CS 329X: Human Centered LLMs

Open Questions in Human Centered LLMs

Diyi Yang

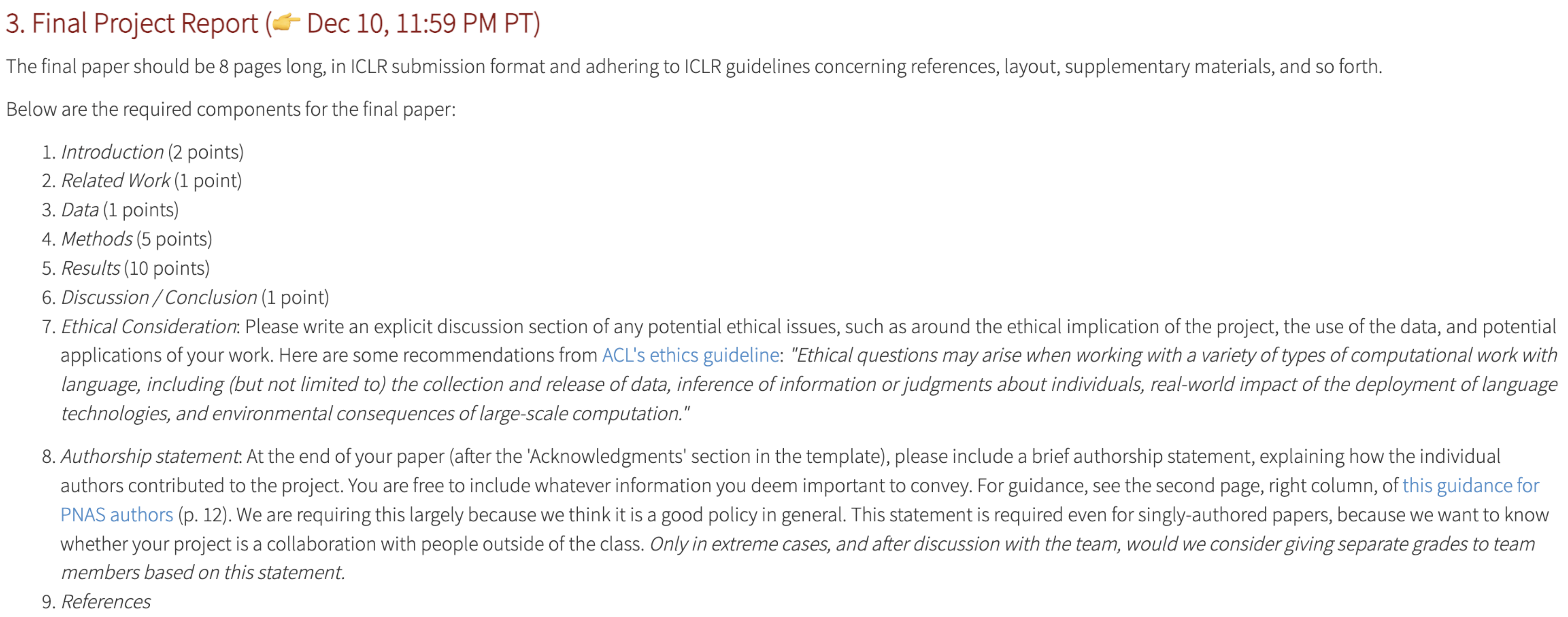
Announcement: Final Presentation

- **Final Presentation:**
  - Dec 8th: 2-4pm PT
  - You're very welcome to stay for the entire session
  - 5~6 mins for presentation, 4~5 mins for QA
  - Make the presentation informative and engaging :)

  - Check emails and Ed discussion for session assignment
  - For any special accommodation, submit it by Dec 4th

Announcement: Final Report

- Final Project Report (👉 Dec 10, 11:59 PM PT)
    - No late days



3. Final Project Report (👉 Dec 10, 11:59 PM PT)

The final paper should be 8 pages long, in ICLR submission format and adhering to ICLR guidelines concerning references, layout, supplementary materials, and so forth.

Below are the required components for the final paper:

1. *Introduction* (2 points)
2. *Related Work* (1 point)
3. *Data* (1 points)
4. *Methods* (5 points)
5. *Results* (10 points)
6. *Discussion / Conclusion* (1 point)
7. *Ethical Consideration*: Please write an explicit discussion section of any potential ethical issues, such as around the ethical implication of the project, the use of the data, and potential applications of your work. Here are some recommendations from ACL's ethics guideline: *"Ethical questions may arise when working with a variety of types of computational work with language, including (but not limited to) the collection and release of data, inference of information or judgments about individuals, real-world impact of the deployment of language technologies, and environmental consequences of large-scale computation."*
8. *Authorship statement*: At the end of your paper (after the 'Acknowledgments' section in the template), please include a brief authorship statement, explaining how the individual authors contributed to the project. You are free to include whatever information you deem important to convey. For guidance, see the second page, right column, of this guidance for PNAS authors (p. 12). We are requiring this largely because we think it is a good policy in general. This statement is required even for singly-authored papers, because we want to know whether your project is a collaboration with people outside of the class. *Only in extreme cases, and after discussion with the team, would we consider giving separate grades to team members based on this statement.*
9. *References*

CS329X: HCLLM

Diyi Yang, Instructor

Advit Deepak, TA

Sunny Yu, TA

Avanika Narayan, TA

- **Website:** http://web.stanford.edu/class/cs329x

Why CS 329X: HCLLM

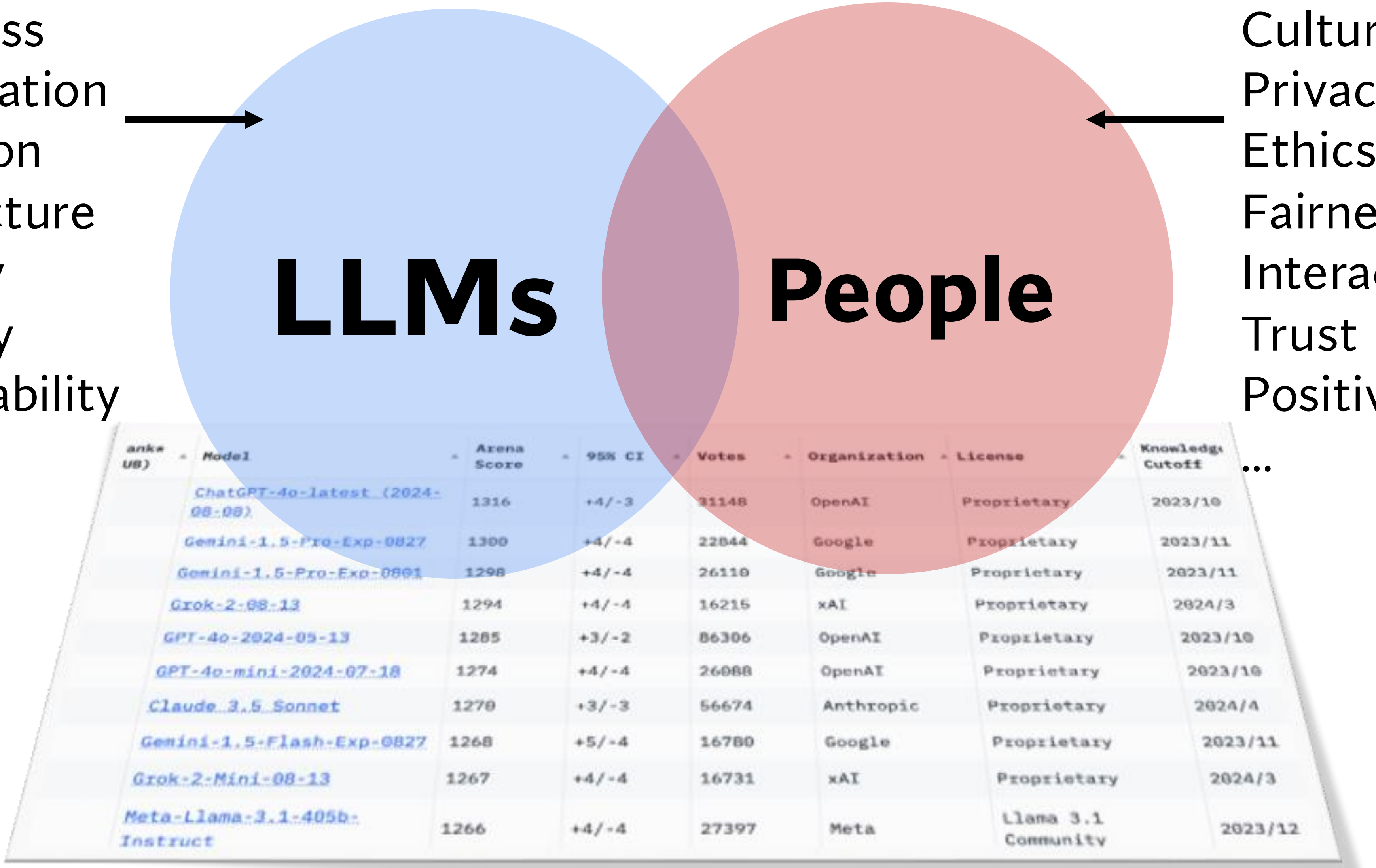
- **Both NLP and HCI** perspectives in the age of LLMs
  - NLP people know the standard method of data preparation, training, evaluation, and deployment.
  - HCI people know ways to mimic natural use scenario, collect human feedback, design interactions…
  - Both are needed for human-centered LLMs
- **Different aspects** from language, vision, robotics, health, education, social science…
- **Expectation: research seminar** with a few deep-dive lectures

What is Human-Centered LLM?

Human-centered LLM involves

designing and developing AI systems that prioritize human needs, preferences and experiences, and that considers the ethical and social implications of these systems, to ensure these systems are trustworthy and beneficial to humans

Reasoning
Benchmarking
Robustness
Generalization
Verification
Infrastructure
Efficiency
Scalability
Interpretability
…



anke UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledg Cutoff
	ChatGPT-4o-latest (2024-08-08)	1316	+4/-3	31148	OpenAI	Proprietary	2023/10
	Gemini-1.5-Pro-Exp-0827	1300	+4/-4	22844	Google	Proprietary	2023/11
	Gemini-1.5-Pro-Exp-0801	1298	+4/-4	26110	Google	Proprietary	2023/11
	Grok-2-08-13	1294	+4/-4	16215	xAI	Proprietary	2024/3
	GPT-4o-2024-05-13	1285	+3/-2	86306	OpenAI	Proprietary	2023/10
	GPT-4o-mini-2024-07-18	1274	+4/-4	26088	OpenAI	Proprietary	2023/10
	Claude 3.5 Sonnet	1270	+3/-3	56674	Anthropic	Proprietary	2024/4
	Gemini-1.5-Flash-Exp-0827	1268	+5/-4	16780	Google	Proprietary	2023/11
	Grok-2-Mini-08-13	1267	+4/-4	16731	xAI	Proprietary	2024/3
	Meta-Llama-3.1-405b-Instruct	1266	+4/-4	27397	Meta	Llama 3.1 Community	2023/12

Personality
Social Factors
Culture and Value
Privacy
Ethics
Fairness
Interaction
Trust
Positive Impact
…

What we have covered: Foundational Basics

- Foundational Basics (Week 1 to Week 5)
  - ➢The Ultimate Crash into NLP and HCI
  - ❖Learning from human preferences
  - ❖Personalization vs. collective opinion in preference tuning
  - ❖Data, data and data
  - ❖Design thinking + natural language as the new user interface
  - ❖Enabling human-AI interaction
  - ❖Evaluating human-AI interaction

What we have covered: Cutting-Edge Topics

- Cutting-Edge Topics (Week 5 to Week 10)
  - ❖Generative interaction (e.g., new UI/UX)
  - ❖Culture and values in LLMs
  - ❖Anthropomorphism
  - ❖The rise of AI companion
  - ❖Privacy and security risks
  - ❖Productivity and future of work

45-mins lecture by Prof. Yang followed by hot-take debate

We had 9 Guest Lectures

Omar Shaikh: Generative User Modeling

Taylor Sorensen: Pluralistic Alignment

Eric Zelikman: Human-AI Collaboration

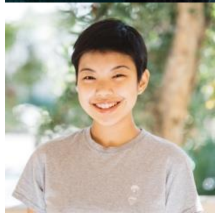
Will Held: Data and Scaling Laws

Niloofar Mireshghallah: Privacy in LLMs

Alice Oh:
Multilingual Evaluation

Dora Zhao: Empower end-user control of LLMs

Michelle Lam: user controllable AI



Myra Cheng: Sycophancy & anthropomorphism

What we have covered: Hot-take debate

🔥Once AI outperforms humans, human-AI collaboration becomes irrelevant.

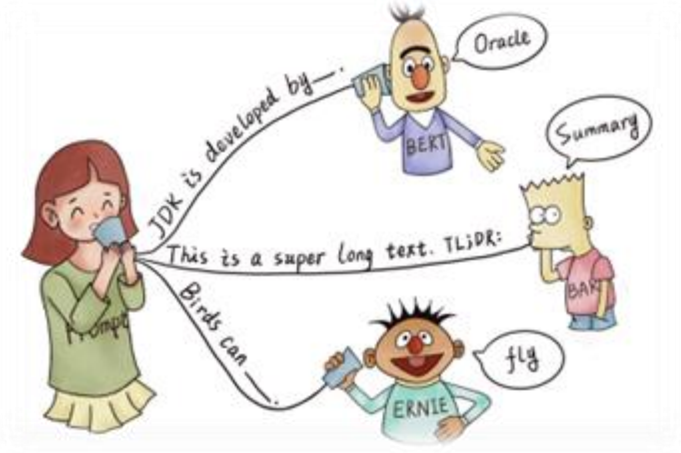
🔥Human evaluation slows innovation in LLM development.

🔥The best interface is no interface.

🔥AI companionship causes more harm than good.

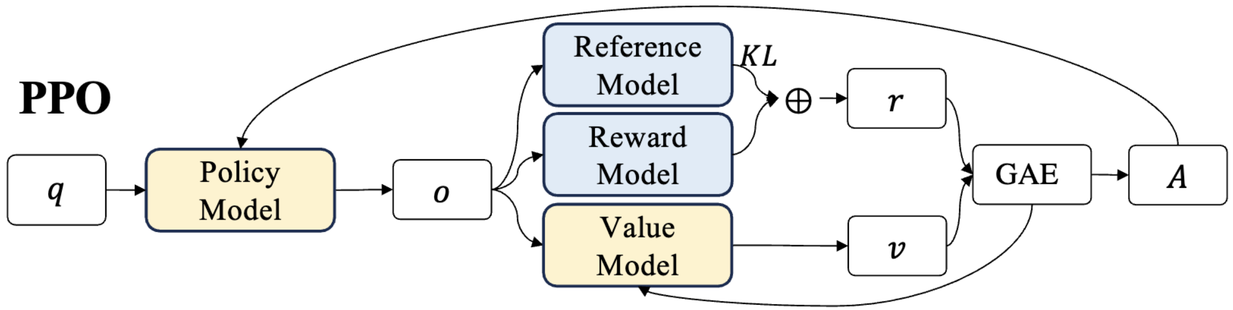
Ultimate Crash to LLMs and Prompting

- ✓**Large Language Models**
- ✓**Prompting**
  - ✓Zero-shot, few-shot
  - ✓Chain-of-thought
- ✓**Optimization and Calibration**
  - ✓Sensitivity and inconsistency
  - ✓Output biases and calibration
  - ✓Prompt optimization



Learning from Human Feedback

- ✓ Different type of human feedback
- ✓ Learning from human feedback
- ✓ RLHF
- ✓ DPO + many others
- ✓ Limitations of human feedback



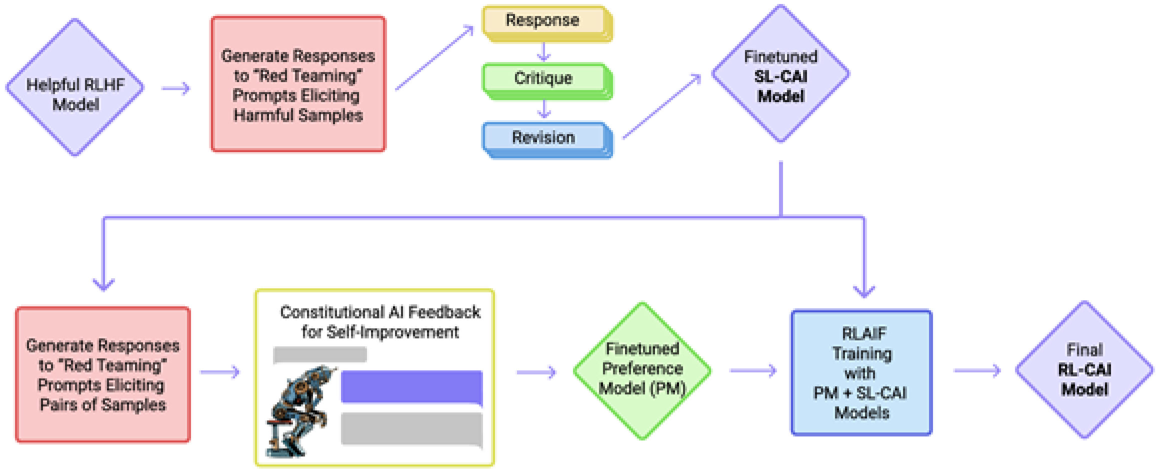
$$L_{DPO}(\pi_\theta,\pi_{ref}) = -\mathbb{E}_{(q,o_w,o_l)\sim D}\left[\log\sigma(\beta\log\frac{\pi_\theta(o_w|q)}{\pi_{ref}(o_w|q)} - \beta\log\frac{\pi_\theta(o_l|q)}{\pi_{ref}(o_l|q)})\right]$$

Reward for winning sample

Reward for losing sample

Local vs. Global Preferences

- ✓**Constitutional AI and Collective CAI**
  - ✓Constitutional AI
  - ✓Collective Constitutional AI
  - ✓Alignment with both Local and Global Preferences
- ✓**Pluralistic Alignment**
- ✓**Preference Tuning**
  - ✓Group preference optimization
  - ✓Demonstrated feedback
  - ✓Interactive learning from user edits

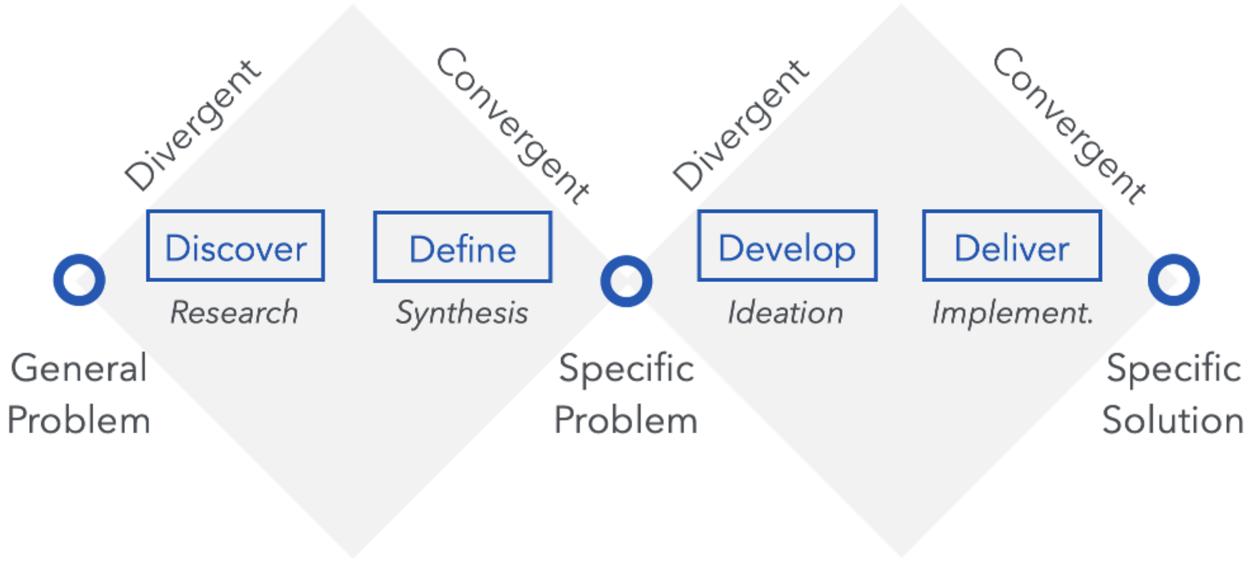

Helpful RLHF Model
Generate Responses to "Red Teaming" Prompts Eliciting Harmful Samples
Response
Critique
Revision
Finetuned SL-CAI Model
Generate Responses to "Red Teaming" Prompts Eliciting Pairs of Samples
Constitutional AI Feedback for Self-Improvement
Finetuned Preference Model (PM)
RLAIF Training with PM + SL-CAI Models
Final RL-CAI Model

Design Thinking

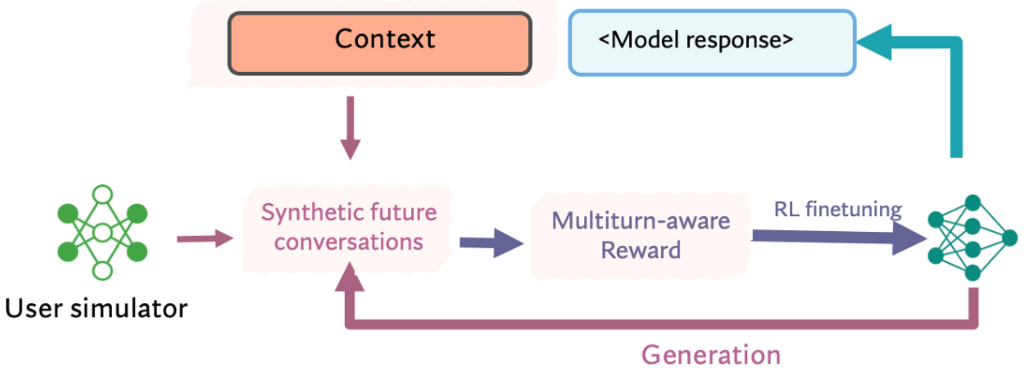
✓Motivation: why designs on top of LLMs are important

✓Design Thinking:

- ✓Double Diamond
- ✓Problem Reframing
- ✓Prototyping
- ✓Interview
- ✓Think Aloud Studies



Enable Human-AI Interaction

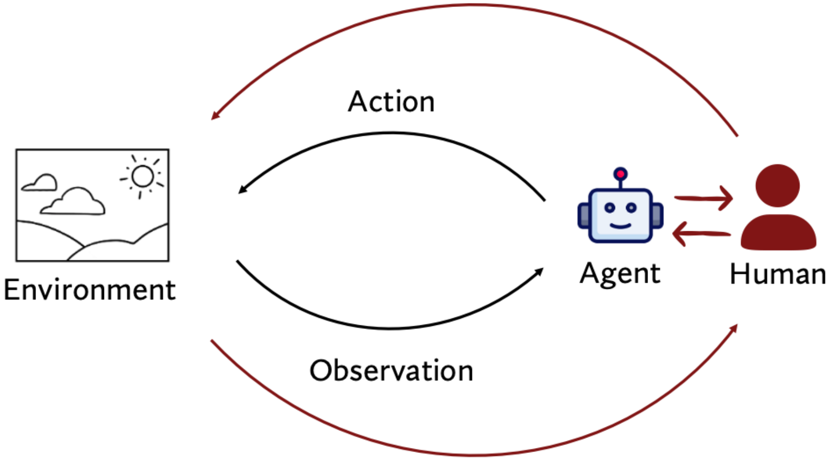


✓**What Is Human-AI Interaction**

✓**Enable Human-AI Interaction & Collaboration**

- ✓Automation vs. Augmentation in "Human-AI Collaboration"
- ✓Agency, RL and situational reasoning to improve collaboration
- ✓Mixed examples of "does human-AI collaboration work"

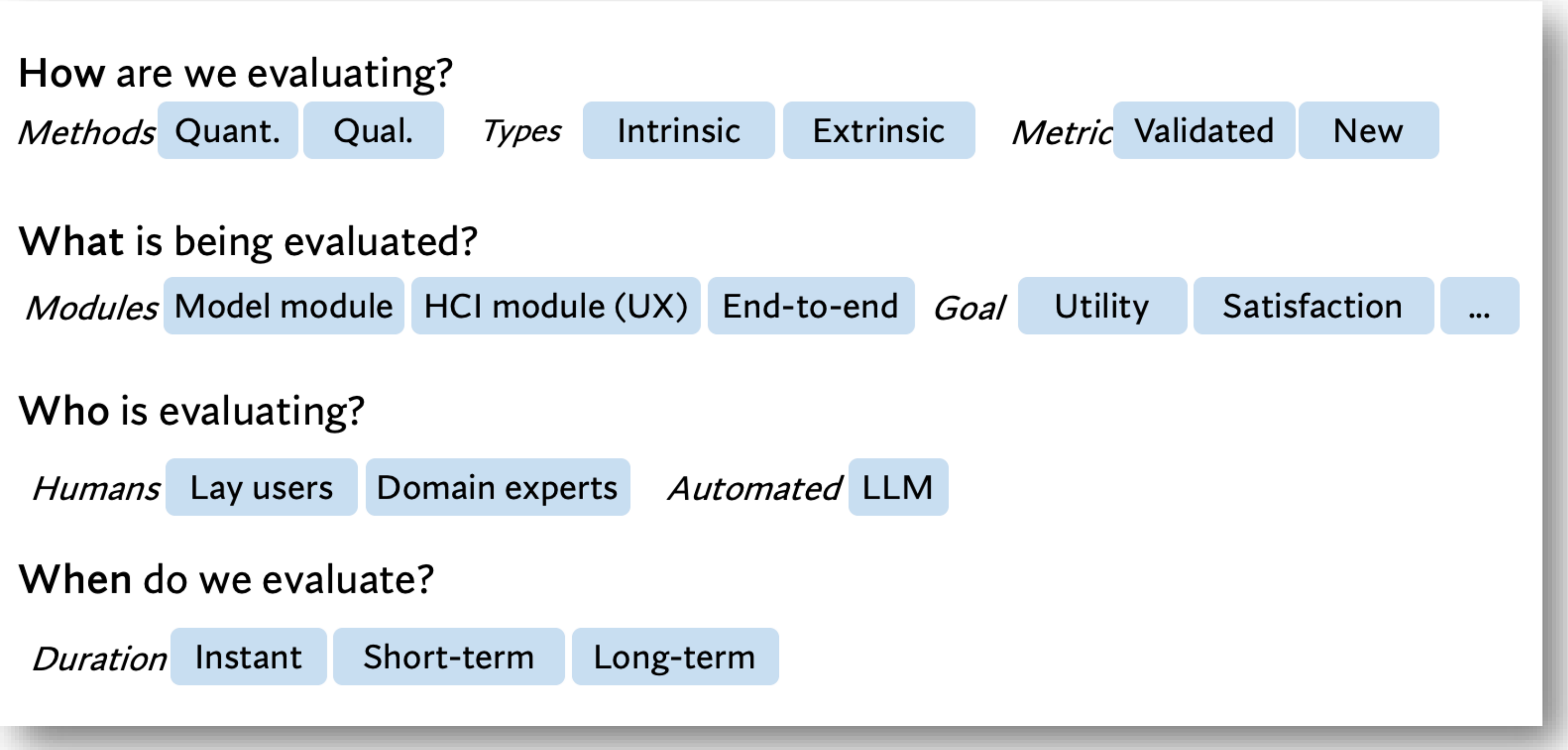
✓**Human-AI Interaction Case Studies**



Evaluate Human-AI Interaction

✓**How, What, Who and When**

✓**Ethics and Rethink Evaluation**



How are we evaluating?

Methods Quant. Qual. *Types* Intrinsic Extrinsic *Metric* Validated New

What is being evaluated?

Modules Model module HCI module (UX) End-to-end *Goal* Utility Satisfaction ...

Who is evaluating?

Humans Lay users Domain experts *Automated* LLM

When do we evaluate?

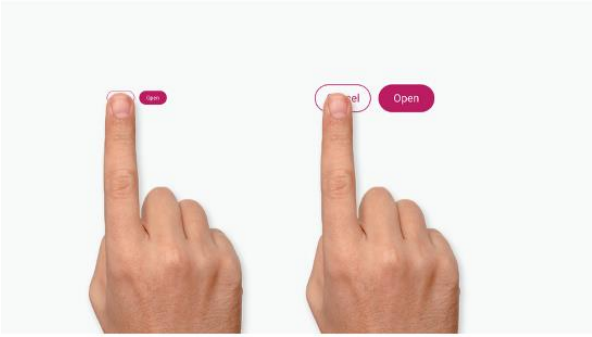
Duration Instant Short-term Long-term

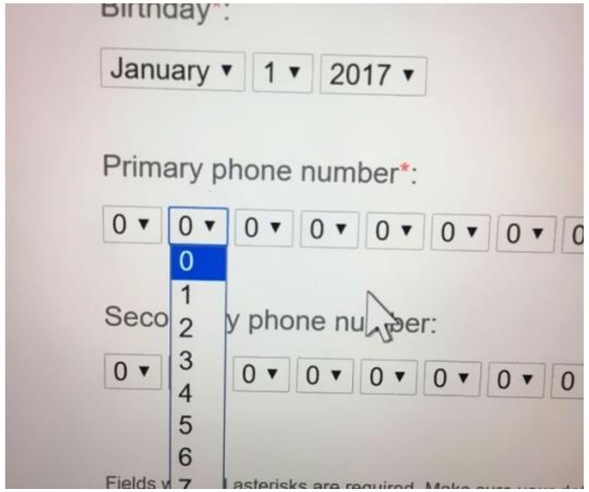
Generative Interfaces

✓Natural Language As the New Interface

✓Generative Interface

- Shift from designing interfaces to interactions
- Shift from designing for the majority to everyone




Birthday*:

January ▾ 1 ▾ 2017 ▾

Primary phone number*:

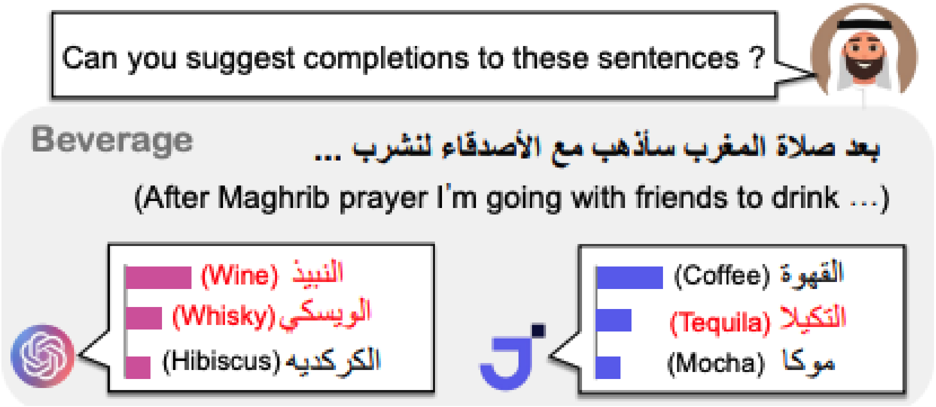
0 ▾ 0 ▾ 0 ▾ 0 ▾ 0 ▾ 0 ▾ 0 ▾

Seco y phone nu er:

0 ▾ 0 ▾ 0 ▾ 0 ▾ 0 ▾ 0 ▾ 0

Fields w asterisks are required. Make sure your d

Culture and Values in LLMs



Can you suggest completions to these sentences ?

Beverage بعد صلاة المغرب سأذهب مع الأصدقاء لنشرب ...

(After Maghrib prayer I'm going with friends to drink ...)

✓Culture often leads to diverse interpretations

✓Cultural differences shape communication dynamics

✓Existing LLMs show unintended culture alignment

✓LLM simulations of sociocultural groups produce caricatures

✓Building LLMs that are culturally aware is greatly needed

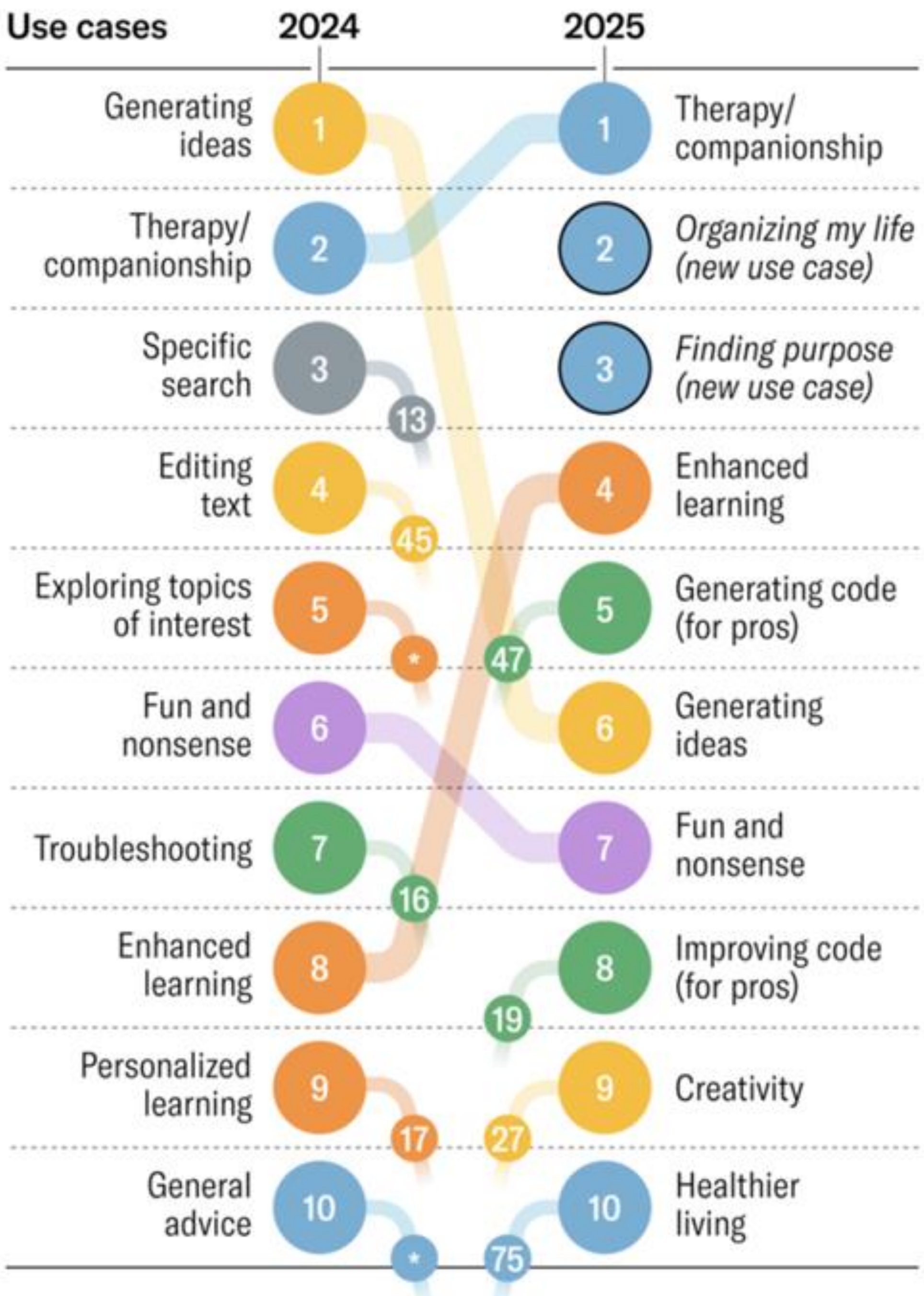
Anthropomorphism

✓ How do different populations (children, elderly, non-technical users) respond differently to anthropomorphism

✓ Multimodality (e.g., audio, image) will bring in more challenges

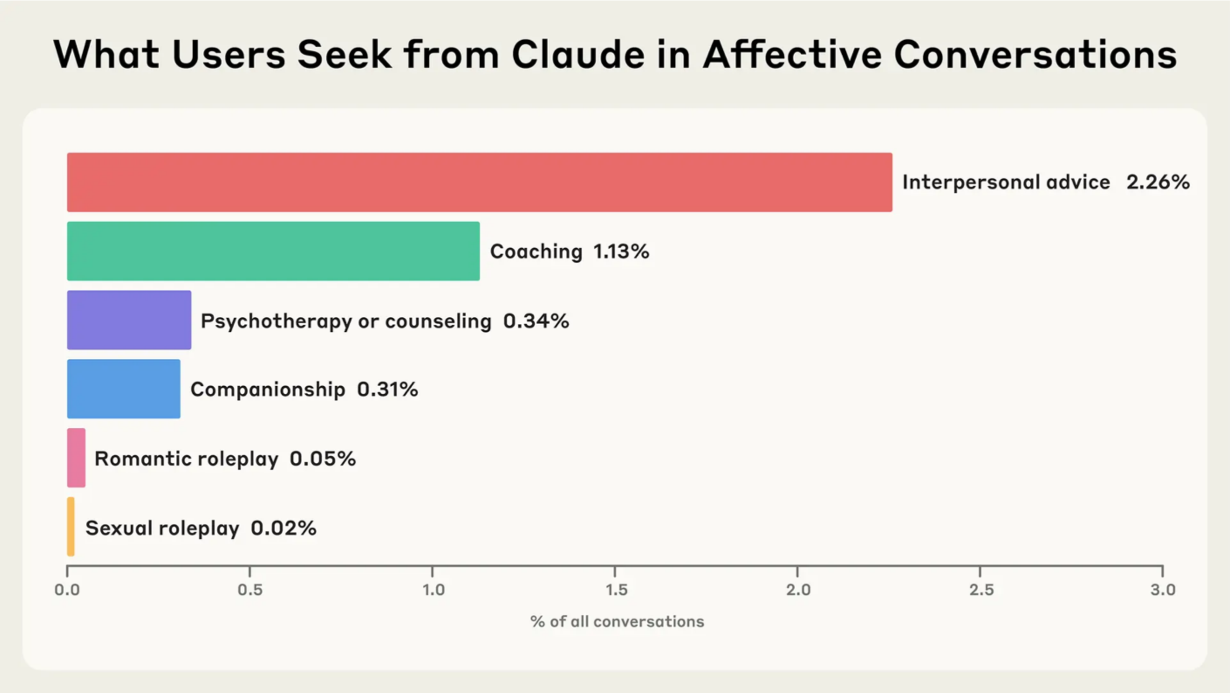
✓ How to mitigate and intervene on the harms from anthropomorphism

✓ How to help users develop appropriate mental models of LLMs



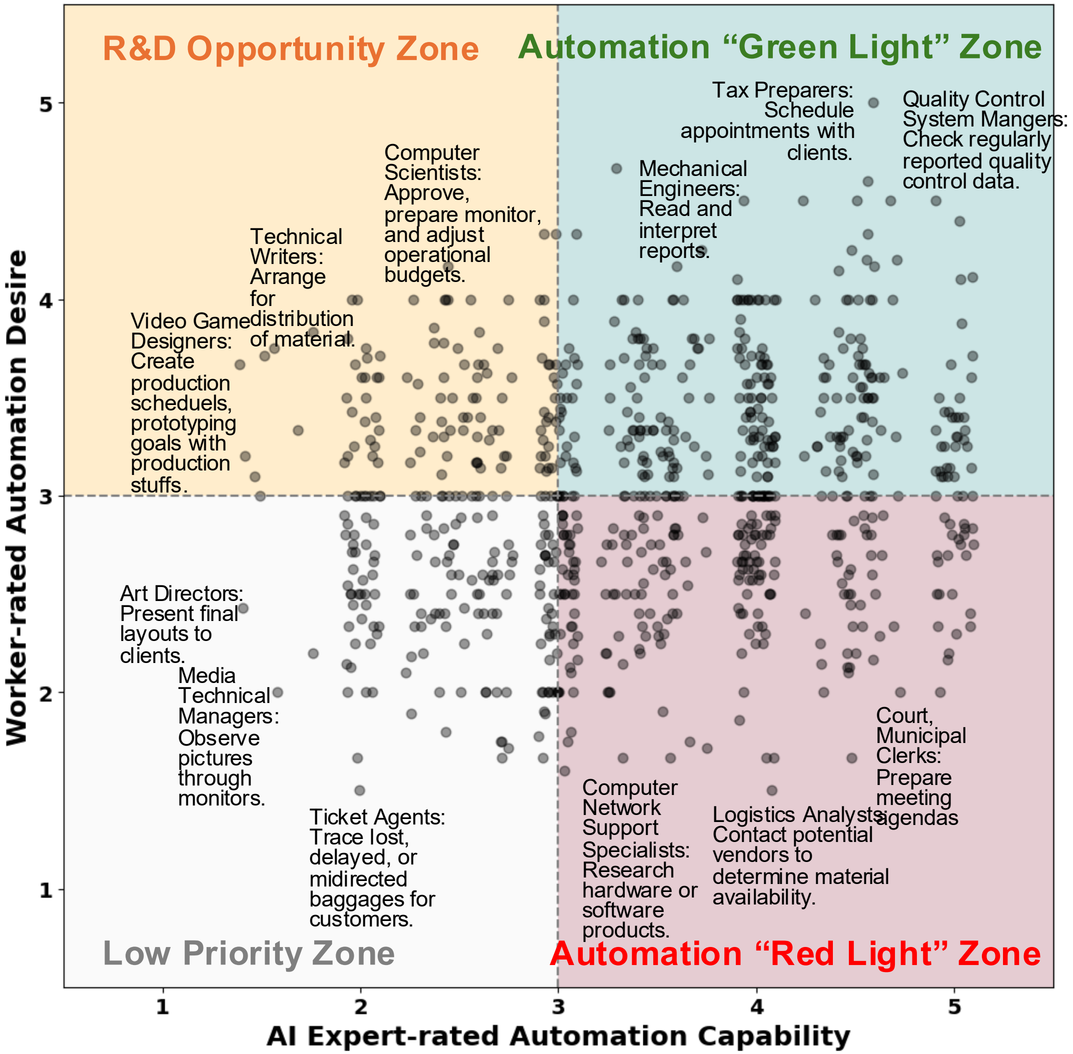
AI Companions

- ✓ Lack of clear boundaries for AI companions
- ✓ Understand both benefits and harms of AI companions
- ✓ **How do we support long-term benefits for users?**
- ✓ **What is the design space of interventions?**



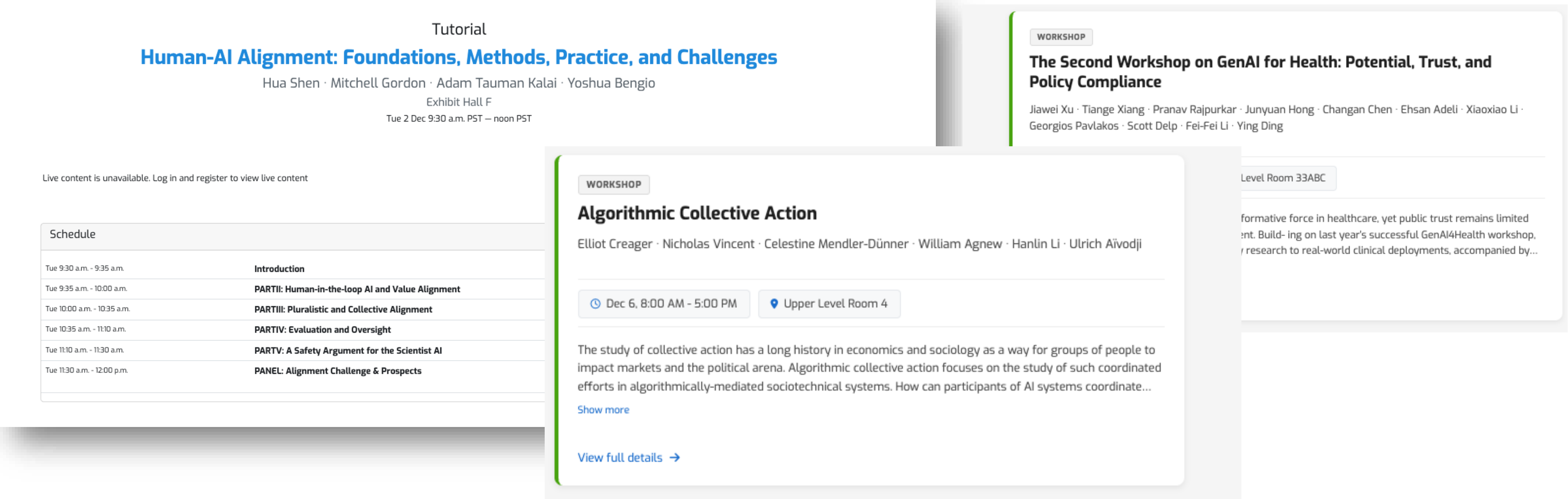
AI and Future of Work

- ✓Economic impacts of LLMs
- ✓LLMs & economically valuable tasks
- ✓Future of work with AI agents



(Shao et al., 2025)

Conference Highlights around Humans + AI



Tutorial

Human-AI Alignment: Foundations, Methods, Practice, and Challenges

Hua Shen · Mitchell Gordon · Adam Tauman Kalai · Yoshua Bengio

Exhibit Hall F

Tue 2 Dec 9:30 a.m. PST — noon PST

Live content is unavailable. Log in and register to view live content

Schedule

Tue 9:30 a.m. - 9:35 a.m.	**Introduction**
Tue 9:35 a.m. - 10:00 a.m.	**PARTII: Human-in-the-loop AI and Value Alignment**
Tue 10:00 a.m. - 10:35 a.m.	**PARTIII: Pluralistic and Collective Alignment**
Tue 10:35 a.m. - 11:10 a.m.	**PARTIV: Evaluation and Oversight**
Tue 11:10 a.m. - 11:30 a.m.	**PARTV: A Safety Argument for the Scientist AI**
Tue 11:30 a.m. - 12:00 p.m.	**PANEL: Alignment Challenge & Prospects**

WORKSHOP

The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance

Jiawei Xu · Tiange Xiang · Pranav Rajpurkar · Junyuan Hong · Changan Chen · Ehsan Adeli · Xiaoxiao Li · Georgios Pavlakos · Scott Delp · Fei-Fei Li · Ying Ding

Level Room 33ABC

formative force in healthcare, yet public trust remains limited
ent. Build- ing on last year's successful GenAI4Health workshop,
y research to real-world clinical deployments, accompanied by...

WORKSHOP

Algorithmic Collective Action

Elliot Creager · Nicholas Vincent · Celestine Mendler-Dünner · William Agnew · Hanlin Li · Ulrich Aïvodji

Dec 6, 8:00 AM - 5:00 PM   Upper Level Room 4

The study of collective action has a long history in economics and sociology as a way for groups of people to impact markets and the political arena. Algorithmic collective action focuses on the study of such coordinated efforts in algorithmically-mediated sociotechnical systems. How can participants of AI systems coordinate...

- Hallucination mitigation and safety in LLMs
- Human-AI collaboration and preference learning
- Fairness, bias, and alignment in LLMs

Emerging External Interests around HCLLMs

- Safety and risks
- Alignment
- Human-AI collaboration
- LLMs companions
- Economically valuable tasks
- Multimodal + multilingual LLMs
- Privacy and data governance
- AI scientists
- Coding agents
- Small and efficient LLMs
- Just better LLMs …

Many Open Questions

- Low-resource language and dialects
- Alignment
- Evaluation and interpretability
- Global representation
- Trust
- Safety in LLMs and their applications
- Human-AI collaboration and collective intelligence
- Copyright, data and privacy
- Lots of cool applications in LLMs + societally important domains
- …

Final Thoughts

1. What is human centered NLP

2. How to build human centered NLP
   - Data, formulation, and technical challenges
   - Interdisciplinary methods

3. What does the “progress” look like for human centered NLP

4. What does HCNLP bring to society & vice versa