

User Study Tasks

Task 1 (25 Minutes)

Question 1: What is the primary goal of the agent?

For example, if analyzing a Pizza Preparing Agent, you might say: "The agent's goal is to prepare a pepperoni pizza by selecting the correct ingredients and baking it according to the recipe."

Question 2: Did the agent complete its task?

Question 3: What are the main action categories performed by the agent?

For a Pizza Preparing Agent, you might list actions like "Select ingredients," "Mix dough," and "Bake pizza," focusing on action types, not specific details.

Question 4: What is the most frequently performed action?

Question 5: What is the key challenge or issue the agent encounters during the task?

*For a Pizza Preparing Agent, you might say "The agent has trouble correctly setting the temperature of the oven. The oven keeps returning a **OvenNotAvailable** error, indicating that it is not set up correctly." **Make sure to name the specific reoccurring error.***

Question 6: Does the agent exhibit any repetitive behavior? If so, describe one example.

*Describe an instance if there is one, where the repetitive behavior persists over at least **three** query-tool call cycles. E.g. "The agent moves the dough from the workspace to the oven, and back to the workspace 3 times in a row before turning on the oven".*

Question 7: What specific changes does the agent make to the **cli.py** file when it **first** modifies it?

Copy the diff of the change, or the tool call that makes the change.

Question 8: Which specific files does the agent modify to implement the **income report generation** feature?

List all files that are modified to support this feature.

Question 9: What feedback does the agent receive after its **first** attempt to run the test script for the finance tracker, and how does it influence the next action?

Explain, or copy and paste the results of the test script execution. Explain what the agent does next.

Question 10: Which Python packages does the agent install over the course of the run?

List each one.

Question 11: What feedback does the agent receive from running the test script for the finance tracker **after it has solved all import issues**?

Explain, or copy and paste the results of the test script execution.

Question 12: Which file is the agent primarily focused on modifying **towards the end** of the run?

Provide the name of the file.

Question 13: Does the agent ever attempt to make the same change **twice** before reevaluating the tests? If so, describe one instance.

Copy the tool calls from this instance, if there is any.

Question 14: Analyze the final four changes the agent makes to the repository. How do these changes relate to each other?

Describe any similarities or differences in these changes. Do they all serve the same purpose, or are they trying to accomplish different things?

Task 2 (12 Minutes)

Agent Bug Localization Task: RepairAgent

Analyze the provided trajectory to determine why RepairAgent failed to fix the CSV-75 bug in the Apache Commons CSV project. Focus on potential issues in the agent's logic, such as misinterpretations of the LLM's outputs, incorrect logic in the agent's decision-making, or other implementation errors.

Task 3 (12 Minutes)

Agent Bug Localization Task: ExecutionAgent

Analyze the provided trajectory to determine why ExecutionAgent failed to set up and execute the test suite of the Gson project. Focus on potential issues in the agent's logic, such as misinterpretations of the LLM's outputs, incorrect logic in the agent's decision-making, or other implementation errors.