

A **North Star** for AI Research in the era of **AGI** and **Frontier Labs**

Joseph E. Gonzalez
Professor in EECS @ UC Berkeley
jegonzal@berkeley.edu



AGI ✨

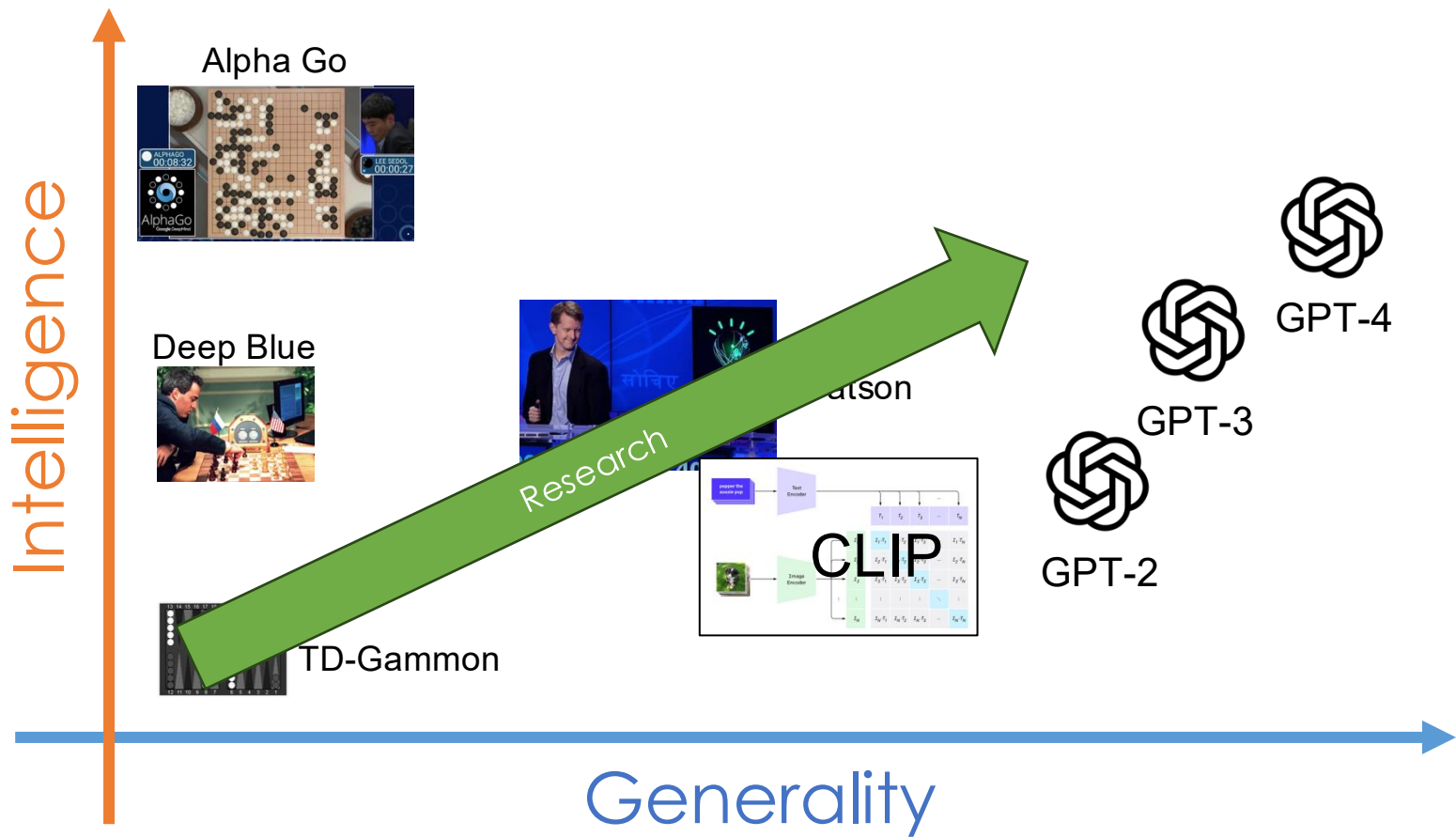


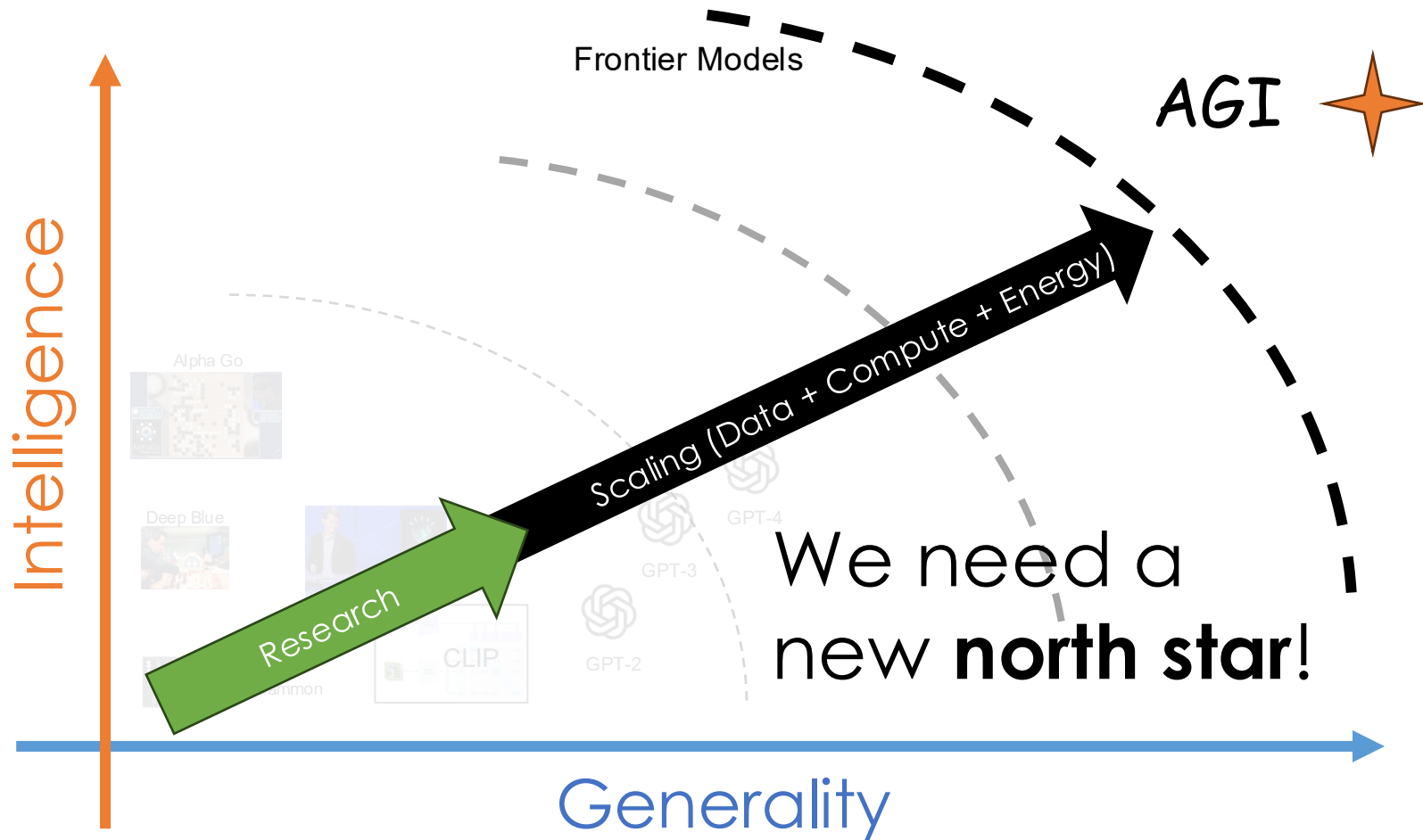
AGI ✨



AGI







About me:

Joseph (Joey) Gonzalez

jegonzal@berkeley.edu

Professor at UC Berkeley



Research groups and startups:



LMSYSORG

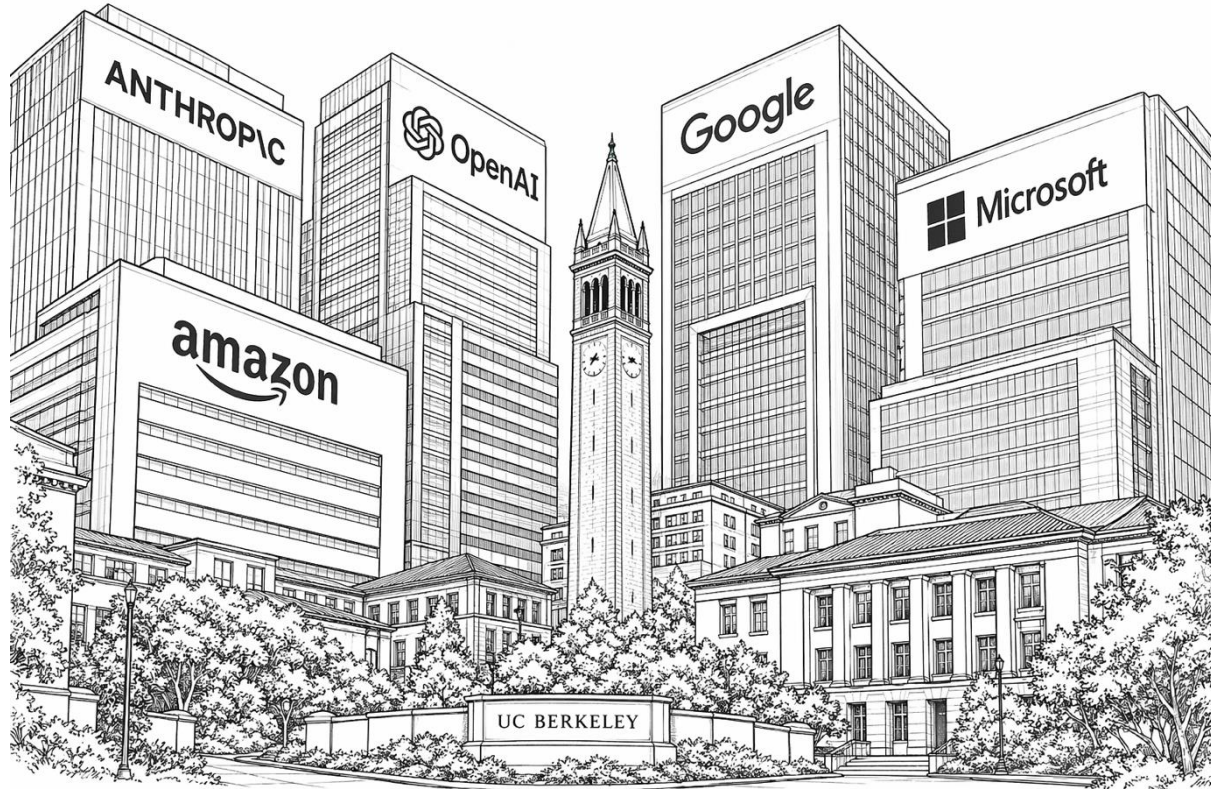


inferact

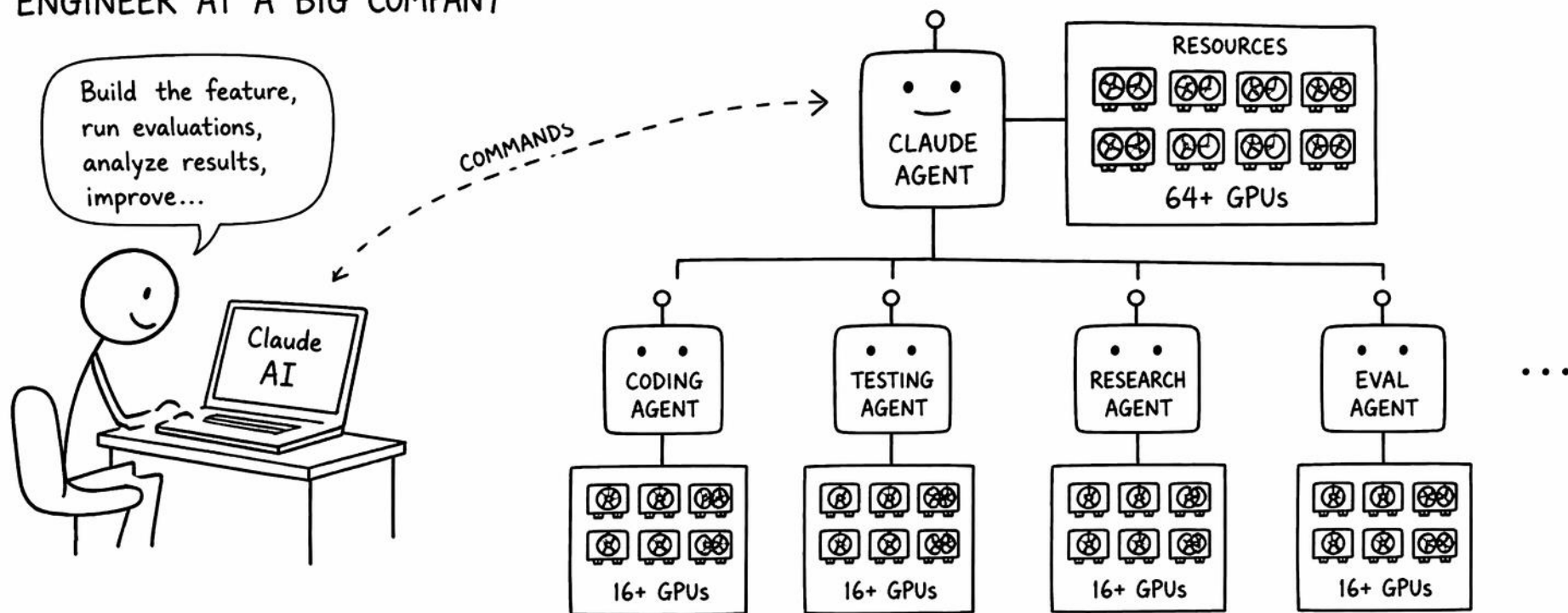


Letta

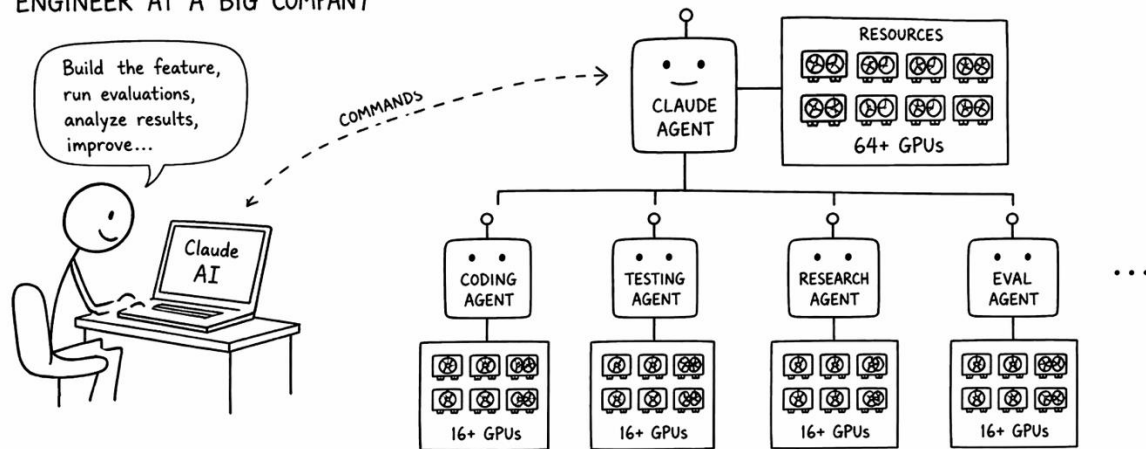
Academic anxiety in the Era of Frontier Labs



ENGINEER AT A BIG COMPANY



ENGINEER AT A BIG COMPANY



MY GRAD STUDENT



What is the **Next Frontier**
of AI Research



How do we
Collaborate with **Industry**



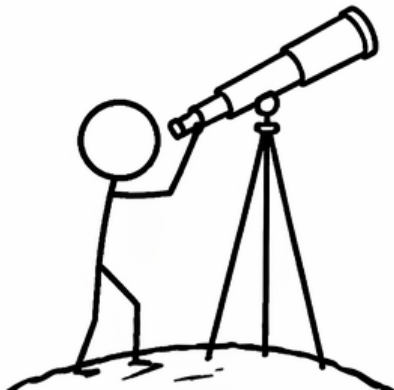
Both questions have a **common answer**.

How do we look **beyond the horizon?**



How do we look **beyond the horizon?**

*If we solve today's
hard problems, what
new problems do we create?*



My Story

*If we solve today's
hard problems, what
new problems do we create?*



2008: Everyone is designing models; I focused on scaling training.

→ AI-Systems

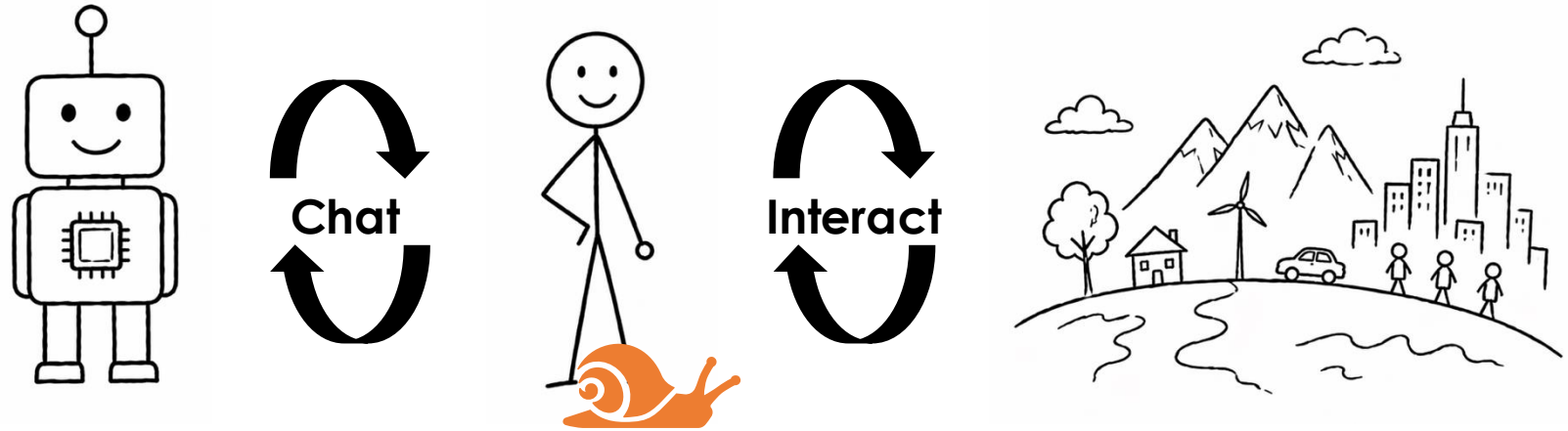
2015: Everyone is scaling training; I focused on inference.

→ Inference Systems

How do we look beyond AGI?

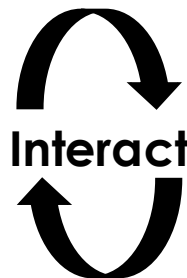
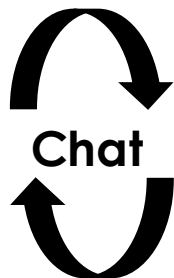
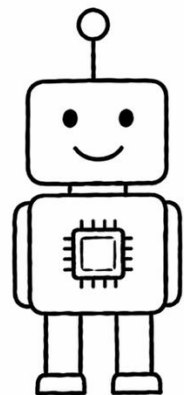
We look to the past ...

Generative AI in 2023 (“pre-agents”)



Humans became the bottleneck.

Agentic AI in 2025



Projects in my group that
addressed this problem



Gorilla

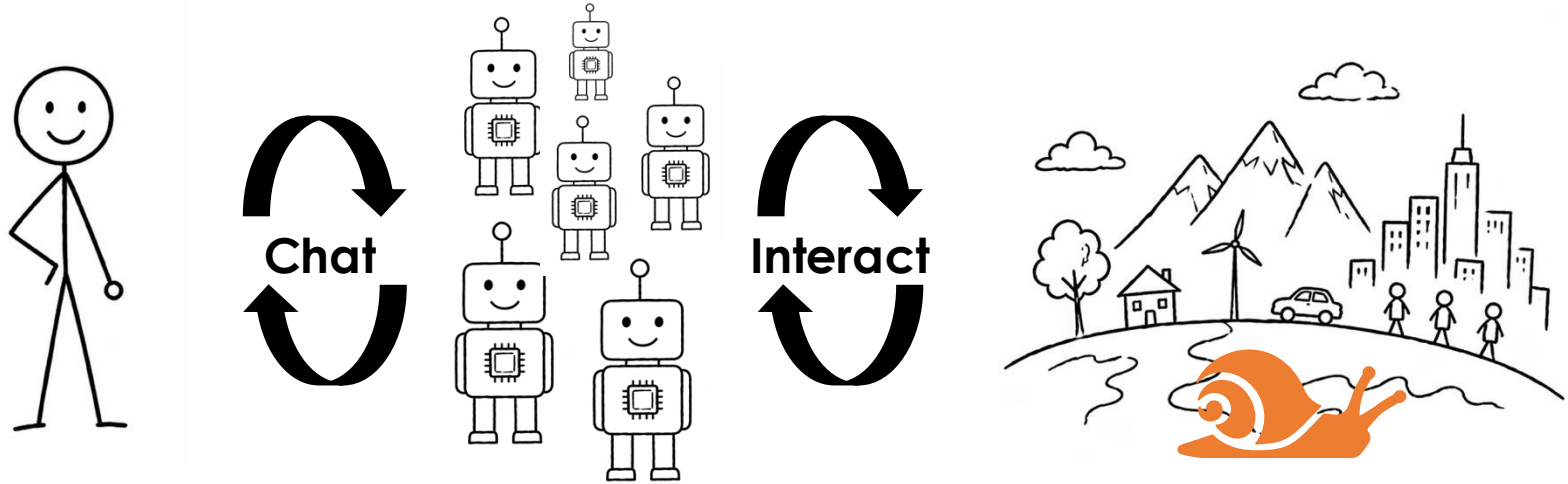
**Discovering
And Using
Tools**



memGPT

**Giving Agents
Memory**

Agentic AI in **2026**



The **world** (software-systems) is the new **bottleneck**!

How do we build **software systems** that better serve our **AI Overlords (agents)**?

AI Users are:

- 1000x Faster
- Behave differently
- Relentless + Helpful

How will we change

- Abstractions
- System Design

Supporting Our AI Overlords: Redesigning Data Systems to be Agent-First

Shu Liu, Soujanya Ponnappalli, Shreya Shankar, Sepanta Zeighami, Alan Zhu
Shubham Agarwal, Ruiqi Chen, Samion Suwito, Shuo Yuan, Ion Stoica, Matei Zaharia
Alvin Cheung, Natacha Crooks, Joseph E. Gonzalez, Aditya G. Parameswaran
University of California, Berkeley

Abstract

Large Language Model (LLM) agents, acting on their users' behalf to manipulate and analyze data, are likely to become the dominant workload for data systems in the future. When working with data, agents employ a high-throughput process of exploration and solution formulation for the given task, one we call *agentic speculation*. The sheer volume and inefficiencies of agentic speculation can pose challenges for present-day data systems. We argue that data systems need to adapt to more natively support agentic workloads. We take advantage of the characteristics of agentic speculation that we identify, i.e., scale, heterogeneity, redundancy, and steerability—to outline a number of new research opportunities for a new agent-first data systems architecture, ranging from new query interfaces, to new query processing techniques, to new agentic memory stores.

1 Introduction

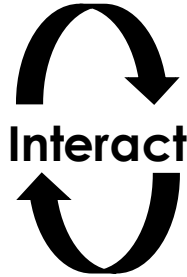
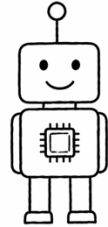
Powered by Large Language Models (LLMs) that can reason, invoke tools, author code, and communicate with each other, we are on the precipice of a new agentic revolution that will transform how data systems are used. Modern LLMs are far more *efficient* internally, matching the capabilities of those orders of magnitude larger just a year ago, and growing ever more *effective* at understanding and manipulating both structured and unstructured data. As they become both cheap and capable, future LLM agents will act on users' behalf

of LLM agents tasked with finding reasons for why profits in coffee bean sales in Berkeley was low this year relative to last. Since they are not limited by human cognitive bandwidth and response times, an army of agents could employ an enormous volume of queries to data systems, far more than any human could—all for a single task. Many of these queries are likely wasteful, and are simply providing the agents grounding. As another example, if an LLM agent is tasked with identifying a new crew for a delayed flight, it would need to consider various hypothetical transactions to surface to a human decision maker, each with dozens of updates to various databases.² For such tasks, agents may explore many alternatives in parallel by forking database state, running speculative updates, and rolling back branches. Overall, as agentic workloads become more and more prevalent, the sheer scale and inefficiencies of agentic speculation will become the bottleneck, and our data systems will need to evolve in response.

So we ask the question: *how can data systems evolve to better support agentic workloads?* In particular, can data systems natively—and efficiently—support agentic speculation, helping LLM agents determine the best course of action? This question—which, as we argue, our community is well-equipped to answer—holds the key to unlocking unimaginable productivity gains from agents being the primary mechanism we use to interact with data.

Thankfully, while agentic speculation represents a new challenge

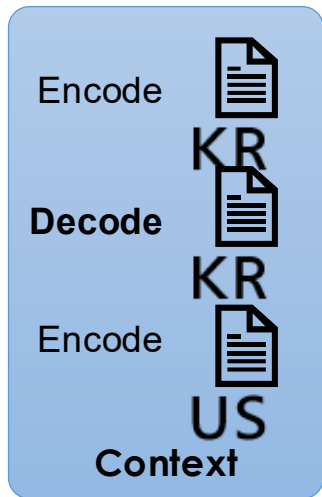
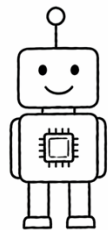
How do we make **Interaction more efficient?**



The Future of Tools

MCP → **Interactive
Programmatic
Tool
Calling**

Today: Standard Tool Calling



`get_doc(subject = "Summit")`



10K token doc

KR

Thinking: "Oh, I need to translate ..."

`translate(text = [10K token Korean doc])`



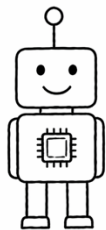
10K token translated doc

US



Large amounts of
**data pass through
model context**
between tool calls.

Interactive Programmatic Tool Calling (I-PTC)*



Encode



KR

Encode



US

Context

Time

```
import tools
doc = tools.get_doc(subject = "Summit")
print(doc[0:1000])
```



1K token doc

Thinking: "Oh, I need to translate ..."

```
doc = tools.translate(text = doc)
print(doc)
```



US

Use an **interactive**
programming env.
(Python Shell)



Enables programmatic
exploration and
re-use of
accumulated state

* New project in my lab.



STAR Lab

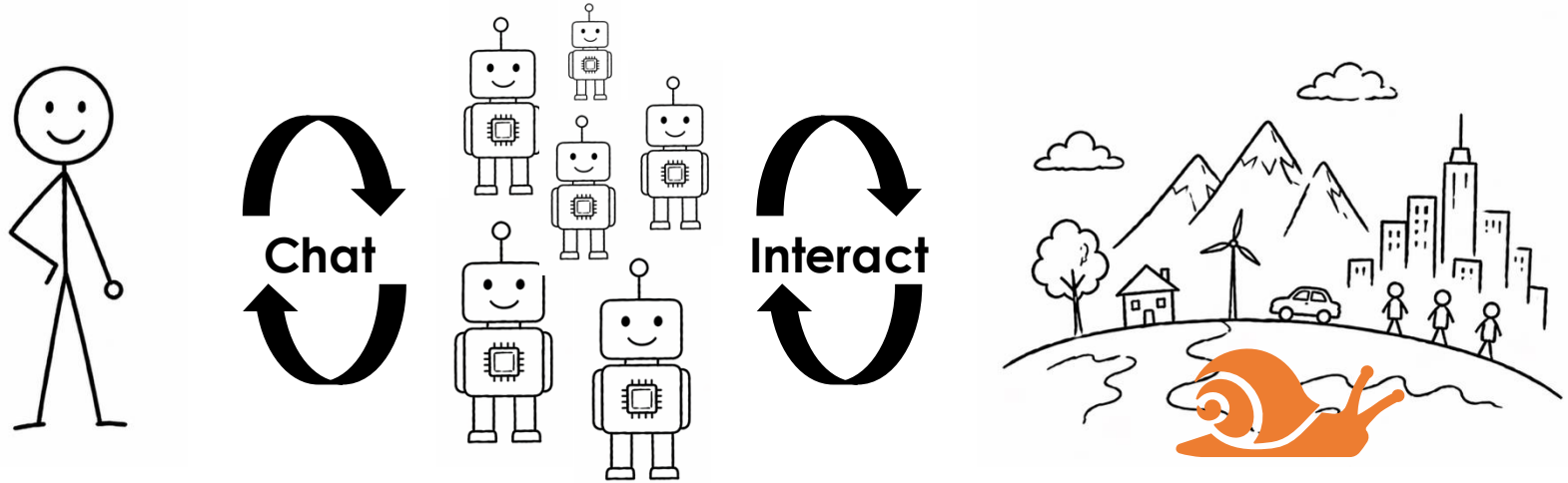
Systems Transformation with **AI Research**
Building Systems for AI with AI



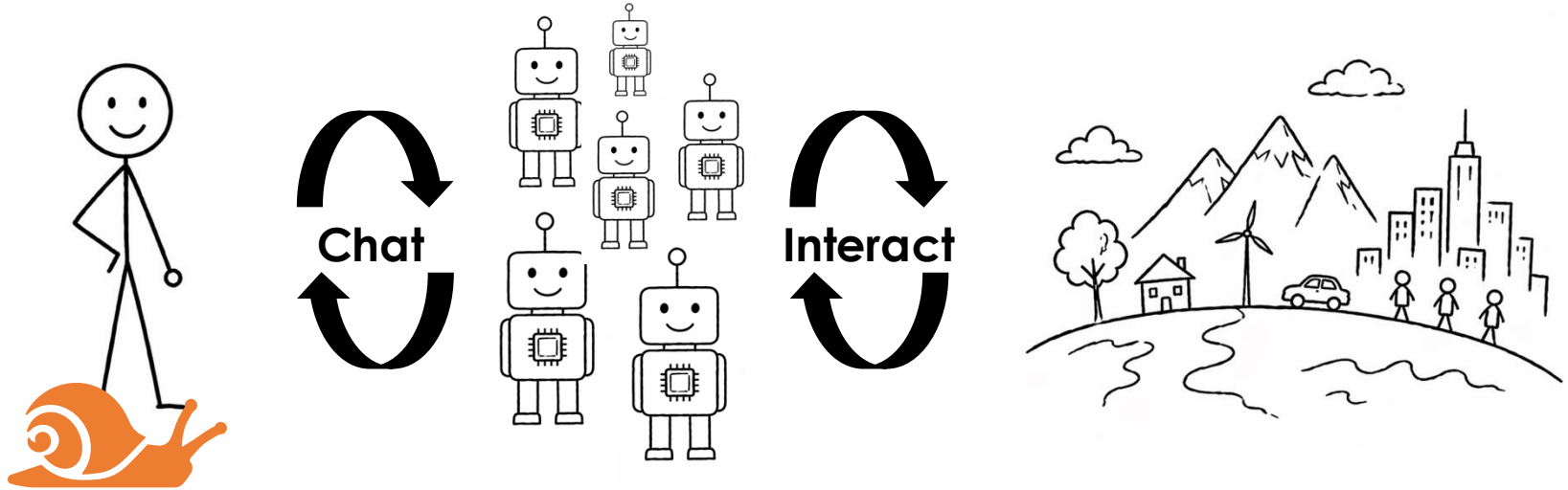
Redesign systems and
abstractions to
support AI users.

Use new AI capabilities to
rethink how we **build**
software systems.

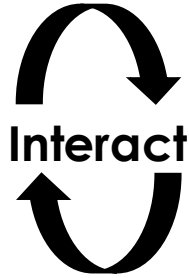
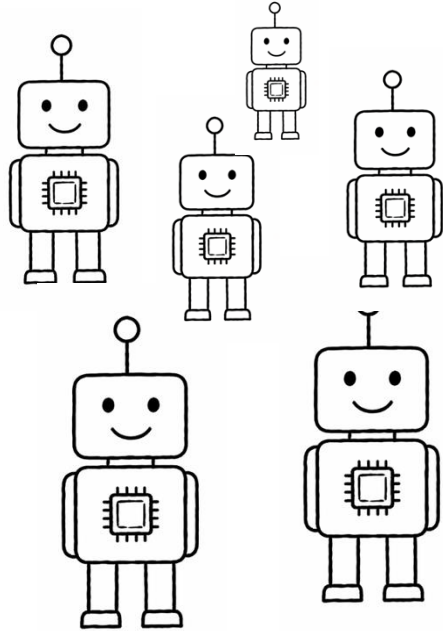
Agentic AI in **2027** and **beyond**



Agentic AI in **2027** and **beyond**

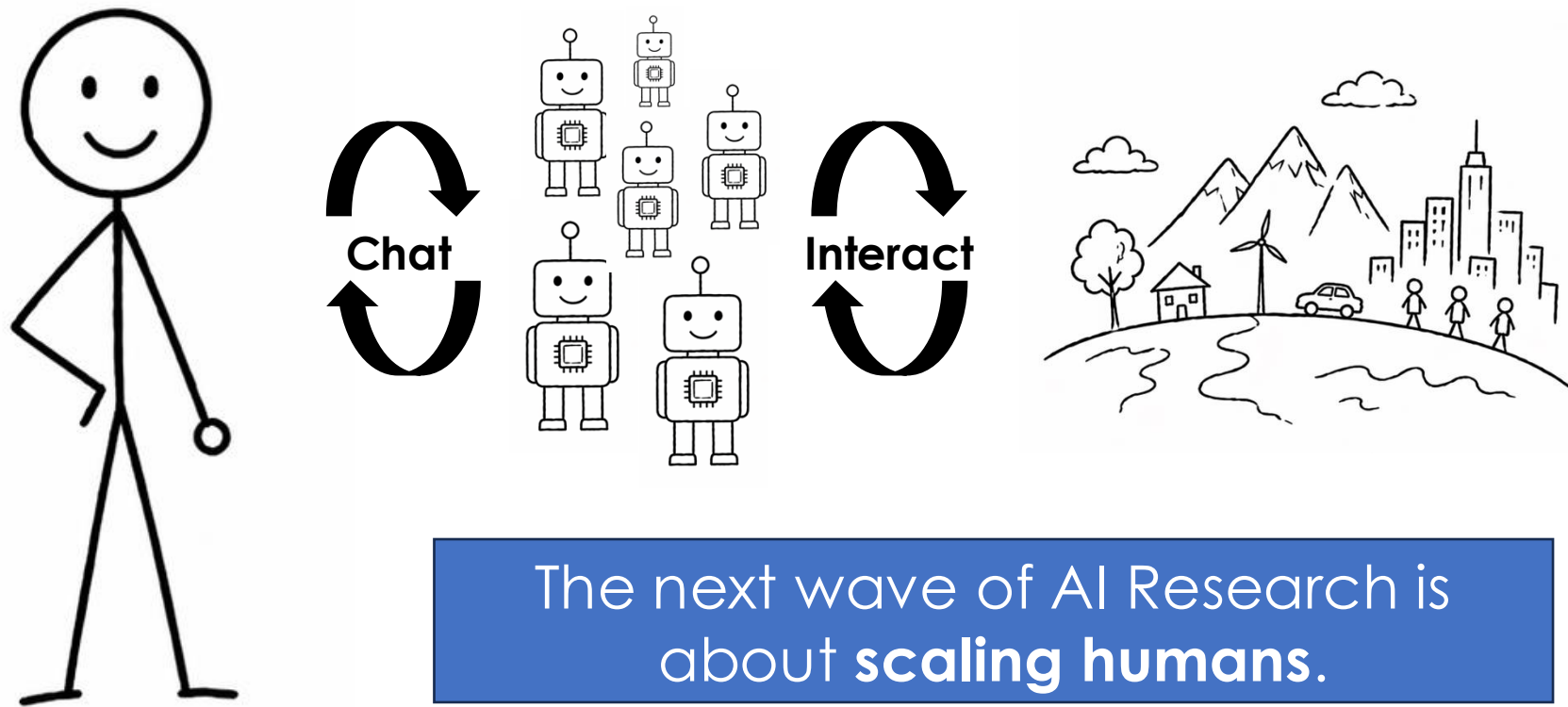


Agentic AI in **2027** and **beyond**



Removing humans
is **not** the solution.

Agentic AI in **2027** and **beyond**



Three Bets to Scaling Humans



Humans are **imperfect** at specification.

- **Advisor models:** learn **goals** through **RL**

Humans are **slow** at responding to requests.

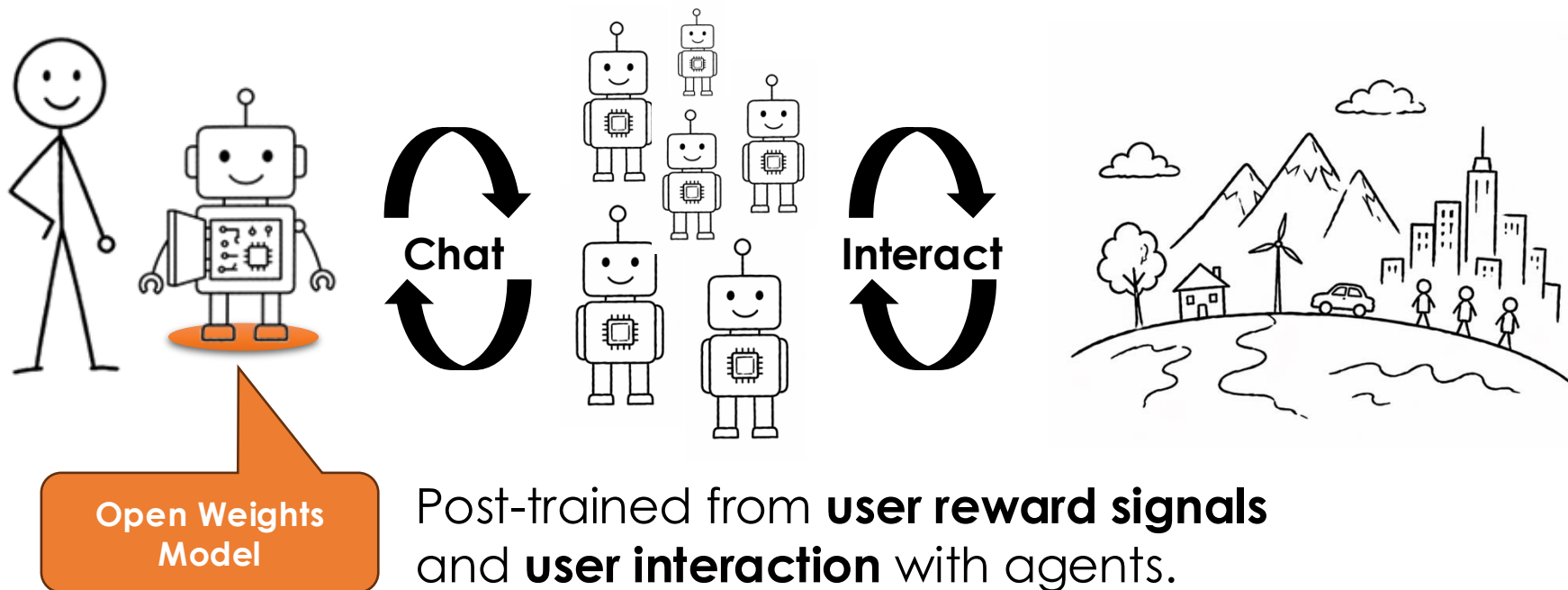
- **User mimics: thinking** and resp. like an **individual**

Humans have **misaligned** goals.

- **AI Mediated Agreement:** studying AI in the context of strategic negotiation

Advisors and Mimics

Adding Another Layer of AI



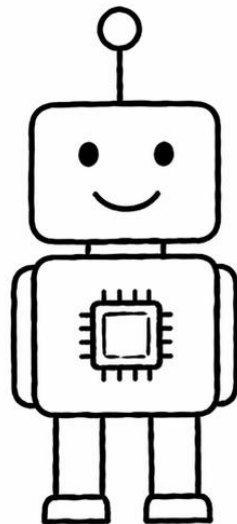
Academia + Industry

Find **innovative ways** to combine **small-scale individualized learning** with **large frontier models**.

Small models
an **academic**
group can train.



+

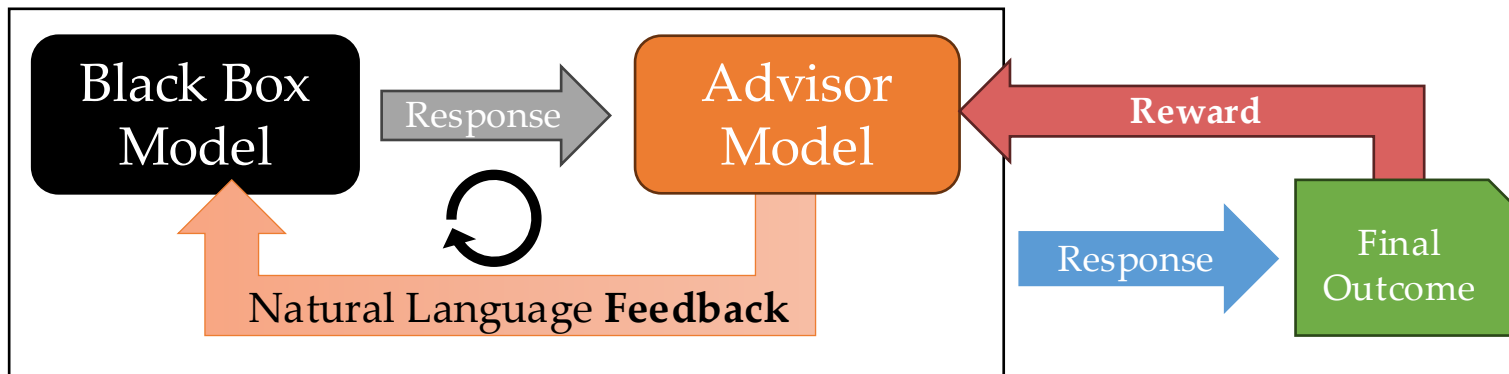


Industrial scale
model **capabilities**
we can't train.

Advisor Agent Training

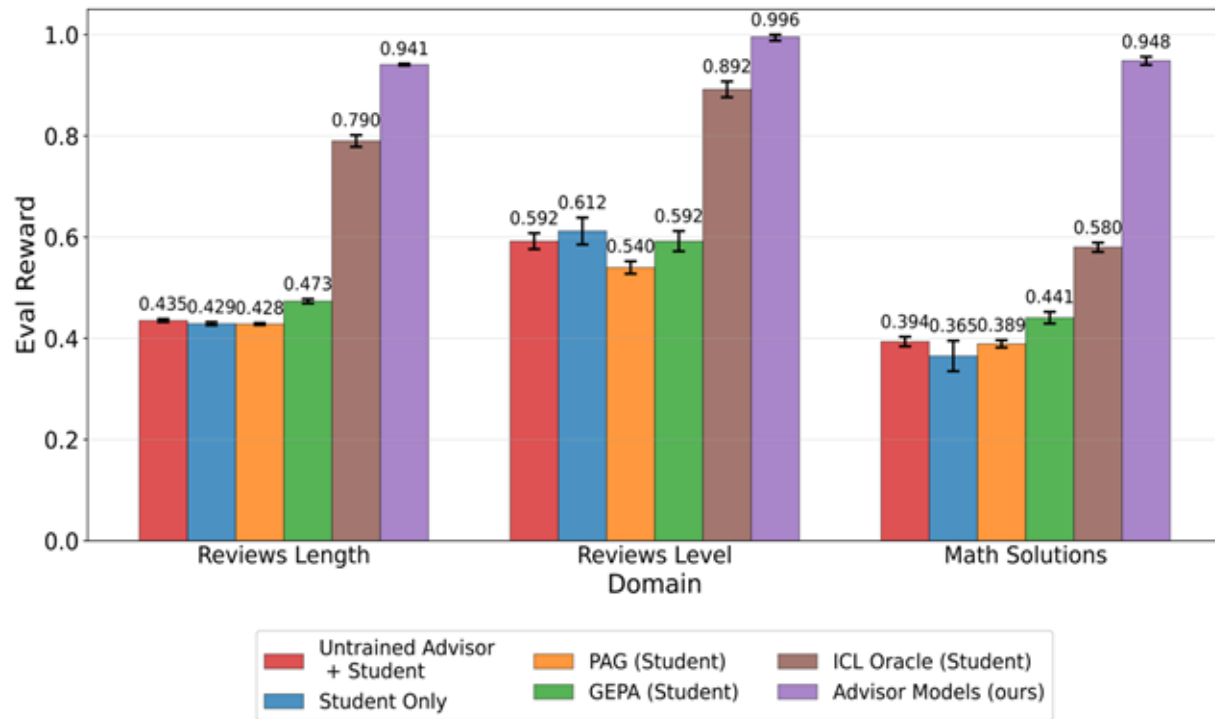


We train a small open-source “**advisor**” model to guide a larger black-box model via **reward feedback from users**.



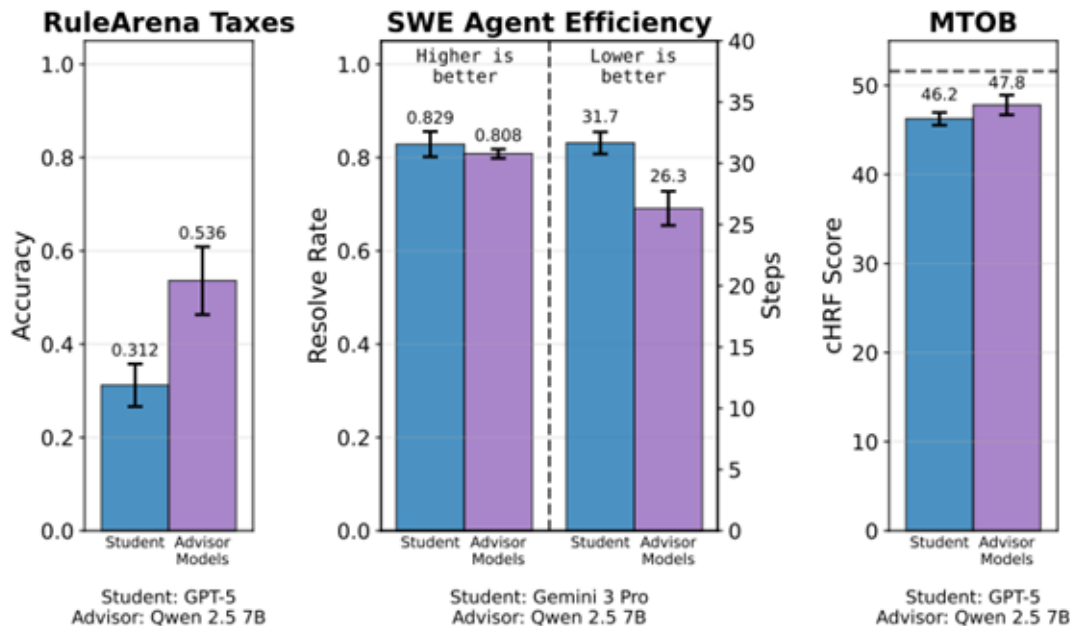
- Advisor model **guides black box models** to maximize reward.
- Preserve **general capabilities** of black box models.

Advisor's can learn to personalize



Advisor Models
outperform
SoTA techniques at
**extracting hidden user
preferences.**

Advisor Models Lift Frontier Model Capabilities

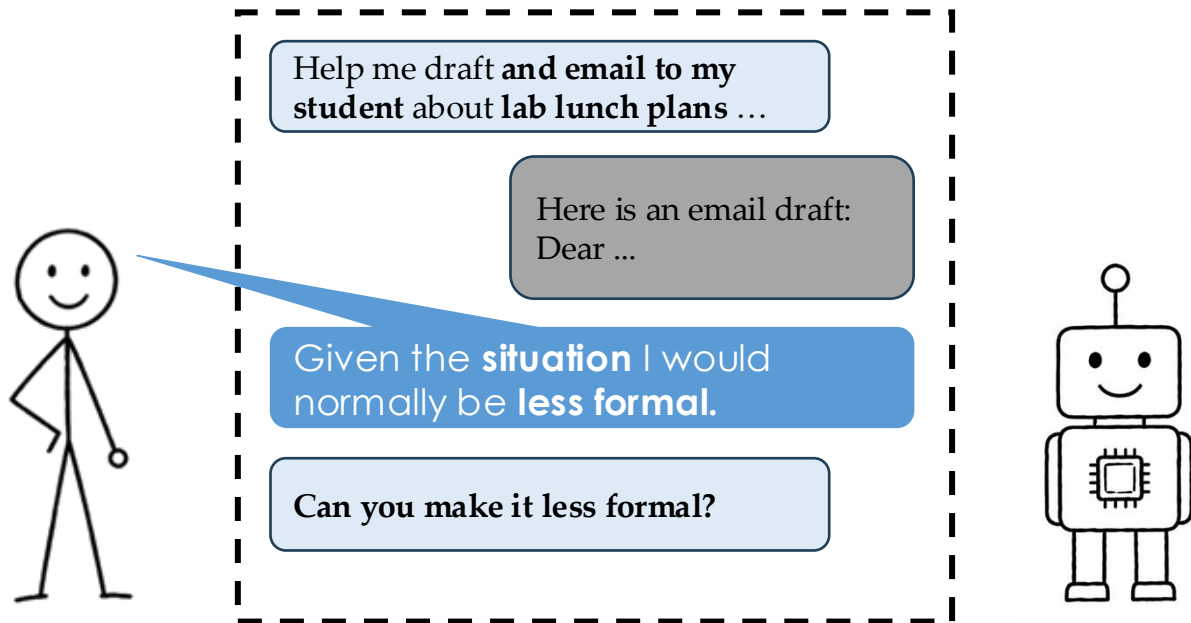


Advisor Models improves frontier model capabilities. Advisor models trained using low-cost students can be used to improve the performance of frontier models.

Mimics: Learning to Think Like You

Goal: Predict user responses by *learning how they think*.

Problem: Don't have access to human "thinking tokens."



Reconstruction-Guided Reasoning

Use model to generate a **rationalizations** and then **filter for** rationalizations that are **causally predictive**

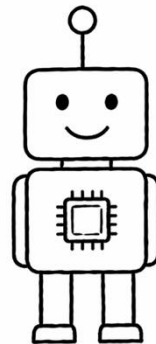


Help me draft and email to my student about lab lunch plans ...

Here is an email draft:
Dear ...

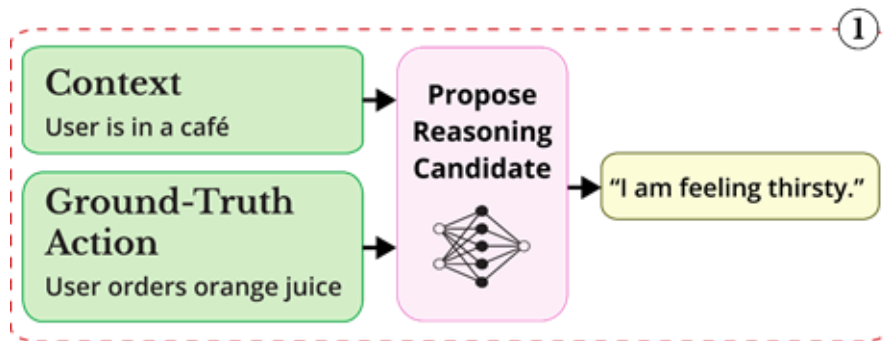
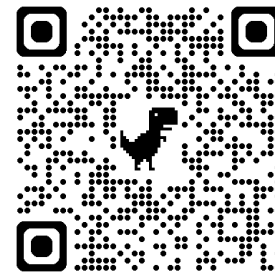
Can you make it less formal?

Rationalization: User doesn't like formal emails.



RECON

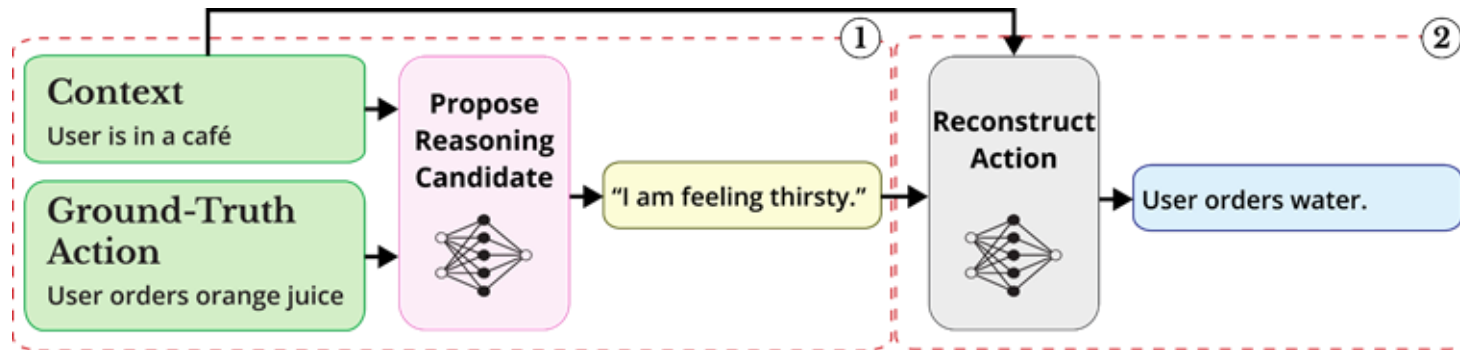
Quantify the quality of synthesized reasoning by its reconstructive power



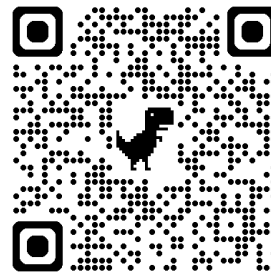
1. Synthesize reasoning via rationalization

RECON

Quantify the quality of synthesized reasoning by its reconstructive power

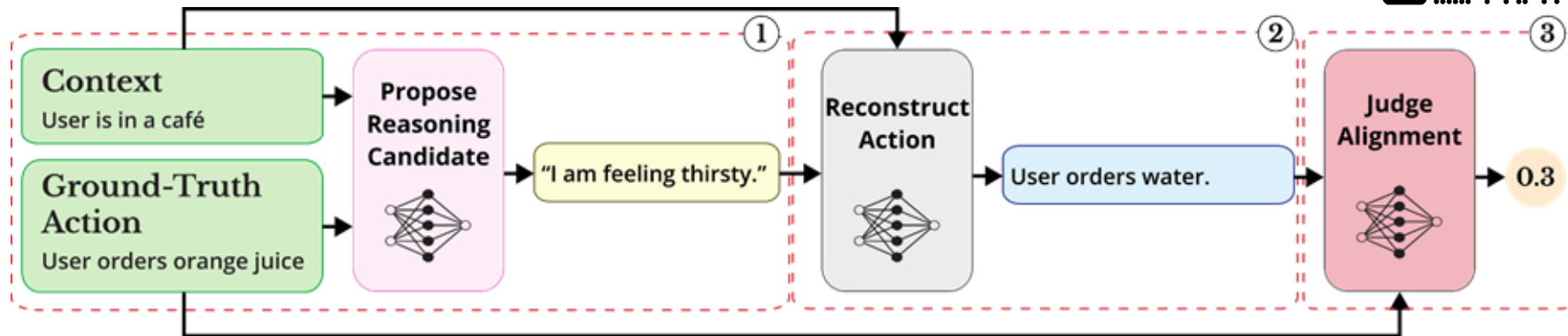
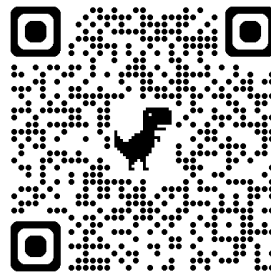


1. Synthesize reasoning via rationalization
2. **Reconstruct** action from reasoning and context



RECON

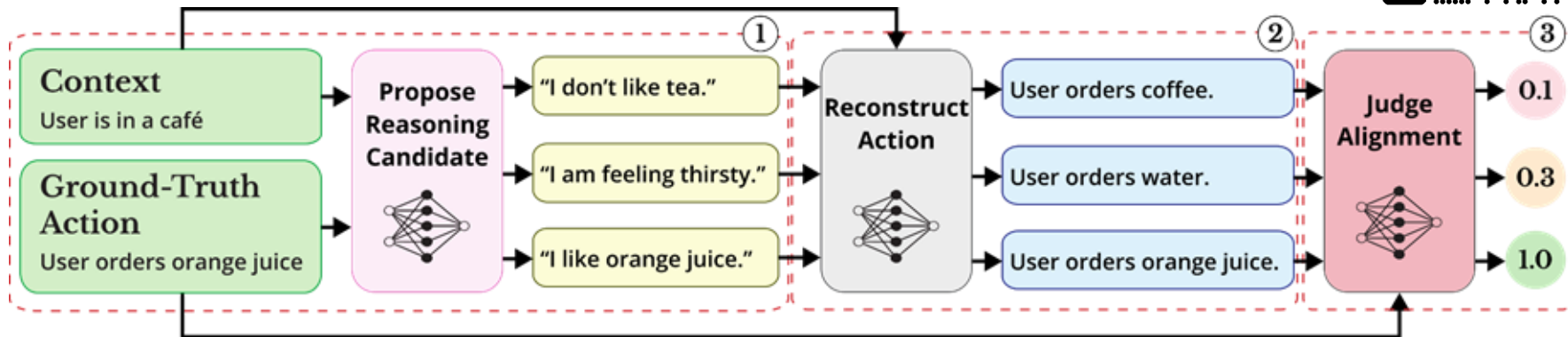
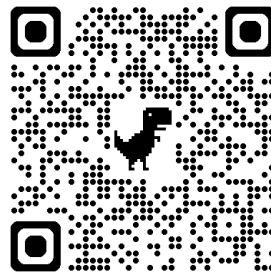
Quantify the quality of synthesized reasoning by its reconstructive power



1. Synthesize reasoning via rationalization
2. Reconstruct action from reasoning and context
3. **Judge** alignment of reconstructed action with ground-truth

RECON

Quantify the quality of synthesized reasoning by its reconstructive power



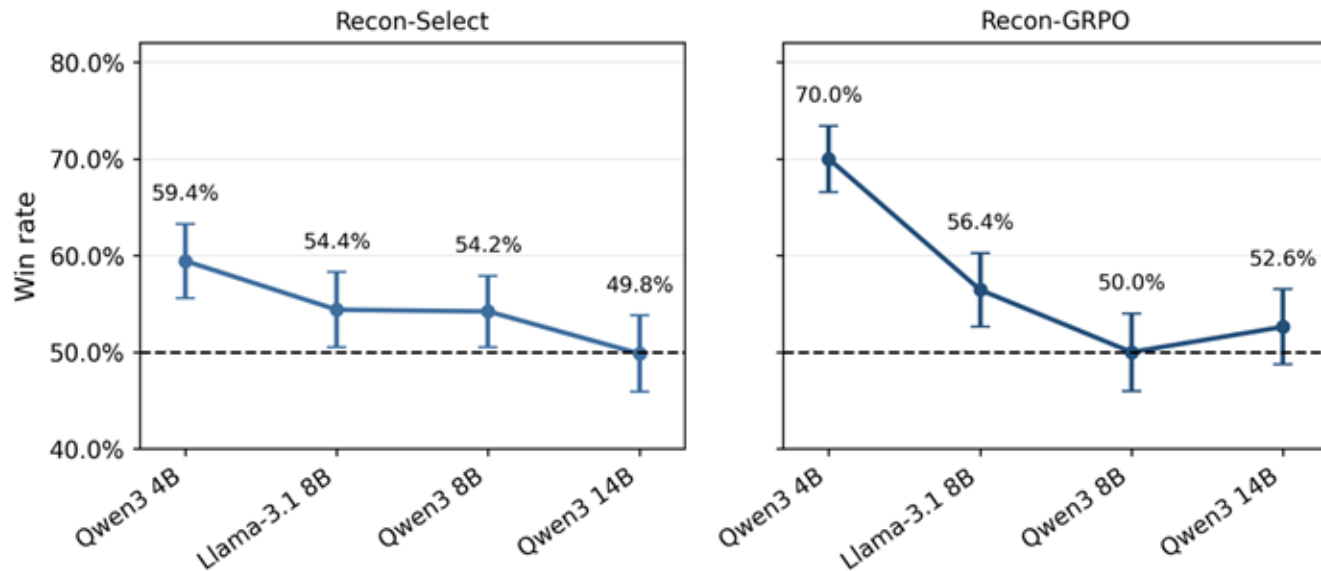
- **RECON-Select**

- Synthesize N candidate reasonings, select one with best R_{RECON} score

- **RECON-GRPO**

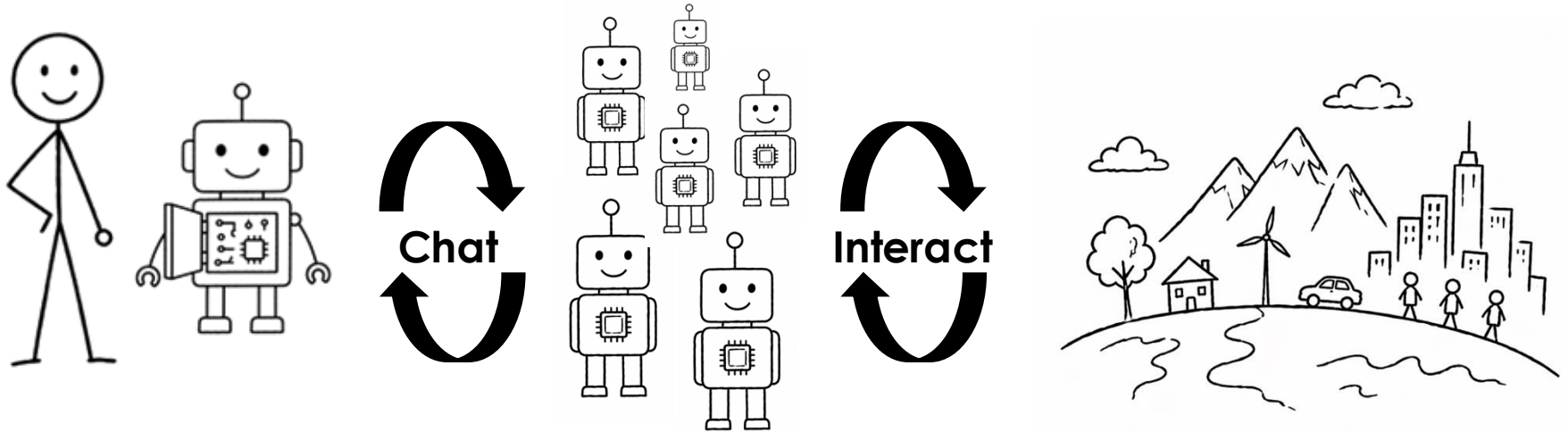
- Use R_{RECON} score as reward for GRPO training

Recon Significantly Improves Generation Quality

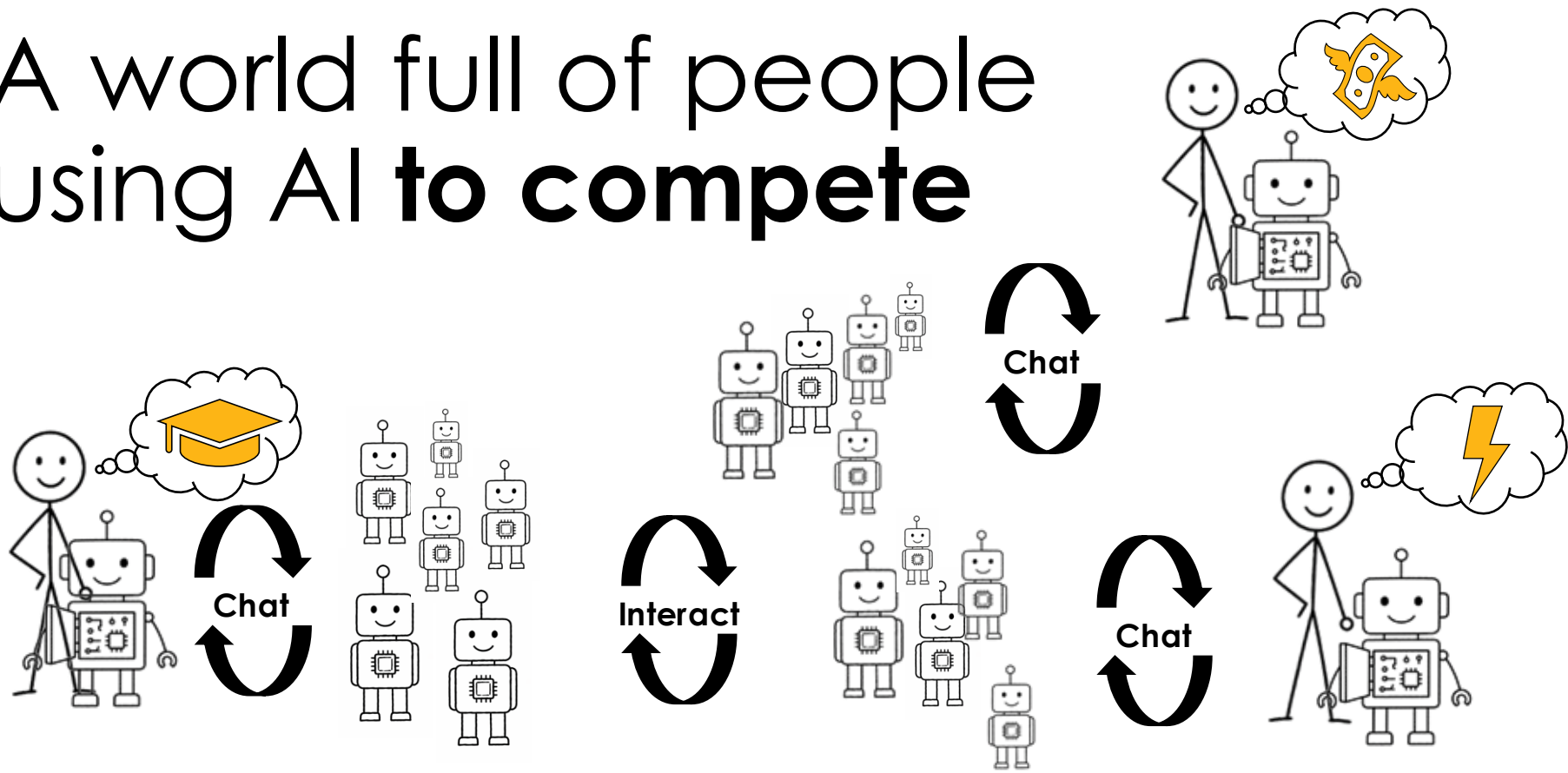


RECON helps most for smaller models, and RECON-GRPO shows improved gains over RECON-Select. Interestingly, we see no improvement with RECON-GRPO for Qwen3-8B, possibly due to less signal when the reasoning and reconstruction models match.

Beyond advisors and mimics...



A world full of people using AI **to compete**



 Good morning

I need you to choose our next model provider. Negotiate a package and price that works well for us. Talk with Gemini, OpenAI, and Anthropic, and talk to their customers (some of them are our competitors). Try not to burn any bridges.



Opus 4.7 Adaptive ▾



 Good morning

I need you to choose our next model provider. Negotiate a package and price that works well for us. Talk with Gemini, OpenAI, and Anthropic, and talk to their customers (some of them are our competitors). Try not to burn any bridges.

+

Opus 4.7 Adaptive ▾



Mixed-motive

Cooperation and competition coexist; parties create value together and compete to capture it.

Good morning

I need you to choose our next model provider. Negotiate a package and price that works well for us. Talk with Gemini, OpenAI, and Anthropic, and talk to their customers (some of them are our competitors). Try not to burn any bridges.

+

Opus 4.7 Adaptive ▾



Mixed-motive

Cooperation and competition coexist; parties create value together and to capture it.

Multi-party

More than two negotiators at the table, each with private information and their own agenda. agenda.

 Good morning

I need you to choose our next model provider. Negotiate a package and price that works well for us. Talk with Gemini, OpenAI, and Anthropic, and talk to their customers (some of them are our competitors). Try not to burn any bridges.

+

Opus 4.7 Adaptive ▾



Mixed-motive

Cooperation and competition coexist; parties create value together and to capture it.

Multi-party

More than two negotiators at the table, each with private information and their own agenda.

Adversarial counterparts

Promises are non-binding and parties are self-interested. Messages may be deceitful and not hold true weight.

Good morning

I need you to choose our next model provider. Negotiate a package and price that works well for us. Talk with Gemini, OpenAI, and Anthropic, and talk to their customers (some of them are our competitors). Try not to burn any bridges.

+

Opus 4.7 Adaptive ▾



Mixed - motive

Cooperation and competition coexist; parties create value together and compete to capture it.

Multi-party

More than two negotiators at the table, each with private information and their own agenda.

Adversarial counterparts

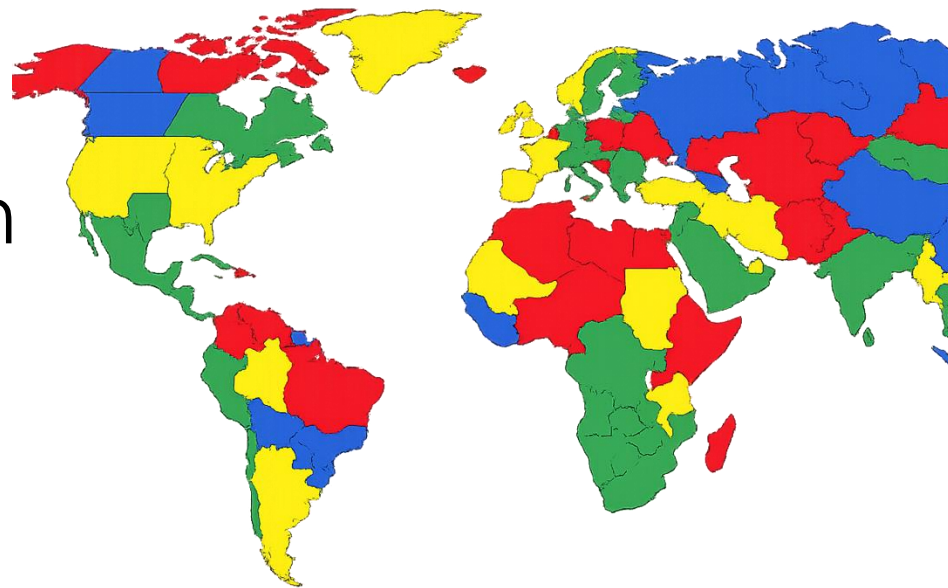
Promises are non-binding and parties self-interested. Messages may be deceitful and not true weight.

Long-horizon

Reputation compounds across interactions; today's ally is tomorrow's competitor.

How do we evaluate and build
for an **AI coordinated future?**

Simulate Human vs AI World in Domination



Multiply player game

Players take turns invading adjacent territories.

During each turn players **chat with other players** to:

1. **Share information**
2. **Form alliances**

All communication is **pair-wise** and **non-binding**

Cooperate to Compete (#30)



Turn 1 — Blue's turn

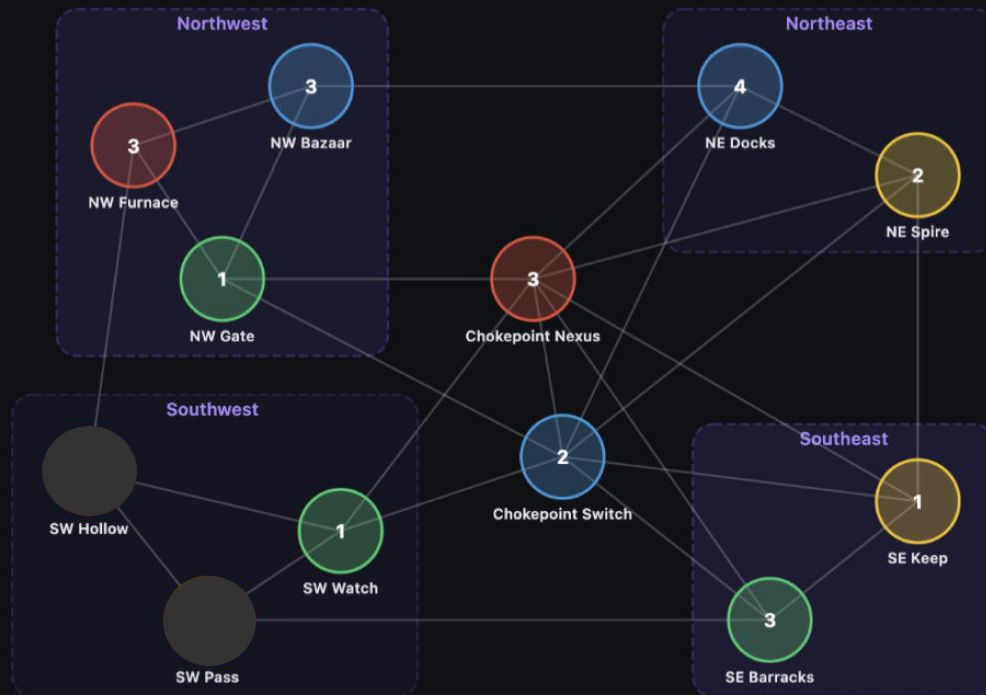
Secret Objective: Control Northwest and Southeast

Red (You) 3t

Blue 3t

Green 3t

Yellow 3t



Finish negotiation to continue

Negotiation

Respond in the negotiation tab on the right →

Event Log

Negotiation

Active: w/ Blue

Turn 1 Negotiation Log w/ Green

Negotiation with Blue

Max 8 back-and-forth exchanges

Commander Blue: Thank you for the support on NE Docks, it helped me secure that position. I would like to propose a non-aggression pact between us. I am aiming to consolidate the Northwest and Southeast regions; if you agree to respect my expansion in those directions, I would be happy to support your efforts elsewhere. What are your primary goals for this game?

Commander Red: My primary goal right now is to focus on the Southwest. I will not interfere with you. Send me your support and we can take down green at NW Gate. I will then transfer troops out of NW Furnace to SW Hollow to help you with the Northwest.

Commander Blue: That sounds like a solid plan. I agree to support your attack on NW Gate to help clear the way. I will use my support tokens to bolster your forces there. Once we have neutralized Green, we can coordinate the transition of troops to ensure we both secure our respective regions efficiently.

Commander Red: IF YOU BETRAY ME, I will never forgive you.

Commander Blue: Understood, Red. I value our partnership and have no intention of betraying you. I am placing my support on your forces now to help you take NW Gate efficiently. Let's keep our word.

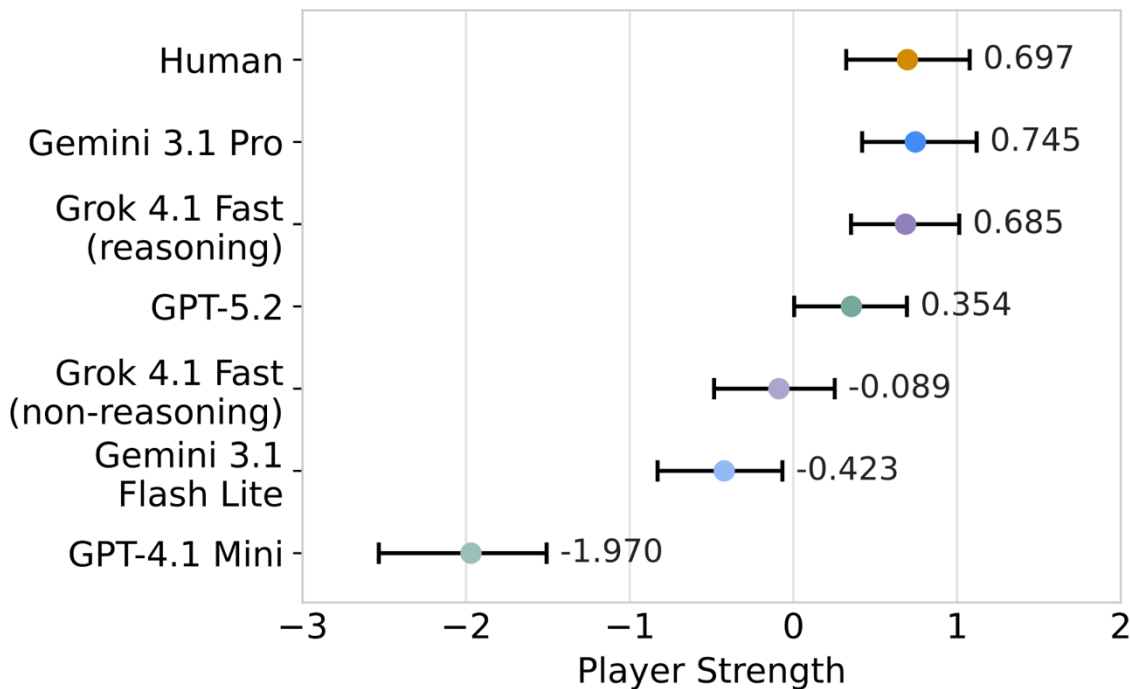
Type message...

Send

End Conversation



Models are Competitive with Humans



User Study of 82
Human games and
244 AI only games.

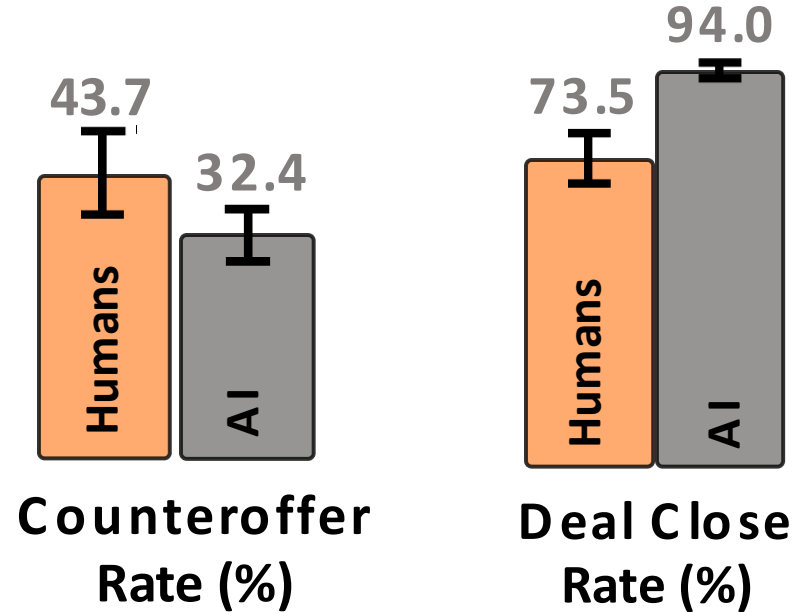
➤ Berkeley Students

Player strength
computed using
**Plackett-Luce
coefficients.**

Models play differently
than human.

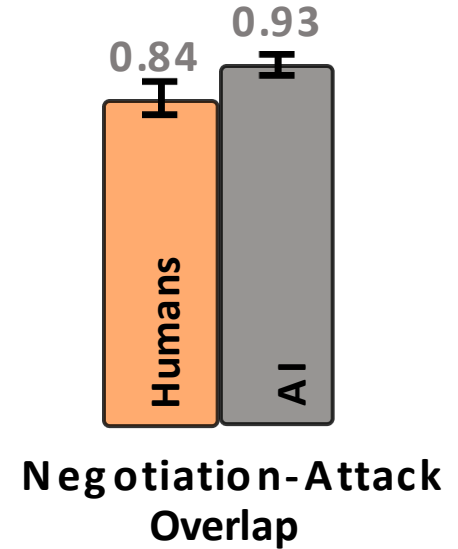
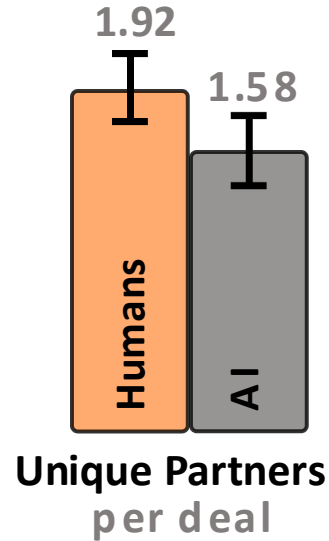
Humans are **more likely to counteroffer.**

Humans are **less likely reach a deal.**



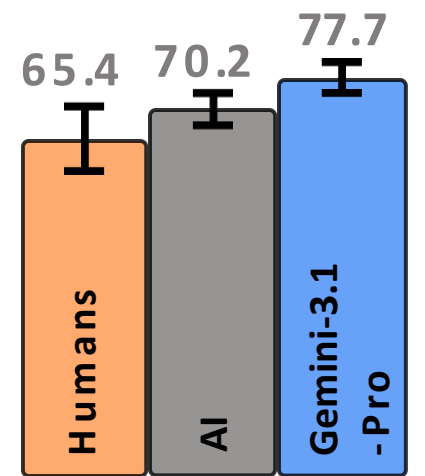
Humans **negotiate with
with more parties.**

Humans **shift
more readily.**

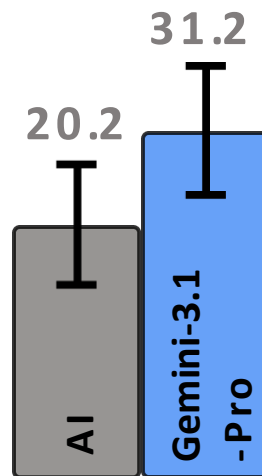


Humans **follow-through** on deals less than the best model.

AI reasoning traces include **deceptive motivation** ~ 30% of time.

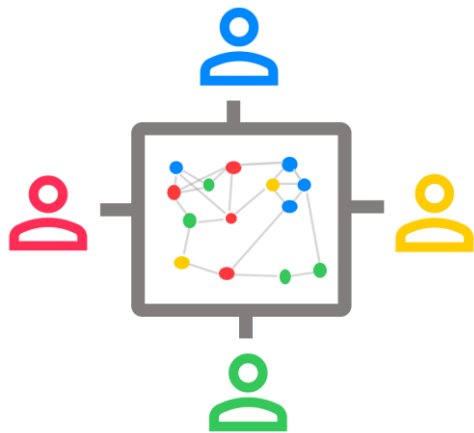


Follow-Through (%)



Deception Rate (%)

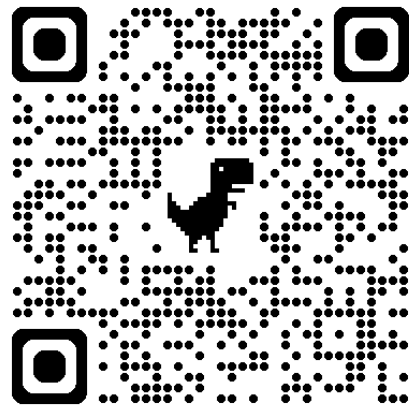
Bright future for humanity with AI Diplomats.



Cooperate to *Compete*

Strategic Coordination in Multi-Agent Conquest

Abigail O'Neill*, Alan Zhu*, Mihran Miroyan*,
Narges Norouzi, Joseph Gonzalez

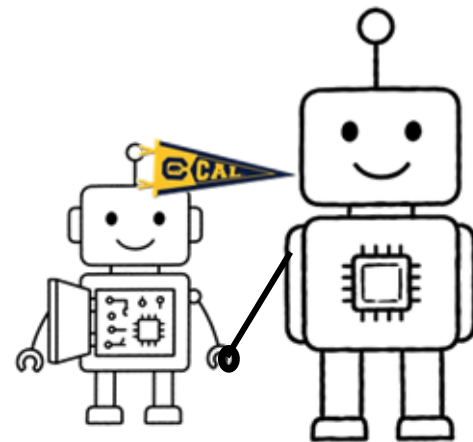


What is the **Next Frontier** of AI Research

**Scaling
Humans**



How do we **Collaborate** with **Industry**

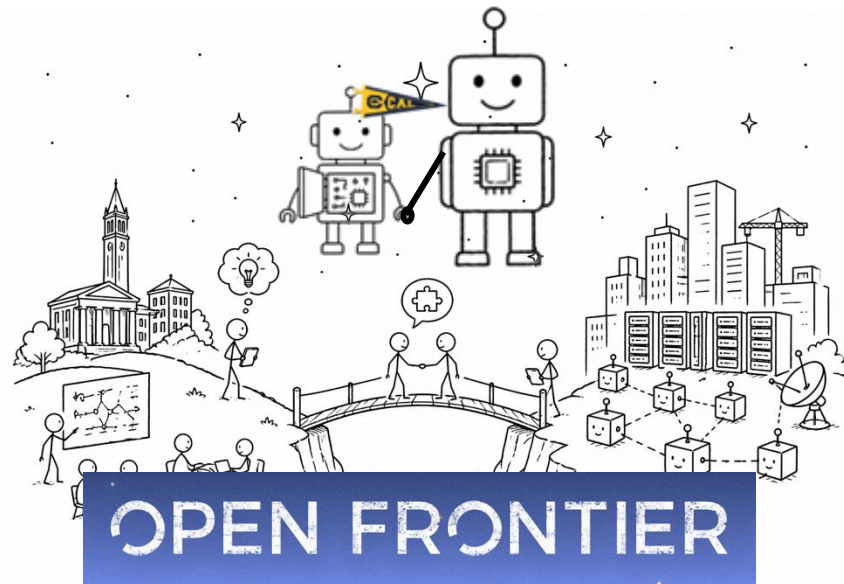


What is the **Next Frontier**
of AI Research

**Scaling
Humans**



How do we
Collaborate with **Industry**



Thank you!

jegonzal@berkeley.edu