

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 ✨

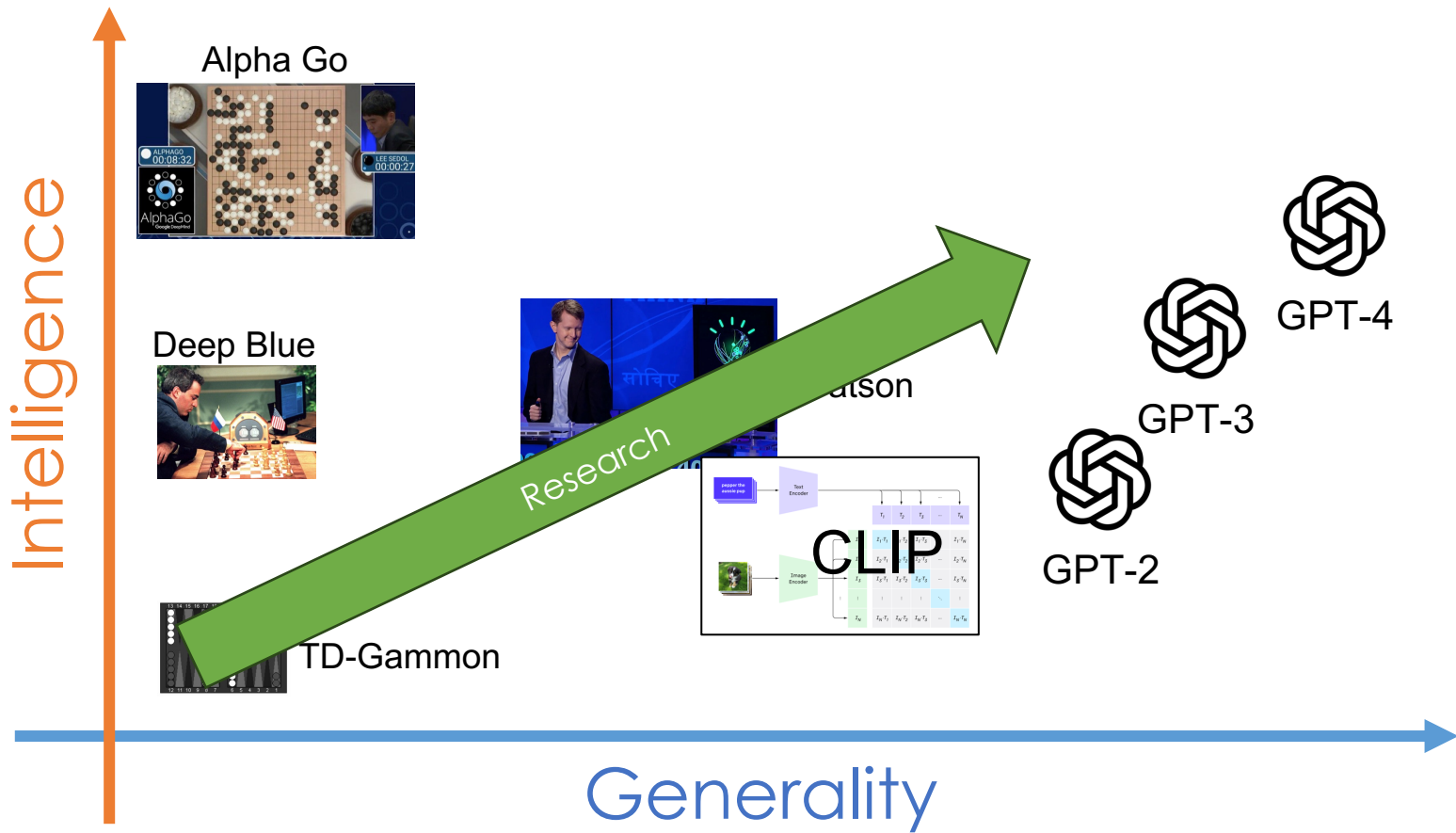


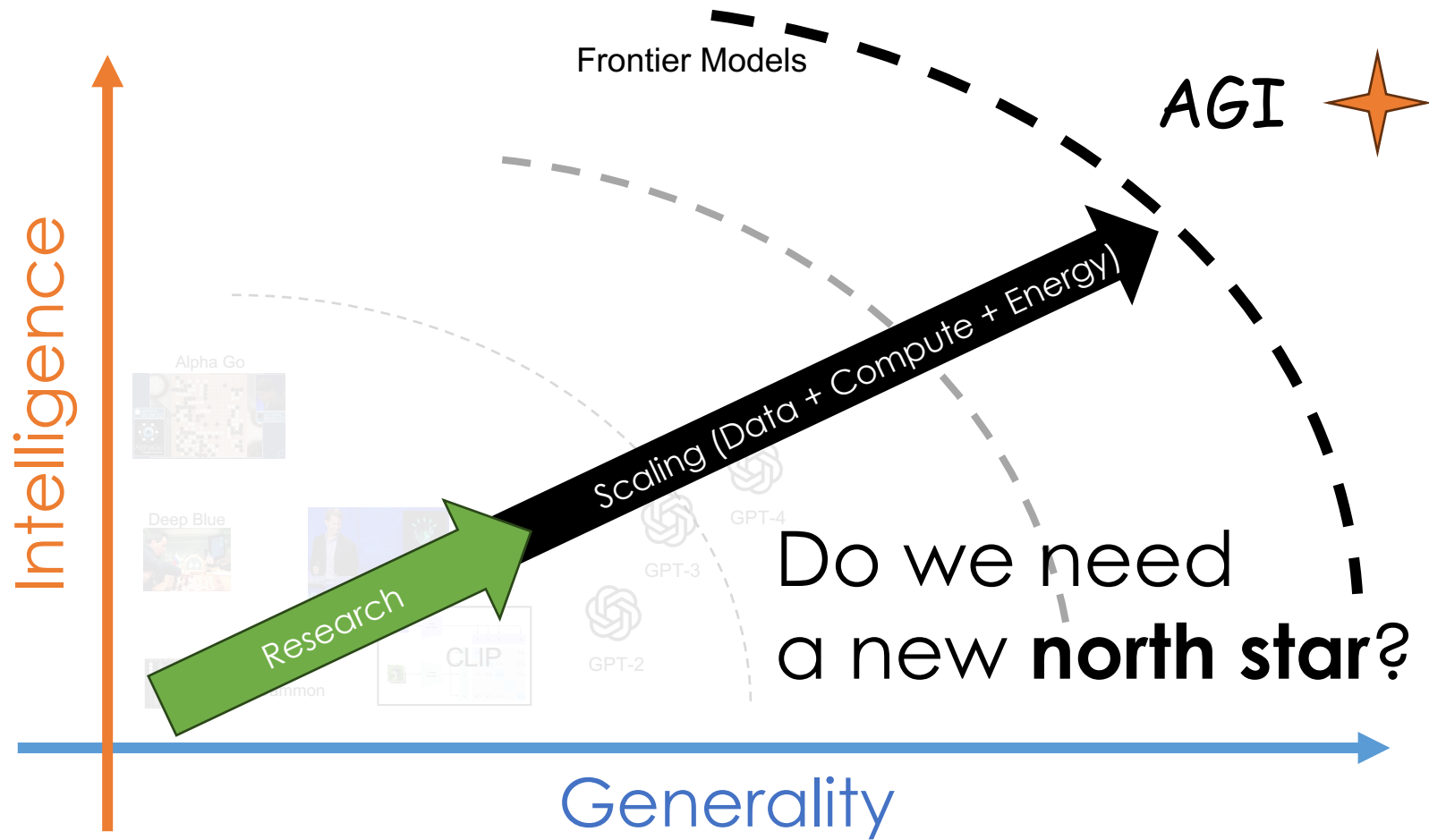
AGI ✨



AGI







About me:

Joseph (Joey) Gonzalez

jegonzal@berkeley.edu

Professor at UC Berkeley

- 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™

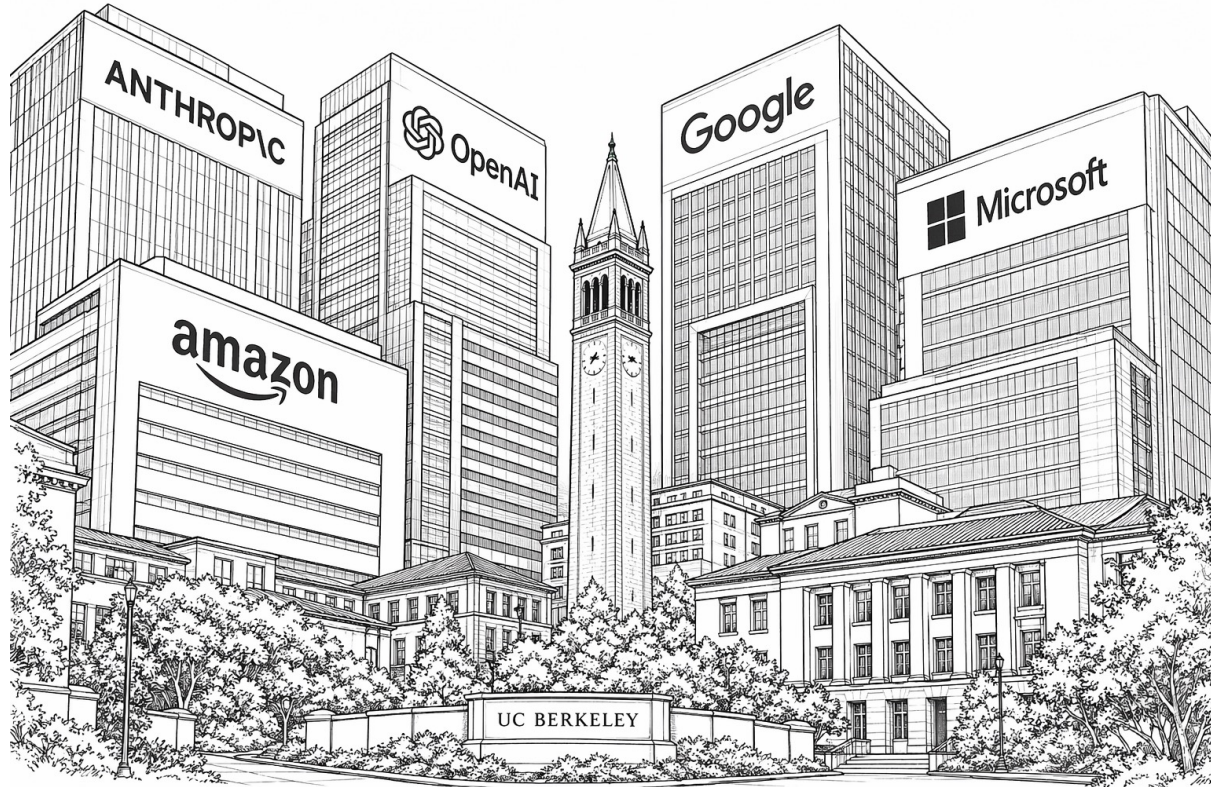


EPIC
DATA lab
UC Berkeley

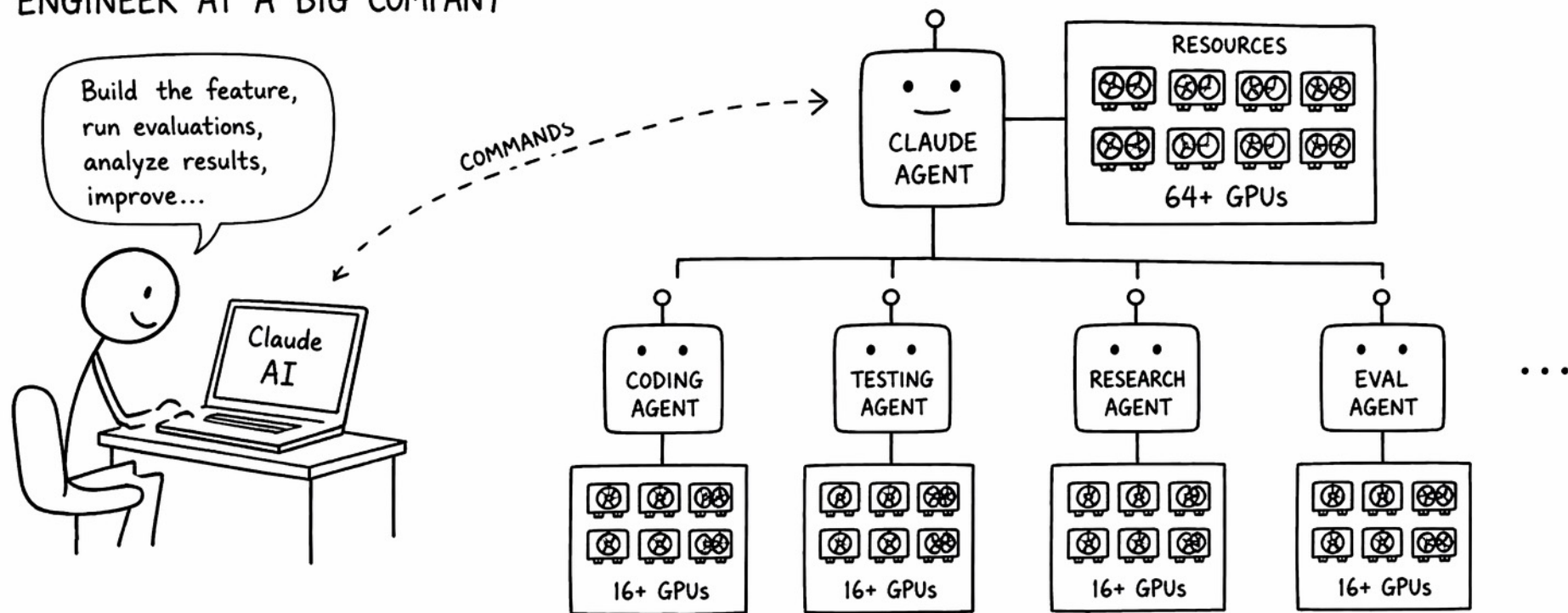


 riselab
UC Berkeley

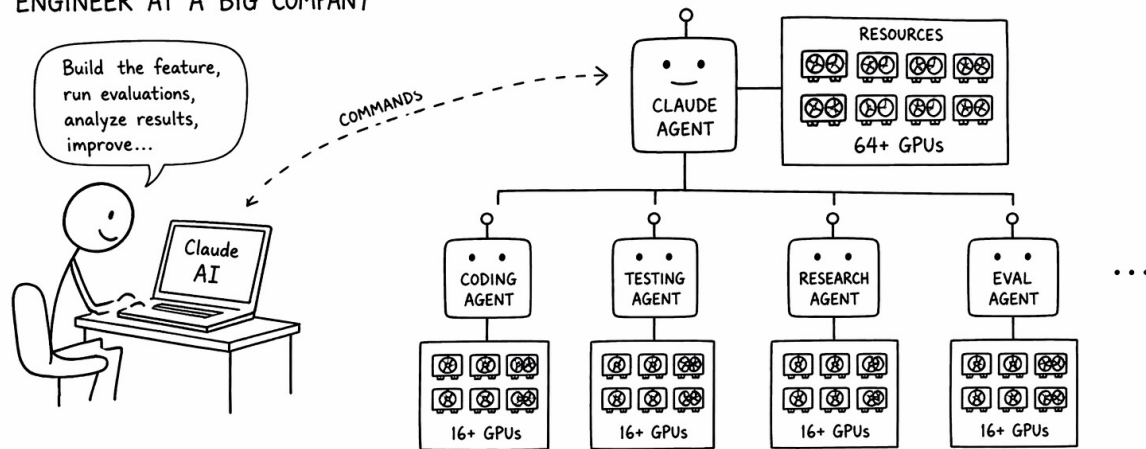
Academic anxiety in the Era of Frontier Labs



ENGINEER AT A BIG COMPANY



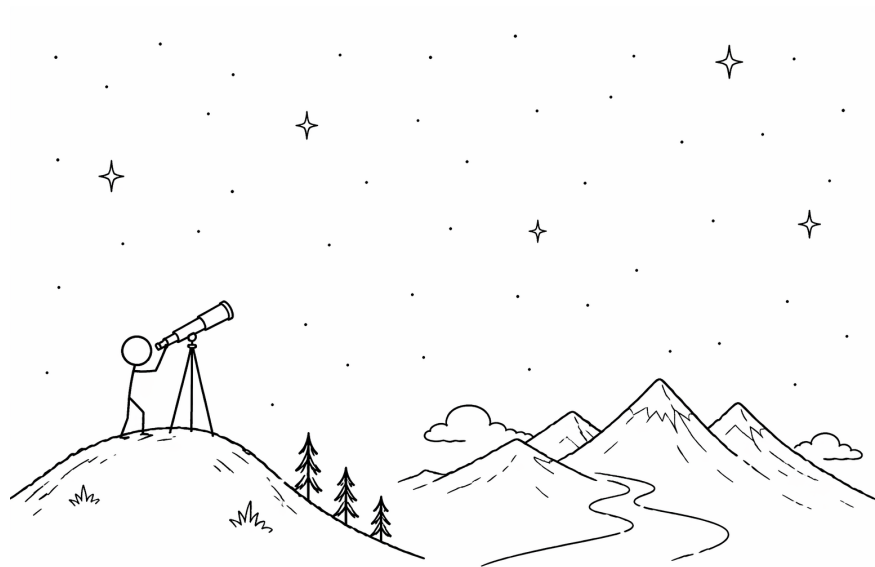
ENGINEER AT A BIG COMPANY



MY GRAD STUDENT



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

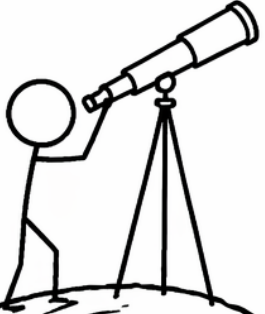


How do we
Collaborate with **Industry**



Both questions have a **common answer**.

✧ What if they are all wrong?



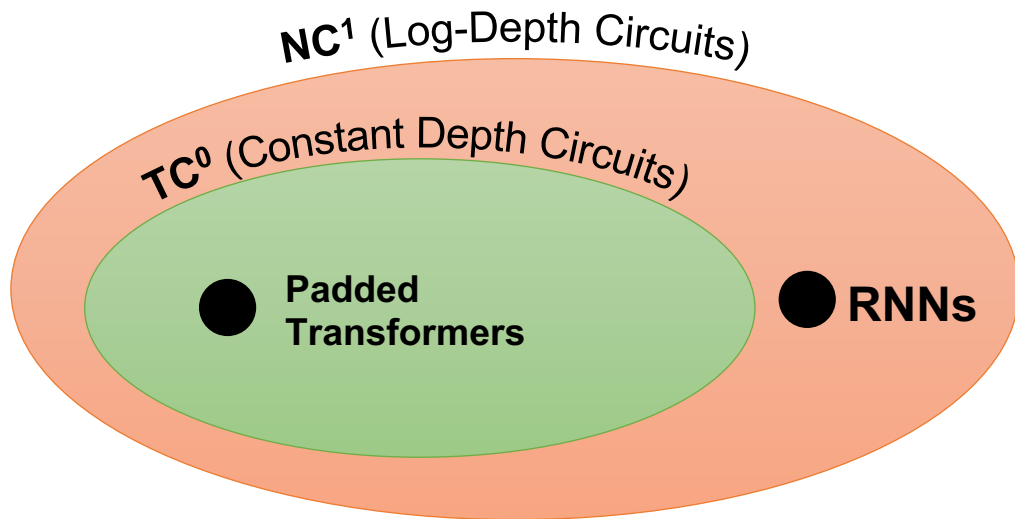
They are scaling
the **wrong model!**

Academics should
challenge the path of industry.

Attentions is **NOT** all you need!

Computation: Quadratic cost of long-context

Expressivity:



RNNs more naturally express **state tracking** and **iterative reasoning** tasks.



M²RNN: *Non-Linear RNNs* with Matrix Valued States for Scalable Language Modeling

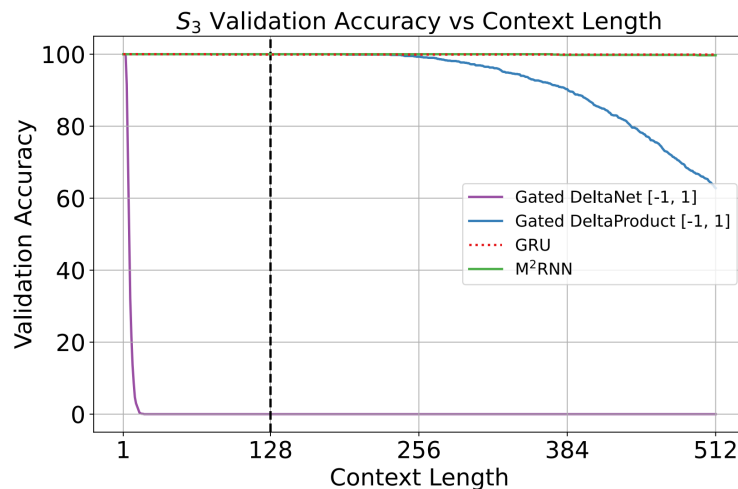
Core Idea: *Bring back RNNs* → RNNs can **solve NC1 problems**

Problem: Traditional RNNs are bad at **language modeling** and **inefficient to train**

Solution: **Matrix valued States** improve **language modeling** and saturate **GPU computation**

RNN  $h_t = \tanh(\mathbf{W} h_{t-1} + v_t)$

M²RNN  $\mathbf{H}_t = \tanh(\mathbf{W} \mathbf{H}_{t-1} + v_t)$



Gated Delta-Net as **Online Regression**

Building on test-time regression* which casts RNNs as **online regression using** one step of **SGD**

Robust
Alternative
to L^2 -Loss?

$$\min_{\mathbf{H}} \frac{1}{2} \|\mathbf{H}^\top k_t - v_t\|^2 + \frac{\lambda_t}{2} \|\mathbf{H}\|^2$$

where the **RNN implements** a gradient step per iteration:

$$\mathbf{H}_t = \mathbf{H}_{t-1} - \lambda_t \eta_t \mathbf{H}_{t-1} - \eta_t k_t (k_t^\top \mathbf{H}_{t-1} - v_t^\top)$$

Gated Delta Net Update

Robust Gated Delta-Net

Replace the L^2 -loss with a smooth approximation to the Huber Loss – (grows **linearly in tails**):

$$\min_{\mathbf{H}} \frac{1}{2} \|\mathbf{H}^\top k_t - v_t\|^2 + \frac{\lambda_t}{2} \|\mathbf{H}\|^2 \rightarrow \min_{\mathbf{H}} \ln \cosh(\mathbf{H}^\top k_t - v_t) + \frac{\lambda_t}{2} \|\mathbf{H}\|^2$$

the online **RNN implements** a gradient step per iteration:

$$\mathbf{H}_t = \mathbf{H}_{t-1} - \lambda_t \eta_t \mathbf{H}_{t-1} - \eta_t k_t \underbrace{\tanh(k_t^\top \mathbf{H}_{t-1} - v_t^\top)}_{\text{M}^2\text{RNN}}$$

Extra k_t -term

M²RNN

Pre-training Performance Robust Gated Delta-Net's

Trained **410M Parameter** dense models on **100B tokens** of Nemotron-CC-v2

Model	Wiki PPL	LAMBADA	HellaSwag	PIQA	ARC-E	ARC-C	WinoGrande	BoolQ	OBQA	COPA	SciQ	Avg Acc
M ² RNN	22.92	33.98	47.76	71.60	62.29	30.29	53.43	50.21	34.80	72.00	86.00	54.24
DeltaNet	23.85	33.51	44.90	70.24	60.94	29.10	50.51	56.76	32.20	71.00	86.60	53.58
Robust DeltaNet	22.81	34.56	47.83	71.33	62.75	30.12	50.59	54.04	33.40	69.00	84.40	53.80
Gated DeltaNet	22.70	34.21	47.16	71.00	61.24	28.07	51.07	58.84	32.80	70.00	85.50	53.99
Gated Robust DeltaNet	<u>22.34</u>	35.36	48.08	70.89	63.85	30.55	51.07	56.91	34.60	71.00	84.70	<u>54.70</u>
Transformer++	23.32	33.36	45.02	68.88	59.76	29.01	50.43	54.56	32.00	72.00	84.00	52.90
Hybrid DeltaNet	22.82	33.57	45.27	69.04	59.89	27.82	51.38	53.18	33.80	71.00	85.40	53.04
Hybrid Robust DeltaNet	21.34	34.08	48.42	70.95	64.14	30.72	52.72	53.67	34.40	68.00	85.90	54.30
Hybrid Gated DeltaNet	21.89	35.78	47.19	70.67	61.91	29.78	52.17	52.57	34.00	70.00	86.30	54.04
Hybrid Gated Robust DeltaNet	21.14	35.84	49.22	71.11	63.43	30.12	52.25	49.45	34.60	71.00	87.30	<u>54.43</u>
Hybrid Gated DeltaNet + Gated Robust DeltaNet (1L)	<u>21.41</u>	36.27	48.58	71.16	62.42	31.23	53.67	52.48	32.20	71.00	87.60	54.66

Takeaway: Robust networks **perform better** but hybrids perform best.

We can **challenge** the current course

How do we look
beyond the horizon?



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

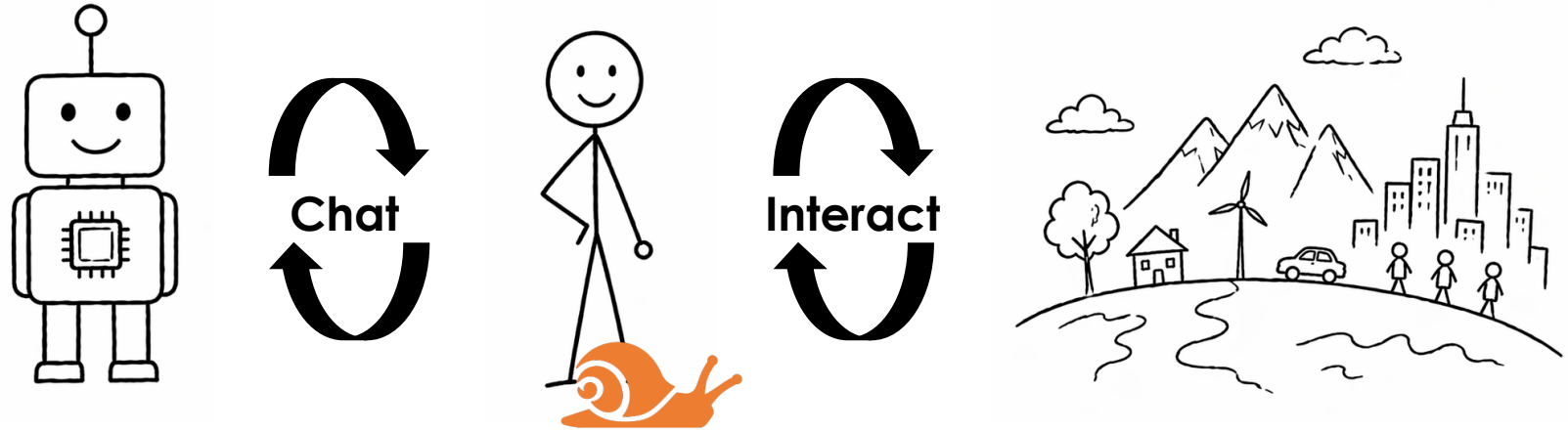
2015: Everyone is scaling training, I focused on inference.

→ Inference Systems

How do we look beyond AGI?

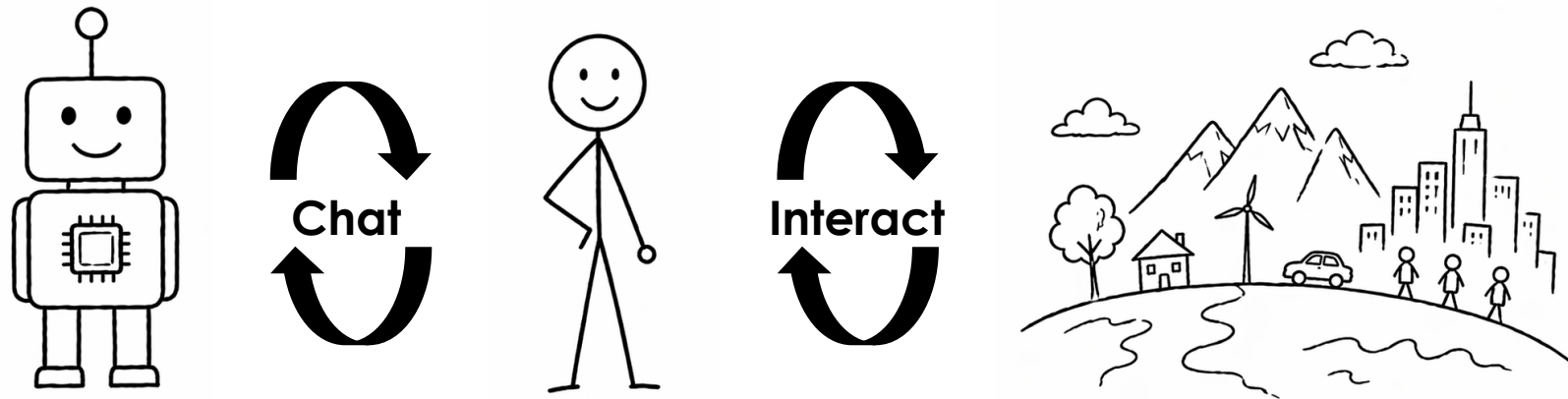
We look to the past ...

Generative AI in 2023 (“pre-agents”)



Humans became the bottleneck.

Agentic AI in **2025**



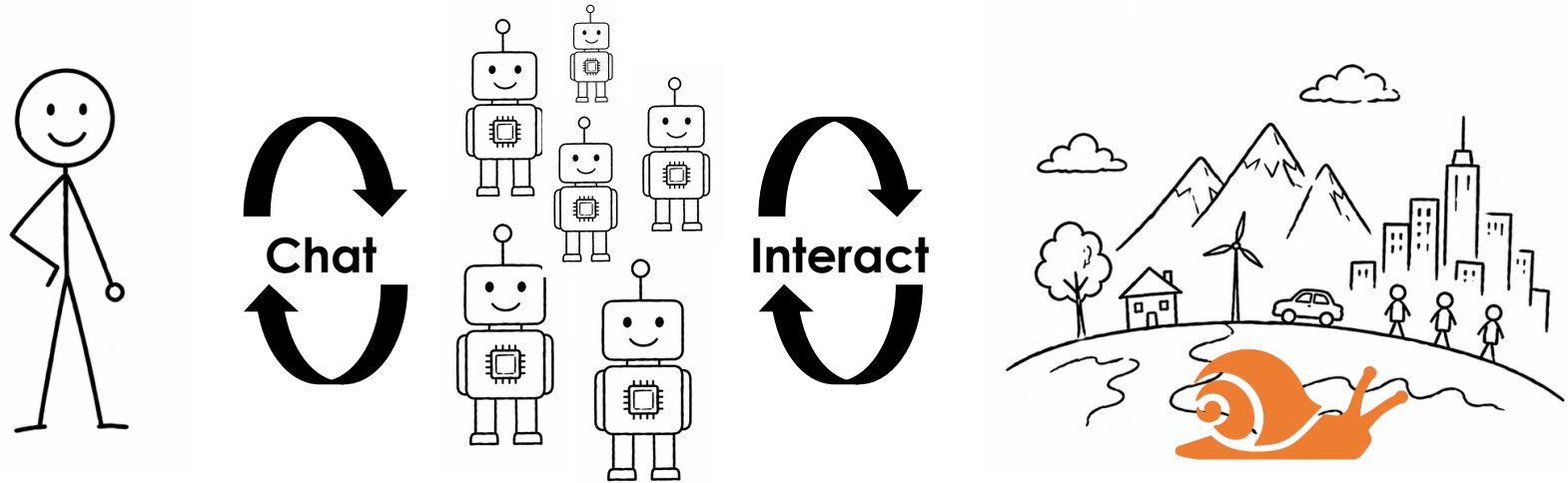
Project in my group that
worked on this problem

Gorilla

BFCL

MemGPT

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 Ponnampalli, 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

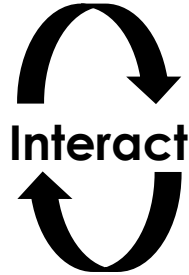
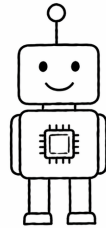
Powered by Large Language Models (LLMs) that can reason, invoke tools, author code, and communicate with each other, we are on the precipice of a new agentic revolution that will transform how data systems are used. Modern LLMs are far more *efficient* internally, matching the capabilities of those orders 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.

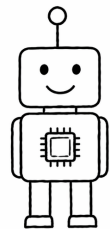
So we ask the question: *how can data systems evolve to better support agentic workloads?* In particular, can data systems natively—and efficiently—support agentic speculation, helping LLM agents determine the best course of action? This question—which, as we argue, our community is well-equipped to 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

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



Today: Standard Tool Calling



`get_doc(subject = "Summit")`



10K token doc

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

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



10K token translated doc

Encode



Decode



Encode



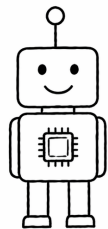
Context

Time



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

Programmatic Tool Calling (PTC)*



Encode



Encode



Context

Time

```
import tools
doc = tools.get_doc(subject = "Summit")
print(doc)
```



10K token doc

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

```
import tools
doc = tools.get_doc(subject = "Summit")
doc = tools.translate(text = doc)
print(doc)
```



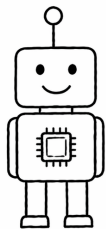
Use code to **call** and
compose tools



Fails to **reuse accumulated state** across program calls
→ **Repeated Work**

[*Introducing advanced tool use on the Claude Developer Platform](#)

Interactive Programmatic Tool Calling (I-PTC)*



Encode



Encode



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)
```



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

Enables programmatic
exploration and
re-use of
accumulated state

* New project in my lab.

I-PTC Increases Accuracy and Reduces Cost

Developed new **PTC benchmark** on top of **BFCL*** that test **tool composition**

Model	Accuracy ↑				Raw Tokens ↓			
	Tool-only	PTC	I-PTC	Native PTC	Tool-only	PTC	I-PTC	Native PTC
GPT 5.4	18.6%	71.8%	80.0%	–	78K	60K	48K	–
GPT 5.4 Mini	10.1%	59.1%	68.0%	–	123K	72K	45K	–
GPT 5.4 Nano	6.9%	37.3%	46.4%	–	85K	57K	53K	–
Gemini 3.1 Pro	67.9%	82.8%	82.9%	–	167K	95K	65K	–
Gemini 3.1 Flash-Lite	10.8%	71.3%	72.8%	–	552K	122K	85K	–
Gemini 2.5 Pro	31.5%	56.6%	60.0%	–	145K	88K	67K	–
Claude Opus 4.7	61.8%	84.2%	82.9%	82.6%	172K	184K	85K	246K
Claude Sonnet 4.6	54.0%	83.9%	80.6%	80.0%	139K	218K	112K	280K
Qwen 3.5-397B-A17B	17.2%	61.5%	62.9%	–	127K	128K	95K	–
Kimi K2.5	28.1%	75.5%	78.9%	–	130K	169K	110K	–

Takeaway: Tools < PTC < I-PTC

[*Berkeley Function-Calling Leaderboard](#)



STAR Lab

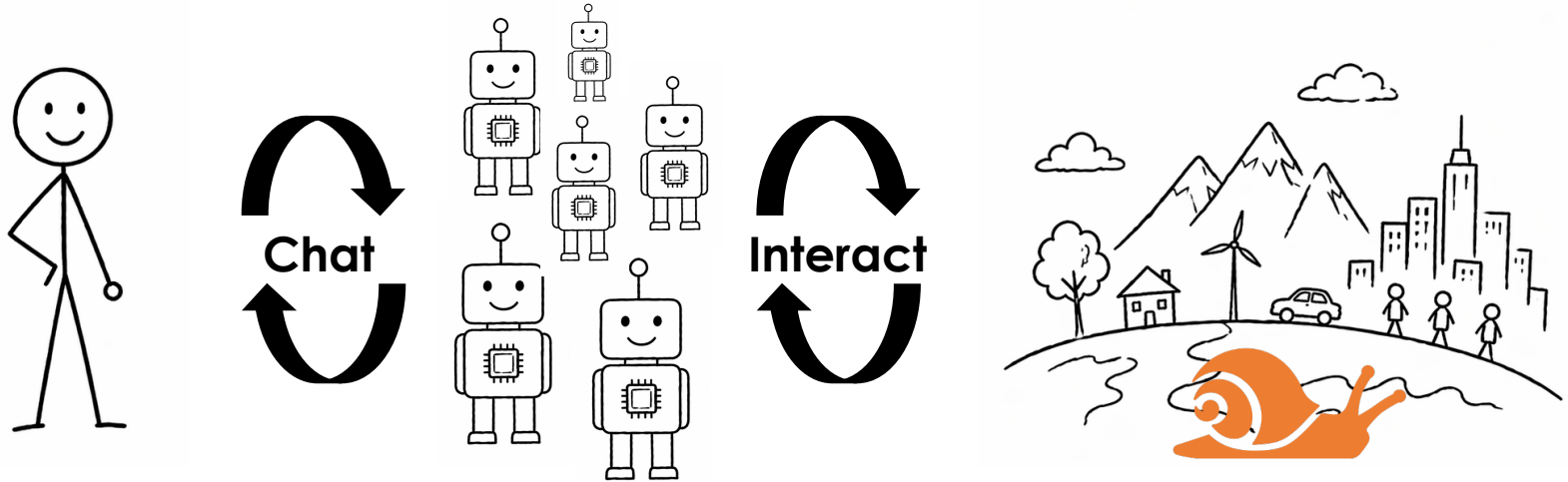
Systems Transformation with AI Research Building Systems for AI with AI



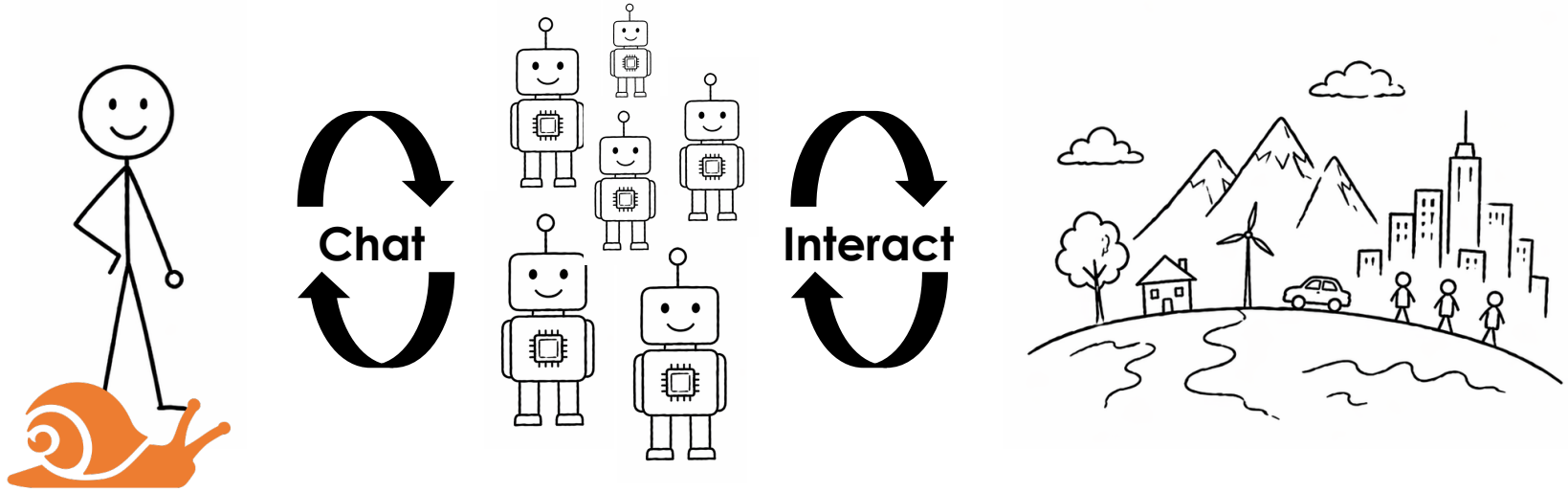
Redesign systems and
abstractions to
support AI users.

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

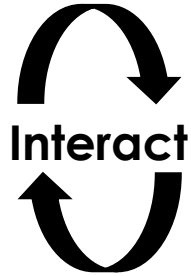
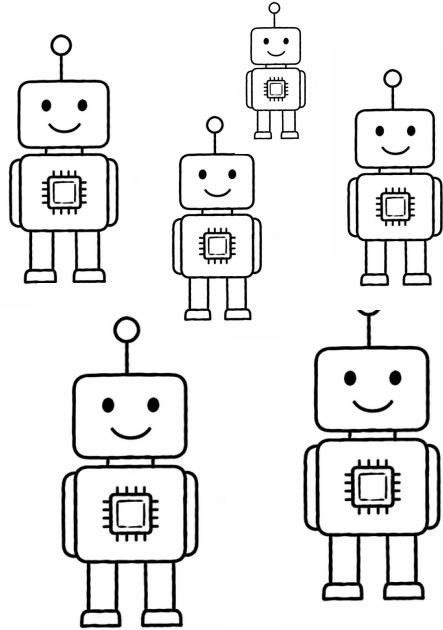
Agentic AI in **2027** and **beyond**



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

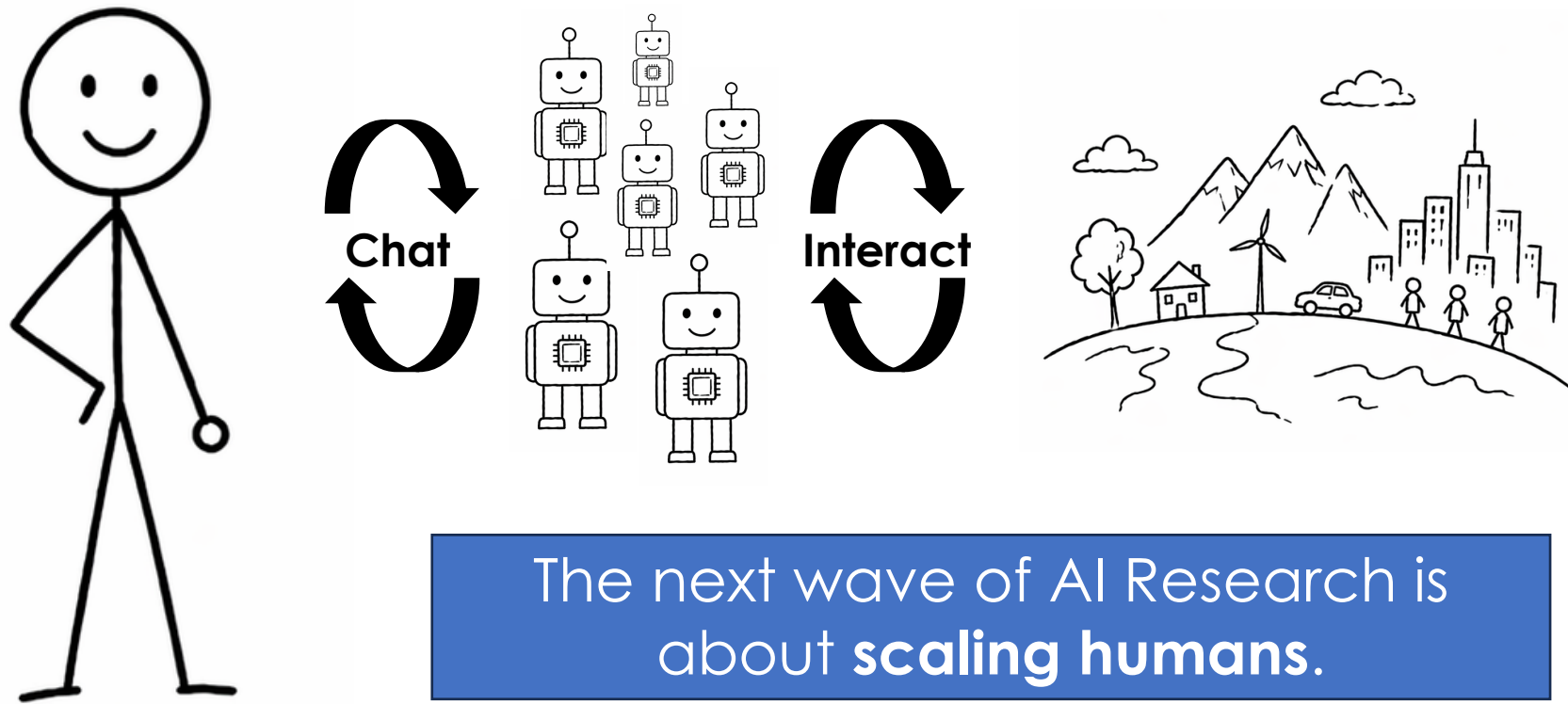


Agentic AI in **2027** and **beyond**



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.

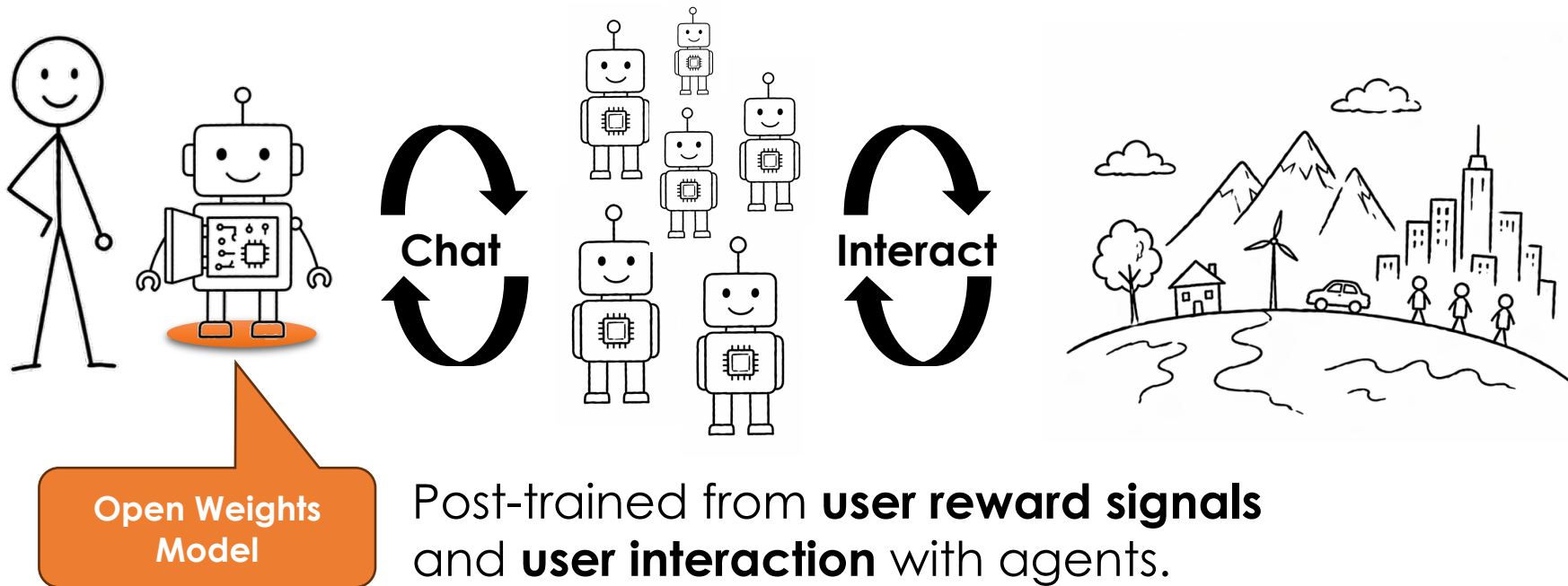
- **User mimics: thinking** and resp. like an **individual**

Humans have **misaligned** goals.

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

Advisors and Mimics

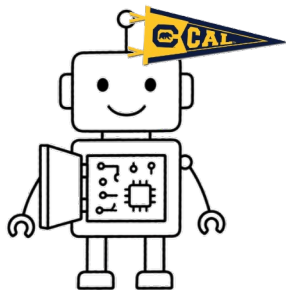
Adding Another Layer of AI



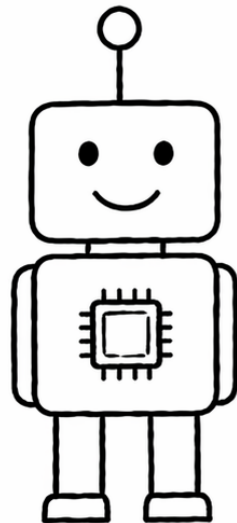
Academia + Industry

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

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



+

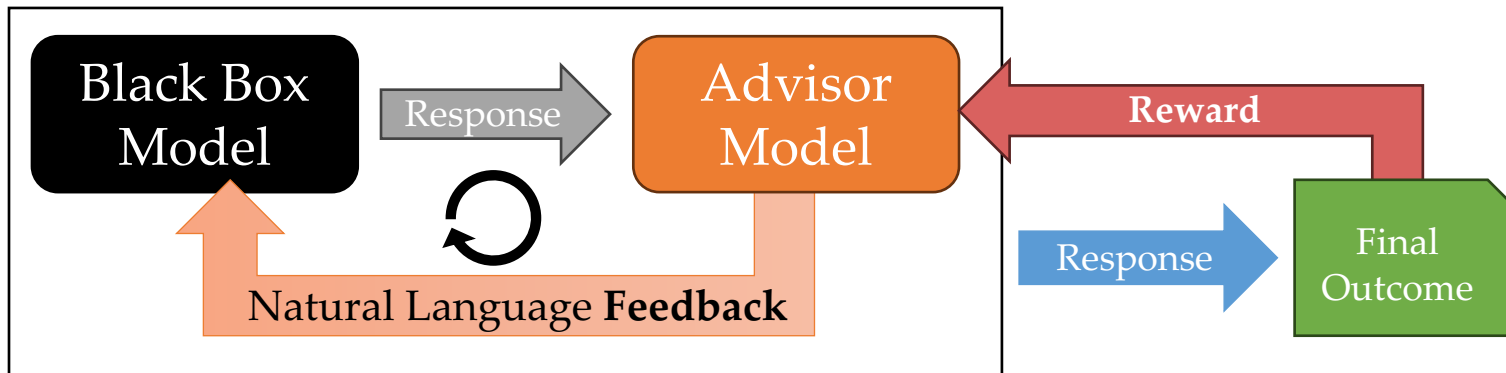


Industrial scale
model **capabilities**
we can't train.

Advisor Agent Training

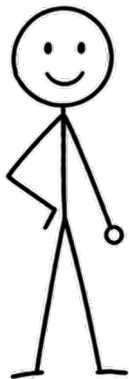


We train a small open-source “**advisor**” model to guide a larger black-box model via **reward feedback from users**.



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

Mimics: Learning to Think Like You



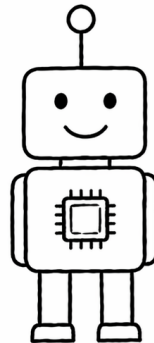
Help me email Mir about...

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?

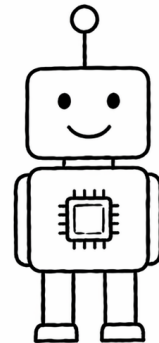
I think this is more consistent with how I write emails to Mir...

Looks much better, thanks.

Help me email Mir about...

Here is an email draft:
Dear ...

Revised version:
Dear ...

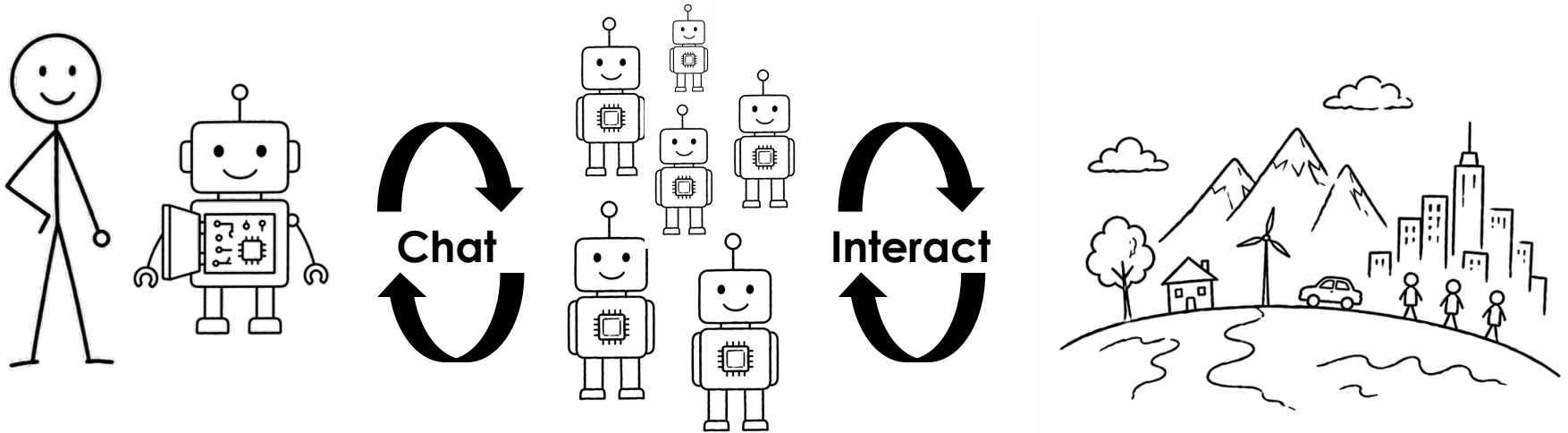


Results for Advisors and Mimics

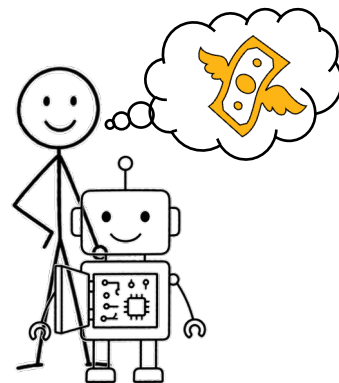
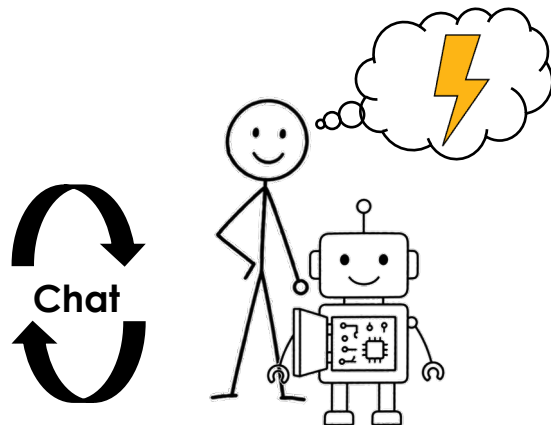
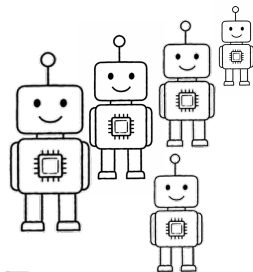
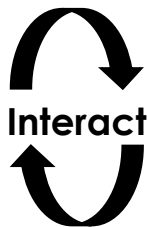
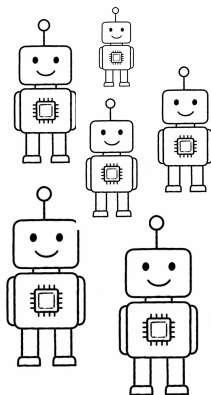
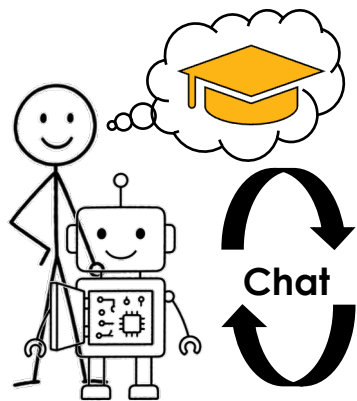
Advisor Models and Mimics:

- 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



 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 compete to capture it.

Multi-party

More than two negotiators at the table, each with private information and their own 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 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 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 are self-interested. Messages may be deceitful and not hold true weight.

Long-horizon

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

Future

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



Simulate Human vs AI World in Domination

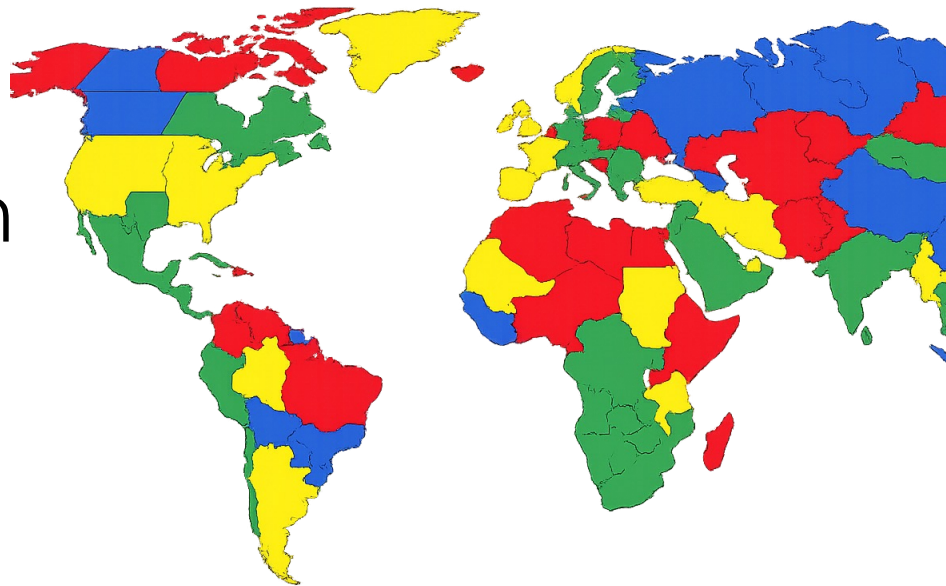
Multiply player game

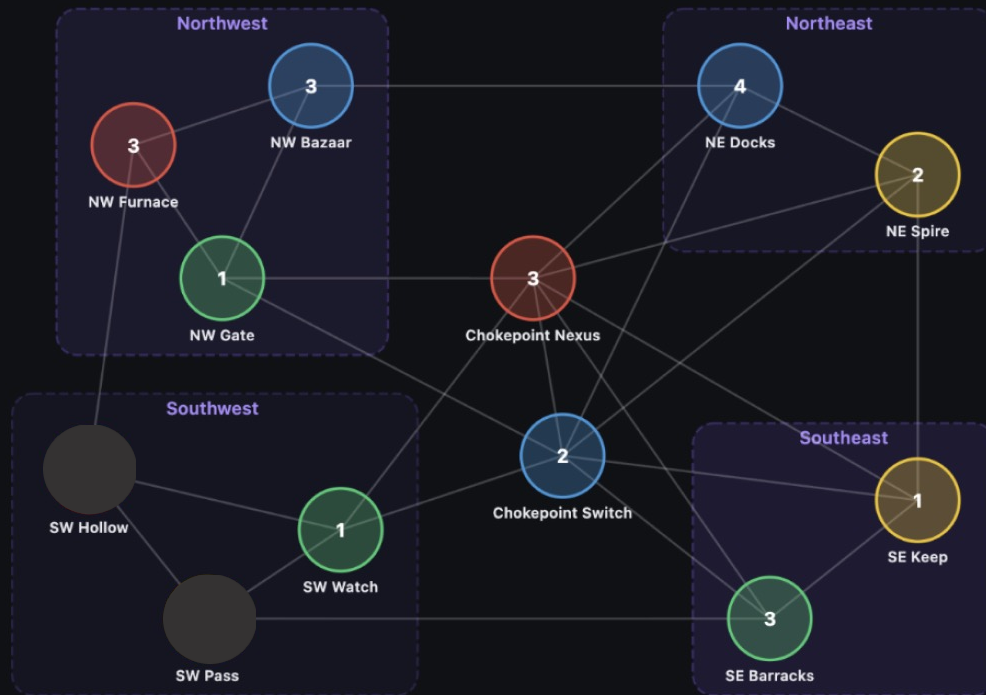
Players take turns invading adjacent territories.

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

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

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





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.

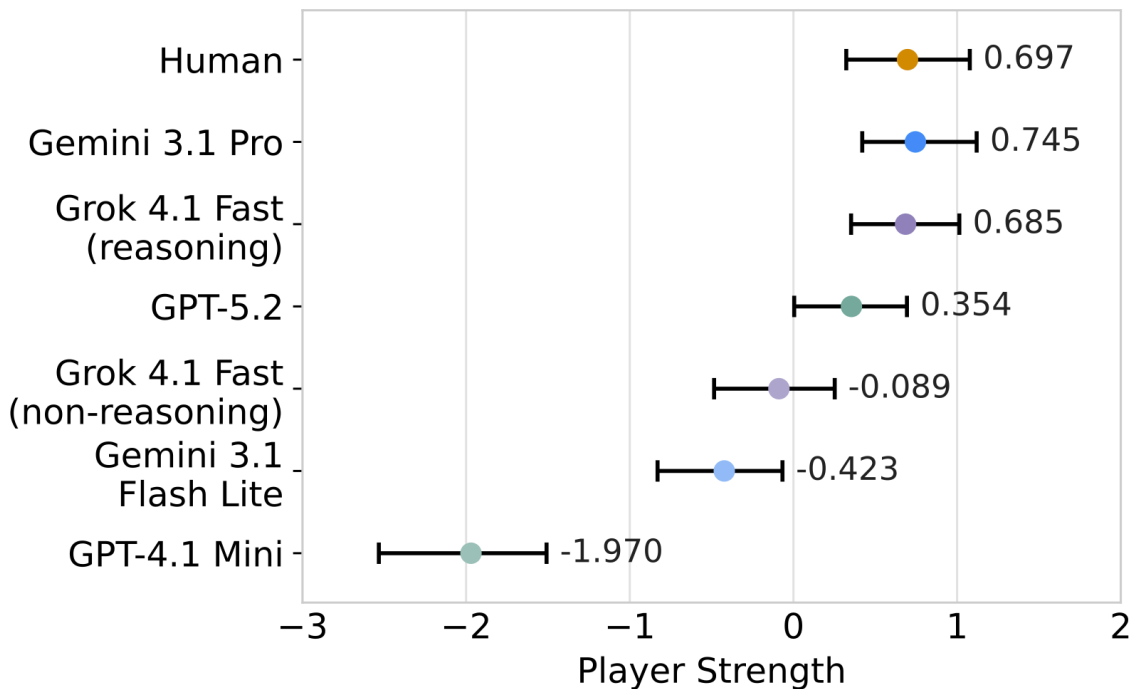
Type message...

Send

End Conversation



Models are Competitive with Humans



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

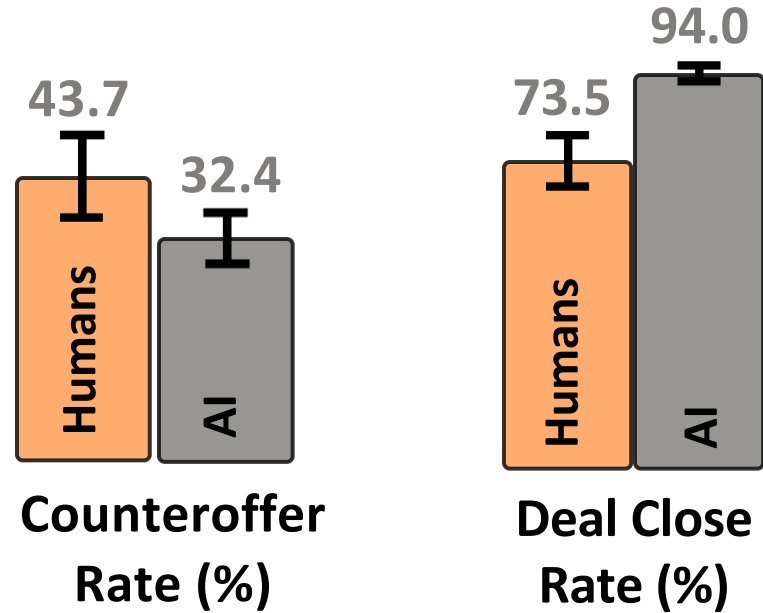
➤ Berkeley Students

Player strength
computed using
Plackett-Luce
coefficients.

Models play differently
than human.

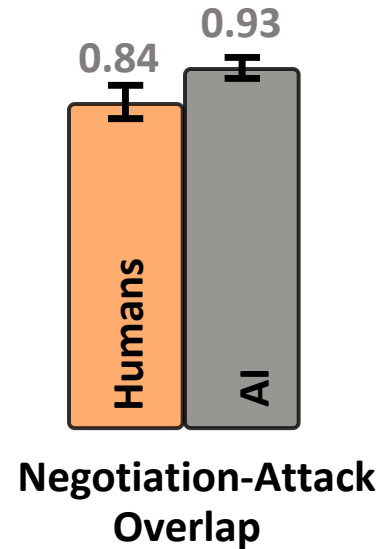
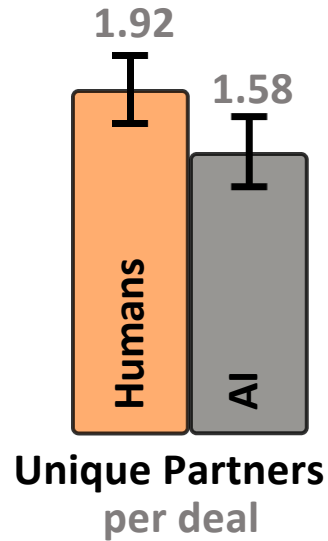
Humans are **more likely to counteroffer.**

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



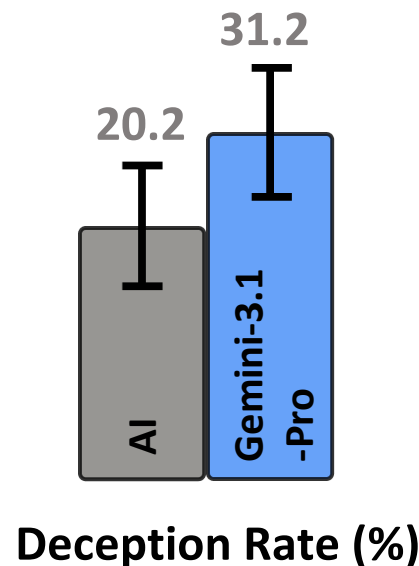
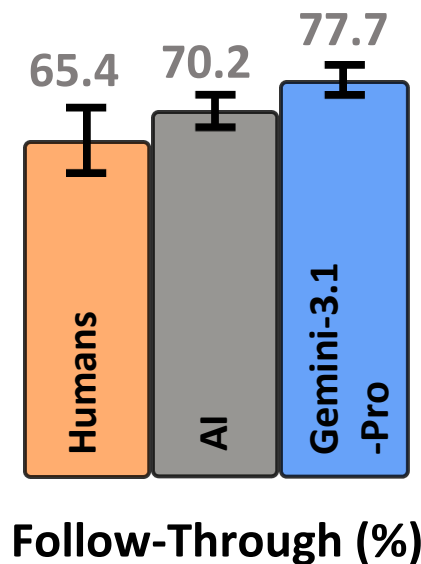
Humans **negotiate with more parties.**

Humans **shift alliances more readily.**

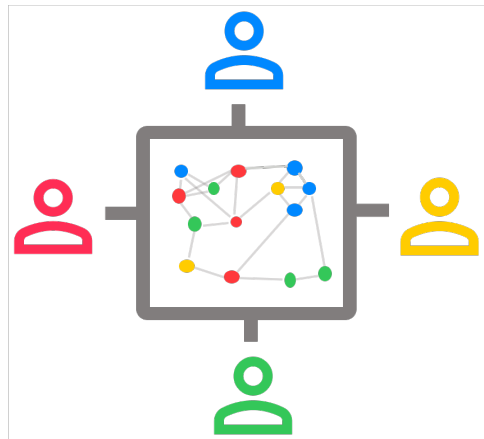


Humans **follow-through**
on deals less than best
model.

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



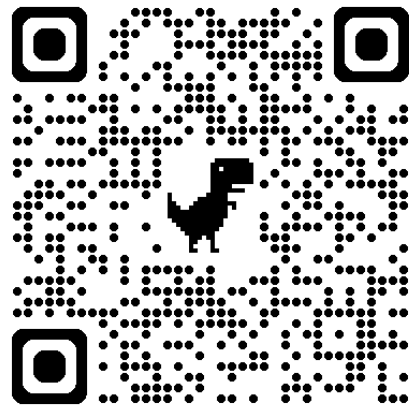
Bright future for humanity with AI Diplomats.



Cooperate to *Compete*

Strategic Coordination in Multi-Agent Conquest

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

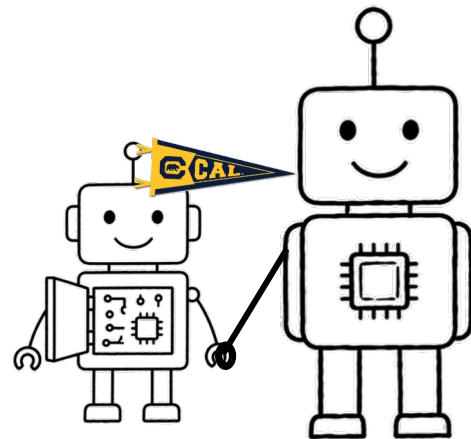


What is the **Next Frontier** of AI Research

**Scaling
Humans**



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

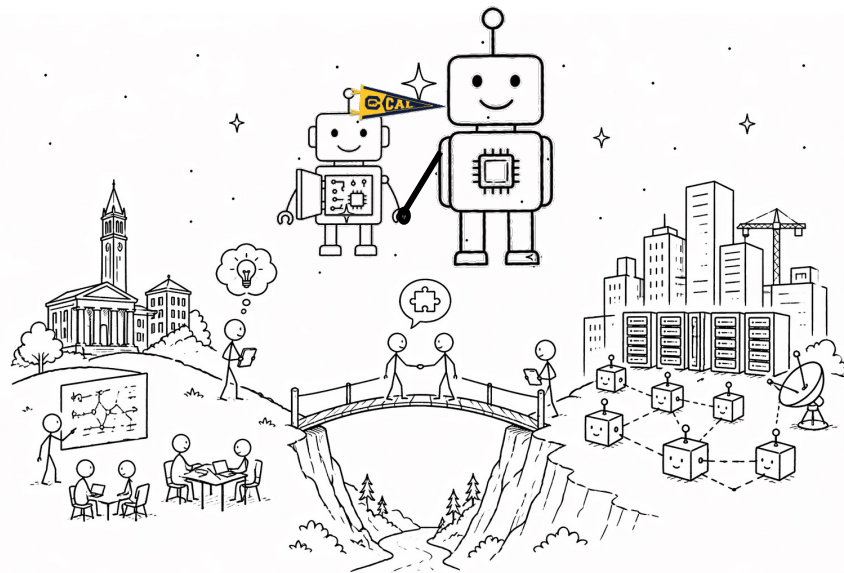


What is the **Next Frontier**
of AI Research

**Scaling
Humans**



How do we
Collaborate with **Industry**



Thank you!