A DATASETS

**WebShop.** WebShop (Yao et al., 2022) is a goal-oriented web navigation environment where an agent interacts with an e-commerce platform to fulfill natural language shopping instructions. A final reward in [0,1] is assigned based on how well the purchased item matches the instruction in terms of attributes and price. For training, we use the expert and GPT-4 ReAct-style trajectories provided by Song et al. (2024), retaining only those with final rewards above 0.7 to ensure quality.

# Example WebShop Trajectory

## **Instruction:**

I need a long clip-in hair extension which is natural looking, and price lower than 40.00 dollars.

#### **Conversation:**

- **Human:** You are web shopping. I will give you instructions about what to do. Every round I will give you an observation and a list of available actions...
- GPT: OK
- **Human:** WebShop [SEP] Instruction: [SEP] i need a long clip-in hair extension ...
- **GPT:** Thought: I need to find a long clip-in hair extension ... **Action:** search[long clip-in hair extension]
- Human: Observation: ... (product list with IDs and prices)
- **GPT:** Thought: The first option B09C337K8S fits the budget and is natural looking... **Action:** click[b09c337k8s]
- **Human:** Observation: ... (product details shown)
- **GPT:** Thought: The details confirm it is natural looking and affordable. **Action:** click[buy now]

Reward: 1.0

Figure 4: WebShop trajectory with instruction, dialogue turns, and reward.

**ScienceWorld.** ScienceWorld (Wang et al., 2022) is a text-based embodied environment for conducting elementary science experiments across ten domains, including thermodynamics and electricity. Each task consists of multiple optional subgoals, and the final reward in [0, 1] reflects the proportion of subgoals completed. Following Song et al. (2024), we exclude the two longest task types and subsample test instances with more than ten variations. We evaluate generalization using the original test set of unseen variations (e.g., training on boiling water and testing on boiling lead) and use the development set for seen scenarios. Expert trajectories are drawn from the official dataset, with additional GPT-4—generated trajectories filtered by final rewards, as in Song et al. (2024).

**InterCodeSQL.** InterCodeSQL (Yang et al., 2023) is an interactive SQL querying environment where an agent must translate natural language questions into executable SQL programs through multi-turn interaction. The environment provides a binary reward, which we normalize to [0,1], indicating whether the final query produces the correct result. We use the expert trajectories collected by **?** with GPT-4, retaining only those with reward 1.0. Since the original dataset does not contain intermediate reasoning, each trajectory is augmented with ReAct-style (Yao et al., 2023) thought annotations generated by GPT-4, as in **?**.

#### Example ScienceWorld Trajectory

# **Instruction:**

Your task is to determine if a metal pot is electrically conductive. The metal pot is located in the kitchen. If conductive, place it in the blue box; if nonconductive, place it in the orange box.

# **Conversation:**

- **Human:** You are a helpful assistant to do some scientific experiment in an environment. (Lists available rooms and actions.)
- · GPT: OK
- **Human:** Your task is to determine if metal pot is electrically conductive...
- **GPT:** *Thought: I should locate the pot, probably in the kitchen.* **Action:** look around
- Human: Observation: You are in the foundry...
- **GPT:** Thought: Wrong room, I need to go to the kitchen. **Action:** teleport to kitchen
- **Human:** Observation: You teleport to the kitchen.
- **GPT:** Thought: I should pick up the metal pot. **Action:** pick up metal pot
- **Human:** Observation: Pot moved to inventory.
- **GPT:** Thought: Now focus on the pot to test it. **Action:** focus on metal pot
- Human: Observation: You focus on the pot.
- **GPT:** *Thought: I need to set up a circuit in the workshop.* **Action:** teleport to workshop
- **Human:** Observation: You teleport to the workshop (battery, wires, bulbs available).
- **GPT:** Thought: Place the pot and build the circuit. **Action:** drop metal pot
- **Human:** Observation: Pot placed in workshop.
- **GPT:** Thought: Connect wires, battery, and pot in circuit. **Action:** connect battery anode to orange wire ... (multiple connections omitted)
- **Human:** Observation: Circuit components connected.
- **GPT:** *Thought: Wait and check if the bulb lights up.* **Action:** wait1
- **Human:** Observation: Blue light bulb is on.
- **GPT:** Thought: The pot is conductive. Place it in the blue box. **Action:** move metal pot to blue box

Figure 5: ScienceWorld trajectory with instruction, dialogue turns, and final action.

#### Example InterCodeSOL Trajectory

## **Instruction:**

What are the enrollment and primary conference for the university which was founded the earliest?

## **Conversation:**

- **Human:** You are a helpful assistant assigned with the task of problem-solving. To achieve this, you will interact with a MySQL Database system using SQL queries to answer a question. (Gives detailed formatting rules.)
- GPT: OK
- **Human:** What are the enrollment and primary conference for the university which was founded the earliest?
- **GPT:** Thought: I need to check what tables are available. **Action:** SHOW TABLES
- **Human:** Observation: [('basketball\_match',), ('university',)]
- **GPT:** Thought: The university table should have the info. I should check its schema. **Action:** DESC university
- Human: Observation: [('School', ...), ('Founded', ...), ('Enrollment', ...), ('Primary\_conference', ...)]
- **GPT:** Thought: The table has Founded, Enrollment, Primary\_conference.

  I can query the earliest Founded. Action: SELECT Enrollment,
  Primary\_conference FROM university ORDER BY Founded
  ASC LIMIT 1
- Human: Observation: [(19067.0, 'Colonial Athletic Association (D-I)')]
- **GPT:** Thought: The earliest university has enrollment 19067 and primary conference Colonial Athletic Association (D-I). This is the final answer. **Action:** submit

Figure 6: Illustrative InterCodeSQL trajectory with instruction, SQL reasoning steps, and final submission.

# **B** IMPLEMENTATION DETAILS

We adopt Llama-2-7B-Chat (Touvron et al., 2023) as the primary base model for constructing LLM agents, following prior work (Song et al., 2024). All models are optimized with AdamW (Loshchilov & Hutter, 2017), and we use the same hyperparameter settings across datasets for fair comparison. During the supervised fine-tuning (SFT) stage, we train for 3 epochs with a batch size of 48 and a learning rate of 1e-5, using 3% linear warmup followed by a cosine decay schedule. The base agent then explores each training instance once to collect failure trajectories. In the subsequent optimization phase, we train with a batch size of 32 and a learning rate of 1e-6 to 5e-7, with the DPO scaling factor  $\beta$  set to 0.1-0.5. The number of optimization epochs is fixed to 3, and the number of co-evolution iterations is set to 3 for WebShop and ScienceWorld and 5 for InterCodeSQL. All experiments are conducted on 8 NVIDIA H100 GPUs with 80GB.

## C ANALYSIS ON GENERATED TRAJECTORIES