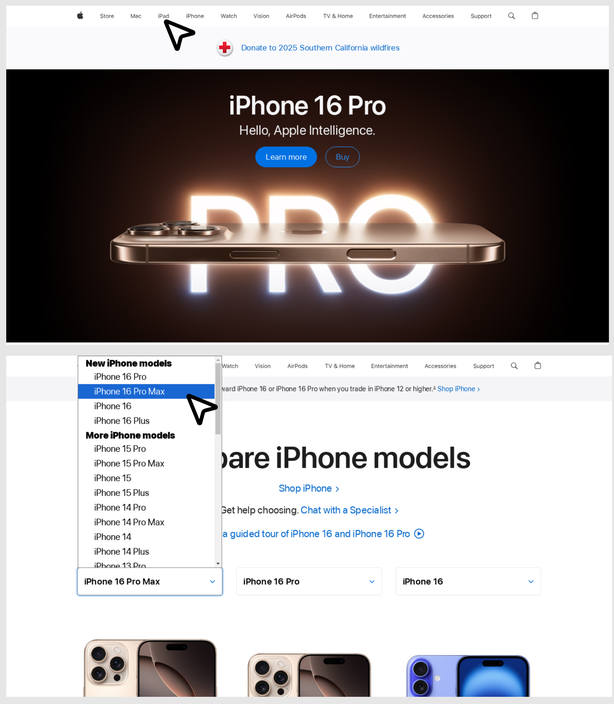
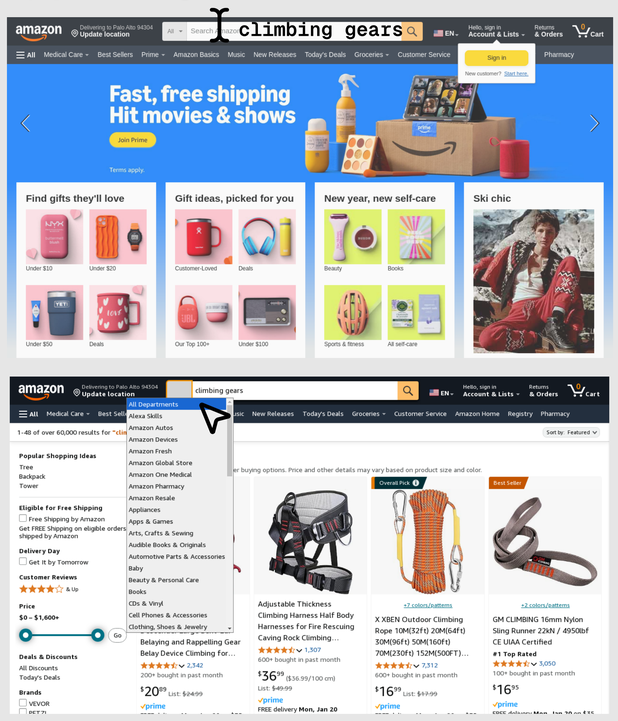
Collect human feedback

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



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...



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 — ASPECT

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.

IS IT A POSITIVE ASPECT?

⊗ **No.**

AGENT BEHAVIOR — ASPECT

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

HUMAN FEEDBACK ASPECT

The agent **did not use the correct** ... for price sorting.

IS IT A POSITIVE ASPECT?

⊗ **No.**

AGENT BEHAVIOR — ASPECT

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

HUMAN FEEDBACK ASPECT

The agent used the right query.

IS IT A POSITIVE ASPECT?

✓ **Yes.**

AGENT BEHAVIOR — ASPECT

Agent input “innovative and ... open-source NLP models ... released in the past month” in the search bar...

HUMAN FEEDBACK ASPECT

The agent used a specific query that is not supported by HF.

IS IT A POSITIVE ASPECT?

⊗ No.

Other feedback aspects.

METRIC

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**, ... 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. ...

METRIC

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 ...

GOOD BEHAVIORS

1. Formulated precise queries such as 'derivative of x^2 when x=5.6', 'Introduction to Psychology' ...
2. Refined queries ...
3. ...

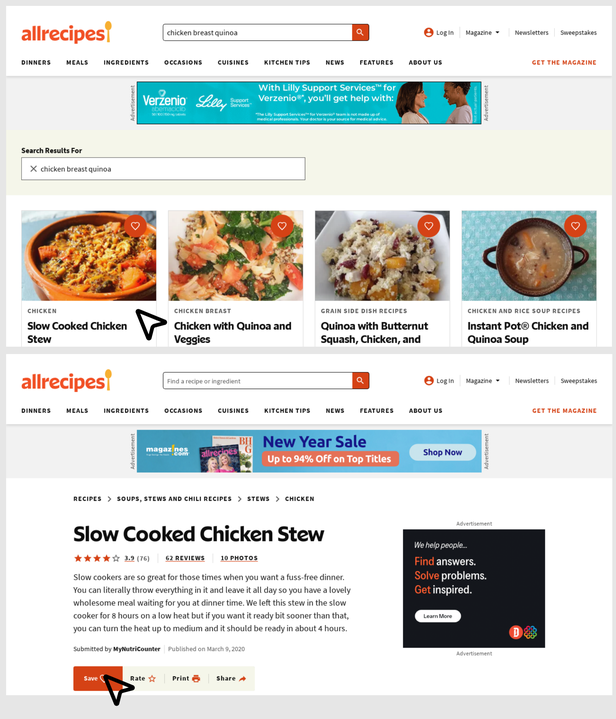
BAD BEHAVIORS

1. ...

Other induced metrics.

Evaluate agents w/ induced metrics

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



Judgement with LLMs

POSITIVE TRAITS

- Search query is correct.
- Using the correct buttons...

NEGATIVE TRAITS

- should choose “Chicken w/ ...
- recipe not desired w/o quinoa

NOT APPLICABLE METRICS

Evaluate induced metrics

Unseen human feedback

Meta Evaluation

recipe by the agent contained no chicken breast or quinoa.

Covered by 💯

The agent efficiently found a recipe.

Not covered since is judged as N/A.

Aggregated Feedback Coverage: **89%**