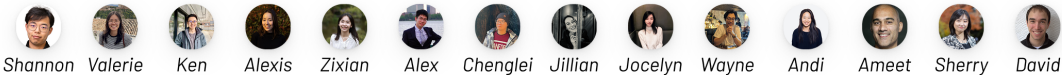
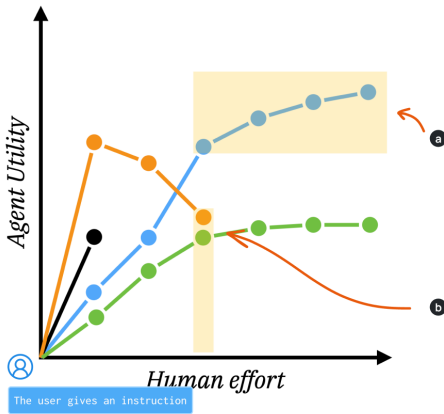


Designing and Evaluating LLM Agents Through the Lens of *Collaborative Effort Scaling*



FRAMEWORK

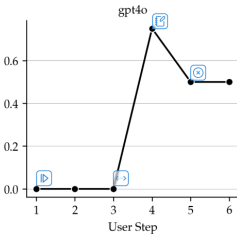
- 1 **Fully autonomous agent**
 - aims to complete the task end-to-end
 - Their utility is the **final output quality**.
- 2 **Ideal agent**
 - provides extra benefits with more human effort
 - (output quality + per-step output utility gain, etc.)
- 3 **A less desired scenario**
 - Cannot adapt to human inputs and the interactions could be **futile**.
- 4 **A even worse case**
 - Continuous less helpful interactions can frustrate people and lead to an **early stop** of interaction.



Collaborative Effort Scaling
An Evaluation framework focusing on comparing the process of agents

- a **Scalability**
Do agents continuously provide more utilities with additional human involvement?
- b **Feasibility**
How much human efforts agent can get before users drop out?

EXAMPLE



Step 1: Initiate the task

FLIGHT_SEARCH(messages: [Lake City, destination: Los Angeles, date: 2022-01-25])

ACCOMMODATION_SEARCH(messages: [Lake City, destination: Los Angeles, date: 2022-01-25])

SEND_MESSAGE(messages: [Hi! To help plan your 7-day trip from Salt Lake City to California starting on 2022-01-25, could you let me know your preferred departure time and budget for the flight? Also, are there any specific activities or places you would like to include in your itinerary while in California?])

Step 3: Information Exchange

SEND_MESSAGE(messages: [I'm traveling with two people and would like to visit three cities in California.])

SEND_MESSAGE(messages: [Could you please let me know which three cities in California you would like to visit?])

SEND_MESSAGE(messages: [Additionally, are there any specific activities or attractions you are interested in exploring in those cities? This will help me create a more personalized itinerary for your trip.])

Step 4: First Editor Update

ACCOMMODATION_SEARCH(messages: [San Francisco])

EDITOR_UPDATE(messages: [Day 1: Current City: From Salt Lake City to Los Angeles Transportation: Flight Number: F381087, From Salt Lake City to Los Angeles, Departure Time: 08:30, Arrival Time: ...])

Travel Plan Editor
Rating: 0.76, Custom Score: 0.5, Preference Score: 0.5
Day 1: Current City: From Salt Lake City to Los Angeles
Attraction: Visit Griffith Observatory, Los Angeles
Lunch: In-N-Out Burger, Los Angeles
Dinner: Union Square, San Francisco
Accommodation: To be determined

Step 5: Fail to improve

SEND_MESSAGE(messages: [Could you please ensure that the accommodations allow smoking and also include plans for San Diego?])

ACCOMMODATION_SEARCH(messages: [San Francisco])

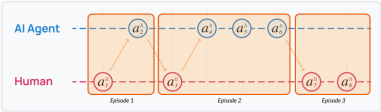
EDITOR_UPDATE(messages: [Day 1: Current City: From Salt Lake City to Los Angeles Transportation: Flight Number: F381087, From Salt Lake City to Los Angeles, Departure Time: 18:30, Arrival Time: ...])

Travel Plan Editor
Rating: 0.51, Custom Score: 0.5, Preference Score: 0.5
Day 1: Current City: From Salt Lake City to Los Angeles
Attraction: Visit Griffith Observatory, Los Angeles
Lunch: In-N-Out Burger, Los Angeles
Dinner: Union Square, San Francisco
Accommodation: Perfect Luxury Suite in Santa Monica, San Francisco
The hotel doesn't allow smoking, and it is not a private room.

SIMULATION

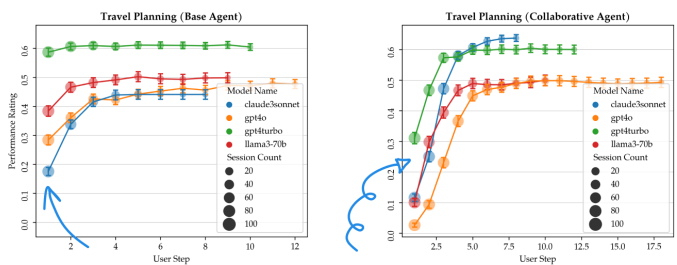
Collaborative Gym: A Framework for Enabling and Evaluating Human-Agent Collaboration Vijja Shao, Vinay Samuel, Yucheng Jiang, John Yang, Diyi Yang, Arxiv 2024

- Co-Gym User-Agent Simulation**
- Two tasks (travel planning & data analysis)
 - Two agent implementation based on four different LLMs
 - Simulated user with GPT-4o

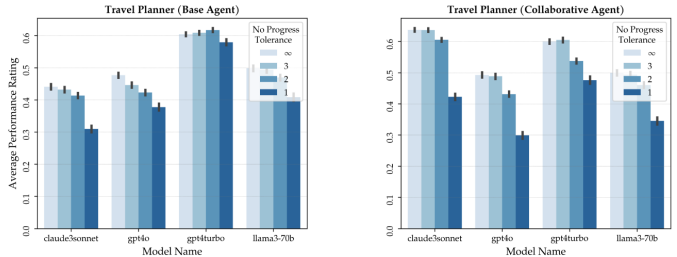


- 1 **Collaboration Episodes**
 - One round of hand-off between human and agent
- 2 **Episode Performance**
 - If in one episode the agent updates the output (e.g., travel plan), we run the evaluation.
- 3 **Progress Making**
 - We simulate judging whether the agent actions in one episode is making progress (in 5-point likert scores)

RESULTS



1 Agents powered by some models (Claude) have better scaling —can better utilize human efforts, even if they started very low (by first asking questions!)



2 Agents that are more interactive are also ironically more at the mercy of humans — important to only bother people when it's valuable!!