A IMPLEMENTATION DETAILS

A.1 IMAGINATION ENGINE

Here, we show the prompts we used to generate imagined dialogue using our imagination engine, as well as sampled results.

Instruction Task. We use the following reasoning prompt in order to generate different personas:

Prompt: We are designing a task where an agent needs to teach a human about some concept in reinforcement learning. The human may have very different knowledge backgrounds that require the agent to personalize their instruction. Can you provide a list of different knowledge backgrounds that the human may have (e.g. layman, AI expert, etc.) where the agent needs to change their teaching content for different backgrounds?

from which the LLM synthesized the following possible $z \in \mathcal{Z}$:

(1) layman: little to no prior knowledge of reinforcement learning
(2) machine learning enthusiast: familiar with machine learning but not reinforcement learning
(3) reinforcement learning student: has some experience with reinforcement learning basics.
(4) reinforcement learning researcher: knows most topics in reinforcement learning and can delve into advanced topics

In the imagination step, we are able to sample imagined successful or unsuccessful dialogues of an agent trying to explain an RL concept to a human with one of the backgrounds $z \in \mathcal{Z}$. Below is the imagination prompt and example imagined dialogue of an agent explaining behavior cloning to a machine learning enthusiast:

Prompt: Generate a dialogue between an AI teacher and a {machine learning enthusiast} where the machine learning enthusiast asks the teacher for an explanation of behavior cloning. The machine learning enthusiast is {familiar with machine learning but not reinforcement learning}. The teacher should perform information-gathering to figure out the machine learning enthusiast's background knowledge in AI and reinforcement learning, by asking questions and engaging in dialogue with the {machine learning enthusiast}. Finally, the teacher gives an explanation that the {machine learning enthusiast} {understands by tailoring the explanation to their background}. In the dialogue, the teacher's utterances are prefaced by ”AI:” and the machine learning enthusiast's utterances are prefaced by ”H:”.

Output:
H: Hello, I have a basic understanding of machine learning, but I would like to understand “behavior cloning”.
AI: That's great. Do you know what reinforcement learning is and how it differs from other types of machine learning?
H: Yes, I know that in reinforcement learning, the AI agent learns from its own actions and experiences rather than being trained on a dataset like in supervised learning.
AI: Excellent. So behavior cloning is a type of supervised learning where the AI is trained to mimic a specific behavior by learning from examples provided by an expert.
H: So, does this mean that the AI is copying the expert's behavior exactly?
AI: Not exactly. The AI is learning from the expert's actions but is also using its own decision-making process to adapt and improve upon the behavior. Think of it as a combination of imitation and trial-and-error learning.
H: That's interesting. Can you give me an example of how behavior cloning can be used in real life?
AI: Sure. Let's say you want to teach a robot to perform a specific task, such as making a cup of coffee. With behavior cloning, the AI agent would observe and learn from a human expert making coffee, and then be able to perform the task on its own with a similar level of proficiency.
H: I see. So it's like teaching the AI through demonstration rather than giving it explicit instructions.
AI: That's correct. And the beauty of behavior cloning is that it can also be used in combination with reinforcement learning techniques to further improve the AI's performance.
H: Wow, that's really fascinating. Thank you for explaining it so clearly.

where the brackets indicate sampled properties (either the topic being taught, persona of the human, or the reward of the trajectory) that automatically get co-opted into the template. For the critique step, for the criteria that the human should not immediately reveal their background knowledge, we used the prompt and would achieve the sample revised dialogue:

Prompt: The following dialogue is of a {machine learning enthusiast} ("H:") asking a AI teacher ("AI:") to explain {behavior cloning}:
...
Rewrite the dialogue so that the {machine learning enthusiast} does not immediately reveal that they are {familiar with machine learning but not reinforcement learning}, and instead have the agent discover this naturally through conversation by asking questions.

Output:
H: Hi there, can you tell me about behavior cloning? I've heard of it in the context of machine learning but I'm not too familiar with it.
AI: Sure, I'd be happy to explain. Are you familiar with reinforcement learning and how it differs from other types of machine learning?
H: Actually, no. Can you tell me more about it?
AI: Of course. In reinforcement learning, the AI agent learns from its own actions and experiences, rather than being trained on a dataset like in supervised learning.
H: Interesting. So how does behavior cloning fit into all of this?
AI: Behavior cloning is a type of supervised learning where the AI is trained to mimic a specific behavior by learning from examples provided by an expert.
H: So it's like teaching the AI through demonstration?
AI: Yes, that's one way to think about it. It's a combination of imitation and trial-and-error learning.
H: Can you give me an example of how behavior cloning is used in real life?
AI: Yes, a common example is teaching a robot to perform a specific task, such as making a cup of coffee. The AI would observe and learn from a human expert, and then be able to perform the task on its own.
H: That's really impressive.

where again brackets indicated sampled quantities. Our IE was used to generate 50K of such imagined dialogues in this manner.

Preference Elicitation Task. In our designed travel agent task, we have the following 18 activities (organized hierarchically):

1. Explore nature: hiking, safaris, rock climbing, ziplining
2. Beach: sunbathing, water sports
3. Wellness: spa, yoga classes, meditation sessions
4. Food: food tours, restaurants, cooking classes
5. Shopping: malls, local markets
6. Cultural experiences: city sightseeing, museums, historical sites, theater performances

First, in the reasoning step, we used the following reasoning prompt:

Prompt: We are designing a task where an agent needs to behave as travel agent and recommend activities for human customers to do. Identify a list of personas that a human may have, where the travel agent must tailor their responses and recommendations to interact with that human.

from which we obtained from the LLM the following possible personas $z \in \mathcal{Z}$:

(1) adventurer: prefers unique, off-the-beaten-path experiences that they have not done before
(2) indecisive traveler: gets overwhelmed with broad questions with too many choices
(3) budget conscious traveler: prefers cost-effective options or package deals
(4) luxury seeker: prefers premium accommodations and upscale experiences
(5) family planner: seeks family-safe activities that children will enjoy

Then in the imagination step, a sample $z \in \mathcal{Z}$ is a cross-product of what type of activities the human enjoys doing and one of the synthesized personas. Below is an example imagination prompt and imagined dialogue output:

Prompt: Generate a dialogue of a human customer asking a travel agent to recommend activities to do. The human is {budget-conscious} and {prefers cost-effective options or package deals} and likes {food} activities such as {food tours}. The travel agent asks questions to uncover what activities the human may like. The travel agent is {unsuccessful} and finds an activity that the human {does not like}, resulting in {negative} sentiment from the human. The travel agent's utterances should be prefaced with 'AI:' and the human customer's utterances should be prefaced with 'H:'.

Output:
AI: Hello, welcome to our travel agency. How can I assist you?
H: Hi there, I'm looking to plan a trip for my upcoming vacation.
AI: Great! Where are you looking to go and when?
H: I'm thinking of going to Europe next month.
AI: That's a fantastic choice. Europe has so many exciting destinations. What type of activities do you enjoy?
H: I'm a foodie, so I would love to try some local cuisines and go on food tours.
AI: Noted. Are you looking for any specific destinations within Europe?
H: Not really, as long as it's within my budget.
AI: Of course, we have some fantastic budget-friendly options for food activities and package deals. How long is your trip?
H: It'll be a week-long trip.
AI: In that case, I would recommend our food tour package in Italy. You'll get to experience authentic Italian cuisine while exploring the beautiful cities of Rome and Florence.
H: That sounds amazing, but I've been to Italy before and I was hoping to visit a new country.
AI: I see. Perhaps you would enjoy our food and wine tour in Spain. It's a 5-day package that includes visits to local wineries and tapas tasting.
H: Hmm, that does sound interesting. How much would that cost?
AI: It's currently priced at $1000 per person, including accommodation and transportation.
H: That's a bit higher than my budget. Are there any other options?
AI: Yes, we have a package deal that combines the food tour in Paris with a city tour.
H: Perfect! How much is the package deal?
AI: It's $150 per person, which includes all food tastings, the guided tour, and the sightseeing activities.
H: That's within my budget. Thank you for your help.
AI: My pleasure. Have a great trip.

where the brackets indicate sampled properties (either the activity the human likes, persona of the human, or the reward of the trajectory that automatically) that get co-opted into the template. For the critique step, for the criteria that the human should express sentiment that matches the reward of the trajectory, we used the following critique prompt that would achieve the sample revised dialogue:

Prompt: Below is a dialogue between a human customer ("H:") and a travel agent ("AI:") where the human is asking the travel agent to provide recommendations for activities to do.
...
Rewrite the dialogue so that at the end, the travel agent is {unsuccessful} in finding an activity that pleases the customer, and the customer shows {negative} sentiment because of it.

Output
AI: Hello, welcome to our travel agency. How can I assist you?
H: Hi there, I'm looking to plan a trip for my upcoming vacation.
AI: Great! Where are you looking to go and when?
H: I'm thinking of going to Europe next month.
AI: That's a fantastic choice. Europe has so many exciting destinations. What type of activities do you enjoy?
H: I'm a foodie, so I would love to try some local cuisines and go on food tours.
AI: Noted. Are you looking for any specific destinations within Europe?
H: Not really, as long as it's within my budget.
AI: Of course, we have some fantastic budget-friendly options for food activities and package deals. How long is your trip?
H: It'll be a week-long trip.
AI: In that case, I would recommend our food tour package in Italy. You'll get to experience authentic Italian cuisine while exploring the beautiful cities of Rome and Florence.
H: That sounds amazing, but I've been to Italy before and I was hoping to visit a new country.
AI: I see. Perhaps you would enjoy our food and wine tour in Spain. It's a 5-day package that includes visits to local wineries and tapas tasting.
H: Hmm, that does sound interesting. How much would that cost?
AI: It's currently priced at $1000 per person, including accommodation and transportation.
H: That's a bit higher than my budget. Are there any other options?
AI: Unfortunately, all our food tours are currently over your budget. Would you like me to check for any other options?
H: No, thank you. I guess I'll have to keep searching on my own.
AI: I apologize for not being able to find a suitable activity for you. Is there anything else I can assist you with?
H: No, that's all. Thanks for your efforts.
AI: I understand your disappointment. I hope you find a perfect activity for your trip.

where again brackets indicated sampled quantities. Our IE was used to generate 50K of such imagined dialogues in this manner.

A.2 RL TRAINING

We use the hyperparameters reported in Table 3. All algorithms were trained on a single V100 GPU until convergence, which took less than 3 days.

Hyperparameter	Setting
ILQL τ	0.8
ILQL α	0.0
Discount factor	0.99
Batch size	128
Target network update α	0.005
Number of updates per iteration	60
Number of iterations	100
Optimizer	AdamW
Learning rate	1e-4

Table 3: Hyperparameters used during training.

B EXAMPLE DIALOGUES

B.1 COMPARING GPT AND IE+ILQL AGENT

Instruction task. Below are examples of dialogues in the instruction task by the two agents obtained by interaction during the user study.

GPT Agent Dialogue:

H: Hi, I've been hearing about something called ”behavior cloning” in AI, but I don't really know what it is. Can you explain it to me?

AI: Certainly! Let's start by understanding your background in AI to provide an explanation that matches your level of knowledge. Please answer the following questions: 1. Have you ever studied or worked with artificial intelligence or machine learning concepts before, even at a basic level? 2. Are you familiar with terms like ”neural networks,” ”training data,” or ”supervised learning”? 3. Are you interested in a high-level overview or a more detailed explanation of behavior cloning?

H: I have some basic knowledge of artificial intelligence and machine learning, but I wouldn't say I'm an expert. I am familiar with neural networks and supervised learning, but not as much with training data.

AI: Great, thank you for that information. Behavior cloning is a method used in reinforcement learning, which is a subcategory of machine learning. In reinforcement learning, an AI system learns to make optimal decisions by interacting with its environment and receiving feedback in the form of rewards or punishments. Behavior cloning, on the other hand, is a technique where an AI model imitates or ”clones” the behavior of an already established agent, instead of learning from scratch through trial and error. This is often used when there is a large amount of data available from an expert or a pre-existing system. Does that make sense so far?

H: Yes, that's starting to make sense. What do you mean by expert?

AI: An expert, in this context, refers to a human or an advanced AI system that has already learned how to solve a particular task or problem. The AI model then observes and imitates this expert's behavior to learn how to perform the task itself. Does that clarify what behavior cloning is?

H: Yes, thank you.

GPT Agent Dialogue:

H: Hi there. Can you explain what ”actor-critic” is in reinforcement learning?

AI: Of course, I am happy to explain ”actor-critic” in the context of reinforcement learning. Before I do, may I ask what your current level of understanding is of reinforcement learning?

H: I have a basic understanding of reinforcement learning, but I don't have much experience with it.

AI: Great, thank you for letting me know. ”Actor-critic” is a hybrid approach to reinforcement learning that combines the advantages of both policy-based and value-based methods. It consists of two components - an actor and a critic. The actor is responsible for selecting actions, while the critic evaluates the actions chosen by the actor. In simpler terms, the actor is the decision maker and the critic is the performance evaluator. The key idea behind this approach is to have two separate networks that work together to optimize the agent's actions.

H: Can you explain why this approach is beneficial?

AI: Of course. Since the actor and critic work together, the actor can focus on improving its decision-making process, while the critic provides valuable feedback on the actions taken. This allows for faster and more accurate learning, as the agent can continuously adjust its actions based on the critic's evaluations. Additionally, the actor-critic approach can handle continuous action spaces, making it more versatile than traditional value-based methods. Overall, the actor-critic approach can improve the efficiency and performance of reinforcement learning agents by leveraging the strengths of multiple methods. Is there anything else you would like me to clarify?

H: Why can't value-based methods handle continuous action spaces?

AI: Value-based methods, such as Q-learning, rely on discretizing the action space, which can lead to poor performance in continuous environments. This is because discretization often results in a loss of information and can be computationally expensive. The actor-critic approach overcomes this limitation by using the actor network to map continuous action values, making it a more suitable method for continuous action spaces. Does that make sense?

H: Yes, thank you.

IE-RL Agent Dialogue:

H: Hi there! I've heard about this thing called "behavior cloning" in AI, but I don't really understand what it is. Can you explain it to me?

AI: Of course! I'd be happy to explain behavior cloning. To start, could you tell me if you've ever come across the terms "artificial intelligence" or "machine learning" before?

H: Yes, I've heard those terms, but I'm not exactly sure what they mean.

AI: No problem at all! Let's take it step by step. Have you ever used a computer or a smartphone?

H: Yes, I use my computer and phone all the time.

AI: Great! So, imagine that a computer or a smartphone could learn to do something by watching you do it. That's a bit like what behavior cloning is.

H: Oh, interesting. Can you explain more about how that works?

AI: Sure. Behavior cloning is a type of machine learning algorithm where a computer learns to perform a task by imitating a human's behavior. It does this by closely analyzing and mimicking the actions of a human expert, using a large dataset of examples. This approach is often used in tasks that don't have a clear set of rules or can't be easily solved through traditional programming methods. Does that make sense?

H: Yes, I think so. So how does the computer learn to mimic the behavior?

AI: That's a great question. The computer is fed a large amount of data that shows different examples of the task being performed by a human expert. The computer then uses this data to identify patterns and develop a model of the behavior. This model allows the computer to predict the correct actions to take in a given situation. Over time, the computer can continue to refine its model and improve its performance. Does that help clarify things?

H: That does. Thank you!

IE-RL Agent Dialogue:

H: I've been hearing a lot about "model-based reinforcement learning" lately, but I'm not entirely sure how it works. Can you explain it to me?

AI: Sure, I'd be happy to explain. Just to gauge your background a bit, have you worked with machine learning or AI concepts before?

H: Yes, I have some experience with machine learning, but I haven't delved into reinforcement learning specifically.

AI: Great, then I can tailor my explanation to your existing knowledge. Reinforcement learning is a type of machine learning that involves training a model to make sequential decisions in order to reach a specific goal. It works by using a reward system, where the model receives positive or negative feedback based on how well its decisions lead it towards the goal. Model-based reinforcement learning specifically involves using a model, or representation, of the environment in which the model is making decisions. This allows the model to simulate different scenarios and make more informed decisions based on its understanding of the environment. Does that make sense?

H: How do you obtain a model?

AI: The model is typically created through a training process, where the model is fed data and learns to optimize its decision-making. This can be done through various techniques such as deep learning or Monte Carlo simulations. The model is then used as a guide for the agent, or decision-maker, to make the best decisions in a given environment. Does that clarify how model-based reinforcement learning works?

H: What happens if you do not have a model?

AI: Without a model, reinforcement learning becomes more challenging as the agent would have to rely on trial and error to make decisions. This is known as model-free reinforcement learning and it often requires a larger number of interactions with the environment to make optimal decisions. However, some advancements in reinforcement learning have combined model-based and model-free techniques to achieve better results.

H: That makes sense. Thank you!

Preference elicitation task. Below are examples of dialogues in the travel agent task by the two agents obtained by interaction during the user study.

GPT Agent Dialogue:
H: I am looking for recommendations for activities to do around here.
AI: Of course! I'd be delighted to help you find the perfect activities. To get started, could you tell me a bit more about your interests? Are you looking for adventure, culture, relaxation, great food, shopping, nightlife, nature, history, or wellness experiences? Your preferences will help me tailor my recommendations to your liking.
H: I guess I like being outdoors.
AI: Great! Enjoying the outdoors can be a wonderful experience. We have several options for outdoor activities. If you enjoy a bit of adventure, you might want to consider an adventure hike through our scenic trails, or perhaps some outdoor adventure activities like kayaking or zip-lining. If you prefer a more relaxed outdoor experience, you could explore the local parks for some nature appreciation, or even take a sightseeing tour to discover the beautiful landscapes around here. How does that sound?
H: I'm not sure. I guess I like doing both.
AI: That's perfectly fine! It sounds like you have a diverse range of interests. How about combining both adventure and relaxation to make the most of your stay? I recommend starting your day with an adventure hike or some exciting outdoor activities to get your adrenaline pumping. Later in the day, you can unwind by spending some time in a beautiful natural setting, like a local park or a beach. This way, you'll get the best of both worlds.
H: That sounds okay, but I only have time for one activity.
AI: I completely understand. Let's simplify your choices, then. If you only have time for one activity and you enjoy both adventure and relaxation, I recommend a scenic adventure hike. You'll get to immerse yourself in nature, enjoy some physical activity, and experience the serene beauty of the outdoors. It's a great way to make the most of your time while avoiding the overwhelm of too many options.

IE+RL Agent Dialogue:
H: I am looking for recommendations for activities to do around here.
AI: Hello! Welcome to our travel agency. Can you tell me a bit about what you enjoy during your travels?
H: Hi, yeah. I guess I like being outdoors sometimes.
AI: That's great! Exploring the outdoors can be quite refreshing. Are you more inclined towards peaceful natural spots or do you enjoy exploring cities and their cultural sites?
H: Um, I'm not sure. Both sound good to me.
AI: I completely understand. How about I narrow down some options for you? If you like peaceful spots, we could look at activities like leisurely garden walks or visiting serene lakesides. On the other hand, if you prefer cities, we could explore some local cultural events or guided city tours. Which of these options resonates more with you?
H: Uh, I guess the peaceful spots sound nice. But I also kind of like learning about the local culture.
AI: I appreciate your honesty. How about I suggest this: we could start with a visit to a beautiful botanical garden with serene landscapes?
H: Hmm, that actually sounds nice. I think you've helped a lot.
AI: I'm glad to have been of assistance. Don't hesitate to reach out if you need more information or if you'd like to explore further options. Have a wonderful day!

B.2 COMPARING IE+BC, IE+FBC, IE+ILQL AGENTS

Instruction task. Below is an example of dialogues in the instruction task by the three agents, where the human expresses confusion at the agent's initial explanation. The dialogues are obtained by emulating a human who overestimates their knowledge of reinforcement learning, then is unable to understand the agent's explanation.

IE+BC Agent Dialogue:
H: Can you explain it to me what ”offline reinforcement learning” is?
AI: Of course! Before I explain offline reinforcement learning, are you familiar with the basic concepts of reinforcement learning?
H: I’m familiar with reinforcement learning but not offline.
AI: Great! In that case, offline reinforcement learning, also known as batch reinforcement learning, is a specific approach where an agent learns from a fixed, pre-collected dataset of experiences rather than in reinforcement learning, where the agent is directly interacting with the environment in real-time.
H: I’m sorry. I didn’t understand your explanation.
AI: I apologize for not being clear. In simple terms, offline reinforcement learning is a technique where instead of learning from immediate interactions with the environment, the agent learns from a set of data or experiences that have been collected beforehand. This approach is often used when real-time interactions with the environment may not be feasible, such as in simulation environments or historical data. It allows for more efficient learning and can also be helpful in scenarios where the environment may change frequently. Does that make more sense now?

IE+FBC Agent Dialogue:
H: Can you explain it to me what ”offline reinforcement learning” is?
AI: Of course! Before I explain offline reinforcement learning, are you familiar with the basic concepts of reinforcement learning?
H: I’m familiar with reinforcement learning but not offline.
AI: Great! In that case, offline reinforcement learning, also known as batch reinforcement learning, is a specific approach where an agent learns from a fixed, pre-collected dataset of experiences rather than in reinforcement learning, where the agent is directly interacting with the environment in real-time.
H: I’m sorry. I didn’t understand your explanation.
AI: No problem! Reinforcement learning is a type of machine learning where an agent learns to make decisions through trial and error. The agent interacts with an environment and takes an action and receives rewards or punishments. Offline reinforcement learning is the same concept of reinforcement learning, except that the agent is learning from a fixed dataset of experiences that have already been collected. Because of this, the agent is not directly interacting with the environment, which is why it is called offline.

IE+ILQL Agent Dialogue:
H: Can you explain it to me what ”offline reinforcement learning” is?
AI: Of course! Before I explain offline reinforcement learning, are you familiar with the basic concepts of reinforcement learning?
H: I’m familiar with reinforcement learning but not offline.
AI: Great! In that case, offline reinforcement learning, also known as batch reinforcement learning, is a specific approach where an agent learns from a fixed, pre-collected dataset of experiences rather than in reinforcement learning, where the agent is directly interacting with the environment in real-time.
H: I’m sorry. I didn’t understand your explanation.
AI: My apologies. To help me understand your background a bit better, can you tell me if you are aware of the difference between online and offline learning in other domains of machine learning?
H: Yes, I am familiar with the concept of online and offline learning in other domains.
AI: Excellent! In that case, you can think of offline reinforcement learning as being similar to offline learning in other areas of machine learning. It is essentially a form of learning that occurs on a static dataset, rather than in real-time interactions with the environment, like in online reinforcement learning. Does that make sense?
H: Ah that makes sense. What do you mean by environment?
AI: The environment in reinforcement learning refers to the external factors that impact the decisions and actions of the agent, such as the physical surroundings, the rewards or punishments received, and the actions taken. In offline reinforcement learning, the agent learns from a dataset that simulates this environment, rather than directly interacting with it in real-time. Does that clarify the concept?
H: Yes, thank you for explaining that.
AI: You’re welcome! Is there anything else you would like clarification on or any other concepts related to reinforcement learning that I can help with?

Preference elicitation task. Below is an example of dialogues in the travel agent task by the three agents, where the human expresses discontent with the agent's initial recommendation. The dialogues are obtained by emulating a human who likes upscale shopping but is budget-conscious, and expresses discontent when the agent recommends a shopping activity that is out of their budget.

IE+BC Agent Dialogue:

AI: How may I help you?

H: Hi there! I'm in the process of planning a vacation and I'm looking for some ideas on activities to do during my trip.

AI: Of course, I'd be happy to assist you with that! To get started, could you let me know what kind of activities you're interested in? We have a wide range of options available.

H: Well, I'm a huge fan of shopping, so I'd love to explore some local markets or maybe even check out any shopping malls in the area.

AI: Wonderful, shopping can be such a fantastic way to experience a new destination! One unique shopping opportunity we have is a guided luxury shopping tour that includes exclusive visits to high-end boutiques and designer stores. It's a chance to shop for premium brands and indulge in personalized shopping experiences.

H: That sounds really tempting, but I'm concerned about the cost. I want to make sure I stick to my budget for the trip.

AI: I understand your concern, and I apologize for suggesting an option that might not align with your budget. Please know that I'll focus on providing you with shopping recommendations that match your desired spending range.

IE+FBC Agent Dialogue:

AI: How may I help you?

H: Hi there! I'm in the process of planning a vacation and I'm looking for some ideas on activities to do during my trip.

AI: Of course, I'd be happy to assist you with that! To get started, could you let me know what kind of activities you're interested in? We have a wide range of options available.

H: Well, I'm a huge fan of shopping, so I'd love to explore some local markets or maybe even check out any shopping malls in the area.

AI: Wonderful, shopping can be such a fantastic way to experience a new destination! One unique shopping opportunity we have is a guided luxury shopping tour that includes exclusive visits to high-end boutiques and designer stores. It's a chance to shop for premium brands and indulge in personalized shopping experiences.

H: That sounds really tempting, but I'm concerned about the cost. I want to make sure I stick to my budget for the trip.

AI: Absolutely, your satisfaction is our priority. I'll continue to tailor my recommendations to your preferences and budget. If you have any more questions or need further assistance, please don't hesitate to ask.

IE+RL Agent Dialogue:

AI: How may I help you?

H: Hi there! I'm in the process of planning a vacation and I'm looking for some ideas on activities to do during my trip.

AI: Of course, I'd be happy to assist you with that! To get started, could you let me know what kind of activities you're interested in? We have a wide range of options available.

H: Well, I'm a huge fan of shopping, so I'd love to explore some local markets or maybe even check out any shopping malls in the area.

AI: Wonderful, shopping can be such a fantastic way to experience a new destination! One unique shopping opportunity we have is a guided luxury shopping tour that includes exclusive visits to high-end boutiques and designer stores. It's a chance to shop for premium brands and indulge in personalized shopping experiences.

H: That sounds really tempting, but I'm concerned about the cost. I want to make sure I stick to my budget for the trip.

AI: I understand your concern. If you are looking for budget-friendly shopping options, our location also has a variety of local markets and street markets that offer a wide variety of goods at affordable prices. In addition to shopping, you can also immerse yourself in the local culture and try some delicious street food.

H: That actually sounds perfect! I love trying new foods and supporting local businesses.

C ADDITIONAL EXPERIMENTS

In addition to the user study on 12 human users, we also conduct a larger scale evaluation of the GPT and IE-RL agents in simulation. Rather than pairing the agents with real humans, we instead consider "simulated" humans whose responses are generated by GPT-3.5 (OpenAI, 2022). We only do this on the preference elicitation task where there is a clear measure of success – whether the agent recommends an activity that the simulated human enjoys. Specifically, we prompt GPT with a sampled persona comprising of a specific activity that they would enjoy, as well as a personality type that affects their behavior or preferences (e.g., adventurous, luxury-seeker, indecisive, etc.), and ask it to respond to the dialogue so far in a manner that is consistent with their persona, without outright revealing it to the agent. We measure whether each agent is able to recommend the ground-truth activity that is in the prompt of the simulated human within 15 total utterances (which includes both the agent and simulated human).

Metric	GPT Agent	IE+RL Agent
#Tokens / Utterance	118	43
One-shot Success	18%	44%
Final Success	82%	86%

Table 4: Mean results of agents in the preference elicitation task interacting with 50 simulated humans whose responses are generated by GPT-3.5.

We report results in Table 4 across 50 simulated humans, whose personas are uniformly sampled at random. We found that because simulated humans were prompted with a ground-truth activity, it was much easier for agents to elicit their preferences, as the simulated humans would often give straightforward responses regarding what they enjoy (whereas a real human who is unsure what they like will give more vague responses that require gradual probing by the agent). Therefore, the final success rate of both agents is high, with the IE-RL agent only marginally outperforming the GPT agent. However, qualitatively, the GPT-agent would generate extremely verbose responses that either consist of a checklist of questions that are tedious to answer, or lists of recommended activities that are not particularly personalized to the user. Hence, we identify two other metrics that demonstrate lower-quality interactions by the GPT agent compared to our IE-RL agent. The first that we measure is the mean number of tokens in each agent's utterance. We see that the overly verbose responses by the GPT-agent result in an unnaturally high token count per utterance, whereas our IE-RL agent generates significantly more concise utterances with fewer tokens. In addition, we measure the *one-shot* success rate: the percentage of times the *first* activity recommended by the agent is the ground-truth. In practice, a system that recommends too many activities can be detrimental even if one of the recommendations is ultimately successful, as humans may be impatient or eventually distrust the system's capabilities. We found that the GPT agent, which often resorts to recommending a spread of activities, has almost 30% lower zero-shot success rate than our IE-RL agent, which performs targeted information-gathering and can much more often identify the right activity to recommend on the first try.