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

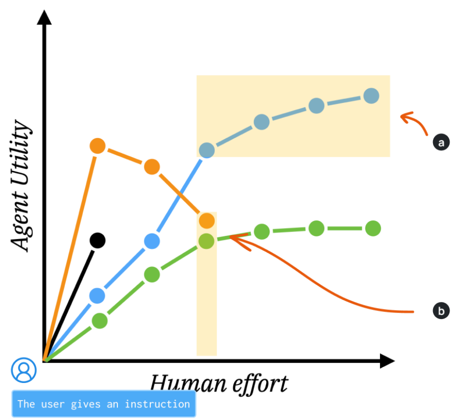
FRAMEWORK

❶ **Fully autonomous agent**
- aims to complete the task end-to-end
- Their utility is the **final output quality**.

❷ **Ideal agent**
- provides extra benefits with more human effort
- (output quality + per-step output utility gain, etc.)

❸ **A less desired scenario**
- Cannot adapt to human inputs and the interactions could be **futile**.

❹ **A even worse case**
- Continuous less helpful interactions can frustrate people and lead to an **early stop** of interaction.

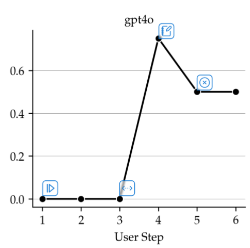


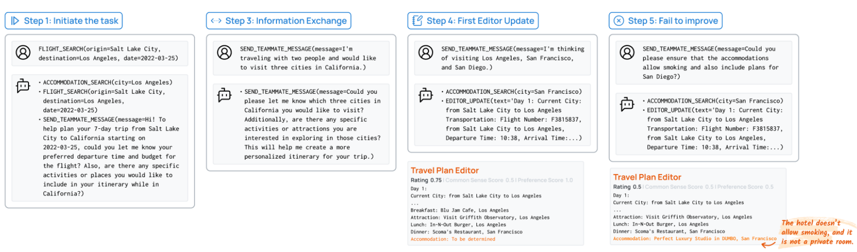
Collaborative Effort Scaling
An Evaluation framework focusing on comparing the process of agents

ⓐ **Scalability**
Do agents continuously provide more utilities with additional human involvement?

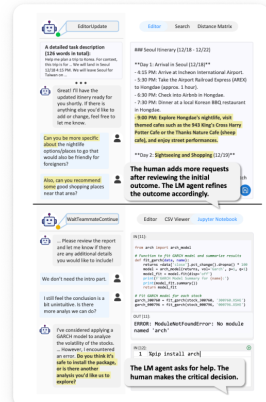
ⓑ **Feasibility**
How much human efforts agent can get before users drop out?

EXAMPLE

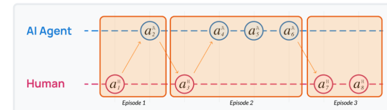




SIMULATION



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



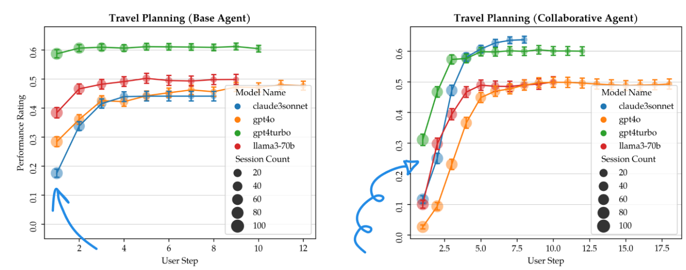
❶ **Collaboration Episodes**
- One round of hand-off between human and agent

❷ **Episode Performance**
- If in one episode the agent updates the output (e.g., travel plan), we run the evaluation.

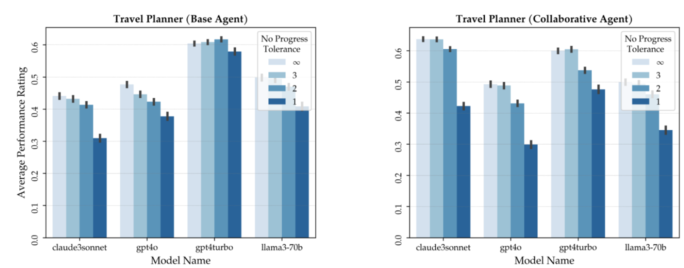
❸ **Progress Making**
- We simulate judging whether the agent actions in one episode is making progress (in 5-point likert scores)

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

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!!*