Extended Abstract

Motivation Many real-world conversational tasks involve multi-turn reasoning with sparse or delayed feedback, which makes learning effective strategies difficult. This project addresses these challenges by applying multi-turn Reinforcement Learning with Human Feedback (MT RLHF) with proxy rewards–interactivity and information gain–to guide strategic question-asking and uncertainty reduction in the sparse-reward 20 Questions game.

MT RLHF works by collecting assistant-user dialogues, scoring them with a judge model on proxy metrics like interactivity and information gain, and fine-tuning the assistant via preference optimization to prefer higher-reward trajectories without requiring explicit reward functions. This is suitable to the 20 questions task due to its multi-turn strategic and long-horizon nature.

Method Our novelty lies in combining MT RLHF and sparse reward spaces. We formulate two novel proxy metrics for accuracy: (1) interactivity, which measures the assistant's ability to maintain engaging, coherent, on-topic conversation, and (2) information gain, which quantifies how much each question reduces uncertainty, encouraging strategic inquiry before the final answer. These custom reward formulations complement the sparse, long-horizon accuracy signal of the task, enabling better credit assignment and guiding the model's learning over full multi-turn trajectories. Using a judge model to score rollouts on these metrics, the assistant is fine-tuned via Direct Preference Optimization (DPO) to prefer trajectories that maximize these rewards, improving its question-asking strategy. This reward structure aims to overcome sparse feedback challenges by shaping behavior with interpretable, intermediate signals that align with long-term task success.

Implementation Our contributions to the CollabLLM framework include applying the novel proxy rewards for interactivity and information gain, generating large multi-turn preference datasets (5000 and 1200 examples respectively), fine-tuning a 1B-parameter Llama 3.2 Instruct model via DPO on each dataset, and evaluating the resulting reward-specific models on 110 diverse held-out objects to assess accuracy gains over the baseline.

Results We evaluated the base, interactive, and infogain models on 20-turn games across 110 objects, finding that both the interactive and infogain models significantly outperform the base model in accuracy–22% and 26% versus 11%. While all models take about the same number of turns per game, the infogain model achieves higher accuracy with fewer turns when correct, indicating more efficient questioning. The information gain model outperforms the baseline model by 130% and is statistically significant on an McNemar contingency table test with a p value of $0.0066 < 0.01$. Proxy metrics show that information gain correlates more strongly with accuracy (over 65%) than interactivity, suggesting it as a better reward signal for training. Qualitative case studies on target objects like camera, lake, and cactus reveal that the interactive and infogain models ask more targeted, high-yield questions, leading to more accurate and systematic narrowing of the search space, focusing on reducing scope at a reasonable rate compared to the base model. Overall, training with reward signals focused on interactivity or information gain improves dialogue strategies, with information gain providing the highest gains.

Discussion Our results demonstrate that training with a reward on the information gain metric leads to more effective and efficient question strategies in sparse multi-turn tasks, outperforming interactivity-based rewards. This suggests that leveraging information gain can significantly enhance dialogue models' ability to systematically narrow down the search space with fewer, higher-value questions. Limitations include small training data size and training model size. In addition, we assume the judge model always labels metrics correctly and the user model always responds reasonably–this could be checked more rigorously in future tests.

Conclusion This work shows that MT RLHF with proxy rewards–especially using information gain–can significantly boost model performance in sparse multi-turn tasks like 20 Questions, yielding a 2.3× accuracy improvement over the base, and a more efficient number of turns and tokens to reach the target object. Future work could explore combining reward signals and applying this approach to other interactive domains or online learning settings for continual improvement.

20 Questions: Multi-turn RLHF for Sparse Rewards

Aditi Bhaskar
Department of Computer Science
Stanford University
aditijb@cs.stanford.edu

Abstract

Long-horizon reasoning tasks with sparse feedback, such as the 20 Questions game, challenge reinforcement learning from human feedback (RLHF) due to sparse reward signals and need for strategy across turns. We propose to apply a multi-turn RLHF (MT RLHF) framework leveraging proxy rewards (interactivity and information gain) to guide learning in this sparse-reward setting. Using a user simulator, assistant, and judge models, we train agents to ask strategic yes/no questions to identify hidden objects. Our experiments show that optimizing for information gain (26% accuracy) significantly improves final accuracy over baseline (11% accuracy) and interactive tuning (22% accuracy). Manually analyzed rollouts show the information gain model has more efficient search-space reduction and that while the interactive model maintains engagement, it yields smaller accuracy gains. We find over a 65% correlation between the infogain metric and model accuracy, suggesting the infogain metric could lead to long-term task success in multi-turn dialogue. This work shows promising results towards strategically solving sparse reward discussion challenges and can be applied to many similar situations including online tech-support chatbots.

Code can be found on github at aditi-bhaskar/multiturn-20q.

1 Introduction

Have you ever chatted with a tech support bot that seemed clueless about your problem and terrible at understanding what the scope of your problem was? Long-horizon reasoning tasks with sparse or delayed feedback remain a significant challenge for reinforcement learning from human feedback (RLHF). Many real-world applications, such as conversational agents or interactive assistants, require a sequence of actions before meaningful reward signals are observed, complicating the learning process.

This project investigates how to design effective reward functions for such sparse-reward, multi-turn settings using multi-turn RLHF (MT RLHF) methods Wu et al. (2025). We focus on the 20 Questions benchmark Abdulhai et al. (2023), where a language model ("guesser") must identify a hidden object by asking yes/no questions to an oracle over multiple turns. This task exemplifies the importance in attributing credit to intermediate actions when final rewards are sparse.

Our objective is to improve the guesser's ability to ask strategically informative questions by leveraging proxy rewards such as interactivity and information gain. Specifically, we evaluate whether MT RLHF can increase model accuracy in this sparse-reward environment and whether proxy metrics that emphasize uncertainty reduction better guide learning compared to traditional engagement-based rewards.

2 Related Work

2.1 Multi-turn Reinforcement Learning with Human Feedback

Although reinforcement learning with human feedback (RLHF) has succeeded in single-turn tasks such as instruction following Ouyang et al. (2022), many real-world tasks such as interactive games or teaching require reasoning over a sequence of turns. Recent work extends RLHF to these multi-turn (MT) settings by using simulated users to supervise full dialogue trajectories Shani et al. (2024); Zhou et al. (2024). CollabLLM Wu et al. (2025) further advances this by introducing intrinsic rewards such as brevity and causal influence to encourage efficient collaboration. Brevity was measured by tokens generated, and causal influence was used to measure the impact of an assistant on future dialogue trajectories. However, these methods focus on dense-reward tasks (e.g., coding or editing) and often optimize for outcomes like conciseness or response fluency.

Our work departs from these efforts by focusing on sparse-reward exploratory dialogue, where progress is gradual and guided by asking informative questions. We specifically study the 20 Questions game, where the model must refine its hypothesis through a sequence of yes/no queries. In this domain, rewarding brevity or individual question quality (as in DPO-style supervision) may not capture the chain of reasoning across dialogue turns. Instead, we use the MT-RLHF framework precisely because it supports trajectory-level supervision without requiring a hand-crafted reward function. It also enables us to reward high-level properties such as interactivity and information gain, that might be better proxies to enable strategic progress in sparse tasks like 20Q.

2.2 Sparse Rewards and the 20 Questions Task

Sparse reward environments challenge standard RL methods, as feedback is delayed until task completion. Previous approaches to the 20 Questions task typically employ heuristic methods, including LSTM-based dialogue policies Zhao and Eskenazi (2016) and post-hoc global entropy reduction Hu et al. (2019).Although these recognize the importance of uncertainty reduction, they do not model the evolving role of each question within a dialogue. LMRLGym Abdulhai et al. (2023) includes 20Q as a benchmark, but current solutions rely on supervised fine-tuning of single questions, not multi-turn strategy.

We build on this line of work by introducing proxy reward metrics that reflect intermediate progress, namely, interactivity (how much the assistant engages in clarifying or narrowing questions) and information gain (how much uncertainty is reduced given the dialogue history). These metrics are scored by a learned judge model and used to fine-tune a small Llama 3.2 1B assistant via MT-RLHF. In contrast to DPO, which treats question quality in isolation and optimizes individual turns independently, our approach leverages multi-turn RLHF to reflect the sequential nature of 20Q, where each question is context-dependent and directed toward a cumulative goal.

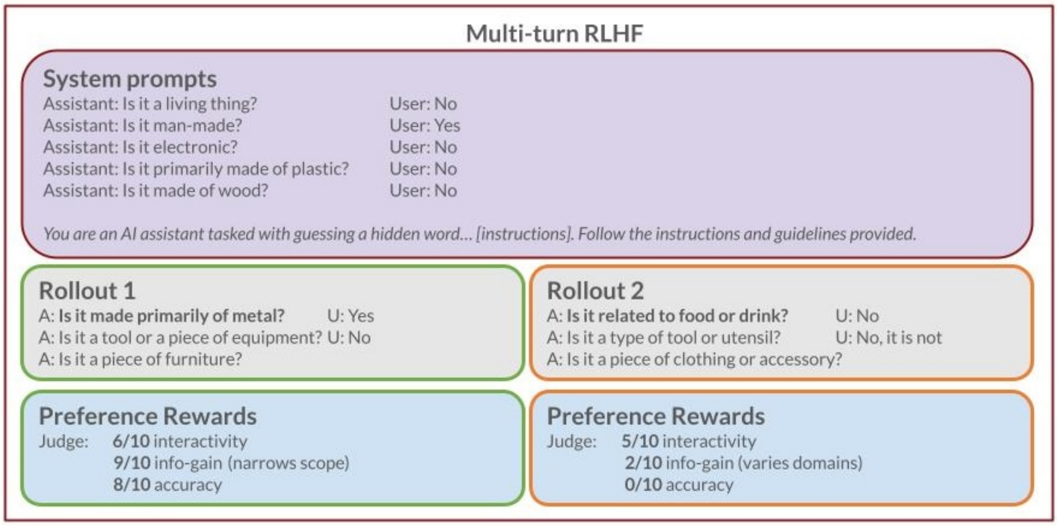
The combination of MT RLHF and a sparse reward space like the 20 questions game has not yet been investigated in past literature.

3 Method

The MT RLHF method is inspired by CollabLLM Wu et al. (2025), which uses intrinsic metrics such as brevity and extrinsic metrics such as causal influence for dense-reward tasks. We adapt the MT RLHF framework to sparse-reward, exploratory dialogue. We evaluate our approach on a 20 Questions (20Q) guessing game, chosen for its sparse-reward structure and need for multi-turn reasoning.

We construct a dataset where a user simulator, conditioned on a target object, engages in 20-round dialogues with an assistant model, whose goal is to identify the object via yes/no questions. Each assistant turn is scored by a judge model along three metrics: interactivity, information gain, and final answer accuracy. Since accuracy is a sparse, terminal reward, we leverage interactivity and information gain as dense proxy signals to assign multi-turn rewards.

Interactivity reflects the assistant's ability to stay on-topic, follow game rules, and sustain coherent, engaging dialogue. It is essential for maintaining natural, user-aligned conversations–central to CollabLLM's goal of optimizing multi-turn user satisfaction. Accuracy is the sparse ground-truth



Assistant: Is it a living thing? User: No
Assistant: Is it man-made? User: Yes
Assistant: Is it electronic? User: No
Assistant: Is it primarily made of plastic? User: No
Assistant: Is it made of wood? User: No

You are an AI assistant tasked with guessing a hidden word... [instructions]. Follow the instructions and guidelines provided.

Rollout 1
A: **Is it made primarily of metal?** U: Yes
A: Is it a tool or a piece of equipment? U: No
A: Is it a piece of furniture?

Rollout 2
A: **Is it related to food or drink?** U: No
A: Is it a type of tool or utensil? U: No, it is not
A: Is it a piece of clothing or accessory?

Preference Rewards
Judge: **6/10** interactivity
9/10 info-gain (narrows scope)
8/10 accuracy

Preference Rewards
Judge: **5/10** interactivity
2/10 info-gain (varies domains)
0/10 accuracy

Figure 1: Multi-turn RLHF setup for 20 Questions. The assistant interacts with a user simulator to identify the user's hidden object through yes/no questions. A judge model, who knows the hidden object, scores each rollout on interactivity, information gain, and final accuracy. These scores enable preference-based learning without a handcrafted reward function. Shown are two rollouts with contrasting reward profiles.

objective that evaluates whether the assistant correctly identifies the target object by the end of the game; it reflects final task success but provides no learning signal during intermediate turns. Information gain captures how much each question reduces the hypothesis space, functioning as a reward shaping mechanism that encourages strategic inquiry even before the final answer. It serves as a dense approximation to accuracy and directly incentivizes efficient reasoning. In all three cases, higher scores (not necessarily magnitudes) indicate stronger alignment with their respective behavioral objectives.

We use `gpt-4o-mini` OpenAI (2024) as the user, assistant, and judge models for data generation. It was selected for its low cost and strong performance. We assume that the user and judge produce reasonable, rule-abiding responses, and that the assistant adheres to game constraints. For instance, the assistant avoids asking multiple questions per turn, and the user responds strictly with "yes" or "no," maintaining the structured turn-taking required by the 20Q format. We manually inspect a subset of rollouts to verify that the judge produces reliable scores and that the user model generates realistic, human-like answers; it does.

We define two proxy reward functions to induce preferences:

$$R_{\text{interactivity}} = 2 \cdot (\text{Interactivity} - 2.5) + \text{Accuracy} \qquad \text{(generic / math-style)}$$

$$R_{\text{infogain}} = 2 \cdot \text{InfoGain} + \text{Accuracy} \qquad \text{(20Q-specific)}$$

$R_{\text{interactivity}}$: centers interactivity (from $[1, 5] \to [-3, +5]$) and combines it with accuracy. R_{infogain}: prioritizes information gain (scaled from $[0, 1] \to [0, 2]$) and increases accuracy.

These task-specific rewards extend the CollabLLM framework to sparse-reward settings. They allow us to generate multi-turn preference data with a judge model scoring both reward types.

For the assistant, we fine-tune `llama-3.2-1b-instruct` Meta_AI (2024) using DPO over full multi-turn trajectories. Each training example consists of two assistant rollouts (preferred and rejected) conditioned on the same user and object. Each trajectory includes the full turn history, the chosen vs rejected assistant turn, and the judge's explanation. The judge determines preference between each pair based on the three metrics.

DPO optimizes the assistant by increasing the likelihood of preferred trajectories while penalizing rejected ones. The objective is:

$$\mathcal{L}_{\text{DPO}} = \log \sigma \left(\beta \left[\log \pi(x, y_{\text{preferred}}) - \log \pi(x, y_{\text{rejected}}) \right] \right) \tag{1}$$

where π is the assistant policy, β is a temperature parameter, and σ is the sigmoid function. In this paper, we use a temperature of 0.7 to generate data, train models, and evaluate. We ran ablations at $\beta = 0.3, 0.5, 0.9$, which showed similar trends.

The novel contributions to the multi-turn RLHF research space include:

- Designing novel reward functions for sparse task structure (information gain, interactivity).
- Generating multi-turn preference data using these rewards and a judge model.
- Fine-tuning a 1B parameter assistant using DPO across both reward types.
- Evaluating trained models on 110 held-out objects for both reward tasks.

4 Experimental Setup

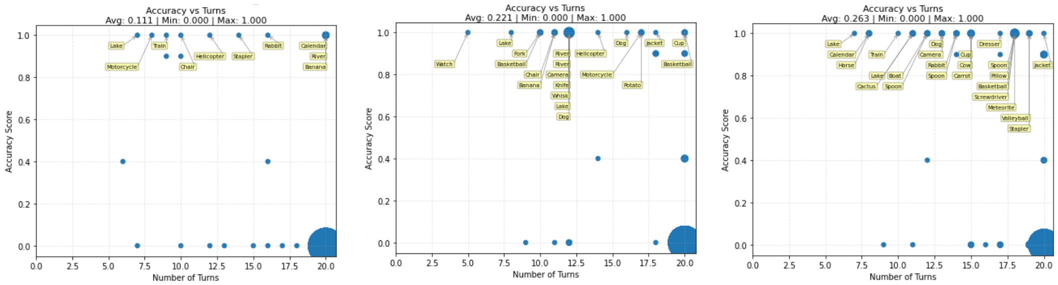
We considered several candidate reward function designs, including entropy reduction (measuring how much a question narrows the hypothesis space), search space coverage (whether the object could be identified within 20 questions), and multi-part question facilitation. While the base model already showed reasonable search space coverage and understood to play the game, it still had a low accuracy score. To increase the accuracy score, we focused on information gain and interactivity as primary proxy rewards. We compare to the base model Meta Llama 3.2 1B Instruct to evaluate the improvement of these reward models.

Our publicly available MT RLHF datasets contain over 5200 and 1200 preference pairs respectively, released as the 20q interactive dataset and 20q infogain dataset. These rollouts exhibit consistent and strategic behavior, providing a strong foundation for the next step: training.

We finetuned 2 models from our base model using DPO on each of the datasets, respectively found at 20q interactive model and 20q infogain model. A difficulty included overlapping rollouts between the rejected and chosen DPO pairs. To reduce the negative impact of this type of data, the DPO pairs with a score difference of less than 0.2 were removed from the training data to improve the model's learning.

5 Results

Evaluating each of the base model, interactive model, and infogain model on 20-turn games over 110 objects, we find a statistically significant increase in accuracy of the interactive and infogain tuned models over the base model.



(a) Base model

(b) Interactive model

(c) InfoGain model

Figure 2: Evaluation results over 110 objects for each model. Tested at temp=0.7 (same as when training). The points' radii logarithmically correspond to how many objects categorize to that position.

Method	Interactivity Metric	Accuracy Metric	Information Gain Metric
Baseline	-3.4	0.11	-3.9
Interactive Model	-3.3	0.22	-3.7
Infogain Model A	-3.3	0.26	-3.7

Method	Avg turns (overall)	Avg turns (accurate rollouts)	Avg num tokens per question
Baseline	18.64	13.6	105
Interactive model	18.26	14.29	111
Infogain model A	18.36	13.05	107

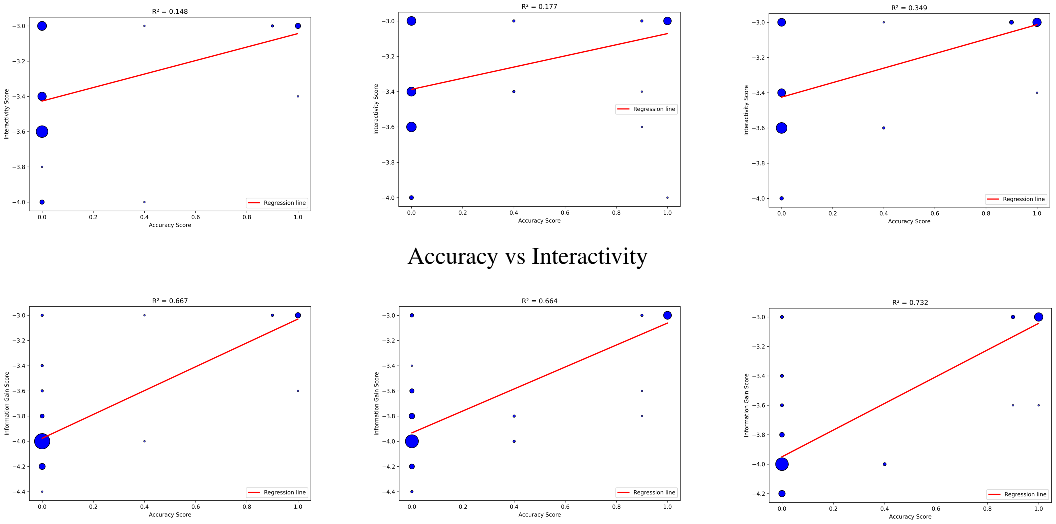
Table 1: Evaluation on 20 Questions: Performance Comparison of Trained Models

5.1 Quantitative Evaluation

Table 1 and Figure 2 reports performance metrics across the baseline, interactive, and infogain-trained models. Accuracy improves substantially with each level of training: the interactive model achieves 22% accuracy (a 100% relative increase from the baseline's 11%), while the infogain model reaches 26% accuracy (a 136% increase). We run a McNemar contingency table test to show that these gains are statistically significant: the interactive model outperforms the baseline with $p = 0.0339$, and the infogain model outperforms the baseline with $p = 0.0066$.

All models take roughly the same number of total turns per game (about 18), but the interactive model uses more turns (14.29) than the baseline (13.6) when accurate, suggesting each question might not be as powerful. The infogain model, on the other hand, uses 13.05 turns on average when accurate, suggesting each question is slightly more powerful than the baseline.

The average number of tokens per question represents how much text the assistant generates per each of its questions. Token usage slightly increases for the interactive model (111 tokens/rollout) compared to baseline (105), which suggests that, although the interactive model is double as accurate as the baseline, it is more loquacious.



Accuracy vs Interactivity

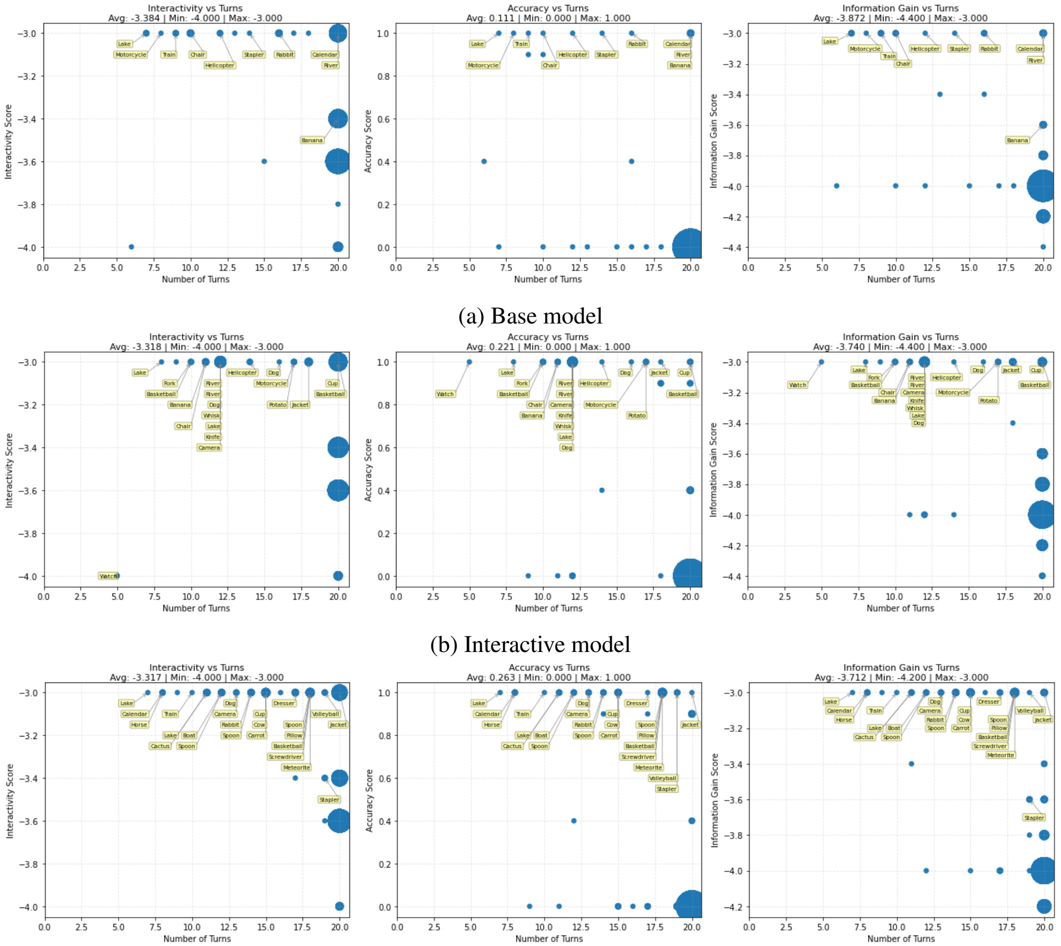
Accuracy vs Information Gain

(a) Over 110 trials, the proxy measures of interactivity (top) and 2. information gain (bottom) are plotted against the accuracy of the model. To the left is the base model, in the middle column is the interactive model, and to the right is the infogain model. The points' radii linearly correspond to how many objects categorize to that position. Information gain has a much stronger correlation with accuracy (r^2 = .66-.73) than interactivity and accuracy (r^2 = .14-.34), which has no correlation. It is notable that there are non-binary accuracy reward signals. Accuracy scores that aren't 0 and 1 correspond to the model final response being very close to the correct final reward (eg. guessing yam instead of potato).

Aside from the most important accuracy scores, we can also look at the interactivity and information gain metrics in figure 4. Interactivity scores show marginal improvement across models (from -3.4 to -3.3), while the information gain metric improves slightly from -3.9 (baseline) to -3.7 (infogain). Despite modest shifts in proxy metrics, the infogain metric correlates strongly, $r^2 > .65$, with accuracy (see figure 3a). This tells us that for scaling this reward function to other objectives like car dealership, information gain is more likely transferable since it is more directly correlated to accuracy as compared to interactivity, which had a weak correlation with $r^2 < 0.35$.

Overall, we see that the infogain model is over double as accurate, and marginally quicker at getting to the correct answer. Meanwhile, the interactive model is more loquacious, yet much more accurate than the baseline model.

These results suggest that training with reward signals targeting either interactivity or information gain leads to more accurate and effective dialogue strategies, with information gain providing the most consistent improvements across metrics.



(a) Base model

(b) Interactive model

(c) InfoGain model

Figure 4: Full evaluation results on all three metrics. Similar to the figure in the paper, models are evaluated over 110 objects for each model. All hyperparameters fixed.

5.2 Qualitative Analysis

In figure 4, we can see the interactivity, accuracy, and information gain of all three models, and all the accurate objects are also labeled. In this following section, we delve into some of those objects to understand their scores.

Case study 1: Target Object is Camera

Both the interactive and infogain model correctly guess "camera" using slightly different strategies.

All models begin with common foundational questions such as “Is it man-made?”, “Can you hold it in your hand?”, and “Does it have moving parts?”. The base model demonstrates a logical progression and builds on previous answers, but it sometimes diverges with questions that revisit irrelevant or categorical (as opposed to binary) sections of the search space (e.g., “Is it found in the kitchen?”, “in the bathroom?”, “in the garage?”). This limits its ability to systematically reduce the search space.

The interactive model, in contrast, prioritizes questions that are both engaging and relevant–e.g., “Is it man-made?”, “Does it have buttons?”, and “Is it related to a hobby?”–which collectively narrow the hypothesis space more efficiently. It does not always demonstrate maximal information gain.

The infogain model focuses even more tightly on uncertainty reduction. For instance, it pursues material properties after knowing the object is a tool (“Is it metal?”, “Is it plastic?”) to rule out large branches of the object space. Unlike the Interactive model, which shifts categories (e.g., “Is it jewelry?” after “Is it metal?”), the infogain model continues down consistent semantic paths, leading to faster elimination of unlikely options.

Ultimately, both the interactive and infogain models correctly identify camera, while the base model does not. The base narrows in on vague descriptors–e.g., “a tool used by a profession,” or “found in a garage”–that are technically true of a camera, but not uniquely informative. This case highlights how proxy rewards like interactivity and information gain guide the model toward more targeted, high-yield questioning strategies.

Case study 2: Target Object is Lake

All three models correctly guessed “lake,” but used different strategies. The base model started with low-yield questions like “can it be held in your hand” and jumped to specific guesses–like “mountain” and then “lake”–within 7 turns. The Interactive model ruled out living and land-based objects, then focused on water, guessing “lake” as its first specific object in 8 turns. The infogain model asked about man-made vs. natural and geological types, reaching “lake” in 7 turns through deliberate elimination. Each model approached the space differently in terms of structure, pacing, and how it handled narrowing.

Case study 3: Target Object is Cactus

All three models attempted to guess "cactus," but only the infogain model was correct. The base model identified the object as a plant but quickly moved into subcategories like seeds and mushrooms, missing the scope of the true object. The interactive model also focused too quickly on features like the smell of the plant and types of leaves, skipping over the object’s scope. Contrastingly, infogain methodically asked about plant types–flowers, fruits, trees–and narrowed the space at a slightly lower rate, which proved to be successful. This demonstrates the infogain model’s ability to consistently reduce scope without overshooting the answer.

6 Discussion

Our results indicate that information gain serves as a more faithful proxy for downstream task success than interactivity in sparse, multi-turn environments. However, the accuracy gains remain relatively modest in absolute terms, suggesting that stronger reward weights or more epochs training may be necessary.

One limitation is the cost of generating multi-turn preference data: while we used DPO over 2000-5000 dialogues to train, many pairs lacked strong preference contrast, particularly when using similarly capable models (e.g., Llama 3.2 1B and GPT-4o-mini). Additionally, our information gain metric requires access to the true object label, which may not be available in other sparse tasks. We also assume that proxy metrics are robust across variations in user behavior, but this assumption may not hold when question styles or exploration strategies differ significantly. For example, when we evaluated some target objects like "lake" and "dog" multiple times (they reoccurred in LMRL Gym’s evaluation dataset), some models only got it correct one of the times. This could be studied further

Effective search space reduction via information gain relies on a model with strong world knowledge. Given that the infogain model used only 2,000 training points, the substantial improvements suggest that the base model already possessed much of the general knowledge required for the 20Q task and only missed some strategic insights.

7 Conclusion

This work demonstrates that multi-turn reinforcement learning from human feedback (MT RLHF) with proxy rewards can significantly improve model alignment in sparse-reward, long-horizon tasks. On the 20 Questions benchmark, optimizing for information gain yields a 2.3× increase in final accuracy compared to the base model, outperforming interactivity-based tuning. Information gain rewards effectively accelerate hypothesis space reduction, while interactivity rewards maintain engagement and coherence–both essential for multi-turn reasoning. Case studies further show how different reward types induce distinct questioning strategies and patterns of search-space reduction. Our correlation plots indicate that rewards targeting uncertainty reduction better correlate with accuracy and overall task success than those focusing solely on conversational fluency. This infogain-proxy driven MT RLHF seems promising to scale for similar sparse tasks without reliance on costly ground-truth supervision.

Future research could explore adaptive combinations of accuracy, interactivity, and information gain rewards. We could apply MT RLHF to other sparse-reward interactive domains such as LMRL Gym's Car Dealer. Such models have potential to enhance consumer-facing assistants by improving long-term problem-solving efficiency and significantly reducing computational costs.

In addition, this work opens the door to online reinforcement learning extensions, where evaluation rollouts could be added to a replay buffer to enable continual improvement. By retrospectively tagging these rollouts as successful or unsuccessful, we can generate additional training signals in otherwise sparse-reward settings. Such an approach could be particularly valuable as MT RLHF methods are deployed in real-world chatbot applications, enabling systems to learn dynamically from ongoing interactions.

8 Team Contributions

- **Aditi:** All methods, experiments, evaluations and writeups. I used the royal "we" in this paper for formality.

Changes from Proposal No significant deviations were made from the proposal.

References

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. 2023. LMRL Gym: Benchmarks for Multi-Turn Reinforcement Learning with Language Models. arXiv:2311.18232 [cs.CL] https://arxiv.org/abs/2311.18232

Huang Hu, Xianchao Wu, Bingfeng Luo, Chongyang Tao, Can Xu, Wei Wu, and Zhan Chen. 2019. Playing 20 Question Game with Policy-Based Reinforcement Learning. arXiv:1808.07645 [cs.HC] https://arxiv.org/abs/1808.07645

Meta_AI. 2024. Llama 3.2-1B. https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct. Accessed: 2025-05-21.

OpenAI. 2024. GPT-4o-mini. https://docs.aimlapi.com/api-references/text-models-llm/openai/gpt-4o-mini. Accessed: 2025-05-16.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Rémi Munos. 2024. Multi-turn Reinforcement Learning from Preference Human Feedback. arXiv:2405.14655 [cs.LG] https://arxiv.org/abs/2405.14655

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. CollabLLM: From Passive Responders to Active Collaborators. arXiv:2502.00640 [cs.AI] https://arxiv.org/abs/2502.00640

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning. arXiv:1606.02560 [cs.AI] https://arxiv.org/abs/1606.02560

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. ArCHer: Training Language Model Agents via Hierarchical Multi-Turn RL. arXiv:2402.19446 [cs.LG] https://arxiv.org/abs/2402.19446