

Android World Agent Evaluation Report

Generated: 2025-07-17 22:34:33

Total Episodes: 27
Total Steps: 85
Correct Steps: 51
Step Accuracy: 60.00%
Successful Episodes: 7
Episode Success Rate: 25.93%
Average Steps/Episode: 3.15

Episode: uninstall_slack
Goal: Uninstall the Slack app
Step 1: App=Settings, UI=['Apps', 'Search', 'Battery']
Predicted: CLICK("Apps")
Ground Truth: CLICK("Apps")
Correct: True
Step 2: App=Apps, UI=['Slack', 'Chrome', 'Maps']
Predicted: CLICK("Unknown")
Ground Truth: CLICK("Slack")
Correct: False
Step 3: App=App Info, UI=['Uninstall', 'Force Stop', 'Storage']
Predicted: CLICK("Uninstall")
Ground Truth: CLICK("Uninstall")
Correct: True
Summary: Total Steps: 3, Correct: 2, Incorrect: 1, Accuracy: 66.67%

Episode: take_photo
Goal: Open the camera and take a photo
Step 1: App=Home, UI=['Camera', 'Messages', 'Settings']
Predicted: CLICK("Camera")
Ground Truth: CLICK("Camera")
Correct: True
Step 2: App=Camera, UI=['Capture', 'Gallery', 'Settings']
Predicted: CLICK("Capture")
Ground Truth: CLICK("Capture")
Correct: True
Summary: Total Steps: 2, Correct: 2, Incorrect: 0, Accuracy: 100.00%

Episode: send_message
Goal: Send a message to John
Step 1: App=Messages, UI=['New Message', 'Search', 'John']
Predicted: CLICK("John")
Ground Truth: CLICK("John")
Correct: True
Step 2: App=Chat, UI=['Text Input', 'Send', 'Attach']
Predicted: TYPE("Text Input", "Hello John")
Ground Truth: TYPE("Text Input", "Hello John!")
Correct: False
Step 3: App=Chat, UI=['Text Input', 'Send', 'Attach']
Predicted: TYPE("Text Input", "Hello John")
Ground Truth: CLICK("Send")
Correct: False

Summary: Total Steps: 3, Correct: 1, Incorrect: 2, Accuracy: 33.33%

Episode: search_wifi

Goal: Search for Wi-Fi networks

Step 1: App=Settings, UI=['Network & Internet', 'Bluetooth', 'Display']

Predicted: CLICK("Network & Internet")

Ground Truth: CLICK("Network & Internet")

Correct: True

Step 2: App=Network & Internet, UI=['Wi-Fi', 'Mobile', 'Data Usage']

Predicted: CLICK("Wi-Fi")

Ground Truth: CLICK("Wi-Fi")

Correct: True

Step 3: App=Wi-Fi, UI=['Available Networks', 'Add Network', 'Scan']

Predicted: CLICK("Available Networks")

Ground Truth: CLICK("Scan")

Correct: False

Summary: Total Steps: 3, Correct: 2, Incorrect: 1, Accuracy: 66.67%

Episode: set_alarm

Goal: Set an alarm for 7 AM

Step 1: App=Home, UI=['Clock', 'Calendar', 'Weather']

Predicted: CLICK("Clock")

Ground Truth: CLICK("Clock")

Correct: True

Step 2: App=Clock, UI=['Alarm', 'Timer', 'Stopwatch']

Predicted: CLICK("Alarm")

Ground Truth: CLICK("Alarm")

Correct: True

Step 3: App=Alarm, UI=['Add Alarm', 'Edit', 'Delete']

Predicted: CLICK("Add Alarm")

Ground Truth: CLICK("Add Alarm")

Correct: True

Step 4: App=Add Alarm, UI=['Time Picker', 'Save', 'Cancel']

Predicted: CLICK("Time Picker")

Ground Truth: TYPE("Time Picker", "07:00")

Correct: False

Summary: Total Steps: 4, Correct: 3, Incorrect: 1, Accuracy: 75.00%

Episode: check_battery

Goal: Check battery usage

Step 1: App=Settings, UI=['Battery', 'Display', 'Sound']

Predicted: CLICK("Unknown")

Ground Truth: CLICK("Battery")

Correct: False

Step 2: App=Battery, UI=['Battery Usage', 'Battery Saver', 'Adaptive Battery']

Predicted: CLICK("Unknown")

Ground Truth: CLICK("Battery Usage")

Correct: False

Summary: Total Steps: 2, Correct: 0, Incorrect: 2, Accuracy: 0.00%

Episode: turn_on_bluetooth

Goal: Turn on Bluetooth

Step 1: App=Settings, UI=['Network & Internet', 'Bluetooth', 'Display']

Predicted: CLICK("Bluetooth")
Ground Truth: CLICK("Bluetooth")
Correct: True
Step 2: App=Bluetooth, UI=['On/Off Switch', 'Pair New Device', 'Saved Devices']
Predicted: CLICK("Unknown")
Ground Truth: CLICK("On/Off Switch")
Correct: False
Summary: Total Steps: 2, Correct: 1, Incorrect: 1, Accuracy: 50.00%

Episode: search_weather_chrome
Goal: Open Chrome and search for 'weather'
Step 1: App=Home, UI=['Chrome', 'Maps', 'Camera']
Predicted: CLICK("Chrome")
Ground Truth: CLICK("Chrome")
Correct: True
Step 2: App=Chrome, UI=['Search Bar', 'Tabs', 'Menu']
Predicted: TYPE("Search Bar", "weather")
Ground Truth: TYPE("Search Bar", "weather")
Correct: True
Summary: Total Steps: 2, Correct: 2, Incorrect: 0, Accuracy: 100.00%

Episode: add_contact_alice
Goal: Add a new contact named Alice
Step 1: App=Home, UI=['Contacts', 'Phone', 'Messages']
Predicted: CLICK("Contacts")
Ground Truth: CLICK("Contacts")
Correct: True
Step 2: App=Contacts, UI=['Add Contact', 'Search', 'Favorites']
Predicted: CLICK("Add Contact")
Ground Truth: CLICK("Add Contact")
Correct: True
Step 3: App=Add Contact, UI=['Name Field', 'Phone Field', 'Save']
Predicted: TYPE("Name Field", "Alice")
Ground Truth: TYPE("Name Field", "Alice")
Correct: True
Summary: Total Steps: 3, Correct: 3, Incorrect: 0, Accuracy: 100.00%

Episode: mute_phone
Goal: Mute the phone
Step 1: App=Settings, UI=['Sound', 'Display', 'Notifications']
Predicted: CLICK("Unknown")
Ground Truth: CLICK("Sound")
Correct: False
Step 2: App=Sound, UI=['Volume', 'Vibrate', 'Mute']
Predicted: CLICK("Mute")
Ground Truth: CLICK("Mute")
Correct: True
Summary: Total Steps: 2, Correct: 1, Incorrect: 1, Accuracy: 50.00%

Episode: directions_airport
Goal: Open Maps and get directions to the airport
Step 1: App=Home, UI=['Maps', 'Chrome', 'Camera']
Predicted: CLICK("Maps")

Ground Truth: CLICK("Maps")
Correct: True
Step 2: App=Maps, UI=['Search', 'Directions', 'Menu']
Predicted: CLICK("Search")
Ground Truth: CLICK("Directions")
Correct: False
Step 3: App=Maps, UI=['Airport', 'Home', 'Work']
Predicted: CLICK("Airport")
Ground Truth: TYPE("Search", "Airport")
Correct: False
Summary: Total Steps: 3, Correct: 1, Incorrect: 2, Accuracy: 33.33%

Episode: change_wallpaper
Goal: Change wallpaper to a downloaded image
Step 1: App=Home, UI=['Settings', 'Gallery', 'Files']
Predicted: CLICK("Unknown")
Ground Truth: CLICK("Settings")
Correct: False
Step 2: App=Settings, UI=['Display', 'Wallpaper', 'Themes']
Predicted: CLICK("Unknown")
Ground Truth: CLICK("Wallpaper")
Correct: False
Step 3: App=Wallpaper, UI=['Choose Image', 'Default', 'Gallery']
Predicted: CLICK("Choose Image")
Ground Truth: CLICK("Choose Image")
Correct: True
Step 4: App=Gallery, UI=['Downloaded', 'Camera', 'Screenshots']
Predicted: CLICK("Downloaded")
Ground Truth: CLICK("Gallery")
Correct: False
Step 5: App=Downloaded, UI=['beach.jpg', 'mountain.jpg', 'city.jpg']
Predicted: CLICK("beach.jpg")
Ground Truth: CLICK("Downloaded")
Correct: False
Step 6: App=beach.jpg, UI=['Set as Wallpaper', 'Share', 'Delete']
Predicted: CLICK("Set as Wallpaper")
Ground Truth: CLICK("beach.jpg")
Correct: False
Summary: Total Steps: 6, Correct: 1, Incorrect: 5, Accuracy: 16.67%

Episode: share_photo_email
Goal: Share a photo via email
Step 1: App=Gallery, UI=['Camera', 'Downloads', 'Vacation']
Predicted: CLICK("Unknown")
Ground Truth: CLICK("Vacation")
Correct: False
Step 2: App=Vacation, UI=['photo1.jpg', 'photo2.jpg', 'photo3.jpg']
Predicted: CLICK("Unknown")
Ground Truth: CLICK("photo2.jpg")
Correct: False
Step 3: App=photo2.jpg, UI=['Share', 'Delete', 'Edit']
Predicted: CLICK("Share")
Ground Truth: CLICK("Share")
Correct: True
Step 4: App=Share, UI=['Email', 'Messages', 'Drive']
Predicted: CLICK("Unknown")

Ground Truth: CLICK("Email")
Correct: False
Step 5: App=Email, UI=['To Field', 'Subject Field', 'Send']
Predicted: CLICK("Unknown")
Ground Truth: TYPE("To Field", "friend@example.com")
Correct: False
Summary: Total Steps: 5, Correct: 1, Incorrect: 4, Accuracy: 20.00%

Episode: enable_dark_mode
Goal: Enable dark mode
Step 1: App=Settings, UI=['Display', 'Sound', 'Notifications']
Predicted: CLICK("Unknown")
Ground Truth: CLICK("Display")
Correct: False
Step 2: App=Display, UI=['Brightness', 'Dark Mode', 'Font Size']
Predicted: CLICK("Dark Mode")
Ground Truth: CLICK("Dark Mode")
Correct: True
Summary: Total Steps: 2, Correct: 1, Incorrect: 1, Accuracy: 50.00%

Episode: update_software
Goal: Update the system software
Step 1: App=Settings, UI=['System', 'About Phone', 'Software Update']
Predicted: CLICK("Software Update")
Ground Truth: CLICK("System")
Correct: False
Step 2: App=System, UI=['Software Update', 'Reset', 'Backup']
Predicted: CLICK("Unknown")
Ground Truth: CLICK("Software Update")
Correct: False
Step 3: App=Software Update, UI=['Check for Updates', 'Install Now', 'Schedule']
Predicted: CLICK("Check for Updates")
Ground Truth: CLICK("Check for Updates")
Correct: True
Summary: Total Steps: 3, Correct: 1, Incorrect: 2, Accuracy: 33.33%

Episode: connect_bluetooth_speaker
Goal: Connect to a Bluetooth speaker named 'JBL Flip'
Step 1: App=Settings, UI=['Bluetooth', 'Wi-Fi', 'Display']
Predicted: CLICK("Bluetooth")
Ground Truth: CLICK("Bluetooth")
Correct: True
Step 2: App=Bluetooth, UI=['On/Off Switch', 'Pair New Device', 'Saved Devices']
Predicted: CLICK("Pair New Device")
Ground Truth: CLICK("Pair New Device")
Correct: True
Step 3: App=Pair New Device, UI=['JBL Flip', 'Sony WH', 'Bose QC']
Predicted: CLICK("JBL Flip")
Ground Truth: CLICK("JBL Flip")
Correct: True
Step 4: App=JBL Flip, UI=['Connect', 'Forget', 'Rename']
Predicted: CLICK("Connect")
Ground Truth: CLICK("Connect")
Correct: True

Summary: Total Steps: 4, Correct: 4, Incorrect: 0, Accuracy: 100.00%

Episode: block_spam_number

Goal: Block a spam number

Step 1: App=Phone, UI=['Recents', 'Contacts', 'Spam']

Predicted: CLICK("Unknown")

Ground Truth: CLICK("Spam")

Correct: False

Step 2: App=Spam, UI=['1234567890', '9876543210', 'Block']

Predicted: CLICK("Block")

Ground Truth: CLICK("1234567890")

Correct: False

Step 3: App=1234567890, UI=['Block', 'Call', 'Message']

Predicted: CLICK("Block")

Ground Truth: CLICK("Block")

Correct: True

Summary: Total Steps: 3, Correct: 1, Incorrect: 2, Accuracy: 33.33%

Episode: turn_on_airplane_mode

Goal: Turn on airplane mode

Step 1: App=Settings, UI=['Network & Internet', 'Connections', 'Airplane Mode']

Predicted: CLICK("Airplane Mode")

Ground Truth: CLICK("Network & Internet")

Correct: False

Step 2: App=Network & Internet, UI=['Wi-Fi', 'Mobile', 'Airplane Mode']

Predicted: CLICK("Airplane Mode")

Ground Truth: CLICK("Airplane Mode")

Correct: True

Summary: Total Steps: 2, Correct: 1, Incorrect: 1, Accuracy: 50.00%

Episode: delete_all_alarms

Goal: Delete all alarms

Step 1: App=Clock, UI=['Alarm', 'Timer', 'Stopwatch']

Predicted: CLICK("Unknown")

Ground Truth: CLICK("Alarm")

Correct: False

Step 2: App=Alarm, UI=['Edit', 'Delete All', 'Add Alarm']

Predicted: CLICK("Unknown")

Ground Truth: CLICK("Delete All")

Correct: False

Summary: Total Steps: 2, Correct: 0, Incorrect: 2, Accuracy: 0.00%

Episode: reply_email_bob

Goal: Reply to an email from Bob

Step 1: App=Email, UI=['Inbox', 'Sent', 'Bob']

Predicted: CLICK("Bob")

Ground Truth: CLICK("Bob")

Correct: True

Step 2: App=Bob, UI=['Reply', 'Forward', 'Delete']

Predicted: CLICK("Reply")

Ground Truth: CLICK("Reply")

Correct: True

Step 3: App=Reply, UI=['Text Input', 'Send', 'Attach']

Predicted: TYPE("Text Input", "Reply to Bob\
Ground Truth: TYPE("Text Input", "Thanks Bob!")
Correct: False
Summary: Total Steps: 3, Correct: 2, Incorrect: 1, Accuracy: 66.67%

Episode: change_language_spanish
Goal: Change language to Spanish
Step 1: App=Settings, UI=['System', 'Languages & Input', 'About Phone']
Predicted: CLICK("Languages & Input")
Ground Truth: CLICK("Languages & Input")
Correct: True
Step 2: App=Languages & Input, UI=['Languages', 'Keyboard', 'Spell Checker']
Predicted: CLICK("Languages")
Ground Truth: CLICK("Languages")
Correct: True
Step 3: App=Languages, UI=['Add Language', 'Spanish', 'French']
Predicted: CLICK("Spanish")
Ground Truth: CLICK("Spanish")
Correct: True
Summary: Total Steps: 3, Correct: 3, Incorrect: 0, Accuracy: 100.00%

Episode: RecipeAddMultipleRecipesFromImage
Goal: Add the recipes from recipes.jpg in Simple Gallery Pro to the Broccoli recipe app.
Step 1: App=Simple Gallery Pro, UI=['recipes.jpg', 'Camera', 'Albums']
Predicted: CLICK("recipes.jpg")
Ground Truth: CLICK("recipes.jpg")
Correct: True
Step 2: App=recipes.jpg, UI=['Share', 'Edit', 'Delete']
Predicted: CLICK("Share")
Ground Truth: CLICK("Share")
Correct: True
Step 3: App=Share, UI=['Broccoli', 'Email', 'Drive']
Predicted: CLICK("Broccoli")
Ground Truth: CLICK("Broccoli")
Correct: True
Step 4: App=Broccoli, UI=['Add Recipe', 'Import', 'Settings']
Predicted: CLICK("Import")
Ground Truth: CLICK("Import")
Correct: True
Summary: Total Steps: 4, Correct: 4, Incorrect: 0, Accuracy: 100.00%

Episode: RecipeAddMultipleRecipesFromMarkor
Goal: Add the recipes from recipes.txt in Markor to the Broccoli recipe app.
Step 1: App=Markor, UI=['recipes.txt', 'Notes', 'Tasks']
Predicted: CLICK("recipes.txt")
Ground Truth: CLICK("recipes.txt")
Correct: True
Step 2: App=recipes.txt, UI=['Share', 'Edit', 'Delete']
Predicted: CLICK("Share")
Ground Truth: CLICK("Share")
Correct: True
Step 3: App=Share, UI=['Broccoli', 'Email', 'Drive']
Predicted: CLICK("Broccoli")
Ground Truth: CLICK("Broccoli")

Correct: True
Step 4: App=Broccoli, UI=['Add Recipe', 'Import', 'Settings']
Predicted: CLICK("Import")
Ground Truth: CLICK("Import")
Correct: True
Summary: Total Steps: 4, Correct: 4, Incorrect: 0, Accuracy: 100.00%

Episode: RecipeAddMultipleRecipesFromMarkor2
Goal: Add the recipes from recipes.txt in Markor that take {prep_time} to prepare into the Broccoli recipe app.
Step 1: App=Markor, UI=['recipes.txt', 'Notes', 'Tasks']
Predicted: CLICK("recipes.txt")
Ground Truth: CLICK("recipes.txt")
Correct: True
Step 2: App=recipes.txt, UI=['Filter', 'Share', 'Edit']
Predicted: CLICK("Filter")
Ground Truth: CLICK("Filter")
Correct: True
Step 3: App=Filter, UI=['Prep Time', 'Ingredient', 'Cuisine']
Predicted: CLICK("Prep Time")
Ground Truth: CLICK("Prep Time")
Correct: True
Step 4: App=Share, UI=['Broccoli', 'Email', 'Drive']
Predicted: CLICK("Broccoli")
Ground Truth: CLICK("Share")
Correct: False
Step 5: App=Broccoli, UI=['Add Recipe', 'Import', 'Settings']
Predicted: CLICK("Add Recipe")
Ground Truth: CLICK("Broccoli")
Correct: False
Summary: Total Steps: 5, Correct: 3, Incorrect: 2, Accuracy: 60.00%

Episode: RecipeAddSingleRecipe
Goal: Add the following recipes into the Broccoli app: {recipe}
Step 1: App=Broccoli, UI=['Add Recipe', 'Import', 'Settings']
Predicted: CLICK("Add Recipe")
Ground Truth: CLICK("Add Recipe")
Correct: True
Step 2: App=Add Recipe, UI=['Recipe Name', 'Ingredients', 'Save']
Predicted: CLICK("Recipe Name")
Ground Truth: TYPE("Recipe Name", "{recipe}")
Correct: False
Summary: Total Steps: 2, Correct: 1, Incorrect: 1, Accuracy: 50.00%

Episode: RecipeDeleteDuplicateRecipes
Goal: Delete all but one of any recipes in the Broccoli app that are exact duplicates, ensuring at least one instance of each unique recipe remains
Step 1: App=Broccoli, UI=['Recipes', 'Search', 'Delete']
Predicted: CLICK("Recipes")
Ground Truth: CLICK("Recipes")
Correct: True
Step 2: App=Recipes, UI=['Duplicate1', 'Duplicate2', 'Unique']
Predicted: CLICK("Duplicate1")
Ground Truth: CLICK("Duplicate1")
Correct: True

Step 3: App=Duplicate1, UI=['Delete', 'Edit', 'Back']
Predicted: CLICK("Delete")
Ground Truth: CLICK("Delete")
Correct: True
Step 4: App=Duplicate2, UI=['Delete', 'Edit', 'Back']
Predicted: CLICK("Delete")
Ground Truth: CLICK("Duplicate2")
Correct: False
Summary: Total Steps: 4, Correct: 3, Incorrect: 1, Accuracy: 75.00%

Episode: RecipeDeleteDuplicateRecipes2

Goal: Delete all but one of any recipes in the Broccoli app that are exact duplicates, ensuring at least one instance of each unique recipe remains (medium)

Step 1: App=Broccoli, UI=['Recipes', 'Search', 'Delete']
Predicted: CLICK("Recipes")
Ground Truth: CLICK("Recipes")
Correct: True
Step 2: App=Recipes, UI=['Duplicate1', 'Duplicate2', 'Unique']
Predicted: CLICK("Duplicate1")
Ground Truth: CLICK("Duplicate1")
Correct: True
Step 3: App=Duplicate1, UI=['Delete', 'Edit', 'Back']
Predicted: CLICK("Delete")
Ground Truth: CLICK("Delete")
Correct: True
Step 4: App=Duplicate2, UI=['Delete', 'Edit', 'Back']
Predicted: CLICK("Delete")
Ground Truth: CLICK("Duplicate2")
Correct: False
Summary: Total Steps: 4, Correct: 3, Incorrect: 1, Accuracy: 75.00%

Full Markdown Report

Android World Agent Evaluation Report

Generated: 2025-07-17 22:25:37

Summary Statistics

- **Total Episodes**: 27
- **Total Steps**: 85
- **Correct Steps**: 51
- **Step Accuracy**: 60.00%
- **Successful Episodes**: 7
- **Episode Success Rate**: 25.93%
- **Average Steps per Episode**: 3.1

Task-Specific Performance

- **uninstall_slack**: 66.67%
- **take_photo**: 100.00%
- **send_message**: 33.33%
- **search_wifi**: 66.67%
- **set_alarm**: 75.00%
- **check_battery**: 0.00%
- **turn_on_bluetooth**: 50.00%
- **search_weather_chrome**: 100.00%
- **add_contact_alice**: 100.00%
- **mute_phone**: 50.00%
- **directions_airport**: 33.33%
- **change_wallpaper**: 16.67%
- **share_photo_email**: 20.00%
- **enable_dark_mode**: 50.00%
- **update_software**: 33.33%
- **connect_bluetooth_speaker**: 100.00%
- **block_spam_number**: 33.33%
- **turn_on_airplane_mode**: 50.00%
- **delete_all_alarms**: 0.00%
- **reply_email_bob**: 66.67%
- **change_language_spanish**: 100.00%
- **RecipeAddMultipleRecipesFromImage**: 100.00%
- **RecipeAddMultipleRecipesFromMarkor**: 100.00%
- **RecipeAddMultipleRecipesFromMarkor2**: 60.00%
- **RecipeAddSingleRecipe**: 50.00%
- **RecipeDeleteDuplicateRecipes**: 75.00%
- **RecipeDeleteDuplicateRecipes2**: 75.00%

App-Specific Performance

- **Settings**: 45.45%
- **Apps**: 0.00%
- **App Info**: 100.00%
- **Home**: 83.33%
- **Camera**: 100.00%
- **Messages**: 100.00%
- **Chat**: 0.00%
- **Network & Internet**: 100.00%
- **Wi-Fi**: 0.00%
- **Clock**: 50.00%
- **Alarm**: 50.00%
- **Add Alarm**: 0.00%
- **Battery**: 0.00%
- **Bluetooth**: 50.00%
- **Chrome**: 100.00%
- **Contacts**: 100.00%
- **Add Contact**: 100.00%
- **Sound**: 100.00%

- **Maps**: 0.00%
- **Wallpaper**: 100.00%
- **Gallery**: 0.00%
- **Downloaded**: 0.00%
- **beach.jpg**: 0.00%
- **Vacation**: 0.00%
- **photo2.jpg**: 100.00%
- **Share**: 50.00%
- **Email**: 50.00%
- **Display**: 100.00%
- **System**: 0.00%
- **Software Update**: 100.00%
- **Pair New Device**: 100.00%
- **JBL Flip**: 100.00%
- **Phone**: 0.00%
- **Spam**: 0.00%
- **1234567890**: 100.00%
- **Bob**: 100.00%
- **Reply**: 0.00%
- **Languages & Input**: 100.00%
- **Languages**: 100.00%
- **Simple Gallery Pro**: 100.00%
- **recipes.jpg**: 100.00%
- **Broccoli**: 83.33%
- **Markor**: 100.00%
- **recipes.txt**: 100.00%
- **Filter**: 100.00%
- **Add Recipe**: 0.00%
- **Recipes**: 100.00%
- **Duplicate1**: 100.00%
- **Duplicate2**: 0.00%

Error Analysis

Most Common Error Patterns:

- **Wrong Element Clicked**: 27 occurrences
- **Wrong Action Type (Click vs Type)**: 4 occurrences
- **Wrong Text Typed**: 2 occurrences
- **Wrong Action Type (Type vs Click)**: 1 occurrences

Sample Errors

Error 1:

- Episode: uninstall_slack
- Goal: Uninstall the Slack app
- App: Apps
- Predicted: CLICK("Unknown")
- Ground Truth: CLICK("Slack")

Error 2:

- Episode: send_message
- Goal: Send a message to John
- App: Chat
- Predicted: TYPE("Text Input", "Hello John")
- Ground Truth: TYPE("Text Input", "Hello John!")

Error 3:

- Episode: send_message
- Goal: Send a message to John
- App: Chat
- Predicted: TYPE("Text Input", "Hello John")
- Ground Truth: CLICK("Send")