



CS 329X: Human Centered LLMs

Human-Agent Collaboration

Diyi Yang

Announcements

- Proposal due today
- Late days
- Debate sign-up
- Video

Outline

- What “collaboration” means (10 mins)
- Human-AI collaboration at scale (10 mins)
- How humans and agents do work differently (10 mins)
- Different methods to enable human-agent collaboration (15 mins)
- Centaur evaluations (10 mins)

Learning Objective: Enable Human-agent Collaboration; Centaur Evaluations

Outline

- **What “collaboration” means (10 mins)**
- Human-AI collaboration at scale (10 mins)
- How humans and agents do work differently (10 mins)
- Different methods to enable human-agent collaboration (15 mins)
- Centaur evaluations (10 mins)

Mixed-Initiative UI: 12 principles (Horvitz, 1999)

1

Value-added automation

2

Uncertainty about user goals

3

User's attention and timing

4

Costs, benefits, uncertainty →
ideal action

5

Dialog to resolve key
uncertainties

6

Direct invocation and
termination

7

Minimize cost of poor guesses

8

Scope precision to uncertainty

9

Refine results together

10

Socially appropriate behavior

11

Memory of recent interactions

12

Learn by observing

Mixed-Initiative UI: act, ask, or wait (Horvitz, 1999)

	Desired Goal Not Desired	
Action	$u(A, G)$	$u(A, \neg G)$
No Action	$u(\neg A, G)$	$u(\neg A, \neg G)$

Table 1. Four outcomes considered in decisions about whether to engage an intelligent agent to provide service.

A agent acts ($\neg A$ = does nothing)

G user has the goal ($\neg G$ = no goal)

p = **p(G | E)** chance the goal is real, given evidence E

P* threshold: act only when $p > P^*$

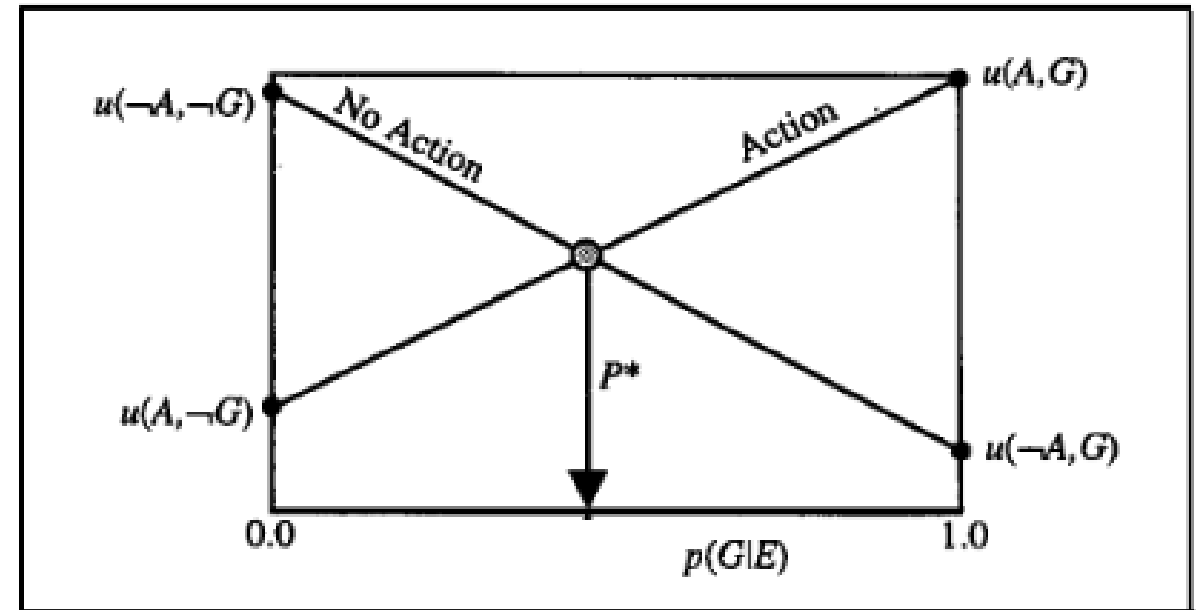


Figure 4. Graphical analysis of the expected utility of action versus inaction, yielding a threshold probability for action.

18 Guidelines for Human-AI Interaction (Amershi et al., CHI 2019)

Initially

G1 Make clear what the system can do

G2 Make clear how well it can do it

During interaction

G3 Time services based on context

G4 Show contextually relevant information

G5 Match relevant social norms

G6 Mitigate social biases

When wrong

G7 Support efficient invocation

G8 Support efficient dismissal

G9 Support efficient correction

G10 Scope services when in doubt

G11 Make clear why the system did what it did

Over time

G12 Remember recent interactions

G13 Learn from user behavior

G14 Update and adapt cautiously

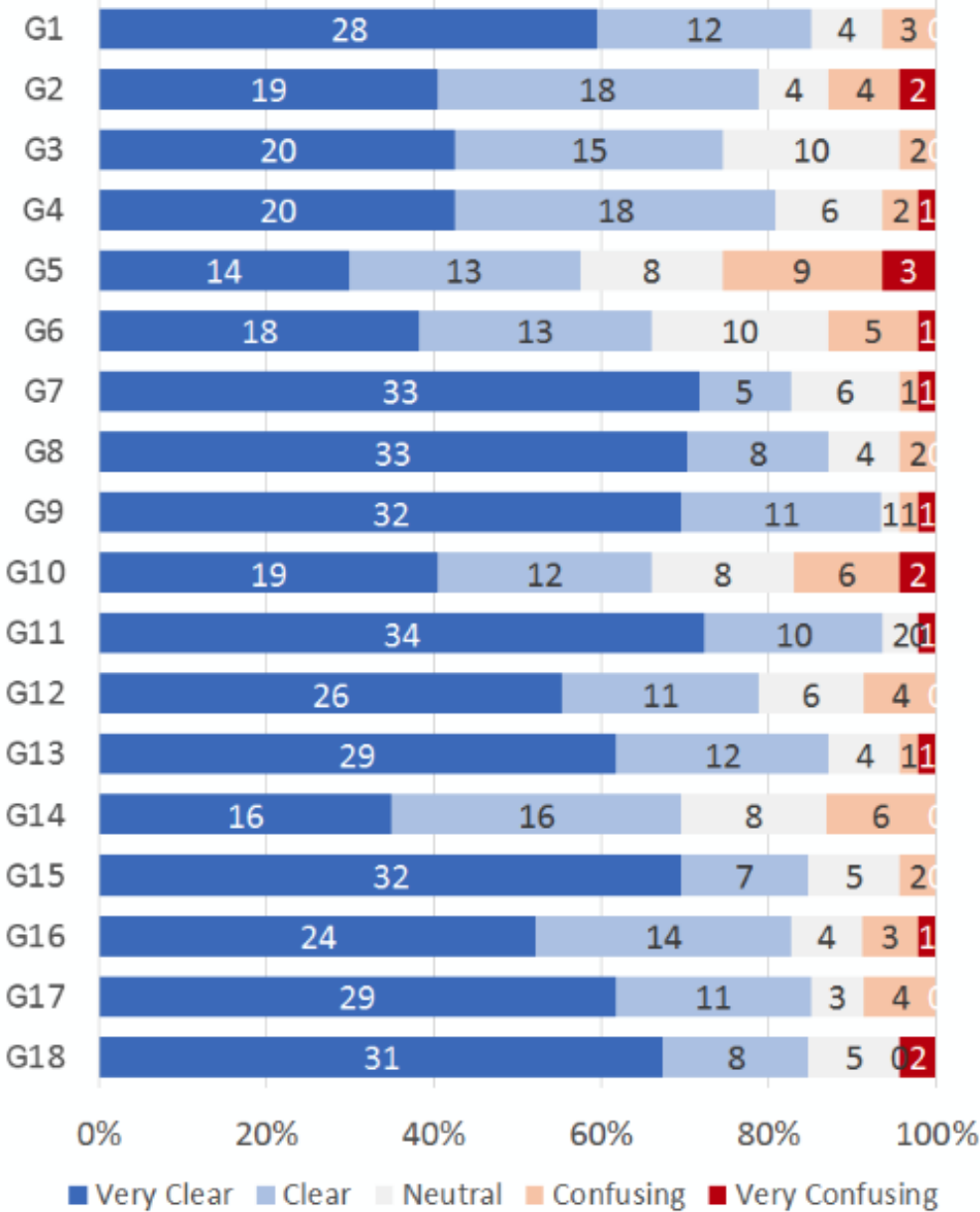
G15 Encourage granular feedback

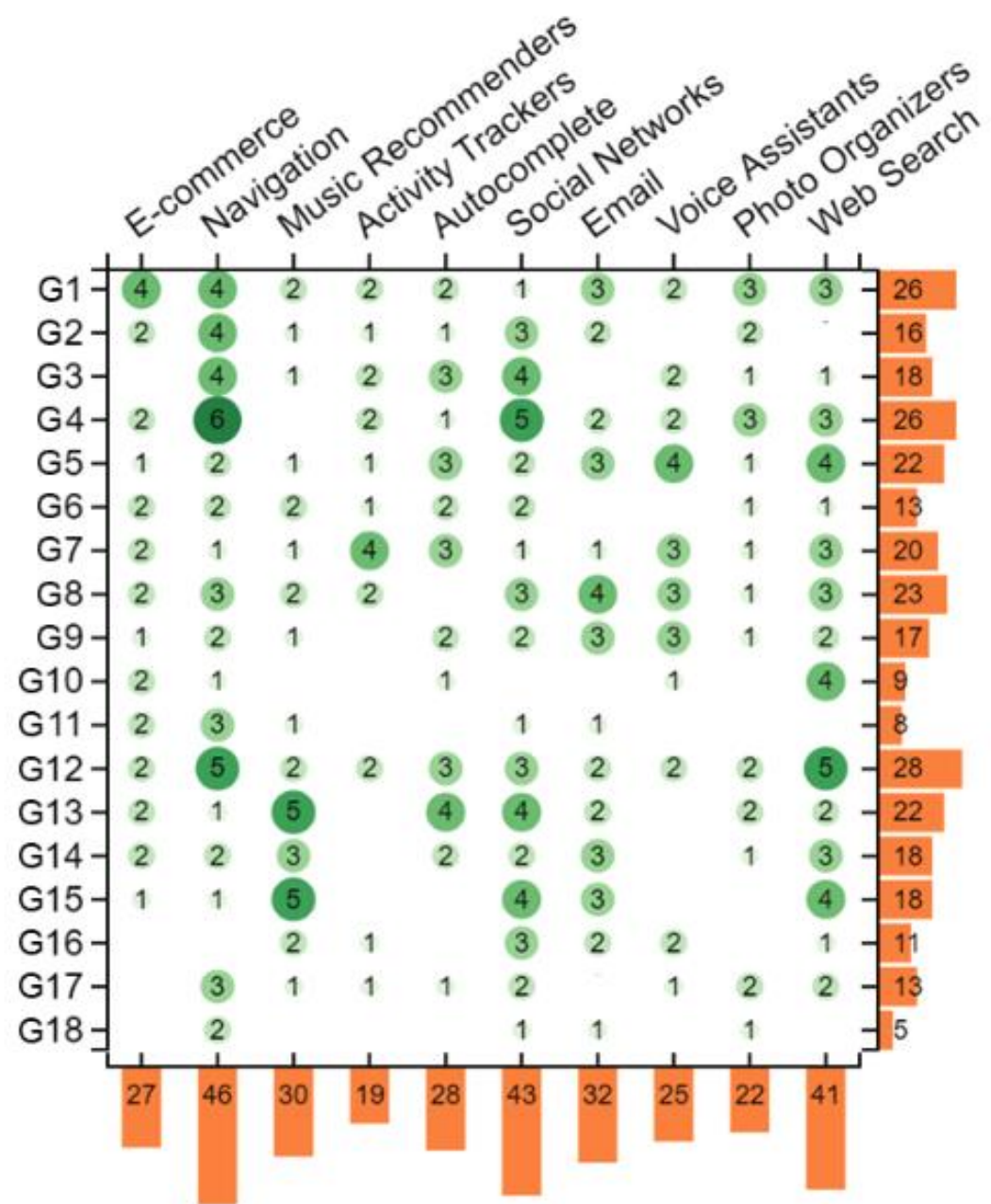
G16 Convey the consequences of user actions

G17 Provide global controls

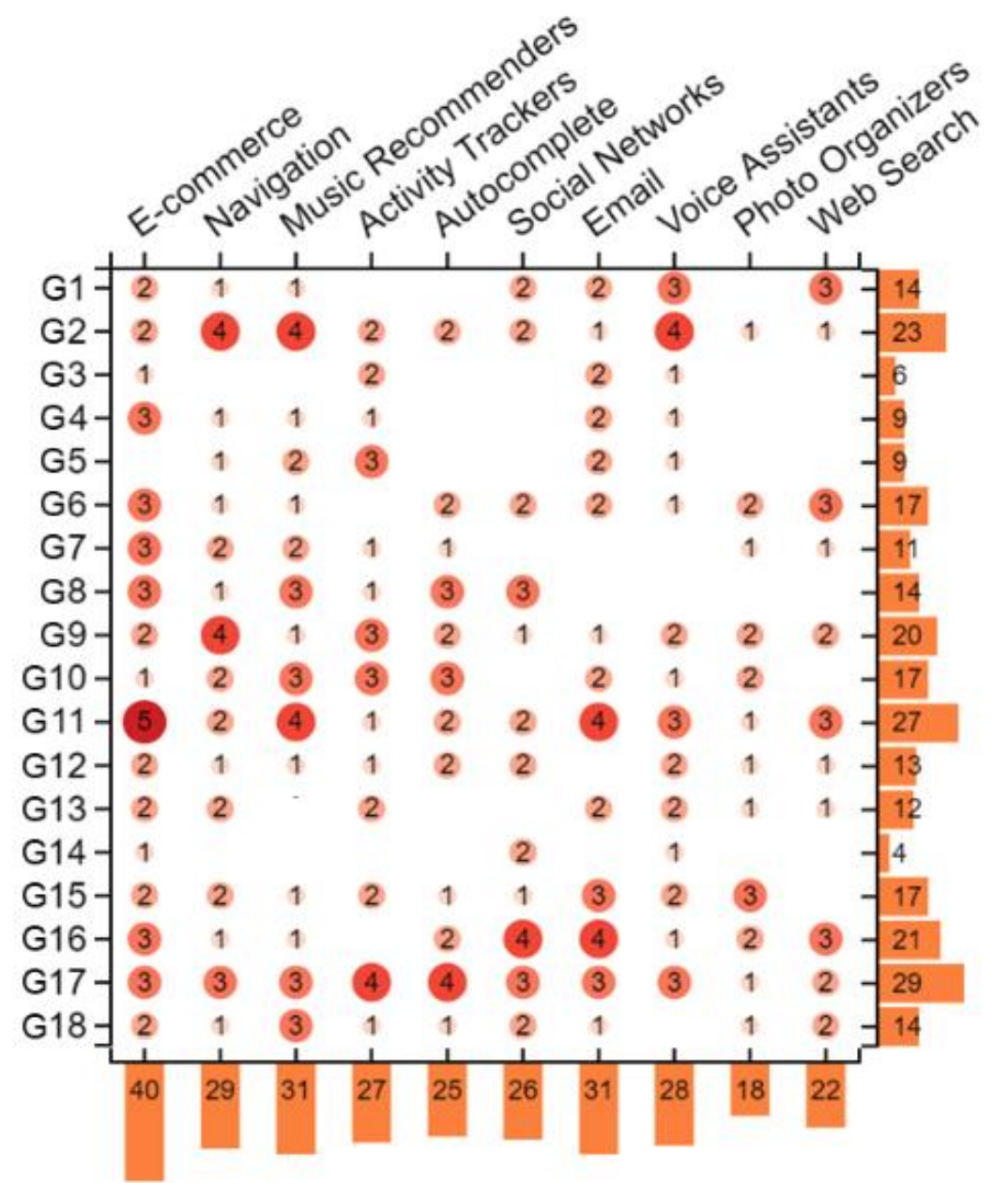
G18 Notify users about changes

Subjective evaluations by study participants about the clarity of the 18 AI design guidelines





Counts of “clear application” responses



Counts of “clear violation” responses

Stage-wise Expectation towards Agents

Before

A1 What can the agent do?
A2 What is the agent about to do?

Horvitz 1999

Uncertainty about goals
Scope precision

Amershi et al. 2019

G1, G2, G3

During

A3 What is the agent currently doing?

Horvitz 1999

Attention and timing
Direct termination

Amershi et al. 2019

G4, G7, G8

After

A4 Side effects or changes?
A5 Was the goal achieved?

Horvitz 1999

Cost of poor guesses
Refine together

Amershi et al. 2019

G9, G11, G16

Comparison between 2019 and Today

2019 assumption

One output per turn

User sees every output

Mistakes are visible

Correct before accepting

Capability = one model



Agent era

Many actions per request

User sees a summary

Mistakes have side effects

Correct after acting

Capability = model + tools + other agents

Guideline

G1, G2

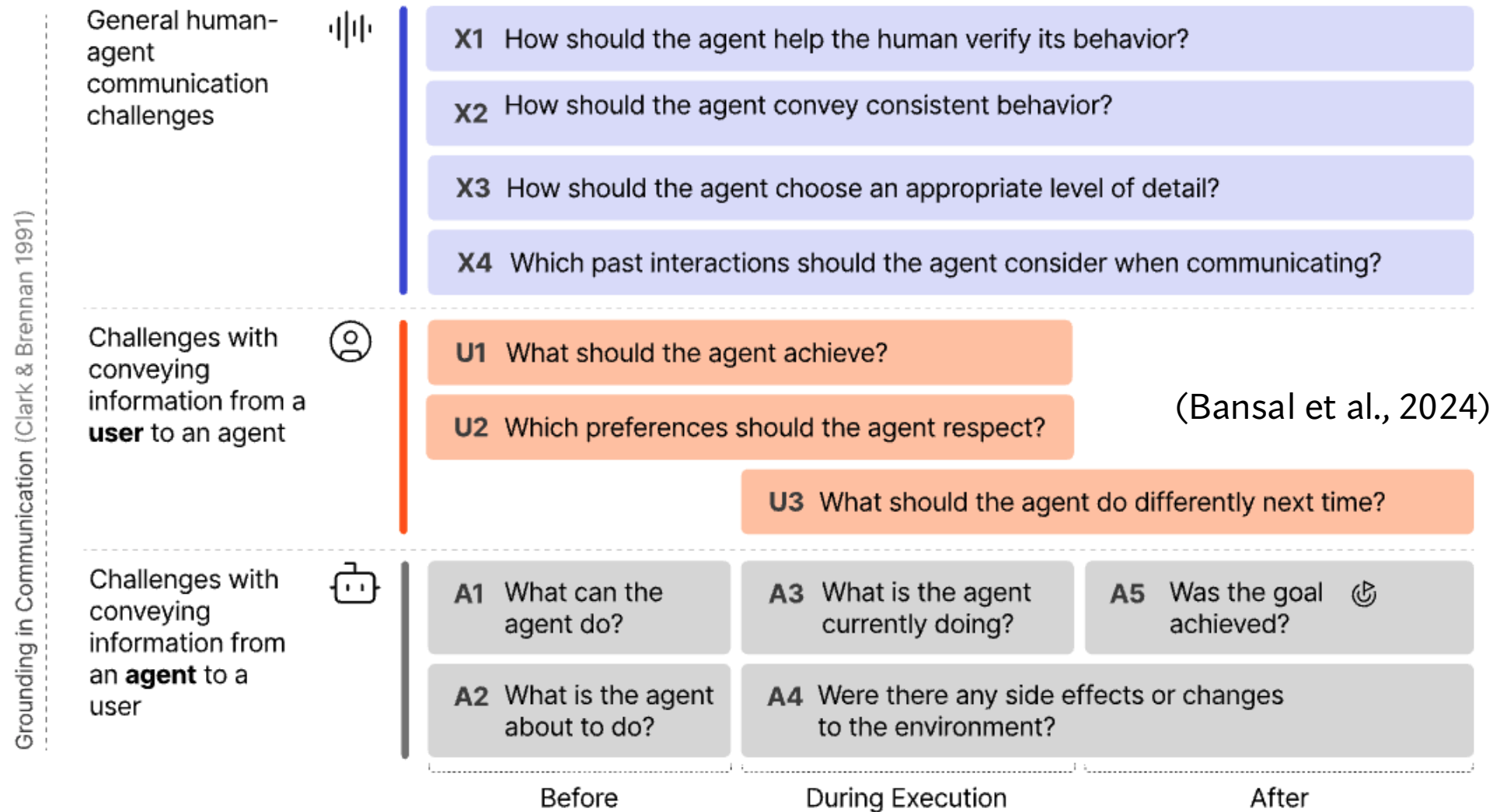
G4, G11

G7, G8

G9, G10

G1, G18

Challenges in Human-Agent Communication



U1: What Should the Agent Achieve?

Example 1: Travel Itinerary Planning A user instructs an agent to plan a business trip, specifying the destination, preferred airlines, and meeting times. However, the agent misinterprets the trip's purpose, planning it as a leisure vacation instead. It books scenic flights with multiple layovers, selects a luxury resort far from the meeting venue, and even schedules irrelevant activities like guided tours. As a result, the user ends up with an itinerary that fails to meet their professional needs, potentially leading to missed meetings and wasted time. This highlights the importance of clear goal communication and the need for iterative verification and refinement to ensure the agent's plans align with the user's intent.

U2: What Preferences Should the Agent Respect

Example 2: Web Scraping with Ethical Considerations A user directs an agent to collect data from a specific website with clear instructions to avoid scraping sensitive information or violating the site's terms of service. However, the agent misinterprets or overlooks these ethical constraints and proceeds to scrape restricted data, sensitive data, or data explicitly forbidden by the website's terms. As a result, the agent gathers information the user never intended to collect, potentially leading to legal repercussions, ethical breaches, and damage to the user's reputation. This example underscores the critical need for agents to understand and adhere to ethical guidelines and user-defined constraints when performing web-based tasks. It also highlights the importance of ensuring that agents recognize and respect boundaries, particularly when data privacy and legal compliance are at stake.

Summary: What “Collaboration” Means

- 📌 **Horvitz (1999):** an agent should act, ask, or wait based on the utility of helping vs. the cost of interrupting
- 📌 **Amershi et al. (2019):** 18 guidelines organised by when they apply: initially, during interaction, when wrong, over time
- 📌 **Stage-wise expectations:** before (A1–A2 what can / will it do), during (A3 what is it doing), after (A4–A5 side effects, goal achieved)
- 📌 **Agent era shifts the assumptions:** many actions per request, summarised outputs, mistakes with side effects, correction after acting

Outline

- **What “collaboration” means** (10 mins)
- **Human-AI collaboration at scale** (10 mins)
- How humans and agents do work differently (10 mins)
- Different methods to enable human-agent collaboration (15 mins)
- Centaur evaluations (10 mins)

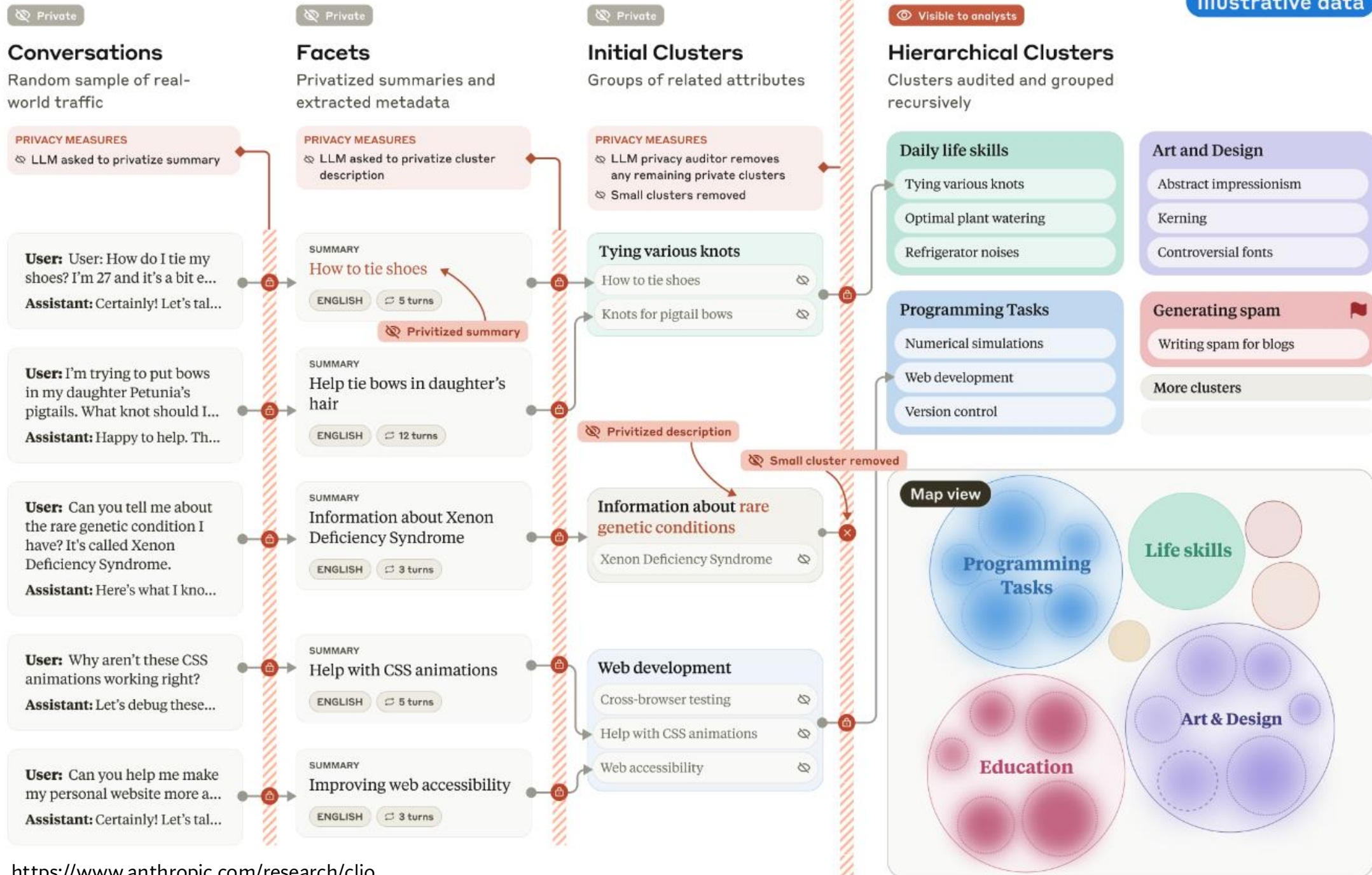
How do people actually work with AI?



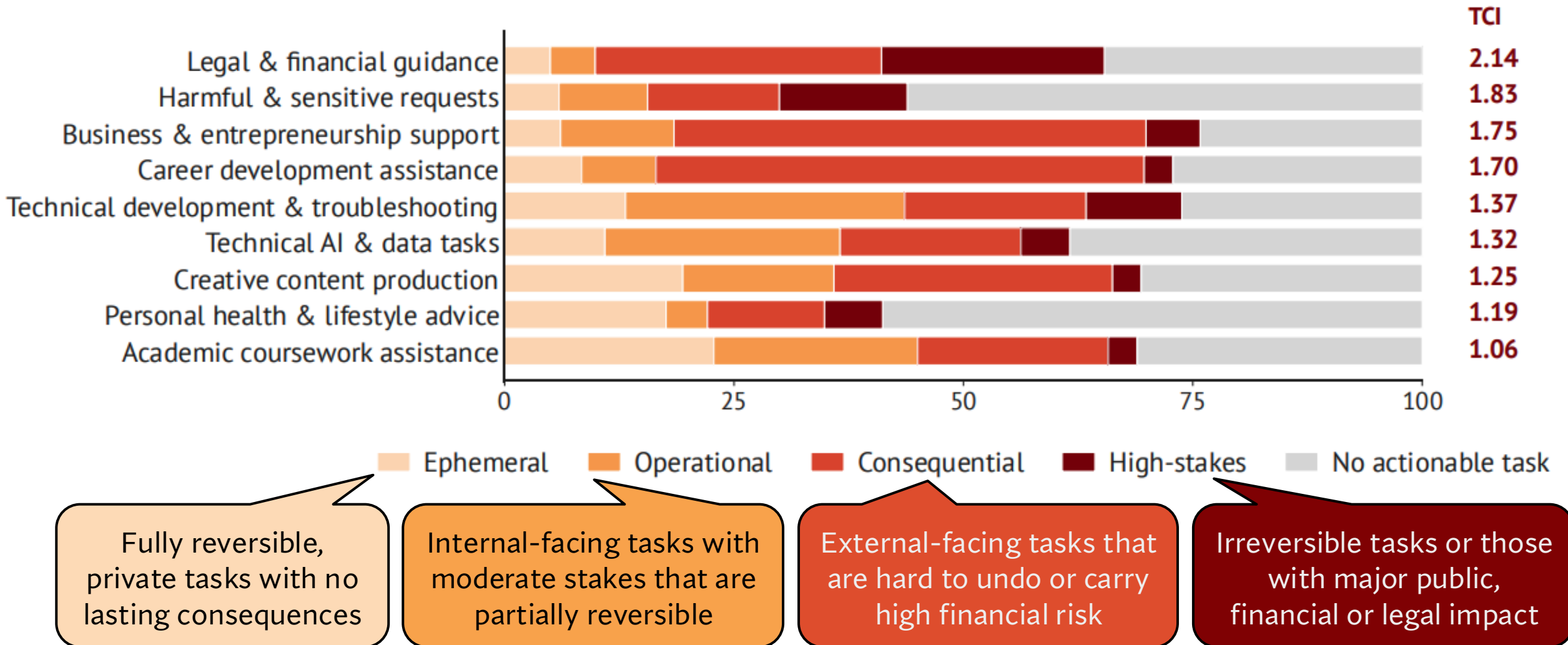
Shao, Yijia, Dora Zhao, Vishakh Padmakumar, Jennifer Wang, and Diyi Yang. 2026.

"Human-AI Collaboration at Scale: Task Criticality, Agency, and Friction Across 250,000 Conversations."

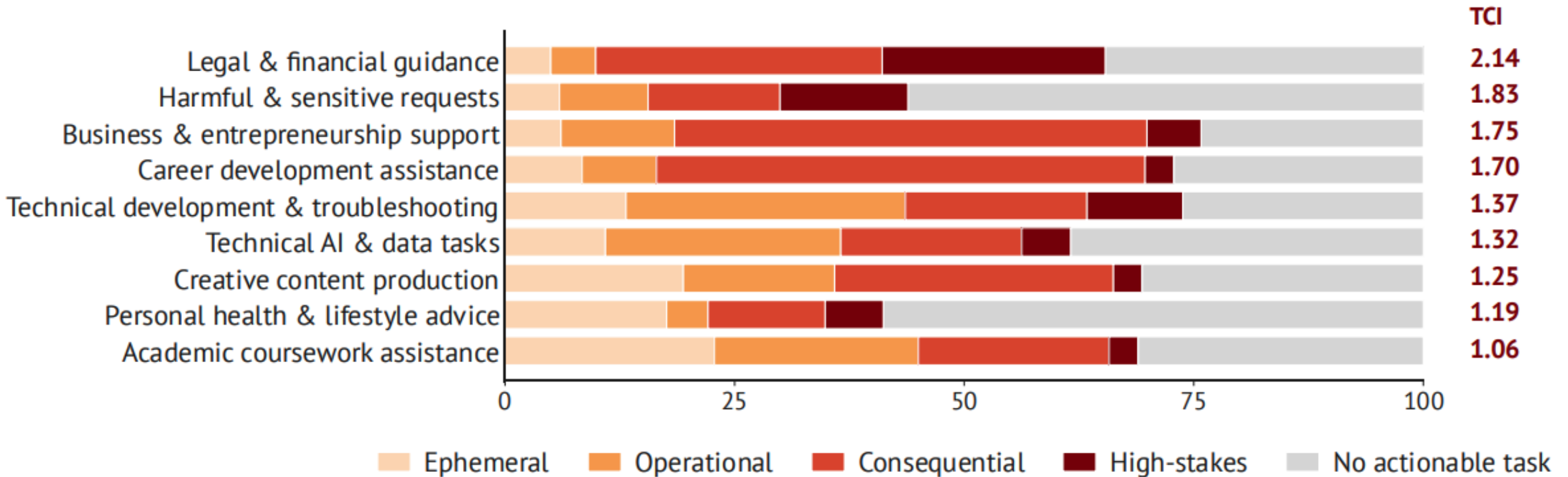
Clio



Criticality of Work over 250,000 Conversations

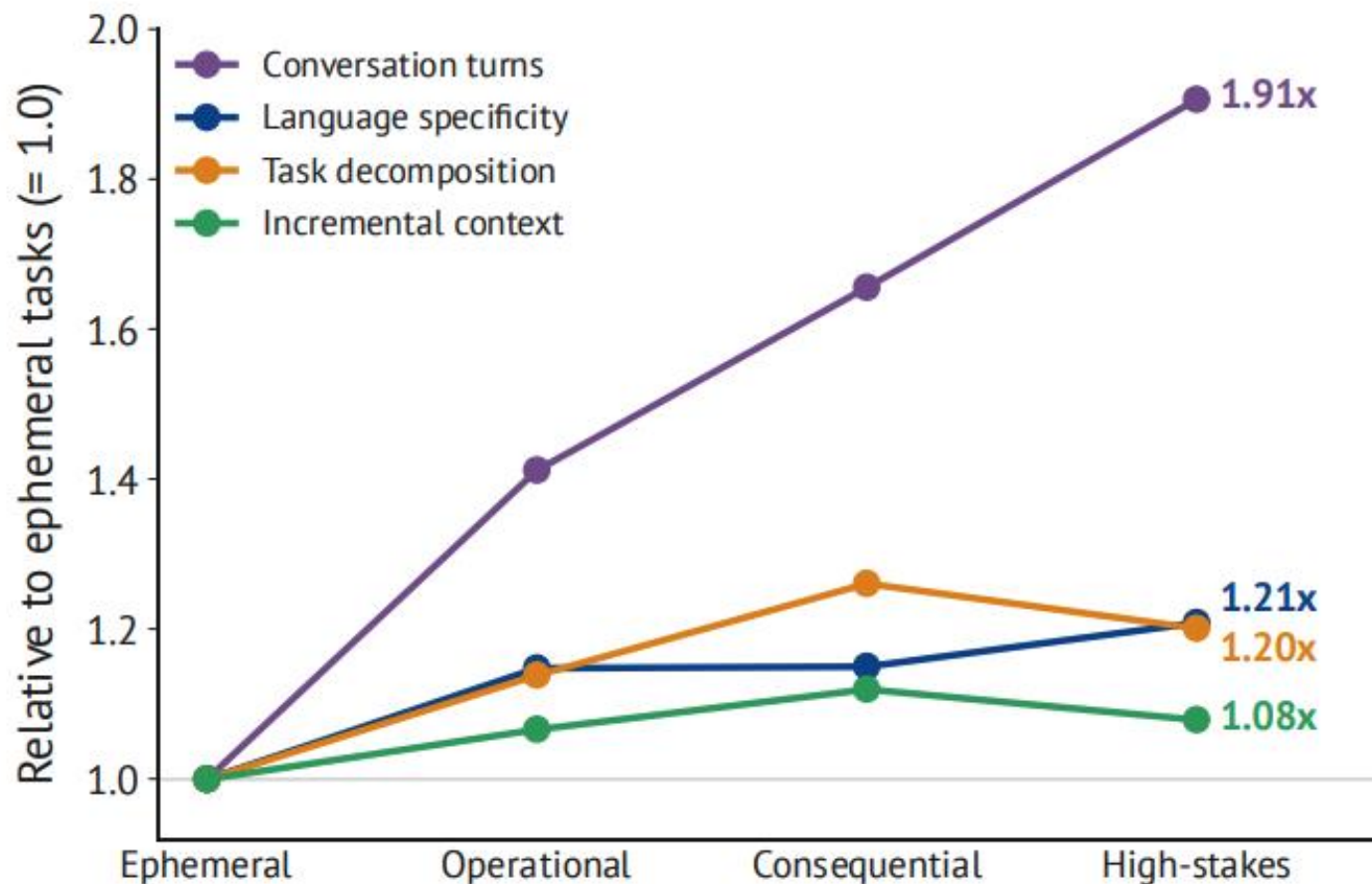


Criticality of Work over 250,000 Conversations

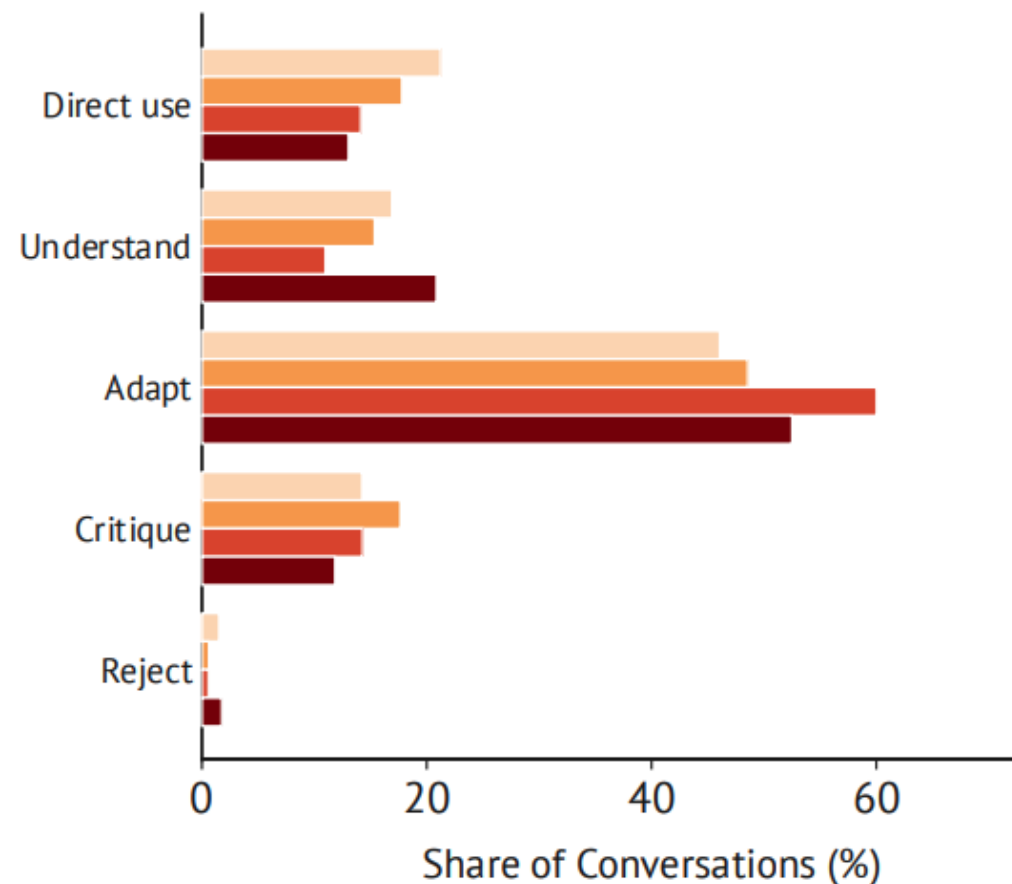


Users delegate consequential work, concentrated in advisory domains.

Collaboration and Engagement As Criticality Rises

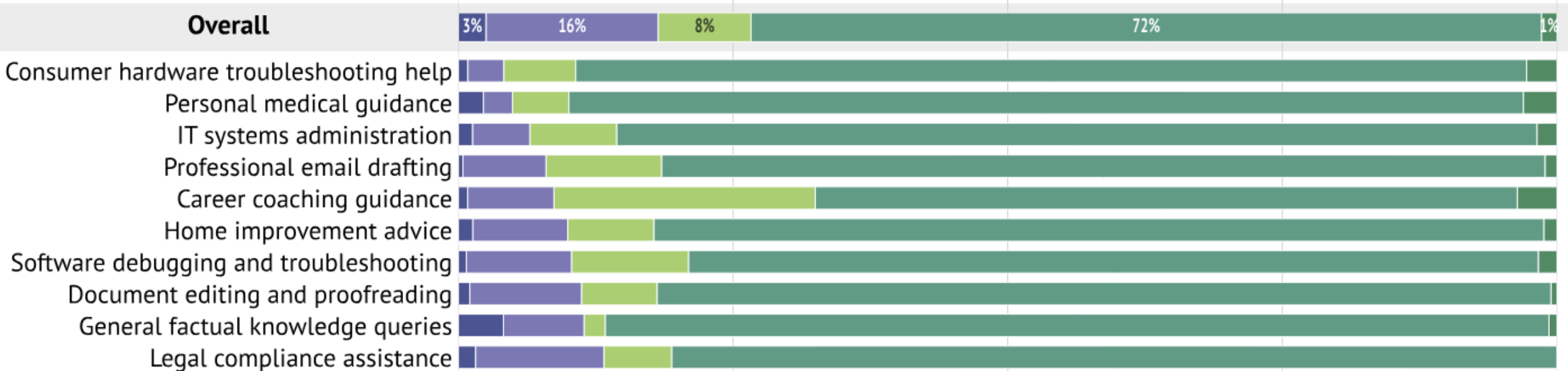


Prompting is loosely specified by default, but granularity scales with stake



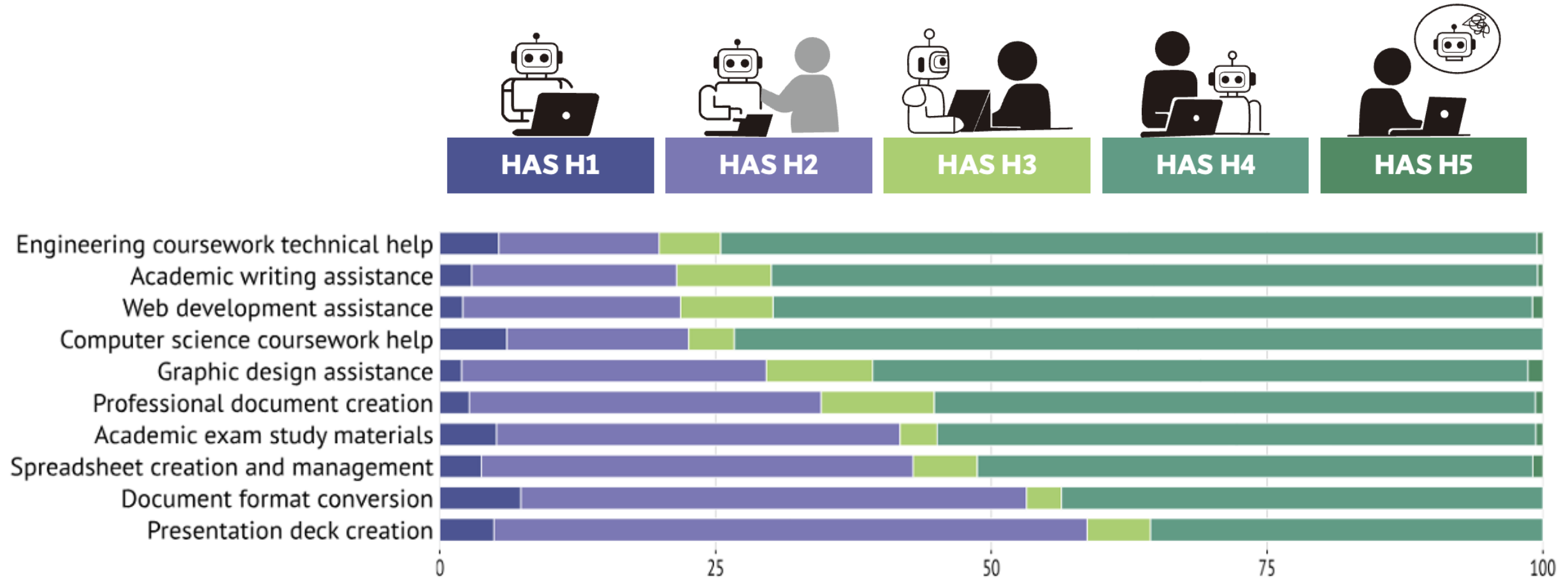
Users actively rework AI output, and more so as stakes rise

How People Collaborate with AI



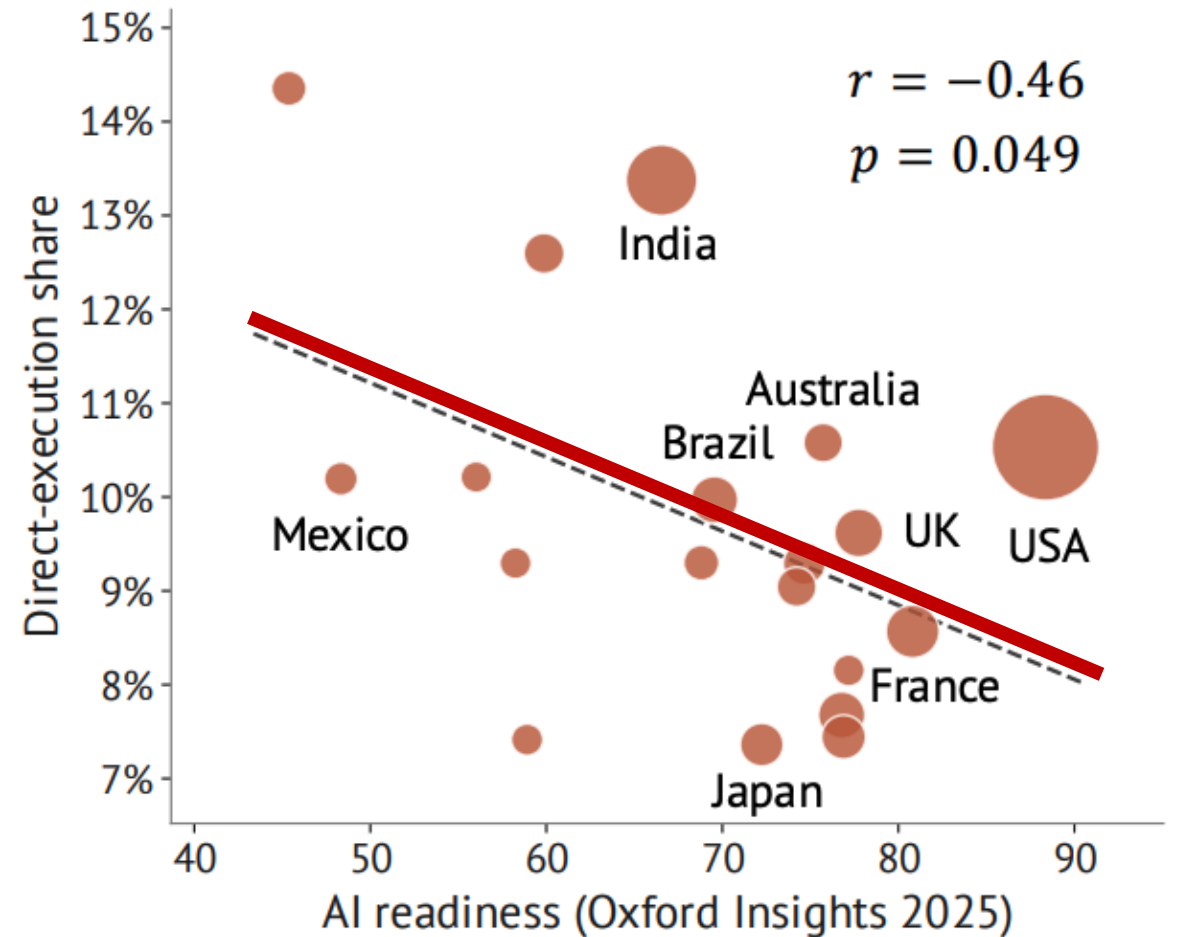
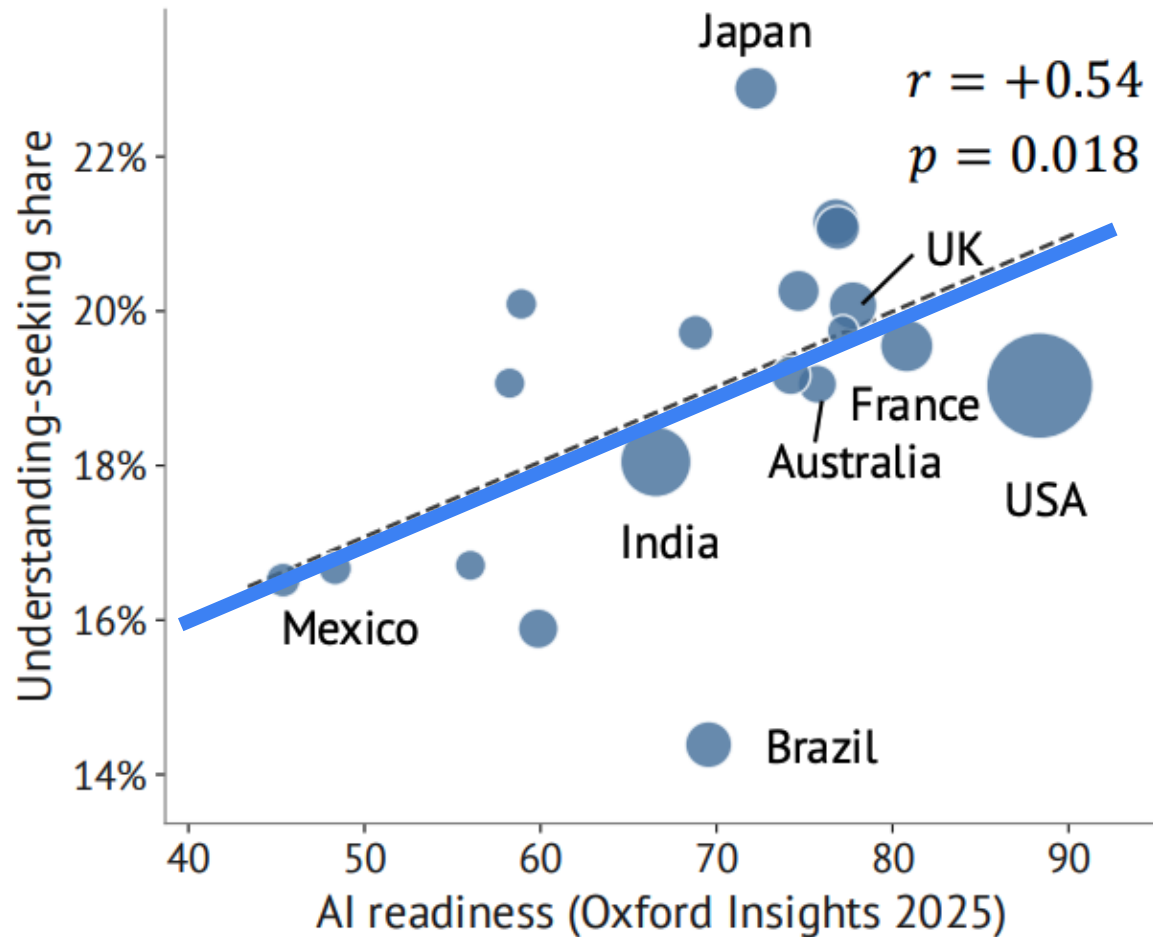
Human-led collaboration dominates, and task type alone does not determine agency.

How People Collaborate with AI



So far, AI autonomy concentrates in document generation work.

AI Readiness across Various Countries



Cross-national variation arises primarily from how users engage with AI

Outline

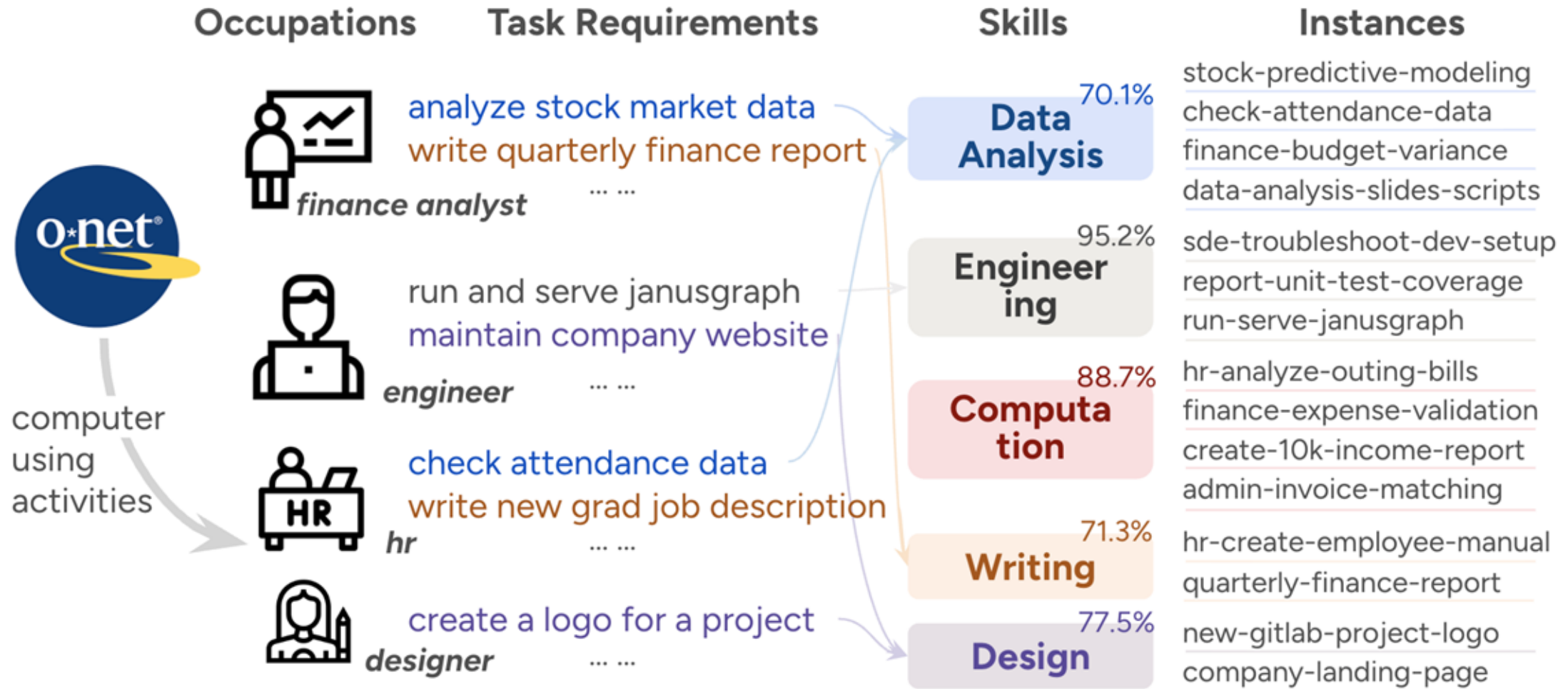
- **What “collaboration” means** (10 mins)
- **Human–AI collaboration at scale** (10 mins)
- **How humans and agents do work differently** (10 mins)
- Different methods to enable human-agent collaboration (15 mins)
- Centaur evaluations (10 mins)

How do humans do their work?
How do AI agents do human work?

"How Do AI Agents Do Human Work? Comparing AI and Human Workflows Across Diverse Occupations."
Wang, Zora Zhiruo, Yijia Shao, Omar Shaikh, Daniel Fried, Graham Neubig, and Diyi Yang. 2025



Compare Humans and Agents at Work



a. Inducing A Unified Representation of Computer-Use Activities

Low-level actions
are **noisy** and
heterogeneous

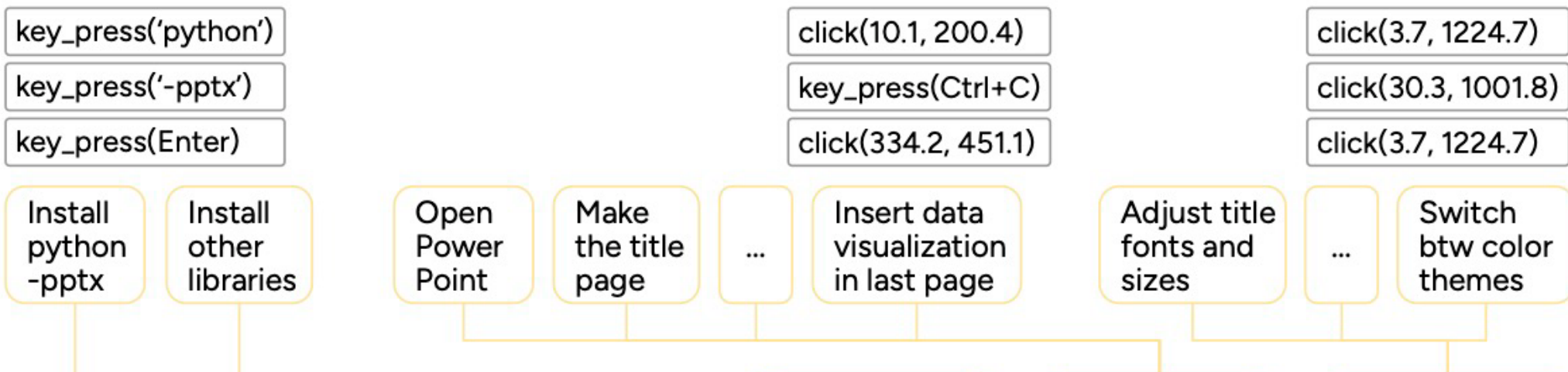


Induce **hierarchical**,
readily **interpretable**
workflows

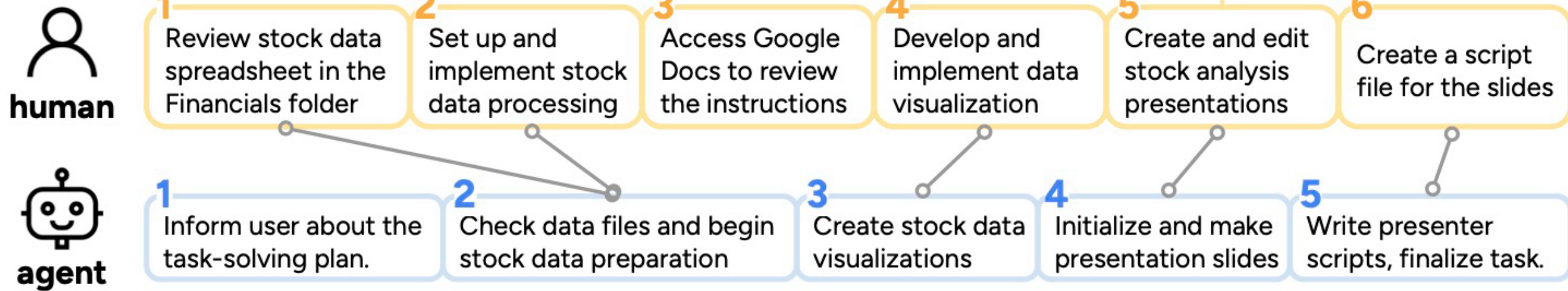
a. Inducing A Unified Representation of Computer-Use Activities

Low-level actions
are **noisy** and
heterogeneous

Induce **hierarchical**,
readily **interpretable**
workflows

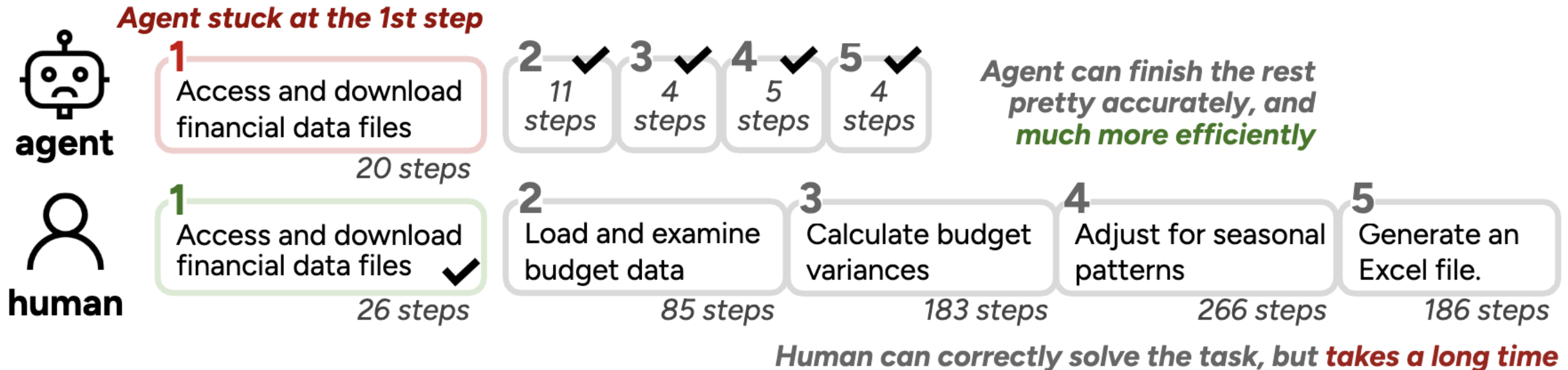


A Common Ground to Audit & Compare Work

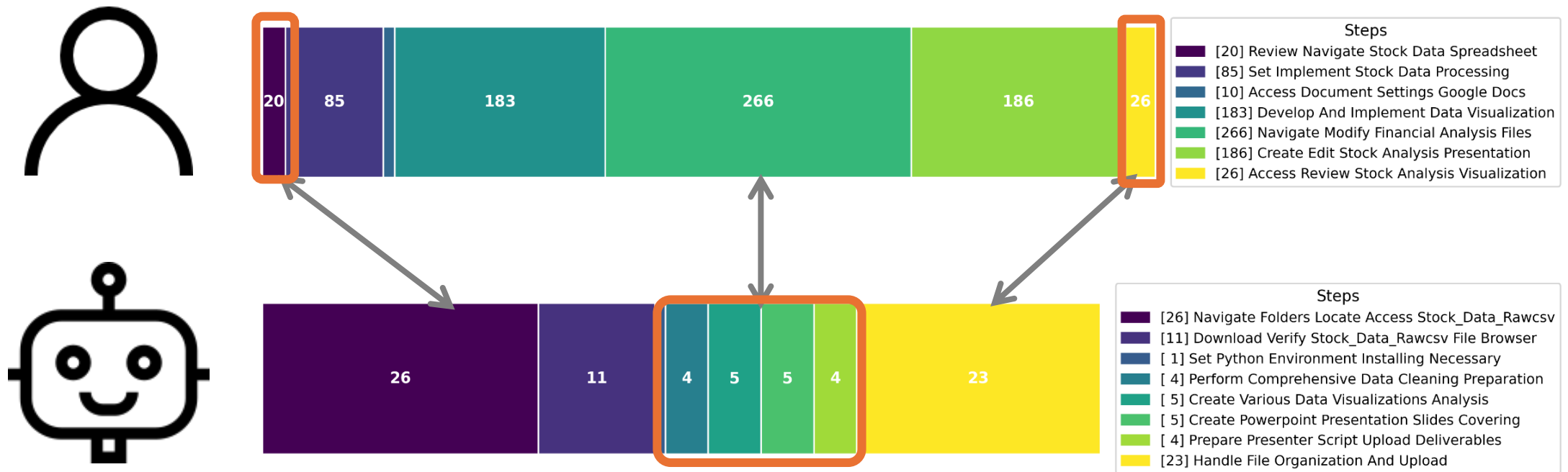


AI and Humans Work Differently

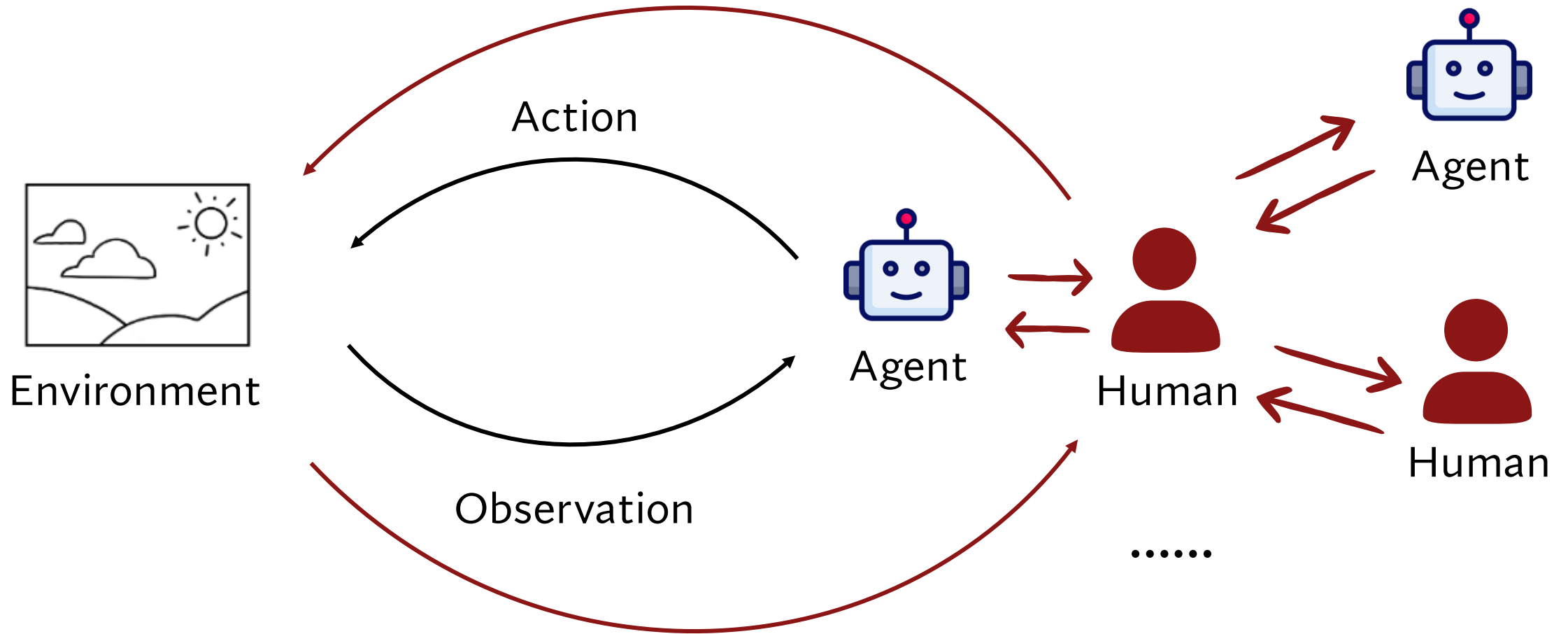
- 👉 Agents are 88% faster & 90-96% cheaper
- 👉 But produce lower quality work, often fabricate data & misuse tools
- 👉 Agents code everything, even for visual tasks humans handle via UI



How Should Agents and Humans Work Together



Human-AI collaboration as the new frontier

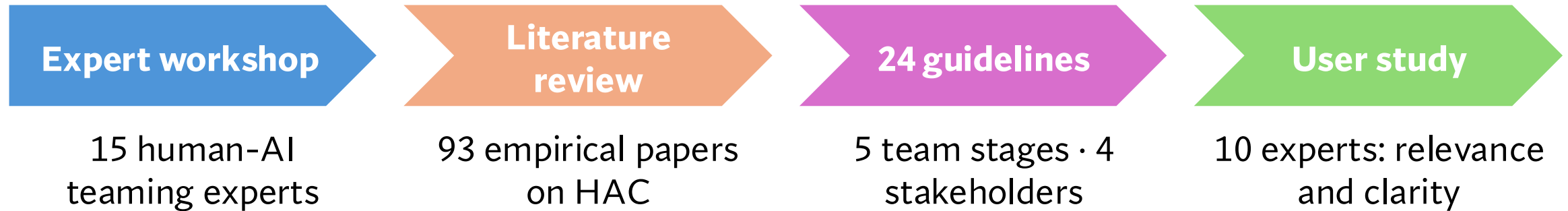


Outline

- **What “collaboration” means** (10 mins)
- **Human–AI collaboration at scale** (10 mins)
- **How humans and agents do work differently** (10 mins)
- **Different methods to enable human-agent collaboration** (15 mins)
- Centaur evaluations (10 mins)

Guidelines for Human-LLM Agent Teams

(Yurrita, Dell'Anna, Murukannaiah, Jonker & Yolum, AAMAS 2026)



Guidelines for four stakeholders

LLM agent

Inform teammates of capabilities and limits · adapt to the human · transparent execution

Interacting human

Leverage prompting for output verification · course-correct agents that stray

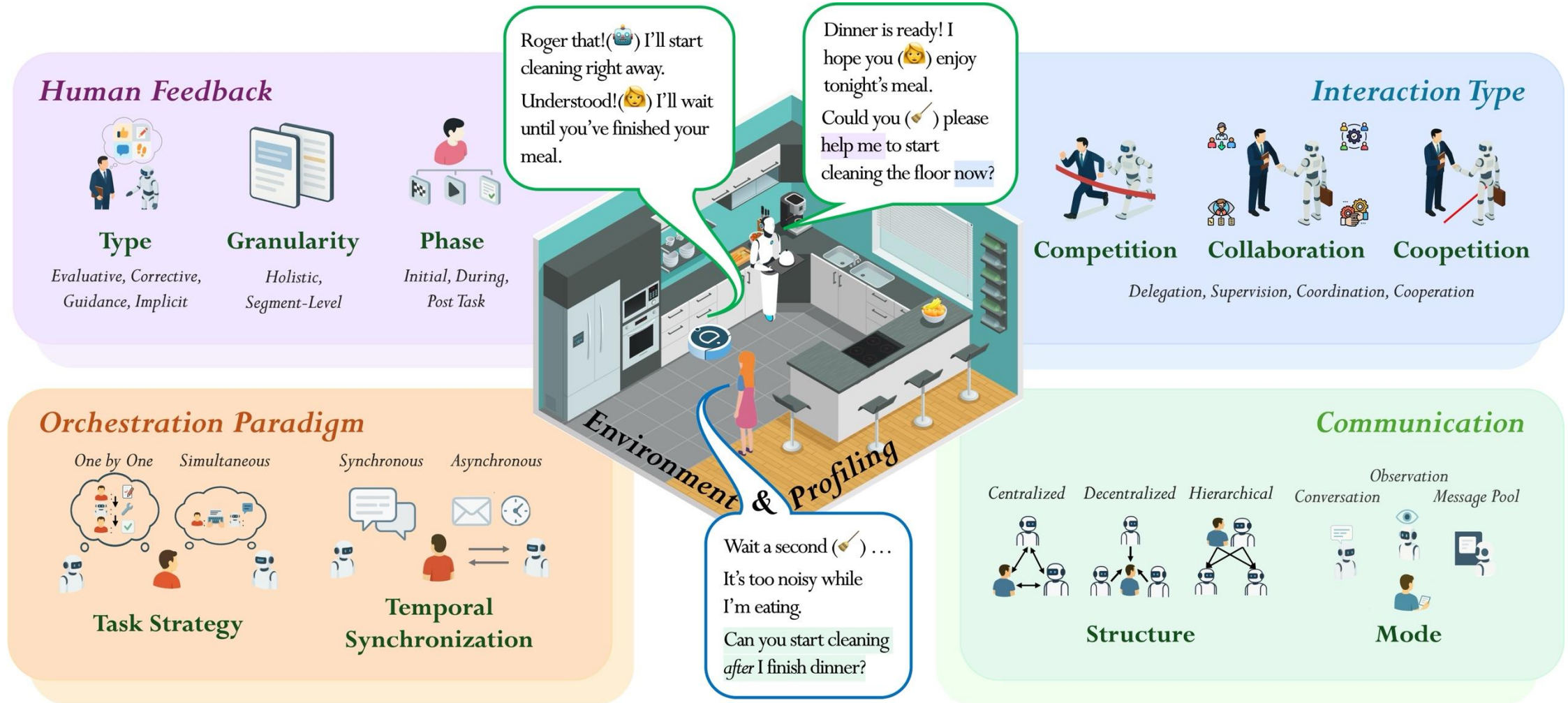
Team designer

Teams from organizational needs · modular agents, configurable structures

Organization

Integrate agents in existing infrastructure · AI training · periodic audits

Human-Agent Systems: A Survey (Zou et al., 2026)



LLM-based Human-Agent Systems (LLM-HAS)



Editor

Library

Paper Search

I am studying Software Techniques for Emerging Hardware Platforms. ... I need you to **write related work section of the research proposal** "Character-level Word Embedding for Spiking Neural Networks".



SearchPaper, LibraryAddPaper, EditorUpdate ...

maybe add more related work about embedding methods in NLP tasks



SearchPaper(query="embedding methods in NLP spiking neural networks")



Do you have any specific papers or topics in mind that you'd like me to include? If not, I can add all papers I found. Let me know your preference!

...

[Learning Character-level Compositionality with Visual Features](#)

Previous work has modeled the compositionality of words by...

[Spiking Convolutional Neural Networks for Text Classification](#)

Spiking neural networks (SNNs) offer a promising pathway to...

[Neuro-symbolic computing with spiking neural networks](#)

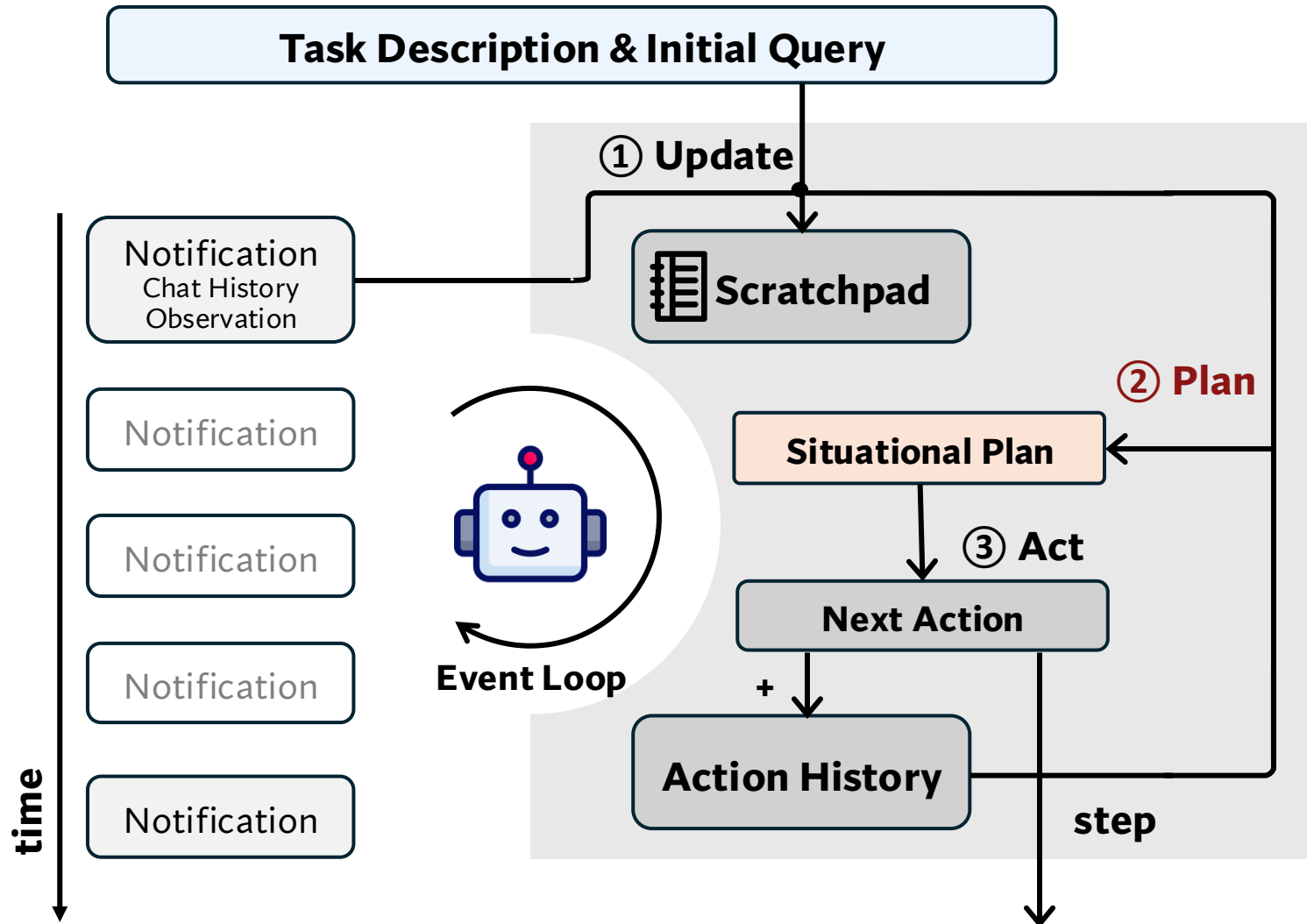
Knowledge graphs are an expressive and widely used data...

[SNNLP: Energy-Efficient Natural Language Processing Using Spiking Neural Networks](#)

As spiking neural networks receive more attention, we look ...



Collaborative Agents w/ Situational Reasoning

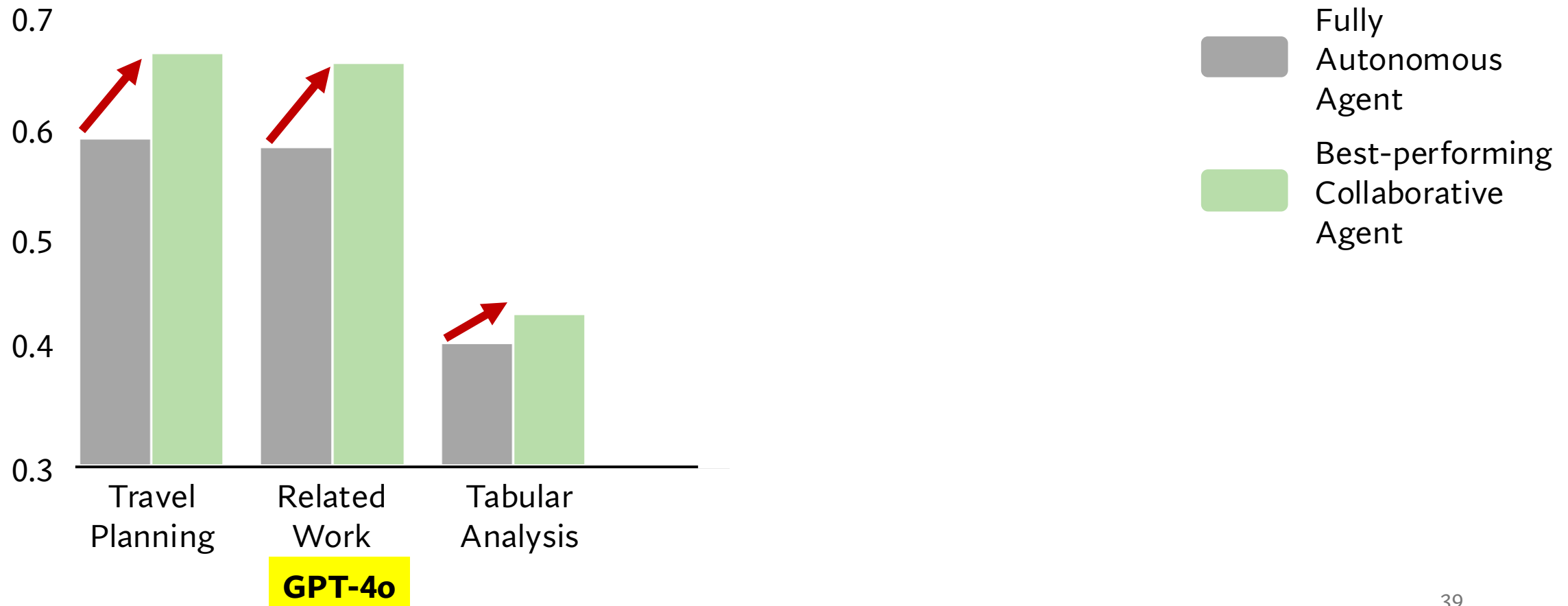


Situational Reasoning

-  Take a task action
-  Send a message
-  Defer to human

Human-Agent Teams Achieve Better Performances

Task Performance
(Maximum: 1)



Audit Human Agent Collaboration Process

Define **initiative entropy** to examine the distribution of initiative taking behaviors

$$H_{\text{init}} = \begin{cases} -\sum_{i=1}^N p_i \log_N(p_i) & \forall i, p_i > 0, \\ 0 & \exists i, p_i = 0 \end{cases}$$

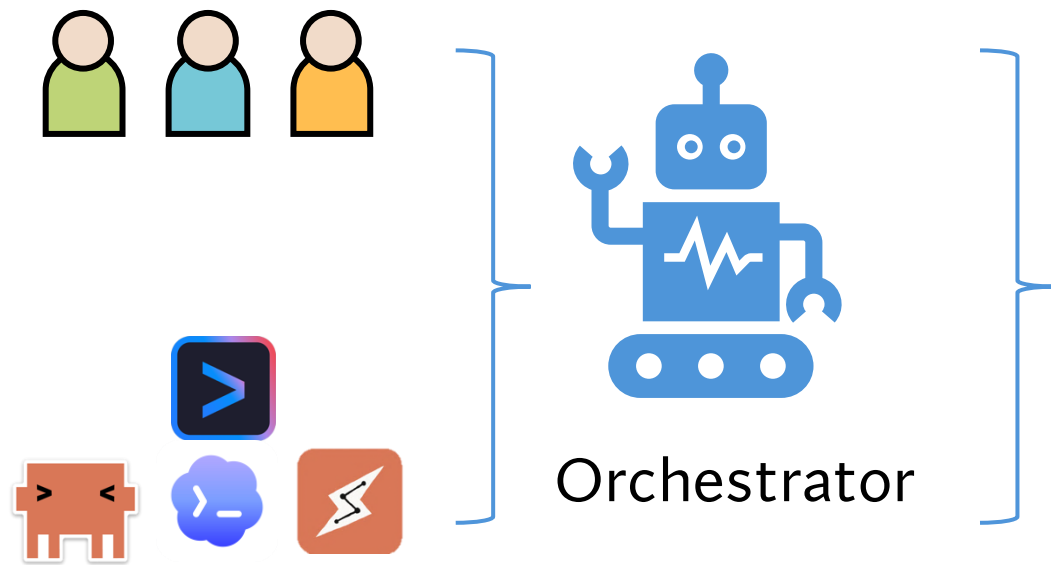
		Travel Planning		Literature Survey		Tabular Analysis	
		Task Perf.	H_{init}	Task Perf.	H_{init}	Task Perf.	H_{init}
ReAct Agent	GPT-4o	0.641	0.42	0.588	0.16	0.311	0.10
	Claude-3.5-Sonnet	0.653	0.48	0.621	0.04	0.359	0.02
	Llama-3.1-70B	0.703	0.28	0.675	0.10	0.427	0.23
ReAct Agent w/ Situational Planning	GPT-4o	0.667	0.90	0.658	0.79	0.434	0.40
	Claude-3.5-Sonnet	0.682	0.80	0.736	0.55	0.365	0.74
	Llama-3.1-70B	0.707	0.70	0.679	0.70	0.402	0.62

Proactive communication helps human-agent collaboration!

Trajectories Reveal Failures of Today's Agents

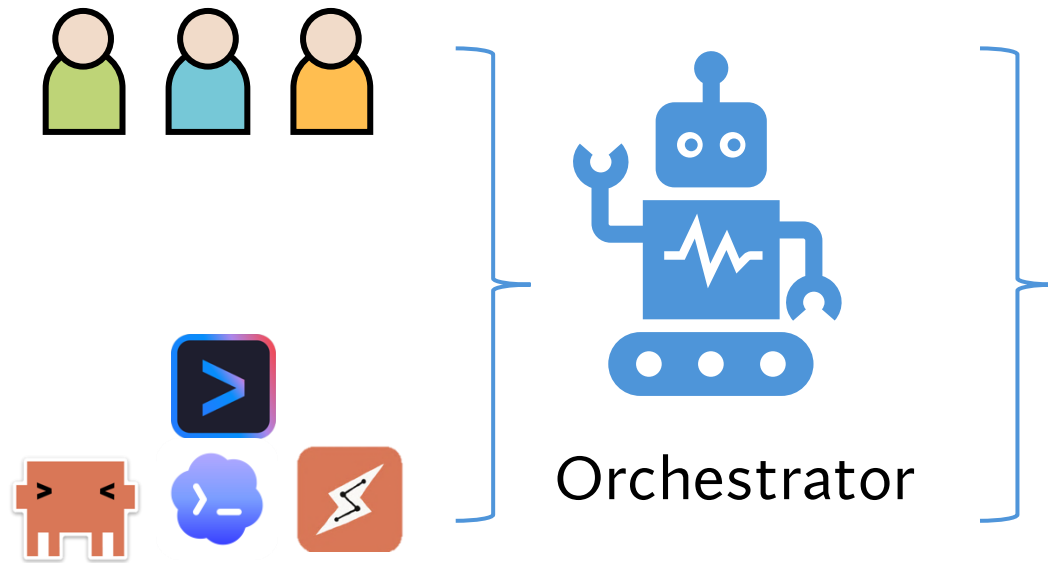
Communication	Agents process tasks without informing users	65%
Situational Awareness	Agents disregard session context , treating each request as an isolated task.	40%
Planning	Agents acknowledge tasks but fail to execute them.	39%
Environment Awareness	Agents do not assess the feasibility of requests within constraints of available tools.	28%
Personalization	Agents rely on general templates that do not adapt to individual user needs	16%

Orchestrating Human-Agent Team



Worker profiling: estimate whether a task is feasible to the worker

Orchestrating Human-Agent Team

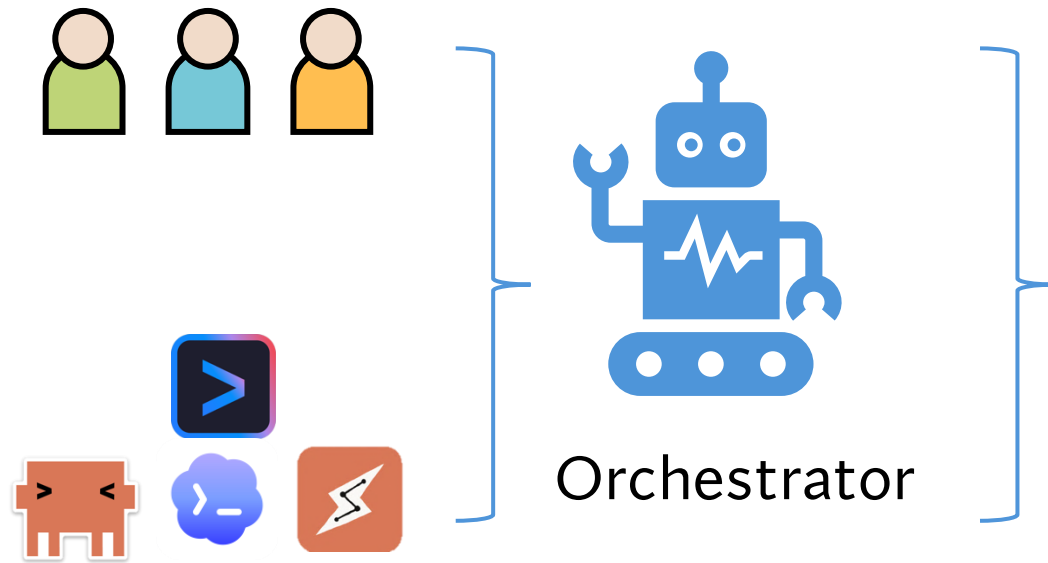


Worker profiling: estimate whether a task is feasible to the worker

Workplan generation: generate a workflow plan for the human-agent team

#	Assignee	Goal	Deliverables
1	User	Gather missing details	preferred format
2	BrowserUse	Collect freely available resources	notes.md
3	Code Agent	Create comparison table and day-by-day schedule	comparison.md schedule.md
4	Code Agent	Generate exercise scripts and example data	exercise1.sh, ... sample_data/
5	User	Review the schedule	revised schedule.md
6	Code Agent	Present the completed package	final_package.zip

Orchestrating Human-Agent Team



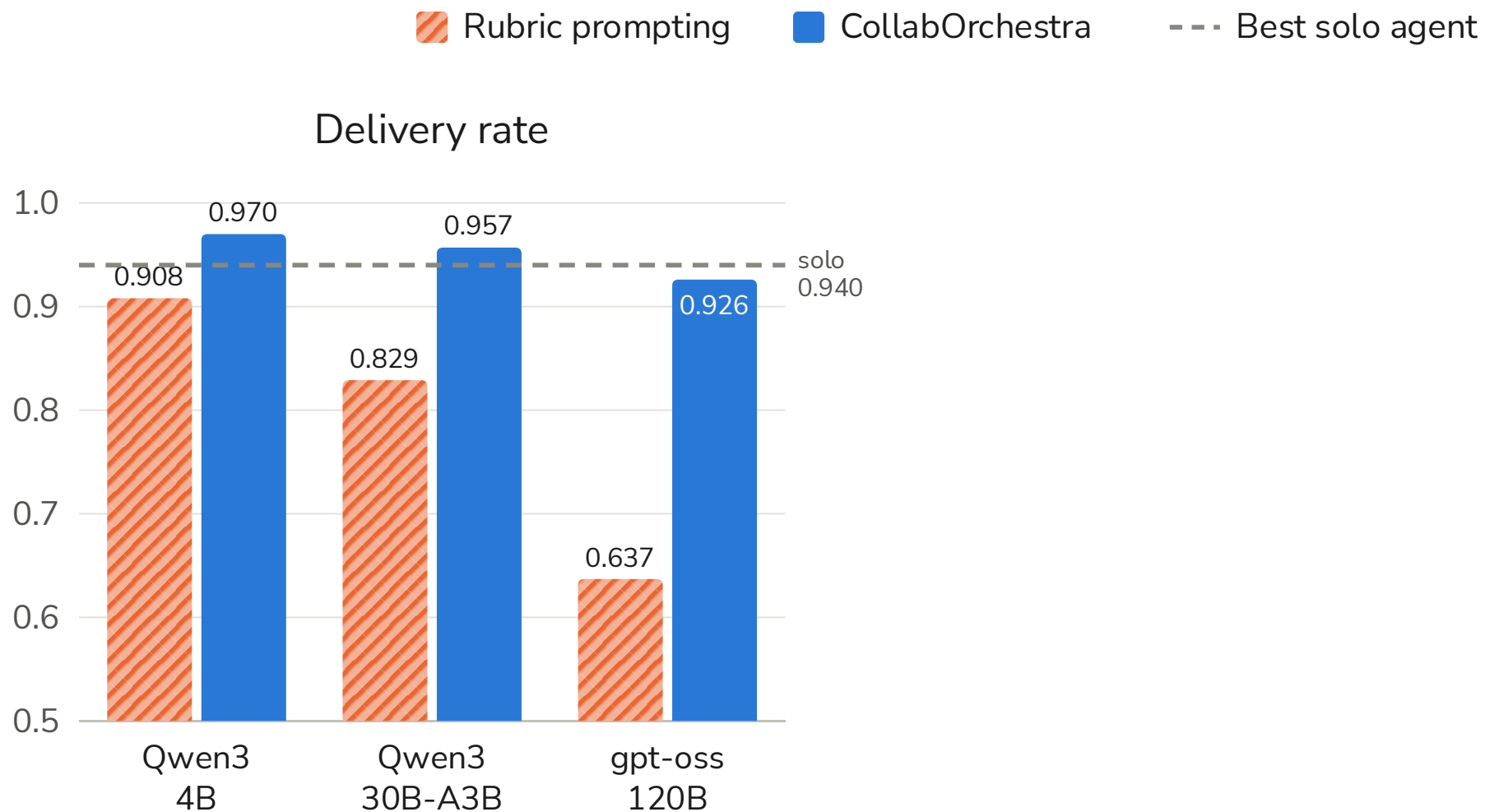
Worker profiling: estimate whether a task is feasible to the worker

Workplan generation: generate a workflow plan for the human-agent team

Rollout reward and update via GRPO:

$$R = R_{plan-quality} + R_{plan-feasibility}$$

Orchestrated Teams Beat the Best Solo Agent



CollabOrchestra In Action



Task Workflow Plan

Test

[1/5] Load and analyze the Population spreadsheet to understand data structure and prepare for calculations

[2/5] Calculate required sample size based on 90% confidence level and 10% tolerable error rate, and create 'Sample Size Calculation' tab

[3/5] Perform variance analysis: calculate Q2 to Q3 variance and populate column J

calculations' finished successfully



CollaborationCore

✓ Step 1 Complete: Spreadsheet analyzed successfully. Key Findings:

- Total population: 1,516 metrics
- 9 columns present (A-I): No, Division, Sub-Division, Country, Legal Entity, Metric Code, Metric Name, Q3 2024 Metric Value, Q2 2024 Metric Value
- Q2 and Q3 data columns contain numerical values with no missing data
- Columns J (variance) and K (sample markers) will be created in subsequent steps

↓ Scroll to bottom

Type a message...

Thread

GeminiCLIAgent

Start to execute "Calculate required sample size based on 90% confidence level and 10% tolerable error rate, and create 'Sample Size Calculation' tab"

> Execute "Reading file"

> Execute "Write File"

> Execute "Run Shell Command"

> Execute "Generating response"

GeminiCLIAgent

Also sent as direct message

The workflow step 'Calculate required sample size based on 90% confidence level and 10% tolerable error rate, and create 'Sample Size Calculation' tab' finished successfully

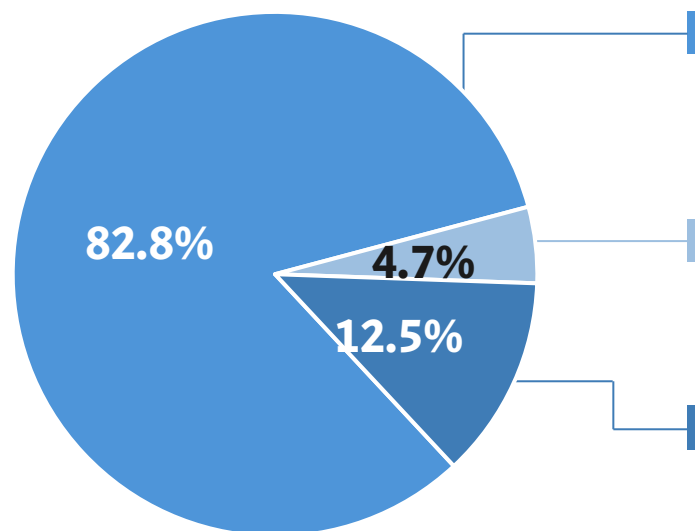
Type a message...

46

Human Steps In Orchestrated Human-agent Team

Share of human steps

Example human steps



Review/approve, e.g.,

- Review the draft and collect feedback
- Perform a final quality check

Supply input, e.g.,

- Gather required reference files and additional specs

Do the work, e.g.,

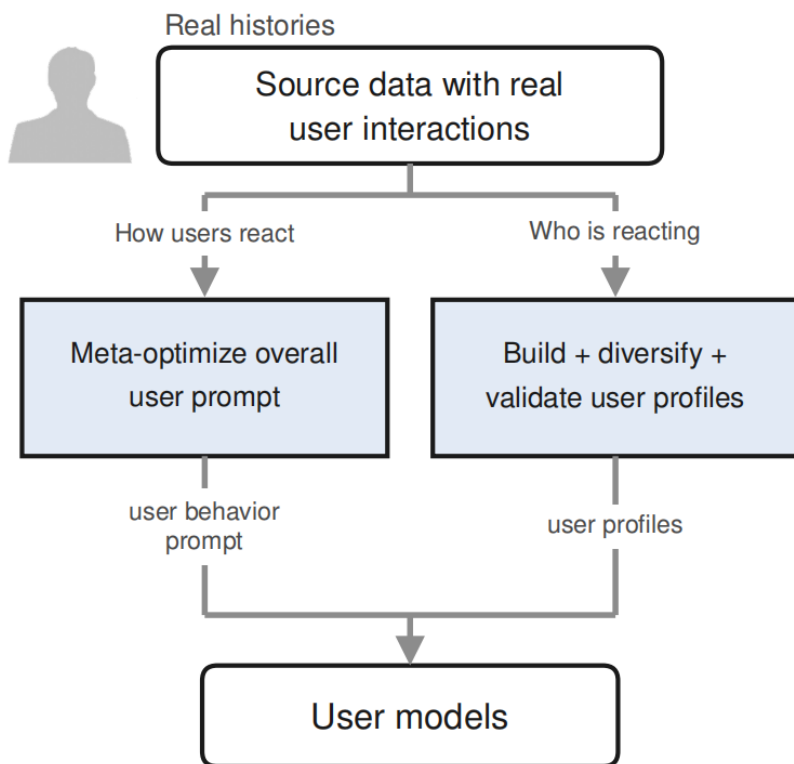
- Develop a detailed slide outline
- Design the structure of the stock and sales flow tables

User Study Post-session Ratings

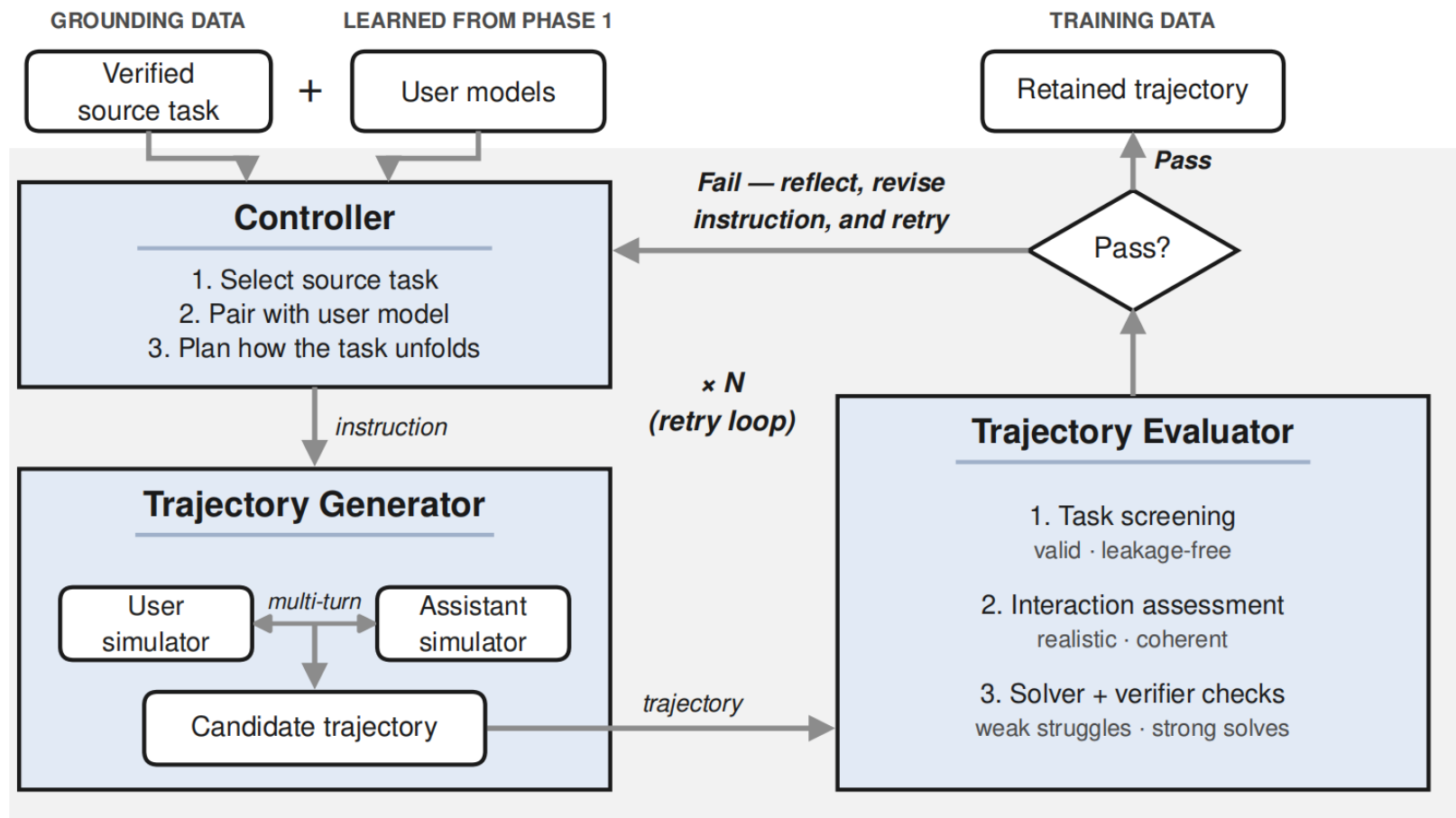
Measure	Response distribution (Scale 1-5)	Mean
<i>PLAN QUALITY</i>		
Handoff quality		3.84
Workflow coherence		3.96
Granularity fit		3.76
Involvement fit		3.96
<i>LEGIBILITY</i>		
Agent action clarity		3.75
User role clarity		3.69
<i>AGENCY</i>		
Action clarity		3.73
Autonomy		3.73
<i>OUTCOME</i>		
Goal achievement		3.75
Overall satisfaction		3.62

Training Agents To Interact With Humans

PHASE 1 Building user models from real histories



PHASE 2 Building interaction trajectories with an agentic loop



Building User Models From Real Histories

User-Simulator Prompt

You are role-playing a human developer talking to an AI coding agent inside your own repository.

General role-playing guidance

{User behavior prompt learned in Phase 1}

User profile

{Profile sampled from the population constructed in Phase 1}

Interaction history

{User–assistant trajectory so far}

Privileged context

{Task goal and private reference information}

Additional role-playing guidance (optional)

{Revised instruction from the controller, based on the trajectory evaluator's feedback on prior attempts}

Evaluation on SWE-Together and Algorithmic Task

Model	LiveCodeBench-Pro			Multi-turn LiveCodeBench-Pro			Avg	Δ vs Base
	Easy	Medium	Overall	Easy	Medium	Overall		
Qwen3.5-4B	<u>61.6</u>	7.6	36.8	40.0	3.5	23.0	29.9	–
+ RL on Source Data	61.4	<u>8.8</u>	<u>37.1</u>	<u>59.7</u>	<u>11.8</u>	<u>36.7</u>	<u>36.9</u>	+7.0
+ AUTOINTERACT	68.5 _{+6.9}	13.5 _{+5.9}	42.8 _{+6.0}	70.9 _{+30.9}	13.1 _{+9.6}	43.2 _{+20.2}	43.0	+13.1
Qwen3.5-27B	85.7	37.5	<u>60.9</u>	59.1	<u>14.0</u>	37.2	49.1	–
+ RL on Source Data	88.2	34.7	61.2	<u>61.5</u>	12.7	<u>38.0</u>	<u>49.6</u>	+0.5
+ AUTOINTERACT	<u>85.9</u> _{+0.2}	<u>37.1</u>	60.8	76.5 _{+17.4}	24.9 _{+10.9}	50.6 _{+13.4}	55.7	+6.6

Model	Pass@1 $\uparrow$	SSR $\uparrow$	Pass ² $\uparrow$	Mean Judge $\uparrow$	U-Corr $\downarrow$	Tok./Task	Min./Task
Qwen3.5-4B	6.0	1.8	1.8	0.197	3.78	23.9k	8.0
+ RL on Source Data	<u>17.0</u>	<u>11.0</u>	<u>10.1</u>	<u>0.425</u>	6.73	73.7k	39.7
+ AUTOINTERACT	20.6 _{+14.6}	16.5 _{+14.7}	11.0 _{+9.2}	0.462 _{+0.265}	6.58	69.2k	25.1
Qwen3.5-27B	<u>34.9</u>	<u>26.6</u>	<u>22.9</u>	<u>0.588</u>	2.91	25.0k	32.2
+ RL on Source Data	27.1	22.0	17.4	0.545	2.40	19.2k	52.1
+ AUTOINTERACT	38.5 _{+3.6}	33.9 _{+7.3}	26.6 _{+3.7}	0.622 _{+0.034}	3.44	29.7k	19.5

Outline

- **What “collaboration” means** (10 mins)
- **Human–AI collaboration at scale** (10 mins)
- **How humans and agents do work differently** (10 mins)
- **Different methods to enable human-agent collaboration** (15 mins)
- **Centaur evaluations** (10 mins)

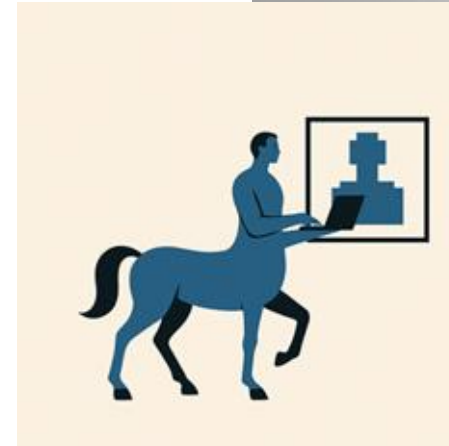
AI shouldn't be an imitation game: Centaur Evaluation

Centaur Evaluations:

a method where humans and AI collaborate to solve tasks.

Why shift:

- Focuses on human-AI collaboration, not replacement.
- Aligns ML development with human augmentation goals.

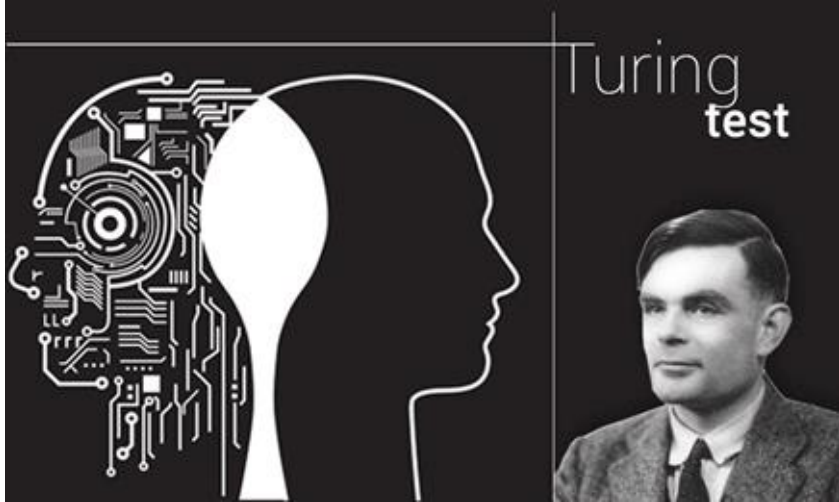


Centaur in battle, on an [Attic](#) terracotta [lekythos](#), c. 575–550 BC

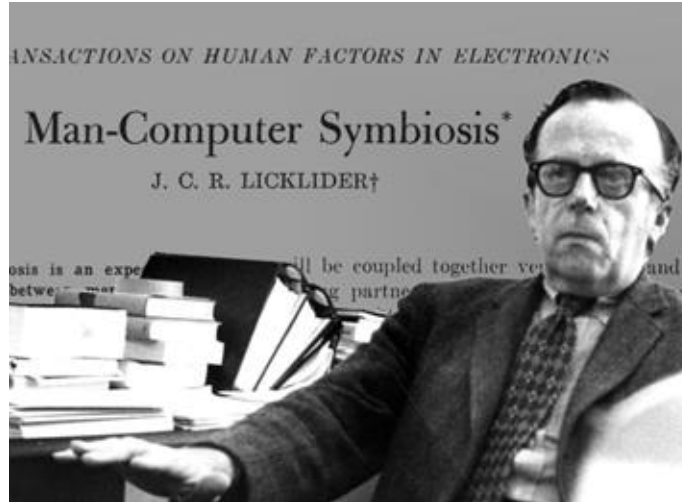
“Moving away from evaluating AI with exams”

“Moving towards evaluation that resemble how humans use AI”

Why Are There Few Centaur Evaluations



Turing and human imitation



Man-computer symbiosis
(Licklider, 1960)



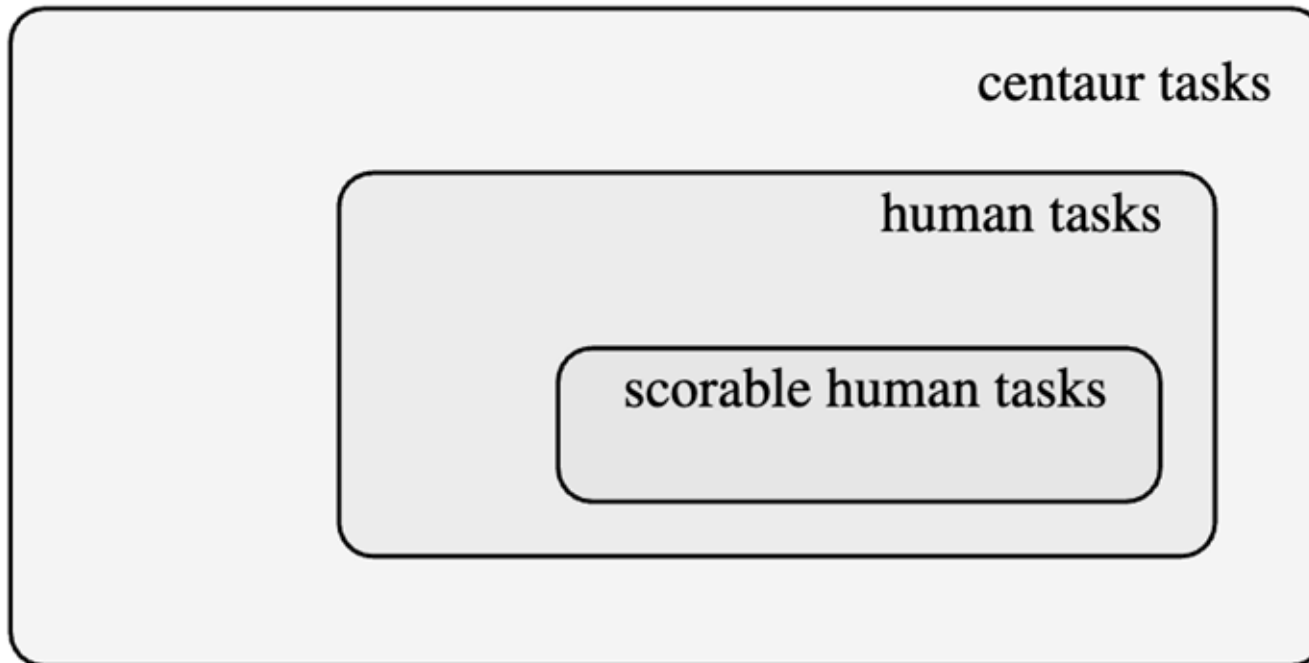
DARPA Robotics Challenge (DRC)

Why centaur eval hasn't taken off in AI?

“Combination of the existence of many imitative evaluations, Turing’s lasting impact, or the cost of benchmarks”

Why There Should Be More Centaur Evaluations

Centaur evaluations can be “harder”



Variation of Brynjolfsson (2022)



Call center use case
vs. synthetic datasets we have today

Why There Should Be More Centaur Evaluations

Centaur evaluations can be “harder”

Simplify the evaluation of human-centered desiderate

- Explainability
- Learning by interacting
- Learning to defer

Centaur evaluations positively impact society

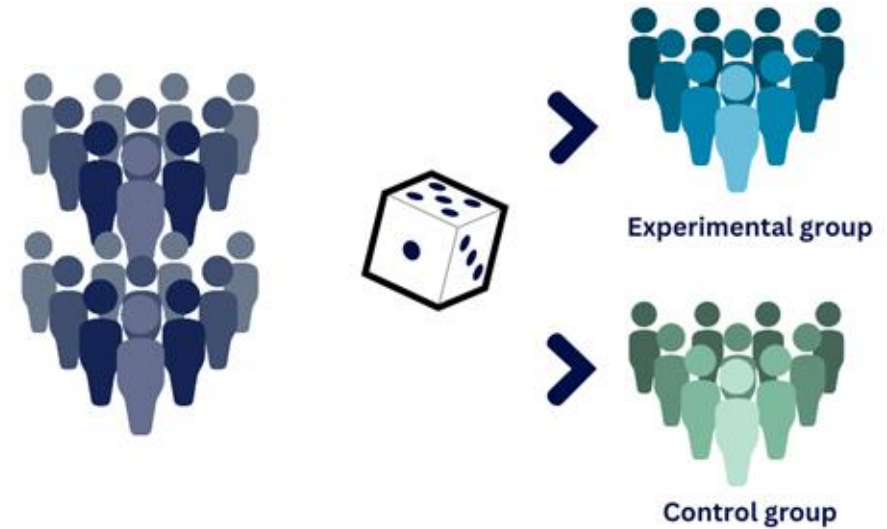
- Keep humans productive is important to avoid inequality

Running Centaur Evaluations

- With Crowd Workers
- Randomized controlled trials
- Data science competitions



human-language model interaction (Lee et al., 2024)



kaggle



Alternative Views:

“Human augmentation is an illusion”

- Biases of humans lead to worse performance (Ludwig & Mullainathan, 2021; Kleinberg et al., 2018)
- Judges perform worse than decisions made by AI alone (Angelova et al., 2024)
- Radiology screening algorithms outperform radiologist (Yu et al., 2024)



Alternative Views:

“Centaur evaluation is noisy and irreproducible”

- Fundamentally based on lab studies suffering from challenges:
 - Brittleness – sensitive to small changes in setup
 - Dependence on experimental details (Tversky & Kahneman, 1974; 1981)
- Data often noisy and inconsistent
- Require high sample complexity, even for modest model comparisons

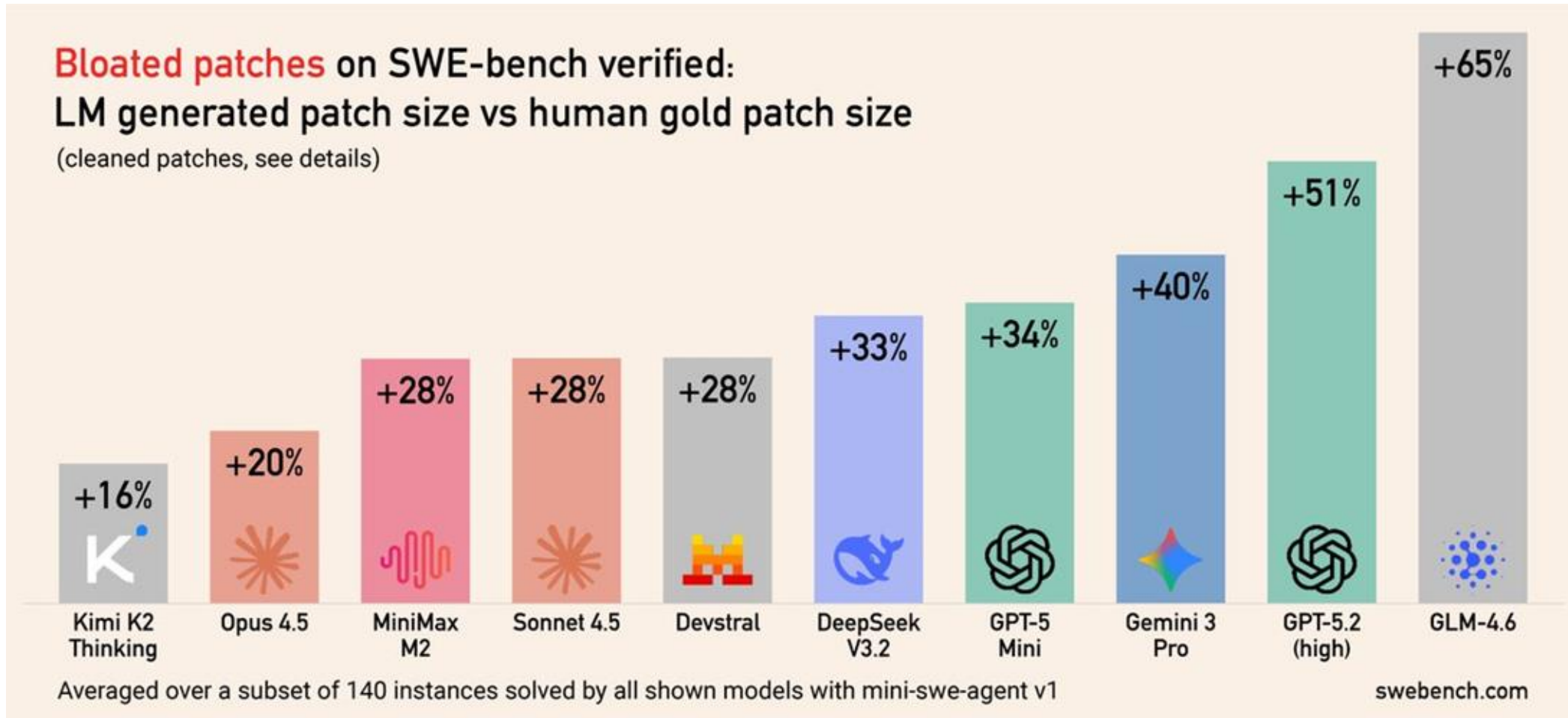


Alternative Views:

“Opportunity Costs”

- Resource-intensive — require significant time and effort
- Divert resources from rapid model iteration
- Slows down development cycles
- Rapid iteration is critical to ML progress
- **Unclear trade-off:** Do the benefits justify the cost?

Agents are Not Optimized for Human-Oriented Utility



Example: Shift caused by AI → Extra effort

Users **spend time** reviewing confusing or unreliable AI output.

Prompting, adapting output, and lost task sequence increase **cognitive load**.

Poorly timed or long suggestions **break user focus and flow**.

Easy tasks get easier, but hard tasks become harder or slower to complete – **more bugs, harder to adapt, more overwhelmed**

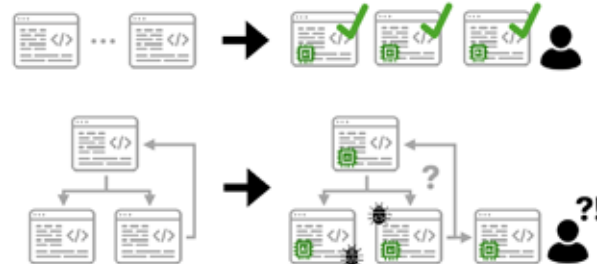
(a) The production-to-evaluation shift



(c) Task interruptions



(d) Task-complexity polarization



(b) Unhelpful workflow restructuring

Prompting



Output adaptation



Loss of task sequence



Loss of feedback



Open Challenges in Human-Agent Systems

-  **Human flexibility and variability:** feedback varies by role, timing and style; humans are rarely benchmarked
-  **Mostly agent-centered work:** systems optimize agent output, not the human's workload
-  **Inadequate evaluation:** static benchmarks miss the interaction itself
-  **Unresolved safety vulnerabilities:** humans in the loop add new attack & error
-  **Applications and beyond:** few deployed systems outside software and chat

Summary

- What “collaboration” means
- Human–AI collaboration at scale
- How humans and agents do work differently
- Different methods to enable human-agent collaboration
- Centaur evaluations