Integration of Language and Experience via the Instructed Bandit Task

Ellen Su (ellensu@nyu.edu)
Mark K. Ho (mkh260@nyu.edu)
Todd Gureckis (todd.gureckis@nyu.edu)
Department of Psychology, 6 Washington Pl
New York, NY 10003

Abstract

Humans learn by interacting directly with their environments and by communicating via language. In this project, we explore this interaction between language and experiential learning through a novel sequential decision-making task, the "instructed bandit task" (IBT). In the IBT, agents make choices and receive rewards sampled from an unknown Gaussian distributions, after being given linguistic hints. The IBT assesses how linguistic input and experienced reward values combine to determine choice behavior. We additionally propose a novel Bayesian reinforcement learning model that combines Bayesian updating from experience with propositional constraints that capture the meaning of the linguistic hints. As a point of comparison, we evaluate both human participants and Centaur, a LLaMA-based model fine-tuned to mimic human behavior, on the IBT. Our results show that all agents converge with the Bayesian model, and the granular difference in choice sequences reveal the varied role instruction plays in decision-making tasks.

Keywords: reinforcement learning, language, Bayesian cognition, large language models

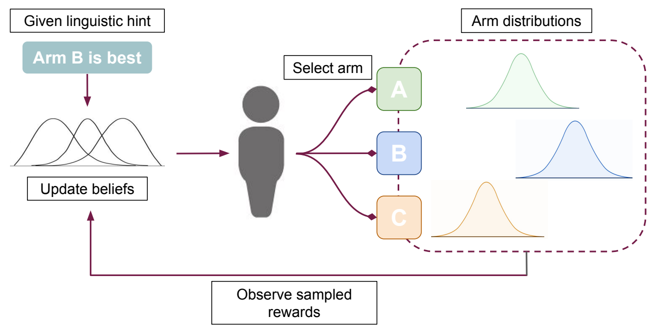

Figure 1: Overview of the instructed bandit task (IBT). Agents are given qualitative hints in the form of verbal instructions about the task then repeatedly make arm choices in a standard multi-armed bandit. Their behavior will depend on both the verbal, instructed information and direct experienced rewards.

Introduction

A key feature of human cognition is the ability to adapt behavior to experience. The actions people take determine the feedback they receive from the surrounding world, and these rewards, consequences, or other variables then update their beliefs and inform future actions. For instance, a huge literature on value-based decision-making and reinforcement learning (RL) explores how humans make choices based on the history of experienced rewards.

Much of what people learn is also derived indirectly by communicating with others using language. For example, an individual may receive recommendations from another person (e.g. "avoid the campus food trucks, they are not that great") that alter their beliefs. Language is powerful in allowing humans to communicate goals, rules, and limitations in an abstract and general way (Gopnik & Meltzoff, 1987; Lupyan & Bergen, 2015). The use of language as instruction, then, has the ability to greatly expedite human learning and decision-making in some cases. In others, this communication may only provide partial information that is vague, incomplete, or qualitative.

Most importantly, these two types of information, both relevant for choice, rely on seemingly incommensurate cognitive processes. Language comprehension exemplifies higher-level cognitive processing; it is thought to be rapid and symbolic, and it can incorporate explicit inferences about a speaker's intentions among other factors (Goodman & Frank, 2016). On the other hand, experiential learning from environmental events lies at the opposite extreme; it is slow, statistical, and often implicit (Dayan & Niv, 2008; Sutton & Barto, 2018). Language can have varied effects on experience-based tasks. It might provide specifications of goals, alter patterns of exploration, suggest optimal strategies, or, in some cases, lead people astray. One method to explore this dynamic is therefore to study tasks where people are given instructions or linguistic hints and need to combine that information with their direct experience in making a series of value-based choices. Despite task instructions being critical to behavior, such linguistic information is often ignored in modeling choice behavior.

In this paper, we propose a new class of experiential learning tasks (based on the classic multi-armed bandit task) which incorporate linguistic information in the form of hints and instructions. In the instructed bandit task (IBT), an agent makes a series of choices and receives evaluative feedback on those choices in the form of rewards. The agent's goal is to maximize their rewards over the course of a game. Each agent is also given a hint which provides qualitative information about how to approach the task. This paradigm sheds light into the relative weight that agents give to their own experience versus the hints, and suggests how the specific semantic content of a hint may structure choice and exploration. It differs from past work regarding social communication in multi-armed bandit tasks (Sankararaman, Ganesh, & Shakkottai, 2019) by including natural language instruction in accounting for choice behavior.

Bayesian RL and language modeling

We model the choices made by the participants in our experiment with two different approaches. The first leverages Bayesian RL and treats the verbal instructions as prior constraints on the distribution of possible rewards (Chapelle & Li, 2011). The second makes use of recent advances in large language models (LLM) and samples choice data from the Centaur model (Binz et al., 2024). Binz et al. recently released the Centaur model, a large language model (LLM) fine-tuned on Psych-101, a large-scale dataset of past psychological experiments translated into a text-based environment (Binz et al., 2024). The fine-tuning objective was to align the base model, a LLaMA 3.1 70B model (Grattafiori et al., 2024), to human behavior in a domain-general manner . One interesting claim in the paper is that Centaur is able to predict human behavior for entirely new behavioral tasks with only the relevant task instructions. We were interested in evaluating it on IBT to carefully measure the impact that different hints would have on the prediction behavior for a model that relies heavily on linguistic input.

We begin by describing the IBT and proposing the Bayesian RL model as a framework for how people might incorporate both linguistic instruction and experiential learning in sequential decision-making. Next, we report a novel experiment where human participants completed the IBT with a dataset of different arm values and hints. Finally, we compare the human choices with those predicted by both the Bayesian RL agent and Centaur.

The instructed bandit task (IBT)

We chose to study how language integrates with experience through a variant of the stationary multi-armed bandit problem, which is a simple yet powerful framework that formalizes the process of an agent reasoning through a sequential decision-making task as an optimization problem (Sutton & Barto, 2018). In our N-armed bandit problem, an agent repeatedly chooses between N arms, where each arm returns a reward sampled from a fixed Gaussian distribution with unknown mean μ and known variance σ^2. The agent's objective is to maximize their rewards, and they must do so by learning an estimate of the arm values, $\mu_{1:N}$, in order to choose optimally. Critically, agents are given a hint in English which provides qualitative information relevant for the task. For example an agent might be told "One of the arms is more than 25". Such information provides partial constraints on the task and should affect choice behavior in systematic ways.

Bayesian agent

To model the learning process of an idealized agent, we take a Bayesian approach as it allows for natural integration of information from different sources in the posterior inference. At each time step t, the agent selects an action a_t and receives a reward $r_t \sim \mathcal{N}(\mu_a, \sigma_a^2)$. The learning history up to time step t is given by the full sequence of actions taken and rewards received, $h_t = (a_1, r_1, a_2, r_2, \ldots, a_t, r_t)$. In the base case of a Gaussian bandit, the agent's posterior distribution over the N arm means conditioned on the learning history is,

$$P(\mu_{1:N} \mid h_t) \propto \underbrace{P_0(\mu_{1:N})}_{Prior} \underbrace{P(h_t \mid \mu_{1:N})}_{Experience} \tag{1}$$

where $P(\mu_{1:N})$ is the prior distribution—uniform under total uncertainty—over the arm means. Without any instruction, RL is a purely evaluative process (Sutton & Barto, 2018). The Bayesian agent will explore all possible actions in order to receive feedback on each arm and iteratively learn the arm values. More specifically, we assume the agent samples actions using Thompson Sampling at each time step and actively updates the posteriors according to the new data (Thompson, 1933; Wilson, Bonawitz, Costa, & Ebitz, 2021).

Linguistic hints

Linguistic instruction often guides exploratory learning in providing partial supervision on the arm values and subsequently narrowing the search space for the agent. In the context of multi-armed bandits, the instructions or hints provided contribute some information about the bandit arm values or relationships, and the learning agent may use the information from the hint in a similar manner as the prior. Following previous work in formal semantics and Bayesian language modeling (Cresswell, 2006; Goodman & Frank, 2016), we formalize the meaning of each linguistic hint as a function over the parameter space. By mapping each hint to its functional form, we can then model how language and instruction combines with evaluative exploration and learning.

In the N-armed Gaussian bandit setting, the parameter space is defined as $\Theta = \mathbb{R}^N$ and captures all possible arm mean configurations. Each linguistic hint l with meaning f_l is a mapping from the task parameterization to the non-positive real numbers extended with negative infinity, $f_l : \Theta \to \mathbb{R}_{\leq 0} \cup \{-\infty\}$. The extended real-valued functions allows us to capture categorically true or false hints, represented by 0 or $-\infty$ respectively, as well as graded similarity judgements between parameter values in a manner analogous to fuzzy logic used in control engineering (Zadeh, 1965). Examples of linguistic hints and their meaning functions are provided in table 1.

Each meaning function is represented as an arithmetic and logical term which is evaluated on a particular set of parameter values $\mu_{1:N} \in \Theta$. The terms adhere to a formal grammar similar to those used in rule induction (Goodman, Tenenbaum, Feldman, & Griffiths, 2008) and are hand encoded to the N-armed bandit setting. For example, we encode the hint

Table 1: Example hint encodings

Linguistic Hint (l)	Meaning function (f_l)
"Arm 1 is more than 25"	$\ln \mathbf{1}(\mu_1 > 25)$
"Arm 1 is around 40"	$-\lvert\mu_1 - 40\rvert$
"Arm 1 is similar to arm 2"	$-\lvert\mu_1 - \mu_2\rvert$
"Arm 3 is the best"	$\ln \mathbf{1}(\mathrm{argmax}_a \mu_a = 3)$
"One of the arms is more than 25"	$\max_a\{\ln_1(\mu_a > 25)\}$
"Two arms are similar"	$\max_{i,j:i\neq j}\{-\lvert\mu_i - \mu_j\rvert\}$

"Arm 1 is more than 25" as $\ln \mathbf{1}(\mu_1 > 25)$, which includes a Boolean sub-term $(\mu_1 > 25)$, an indicator function $\mathbf{1}(\cdot)$, and the natural log function $\ln(\cdot)$. If arm 1 indeed has mean over 25, the meaning function will evaluate to 0, indicating truth. The final log transformation is necessary for introducing the meaning function into the Bayesian model.

Integrating language and experience

To incorporate the information from the linguistic hints into the Bayesian decision-making agent, we adapt equation 1 as follows,

$$P(\mu_{1:N}|h_t,l) \propto \underbrace{P_0(\mu_{1:N})}_{Prior}\underbrace{P(h_t|\mu_{1:N})}_{Experience}\underbrace{\exp(\gamma f_l(\mu_{1:N})\})}_{Hint} \quad (2)$$

The addition of the term $\exp(\gamma f_l(\mu_{1:N})\})$ introduces a language dependency in the agent's posterior beliefs. The exponential term transforms the linguistic hint function value, which takes on values between $[-\infty, 0]$, into a probabilistic range of values in $[0,1]$. Intuitively, the contribution from the linguistic hint can be considered a second prior term. Under total uncertainty, the agent's prior may be uniformly distributed over arm means, but the linguistic hint should sway the agent's initial beliefs in the absence of any experience data. The γ term also serves as a weighting for this dependency and is necessary to capture variation among individuals, among other factors.

Continuing the example from above, at each time step, the Bayesian agent will maintain a posterior distribution for each of the arms. Recall that the hint "Arm 1 is more than 25" will have a meaning function value of $f(\text{"Arm 1 is more than 25"}) = 0$. Then, $\exp(0) = 1$. In context, the effect of this hint will be to keep all distribution values greater than 25 for arm 1, and zero out any values less than 25 for arm 1. It will have no effect on the distributions of the other arms. As the arm distributions are normalized at each update step, the practical effect of this hint is that the agent will more efficiently converge on a posterior distribution for arm 1 that has mean greater than 25.

5-armed instructed bandit experiment

We created a dataset of 5-armed bandit scenarios alongside distinct sets of linguistic hints. Each hint contains varying informational value with respect to the parameter space, and some hints are misleading. Each of the Gaussian arms in the dataset abides by the constraints $\mu \in [0,100]$ and $\sigma^2 = 10$. An example of a set of arm values and its corresponding hints is given by:

```
(10, 20, 40, 60, 70): [

 "{max2} is the best",            # helpful
 "{max2} is sum of three arms",  # helpful
 "The range of all arms is 60",  # helpful
 "{min1} is better than {max1}", # misleading
 "no hint",                       # neutral

]
```

For each set of arm values, the order of the arms is fixed as (min2, min1, mid, max1, max2). The variable hints are necessary for implementing rotations of the arm values across trials in order to avoid biases in their labels or memory of their values. For each rotation and scenario, the hints self-updated to reflect the new indices of the arms. Each scenario in the experiment paired a set of rotated arm values and a hint and was replicated at least 60 times in the experiment. For each trial, the agent was given 20 arm choices to maximize their rewards.

Human behavioral experiment

We recruited 63 participants from Prolific (Palan & Schitter, 2018). We chose the number of participants to ensure a minimum of 60 replicates per scenario. We additionally filtered out any participants who did not complete all 15 games (2 people) or who failed the instructions quiz more than twice (0 people) in our final analysis. We paid participants at a rate of $15/hour and a bonus of up to $3.00 based on a random sample from their points earned.

We gave the participants instructions to maximize their point earnings over 20 arm choices, and each participant played a total of 15 games. We followed a within-subjects design and each game consisted of a unique scenario of linguistic hint and arm values sampled from the dataset. The hint was visible to the participant through all trials of each game and fell into the three categories of "no hint", helpful hints, or misleading hints. When the participant was given "no hint" in the control condition, they were additionally asked to provide hints which may help another participant in that game. These collected hints formed a rich hint dataset which we will encode into meaning functions for future iterations of instructed choice tasks.

Centaur model

We configured the Centaur model with default hyperparameters for sequence length, new tokens, and temperature (Binz et al., 2024; Grattafiori et al., 2024). We performed a temperature exploration experiment to ensure that the results presented were not determined by the temperature of the model, and we found the average choice sequences to be consistent across temperatures. We additionally experimented with whether the instructions given to the model should be taken directly from the behavioral experiment or be adapted to the wording of the multi-armed bandit experiments in the Psych-101 dataset. Both led to similar model results, and thus we made only the minimal edits to the instructions given to the participants (such as adding in brackets to indicate model choices) recommended by the Centaur paper (Binz et al., 2024). Centaur was run 60 times on each distinct data scenario and was also given rotated arm values across trials. The model is given the trial history over the course of a game, but each game serves as an independent simulation. Our exact configurations and code to run the Centaur model experiments can be found in `https://github.com/eysu35/Su_IBT_CogSci_25`.

Bayesian RL agent

The Bayesian RL agent approximates equation 2 by algorithmically integrating choice feedback with the linguistic hints. Actions are first sampled using Thompson sampling (Thompson, 1933). A generative model then computes the experience-only posteriors by taking into account the learning history h_t and updating the posterior means using conjugate priors (Murphy, 2022). Next, K samples are taken from the posterior distributions and filtered by the weights corresponding to a hint's meaning function (Shachter & Peot, 1990). Finally, the normalized weights approximates the probability of each sample conditioned on the linguistic hint, or the left hand side of equation 2, from which Thompson sampling can be applied again to generate new actions. We ran a series of simulations using this algorithm in which the Bayesian RL agents were given no hint, informative hints, or misleading hints. Each scenario was replicated for 600 trials. Our experiments were done using default sampling parameters detailed in (Ho & Gureckis, 2023).

Results

Linguistic hints aid learning

Panel A of figure 2 summarizes the results from the experiments with human participants, the centaur model, and Bayesian RL agent simulations across all hint conditions. We measured the proportion of trials (out of 20 total trials) where the agent made the optimal arm choice for each of the hint conditions. Each line is averaged across all arm values, hints within each category, and game trials of each scenario.

We performed a one-way ANOVA statistical test on the collected data for each agent to evaluate the effect of hint type on performance. We found that the variation between the mean performance in different hint categories was significantly larger than the variation within each hint category in all cases (Human: $F(2,899) = 46.12$, $p < 10^{-3}$, Centaur: $F(2,597),= 58.96$, $p < 10^{-10}$, Bayesian RL agent: $F(2,2997 = 714.92$, $p < 10^{-10}$), indicating that the addition of linguistic hints had a significant effect on learning performance. Further, in the post-hoc comparison using Turkey's HSD test, we found that all three hint conditions differed significantly from one another. Across human participant results and simulated results from the Centaur model and Bayesian RL agent, performance in the helpful hint condition was significantly higher than the no hint condition (Human: $p < 10^{-3}$, Centaur: $p < 10^{-3}$, Bayesian RL agent: $p < 10^{-10}$), and performance in the no hint condition was significantly higher than the misleading hint condition (Human: $p < 5^{-2}$, Centaur: $p < 10^{-3}$, Bayesian RL agent: $p < 10^{-10}$).

These results, that helpful hints meaningfully improve performance and misleading hints undermine performance, support the idea that humans integrate their direct experiences with the provided linguistic information to make their choices in the multi-armed bandit task. The consistency of the effect across humans and models also indicate that both the Centaur model and the Bayesian RL agent are able to simulate this behavior and integrate the linguistic and experiential information as well.

Learning trajectories

Next, we observed the individual learning trajectories of all three agents when provided no hint, helpful, or misleading hints. Both quantitatively shown in section and qualitatively observed in panel B of figure 2, all three learning agents are able to make more optimal choices over trials and when receiving helpful linguistic hints as opposed to when receiving no hint or a misleading hint.

First, the"no hint" case (gray dashed lines in all subplots) represents a control scenario. Here, no information on the arm means was given by the linguistic hint, and the agents had to make their first choice solely based on their prior, which is uniformly distributed over the arm values in this setting. As the first choice is randomly sampled from the 5 arms, the average proportion that the optimal arm is chosen should be 0.2 across all game trials. This behavior is observed across all three learning agents and in all scenarios.

A helpful hint (blue lines in all subplots in panel B) narrows the search space of the arm means by upweighting the agent's prior in the direction of the optimal arm. For example, the most helpful hint, "{max2} is best" (darkest blue line), reduces the problem completely and directly provides the agent with the optimal policy. Thus, in the Bayesian model simulation, the optimal arm is chosen in nearly every trial. Panel B of figure 2 shows that the human participants, Centaur model, and Bayesian RL agent all achieved a high proportion of optimal action taken at trial 0 when they received this hint. Even when provided the maximally helpful hint, we observe a striking pattern in both the human and the Centaur model data where, after selecting the optimal arm, there is an exploratory period of 5-7 trials while the agent explored other arm values. An agent that relies entirely on instruction should omit exploration and chose the optimal arm in every trial. Thus, this pattern is evidence that agents require both experiential and evaluative learning to supplement and validate the linguistic input.

The 5-armed bandit problem is a fairly simple task; it takes around 5 trials to determine the optimal action to take, with slight variation due to stochasticity and the overlapping arm

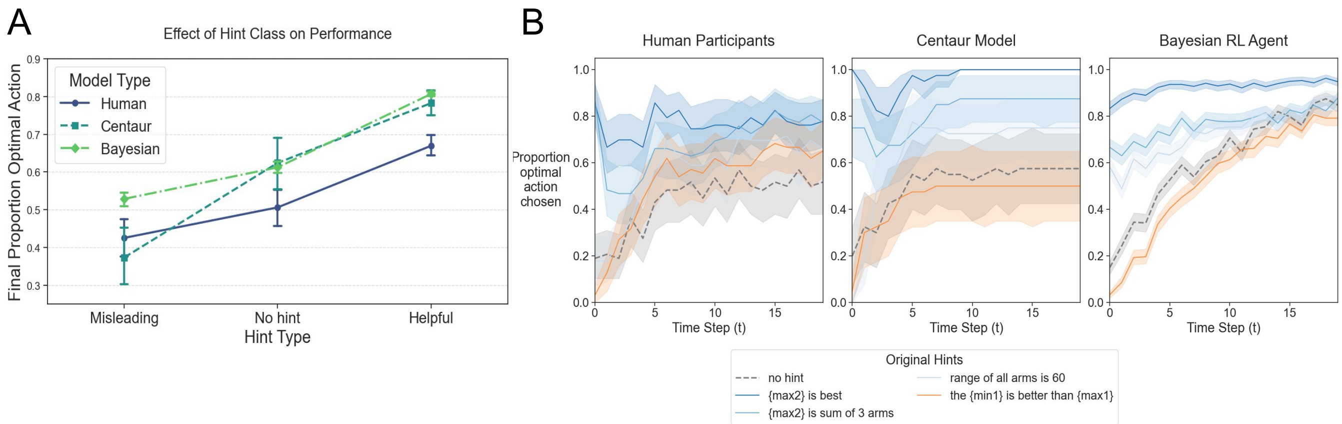

Figure 2: **A:** Summary of the performance, as measured by proportion of choices optimal action was chosen at the 20th trial, for all three models. The neutral condition indicates the ”No hint” baselines, the misleading and helpful conditions indicate the hint type. All results are averaged across bandit scenarios and game trials. Error bars show 95% confidence intervals. **B:** Learning trajectories for human participants, centaur model, and Bayesian RL agent over time trials. The dotted gray line indicates the ”No hint” baseline, blue lines indicate helpful hints, and the orange line indicates the misleading hint. All results are taken from the scenario with arm values (10, 20, 40, 60, 70) and are averaged over all game trials with these values. Error bars show 95% confidence intervals.

distributions. Still, even with an already efficient exploration strategy, we observe in the lighter blue lines in panel B of figure 2 that even partial information on the arm values can improve choice performance. This result supports the claim that language can improve sample efficiency in when making value-based choices.

In each scenario of arm values, the agents were also provided one misleading hint whose meaning function contributed information that contradicted the true arm means. In panel B of figure 2, the misleading hint is shown in the orange line: “The $\{\text{min1}\}$ is better than $\{\text{max1}\}$”. In this case, the agents’ choices underperform the “no hint” baseline at trial 0. However, the net positive slopes of the orange lines indicate that every agent learned over trials that the hint was untrue through their experienced rewards. The exact shape of the learning curve may also reveal the relative weight that agents place on linguistic and experienced data.

While prior work has shown that verbal instruction can have a strong influence on choice and lead to confirmation bias in the interpretation of collected evidence, the results from the IBT demonstrate the agents overcoming this phenomena combining both instructions and experienced reward (Doll, Hutchison, & Frank, 2011; Nickerson, 1998).

Language models for simulating human behavior

We additionally analyzed the fine-grained choice behaviors across trials in figure 3. First, we generated a behavioral signature for each agent by computing their average arm choice probabilities at each trial for a given scenario and flattening these into a single vector. We then computed the Pearson correlation coefficient between these summary vectors for each agent to get a dissimilarity score for their choices. We averaged these scores across hints within the same condition (no hint, helpful, or misleading) to arrive at the final scores between agents in table 2. A score of 0 means identical choice patterns and higher scores mean more dissimilar choice patterns.

Condition	Centaur	Bayesian RL
No hint	0.117	0.677
Helpful hint	0.060	0.677
Misleading hint	0.112	0.801

Table 2: Dissimilarity scores between model and human choice sequences. Each row represents a different hint condition. Each column represents the a model, Centaur model or Bayesian RL agent. First, a correlation score is computed between the model-predicted choice sequences and the human choice sequences. The dissimilarity score is then computed by taking $1-$ Pearson correlation. The dissimilarity score takes on values in $[0,2]$, with higher values being more dissimilar.

Table 2 shows that the choice sequences predicted by the Centaur model align much more closely with the human data than those predicted by the Bayesian model across all hint conditions. This result is further supported in the raster plots shown in figure 3, which demonstrate a similarity between the pattern of choices taken by human participants and the Centaur model. As previously discussed in in section, the Centaur

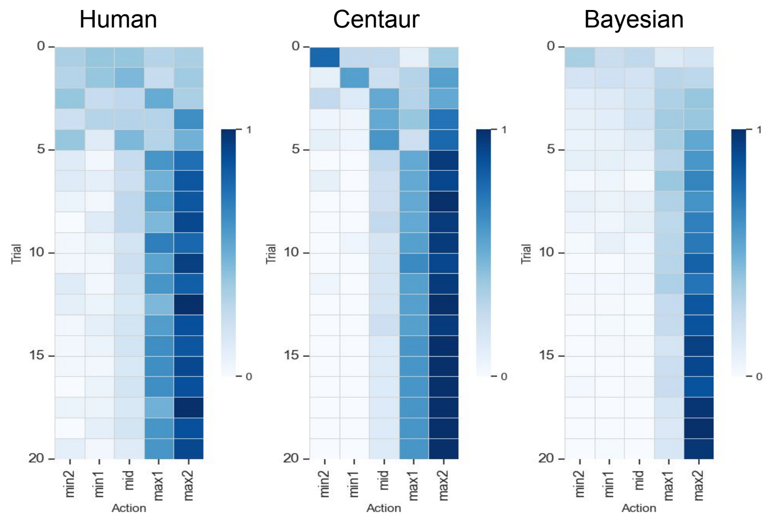

Figure 3: Each subplot shows the normalized frequencies of the actions chosen (columns in each subplot) over the 20 trials (rows in each subplot). All plots show the choice frequency data from the no hint scenario for the same arm values: (10, 20, 40, 60, 70). For each agent, the choice sequences are averaged over all game trials of that scenario.

model is able to simulate the human-like tendency to explore arm choices in early trials even when given a maximally helpful linguistic hint.

These results suggest that there is value associated with language-model based models of cognition. While the Bayesian model is able to provide an explanatory framework for how the information from language and experience integrate in the learning process, the Centaur model is better able to generate data that is predictive of human action. Perhaps the performance gap between the Centaur and Bayesian model in correlation with the human data suggests that the Bayesian theory of integrating language and experience is still missing critical components. Thus, from a behavioral standpoint, these results support the idea that Centaur and future language models can be useful for studying how people learn from both direct experience and linguistic information when making value-based choices.

Discussion

In this paper, we introduced the instructed bandit task (IBT), in which a simple value-based sequential decision-making problem was supplemented by the addition of linguistic hints. Our experiments reveal that agents are able to learn more efficiently and effectively in the multi-armed bandit task when provided helpful linguistic information. The contribution of our work lies in taking the first steps towards modeling the interplay of instructed and experienced information on learning and exploration.

In particular, we provided a novel Bayesian framework of mapping linguistic instruction to functions over the task parameters. This model effectively quantifies the amount of information provided by the hint relevant for the task. We also compared human behavior with simulated choices from the Centaur model. As language models are able to compute over natural language stimuli and thus do not require any hard coding of the linguistic input, they bring new opportunities to aligning models and human systems (Carvalho & Lampinen, 2025; Frank, 2025). Interestingly, Centaur model was able to mimic human choice sequences with surprising accuracy and with no additional parameter fitting. In future work, we aim to extend the Bayesian framework to allow us to model in a trial-by-trial fashion the impact and relative balance of instructed and experienced information on choice.

The version of the hints we explored were relatively simple and gave agents direct information about the reward distributions with varying accuracy and completeness. However, the IBT framework can be generalized and extended to tasks with sequential dependencies between actions, more complex cues (e.g., contextual or non-stationary bandits), and other dynamics. Furthermore, interesting issues about the context under which instructions were provided may be informative, such as the trustworthiness or credibility of the instruction provider. We additionally plan to model the hints collected from participants in later work.

Much of the work in psychology and neuroscience that explores experience-based choice and learning ignores the fundamental role that task instructions and language play in shaping human decision-making. Moreover, combining language and reinforcement learning is a growing intersection of interest in computer science, reflecting the important role that linguistic and semantic representations play in guiding complex behaviors (Luketina et al., 2019). This project and the IBT thus respond to these gaps and interests and moves forward our understanding of how humans integrate of language and experience.

References

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... Schulz, E. (2024). *Centaur: a foundation model of human cognition.* Retrieved from https://arxiv.org/abs/2410.20268

Carvalho, W., & Lampinen, A. (2025). *Naturalistic computational cognitive science: Towards generalizable models and theories that capture the full range of natural behavior.* Retrieved from https://arxiv.org/abs/2502.20349

Chapelle, O., & Li, L. (2011). An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 24). Curran Associates, Inc.

Cresswell, M. (2006). Formal semantics. In M. Devitt & R. Hanley (Eds.), *The blackwell guide to the philosophy of language* (pp. 131–146). Wiley-Blackwell.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185-196. (Cognitive neuroscience) doi: https://doi.org/10.1016/j.conb.2008.08.003

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *PLoS ONE*, *6*(9), e24566. doi: 10.1371/journal.pone.0024566

Frank, M. C. (2025, Mar). *Cognitive modeling using artificial intelligence.* PsyArXiv. Retrieved from osf.io/preprints/psyarxiv/wv7mg_v1 doi: 10.31234/osf.io/wv7mg_v1

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818-829. doi: https://doi.org/10.1016/j.tics.2016.08.005

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32 1*, 108-154.

Gopnik, A., & Meltzoff, A. N. (1987). The development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child Development*, *58*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... Ma, Z. (2024). *The llama 3 herd of models.* Retrieved from https://arxiv.org/abs/2407.21783

Ho, M. K., & Gureckis, T. (2023). Learning from language and experience. In *Cognitive computational neuroscience.*

Luketina, J., Nardelli, N., Farquhar, G., Foerster, J. N., Andreas, J., Grefenstette, E., ... Rocktäschel, T. (2019). A survey of reinforcement learning informed by natural language. *CoRR*, *abs/1906.03926*. Retrieved from http://arxiv.org/abs/1906.03926

Lupyan, G., & Bergen, B. (2015, 07). How language programs the mind. *Topics in cognitive science*, *8*. doi: 10.1111/tops.12155

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press. Retrieved from http://probml.github.io/book1

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175 - 220.

Palan, S., & Schitter, C. (2018). Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22-27. doi: https://doi.org/10.1016/j.jbef.2017.12.004

Sankararaman, A., Ganesh, A., & Shakkottai, S. (2019, December). Social learning in multi agent multi armed bandits. *Proc. ACM Meas. Anal. Comput. Syst.*, *3*(3). Retrieved from https://doi.org/10.1145/3366701 doi: 10.1145/3366701

Shachter, R. D., & Peot, M. A. (1990). Simulation approaches to general probabilistic inference on belief networks. In M. Henrion, R. D. Shachter, L. N. Kanal, & J. F. Lemmer (Eds.), *Uncertainty in artificial intelligence* (Vol. 10, p. 221-231). North-Holland. doi: https://doi.org/10.1016/B978-0-444-88738-2.50024-5

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, MA: MIT Press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294.

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338-353. doi: https://doi.org/10.1016/S0019-9958(65)90241-X