

# Calibrate-Then-Act: Cost-Aware Exploration in LLM Agents

Wenxuan Ding<sup>1</sup> Nicholas Tomlin<sup>1</sup> Greg Durrett<sup>1</sup>

## Abstract

LLMs are increasingly being used for complex problems which are not necessarily resolved in a single response, but require interacting with an environment to acquire information. In these scenarios, LLMs must reason about inherent cost-uncertainty tradeoffs in when to stop exploring and commit to an answer. For instance, on a programming task, an LLM should test a generated code snippet if it is uncertain about the correctness of that code; the cost of writing a test is nonzero, but typically lower than the cost of making a mistake. In this work, we show that we can induce LLMs to explicitly reason about balancing these cost-uncertainty tradeoffs, then perform more optimal environment exploration. We formalize multiple tasks, including information retrieval and coding, as sequential decision-making problems under uncertainty. Each problem has latent environment state that can be reasoned about via a prior which is passed to the LLM agent. We introduce a framework called Calibrate-Then-Act (CTA), where we feed the LLM this additional context to enable it to act more optimally. This improvement is preserved even under RL training of both the baseline and CTA. Our results on information-seeking QA and on a simplified coding task show that making cost-benefit tradeoffs explicit with CTA can help agents discover more optimal decision-making strategies.

## 1. Introduction

Large language model (LLM) agents are increasingly tasked with operating in environments where information is incomplete. Behaving rationally requires gaining information by exploring the environment. However, exploration comes with a cost: every additional step increases API costs, interaction latency, and user burden.

<sup>1</sup>New York University. Correspondence to: Wenxuan Ding <wd2403@nyu.edu>. Code and data available at <https://github.com/Wenwen-D/env-explorer>.

Preprint. February 19, 2026.

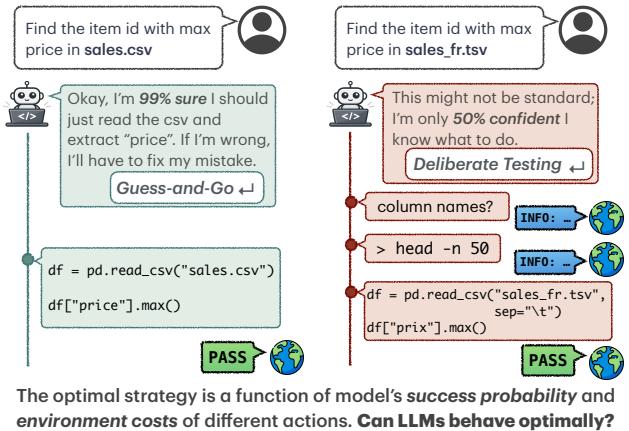


Figure 1. Given the same task, a coding agent may either verify assumptions via intermediate checks carefully (right) or attempt a direct solution as soon as possible (left). The optimal choice depends on uncertainty and specific cost constraints. Calibrate-Then-Act (CTA) materializes this information for better decision-making.

This exploration and its cost come in many forms. In software development and debugging, agents must decide whether to run targeted checks or perform full execution before committing to a solution (Zhou et al., 2025). In machine learning experimentation, practitioners balance inexpensive proxy evaluations against costly full training runs under limited compute budgets (Ji & Carin, 2007; Hennig et al., 2024; Xu et al., 2025). In diagnosis and scientific discovery, additional tests or experiments reduce uncertainty but incur monetary, temporal, or safety costs (Kärkkäinen et al., 2019; Li & Oliva, 2025; Gupta et al., 2025). Online decision-making settings such as shopping (Yang et al., 2023; Wang et al., 2025d), recommendation (Herlihy et al., 2024), and tool-augmented question answering (Yao et al., 2023) exhibit the same structure, as agents weigh further information gathering against acting with partial information.

Agent policies for this exploration depend in a complex way on their prompt, their inputs, and their training data. However, these policies are frequently *static*. For instance, ChatGPT Deep Research always asks a single round of clarifying questions before searching (Deng et al., 2025), and coding agents like SWE-agent (Yang et al., 2024) start by reading through an existing codebase. This contrasts

with settings like Figure 1, where we see that a model can act more efficiently if it has confidence that it understands the problem setup and appropriately trades off exploration against action costs.

In this work, we ask: **how can we develop LLMs that explore in a Pareto-optimal way under varying cost and uncertainty profiles?** Leveraging the strong reasoning capabilities of modern LLMs, we propose a framework called Calibrate-Then-Act (CTA), which decouples the calibration of uncertainty from the reasoning of action selection. The key insight is that by presenting priors explicitly to the model, we induce the model to reason about an underlying sequential decision-making problem abstractly and discover the optimal action.

We first study a synthetic task to illustrate this. On these “Pandora’s Box” problems (Weitzman, 1979), an agent faces multiple hidden boxes with known prior reward distributions and must decide which boxes to open (each incurring a cost) and when to stop and commit to the best observed option. Even a small thinking model (Qwen3-8B) can compute the optimal action. We then study two more realistic settings: (1) *knowledge QA with optional retrieval*, where uncertainty is inferred from the model’s own confidence; and (2) *coding tasks* where priors regarding environmental structure (e.g., file schemas) are derived from cues learned through past experience. Feeding calibration information via a prompt (CTA-PROMPTED) enables dynamic decision-making that is lacking from basic LLMs. Crucially, this behavior does not emerge from basic RL alone: an LLM trained with RL on the coding task fails to internalize the relevant priors from end-to-end training and does not reliably adopt the correct adaptive strategy. In contrast, CTA explicitly exposes these priors to the policy and can be combined with RL to yield further, consistent performance gains.

Our contributions are: (1) Framing environment exploration as a sequential decision-making problem; (2) Introducing Calibrate-Then-Act, which induces LLMs to reason about the optimality of their actions and achieve better cost-performance tradeoffs than baselines.

## 2. The Environment Exploration Problem

We formalize cost-aware environment exploration task as sequential decision-making problem. Our agent, which for the purposes of this work will be an LLM, is given a query  $\mathbf{x}$  and operates in some environment, which can be defined as a partially-observable Markov Decision Process  $\mathcal{W} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, O, T, R, D_\theta)$ , a tuple of states  $\mathcal{S}$ , actions  $\mathcal{A}$ , observations  $\mathcal{O}$ , observation function  $O$ , transition function  $T$ , reward function  $R$ , and parameterized discount function  $D_\theta$ , which integrates the cost.

In the settings we consider,  $\mathcal{A}$  and  $\mathcal{O}$  are both string-valued

spaces; LLMs produce string actions (code, API calls, etc.) and receive string-valued responses from the environment. The observation function  $O$  produces string realizations of the underlying environment; e.g., in Figure 1, the results of executing commands return string output in the terminal reflecting the underlying state of the environment.

The environment contains problem-critical unobserved features that will determine the agent’s performance, e.g., details about the formatting of the unobserved file in Figure 1. These can be thought of as a subset of the information in  $\mathcal{S}$ . We represent these as a random variable  $Z$  taking values  $\mathbf{z} \in \mathcal{Z}$ .

The agent interacts with the environment over multiple timesteps before terminating. At each timestep  $t$ , the agent selects an action  $a_t \in \mathcal{A}$  and receives an observation  $o_t \in \mathcal{O}$ ; for simplicity, we assume that  $o_t$  encodes  $a_t$ . Based on this information, we can form an idealized posterior distribution  $b_t(Z) = p(Z | \mathbf{x}, o_{0:t})$  which reflects remaining uncertainty over the latent variables. The action space generally consists of multiple *exploration* actions and a *commit* action, which terminates the episode by producing a final result. In Figure 1, exploration actions include querying aspects of the input data file; other actions (not shown) include running the code or writing and running unit tests.

Each action incurs different costs depending on the setting, while the commit action corresponds to returning a final solution. These costs are reflected by a discount factor  $D_\theta(a_{1:T}) \in [0, 1]$ , which discounts the value of successful task completion based on the exploration actions taken prior to commitment.

Finally, the agent receives reward upon committing:

$$R = \mathbb{I}[\text{task completed at } a_t] \cdot D_\theta(a_{1:T}),$$

Overall, the agent’s objective is to maximize the expected discounted reward  $R$  by carefully selecting actions that adaptively balancing exploration and commitment in response to uncertainty and cost constraints of the environment.

## 3. Calibrating Agent Environment Exploration

To behave Pareto-optimally in a sequential decision-making problem, an agent must jointly compare the cost of additional exploration against the expected value of additional information to decide whether to continue exploring or to commit based on its current partial information. The value of additional information depends on reasoning over current beliefs about the underlying world state via the prior  $p(\mathbf{z} | \mathbf{x})$  and posterior  $b_t$ .

We define our LLM agent as  $\pi(a_t | \mathbf{x}, \mathcal{A}, D_\theta(\cdot), o_{0:t})$ , placing a distribution over the next action in a given state. Figure 2 (left side) shows the basic form of this agent.  $\pi$  can be

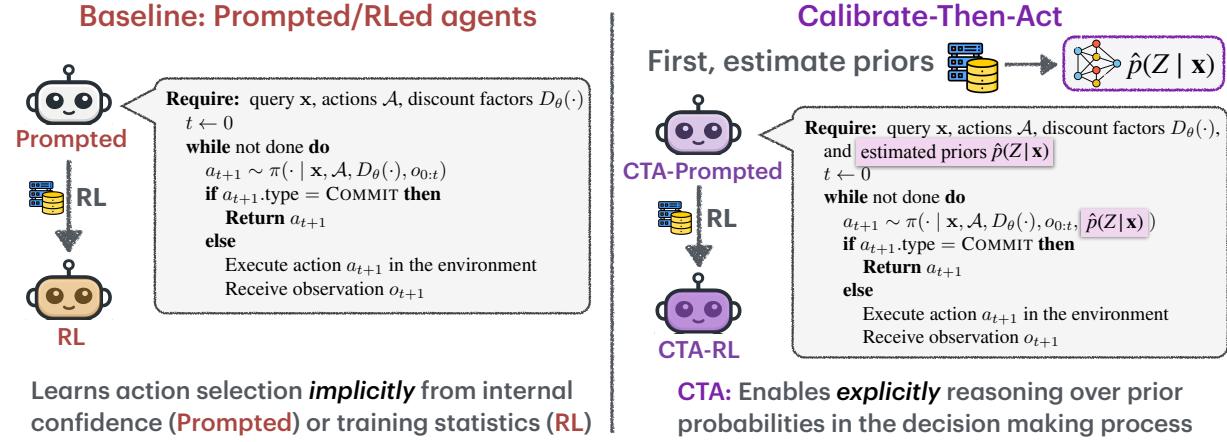


Figure 2. Standard agentic decision loop (left) and proposed method CTA with estimated priors (right). In CTA, we learn a prior estimator from training data and condition the agent on estimated  $\hat{p}$  at inference and/or training time, inducing more optimal decision making through explicit reasoning over prior probabilities.

implemented either via a prompted LLM or through a model trained with reinforcement learning. However, in practice, it is extremely difficult for  $\pi$  to learn to do the right reasoning in the environment, as we discuss later.

Our key methodological contribution is to *explicitly* provide estimates of the prior, denoted as  $\hat{p}(Z | x)$ , and optionally the posterior  $\hat{b}_i$  (not required in our settings). Figure 2 (right side) illustrates the role of a prior estimation in the agentic decision loop. This model can again be used zero-shot or trained with RL, but in either case, the prior distills key summary information from the training dataset. Conditioned on this input, the model iteratively explores the environment to acquire information until it commits to a final solution.

In Section 4, we present a proof-of-concept abstract example showing that models can reason appropriately when explicit priors are given. Then in Section 5, we introduce two more realistic tasks to study in the remainder of the paper.

## 4. Toy Setting: Pandora’s Box Problem

We begin with an abstract setting with explicit and well-defined priors to demonstrate that LLMs are capable of reasoning with the uncertainty and cost parameters to follow Pareto-optimal exploration strategies. Our only goal is to show that LLMs can gainfully employ priors in settings like this; we do **not** focus on demonstrating any learning of the priors, nor proving their use for other tasks.

### 4.1. Formalization

We consider a variant of the classic Pandora’s Box Problem with discounted reward over time (Weitzman, 1979). In this setting, a decision maker is presented with  $n$  boxes where one of them contains a prize and they need to pick

the correct one to receive the reward. They may delay the decision to inspect the boxes sequentially at a cost, but their final reward will be discounted based on the time at which the commitment is made.

Formally, the task involves a finite set of boxes  $\{z_1, z_2, \dots, z_K\}$ , among which exactly one box  $z^*$  contains a prize of value 1. This box is unknown to the agent and is drawn from a prior distribution

$$p(z_k = z^*) = p_k, \quad \sum_{k=1}^K p_k = 1.$$

At each timestep  $t$ , the model can either *verify* a box  $z_{k_t}$  of their choice to check if it contains the prize and discount their final reward by  $\gamma \in [0, 1]$ , or *commit* to a box given its current information and receive the reward

$$R = \gamma^t \cdot \mathbb{I}(z_{k_t} = z^*).$$

Intuitively, the model is rewarded for committing to the correct box, with a penalty for delayed commitment controlled by  $\gamma$ . This task can be viewed as a toy version of real-world use cases for LLM agents. For instance, if trying to identify a bug in a piece of code, we can view committing as directly telling a user where the bug is, and verifying as writing a unit test to check the correctness of the code.

Knowledge of the prior probabilities is necessary to behave optimally. Under this formulation, the optimal policy proceeds as follows. Boxes are verified in decreasing order of prior probability. A box is committed to if its posterior probability is greater than  $\gamma$ , which gives better expected value for guessing than for verifying. Otherwise, the model continues to verify. The full algorithm is shown in Algorithm 1 and proof is provided in Appendix D.

**Table 1.** Performance comparison on the Pandora’s Box task. Methods without knowledge of the priors (*Prompted-NT*, *Prompted*) fail to recover the optimal decision rule and achieve near-zero optimal match rate. With explicit priors (CTA) and thinking enabled, the agent’s decision-making behavior closely aligns with the oracle policy from Algorithm 1.

Method	Optimal Match Rate (%)	Avg. Reward
Oracle policy	100.0	0.649
PROMPTED-NT	11.0	0.441
PROMPTED	23.0	0.476
CTA-PROMPTED-NT	20.0	0.436
CTA-PROMPTED	<b>94.0</b>	<b>0.625</b>

## 4.2. LLM Agent Performance

We instantiate an LLM agent using the framework in Figure 2. Prompts for this setting are shown in Figure 9. We use Qwen3-8B as the model implementing the agent.

We evaluate on a collection of 100 examples with  $K$  set to 3 and discount factors sampled from  $[0, 0.1, 0.2, \dots, 1.0]$ . We sample priors independently from a symmetric Dirichlet distribution with concentration parameter  $\alpha = 0.5$ .

Table 1 reports the performance on the Pandora’s Box task under two metrics: average reward and optimal policy match rate, measuring whether the model’s interaction trace aligns with the oracle strategy. In PROMPTED-NT and PROMPTED, the model does not have access to prior probabilities; “NT” denotes *no thinking*, while thinking is enabled by default otherwise.

Without explicit priors or when the thinking mode is disabled, the agent exhibits near-zero optimal match rates, indicating failure to recover the optimal decision rule. In contrast, CTA-PROMPTED achieves 94.0% optimal match rate and significantly higher reward, indicating that the model is capable of reasoning about optimal exploration-commitment tradeoffs and adapting to different discount factors and prior distributions when the environment constraints are made explicit. Appendix A compares interaction traces across settings, showing how explicit reasoning with priors leads to better alignment with the oracle.

**Crucially, the presence of explicit prior information triggers the model to cast the problem in a different light, under which it can make optimal decisions.** We carry this intuition forward into our implementation of LLM agents for two more realistic settings.

## 5. Real Exploration Scenarios

In real-world LLM applications, it is typically less clear what the potential reward from an action is. Instead, models need to reason about decisions with implicit uncertainty and potential information gain against the cost of tool calling.

In this section, we study task settings that more closely reflect practical LLM deployment scenarios, where the priors over the underlying world may come from the models’ internal confidence or be derived from cues based on past experience.

### 5.1. Task QA: Knowledge QA with Optional Retrieval

We study a knowledge question answering setting in which an LLM can optionally acquire external information at a cost (Eisenstein et al., 2025). Given a factual query, the model must decide whether to rely on its parametric knowledge or defer commitment and retrieve additional evidence, trading off potential accuracy gains against latency and API costs.

**Formalization** Given a question  $x$ , a discount factor  $\gamma \in [0, 1]$ , and a retriever with quality  $p_{\text{ret}}$  (defined as the probability that the model can answer  $x$  correctly given its retrieved document), the model chooses between two actions at each timestep  $t$ : *retrieve* or *answer*. A retrieve action queries a retrieval system with the input  $x$ , while an answer action invokes the LLM to answer the question given the context so far, producing an answer  $a$ . The model receives reward  $R = \gamma^t \cdot \mathbb{I}(a = a^*)$ , where  $a^*$  is the ground-truth answer to  $x$ .

**Latent structure and prior** There are two relevant latent variables for this problem. First, we define  $p_{\text{da}} = p(a = a^* | x)$  that the model will correctly answer the question if asked directly, without retrieval. We model  $p_{\text{da}}$  with a distribution  $p(p_{\text{da}} | x) = \delta(z = k_{\text{da}}(x))$  for an estimate  $k_{\text{da}}(x)$  that the LLM returns the correct answer, where  $\delta$  is the Dirac delta function.

Additionally, to make the decision of whether to call the retriever, one needs to have information about the retriever quality  $p_{\text{ret}} = p(a' = a^* | x, c)$ , where  $c$  is the retrieved context. We similarly represent this as a delta function  $p(p_{\text{ret}} | x) = \delta(z = k_{\text{ret}})$

Under this abstraction, the oracle policy retrieves whenever the expected discounted accuracy after retrieval exceeds that of direct answering:  $p_{\text{ret}} \cdot \gamma \geq p_{\text{da}}$ .

### 5.2. Task CODE: Coding with Selective Testing

As shown in Figure 1, coding agents can choose to perform careful verification, such as running unit tests or partial execution, or directly return a solution promptly based on their current information and belief. We design a coding setting to encapsulate these considerations. Concretely, the model needs to write code to load a data file with correct formatting options and return the answer to a task query (e.g., identifying the `user_id` associated with the maximum `score`). The true file schema is not specified in the input; instead, the model may infer likely formats from filename cues and

Table 2. Unified formalization of cost-aware decision problems and their instantiations across tasks. We characterize each by latent variables  $z^*$ , a prior belief  $\pi$  over  $\mathbf{z}$ , an action space  $\mathcal{A}$  consisting of exploration and commit actions, observations  $\mathcal{O}$  revealed through exploration, costs  $\theta$  associated with the exploration actions, and a final reward  $R$  that discounts task success by incurred costs.

Setting	$z^* \in \mathcal{Z}$	$p$	$\mathcal{A}$	$\mathcal{O}$	$\theta$	$R$
UNIFIED (§2)	Latent env. state	Prior over relevant latent variables	Actions	Observations	Cost	Reward
PANDORA (§4)	$z^* \in \{z_1, \dots, z_K\}$ Prize-containing box	$\{p_k = \pi(z^* = z_k)\}$	Verify( $k$ ), Commit( $k$ )	Whether $z_k$ has prize	$\gamma$	$\gamma^T \mathbb{I}(z_k = z^*)$
QA (§5.1)	$(p_{da}, p_{ret})$ Answer probabilities with and without retrieval	$p_{da} = \delta(z = k_{da}(\mathbf{x}))$ , $p_{ret} = \delta(z = k_{ret})$	Retrieve, Answer( $a'$ )	Retrieved context	$\gamma$	$\gamma^T \mathbb{I}(a' = a^*)$
CODE (§5.2)	$\mathbf{z}^* = (z_d^*, z_q^*, z_s^*)$ $\in \mathcal{Z}_d \times \mathcal{Z}_q \times \mathcal{Z}_s$	$p(\mathbf{z}   n)$	UNIT_TEST( $f$ ), CODE( $d, q, s$ ), ANSWER( $a'$ )	Format value, stdout + stderr	$(d_u, d_c)$	$d_u^U d_c^C \mathbb{I}(a' = a^*)$

past experience. Unlike the QA setting, these priors are not known to the model a priori, but must be learned from training on this task.

**Formalization** Given a query  $\mathbf{x}$  that contains a task specification and a CSV filename  $n$ , the agent needs to write code to load the file correctly to answer the query. Specifically, each file is associated with latent formatting attributes

$$\mathbf{z} = (z_d, z_q, z_s) \in \mathcal{Z},$$

where  $z_d \in \{\text{,}, \text{;}, \text{\t}\}$  denotes the delimiter,  $z_q \in \{\text{'}, \text{"}\}$  the quote character, and  $z_s \in \{0, 1\}$  the number of skipped header rows. Without making a correct inference about  $\mathbf{z}$ , the task is not solvable.

At each timestep  $t$ , the agent selects one action  $a_t \in \mathcal{A}$  from three types: UNIT\_TEST( $f$ ), CODE( $d, q, s$ ), or ANSWER. A UNIT\_TEST probes a chosen formatting attribute  $f$  and reveals its true value. A CODE action executes code written under the agent’s current belief  $(d, q, s)$  and returns feedback via `stdout` and `stderr`, which may contain the answer or signals useful for debugging and refinement. An ANSWER action commits to a final answer  $a'$  and terminates the episode. The agent may interleave UNIT\_TEST and CODE actions in any order, and may perform multiple CODE actions to refine its solution based previous execution feedback.

Each UNIT\_TEST and CODE action incurs multiplicative discounts  $d_u$  and  $d_c$ , respectively. Upon committing at time  $T$  by ANSWER, the agent receives reward  $R = d_u^U \cdot d_c^C \cdot \mathbb{I}(a' = a^*)$ , where  $a^*$  denotes the ground-truth answer.

**Prior** The prior distribution  $p(\mathbf{z} | n)$  over formatting attributes may be inferred from conventions or past experience, or provided explicitly by a format predictor.

The prompt templates for this task are provided in Appendix E.3.

## 6. Calibrate-Then-Act: Inducing More Optimal Exploration with Explicit Priors

Section 3 established a methodological framework, which Section 4 showed can work in a toy problem. Section 5 defined realistic problems in the same sequential decision-making framework. Equipped with these ingredients, we now describe how to implement our framework from Section 3 and Figure 2 for these problems. This implementation boils down to estimating the priors described in Table 2, at which point the approach in Figure 2 can be employed.

**Prior Estimator From Model Internal Confidence** In QA, the true  $k_{da}(\mathbf{x})$  that the model can answer  $\mathbf{x}$  correctly without retrieval is not directly observable. There are several ways to obtain model predictions of confidence, including inspecting logits, probe-based methods, and verbalized confidence. For simplicity and generality, we obtain verbalized confidence as follows (Mohri & Hashimoto, 2024). Given a question  $\mathbf{x}$ , we prompt the model to produce a verbalized confidence label  $p_v(\mathbf{x})$ , and apply an isotonic regression model (Zadrozny & Elkan, 2002) ISO trained on the validation set to obtain a calibrated estimate

$$\hat{k}_{da}(\mathbf{x}) = \text{ISO}(p_v(q)).$$

This calibration step is necessary for good performance. After calibration, the expected calibration error (ECE) is reduced from 0.618 to 0.029 on PopQA dataset (Mallen et al., 2023), which is a long-tail knowledge answering benchmark. This is reflective of initial poor calibration (Guo et al., 2017; Xiong et al., 2024; Shen et al., 2024; Wang et al., 2025b) and demonstrates that rescaling can help (Desai & Durrett, 2020).

**Prior Estimator From Training Data** In the CODE task, an agent may implicitly infer file formats from prior experience, or acquire such priors through end-to-end training. We additionally study a decoupled setting in which explicit priors are provided with a format predictor based on training data. We train a filename-to-format predictor, denoted as  $\mathcal{M}_{\text{BERT}}$ , to estimate the distribution  $p(\mathbf{z} \mid n)$  from the filename. The predictor is based on a lightweight BERT-tiny encoder (4.4M parameters) (Bhargava et al., 2021; Turc et al., 2019).

Given a filename  $n$ , the model encodes the tokenized string and uses the [CLS] representation to produce three independent categorical distributions via linear heads: delimiter, quote character, and skiprows. The model is trained with a summed cross-entropy objective across the three heads for one epoch on the training split. On the validation split,  $\mathcal{M}_{\text{BERT}}$  achieves an average classification accuracy of 67% across the three formatting attributes.

After training,  $\mathcal{M}_{\text{BERT}}$  outputs marginal probabilities  $\{p(z_d \mid n), p(z_q \mid n), p(z_s \mid n)\}$ , which are provided to the agent associated with each task during RL training or test time, thereby decoupling uncertainty estimation from action selection.

**Reinforcement Learning Conditioned on Priors** In baseline RL, where the model is trained end-to-end with a reward objective and must learn a policy from priors implicitly encoded in training data, uncertainty estimation and action selection are entangled, making it difficult for the model to reliably learn cost-aware behavior. However, such a model may still benefit from RL on the training data. We therefore explore a variant of our method that optimizes the CTA-PROMPTED system with RL. This method, CTA-RL, is identical to the baseline RL method but with the CTA prompt swapped in.

## 7. Experiment Setup

**Datasets** For QA, we evaluate on PopQA (Mallen et al., 2023), a QA benchmark that covers long tail factual knowledge and benefits from retrieval. We sample 1,000 questions for evaluation and build the retriever based on **Contriever** (Izacard et al., 2022; Bajaj et al., 2016). For each question, we sample a discount factor  $\gamma \sim \mathcal{U}[0.1, 0.65]$  to study model behavior across varying retrieval costs.

For CODE, we construct a CSV-based question-answering dataset FILEREADING where filename cues provide informative signals about file formats and correct parsing requires executing code with appropriate format values. At test time, the true file format is hidden and only the filename is provided as part of the task query  $\mathbf{x}$ . FILEREADING contains 2,000 tasks, split into 1,400 training, 300 validation, and

300 test examples. For each task, we randomly sample a unit-test discount  $d_u$  from  $[0.5, 1]$  and duplicate the instance across four code discount settings  $d_c = d_u^\rho$ , with  $\rho \in \{0.5, 1.5, 2.0, 4.0\}$ , varying the relative cost of code execution while holding the task fixed. Details of the dataset construction are provided in Appendix C.

**Metrics** We evaluate the model performance across three set of metrics. (1) Exploration statistics: for QA, we report retrieval rate **Retrieve%** (the fraction of questions for which retrieval is invoked), and for CODE, we report the number of unit tests  $U$  and code attempts  $C$ . (2) Task accuracy, which measures whether the final model output matches the ground-truth answer for a task query. (3) Reward, which discounts correctness with exploration costs.

**Models and Baselines** We use Qwen3-8B (Qwen Team, 2025) as the base model for both the QA and CODE tasks. We compare the baselines against our methods (CTA) that condition the model on  $\hat{p}$  from prior estimators at inference and/or training time.

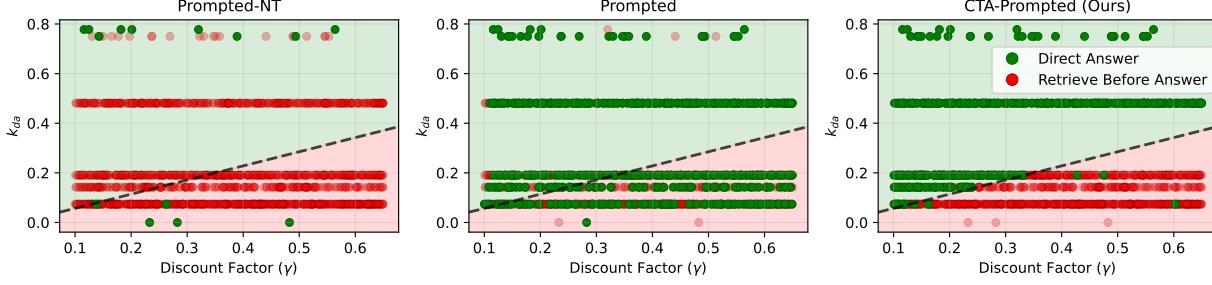
- **PROMPTED:** We prompt the base model directly with task description and query  $\mathbf{x}$ .
- **PROMPTED-NT (no thinking):** Similar to PROMPTED, but with thinking mode disabled by prepending <think></think> tags without thinking content. Unless specified with NT, we enable thinking mode by default for all other settings.
- **RL:** We fine-tune the model end-to-end using GRPO (Shao et al., 2024) with discounted reward objective, and evaluate it by prompting with the task description and query  $\mathbf{x}$ .
- **CTA-PROMPTED (Ours):** We prompt the model with estimated priors  $\hat{p}(Z \mid \mathbf{x})$  together with  $\mathbf{x}$ .
- **CTA-RL (Ours):** During both training and inference, we condition the model on both  $\mathbf{x}$  and estimated  $\hat{p}$ .

To expose both RL settings to diverse cost trade-offs in CODE, we duplicate each training instance across multiple relative cost values  $\rho \in \{0.5, 1.0, 2.0, 4.0\}$ , yielding a 4× larger RL training set than that used for training  $\mathcal{M}_{\text{BERT}}$ .

## 8. Results

We report the performance of Qwen3-8B (Qwen Team, 2025) on QA and CODE in Table 3, 4 respectively.

**CTA pushes the model towards rational decision making.** In QA, fixed one-turn strategies provide useful reference points. Directly answering without retrieval yields low accuracy (0.226), while always retrieving improves accuracy to 0.578 but incurs unnecessary cost, resulting in lower discounted reward. These baselines highlight the trade-off between correctness and cost that the model must navigate.



*Figure 3.* Model’s retrieval decision with respect to their confidence level  $k_{da}$  and retrieval discount factor  $\gamma$ . Each dot corresponds to one question: green indicates the model directly answers, and red indicates it retrieves. The dashed line marks the oracle threshold: red region retrieves, green region directly answers. Models with calibrated priors closely align with the oracle decision rule, exhibiting more cost-aware retrieval behavior.

*Table 3.* Performance on QA. We focus on discounted reward, which captures the trade-off between accuracy and retrieval cost. One-turn baselines use fixed strategies, while multi-turn agents adaptively decide when to retrieve. Across all settings, CTA-PROMPTED achieves the highest discounted reward.

Method	Retrieve %	Acc.	Reward
<i>Single-turn baselines</i>			
Never Retrieve	0.0	0.226	0.226
Always Retrieve	100.0	0.578	0.213
<i>Multi-turn agents</i>			
PROMPTED-NT	97.7	0.619	0.244
PROMPTED	61.4	0.501	0.283
CTA-PROMPTED (Ours)	65.3	0.512	<b>0.293</b>

Across settings, CTA-PROMPTED achieves the highest discounted reward by balancing task accuracy and retrieval cost.

Notably, in the multi-turn setup, where the model gets to decide when to retrieve, behavior differs substantially across settings. Figure 3 shows a scatter plot visualizing the models’ decisions. The x-axis represents the discount factor for a retrieve action, and the y-axis represents the estimated prior  $k_{da}$ . The space is colored based on the optimal decision and the dashed line marks the oracle threshold: red region retrieves, green region directly answers. Each dot corresponds to the agent’s decision for a query  $x$ : red indicates that the agent chooses to call retriever before answering, while green indicates direct answering based on their parametric knowledge. In the left subplot, the model with thinking mode disabled (PROMPTED-NT) almost always retrieves before answering (98.4% retrieval rate), leading to suboptimal reward. Enabling thinking mode reduces retrieval by 35.0% and improves overall reward, demonstrating that agent is more aware of retrieval costs in its decision making. While PROMPTED (middle) exhibits a largely unstructured decision-making pattern, the agent’s decisions in CTA-PROMPTED (right) are more closely aligned with the oracle strategy and display a clear decision boundary with

*Table 4.* Performance on CODE, averaged across relative unit-test and code-execution cost ratios  $\rho \in \{0.5, 1.0, 2.0, 4.0\}$ . We report the average number of turns, unit-test calls (U), code executions (C), accuracy, and discounted reward.

Method	# Turns	U	C	Acc.	Reward
<i>Without Training</i>					
PROMPTED	3.62	2.67	1.42	0.958	0.229
CTA-PROMPTED (Ours)	3.47	2.51	1.41	0.945	0.240
<i>With RL Training</i>					
RL	3.51	2.13	1.39	0.997	0.259
CTA-RL (Ours)	3.46	1.98	1.46	0.991	<b>0.268</b>

respect to both confidence and retrieval costs.

### CTA-RL generalizes better in domain than baseline end-to-end RL.

Table 4 shows the model performance on CODE aggregated across different cost settings. In the with RL training settings, both RL and CTA-RL have access to the same set of training data which encodes the distribution of format given filenames implicitly. The agent achieves a discounted reward of 0.259 by training end-to-end (RL) with the discounted reward as objective, while conditioning the training on explicit estimated priors (CTA-RL) further improves by 3.5% with a overall reward of 0.268. This shows that incorporating estimated priors help the model generalize better on unseen test data.

**Conditioning the training on estimated priors reinforces the adaptive decision reasoning and induces Pareto-optimal behavior.** We examine the agents’ decision making behavior across varying cost regimes. We categorize the tasks in CODE by the relative cost of coding attempts against unit tests:  $\rho = \log d_c / \log d_u$ . For example,  $\rho = 3$  means one code attempt costs as much as three unit tests, making tests more optimal. When  $\rho$  is relatively small, an agent should behave more aggressively by attempting full code execution early.

Figure 4 reveals systematic differences in decision making across different methods. Each stacked bar represents the

collection of tasks with a specific  $\rho$ , and each color represents the proportion of action trace pattern in the collection. The reward is labeled above each bar, and the percentage of “guess-and-go”  $x\%$  (attempting code without any preceding unit tests) are also labeled in each bar. From the left column of Figure 4, while RL improves the discounted reward of the agent compared to its non-training counterpart PROMPTED, its decision making behavior collapses to a static policy which always triggers unit tests before the first code attempt (0% “guess-and-go”). RL without prior cannot internalize the structure of the data with end-to-end training and instead defaults to suboptimal exploration policy.

In contrast, the subplots on the right shows that CTA-PROMPTED, which conditions the agentic decision-making on estimated priors without training, already exhibits adaptive behavior in response to costs. Specifically, they become more conservative with higher  $\rho$ . After training, such adaptive behavior remains pronounced in CTA-RL. As a result, even with imperfect prior estimates, CTA-RL consistently perform better than RL baseline across cost regimes.

To demonstrate the Pareto-optimality of CTA-RL, we plot the  $\Delta$  Reward against the naive baseline “3\*Tests→Code” for each methods in Figure 5. While RL only shows advantage with large  $\rho$ , and static policy such as **Code first** only performs well with small  $\rho$ , our method **CTA-RL** stays at the Pareto-optimal front across  $\rho$  values. This suggests that conditioning training with explicit priors effectively reinforces the adaptive decision reasoning.

## 9. Related Work

**Decision making under incomplete information** LLMs are increasingly being applied to tasks with incomplete information, arising from underspecified user queries (Cole et al., 2023; Zhang et al., 2025; Zhang & Choi, 2025; Li et al., 2025; Shaikh et al., 2025), ambiguity (Min et al., 2020; Choi et al., 2025; Deng et al., 2025), and partially observed environments (Wong et al., 2023; Lin et al., 2024; Dwaracherla et al., 2024; Chen et al., 2025a; Grand et al., 2025). To resolve uncertainty, models often need to ask clarifying questions (Rao & Daumé III, 2018; Handa et al., 2024; Lalai et al., 2025), query the environments (Charikar et al., 2002; Nadimpalli et al., 2025; Monea et al., 2024), or engage in collaboration (Wu et al., 2025; Chen et al., 2025b). While prior work has typically focus on training or prompting strategies, we show that the models can abstractly reason about the optimal solution when provided with explicit priors, which we use to induce such reasoning.

**Agents in cost-aware deployment** LLM-based agents are increasingly deployed in real-world settings that require multi-step reasoning and tool use, including interactive coding (Tang et al., 2024; Zhou et al., 2025; Wang et al., 2025c;

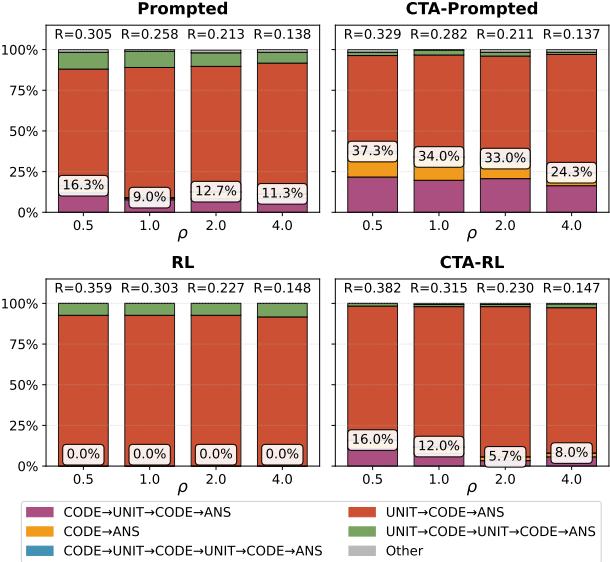


Figure 4. Action pattern distribution for prompting and RL-trained agents, with and without calibrated priors, across relative cost parameters  $\rho$ . Each stacked bar shows the proportion of decision traces corresponding to different action patterns, with the reward  $R$  labeled above. Annotated percentages indicate the fraction of tasks where the agent attempts code execution before any unit tests.

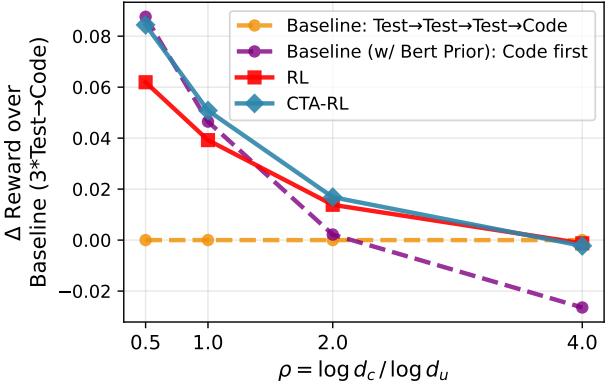


Figure 5. Pareto frontier of average reward under varying costs. Static strategies (test-first or code-first) achieve high reward only in limited regimes, whereas CTA-RL with estimated priors consistently attains Pareto-optimal performance across cost settings.

Jain et al., 2025), planning (Zhou et al., 2024; Liu et al., 2025), question answering (Yao et al., 2023; Eisenstein et al., 2025), and scientific research (Schwettmann et al., 2023; GX-Chen et al., 2025; Khan et al., 2025; Abaskohi et al., 2025; Agarwal et al., 2025). While tool use expands agents’ capabilities and reliability, interacting with external environments often incurs latency (Guan et al., 2025), resources (Damani et al., 2024), and overhead that can negatively affect user experience (Elfleet & Chollet, 2024; Herlihy et al., 2024). In response, several lines of work have emerged to study agent behavior under explicit cost constraints. Liu

et al. (2025) introduce a cost-centric benchmark for evaluating agents’ tool-planning abilities. Wang et al. (2025a); Gul et al. (2025); Wang et al. (2025b); Lin et al. (2025) study how models can reduce unnecessary retrieval or tool use while maintaining answer quality, for example through abstention, selective search, or efficiency-oriented action policies. Berant et al. (2025) train steerable clarification policies that adapt to cost coefficients. An underexplored aspect of efficient exploration is the joint treatment of uncertainty priors and cost constraints, which together determine Pareto-optimal decisions. We propose a unified framework for interactive agentic tasks and show that calibrated priors are key to inducing appropriate decision-making in LLM agents.

## 10. Conclusion

This paper presents a method for having LLMs balance uncertainty-cost tradeoffs in their environment interaction. By presenting an LLM with priors over unobserved features of the environment, the LLM can successfully reason about Pareto-optimal behavior and navigate action costs effectively. This work illustrates new ways of inducing agents to think optimally, and suggests that meta-level information (priors about capabilities) may have a role to play in shaping agent policies.

## Acknowledgments

This work was supported by the NSF under Cooperative Agreement 2421782 and the Simons Foundation grant MPS-AI-00010515 awarded to the NSF-Simons AI Institute for Cosmic Origins — CosmicAI, <https://www.cosmicaid.org/>. This was also partially supported by NSF CAREER Award IIS-2145280, NSF grant IIS-2433071, by the Sloan Foundation, and by grants from Amazon and Open Philanthropy. This research has been supported by computing support on the Vista GPU Cluster through the Center for Generative AI (CGAI) and the Texas Advanced Computing Center (TACC) at the University of Texas at Austin, through the Torch cluster at NYU, and through a compute grant from NVIDIA.

## Impact Statement

This paper presents a method for, broadly speaking, improving the cost-benefit tradeoffs of LLM agents. Although this is not yet integrated into production agent systems, we envision that this approach or one derived from it could be, which would ideally lead to cost savings and increased efficiency. We do not foresee specific drawbacks of this approach relative to other advancements in LLMs, agents, and machine learning more broadly. Potential broad drawbacks are inherited from the drawbacks of capability advance-

ments for LLMs and agents more broadly, such as enabling the further propagation of AI technology in society.

## References

- Abaskohi, A., Chen, T., Munoz-M’armol, M., Fox, C., Ramesh, A. V., Marcotte, E., Lù, X. H., Chapados, N., Gella, S., Pal, C., Drouin, A., and Laradji, I. H. DRBench: A Realistic Benchmark for Enterprise Deep Research. *ArXiv*, abs/2510.00172, 2025. URL <https://api.semanticscholar.org/CorpusID:281705844>.
- Agarwal, D., Majumder, B. P., Adamson, R., Chakravorty, M., Gavireddy, S. R., Parashar, A., Surana, H., Mishra, B. D., McCallum, A., Sabharwal, A., et al. Open-ended Scientific Discovery via Bayesian Surprise. *arXiv preprint arXiv:2507.00310*, 2025.
- Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.
- Berant, J., Chen, M., Fisch, A., Aghajani, R., Huot, F., Lapata, M., and Eisenstein, J. Learning steerable clarification policies with collaborative self-play. *arXiv preprint arXiv:2512.04068*, 2025.
- Bhargava, P., Drozd, A., and Rogers, A. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics, 2021.
- Charikar, M., Fagin, R., Guruswami, V., Kleinberg, J., Raghavan, P., and Sahai, A. Query strategies for priced information. *Journal of Computer and System Sciences*, 64(4):785–819, 2002. ISSN 0022-0000. doi: <https://doi.org/10.1006/jcss.2002.1828>. URL <https://www.sciencedirect.com/science/article/pii/S0022000002918283>.
- Chen, S., Chen, X., Huang, Y., Xie, R., and Dhingra, B. When greedy wins: Emergent exploitation bias in meta-bandit lilm training. *ArXiv*, abs/2509.24923, 2025a. URL <https://api.semanticscholar.org/CorpusID:281674231>.
- Chen, W., Yuan, J., Qian, C., Yang, C., Liu, Z., and Sun, M. Optima: Optimizing effectiveness and efficiency for LLM-based multi-agent system. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11534–11557, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.601. URL <https://aclanthology.org/2025.findings-acl.601/>.

- Choi, J., Bansal, M., and Stengel-Eskin, E. Language models identify ambiguities and exploit loopholes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32991–33006, 2025.
- Cole, J. R., Zhang, M. J., Gillick, D., Eisenschlos, J. M., Dhingra, B., and Eisenstein, J. Selectively answering ambiguous questions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=x2W2dKdNI8>.
- Damani, M., Shenfeld, I., Peng, A., Bobu, A., and Andreas, J. Learning how hard to think: Input-adaptive allocation of lm computation. *ArXiv*, abs/2410.04707, 2024. URL <https://api.semanticscholar.org/CorpusID:273186996>.
- Deng, M., Huang, L., Fan, Y., Zhang, J., Ren, F., Bai, J., Yang, F., Miao, D., Yu, Z., Wu, Y., Zhang, Y., Teng, F., Wan, Y., Hu, S., Li, Y., Jin, X., Hu, C., Li, H., Fu, Q., Zhong, T., Wang, X., Tang, X., Tang, N., Wu, C., and Luo, Y. InteractComp: Evaluating Search Agents With Ambiguous Queries. *ArXiv*, abs/2510.24668, 2025. URL <https://api.semanticscholar.org/CorpusID:282401680>.
- Desai, S. and Durrett, G. Calibration of pre-trained transformers. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.21. URL <https://aclanthology.org/2020.emnlp-main.21/>.
- Dwaracherla, V., Asghari, S. M., Hao, B., and Van Roy, B. Efficient exploration for LLMs. *arXiv preprint arXiv:2402.00396*, 2024.
- Eisenstein, J., Aghajani, R., Fisch, A., Dua, D., Huot, F., Lapata, M., Zayats, V., and Berant, J. Don’t lie to your friends: Learning what you know from collaborative self-play. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=2vDJiGUfhV>.
- Elfleet, M. and Chollet, M. Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM-Powered Embodied Conversational Agents in Virtual Reality. In *IVA*, pp. 12:1–12:9, 2024. URL <https://doi.org/10.1145/3652988.3673965>.
- Grand, G., Pepe, V., Andreas, J., and Tenenbaum, J. B. Shoot First, Ask Questions Later? Building Rational Agents that Explore and Act Like People. *arXiv preprint arXiv:2510.20886*, 2025.
- Guan, Y., Lan, Q., Fei, S., Ding, D., Acharya, D., Wang, C., Wang, W. Y., and Hua, W. Dynamic speculative agent planning. *arXiv preprint arXiv:2509.01920*, 2025.
- Gul, M. O., Cardie, C., and Goyal, T. Pay-per-search models are abstention models. *arXiv preprint arXiv:2510.01152*, 2025.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.
- Gupta, R., Hartford, J., and Liu, B. LLMs for experiment design in scientific domains: Are we there yet? In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025. URL <https://openreview.net/forum?id=dIEeOwrmOe>.
- GX-Chen, A., Lin, D., Samiei, M., Precup, D., Richards, B. A., Fergus, R., and Marino, K. Language agents mirror human causal reasoning biases. how can we help them think like scientists? *ArXiv*, abs/2505.09614, 2025. URL <https://api.semanticscholar.org/CorpusID:278602122>.
- Handa, K., Gal, Y., Pavlick, E., Goodman, N., Andreas, J., Tamkin, A., and Li, B. Z. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*, 2024.
- Hennig, L., Tornede, T., and Lindauer, M. Towards leveraging AutoML for sustainable deep learning: A multi-objective HPO approach on deep shift neural networks. *arXiv preprint arXiv:2404.01965*, 2024.
- Herlihy, C., Neville, J., Schnabel, T., and Swaminathan, A. On Overcoming Miscalibrated Conversational Priors in LLM-based ChatBots. In *Uncertainty in Artificial Intelligence*, pp. 1599–1620. PMLR, 2024.
- Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=jKN1pXi7b0>.
- Jain, A. K., Gonzalez-Pumariega, G., Chen, W., Rush, A. M., Zhao, W., and Choudhury, S. Multi-turn code generation through single-step rewards. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=aJeLhLcsh0>.
- Ji, S. and Carin, L. Cost-sensitive feature acquisition and classification. *Pattern Recognition*, 40(5):1474–1485, 2007.

- Kärkkäinen, K., Kachuee, M., Goldstein, O., and Sarafzadeh, M. Cost-sensitive feature-value acquisition using feature relevance. *arXiv preprint arXiv:1912.08281*, 2019.
- Khan, Z., Prasad, A., Stengel-Eskin, E., Cho, J., and Bansal, M. One life to learn: Inferring symbolic world models for stochastic environments from unguided exploration. *ArXiv*, abs/2510.12088, 2025. URL <https://api.semanticscholar.org/CorpusID:282064346>.
- Lalai, H. N., Shah, R. S., Pei, J., Varma, S., Wang, Y.-C., and Emami, A. The world according to LLMs: How geographic origin influences LLMs' entity deduction capabilities. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=hJtvCfDfs1>.
- Li, B. Z., Kim, B., and Wang, Z. Questbench: Can LLMs ask the right question to acquire information in reasoning tasks? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL <https://openreview.net/forum?id=gpwA9aZLTZ>.
- Li, Y. and Oliva, J. Towards cost sensitive decision making. In Li, Y., Mandt, S., Agrawal, S., and Khan, E. (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3601–3609. PMLR, 03–05 May 2025. URL <https://proceedings.mlr.press/v258/li25h.html>.
- Lin, J., Tomlin, N., Andreas, J., and Eisner, J. Decision-oriented dialogue for human-AI collaboration. *Transactions of the Association for Computational Linguistics*, 12:892–911, 2024. doi: 10.1162/tacl\_a.00679. URL <https://aclanthology.org/2024.tacl-1.50/>.
- Lin, T.-H., Chen, W.-L., Li, C.-A., yi Lee, H., Chen, Y.-N., and Meng, Y. AdaSearch: Balancing Parametric Knowledge and Search in Large Language Models via Reinforcement Learning. *ArXiv*, abs/2512.16883, 2025. URL <https://api.semanticscholar.org/CorpusID:283933928>.
- Liu, J., Qian, C., Su, Z., Zong, Q., Huang, S., He, B., and Fung, Y. R. CostBench: Evaluating Multi-Turn Cost-Optimal Planning and Adaptation in Dynamic Environments for LLM Tool-Use Agents. *arXiv preprint arXiv:2511.02734*, 2025.
- Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL <https://aclanthology.org/2023.acl-long.546/>.
- Min, S., Michael, J., Hajishirzi, H., and Zettlemoyer, L. AmbigQA: Answering ambiguous open-domain questions. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL <https://aclanthology.org/2020.emnlp-main.466/>.
- Mohri, C. and Hashimoto, T. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 36029–36047, 2024.
- Monea, G., Bosselut, A., Brantley, K., and Artzi, Y. LLMs Are In-Context Bandit Reinforcement Learners. *arXiv preprint arXiv:2410.05362*, 2024.
- Nadimpalli, S., Qiao, M., and Rubinfeld, R. No price tags? no problem: Query strategies for unpriced information. *ArXiv*, abs/2511.06170, 2025. URL <https://api.semanticscholar.org/CorpusID:282911938>.
- Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Rao, S. and Daumé III, H. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In Gurevych, I. and Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2737–2746, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1255. URL <https://aclanthology.org/P18-1255/>.
- Schwettmann, S., Shaham, T., Materzynska, J., Chowdhury, N., Li, S., Andreas, J., Bau, D., and Torralba, A. FIND: A Function Description Benchmark for Evaluating Interpretability Methods. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 75688–75715. Curran Associates, Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/ef0164c1112f56246224af540857348f-Paper-Datasets\\_and\\_Benchmarks.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/ef0164c1112f56246224af540857348f-Paper-Datasets_and_Benchmarks.pdf).

- Shaikh, O., Mozannar, H., Bansal, G., Fourney, A., and Horvitz, E. Navigating rifts in human-LLM grounding: Study and benchmark. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20832–20847, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1016. URL <https://aclanthology.org/2025.acl-long.1016/>.
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Shen, Y., Zhu, X., and Chen, L. SMARTCAL: An approach to self-aware tool-use evaluation and calibration. *arXiv preprint arXiv:2412.12151*, 2024.
- Tang, H., Hu, K., Zhou, J. P., Zhong, S., Zheng, W.-L., Si, X., and Ellis, K. Code Repair with LLMs gives an Exploration-Exploitation Tradeoff. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 117954–117996. Curran Associates, Inc., 2024. doi: 10.5220/079017-3746. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/d5c56ec4f69c9a473089b16000d3f8cd-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/d5c56ec4f69c9a473089b16000d3f8cd-Paper-Conference.pdf).
- Turc, I., Chang, M., Lee, K., and Toutanova, K. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962, 2019. URL <http://arxiv.org/abs/1908.08962>.
- Wang, H., Qian, C., Zhong, W., Chen, X., Qiu, J., Huang, S., Jin, B., Wang, M., Wong, K.-F., and Ji, H. Acting Less is Reasoning More! Teaching Model to Act Efficiently. *arXiv preprint arXiv:2504.14870*, 2025a.
- Wang, H., Xue, B., Zhou, B., Zhang, T., Wang, C., Wang, H., Chen, G., and Wong, K.-F. Self-DC: When to reason and when to act? self divide-and-conquer for compositional unknown questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6510–6525, 2025b.
- Wang, Y., Zhang, Y., Qin, Z., Zhi, C., Li, B., Huang, F., Li, Y., and Deng, S. ExploraCoder: Advancing code generation for multiple unseen APIs via planning and chained exploration. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18124–18145, Vienna, Austria, July 2025c. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.887. URL <https://aclanthology.org/2025.acl-long.887/>.
- Wang, Z., Wang, K., Wang, Q., Zhang, P., Li, L., Yang, Z., Yu, K., Nguyen, M. N., Liu, L., Gotlieb, E., Lam, M., Lu, Y., Cho, K., Wu, J., Li, F.-F., Wang, L., Choi, Y., and Li, M. RAGEN: Understanding Self-Evolution in LLM Agents via Multi-Turn Reinforcement Learning. *ArXiv*, abs/2504.20073, 2025d. URL <https://api.semanticscholar.org/CorpusID:278170861>.
- Weitzman, M. L. Optimal search for the best alternative. *Econometrica*, 47(3):641–654, 1979. ISSN 00129682, 14680262. URL <http://www.jstor.org/stable/1910412>.
- Wong, L. S., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., and Tenenbaum, J. B. From word models to world models: Translating from natural language to the probabilistic language of thought. *ArXiv*, abs/2306.12672, 2023. URL <https://api.semanticscholar.org/CorpusID:259224900>.
- Wu, S., Galley, M., Peng, B., Cheng, H., Li, G., Dou, Y., Cai, W., Zou, J., Leskovec, J., and Gao, J. CollabLLM: From passive responders to active collaborators. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=DmH4HHVb3y>.
- Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=gjeQKFxFpZ>.
- Xu, Y., Chen, Z., and Wen, Z. EcoTune: Token-efficient multi-fidelity hyperparameter optimization for large language model inference. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7735–7745, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.394. URL [https://aclanthology.org/2025.emnlp-main.394/](https://aclanthology.org/2025.emnlp-main.394).

- Yang, H., Yue, S., and He, Y. Auto-GPT for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*, 2023.
- Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., and Press, O. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 50528–50652. Curran Associates, Inc., 2024. doi: 10.52202/079017-1601. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/5a7c947568c1b1328ccc5230172e1e7c-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/5a7c947568c1b1328ccc5230172e1e7c-Paper-Conference.pdf).
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023. URL [https://openreview.net/forum?id=WE\\_vluYUL-X](https://openreview.net/forum?id=WE_vluYUL-X).
- Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.
- Zhang, M. J. and Choi, E. Clarify when necessary: Resolving ambiguity through interaction with LMs. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5526–5543, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.306. URL <https://aclanthology.org/2025.findings-naacl.306/>.
- Zhang, M. J., Knox, W. B., and Choi, E. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=cwuSAR7EKd>.
- Zhou, L., Shi, J., Gao, J., and Wang, D. Credit-budgeted ICPC-style coding: When LLM agents must pay for every decision. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*, 2025. URL <https://openreview.net/forum?id=JEdeMSKvbT>.
- Zhou, Y., Zanette, A., Pan, J., Levine, S., and Kumar, A. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=b6rA0kAHT1>.

## Case Study: Pandora’s Box Problem, CTA-PROMPTED-NT

### Problem Parameters

- Bag Labels: A, B, C
- Prior Probabilities: A: 0.04, B: 0.68, C: 0.28
- Discount Factor r: 0.2

### Interaction Trace

*Turn 0*

**User:** {Instruction instantiated with the problem parameters}

**Assistant:** VERIFY A

*Turn 1*

**User:** The verification result is: NO, A is incorrect. Given this, please provide your next action.

**Assistant:** VERIFY B

*Turn 2*

**User:** The verification result is: NO, B is incorrect. Given this, please provide your next action.

**Assistant:** GUESS C

*Figure 6.* Example interaction trace on a 3-bag Pandora’s Box instance with priors (0.04, 0.68, 0.28) and discount factor  $\gamma = 0.2$  with thinking mode **disabled**. In this setting, the model explores all bags before committing and follows a suboptimal verification order, rather than prioritizing the highest-probability option.

## A. Qualitative trace analysis of Pandora’s Box Problem

We present representative interaction traces from three settings: CTA-PROMPTED-NT, PROMPTED, and CTA-PROMPTED. In CTA-PROMPTED-NT (with thinking mode disabled; Figure 6), the model does not appear to compare the expected value of additional information against the exploration cost. As a result, it tends to verify all options before committing, regardless of the prior distribution, leading to unnecessary exploration. In PROMPTED (Figure 7), the model does not have access to the prior probabilities. Lacking calibrated uncertainty information, it effectively operates under an implicit uniform prior and consequently follows a suboptimal verification strategy.

In contrast, in CTA-PROMPTED (Figure 8), the model is provided with prior probabilities and has thinking mode enabled. In this setting, it explicitly reasons about the trade-off between expected reward and exploration cost by comparing the value of immediate commitment with the discounted value of further verification. The resulting behavior aligns with the oracle policy.

These qualitative examples illustrate how explicit prior information induces the correct reasoning of the model over value of additional information against action cost and making the optimal decision accordingly.

## B. Experiment details for QA

For task QA, we evaluate on the POPQA dataset. We build the retriever based on CONTRIEVER (Izacard et al., 2022; Bajaj et al., 2016). The retriever quality  $p_{\text{ret}}$ , defined as the probability that the model can answer a question correctly when conditioned on the document retrieved from the retriever, is estimated on a validation set and provided to the as a part of  $\mathbf{x}$  LLM at inference time. Note that  $p_{\text{ret}}$  depends on the retriever and the agent being used, but not on individual questions.

## C. Dataset Construction Details for Task: Coding with Selective Testing

This appendix describes the oracle filename-to-format model and the procedure used to generate the coding task dataset FILEREADING.

Each task consists of a structured filename, a CSV file generated under a latent formatting configuration, and an associated query. Filenames combine a small set of indicative tokens (e.g., `_eu`, `_tab`, `_sas`, `_cn`) with additional irrelevant strings, inducing a prior over possible parsing configurations.

To capture scenarios in which filename cues provide informative signals about file formats, we construct a synthetic dataset of CSV-based question-answering tasks with structured filenames. Each filename is generated by combining a small set of

## Case Study: Pandora's Box Problem, PROMPTED

### Problem Parameters

- Bag Labels: A, B, C
- Prior Probabilities: A: 0.04, B: 0.68, C: 0.28
- Discount Factor r: 0.2

### Interaction Trace

*Turn 0*

**User:** {Instruction instantiated with the problem parameters, **but the prior probabilities are not provided to the model**}

**Assistant:** <think> Okay, let's see. I need to figure out the best action here. So, there are three bags: A, B, and C. Only one of them has the prize worth 1. The discount factor is 0.2. At each step, I can either verify a bag to get information (which takes a timestep) or guess a bag to end the game and get the reward, which is discounted by  $r^t$  where t is the timestep.

Starting at t=0. If I guess now, the reward would be  $1 * 0.2^0 = 1$ . But if I guess wrong, I get nothing. Since there's a 1/3 chance of being correct, the expected reward for guessing at t=0 is  $(1/3) * 1 + (2/3) * 0 = 1/3 \approx 0.333$ .

Alternatively, if I verify a bag. Let's say I verify bag A. It takes one timestep (t=0 to t=1). If the verification says "YES", then I know A has the prize. Then, I can guess A at t=1, getting  $1 * 0.2^1 = 0.2$ . If the verification says "NO", then the prize is in B or C. Then, I have two bags left. But I need to think about the next steps.

Wait, but verifying takes a timestep. So, if I verify at t=0, then the next step is t=1. Let's think about the expected value of verifying. Suppose I verify A at t=0. Then, the result is YES with 1/3, NO with 2/3.

If YES: Then at t=1, I can guess A and get 0.2. If NO: Then the prize is in B or C. Now, at t=1, I have two bags left. If I guess at t=1, the expected reward would be  $(1/2) * 0.2^1 + (1/2) * 0 = 0.1$ . Alternatively, I can verify another bag. Let's say I verify B at t=1. Then, if it's YES, I can guess B at t=2 (reward  $0.2^2 = 0.04$ ), else verify C at t=2, etc. But this might take longer.

Alternatively, maybe it's better to guess at t=1 if I have two bags left. The expected value would be 0.1 as above. But if I verify, maybe I can get more information. Let me calculate the expected value of verifying A at t=0.

So, if I verify A at t=0, then at t=1, if it's YES (prob 1/3), I get 0.2. If NO (prob 2/3), then I have two bags left. Then, the expected value from that point is the maximum between guessing or verifying again. Let's assume that the optimal strategy is to guess when there are two bags left. Then, the expected value would be 0.1 as before. So the total expected value of verifying A at t=0 is  $(1/3) * 0.2 + (2/3) * 0.1 = 0.2/3 + 0.2/3 = 0.4/3 \approx 0.133$ . Which is less than the 0.333 from guessing at t=0. So, it seems better to guess at t=0.

Wait, but maybe I'm missing something. Because if I verify A at t=0, and it's NO, then at t=1, I can verify another bag. Let's say I verify B at t=1. Then, if it's YES (prob 1/2), then I can guess B at t=2 (reward  $0.2^2 = 0.04$ ). If NO (prob 1/2), then the prize is in C, so I can guess C at t=2, getting 0.04. So the expected value here would be 0.04. So, the total expected value if I verify A at t=0, then B at t=1, would be  $(1/3) * 0.2 + (2/3) * [(1/2) * 0.04 + (1/2) * 0.04] = (1/3) * 0.2 + (2/3) * 0.04 = 0.0667 + 0.0267 \approx 0.0933$ . Which is worse than the 0.333 from guessing at t=0. So, in that case, guessing at t=0 is better.

Alternatively, maybe there's a better strategy. Let me think again. Suppose I verify one bag, and if it's no, then verify another. But with discounting, the later rewards are smaller. So, perhaps the optimal strategy is to guess immediately. Because the expected reward from guessing at t=0 is 1/3, which is higher than the expected value of verifying and then proceeding. So, maybe the best action is to GUESS at t=0.

But wait, what if I verify one bag and then guess? Let's say I verify A at t=0. If it's YES, then I can guess at t=1 (0.2). If it's NO, then I have two bags left. Then, if I guess at t=1, the expected reward is 0.1. So the total expected value is  $(1/3) * 0.2 + (2/3) * 0.1 = 0.2/3 + 0.2/3 = 0.4/3 \approx 0.133$ , which is less than 0.333. So, indeed, guessing at t=0 is better.

Alternatively, if I verify two bags, but that would take more steps. For example, verify A at t=0, then B at t=1. If both are NO, then the prize is in C. Then at t=2, guess C and get  $0.2^2 = 0.04$ . So the expected value would be (probability that A is NO and B is NO) \* 0.04. The probability that A is NO is 2/3, then B is NO is 1/2, so  $2/3 * 1/2 = 1/3$ . So expected value is  $1/3 * 0.04 = 0.0133$ . Which is worse than guessing at t=0. So, that's even worse.

So, the conclusion is that guessing at t=0 gives a higher expected value than any strategy involving verification. Therefore, the optimal action is to GUESS one of the bags. But since all bags are symmetric, it doesn't matter which one. So, the answer should be GUESS A, B, or C. But the user probably expects me to choose one. Let me check if there's any other consideration. For example, maybe the discount factor is 0.2, which is quite low. So, waiting to get more information might not be worth it. So, the best action is to GUESS now. </think>

GUESS A

Figure 7. Example interaction trace on a 3-bag Pandora's Box instance with priors (0.04, 0.68, 0.28) and discount factor  $\gamma = 0.2$ , where the model is not given access to the prior probabilities. In this setting, the model implicitly treats the bags as equally likely and follows a suboptimal strategy that deviates from the optimal policy.

## Case Study: Pandora's Box Problem, CTA-PROMPTED

### Problem Parameters

- Bag Labels: A, B, C
- Prior Probabilities: A: 0.04, B: 0.68, C: 0.28
- Discount Factor r: 0.2

### Interaction Trace

*Turn 0*

**User:** {Instruction instantiated with the problem parameters}

**Assistant:**

<think> Okay, let's see. I need to figure out whether to verify or guess in this scenario. The problem has three bags: A, B, and C with prior probabilities 0.04, 0.68, and 0.28 respectively. The discount factor is 0.2.

First, I should remember that each VERIFY action costs a timestep, and GUESS ends the game immediately. The reward for a correct guess is discounted by  $r^t$ , where t is the timestep. So, if I guess at  $t = 0$ , the discount is 1 (since  $0.2^0 = 1$ ). But if I verify, I spend a timestep and then maybe have more information to make a better decision.

The key here is to balance between getting more information (which might lead to a higher reward) and the risk of waiting too long, which could reduce the discounted reward.

Let me think about the expected value of guessing immediately versus verifying.

If I guess at  $t = 0$ , the expected reward would be the sum of the probabilities of each bag multiplied by the discount factor. Wait, no. Wait, the reward is 1 if the guessed bag has the prize, 0 otherwise. So the expected reward for guessing at  $t=0$  would be the probability of the chosen bag having the prize. **But since I can choose any bag to guess, I should pick the one with the highest prior probability.**

Looking at the priors: B is 0.68, C is 0.28, A is 0.04. So if I guess B, the expected reward is  $0.68 * 1$  (since if it's correct, I get 1, else 0). But since the discount is  $0.2^0 = 1$ , the expected reward is 0.68.

Alternatively, if I verify a bag, I can get more information. Let's say I verify B first. If I verify B, then I spend a timestep ( $t=1$ ). Then, depending on the result, I can decide whether to guess or verify again.

...

But the problem is that the discount factor is 0.2. So, each time you wait, the reward is discounted. So, the longer you wait, the less valuable the reward becomes.

So, the question is **whether the information gained from verifying is worth the cost of waiting.**

Let me think about the expected value of verifying versus guessing.

**If I guess at  $t=0$ , the expected reward is the maximum prior probability, which is 0.68.**

**If I verify a bag, say B, then I spend one time step ( $t=1$ ). Let's calculate the expected reward if I verify B.**

**After verifying B, there are two possibilities:**

1. The result is YES (probability 0.68). Then, I know B has the prize. So, I can guess B at  $t=1$ , which would give me a reward of  $1 * 0.2^1 = 0.2$ .

2. The result is NO (probability  $1 - 0.68 = 0.32$ ). Then, B does not have the prize. Now, I have to decide whether to guess or verify another bag.

But if I get NO on B, then the remaining bags are A and C. Their probabilities would be updated based on the fact that B is not the correct one. Wait, but the problem says that the bags are independent. Wait, actually, the problem states that the prior probabilities are for each bag. So, if B is not the correct one, then the prize is in A or C. However, the original priors are A:0.04, B:0.68, C:0.28. But if we know that B is not the correct one, then the total probability of the prize being in A or C is  $0.04 + 0.28 = 0.32$ .

**So, the posterior probabilities would be A:  $0.04/0.32 = 0.125$ , and C:  $0.28/0.32 = 0.875$ .**

So, in this case, if we get a NO on B, then the optimal choice would be to guess C, which has a 0.875 chance of being correct. Then, the reward would be  $0.875 * 0.2^1$  (since  $t=1$ ).

But wait, after verifying B at  $t=0$ , we spend a timestep ( $t=1$ ), and then we can guess. So, if we get NO on B, then we can guess C at  $t=1$ , which gives us  $0.875 * 0.2 = 0.175$ .

Alternatively, if we get YES on B, then we can guess B at  $t=1$ , which gives  $1 * 0.2 = 0.2$ .

**So, the expected reward of verifying B is:**

$$0.68 * 0.2 + 0.32 * 0.175 = 0.136 + 0.056 = 0.192.$$

Compare that to the expected reward of guessing immediately (0.68).

**So, 0.68 is higher than 0.192. Therefore, verifying B is worse than guessing immediately.**

</think>

Action: GUESS B

Figure 8. Example model reasoning trace on a 3-bag Pandora's Box instance with priors (0.04, 0.68, 0.28) and discount factor  $\gamma = 0.2$ . The model explicitly compares the expected value of immediate guessing versus verification and then chooses to guess B immediately, which is the optimal strategy in this case. Key reasoning steps, including the explicit comparison between action value and exploration cost, are highlighted in yellow. This example illustrates that when priors are provided explicitly, the model can reason about uncertainty and exploration cost to select an optimal strategy.

indicative tokens (e.g., `_eu`, `_tab`, `_sas`, `_cn`) with additional irrelevant strings.

We define an oracle filename-to-format model

$$\mathcal{M}_{\text{oracle}} : n \mapsto \pi(\mathbf{z} \mid n),$$

which maps a filename  $n$  to a prior distribution over formatting configurations  $\mathbf{z} = (z_d, z_q, z_s)$ . For example, filename tokens like `_tsv` substantially increase the probability of a tab delimiter relative to a default comma delimiter.

We then sample the true formatting configuration  $\mathbf{z}^* \sim \pi(\mathbf{z} \mid n)$  and generate the corresponding CSV content and task query.

The dataset is constructed so that the correct answer is obtainable only when the file is parsed with the correct configuration; incorrect formatting assumptions lead to parsing failures or misaligned columns. Filenames are represented by four binary features, yielding  $2^4 = 16$  distinct filename feature configurations and corresponding prior distributions over formatting attributes. We generate 2,000 task instances, each consisting of a filename, a CSV file generated under a sampled formatting configuration, and an associated query. The dataset is split into 1,400 training examples, 300 validation examples, and 300 test examples. Details of the oracle filename-to-format model, feature templates.

**Latent Formatting Variables** Each task instance is associated with a latent formatting configuration

$$\mathbf{z} = (z_d, z_q, z_s),$$

where  $z_d \in \{ , ; , \backslash t \}$  denotes the delimiter,  $z_q \in \{ " , ' \}$  the quote character, and  $z_s \in \{0, 1\}$  the number of header rows to skip. The correct answer can be obtained if and only if the file is parsed using the fully correct configuration  $\mathbf{z}^*$ .

**Filename Features** We extract four binary features from each filename  $n$ , each indicating the presence or absence of a specific substring: `has_eu`, `has_tsv`, `has_sas`, and `has_cn`. Each feature is either on or off, resulting in  $2^4 = 16$  possible filename feature configurations. Each configuration corresponds to a distinct prior distribution over formatting attributes induced by the oracle model.

**Oracle Filename-to-Format Model** We define an oracle filename-to-format model

$$\mathcal{M}_{\text{oracle}} : n \mapsto \pi(\mathbf{z} \mid n),$$

which maps a filename  $n$  to a prior distribution over formatting configurations  $\mathbf{z} = (z_d, z_q, z_s)$ . For each of the 16 possible filename feature configurations, the model induces a corresponding prior over formatting attributes. The prior factorizes as

$$\pi(\mathbf{z} \mid n) = [\pi(z_d \mid n), \pi(z_q \mid n), \pi(z_s \mid n)],$$

where each factor is parameterized as a log-linear model over the filename features.

**Sampling and File Generation** For each filename  $n$ , we sample a formatting configuration  $\mathbf{z} \sim \pi(\mathbf{z} \mid n)$  from the oracle model  $\mathcal{M}_{\text{oracle}}$ . We then generate a CSV file whose content conforms to  $\mathbf{z}$ . The data are constructed such that incorrect parsing—due to an incorrect delimiter, quote character, or number of skipped rows—either produces malformed outputs or prevents access to the correct answer.

**Task Instances and Splits** Each task instance consists of a filename  $n$ , a generated CSV file, and a query requiring the agent to compute an answer from the file. We generate 2,000 task instances in total, split into 1,400 training examples, 300 validation examples, and 300 test examples.

## D. Oracle Strategy for Pandora’s Box Problem

Algorithm 1 presents the optimal policy for the Pandora’s Box problem. In this section, we prove its optimality.

We begin by characterizing the structure of an optimal policy. At any state, let  $S$  denote the remaining set of boxes and let  $q_i(S)$  denote the posterior probability that box  $i \in S$  contains the prize. An optimal policy only needs to consider the box with maximum posterior probability at each step.

---

**Algorithm 1** Oracle policy for Pandora's box with  $n$  boxes

---

```

function Solve( $S$ )
    if  $|S| = 1$  then
        return (1, COMMIT(only element of  $S$ ))                                 $\triangleright$  Base case: Final box selection
    end if

     $W \leftarrow \sum_{j \in S} p_j$ 
     $i^* \leftarrow \arg \max_{i \in S} p_i$                                                $\triangleright$  Select candidate with highest success posterior
     $q \leftarrow p_{i^*}/W$ 

     $V_{\text{guess}} \leftarrow q$ 
     $(V_{\text{fail}}, -) \leftarrow \text{SOLVE}(S \setminus \{i^*\})$                           $\triangleright$  Expected value if committing to box  $i^*$  now
     $V_{\text{verify}} \leftarrow \gamma(q + (1 - q) \cdot V_{\text{fail}})$                             $\triangleright$  Recurse to find value if box  $i^*$  is empty
                                                                 $\triangleright$  The value of verifying  $i^*$  first

    if  $V_{\text{guess}} \geq V_{\text{verify}}$  then
        return ( $V_{\text{guess}}$ , COMMIT( $i^*$ ))
    else
        return ( $V_{\text{verify}}$ , VERIFY( $i^*$ ))
    end if
end function

```

---

First, if the agent commits, the expected reward equals the posterior success probability of the chosen box. Hence, committing to any box other than a maximum-posterior box is suboptimal. Second, verification is beneficial only insofar as it increases the probability of early termination before further discounting. Verifying a higher-posterior box increases the chance of immediate success and therefore weakly dominates verifying a lower-posterior box. Consequently, it suffices to consider the box  $i^* \in \arg \max_{i \in S} q_i(S)$  at each decision step.

It remains to determine whether the agent should commit to  $i^*$  immediately or verify it first. Let  $q = q_{i^*}(S) = \max_{i \in S} q_i(S)$  be the posterior probability of the most likely box  $i^*$ .

If the agent commits immediately, the expected value is

$$V_{\text{guess}}(S) = q.$$

If it verifies  $i^*$ , then with probability  $q$  verification succeeds and yields reward 1, and with probability  $1 - q$  it fails and the problem reduces to the smaller set  $S \setminus \{i^*\}$ . Since verification incurs one multiplicative discount factor  $\gamma$ , the expected value of verifying is

$$V_{\text{verify}}(S) = \gamma \left( q + (1 - q) V(S \setminus \{i^*\}) \right).$$

Therefore, optimality implies the Bellman recursion

$$V(S) = \max \{V_{\text{guess}}(S), V_{\text{verify}}(S)\},$$

with base case  $V(\{i\}) = 1$ . This recursion is exactly implemented by Algorithm 1.

## E. Prompt templates

### E.1. Prompts for PANDORA

Prompt templates used for PANDORA are shown in Figure 9.

### E.2. Prompts for QA

Prompts used in QA are provided in Figure 10.

## Prompts: Pandora's Box

### SYSTEM

You are a rational agent tasked with solving sequential decision-making problems under uncertainty. You are given a set of options (bags) with prior probabilities of containing a prize with value 1. You can either VERIFY an option to get information (YES/NO) or GUESS an option to end the game and collect the reward.

- Each VERIFY action consumes one timestep.
- The reward for a correct GUESS is discounted by a factor  $r^t$ , where  $t$  is the timestep when you GUESS.
- You must balance information gathering (VERIFY) with timely exploitation (GUESS) to maximize expected discounted reward.

Always respond with exactly one action token per step, using the format:  
VERIFY <Option> or GUESS <Option>.

### Instruction

```
--- NEW GAME ---
Timestep: t=0
```

#### PROBLEM PARAMETERS:

- Bag Labels: {labels\_str}
- Prior Probabilities: {priors\_str} //only provided to the model in CTA- settings
- Discount Factor r: {r}

Choose your action.

*Figure 9.* Prompt template for Pandora's Box setting.

### E.3. Prompts for CODE

Prompts used in the CODE setting are provided in Figures 11, 12, 13, and 14.

## F. Case study: Cost-Aware Decision Traces in CODE with CTA-RL and RL

Figures 15 and 16 compare representative traces under a high relative code cost setting ( $\rho = 4.0$ ). The RL model (trained without conditioning on explicit priors) tends to default to running unit tests before attempting any code, and does not explicitly reason about the relative costs of UNIT\_TESTS versus CODE when choosing its next action. In contrast, the CTA-RL model exhibits the intended cost-aware behavior: it reasons about both (i) uncertainty over the CSV format and (ii) the relative cost of unit tests and code execution, and uses these factors to decide whether verification is worth performing before committing to a code attempt.

## Prompts: QA with optional retriever

### SYSTEM

You are a rational agent tasked with answering factual questions under uncertainty. At each step, you can either directly answer the question or retrieve additional context before answering.

Available actions:

- RETRIEVE: request a related context to consult before deciding your final answer. This consumes one timestep.
- ANSWER: <your short factual answer>, provide your final answer and end the interaction.

Each question comes with your estimated probabilities of answering correctly:

- p\_no\_context: probability you can answer correctly without retrieving.
- p\_with\_context: probability you can answer correctly after retrieval.

Your goal is to maximize expected discounted reward:

Reward =  $r^t * \text{correctness}$ , where t is the timestep when you issue ANSWER and correctness  $\in \{0,1\}$ .

Retrieval information:

If you choose RETRIEVE, your expected answer accuracy after retrieval is p\_with\_context. If you choose not to retrieve, you must rely only on your current knowledge.

Be deliberate — retrieving may improve accuracy but reduces reward due to time discounting. Balance speed and correctness carefully.

Always respond with exactly one action token per step, using the format:

RETRIEVE or ANSWER: <short factual answer>.

### Instruction - turn 1

--- NEW QUESTION ---

TIMESTEP: t=0

Question: question

Parameters:

- Discount factor (r): {r}
- Success probability with retrieval (p\_with\_context): {p\_with\_context}

Choose your action:

RETRIEVE or ANSWER: <short factual answer>.

### Instruction - turn 2

TIMESTEP: t=1

You have retrieved the following context:

{context}

Question: {question}

Now decide whether to answer:

Respond with:

ANSWER: <short factual answer>

*Figure 10.* Prompt templates for QA.

## System Prompt: CODE

You are an agent tasked with writing code to fulfill an instruction about a CSV file (e.g., answering a question using its contents). Your goal is to produce a correct answer while efficiently using available resources, as measured by discounted reward.

The exact CSV formatting may not be fully known. In practice, you can either proceed using reasonable default assumptions about the format, or run unit tests to verify specific formatting details you are unsure about before committing to a final answer.

### Allowed actions (choose exactly ONE per turn):

#### 1) UNIT\_TESTS

Run unit tests to debug CSV formatting assumptions. Unit test outputs are perfectly reliable.

Available unit tests:

- `test_delimiter(path) → {';', ';', '\t'}`
- `test_quotechar(path) → {"\"", "\""}`
- `test_skiprows(path) → {0, 1}`

Format (NO code fences):

```
UNIT_TESTS: test_delimiter("file.csv"), test_quotechar("file.csv")
```

You may include multiple unit tests in a single UNIT\_TESTS action. Each individual unit test counts toward the total number of unit tests used.

#### 2) CODE

Write Python code toward solving the task using your current assumptions about the CSV format.

- Enclose code in `'''python ...'''`
- You may import pandas as pd and read the file with:  
`pd.read_csv(filepath, delimiter=..., quotechar=..., skiprows=...)`
- Do NOT print the entire CSV.
- If your code computes the final result, print it to stdout so it can be read from the output.

After submission, the code will be executed and its stdout and stderr will be returned. You may use this feedback to extract the answer, debug, run additional unit tests, refine, or write additional CODE.

#### 3) ANSWER

Provide the final answer to the task and end the conversation.

Format exactly: ANSWER: <your\_answer>

The conversation ends immediately after you provide ANSWER.

#### Reward:

- Let  $U$  be the total number of unit tests used.
- Let  $C$  be the total number of CODE actions taken.
- Final reward = correctness  $\times (d_{\text{unit}})^U \times (d_{\text{code}})^C$ .
- Discount factors represent cost multiplicatively.
- A smaller discount factor means a MORE expensive action.
- If  $d_{\text{code}} = d_{\text{unit}}^k$ , one CODE attempt costs about as much as  $k$  UNIT\_TESTS.

*Figure 11.* System prompt for CODE.

### System Prompt: CODE (continuation)

**General guidance:**

- Start from reasonable default beliefs about the CSV format based on common conventions or provided likelihoods.
- Both UNIT\_TESTS and CODE are costly actions; neither should be treated as free.
- Use UNIT\_TESTS to reduce uncertainty when the expected benefit outweighs their cost.
- Use CODE to make progress toward solving the task, but recognize that failed or repeated CODE attempts are also costly.
- Decide when it is better to verify assumptions with UNIT\_TESTS versus attempting CODE earlier, taking into account your confidence and the relative cost of these actions.
- Decide rationally how much debugging and iteration is worthwhile before committing to a final ANSWER.

*Figure 12.* Continuation of the system prompt for CODE.

### Instruction Prompt Template for CODE (without estimated prior)

You are given a CSV file {csv\_name}.

Your task: {task\_description}

**Additional context:**

- No format likelihoods are provided.
- Make reasonable default assumptions about the CSV format based on common conventions, unless you choose to verify them with unit tests.

**Reward parameters:**

- Unit test discount  $d_{unit}$ : {d\_unit}
- Code iteration discount  $d_{code}$ : {d\_code}

**Constraints:**

- You should never print all rows of the CSV or you will get zero reward.
- You may use UNIT\_TESTS, CODE, or ANSWER as described in the system instructions in any order; only the final ANSWER ends the conversation.
- Incorrect intermediate CODE does not end the episode; only the final ANSWER determines correctness.

*Figure 13.* Instruction prompt template specifying the CSV task, reward parameters, and constraints provided to the agent.

### Instruction Prompt Template for CODE (with estimated prior)

You are given a CSV file `{csv_name}`.

Your task: `{task_description}`

**Additional context:**

- Estimated format likelihoods are provided below.
- These likelihoods reflect how likely each formatting option is in practice and can be used as default assumptions.

**Format likelihoods:**

`{prior}`

**Reward parameters:**

- Unit test discount  $d_{unit}$ : `{d_unit}`
- Code iteration discount  $d_{code}$ : `{d_code}`

**Constraints:**

- You should never print all rows of the CSV or you will get zero reward.
- You may use UNIT\_TESTS, CODE, or ANSWER as described in the system instructions.

Figure 14. Instruction prompt template with estimated CSV format likelihoods, enabling the agent to use probabilistic defaults when trading off unit tests, code execution, and early commitment.

### Case study: Example thinking trace of RL model (without estimated priors) ( $\rho = 4.0$ )

**Content:**

`<think>`

Okay, let's tackle this problem. The user has a CSV file named 'race\_tsv\_sas.tsv' and they want the minimum salary value, excluding any None entries. The file's name ends with .tsv, which usually stands for Tab-Separated Values. So, the delimiter is likely a tab character. But I should confirm that.

First, I need to figure out the CSV format. Since it's a TSV, the delimiter is probably a tab. But maybe the file uses a different delimiter. To be safe, I should run a unit test to check the delimiter. The `test_delimiter` function can help here. If the delimiter is a tab, then the test should return '\t'.

Next, the quote character. In TSVs, sometimes fields are enclosed in quotes, but it's not always the case. The default quotechar for pandas is double quotes, but maybe this file uses single quotes. Testing with `test_quotechar` would help.

Also, sometimes CSV files have header rows that need to be skipped. The `test_skiprows` function can check if there's a need to skip rows. But since the file is a TSV, maybe the first row is the header. However, without knowing the exact structure, it's safer to assume that the first row is the header and set `skiprows=0`. But maybe the file has some leading rows to skip.

But since the user mentioned it's a TSV, I can start by assuming the delimiter is a tab. Let me run the `test_delimiter` first. If the delimiter is a tab, then I can proceed. If not, I'll adjust accordingly.

Once the delimiter is confirmed, I can use pandas to read the CSV. Then, check the 'salary' column, filter out None values, and find the minimum.

But to avoid errors, maybe I should first run the unit tests to confirm the delimiter and quotechar. Let's start with that.

`</think>`

**Action:**

UNIT\_TESTS: `test_delimiter("race_tsv_sas.tsv")`, `test_quotechar("race_tsv_sas.tsv")`

Figure 15. Example reasoning trace of an RL-trained model without explicit prior conditioning in the CSV exploration task. Despite operating under the same high relative code cost setting ( $\rho = 4.0$ ), the model defaults to verification-first behavior based on surface cues (e.g., file extension) and does not explicitly reason about uncertainty or cost trade-offs, illustrating a lack of adaptive decision-making compared to the CTA-RL model.

## Case study: Example thinking trace of CTA-RL model ( $\rho = 4.0$ )

### Content:

<think>

Okay, let's tackle this problem. The task is to compute the average of the 'salary' column from the CSV file 'must\_eu.csv', excluding any None values. The challenge is figuring out the correct CSV format parameters to read the file properly.

First, I need to determine the delimiter, quotechar, and skiprows parameters. The format likelihoods are given, so I should start with the most probable options. The delimiter is most likely to be ';' with a probability of  $\sim 0.85$ . The quotechar is probably '"' with 0.589, and skiprows is 0 with 0.607. But I'm not 100% sure, so maybe I should run some unit tests to confirm.

But wait, the reward parameters are important. Each unit test costs 0.77 per use, and code attempts cost 0.3515 per use. Since  $d_{\text{code}} = d_{\text{unit}}^4$ , one code attempt is equivalent to 4 unit tests. So, if I'm uncertain, it's cheaper to run unit tests first. However, if I can make an educated guess, maybe I can proceed with code.

Let me think: the delimiter is most likely ';', so maybe I can try reading the file with that. But if the actual delimiter is different, the code will fail. Alternatively, running a unit test for the delimiter would be better. Let's check the delimiter first.

So, I'll run UNIT\_TESTS to check the delimiter. The `test_delimiter` function will tell me if it's comma, semicolon, or tab. Since the likelihood is high for semicolon, but I need to confirm. Let's do that.

Once I know the delimiter, I can check the quotechar. The quotechar is either '"' or "'", with higher likelihood for '"' (0.589). But again, maybe a unit test is better. However, if I proceed with code, maybe I can handle the possible quotechar. But if the code fails, I'll have to adjust.

Alternatively, maybe I can proceed with the most likely parameters and see. Let's try that. Let's assume delimiter is ';', quotechar is '"', and skiprows 0. Then, read the CSV with these parameters, compute the average of salary, skipping None values.

But if the actual delimiter is different, the code will fail. So, perhaps it's better to run unit tests first. Let's start with the delimiter. So, the first action is to run UNIT\_TESTS to determine the delimiter and quotechar. Let's do that.

</think>

### Action:

```
UNIT_TESTS: test_delimiter("must_eu.csv"), test_quotechar("must_eu.csv")
```

*Figure 16.* Example reasoning trace of the CTA-RL model on the CODE task ( $\rho = 4.0$ ), illustrating cost-aware trade-offs between unit tests and code execution under a high relative code cost setting, while jointly reasoning about format uncertainty.