

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



We need a new **north star**!



About me:

Joseph (Joey) Gonzalez

jegonzal@berkeley.edu

Professor at UC Berkeley

- Co-director of the Sky Computing Lab
- Co-founded of RunLLM, InferAct, and Turi

Research in AI and Data Systems

Teach the Data Science and AI



RunLLM

turi 
create intelligence™

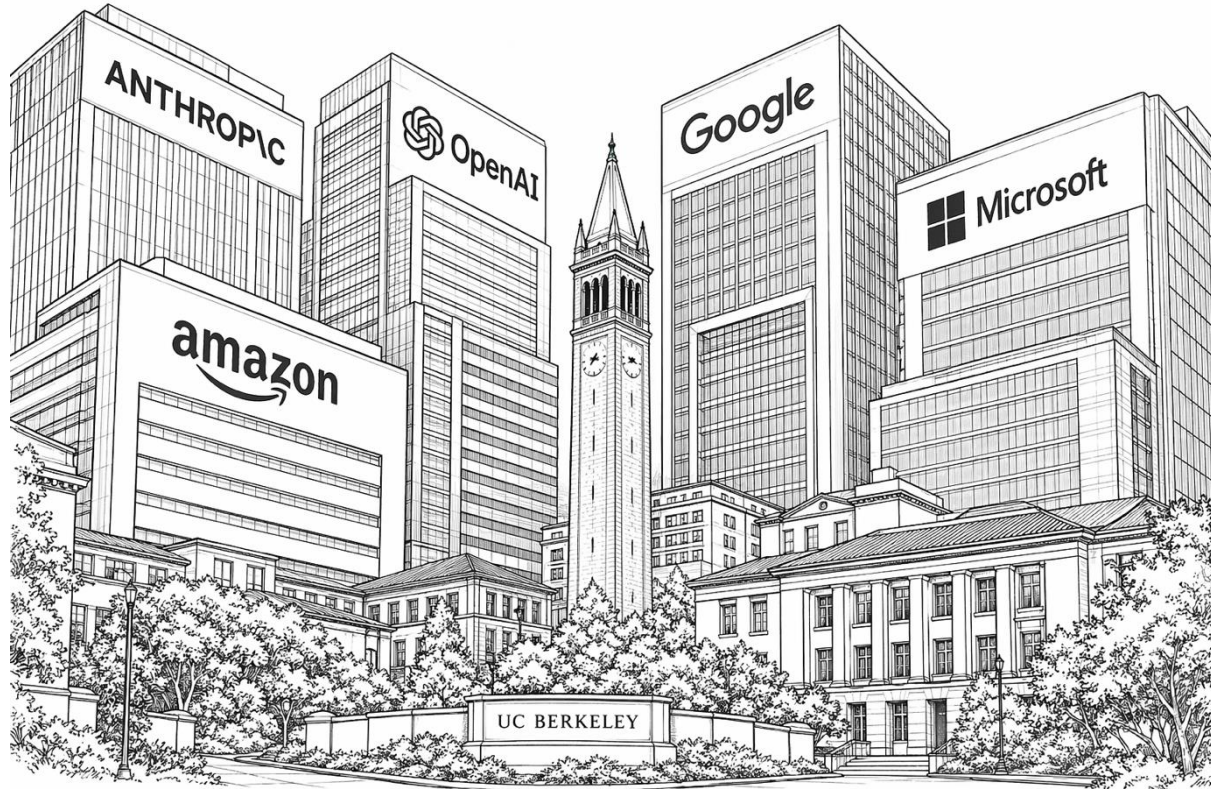


EPIC
DATA lab
UC Berkeley

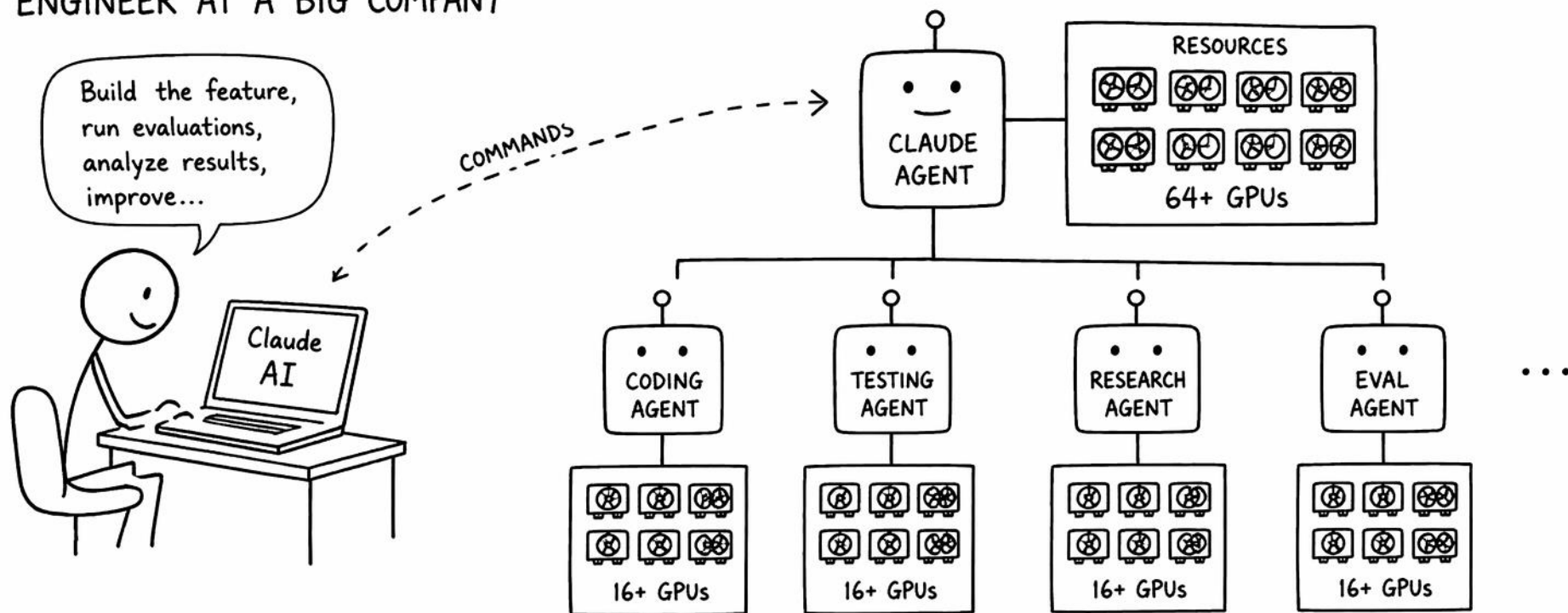


 **riselab**
UC Berkeley

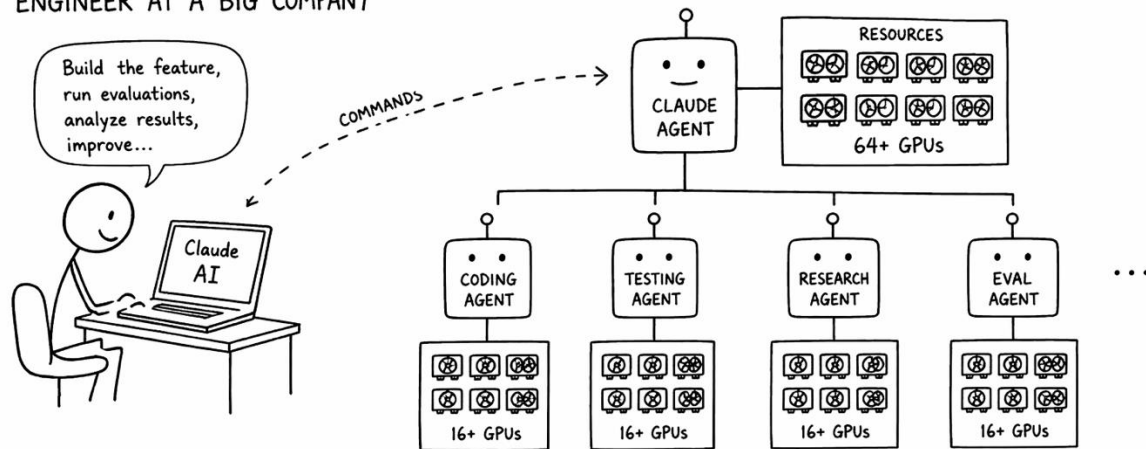
Academic anxiety in the Era of Frontier Labs



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**.

Looking beyond the horizon

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



My story:

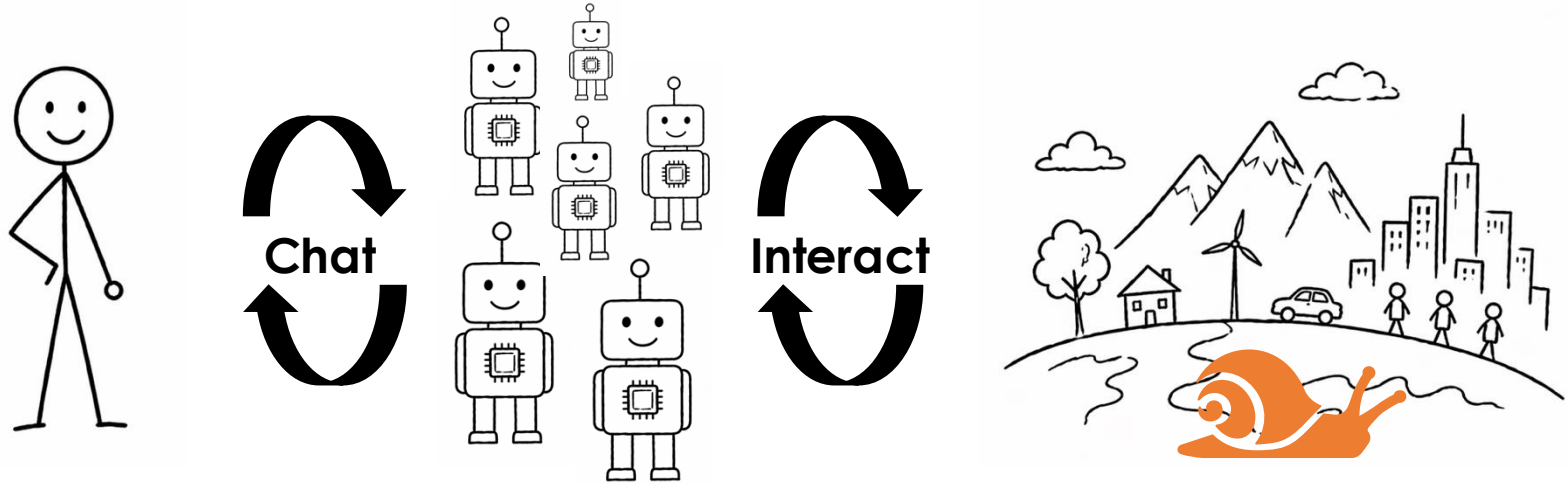
2008: Everyone is designing models, I focused on scaling learning.

→ AI-Systems

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

→ Inference Systems

Agentic AI in **2026**



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

How do we build **software systems** that better serves 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 *cheaper* 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



STAR Lab

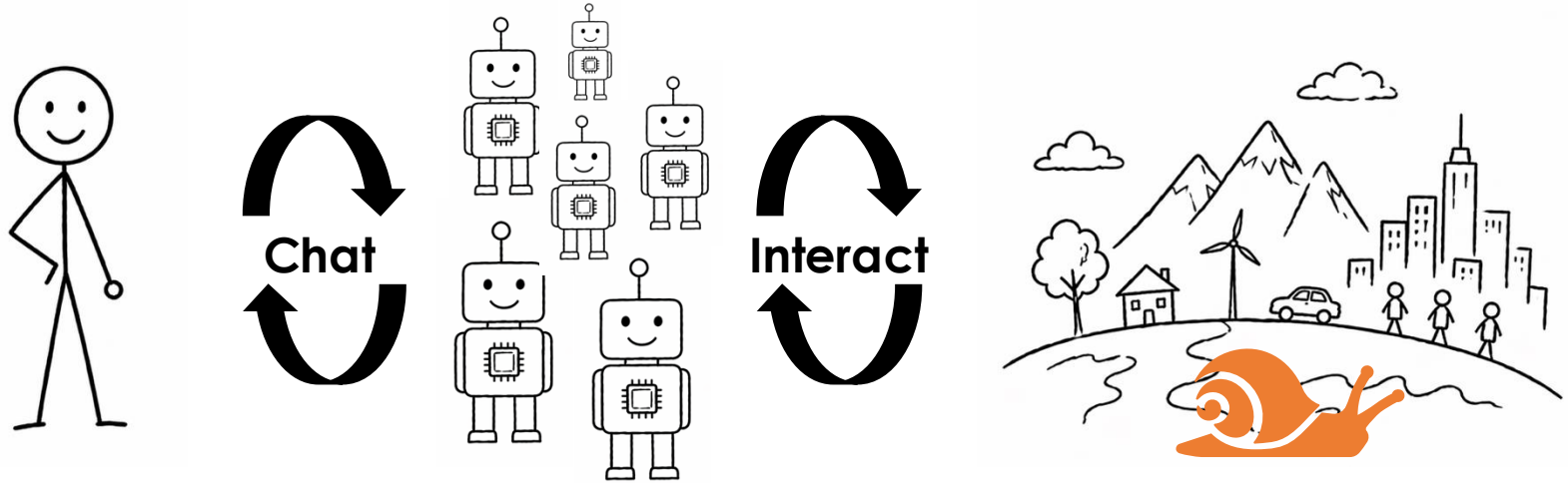
Systems Transformation with **AI Research**
Building Systems for AI with AI



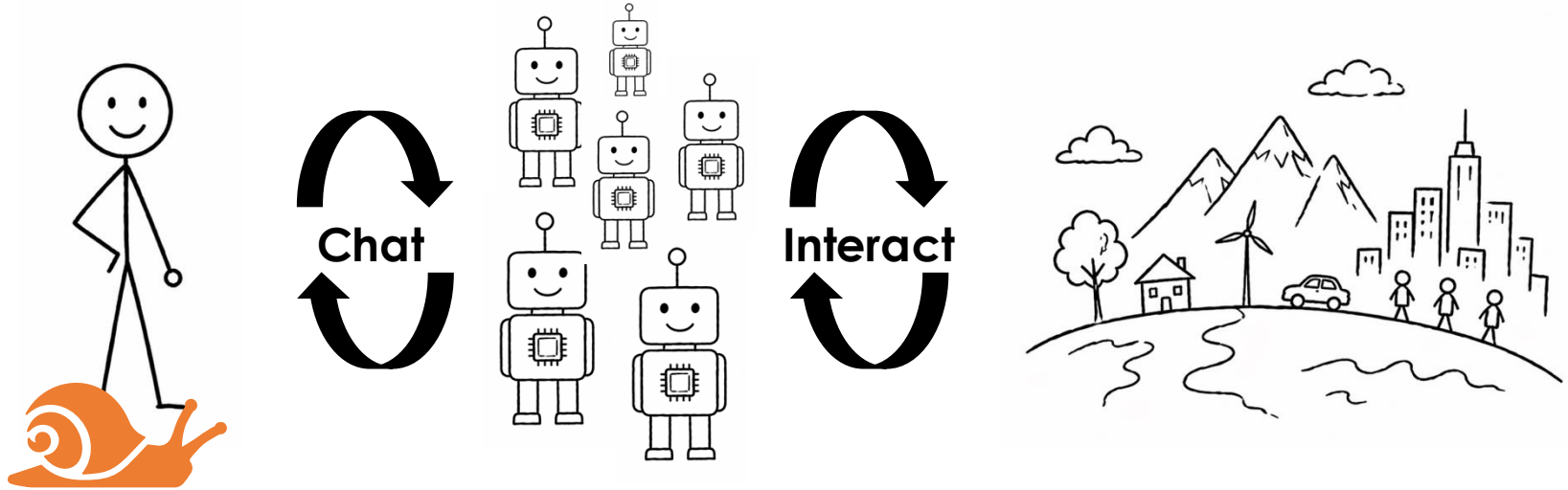
Redesign systems and
abstractions to
support AI users.

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

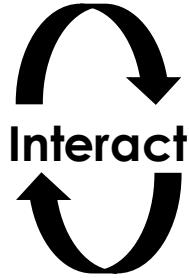
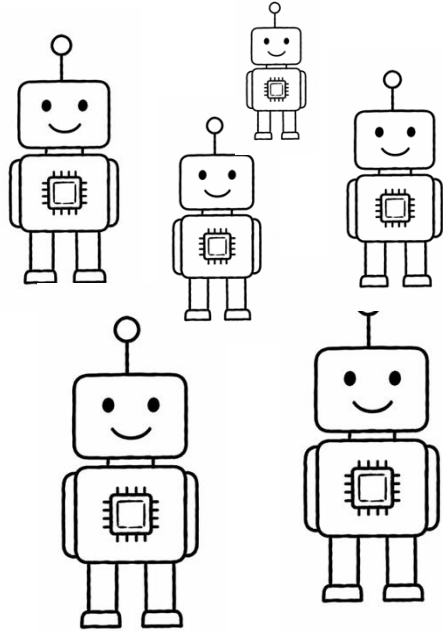
Agentic AI in **2027** and **beyond**



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

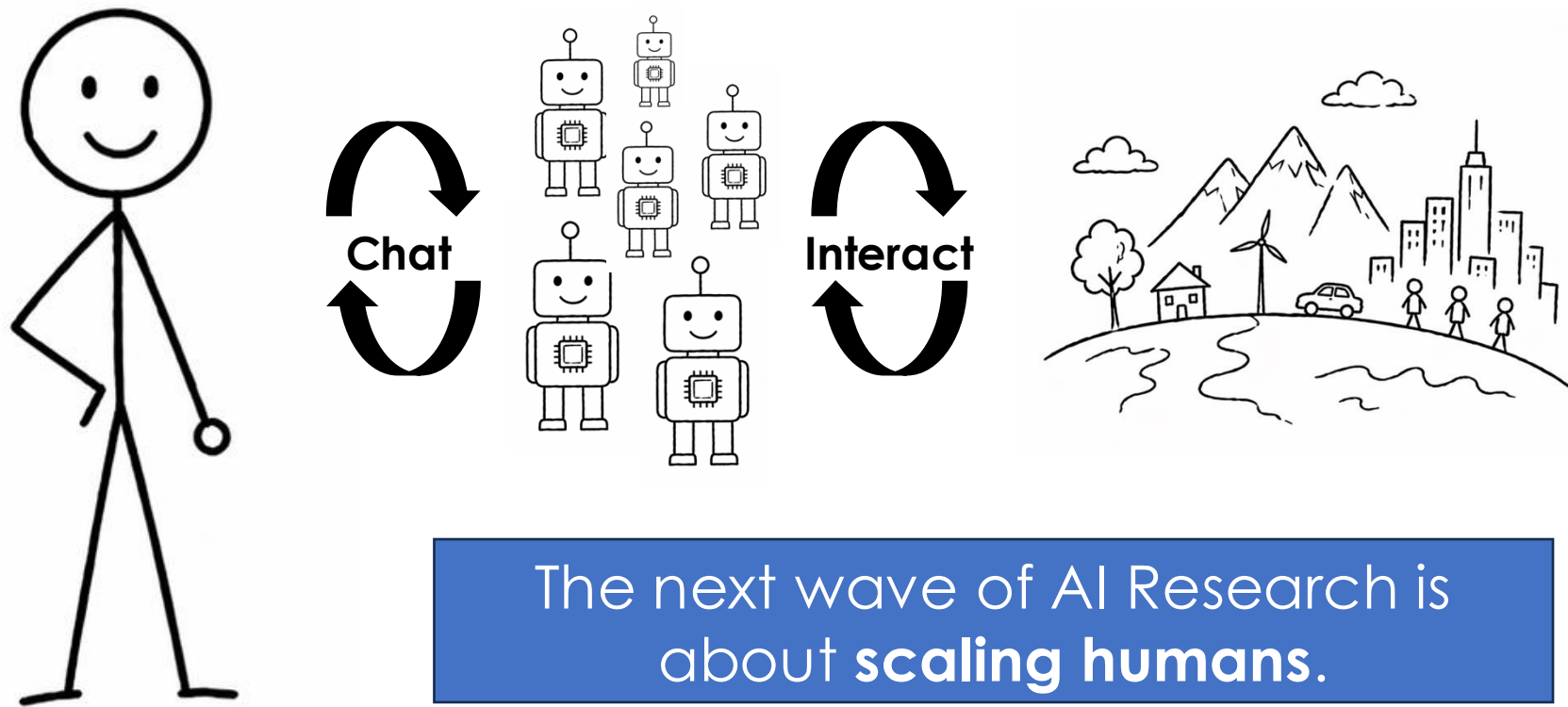


Agentic AI in **2027** and **beyond**



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.

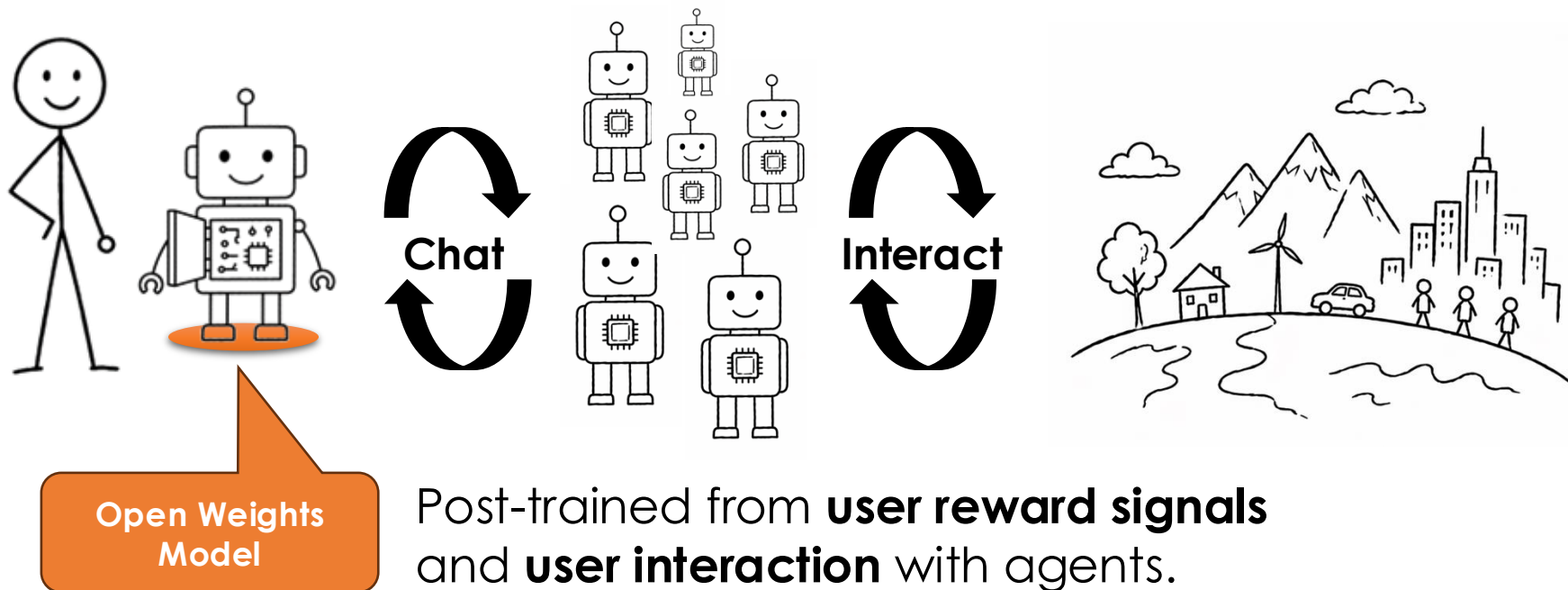
- **User mimics: thinking** and resp. like an **individual**

Humans have **misaligned** goals.

- **Cooperate to Compete:** studying AI in the context of strategic negotiation

Advisors and Mimics

Adding Another Layer of AI



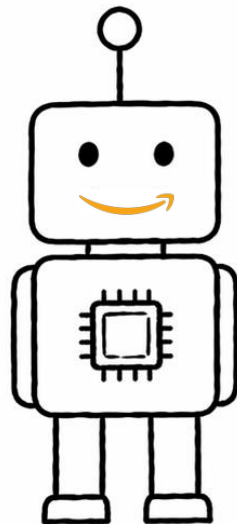
Academia + Industry

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

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



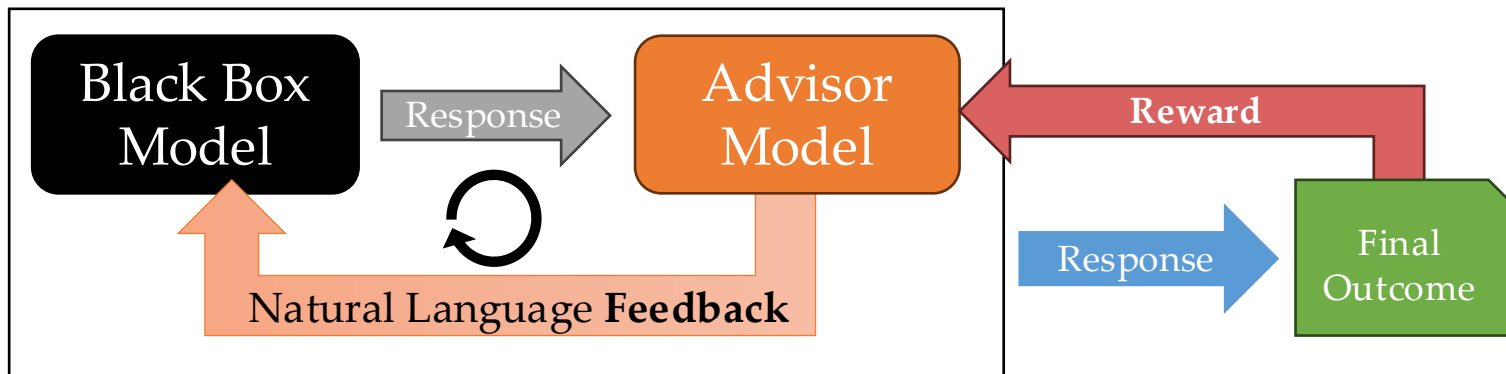
+



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.

Mimics: Learning to Think Like You



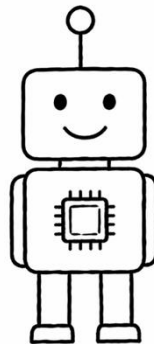
Help me email Mir about...

Here is an email draft:
Dear ...

Can you make it less formal and
more concise?

Revised version:
Dear ...

Looks much better, thanks.



Mimics: Learning to Think Like You

Use frontier models to generate (infer) thinking **conditioned on the response**

Mimic:

Train open-weight model to **generate thinking** and **response** using (SFT + RL).



Mir is a student and this isn't a formal concern.

Can you make it less formal and more concise?

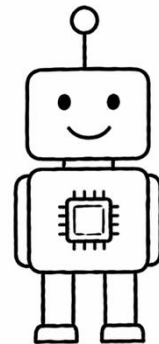
I think this is more consistent with how I write emails to Mir...

Looks much better, thanks.

Help me email Mir about...

Here is an email draft:
Dear ...

Revised version:
Dear ...

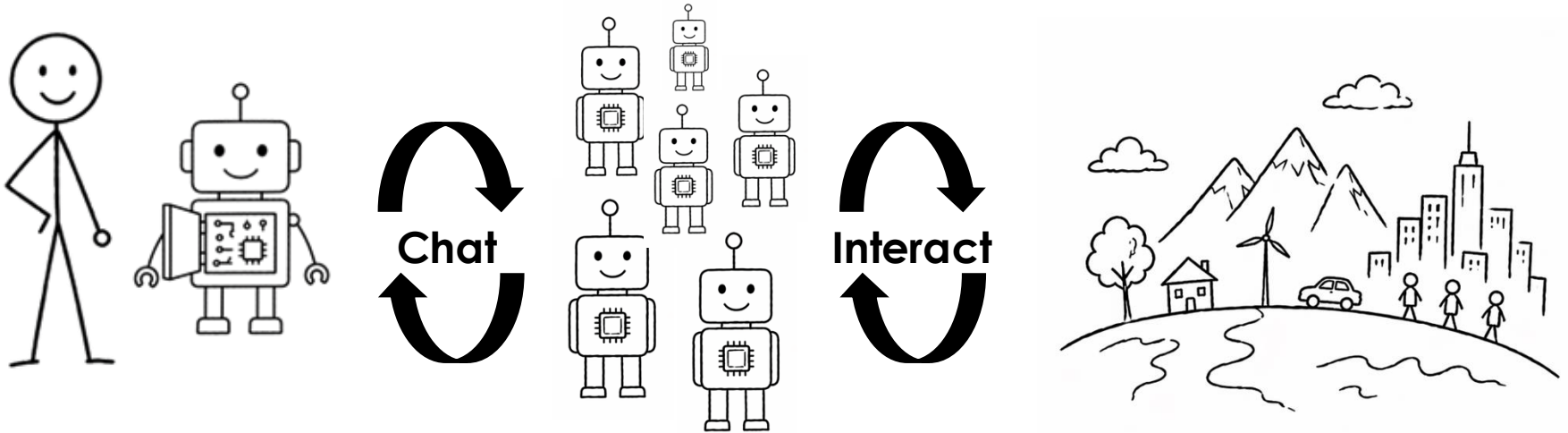


Results for Advisors and Mimics

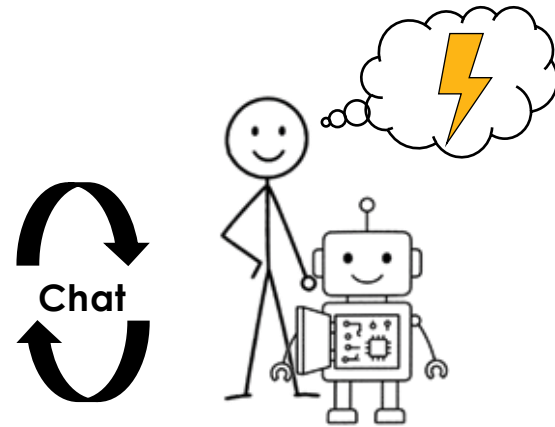
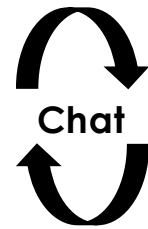
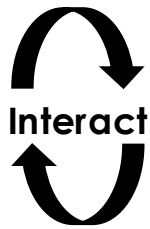
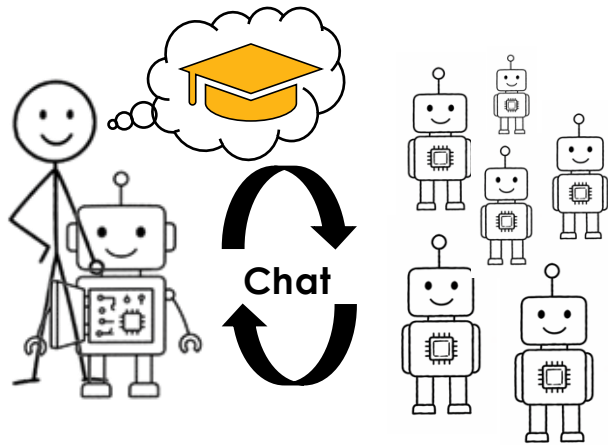
Advisor Models and Mimics:

- outperform in-context techniques at **extracting hidden user preferences**.
- **lift frontier model performance** on specialized tasks.
- are **robust** to domain shift; don't degrade frontier model performance when out of domain.

Beyond advisors and mimics...



The world is other people using AI to compete



Studying how to Cooperate to Compete

Studying **AI agents** and **humans** playing negotiation games

- Based on **Risk** and **Diplomacy**
- Designed to balance **negotiation** and **tactics**

Frontier agents can **defeat** “average” humans (grad. students).

Agents are more likely to **collaborate, make, and keep** deals.

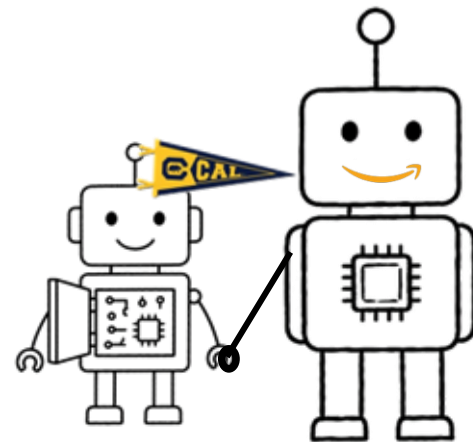
Encouraging agents to be **more deceptive** increases win-rate.

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

**Scaling
Humans**



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

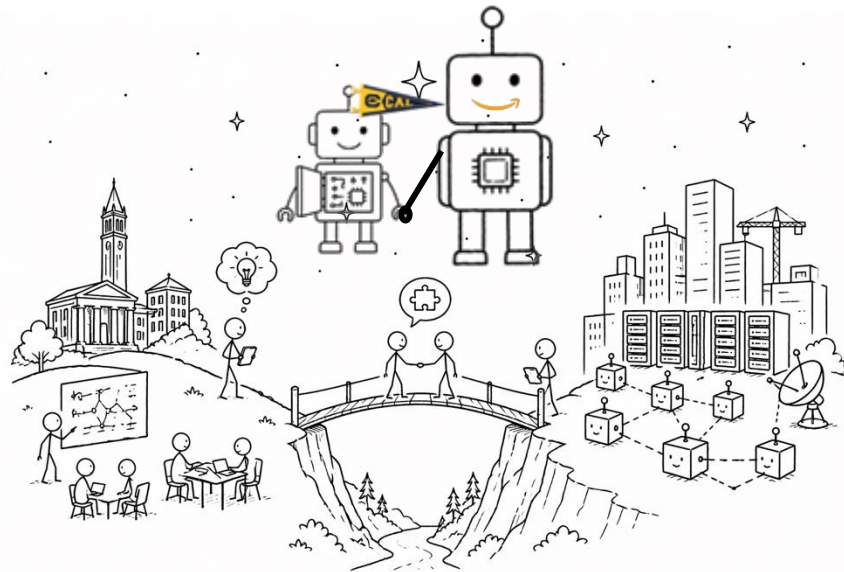


What is the **Next Frontier**
of AI Research

Scaling
Humans



How do we
Collaborate with **Industry**



Thank you!