Imitating Interactive Intelligence

Stanford CS379c Lecture
April 20, 2021

Greg Wayne, DeepMind

Interactive Agents Group

Interactive Agents

Creating agents that cooperatively interact with humans.

Why *interaction*?

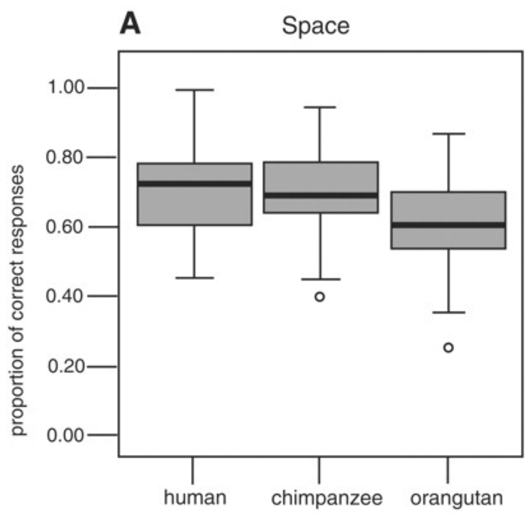
1. Interaction with humans is the best test of intelligence (Turing, 1951).

2. Agents that interact with humans (answering questions, helping, learning socially) could profoundly enable people.

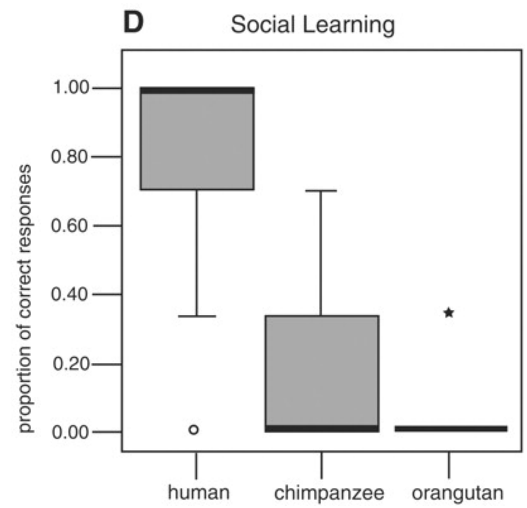
Our long-term goal: produce agents that can learn socially from humans as does a child or peer.

Social Learning is the Source of Human Intelligence

(Non-verbal intelligence test)



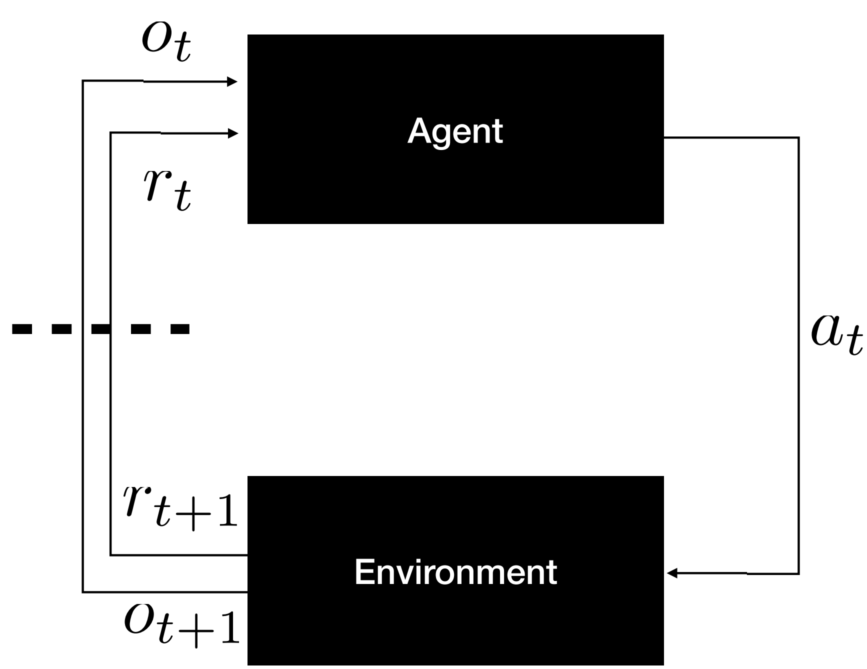
Problem solving involving spatial understanding



involving e.g. non-verbal imitation

Herrmann et al., Science, 2007: Cited in Tomasello, 2019

What's an Agent?



Typical Training Paradigm

Agent interacts with environment and receives programmed reward for successes.

Example: play Go and receive reward = 1 upon winning, reward = -1 upon losing, reward = 0 at other moments.

$$\sum_{a_t} \pi_\theta(a_t \mid o_{\leq t}) \mathbb{E}_{\pi_\theta}[R_t \mid o_{\leq t}, a_t]$$

$$R_t = \sum_{t' \geq t} r_{t'} = r_t + r_{t+1} + r_{t+2} \ldots$$

From Untrained Agents to Agents that Interact

Human children: organic process of nurtured and self-directed learning

Untrained AIs: nurtured and self-directed learning hard to implement

- [Lack human objectives]: We do not (yet) understand human drives and motivations at an algorithmic level: complicated, species-specific, hard-to-guess.
- [Feedback from scratch]: Agents begin at *tabula rasa* (blank slate / monkeys typing on typewriters). Intractable for humans to watch untrained agents and give reinforcing feedback until agents reach competence in practical amounts of time.
- [Ambiguity in communication]: Even simple instructions can be ambiguous. “Go near the door.” What is “near”?

Therefore, it is difficult to (a) write down an objective for agent development; (b) provide feedback for untrained agents; and (c) formalize reward for even very simple communicative interactions.

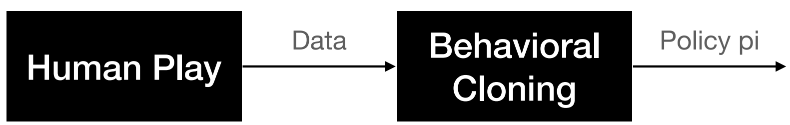
Imitation Learning for Creating Behavioral *Priors*

Increasingly commonly: use supervised learning as an initial basis for behavior. Then improve from there.

GPT-3

AlphaGo

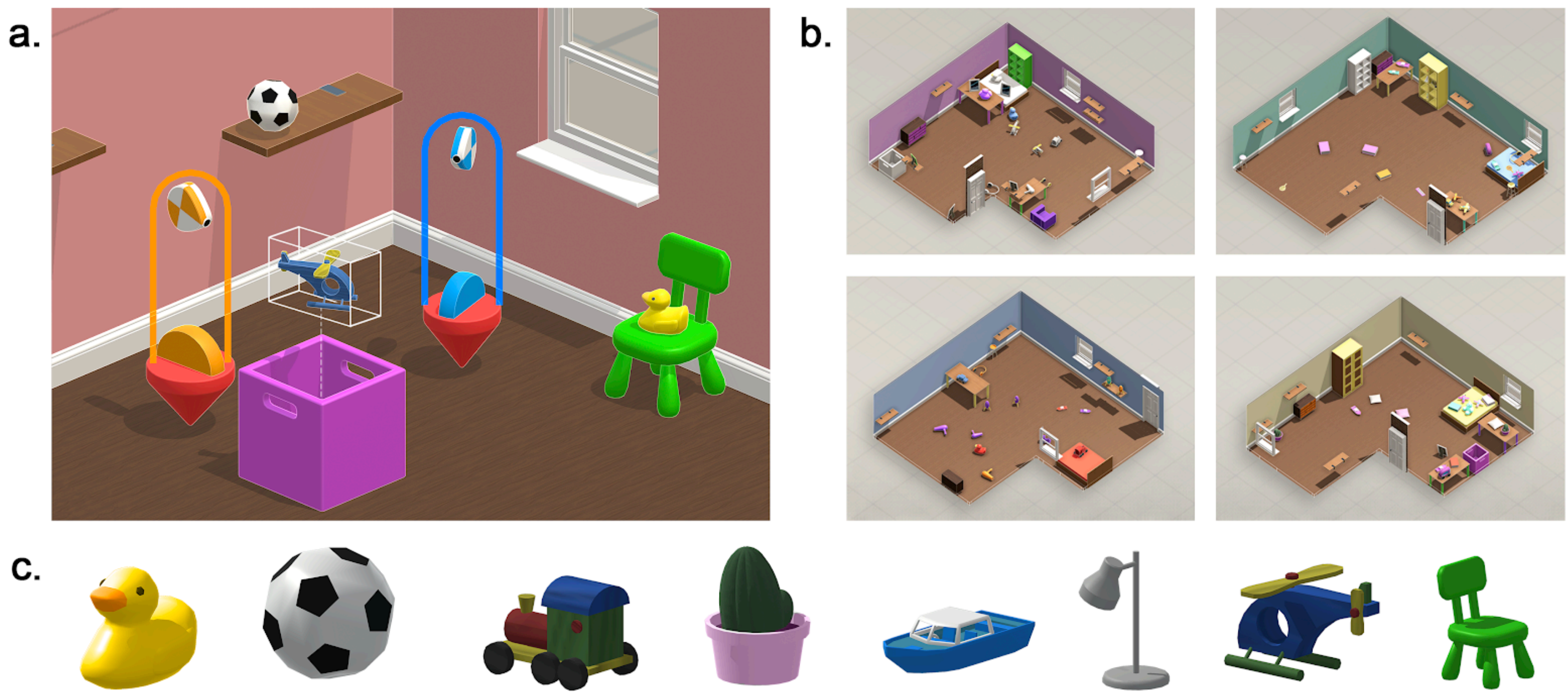
- In AlphaGo, dataset of human play was later replaced
    - Self-play in a win-lose game is a good curriculum
    - For general cooperative interactions, don't have win-lose or a dataset



$$\mathbb{E}_{s,a\sim\mathcal{D}}[\log \pi_\theta(a \mid s)]$$

Strategy: Create Dataset of Interactions

Playroom Virtual Environment



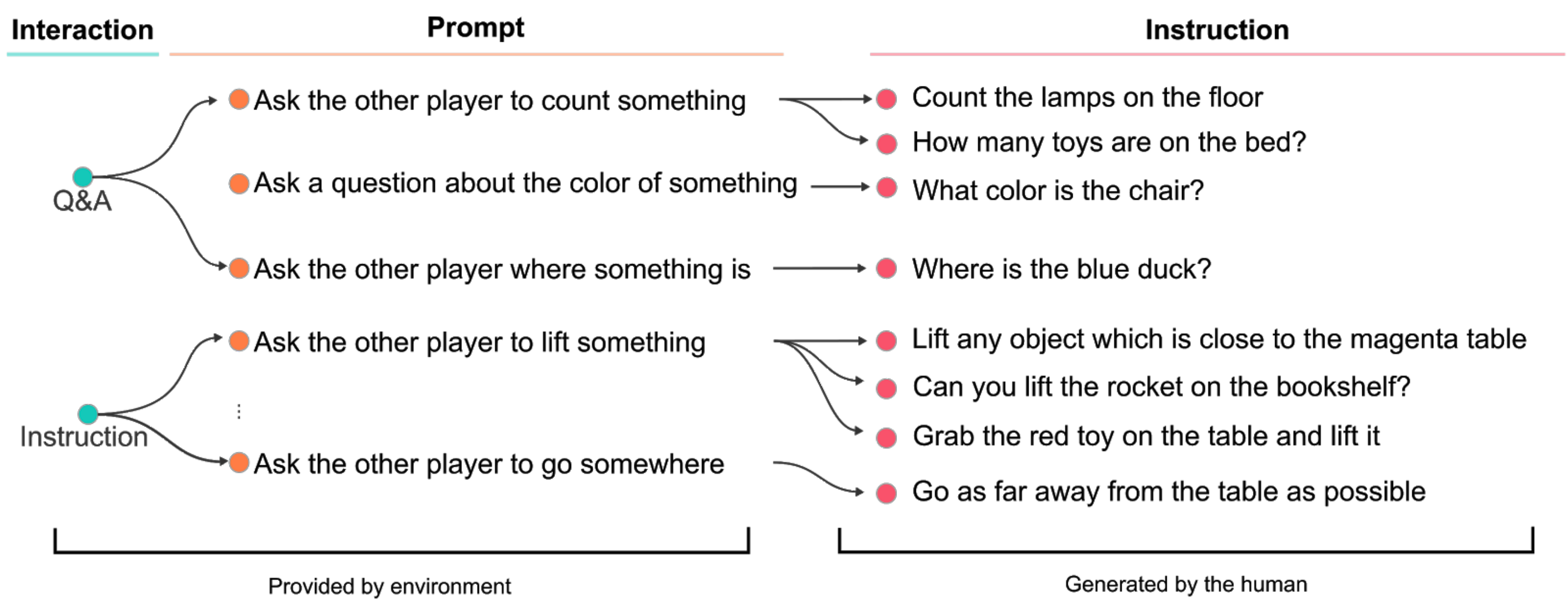
Eliciting Diverse Interactions

Setter: What color are the shelves in the corner?

Solver: They are red.

Prompt:
Ask a question about the color of an object.

From Prompts to Instructions



Prompt	Full text
go	Ask the other player to go somewhere
lift	Ask the other player to lift something
position object	Ask the other player to position something relative to something else
position yourself	Ask the other player to stand in some position relative to you
bring me	Ask the other player to bring you one or more objects
touch	Ask the other player to touch an object using another object
push object	Ask the other player to push an object around using another object
make a row	Ask the other player to put three or more specific objects in a row
arrange	Ask the other player to move a group of objects into a simple arrangement
put on top	Ask the other player to put something on top of something else
put underneath	Ask the other player to put something underneath something else
freestyle activity	Ask the other player to perform an activity of your choice
say what you see	Ask the other player to say what they are looking at or noticing right now
question about colour	Ask a question about the colour of something
question about existence	Ask the other player whether a particular thing exists in the room
describe location	Ask the other player to describe where something is
count	Ask the other player to count something

Table 4: Prompts used in language games.

Modifier	Full text
refer to objects by colour	Try to refer to objects by colour
refer to location by colour	Try to refer to the location by colour
use shape words	Try to use shape words like: circular, rectangular, round, pointy, long
refer to objects by location	Try to refer to objects by location
use proximity words	Try to use words like: near, far, close to, next to
use horizontal position words	Try to use words like: in front, behind, left of, right of, between
use vertical position words	Try to use words like: on top, beneath, above, below
use negation words	Try to use words like: not, isn't
use quantifier words	Try to use words like: some, all, most, many, none
not bed, door, or window	Do not use the words: bed, door, window

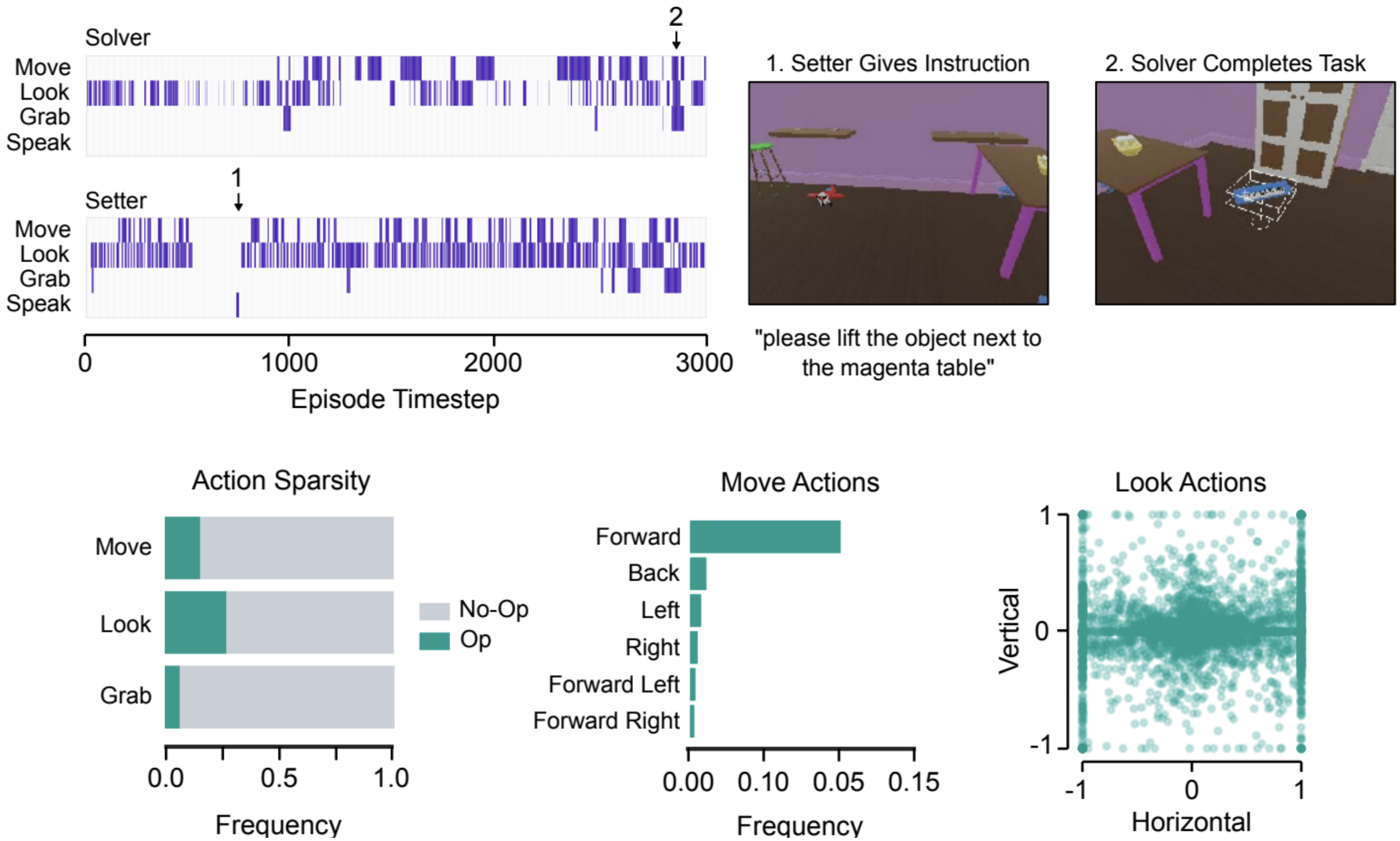
Table 5: Modifiers used in language games

24 base prompts, ~10 modifiers

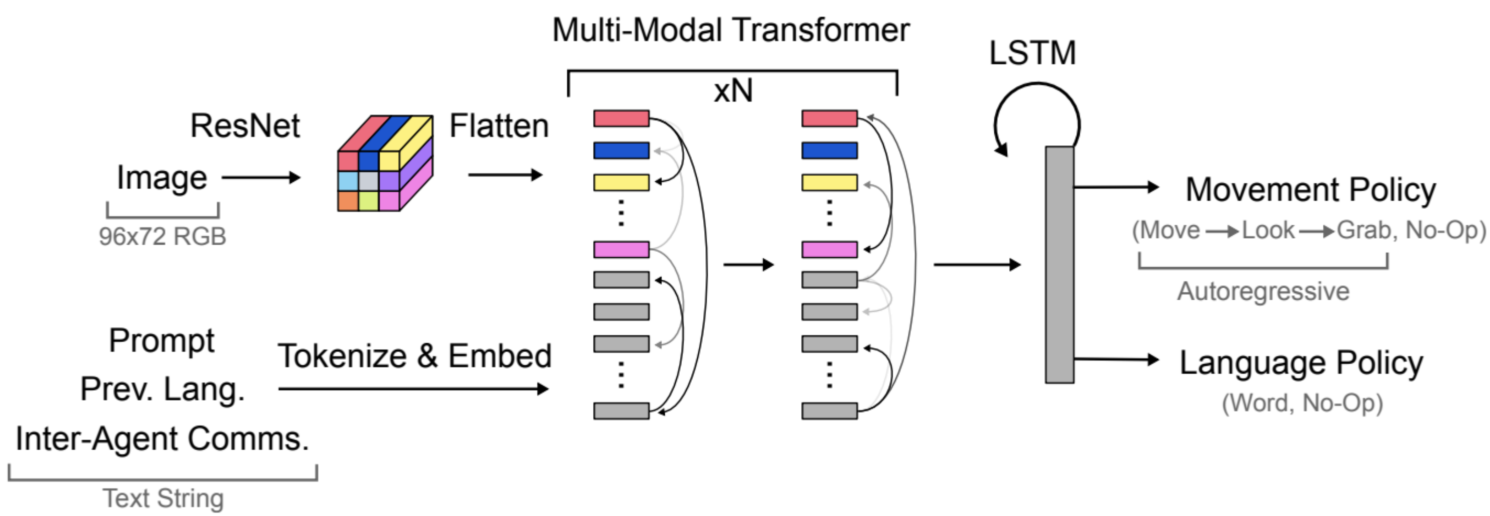
More than one year of video data
610,608 episodes
320,144 unique setter instructions of length 7.5 +/- 2 words

Human-Human Interaction
(dataset example)

Recorded Data

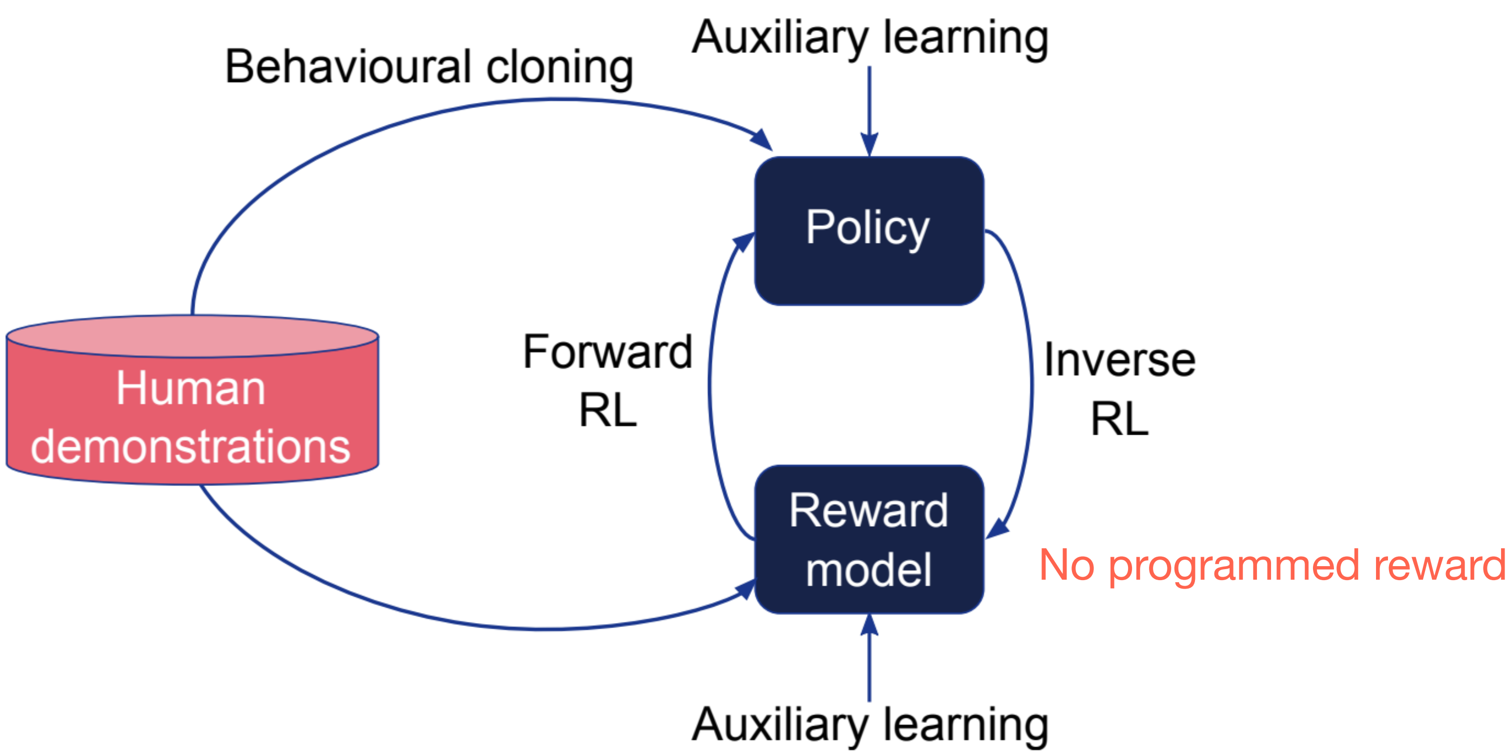


Agent Model



$$\mathbb{E}_{s,a\sim\mathcal{D}}[\log \pi_\theta(a \mid s)]$$

Learning from Data



Weaknesses of Behavioral Cloning

- Does not utilize environment interaction to learn how to respond to unusual contingencies
- Provides a relatively weak signal to train perceptual similarity

Seen this $$\pi_\theta(a_t \mid o_{\leq t})$$

What to do now? $$\pi_\theta(a_t \mid o_{\mathrm{novel}, \leq t})$$

Are novel observations similar to previously seen ones?

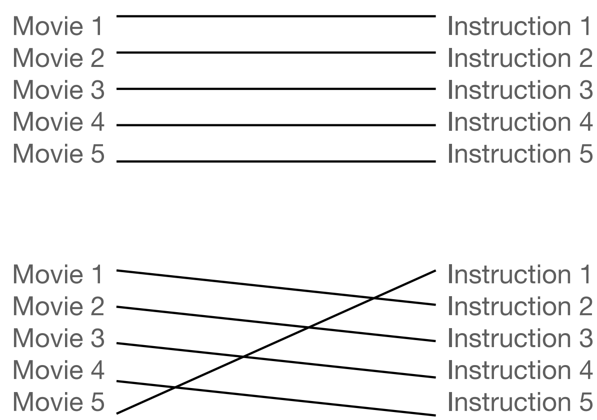
How to act?

Use Language to Instruct Similarity

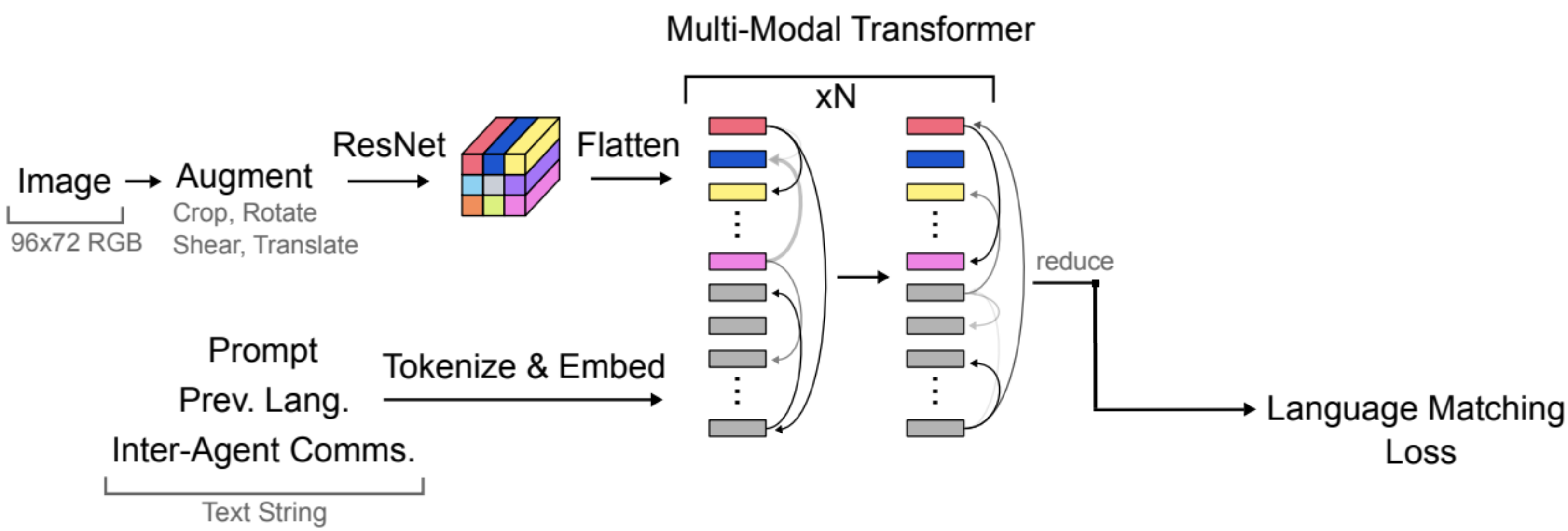
Consider two visual movies similar if it is not possible to distinguish their instructions.

$D(\text{movie}, \text{instruction})$

Classify plausibility of
movie - instruction pair
from real versus shuffled dataset



Language Matching Objective



Weaknesses of Behavioral Cloning (2)

Goals can be more compact than policies.

Consider: robot designed to climb Mount Everest.

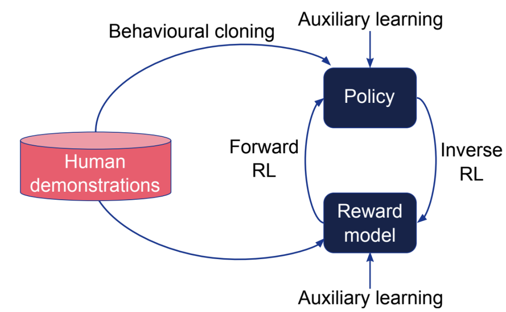
Policy is arguably very complicated. Must prescribe what to do in each scenario.

But the goal is simple: maximize altitude.

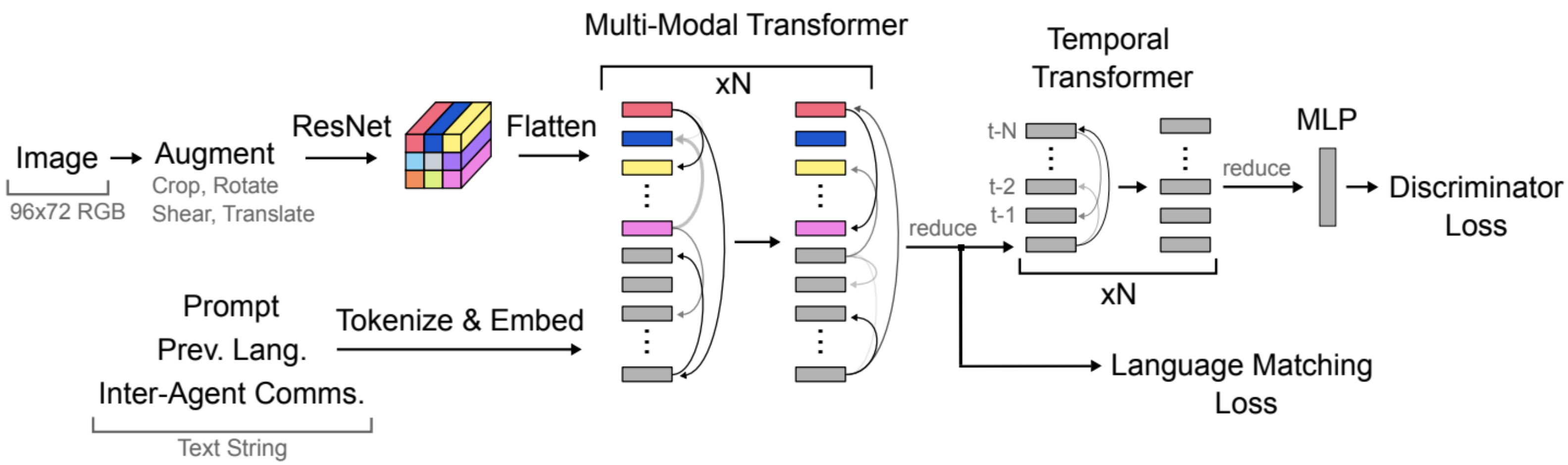
If the goal is known and success is measurable, then it is possible to practice with goal to acquire the policy.

Learning a Reward Model

version of GAIL (Ho and Ermon, 2016)

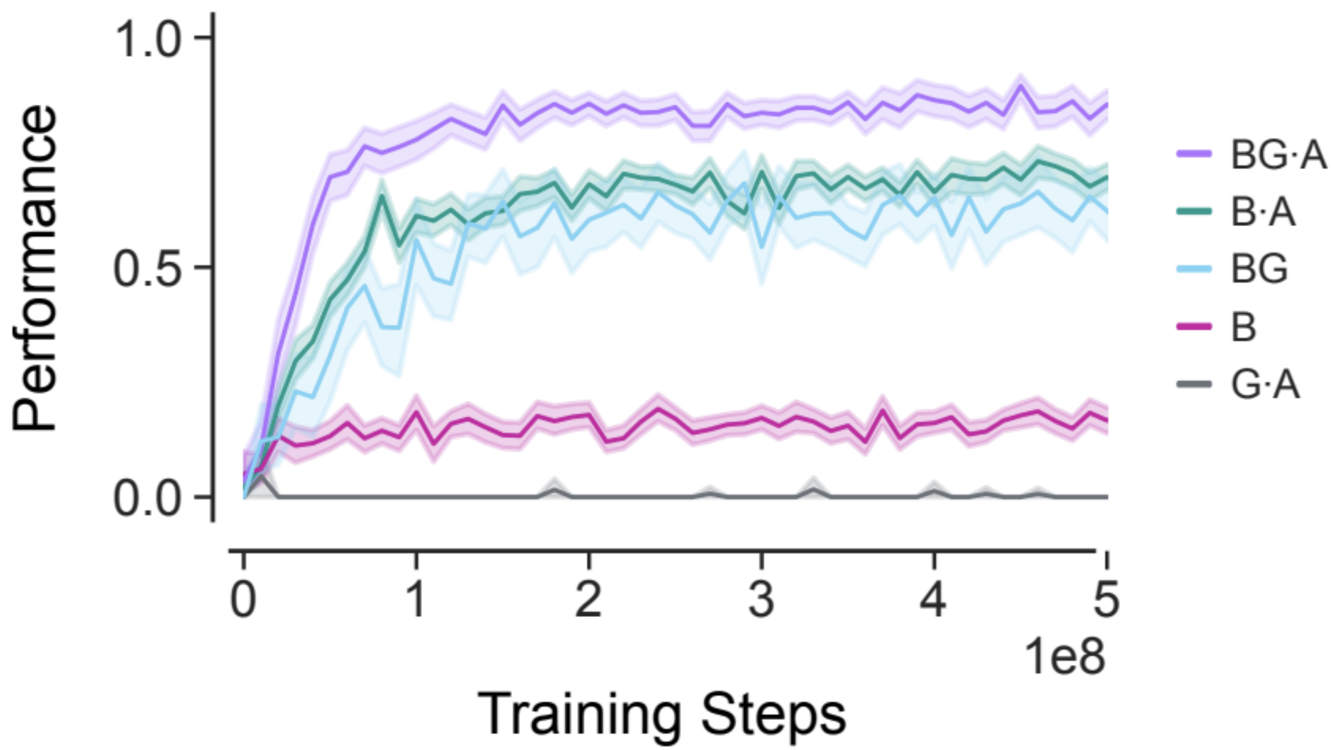


Discriminate between agent and human behavior using features from language matching.



Comparing Contributions on a Simple Task

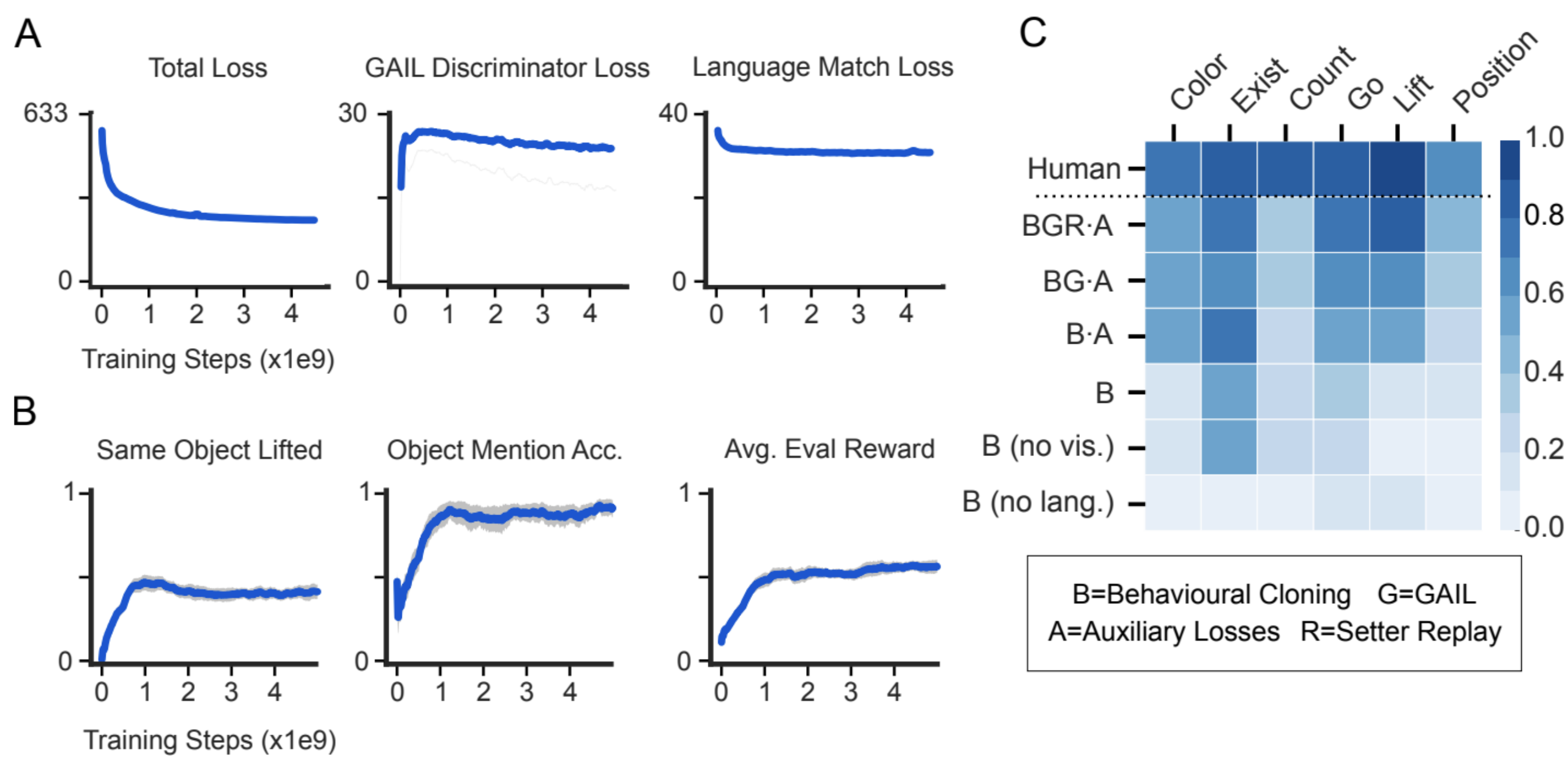
“Put X on Bed”



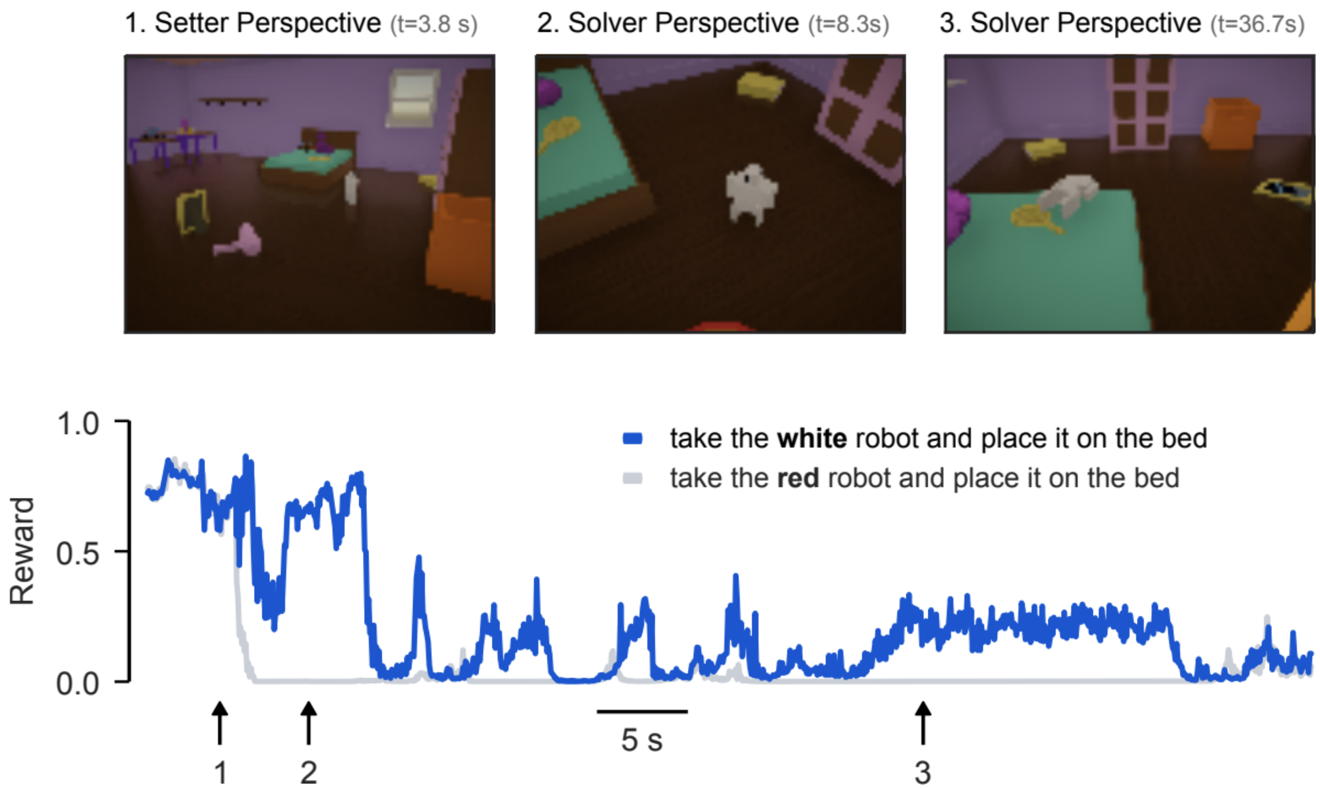
Interactive Training

Evaluation

From code to human interaction

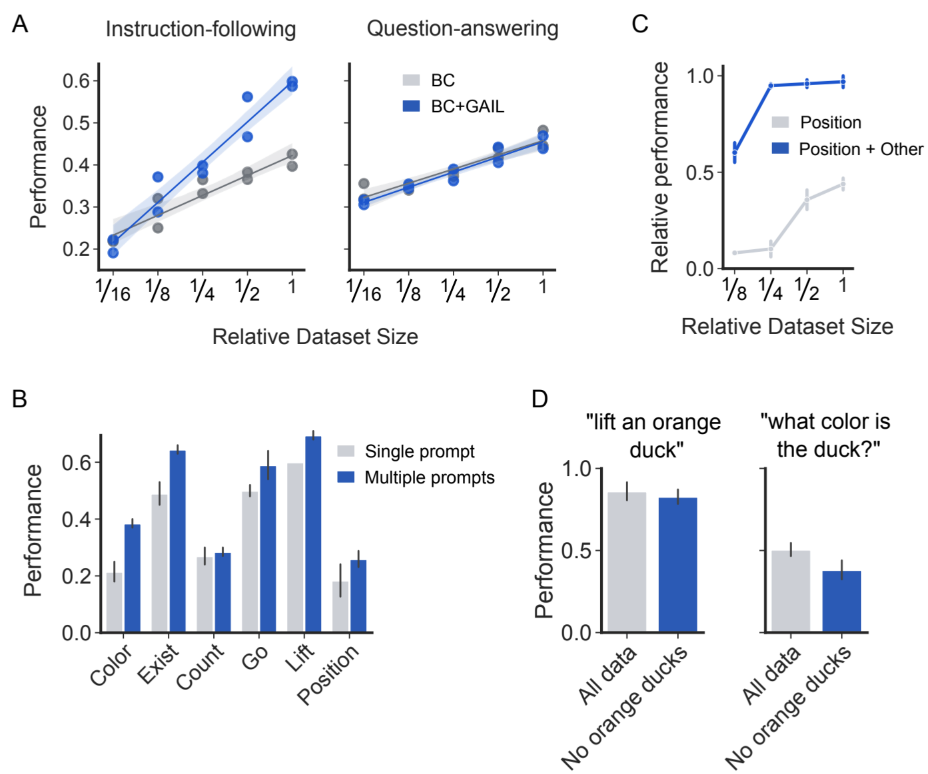


Inspecting Reward Model

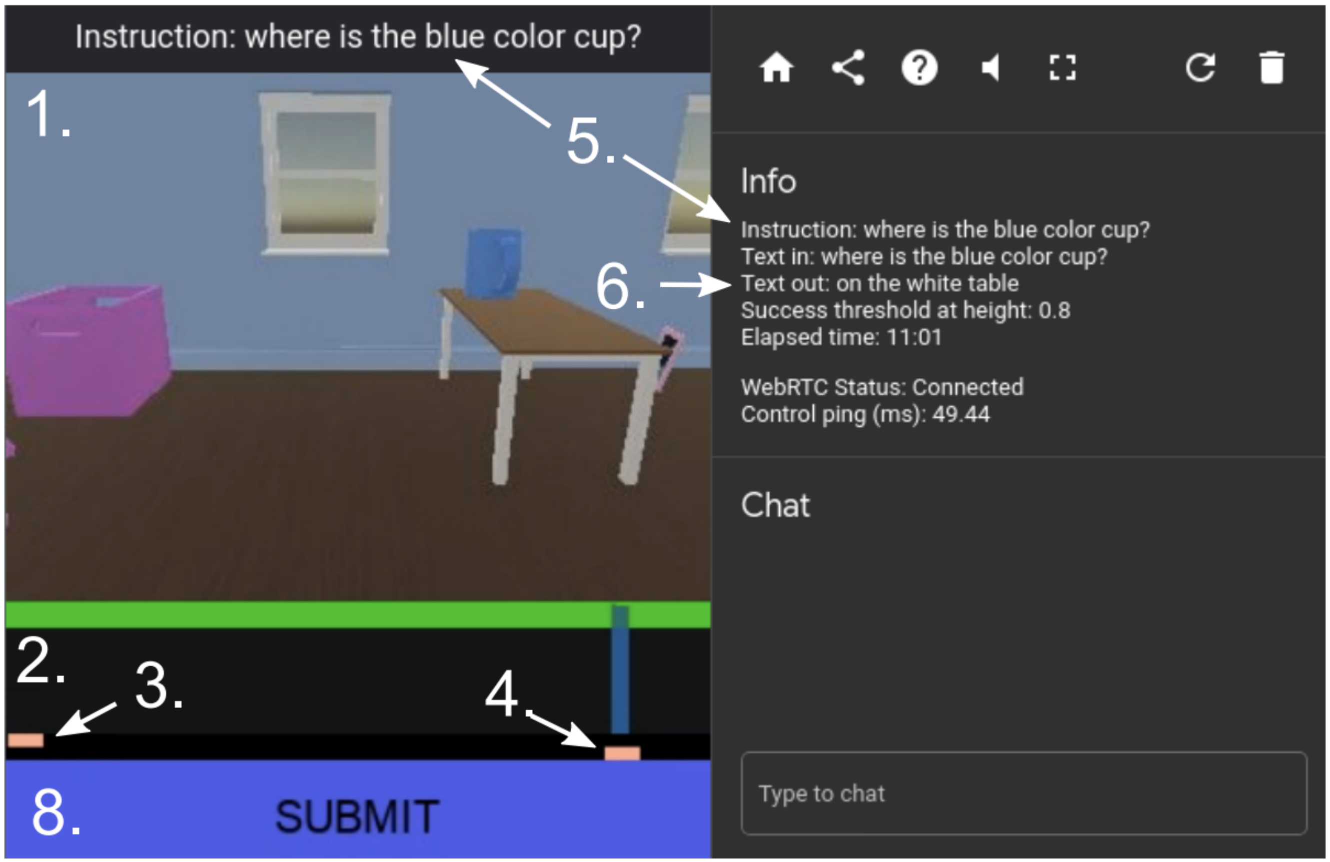


Scaling and Transfer Performance

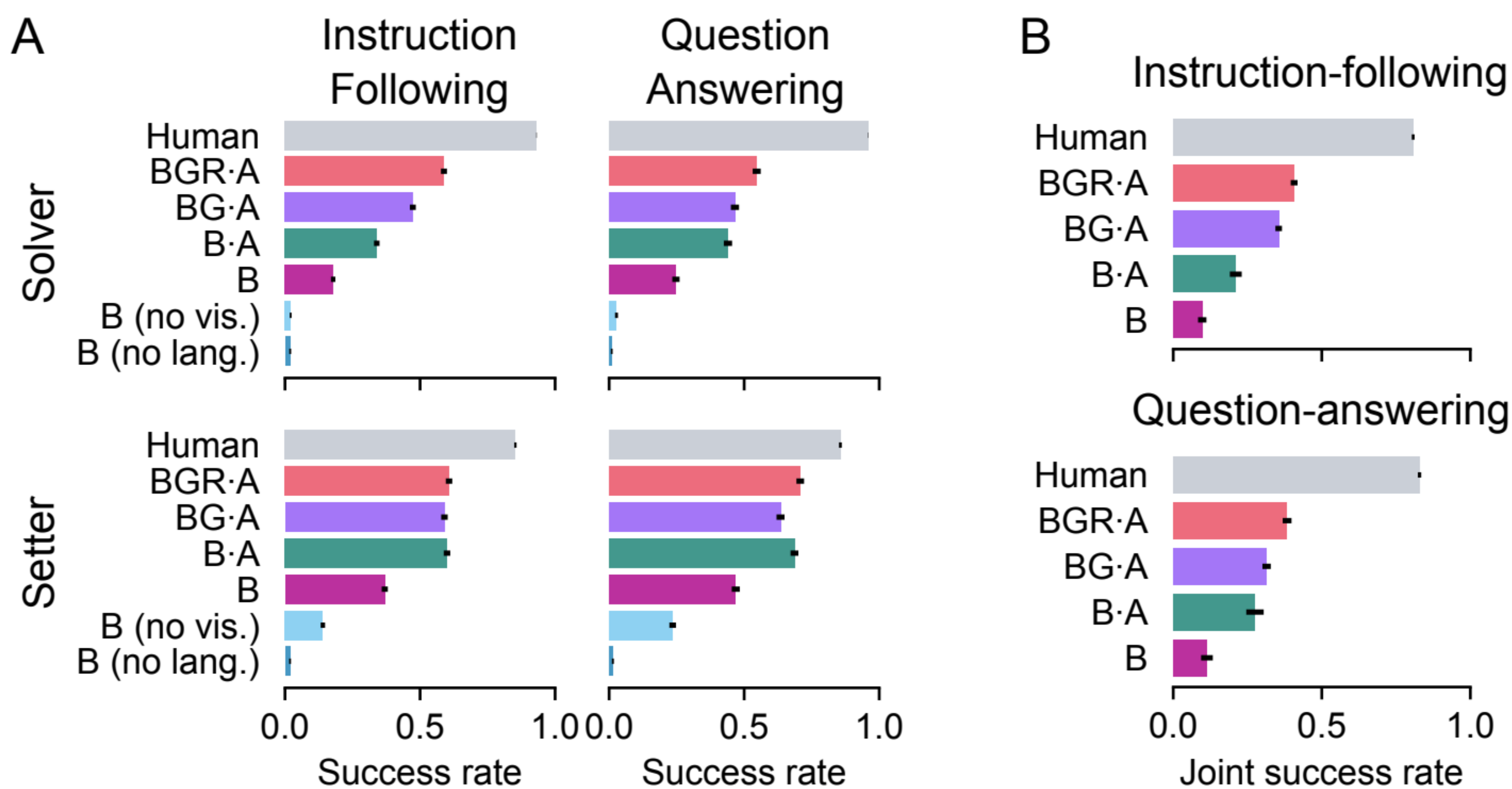
See Scaling Laws for Neural Language Models (Kaplan et al., 2019)



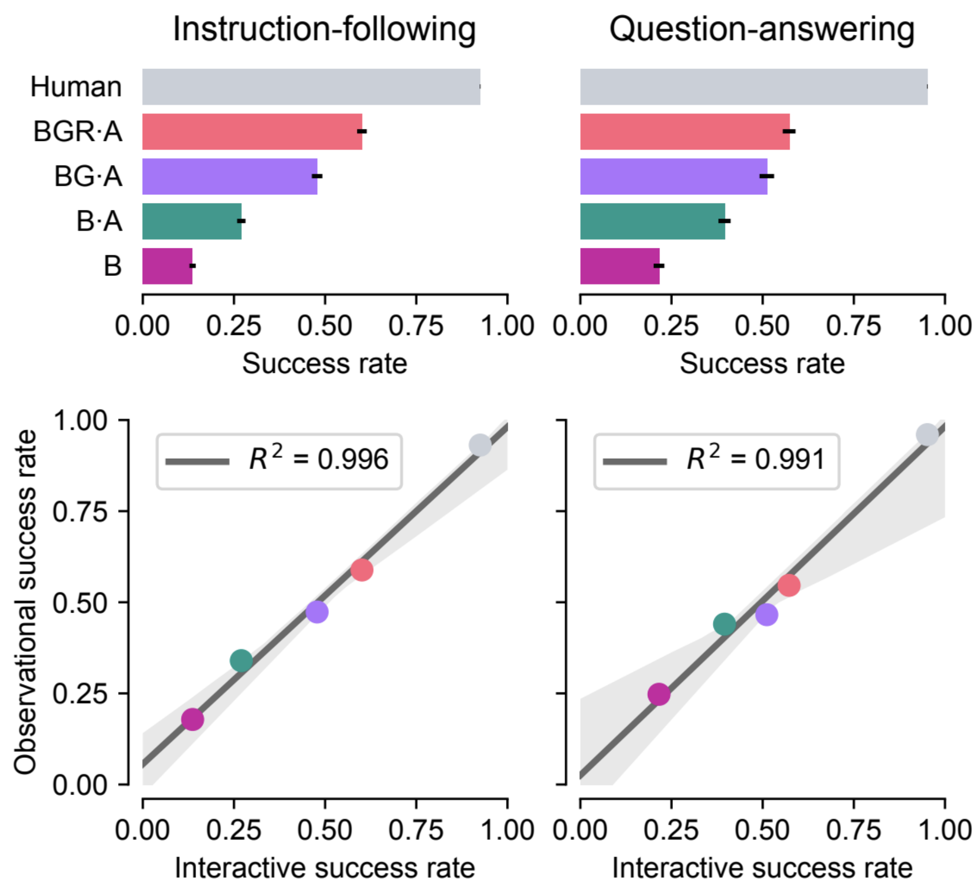
Human Evaluation Techniques

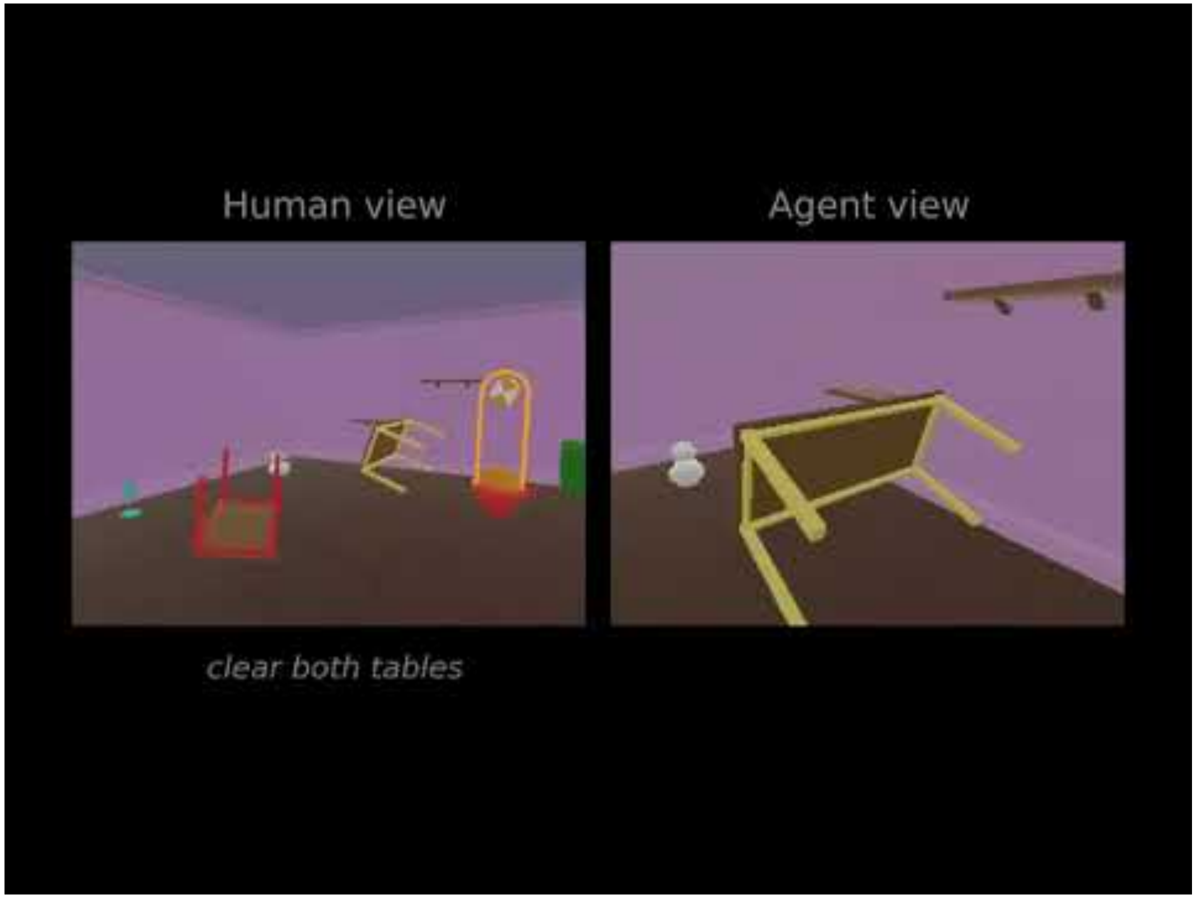


Human Evaluation (Observational)



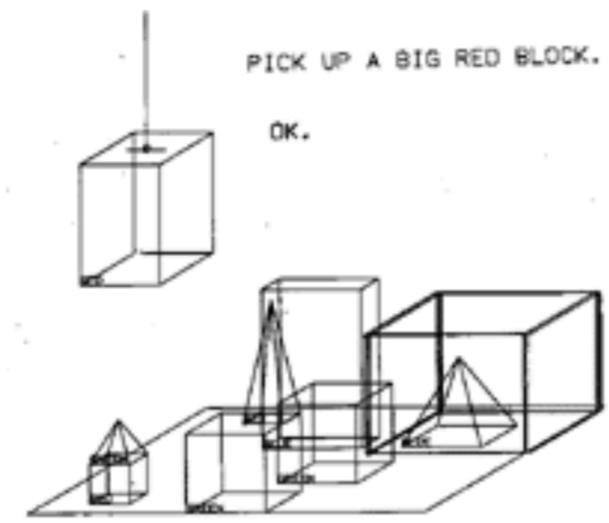
Human Evaluation (Interactive)



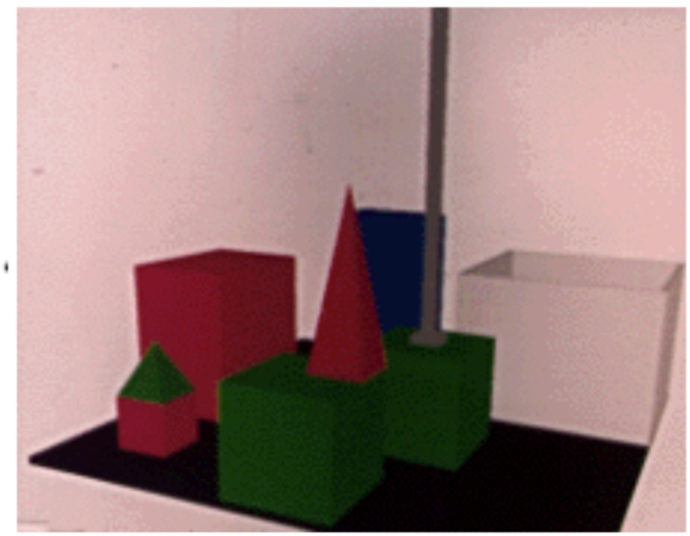

Human view
Agent view
clear both tables

A Small Callback to the Programmer's Apprentice

SHRDLU (Winograd, 1968)



Original screen display



Later color rendering (Univ. of Utah)

Human-Computer Interaction

Computers are being used today to take over many of our jobs. They can perform millions of calculations in a second, handle mountains of data, and perform routine office work much more efficiently and accurately than humans. But when it comes to telling them what to do, they are tyrants. They insist on being spoken to in special computer languages, and act as though they can't even understand a simple English sentence.

Let us envision a new way of using computers so they can take instructions in a way suited to their jobs. We will talk to them just as we talk to a research assistant, librarian, or secretary, and they will carry out our commands and provide us with the information we ask for. If our instructions aren't clear enough, they will ask for more information before they do what we want, and this dialog will all be in English.

Winograd, 1971

Discussion