B_t is conditionally independent of $(A_t,Y_{A_t,t})$ given $(\theta^\star,\mathcal{H}_\sqcup)$. The conditional data-processing inequality therefore gives

$$I(B_t;(A_t,Y_{t,A_t})\mid\mathcal{H}_t)\leq I(\theta^\star;(A_t,Y_{t,A_t})\mid\mathcal{H}_t).$$

Because $\mathcal{H}_t=(\mathcal{H}_t,A_t,Y_{t,A_t})$, the chain rule yields

$$\begin{aligned}\sum_{t=1}^{T} I(B_t;(A_t,Y_{t,A_t})\mid\mathcal{H}_t) &\leq \sum_{t=1}^{T} I(\theta^\star;(A_t,Y_{t,A_t})\mid\mathcal{H}_t)\\ &= I(\theta^\star;\mathcal{H}_{T+1}).\end{aligned} \tag{37}$$

Substitution into equation 36 gives

$$\mathbb{E}\left[\sum_{t=1}^{T}(\mu_{B_t}(\theta^\star)-\mu_{A_t}(\theta^\star))\right]\leq\sqrt{\frac{KT}{2}I(\theta^\star;\mathcal{H}_{T+1})}. \tag{38}$$

If $\theta^\star$ is discrete with finite entropy, then $A^\star$ is a function of $\theta^\star$, so

$$\begin{aligned}I(\theta^\star;\mathcal{H}_{T+1}) &= I(A^\star;\mathcal{H}_{T+1})+I(\theta^\star;\mathcal{H}_{T+1}\mid A^\star)\\ &\leq H(A^\star)+H(\theta^\star\mid A^\star)\\ &= H(\theta^\star).\end{aligned} \tag{39}$$

Step 5: IRTS-C approximation error. Fix $\mathcal{H}_t=h$. For the auxiliary dataset $\widetilde{D}_t^B$, define the empirical mean rewards and their maximiser by

$$\overline{R}_{t,a}^B=\frac{1}{\tau}\sum_{i=1}^{\tau}R(\widetilde{Y}_{t,a}^{B,(i)}),\qquad \widehat{A}_t\in\arg\max_{a\in[K]}\overline{R}_{t,a}^B.$$

By construction,

$$B_t\in\arg\max_{a\in[K]}\mathbb{E}\left[R(Y_{t,a})\,\middle|\,\mathcal{H}_t=h,\widetilde{D}_t^B\right].$$

In addition, since $\widehat{A}_t$ is deterministic given $\widetilde{D}_t^B$, the conditional optimality of B_t gives

$$\mathbb{E}\left[R(Y_{t,B_t})\,\middle|\,\mathcal{H}_t=h,\widetilde{D}_t^B\right]\geq\mathbb{E}\left[R(Y_{t,\widehat{A}_t})\,\middle|\,\mathcal{H}_t=h,\widetilde{D}_t^B\right].$$

Subtracting both sides from the same conditional expectation of $\mu^\star(\theta^\star)$ and averaging over $\widetilde{D}_t^B$ yields

$$\begin{aligned}&\mathbb{E}\left[\mu^\star(\theta^\star)-\mu_{B_t}(\theta^\star)\,|\,\mathcal{H}_t=h\right]\\ &\quad\leq\mathbb{E}\left[\mu^\star(\theta^\star)-\mu_{\widehat{A}_t}(\theta^\star)\,\middle|\,\mathcal{H}_t=h\right].\end{aligned} \tag{40}$$

It remains to bound the right-hand side. Fix θ, condition further on $\theta^\star=\theta$, and define

$$a^\star(\theta)\in\arg\max_{a\in[K]}\mu_a(\theta),\qquad \xi_a=\overline{R}_{t,a}^B-\mu_a(\theta).$$

The same selection inequality used in equation 24 gives

$$\mu^\star(\theta) - \mu_{\widehat{A}_t}(\theta) \le \max_{a\in[K]} \xi_a - \xi_{a^\star(\theta)}.$$

Conditional on $(\mathcal{H}_t = h, \theta^\star = \theta)$, the empirical mean rewards are unbiased and $a^\star(\theta)$ is nonrandom. Hoeffding's lemma and the same log-sum-exp calculation used for IRTS-A therefore give

$$\mathbb{E}\left[\mu^\star(\theta) - \mu_{\widehat{A}_t}(\theta) \,\middle|\, \mathcal{H}_t = h, \theta^\star = \theta\right] \le \sqrt{\frac{\log K}{2\tau}}.$$

The calculation does not require independence across arms. Taking expectations over θ and applying equation 40 proves equation 30.

Finally, equation 31 and equation 38 prove equation 7. □

A.6 Proof of Theorem 3.5

The proof for the contextual regret bound of IRTS-A and IRTS-C under i.i.d. contexts is almost identical to that of IRTS-C, as the conditioning on i.i.d. contexts does not affect the key probability matching properties or the conditioning structure that the proof relies on. In the contextual setting, the history includes contexts, so $\mathcal{H}_t = (X_1, A_1, Y_{1,A_1}, \cdots, X_{t-1}, A_{t-1}, Y_{t-1,A_{t-1}})$.

Proof. Let $M \in \{A, C\}$ denote either contextual IRTS variant. We assume that $X_1, \ldots, X_T \overset{\text{i.i.d.}}{\sim} \rho$, independently of $\theta^\star$, and write

$$\mu_a(x;\theta) = \mathbb{E}[R(Y_{t,a}) \,|\, X_t = x, \theta^\star = \theta], \qquad \mu^\star(x;\theta) = \max_{a\in[K]} \mu_a(x;\theta).$$

Step 1: posterior matching conditional on the context. Fix a round t. Conditional on $(\theta^\star, \mathcal{H}_t, X_t)$, generate an auxiliary synthetic block using the same construction as contextual IRTS-M, but with fresh randomness independent of the algorithm and outcome process. Let B_t denote the arm selected from this auxiliary block.

Equivalently, for some randomised decision rule $q^M_{t,\tau}$,

$$B_t \sim q^M_{t,\tau}(\cdot \mid \theta^\star, \mathcal{H}_t, X_t).$$

Since X_t is independent of $(\theta^\star, \mathcal{H}_t)$, $p(\theta^\star \mid \mathcal{H}_t, X_t) = p(\theta^\star \mid \mathcal{H}_t) = \pi_t(\theta^\star)$. Hence, for every realised $(\mathcal{H}_t, X_t) = (h, x)$,

$$\mathbb{P}(B_t = a \mid \mathcal{H}_t = h, X_t = x) = \int q^M_{t,\tau}(a \mid \theta, h, x)\, \pi_t(\theta)\, d\theta, \tag{41}$$

$$\mathbb{P}(A_t = a \mid \mathcal{H}_t = h, X_t = x) = \int q^M_{t,\tau}(a \mid \theta, h, x)\, \pi_t(\theta)\, d\theta. \tag{42}$$

Consequently,

$$\mathbb{P}(A_t = a \mid \mathcal{H}_t, X_t) = \mathbb{P}(B_t = a \mid \mathcal{H}_t, X_t). \tag{43}$$

Moreover, since A_t is generated with fresh randomness independent of $\theta^\star$ and B_t, we have

$$A_t \perp\!\!\!\perp \left(\theta^\star, B_t, Y_{t,1}, \ldots, Y_{t,K}\right) \mid (\mathcal{H}_t, X_t). \tag{44}$$

Step 2: regret decomposition. For each t,

$$\mu^\star(X_t;\theta^\star) - \mu_{A_t}(X_t;\theta^\star) = \Big(\mu^\star(X_t;\theta^\star) - \mu_{B_t}(X_t;\theta^\star)\Big) + \Big(\mu_{B_t}(X_t;\theta^\star) - \mu_{A_t}(X_t;\theta^\star)\Big).$$

Step 5 proves, for both $M = A$ and $M = C$, that

$$\mathbb{E}[\mu^\star(X_t;\theta^\star) - \mu_{B_t}(X_t;\theta^\star) \,|\, \mathcal{H}_t, X_t] \le \sqrt{\frac{\log K}{2\tau}}. \tag{45}$$

Therefore,

$$\mathcal{R}(T) \le T\sqrt{\frac{\log K}{2\tau}} + \mathbb{E}\left[\sum_{t=1}^{T} (\mu_{B_t}(X_t;\theta^\star) - \mu_{A_t}(X_t;\theta^\star))\right]. \tag{46}$$

Step 3: one-round information bound. Fix t and $(\mathcal{H}_t, X_t) = (h, x)$. By equation 43 and equation 44,

$$\begin{aligned}
&\mathbb{E}[\mu_{B_t}(x;\theta^\star) - \mu_{A_t}(x;\theta^\star) \,|\, \mathcal{H}_t = h, X_t = x] \\
&\quad = \sum_{a=1}^{K} \mathbb{P}(B_t = a \mid \mathcal{H}_t = h, X_t = x)\Big(\mathbb{E}[\mu_a(x;\theta^\star) \mid B_t = a, \mathcal{H}_t = h, X_t = x] \\
&\qquad\qquad - \mathbb{E}[\mu_a(x;\theta^\star) \mid \mathcal{H}_t = h, X_t = x]\Big).
\end{aligned} \tag{47}$$

For any arm having positive conditional probability, Pinsker's inequality and the KL representation of mutual information give

$$\begin{aligned}
&|\mathbb{E}[\mu_a(x;\theta^\star) \mid B_t = a, \mathcal{H}_t = h, X_t = x] - \mathbb{E}[\mu_a(x;\theta^\star) \mid \mathcal{H}_t = h, X_t = x]| \\
&\quad \le \sqrt{\frac{I(B_t; Y_{t,a} \mid \mathcal{H}_t = h, X_t = x)}{2\mathbb{P}(B_t = a \mid \mathcal{H}_t = h, X_t = x)}}.
\end{aligned} \tag{48}$$

Substituting equation 48 into equation 47, followed by the triangle inequality and Cauchy–Schwarz, yields

$$\begin{aligned}
&|\mathbb{E}[\mu_{B_t}(x;\theta^\star) - \mu_{A_t}(x;\theta^\star) \,|\, \mathcal{H}_t = h, X_t = x]| \\
&\quad \le \sqrt{\frac{K}{2}\sum_{a=1}^{K} \mathbb{P}(B_t = a \mid \mathcal{H}_t = h, X_t = x) I(B_t; Y_{t,a} \mid \mathcal{H}_t = h, X_t = x)}.
\end{aligned} \tag{49}$$

The mutual-information chain rule, conditional independence equation 44, and probability matching give

$$\begin{aligned}
&I(B_t; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t = h, X_t = x) \\
&\quad = I(B_t; A_t \mid \mathcal{H}_t = h, X_t = x) + I(B_t; Y_{t,A_t} \mid A_t, \mathcal{H}_t = h, X_t = x) \\
&\quad = \sum_{a=1}^{K} \mathbb{P}(B_t = a \mid \mathcal{H}_t = h, X_t = x) I(B_t; Y_{t,a} \mid \mathcal{H}_t = h, X_t = x).
\end{aligned} \tag{50}$$

Combining equation 49 and equation 50,

$$|\mathbb{E}[\mu_{B_t}(X_t;\theta^\star) - \mu_{A_t}(X_t;\theta^\star) \,|\, \mathcal{H}_t, X_t]| \le \sqrt{\frac{K}{2} I(B_t; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t, X_t)}. \tag{51}$$

Step 4: summing and telescoping the information gains. Taking expectations in equation 51, followed by Jensen's inequality and Cauchy–Schwarz over time, gives

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} (\mu_{B_t}(X_t;\theta^\star) - \mu_{A_t}(X_t;\theta^\star))\right] \\
&\qquad \le \sqrt{\frac{KT}{2}\sum_{t=1}^{T} I(B_t; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t, X_t)}.
\end{aligned} \tag{52}$$

The target B_t is freshly generated at each round, so its information gains do not themselves telescope. Conditional on $(\theta^\star, \mathcal{H}_t, X_t)$, however, the auxiliary randomisation generating B_t is independent of (A_t, Y_{t,A_t}). Hence

$$B_t \perp\!\!\!\perp (A_t, Y_{t,A_t}) \mid (\theta^\star, \mathcal{H}_t, X_t),$$

and the conditional data-processing inequality gives

$$I(B_t; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t, X_t) \le I(\theta^\star; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t, X_t)\,. \tag{53}$$

Since the contexts are i.i.d. and independent of $\theta^\star$,

$$I(\theta^\star; X_t \mid \mathcal{H}_t) = 0.$$

Therefore, using $\mathcal{H}_{t+1} = (\mathcal{H}_t, X_t, A_t, Y_{t,A_t})$, the chain rule gives

$$\begin{aligned} I(\theta^\star; \mathcal{H}_{T+1}) &= \sum_{t=1}^{T} I(\theta^\star; (X_t, A_t, Y_{t,A_t}) \mid \mathcal{H}_t) \\ &= \sum_{t=1}^{T} I(\theta^\star; X_t \mid \mathcal{H}_t) + \sum_{t=1}^{T} I(\theta^\star; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t, X_t) \\ &= \sum_{t=1}^{T} I(\theta^\star; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t, X_t)\,. \end{aligned} \tag{54}$$

Combining equation 53 and equation 54,

$$\sum_{t=1}^{T} I(B_t; (A_t, Y_{t,A_t}) \mid \mathcal{H}_t, X_t) \le I(\theta^\star; \mathcal{H}_{T+1}). \tag{55}$$

Substitution into equation 52 gives

$$\mathbb{E}\left[\sum_{t=1}^{T} (\mu_{B_t}(X_t; \theta^\star) - \mu_{A_t}(X_t; \theta^\star))\right] \le \sqrt{\frac{KT}{2} I(\theta^\star; \mathcal{H}_{T+1})}. \tag{56}$$

Step 5: approximation error. Fix $(\mathcal{H}_t, X_t) = (h, x)$. Conditional on $(\theta^\star, \mathcal{H}_t, X_t)$, generate τ synthetic outcome vectors

$$\widetilde{Y}_t^{B,(1)}, \ldots, \widetilde{Y}_t^{B,(\tau)} \overset{\text{i.i.d.}}{\sim} p_Y(\cdot \mid x, \theta^\star),$$

which, as usual, we denote as $\widetilde{D}_t^B$. Define

$$\overline{R}_{t,a}^B = \frac{1}{\tau} \sum_{i=1}^{\tau} R(\widetilde{Y}_{t,a}^{B,(i)}), \qquad \widehat{A}_t \in \arg\max_{a \in [K]} \overline{R}_{t,a}^B.$$

For IRTS-A, $B_t = \widehat{A}_t$ by construction. For IRTS-C,

$$B_t \in \arg\max_{a \in [K]} \mathbb{E}\Big[R(Y_{t,a}) \,\Big|\, \mathcal{H}_t = h, X_t = x, \widetilde{D}_t^B\Big]\,.$$

Since $\widehat{A}_t$ is measurable with respect to $\widetilde{D}_t^B$, conditional optimality of B_t gives

$$\begin{aligned} &\mathbb{E}\Big[R(Y_{t,B_t}) \,\Big|\, \mathcal{H}_t = h, X_t = x, \widetilde{D}_t^B\Big] \\ &\quad \ge \mathbb{E}\Big[R(Y_{t,\widehat{A}_t}) \,\Big|\, \mathcal{H}_t = h, X_t = x, \widetilde{D}_t^B\Big]\,. \end{aligned}$$

Consequently, for IRTS-C,

$$\begin{aligned} &\mathbb{E}[\mu^\star(x; \theta^\star) - \mu_{B_t}(x; \theta^\star) \,|\, \mathcal{H}_t = h, X_t = x] \\ &\quad \le \mathbb{E}\Big[\mu^\star(x; \theta^\star) - \mu_{\widehat{A}_t}(x; \theta^\star) \,\Big|\, \mathcal{H}_t = h, X_t = x\Big]\,. \end{aligned} \tag{57}$$

For IRTS-A, the two sides of equation 57 are equal.

It remains to bound the right-hand side. Fix θ and condition further on $\theta^\star = \theta$. Let

$$a^\star(x,\theta) \in \arg\max_{a\in[K]} \mu_a(x;\theta), \qquad \xi_a = \overline{R}^B_{t,a} - \mu_a(x;\theta).$$

Since $\widehat{A}_t$ maximises the empirical mean rewards,

$$\mu^\star(x;\theta) - \mu_{\widehat{A}_t}(x;\theta) \le \max_{a\in[K]} \xi_a - \xi_{a^\star(x,\theta)}. \tag{58}$$

Conditional on $(\mathcal{H}_t = h, X_t = x, \theta^\star = \theta)$, the empirical means are unbiased and $a^\star(x,\theta)$ is nonrandom. Hence

$$\mathbb{E}[\xi_{a^\star(x,\theta)} \mid \mathcal{H}_t = h, X_t = x, \theta^\star = \theta] = 0.$$

Since the rewards lie in $[0,1]$, Hoeffding's lemma gives, for every $\lambda > 0$,

$$\mathbb{E}\big[e^{\lambda\xi_a} \,\big|\, \mathcal{H}_t = h, X_t = x, \theta^\star = \theta\big] \le \exp\left(\frac{\lambda^2}{8\tau}\right).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\max_{a\in[K]} \xi_a \,\middle|\, \mathcal{H}_t = h, X_t = x, \theta^\star = \theta\right] \\
&\quad\le \frac{1}{\lambda}\log\left(\sum_{a=1}^{K} \mathbb{E}\big[e^{\lambda\xi_a} \,\big|\, \mathcal{H}_t = h, X_t = x, \theta^\star = \theta\big]\right) \\
&\quad\le \frac{\log K}{\lambda} + \frac{\lambda}{8\tau}.
\end{aligned}$$

Choosing $\lambda = \sqrt{8\tau \log K}$ yields

$$\mathbb{E}\Big[\mu^\star(x;\theta) - \mu_{\widehat{A}_t}(x;\theta) \,\Big|\, \mathcal{H}_t = h, X_t = x, \theta^\star = \theta\Big] \le \sqrt{\frac{\log K}{2\tau}}.$$

This calculation requires independence across the τ sample indices, but not independence across arms. Integrating over $\theta^\star \mid \mathcal{H}_t = h, X_t = x$ and applying equation 57 proves equation 45 for both IRTS-A and IRTS-C.

Finally, combining equation 46 and equation 56 gives

$$\mathcal{R}(T) \le T\sqrt{\frac{\log K}{2\tau}} + \sqrt{\frac{KT}{2} I(\theta^\star; \mathcal{H}_{T+1})}.$$

□

Algorithm 3 Contextual IRTS-A/C

Require: Arms $[K]$, horizon T, budget τ, mode $M \in \{A, C\}$

for $t = 1, \dots, T$ **do**
  Observe X_t; let $\mathcal{H}_t$ be the observed history
  Initialise $\widetilde{D}_t \leftarrow \varnothing$
  for $i = 1, \dots, \tau$ **do**
    Draw the joint outcome vector
    $\widetilde{Y}_t^{(i)} \leftarrow \left(\widetilde{Y}_{t,a}^{(i)}\right)_{a\in[K]} \sim p_Y\Big(\cdot \mid \mathcal{H}_t, \widetilde{D}_t, X_t\Big)$
    $\widetilde{D}_t \leftarrow \widetilde{D}_t \cup \{X_t, \widetilde{Y}_t^{(i)}\}$
  end for
  for $a = 1, \dots, K$ **do**
    if $M = A$: $\widehat{\mu}^A_{t,a}(X_t) \leftarrow \tau^{-1} \sum_{i=1}^{\tau} R(\widetilde{Y}_{t,a}^{(i)})$
    if $M = C$: $\widehat{\mu}^C_{t,a}(X_t) \leftarrow \mathbb{E}[R(Y_{t,a}) \mid \mathcal{H}_t, \widetilde{D}_t, X_t]$
  end for
  $A_t \in \arg\max_{a\in[K]} \widehat{\mu}^M_{t,a}(X_t)$
  Play A_t and observe Y_{t,A_t}
end for

B Implicit Thompson Sampling for Contextual Bandits

The contextual variants use the same shared-block construction. At round t, after observing X_t, the learner generates τ imagined outcomes for every arm at that context. Each pair $(X_t, a, \widetilde{Y}_{t,a}^{(i)})$ is appended to the common synthetic dataset before the next outcome is drawn. Thus the entire block again corresponds to one posterior draw of the environment. This is summarised in Algorithm 3.

C Additional Experimental Details

C.1 Analytic results

For the analytic settings, where the agents have conjugate Bayesian distributions, we consider four bandit settings: Bernoulli, Gaussian, Mismatch, and Contextual.

C.1.1 Bandit instances.

Bernoulli. *Environment.* Each arm a has a mean reward drawn independently from $\mu_a \sim \mathrm{Beta}(1, 1)$, with Bernoulli rewards:

$$Y_{t,a} \mid \mu_a \sim \mathrm{Bernoulli}(\mu_a).$$

Predictive model. The agent uses the correctly-specified model: a $\mathrm{Beta}(1, 1)$ prior with Bernoulli likelihood. The posterior before round t is $p(\mu_a \mid \mathcal{H}_t) = \mathrm{Beta}(1 + n_a^+, 1 + n_a^-)$, where n_a^+ and n_a^- denote the number of successes and failures for arm a.

Gaussian. *Environment.* Each arm a has a mean reward drawn independently from $\mu_a \sim \mathcal{N}(0, 1)$, with Gaussian rewards:

$$Y_{t,a} \mid \mu_a \sim \mathcal{N}(\mu_a, 1).$$

Predictive model. The agent uses the correctly-specified model: a $\mathcal{N}(0, 1)$ prior with Gaussian likelihood. After n_a observations with sample mean $\overline{Y}_a$, the posterior is

$$p(\mu_a \mid \mathcal{H}_t) = \mathcal{N}\left(\frac{n_a \overline{Y}_a}{1 + n_a}, \frac{1}{1 + n_a}\right).$$

Mismatch. *Environment.* Identical to the Bernoulli setting: $\mu_a \sim \mathrm{Beta}(1, 1)$ with $Y_{t,a} \mid \mu_a \sim \mathrm{Bernoulli}(\mu_a)$.

Table 2: Average regret and standard error in the mean after 1 million time steps across 100 Bayesian bandit seeds. We bold methods whose mean regret is within one standard error of the lowest mean regret, using the standard error of the method attaining that minimum.

Method	Bernoulli	Gaussian	Mismatch	Contextual
TS	$\mathbf{60.8 \pm 7.4}$	$\mathbf{118 \pm 11}$	$\mathbf{275 \pm 18}$	305.2 ± 4.5
Greedy	46644 ± 8146	234083 ± 42911	224713 ± 22353	15986 ± 3713
IRTS-A, $\tau = 1$	177462 ± 12302	352464 ± 9523	250480 ± 8091	220401 ± 2991
IRTS-A, $\tau = 5$	57116 ± 3273	97678 ± 5553	158211 ± 4519	49866 ± 1064
IRTS-A, $\tau = 10$	29688 ± 2011	51843 ± 4251	110666 ± 2950	25554 ± 584
IRTS-C, $\tau = 1$	23406 ± 5794	6059 ± 2867	539 ± 287	$\mathbf{201.9 \pm 3.5}$
IRTS-C, $\tau = 5$	5058 ± 1693	3152 ± 2370	324 ± 212	212.3 ± 3.0
IRTS-C, $\tau = 10$	1805 ± 1140	616 ± 388	337 ± 199	226.6 ± 3.3

Predictive model. The agent uses a *misspecified* model: the Gaussian model from the Gaussian setting above, despite the true rewards being Bernoulli. This setting allows us to assess algorithmic robustness to model misspecification.

Contextual. *Environment.* A Bayesian linear contextual bandit. Each arm a has an unknown weight vector $\theta_a \sim \mathcal{N}(\mathbf{0}, I_5)$ and scalar bias $\theta_a^{(0)} \sim \mathcal{N}(0, 1)$. At each step t, the agent observes a context $X_t \sim \mathcal{N}(\mathbf{0}, I_5)$ and receives reward

$$Y_{t,a} \mid X_t, \theta_a, \theta_a^{(0)} \sim \mathcal{N}\Big(\theta_a^\top X_t + \theta_a^{(0)}, 1\Big).$$

Predictive model. The agent uses the correctly-specified Bayesian linear model: independent $\mathcal{N}(\mathbf{0}, I_5)$ priors on each θ_a, $\mathcal{N}(0, 1)$ priors on each $\theta_a^{(0)}$, and likelihood

$$p_{Y_a}(Y_{t,a} \mid X_t, \theta_a, \theta_a^{(0)}) = \mathcal{N}\Big(\theta_a^\top X_t + \theta_a^{(0)},\ 1\Big).$$

C.1.2 Methods

We evaluate IRTS-A and IRTS-C for $\tau \in \{1, 5, 10\}$ across all four settings, alongside two baselines: Greedy and Thompson Sampling. We run each algorithm for 1 million time steps across 100 seeds.

Greedy. Selects the arm with the highest posterior predictive mean. In non-contextual settings (Bernoulli, Gaussian, Mismatch), $A_t = \arg\max_a \mathbb{E}[\mu_a \mid \mathcal{H}_t]$; in the contextual setting, the mean is additionally conditioned on the current context, $A_t = \arg\max_a \mathbb{E}[\theta_a^\top X_t + \theta_a^{(0)} \mid \mathcal{H}_t, X_t]$.

Thompson Sampling (TS). Selects the arm with the highest value under a posterior sample. In non-contextual settings, $A_t = \arg\max_a \widetilde{\mu}_a$, where $\widetilde{\mu}_a \sim p(\mu_a \mid \mathcal{H}_t)$; in the contextual setting, $A_t = \arg\max_a(\widetilde{\theta}_a^\top X_t + \widetilde{\theta}_a^{(0)})$, where $(\widetilde{\theta}_a, \widetilde{\theta}_a^{(0)}) \sim p(\theta_a, \theta_a^{(0)} \mid \mathcal{H}_t)$. Note that in the Contextual setting, this is equivalent to Linear TS (Agrawal & Goyal, 2013), the method used as a baseline in the TabPFN experiments (Section C.2).

C.1.3 Results

We provide the average regret and standard error across our experiments in Table 2.

C.2 TabPFN

For the TabPFN experiments, we use `TabPFN v2.6` (Hollmann et al., 2025), licensed under the Prior Labs License (Apache 2.0 with additional attribution requirement; `https://github.com/PriorLabs/TabPFN`).

TabPFN joint and independent action sampling. As part of Algorithm 2, we require an autoregressive predictive sample sampled from the joint over actions $\widetilde{Y}_t^{(i)} \sim p_Y\left(\cdot \mid \mathcal{H}_t, \widetilde{D}_t\right)$, which is a vector of size K. However, TabPFN v2.6 does not natively provide the option to sample a joint

Table 3: Average cumulative regret and standard error of the mean on the Wheel bandit task after 2500 time steps across 10 bandit seeds. IRTS variants use $\tau = 1$. We bold methods whose mean regret is within one standard error of the lowest mean regret, using the standard error of the method attaining that minimum.

Linear TS	Greedy	IRTS-A	IRTS-C
20831 ± 6360	59939 ± 13312	$\mathbf{12243 \pm 1053}$	36434 ± 12379

vector, and multivariate outputs are processed independently across the output dimension. Hence, we construct a joint sample in the following way sequentially for each index $i \in 1, \cdots, \tau$. First, we randomly order the actions. Without loss of generality, denote this ordering as $a_1, \cdots a_K$. Then, we sample a value for Y_{t,a_1} with $\mathcal{H}_t, \widetilde{D}_t$ in the context, and append (a_1, Y_{t,a_1}) to $\widetilde{D}_t$. We then sample Y_{t,a_2} with $\mathcal{H}_t, \widetilde{D}_t$ in the context, and append (a_2, Y_{t,a_2}) to $\widetilde{D}_t$. This is done iteratively, until we have exhausted all K actions and produced an joint outcome vector. We refer to this scheme as 'joint-action sampling'.

On the other hand, an alternative sampling scheme is to independently construct samples for each action. That is, for each action pass τ autoregressively generated samples conditioned only on that one action's imagined rewards. This would be equivalent to a joint sample in the case where the arms are independent, and we therefore refer to it as 'independent-action sampling'.

C.2.1 Wheel bandit

The Wheel bandit is a bandit task introduced in Riquelme et al. (2018) to address the issue that some bandit tasks derived from real data do not require algorithms to carry out sophisticated or thorough exploration to be successful. In this task, the context X is sampled uniformly from the unit disk. A parameter $\delta \in (0, 1)$ divides the context space into a central and an outer region. For $\|X\| \leq \delta$, action 1 is optimal, with $R_1 \sim \mathcal{N}(1.2, 0.01^2)$, while actions $a \in \{2, 3, 4, 5\}$ yield $R_a \sim \mathcal{N}(1.0, 0.01^2)$.

For $\|X\| > \delta$, the optimal action among $\{2, 3, 4, 5\}$ depends on the quadrant containing X. The corresponding optimal action yields $R_a \sim \mathcal{N}(50, 0.01^2)$, while the remaining three outer actions yield $R_a \sim \mathcal{N}(1.0, 0.01^2)$. Action 1 continues to yield $R_1 \sim \mathcal{N}(1.2, 0.01^2)$. Thus, action R_1 can be interpreted as a 'safe' action, but an agent with strong exploration capabilities should be able to determine that the other actions can be highly rewarding in their corresponding regions. Larger values of δ make the high-reward outer region rarer and the exploration problem more difficult. We use $\delta = 0.5$ as a default. A visual representation of the wheel bandit is provided in Fig. 5.

C.2.2 Methods

The non-IRTS baselines we evaluate are the following:

- **Greedy.** We select the action with the highest TabPFN predictive mean.
- **Linear TS.** maintains a per-action Gaussian posterior over a linear reward model; the model is linear in the x and y co-ordinates. We employ an independent Gaussian prior with mean zero and unit standard deviation and an observation noise with standard deviation of of 0.01, matching the ground-truth environment reward generation noise. As discussed in the main text, the linear model is not sufficiently expressive to accurately model the non-linear underlying reward function.
- **IRTS.** We apply both IRTS with $\tau = 1$ and the joint-action sampling scheme.

C.2.3 Wheel Bandit Results

We provide the average regret and standard error across experiments originally presented in the main text's Fig. 3. The time horizon is $T = 2500$ steps and reuslts are averaged over 10 seeds. The numeric mean and standard errors are presented in Table 3, and the regret curves and distributions during learning are shown in Fig. 6.

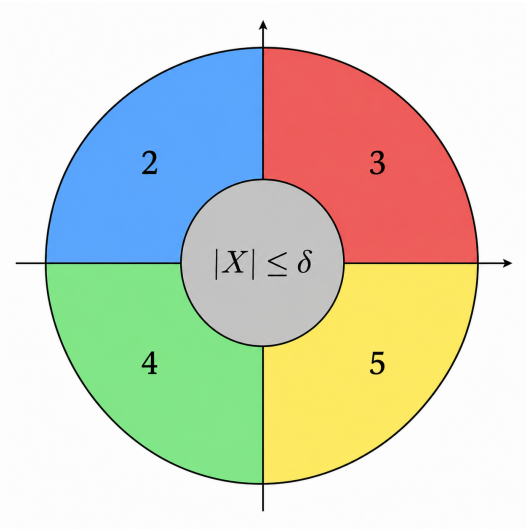

Figure 5: Visual representation of the wheel bandit. In the numbered regions, the corresponding-number action is optimal. In the middle grey disk, action 1 is optimal. We use $\delta = 0.5$.

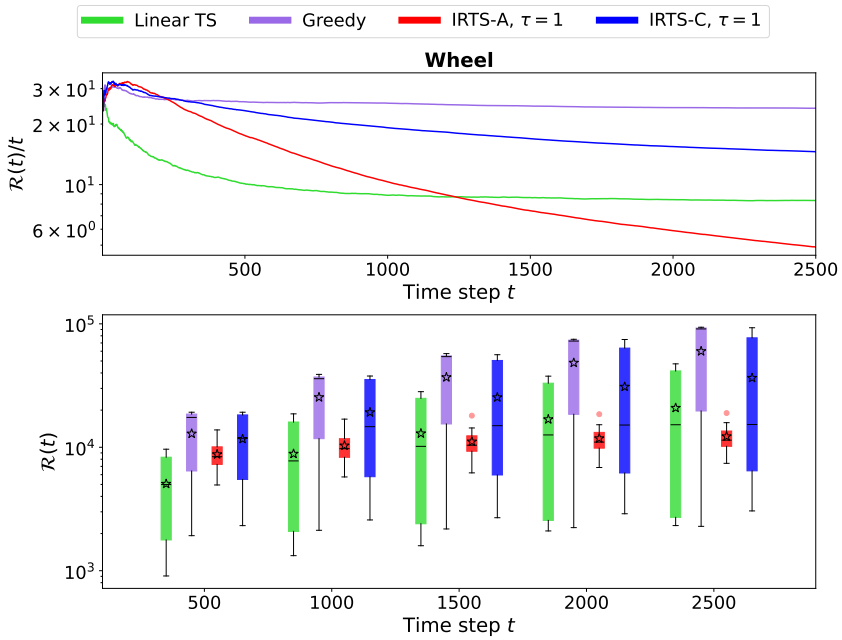

Figure 6: Average-regret-so-far $\mathcal{R}(t)/t$ (top) and cumulative-regret box plots (bottom, with means marked by $\star$) for the Wheel bandit task. Results use 10 runs over $T = 2500$ steps, with $\tau = 1$ for both IRTS variants.

Table 4: Average regret and standard error in the mean on Movielens bandit task after 1500 time steps across 10 bandit seeds. We bold methods whose mean regret is within one standard error of the lowest mean regret, using the standard error of the method attaining that minimum.

Experiment	Linear TS (untuned)	Linear TS	Greedy	IRTS-A	IRTS-C
Deterministic	1922 ± 110	918 ± 69	1148 ± 169	735 ± 22	$\mathbf{672 \pm 14}$
Stochastic	476.4 ± 8.9	$\mathbf{159 \pm 15}$	235 ± 14	306.8 ± 4.9	201.3 ± 6.5

C.2.4 MOVIELENS

On this task, we evaluated IRTS-A and IRTS-C algorithms with $\tau = 1$. We consider the standard greedy algorithm and linear thompson sampling (Linear TS) Agrawal & Goyal (2013) as baselines. We evaluate on a semi-synthetic contextual bandit derived from the MovieLens 100k-ratings dataset (Nie et al., 2024). At each step, the agent is presented with $K = 10$ candidate movies and a user context, and must select one to recommend. The context is a 5-dimensional numeric feature vector obtained from a rank $k = 5$ SVD decomposition of the user-movie rating matrix (i.e. the latent user embedding). Each movie is similarly represented by its 5-dimensional item embedding.

All experiments in the main text run for $T = 1500$ steps and are averaged over 10 random seeds.

We consider two variants of the reward signal:

- **Deterministic.** The reward is the raw 1-5 star rating predicted by the low-rank factorisation, treated as a continuous signal (identity conversion). TabPFN is instantiated as a *regressor* (`TabPFNRegressor`).
- **Stochastic.** The continuous rating is converted to a binary Bernoulli reward by treating the normalised rating $r/5$ as a success probability and drawing a Bernoulli sample (proportional conversion). TabPFN is instantiated as a *classifier* (`TabPFNClassifier`).

C.2.5 METHODS

- **Greedy.** We select the action with the highest TabPFN predictive mean.
- **Linear TS.** maintains a per-action Gaussian posterior over a linear reward model. The model uses a 5-dimensional context (matching the rank-5 SVD features), Gaussian prior with mean $\mu_0 = 0$ and variance $\sigma_0^2 = 1$ on all weight components (including the intercept), and observation noise $\sigma_r^2 \in \{0.01, 0.1, 1\}$ is selected to be the best-performing value. We also report results for an untuned TS with default parameter, $\sigma_r^2 = 1$. As discussed in the main text (Section 5.2), the semi-synthetic data-generation process is well-aligned with the Linear TS model assumptions, making it a particularly strong baseline for this task.
- **IRTS.** We apply both IRTS with $\tau = 1$ for the full horizon in the main text of $T = 1500$ and $\tau = 1, 5, 10$ for a shorter time horizon of $T = 1000$. For these experiments, we employ the independent-action sampling scheme.

C.2.6 MOVIELENS RESULTS

We provide the average regret and standard error across our TabPFN experiments presented in the main text in Table 2.

We additionally report results at $T = 1000$ steps for all $\tau \in \{1, 5, 10\}$ in Figure 7, which corroborates the findings above and shows that the relative ordering of methods is stable across the full range of τ values.

C.2.7 OPENML

We evaluate Greedy, LinTS, IRTS-A IRTS-C on contextual bandit environments derived from eight supervised learning classification OpenML datasets. For each dataset, we use the features of the data samples as the context. The reward then depends on the action taken. Each dataset has number of actions corresponding to number of classes, and if the bandit takes the action corresponding to

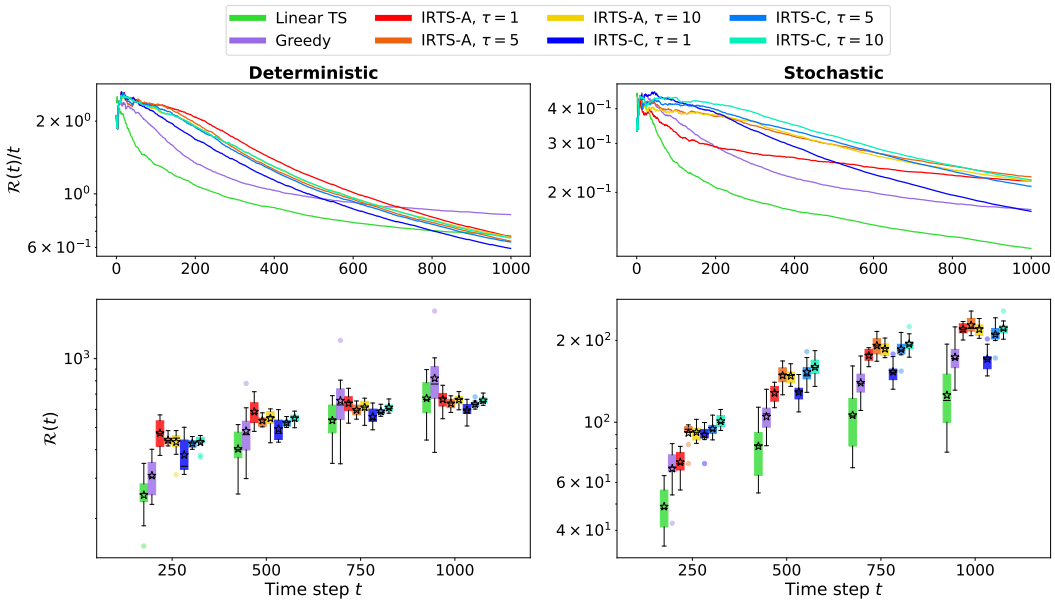

Figure 7: Average-regret-so-far $\mathcal{R}(t)/t$ (top row) and corresponding box-plots (bottom row, with mean regrets displayed with a star $\star$ symbol) for the MovieLens experiments with TabPFN across all values of $\tau \in \{1, 5, 10\}$ (10 runs, $T = 1000$ steps). Left and right columns show the Deterministic and Stochastic variants respectively. Compared with the main-text results (Figure 3, $T = 1500$), IRTS-C remains competitive with tuned Linear TS across all τ, while IRTS-A degrades in the Stochastic setting for all τ.

the correct class it sees a reward of 1 with probability 0.75, and otherwise sees a reward of 0. On the other hand, if the action is incorrect, the reward is 1 with probability 0.25 and 0 otherwise. The stochasticity was introduced to align the supervised learning task slightly towards more classic bandit tasks, where stochasticity in the reward is a key challenge that needs to be accounted for.

C.2.8 Methods

- **Greedy.** We select the action with the highest TabPFN predictive mean.
- **Linear TS.** maintains a per-action Gaussian posterior over a linear reward model; the model is linear in the data features. We employ an independent Gaussian prior with mean zero and unit standard deviation and an observation noise with standard deviation of of $0.43 \approx \sqrt{0.75 \times 0.25}$, matching the ground-truth environment reward generation noise from the Bernoulli reward random variable.
- **IRTS.** We evaluate both IRTS variants with $\tau = 1$ under the join-action sampling scheme.

C.2.9 OpenML Results

We report results across 5 seeds per task for eight datasets with time horizons of $T = 2000$. In Fig. 8 we report the individual regret curves throughout training for each task and the final cumulative regret and standard error is summarised

C.3 LLMs

C.3.1 LLM buttons-bandit experimental details

Bandit instance. The buttons bandit is a Bernoulli multi-armed bandit with $K = 5$ arms ("Buttons"), which was used as a bandit problem in Krishnamurthy et al. (2024). Each instance is parameterised by a base success rate p and a gap δ: a single optimal arm, drawn uniformly at random per seed, has Bernoulli parameter $p + \delta/2$, while the remaining four sub-optimal arms share Bernoulli

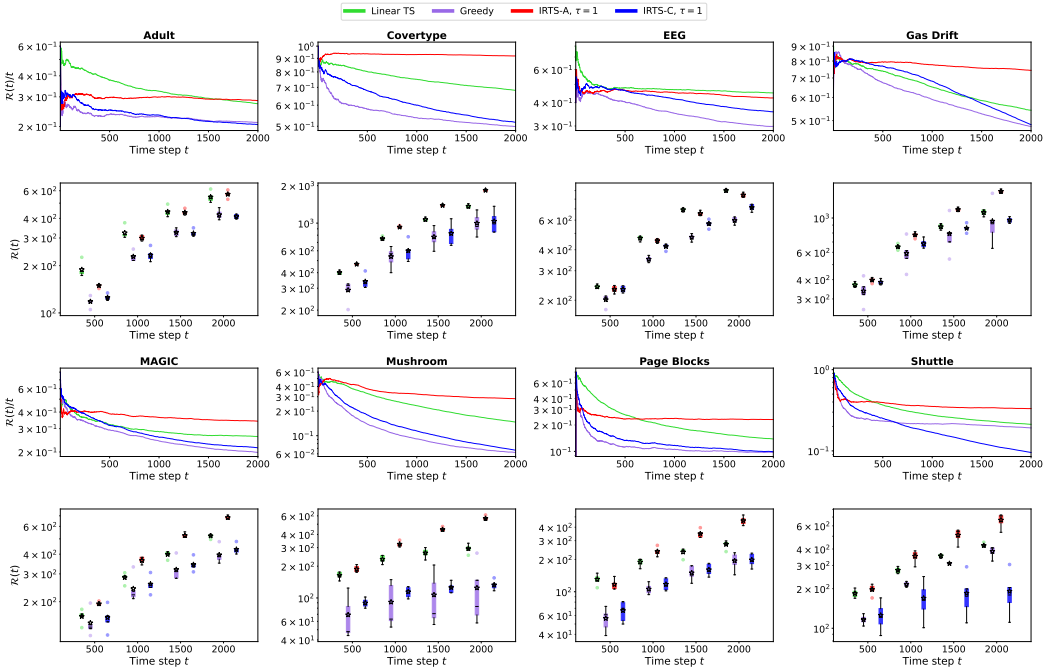

Figure 8: Average-regret-so-far $\mathcal{R}(t)/t$ and cumulative-regret box plots for the eight OpenML bandit tasks, with curves above the corresponding box plots and means marked by $\star$. Results use 5 runs per task over $T = 2000$ steps, with $\tau = 1$ for both IRTS variants.

Table 5: Average cumulative regret and standard error of the mean on the OpenML bandit tasks after 2000 time steps across 5 bandit seeds per task. IRTS variants use $\tau = 1$. We bold methods whose mean regret is within one standard error of the lowest mean regret, using the standard error of the method attaining that minimum.

Experiment	Linear TS	Greedy	IRTS-A	IRTS-C
Adult	547 ± 19	422 ± 14	570 ± 13	$\mathbf{411 \pm 6}$
Covertype	1364 ± 21	$\mathbf{998 \pm 88}$	1837 ± 14	$\mathbf{1036 \pm 97}$
EEG	901 ± 7	$\mathbf{598 \pm 12}$	847 ± 11	714 ± 15
Gas Drift	$\mathbf{1085 \pm 21}$	$\mathbf{954 \pm 135}$	1487 ± 17	$\mathbf{968 \pm 17}$
MAGIC	522 ± 7	$\mathbf{396 \pm 22}$	682 ± 8	429 ± 14
Mushroom	293 ± 12	$\mathbf{125 \pm 39}$	567 ± 12	$\mathbf{133 \pm 6}$
Page Blocks	277 ± 11	$\mathbf{195 \pm 15}$	461 ± 18	$\mathbf{198 \pm 12}$
Shuttle	425 ± 6	384 ± 16	664 ± 36	$\mathbf{191 \pm 32}$

parameter $p - \delta/2$. We use $p = 0.7$ and $\delta = 0.2$ throughout, so that $\Pr(Y_{t,A^\star} = 1) = 0.8$ and $\Pr(Y_{t,a} = 1) = 0.6$ for each sub-optimal arm a. Each run consists of $T = 500$ interaction steps and we average over 10 random seeds (each seed re-shuffles the index of the optimal arm).

Models. We use three open-weight base language models, served through a `vLLM` (Kwon et al., 2023) (Apache 2.0) HTTP completions endpoint:

- `mistralai/Mistral-7B-v0.1` (Jiang et al., 2023), Apache 2.0 license (`https://huggingface.co/mistralai/Mistral-7B-v0.1`);
- `Qwen/Qwen3-14B` (Yang et al., 2025), Apache 2.0 license (`https://huggingface.co/Qwen/Qwen3-14B`);
- `meta-llama/Llama-3.1-8B` (Grattafiori et al., 2024), Llama 3.1 Community License (`https://huggingface.co/meta-llama/Llama-3.1-8B`).

Predictive action-selection methods (greedy, IRTS-A, IRTS-C). For these three methods we use the LLM purely as an in-context predictor of the next outcome, and apply the action-selection rules of Section 3. The prompt is constructed as a plain transcript of actions and observed outcomes with no natural-language scaffolding; concretely, the prompt fed to the LLM at time t to query action a has the form

Predictive action-selection methods prompt

```
action = a_1 <reward>r_1</reward>
action = a_2 <reward>r_2</reward>
...
action = a_t <reward>r_t</reward>
action = a <reward>
```

where each $a_i \in \{0, 1, 2, 3, 4\}$ is the integer button index and each $r_i \in \{0, 1\}$ is the binary outcome observed previously. Here $R(y) = y$, so the outcome and reward coincide. The model is queried with `max_tokens=1` so that completions consist of a single token immediately after the open `<reward>` tag. From each completion we read out the top-K token logprobs and:

- For the *predictive mean* (used by greedy and by the conditional mean of IRTS-C), we extract the logprobs of the tokens "0" and "1", renormalise them to a Bernoulli distribution, and take its mean as $\mathbb{E}[R(Y_{t,a}) \mid \mathcal{H}_t]$.
- For *imagined-reward sampling* (used by IRTS-A and IRTS-C), we read the sampled token directly and parse it as 0 or 1.

The buttons model factorises across arms, so the shared synthetic block of Algorithm 2 can be generated arm by arm, under the independent-action sampling scheme. For each $a \in [K]$, we generate τ rewards autoregressively, appending each sampled outcome to the prompt before drawing the next. The three methods then differ only in how they aggregate the resulting rewards:

- **Greedy** corresponds to $\tau = 0$: no imagined rewards are drawn, and the score for arm a is the predictive mean reward $\mathbb{E}[R(Y_{t,a}) \mid \mathcal{H}_t]$ read directly from the logprobs.
- **IRTS-A** ($\tau = 1$): the score is the empirical mean reward of the imagined outcome(s); with $\tau = 1$ this reduces to the reward of the single sampled outcome, recovering the TS-LLM scheme of Sun et al. (2025) without temperature scheduling.
- **IRTS-C** ($\tau = 1$): one imagined outcome is drawn and appended to the prompt, and the score is the conditional mean reward read from the logprobs of the resulting extended prompt.

The first $K = 5$ steps of every run are used to play each arm once (warm-up), so that every arm has at least one observation in the history before any LLM-based scoring is performed. Initial Bernoulli parameters and indexing of the optimal arm are determined by the seed.

Table 6: Regret from the LLM Buttons Experiments. We bold those methods within standard error of the one with the lowest mean regret; reported uncertainties are standard errors.

Method	Mistral 7B	Qwen 3	LLaMA 3.1 8B
Greedy	28.3 ± 7.1	$\mathbf{43 \pm 15}$	$\mathbf{19.6 \pm 8.8}$
Instruction	83.2 ± 1.1	80.0 ± 5.2	83.68 ± 0.86
IRTS-C	$\mathbf{23.8 \pm 3.0}$	$\mathbf{36 \pm 12}$	$\mathbf{22.6 \pm 8.8}$
IRTS-A	86.2 ± 7.0	86.4 ± 7.5	83.7 ± 8.4

Instruction baseline. The Instruction baseline asks the LLM to directly choose the next action, mirroring the prompting style used in Song et al. (2025); Nie et al. (2024); Lim et al. (2025); Xia et al. (2025). We use the following prompt template:

Instruction baseline prompt

```
You are interacting with a multi-armed bandit. Each arm
yields a reward of 0 or 1. Choose the action you predict will
 yield the highest reward in the long run.
Available actions:
1. 0
2. 1
3. 2
4. 3
5. 4
Past history:
action = a_1 <reward>r_1</reward>
...
action = a_t <reward>r_t</reward>

Which action will you choose next? Respond with the action's
name only.
```

Generation uses `max_tokens=30` and we parse the response by (i) exact match to an action name, (ii) interpretation as a 1-indexed integer, or (iii) a unique substring match against the action set. If the response cannot be parsed into a valid action, the prompt is re-issued with a fresh sampling seed, up to 5 retries; if all retries fail, an action is chosen uniformly at random. As with the predictive methods, the first K steps play each arm once before any LLM call is made.

Inference setup. All LLMs are served by a single `vLLM` instance per model on the cluster, queried over HTTP using the OpenAI-compatible `/v1/completions` endpoint. The predictive methods use `logprobs=1` (which, for the `vLLM` server we use, returns the most likely token together with its logprob and the logprob of the sampled token) and the Instruction baseline uses `logprobs=1` with `max_tokens=30`. No temperature, top-p, or repetition-penalty scheduling is applied; all decoding parameters are left at their server defaults so that comparisons isolate the effect of the action-selection rule. Per-request seeds are drawn from a per-run `numpy` random generator so that runs are reproducible while individual LLM calls within a run are stochastic.

C.3.2 LLM Results

Table 6 reports cumulative regret for the LLM experiments.

D Code Availability and Computational Resources

Code: The anonymised codebase, including all model implementations and scripts to reproduce the experiments presented in this paper, is publicly available at: `https://anonymous.4open.science/r/Implicit-Thompson-Sampling-Review-Version-CB18`.

Computational Resources: We use the following computational resources for our experiments

- Analytic Experiments: All experiments were conducted on a university high-performance computing cluster with heterogeneous nodes. Each job used up to 32 CPU cores on x86_64 server-class processors (Intel Xeon family) with 128 GB RAM per node. Experiments were run on single nodes. Approximately 20 CPU hours were required to run all analytic experiments. Each analytic run (1 seed, 1M steps) required approximately 5-10 seconds.
- TabPFN Experiments: All experiments were conducted on a university high-performance computing cluster with heterogeneous nodes. Each job used up to 32 CPU cores on x86_64 server-class processors (Intel Xeon family) with 128 GB RAM per node. Experiments were run on single nodes. Approximately 200 CPU hours were required to run all TabPFN experiments. Each TabPFN run (1 seed, 1500 steps) required approximately 30-90 minutes.
- LLM Experiments: All models were hosted on rented RTX 3090 GPUs (24GB VRAM each). `Qwen/Qwen3-14B` required 2 RTX 3090 GPUs to host. `meta-llama/Llama-3.1-8B` and `mistralai/Mistral-7B-v0.1` required 1 GPU each. Approximately 300 GPU hours were required to run all experiments across the 3 models. Each run (1 seed, 500 steps) required approximately 30-60 minutes. This baseline efficiency can be improved across multiple seeds by running multiple models in parallel on the same GPU.

The reported totals only reflect the final experiments included in the manuscript. We estimate this represents about 20-40% of the overall computational budget for this project.