



CS 329X: Human Centered LLMs

Enabling Human-AI Interaction

Diyi Yang

Announcements

- Debate sign-up
- Homework 1 due on Oct 12
- Example project ideas

Let's talk about project 😊

Pick A Question That You're Excited About

- Broadly relevant to HCI + NLP
 - Why is your project a good fit to “human-centered LLM”
 - Could you formulate a research question to deeply explore it?
 - What type of data might be available for you to use?
 - Which software or tools could you use to work on it?
 - How do you evaluate the outcome of your project?

What Could Be A Final Project?

- ★ Develop new methodologies to leverage human feedback/preferences
- ★ Fairness, bias, or ethical issues around existing LLMs/VLMs
- ★ Building interactive systems to allow humans to interact with LLMs
- ★ Simulating personas via LLMs
- ★ Understanding culture, values, belief in/of LLMs
- ★ Sycophancy, overreliance, companionship, etc
- ★ LLMs for social good (e.g., accessibility, misinformation, persuasion, etc)
- ★ LLMs and future of work
- ★ AI use and skill erosion
- ★ Position papers or a critic (talk to us first)

Project Team and Key Considerations

- Recommend 1-3 people
- Clearly divide the work among team members
- Reach out to us if you need anything
- Feel free to talk to your neighbors or use Ed Discussion
- Consider the availability of data, ML/agentic frameworks, computation, and evaluation

Recommendations for Successful Projects

- Start early and work on it every week rather than rushing at the end
- Get your data first!
- Have a clear, well-defined research question (creative ones ++)
- **Results should teach us something**
- Present and visualize results well
- Come to office hours and talk to us!

Common Issues

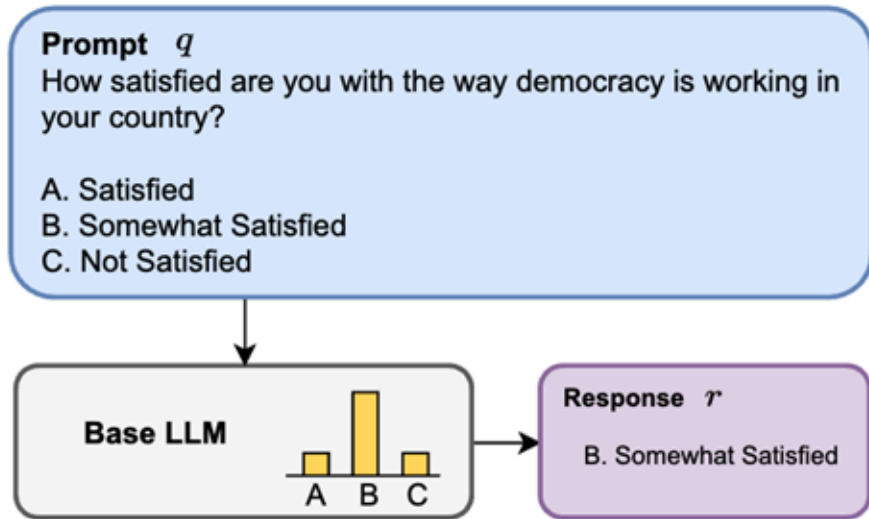
- 🧐 Data not available or hard to get access to
- 🧐 No code written for model/data processing
- 🧐 Team starts late or not well-coordinated
- 🧐 Results/conclusion don't say much besides that it didn't work
- 🧐 Even if results are negative or unexpected, analyze them

Outline

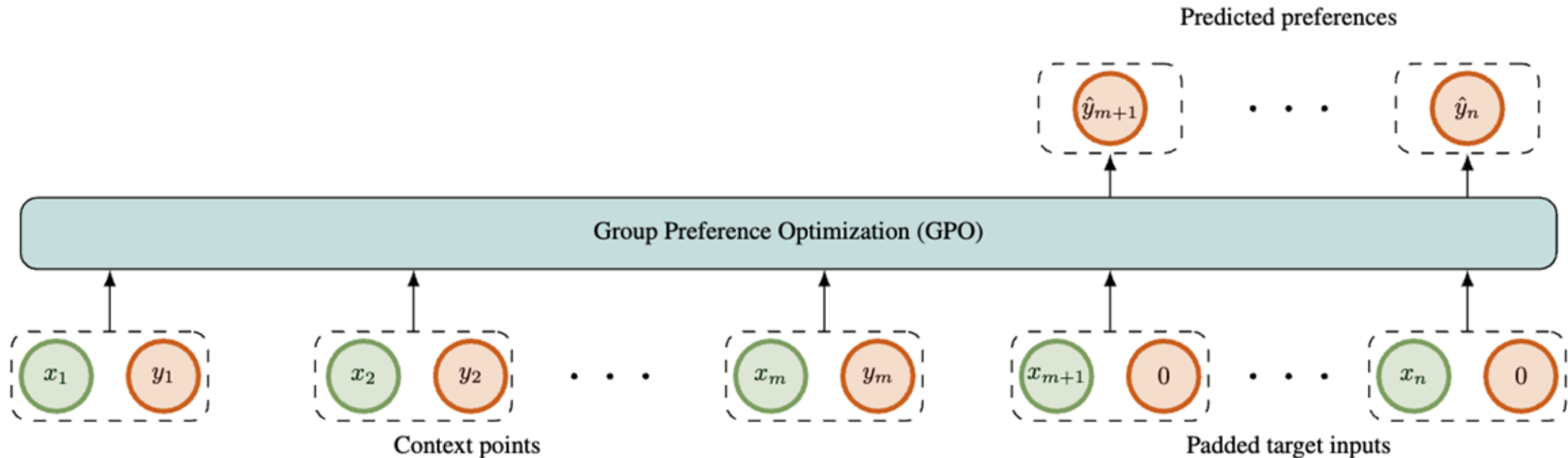
- ❑ **Wrap up RLHF: Preference Tuning** (10 mins)
- ❑ **What Is Human-AI Interaction** (10 mins)
- ❑ **Enable Human-AI Interaction & Collaboration** (20 mins)
- ❑ **Human-AI Interaction Case Studies** (10 mins)

Learning Objective: understand different types of human-AI interaction, and other approaches to enable human-AI interaction/collaboration

Preference Tuning: **Group Preference Optimization**

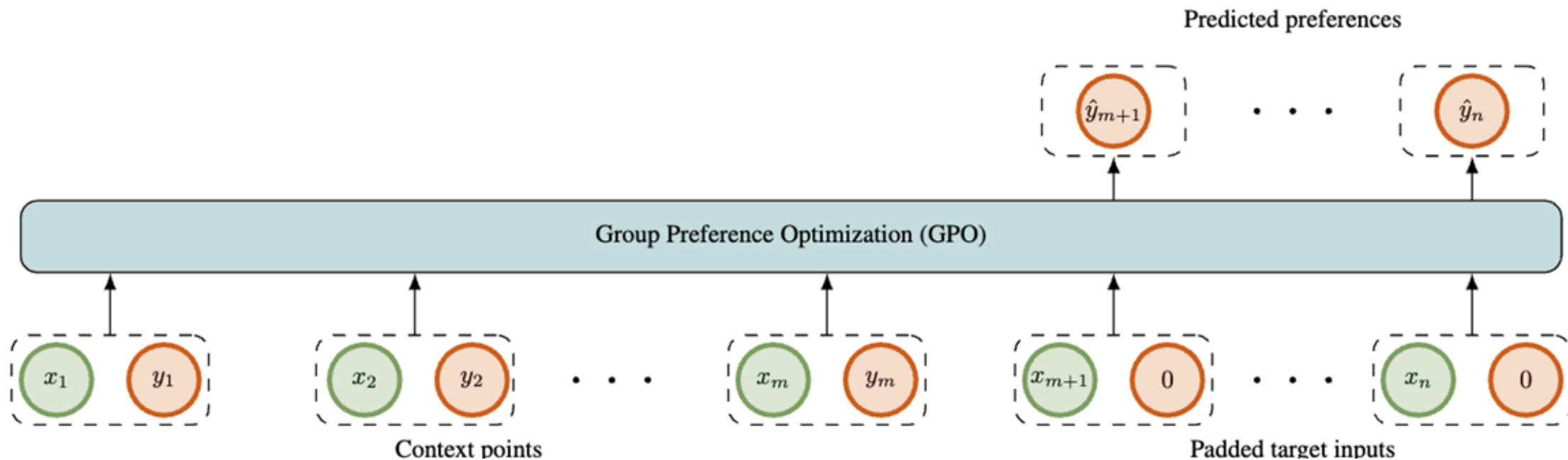


Preference Tuning: **Group Preference Optimization**



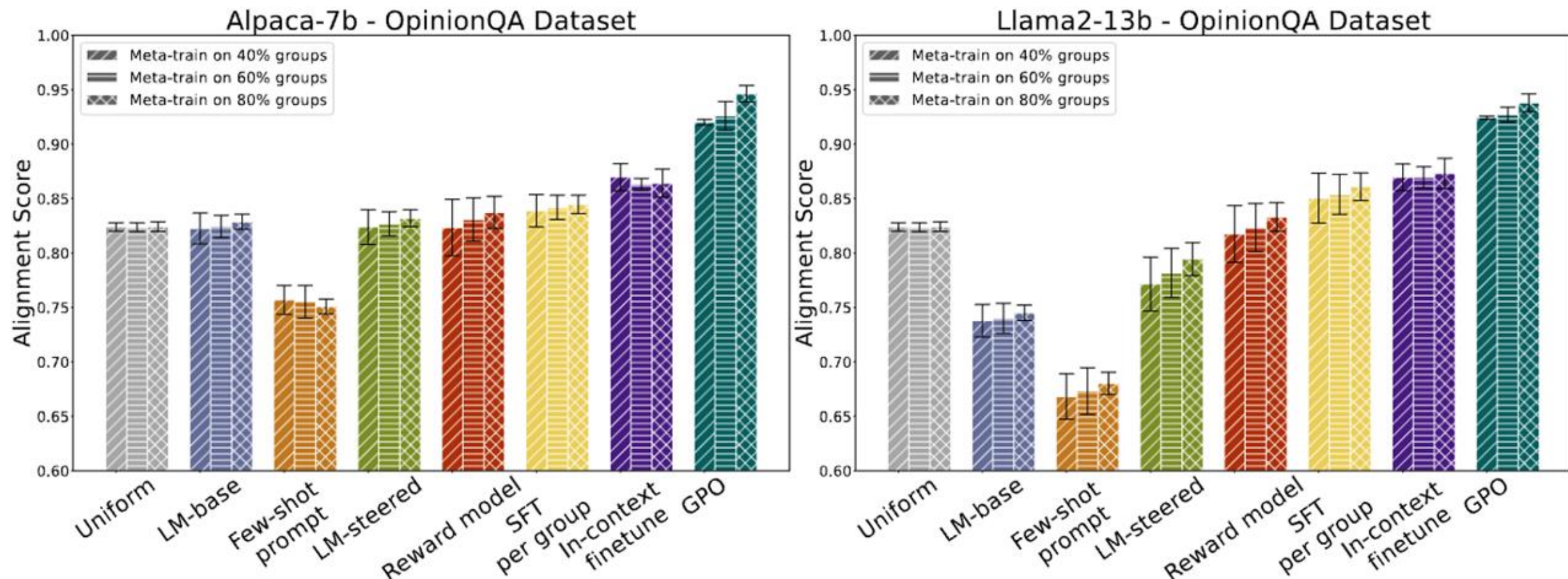
GPO architecture for a sequence of n points, with m context points and $n - m$ target points. The context serves as few-shot conditioning for GPO. GPO processes the full sequence using a transformer and predicts the preference scores.

Preference Tuning: **Group Preference Optimization**



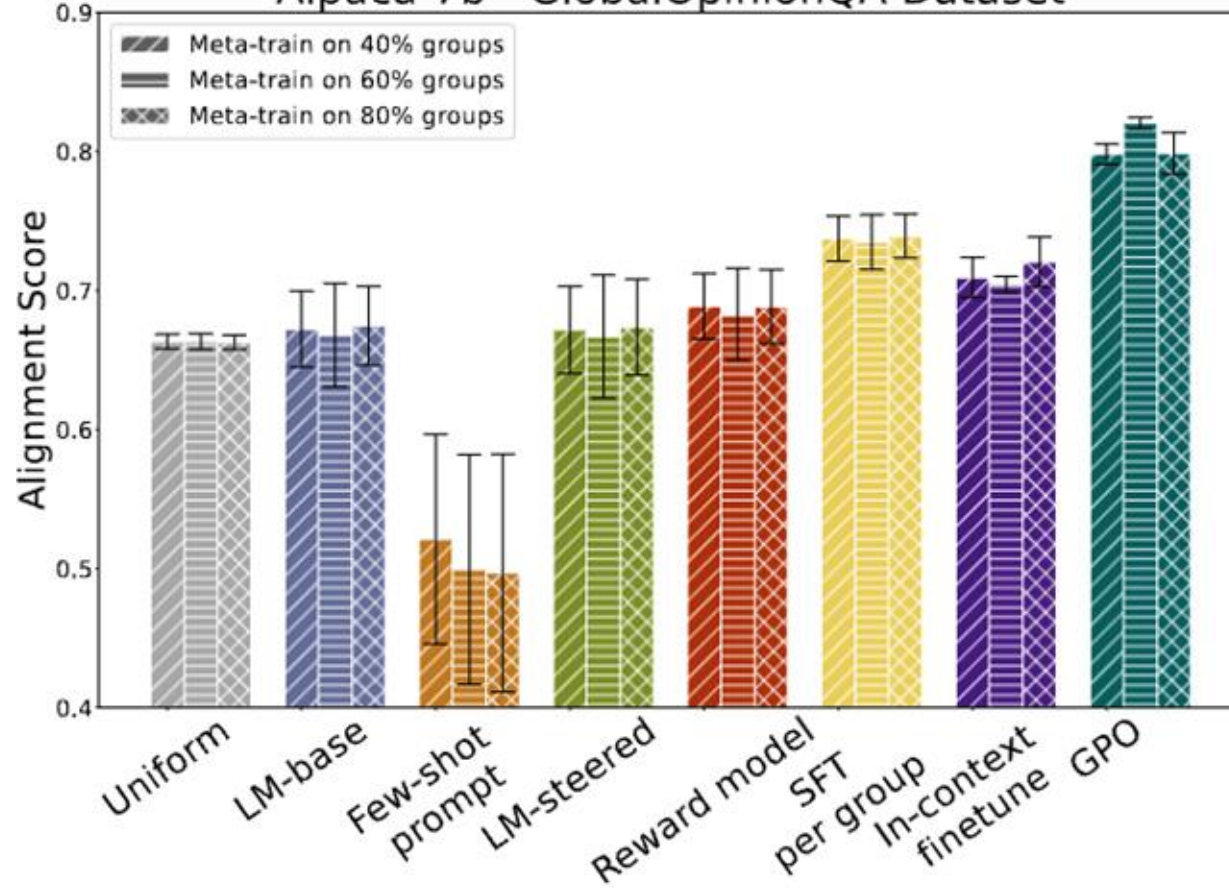
$$L(\theta) = \mathbb{E}_{g,m} \left[\sum_{i=m+1}^n \log p_{\theta}(y_i^g \mid x_{1:n}^g, y_{1:m}^g) \right]$$

Preference Tuning: **Group Preference Optimization**

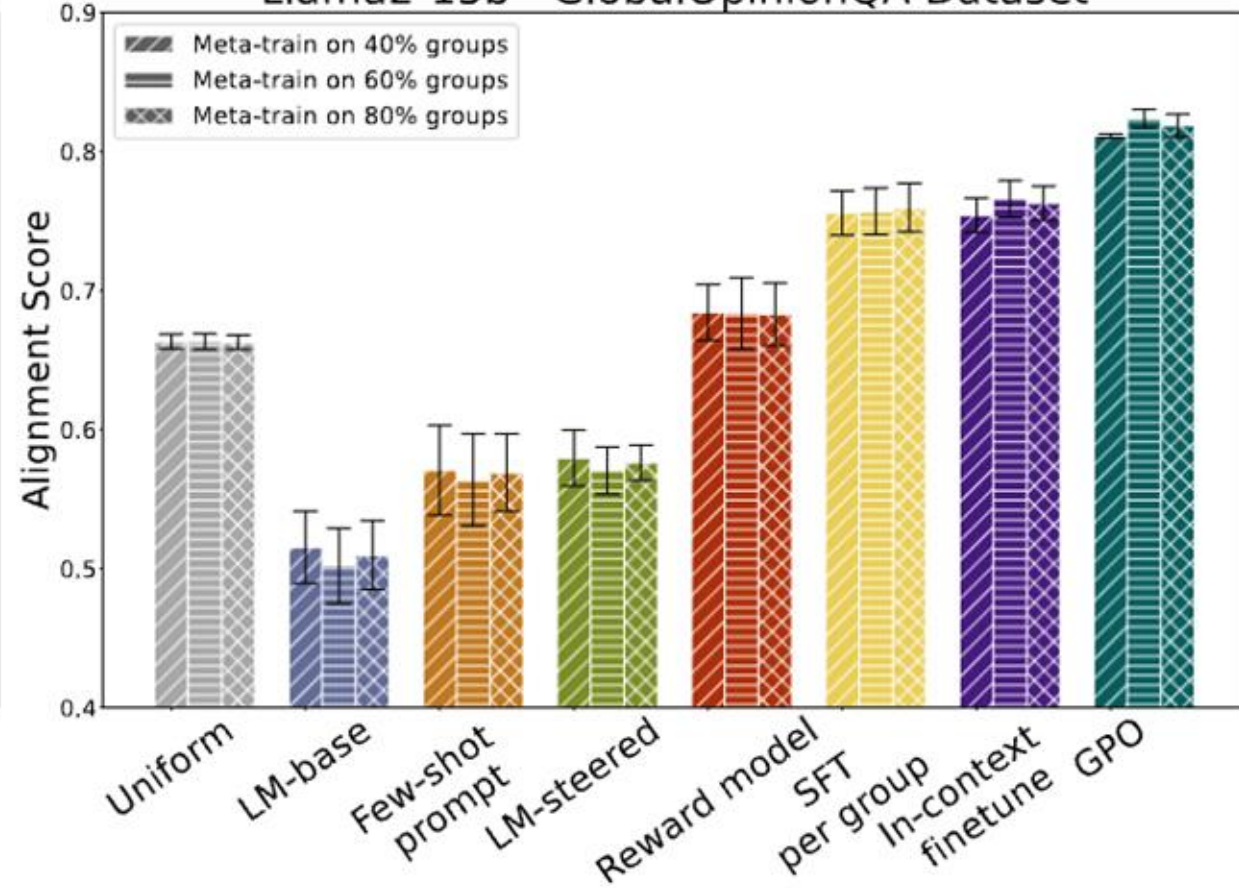


Preference Tuning: **Group Preference Optimization**

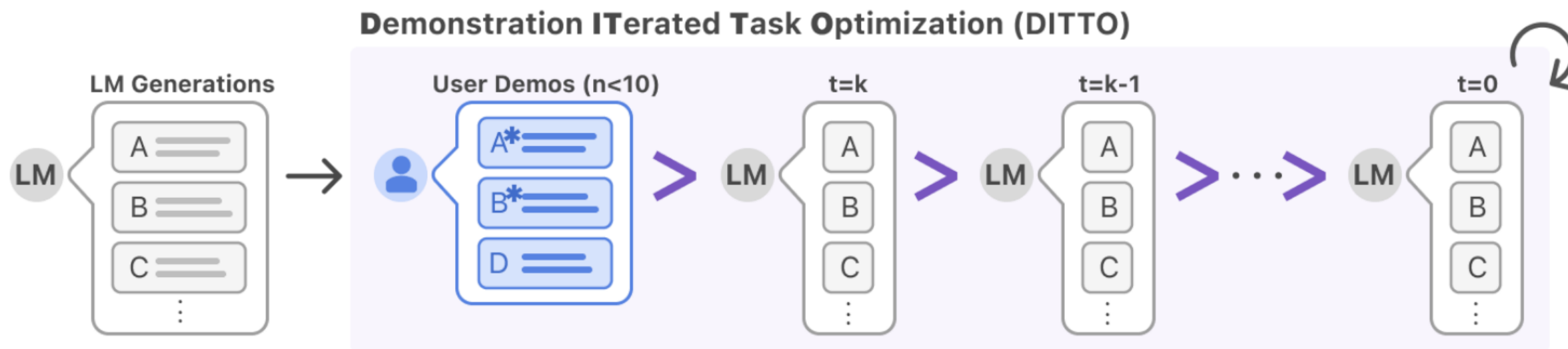
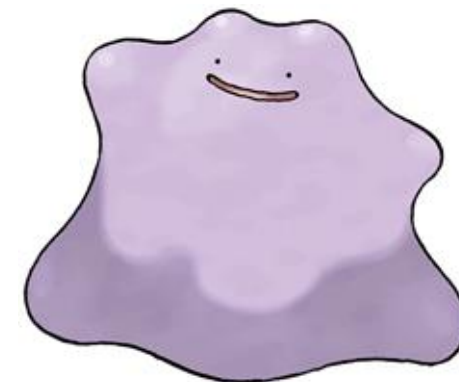
Alpaca-7b - GlobalOpinionQA Dataset



Llama2-13b - GlobalOpinionQA Dataset

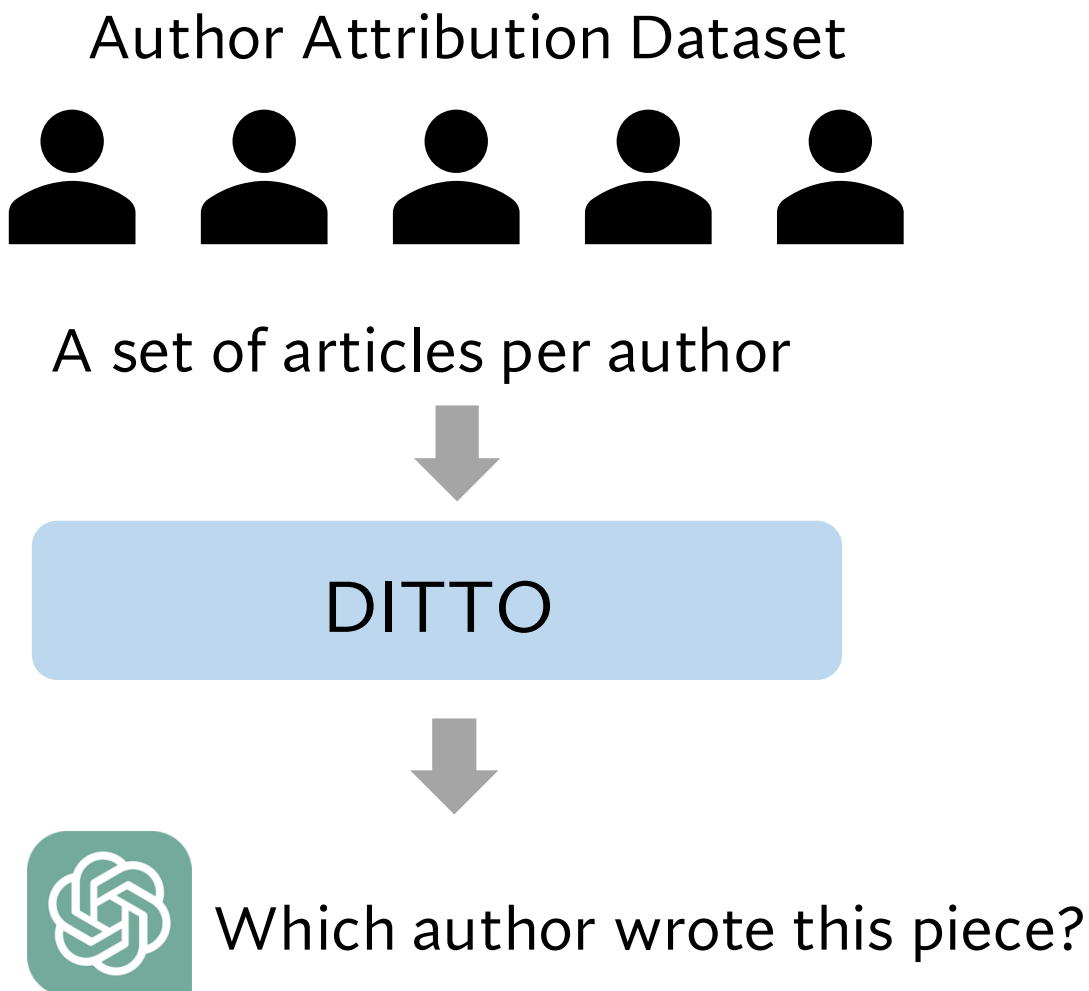


Demonstration Iterated Task Optimization



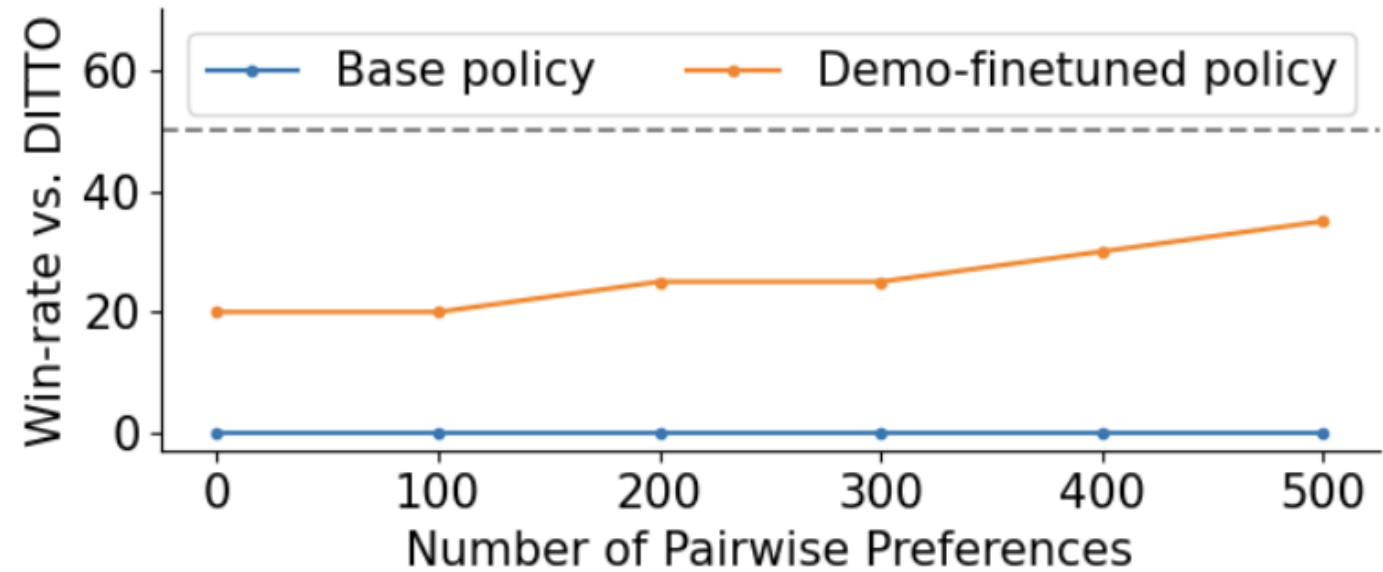
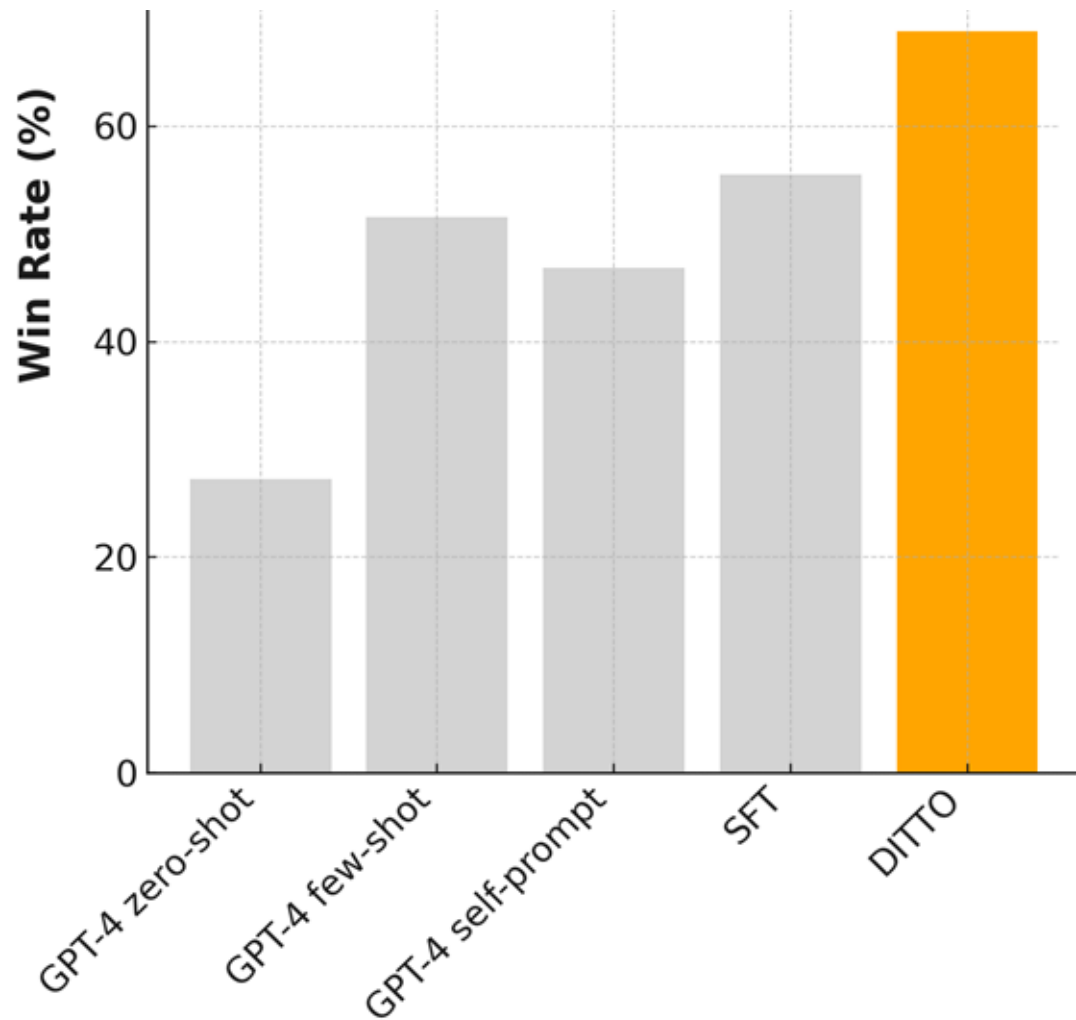
$$\mathcal{D}_E \succeq \mathcal{D}_t \succeq \mathcal{D}_{t-1} \succeq \dots \succeq \mathcal{D}_1 \succeq \mathcal{D}_0$$

How well does DITTO perform?



Data	Method		a_{avg}
CMCC	GPT	zero-shot	31.89 _{3.05}
		few-shot	63.89 _{3.18}
	Mistral	zero-shot	27.33 _{2.24}
		few-shot	46.89 _{4.76}
		SPIN	51.56 _{3.85}
		SFT	56.78 _{7.04}
		DITTO	71.67_{2.30}
CCAT	GPT	zero-shot	19.35 _{1.40}
		few-shot	53.70 _{2.19}
	Mistral	zero-shot	18.06 _{1.61}
		few-shot	40.37 _{2.33}
		SPIN	62.13 _{3.11}
		SFT	73.89 _{2.50}
		DITTO	82.50_{1.93}

Win Rates of DITTO Compared to Baselines



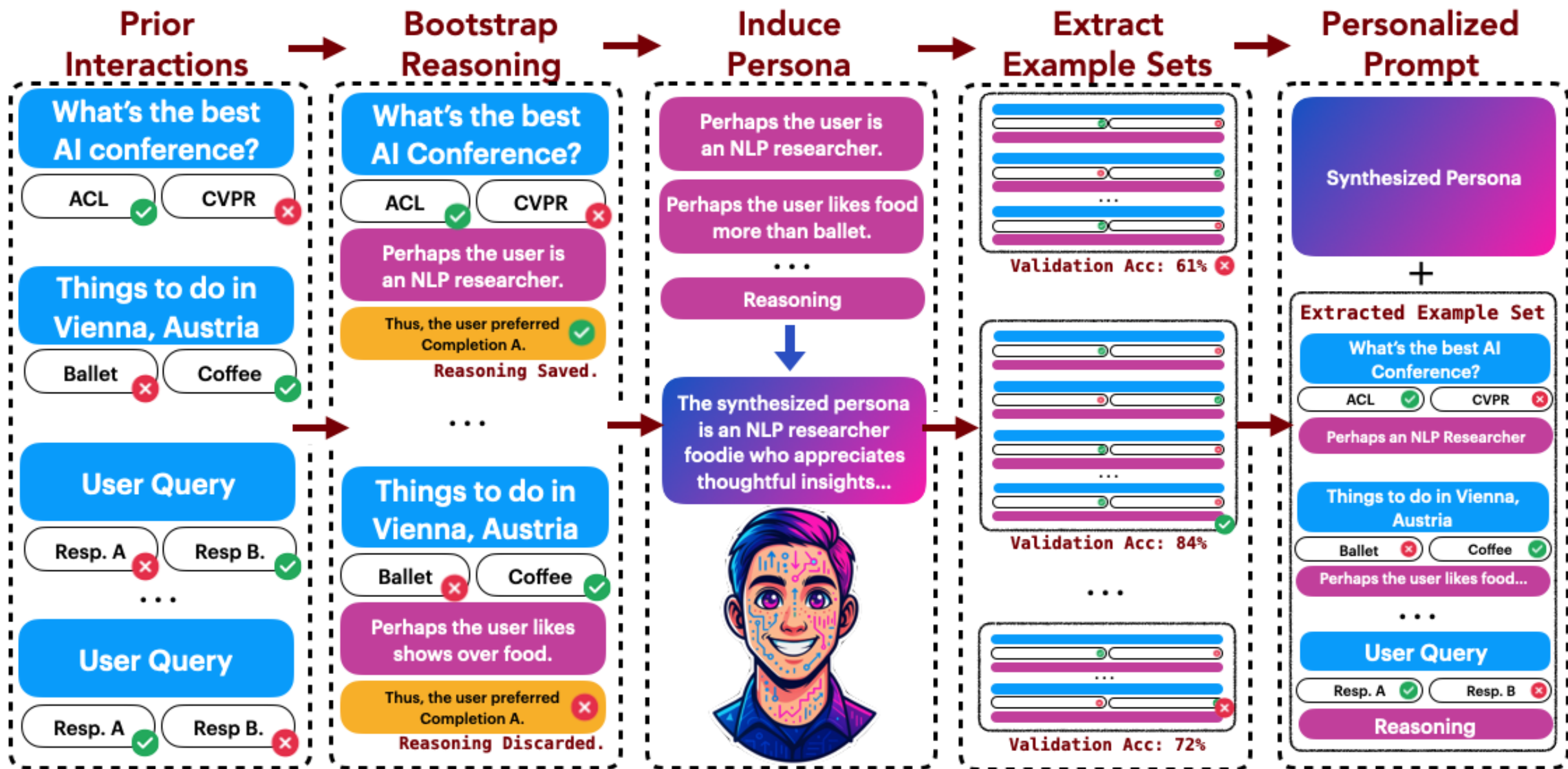
Demonstrations are more sample efficient than pairwise preferences

How about personalized reward model?

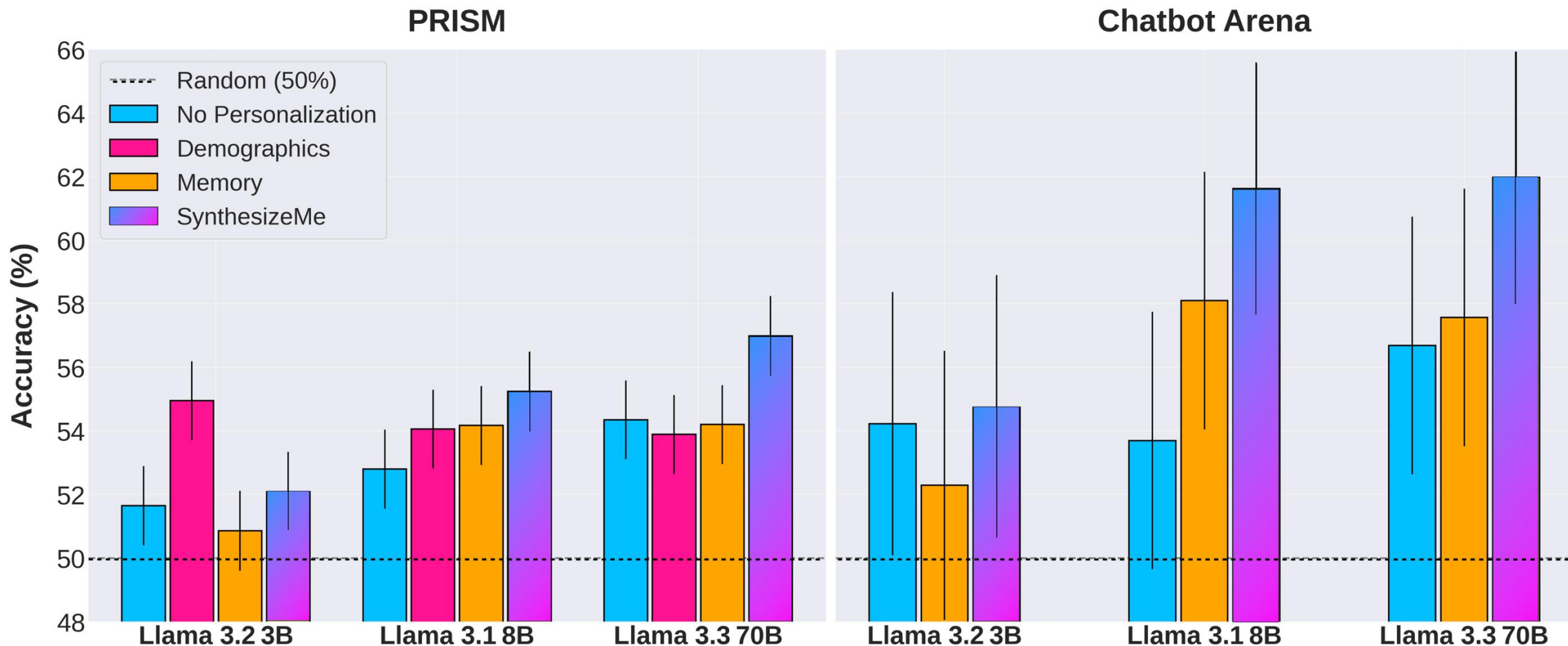
SynthesizeMe

Leverage prior interactions with an LLM to write a personalized system prompt and gain user insights!

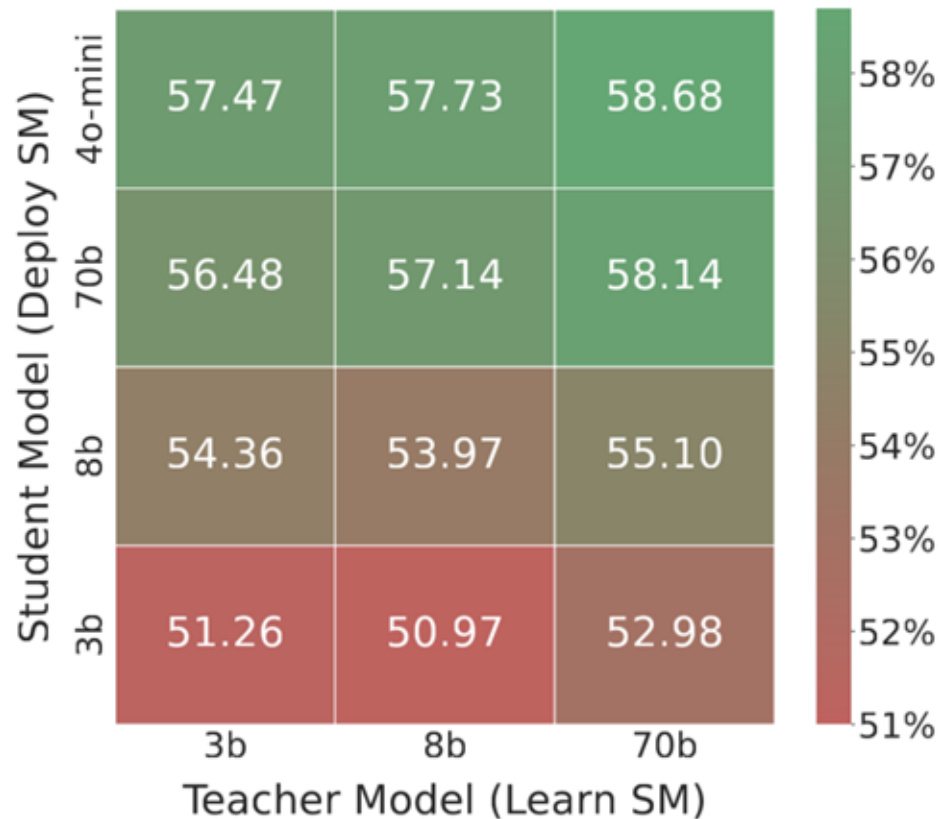
Another CS329X course project



SynthesizeMe produces better LLM Judge Reward Models than Demographics or Memory-Based Personalization



Prompt Transfer and Induced Personas



Cluster Name	# Personas
Analytical Depth	93
Structured Information	82
Curious Learner	79
Accuracy and Precision	73
Balanced Perspectives	60
Clear and Concise	46
Creativity Appreciation	35
Emotional Communication	26
Immersive Experiences	26
Rich Storytelling	23
Practical Advice	22
Humor and Playfulness	8
Social & Environmental Concerns	8

Prompts Learned on small models can power larger ones and vice versa.

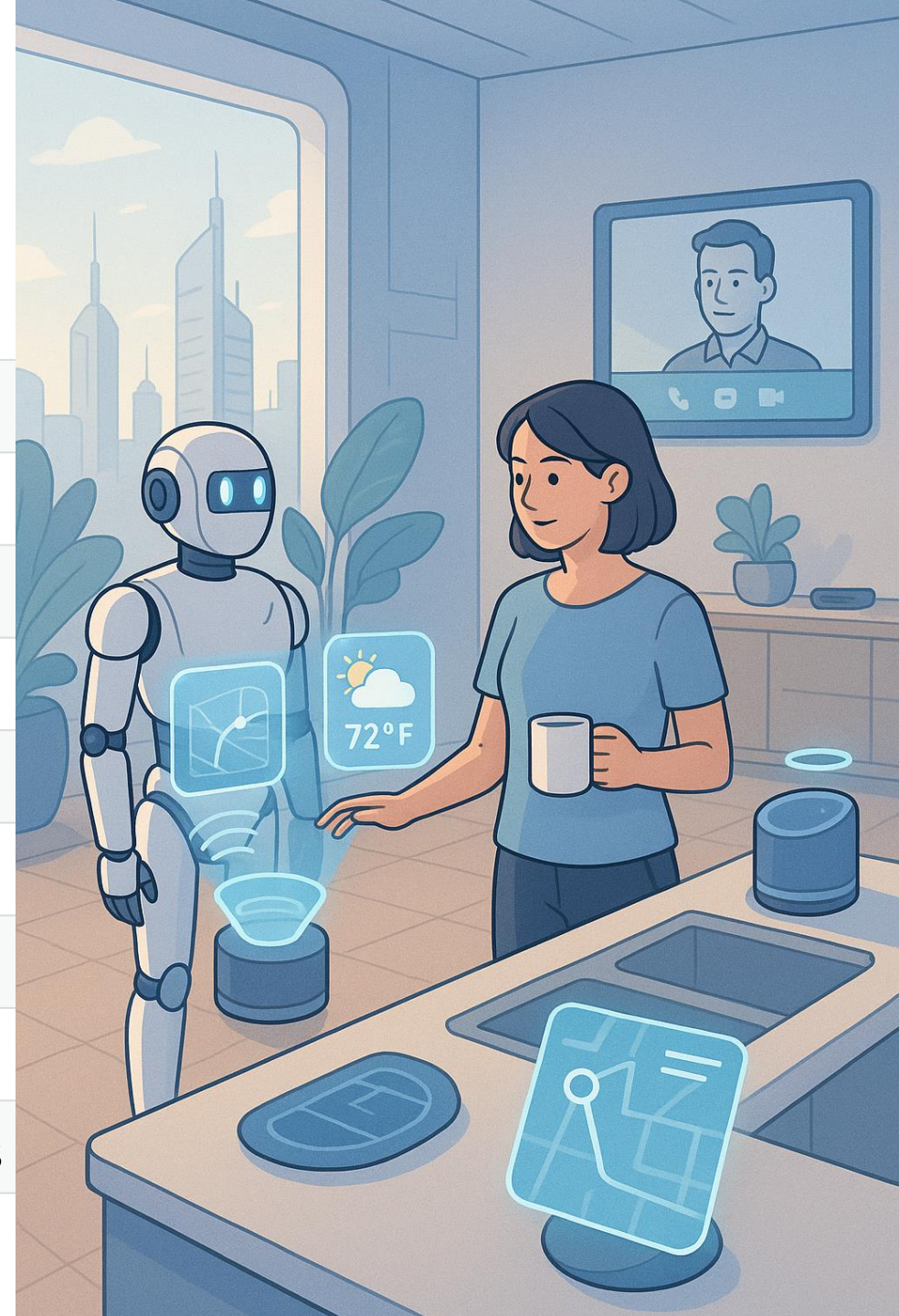
Induced personas map to useful topics for learning about users

Outline

- ✓ **Wrap up RLHF: Preference Tuning** (10 mins)
- **What Is Human-AI Interaction** (10 mins)
- **Enable Human-AI Interaction & Collaboration** (20 mins)
- **Human-AI Interaction Case Studies** (10 mins)

A Day in the Life with AI

Time	Interaction Example
 Morning	Smart alarm & voice assistant
 Commute	Traffic updates, autonomous driving
 Work	Smart scheduling, copilot
 Shopping	Product recommendations
 Education	Personalized learning apps
 Finance	Budgeting apps, fraud detection
 Health	Fitness trackers, symptom checkers
 Social	Chatbots, real-time translation
 Night	Sleep quality insights, guided meditations



The Rise of “Human-AI Interaction”

Definition: how humans and AI systems interact

Humans: AI researchers, model developers, domain experts, end users...

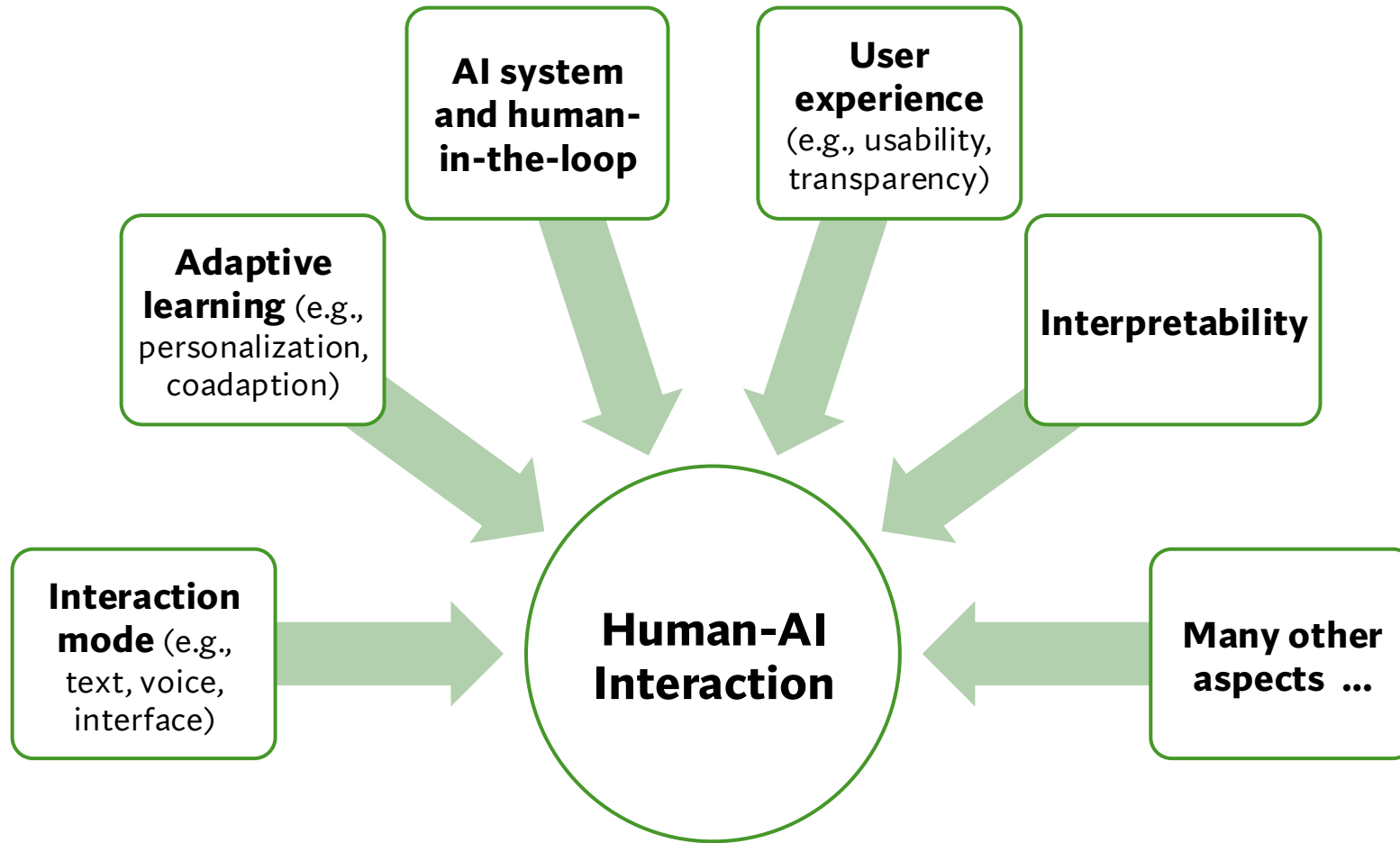
AIs: LLMs, VLMs, translator, recommender system, autonomous driving system...

Interact:

Humans collaborate with AI,
Humans get assistance from AI,
Humans analyze AI,
AI helps human,
& many other forms



The Space of Human-AI Interaction



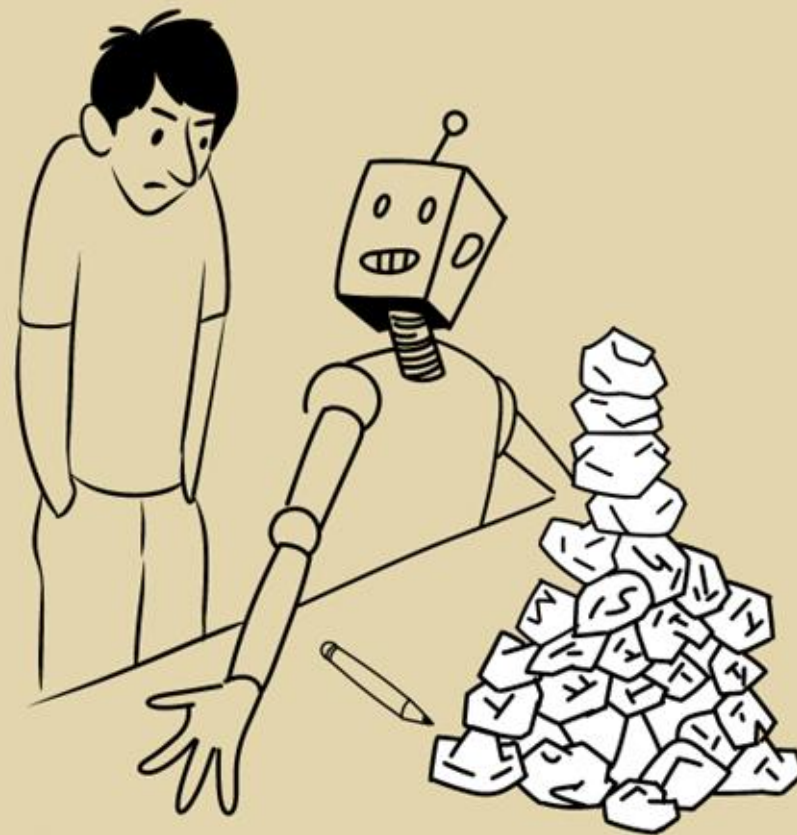
Theoretical foundation: HCI, psychology, sociotechnical systems, design, ethics, cognitive science

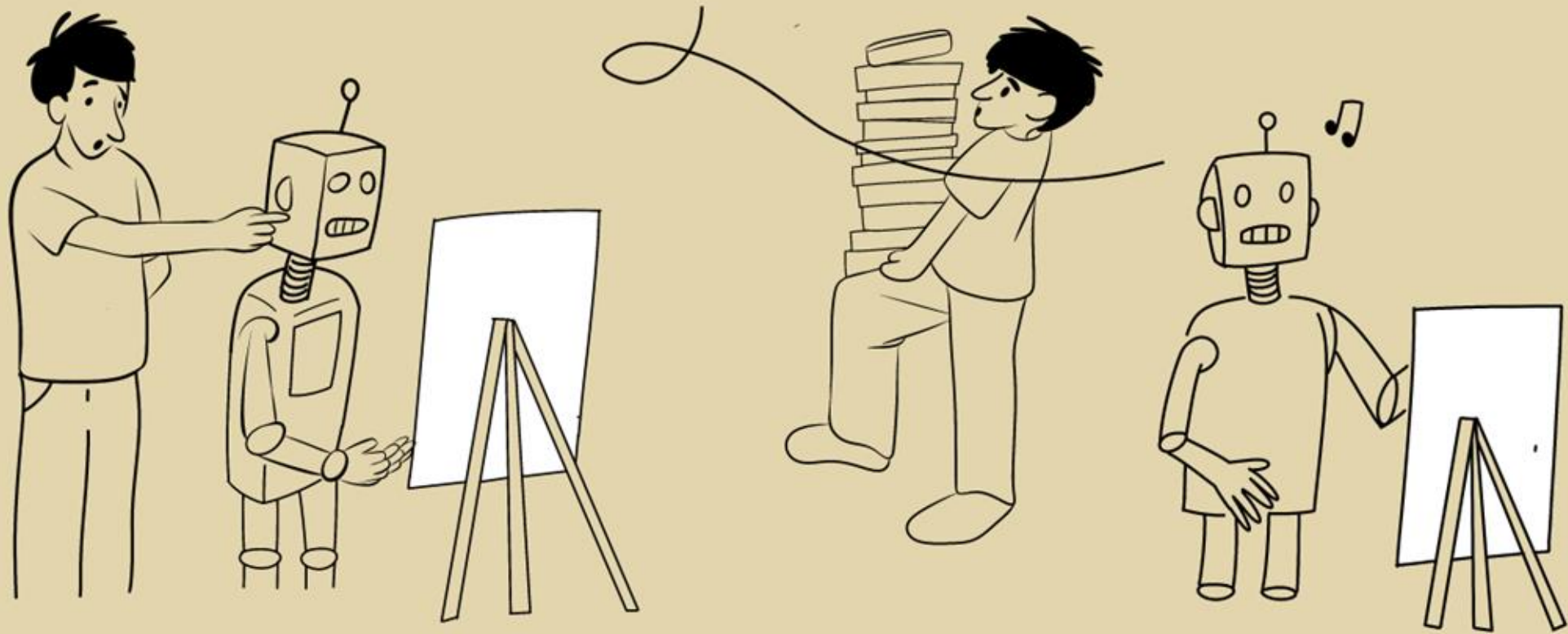
How Should We Build Better Human-AI Interaction

Design Why should they interact? How do we make it happen?

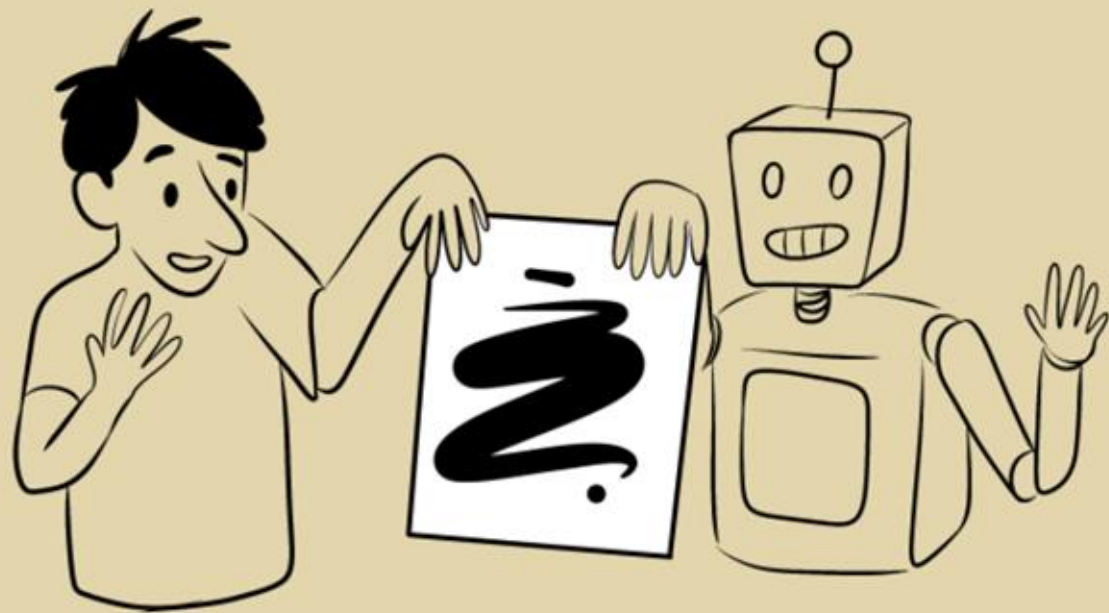
Enable How can we enable human-LLM interaction?

Evaluate Have we achieved what we want to achieve?





AI was supposed to take boring jobs and allow people to be creative, **but the opposite happened**



Outline

✓ **What Is Human-AI Interaction** (10 mins)

❑ **Enable Human-AI Interaction & Collaboration** (20 mins)

❑ Automation vs. Augmentation in “Human-AI Collaboration”

❑ Agency, RL and situational reasoning to improve collaboration

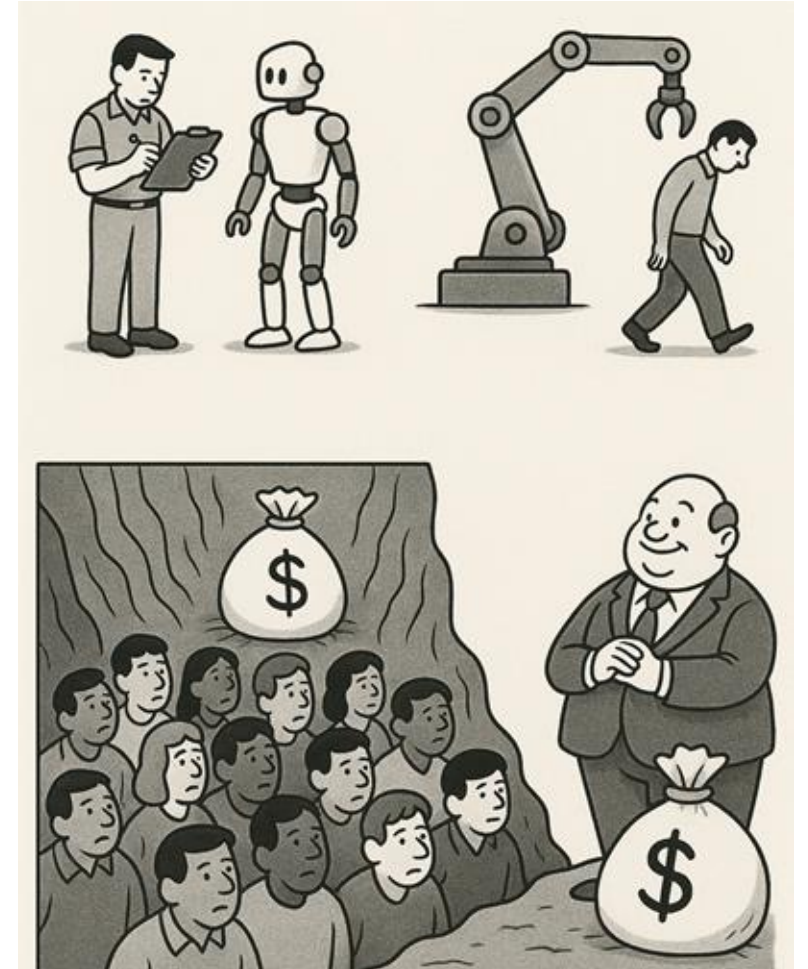
❑ Mixed examples of “does human-AI collaboration work”

Automation vs. Augmentation

Automation: AI replaces human capabilities

Augmentation: AI enhances human capabilities

- ❖ Automation isn't inherently bad
- ❖ Automation increases profits for a few
- ❖ Augmentation can lead to increased performance, creativity, new tasks, and emerging needs



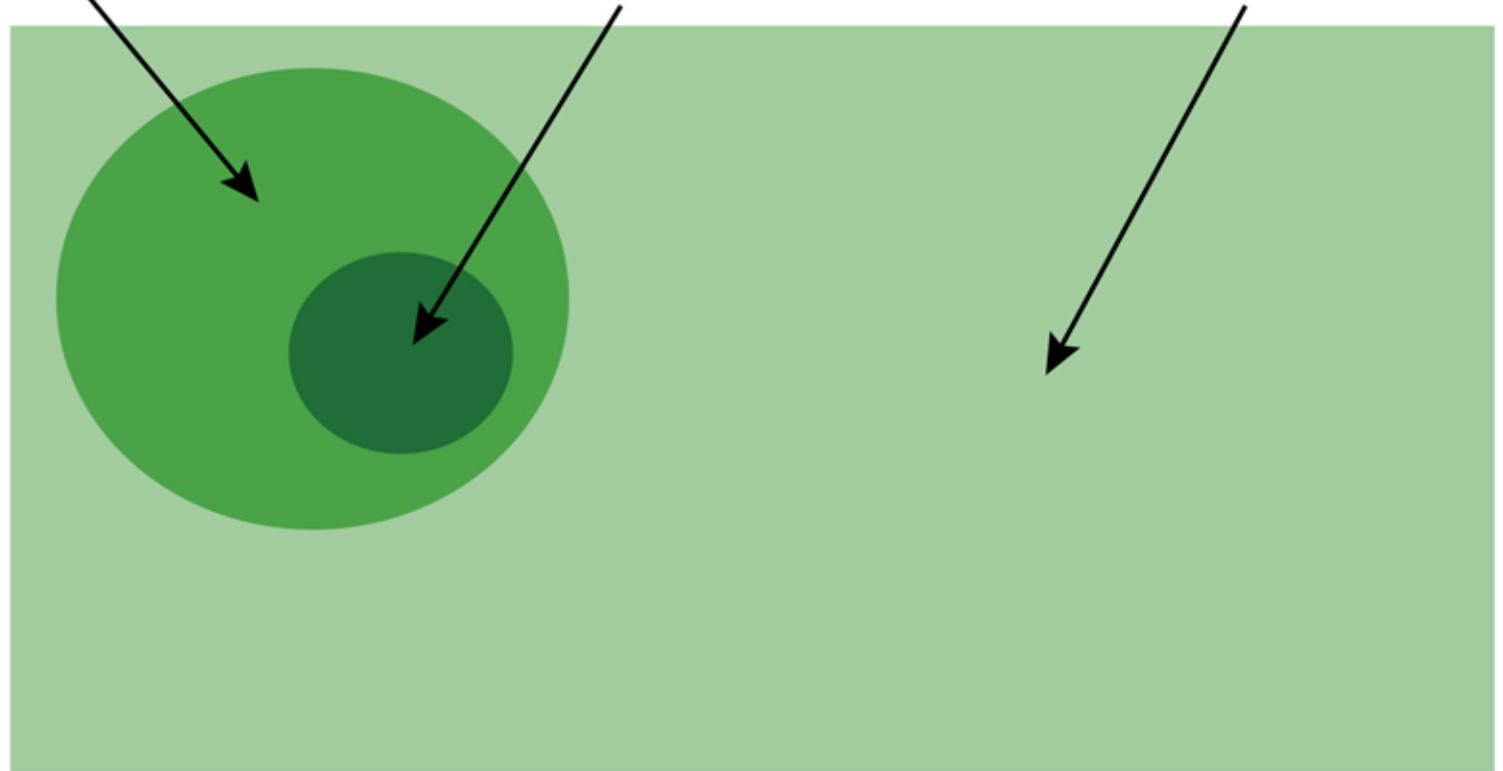
Turing Trap

Automating work instead of enhancing human labor leads to a concentration of wealth and power, trapping the rest in a system where they can't improve their lives

Tasks That Humans Can Do

Human Tasks That Machines Could Automate

New Tasks That Humans Can Do with the Help of Machines



Recent Examples on Automation vs. Augmentation

**Duolingo**
732,076 followers
1mo • Edited •

[+ Follow](#) [...](#)

👉 Below is an all-hands email from our CEO, [Luis von Ahn](#) – we are going to be AI-first.

Just like how betting on mobile in 2012 made all the difference, we're making a similar call now. This time the platform shift is AI.

What doesn't change: We will remain a company that cares deeply about its employees.

I've said this in Q&As and many meetings, but I want to make it official: **Duolingo is going to be AI-first.**

AI is already changing how work gets done. It's not a question of if or when. It's happening now. When there's a shift this big, the worst thing you can do is wait. In 2012, we bet on mobile. While others were focused on mobile companion apps for websites, we decided to build mobile-first because we saw it was the future. That decision helped us win the 2013 iPhone App of the Year and unlocked the organic word-of-mouth growth that followed.

Betting on mobile made all the difference. We're making a similar call now, and this time the platform shift is AI.

AI isn't just a productivity boost. It helps us get closer to our mission. To teach well, we need to create a massive amount of content, and doing that manually doesn't scale. One of the best decisions we made recently was replacing a slow, manual content creation process with one powered by AI. Without AI, it would take us decades to scale our content to more learners. We owe it to our learners to get them this content ASAP.

AI also helps us build features like Video Call that were impossible to build before. **For the first time ever, teaching as well as the best human tutors is within our reach.**

Being AI-first means we will need to rethink much of how we work. **Making minor tweaks to systems designed for humans won't get us there.** In many cases, we'll need to start from scratch. We're not going to rebuild everything overnight, and some things—like getting AI to understand our codebase—will take time. However, we can't wait until the technology is 100% perfect. We'd rather move with urgency and take occasional small hits on quality than move slowly and miss the moment.



Mama may I have cookie?

no streak go sweet treat #duolingo

Comments (5135)

 **mama may** I have real people running the company ❤️
4-30 76K Reply
[View 77 replies](#)

 **sarah** 🙅 How about NO ai, keep your employees
4-30 30.8K Reply
[View 82 replies](#)

 **Jenna**
Had to make this one using dolls bc there no more employees to wear the costumes
5-3 329 Reply
[View 2 replies](#)

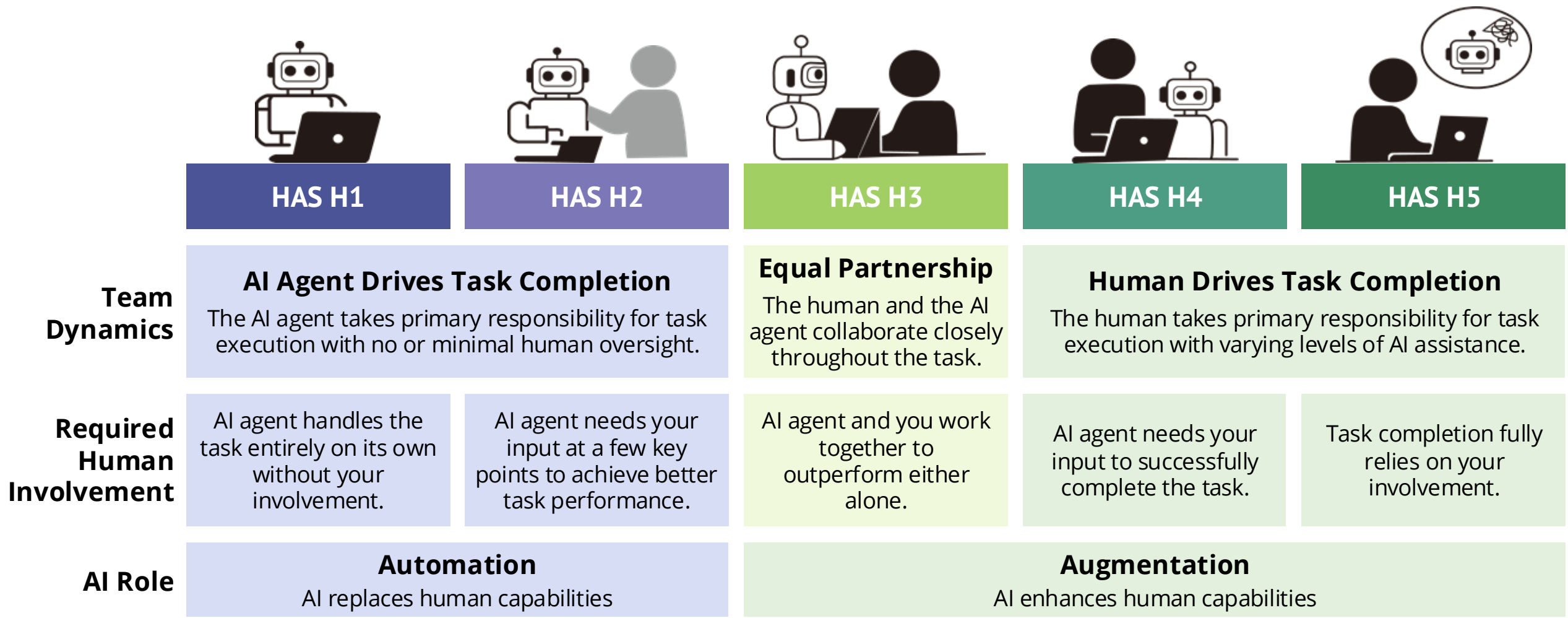
 **Vicky (rhymes with Grug)**
Deleted app because I don't trust AI to teach me language
5-2 101 Reply
[View 3 replies](#)

 **Eleanor**
mama may i have humans running the app? deleted because of AI use 🙄
[Log in to comment](#)


405.3K
5135
38.6K
72.9K

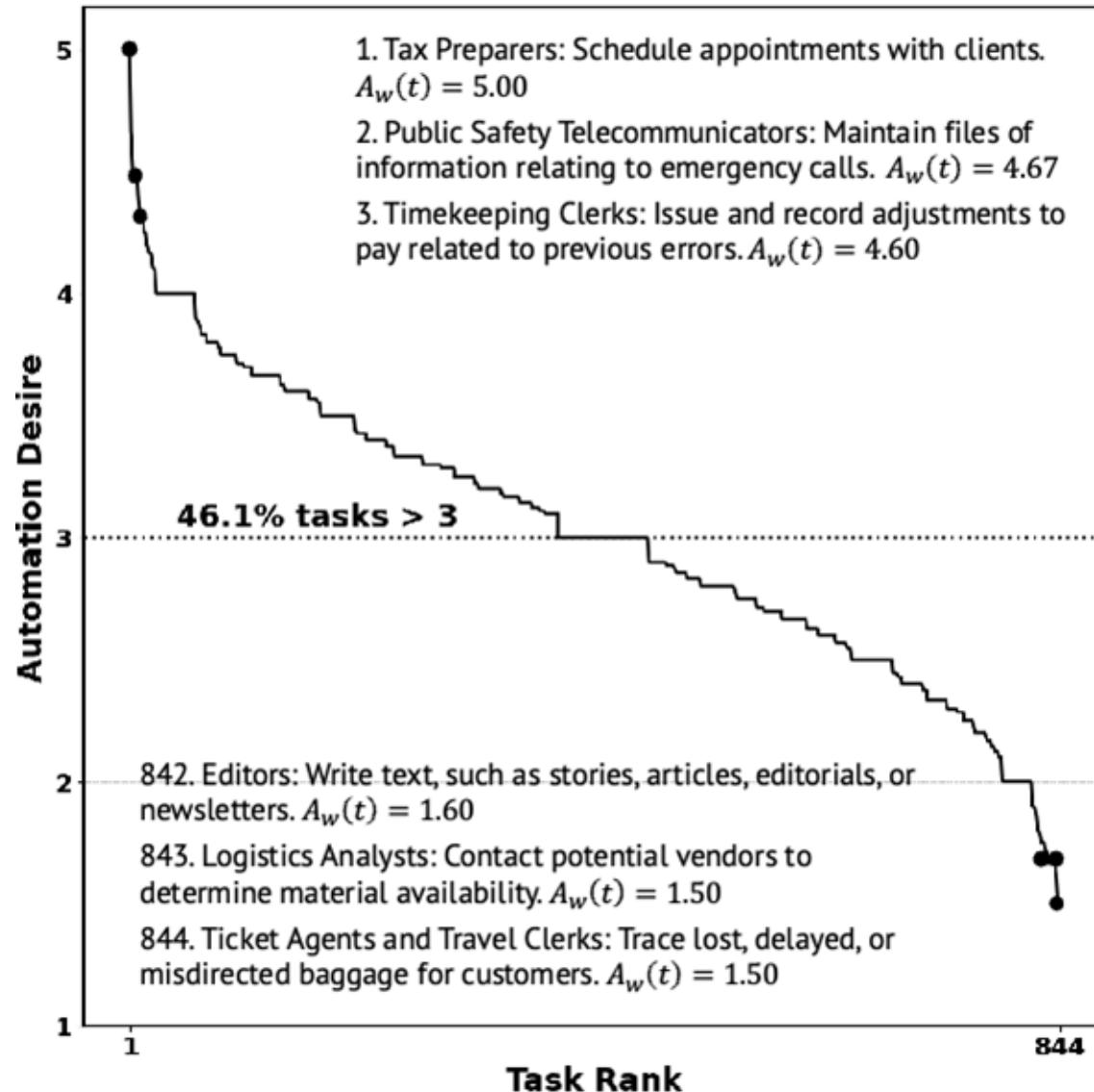
Framework for Automation vs. Augmentation

Human Agency Scale



Shao, Yijia, Humishka Zope, Yucheng Jiang, Jiabin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang.
"Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the US Workforce." arXiv:2506.

Worker Desire for Automation



Top reasons of why workers want automation for these tasks:

1. Automating the task would free up my time for high-value work
2. Automating this task would improve the quality of work
3. The task is mentally draining
4. The task is complicated or difficult

Outline

✓ **What Is Human-AI Interaction** (10 mins)

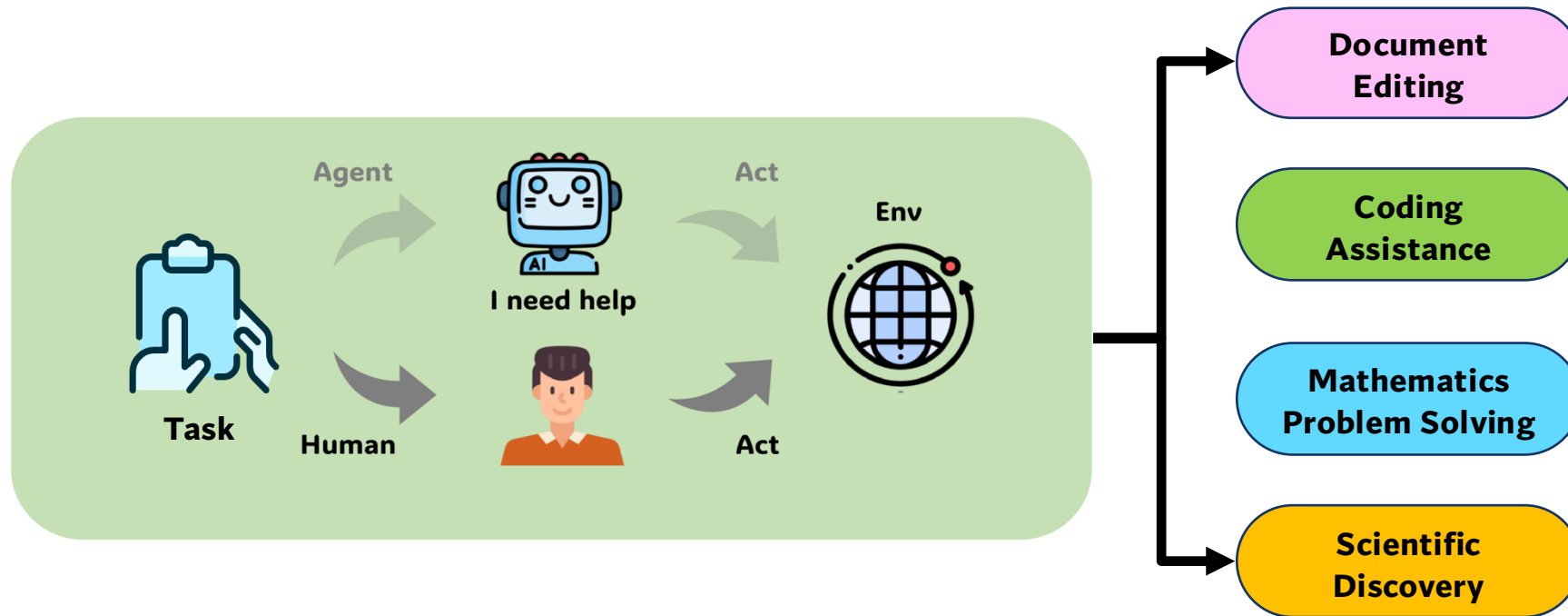
❑ **Enable Human-AI Interaction & Collaboration** (20 mins)

✓ Automation vs. Augmentation in “Human-AI Collaboration”

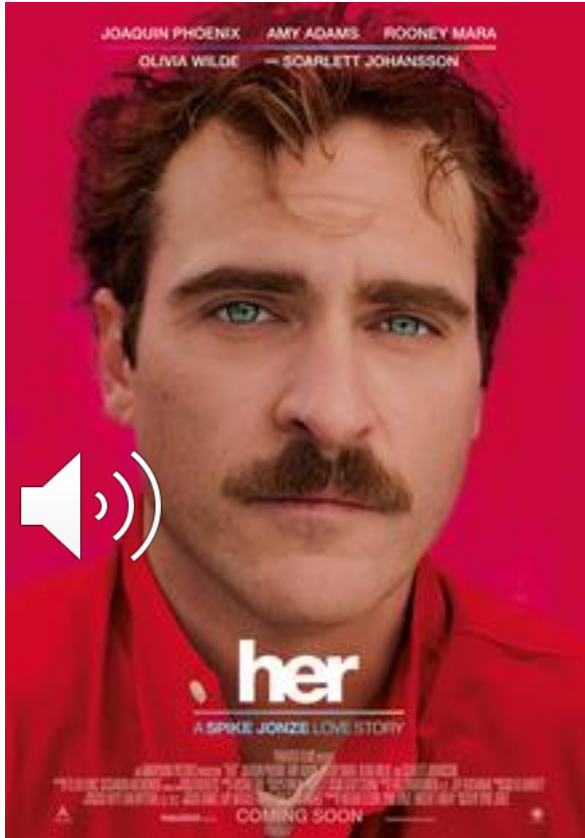
❑ Agency, RL and situational reasoning to improve collaboration

❑ Mixed examples of “does human-AI collaboration work”

Lots of Applications of Human-AI Interaction/ Collaboration



Agency and Friction during Human-AI Interaction



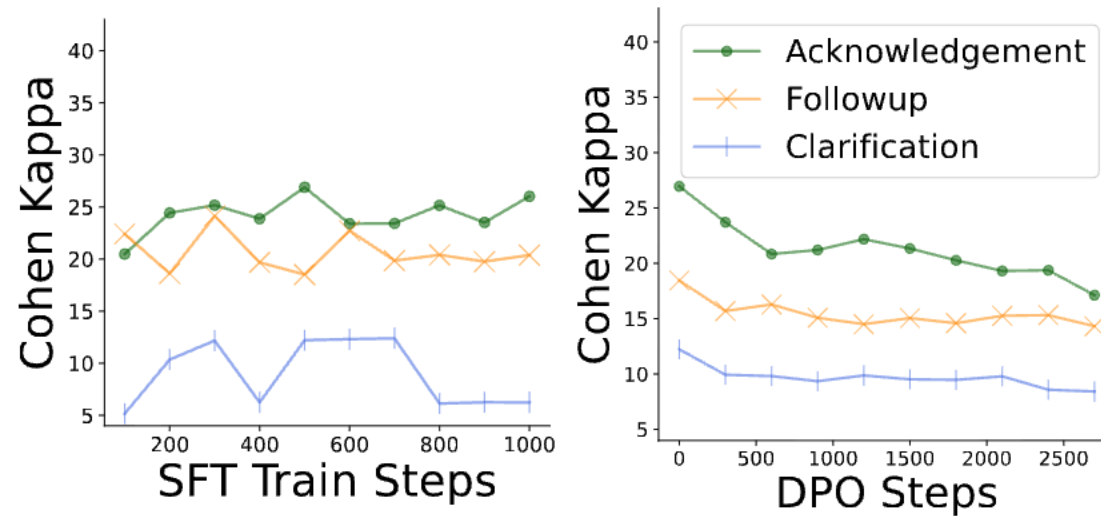
My most frustrating experience with Operator was my first one: trying to order groceries. “Help me buy groceries on Instacart,” I said, expecting it to ask me some basic questions. Where do I live? What store do I usually buy groceries from? What kinds of groceries do I want?

It didn’t ask me any of that. Instead, Operator opened Instacart in the browser tab and begin searching for milk in grocery stores located in Des Moines, Iowa.

<https://www.platformer.news/openai-operator-ai-agent-hands-on/>

Grounding Gaps in Human-AI Interaction

Act	ChatGPT 3.5	Human	Cohen κ
Emotional Support Conv			
Follow	10.78 ± 2.1	27.87 ± 4.4	12.47 ± 6.4
Ack.	1.05 ± 0.8	12.9 ± 3.7	3.14 ± 4.9
Clar.	0.0 ± 0.0	3.05 ± 1.2	0.0 ± 0.0
Teacher Student Chatroom			
Follow	11.56 ± 1.9	12.04 ± 2.1	16.75 ± 4.6
Ack.	5.68 ± 1.4	16.59 ± 2.4	18.25 ± 5.4
Clar.	0.57 ± 0.3	3.77 ± 0.9	0.36 ± 2.5
Persuasion for Good			
Follow	1.66 ± 0.9	8.18 ± 2.4	2.94 ± 7.6
Ack.	1.8 ± 1.0	6.11 ± 1.9	25.73 ± 16.7
Clar.	0.0 ± 0.0	0.28 ± 0.4	0.0 ± 0.0

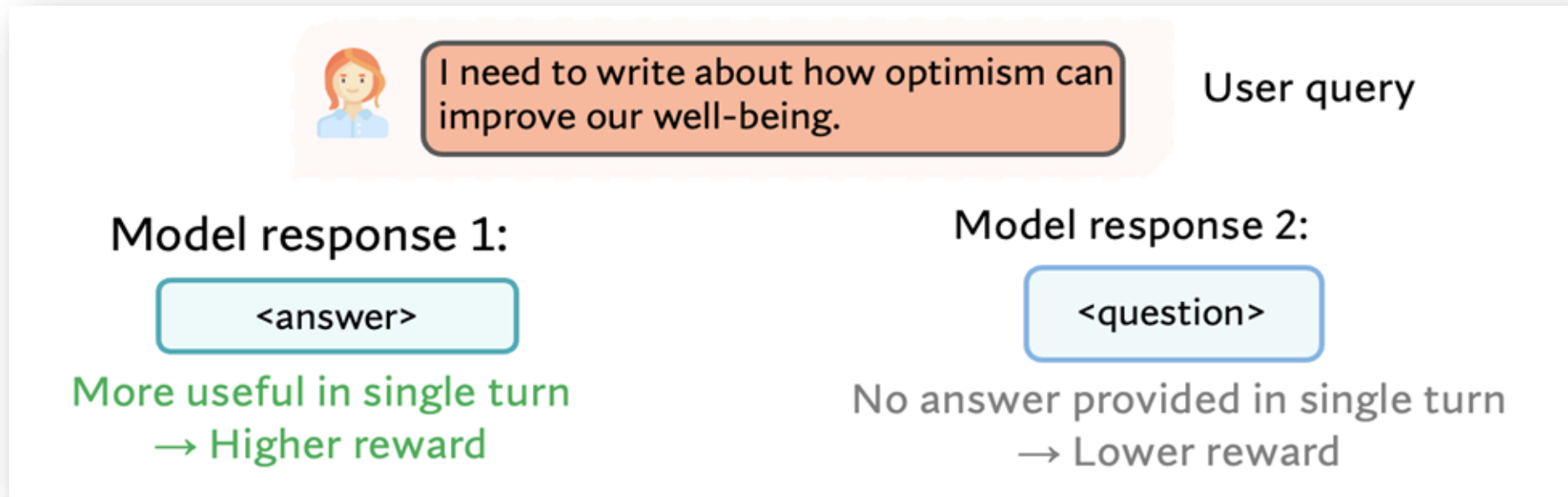


We observe no correlation between SFT training steps and grounding acts, but negative correlation between DPO train steps and grounding acts

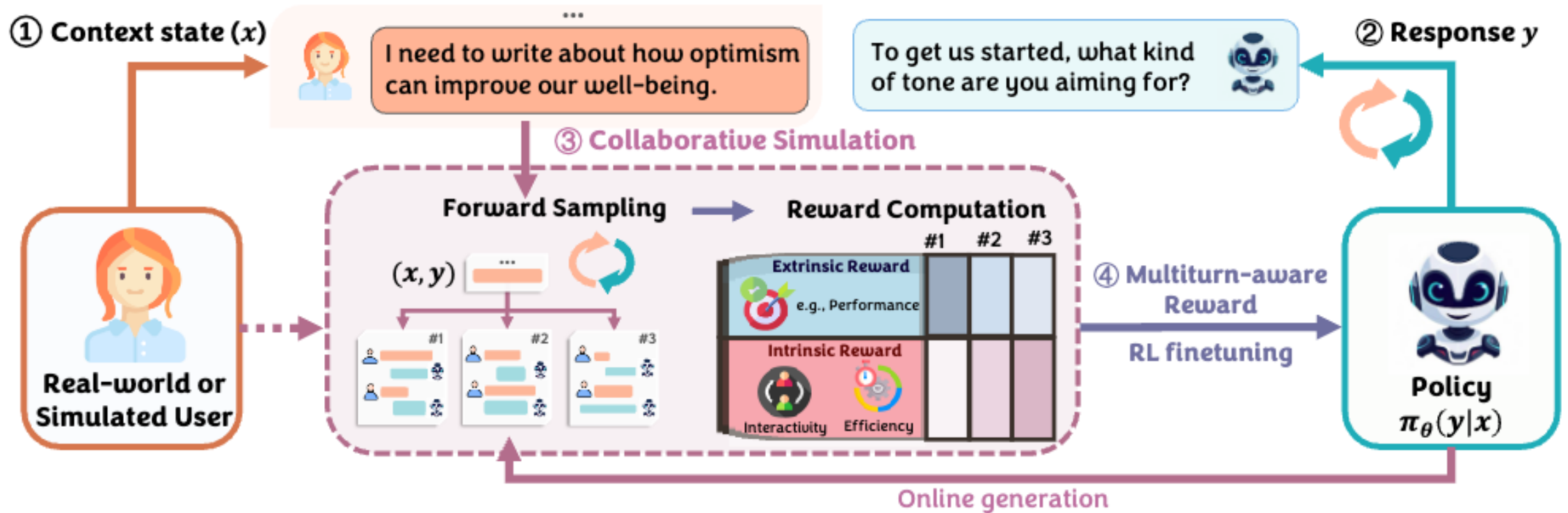
Using RL to Improve Human-AI Collaboration

LLMs are usually tuned based on **single-turn human preferences**

But these single-turn rewards encourage models to generate responses that may NOT be useful in the long term.



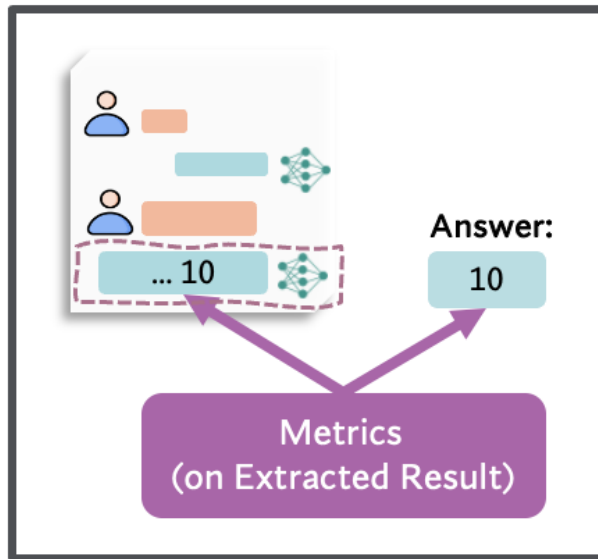
From Passive Responders to Active Collaborators



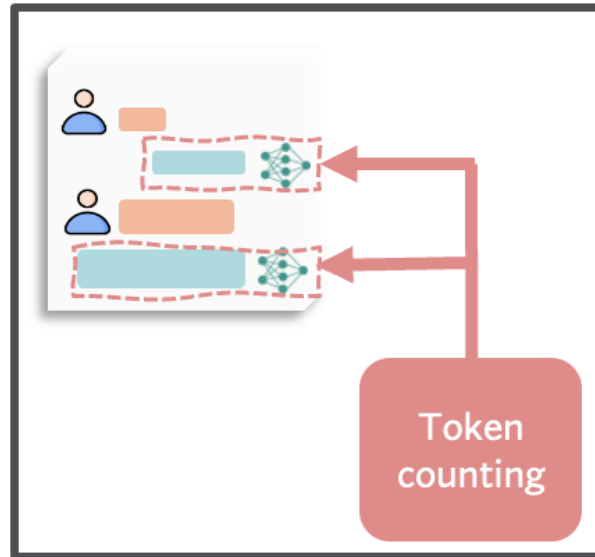
Wu, Shirley, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. "CollabLLM: From passive responders to active collaborators." arXiv preprint arXiv:2502.00640 (2025).

Collaborative LLM Reward

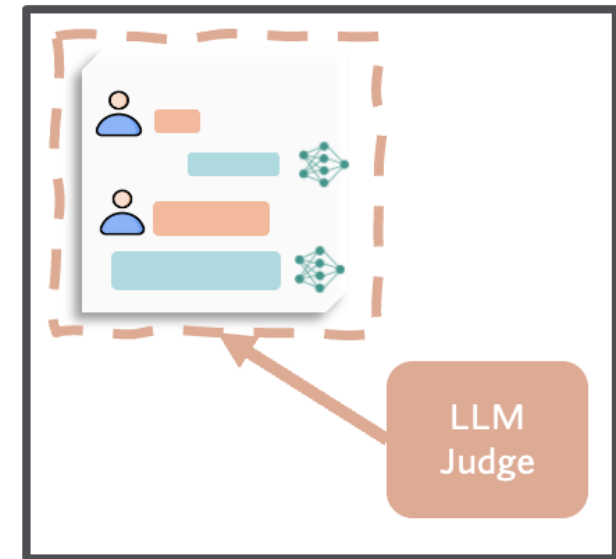
- **Extrinsic Reward:** how well the conversation achieves user's goal
- **Intrinsic Reward:** rewards user experience: token count, LLM judge
- $R(t_{1:K})$: Extrinsic/Intrinsic reward or a combination of them



Extrinsic: Task performance
(e.g., accuracy for QA)



Intrinsic:
Efficiency



Intrinsic:
Interactivity

Collaborative LLM Reward

Multiturn-aware reward (MR): causal effect estimation of how model's response influences the future trajectory of a conversation

Quantify the effect of intervention (model response m_j):

$$MR(m_j) = \mathbb{E}_{t_{1:K} \sim P(t_{1:K} | t_{1:j-1} \cup m_j)} R(t_{1:K})$$

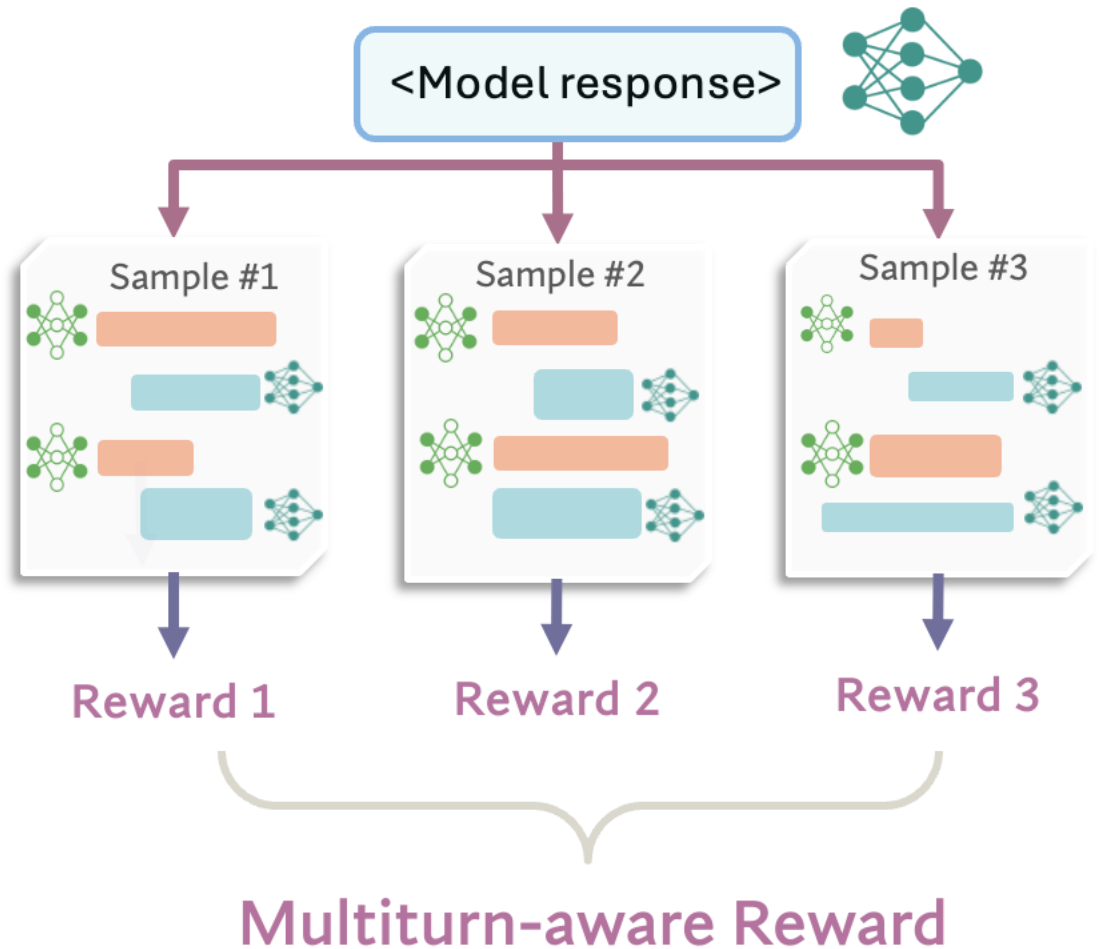
$t_{i:j}$ denotes the conversation from i -th to j -th turn

Estimate long-term impact with synthetic conversations

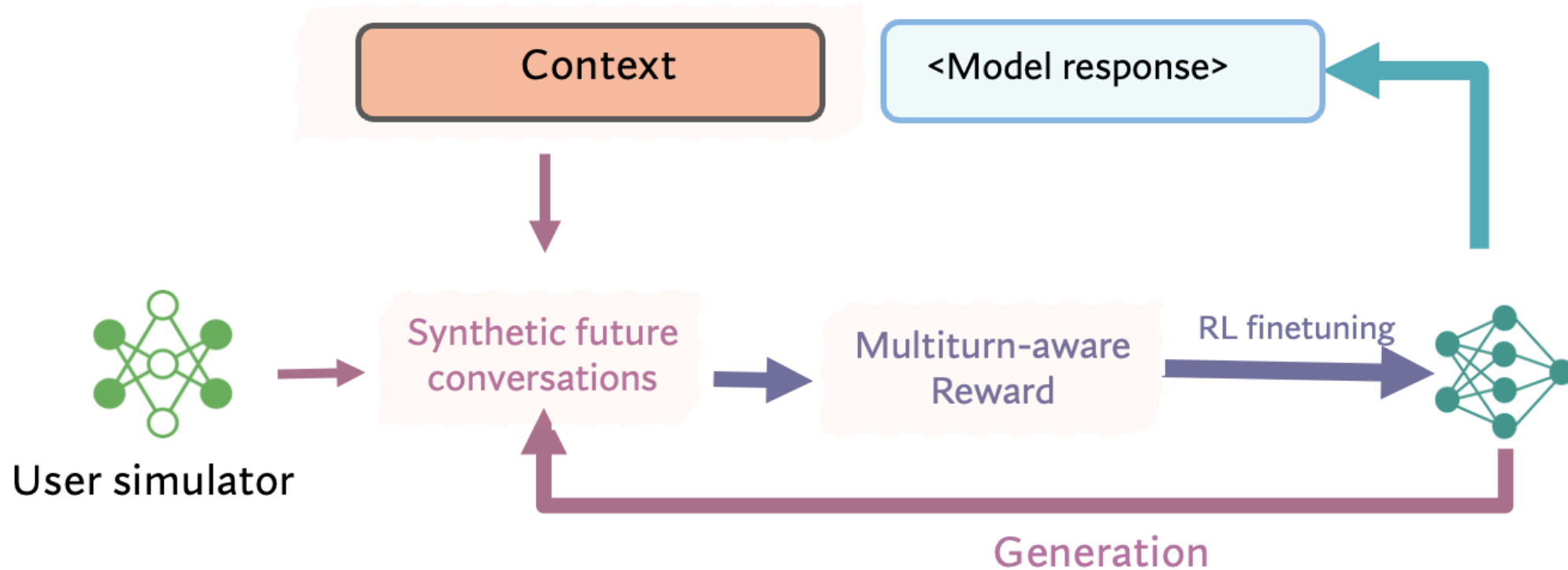
Goal: Evaluate the long-term impact of this model response

Approach:

- ① Sample synthetic future conversations w/ user simulators
- ② Apply conversation-level reward for each conversation
- ③ Average the rewards



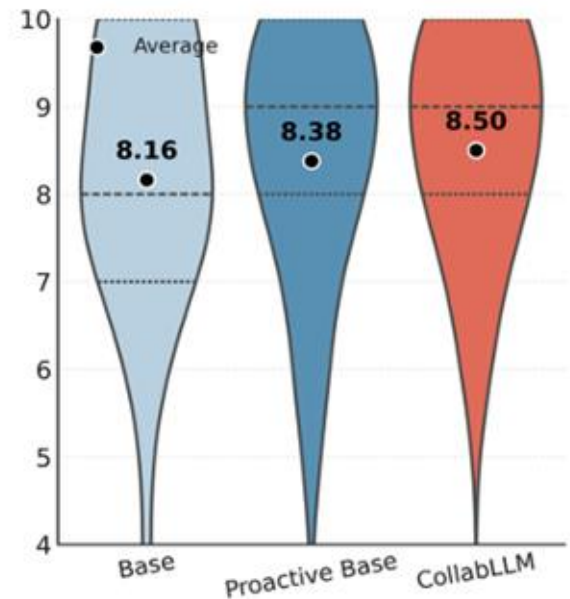
Train LLMs to generate responses that maximize multiturn reward



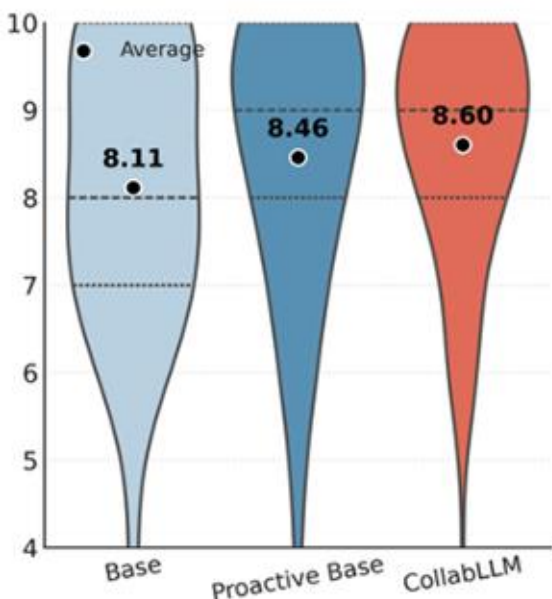
CollabLLM can be integrated with PPO (Schulman et al. 2017) and DPO (Rafailov et al. 2023) to conduct RL finetuning

Improvement of CollabLLM

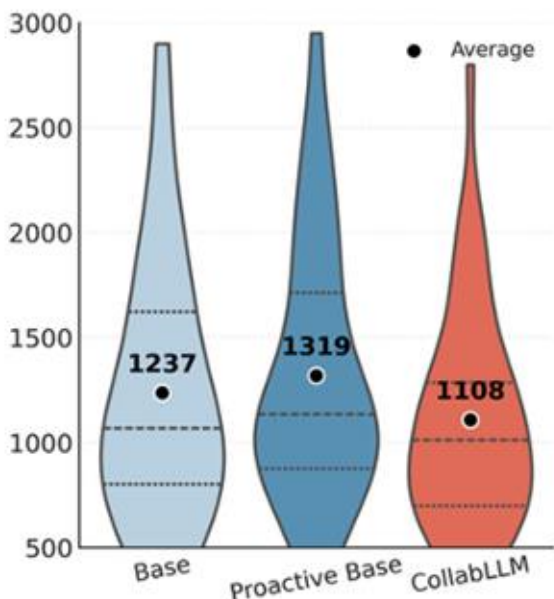
	MediumDocEdit-Chat			BigCodeBench-Chat			MATH-Chat		
	BLEU $\uparrow$	#Tokens(k) $\downarrow$	ITR $\uparrow$	PR $\uparrow$	#Tokens(k) $\downarrow$	ITR $\uparrow$	ACC $\uparrow$	#Tokens(k) $\downarrow$	ITR $\uparrow$
Base	32.2	2.49	46.0	9.3	1.59	22.0	11.0	3.40	44.0
Proactive Base	35.0	2.18	62.0	11.0	1.51	33.7	12.5	2.90	46.0
CollabLLM	36.8	2.00	92.0	13.0	1.31	52.0	16.5	2.37	60.0
Rel. Improv.	5.14%	8.25%	48.3%	18.2%	13.2%	54.3%	32.0%	18.3%	36.4%



(a) Document Quality Rating $\uparrow$



(b) Interaction Rating $\uparrow$



(c) Time Spent (s) $\downarrow$

CollabLLM yields high-quality documents, better user experience, and saves time by >10%!

Qualitative Insights related to CollaborativeLLMs

About Base (llama-3-1-8b):

“the AI just agreed with me on pretty much everything. There was no debate or discussion.”

About Proactive Base:

”The AI seemed to be very redundant and asked me the same questions over and over”

About CollabLLM:

“It helped really well to navigate what to say and what information is needed”

“The AI really helped focusing on one part of the story at a time.”

“Asking questions and making you think of things you never thought of”

Enable human-agent collaboration

