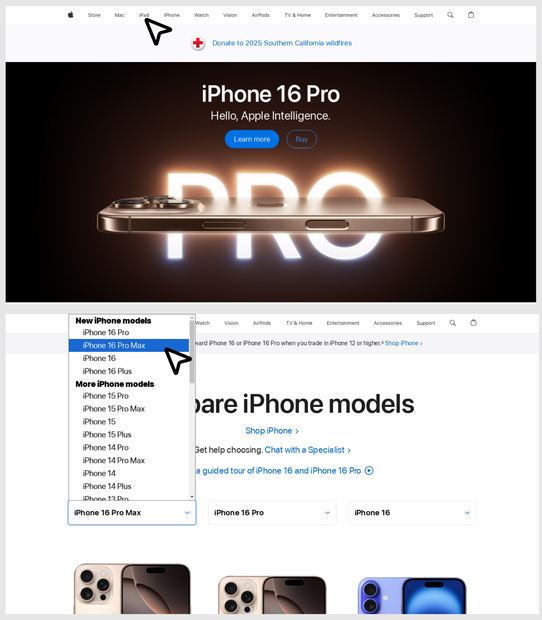
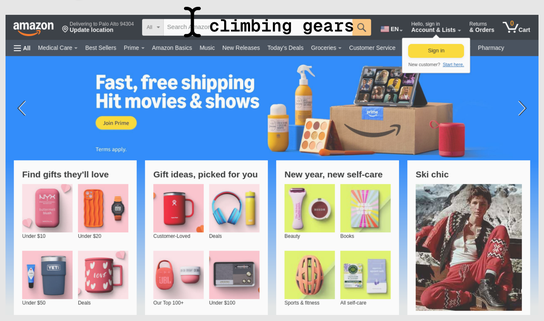
Collect human feedback

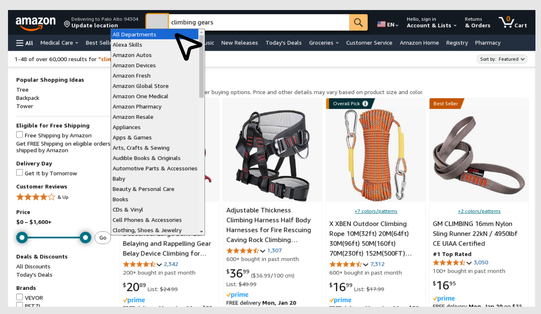
Compare the prices and chips for the **iPhone 14 Pro** and **iPhone 15 Pro** models ...



The agent did not interact with the drop-down correctly to choose iPhone 14 pro or 15 pro from choices.

Find climbing gears and sort the results by **price high to low**. Answer the first 3 results





The agent used the right search query, but failed to use “Sort by” drop-down menu to sort by price.

Other trajectory-feedback pairs.

Induce agent evaluation metrics

AGENT BEHAVIOR

Agent clicked on drop-down menu for iPhone models; Agent selected iPhone 16 Pro Max.

HUMAN FEEDBACK ASPECT

The agent did not select the iPhone 14 or 16 in the task description.

IS IT A POSITIVE ASPECT?

No.

AGENT BEHAVIOR

Agent clicked on drop-down menu for product categories; Agent chose All Department.

HUMAN FEEDBACK ASPECT

The agent did not use the correct drop-down menu for price sorting.

IS IT A POSITIVE ASPECT?

No.

AGENT BEHAVIOR

Agent input “climbing gears” in the search bar; Agent clicked on the search button.

HUMAN FEEDBACK ASPECT

The agent used the right search query.

IS IT A POSITIVE ASPECT?

Yes.

AGENT BEHAVIOR

Agent input “innovative and widely recognized open-source NLP models for automatic speech recognition released in the past month” in the search bar; Agent clicked on the search button.

HUMAN FEEDBACK ASPECT

The agent used a very specific search query that is not supported by Huggingface.

IS IT A POSITIVE ASPECT?

No.

Other grounded feedback aspects.

Element Interaction Accuracy

METRIC DESCRIPTION

This metric evaluates if the agent interacts with the correct UI elements for each task. Good behaviors show **accurate targeting of links, buttons, and textboxes**, ensuring that actions have the intended effect, while bad behaviors involve misidentification or misuse of elements that lead to errors.

GOOD BEHAVIORS

1. Agent correctly uses the search bar to search for news related to Brexit.
2. Agent uses the filter feature to check for audio datasets.
3. ...

BAD BEHAVIORS

1. Agent did not correctly use drop-down menu to find the correct iPhone model.
2. Agent did not correctly use the “Sort by” drop-down menu to sort by price.
3. ...

Query and Search Strategy

METRIC DESCRIPTION

This metric assesses the agent's ability to **craft and refine search queries**... Good behaviors ... clear, targeted queries that return relevant results, whereas poor behaviors ... fail to align with the task requirements.

GOOD BEHAVIORS

1. Formulated precise queries such as 'derivative of x^2 when x=5.6', 'Introduction to Psychology' ...
2. Refined queries when necessary (e.g., from 'travel guide books' to 'travel guide books for Japan 2024')
3. ...

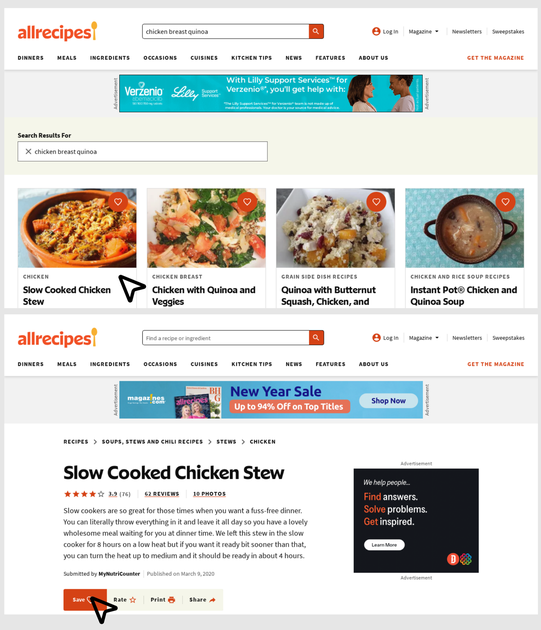
BAD BEHAVIORS

1. ...

Other induced metrics.

Evaluate agents

Could you find a recipe with **Chicken** and **Quinoa** and save it?



Evaluate metrics

Unseen human feedback

LLM-as-a-Judge Results

POSITIVE TRAITS

Search query is correct.
Using the correct buttons and search boxes.

NEGATIVE TRAITS

Information is **misunderstood**. The agent should choose “Chicken with Quinoa and Veggies” Instead.
The agent **didn’t reach the desired goal**. The recipe didn’t contain quinoa.

NOT APPLICABLE METRICS

Meta Evaluation

The recipe found by the agent contained neither chicken breast or quinoa.

Covered by

The agent efficiently found a recipe.

Not covered since is judged as N/A.

Coverage is aggregated across all trajectories.