

---

# Active Preference Inference using Language Models and Probabilistic Reasoning

---

**Top Piriyaikulij**  
Cornell University  
wp237@cornell.edu

**Volodymyr Kuleshov**  
Cornell Tech  
kuleshov@cornell.edu

**Kevin Ellis**  
Cornell University  
kellis@cornell.edu

## Abstract

Actively inferring user preferences, for example by asking good questions, is important for any human-facing decision-making system. Active inference allows such systems to adapt and personalize themselves to nuanced individual preferences. To enable this ability for instruction-tuned large language models (LLMs), one may prompt them to ask users questions to infer their preferences, transforming the language models into more robust, interactive systems. However, out of the box, these models are not efficient at extracting preferences: the questions they generate are not informative, requiring a high number of user interactions and impeding the usability of the downstream system. In this work, we introduce an inference-time algorithm that helps LLMs quickly infer preferences by using more informative questions. Our algorithm uses a probabilistic model whose conditional distributions are defined by prompting an LLM, and returns questions that optimize expected entropy and expected model change. Results in a simplified interactive web shopping setting with real product items show that an LLM equipped with our entropy reduction algorithm outperforms baselines with the same underlying LLM on task performance while using fewer user interactions.

## 1 Introduction

For any decision-making AI system, knowing each user’s task preferences is crucial for the system to be most helpful to individual users. Ideally, in the context of language-based AI systems, we might naively hope for users to fully specify a task in natural language. However, while human speakers can be very effective at selecting their utterances (Grice, 1991), they often fail to fully specify their task preferences to human or AI listeners. Designing language-supported AI systems that can actively infer users’ task preferences (e.g., by asking questions when they are uncertain) could help make these technologies more aligned, more personalized, and ultimately more broadly applicable.

Recently, large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023), enhanced with instruction tuning (Christiano et al., 2017; Ouyang et al., 2022), have shown impressive performance at following natural language task instructions. While LLMs are not natively trained to be good active learners, we can prompt them, leveraging their in-context learning ability, to ask users questions to infer their preferences. Concurrent work (Li et al., 2023) has shown that prompting LLMs to elicit user preferences indeed helps improve performances over various tasks. Using language models to first elicit user preferences before carrying out the tasks provides a promising path toward enabling these models in domains that require fine-grained understanding of user preferences.

Nevertheless, a simple conversation with these off-the-shelf LLMs (Figure 4) reveals their poor ability to efficiently learn user preferences from few questions. Although these LLMs are able to ask relevant questions when prompted, they do not ask the more informative questions, given the current context. This increases the amount of model-human interaction needed to infer the user’s preferences.

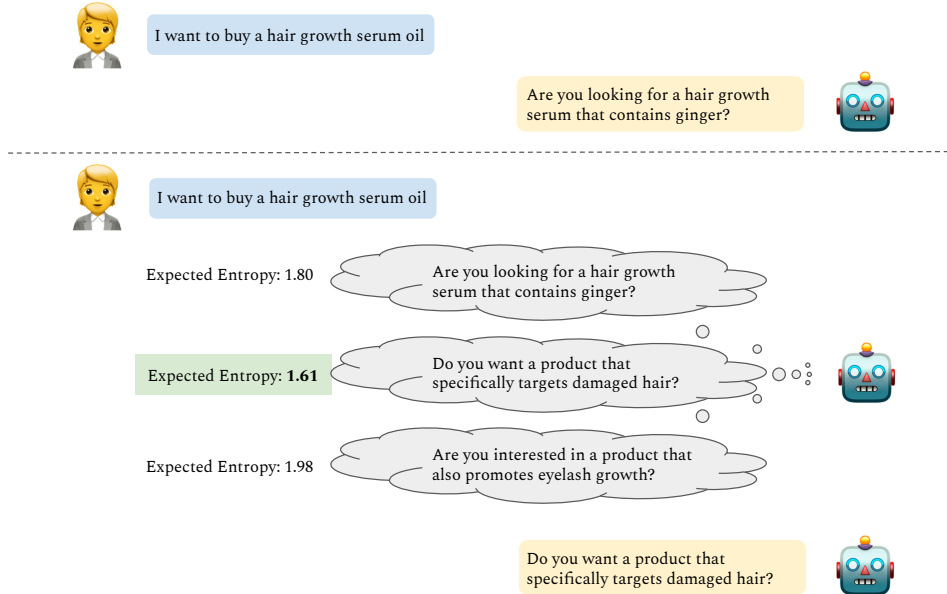


Figure 1: (Top) Vanilla instruction-tuned LLM prompted to be a hair growth serum oil seller. (Bottom) LLM with inference-time expected entropy reduction algorithm.

We propose an inference-time algorithm for active inference of user preferences by LLMs. Our algorithm is motivated by the observation that while out-of-the-box LLMs are suboptimal at interactively inferring user preferences, they are decent at generating diverse questions for extracting preferences, and can generate plausible answers to each such question. Moreover, given an ongoing interaction with a user, LLMs can reasonably score how likely the user will answer a question in certain ways. Thus, we can use the LLM-defined scores to define a full probabilistic generative model of what question to ask, how the user might respond, and how those responses update the system’s probabilistic beliefs. Concretely, our algorithm samples a finite set of questions from the LLM and chooses the one that maximizes information gain.

We evaluate our approach in a simplified web shopping setting using data from a realistic web shopping benchmark (Yao et al., 2022). Each task consists of a list of products, together with a user who has a latent preference for a specific target product from that list. LLMs will ask a series of yes/no questions to infer the target product. We compare our inference-time active inference method against a vanilla instruction-tuned LLM (Christiano et al., 2017; Ouyang et al., 2022) and ReAct (Yao et al., 2023), finding that our approach outperforms both baselines while using fewer questions. In line with recent work (Dohan et al., 2022; Zhao et al., 2023; Ellis, 2023), our results should be seen as a way of augmenting LLMs with inference-time probabilistic reasoning.

## 2 Active Preference Inference

### 2.1 Task formulation

We first introduce the active preference inference formulation. For decision-making tasks, human preferences can be viewed as reward functions that map decisions to a scalar,  $R : \mathcal{X} \rightarrow \mathcal{R}$  where  $\mathcal{X}$  is the space of decisions. The human user can be described with a reward function and an answer model  $(R^*, p_{\text{human}}(a|q))$  where  $q, a \in \Sigma^*$ .  $q$  and  $a$  are a natural language question and answer, and  $\Sigma^*$  is the space of natural language. Active preference inference task boils down to a task of asking a series of questions  $q_1, \dots, q_k$ , to which the answers  $a_i \sim p_{\text{human}}(a|q_i)$  will be generated, to infer the unknown target reward  $R^*$ .

Active preference inference resembles active learning but differs in a few key ways. First, active preference inference separates the answer model  $p_{\text{human}}(a|q)$  from the reward function  $R^*$ . This

	Task-specific data collection	Informative-question-asking ability	Inference cost
Traditional interactive NLP system (Rao & Daumé III, 2018; Yu et al., 2020)	Required	Optimized during training	Low
LLM w/o probabilistic reasoning (Li et al., 2023)	Not required	Not optimized	Medium
<b>LLM w/ probabilistic reasoning (ours)</b>	Not required	Optimized during inference	High

Table 1: Comparison between different existing approaches for active preference inference formulation allows the flexibility of asking arbitrary questions to the user. We also note that the question  $q$  can be of the form "Is the following example positive or negative: {example}", which means active preference inference can be seen as a generalization of active learning. Second, active learning usually assumes access to unlabeled data points that an algorithm can query, while active preference inference asks an algorithm to generate its own questions.

## 2.2 Existing approaches

Prior to the advent of LLMs, there were attempts at training NLP systems specifically for interactive inference (Rao & Daumé III, 2018; Yu et al., 2020). Although these works do not consider the question of how to infer the reward function from the final result conversation, we can easily define the predicted reward function as LLM conditioned with that conversation. The main downside of these methods, however, is that they require expensive task-specific data collection for training. Having to gather data for every new task severely limits the usability of these methods.

LLM-based approaches to active preference inference promise to bypass the data collection process with its zero-shot learning capability. With instruction-tuned LLMs, we can directly prompt the models to ask users questions to infer preferences. However, as mentioned and confirmed later in our experiments, these models turn out to be ineffective at preference inference off-the-shelf and require many interactions with the users before arriving at a good prediction of the reward function.

In this work, we equip LLMs with inference-time probabilistic reasoning algorithm defined in Section 3 to address this core issue, extracting as much information from the user as possible in as few questions as needed. This essentially minimizes human-user effort, at the expense of more LLM computation. As LLMs inference will likely become cheaper in the future, we believe this is the right tradeoff to make. Table 1 compares other methods that with alternative tradeoffs.

## 3 Inference-time probabilistic reasoning for asking informative question

In this section, we define the relevant models and present two measures from the active learning literature (Settles, 2009) that one may use to pick an informative question.

### 3.1 Model definition

To develop our model, we make an assumption that the target reward function  $R^*$  is a binary function that is 1 only at the target decision  $x^* \in \mathcal{X}$  and 0 everywhere else. With this assumption, we only need to infer the target decision  $x^*$  to know  $R^*$ .

We define the joint distribution as  $p(x, q, a) = p(a|x, q)p(x)p(q)$ . The distributions needed to compute the joint are defined by prompting an LLM:

$$\begin{aligned}
 p(q) &= \text{Uniform}(\Sigma^*) \\
 p(x) &= \text{Uniform}(\mathcal{X}) \\
 p(a|x, q) &= \left[ \begin{array}{c} \text{a binary 0/1 score output by prompting an LLM} \\ \text{whether the answer } a \text{ to the question } q \text{ is consistent} \\ \text{with the product description of } x \end{array} \right]
 \end{aligned}$$

With these defined distributions, we know that  $p(x|q, a) \propto p(a|x, q)$  because  $p(x|q) = p(x) = \mathcal{U}(X)$ . We can also easily calculate  $p(a|q) = \sum_{x \in X} p(a|x, q)p(x)$ . In addition, we define a question proposal distribution  $r(q)$  by asking an LLM to generate a finite number of possible questions given all product information  $X$ . Prompts used for obtaining  $p(a|x, q)$  and  $r(q)$  can be founded in Appendix A.3.

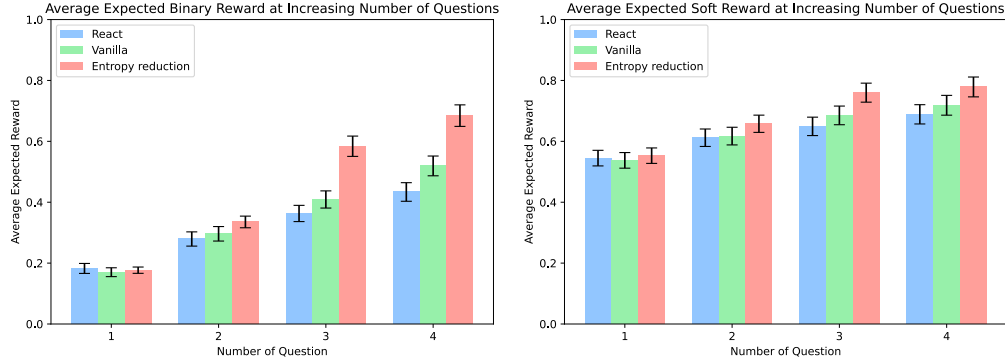


Figure 2: (Left) Average expected binary reward at increasing number of number of questions. (Right) Average expected soft reward at increasing number of number of questions.

We can easily extend our model to the setting where we have multiple pairs of questions and answers. Let  $\mathbf{c}$  denote a conversation containing a sequence of questions and answers  $\mathbf{c} = \{(q_1^{(c)}, a_1^{(c)}), \dots, (q_m^{(c)}, a_m^{(c)})\}$ . Our model becomes  $p(x, \mathbf{c}) = p(x) \prod_i p(a_i^{(c)} | x, q_i^{(c)}) p(q_i^{(c)})$ . With this new model,  $p(x | \mathbf{c}) \propto \prod_i p(a_i^{(c)} | x, q_i^{(c)})$ , and  $p(a | \mathbf{c}, q) \propto \sum_{x \in \mathcal{X}} (p(a | x, q) \prod_i p(a_i^{(c)} | x, q_i^{(c)}))$ . We also modify the prompt for  $r(q)$  to take into account the conversation, notated  $r(q | \mathbf{c})$ .

### 3.2 Objectives for choosing informative questions

With the distributions defined in Section 3.1, we sample a finite set of questions from the proposal distribution  $r(q | \mathbf{c})$ , calculate the amount of information each question would reveal in expectation, and subsequently pick the question with the highest amount of expected information. We consider two objectives, commonly used in the active learning field (Settles, 2009), to measure how informative a question is: expected entropy minimization and expected model change maximization.

**Expected Entropy Minimization.** The most informative question could be thought as the question which reduces the most amount of uncertainty in expectation. This notion can be captured with the expected entropy minimization objective. The objective is as follows:

$$\arg \min_{q \sim r(q | \mathbf{c})} E_{p(a | \mathbf{c}, q)} [H(p(x | \mathbf{c}, q, a))] \quad (1)$$

**Expected Model Change Maximization.** Another approach is to think about model parameter/belief and consider the most informative question as the question that would make the most change to the model parameter/belief. For our tasks, we can let belief be a decision  $x \in \mathcal{X}$ , representing the notion that we believe  $x$  is the decision with a positive reward. We can use KL divergence to measure the change in belief. The objective is as follows:

$$\arg \max_{q \sim r(q | \mathbf{c})} E_{p(a | \mathbf{c}, q)} [D_{\text{KL}}(p(x | \mathbf{c}, q, a) || p(x | \mathbf{c}))] \quad (2)$$

**Equivalence between the two objectives.** In our specific setting, it can be shown that expected entropy minimization and expected model change maximization are in fact equivalent. This is due to the fact that the expected cross-entropy  $E_{p_{\text{model}}(a | \mathbf{c}, q)} [E_{p(h | \mathbf{c}, q, a)} [-\log p(h | \mathbf{c})]]$  is the same for all  $q, a$  because by our definition,  $p(h | \mathbf{c}, q, a) > 0$  if  $p(h | \mathbf{c}) > 0$ , and if  $p(h | \mathbf{c}) > 0$ , then  $p(h | \mathbf{c}) = f(\mathbf{c})$  is constant to  $h$ . More details can be founded at Appendix A.1

## 4 Experiments

We evaluate an LLM augmented with our entropy reduction algorithm (entropy reduction LLM) on 150 simplified web shopping tasks, each task consisting of 10 product items from WebShop (Yao et al., 2022). The products in each task come from the same product type, e.g., hair growth serum oil, HDMI cable, phone case, etc. While the goal of active preference inference is to infer the target

reward function  $R^*$ , it is more realistic in web-shopping settings to make algorithms return a single product  $\hat{x}$  and evaluate them based on the reward of the returned product  $R^*(\hat{x})$ , so we employ this evaluation in this work. This evaluation can still be seen as implicitly inferring reward function  $\hat{R}$  and then return  $\hat{x} = \arg \max_{x \in \mathcal{X}} \hat{R}(x)$ . We consider two types of reward functions: binary and soft reward. Both are constructed based on a target product  $x^*$  randomly chosen from the task’s list of products. More details on task creation can be found at Appendix A.2. We simulate human users  $p_{\text{human}}(a|q)$  with GPT-4 (OpenAI, 2023) by inputting target product  $x^*$  in the prompt. We make a simplifying assumption that only yes/no questions to users are allowed.

Two baselines we consider are vanilla instruction-tuned LLM (Christiano et al., 2017; Ouyang et al., 2022) and ReAct LLM (Yao et al., 2023). All methods use the same  $p(x|q, a)$  to choose a product given questions and answers; the differences in these methods lie in the questions they choose to ask. Vanilla instruction-tuned LLM is prompted to roleplay to be a product seller and ask questions. Our entropy reduction method samples 8 questions from the proposal distribution  $r(q|c)$  but would reduce down to vanilla instruction-tuned LLM if we only sample one question from the proposal distribution. ReAct is a prompting method for LLM Agents where thinking is part of the action space. In our question-asking scenario, asking a question can be seen as an action. We write a prompt that has a thinking process on what question is best to ask at each point. Prompts for both of these methods can be found at Appendix A.3. When more than one product has positive probability  $p(x|q, a)$  at the end of the questioning, we assume the methods return the positive-probability products at random and actually calculate the expected reward on them. The underlying LLM for all methods is GPT-4. We sample at zero temperature (For the proposal distribution, we actually prompt LLM to return a list of diverse questions instead of sampling questions from LLM with positive temperature).

#### 4.1 Binary Reward

First, we consider a binary reward setting where the reward is 1 if the target product is retrieved and 0 otherwise. This binary reward is consistent with our modeling assumption. Figure 2 (left) shows that entropy reduction LLM significantly outperforms both baselines when the number of questions is greater than one. In fact, our algorithm after three questions even surpasses the two baselines after four questions.

Figure 3 reports average information gain at each round of questioning. For the first three questions, we can see that entropy reduction LLM gains more information than both baselines. At the last question, vanilla instruction-tuned LLM gets more information, but all methods do not receive much information gain at this point. There are 10 products in the list of products for each task, and at the fourth question, not much information is left to be extracted from human users.

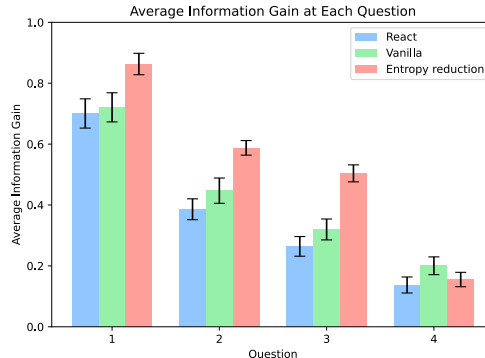


Figure 3: Average information gain at each question. Confidence intervals are computed over 150 different tasks.

#### 4.2 Soft Reward

We also consider a more realistic, soft reward setting. In practical scenarios, it is likely that user preferences for product items are not all-or-nothing, and user satisfaction upon receiving each product item as output is more fine-grained. We adopt the reward used in the Webshop benchmark (Yao et al., 2022). We exclude product price and options from the original reward calculation as they are not relevant in our task setting. The reward is then calculated based on Webshop’s text matching heuristics and the attribute overlap between the target product and the returned product. The reward falls within the range of 0 to 1, with an exact match yielding a reward of 1.

Similar to the binary reward, Figure 2 (right) shows that entropy reduction LLM outperforms both baselines when the number of questions is greater than one. Even though the soft reward is different

from our modeling assumption, the inference-time entropy reduction algorithm is still better at preference inference than the two baselines.

## 5 Related Work

**Learning to ask clarifying question.** There are many prior works in learning to ask clarification questions (Rao & Daumé III, 2018; Aliannejadi et al., 2019; Yu et al., 2020; Zamani et al., 2020). Rao & Daumé III (2018) relies on the idea that a good clarification question is the one whose answer will be useful. They use Q&A forum data to train neural networks to be an answer model and an answer utility calculator and choose question that maximizes expected utility. Aliannejadi et al. (2019) looks at the problem of asking good clarifying questions from an information retrieval perspective, and train models with question-answer pairs data to retrieve good questions from a crowdsourced dataset. Yu et al. (2020) studies interactive classification where questions are selected based on the entropy minimization principle similar to our work. They crowdsource non-interactive data to train neural networks used to define their probability distributions. The main difference between our work and these works is that we do not require any sort of training data and instead rely on the powerful zero-shot learning capability of LLMs to define rich distributions, which we can probabilistically reason over at inference time.

**Probabilistic reasoning with LLMs.** Dohan et al. (2022) provides a unifying framework of seeing recent inference-time LLM algorithms (Nye et al., 2021; Wei et al., 2022; Creswell et al., 2022) as performing reasoning with probabilistic programs. Zhao et al. (2023) utilizes LLM to define prior distributions in a probabilistic graphical model and performs inference in the model to solve navigation, action recognition, and image segmentation. Our work studies inference-time probabilistic reasoning with LLM in the context of active preference inference.

## 6 Conclusion and Discussion

In this work, we present an inference-time algorithm to enhance the efficiency of actively inferring preferences of LLMs and demonstrate that in our web shopping setting, our method performs better than the baselines using fewer user interactions.

There are several future directions for this work. The most obvious one is to let the allowed questions be open-ended. Even though this modification will not call for any change to the model definition, this generalization will necessitate an approximation to the objective calculation as there might be infinitely many answers to an open-ended question. One needs to test empirically how much computation is needed before the Monte Carlo approximation is close enough to the true expected value. Another direction is to come up with a more flexible probabilistic model that does not assume the target reward function is binary. While this assumption works well in the web shopping setting, as evidently shown in both binary and soft reward evaluation, it may result in poor performance for other tasks that involve a full prediction of the reward function.

## References

- Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pp. 475–484, 2019.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.

- David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A Saurous, Jascha Sohl-Dickstein, et al. Language model cascades. *arXiv preprint arXiv:2207.10342*, 2022.
- Kevin Ellis. Human-like few-shot learning via bayesian reasoning over natural language. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1991.
- Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*, 2023.
- Maxwell Nye, Anders Andreassen, Guy Gur-Ari, Henryk Witold Michalewski, Jacob Austin, David Bieber, David Martin Dohan, Aitor Lewkowycz, Maarten Paul Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2021. <https://arxiv.org/abs/2112.00114>.
- OpenAI. Gpt-4 technical report, 2023.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.
- Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2737–2746, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1255. URL <https://aclanthology.org/P18-1255>.
- Burr Settles. Active learning literature survey. 2009.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *ArXiv*, 2022.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
- Lili Yu, Howard Chen, Sida I. Wang, Tao Lei, and Yoav Artzi. Interactive classification by asking informative questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2664–2680, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.237. URL <https://aclanthology.org/2020.acl-main.237>.
- Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*, pp. 418–428, 2020.
- Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *arXiv preprint arXiv:2305.14078*, 2023.

## A Appendix

User: Consider the following products:

1. Product 1 - blue, plastic, heavy, dust-proof, iPhone phone case
2. Product 2 - green, plastic, heavy, dust-proof, iPhone phone case
3. Product 3 - purple, plastic, heavy, dust-proof, iPhone phone case
4. Product 4 - green, leather, light, water-proof, iPhone phone case
5. Product 5 - red, plastic, heavy, dust-proof, iPhone phone case
6. Product 6 - red, plastic, heavy, dust-proof, android phone case

Please ask yes or no question to figure which product I want, one at a time.

- ↪ You will only get to ask three questions, so choose wisely. If you're
- ↪ ready, you can go ahead and ask the first question

GPT-4: Is the product you want made of plastic?

User: yes

GPT-4: Is the product you want for an iPhone?

User: yes

GPT-4: Is the product you want blue?

User: no

GPT-4: Is the product you want green?

User: no

GPT-4: Is the product you want red?

User: no

GPT-4: So, the product you want is the purple, plastic, heavy, dust-proof,  
↪ iPhone phone case. Is that correct?

Figure 4: An example conversation with GPT-4 (zero temperature) demonstrating how GPT-4 is lacking efficiency when prompted to ask questions. In this example, one can always figure out the target product within 3 questions, but GPT-4 is using 5 questions, which is as bad as asking if the target product is Product 1 or 2 or 3 and so on.

### A.1 Equivalence between expected entropy minimization and expected model change maximization

Expected entropy minimization and expected model change maximization can be shown to be equivalent for the setting in this work. The proof is straightforward:

$$\begin{aligned} & \arg \max_{q \sim r(q|\mathbf{c})} E_{p_{\text{model}}(a|\mathbf{c},q)} [D_{\text{KL}}(p(h|\mathbf{c},q,a) || p(h|\mathbf{c}))] \\ &= \arg \max_{q \sim r(q|\mathbf{c})} E_{p_{\text{model}}(a|\mathbf{c},q)} [E_{p(h|\mathbf{c},q,a)} [\log p(h|\mathbf{c})] - H(p(h|\mathbf{c},q,a))] \\ &= \arg \max_{q \sim r(q|\mathbf{c})} -H(p(h|\mathbf{c},q,a)) \\ &= \arg \min_{q \sim r(q|\mathbf{c})} H(p(x|\mathbf{c},q,a)) \end{aligned}$$

### A.2 Task Creation

To create the tasks we evaluate in our work, we first use GPT-4 to extract product types/categories from actual products from WebShop (Yao et al., 2022) (see Figure 5). Each task has its own product



type, and the list of products for the task is retrieved by searching WebShop website for the top 10 products with the task’s product type as the search keyword. The target product is sampled randomly from the list of products.

Given a product, please tell me what's the type of the product.

For example,

Product: Bright Citrus Deodorant by Earth Mama | Natural and Safe for  
→ Sensitive Skin, Pregnancy and Breastfeeding, Contains Organic Calendula  
→ 3-Ounce  
Product type: Deodorant

Here's the product,  
Product: {product}

Please put your answer in the format 'Product type: '. Do not say anything  
→ else.

Figure 5: Prompt used for extracting product type

### A.3 Prompts

Figure 6, Figure 7, Figure 8, Figure 9, Figure 10 contain the prompts for  $r(q|c)$ ,  $p_{\text{human}}(a|q)$ , vanilla baseline, ReAct baseline, and soft reward calculation respectively. For  $p(x|q, a)$ , we actually use the same prompt as  $p_{\text{human}}(a|q)$  to calculate it by simulating a human that prefers  $x$  most and ask question  $q$  to the simulated human. Suppose  $a^{(\text{human})}$  is the answer that the simulated human outputs. The binary score that  $p(x|q, a)$  is proportional to is 1 when the answer matches human answer and 0 when it is not.

Suppose you are a seller, and you need to determine which of the following  
→ products the customer wants to buy:  
{products}  
The conversation you've had with the customer so far is as follows:  
→ \n{conversation}\n. \  
Generate a list of 10 yes/no questions you would ask the customer to  
→ further figure out which of the products the customer want.  
Keep in mind that:  
1. The question must ask for a yes or no answer only.  
2. The question must ask about one specific thing, do not use 'and' or 'or'  
→ to put many features/properties/attributes together.  
3. Do not explain.

Figure 6: Prompt used for  $r(q|c)$

Suppose you want to buy the following product:

{product\_txt}

You will interact with a salesperson. They will ask you a question about  
→ the product you want. Please answer only either yes or no based on the  
→ product information.  
If you are asked about features not mentioned in the product description,  
→ then say No.

Figure 7: Prompt used for simulating  $p_{\text{human}}(a|q)$  and also for  $p(x|q, a)$

Suppose you are a seller, and you need to determine which of the following  
→ products the customer wants to buy:  
{products}  
The conversation you've had with the customer so far is as follows:  
→ \n{conversation}\n. \  
What's the next yes/no question you would ask the customer to further  
→ figure out which of the products the customer want?  
Keep in mind that:  
1. The question must ask for a yes or no answer only.  
2. The question must ask about one specific thing, do not use 'and' or 'or'  
→ to put many features/properties/attributes together.  
3. Do not explain.

Figure 8: Prompt used for roleplaying Vanilla LLM

Please choose the appropriate action (think, ask\_question, get\_products, ↪ choose\_products) based on the observation and given example runs.

Important: the question must be a yes/no question

Example 1:

Instruction: i want to buy a phone case

Action: show\_products[]

Observation:

1. ID: B09G9D18YS | Product Name: BURGA Phone Case Compatible with iPhone  
↪ 13 - Hybrid 2-Layer Hard Shell + Silicone Protective Case - Black Polka  
↪ Dots Pattern Nude Almond Latte Fashion - Scratch-Resistant Shockproof  
↪ Cover  
- Price: \$19.95  
- Attributes: phone case, wireless charging, heavy duty  
- Options: color (almond latte, black & gold onyx, dazzling glow, emerald  
↪ pool, fatal contradiction, gentle wind, gold dust, hidden beauty,  
↪ iconic ruby, mystic river)

... # Full list of products in the actual prompt not included here for  
↪ brevity

Action: think[I'll ask a few yes/no questions to determine which one of  
↪ these products the customer wants. Some of these phone cases are  
↪ holster phone cases. I'll ask a yes/no question about if they want a  
↪ holster]

Observation: OK.

Action: ask\_question[Are you looking for a holster phone case? (Yes/No)]

Observation: Answer: No.

Action: think[Ok, we're looking for non-holster phone case. Among these,  
↪ there are some clear phone cases. Let's ask a yes/no question if they  
↪ want the phone case to be clear]

Observation: OK.

Action: ask\_question[Are you looking for a clear phone case? (Yes/No)]

Observation: Answer: No.

... # Full question asking in the actual prompt not included for brevity

IMPORTANT: No more question can be asked.

Action: choose\_product[]

Observation: OK.

Example 2:

Instruction: {instruction}

Figure 9: Prompt used for ReAct

Given {product\_type} with the following description:

{target\_product\_txt}

Please rate each of the following {product\_type} on a scale of 1-10 based  
↪ on how similar it is to the given target {product\_type}

{products\_txt}

The {product\_type} least similar to the given target {product\_type} should  
↪ receive a score of 1. Please give you answer in the format '1.  
↪ Explanation: one sentence, Rating: x/10' and so on. Always explain and  
↪ give rating to all answers. At the very end, please put all output  
↪ ratings in the form 'All ratings: score\_1, score\_2, ...

Figure 10: Prompt used to calculate soft reward