Should You Use Your Large Language Model to Explore or Exploit?

Anonymous Authors[1]

Abstract

We evaluate the ability of the current generation of large language models (LLMs) to help a decision-making agent facing an exploration-exploitation tradeoff. While previous work has largely study the ability of LLMs to solve combined exploration-exploitation tasks, we take a more systematic approach and use LLMs to explore and exploit *in silos* in various (contextual) bandit tasks. We find that reasoning models show the most promise for solving exploitation tasks, although they are still too expensive or too slow to be used in many practical settings. Motivated by this, we study tool use and in-context summarization using non-reasoning models. We find that these mitigations may be used to substantially improve performance on medium-difficulty tasks, however even then, all LLMs we study perform worse than a simple linear regression, even in non-linear settings. On the other hand, we find that LLMs do help at exploring large action spaces with inherent semantics, by suggesting suitable candidates to explore.

1. Introduction

The machine learning community is increasingly interested to apply advances in generative AI and large language models (LLMs) to decision-making problems. Early work in this direction has already produced impressive agentic behavior in both virtual (e.g., Wang et al., 2023; OpenAI, 2025) and physical-world environments (e.g., Black et al., 2024). Beyond generalization (needed for supervised learning), decision-making under uncertainty requires two additional capabilities: *exploitation* (making best decisions given the current data) and *exploration* (for long-term benefit). Balancing the two has led to a large literature (e.g., Slivkins, 2019; Lattimore & Szepesvári, 2020; Agarwal et al., 2020).

A recent line of work (e.g., Krishnamurthy et al., 2024; Nie et al., 2024) evaluates the ability of LLMs to balance exploration and exploitation entirely *in-context*, i.e., specifying the problem description, parameters, and history in the LLM prompt. Focused on simple tasks in reinforcement learning (RL), these results are mixed. LLMs fail to solve these tasks adequately out-of-the-box, but they can be prompted to do so by providing succinct summary statistics in-context. However, such statistics do not exist beyond simple RL problems, e.g., for contextual bandits. RL-specific pre-training or fine-tuning (on data from algorithmic baselines on similar problem instances) tends to work well (see Related Work), but may be prohibitive due to cost or insufficient training data. Besides, using a commonly available LLM would leverage its "generalist" intelligence and may be much easier in terms of logistics and required expertise.

Motivated by these observations, we study the ability of LLMs to explore and exploit in-context *in silos*, with an eye towards leveraging a pre-trained LLM (and the inductive bias therein) as a part of a larger decision-making agent. We focus on (contextual) bandits, as a standard abstraction for the explore-exploit tradeoff. We experiment with many LLMs: GPT-5.2, GPT-5-MINI, GPT-5-NANO, GPT-4, GPT-4O, GPT-3.5, QWEN-2.5, GEMMA-3, MISTRAL-7B, and DEEPSEEK-R1-DISTILL-QWEN, a reasoning model. [1]

In Section 2, we evaluate LLMs as *exploitation oracles* for contextual bandits. Given a history of (context, action, reward) tuples, the LLM is tasked with identifying the best action to take given a new context. Our results here are mixed. We show that LLMs can effectively exploit in-context for small-sized tasks but their performance degrades when the tasks become moderately sized. We find that frontier reasoning models outperform non-reasoning models (fixing model size and provider), but tend to be slow and/or expensive. We investigate to mitigations: in-context summary techniques and tool use (namely, a Python code interpreter). We find that both mitigations improve performance in a linear task, but on on-linear tasks LLMs with these mitigations still perform worse than a simple linear regression baseline.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Some other LLMs did not work for our purposes: LLAMA-3.2-3B would not follow instructions. Two reasoning models, DEEPSEEK-R1-DISTILL-LLAMA-70B and QWEN-2.5-MATH-70B have a context window that is too short.

In Section 3, we evaluate LLMs as an *exploration oracle* which suggests a small set of candidate actions by discretizing a large action space. (This set can then be used to instantiate an off-the-shelf bandit algorithm.) On various text-based multi-armed bandit tasks that we design, we find this approach far superior to a non-LLM baseline. Our first task is based on the MovieLens dataset (Harper & Konstan, 2015), where actions are movie recommendations, rewards are random draws parameterized by ground-truth movie ratings, and the baseline is a hand-crafted discretization based on movie genre. Second, we design a larger-scale bandit task based on paper titles/abstracts from arXiv: find a suitable title for a given abstract. Given the high dimensionality of the action space, discretization approaches from continuous bandit problems are inapplicable. Several prompting strategies all lead to relatively good exploration compared to natural baselines. Our third task is based on open-ended "philosophical" questions and contrarian answers.

Related work. Our results belong to a growing line of work on using pre-trained LLMs for in-context reinforcement learning (RL). Coda-Forno et al. (2023); Krishnamurthy et al. (2024); Nie et al. (2024); Monea et al. (2025); Xia et al. (2024); Park et al. (2024); Wu et al. (2024) evaluate the ability of LLMs to solve various multi-armed bandit and contextual bandit tasks, and find that the current generation of LLMs largely fail to solve these tasks in-context. Indeed, positive findings are restricted to very simple tasks and/or require substantial mitigations (which in turn do not readily extend beyond simple settings). Xia et al. (2024) use LLMs to solve dueling bandit tasks, and Park et al. (2024) also evaluate the ability of LLMs to learn in games. While our paper is primarily concerned with whether LLMs succeed as algorithms, several others (e.g., Schubert et al., 2024; Hayes et al., 2024; Coda-Forno et al., 2024) use in-context bandits (and many other tasks) to study whether LLMs exhibit human-like behavior/biases in decision-making.

A broader literature on in-context learning (starting from Brown et al. (2020)) aims to solve various tasks by providing all relevant information in the LLM prompt. The work on *exemplar selection* (selecting examples and other information to present in-context) (e.g., Khalifa et al., 2023; Zhang et al., 2022; Xiong et al., 2023; Tonglet et al., 2023) is relevant to our exploitation experiments.

A growing line of work aims to use LLMs as a part of a larger decision-making agent (e.g., Li et al., 2024; Zhou et al., 2023; Zhao et al., 2024). Our exploration experiments take inspiration from the work on using LLMs as "action priors" inside of a larger RL algorithm (Yan et al., 2024; Carta et al., 2023; Yao et al., 2024; Hao et al., 2023). Much of this work falls under the proposer-verifier framework of Snell et al. (2024), where an LLM proposes several possible sequences from which a verifier selects suitable candidates. In comparison, our goal is a more systematic evaluation of LLMs' abilities to explore large action spaces, in isolation from other components of the decision-making task.

Finally, a parallel line of work trains transformers to solve various RL tasks (e.g., Laskin et al., 2022; Lin et al., 2023; Lee et al., 2024; Raparthy et al., 2023; Xu et al., 2022; Lehnert et al., 2024; Mukherjee et al., 2024). Our negative results in Section 2 provide an additional foundation for this work, emphasizing the shortcomings on frontier LLMs that are not custom-trained for RL.

Background on multi-armed bandits (MAB). We consider tasks based on MAB and contextual bandits (CB), well-studied special cases of RL that abstract explore-exploit tradeoff (Slivkins, 2019; Lattimore & Szepesvári, 2020). In MAB, there are T rounds and K arms. In each round $t \in [T]$, the learner chooses an action (*arm*) $a_t \in [K]$ and observes reward r_t drawn from some (fixed, unknown) sub-Gaussian reward distribution $D(a_t)$ with mean $\mu(a_t)$. In CB, the learner also observes a context z_t before each round t. The expected reward $\mu(z_t, a_t)$ depends on both the context and the arm. The learner's goal is to balance exploration and exploitation to maximize cumulative reward.

An "exploitation oracle" (which optimizes for the current round given the history) naturally plugs into standard bandit algorithms such as Epsilon-Greedy, Explore-then-Commit, and Follow-The-Perturbed-Leader. Typical implementations in CB involve model-based (e.g., linear) regression or cost-sensitive classification (Slivkins, 2019, Ch.8). Designing CB exploitation oracles for their own sake, a.k.a. *offline policy optimization*, is well-studied (starting from, e.g., Beygelzimer & Langford (2009); Dudík et al. (2014)). Usage of an LLM as an "exploration oracle" is closely related to the literature on dynamic pricing and Lipschitz bandits (e.g., Kleinberg & Leighton, 2003; Kleinberg et al., 2008; 2019; Bubeck et al., 2011), as we elaborate in Sec. 3.

2. LLMs as exploitation oracles

We evaluate the ability of LLMs to *exploit* in decision-making tasks with statistical uncertainty on the outcomes. We present LLMs with in-context exploit tasks inspired by multi-armed bandits (MAB) and contextual bandits (CB). In a CB exploit task, an LLM is given a history consisting of context-arm-reward tuples, and is instructed to take the best arm given the current history and the current context. A MAB exploit task is the same, but without contexts. These tasks are generated from some parameterized distributions called *exploit puzzles*. Due to computational constraints, our experiments focus mostly on GPT model family, with additional LLMs evaluated as a robustness check in Appendix B.4, with similar findings. Across all experiments, varying the model temperature did not significantly affect

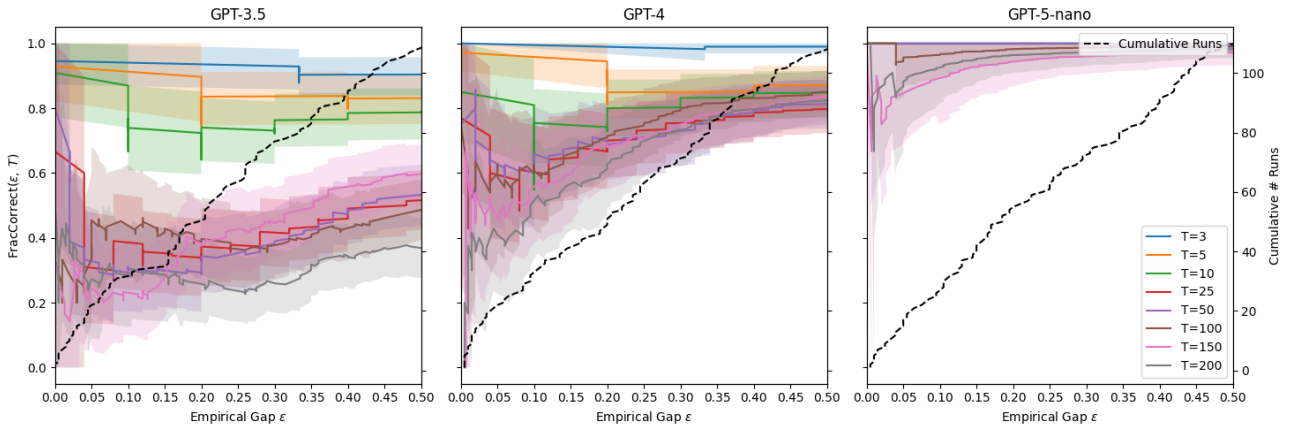


Figure 1. MAB exploit puzzle for GPT-3.5 (left), GPT-4 (middle), and GPT-5-NANO (right), all with the "buttons" prompt and temperature 1. Results for GPT-5-MINI are not visualized, as $\text{FracCorrect}(\epsilon, T) = 1$ for all ϵ and T. The following conventions apply to all figures in this section. Each line corresponds to a particular value of #rounds T and plots $\text{FracCorrect}(\epsilon, T)$ against empirical gap ϵ on the X-axis. The shaded band around the line represents a 95% confidence interval. The dashed line is the number of tasks ("runs") with empirical gap $\leq \epsilon$; the resp. Y-scale is on the right.

performance. We include results for temperature 0 and 1.

MAB exploit puzzles. We begin by testing the GPT model family on simple MAB exploit puzzles. We find that performance significantly improves as the models progress from GPT-3.5 to GPT-4 to GPT-5. In fact, GPT-5-MINI achieves perfect performance. Our experiments on GPT-4 and GPT-3.5 provide a partial explanation for the observation in (Krishnamurthy et al., 2024; Nie et al., 2024), that these models fail to solve end-to-end MAB tasks in-context when presented with raw (non-summarized) history.

Following these two papers, we try two prompts, where arms correspond, resp., to pushing different colored buttons and to showing different ads to users. The LLM is asked to choose the arm with the highest empirical reward in the next round. We also try chain-of-thought (CoT) prompts, for the total of 4 prompt designs: $\{$ buttons, adverts $\} \times \{$ CoT, no-CoT $\}$. See Appendix B.1 for more details on this setup.

Our MAB exploit puzzle is parametrized by gap $\Delta \in [0, 1]$ and history size T. The tasks, all with 5 arms, are constructed as follows. We pick an arm a^* uniformly-at-random (u.a.r.). Expected rewards are $\mu(a^*) = 1/2 + \Delta/2$ and $\mu(a) = 1/2 - \Delta/2$ for all other arms a. Then, we generate a history of T rounds for each arm a, where the reward $r_t(a)$ at each round $t \in [T]$ is an independent Bernoulli draw with mean $\mu(a)$. For a given T, we generate 10 tasks from this puzzle for each value of $\Delta \in \{0, .05, .1, .15, .2, .25, .3, .4, .45, .5\}$.

Given a (realized) exploit task, we measure its difficulty via *empirical gap* ϵ: the difference between the largest and second-largest average reward $\bar{r}_t(a) := \frac{1}{T}\sum_{t\in[T]} r_t(a)$ among all arms a. Intuitively, puzzle difficulty decreases with ϵ (as it gets harder to distinguish the top two arms). Empirical gap mirrors the "gap" between the top two expected rewards, a standard notion of difficulty in MAB.

We measure an LLM's performance over a given set S of tasks as the fraction of tasks for which the LLM returns a "correct answer": an arm with the largest empirical reward; denoted by $\text{FracCorrect}(S)$. We are interested in how FracCorrect varies depending on the difficulty level. Hence, let $S(\epsilon, T)$ be the set of all tasks with empirical gap at most ϵ and history size T. We plot $\text{FracCorrect}(\epsilon, T) := \text{FracCorrect}\,(\,S(\epsilon, T)\,)$ against ϵ.

We find that while GPT-4 and GPT-3.5 do not perform well, GPT-5 models do, see Figure 1. For GPT-4 and GPT-3.5, performance tends to degrade as (1) history size T increases, with more data points to process in-context and (2) empirical gap decreases, making the bandit instance more difficult.

CB exploit puzzles. While the history in K-armed bandits can be summarized with $2K$ numbers (for each arm, the average reward and #plays), such succinct summary statistics may not be readily available (or even exist) in more complicated decision-making tasks such as CB.[2]

We begin with linear CB, where the expected reward of each arm a is linear in the context $z \in \mathbb{R}^d$: $\mu(z, a) = \langle z, \theta^*_a \rangle$ for some fixed (but unknown) vector $\theta^*_a \in \mathbb{R}^d$. [3] (We turn to non-linear CB later in the section.)

We consider a CB exploit puzzle parameterized by #arms K, dimension d, and history size T. The tasks are constructed

[2]This consideration also motivates MAB exploit puzzles with raw (non-summarized) history, as a simpler special case of the general scenario when succinct summarization is unavailable.

[3]Linear CB are well-studied, starting from Li et al. (2010); Chu et al. (2011); Abbasi-Yadkori et al. (2011).

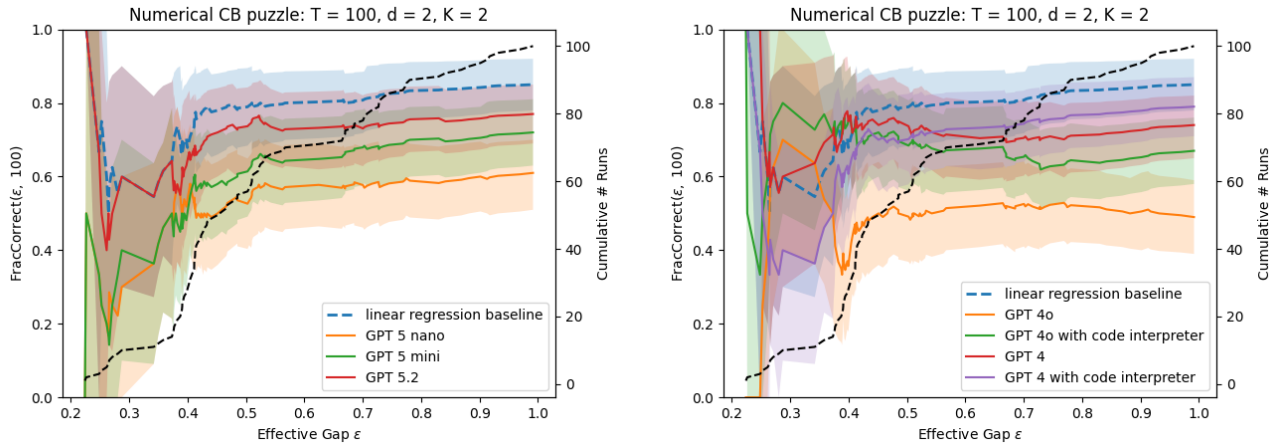


Figure 2. GPT-5 (left) and GPT-4 (right) model families on a numerical CB puzzle (all with temperature 1). GPT-5 models generally perform better than GPT-4 models of similar size, although letting GPT-4 and GPT-4O use a code interpreter improves performance.

as follows. We sample parameters $\theta_a \in [-1,1]^d$ and $\gamma_a \in [-\frac{1}{4}, \frac{1}{4}]$ independently u.a.r. for each arm a. Given context $z \in \mathbb{R}^d$, expected reward for arm a is $\mu(z,a) = \langle z, \theta_a \rangle + \gamma_a$. We generate a history of T rounds $t \in [T]$. Contexts z_t are sampled independently u.a.r. from $[-1,1]^d$. For simplicity, the history contains rewards of all arms a in all rounds t, where the reward equals $\mu(z_t, a)$ plus an independent unit-variance Gaussian. Given the history and a new context z_{T+1} (drawn in the same way), the LLM is asked to select the action for round $T+1$ which appears best. This gives one exploit task. We generate N tasks for the same K, T, d.

We use a modified "buttons" prompt, where contexts correspond to "numbers on a screen" which affect the payoffs for each button. The prompt does not mention linearity of the CB instance (because such model-based information is typically unavailable in applications).

Given an exploit task, we now define a "correct answer" as arm a which maximizes expected reward $\mu(z_{T+1}, a)$.[4] $\mathrm{FracCorrect}(S)$ is the fraction of correct answers in a given set S of tasks. Task difficulty is also not easily defined in terms of realized rewards. Instead, we focus on *effective gap* ϵ: the difference in expected reward $\mu(a,z)$ between the best and second-best arm given the current context $z = z_{T+1}$. Intuitively, smaller ϵ corresponds to increased difficulty.

We study how $\mathrm{FracCorrect}$ varies with effective gap. In each plot, we let $S(\epsilon, T)$ be the set of all tasks with given K, d, T and effective gap at most ϵ. We plot $\mathrm{FracCorrect}(\epsilon, T) := \mathrm{FracCorrect}\,(\,S(\epsilon, T)\,)$ against ϵ.

Figure 2 considers a setting with $d = K = 2$ for the GPT-5 family (left) and the GPT-4 family (right). We find that larger models generally perform better than their smaller, distilled counterparts and, conditioned on model size, GPT-5 models outperform GPT-4 models. We also plot a linear regression baseline (blue dashed line), which effectively serves as an upper-bound on the performance of any model, due to the underlying linear structure of our CB puzzle.

While not pictured, all models tend to perform strictly better on smaller problem instances. For example, GPT-4 performed near-perfectly when $d = 1$, $K = 2$, and $T = 50$. However, performance degrades as the problem size grows (e.g., see GPT-4O in Figure 4). Moreover, limited prompt size may prevent processing larger histories.[5]

Motivated by these observations, we implement two types of *mitigations*: tool use (providing the model with access to a Python interpreter to execute code in) and in-context summarization (inspired by the literature on exemplar selection for in-context learning, see Related Work). Evaluating these mitigations, we focus on the GPT-4 family as (1) at the time of writing GPT-5 models are not configured for tool use and (2) we found the GPT-5 models to be either prohibitively expensive (GPT-5.2) or prohibitively slow (GPT-5-MINI and GPT-5-NANO) on larger-scale experiments.

See Figure 2 (right) for the performance of GPT-4 and GPT-4O with access to a code interpreter (CI). Both see significantly improved performance with CI, with GPT-4 even surpassing GPT-5.2. However, enabling the CI made GPT-4 and GPT-4O much slower and much more expensive.[6]

We consider the following summarization mitigations:

[4] Note that it is unclear how to define an "empirically best arm" given a CB history and the current context.

[5] E.g., our LLM access points bottomed out at $T \approx$ 100-200 for GPT-4 and $T \approx$ 1000-2000 for GPT-4O.

[6] E.g., our GPT-4 experiment with CI in Figure 2 cost approximately \$70 and took 6+ hours to run. Our GPT-4 experiment in Figure 2 without CI cost a few dollars and took under 5 minutes.

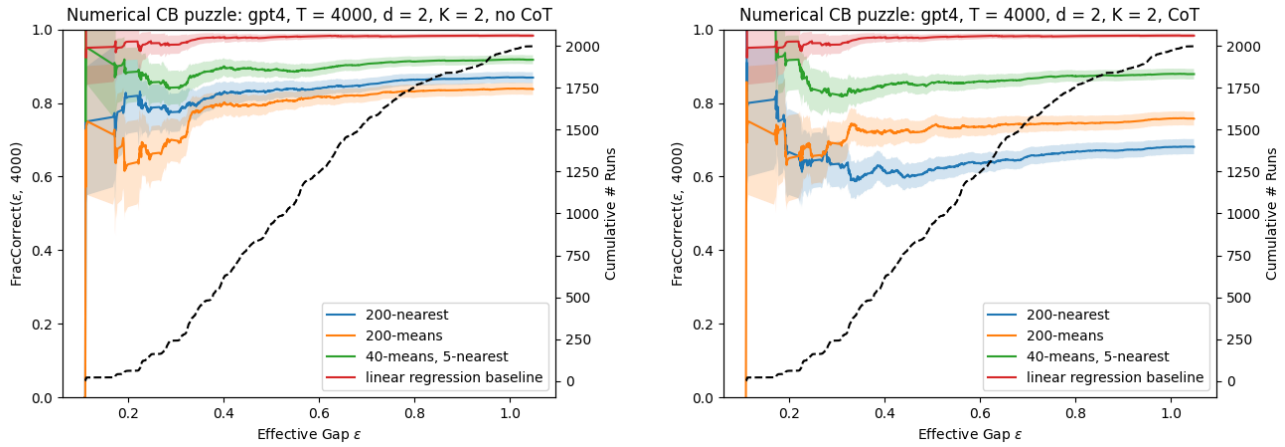


Figure 3. CB exploit puzzle with $d = K = 2$ and $T = 4000$: **mitigations help substantially**. GPT-4 without CoT (left) and GPT-4 with CoT (right). (Full history with this T vastly exceeds the context window for GPT-4, GPT-4O, and GPT-3.5.)

1. *k-nearest:* Among the observed contexts, pick distinct k contexts closest to z_{T+1}, according to the ℓ_2 metric. Limit the reported history to (the rounds with) these k contexts.

2. *k-means:* Run an off-the-shelf algorithm for k-means clustering on contexts $\{ z_1, \ldots, z_T \}$, obtaining k centroid contexts z_i^* and their respective clusters Z_i^*, $i \in [k]$. For each centroid z_i^* and each arm a, let $\bar{r}(z_i^*, a)$ be the average reward for this arm over all rounds t with contexts $z_t \in Z_i^*$. Report $(z_i^*, a, \bar{r}(z_i^*, a))$ as a context-arm-reward triple.

3. *k-means, $k' < k$-nearest:* First, run the k-means mitigation. Report $(z^*, a, \bar{r}(z^*, a))$ as a context-arm-reward triple, for each arm a and each centeroid context z^* among the k' centroids closest to z_{T+1} (according to the ℓ_2 metric).

We do not explain the "mitigation strategy" in the prompt: we present the context-arm-reward tuples as if it were the entire history without mentioning clustering or averaging.

Figure 3 shows the performance of our mitigations on a slightly larger puzzle with $K = d = 2$, $T = 4000$. We use GPT-4 with and without chain-of-thought (CoT) prompting. As before, we compare against a linear regression baseline. Without CoT, all three mitigations achieve FracCorrect $\approx 80\% - 90\%$, although this dips to $\approx 60\% - 85\%$ with CoT. Performance aside, the mitigations offer a practical way to handle large histories; e.g., when $T = 4,000$, our prompt exceeds the context window of all models we had access to.

However, in-context mitigations can only help so much. In Figure 4, we plot their performance under various hyper-parameters, on GPT-4O. While k-means (left) is almost as bad as random guessing, k-nearest (center) achieves $\approx 50\%$ FracCorrect, and the k-means, k'-nearest mitigation (right) obtains $\approx 60\%$. The latter two out-perform unmitigated GPT-4O, but fall significantly short of the linear baseline.

CB exploit puzzles (text-based & non-linear). As a robustness check, we repeat our CB experiments on a text-based exploit puzzle. In this puzzle, contexts are items in a room (e.g. animals, objects on a table), and actions have an associated semantic meaning (e.g. eat the food item, leave the room). Rewards are still presented numerically, and are non-linear functions of both the context and action. See Appendix B for full details on our experimental setup.

Figure 5 shows the performance of GPT-5.2 (left) and GPT-4O (right) with mitigations on this puzzle. While the reward function is non-linear (and thus the linear baseline only achieves 70% FracCorrect), we find that all configurations are still significantly out-performed by the linear baseline. Interestingly, GPT-4O with code interpreter access (orange) performs the worst out of all configurations.

Our intuition for these findings is as follows: While LLMs are generally good with textual inputs, the reward data is numerical, so the LLM faces same challenges as before (choosing a "solution strategy" and executing it correctly), exacerbated by non-linearity. Meanwhile, linear regression is known to often work fairly well even on non-linear data.

Takeaways. Frontier models can now solve fairly complicated MAB exploit puzzles. GPT-4 and GPT-3.5 struggle, partially explaining their failure on end-to-end MAB tasks in (Krishnamurthy et al., 2024; Nie et al., 2024).

In numerical CB puzzles, unmitigated performance drops as d, K, T grow, but tool use can help and increases in T can largely be taken care of by our in-context mitigations. In contrast, our mitigations were largely unhelpful in a non-linear, text-based CB puzzle. Taken together, our findings suggest that while current LLMs are effective at exploiting in small-scale, numerical decision-making tasks, they still struggle in more complicated settings.

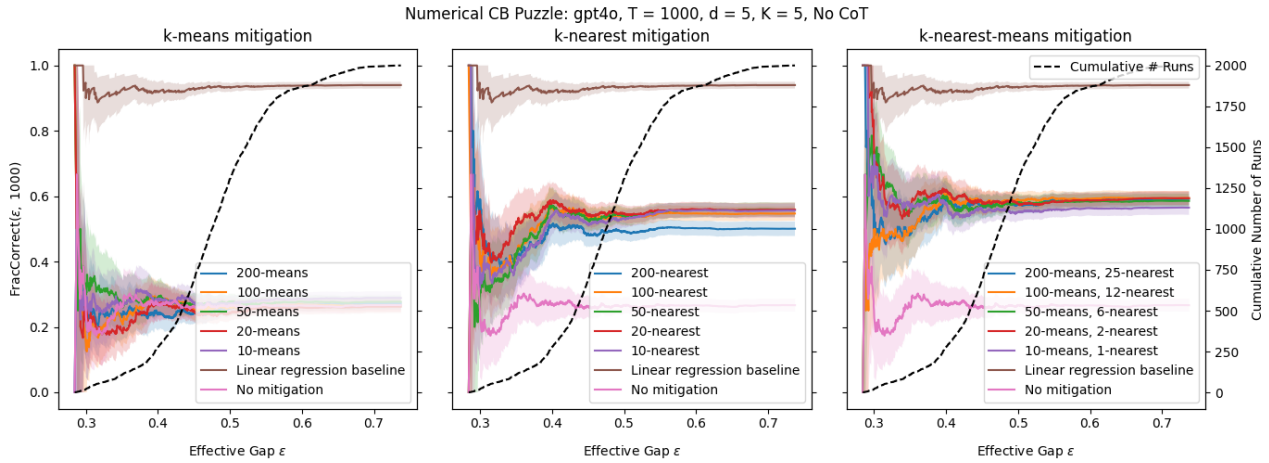


Figure 4. CB exploit puzzle with $d = K = 5$ and $T = 1000$: **in-context mitigations perform badly**, but (mostly) much better than the no-mitigation baseline. GPT-4O without CoT.

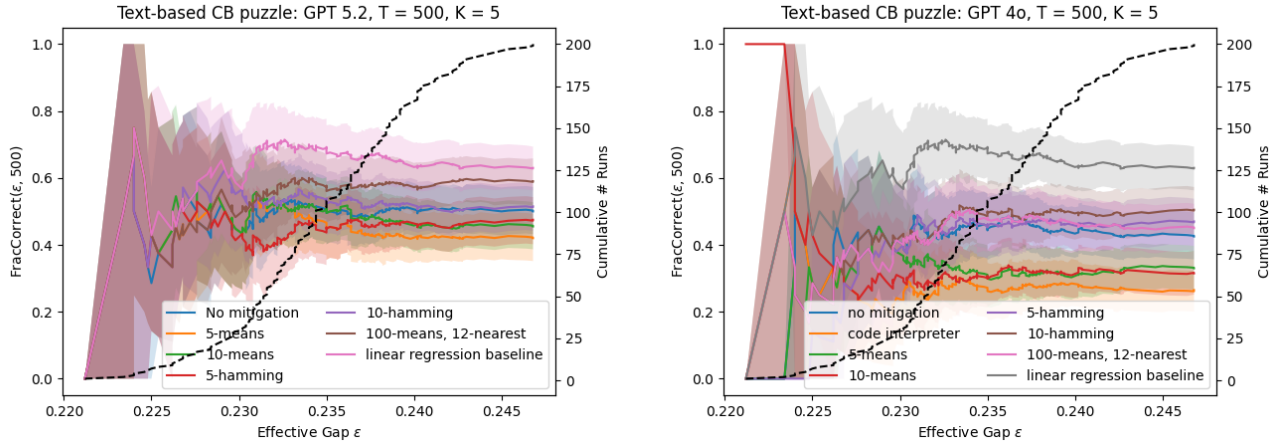


Figure 5. Performance of the GPT-5.2 (left) and GPT-4O on a text-based CB puzzle. In this more challenging non-linear setting, some mitigations help, but all are **outperformed by linear regression**.

3. LLMs as exploration oracles

We now study the ability of LLMs to explore large action spaces. We leverage the inductive bias of an LLM to generate a small set of candidate actions from a text-based action space, before running an off-the-shelf MAB algorithm on this set. We refer to this LLM usage as *exploration oracle*.

This usage of LLMs is closely aligned with "discretization" in dynamic pricing and Lipschitz bandits (e.g., Kleinberg & Leighton, 2003; Kleinberg et al., 2008; 2019; Bubeck et al., 2011; Slivkins, 2014). Given a very large action space, it may help to focus on a much smaller set of candidate arms. A naively constructed discretization, e.g., uniform w.r.t. some known distance notion on the arms, may have a poor tradeoff between size (#candidates) and quality (of the best candidate), placing all but a few candidates in low-performing regions of the action space. A crucial theme in this literature is "smarter" discretization algorithms which gradually "zoom in" on better-performing arms. If the latter comprise a low-dimensional region of the action space, covering this region should require much fewer samples for the same "discretization quality". Our hope is that LLMs can suggest candidate actions from this region, if there is sufficient semantic meaning encoded in the bandit task.

Non-LLM approaches appear hopeless for our tasks. Standard bandit algorithms (Thompson Sampling, UCB1, etc.) fail for a very large #arms, and so do bandit algorithms which randomly subsample the arms. Another approach embeds the arms' text labels in $\mathbb{R}^d$. If a suitable reward-distance relation were known to the algorithm/agent (assumption we do not make), one could apply the "smarter" discretization algorithms mentioned above. However, their regret scales exponentially in d, making them impractical for state-of-art text embeddings (e.g., one we use has $d = 384$).

Our tasks. We consider stylized exploration tasks of the following form. In each task, we define an MAB instance with Bernoulli rewards and a very large #arms, each labeled with a text string. The mean rewards $\mu(\cdot)$ and the meaning of the arms' labels are based on real-world data. We call an LLM once to generate a small representative subset of K arms. To grade the entire set (not just the best arm therein), we run a standard bandit algorithm (UCB1; (Auer et al., 2002)) over these K arms, for $T = 1000$ rounds. We record the algorithm's average cumulative reward, "denoised" as $\texttt{rew} := \frac{1}{T}\sum_{t\in[T]} \mu(a_t)$.

We consider three task types: (1) recommend a movie, based on the MovieLens 100K dataset (Harper & Konstan, 2015), a standard benchmark for evaluating CB (MovieLens task), (2) suggest a title for an arXiv research paper given its abstract (arXiv task), and (3) answer an open-ended "philosophical" question (Q/A task). Particular workloads within each task type are called *explore puzzles*. Our findings are similar across all task types and all LLMs.

The MovieLens 100K dataset contains ratings from 943 users across 1,682 movies. In our tasks, arms correspond to the movies titles. Each task corresponds to a particular user, with mean rewards given by this user's ratings. Since every user did not rate every movie, we used a standard algorithm to complete the data matrix (the soft-impute low-rank matrix completion method from the fancyimpute Python package). We then normalize this matrix so that all values lie $[0, 1]$.

The arXiv task (resp. Q/A task) is constructed as follows. Each task corresponds to an abstract-title pair from arXiv (resp., a question-answer pair). The arms are all possible titles (resp., answers). The "best arm" a^* is the actual title (resp. a contrarian answer generated by another LLM). The expected reward $\mu(a)$ of an arm a is the cosine similarity between a and a^* in the embedding generated by the Sentence-BERT embedding model (Reimers, 2019). [7] [8]

As a robustness check, we try several prompting strategies (detailed in Appendix C.1). We either ask the LLM to generate suggestions "all-at-once" with temperature 0 or "1-by-1" with temperature 1 (we repeatedly show the LLM the candidate arms so far and ask it to generate one more). We also try to explicitly encourage the LLM to provide a diverse candidate set. Thus, we have 4 prompting strategies: $\{$ all-at-once, 1-by-1 $\} \times \{$ with, without $\}$ encouragement. However, all 4 strategies performed similarly.[9]

Explore puzzle: MovieLens. We run a MovieLens task for each of the 943 users, with $K \in \{10, 19\}$. We re-run each task 5 times to denoise. For a given K, we record $\overline{\texttt{rew}}(K)$: reward `rew`, defined above, averaged over all users and runs. For baselines, we compare against running UCB1 for T rounds over (1) a random subset of K movies and (2) a random movie from each of the 19 movie genres in the dataset. (Different random draws for each user.)

Table 1. Average Reward for QWEN-2.5, MISTRAL-7B, and Baselines for $K \in \{10, 19\}$ in the MovieLens Task.

$\overline{\texttt{rew}}$	$K = 10$	$K = 19$
QWEN-2.5	671.88	**678.25**
MISTRAL-7B	**735.38**	675.16
Random	570.39	599.73
Genre-based	—	617.27

Table 2. Performance using all-at-once prompting on select arXiv tasks. Results are averaged over 10 abstract-title pairs. The full table (Table 3) may be found in the Appendix.

	$K = 1$	$K = 2$	$K = 5$
gr-qc	0.63	0.64	**0.68**
hep-ex	0.81	0.81	**0.83**
hep-th	0.65	0.71	**0.73**
math-ph	0.64	0.73	**0.74**
nucl-ex	0.73	**0.79**	0.75

Our results for MISTRAL-7B and QWEN-2.5 are in Table 1 Both LLMs significantly outperform all baselines for both values of K. We also tried running the experiment using LLAMA-3-8B and GEMMA-3, but the list of all movies was too long to fit in the context window.

The LLM has implicit knowledge about each movie which it may leverage when selecting the discretization. To mitigate overfitting to "popular" movies, we prompted the LLM to not take movie popularity into account. However, this change did not significantly affect the rewards.

Explore puzzle: arXiv. Using the arXiv API (arXiv.org, 2025), we collected 10 abstract-title pairs from each of the 41 arXiv categories.[10] Each abstract-title pair yields a task, with $K = 1, 2, 5$ suggestions. Each task is re-run 5 times to denoise. Given K, we record $\overline{\texttt{rew}}(\text{category}, K)$: reward `rew`, as defined above, averaged over all tasks within the same arXiv category and all runs.

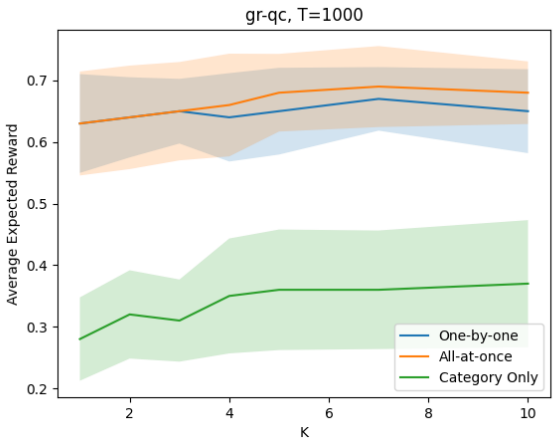
We evaluate GPT-4O as an exploration oracle for these tasks. We visualize our findings in Figure 6 , focusing on six arXiv categories. To assess the LLM's ability to specialize to a task, we consider a baseline where the candidate arms

[7] While cosine similarity ranges on $[-1, 1]$, it was usually strictly positive in our experiments. In the (very rare) cases where it was negative we defined the expected reward as zero.

[8] Our results are largely unchanged under the Universal Sentence Encoder (Cer, 2018), see Section C.4.

[9] We only tried the 4 strategies for the arXiv and Q/A tasks.

[10] To minimize the likelihood that these papers appear in the training corpus, we only use papers uploaded after June'24 (these experiments were done in Sept'24).



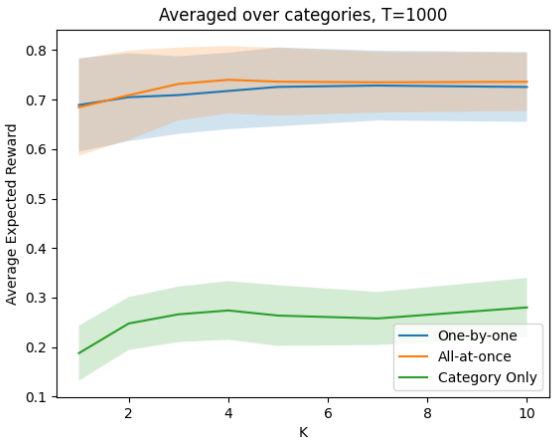


Figure 6. Average expected reward $\overline{\texttt{rew}}(\text{category}, K)$ (averaged over rounds and tasks), against K, the number of candidates. Each line corresponds to a prompting strategy or the Category-Only baseline. The shaded regions represent a 95% confidence interval. A single arXiv category ("General Relativity and Quantum Cosmology", left), averages over 6 categories (right).

are generated by GPT-4O given only the category, *not the abstract*. For the other arXiv categories, we provide a table with $\overline{\texttt{rew}}(\text{category}, K)$ for $K = 1, 2, 5$.

Explore puzzle: Q/A task. We used GPT-4 to generate a dataset of 10 open-ended questions with many reasonable answers, along with an intentionally contrarian answer for each question to serve as the ground truth. (E.g., "What does it mean to live a fulfilling life?" "Fulfillment comes from embracing discomfort.") Each question-answer pair yields a task, with $K \in \{1, 2, 3, 4, 5, 7, 10\}$ candidate answers. For a given K, we run each task 10 times and record $\overline{\texttt{rew}}(\text{task}, K)$: reward `rew` averaged over all runs.

We evaluate GPT-4O, QWEN-2.5, GEMMA-3, and MISTRAL-7B as exploration oracles. We also consider a non-LLM baseline (`Random`), which selects K arms independently and u.a.r. in the embedding space (leveraging the AI-based embedding but suffering from its high dimension).

The LLMs perform similarly, with $\overline{\texttt{rew}}$ ≈0.5-0.6. and typically work best for K=3,4. see Section A. In contrast, `Random` catastrophically fails, with $\overline{\texttt{rew}} < 0.1$. Moreover, we observe that LLM-generated answers pass the "eye test" (they are reasonable, yet spiritually and semantically different), and yield substantially different expected rewards. We conclude that the LLM does succeed as an exploration oracle for these tasks.

4. Conclusions

Our work adds to (and provides partial explanations for) the growing literature on in-context learning for decision-making tasks. We find that LLMs are useful as *exploration oracles* that propose high-quality candidate actions in large, semantically meaningful action spaces. However, current LLMs are not that good at *exploitation*, particularly in larger or more complex tasks, although reasoning models (specifically, the GPT-5 family of models) show the most promise. While we experiment with several mitigations including tool use and in-context summaries, they consistently underperform relative to a simple linear regression baseline, even in inherently non-linear tasks.

We provide a systematic evaluation of LLMs' capabilities as exploration oracles, in isolation from other components of a decision-making agent. We find robust success across several different task types, LLMs, and prompting strategies. In contrast, non-LLM baselines are often quite weak, absent known problem structure to be leveraged by an algorithm.

LLMs' continued difficulties at exploitation tasks should be contrasted with the rapid progress in LLMs and the extreme optimism associated with this progress. Ever since the GPT-3.5 release in 2022 (and especially since GPT-4 frenzy of early 2023), it was commonly believed that "generalist" frontier LLMs will soon master many/most relatively simple tasks: if not yet, then more likely in several months than in several years. However, our results show that these difficulties persist longer than anticipated.

Future work. LLM-based exploration oracles may potentially help with "smart discretization" for rich text-based action spaces. Non-LLM bandit algorithms tend to fail in such scenarios, as discussed in Section 3. The hope is to "zoom in" entirely in the space of "potentially relevant" actions, rather than in the space of *all* actions.

Second, while we experiment with tool use for exploitation tasks, we find that it does not improve performance across the board. One could investigate various approaches for tool use in these tasks, and whether/when they help.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *25th Advances in Neural Information Processing Systems (NIPS)*, pp. 2312–2320, 2011.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms, 2020. Book draft, circulated since 2019. Available at https://rltheorybook.github.io.

arXiv.org. arxiv api, 2025. URL https://arxiv.org/help/api. Accessed: 2025-01-28.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Beygelzimer, A. and Langford, J. The offset tree for learning with partial labels. In *15th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2009.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvari, C. Online Optimization in X-Armed Bandits. *J. of Machine Learning Research (JMLR)*, 12:1587–1627, 2011. Preliminary version in *NIPS 2008*.

Carta, T., Romac, C., Wolf, T., Lamprier, S., Sigaud, O., and Oudeyer, P.-Y. Grounding large language models in interactive environments with online reinforcement learning. In *International Conference on Machine Learning*, pp. 3676–3713. PMLR, 2023.

Cer, D. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual Bandits with Linear Payoff Functions. In *14th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Coda-Forno, J., Binz, M., Akata, Z., Botvinick, M., Wang, J., and Schulz, E. Meta-in-context learning in large language models. *Advances in Neural Information Processing Systems*, 36:65189–65201, 2023.

Coda-Forno, J., Binz, M., Wang, J. X., and Schulz, E. Cogbench: a large language model walks into a psychology lab. *arXiv:2402.18225*, 2024.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):1097–1104, 2014.

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Hayes, W. M., Yax, N., and Palminteri, S. Relative value biases in large language models. *arXiv:2401.14530*, 2024.

Khalifa, M., Logeswaran, L., Lee, M., Lee, H., and Wang, L. Exploring demonstration ensembling for in-context learning. *arXiv preprint arXiv:2308.08780*, 2023.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *40th ACM Symp. on Theory of Computing (STOC)*, pp. 681–690, 2008.

Kleinberg, R., Slivkins, A., and Upfal, E. Bandits and experts in metric spaces. *J. of the ACM*, 66(4):30:1–30:77, May 2019. Merged and revised version of conference papers in *ACM STOC 2008* and *ACM-SIAM SODA 2010*. Also available at http://arxiv.org/abs/1312.1277.

Kleinberg, R. D. and Leighton, F. T. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, pp. 594–605, 2003.

Krishnamurthy, A., Harris, K., Foster, D. J., Zhang, C., and Slivkins, A. Can large language models explore in-context? In *NeurIPS*, 2024.

Laskin, M., Wang, L., Oh, J., Parisotto, E., Spencer, S., Steigerwald, R., Strouse, D., Hansen, S., Filos, A., Brooks, E., et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lee, J., Xie, A., Pacchiano, A., Chandak, Y., Finn, C., Nachum, O., and Brunskill, E. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Lehnert, L., Sukhbaatar, S., Mcvay, P., Rabbat, M., and Tian, Y. Beyond a*: Better planning with transformers via search dynamics bootstrapping. *arXiv preprint arXiv:2402.14083*, 2024.

Li, C., Yang, R., Li, T., Bafarassat, M., Sharifi, K., Bergemann, D., and Yang, Z. Stride: A tool-assisted llm agent framework for strategic and interactive decision-making. *arXiv preprint arXiv:2405.16376*, 2024.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *19th Intl. World Wide Web Conf. (WWW)*, 2010.

Lin, L., Bai, Y., and Mei, S. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*, 2023.

Monea, G., Bosselut, A., Brantley, K., and Artzi, Y. Llms are in-context reinforcement learners. *COLM*, 2025.

Mukherjee, S., Hanna, J. P., Xie, Q., and Nowak, R. Pretraining decision transformers with reward prediction for in-context multi-task structured bandit learning. *arXiv preprint arXiv:2406.05064*, 2024.

Nie, A., Su, Y., Chang, B., Lee, J. N., Chi, E. H., Le, Q. V., and Chen, M. Evolve: Evaluating and optimizing llms for exploration. *arXiv preprint arXiv:2410.06238*, 2024.

OpenAI. Introducing operator, 2025. URL https://openai.com/index/introducing-operator/. Accessed: 2025-01-29.

Park, C., Liu, X., Ozdaglar, A., and Zhang, K. Do llm agents have regret? a case study in online learning and games. *arXiv preprint arXiv:2403.16843*, 2024.

Raparthy, S. C., Hambro, E., Kirk, R., Henaff, M., and Raileanu, R. Generalization to new sequential decision making tasks with in-context learning. *arXiv preprint arXiv:2312.03801*, 2023.

Reimers, N. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Schubert, J. A., Jagadish, A. K., Binz, M., and Schulz, E. In-context learning agents are asymmetric belief updaters. *arXiv:2402.03969*, 2024.

Slivkins, A. Contextual bandits with similarity information. *J. of Machine Learning Research (JMLR)*, 15(1):2533–2568, 2014. Preliminary version in *COLT 2011*.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, November 2019. Published with *Now Publishers* (Boston, MA, USA). Also available at https://arxiv.org/abs/1904.07272.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Tonglet, J., Reusens, M., Borchert, P., and Baesens, B. Seer: A knapsack approach to exemplar selection for in-context hybridqa. *arXiv preprint arXiv:2310.06675*, 2023.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Wu, Y., Tang, X., Mitchell, T., and Li, Y. Smartplay: A benchmark for LLMs as intelligent agents. In *International Conference on Learning Representations*, 2024.

Xia, F., Liu, H., Yue, Y., and Li, T. Beyond numeric awards: In-context dueling bandits with llm agents. *arXiv preprint arXiv:2407.01887*, 2024.

Xiong, J., Li, Z., Zheng, C., Guo, Z., Yin, Y., Xie, E., Yang, Z., Cao, Q., Wang, H., Han, X., et al. Dq-lore: Dual queries with low rank approximation re-ranking for in-context learning. *arXiv preprint arXiv:2310.02954*, 2023.

Xu, M., Shen, Y., Zhang, S., Lu, Y., Zhao, D., Tenenbaum, J., and Gan, C. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*, pp. 24631–24645. PMLR, 2022.

Yan, X., Song, Y., Feng, X., Yang, M., Zhang, H., Ammar, H. B., and Wang, J. Efficient reinforcement learning with large language model priors. *arXiv preprint arXiv:2410.07927*, 2024.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhang, Z., Zhang, A., Li, M., and Smola, A. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

Zhao, Z., Lee, W. S., and Hsu, D. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., and Wang, Y.-X. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*, 2023.

Should You Use Your Large Language Model to Explore or Exploit? (Supplementary Material)

May 8, 2026

A. Main Plots for the Q/A explore task

Here we provide the key plots for the Q/A explore task, as defined in Section 3. As stated there, all LLMs we considered perform similarly, with $\overline{\texttt{rew}}$ ≈0.5-0.6. and typically work best for K=3,4, whereas the `Random` baseline catastrophically fails, with $\overline{\texttt{rew}} < 0.1$. Recall that $\overline{\texttt{rew}}(\text{task}, K)$ is the average reward `rew` for a given task with a given K (as defined in Section 3), averaged over all runs of this task.

We visualize our findings in Figure 7. We plot $\overline{\texttt{rew}}(\text{task}, K)$ against K, for one particular task (left) and averaged across all tasks (right). Similar plots for the 9 other tasks can be found in Section C.6.

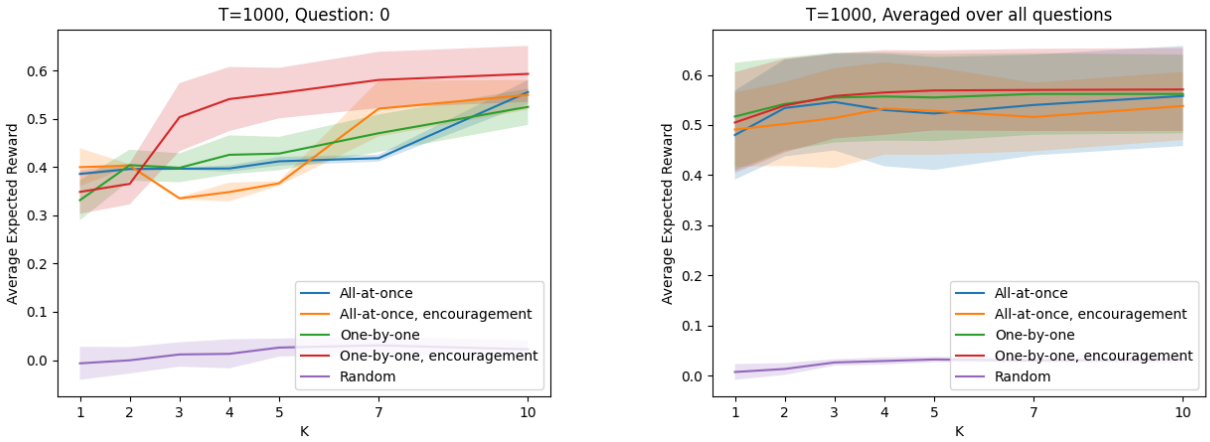


Figure 7. Average expected reward $\overline{\texttt{rew}}(\text{task}, K)$ (averaged over rounds and over runs), against K, #candidates. Each line corresponds to a prompting strategy or the `Random` baseline. The shaded regions represent a 95% confidence interval.

The LLM-generated answers pass the "eye test": they are reasonable, yet spiritually and semantically different. E.g., given the question "What is the role of technology in society", the first $K = 5$ suggestions generated by our 1-by-1 prompt are: (1) Facilitates communication, innovation, and efficiency, (2) Transforms daily life and shapes culture, (3) Drives connectivity and enhances productivity, (4) Facilitates control and surveillance, (5) Disrupts traditional relationships and norms. Finally, we verify that the candidate suggestions are substantially *different* from one another in terms of their rewards, see Figure 8.

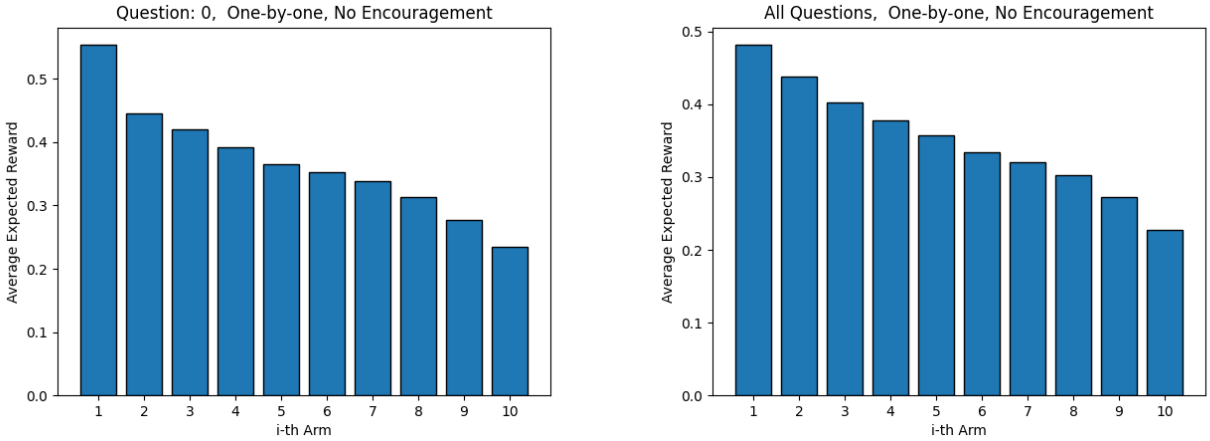


Figure 8. Spread of expected rewards within the candidate set. Arm histogram for one particular question ("What is the meaning of freedom?", left) and averaged over all questions (right). We consider $K = 10$ suggestions generated by our one-by-one prompt without encouragement. For a given "run", the K suggestions are ranked by the expected reward, and then the i-th expected reward is averaged over all runs, for each $i \in [K]$.

B. Details for Section 2: LLMs as exploitation oracles

B.1. Prompts

In this section we give example prompts for each of our experimental setups.

"Buttons" prompt for the MAB puzzle:

> [SYSTEM] You are in a room with 5 buttons labeled blue, green, red, yellow, purple. Each button is associated with a Bernoulli distribution with a fixed but unknown mean; the means for the buttons could be different. For each button, when you press it, you will get a reward that is sampled from the button's associated distribution. Then you must pick the button with the highest empirical average, which must be exactly one of blue, green, red, yellow, or purple. You must provide your final answer immediately within the tags <Answer>COLOR <Answer>where COLOR is one of blue, green, red, yellow, or purple and with no text explanation.
>
> [USER] The past rewards for each button are:
>
> round 1: blue button had reward 1, green button had reward 1, red button had reward 0, yellow button had reward 1, purple button had reward 0
>
> round 2: blue button had reward 0, green button had reward 1, red button had reward 1, yellow button had reward 1, purple button had reward 0
>
> Which button do you choose? Remember, YOU MUST provide your final answer within the tags <Answer>COLOR <Answer>where COLOR is one of blue, green, red, yellow, or purple and with no text explanation.

"Adverts" prompt for the MAB puzzle:

> [SYSTEM] You are recommendation engine that chooses advertisements to display to users when they visit your webpage. There are 5 advertisements you can choose from, named A, B, C, D, E. When a user visits the webpage you can choose an advertisement to display and you will observe whether the user would have clicked each of the ads. You model this by assuming that each advertisement has a certain click rate and users click on advertisements with their corresponding rates. I will show you the past clicks for each advertisement. Then you must pick the advertisement with the highest empirical click rate, which must be exactly one of A, B, C, D, or E. You must provide your final answer immediately and with no text explanation. within the tags <Answer>ADVERTISEMENT <Answer>where ADVERTISEMENT is one of A, B, C, D, or E.
>
> [USER] The past clicks for each advertisement are:
>
> round 1: advertisement A was clicked, advertisement B was clicked, advertisement C was not clicked, advertisement D was clicked, advertisement E was clicked
>
> round 1: advertisement A was not clicked, advertisement B was clicked, advertisement C was clicked, advertisement D was clicked, advertisement E was not clicked
>
> Which advertisement do you choose? Remember, YOU MUST provide your final answer within the tags <Answer>ADVERTISEMENT <Answer>where ADVERTISEMENT is one of A, B, C, D, or E and with no text explanation.

"Buttons" prompt for the numerical CB puzzle:

> [SYSTEM] You are in a room with a television and 2 buttons labeled blue, green. Each button is associated with a Bernoulli distribution with an unknown mean; the means for the buttons could be different from each other and may depend on the list of numbers shown on the screen (i.e. the context). For each button, when you press it, you will get a reward that is sampled from the button's associated distribution, conditioned on the numbers shown on the television screen. I will show you the past numbers shown on the screen and the corresponding rewards for each button. A new list of numbers will then appear on the screen and you must pick the next button in order to maximize your reward in this round only, which must be exactly one of blue or green. You must provide your final answer immediately within the tags ¡Answer¿ COLOR ¡/Answer¿ where COLOR is one of blue or green and with no text explanation.

[USER] The past contexts and rewards for each button are:

In round 1, the context was $[0.3, 0.7]$. The blue button had reward 1, the green button had reward 1

In round 2, the context was $[0.4, 0.6]$. The blue button had reward 0, the green button had reward 1

Which button do you choose? Remember, YOU MUST provide your final answer within the tags <Answer>COLOR <Answer>where COLOR is one of blue or green and with no text explanation.

Prompt for the text-based CB puzzle:

[SYSTEM] You are in a room with a table and a button. There may also be other objects in the room, which I will tell you about. You must then take one of the following actions: ”pet animal”, ”leave room”, ”use tool”, ”eat food”, ”press button”, after which you will receive some reward. The reward you receive is a random function of both the action you take and the information you receive about the objects in the room and time of day. Your goal is to maximize the expected reward you receive. I will show you the past history of play over 2 rounds. For each round, I will show you the state of the room and the corresponding rewards for each action. I will then tell you the current state of the room, and you must pick the next action in order to maximize your reward in this round only, which must be exactly one of ”pet animal”, ”leave room”, ”use tool”, ”eat food”, or ”press button”. Look for patterns in the data and try to estimate the reward of each action, given the information at your disposal. You must provide your final answer immediately within the tags <Answer>ACTION <Answer>where ACTION is one of ”pet animal”, ”leave room”, ”use tool”, ”eat food”, or ”press button” and with no text explanation.

[USER] The past observations and outcomes for each action are:

Round 1 had context time of day: morning, animal: bear, table item: chest, tool: key, food: apple, button color: red. ”pet animal” had reward 0, ”leave room” had reward 1, ”use tool” had reward 1, ”eat food” had reward 0, press button had reward 0

Round 2 had context time of day: afternoon, animal: cat, table item: card, tool: hammer, food: cake, button color: orange. ”pet animal” had reward 1, ”leave room” had reward 0, ”use tool” had reward 0, ”eat food” had reward 1, press button had reward 0

The current state of the room is time of day: evening, animal: bear, table item: envelope, tool: key, food: nut, button color: red.

Which action do you choose? Remember, you must provide your final answer immediately within the tags <Answer>ACTION <Answer>where ACTION is one of ”pet animal”, ”leave room”, ”use tool”, ”eat food”, or ”press button” and with no text explanation.

B.2. Additional Results on MAB exploit puzzle

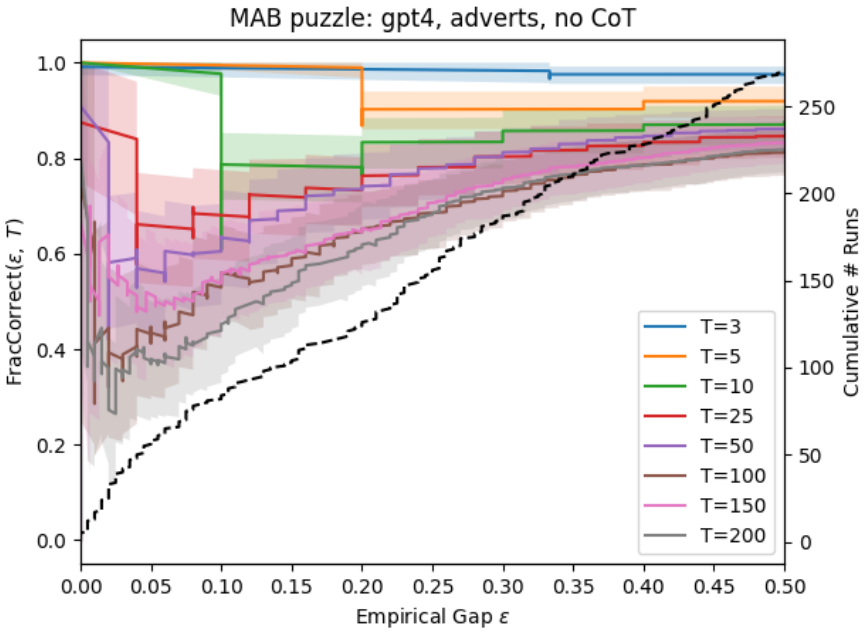


Figure 9. Cumulative fraction correct for GPT-4 in the MAB adverts puzzle.

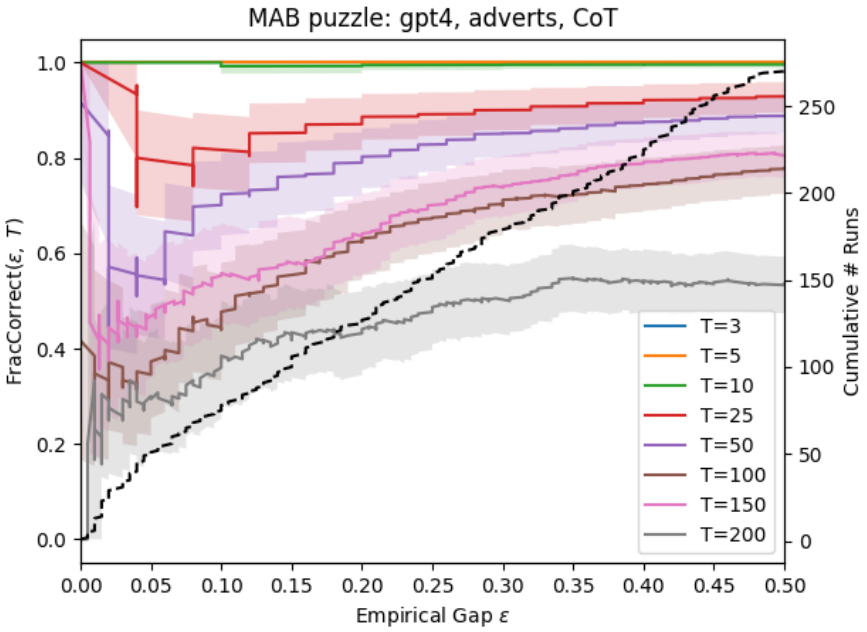


Figure 10. Cumulative fraction correct for GPT-4 with chain-of-thought reasoning in the MAB adverts puzzle.

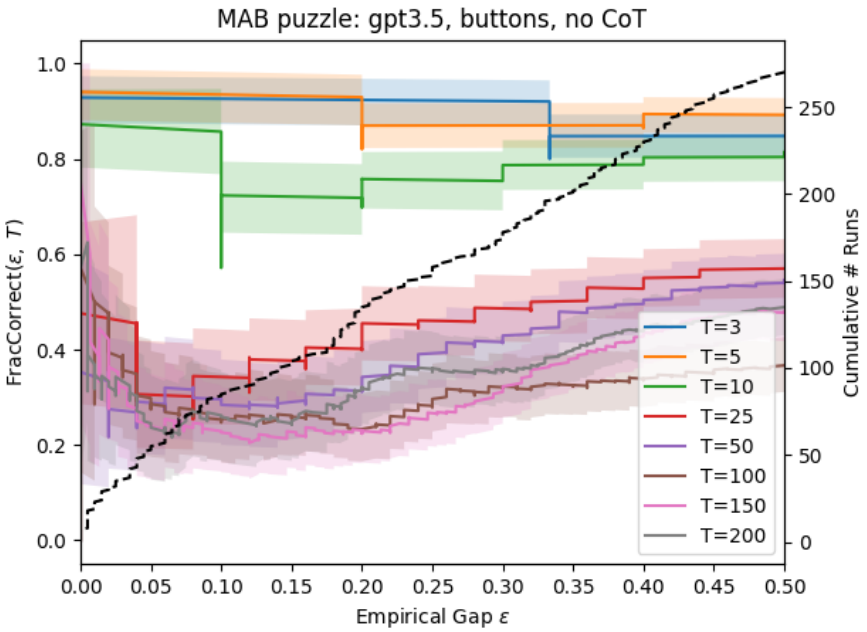


Figure 11. Cumulative fraction correct for GPT-3.5 in the MAB buttons puzzle.

B.3. Additional details for text-based CB puzzles

Each context contains a time of day (belonging to {morning, afternoon, evening, night}), an animal ({bear, dog, cat, None}), a tool ({key, letter opener, hammer, None}), a food item ({cake, apple, nut, None}), and a button with a particular color ({red, orange, yellow, green}). The actions are "pet animal", "leave room", "use tool", "eat food", and "press button".

We considered two reward functions. One was "easy", with expected rewards (ExpRev) as follows:

- ExpRev for petting the animal is 0.01 if the animal is a bear, 0.7 if the animal is a dog, and 0.4 if the animal is a cat. Otherwise, ExpRev if 0.5.
- ExpRev for leaving the room is always 0.5.
- ExpRev for using the tool is 0.75 if it is a key, 0.6 if it is a letter opener, 0.45 if it is a hammer, and 0.2 otherwise.
- ExpRev for eating food is 0.8 if it is cake, 0.6 if it is an apple, 0.2 if it is a nut, and 0.3 otherwise.
- ExpRev for pressing the button is 0.89 if it is green, 0.62 if it is yellow, 0.39 if it is orange, and 0.27 if it is red.

Our results under this reward function are summarized in Figure 12. We used hamming distance to implement our mitigations. Note that in higher-dimensional settings, distance in an embedding space may be used.

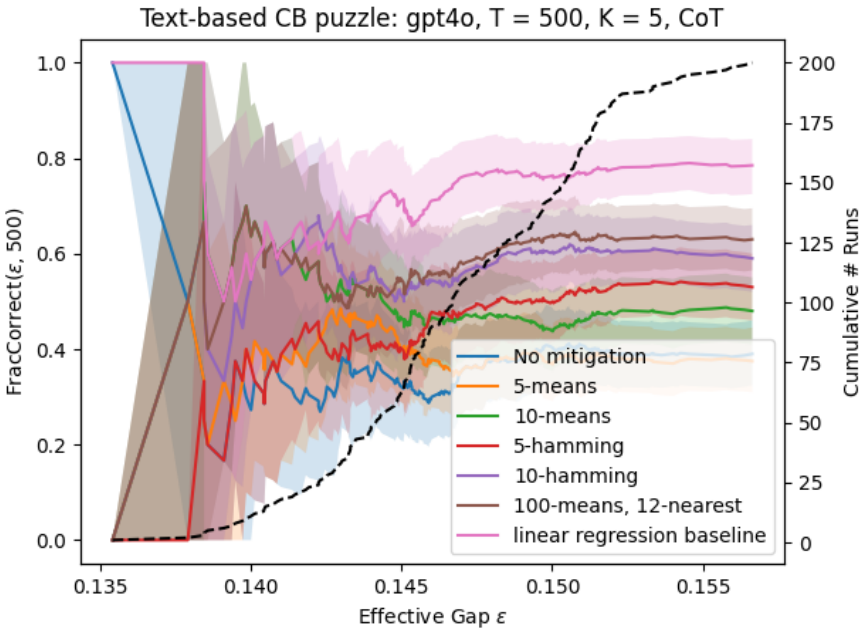


Figure 12. Performance of GPT-4O with mitigations on the words CB puzzle with "easy" rewards.

The reward function we use in the main body is more complicated, and is detailed below:

- ExpRev for petting the animal is 0.01 if it is a bear, 0.7 if it is a dog, 0.3 if it is a cat and the time of day is morning or afternoon, 0.7 if it is a cat and the time of day is evening or night, and otherwise 0.5.
- ExpRev for leaving the room is always 0.5
- If the animal is a bear, ExpRev for using the tool is 0.1. Otherwise, if the tool is a key and the table item is a chest, ExpRev is 0.9. Otherwise, it is 0.4
- If the animal is a bear, ExpRev for eating food is always 0.5. Otherwise, ExpRev is 0.8 for cake, 0.6 for an apple, 0.2 for a nut, and 0.5 otherwise.

- If the animal is a bear, ExpRev for pressing the button is 0.1. Otherwise if the button is green and the time of day is morning, or the button is yellow and the time of day is afternoon, or the button is orange and the time of day is evening, or the button is red and the time of day is night, then ExpRev is 0.9. In all other cases, ExpRev is 0.25.

B.4. Robustness Check with Other Models

In this appendix, we include results for Qwen2.5-7B-Instruct (QWEN-2.5), Gemma 3 12B (GEMMA-3), and Mistral-7B-Instruct-v0.3 (MISTRAL-7B) on our exploit puzzles. Due to the time and costs associated with running our exploit puzzles, we prioritized (1) getting at least one additional set of results for each of our main puzzles (using QWEN-2.5) and (2) getting results for all of our models in our word-based puzzles. Our findings are the same across all models.

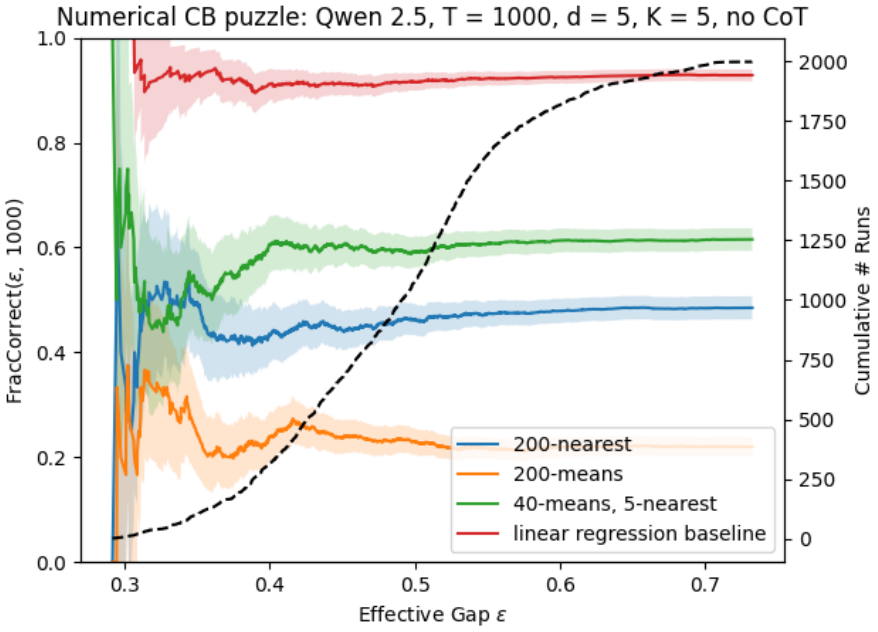


Figure 13. Results for QWEN-2.5 in the setting of Figure 4. Only mitigations are shown, as the unmitigated history is too long to fit in the context window.

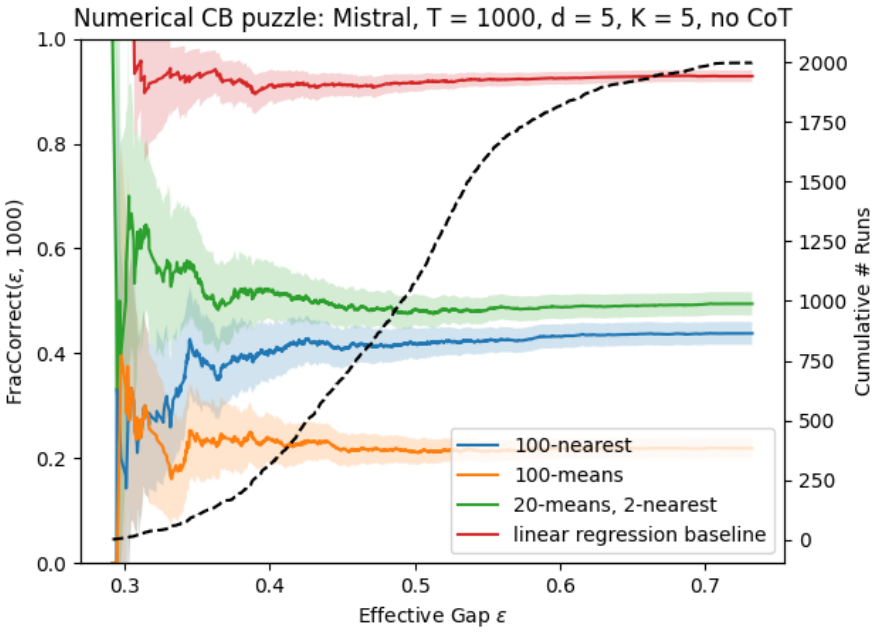


Figure 14. Results for MISTRAL-7B in the setting of Figure 4. Only mitigations are shown, as the unmitigated history is too long to fit in the context window.

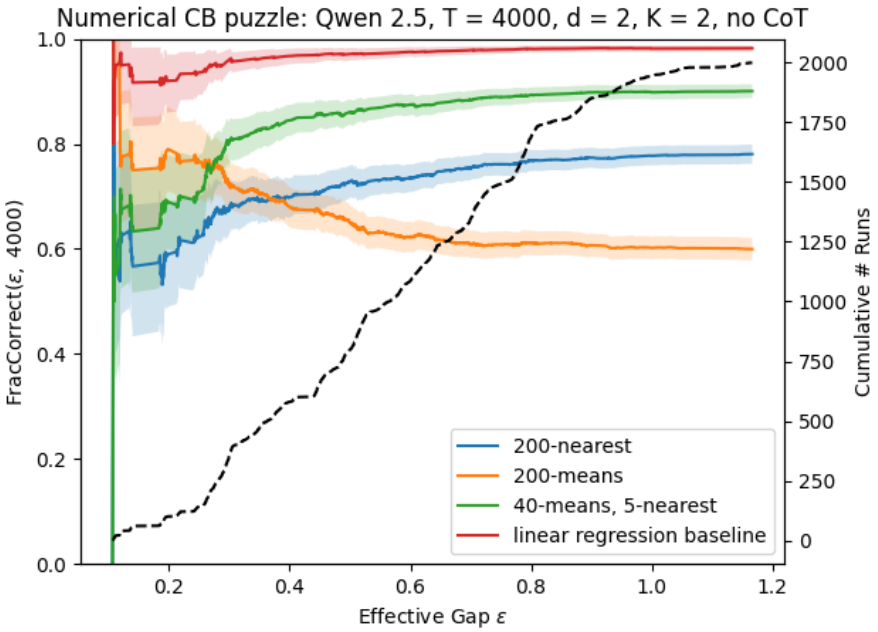


Figure 15. Results for QWEN-2.5 in the setting of Figure 3. Only mitigations are shown, as the unmitigated history is too long to fit in the context window.

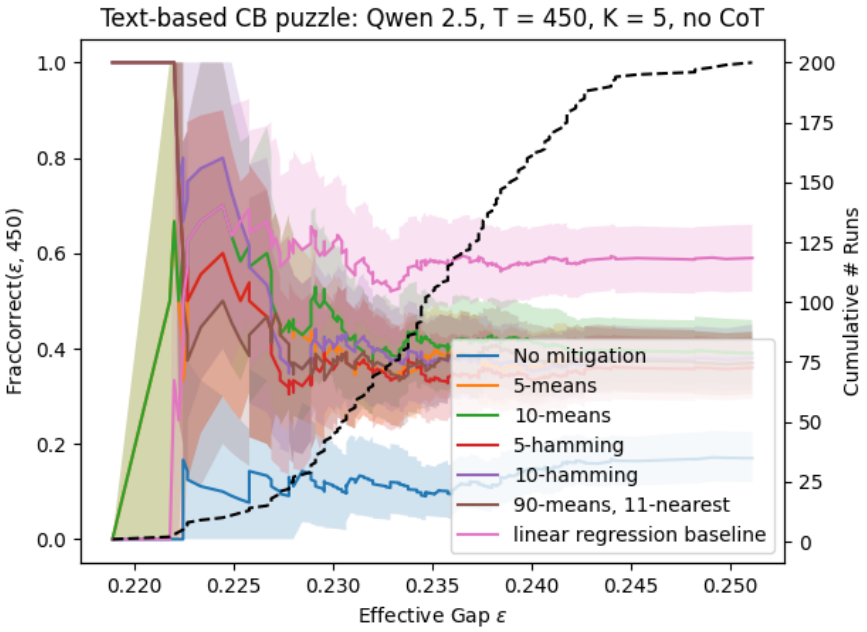


Figure 16. Results for QWEN-2.5 in the setting of Figure 5 (right). We use a slightly smaller history length ($T = 450$ instead of $T = 500$) so that the full history can fit in the context window.

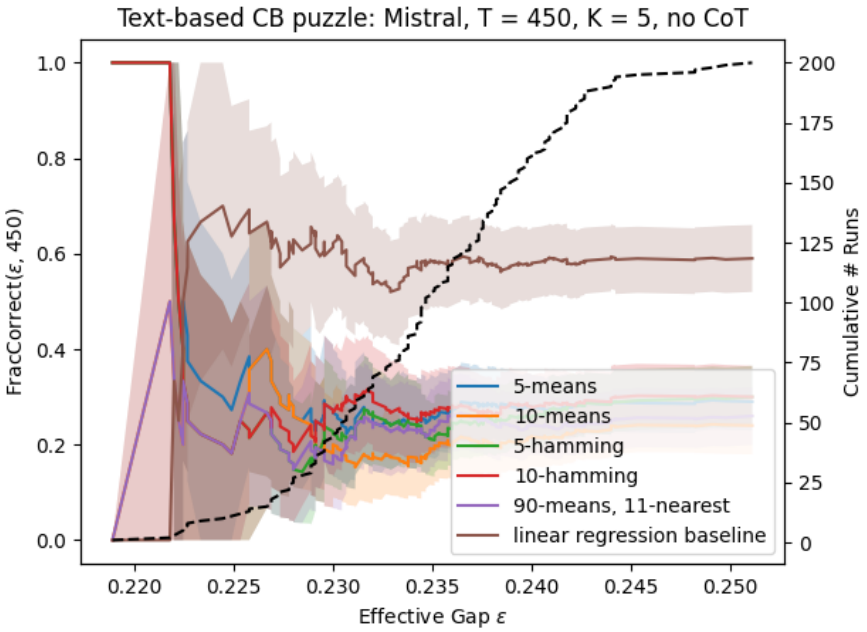


Figure 17. Results for MISTRAL-7B in the setting of Figure 16. Only mitigations are shown, as the full history could not fit in the context window.

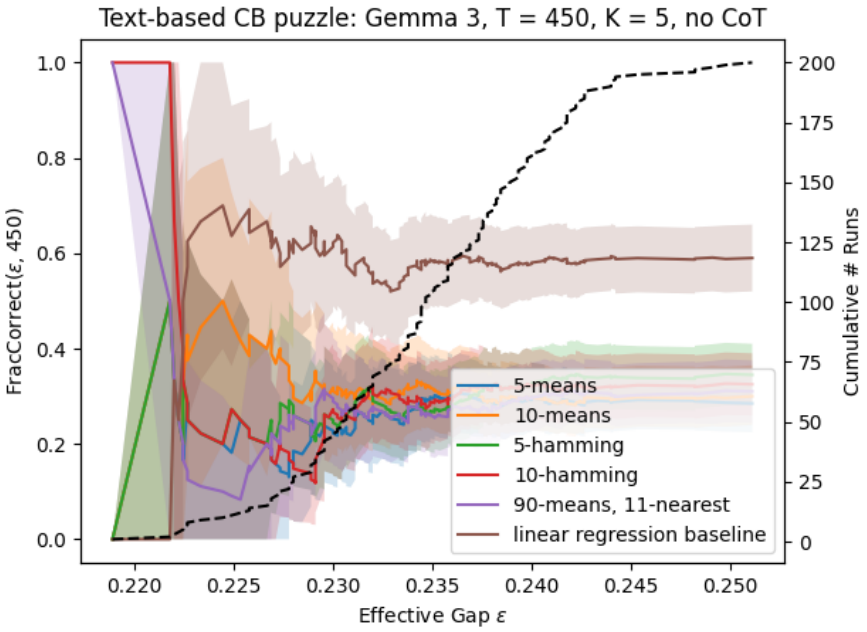


Figure 18. Results for GEMMA-3 in the setting of Figure 16. Only mitigations are shown, as the full history could not fit in the context window.

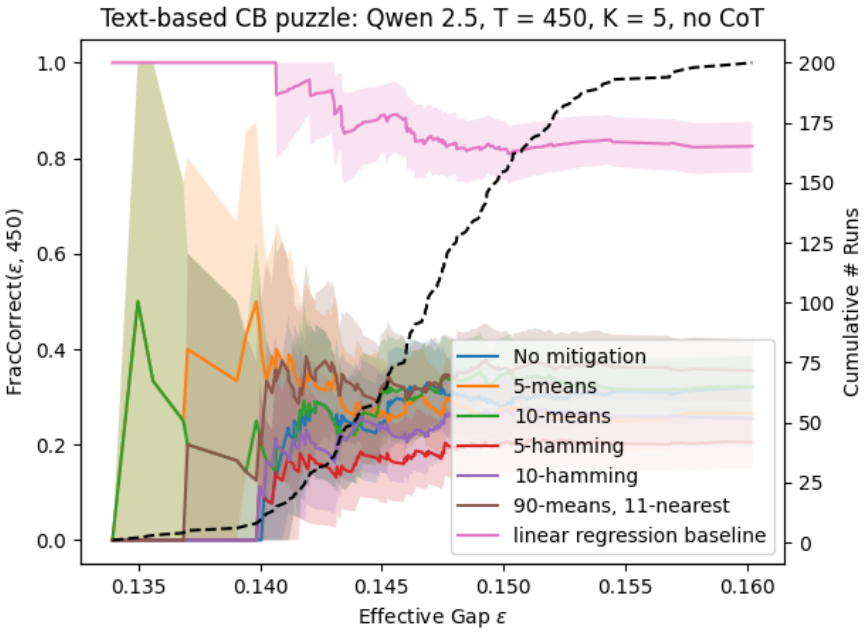


Figure 19. Results for QWEN-2.5 in the setting of Figure 12. We use a slightly smaller history length ($T = 450$ instead of $T = 500$) so that the full history can fit in the context window.

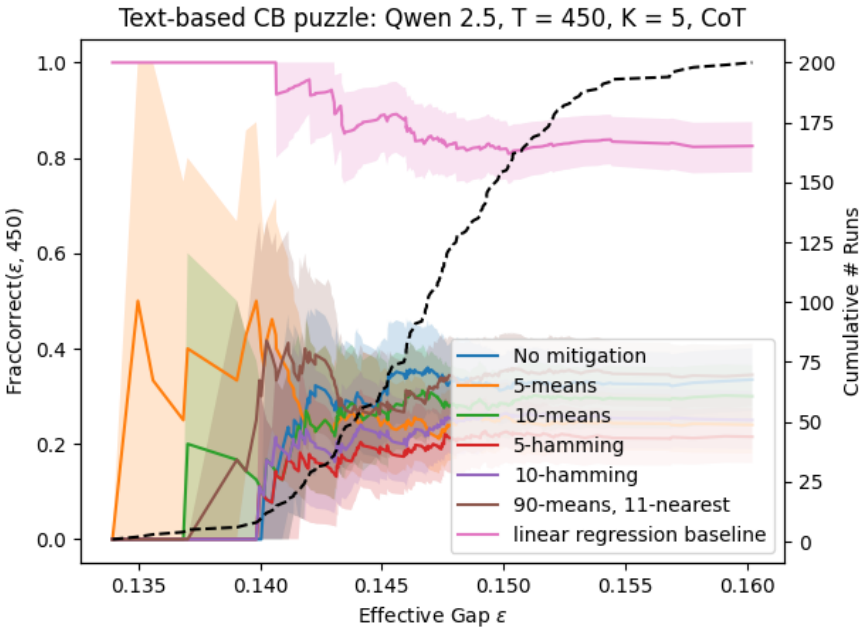


Figure 20. Results for QWEN-2.5 in the setting of Figure 12, using chain-of-thought prompting. We use a slightly smaller history length ($T = 450$ instead of $T = 500$) so that the full history can fit in the context window.

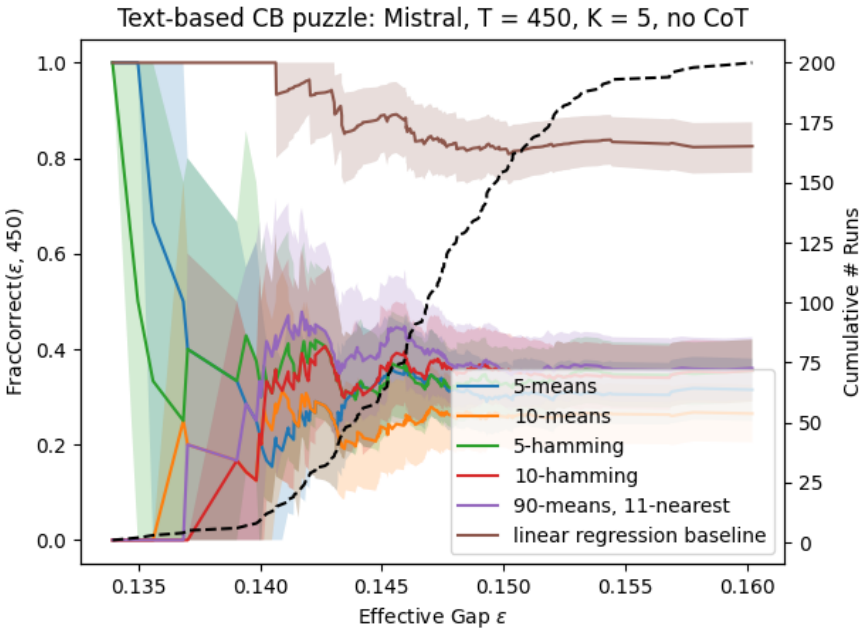


Figure 21. Results for MISTRAL-7B in the setting of Figure 19. Only mitigations are shown, as the full history could not fit in the context window.

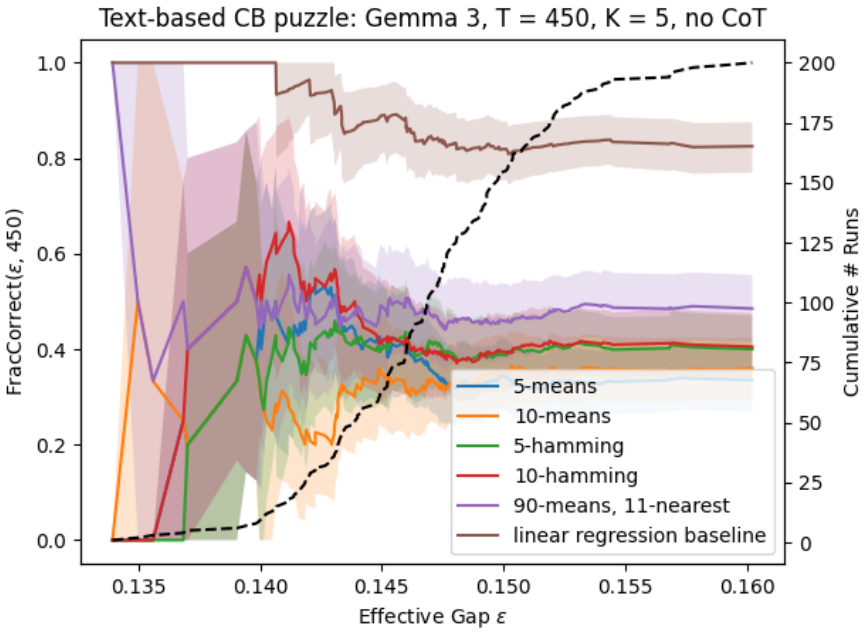


Figure 22. Results for GEMMA-3 in the setting of Figure 19. Only mitigations are shown, as the full history could not fit in the context window.

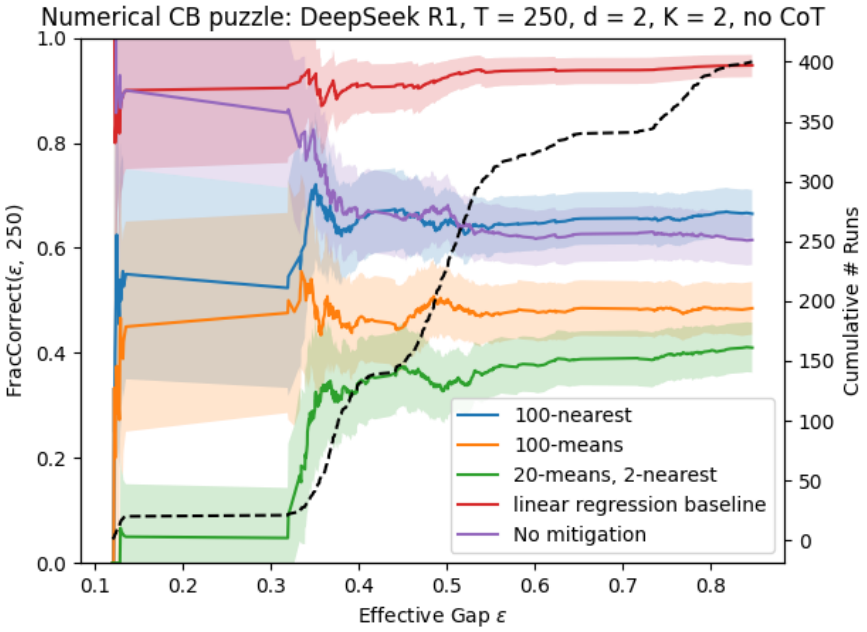


Figure 23. Performance of DEEPSEEK-R1-DISTILL-QWEN on our numerical CB puzzle.

C. Details for Section 3: LLMs as exploration oracles

C.1. Prompts

Prompt for the MovieLens experiments:

> [SYSTEM] You are a movie expert helping a user choose a movie.
>
> [USER] Here is a list of movies with their numeric IDs: ...
>
> From this list, choose {K} movies. You don't know what taste in movies the user has, so select a diverse set of movies from different genres such that they will most likely enjoy at least one of the movies you select. Respond ONLY with the {K} numeric IDs, one per line, with no extra text. Do NOT consider a movie's popularity when deciding whether to select it.

For the arXiv and Q/A tasks, we used a 2x2 space of prompts, as discussed in Section 3: { all-at-once, 1-by-1 } × { with, without } encouragement. We provide two illustrative examples below.

"All-at-once" prompt with encouragement for the Q/A task:

> [SYSTEM] I will give you an open-ended question. Come up with 5 different candidate answers. Reply only with the 5 candidate answers, and put each candidate answer on a separate line. Each answer should only be a few words, skipping any introductory phrasing and going straight to the essence. Try to come up with answers that are very different in spirit from one another.
>
> [USER] Here is the question: "What is the purpose of art?"

"One-by-one" prompt without encouragement for the arXiv task:

> [SYSTEM] I will give you an abstract and some candidate titles for a paper. Come up with a new candidate title that is relevant to the abstract, but different from the other candidate titles. Reply only with the candidate title.
>
> [USER] Here is the abstract: {abstract goes here}
>
> Here are the other candidate titles: {previous suggestions go here}

C.2. Dataset: Q/A task

Our open-ended question dataset consists of the following 10 questions and the corresponding "ground-truth" answers.

Questions:

0. What is the meaning of freedom?
1. How should we define success?
2. What is the role of technology in society?
3. What is the nature of reality?
4. What is the purpose of art?
5. What does it mean to live a fulfilling life?
6. How do cultural differences shape our understanding of morality?
7. What is the relationship between happiness and wealth?
8. How can we balance individuality and community in modern society?
9. What is the role of education in personal and societal growth?

Answers:

0. Freedom is an illusion shaped by societal norms and external influences.
1. Success should be defined as contributing to the greater good rather than personal achievement.
2. Technology disrupts the natural balance of society and often creates more problems than it solves.
3. Reality is subjective, varying entirely based on individual perception and experience.
4. The purpose of art is to challenge conventions and disrupt established ideas.
5. Fulfillment comes from embracing discomfort.
6. Cultural differences create moral superiority.
7. Wealth detracts from true happiness.
8. Individuality thrives when shaped by community.
9. Education's purpose is to challenge authority.

C.3. Dataset: arXiv task

Here is the list of paper titles we used in our arXiv dataset, along with their corresponding categories:

gr-qc

1. There is more to the de Sitter horizon than just the area
2. Mitigating cosmic variance in the Hellings-Downs curve: a Cosmic Microwave Background analogy
3. Calabi-Yau Feynman integrals in gravity: ε-factorized form for apparent singularities
4. QG from SymQRG: AdS_3/CFT_2 Correspondence as Topological Symmetry-Preserving Quantum RG Flow
5. Black hole solutions in theories of supergravity
6. Horndeski in motion
7. Wormholes from beyond
8. Regularizing the Pulsar Timing Array likelihood: A path towards Fourier Space
9. Solutions to the mode equation for a quantized massless scalar field outside a black hole that forms from the collapse of a null shell: Late-time behaviors and computation of the stress-energy tensor
10. Gravitational waves from regular black holes in extreme mass-ratio inspirals

hep-ex

1. Observation of the $K^+ \to \pi^+\nu\bar{\nu}$ decay and measurement of its branching ratio
2. Test of lepton flavour universality in W-boson decays into electrons and τ-leptons using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector
3. Searching for neutrino self-interactions at future muon colliders
4. Quantum Decoherence at ESSnuSB Experiment
5. Test of lepton flavour universality with $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$ decays
6. Cross-section measurements for the production of a W-boson in association with high-transverse-momentum jets in pp collisions at $\sqrt{s}$= 13 TeV with the ATLAS detector
7. Charmful two-body Ω_b decays in the light-front quark model

8. Observation of a spectral hardening in cosmic ray boron spectrum with the DAMPE space mission
9. New BaBar studies of high-order radiation and the new landscape of data-driven HVP predictions of the muon g-2
10. Toponium: the smallest bound state and simplest hadron in quantum mechanics

hep-lat

1. Quantum sampling on a quantum annealer for large volumes in the strong coupling limit for gauge group U(3)
2. Phase diagram of Rydberg atoms in a two-leg rectangular ladder
3. Graph Attention Hamiltonian Neural Networks: A Lattice System Analysis Model Based on Structural Learning
4. What do we know about the confinement mechanism?
5. Designing weight regularizations based on Lefschetz thimbles to stabilize complex Langevin
6. Likelihood of a zero in the proton elastic electric form factor
7. Real-Time Simulation of Asymmetry Generation in Fermion-Bubble Collisions
8. Investigating SU(3) with Nf=8 fundamental fermions at strong renormalized coupling
9. The determination of potential scales in 2+1 flavor QCD
10. Towards the phase diagram of fermions coupled with $SO(3)$ quantum links in $(2+1)$-D

hep-ph

1. Predictions for dimuon production in high-energy neutrino-proton collisions using the color dipole model
2. Extrapolating Jet Radiation with Autoregressive Transformers
3. Accurate Surrogate Amplitudes with Calibrated Uncertainties
4. Calabi-Yau Feynman integrals in gravity: ε-factorized form for apparent singularities
5. The causal structure of the quark propagator
6. Fuzzy Axions and Associated Relics
7. Non-Radial Oscillation Modes in Hybrid Stars with Hyperons and Delta Baryons: Full General Relativity Formalism vs. Cowling Approximation
8. Evidence for the Sombrero Galaxy as an Accelerator of the Highest-Energy Cosmic Rays
9. The cosmic history of Primordial Black Hole accretion and its uncertainties
10. Searching for neutrino self-interactions at future muon colliders

hep-th

1. There is more to the de Sitter horizon than just the area
2. Calabi-Yau Feynman integrals in gravity: ε-factorized form for apparent singularities
3. QG from SymQRG: AdS_3/CFT_2 Correspondence as Topological Symmetry-Preserving Quantum RG Flow
4. Geometrically constrained localized configurations engendering non-topological profile
5. The causal structure of the quark propagator
6. Entanglement Hamiltonian and orthogonal polynomials
7. Black hole solutions in theories of supergravity
8. Fuzzy Axions and Associated Relics

9. Celestial Mellin Amplitudes
10. Evidence for the Sombrero Galaxy as an Accelerator of the Highest-Energy Cosmic Rays

math-ph

1. QG from SymQRG: AdS_3/CFT_2 Correspondence as Topological Symmetry-Preserving Quantum RG Flow
2. Entanglement Hamiltonian and orthogonal polynomials
3. Fermi's golden rule in tunneling models with quantum waveguides perturbed by Kato class measures
4. Semiclassical measure of the propagation between two topological insulators
5. On the Protection Against Noise for Measurement-Based Quantum Computation
6. Calculating Spectra by Sequential High-Pass Filtering
7. Validity of the stochastic Landau approximation for super-pattern forming systems with a spatial 1:3 resonance
8. Multi-component Hamiltonian difference operators
9. Emptiness Instanton in Quantum Polytropic Gas
10. Unitary n-correlations with restricted support in random matrix theory

nucl-ex

1. The evidence of $N = 16$ shell closure and β-delayed neutron emission from $\wedge^{25}$F
2. Isotopic Transparency in Central Xe+Sn Collisions at 100 MeV/nucleon
3. Detecting the Coupling of Axion Dark Matter to Neutron Spins at Spallation Sources via Rabi Oscillation
4. Likelihood of a zero in the proton elastic electric form factor
5. Nuclear structure and direct reaction studies in particle-γ coincidence experiments at the FSU John D. Fox Superconducting Linear Accelerator Laboratory
6. Bottomonium-like states in proton collisions: Fragmentation and resummation
7. Towards a foundation model for heavy-ion collision experiments through point cloud diffusion
8. Study of the energy spectrum of alpha particles in an experiment on irradiation of a boron target with a proton beam at the Prometheus accelerator
9. Staking out the Proton Drip-Line of Thulium at the N=82 Shell Closure
10. Measurements of global and local spin polarization of $\wedge$ and $\bar{\wedge}$ in Au+Au collisions from the RHIC Beam Energy Scan

nucl-th

1. The causal structure of the quark propagator
2. Non-Radial Oscillation Modes in Hybrid Stars with Hyperons and Delta Baryons: Full General Relativity Formalism vs. Cowling Approximation
3. Isotopic Transparency in Central Xe+Sn Collisions at 100 MeV/nucleon
4. Quantum-Corrected Holographic Wilson Loop Correlators and Confinement
5. Dynamics of Hot QCD Matter 2024 – Bulk Properties
6. Spurious Isospin Breaking in the In-medium Similarity Renormalization Group
7. Likelihood of a zero in the proton elastic electric form factor
8. Born-Oppenheimer Renormalization group for High Energy Scattering: the Modified BFKL, or where did it all go?

9. Nuclear structure and direct reaction studies in particle-γ coincidence experiments at the FSU John D. Fox Superconducting Linear Accelerator Laboratory
10. Bottomonium-like states in proton collisions: Fragmentation and resummation

quant-ph

1. Noisy initial-state qubit-channel metrology with additional undesirable noisy evolution
2. The State Preparation of Multivariate Normal Distributions using Tree Tensor Network
3. QG from SymQRG: AdS_3/CFT_2 Correspondence as Topological Symmetry-Preserving Quantum RG Flow
4. Entanglement Hamiltonian and orthogonal polynomials
5. Learning interactions between Rydberg atoms
6. The isoholonomic inequality and tight implementations of holonomic quantum gates
7. Fermi's golden rule in tunneling models with quantum waveguides perturbed by Kato class measures
8. Temporal evolution of a forced optomechanical system with linear and quadratic field – mechanical oscillator couplings
9. Control of a Josephson Digital Phase Detector via an SFQ-based Flux Bias Driver
10. Commentary on the decomposition of universal multiport interferometers: how it works in practice

cs.AI

1. MaxInfoRL: Boosting exploration in reinforcement learning through information gain maximization
2. SepLLM: Accelerate Large Language Models by Compressing One Segment into One Separator
3. Stabilizing Reinforcement Learning in Differentiable Multiphysics Simulation
4. Revelations: A Decidable Class of POMDPs with Omega-Regular Objectives
5. Artificial Intelligence in Traffic Systems
6. The Impact of AI Assistance on Radiology Reporting: A Pilot Study Using Simulated AI Draft Reports
7. Can LLM Prompting Serve as a Proxy for Static Analysis in Vulnerability Detection
8. FSFM: A Generalizable Face Security Foundation Model via Self-Supervised Facial Representation Learning
9. Learning to Navigate in Mazes with Novel Layouts using Abstract Top-down Maps
10. SpeechPrune: Context-aware Token Pruning for Speech Information Retrieval

cs.CL

1. SepLLM: Accelerate Large Language Models by Compressing One Segment into One Separator
2. Making FETCH! Happen: Finding Emergent Dog Whistles Through Common Habitats
3. Semi-automated analysis of audio-recorded lessons: The case of teachers' engaging messages
4. Virtual Agent-Based Communication Skills Training to Facilitate Health Persuasion Among Peers
5. How Private are Language Models in Abstractive Summarization?
6. Can LLM Prompting Serve as a Proxy for Static Analysis in Vulnerability Detection
7. SpeechPrune: Context-aware Token Pruning for Speech Information Retrieval
8. The Open Source Advantage in Large Language Models (LLMs)
9. LLM-RG4: Flexible and Factual Radiology Report Generation across Diverse Input Contexts

10. ExecRepoBench: Multi-level Executable Code Completion Evaluation

cs.CV

1. PanSplat: 4K Panorama Synthesis with Feed-Forward Gaussian Splatting
2. Causal Diffusion Transformers for Generative Modeling
3. CAP4D: Creating Animatable 4D Portrait Avatars with Morphable Multi-View Diffusion Models
4. Wonderland: Navigating 3D Scenes from a Single Image
5. Stabilizing Reinforcement Learning in Differentiable Multiphysics Simulation
6. Instruction-based Image Manipulation by Watching How Things Move
7. IDArb: Intrinsic Decomposition for Arbitrary Number of Input Views and Illuminations
8. UniLoc: Towards Universal Place Recognition Using Any Single Modality
9. CPath-Omni: A Unified Multimodal Foundation Model for Patch and Whole Slide Image Analysis in Computational Pathology
10. CG-Bench: Clue-grounded Question Answering Benchmark for Long Video Understanding

cs.LG

1. MaxInfoRL: Boosting exploration in reinforcement learning through information gain maximization
2. SepLLM: Accelerate Large Language Models by Compressing One Segment into One Separator
3. No More Tuning: Prioritized Multi-Task Learning with Lagrangian Differential Multiplier Methods
4. Stabilizing Reinforcement Learning in Differentiable Multiphysics Simulation
5. Extrapolating Jet Radiation with Autoregressive Transformers
6. Bilevel Learning with Inexact Stochastic Gradients
7. LLMs for Cold-Start Cutting Plane Separator Configuration
8. LeARN: Learnable and Adaptive Representations for Nonlinear Dynamics in System Identification
9. Thermodynamics-informed graph neural networks for real-time simulation of digital human twins
10. Memory-Reduced Meta-Learning with Guaranteed Convergence

cs.NE

1. Deep-learning-based identification of individual motion characteristics from upper-limb trajectories towards disorder stage evaluation
2. Speeding Up the NSGA-II With a Simple Tie-Breaking Rule
3. Optimal Gradient Checkpointing for Sparse and Recurrent Architectures using Off-Chip Memory
4. Runtime Analysis for Multi-Objective Evolutionary Algorithms in Unbounded Integer Spaces
5. Theoretical Analysis of Quality Diversity Algorithms for a Classical Path Planning Problem
6. Populating cellular metamaterials on the extrema of attainable elasticity through neuroevolution
7. Deployment Pipeline from Rockpool to Xylo for Edge Computing
8. Interlocking-free Selective Rationalization Through Genetic-based Learning
9. EVOS: Efficient Implicit Neural Training via EVOlutionary Selector
10. Brain-inspired Chaotic Graph Backpropagation for Large-scale Combinatorial Optimization

cs.RO

1. MaxInfoRL: Boosting exploration in reinforcement learning through information gain maximization
2. Stabilizing Reinforcement Learning in Differentiable Multiphysics Simulation
3. LeARN: Learnable and Adaptive Representations for Nonlinear Dynamics in System Identification
4. Backstepping Control of Tendon-Driven Continuum Robots in Large Deflections Using the Cosserat Rod Model
5. Learning to Navigate in Mazes with Novel Layouts using Abstract Top-down Maps
6. Emma-X: An Embodied Multimodal Action Model with Grounded Chain of Thought and Look-ahead Spatial Reasoning
7. Lightweight Decentralized Neural Network-Based Strategies for Multi-Robot Patrolling
8. Learning Human-Aware Robot Policies for Adaptive Assistance
9. Hardware-in-the-loop Simulation Testbed for Geomagnetic Navigation
10. Sonar-based Deep Learning in Underwater Robotics: Overview, Robustness and Challenges

cs.IT

1. Codes from A_m-invariant polynomials
2. BA-BFL: Barycentric Aggregation for Bayesian Federated Learning
3. Capacity of Hierarchical Secure Coded Gradient Aggregation with Straggling Communication Links
4. Wireless Environmental Information Theory: A New Paradigm towards 6G Online and Proactive Environment Intelligence Communication
5. Quantum search in a dictionary based on fingerprinting-hashing
6. Identification Over Binary Noisy Permutation Channels
7. Iterative Detection and Decoding for Clustered Cell-Free Massive MIMO Networks
8. Structured Sampling for Robust Euclidean Distance Geometry
9. Study of Iterative Detection and Decoding for Multiuser Systems and MMSE Refinements with Active or Passive RIS
10. Shannon information and integrated information: message and meaning

cs.CR

1. Can LLM Prompting Serve as a Proxy for Static Analysis in Vulnerability Detection
2. Efficient Layered New Bit-Flipping QC-MDPC Decoder for BIKE Post-Quantum Cryptography
3. But Can You Use It? Design Recommendations for Differentially Private Interactive Systems
4. Efficiently Achieving Secure Model Training and Secure Aggregation to Ensure Bidirectional Privacy-Preservation in Federated Learning
5. On Large Language Models in Mission-Critical IT Governance: Are We Ready Yet?
6. Just a Simple Transformation is Enough for Data Protection in Vertical Federated Learning
7. SeSeMI: Secure Serverless Model Inference on Sensitive Data
8. DB-PAISA: Discovery-Based Privacy-Agile IoT Sensing+Actuation
9. OTA-Key: Over the Air Key Management for Flexible and Reliable IoT Device Provision
10. Android App Feature Extraction: A review of approaches for malware and app similarity detection

cs.DS

1. Approximating the Top Eigenvector in Random Order Streams

2. Witty: An Efficient Solver for Computing Minimum-Size Decision Trees
3. Adaptive Manipulation for Coalitions in Knockout Tournaments
4. Counting Butterflies over Streaming Bipartite Graphs with Duplicate Edges
5. Quantum search in a dictionary based on fingerprinting-hashing
6. Regularized Dikin Walks for Sampling Truncated Logconcave Measures, Mixed Isoperimetry and Beyond Worst-Case Analysis
7. Proportionally Fair Matching via Randomized Rounding
8. Logarithmic Positional Partition Interval Encoding
9. New results for the detection of bicliques
10. Deterministic Even-Cycle Detection in Broadcast CONGEST

cs.HC

1. Virtual Agent-Based Communication Skills Training to Facilitate Health Persuasion Among Peers
2. The Impact of AI Assistance on Radiology Reporting: A Pilot Study Using Simulated AI Draft Reports
3. Combining Large Language Models with Tutoring System Intelligence: A Case Study in Caregiver Homework Support
4. But Can You Use It? Design Recommendations for Differentially Private Interactive Systems
5. LLMs Can Simulate Standardized Patients via Agent Coevolution
6. LLM-DaaS: LLM-driven Drone-as-a-Service Operations from Text User Requests
7. Private Yet Social: How LLM Chatbots Support and Challenge Eating Disorder Recovery
8. Task-Based Role-Playing VR Game for Supporting Intellectual Disability Therapies
9. Privacy-Preserving Brain-Computer Interfaces: A Systematic Review
10. Accurate, Robust and Privacy-Preserving Brain-Computer Interface Decoding

math.AG

1. Geometry of 3-dimensional del Pezzo fibrations in positive characteristic
2. The Mordell-Schinzel conjecture for cubic diophantine equations
3. The many faces of a logarithmic scheme
4. Lorentzian polynomials and the incidence geometry of tropical linear spaces
5. p-adic Local Langlands Correspondence
6. Real del Pezzo surfaces without points
7. Linearization problem for finite subgroups of the plane Cremona group
8. Groupes de monodromie finie des variétés abéliennes
9. Duality for Arithmetic p-adic Pro-étale Cohomology of Analytic Spaces
10. The external activity complex of a pair of matroids

math.AT

1. Digital $n-$Manifolds With Or Without Boundaries
2. Spatiotemporal Persistence Landscapes
3. Simplifications of finite spaces equipped with sheaves

4. Rational homotopy theory of operad modules through colored operads
5. Algebraic Topology Without Open Sets: A Net Approach to Homotopy Theory in Limit Spaces
6. The geometry of simplicial distributions on suspension scenarios
7. On the Last Kervaire Invariant Problem
8. Machine Proofs for Adams Differentials and Extension Problems among CW Spectra
9. Finite asymptotic dimension and the coarse assembly map
10. Modeling $(\infty, 1)$-categories with Segal spaces

math.AP

1. Decay estimates for massive Dirac equation in a constant magnetic field
2. Semiclassical measure of the propagation between two topological insulators
3. Convex waves grazing convex obstacles to high order
4. A Note on Hyperbolic Relaxation of the Navier-Stokes-Cahn-Hilliard system for incompressible two-phase flow
5. Positive solutions to general semilinear overdetermined boundary problems
6. Capacitary measures in fractional order Sobolev spaces: Compactness and applications to minimization problems
7. Validity of the stochastic Landau approximation for super-pattern forming systems with a spatial 1:3 resonance
8. Spectral bounds for the operator pencil of an elliptic system in an angle
9. Infinite dimensional invariant tori for nonlinear Schrödinger equations
10. A Serrin-type over-determined problem for Hessian equations in the exterior domain

math.CT

1. Open Condensed Subgroups and Mackey's Formula
2. The Relational Quotient Completion
3. Classification of localizing subcategories along t-structures
4. Categorification of modules and construction of schemes
5. Rational RG flow, extension, and Witt class
6. Intrinsically Correct Sorting in Cubical Agda
7. Single and multi-valued Hilbert-bundle renormings
8. Extended (tri)dendriform algebras, pre-Lie algebras and post-Lie algebras as companion structures of extended Rota-Baxter algebras
9. On The Telescopic Picard Group
10. Enhanced 2-categorical structures, two-dimensional limit sketches and the symmetry of internalisation

math.GR

1. F-birestriction monoids in enriched signature
2. Linearization problem for finite subgroups of the plane Cremona group
3. Salter's question on the image of the Burau representation of B_4
4. Averaging groups
5. Enumerating Diagonalizable Matrices over $\mathbb{Z}_{p^k}$

6. The scale function for locally compact groups acting on non-positively curved spaces
7. A computational study of certain Weyl modules for type G_2 in characteristic 2
8. Left-Invariant Riemannian Distances on Higher-Rank Sol-Type Groups
9. Growth Rate Gap for Stable Subgroups
10. Computing Young's Natural Representations for Generalized Symmetric Groups

math.NT

1. The Mordell-Schinzel conjecture for cubic diophantine equations
2. Simultaneous and multiplicative Diophantine approximation on missing-digit fractals
3. Codes from A_m-invariant polynomials
4. Generalised Fermat equation: a survey of solved cases
5. Groupes de monodromie finie des variétés abéliennes
6. Vanishing of Witten zeta function at negative integers
7. Popa's "Recurrent Sequences" and Reciprocity
8. Duality for Arithmetic p-adic Pro-étale Cohomology of Analytic Spaces
9. About Eisenstein's Theorem
10. On the packing dimension of weighted singular matrices on fractals

math.OC

1. Bilevel Learning with Inexact Stochastic Gradients
2. Memory-Reduced Meta-Learning with Guaranteed Convergence
3. On Differential Stability of a Class of Convex Optimization Problems
4. Convergence of trust-region algorithms in compact metric spaces
5. Eckstein-Ferris-Pennanen-Robinson duality revisited: paramonotonicity, total Fenchel-Rockallar duality, and the Chambolle-Pock operator
6. Capacitary measures in fractional order Sobolev spaces: Compactness and applications to minimization problems
7. A monotone block coordinate descent method for solving absolute value equations
8. Bivariate rational approximations of the general temperature integral
9. Toward a Unified Theory of Gradient Descent under Generalized Smoothness
10. A particle system approach towards the global well-posedness of master equations for potential mean field games of control

math.ST

1. Optimality of the Right-Invariant Prior
2. The entropic optimal (self-)transport problem: Limit distributions for decreasing regularization with application to score function estimation
3. Causal Invariance Learning via Efficient Optimization of a Nonconvex Objective
4. A partial likelihood approach to tree-based density modeling and its application in Bayesian inference
5. Dual Unscented Kalman Filter Architecture for Sensor Fusion in Water Networks Leak Localization
6. Learning Massive-scale Partial Correlation Networks in Clinical Multi-omics Studies with HP-ACCORD

7. Well-Posedness and Stability of the Stochastic OGTT Model
8. Posterior asymptotics of high-dimensional spiked covariance model with inverse-Wishart prior
9. Model checking for high dimensional generalized linear models based on random projections
10. The Stein-log-Sobolev inequality and the exponential rate of convergence for the continuous Stein variational gradient descent method

q-bio.BM

1. Category-Specific Topological Learning of Metal-Organic Frameworks
2. Applications of Knot Theory for the Improvement of the AlphaFold Protein Database
3. EquiFlow: Equivariant Conditional Flow Matching with Optimal Transport for 3D Molecular Conformation Prediction
4. FlowDock: Geometric Flow Matching for Generative Protein-Ligand Docking and Affinity Prediction
5. NeuralPLexer3: Physio-Realistic Biomolecular Complex Structure Prediction with Flow Models
6. COMET: Benchmark for Comprehensive Biological Multi-omics Evaluation Tasks and Language Models
7. Quadratic unconstrained binary optimization and constraint programming approaches for lattice-based cyclic peptide docking
8. High-dimensional Statistics Applications to Batch Effects in Metabolomics
9. Precise Antigen-Antibody Structure Predictions Enhance Antibody Development with HelixFold-Multimer
10. Sampling-based Continuous Optimization with Coupled Variables for RNA Design

q-bio.GN

1. BarcodeMamba: State Space Models for Biodiversity Analysis
2. VEPerform: a web resource for evaluating the performance of variant effect predictors
3. A robust, scalable K-statistic for quantifying immune cell clustering in spatial proteomics data
4. Can linguists better understand DNA?
5. A Misclassification Network-Based Method for Comparative Genomic Analysis
6. DNA Fragments in Crude Oil Reveals Earth's Hidden History
7. Ancient DNA from 120-Million-Year-Old Lycoptera Fossils Reveals Evolutionary Insights
8. Emerging Challenges in Molecular Paleontology: Misapplication of Environmental DNA Fragments and Misconception of Deamination as a Key Criterion for In Situ DNA Identification
9. ProtGO: A Transformer based Fusion Model for accurately predicting Gene Ontology (GO) Terms from full scale Protein Sequences
10. DART-Eval: A Comprehensive DNA Language Model Evaluation Benchmark on Regulatory DNA

q-bio.QM

1. Deep-learning-based identification of individual motion characteristics from upper-limb trajectories towards disorder stage evaluation
2. Decoding Drug Discovery: Exploring A-to-Z In silico Methods for Beginners
3. BarcodeMamba: State Space Models for Biodiversity Analysis
4. FlowDock: Geometric Flow Matching for Generative Protein-Ligand Docking and Affinity Prediction
5. Reliable and superior elliptic Fourier descriptor normalization and its application software ElliShape with efficient image processing
6. MEATRD: Multimodal Anomalous Tissue Region Detection Enhanced with Spatial Transcriptomics
7. Cardiovascular Disease Detection By Leveraging Semi-Supervised Learning

8. Predictive Pattern Recognition Techniques Towards Spatiotemporal Representation of Plant Growth in Simulated and Controlled Environments: A Comprehensive Review
9. RAID-Database: human Responses to Affine Image Distortions
10. MiCull2 – simulating mastitis transmission through milking order

q-bio.PE

1. Asymmetric Interactions Shape Survival During Population Range Expansions
2. Quasispecies dynamics with time lags and periodic fluctuations in replication
3. Explicit modeling of density dependence in spatial capture-recapture models
4. Stochastic models in phylogenetic comparative methods: analytical properties and parameter estimation
5. Multivariate Aspects of Phylogenetic Comparative Methods
6. The expensive son hypothesis
7. Self-similarity in pandemic spread and fractal containment policies
8. Estimating excess mortality during the Covid-19 pandemic in Aotearoa New Zealand
9. An assessment of Alberta's strategy for controlling mountain pine beetle outbreaks
10. Mountain pine beetle struggles with jack pine: A mechanistic explanation for slowed range expansion in Alberta

q-fin.CP

1. S&P 500 Trend Prediction
2. Simulation of square-root processes made simple: applications to the Heston model
3. From Votes to Volatility Predicting the Stock Market on Election Day
4. SusGen-GPT: A Data-Centric LLM for Financial NLP and Sustainability Report Generation
5. FinGPT: Enhancing Sentiment-Based Stock Movement Prediction with Dissemination-Aware and Context-Enriched LLMs
6. Reciprocity in Interbank Markets
7. Integrative Analysis of Financial Market Sentiment Using CNN and GRU for Risk Prediction and Alert Systems
8. Financial Fine-tuning a Large Time Series Model
9. Geometric Deep Learning for Realized Covariance Matrix Forecasting
10. Isogeometric Analysis for the Pricing of Financial Derivatives with Nonlinear Models: Convertible Bonds and Options

q-fin.PM

1. Cost-aware Portfolios in a Large Universe of Assets
2. PolyModel for Hedge Funds' Portfolio Construction Using Machine Learning
3. Geometric Deep Learning for Realized Covariance Matrix Forecasting
4. LLMs for Time Series: an Application for Single Stocks and Statistical Arbitrage
5. A Joint Energy and Differentially-Private Smart Meter Data Market
6. Smart leverage? Rethinking the role of Leveraged Exchange Traded Funds in constructing portfolios to beat a benchmark
7. Correlation without Factors in Retail Cryptocurrency Markets
8. Turnover of investment portfolio via covariance matrix of returns

9. MILLION: A General Multi-Objective Framework with Controllable Risk for Portfolio Management
10. Dynamic ETF Portfolio Optimization Using enhanced Transformer-Based Models for Covariance and Semi-Covariance Prediction(Work in Progress)

q-fin.TR

1. Auto-Regressive Control of Execution Costs
2. FinGPT: Enhancing Sentiment-Based Stock Movement Prediction with Dissemination-Aware and Context-Enriched LLMs
3. Efficient and Verified Continuous Double Auctions
4. A Joint Energy and Differentially-Private Smart Meter Data Market
5. A theory of passive market impact
6. Uncertain Regulations, Definite Impacts: The Impact of the US Securities and Exchange Commission's Regulatory Interventions on Crypto Assets
7. Ergodic optimal liquidations in DeFi
8. MarketGPT: Developing a Pre-trained transformer (GPT) for Modeling Financial Time Series
9. Calculating Profits and Losses for Algorithmic Trading Strategies: A Short Guide
10. Market Making without Regret

stat.AP

1. But Can You Use It? Design Recommendations for Differentially Private Interactive Systems
2. Efficient Bayesian inversion for simultaneous estimation of geometry and spatial field using the Karhunen-Loève expansion
3. Chopin: An Open Source R-language Tool to Support Spatial Analysis on Parallelizable Infrastructure
4. Spatial Cross-Recurrence Quantification Analysis for Multi-Platform Contact Tracing and Epidemiology Research
5. P3LS: Point Process Partial Least Squares
6. Missing data imputation for noisy time-series data and applications in healthcare
7. Balancing Accuracy and Costs in Cross-Temporal Hierarchies: Investigating Decision-Based and Validation-Based Reconciliation
8. Statistical Problems in the Diagnosis of Shaken Baby Syndrome/Abusive Head Trauma: Limitations to Algorithms and the Need for Reliable Data
9. CESAR: A Convolutional Echo State AutoencodeR for High-Resolution Wind Forecasting
10. Cardiovascular Disease Detection By Leveraging Semi-Supervised Learning

stat.ML

1. Generalization Analysis for Deep Contrastive Representation Learning
2. Multiplex Dirichlet stochastic block model for clustering multidimensional compositional networks
3. BetaExplainer: A Probabilistic Method to Explain Graph Neural Networks
4. Bayesian Surrogate Training on Multiple Data Sources: A Hybrid Modeling Strategy
5. Scalable Temporal Anomaly Causality Discovery in Large Systems: Achieving Computational Efficiency with Binary Anomaly Flag Data
6. Conditional Diffusion Models Based Conditional Independence Testing
7. Generalized Bayesian deep reinforcement learning
8. A partial likelihood approach to tree-based density modeling and its application in Bayesian inference

9. A Mapper Algorithm with implicit intervals and its optimization
10. Learning Massive-scale Partial Correlation Networks in Clinical Multi-omics Studies with HP-ACCORD

stat.TH

1. Optimality of the Right-Invariant Prior
2. The entropic optimal (self-)transport problem: Limit distributions for decreasing regularization with application to score function estimation
3. Causal Invariance Learning via Efficient Optimization of a Nonconvex Objective
4. A partial likelihood approach to tree-based density modeling and its application in Bayesian inference
5. Dual Unscented Kalman Filter Architecture for Sensor Fusion in Water Networks Leak Localization
6. Learning Massive-scale Partial Correlation Networks in Clinical Multi-omics Studies with HP-ACCORD
7. Well-Posedness and Stability of the Stochastic OGTT Model
8. Posterior asymptotics of high-dimensional spiked covariance model with inverse-Wishart prior
9. Model checking for high dimensional generalized linear models based on random projections
10. The Stein-log-Sobolev inequality and the exponential rate of convergence for the continuous Stein variational gradient descent method

eess.IV

1. Are the Latent Representations of Foundation Models for Pathology Invariant to Rotation?
2. Towards Physically-Based Sky-Modeling
3. Ant Nest Detection Using Underground P-Band TomoSAR
4. Ensemble Learning and 3D Pix2Pix for Comprehensive Brain Tumor Analysis in Multimodal MRI
5. Point Cloud-Assisted Neural Image Compression
6. Flex-PE: Flexible and SIMD Multi-Precision Processing Element for AI Workloads
7. Fast-staged CNN Model for Accurate pulmonary diseases and Lung cancer detection
8. High-speed and High-quality Vision Reconstruction of Spike Camera with Spike Stability Theorem
9. Data-driven Precipitation Nowcasting Using Satellite Imagery
10. Block-Based Multi-Scale Image Rescaling

eess.SP

1. Rate-Splitting Multiple Access for Integrated Sensing and Communications: A First Experimental Study
2. Soil moisture estimation of bare and vegetation-covered areas using a P/L/C-band SAR
3. Ant Nest Detection Using Underground P-Band TomoSAR
4. Scalable Data Transmission Framework for Earth Observation Satellites with Channel Adaptation
5. Sonar-based Deep Learning in Underwater Robotics: Overview, Robustness and Challenges
6. Evaluating the Efficacy of Vectocardiographic and ECG Parameters for Efficient Tertiary Cardiology Care Allocation Using Decision Tree Analysis
7. Acceleration and Parallelization Methods for ISRS EGN Model
8. On-the-Fly Interrogation of Mobile Passive Sensors from the Fusion of Optical and Radar Data

9. Capacity Analysis on OAM-Based Wireless Communications: An Electromagnetic Information Theory Perspective
10. Probabilistic GOSPA: A Metric for Performance Evaluation of Multi-Object Filters with Uncertainties

econ.EM

1. Moderating the Mediation Bootstrap for Causal Inference
2. VAR models with an index structure: A survey with new results
3. Treatment Evaluation at the Intensive and Extensive Margins
4. Forecasting realized covariances using HAR-type models
5. Do LLMs Act as Repositories of Causal Knowledge?
6. An overview of meta-analytic methods for economic research
7. A Neyman-Orthogonalization Approach to the Incidental Parameter Problem
8. Geometric Deep Learning for Realized Covariance Matrix Forecasting
9. A Kernel Score Perspective on Forecast Disagreement and the Linear Pool
10. The Global Carbon Budget as a cointegrated system

econ.GN

1. Multiplexing in Networks and Diffusion
2. Transition dynamics of electricity asset-owning firms
3. Binary or nonbinary? An evolutionary learning approach to gender identity
4. On Prior Confidence and Belief Updating
5. Strategically Acting on Information
6. Is Polarization an Inevitable Outcome of Similarity-Based Content Recommendations? – Mathematical Proofs and Computational Validation
7. Re-examining the social impact of silver monetization in the Ming Dynasty from the perspective of supply and demand
8. Delving into Youth Perspectives on In-game Gambling-like Elements: A Proof-of-Concept Study Utilising Large Language Models for Analysing User-Generated Text Data
9. Does Low Spoilage Under Cold Conditions Foster Cultural Complexity During the Foraging Era? – A Theoretical and Computational Inquiry
10. Emulating the Global Change Analysis Model with Deep Learning

C.4. Robustness Check: Q/A task with another embedding model

We ran a robustness check on the first six open-ended question experiments using the universal sentence encoder of Cer (2018) as our embedding model. Our results remain largely unchanged, and are summarized in Figure 24.

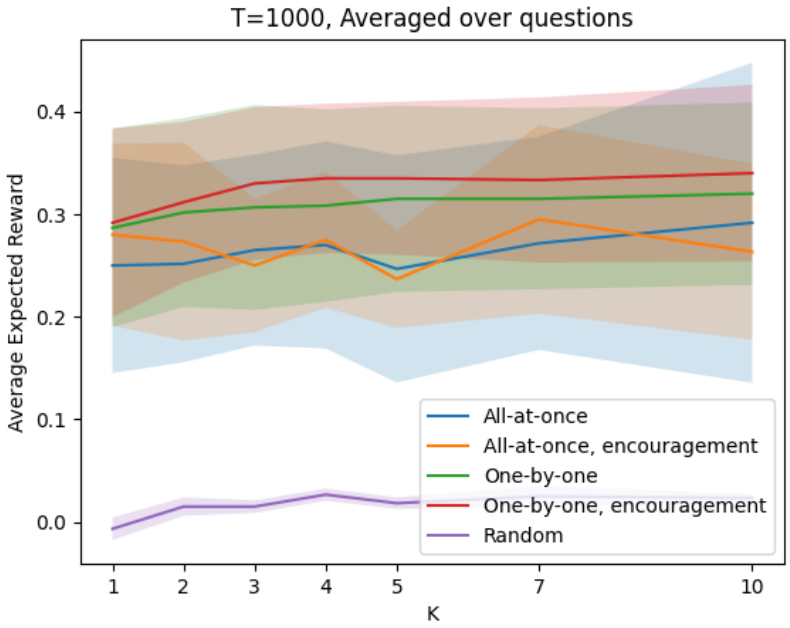


Figure 24. Results averaged over the first six questions, for embeddings generated using the universal sentence encoder.

C.5. Additional Results: arXiv task

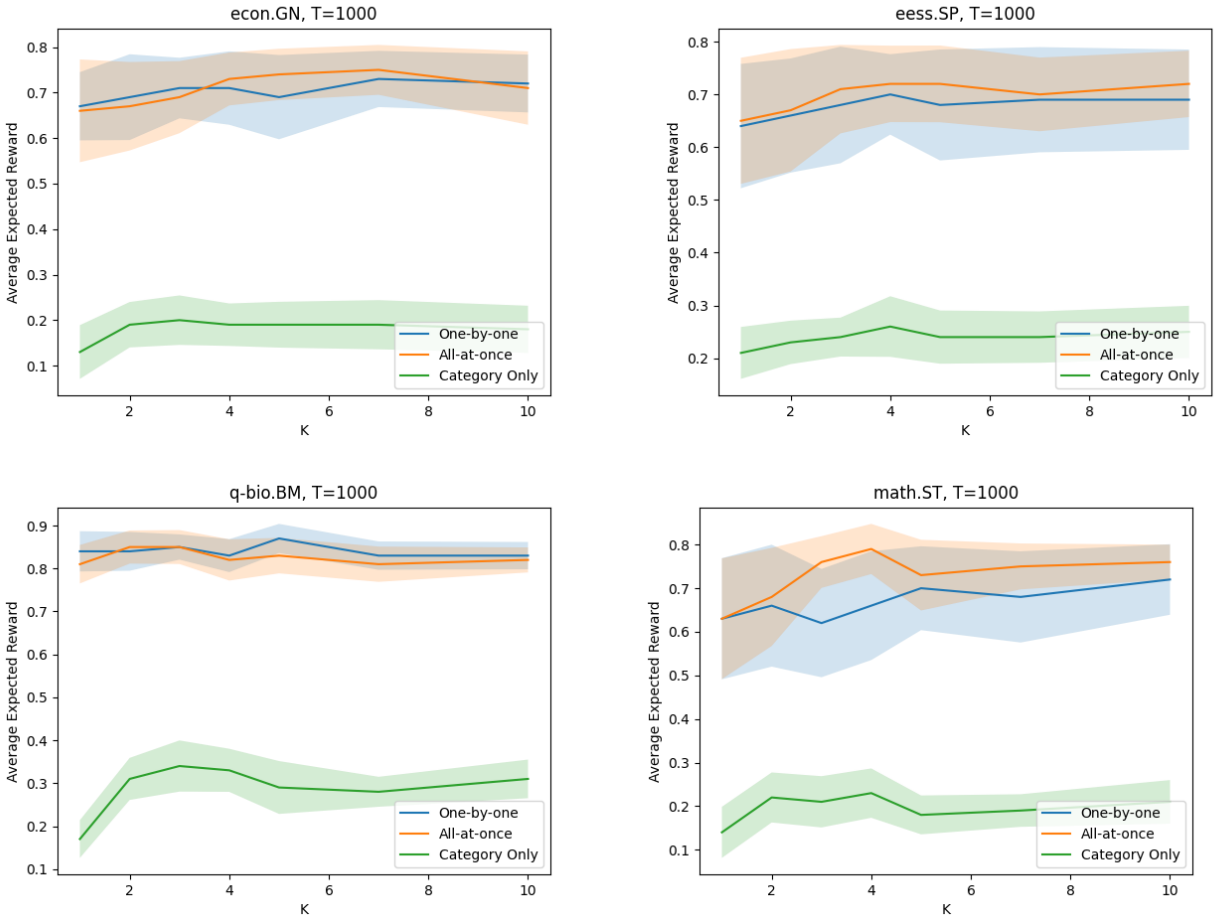


Figure 25. Results for other arXiv categories

Table 3. Performance for all-at-once on arXiv tasks.

	K=1	K=2	K=5
gr-qc	0.63	0.64	0.68
hep-ex	0.81	0.81	0.83
hep-lat	0.72	0.72	0.72
hep-ph	0.7	0.76	0.75
hep-th	0.65	0.71	0.73
math-ph	0.64	0.73	0.74
nucl-ex	0.73	0.79	0.75
nucl-th	0.65	0.69	0.71
quant-ph	0.68	0.71	0.75
cs.AI	0.66	0.71	0.72
cs.CL	0.66	0.71	0.75
cs.CV	0.72	0.74	0.71
cs.LG	0.68	0.72	0.74
cs.NE	0.71	0.78	0.78
cs.RO	0.76	0.79	0.78
cs.IT	0.72	0.73	0.72
cs.CR	0.7	0.72	0.74
cs.DS	0.75	0.77	0.77
cs.HC	0.75	0.75	0.75
math.AG	0.7	0.78	0.78
math.AT	0.68	0.7	0.71
math.AP	0.7	0.79	0.78
math.CT	0.65	0.69	0.71
math.GR	0.73	0.77	0.76
math.NT	0.73	0.79	0.77
math.OC	0.77	0.79	0.76
math.ST	0.63	0.68	0.73
q-bio.BM	0.81	0.85	0.83
q-bio.GN	0.76	0.78	0.79
q-bio.QM	0.76	0.78	0.78
q-bio.PE	0.8	0.82	0.8
q-fin.CP	0.74	0.78	0.77
q-fin.PM	0.74	0.77	0.78
q-fin.TR	0.74	0.78	0.78
stat.AP	0.73	0.69	0.75
stat.ML	0.7	0.73	0.74
stat.TH	0.65	0.67	0.78
eess.IV	0.67	0.73	0.72
eess.SP	0.65	0.67	0.72
econ.EM	0.62	0.68	0.7
econ.GN	0.66	0.67	0.74

Table 4. Performance for one-by-one on arXiv tasks.

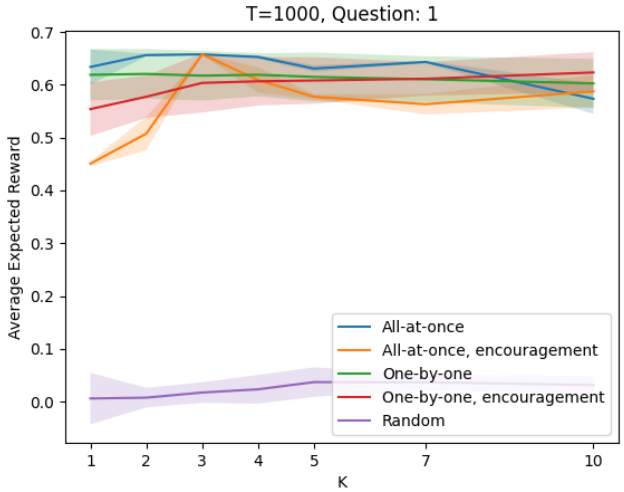
	K=1	K=2	K=5
gr-qc	0.63	0.64	0.65
hep-ex	0.78	0.76	0.81
hep-lat	0.72	0.72	0.74
hep-ph	0.7	0.72	0.73
hep-th	0.64	0.69	0.68
math-ph	0.65	0.71	0.71
nucl-ex	0.72	0.74	0.76
nucl-th	0.64	0.67	0.71
quant-ph	0.7	0.71	0.71
cs.AI	0.7	0.74	0.73
cs.CL	0.69	0.71	0.75
cs.CV	0.73	0.74	0.77
cs.LG	0.67	0.72	0.72
cs.NE	0.74	0.76	0.77
cs.RO	0.78	0.79	0.78
cs.IT	0.76	0.75	0.75
cs.CR	0.72	0.71	0.74
cs.DS	0.75	0.78	0.78
cs.HC	0.72	0.72	0.73
math.AG	0.68	0.77	0.78
math.AT	0.63	0.69	0.7
math.AP	0.72	0.75	0.76
math.CT	0.62	0.71	0.73
math.GR	0.71	0.76	0.74
math.NT	0.73	0.75	0.73
math.OC	0.73	0.77	0.8
math.ST	0.63	0.66	0.7
q-bio.BM	0.84	0.84	0.87
q-bio.GN	0.69	0.76	0.75
q-bio.QM	0.76	0.8	0.77
q-bio.PE	0.8	0.79	0.82
q-fin.CP	0.71	0.72	0.76
q-fin.PM	0.67	0.77	0.73
q-fin.TR	0.72	0.74	0.75
stat.AP	0.73	0.72	0.79
stat.ML	0.74	0.75	0.77
stat.TH	0.61	0.64	0.72
eess.IV	0.71	0.72	0.74
eess.SP	0.64	0.66	0.68
econ.EM	0.66	0.66	0.67
econ.GN	0.67	0.69	0.69

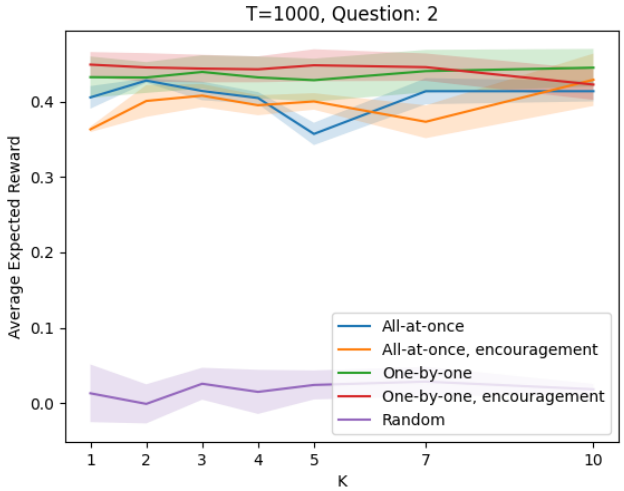
Table 5. Performance for Category Only baseline on arXiv tasks.

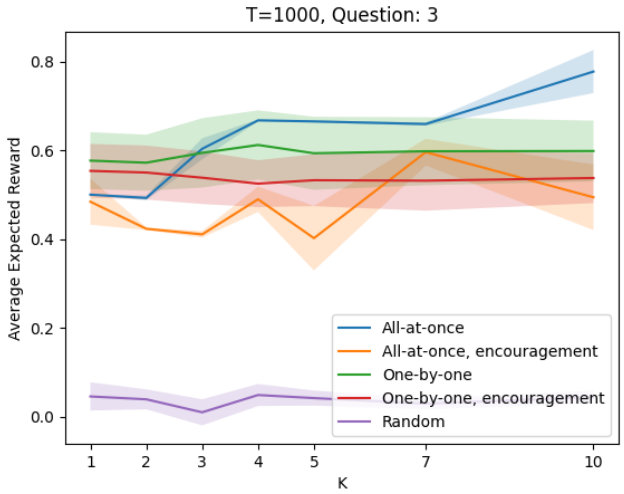
	K=1	K=2	K=5
gr-qc	0.28	0.32	0.36
hep-ex	0.25	0.35	0.38
hep-lat	0.31	0.31	0.38
hep-ph	0.25	0.26	0.32
hep-th	0.24	0.26	0.27
math-ph	0.22	0.3	0.3
nucl-ex	0.37	0.38	0.35
nucl-th	0.31	0.32	0.33
quant-ph	0.23	0.27	0.27
cs.AI	0.15	0.17	0.18
cs.CL	0.14	0.19	0.26
cs.CV	0.19	0.22	0.32
cs.LG	0.21	0.19	0.24
cs.NE	0.27	0.31	0.3
cs.RO	0.27	0.27	0.31
cs.IT	0.29	0.31	0.34
cs.CR	0.21	0.24	0.32
cs.DS	0.2	0.2	0.22
cs.HC	0.12	0.2	0.23
math.AG	0.32	0.33	0.33
math.AT	0.33	0.34	0.4
math.AP	0.19	0.23	0.32
math.CT	0.25	0.23	0.29
math.GR	0.25	0.28	0.34
math.NT	0.21	0.29	0.32
math.OC	0.2	0.27	0.3
math.ST	0.14	0.22	0.18
q-bio.BM	0.17	0.31	0.29
q-bio.GN	0.2	0.25	0.32
q-bio.QM	0.03	0.12	0.1
q-bio.PE	0.28	0.32	0.31
q-fin.CP	0.34	0.36	0.37
q-fin.PM	0.39	0.39	0.43
q-fin.TR	0.33	0.34	0.38
stat.AP	0.07	0.12	0.1
stat.ML	0.19	0.21	0.28
stat.TH	-0.01	0.07	0.12
eess.IV	0.18	0.19	0.24
eess.SP	0.21	0.23	0.24
econ.EM	0.21	0.32	0.37
econ.GN	0.13	0.19	0.19

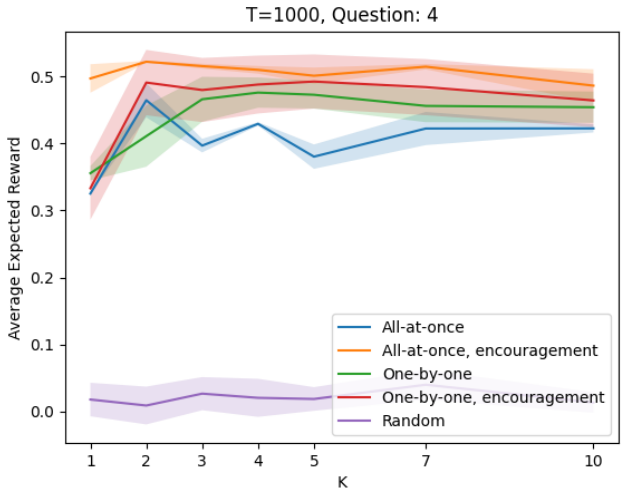
C.6. Additional Results: QA task

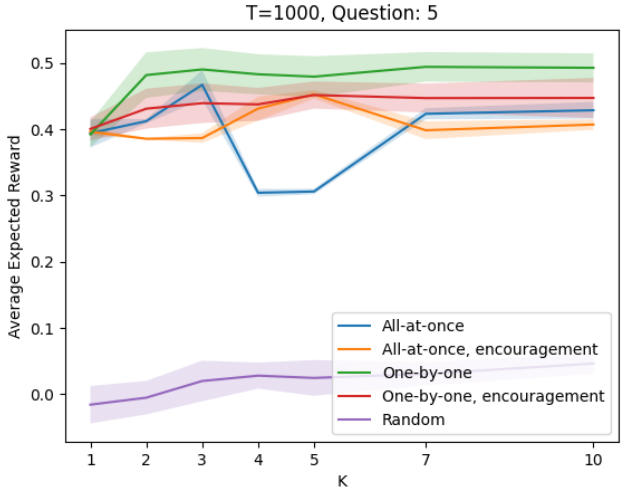
While previously we presented the plot for Question 0 (in Figure 7), below are the individual plots for the remaining 9 questions, in the same notation. Throughout, we used the Sentence-BERT encoder.

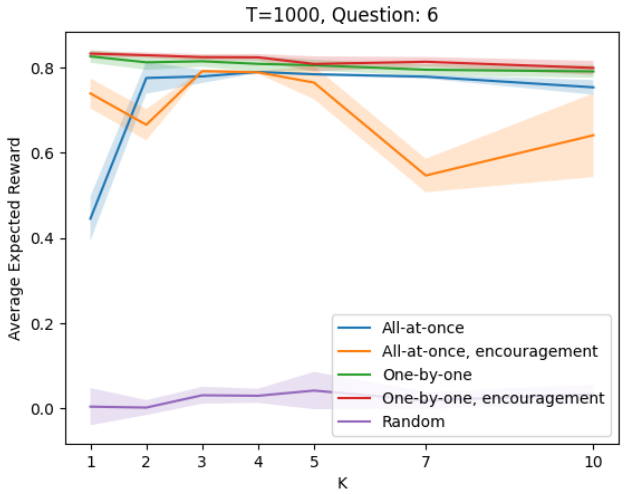


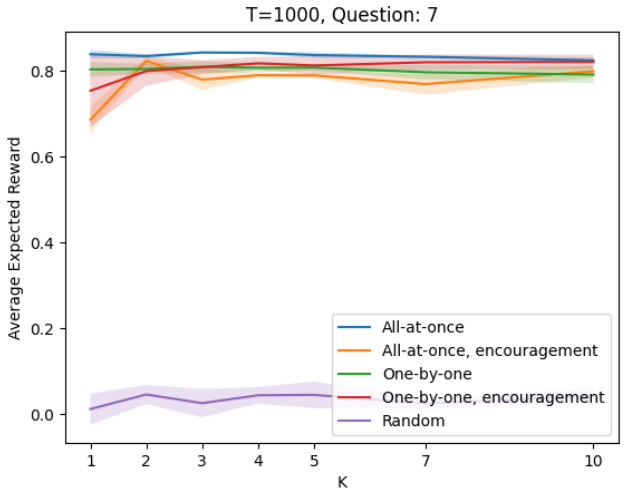


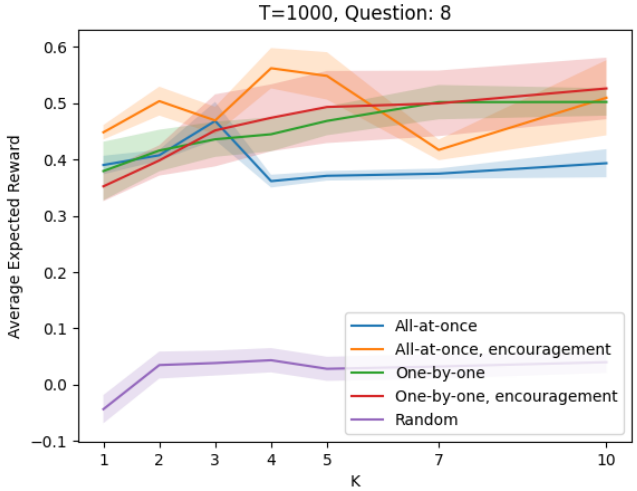

T=1000, Question: 3
Average Expected Reward
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
7
10
K
All-at-once
All-at-once, encouragement
One-by-one
One-by-one, encouragement
Random


T=1000, Question: 4
Average Expected Reward
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
7
10
K
All-at-once
All-at-once, encouragement
One-by-one
One-by-one, encouragement
Random


T=1000, Question: 5
Average Expected Reward
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
7
10
K
All-at-once
All-at-once, encouragement
One-by-one
One-by-one, encouragement
Random


T=1000, Question: 6
Average Expected Reward
K
All-at-once
All-at-once, encouragement
One-by-one
One-by-one, encouragement
Random


T=1000, Question: 7
Average Expected Reward
K
All-at-once
All-at-once, encouragement
One-by-one
One-by-one, encouragement
Random


T=1000, Question: 8
Average Expected Reward
K
All-at-once
All-at-once, encouragement
One-by-one
One-by-one, encouragement
Random

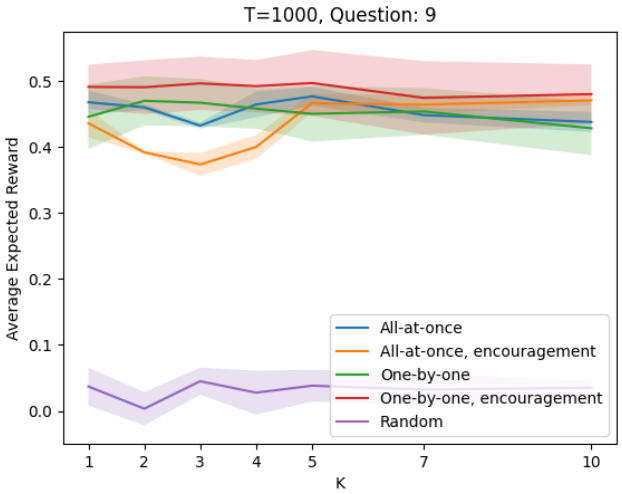

T=1000, Question: 9
Average Expected Reward
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
7
10
K
All-at-once
All-at-once, encouragement
One-by-one
One-by-one, encouragement
Random

Table 6. Performance comparison for all-at-once on open-ended questions.

	K=1	K=2	K=3	K=4	K=5	K=7	K=10
Q0:	0.39	0.4	0.4	0.4	0.41	0.42	0.56
Q1:	0.63	0.66	0.66	0.65	0.63	0.64	0.57
Q2:	0.41	0.43	0.41	0.4	0.36	0.41	0.41
Q3:	0.5	0.49	0.6	0.67	0.67	0.66	0.78
Q4:	0.33	0.46	0.4	0.43	0.38	0.42	0.42
Q5:	0.39	0.41	0.47	0.3	0.31	0.42	0.43
Q6:	0.45	0.78	0.78	0.79	0.78	0.78	0.75
Q7:	0.84	0.84	0.84	0.84	0.84	0.83	0.83
Q8:	0.39	0.41	0.47	0.36	0.37	0.37	0.39
Q9:	0.47	0.46	0.43	0.46	0.48	0.45	0.44

Table 7. Performance comparison for all-at-once with encouragement on open-ended questions.

	K=1	K=2	K=3	K=4	K=5	K=7	K=10
Q0:	0.4	0.4	0.34	0.35	0.37	0.52	0.55
Q1:	0.45	0.51	0.66	0.61	0.58	0.56	0.59
Q2:	0.36	0.4	0.41	0.4	0.4	0.37	0.43
Q3:	0.48	0.42	0.41	0.49	0.4	0.6	0.49
Q4:	0.5	0.52	0.52	0.51	0.5	0.51	0.49
Q5:	0.4	0.39	0.39	0.43	0.45	0.4	0.41
Q6:	0.74	0.67	0.79	0.79	0.77	0.55	0.64
Q7:	0.69	0.82	0.78	0.79	0.79	0.77	0.8
Q8:	0.45	0.5	0.47	0.56	0.55	0.42	0.51
Q9:	0.44	0.39	0.37	0.4	0.47	0.46	0.47

Table 8. Performance comparison for one-by-one on open-ended questions.

	K=1	K=2	K=3	K=4	K=5	K=7	K=10
Q0:	0.33	0.4	0.4	0.43	0.43	0.47	0.52
Q1:	0.62	0.62	0.62	0.62	0.61	0.61	0.6
Q2:	0.43	0.43	0.44	0.43	0.43	0.44	0.45
Q3:	0.58	0.57	0.59	0.61	0.59	0.6	0.6
Q4:	0.36	0.41	0.47	0.48	0.47	0.46	0.45
Q5:	0.39	0.48	0.49	0.48	0.48	0.49	0.49
Q6:	0.83	0.81	0.82	0.81	0.81	0.8	0.79
Q7:	0.8	0.81	0.81	0.81	0.81	0.8	0.79
Q8:	0.38	0.42	0.44	0.44	0.47	0.5	0.5
Q9:	0.45	0.47	0.47	0.46	0.45	0.45	0.43

Table 9. Performance comparison for one-by-one with encouragement on open-ended questions.

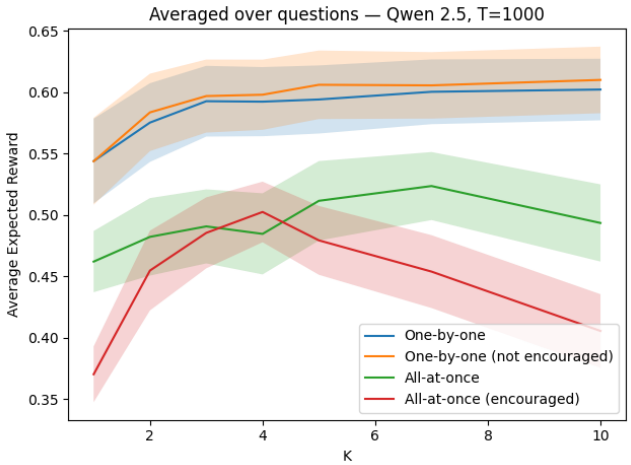
	K=1	K=2	K=3	K=4	K=5	K=7	K=10
Q0:	0.35	0.37	0.5	0.54	0.55	0.58	0.59
Q1:	0.55	0.58	0.6	0.61	0.61	0.61	0.62
Q2:	0.45	0.45	0.44	0.44	0.45	0.45	0.42
Q3:	0.55	0.55	0.54	0.53	0.53	0.53	0.54
Q4:	0.33	0.49	0.48	0.49	0.49	0.48	0.46
Q5:	0.4	0.43	0.44	0.44	0.45	0.45	0.45
Q6:	0.83	0.83	0.82	0.82	0.81	0.81	0.8
Q7:	0.75	0.8	0.81	0.82	0.81	0.82	0.82
Q8:	0.35	0.4	0.45	0.47	0.49	0.5	0.53
Q9:	0.49	0.49	0.5	0.49	0.5	0.47	0.48

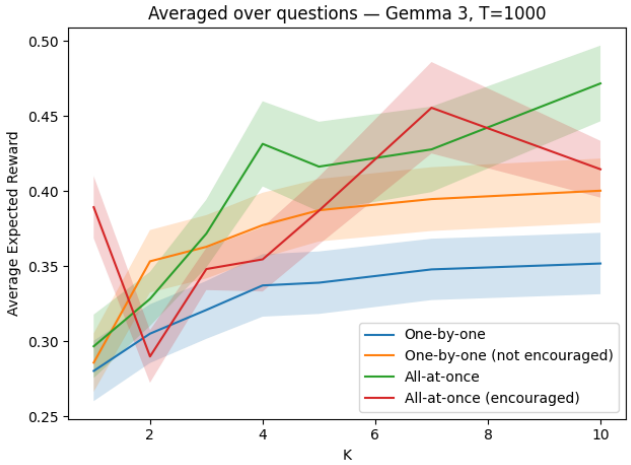
Table 10. Performance comparison for random actions on open-ended questions.

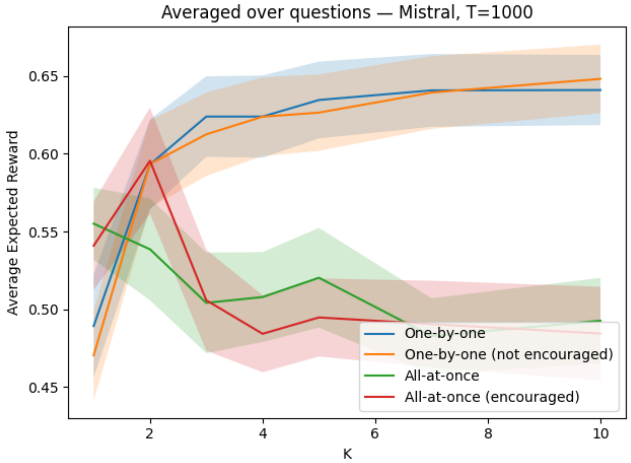
	K=1	K=2	K=3	K=4	K=5	K=7	K=10
Q0:	-0.01	-0.0	0.01	0.01	0.03	0.03	0.02
Q1:	0.01	0.01	0.02	0.02	0.04	0.04	0.03
Q2:	0.01	-0.0	0.03	0.02	0.02	0.03	0.02
Q3:	0.05	0.04	0.01	0.05	0.04	0.03	0.04
Q4:	0.02	0.01	0.03	0.02	0.02	0.04	0.01
Q5:	-0.02	-0.01	0.02	0.03	0.02	0.03	0.05
Q6:	0.0	0.0	0.03	0.03	0.04	0.02	0.03
Q7:	0.01	0.05	0.03	0.04	0.04	0.02	0.04
Q8:	-0.04	0.03	0.04	0.04	0.03	0.03	0.04
Q9:	0.04	0.0	0.04	0.03	0.04	0.03	0.04

C.7. Robustness Check with Other Models

We include results for Qwen2.5-7B-Instruct (QWEN-2.5), Gemma 3 12B (GEMMA-3), and Mistral-7B-Instruct-v0.3 (MISTRAL-7B) on our explore Q/A puzzles, all in the setting of Figure 7. As was the case with the GPT models, performance increased with K when the candidate answers are generated one-by-one. When generating answers all-at-once, we observed that QWEN-2.5 and MISTRAL-7B had trouble generating concise answers, leading to a drop in performance as K increases.







C.8. Benchmarking encoders

Here we benchmark the two encoders we use (Sentence-BERT and the universal sentence encoder) by measuring the cosine similarity between semantically similar/different words.

Table 11. Cosine similarity of different words.

	Sentence-BERT	Universal Sentence Encoder
dog, tacos:	0.25	0.24
Pittsburgh, tiki bar:	0.12	0.17
Honolulu, tiki bar:	0.30	0.25
Pittsburgh, Honolulu:	0.41	0.29
angel, devil:	0.48	0.54
machine learning, artificial intelligence:	0.70	0.58
war, peace:	0.61	0.49
love, hate:	0.49	0.59
love, affection:	0.62	0.56
war, battle:	0.74	0.57
machine learning, battle:	0.25	0.19

The similarity scores of both models in Table 11 suggest that while the embeddings produced by both embedding models are generally "in the ballpark" of what one would consider "similar"/"different", they are still a somewhat coarse measure of distance, which may explain the similar performance of our different prompting strategies.

C.9. Explore Experiments on the Movie Lens Dataset

Prompt:

> [SYSTEM] You are a movie expert helping a user choose a movie.
>
> [USER] Here is a list of movies with their numeric IDs:{movie_list_str}
>
> From this list, choose K movies. You don't know what taste in movies the user has, so select a diverse set of movies from different genres such that they will most likely enjoy at least one of the movies you select. Respond ONLY with the K numeric IDs, one per line, with no extra text. Do NOT consider a movie's popularity when deciding whether to select it.