Can Language Models Perform Implicit Bayesian Inference Over User Preference States?

Linlu Qiu[1*], Fei Sha[2], Kelsey Allen[3], Yoon Kim[1], Tal Linzen[2], Sjoerd van Steenkiste[2]
[1]Massachusetts Institute of Technology, [2]Google Research, [3]Google DeepMind

Abstract

To successfully interact with the world, both humans and machines need to construct models of the world and form beliefs about these models. These beliefs need to be updated as new information comes in. We formalize this problem using a simple flight recommendation task, where in order to provide useful recommendations the assistant needs to infer the user's preferences as it interacts with the user. We evaluate the Gemma 2 family of instruction-tuned language models in this setting, and find that they perform poorly compared to an optimal Bayesian model. Most importantly, Gemma 2's performance remains constant even as more information becomes available. Overall, we identify probabilistic belief updating as a central challenge for interactive language models.

1 Introduction

Humans reason about the world based on our *beliefs* about it. To effectively support decision making, our beliefs need to take into consideration the structure of the world: they need to be supported by appropriate "world models" (Johnson-Laird, 1980; Wong et al., 2023). To the extent that we are uncertain about our environment, our beliefs need to be probabilistic and reflect this uncertainty. Finally, when the situation changes, or new information becomes available, we need to update these probabilistic beliefs to reflect the new information.

In this paper, we propose a paradigm that evaluates whether a system is able to perform such probabilistic inference over well-calibrated models of the world, and apply this paradigm to evaluate interactive language models (LMs). As a proof of concept, we instantiate this paradigm in a simple, controllable setting: a flight recommendation task inspired by Lin et al. (2022). This task involves multi-round interactions between a user and a flight booking assistant. In each round, the assistant is expected to recommend a flight out of multiple options. The assistant then receives feedback on the user's preferred option. The user's preferences are shaped by a latent *reward function*, which quantifies the strength of their preferences for, say, longer flights over shorter ones. Because the assistant does not have direct access to the user's preferences, to make optimal recommendations, it must construct an implicit model of the factors that shape the user's preferences, and reason probabilistically about those factors as it learns about the user's choices across different situations.

An optimal strategy to perform this task relies on Bayesian inference (Jern et al., 2017; Tenenbaum et al., 2011; Xu and Tenenbaum, 2007; Baker and Saxe, 2011; Tenenbaum et al., 2006). Such a Bayesian Assistant maintains a probability distribution that reflects its beliefs about the user's preferences, and uses Bayes' rule to update this distribution as new information becomes available. For the simple task we focus on in this paper, exact probabilistic inference is tractable, which makes it straightforward to implement the Bayesian Assistant.

[*] Work done while a student researcher at Google.

The First Workshop on System 2 Reasoning at Scale at NeurIPS 2024.

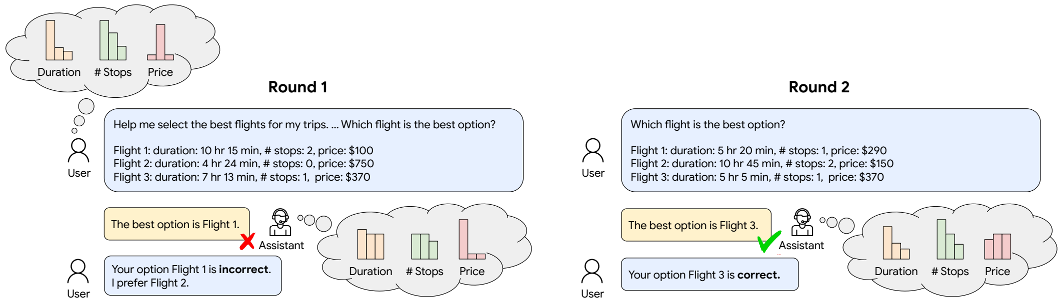


Figure 1: Our task involves multi-round interactions between a user and a flight booking assistant. In each round, the assistant has to make a recommendation to the user out of three available flight options. The assistant then receives the flight that was in fact chosen by the user (based on the user's ground truth reward function). To make good recommendations, the assistant needs to infer the user's reward function from the user's choices.

We find that LMs perform significantly worse than the Bayesian Assistant, and only slightly better than random guessing. Their performance plateaus after a single round, indicating a limited ability to update their beliefs. The Bayesian Assistant, on the other hand, is able to monotonically improve as additional information comes in, and it often infers the correct reward function after a small number of interactions.

A hallmark of a good reasoner is that it can maximally exploit the amount of information it is provided with. Guided by this motivation, we investigate the relationship between the informativeness of the options presented and the extent to which the model's performance improves. We find that informativeness has a large effect on the Bayesian Assistant, but not on the LMs. Overall, then, while transformers have the computational capacity to learn and perform Bayesian inference (Müller et al., 2022), in practice LMs do not implement this strategy when interacting with a user in this task.

This work has two main contributions. First, we propose a framework for evaluating a system's ability to update its beliefs as new information becomes available; and second, we apply this framework to evaluate current LMs. We find that LMs struggle in this setting compared to the optimal Bayesian strategy, and that their performance does not improve as more information becomes available. As LMs are increasingly used in interactive settings where information is provided gradually, in particular in the use case of adaptation to individual users, we see this as a central challenge for LM development.

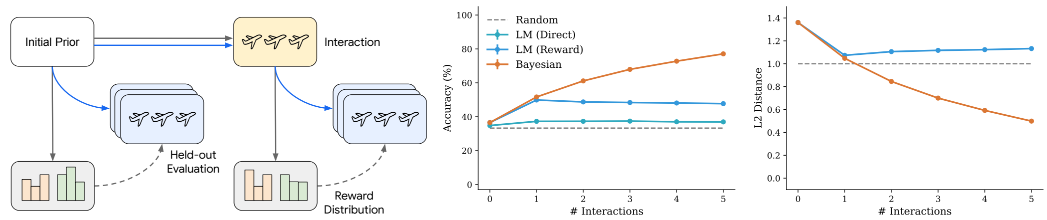
2 Evaluation Framework

2.1 The Flight Recommendation Task

We consider a simple flight recommendation setting similar to Lin et al. (2022), where a user interacts with an assistant for multiple rounds (Figure 1). During each round, a set of k flight options $\mathcal{O} = \{o_1, ..., o_k\}$ is presented to both the user and the assistant. Each flight option is represented by a feature vector $\phi(o) \in \mathbb{R}^4$, which indicates the departure time, the duration of the flight, the number of stops, and the cost of the flight. Each feature can take one of 100 values uniformly distributed between 0 and 1, except for the number of stops, which has 3 values. This defines 3×100^3 unique flight options, which we deterministically map to a textual representation (illustrated in Figure 1). We evaluate an alternative numerical representation in Section 6 .

Each user has their own flight preferences, defined by a reward function θ parameterized by four numbers, which represents their preferences for the aforementioned features. The space Θ of reward functions includes all four-dimensional vectors with the values $\{-1, -0.5, 0, 0.5, 1\}$, where -1 corresponds to a preference for low values of this feature and 1 to a preference for high values. We exclude the reward function $(0, 0, 0, 0)$, which does not provide any information. This results in a total of $5^4 - 1 = 624$ possible reward functions. We then instantiate a user by randomly sampling a vector $\theta \in \Theta$ as their reward function.

Given a set of flight options $\mathcal{O}$, the user determines which flight is the preferred one based on their reward function θ. Concretely, they assign the reward $r(o; \theta) = \theta^T \phi(o)$ to each flight o, and choose



(a) (b)

Figure 2: (a) **Experimental design.** We evaluate on held-out options at the end of each round. The evaluation branches out from the main interactions (that is, the evaluation performed after the first round is not included in the context of the second round). The LM's *direct* evaluation, where we ask the LM directly to choose a flight, follows the blue lines; the *reward* evaluation, where we first extract the LM's predicted reward function and then we use it to choose the flight, follows the gray lines. The dashed lines indicate the deterministic conversion of the LM's reward distribution into flight recommendations. (b) **Main results.** Left: the accuracy of the flight recommendations made by the Bayesian Assistant, directly by the LM, and using the reward function estimated from the LM. Right: the normalized L2 distance between the reward functions estimated from the models and the ground truth reward. The dashed lines indicates random performance.

the preferred flight as follows:

$$o^*(\mathcal{O}, \theta) = \arg\max_{o \in \mathcal{O}} r(o; \theta). \tag{1}$$

The goal of the assistant is to recommend the flight that best matches the user's preferences. At the end of each round, the assistant receives feedback from the user on whether or not it chose correctly; it also receives the correct answer. The next round then proceeds with a new set of options. We instantiate an LM assistant through prompting (see Table 4 for an example). We evaluate its ability to perform the flight recommendation task by *accuracy*, which measures whether the LM's directly predicted output aligns with the user's most preferred option.

2.2 The Bayesian Assistant

In each round the assistant receives a set of options along with the preferred option; this usually provides only partial information about the user's reward function. The optimal strategy is therefore to maintain a probability distribution over possible reward functions: if the assistant made a commitment to a single most likely one, this function could turn out to be incorrect in future rounds. After each round, we expect the optimal model to update its distribution over reward functions using Bayes' rule. This optimal model represents a ceiling on the performance that we can expect from any system, including the LMs we test.

Since the space of reward functions is relatively small, we can perform exact Bayesian updates. At each round, given options $\mathcal{O}$ and the user's preferred option o^*, the Bayesian Assistant updates its posterior using

$$q_B^{i+1}(\theta | \mathcal{O}^{i+1}, o^{*i+1}) = \frac{p(\mathcal{O}^{i+1}, o^{*i+1} | \theta) q_B^i(\theta)}{p(\mathcal{O}^{i+1}, o^{*i+1})}, \tag{2}$$

where the likelihood function indicates whether the reward function is consistent with user's choice:

$$p(\mathcal{O}, o^* | \theta) = \mathbb{1}\big[\max_{o \in \mathcal{O}} r(o; \theta) = o^*\big]. \tag{3}$$

The Bayesian assistant then makes flight recommendations based on its reward posterior mean, $\hat{\theta} = \mathbb{E}_{q(\theta)}[\theta]$, following Equation 1. As more information becomes available, the Bayesian Assistant will rule out more reward functions that are inconsistent with the user's choices, and will assign a higher probability to the user's true reward function, eventually converging on the correct one.

2.3 Assessing the LM's Beliefs

We approximate the LM's internal beliefs about the user's preferences by explicitly prompting it about them, using the user's previous booking history provided as context (see Table 5 for an example).

Specifically, we ask the LM to provide a rating on a scale of 1 to 5, where, for example 1 indicates a strong preference for cheaper flights, 3 indicate no strong preference, and 5 indicates a strong preference for expensive flights. We score the numbers 1, 2, 3, 4, and 5 as possible continuations of the prompt and re-normalize them to form a feature distribution over these five numbers. Then, we approximate the distribution over reward functions as a factorization of these feature distributions:

$$q_{LM}^{i+1}(\theta|\mathcal{O}^{i+1}, o^{*i+1}) \approx \prod_j q_{LM}^{i+1}(\theta_j|\mathcal{O}^{i+1}, o^{*i+1}). \tag{4}$$

where $q_{LM}^{i+1}(\theta_j|\mathcal{O}^{i+1}, o^{*i+1})$ is given by the next-token probability that the LM assigns to predicting each of the user's preferences for feature j followed by normalization. In practice this independence assumption may not be fully justified, as some features are likely to be correlated. We further note that we cannot be sure, of course, that the LM's responses to these prompts provide direct access to beliefs derived from an internal world model that guides its flight recommendations.

For the Bayesian Assistant's prior $q_B^0(\theta)$, we use the LM's initial reward distribution, i.e. the distribution derived from the LM's responses without providing any user-specific information. We take this distribution to capture people's general flight preferences (e.g., most people may prefer shorter flights over long ones).

2.4 Evaluation Metrics

We use three evaluation metrics. First, we record the accuracy of the LM's recommendations—whether the option it selected is the one that most closely matches the user's preferences. We refer to this evaluation metric as "direct" (see Figure 2a). For the Bayesian assistant, we record accuracy using the prediction obtained from the mean of its posterior distribution over reward functions, which is equivalent to making predictions using the posterior predictive distribution as $r(o;\theta)$ is a linear function. For consistency with this evaluation method, we derive a second LM accuracy metric, where we first extract the LM's posterior distribution over reward functions (as described in Section 2.3), and then use the same posterior predictive decision-making rule to obtain a recommendation. Finally, we compute the unit-normalized *L2 distance* between the mean of the posterior distribution over rewards and the ground-truth (GT) reward function, following Lin et al. (2022). We use normalized distance as some reward functions are equivalent. For example, the reward function $[-1, -1]$ is equivalent to the reward function $[-0.5, -0.5]$ as they always lead to the same recommendation.

3 Experimental Setup

For replicability, our experiments focus on open-weights Gemma 2 family of models (Team et al., 2024b). This family performs competitively: in Chatbot Arena leaderboard (Chiang et al., 2024), the instruction-tuned version with 9B parameters,[1] which we use for most of experiments, outperforms the similarly sized version of Llama 3.1 (Dubey et al., 2024) (and also outperforms OpenAI's GPT-4-0314), and the 27B-parameter version of Gemma 2 outperforms the 70B version of Llama 3.

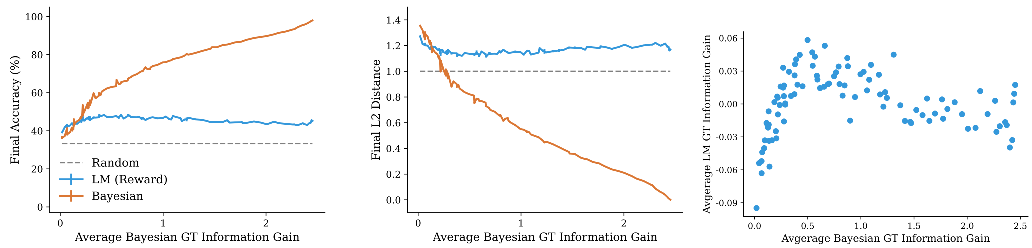
We instantiate users parametrized by each of the 624 reward function in the reward function space Θ (Section 2.1). We have the LM interact with the user for five rounds; in each round we present a set of three randomly sampled flight options. After each round, we evaluate the LM's recommendations up until that point on a held-out set of 100 randomly sampled option sets (see Figure 2a for the evaluation workflow). All experiments are carried out in-context. We evaluate predictions directly outputted from the LM ("direct") and predictions based on the reward distribution extract from the LM ("reward"; see Section 2.4). We use greedy decoding for the LMs. To reduce sensitivity to the specific randomly selected option sets, we average all experiments over three random seeds.[2]

4 Main Results

Overall, the quality of the LM-based flight recommendations is poor, either when we directly evaluate the flight recommendation provided by the LM ("direct"), or when we first extract its distribution

[1] `https://huggingface.co/google/gemma-2-9b-it`.

[2] The results were highly similar across seeds; all plots include error bars, but those are so small to be barely noticeable.



(a) (b)

Figure 3: (a) Effect of option set informativity on model performance. Option set informativity is defined as the increase in the log probability assigned by the Bayesian model to the ground truth reward function after observing the provided options. We plot accuracy (left) and reward L2 distance (right) after five interactions as a function of option set informativity averaged over the five interactions. (b) The relationship between the information gain computed from the Bayesian model and that computed from the LM.

over reward functions and use that distribution to simulate the user's preference ("reward"; results are shown in Figure 2b). Whereas the Bayesian model consistently improves its predictions as it receives more information, the LM's performance remains constant and only slightly better than random guessing, with the exception of a significant improvement after the first round of interaction. The LM's performance also does not improve even with 50-round interactions as shown in Table 10. We show a breakdown of results by reward functions in Appendix A, including a qualitative example. In terms of the L2 distance between the reward function extracted from the model and the ground truth reward function, the LM falls far behind the Bayesian Assistant; in fact, it performs worse than the random baseline.

While both LM-based methods are not very effective, the "reward" method does outperform the "direct" method. This suggests that instead of expecting the LM to directly solve the task, it may be beneficial to use it to compute beliefs about the state of the world, and then use those beliefs in a formal probabilistic framework (Wong et al., 2023). This method is more likely to be effective when the structure of the underlying state of the world is known and easy to specify, as it is in our simplified task.

5 Sensitivity to the Informativeness of Option Sets

Even if the number of interactions with the user is kept constant, the amount of information contained in a single interaction might vary. For example, two options that differ in only one feature dimension could be more informative than two options that differ along multiple dimensions: the minimal pair of options provides direct evidence for the user's preference for this feature. We expect an ideal probabilistic reasoner to be sensitive to this factor: when the user's choice between a particular set of options provides more information, we expect the system to update its beliefs more substantially.

In this section we test whether the LM displays this behavior. Whereas before we sampled the option sets randomly, here we sample them based on their informativeness. To measure the amount of information contained in a set of options $\mathcal{O}$, we define the *ground truth information gain* as

$$\begin{aligned} g(\mathcal{O}, o^*, p(\theta), q(\theta)) &= \mathrm{KL}(p(\theta)||q(\theta)) - \mathrm{KL}(p(\theta)||q(\theta|\mathcal{O}, o^*)) && (5) \\ &= \log q(\theta^*|\mathcal{O}, o^*) - \log q(\theta^*), && (6) \end{aligned}$$

where $p(\theta) = \delta(\theta^*)$ and $q(\theta)$ is either $q_B(\theta)$ or $q_{LM}(\theta)$. This metric captures the increase in the posterior probability of the ground truth reward function after this set of options has been observed. Note that g is relative to the model that is used to update the probability distribution; we use g_B and g_{LM} to refer to the gain derived from the Bayesian model and the LM, respectively.

Experimental setup. We randomly sample 5000 candidate option sets, compute the ground truth information gain of each one based on the Bayesian model, and select the option set that leads to the desirable value of g_B. The performance is evaluated at the end of a 5-round interaction, and the ground truth information gain is averaged over these five rounds.

Results. The Bayesian model's performance consistently improves as option sets become more informative: after observing highly informative options, its performance is almost perfect (Figure 3a).

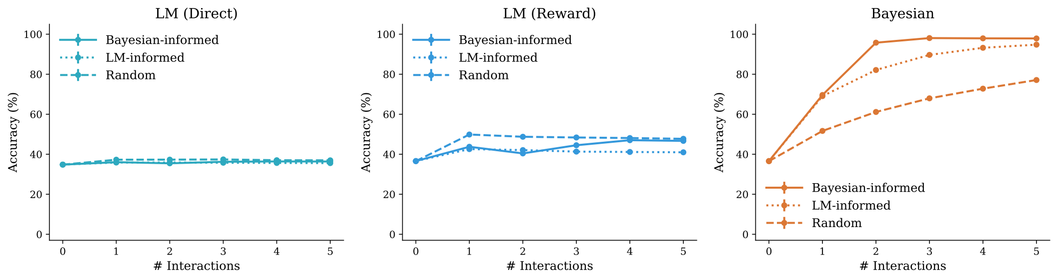


Figure 4: Comparison of using Bayesian-informed, LM-informed, and random options. We show the performance of the LM's direct predictions (left), predictions derived from the LM's reward distribution (middle), and predictions derived from the Bayesian model's reward distribution (right).

By contrast, the LM does not show sensitivity to option set informativity. There appears to be a slight advantage to the LM over the Bayesian model when the options are very uninformative; since this difference is very small we refrain from speculating on its causes.

LM-derived vs. Bayesian information gain. Recall that information gain is relative to the model that is used to update the probability distributions: g_{LM} quantifies the amount of information the LM can absorb from a particular set of options, whereas g_B quantifies the amount that an ideal reasoner can absorb. Figure 3b illustrates how these measures relate to each other. When g_B is small, there is a positive relationship between the two metrics, indicating that more informative options are beneficial for the LM. However, this positive correlation does not persist when g_B is large; if anything, in that part of the range the correlation appears to be slightly negative. We leave an exploration of this finding for future work.

Selecting options based on the LM-derived information gain. Here, we consider a pedagogical setting where the user chooses option sets that would maximally help the assistant to learn their preferences (Shafto et al., 2014; Rafferty et al., 2015; Ross and Andreas, 2024). We assume the user has access to the LM's reward distribution; they use that distribution to identify the option set which would lead to the maximal g_{LM}. We simulate this by giving the LM options that maximize the posterior probability of the ground-truth reward function, using the LM's current reward distribution as the prior. Ideally, to select the most informative options, we would prompt the LM for its posterior reward distributions for all candidates to obtain its reward posteriors. Since this approach is computational expensive, however, we approximate this by using Bayesian updates. Surprisingly, we find that providing LM-informed options hurts the LM's performance (Figure 4). Since we make various assumptions and approximations in this process, it is not guaranteed that these options are optimally informative for the LM. We leave exploring the selection of optimal options for future work.

6 Additional Analyses

Table 1: Results using different flight representations. We compare the textual representation, which uses natural language descriptions deterministically converted from the feature values, and the numerical representation, which directly uses the feature values.

	Direct Accuracy (%)						**Reward Accuracy (%)**						**L2 Distance**					
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
Text	34.8	37.3	37.3	37.4	37.0	37.0	36.5	49.9	48.7	48.4	48.1	47.7	1.4	1.1	1.1	1.1	1.1	1.1
Num	33.6	34.3	34.3	34.2	34.2	34.1	34.9	40.1	39.9	38.9	39.4	38.5	1.4	1.3	1.3	1.3	1.3	1.3

Is the LM's poor performance due to an inability to parse the flight representation? Our main experiments use a textual representation that deterministically maps the feature value of each flight to a text description. While this textual representation is closer to realistic scenarios, and may therefore better align with the LM's training distribution, this setup introduces a potential confounder that complicates the interpretation of our results: the LM's poor performance in the flight recommendation task could be due to its inability to translate the text description into the feature space that is required for reasoning.To control for this factor, we investigate an alternative numerical representation of the flight options, where we directly provide the LM with the feature values (float numbers), in the same way we provide them to the Bayesian Assistant (see Table 6 and Table 7 for examples). We find that, if anything, the textual representation *outperforms* its numerical counterpart in both the task accuracy and the reward inference metrics (Table 1). This suggests that the LM's poor performance cannot be attributed to an inability to parse the textual input to numerical value.

Table 2: Results using different methods to obtain the LM's distribution over reward functions. We compare scoring model continuations with directly asking the LM to generate probability judgments.

	Direct Accuracy (%)						Reward Accuracy (%)						L2 Distance					
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
Scoring	34.8	37.3	37.3	37.4	37.0	37.0	36.5	49.9	48.7	48.4	48.1	47.7	1.4	1.1	1.1	1.1	1.1	1.1
Generation	34.8	37.2	37.4	37.4	37.1	37.0	36.1	47.9	44.3	42.3	43.1	42.5	1.4	1.1	1.2	1.2	1.2	1.2

Scoring continuations leads to more accurate probability estimates than explicit probability judgments. In the main experiment, we estimated the LM's distribution over reward functions by prompting it to rate individual features and scoring the possible continuations; for flight duration, for example, we might ask it what the user's preference is on a scale of 1 to 5. We refer to this method as "scoring". Here, we compare this method to one where we instruct the LM to assign a probability for each of the five ratings on each scale ("generation"; see Table 8 for an example). As in the scoring method, we renormalized the probabilities to sum to 1 (though they typically summed to 1 even before normalization). Overall, the scoring-based reward distribution that is closer the generation-based one to the ground truth distribution (Table 2).

Table 3: We compare instruction-tuned Gemma 2 9B with instruction-tuned Gemma 2 27B.

	Direct Accuracy (%)						Reward Accuracy (%)						L2 Distance					
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
9B	34.8	37.3	37.3	37.4	37.0	37.0	36.5	49.9	48.7	48.4	48.1	47.7	1.4	1.1	1.1	1.1	1.1	1.1
27B	34.2	37.7	39.0	39.6	39.8	40.2	36.3	50.5	50.5	49.1	49.4	48.6	1.4	1.1	1.1	1.1	1.1	1.1

A larger LM does not perform better. Finally, we use our paradigm to evaluate a larger LM, the instruction-tuned Gemma 2 model with 27B parameters, and find that this model performs similarly to the 9B-parameter one (Table 3). There is no evidence, then, that simply scaling up the model size would improve performance in this paradigm.[3] We also include preliminary results using a stronger closed-source model, Gemini 1.5 Pro, in Appendix A.1. While it achieves better performance than the Gemma 2 model, the overall trend remains similar and it lags considerably behind the Bayesian model.

7 Related Work

LMs and probabilistic inference. Existing studies have explored if LMs can perform probabilistic inference from different perspectives. Several prior works have studied how in-context learning can be viewed as implicit Bayesian inference (Xie et al., 2022; Hahn and Goyal, 2023; Jiang, 2023), but see Falck et al. (2024) for a counter example. Many studies have also investigated LMs' probabilistic reasoning capabilities (Nafar et al., 2024; Paruchuri et al., 2024), but most focus on asking LMs to computing statistics explicitly. Other research evaluates LMs' abilities to provide probability judgements (Zhu and Griffiths, 2024; Belem et al., 2024). Our evaluation differs from these settings as we simulate a realistic setting where implicit reasoning about probability is beneficial.

A related line of work has explored leveraging LMs for better probabilistic inference (Feng et al., 2024b; Liu et al., 2024; Piriyakulkij and Ellis, 2024; Grand et al., 2023; Ying et al., 2024; Ellis, 2023). These studies typically adopt a neuro-symbolic approach, where LMs propose and evaluate plausible hypotheses or translate natural language into probabilistic programs, and then use existing inference algorithms to perform probabilistic inference. Our "reward" evaluation is related to this approach.

LMs and world models. Several works have investigated whether LMs learn implicit world models. Some studies find that the representations of pre-trained LMs can be mapped to meaningful conceptual spaces (Patel and Pavlick, 2022; Li et al., 2021; Abdou et al., 2021; Feng et al., 2024a). Other works have trained LMs on synthetic structured data and used probes to reconstruct world states from their internal representations (Li et al., 2023b; Toshniwal et al., 2022; Jin and Rinard, 2024; Hazineh et al.,

[3]According to the Chatbot Arena this model is outperforms Llama 3 70B Instruct (Chiang et al., 2024).

2023; Kuo et al., 2023; Vafa et al., 2024). Most of these studies only probe static world models. We similarly investigate whether LMs construct models of the worlds, but we evaluate them in a dynamic setting where LMs continually receive more information. This setup naturally allows us to further investigate whether LMs can update their beliefs based on new observations. Hase et al. (2024) similarly study belief revision in LMs but focus on model editing.

Preference learning. Inferring reward functions from observations has been widely studied in reinforcement learning (Ng and Russell, 2000; Abbeel and Ng, 2004; Christiano et al., 2017; Ziebart et al., 2008). Recent studies have explored using LMs to elicit user preferences (Li et al., 2023a; Handa et al., 2024; Piriyakulkij et al., 2023; Andukuri et al., 2024; Peng et al., 2024; Aliannejadi et al., 2021; Chen et al., 2024; Lin et al., 2022). Some studies further combine LMs with other approaches, such as Bayesian models, to select informative questions (Handa et al., 2024; Piriyakulkij et al., 2023; Austin et al., 2024). These approaches typically consider an active setting where LMs interact with users. However, in practice, such processes may be time-consuming and expensive. In contrast, we consider a passive setting where LMs only receive observations without asking open-ended questions.

Another line of work has also explored leveraging LMs for recommendation systems (Tsai et al., 2024; Korikov et al., 2024; Ji et al., 2024; Lyu et al., 2023). While we use content recommendation as our task, our focus is on evaluating whether LMs can infer latent models and update their beliefs based on new information. Liu et al. (2024) conducted similar experiments where an LM is asked to infer human preferences from their choices and found strong correlations between LMs and humans (Jern et al., 2017). However, they only evaluated the LM's task performance without assessing its beliefs.

8 Limitations and Future Work

Models. We only evaluate on the Gemma 2 family of models. While these are highly competitive open models (Chiang et al., 2024), we do not claim that our findings will necessarily transfer to other models. Open models are essential for our goals as we need access to log probabilities to obtain the LM's reward distribution; this information is often unavailable for closed-source models. Closed-source models can be evaluated using other methods for estimating the reward distribution, such as directly generating probability judgments (Section 6) or Monte Carlo sampling. We include preliminary experiments on a stronger closed-source model in Appendix A.1 and leave more comprehensive evaluations for future work.

Task. Our main interest is to evaluate whether LMs update their beliefs based on new information. We focus on the flight recommendation task as it provides a controlled setting where LMs continually receive new information, and it allows us to perform exact Bayesian updates, providing an upper bound on performance given current observations. An important direction for future work is to extend our paradigm to settings that involve inference over more complex world states (Wong et al., 2023).

Evaluation. We use only greedy decoding for LM generation to ensure reproducibility. Sampling with higher temperatures might lead to different results. We assess the LM's reward distribution through scoring and generation, but it is unclear whether these are the optimal methods. Other approaches, such as training probes, could also be possible, which we leave for future work. Finally, we only uses standard prompting; other prompting techniques, such as chain-of-thought prompting (Nye et al., 2021; Wei et al., 2022; Kojima et al., 2022), have been shown useful for some reasoning tasks. Future studies could investigate whether such methods improve performance.

9 Conclusion

In this paper, we study whether LMs can update their beliefs based on new observations. We evaluate this using a controllable flight recommendation task, where a booking assistant needs to infer the user's preferences based on previous interactions to provide good recommendations. We compare the LM assistant with an optimal Bayesian model. We find that LMs significantly underperform compared to the Bayesian model, and their performance does not improve as more information becomes available. We further analyze the relationship between performance and the amount of information contained in the option sets. We find that the informativeness of options has a significant impact on the Bayesian model, but not on the LMs. Our study highlights the limitations of LMs in performing implicit Bayesian inference over world states, emphasizing the importance of improving their probabilistic reasoning abilities for future work.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In C. E. Brodley, editor, *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004. doi: 10.1145/1015330.1015430. URL https://doi.org/10.1145/1015330.1015430.

M. Abdou, A. Kulmizev, D. Hershcovich, S. Frank, E. Pavlick, and A. Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.9. URL https://aclanthology.org/2021.conll-1.9.

M. Aliannejadi, J. Kiseleva, A. Chuklin, J. Dalton, and M. Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.367. URL https://aclanthology.org/2021.emnlp-main.367.

C. Andukuri, J.-P. Fränken, T. Gerstenberg, and N. D. Goodman. Star-gate: Teaching language models to ask clarifying questions. *ArXiv preprint*, abs/2403.19154, 2024. URL https://arxiv.org/abs/2403.19154.

D. E. Austin, A. Korikov, A. Toroghi, and S. Sanner. Bayesian optimization with llm-based acquisition functions for natural language preference elicitation. *ArXiv preprint*, abs/2405.00981, 2024. URL https://arxiv.org/abs/2405.00981.

C. Baker and R. Saxe. Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*, 2011.

C. G. Belem, M. Kelly, M. Steyvers, S. Singh, and P. Smyth. Perceptions of linguistic uncertainty by language models and humans. *ArXiv preprint*, abs/2407.15814, 2024. URL https://arxiv.org/abs/2407.15814.

S. Chen, S. Wiseman, and B. Dhingra. Chatshop: Interactive information seeking with language agents. *ArXiv preprint*, abs/2404.09911, 2024. URL https://arxiv.org/abs/2404.09911.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *ArXiv preprint*, abs/2403.04132, 2024. URL https://arxiv.org/abs/2403.04132.

P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun,

J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. E. Tan, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Grattafiori, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Vaughan, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Franco, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Wyatt, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Ozgenel, F. Caggioni, F. Guzmán, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Thattai, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, I. Damlaj, I. Molybog, I. Tufanov, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Prasad, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Huang, K. Chawla, K. Lakhotia, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Tsimpoukelli, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. P. Laptev, N. Dong, N. Zhang, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Li, R. Hogan, R. Battey, R. Wang, R. Maheswari, R. Howes, R. Rinott, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Kohler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Albiero, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wang, X. Wu, X. Wang, X. Xia, X. Wu, X. Gao, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Hao, Y. Qian, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, and Z. Zhao. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

K. Ellis. Human-like few-shot learning via bayesian reasoning over natural language. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 13149–13178, 2023.

F. Falck, Z. Wang, and C. C. Holmes. Are large language models bayesian? a martingale perspective on in-context learning. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language*

Models, 2024.

J. Feng, S. Russell, and J. Steinhardt. Monitoring latent world states in language models with propositional probes. *ArXiv preprint*, abs/2406.19501, 2024a. URL https://arxiv.org/abs/2406.19501.

Y. Feng, B. Zhou, W. Lin, and D. Roth. Bird: A trustworthy bayesian inference framework for large language models. *ArXiv preprint*, abs/2404.12494, 2024b. URL https://arxiv.org/abs/2404.12494.

G. Grand, V. Pepe, J. Andreas, and J. Tenenbaum. Loose lips sink ships: Asking questions in battleship with language-informed program sampling. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2023.

M. Hahn and N. Goyal. A theory of emergent in-context learning as implicit structure induction. *ArXiv preprint*, abs/2303.07971, 2023. URL https://arxiv.org/abs/2303.07971.

K. Handa, Y. Gal, E. Pavlick, N. Goodman, J. Andreas, A. Tamkin, and B. Z. Li. Bayesian preference elicitation with language models. *ArXiv preprint*, abs/2403.05534, 2024. URL https://arxiv.org/abs/2403.05534.

P. Hase, T. Hofweber, X. Zhou, E. Stengel-Eskin, and M. Bansal. Fundamental problems with model editing: How should rational belief revision work in llms? *ArXiv preprint*, abs/2406.19354, 2024. URL https://arxiv.org/abs/2406.19354.

D. S. Hazineh, Z. Zhang, and J. Chiu. Linear latent world models in simple transformers: A case study on othello-gpt. *ArXiv preprint*, abs/2310.07582, 2023. URL https://arxiv.org/abs/2310.07582.

A. Jern, C. G. Lucas, and C. Kemp. People learn other people's preferences through inverse decision-making. *Cognition*, 168:46–64, 2017. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2017.06.017. URL https://www.sciencedirect.com/science/article/pii/S0010027717301798.

J. Ji, Z. Li, S. Xu, W. Hua, Y. Ge, J. Tan, and Y. Zhang. Genrec: Large language model for generative recommendation. In *European Conference on Information Retrieval*, pages 494–502. Springer, 2024.

H. Jiang. A latent space theory for emergent abilities in large language models. *ArXiv preprint*, abs/2304.09960, 2023. URL https://arxiv.org/abs/2304.09960.

C. Jin and M. Rinard. Emergent representations of program semantics in language models trained on programs. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=8PTx4CpNoT.

P. N. Johnson-Laird. Mental models in cognitive science. *Cognitive science*, 4(1):71–115, 1980.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

A. Korikov, S. Sanner, Y. Deldjoo, Z. He, J. McAuley, A. Ramisa, R. Vidal, M. Sathiamoorthy, A. Kasrizadeh, S. Milano, et al. Large language model driven recommendation. *ArXiv preprint*, abs/2408.10946, 2024. URL https://arxiv.org/abs/2408.10946.

M.-T. Kuo, C.-C. Hsueh, and R. T.-H. Tsai. Large language models on the chessboard: A study on chatgpt's formal language comprehension and complex reasoning skills. *ArXiv preprint*, abs/2308.15118, 2023. URL https://arxiv.org/abs/2308.15118.

B. Z. Li, M. Nye, and J. Andreas. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.143. URL https://aclanthology.org/2021.acl-long.143.

B. Z. Li, A. Tamkin, N. Goodman, and J. Andreas. Eliciting human preferences with language models. *ArXiv preprint*, abs/2310.11589, 2023a. URL https://arxiv.org/abs/2310.11589.

K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *International Conference on Learning Representations*, 2023b.

J. Lin, D. Fried, D. Klein, and A. Dragan. Inferring rewards from language in context. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8546–8560, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.585. URL https://aclanthology.org/2022.acl-long.585.

R. Liu, J. Geng, J. C. Peterson, I. Sucholutsky, and T. L. Griffiths. Large language models assume people are more rational than we really are. *ArXiv preprint*, abs/2406.17055, 2024. URL https://arxiv.org/abs/2406.17055.

H. Lyu, S. Jiang, H. Zeng, Y. Xia, Q. Wang, S. Zhang, R. Chen, C. Leung, J. Tang, and J. Luo. Llm-rec: Personalized recommendation via prompting large language models. *ArXiv preprint*, abs/2307.15780, 2023. URL https://arxiv.org/abs/2307.15780.

S. Müller, N. Hollmann, S. Pineda-Arango, J. Grabocka, and F. Hutter. Transformers can do bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

A. Nafar, K. B. Venable, and P. Kordjamshidi. Probabilistic reasoning in generative large language models. *ArXiv preprint*, abs/2402.09614, 2024. URL https://arxiv.org/abs/2402.09614.

A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In P. Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 663–670. Morgan Kaufmann, 2000.

M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *ArXiv preprint*, abs/2112.00114, 2021. URL https://arxiv.org/abs/2112.00114.

A. Paruchuri, J. Garrison, S. Liao, J. Hernandez, J. Sunshine, T. Althoff, X. Liu, and D. McDuff. What are the odds? language models are capable of probabilistic reasoning. *ArXiv preprint*, abs/2406.12830, 2024. URL https://arxiv.org/abs/2406.12830.

R. Patel and E. Pavlick. Mapping language models to grounded conceptual spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=gJcEM8sxHK.

A. Peng, Y. Sun, T. Shu, and D. Abel. Pragmatic feature preferences: Learning reward-relevant preferences from human input. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=OgG0I5toZZ.

T. Piriyakulkij and K. Ellis. Doing experiments and revising rules with natural language and probabilistic reasoning. *ArXiv preprint*, abs/2402.06025, 2024. URL https://arxiv.org/abs/2402.06025.

T. Piriyakulkij, V. Kuleshov, and K. Ellis. Active preference inference using language models and probabilistic reasoning. *ArXiv preprint*, abs/2312.12009, 2023. URL https://arxiv.org/abs/2312.12009.

A. N. Rafferty, M. M. LaMar, and T. L. Griffiths. Inferring learners' knowledge from their actions. *Cognitive Science*, 39(3):584–618, 2015. doi: 10.1111/cogs.12157.

A. Ross and J. Andreas. Toward in-context teaching: Adapting examples to students' misconceptions. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13283–13310, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.acl-long.718`.

P. Shafto, N. D. Goodman, and T. L. Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55–89, 2014. ISSN 0010-0285. doi: https://doi.org/10.1016/j.cogpsych.2013.12.004. URL `https://www.sciencedirect.com/science/article/pii/S0010028514000024`.

G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, S. Mariooryad, Y. Ding, X. Geng, F. Alcober, R. Frostig, M. Omernick, L. Walker, C. Paduraru, C. Sorokin, A. Tacchetti, C. Gaffney, S. Daruki, O. Sercinoglu, Z. Gleicher, J. Love, P. Voigtlaender, R. Jain, G. Surita, K. Mohamed, R. Blevins, J. Ahn, T. Zhu, K. Kawintiranon, O. Firat, Y. Gu, Y. Zhang, M. Rahtz, M. Faruqui, N. Clay, J. Gilmer, J. Co-Reyes, I. Penchev, R. Zhu, N. Morioka, K. Hui, K. Haridasan, V. Campos, M. Mahdieh, M. Guo, S. Hassan, K. Kilgour, A. Vezer, H.-T. Cheng, R. de Liedekerke, S. Goyal, P. Barham, D. Strouse, S. Noury, J. Adler, M. Sundararajan, S. Vikram, D. Lepikhin, M. Paganini, X. Garcia, F. Yang, D. Valter, M. Trebacz, K. Vodrahalli, C. Asawaroengchai, R. Ring, N. Kalb, L. B. Soares, S. Brahma, D. Steiner, T. Yu, F. Mentzer, A. He, L. Gonzalez, B. Xu, R. L. Kaufman, L. E. Shafey, J. Oh, T. Hennigan, G. van den Driessche, S. Odoom, M. Lucic, B. Roelofs, S. Lall, A. Marathe, B. Chan, S. Ontanon, L. He, D. Teplyashin, J. Lai, P. Crone, B. Damoc, L. Ho, S. Riedel, K. Lenc, C.-K. Yeh, A. Chowdhery, Y. Xu, M. Kazemi, E. Amid, A. Petrushkina, K. Swersky, A. Khodaei, G. Chen, C. Larkin, M. Pinto, G. Yan, A. P. Badia, P. Patil, S. Hansen, D. Orr, S. M. R. Arnold, J. Grimstad, A. Dai, S. Douglas, R. Sinha, V. Yadav, X. Chen, E. Gribovskaya, J. Austin, J. Zhao, K. Patel, P. Komarek, S. Austin, S. Borgeaud, L. Friso, A. Goyal, B. Caine, K. Cao, D.-W. Chung, M. Lamm, G. Barth-Maron, T. Kagohara, K. Olszewska, M. Chen, K. Shivakumar, R. Agarwal, H. Godhia, R. Rajwar, J. Snaider, X. Dotiwalla, Y. Liu, A. Barua, V. Ungureanu, Y. Zhang, B.-O. Batsaikhan, M. Wirth, J. Qin, I. Danihelka, T. Doshi, M. Chadwick, J. Chen, S. Jain, Q. Le, A. Kar, M. Gurumurthy, C. Li, R. Sang, F. Liu, L. Lamprou, R. Munoz, N. Lintz, H. Mehta, H. Howard, M. Reynolds, L. Aroyo, Q. Wang, L. Blanco, A. Cassirer, J. Griffith, D. Das, S. Lee, J. Sygnowski, Z. Fisher, J. Besley, R. Powell, Z. Ahmed, D. Paulus, D. Reitter, Z. Borsos, R. Joshi, A. Pope, S. Hand, V. Selo, V. Jain, N. Sethi, M. Goel, T. Makino, R. May, Z. Yang, J. Schalkwyk, C. Butterfield, A. Hauth, A. Goldin, W. Hawkins, E. Senter, S. Brin, O. Woodman, M. Ritter, E. Noland, M. Giang, V. Bolina, L. Lee, T. Blyth, I. Mackinnon, M. Reid, O. Sarvana, D. Silver, A. Chen, L. Wang, L. Maggiore, O. Chang, N. Attaluri, G. Thornton, C.-C. Chiu, O. Bunyan, N. Levine, T. Chung, E. Eltyshev, X. Si, T. Lillicrap, D. Brady, V. Aggarwal, B. Wu, Y. Xu, R. McIlroy, K. Badola, P. Sandhu, E. Moreira, W. Stokowiec, R. Hemsley, D. Li, A. Tudor, P. Shyam, E. Rahimtoroghi, S. Haykal, P. Sprechmann, X. Zhou, D. Mincu, Y. Li, R. Addanki, K. Krishna, X. Wu, A. Frechette, M. Eyal, A. Dafoe, D. Lacey, J. Whang, T. Avrahami, Y. Zhang, E. Taropa, H. Lin, D. Toyama, E. Rutherford, M. Sano, H. Choe, A. Tomala, C. Safranek-Shrader, N. Kassner, M. Pajarskas, M. Harvey, S. Sechrist, M. Fortunato, C. Lyu, G. Elsayed, C. Kuang, J. Lottes, E. Chu, C. Jia, C.-W. Chen, P. Humphreys, K. Baumli, C. Tao, R. Samuel, C. N. dos Santos, A. Andreassen, N. Rakićević, D. Grewe, A. Kumar, S. Winkler, J. Caton, A. Brock, S. Dalmia, H. Sheahan, I. Barr, Y. Miao, P. Natsev, J. Devlin, F. Behbahani, F. Prost, Y. Sun, A. Myaskovsky, T. S. Pillai, D. Hurt, A. Lazaridou, X. Xiong, C. Zheng, F. Pardo, X. Li, D. Horgan, J. Stanton, M. Ambar, F. Xia, A. Lince, M. Wang, B. Mustafa, A. Webson, H. Lee, R. Anil, M. Wicke, T. Dozat, A. Sinha, E. Piqueras, E. Dabir, S. Upadhyay, A. Boral, L. A. Hendricks, C. Fry, J. Djolonga, Y. Su, J. Walker, J. Labanowski, R. Huang, V. Misra, J. Chen, R. Skerry-Ryan, A. Singh, S. Rijhwani, D. Yu, A. Castro-Ros, B. Changpinyo, R. Datta, S. Bagri, A. M. Hrafnkelsson, M. Maggioni, D. Zheng, Y. Sulsky, S. Hou, T. L. Paine, A. Yang, J. Riesa, D. Rogozinska, D. Marcus, D. E. Badawy, Q. Zhang, L. Wang, H. Miller, J. Greer, L. L. Sjos, A. Nova, H. Zen, R. Chaabouni, M. Rosca, J. Jiang, C. Chen, R. Liu, T. Sainath, M. Krikun, A. Polozov, J.-B. Lespiau, J. Newlan, Z. Cankara, S. Kwak, Y. Xu, P. Chen, A. Coenen, C. Meyer, K. Tsihlas, A. Ma, J. Gottweis, J. Xing, C. Gu, J. Miao, C. Frank, Z. Cankara, S. Ganapathy, I. Dasgupta, S. Hughes-Fitt, H. Chen, D. Reid, K. Rong, H. Fan, J. van Amersfoort, V. Zhuang, A. Cohen, S. S. Gu, A. Mohananey, A. Ilic, T. Tobin, J. Wieting, A. Bortsova, P. Thacker, E. Wang, E. Caveness, J. Chiu, E. Sezener, A. Kaskasoli, S. Baker, K. Millican, M. Elhawaty, K. Aisopos, C. Lebsack, N. Byrd, H. Dai, W. Jia, M. Wiethoff, E. Davoodi, A. Weston, L. Yagati, A. Ahuja, I. Gao, G. Pundak, S. Zhang, M. Azzam, K. C. Sim, S. Caelles, J. Keeling, A. Sharma, A. Swing,

Y. Li, C. Liu, C. G. Bostock, Y. Bansal, Z. Nado, A. Anand, J. Lipschultz, A. Karmarkar, L. Proleev, A. Ittycheriah, S. H. Yeganeh, G. Polovets, A. Faust, J. Sun, A. Rrustemi, P. Li, R. Shivanna, J. Liu, C. Welty, F. Lebron, A. Baddepudi, S. Krause, E. Parisotto, R. Soricut, Z. Xu, D. Bloxwich, M. Johnson, B. Neyshabur, J. Mao-Jones, R. Wang, V. Ramasesh, Z. Abbas, A. Guez, C. Segal, D. D. Nguyen, J. Svensson, L. Hou, S. York, K. Milan, S. Bridgers, W. Gworek, M. Tagliasacchi, J. Lee-Thorp, M. Chang, A. Guseynov, A. J. Hartman, M. Kwong, R. Zhao, S. Kashem, E. Cole, A. Miech, R. Tanburn, M. Phuong, F. Pavetic, S. Cevey, R. Comanescu, R. Ives, S. Yang, C. Du, B. Li, Z. Zhang, M. Iinuma, C. H. Hu, A. Roy, S. Bijwadia, Z. Zhu, D. Martins, R. Saputro, A. Gergely, S. Zheng, D. Jia, I. Antonoglou, A. Sadovsky, S. Gu, Y. Bi, A. Andreev, S. Samangooei, M. Khan, T. Kocisky, A. Filos, C. Kumar, C. Bishop, A. Yu, S. Hodkinson, S. Mittal, P. Shah, A. Moufarek, Y. Cheng, A. Bloniarz, J. Lee, P. Pejman, P. Michel, S. Spencer, V. Feinberg, X. Xiong, N. Savinov, C. Smith, S. Shakeri, D. Tran, M. Chesus, B. Bohnet, G. Tucker, T. von Glehn, C. Muir, Y. Mao, H. Kazawa, A. Slone, K. Soparkar, D. Shrivastava, J. Cobon-Kerr, M. Sharman, J. Pavagadhi, C. Araya, K. Misiunas, N. Ghelani, M. Laskin, D. Barker, Q. Li, A. Briukhov, N. Houlsby, M. Glaese, B. Lakshminarayanan, N. Schucher, Y. Tang, E. Collins, H. Lim, F. Feng, A. Recasens, G. Lai, A. Magni, N. D. Cao, A. Siddhant, Z. Ashwood, J. Orbay, M. Dehghani, J. Brennan, Y. He, K. Xu, Y. Gao, C. Saroufim, J. Molloy, X. Wu, S. Arnold, S. Chang, J. Schrittwieser, E. Buchatskaya, S. Radpour, M. Polacek, S. Giordano, A. Bapna, S. Tokumine, V. Hellendoorn, T. Sottiaux, S. Cogan, A. Severyn, M. Saleh, S. Thakoor, L. Shefey, S. Qiao, M. Gaba, S. yiin Chang, C. Swanson, B. Zhang, B. Lee, P. K. Rubenstein, G. Song, T. Kwiatkowski, A. Koop, A. Kannan, D. Kao, P. Schuh, A. Stjerngren, G. Ghiasi, G. Gibson, L. Vilnis, Y. Yuan, F. T. Ferreira, A. Kamath, T. Klimenko, K. Franko, K. Xiao, I. Bhattacharya, M. Patel, R. Wang, A. Morris, R. Strudel, V. Sharma, P. Choy, S. H. Hashemi, J. Landon, M. Finkelstein, P. Jhakra, J. Frye, M. Barnes, M. Mauger, D. Daun, K. Baatarsukh, M. Tung, W. Farhan, H. Michalewski, F. Viola, F. de Chaumont Quitry, C. L. Lan, T. Hudson, Q. Wang, F. Fischer, I. Zheng, E. White, A. Dragan, J. baptiste Alayrac, E. Ni, A. Pritzel, A. Iwanicki, M. Isard, A. Bulanova, L. Zilka, E. Dyer, D. Sachan, S. Srinivasan, H. Muckenhirn, H. Cai, A. Mandhane, M. Tariq, J. W. Rae, G. Wang, K. Ayoub, N. FitzGerald, Y. Zhao, W. Han, C. Alberti, D. Garrette, K. Krishnakumar, M. Gimenez, A. Levskaya, D. Sohn, J. Matak, I. Iturrate, M. B. Chang, J. Xiang, Y. Cao, N. Ranka, G. Brown, A. Hutter, V. Mirrokni, N. Chen, K. Yao, Z. Egyed, F. Galilee, T. Liechty, P. Kallakuri, E. Palmer, S. Ghemawat, J. Liu, D. Tao, C. Thornton, T. Green, M. Jasarevic, S. Lin, V. Cotruta, Y.-X. Tan, N. Fiedel, H. Yu, E. Chi, A. Neitz, J. Heitkaemper, A. Sinha, D. Zhou, Y. Sun, C. Kaed, B. Hulse, S. Mishra, M. Georgaki, S. Kudugunta, C. Farabet, I. Shafran, D. Vlasic, A. Tsitsulin, R. Ananthanarayanan, A. Carin, G. Su, P. Sun, S. V, G. Carvajal, J. Broder, I. Comsa, A. Repina, W. Wong, W. W. Chen, P. Hawkins, E. Filonov, L. Loher, C. Hirnschall, W. Wang, J. Ye, A. Burns, H. Cate, D. G. Wright, F. Piccinini, L. Zhang, C.-C. Lin, I. Gog, Y. Kulizhskaya, A. Sreevatsa, S. Song, L. C. Cobo, A. Iyer, C. Tekur, G. Garrido, Z. Xiao, R. Kemp, H. S. Zheng, H. Li, A. Agarwal, C. Ngani, K. Goshvadi, R. Santamaria-Fernandez, W. Fica, X. Chen, C. Gorgolewski, S. Sun, R. Garg, X. Ye, S. M. A. Eslami, N. Hua, J. Simon, P. Joshi, Y. Kim, I. Tenney, S. Potluri, L. N. Thiet, Q. Yuan, F. Luisier, A. Chronopoulou, S. Scellato, P. Srinivasan, M. Chen, V. Koverkathu, V. Dalibard, Y. Xu, B. Saeta, K. Anderson, T. Sellam, N. Fernando, F. Huot, J. Jung, M. Varadarajan, M. Quinn, A. Raul, M. Le, R. Habalov, J. Clark, K. Jalan, K. Bullard, A. Singhal, T. Luong, B. Wang, S. Rajayogam, J. Eisenschlos, J. Jia, D. Finchelstein, A. Yakubovich, D. Balle, M. Fink, S. Agarwal, J. Li, D. Dvijotham, S. Pal, K. Kang, J. Konzelmann, J. Beattie, O. Dousse, D. Wu, R. Crocker, C. Elkind, S. R. Jonnalagadda, J. Lee, D. Holtmann-Rice, K. Kallarackal, R. Liu, D. Vnukov, N. Vats, L. Invernizzi, M. Jafari, H. Zhou, L. Taylor, J. Prendki, M. Wu, T. Eccles, T. Liu, K. Kopparapu, F. Beaufays, C. Angermueller, A. Marzoca, S. Sarcar, H. Dib, J. Stanway, F. Perbet, N. Trdin, R. Sterneck, A. Khorlin, D. Li, X. Wu, S. Goenka, D. Madras, S. Goldshtein, W. Gierke, T. Zhou, Y. Liu, Y. Liang, A. White, Y. Li, S. Singh, S. Bahargam, M. Epstein, S. Basu, L. Lao, A. Ozturel, C. Crous, A. Zhai, H. Lu, Z. Tung, N. Gaur, A. Walton, L. Dixon, M. Zhang, A. Globerson, G. Uy, A. Bolt, O. Wiles, M. Nasr, I. Shumailov, M. Selvi, F. Piccinno, R. Aguilar, S. McCarthy, M. Khalman, M. Shukla, V. Galic, J. Carpenter, K. Villela, H. Zhang, H. Richardson, J. Martens, M. Bosnjak, S. R. Belle, J. Seibert, M. Alnahlawi, B. McWilliams, S. Singh, A. Louis, W. Ding, D. Popovici, L. Simicich, L. Knight, P. Mehta, N. Gupta, C. Shi, S. Fatehi, J. Mitrovic, A. Grills, J. Pagadora, D. Petrova, D. Eisenbud, Z. Zhang, D. Yates, B. Mittal, N. Tripuraneni, Y. Assael, T. Brovelli, P. Jain, M. Velimirovic, C. Akbulut, J. Mu, W. Macherey, R. Kumar, J. Xu, H. Qureshi, G. Comanici, J. Wiesner, Z. Gong, A. Ruddock, M. Bauer, N. Felt, A. GP, A. Arnab, D. Zelle, J. Rothfuss, B. Rosgen, A. Shenoy, B. Seybold, X. Li, J. Mudigonda, G. Erdogan, J. Xia, J. Simsa, A. Michi,

Y. Yao, C. Yew, S. Kan, I. Caswell, C. Radebaugh, A. Elisseeff, P. Valenzuela, K. McKinney, K. Paterson, A. Cui, E. Latorre-Chimoto, S. Kim, W. Zeng, K. Durden, P. Ponnapalli, T. Sosea, C. A. Choquette-Choo, J. Manyika, B. Robenek, H. Vashisht, S. Pereira, H. Lam, M. Velic, D. Owusu-Afriyie, K. Lee, T. Bolukbasi, A. Parrish, S. Lu, J. Park, B. Venkatraman, A. Talbert, L. Rosique, Y. Cheng, A. Sozanschi, A. Paszke, P. Kumar, J. Austin, L. Li, K. Salama, W. Kim, N. Dukkipati, A. Baryshnikov, C. Kaplanis, X. Sheng, Y. Chervonyi, C. Unlu, D. de Las Casas, H. Askham, K. Tunyasuvunakool, F. Gimeno, S. Poder, C. Kwak, M. Miecnikowski, V. Mirrokni, A. Dimitriev, A. Parisi, D. Liu, T. Tsai, T. Shevlane, C. Kouridi, D. Garmon, A. Goedeckemeyer, A. R. Brown, A. Vijayakumar, A. Elqursh, S. Jazayeri, J. Huang, S. M. Carthy, J. Hoover, L. Kim, S. Kumar, W. Chen, C. Biles, G. Bingham, E. Rosen, L. Wang, Q. Tan, D. Engel, F. Pongetti, D. de Cesare, D. Hwang, L. Yu, J. Pullman, S. Narayanan, K. Levin, S. Gopal, M. Li, A. Aharoni, T. Trinh, J. Lo, N. Casagrande, R. Vij, L. Matthey, B. Ramadhana, A. Matthews, C. Carey, M. Johnson, K. Goranova, R. Shah, S. Ashraf, K. Dasgupta, R. Larsen, Y. Wang, M. R. Vuyyuru, C. Jiang, J. Ijazi, K. Osawa, C. Smith, R. S. Boppana, T. Bilal, Y. Koizumi, Y. Xu, Y. Altun, N. Shabat, B. Bariach, A. Korchemniy, K. Choo, O. Ronneberger, C. Iwuanyanwu, S. Zhao, D. Soergel, C.-J. Hsieh, I. Cai, S. Iqbal, M. Sundermeyer, Z. Chen, E. Bursztein, C. Malaviya, F. Biadsy, P. Shroff, I. Dhillon, T. Latkar, C. Dyer, H. Forbes, M. Nicosia, V. Nikolaev, S. Greene, M. Georgiev, P. Wang, N. Martin, H. Sedghi, J. Zhang, P. Banzal, D. Fritz, V. Rao, X. Wang, J. Zhang, V. Patraucean, D. Du, I. Mordatch, I. Jurin, L. Liu, A. Dubey, A. Mohan, J. Nowakowski, V.-D. Ion, N. Wei, R. Tojo, M. A. Raad, D. A. Hudson, V. Keshava, S. Agrawal, K. Ramirez, Z. Wu, H. Nguyen, J. Liu, M. Sewak, B. Petrini, D. Choi, I. Philips, Z. Wang, I. Bica, A. Garg, J. Wilkiewicz, P. Agrawal, X. Li, D. Guo, E. Xue, N. Shaik, A. Leach, S. M. Khan, J. Wiesinger, S. Jerome, A. Chakladar, A. W. Wang, T. Ornduff, F. Abu, A. Ghaffarkhah, M. Wainwright, M. Cortes, F. Liu, J. Maynez, A. Terzis, P. Samangouei, R. Mansour, T. Kępa, F.-X. Aubet, A. Algymr, D. Banica, A. Weisz, A. Orban, A. Senges, E. Andrejczuk, M. Geller, N. D. Santo, V. Anklin, M. A. Merey, M. Baeuml, T. Strohman, J. Bai, S. Petrov, Y. Wu, D. Hassabis, K. Kavukcuoglu, J. Dean, and O. Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024a. URL `https://arxiv.org/abs/2403.05530`.

G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozińska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucińska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. van Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J. yeong Ji, K. Mohamed, K. Badola, K. Black, K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjoesund, L. Usui, L. Sifre, L. Heuermann, L. Lago, L. McNealus, L. B. Soares, L. Kilpatrick, L. Dixon, L. Martins, M. Reid, M. Singh, M. Iverson, M. Görner, M. Velloso, M. Wirth, M. Davidow, M. Miller, M. Rahtz, M. Watson, M. Risdal, M. Kazemi, M. Moynihan, M. Zhang, M. Kahng, M. Park, M. Rahman, M. Khatwani, N. Dao, N. Bardoliwalla, N. Devanathan, N. Dumai, N. Chauhan, O. Wahltinez, P. Botarda, P. Barnes, P. Barham, P. Michel, P. Jin, P. Georgiev, P. Culliton, P. Kuppala, R. Comanescu, R. Merhej, R. Jana, R. A. Rokni, R. Agarwal, R. Mullins, S. Saadat, S. M. Carthy, S. Perrin, S. M. R. Arnold, S. Krause, S. Dai, S. Garg, S. Sheth, S. Ronstrom, S. Chan, T. Jordan, T. Yu, T. Eccles, T. Hennigan, T. Kocisky, T. Doshi, V. Jain, V. Yadav, V. Meshram, V. Dharmadhikari, W. Barkley, W. Wei, W. Ye, W. Han, W. Kwon, X. Xu, Z. Shen, Z. Gong, Z. Wei, V. Cotruta, P. Kirk, A. Rao, M. Giang, L. Peran, T. Warkentin, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, D. Sculley, J. Banks, A. Dragan, S. Petrov, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, S. Borgeaud, N. Fiedel, A. Joulin, K. Kenealy, R. Dadashi, and A. Andreev. Gemma 2: Improving open language models at a practical size. *ArXiv preprint*, abs/2408.00118, 2024b. URL `https://arxiv.org/abs/2408.00118`.

J. B. Tenenbaum, T. L. Griffiths, and C. Kemp. Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318, 2006. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2006.05.009. URL `https://www.sciencedirect.com/`

science/article/pii/S1364661306001343. Special issue: Probabilistic models of cognition.

J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. doi: 10.1126/science.1192788. URL https://www.science.org/doi/abs/10.1126/science.1192788.

S. Toshniwal, S. Wiseman, K. Livescu, and K. Gimpel. Chess as a testbed for language model state tracking. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11385–11393. AAAI Press, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/21390.

A. Y. Tsai, A. Kraft, L. Jin, C. Cai, A. Hosseini, T. Xu, Z. Zhang, L. Hong, E. H. Chi, and X. Yi. Leveraging llm reasoning enhances personalized recommender systems. *ArXiv preprint*, abs/2408.00802, 2024. URL https://arxiv.org/abs/2408.00802.

K. Vafa, J. Y. Chen, J. Kleinberg, S. Mullainathan, and A. Rambachan. Evaluating the world model implicit in a generative model. *ArXiv preprint*, abs/2406.03689, 2024. URL https://arxiv.org/abs/2406.03689.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

L. Wong, G. Grand, A. K. Lew, N. D. Goodman, V. K. Mansinghka, J. Andreas, and J. B. Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought. *ArXiv preprint*, abs/2306.12672, 2023. URL https://arxiv.org/abs/2306.12672.

S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

F. Xu and J. B. Tenenbaum. Word learning as bayesian inference. *Psychological review*, 114(2):245, 2007.

L. Ying, T. Zhi-Xuan, L. Wong, V. Mansinghka, and J. Tenenbaum. Grounding language about belief in a bayesian theory-of-mind. *ArXiv preprint*, abs/2402.10416, 2024. URL https://arxiv.org/abs/2402.10416.

J.-Q. Zhu and T. L. Griffiths. Incoherent probability judgments in large language models. *ArXiv preprint*, abs/2401.16646, 2024. URL https://arxiv.org/abs/2401.16646.

B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1433–1438. AAAI Press, 2008. ISBN 9781577353683.

A Additional Results

A.1 Other Models

Our main experiments only use the Gemma 2 family of models. Here, we evaluate a stronger closed source model: Gemini 1.5 Pro (Team et al., 2024a). We directly ask the LM to generate probability judgements (Section 6), as we do no have access to the LM's log probabilities. We show results in Figure 5. We observe that the LM achieves non-trivial performance compared to the random baseline, and that the metrics slightly improve as more information becomes available. However, there remains a considerable gap to the optimal Bayesian Assistant.

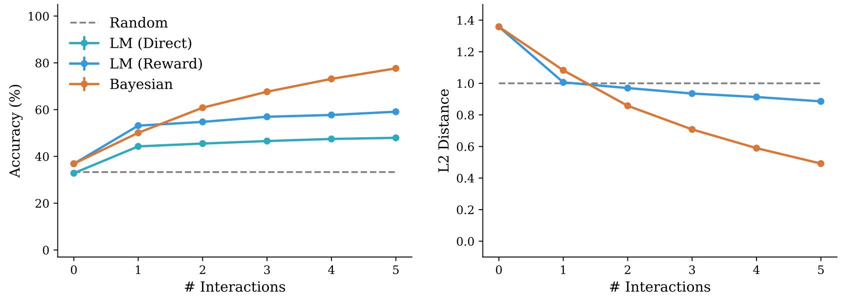


Figure 5: Results using Gemini 1.5 Pro. Left: the accuracy of the flight recommendations made by the Bayesian Assistant, directly by the LM, and using the reward function estimated from the LM. Right: the normalized L2 distance between the reward functions estimated from the models and the ground truth reward. The dashed lines indicates random performance.

A.2 Analysis & Ablations

Robustness to Reward Functions. In Table 2b, we show results averaged over reward functions. However, some reward functions might be harder to infer as it strongly deviates from the LM's prior. For example, the LM may assume most people prefer shorter flights over long ones, then inferring the preferences of a "abnormal" user that prefers longer flights would be more difficult. In practice, we find that the LM generally assigns a high probability to "no preference" when no user-specific information provided. We show a visualization of the LM's prior on each feature in Figure 6.

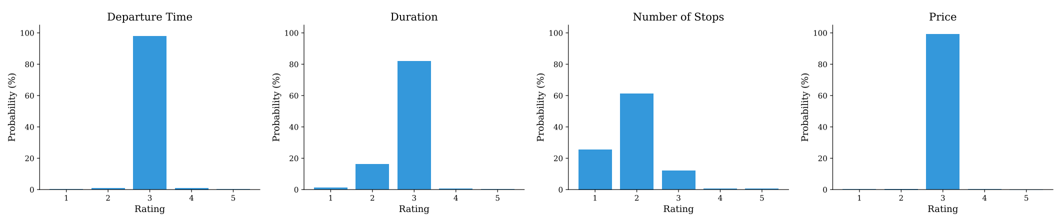


Figure 6: The LM's prior on each feature. A rating of 1 indicates a strongest preference for the earliest departure time, the shortest duration, the fewest number of stops, and the lowest price, while a rating of 5 indicates the opposite. A rating of 3 indicates no preference.

We show results with standard deviation across reward functions in Figure 7. We find that both the LM and the Bayesian model have high variances. However, the variance of the Bayesian Assistant decreases as the number of interactions increases, while the variance of the LM remains almost constant. We provide a further breakdown in Figure 8, where we show the correlation between a reward function's final-round accuracy and its normalized L2 distance to the mean of the prior reward distribution. We observe negative correlations across all settings. The Bayesian model demonstrates greater robustness to a mis-specified prior, with a smaller coefficient when fitted to a linear model, while the "reward" method is the most sensitive to the prior distribution.

Figure 9 shows the relationship between the performance and the reward value for each feature. We do not observe strong correlations except for price, where the LM performs worse as the reward value becomes positive, likely due to the LM's strong prior that people generally prefer cheaper flights, and it struggles to correct this prior based on the given observations.

Increasing the Number of Interactions. Our previous experiments include only 5 rounds of interactions between the user and LMs. Although the performance of LMs does not improve over interactions, this might be due to insufficient information from the limited number of rounds. We investigate this by increasing the number of interactions to 50 and show the results in Figure 10. For the LM, we only show the accuracy using predictions derived from its reward distribution due to computational constraint. While the Bayesian model achieves nearly perfect performance by the end of 50-round interactions, the LM still shows similar performance to our main experiments. This

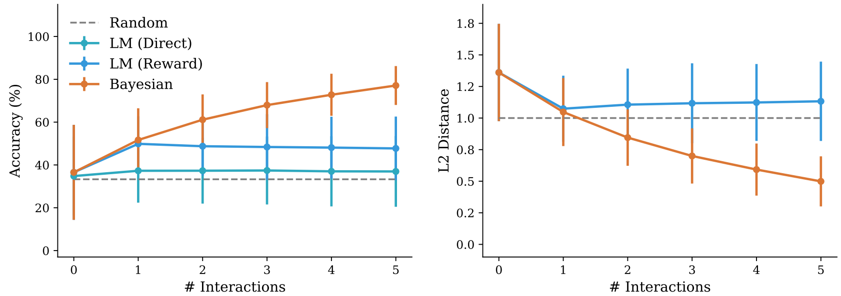


Figure 7: Left: the accuracy of the flight recommendations of the Bayesian Assistant as well as the LM when derived directly from its generated text (“direct”) and when derived from the reward function estimated from it (“reward”). Right: the normalized L2 distance between the reward functions estimated from the models and the ground truth reward. The dashed lines indicates random performance. We show the standard deviations across different reward functions. The results of each reward function are averaged over three random seeds.

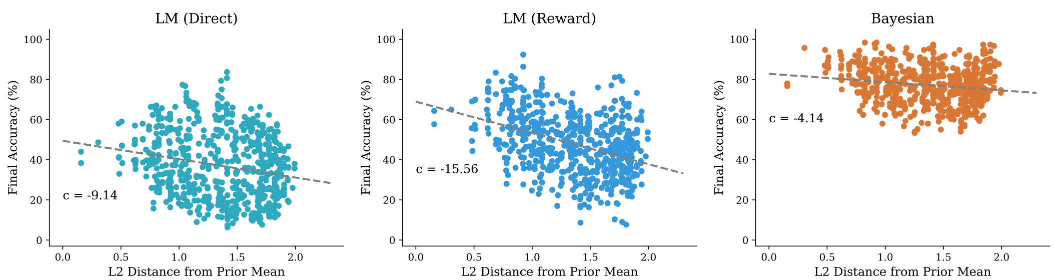


Figure 8: The relationship between the final-round accuracy and the L2 distance to the mean of the prior reward distribution. We show the performance of the LM’s direct predictions (left), predictions derived from the LM’s reward distribution (middle), and predictions derived from the Bayesian model’s reward distribution (right). We also show the coefficients by fitting these data points to a linear model.

suggests that simply increasing the number of interactions is unlikely to significantly improve the LM’s performance.

A.3 Qualitative Example

In Figure 11, we show a qualitative example of how the reward distributions of the LM and the Bayesian model change over interactions. In this case, since the user’s true reward function differs significantly from the LM’s prior, both the LM and the Bayesian model perform poorly at the start of the interactions. However, the Bayesian model gradually converges toward the ground-truth reward function after a few rounds, while the LM continues to assign high probability to reward functions that are inconsistent with its observations.

B Example Interactions

We show example interactions with LMs in Table 4, Table 5, Table 6, Table 7, Table 8.

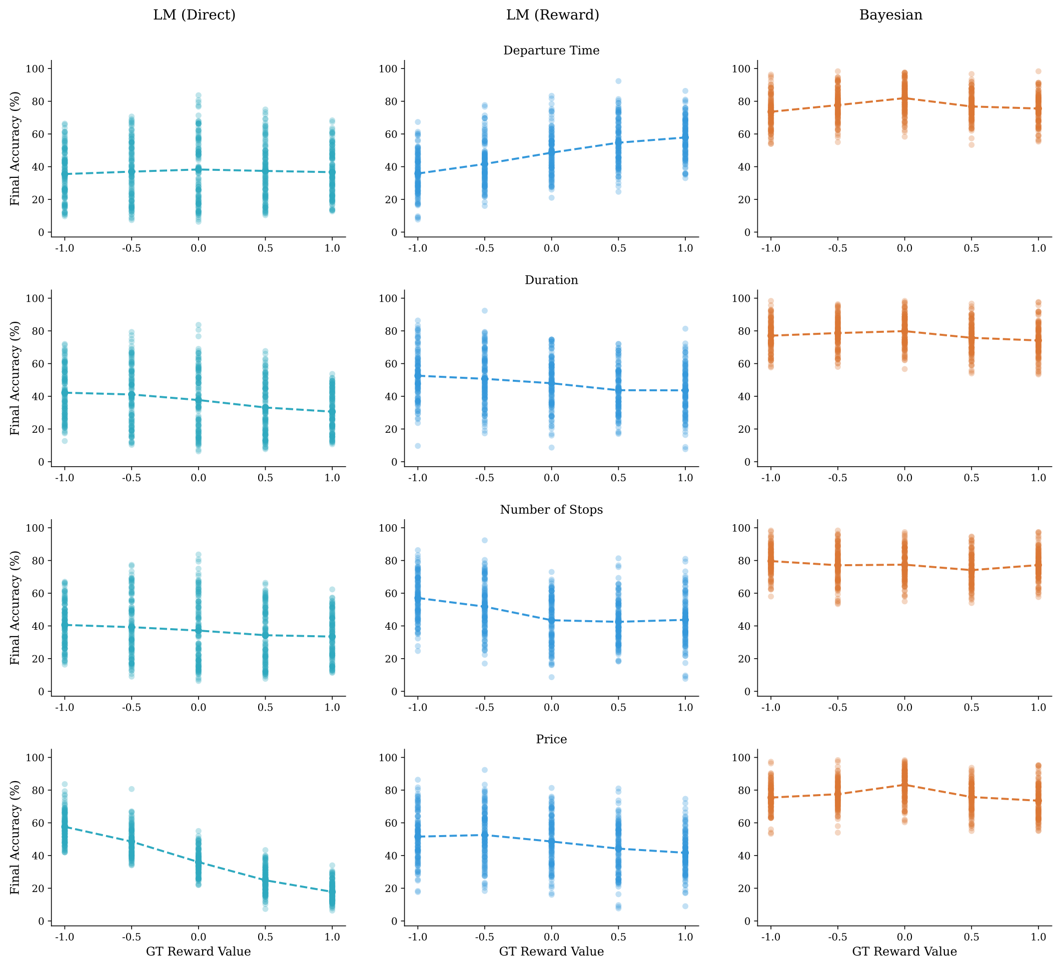


Figure 9: The relationship between the final-round accuracy and the reward function value for each feature.

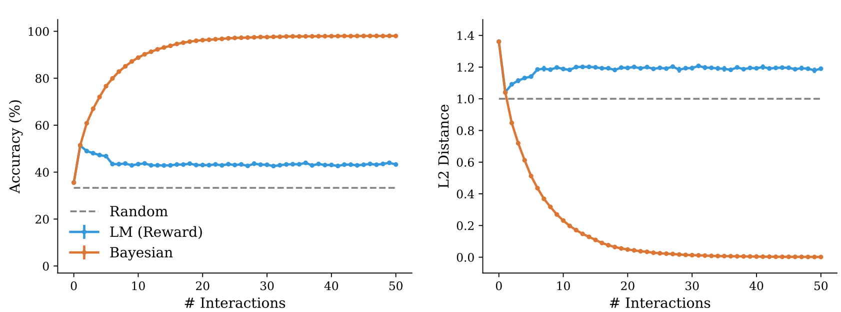


Figure 10: Results using 50-round interactions between the user and the LM.

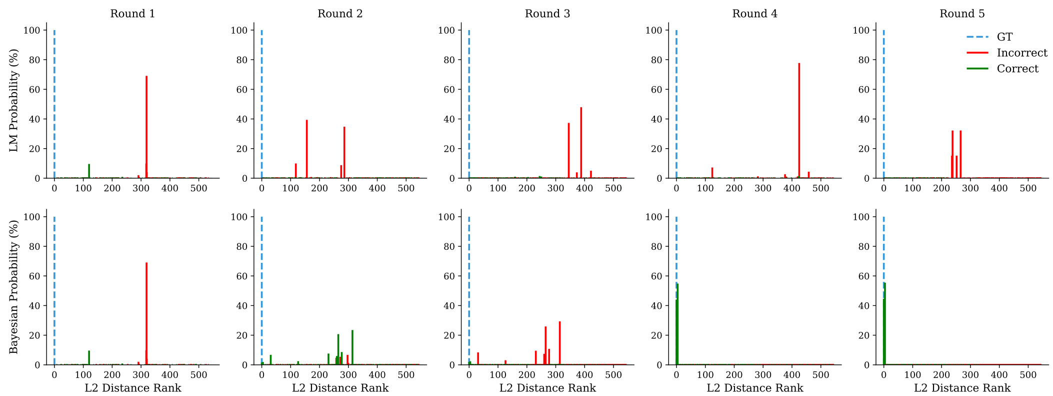


Figure 11: The reward distributions of the LM (top) and the Bayesian model (bottom) over multiple rounds. The reward functions are sorted by their normalized L2 distance from the ground-truth reward function indicated by the blue dashed line at $x = 0$. Red indicates that the reward function's prediction on the given options is incorrect, while green indicates that its prediction is correct.

```
User: Help me select the best flights for my trips. I have specific preferences for what I
like and dislike in a flight, and these preferences remain the same. You need to figure out
my preferences and select the best flights for me. Use your best judgment if you are unsure.
Do not say you need more information.

Which flight is the best option?

Flight 1:
departure time: 02:00 PM, duration: 2 hr 30 min, number of stops: 1, price: $370
Flight 2:
departure time: 10:00 PM, duration: 4 hr 24 min, number of stops: 0, price: $730
Flight 3:
departure time: 03:36 PM, duration: 16 hr 6 min, number of stops: 0, price: $900

Model: The best option is Flight 2.

User: Your option Flight 2 is correct.
```

Table 4: Example of a first-round interaction between the user and the LM.

```
User: Help me select the best flights for my trips. I have specific preferences for what I
like and dislike in a flight, and these preferences remain the same. You need to figure out
my preferences and select the best flights for me. Use your best judgment if you are unsure.
Do not say you need more information.

Which flight is the best option?

Flight 1:
departure time: 02:00 PM, duration: 2 hr 30 min, number of stops: 1, price: $370
Flight 2:
departure time: 10:00 PM, duration: 4 hr 24 min, number of stops: 0, price: $730
Flight 3:
departure time: 03:36 PM, duration: 16 hr 6 min, number of stops: 0, price: $900

Model: The best option is Flight 1.

User: Your option Flight 1 is incorrect. I prefer Flight 2.

On a scale of 1 to 5, what is my preference for price?

- 1: I strongly prefer a cheaper flight
- 2: I prefer a cheaper flight
- 3: I have no strong preference
- 4: I prefer a more expensive flight
- 5: I strongly prefer a more expensive flight

Model: Your preference for price is: 1.
```

Table 5: Example of prompting the LM for the user's preference on a single feature.

```
User: Help me select the best flights for my trips. I have specific preferences for what I
like and dislike in a flight, and these preferences remain the same. You need to figure out
my preferences and select the best flights for me. Use your best judgment if you are unsure.
Do not say you need more information.

Which flight is the best option?

Flight 1:
departure time: 0.7, duration: 0.5, number of stops: 1.0, price: 0.9
Flight 2:
departure time: 0.9, duration: 0.6, number of stops: 0.0, price: 0.7
Flight 3:
departure time: 0.5, duration: 0.9, number of stops: 0.5, price: 0.1

Model: The best option is Flight 1.
```

Table 6: Example of a first-round interaction between the user and the LM using the numerical representation of flights.

```
User: Help me select the best flights for my trips. I have specific preferences for what I
like and dislike in a flight, and these preferences remain the same. You need to figure out
my preferences and select the best flights for me. Use your best judgment if you are unsure.
Do not say you need more information.

Which flight is the best option?

Flight 1:
departure time: 0.7, duration: 0.5, number of stops: 1.0, price: 0.9
Flight 2:
departure time: 0.9, duration: 0.6, number of stops: 0.0, price: 0.7
Flight 3:
departure time: 0.5, duration: 0.9, number of stops: 0.5, price: 0.1

Model: The best option is Flight 1.

User: Your option Flight 1 is incorrect. I prefer Flight 2.

On a scale of 1 to 5, what is my preference for duration?

- 1: I strongly prefer the minimum value
- 2: I prefer the minimum value
- 3: I have no strong preference
- 4: I prefer the maximum value
- 5: I strongly prefer the maximum value
```

Table 7: Example of prompting the LM for the user's preference on a single feature using the numerical representation of flights.

```
User: Help me select the best flights for my trips. I have specific preferences for what I
like and dislike in a flight, and these preferences remain the same. You need to figure out
my preferences and select the best flights for me. Use your best judgment if you are unsure.
Do not say you need more information.

Which flight is the best option?

Flight 1:
departure time: 02:00 PM, duration: 2 hr 30 min, number of stops: 1, price: $370
Flight 2:
departure time: 10:00 PM, duration: 4 hr 24 min, number of stops: 0, price: $730
Flight 3:
departure time: 03:36 PM, duration: 16 hr 6 min, number of stops: 0, price: $900

Model: The best option is Flight 1.

User: Your option Flight 1 is incorrect. I prefer Flight 2.

On a scale of 1 to 5, what is my preference for price?

- 1: I strongly prefer a cheaper flight
- 2: I prefer a cheaper flight
- 3: I have no strong preference
- 4: I prefer a more expensive flight
- 5: I strongly prefer a more expensive flight

Provide an integer between 0 and 100 (%) that reflects the probability of each scale.

Model: The probabilities of each scale are:

- 1: 70%
- 2: 10%
- 3: 15%
- 4: 5%
- 5: 0%
```

Table 8: Example of prompting the LM for the user's preference on a single feature using the model's directly generated probability judgements.