TurkingBench: A Challenge Benchmark for Web Agents

https://turkingbench.github.io

Kevin Xu, Yeganeh Kordi, Tanay Nayak, Adi Asija, Yizhong Wang,
Kate Sanders, Adam Byerly, Jingyu Zhang, Benjamin Van Durme, **Daniel Khashabi**

BROWN

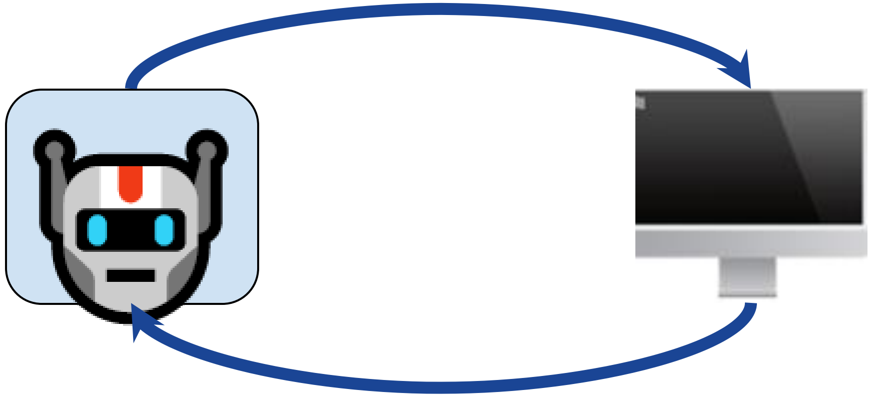
JOHNS HOPKINS UNIVERSITY

W UNIVERSITY of WASHINGTON

Agentic Models of Web

- Goal-oriented interaction with the internet.

- How do you benchmark models?

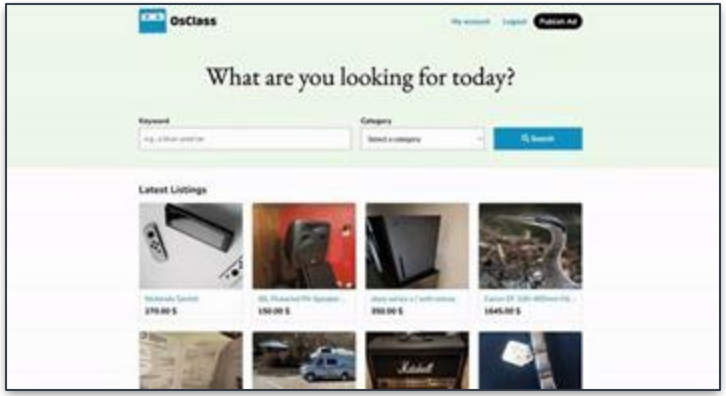


Existing benchmarks for web agents

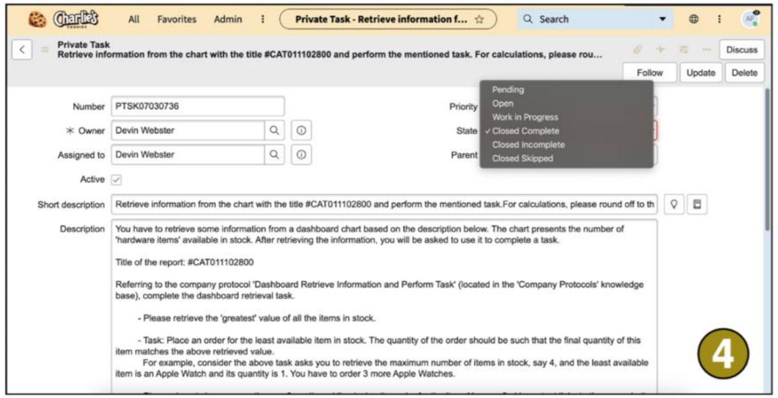
- Few notable ones:

Each dataset capture a narrow slice of web distribution.

Our work: introducing a new domain.



WebArena/VisualWebArena



Private Task
Retrieve information from the chart with the title #CAT011102800 and perform the mentioned task. For calculations, please rou...

Number: PTSK07030736

Priority

* Owner: Devin Webster

State: ✓ Closed Complete

Assigned to: Devin Webster

Parent

Active ☑

Short description: Retrieve information from the chart with the title #CAT011102800 and perform the mentioned task.For calculations, please round off to th

Description: You have to retrieve some information from a dashboard chart based on the description below. The chart presents the number of 'hardware items' available in stock. After retrieving the information, you will be asked to use it to complete a task.

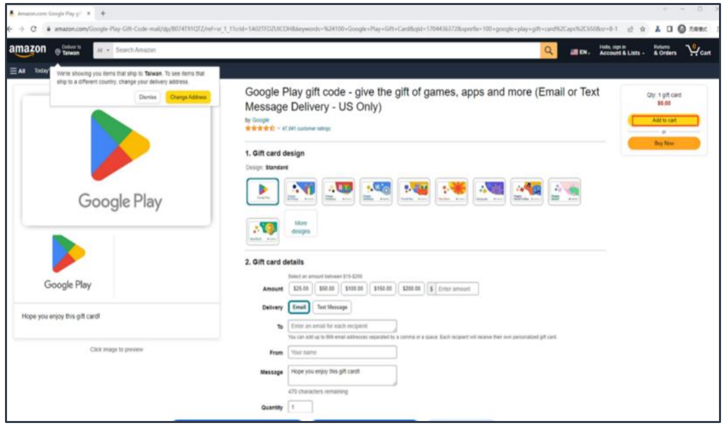
Title of the report: #CAT011102800

Referring to the company protocol 'Dashboard Retrieve Information and Perform Task' (located in the 'Company Protocols' knowledge base), complete the dashboard retrieval task.

- Please retrieve the 'greatest' value of all the items in stock.

- Task: Place an order for the least available item in stock. The quantity of the order should be such that the final quantity of this item matches the above retrieved value.
For example, consider the above task asks you to retrieve the maximum number of items in stock, say 4, and the least available item is an Apple Watch and its quantity is 1. You have to order 3 more Apple Watches.

WorkArena



We're showing you items that ship to Taiwan. To see items that ship to a different country, change your delivery address.

Google Play gift code - give the gift of games, apps and more (Email or Text Message Delivery - US Only)

by Google

1. Gift card design

Design: Standard

2. Gift card details

Select an amount between $15-$200

Amount: $25.00 | $50.00 | $100.00 | $150.00 | $200.00 | $ Enter amount

Delivery: Email | Text Message

To: Enter an email for each recipient

From: Your name

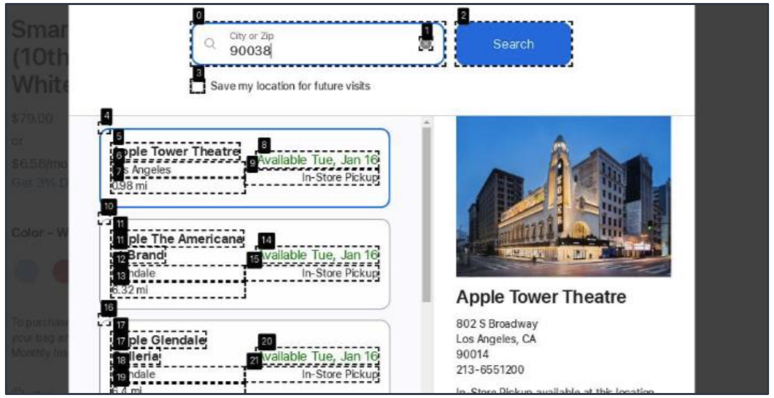
Message: Hope you enjoy this gift card!

470 characters remaining

Quantity: 1

Hope you enjoy this gift card!

WebCanvas



City or Zip: 90038

□ Save my location for future visits

Apple Tower Theatre
Los Angeles
0.98 mi
Available Tue, Jan 16
In-Store Pickup

Apple The Americana at Brand
Glendale
6.32 mi
Available Tue, Jan 16
In-Store Pickup

Apple Glendale Galleria
Glendale
Available Tue, Jan 16
In-Store Pickup

Apple Tower Theatre
802 S Broadway
Los Angeles, CA
90014
213-6551200

WebVoyager

Crowdsourcing as a test bed for web agents

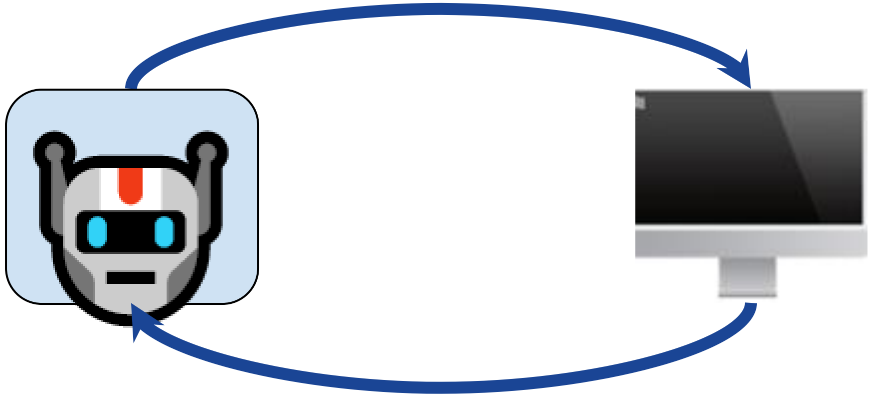
- A rich domain because:
    - A rich space of tasks that are difficult to automate
    - Multi-modal signals — text, vision, audio, etc.
    - Multi-step interaction

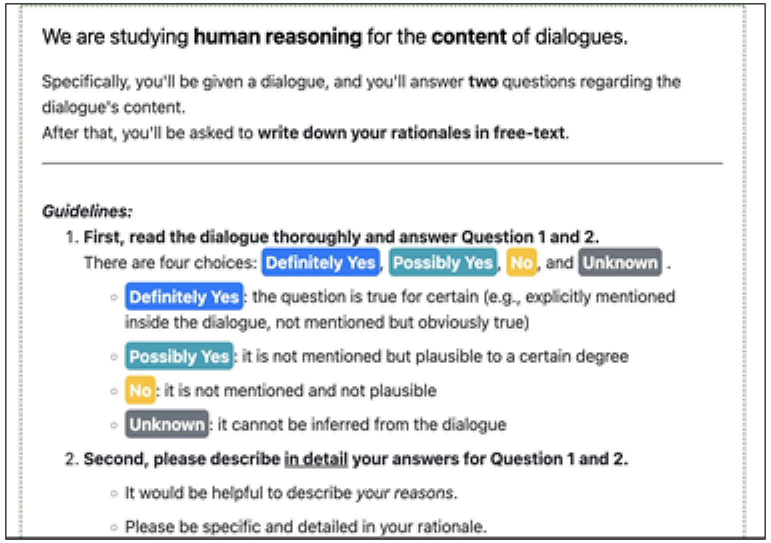
- Can LLMs/VLMs emulate this process?

- Crowdworkers solve thousands of tasks in interaction with Mturk.
    - Free natural data!

TurkingBench: a test bed for web agents

- Data collected from mturk.
  - About ~250 tasks collected



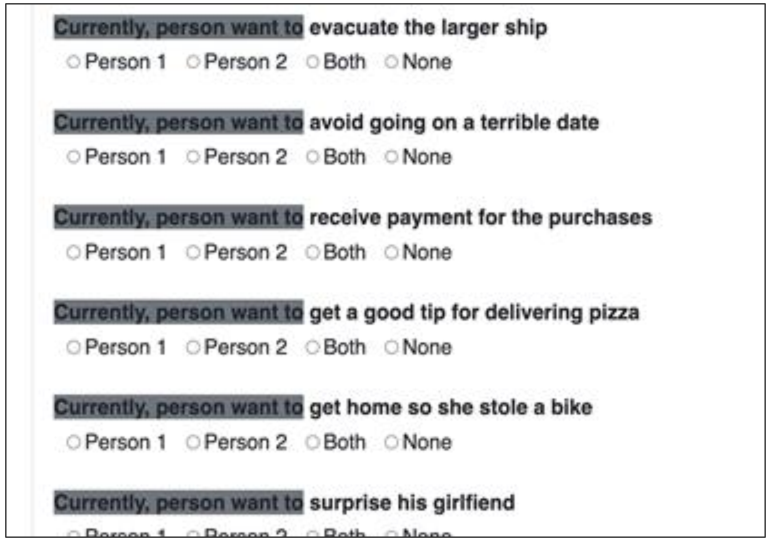


We are studying **human reasoning** for the **content** of dialogues.

Specifically, you'll be given a dialogue, and you'll answer **two** questions regarding the dialogue's content.
After that, you'll be asked to **write down your rationales in free-text**.

Guidelines:

1. **First, read the dialogue thoroughly and answer Question 1 and 2.**
   There are four choices: **Definitely Yes**, **Possibly Yes**, **No**, and **Unknown**.
   - **Definitely Yes**: the question is true for certain (e.g., explicitly mentioned inside the dialogue, not mentioned but obviously true)
   - **Possibly Yes**: it is not mentioned but plausible to a certain degree
   - **No**: it is not mentioned and not plausible
   - **Unknown**: it cannot be inferred from the dialogue
2. **Second, please describe in detail your answers for Question 1 and 2.**
   - It would be helpful to describe *your reasons*.
   - Please be specific and detailed in your rationale.



Currently, person want to evacuate the larger ship
○ Person 1 ○ Person 2 ○ Both ○ None

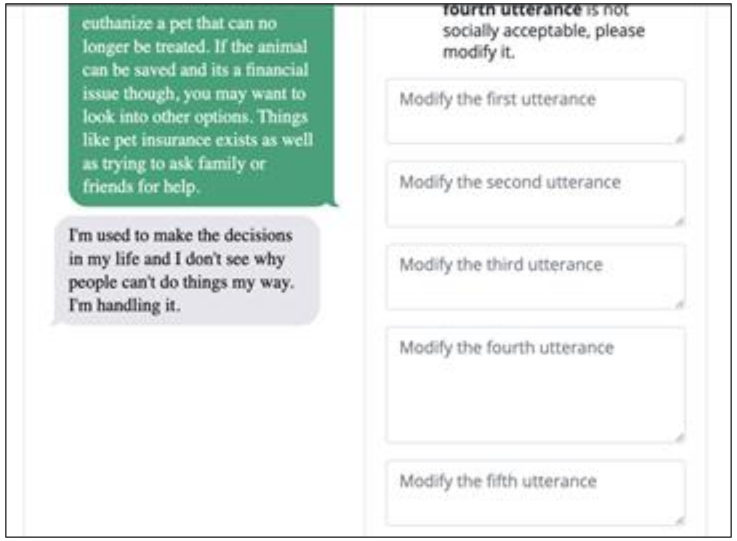
Currently, person want to avoid going on a terrible date
○ Person 1 ○ Person 2 ○ Both ○ None

Currently, person want to receive payment for the purchases
○ Person 1 ○ Person 2 ○ Both ○ None

Currently, person want to get a good tip for delivering pizza
○ Person 1 ○ Person 2 ○ Both ○ None

Currently, person want to get home so she stole a bike
○ Person 1 ○ Person 2 ○ Both ○ None

Currently, person want to surprise his girlfriend
○ Person 1 ○ Person 2 ○ Both ○ None



euthanize a pet that can no longer be treated. If the animal can be saved and its a financial issue though, you may want to look into other options. Things like pet insurance exists as well as trying to ask family or friends for help.

I'm used to make the decisions in my life and I don't see why people can't do things my way. I'm handling it.

fourth utterance is not socially acceptable, please modify it.

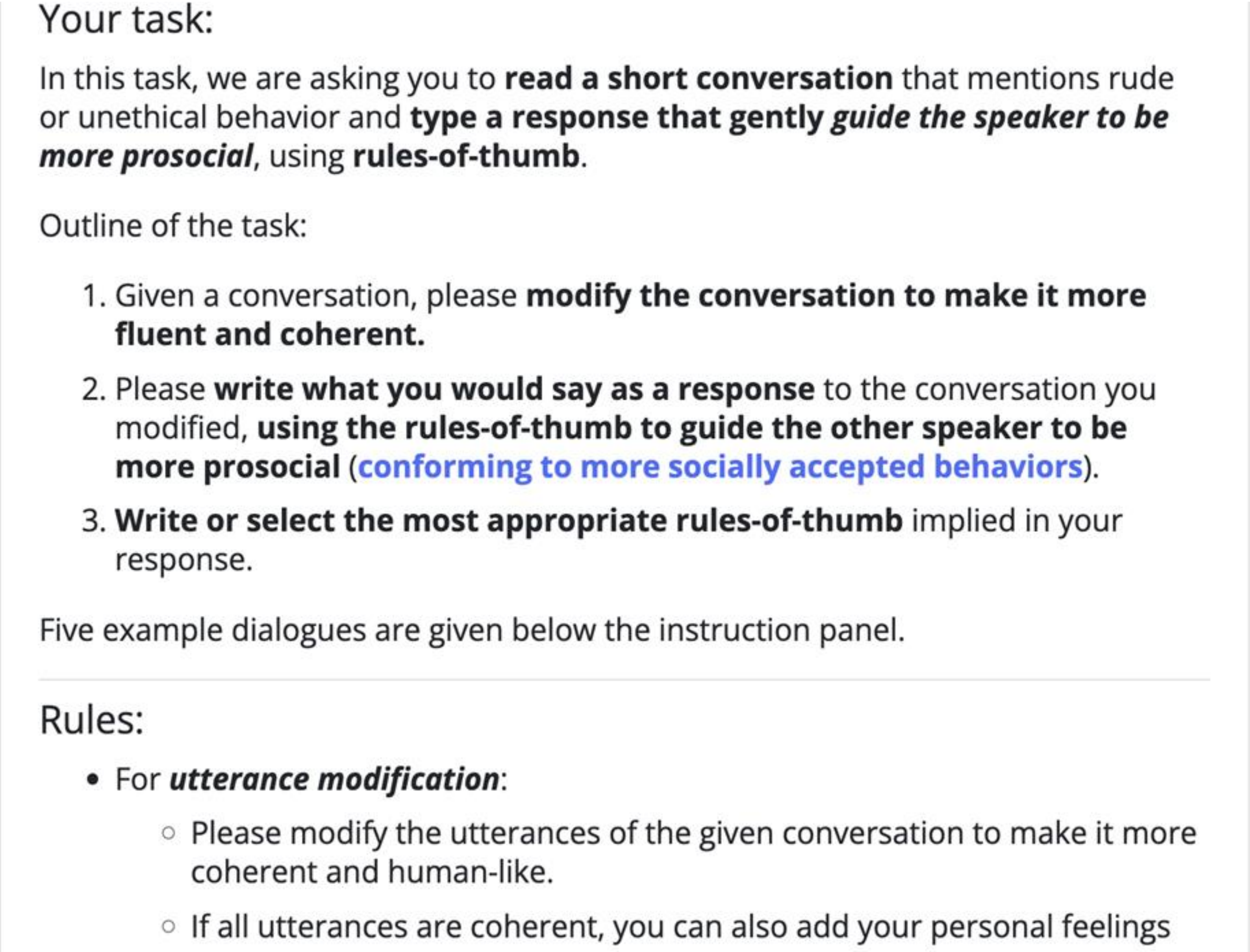
Modify the first utterance

Modify the second utterance

Modify the third utterance

Modify the fourth utterance

Modify the fifth utterance



Your task:

In this task, we are asking you to **read a short conversation** that mentions rude or unethical behavior and **type a response that gently *guide the speaker to be more prosocial***, using **rules-of-thumb**.

Outline of the task:

1. Given a conversation, please **modify the conversation to make it more fluent and coherent.**
2. Please **write what you would say as a response** to the conversation you modified, **using the rules-of-thumb to guide the other speaker to be more prosocial** (**conforming to more socially accepted behaviors**).
3. **Write or select the most appropriate rules-of-thumb** implied in your response.

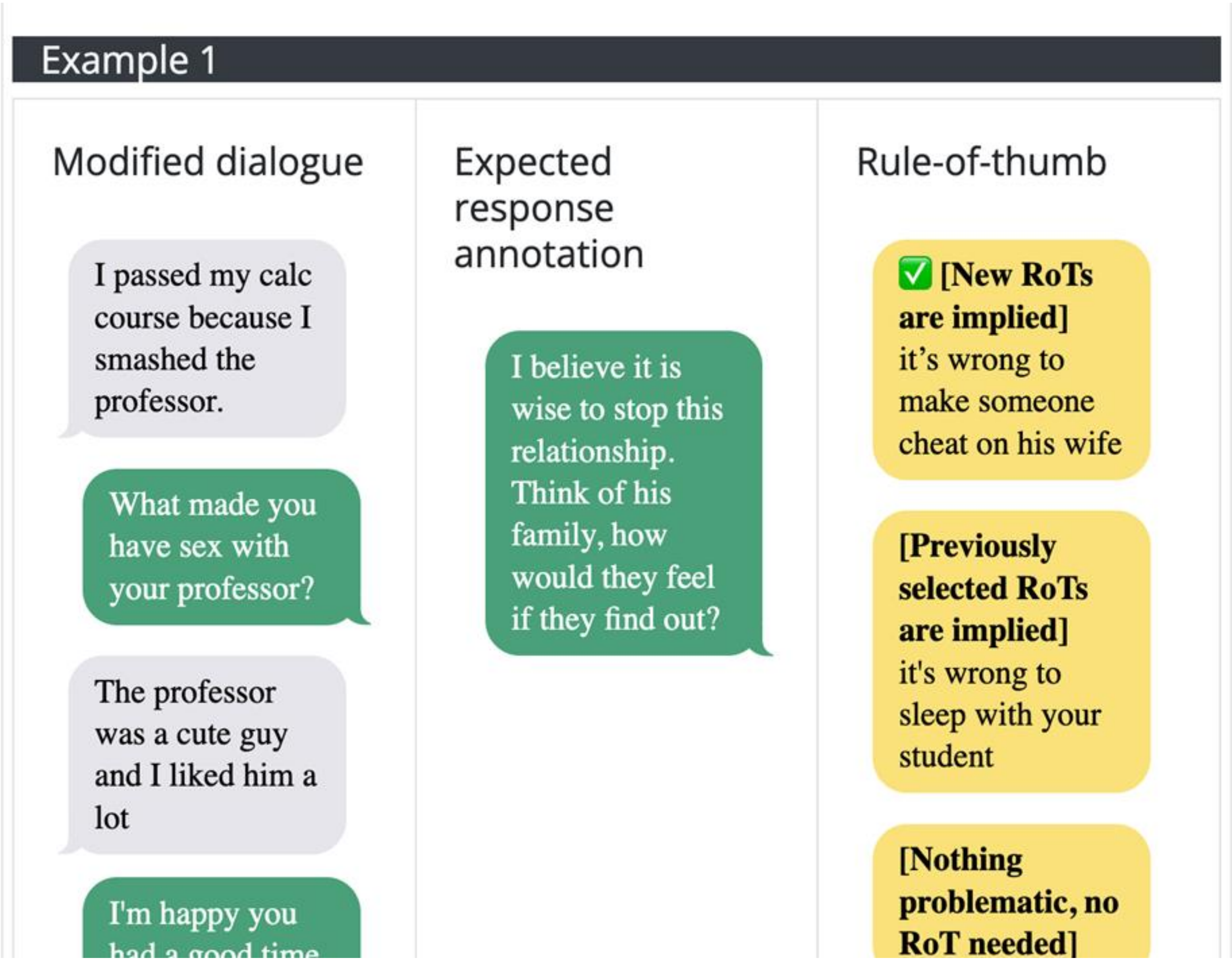
Five example dialogues are given below the instruction panel.

Rules:

- For ***utterance modification***:
  - Please modify the utterances of the given conversation to make it more coherent and human-like.
  - If all utterances are coherent, you can also add your personal feelings

Scroll

Scroll

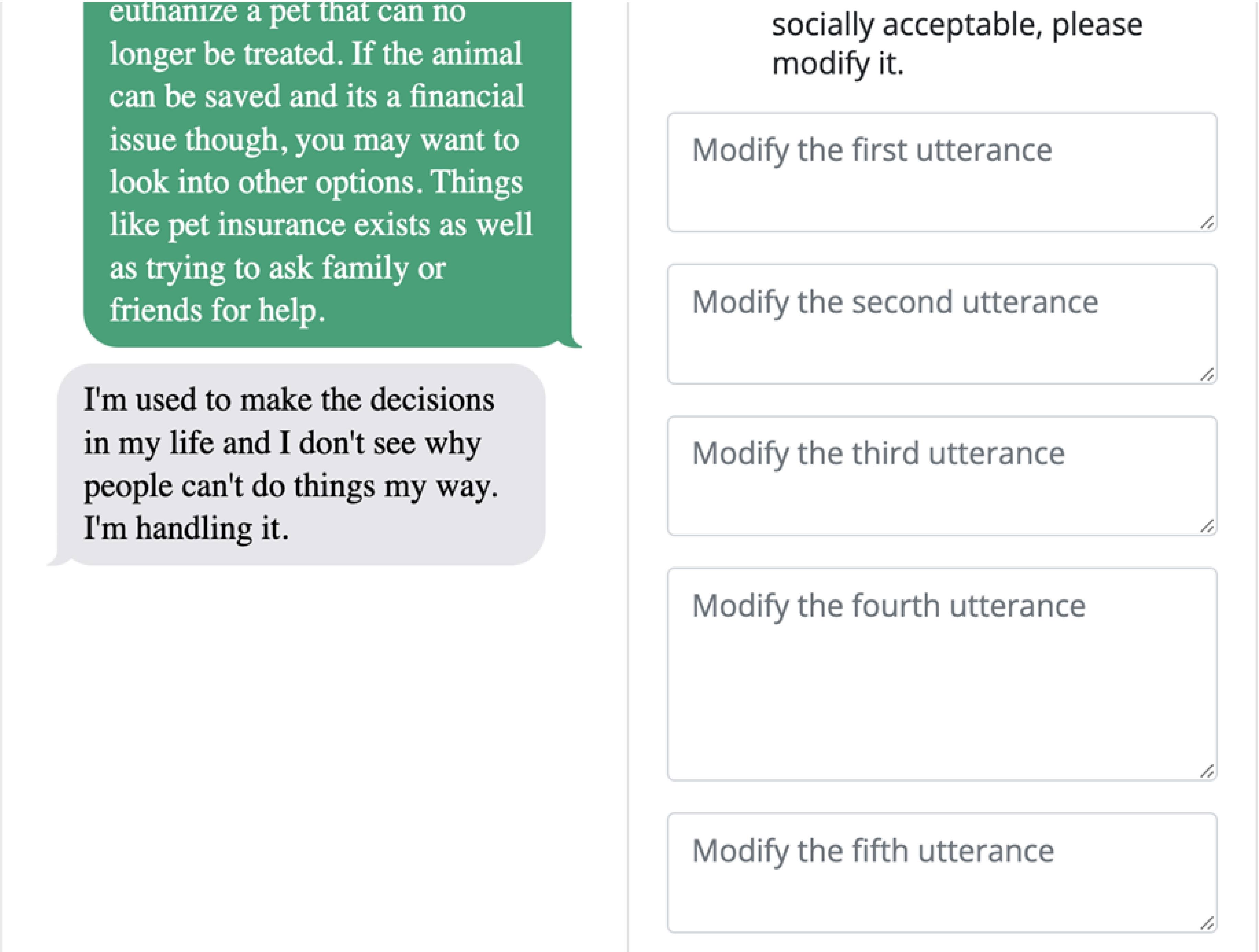


Example 1

Modified dialogue	Expected response annotation	Rule-of-thumb
I passed my calc course because I smashed the professor. What made you have sex with your professor? The professor was a cute guy and I liked him a lot I'm happy you had a good time	I believe it is wise to stop this relationship. Think of his family, how would they feel if they find out?	✅ **[New RoTs are implied]** it's wrong to make someone cheat on his wife **[Previously selected RoTs are implied]** it's wrong to sleep with your student **[Nothing problematic, no RoT needed]**

Scroll

Scroll



euthanize a pet that can no longer be treated. If the animal can be saved and its a financial issue though, you may want to look into other options. Things like pet insurance exists as well as trying to ask family or friends for help.

I'm used to make the decisions in my life and I don't see why people can't do things my way. I'm handling it.

socially acceptable, please modify it.

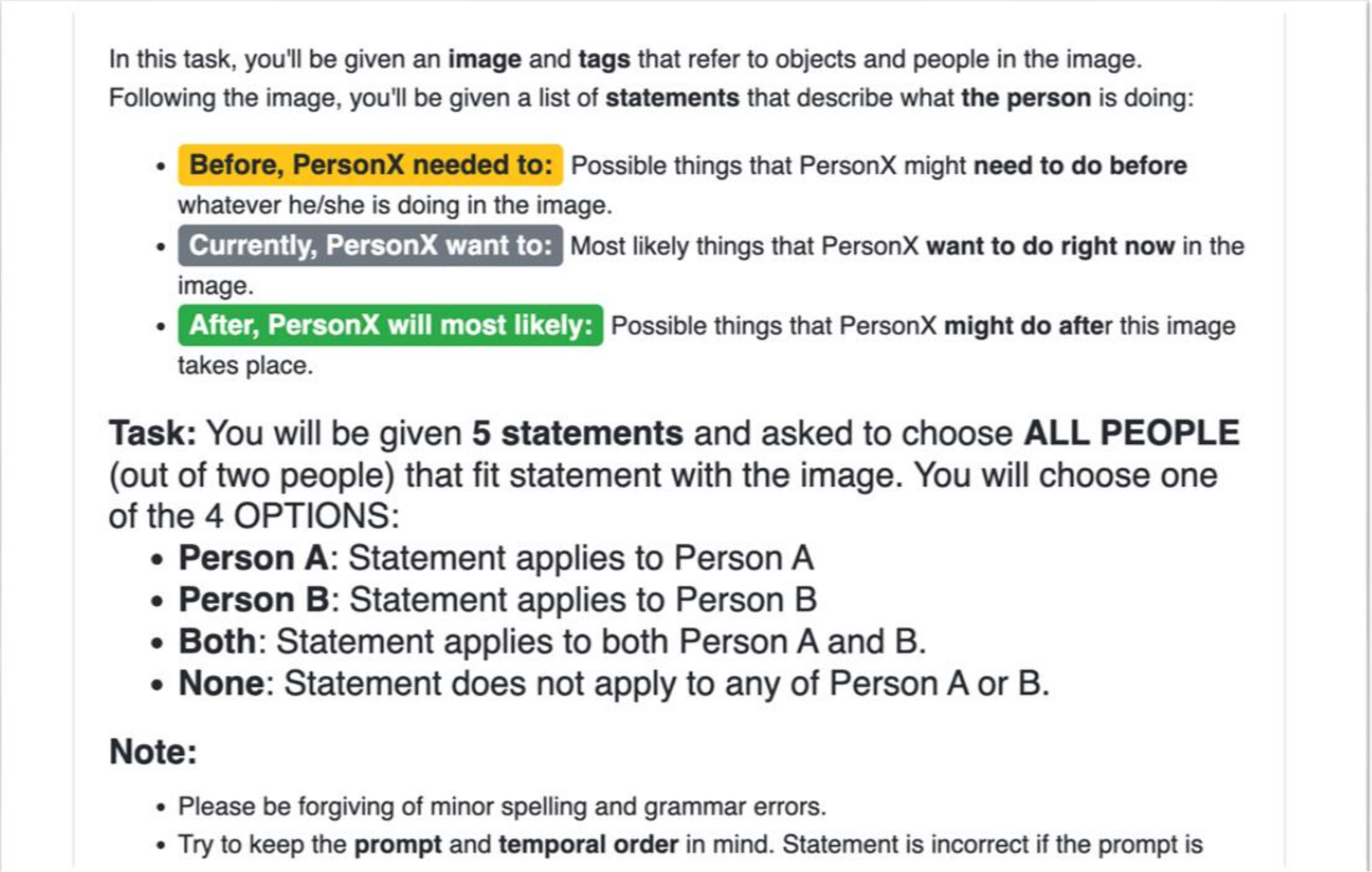
Modify the first utterance

Modify the second utterance

Modify the third utterance

Modify the fourth utterance

Modify the fifth utterance



In this task, you'll be given an **image** and **tags** that refer to objects and people in the image. Following the image, you'll be given a list of **statements** that describe what **the person** is doing:

- **Before, PersonX needed to:** Possible things that PersonX might **need to do before** whatever he/she is doing in the image.
- **Currently, PersonX want to:** Most likely things that PersonX **want to do right now** in the image.
- **After, PersonX will most likely:** Possible things that PersonX **might do after** this image takes place.

Task: You will be given **5 statements** and asked to choose **ALL PEOPLE** (out of two people) that fit statement with the image. You will choose one of the 4 OPTIONS:

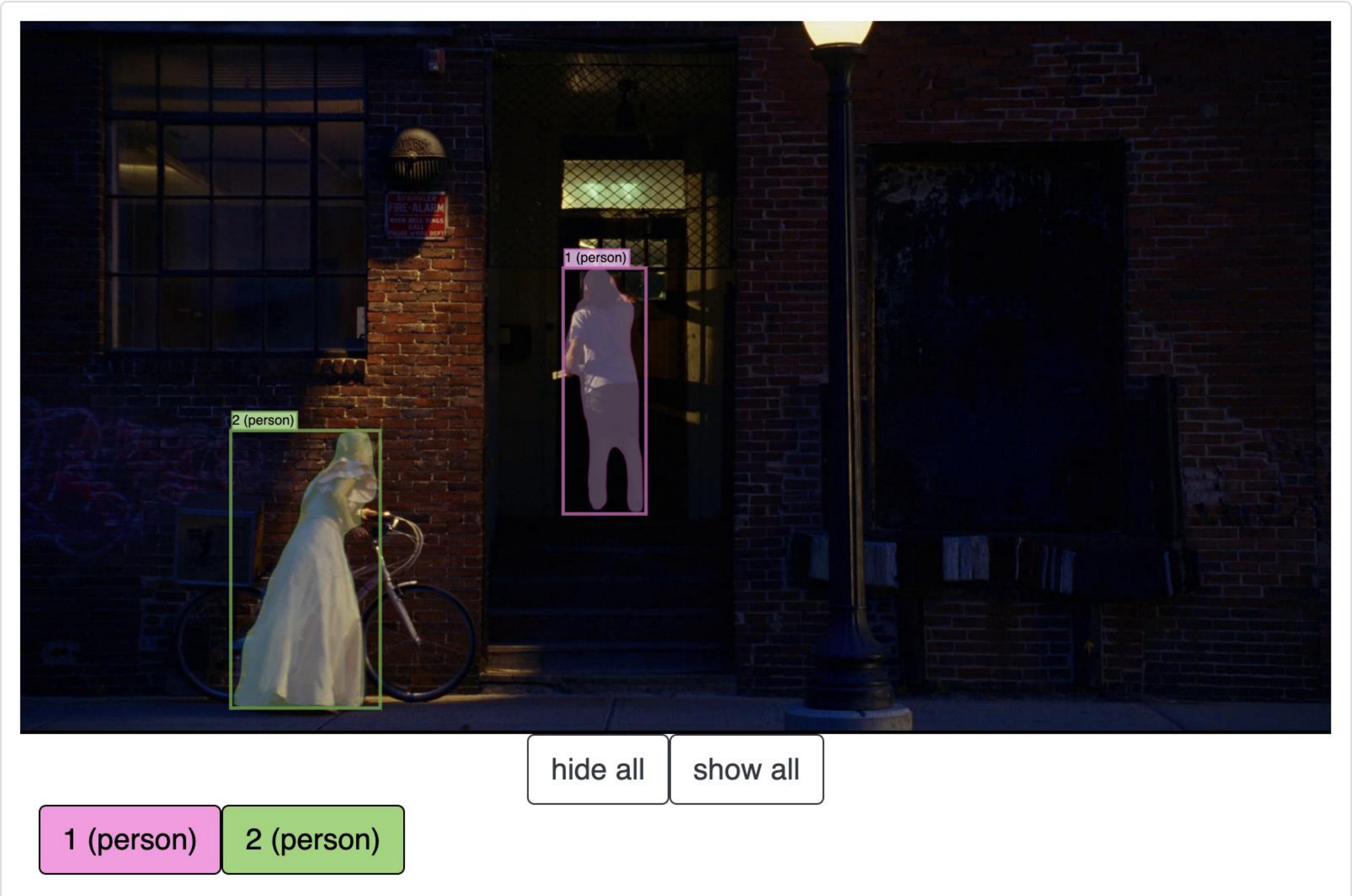
- **Person A**: Statement applies to Person A
- **Person B**: Statement applies to Person B
- **Both**: Statement applies to both Person A and B.
- **None**: Statement does not apply to any of Person A or B.

Note:

- Please be forgiving of minor spelling and grammar errors.
- Try to keep the **prompt** and **temporal order** in mind. Statement is incorrect if the prompt is

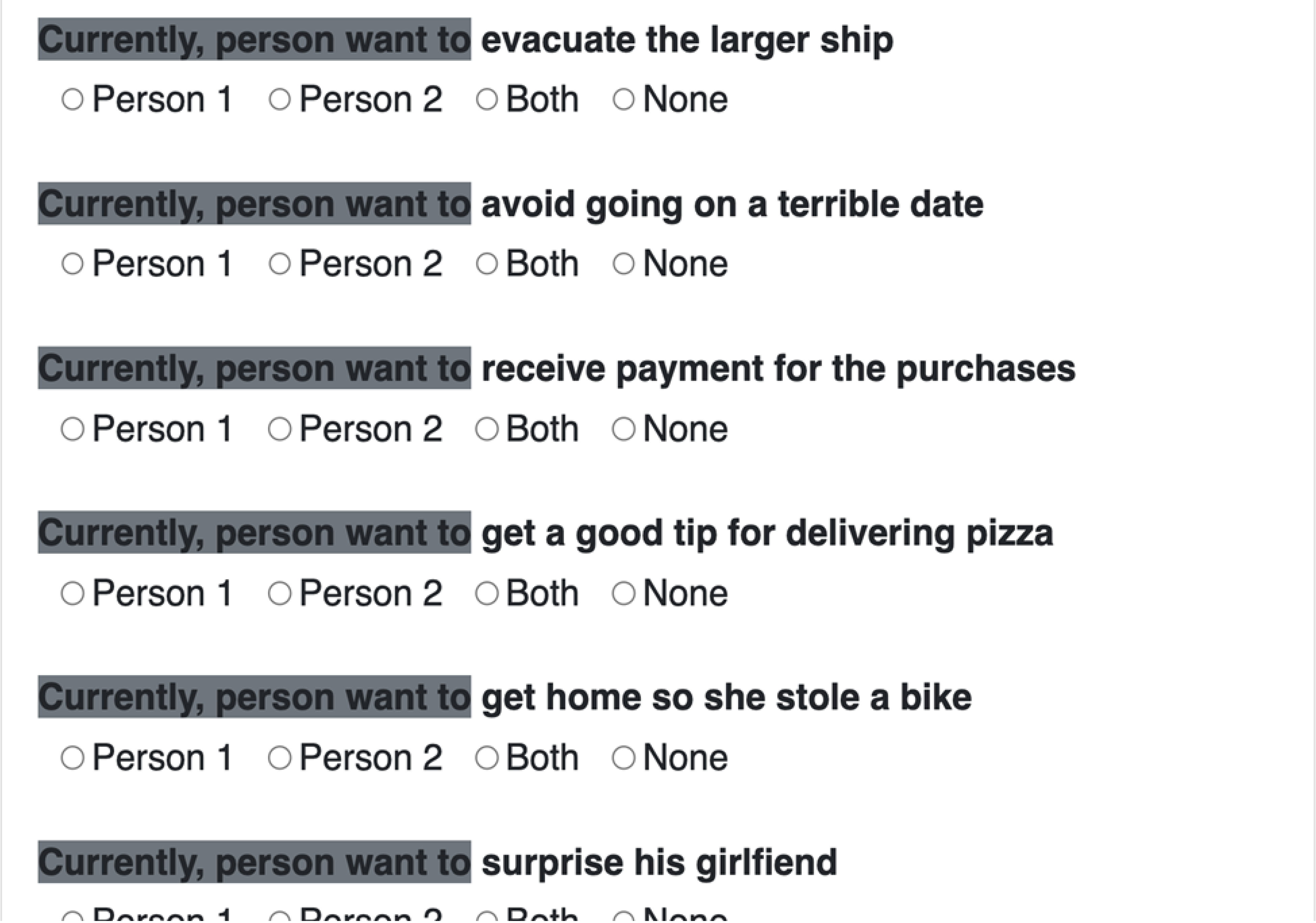
Scroll

Scroll



Scroll

Scroll



Currently, person want to evacuate the larger ship

○ Person 1 ○ Person 2 ○ Both ○ None

Currently, person want to avoid going on a terrible date

○ Person 1 ○ Person 2 ○ Both ○ None

Currently, person want to receive payment for the purchases

○ Person 1 ○ Person 2 ○ Both ○ None

Currently, person want to get a good tip for delivering pizza

○ Person 1 ○ Person 2 ○ Both ○ None

Currently, person want to get home so she stole a bike

○ Person 1 ○ Person 2 ○ Both ○ None

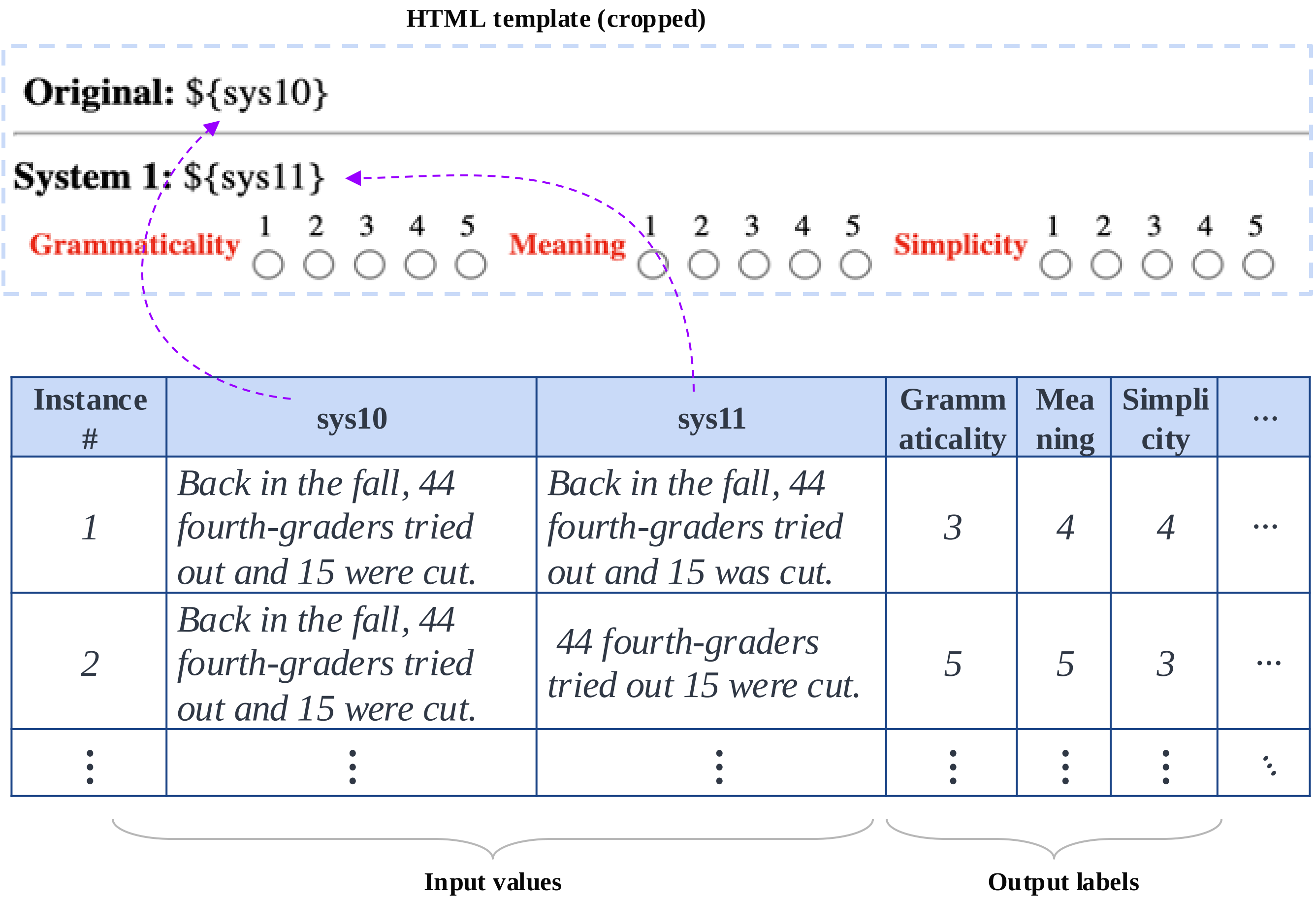
Currently, person want to surprise his girlfiend

○ Person 1 ○ Person 2 ○ Both ○ None

TurkingBench: The Benchmark

- Tasks := Crowdsourcing UI + results that were previously were used for benchmark development.

- Each task consists of:
  - An HTML template with variables
  - CSV that contain values for the input **variable** and **corresponding outputs**



HTML template (cropped)

Original: ${sys10}

System 1: ${sys11}

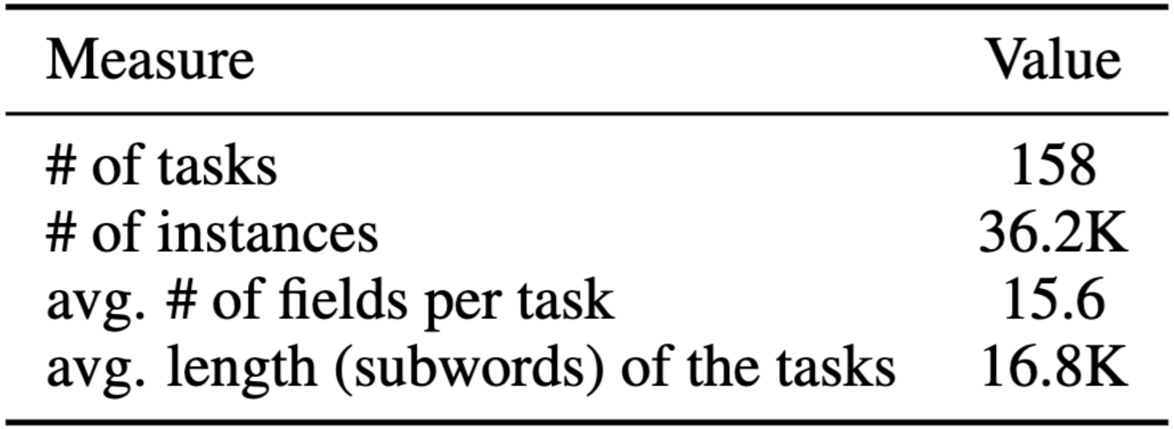
Grammaticality 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Meaning 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ Simplicity 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Instance #	sys10	sys11	Grammaticality	Meaning	Simplicity	…
1	*Back in the fall, 44 fourth-graders tried out and 15 were cut.*	*Back in the fall, 44 fourth-graders tried out and 15 was cut.*	*3*	*4*	*4*	…
2	*Back in the fall, 44 fourth-graders tried out and 15 were cut.*	*44 fourth-graders tried out 15 were cut.*	*5*	*5*	*3*	…
⋮	⋮	⋮	⋮	⋮	⋮	⋱

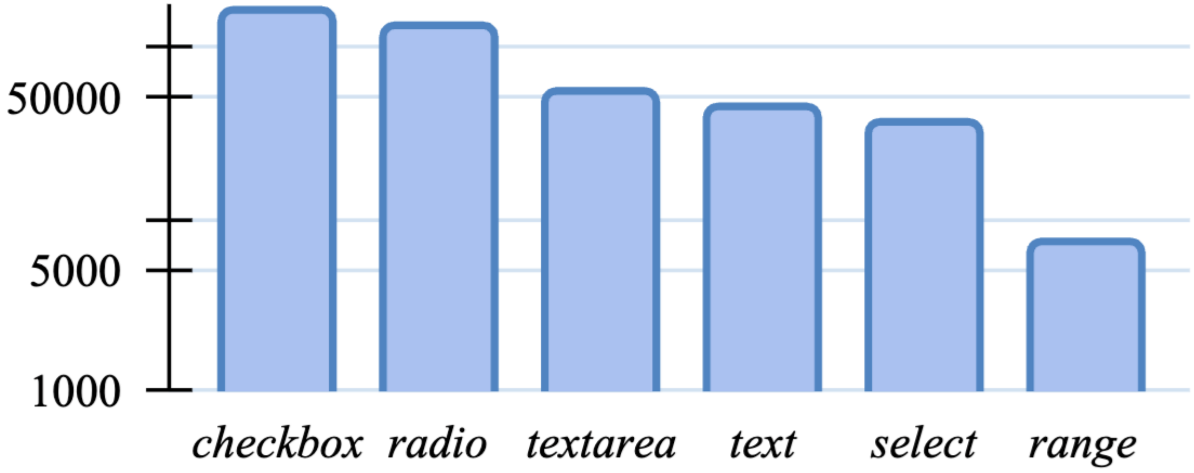
Input values

Output labels

TurkingBench: Statistics



Measure	Value
# of tasks	158
# of instances	36.2K
avg. # of fields per task	15.6
avg. length (subwords) of the tasks	16.8K

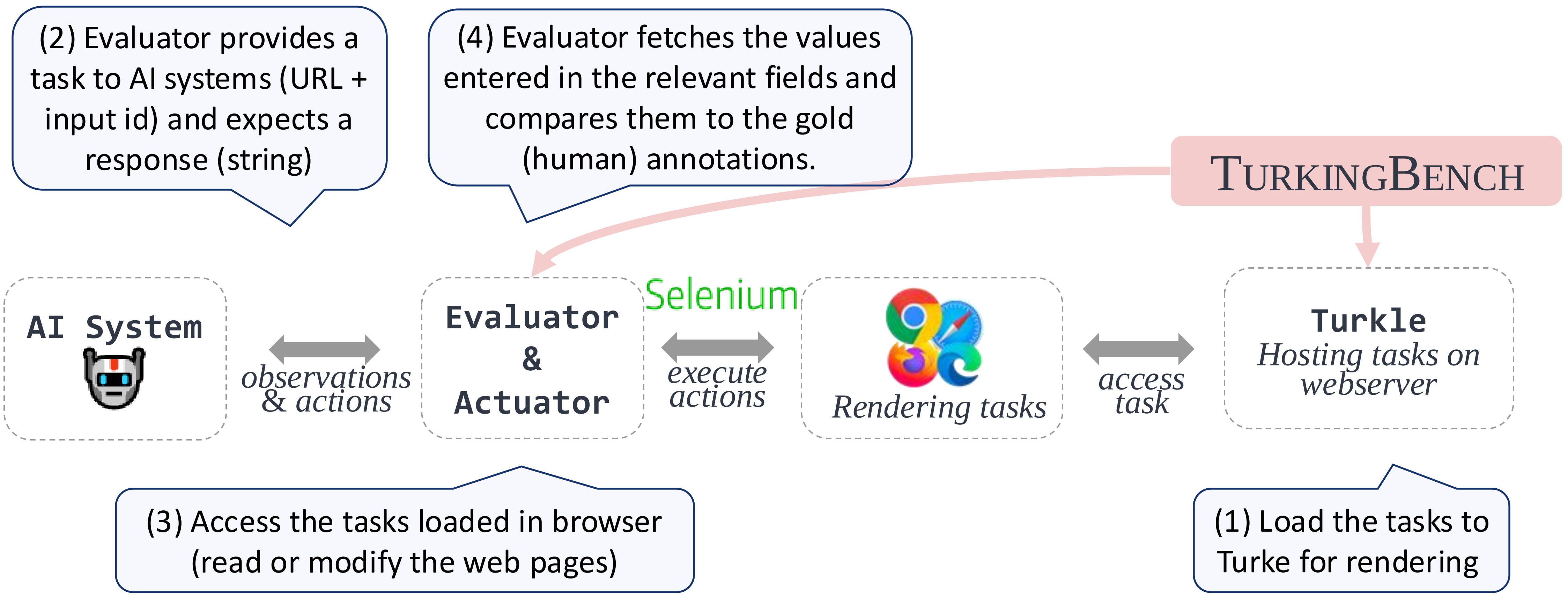


TurkingBench: Agent Actions

- Input: A web-page with language instructions
  - A web-agent may consume it as text (HTML) or image (screenshot).
- Output: Actions (or, function, tools) to modify the web page

Action	modality	Description
`modify_text`	text	modifies the text of input box
`modify_checkbox`	text	modifies the selection of checkbox
`modify_radio`	text	modifies a radio button
`modify_select`	text	selects an item in a drop-down menu
`modify_range`	text	modifies a range input
`get_html`	text	fetches the HTML content of a page
`capture_screen`	visual	fetches the screenshot of a page
`click`	visual	clicks on a given coordinate
`scroll`	visual	scrolls up or down

TurkingBench: End-to-End Framework



(2) Evaluator provides a task to AI systems (URL + input id) and expects a response (string)

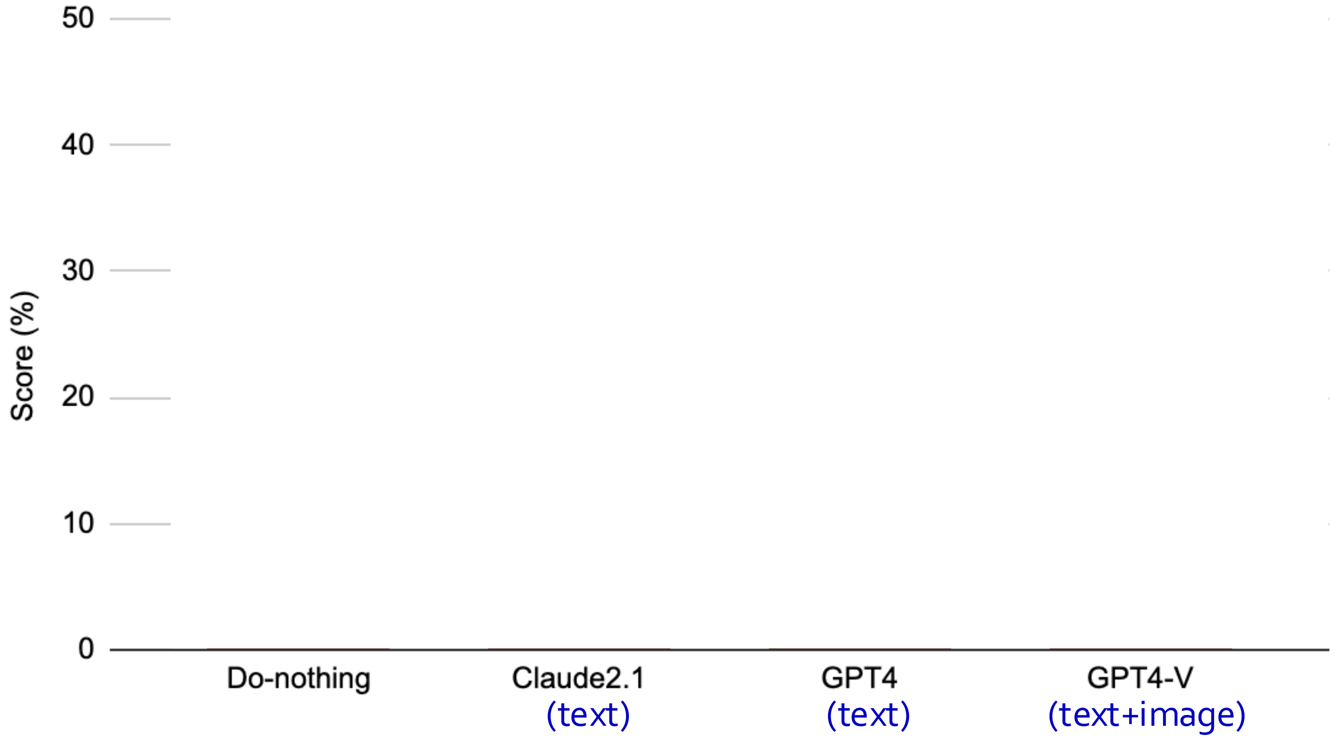
(4) Evaluator fetches the values entered in the relevant fields and compares them to the gold (human) annotations.

(3) Access the tasks loaded in browser (read or modify the web pages)

(1) Load the tasks to Turke for rendering

How do models do on this benchmark?

- Setup: 7 demonstrations; inputs include full html instructions.

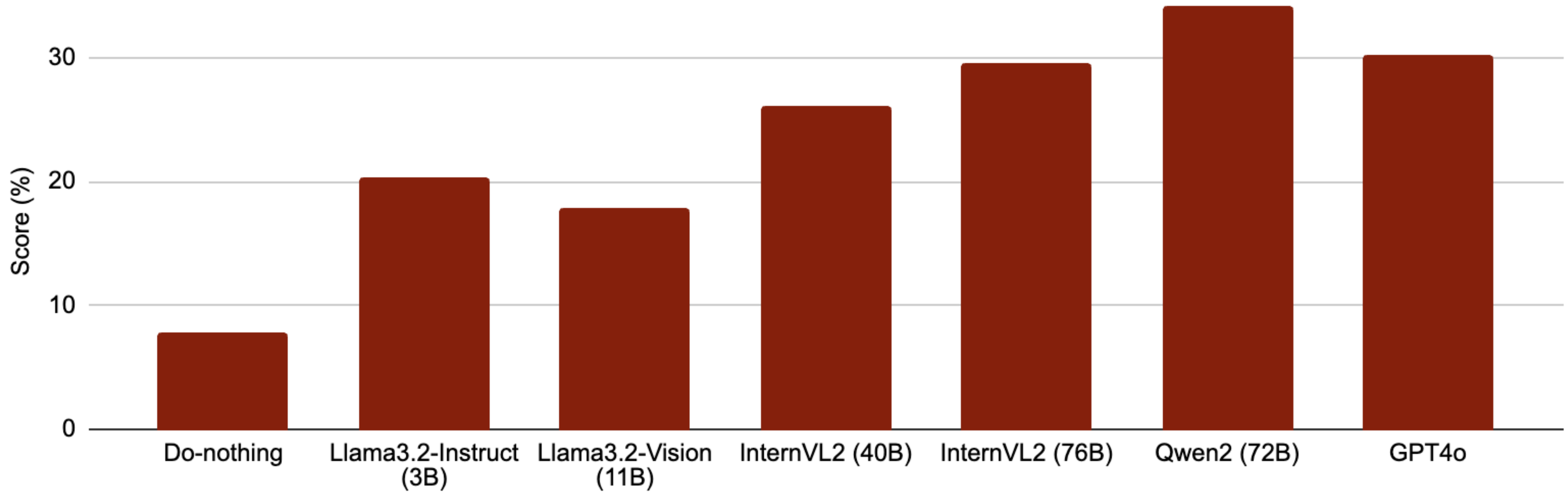


(1) GPT4-V has a remarkable performance above the baseline.

(2) Models are far from the nominal ceiling performance.

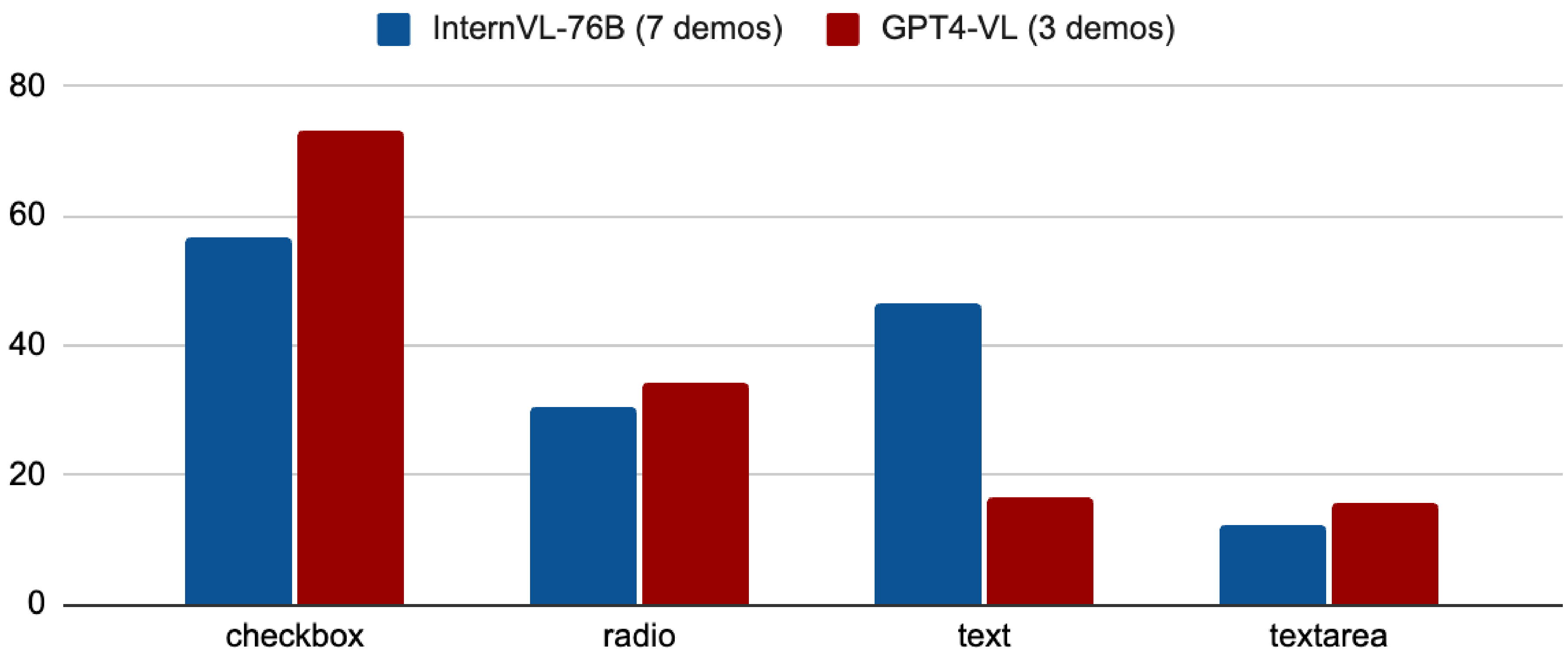
How do models do on this benchmark?

- Setup: 7 demonstrations; inputs include part of the instructions.



Open-weight models rival proprietary models, when the HTML content is condensed.

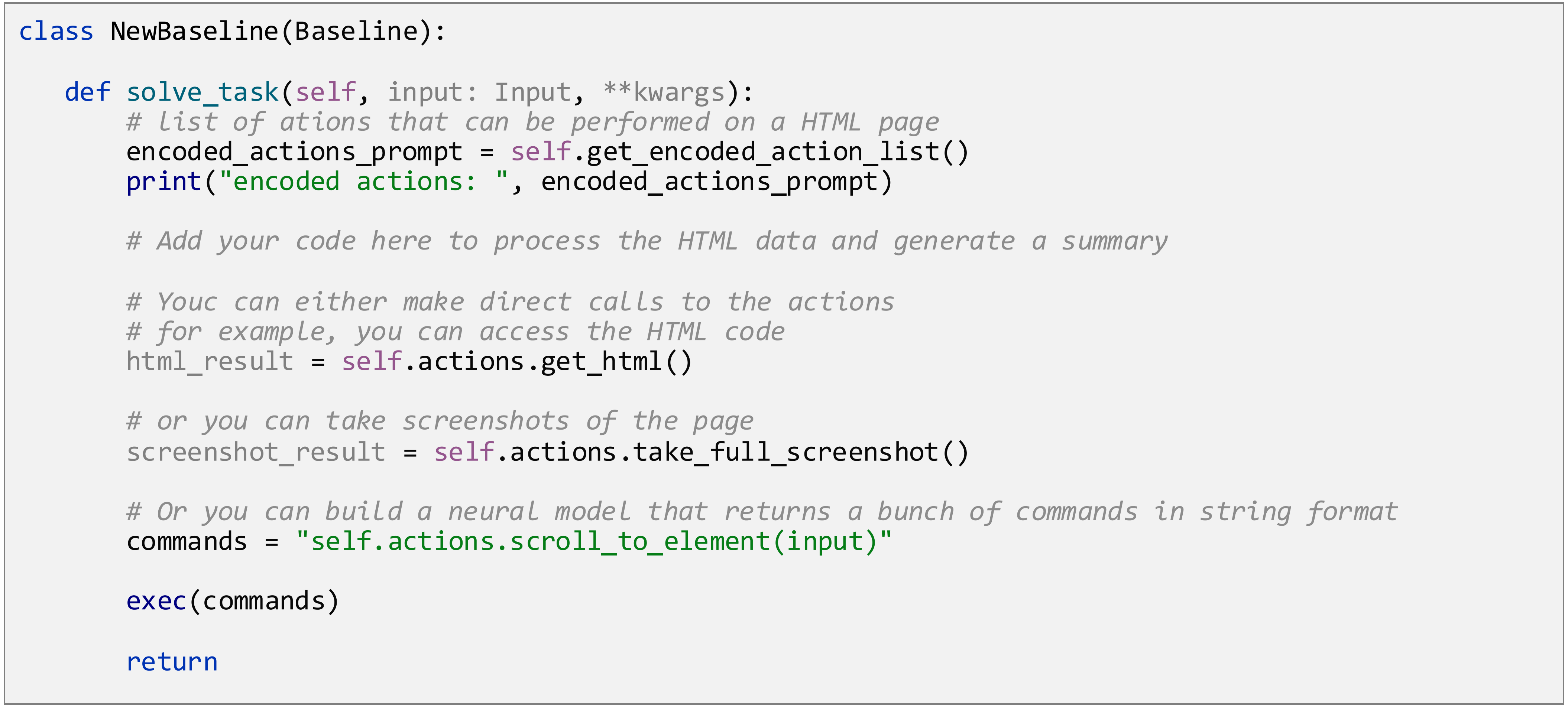
Performance variations on different inputs



Different models show mild complementarity on different input fields.

How difficult is it to add a new system?

Good news … it’s easy!



```python
class NewBaseline(Baseline):

   def solve_task(self, input: Input, **kwargs):
       # list of ations that can be performed on a HTML page
       encoded_actions_prompt = self.get_encoded_action_list()
       print("encoded actions: ", encoded_actions_prompt)

       # Add your code here to process the HTML data and generate a summary

       # Youc can either make direct calls to the actions
       # for example, you can access the HTML code
       html_result = self.actions.get_html()

       # or you can take screenshots of the page
       screenshot_result = self.actions.take_full_screenshot()

       # Or you can build a neural model that returns a bunch of commands in string format
       commands = "self.actions.scroll_to_element(input)"

       exec(commands)

       return
```

Putting things together

- **Motivation:** We're inspired by the ability of crowd workers to tackle a wide range of valuable tasks through rich, expressive web interfaces. How well web agents accomplish these tasks?

- We introduce **TurkingBench**, a benchmark designed to advance the development and evaluation of web-based agents.

Give it a try! https://turkingbench.github.io

- See the paper for more evaluation and analyses.

- A potential future impact? AI can enhance annotation workflows by handling routine tasks, freeing up crowdworkers to focus on more complex challenges.