CollabSkill: Evaluating Human-Agent Collaboration On Real-World Tasks

Yijia Shao[1] **Zora Z. Wang**[2] **Neel Ahuja**[1] **Yicheng Wang**[3] **Bowen Liu**[4] **Diyi Yang**[1]

[1]Stanford University [2]Carnegie Mellon University

[3]University of Rochester [4]Individual Researcher

{shaoyj, diyiy}@cs.stanford.edu

Website: https://cogym.saltlab.stanford.edu

Abstract

AI agents are reshaping the workspace, leading to drastic change of how humans work. Despite the considerable potential of human-agent collaboration both in preserving human agency and generating economic value, this paradigm remains largely absent from occupational task evaluation, hindered by the difficulty of gathering real human data and accounting for inter-human variability. We introduce **CollabSkill**, a framework for evaluating human-agent collaboration on real-world occupational tasks. CollabSkill pairs real human workers with AI agents on tasks matched to their occupational background, collecting data that capture the complexity of economically valuable tasks and the usage patterns of real workers. To account for inter-human variability, CollabSkill employs a Bayesian skill rating system to disentangle and quantify the skill contributions of both humans and AI agents. Drawing on over 1,500 prompts from 386 working sessions contributed by 93 human workers, our analysis yields insights on two fronts: on the agent side, rankings on CollabSkill diverge from those of existing fully autonomous benchmarks where Codex leads, with Claude Code ranking first; on the human side, CollabSkill reveals that practical experience emerges as the primary driver of collaboration skill, with hands-on collaboration meaningfully shifting workers' AI literacy. Together, we hope CollabSkill enables the community to invest in systematic evaluation of human-agent collaboration and spurs development efforts aimed at building AI agents that genuinely augment human workers.

1 Introduction

AI agents are bringing unprecedented opportunities but also massive disruptions to the workspace, raising pressing concerns around job loss, inequality, and reduced human agency (Handa et al., 2025; Demirci et al., 2025; Hazra et al., 2025; Hoffmann et al., 2025). Human-agent collaboration represents a promising yet underexplored paradigm for the future of work with AI (Acemoglu et al., 2026). While recent benchmarks have focused on evaluating fully autonomous agents on occupational tasks (Patwardhan et al., 2025; Mazeika et al., 2025; Vidgen et al., 2025; 2026), effective human augmentation cannot be assumed to follow directly from strong autonomous performance as collaborative outcomes are shaped by many factors, including human AI literacy, the collaborative behaviors of models, and interaction design (Shao et al., 2024; Zou et al., 2025; Mozannar et al., 2025; Haupt & Brynjolfsson, 2025).

Benchmarking agents in collaborative settings, however, is fundamentally more complex than evaluating autonomous agents. A dominant paradigm is to use LLMs as user simulators that generate user turns to drive the interaction (Yao et al., 2024; Zhou et al., 2024;

Vijayvargiya et al., 2026; Zhou et al., 2025; Qian et al., 2025). While simulated users provide a useful proxy for scalable testing, research has shown that LLM simulators tend to be excessively cooperative, stylistically uniform, and devoid of realistic frustration (Zhou et al., 2026; Mehri et al., 2026; Suh et al., 2026). Moreover, simulated users offer limited insight into how real humans engage with AI agents in their actual work or how they perceive them (Hu et al., 2025). For studies that do engage real humans, they reveal more nuanced interaction patterns, but rely on simple tasks or focus narrowly on web navigation, failing to capture how professionals work day to day (Shao et al., 2024; Shen et al., 2025; Huq et al., 2025; Drouin et al., 2024). As AI agents are increasingly deployed in economically valuable settings, this gap becomes consequential (Meimandi et al., 2025). To this end, we introduce **CollabSkill, a framework for evaluating human-agent collaboration on realistic, economically valuable tasks with real human workers**. CollabSkill provides **infrastructure for collecting human-agent collaboration data at scale** (§3): a common task schema that links prompts, reference files, and deliverables to O*NET occupational categories (National Center for O*NET Development) supports heterogeneous tasks across sectors and allows workers to be matched to tasks aligned with their occupational background to mitigate variability in task familiarity; and a reference-free automated grader comprising rubric generation and multi-agent scoring handles open-ended deliverables. From the study period, we collected 386 sessions comprising over 1,500 user prompts across 10 O*NET sectors using this infrastructure.

Beyond data collection, a second challenge is that raw collaboration outcomes conflate agent and human performance. Workers vary substantially in AI literacy and agent familiarity, thus treating human participants as a factor to average over produces unreliable agent rankings. Rather than averaging across sessions from different humans, CollabSkill employs **a Bayesian system for disentangling and quantifying skill contributions from both human and AI agents** (§4). Inspired by multiplayer game rating (Herbrich et al., 2007; Minka et al., 2018) that jointly estimates human and agent collaboration skill from team outcomes, the Bayesian skill rating system explicitly models inter-human variability in AI literacy as a latent human component rather than a source of noise. With the collected data points, it updates the posterior for both AI agents and human workers. The estimated human CollabSkill values also help surface insights into AI literacy across occupational sectors, complemented by pre- and post-task surveys capturing how workers perceive and experience AI agents in their professional workflows (§5).

Our analysis yields insights on two fronts. **On the agent side, CollabSkill produces rankings that diverge from fully autonomous evaluations.** Among terminal-based agents which share the ReAct-based agent scaffolding and differ primarily in their underlying model, Claude Code ranks first (CollabSkill=74.8), a rank reversal relative to autonomous evaluations, where Codex consistently leads Claude Code across our own evaluation in the solo agent setting and public leaderboards like SWE-bench Verified, GDPVal, and SciCode (Patwardhan et al., 2025; Jimenez et al., 2024; Tian et al., 2024). Moreover, Claude Cowork ranks above Claude Code despite sharing the same underlying LLM, isolating interface design as a desiderata of collaboration agents across the broader spectrum of work beyond software engineering. Together, these results show that autonomous evaluation alone is insufficient for capturing human-agent collaboration. **On the human side, CollabSkill reveals that practical experience drives human CollabSkill score, and that hands-on collaboration shifts workers' AI literacy.** Practical LLM experience correlates with collaboration skill (Spearman $\rho = 0.297$, $p = 0.041$), whereas self-reported attitudes largely do not. Hands-on collaboration also shifts workers' perceived AI capability toward greater autonomy ($\Delta = -0.50$, $p < 0.001$), yet preferred autonomy for meaningful work remains unchanged, suggesting that experience updates beliefs about what agents *can* do without changing what workers *want* them to do. Together, these findings call for agent development to focus on the unique demands of human-agent collaboration and suggest CollabSkill can serve as a framework for investigating both AI agents' collaboration capability and workers' AI literacy.

To sum up, we make the following contributions:

- We introduce CollabSkill, a framework for evaluating human-agent collaboration on real-world tasks. CollabSkill employs a Bayesian skill rating system that explicitly

models human collaboration skill to account for inter-human variability, yielding more reliable agent rankings and deeper insights into human worker behavior.
- Across 386 sessions in 10 O*NET sectors, CollabSkill yields agent rankings that diverge from fully autonomous benchmarks and enables comparison across agents with different LLM backbone, agent harness, and interface designs.
- By involving real human workers, CollabSkill contributes insights into AI literacy, revealing how practical experience, attitudes, and collaboration skill interrelate.

2 Related Work

Human-Agent Collaboration Effective human-agent collaboration has been shown to improve reliability (Dong et al., 2025; Mozannar et al., 2025), enable more complex task completion (Huq et al., 2025), and enhance outcome quality (Shao et al., 2024; Shen et al., 2025; Luo et al., 2025) in comparison to full automation. These benefits stem from two underlying assumptions: human agency that humans have a desire for control and implicit requirements over task outcomes (Shao et al., 2025); and complementary synergy, where humans and agents each bring distinct strengths (Wang et al., 2025).

Despite this potential, systematic evaluation of human-agent collaboration remains lacking. On one hand, the HCI community has produced a proliferation of collaboration systems (Feng et al., 2024; Pu et al., 2025; Xu et al., 2025; Yao et al., 2025), yet these focus on system design rather than rigorous benchmarking. On the other hand, the dominant evaluation paradigm in the AI community uses LLM-simulated users (Yao et al., 2024; Zhou et al., 2024; Vijayvargiya et al., 2026; Zhou et al., 2025; Qian et al., 2025), which enable scalable testing but tend to be excessively cooperative and offer limited insight into how real humans engage with agents (Zhou et al., 2026; Mehri et al., 2026; Suh et al., 2026). Moreover, these studies do not account for humans' latent skill when evaluating agents in the human-agent collaboration setting, which confounds the results as prior work reveals that human reliance level and individual success score can significantly affect human-AI collaboration outcome (Guo et al., 2024; Arnaiz-Rodriguez et al., 2025; Davidson et al., 2025).

Agent Ranking Leaderboards and rankings are a driving force for AI advancement, typically following one of two approaches: absolute score-based ranking and comparison-based ranking. Absolute score-based ranking is simple to aggregate and widely adopted in agent leaderboards such as SWE-Bench (Jimenez et al., 2024) and OSWorld (Xie et al., 2024). Comparison-based ranking is preferred when absolute scores are hard to obtain or tasks involve subjective preference (Ammar & Shah, 2011). For example, GDPVal ranks models by average win-rate against expert-produced outcomes (Patwardhan et al., 2025), and Elo-style systems are widely used in both chatbot (Bai et al., 2022; Chiang et al., 2024) and agent leaderboards (Mazeika et al., 2025) when no specific anchor exists. CollabSkill differs from standard leaderboards in that observed outcomes reflect the joint contribution of the agent and the human rather than the agent performance alone. This introduces additional sources of variation including differences in human skill and non-stationarity as users learn or fatigue over time. We therefore adopt a Bayesian skill rating system inspired by multiplayer game platforms like XBox Live (Graepel et al., 2007; Minka et al., 2018) that jointly estimates the skill of both humans and agents from team outcomes.

The Future of Work with AI Agents A broad body of work in digital economics has examined the implications of the recent surge of LLMs and AI agents (Demirci et al., 2025; Eloundou et al., 2024; Handa et al., 2025; Hoffmann et al., 2025), finding significant productivity opportunities alongside substantial economic and societal risks, including job loss, inequality, and reduced human agency (Hazra et al., 2025; Brynjolfsson et al., 2025). Notably, AI augmentation of human workers already accounts for the majority of real-world AI use. Handa et al. (2025) find that augmentation comprises 57% of interactions versus 43% for full automation in Claude usage. Yet this paradigm remains underexploited relative to its transformative potential (Acemoglu et al., 2026). Prior work has approached the augmentation question through analyzing real user data (Handa et al., 2025) and through large-scale worker surveys (Shao et al., 2025). CollabSkill complements these efforts with a performance-based perspective.

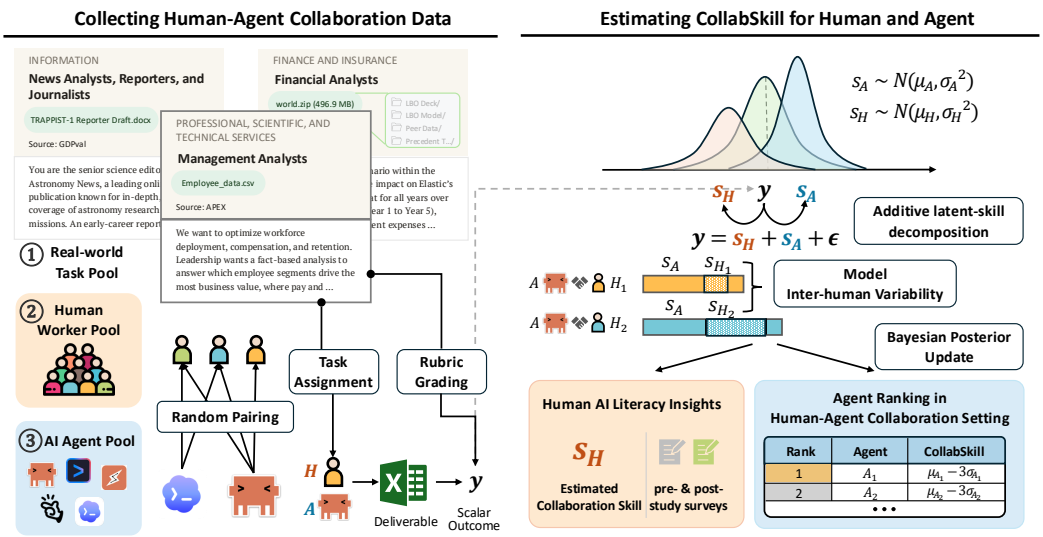

Figure 1: **Overview of CollabSkill.** CollabSkill provides infrastructure for collecting human-agent collaboration data at scale (Left) and a Bayesian skill rating framework to disentangle human and AI contributions (Right). Together, these yield agent rankings in human-agent collaboration setting and insights into human workers' AI literacy.

3 Human-Agent Collaboration Data Collection

In this section, we discuss our infrastructure design for collecting human-agent collaboration data on occupational tasks at scale and present data collection statistics.

3.1 Overview

A datapoint in CollabSkill is one completed evaluation episode in which a human works with an AI agent on a task instance under a fixed scoring function.

- **Task (t):** CollabSkill targets realistic occupational tasks that professionals carry out in their jobs—assignments that require reference files, specialized software, and produce concrete deliverables (Figure 1 ①). We propose `TaskInstance`, a schema with a unique UUID comprising a natural-language prompt, hidden evaluator information (*e.g.*, gold deliverables, domain knowledge), reference files, required software, expected deliverables. Each `TaskInstance` is tagged with O*NET occupational metadata, enabling matching users with tasks by their occupational background.
- **Human (H):** CollabSkill requires users to provide their occupational background, which is used to match them to appropriate tasks (Figure 1 ②).
- **Agent (A):** CollabSkill supports any agent with an installation and usage guide. Our experiment includes (Figure 1 ③): (i) three terminal-based agents—Codex, Gemini CLI, and Claude Code—built on GPT, Gemini, and Claude backbones respectively; (ii) Claude Cowork, a desktop agent sharing the same implementation as Claude Code but with a graphical interface designed for non-developers; and (iii) Manus, a general-purpose browser-accessible agent.[1] This selection enables controlled comparison across LLM backbones, agent harnesses, and interface designs.
- **Scalar Outcome (y):** Once assigned a task and paired with an agent, the user is instructed to set up the agent on their local machine, complete the task, and submit the delivered outcome on CollabSkill platform. CollabSkill then employs a reference-free automated grader (see §3.2) to assign a 0-100 scale score based on the task and submitted deliverables.

[1]The specific model versions used are `gpt-5.2-codex-medium`, `gemini-3-auto`, `claude-sonnet-4.6` (for both Claude Code and Claude Cowork), and `manus-1.6-lite`.

3.2 Grading Task Outcome

Occupational task deliverables are open-ended by nature, such as Excel workbooks, PDFs, slide decks, ZIP archives, and audio files. To obtain a scalar outcome y for each episode, CollabSkill employs a two-stage reference-free automated grading pipeline comprising rubric generation followed by multi-agent scoring.

Rubric Generation Given a task t (comprising the task prompt, reference files, and other available information), we employ an agent[2] to produce a JSON rubric $\mathcal{R}(t)$ that partitions evaluation criteria into weighted top-level categories including *Correctness & Accuracy*, *Deliverable Completeness*, *Task Requirements*, and *Technical Quality* (see Figure 7). Each rubric item specifies not only the criterion and its point value but also the verification procedure (*e.g.*, "Test all links for accessibility and verify they match described services").

Multi-agent Scoring Given rubric $\mathcal{R}(t)$ and the submitted deliverables d, the grader invokes N independent agents $\{f_i\}_{i=1}^{N}$, each receiving $t, \mathcal{R}(t)$, and d, and returning a per-criterion score breakdown summing to a total $\hat{y}_i \in [0, 100]$.[3] The final score y is the average of $\{\hat{y}_i\}_{i=1}^{N}$. This multi-agent design mitigates self-preference bias in LLM-as-a-judge evaluation, whereby a model acting as both generator and evaluator systematically inflates scores for its own outputs (Wataoka et al., 2024).

We validate the automated grading pipeline against official hand-authored GDPVal rubrics which were released after the original GDPval paper (Patwardhan et al., 2025). Manual rubric audit across 27 tasks across all 9 sectors from GDPval shows that our auto-generated rubrics achieve 82.1% recall and 92.2% precision against official criteria, with all generated items judged as important (see Appendix I for details).

3.3 Data Collection Results

Task Sourcing We unify three public datasets: (i) **GDPVal** (Patwardhan et al., 2025), 220 economically valuable tasks across 44 occupations; (ii) **APEX** (Vidgen et al., 2025), professional tasks across investment banking, management consulting, law, and primary care; and (iii) **APEX-Agents** (Vidgen et al., 2026), tasks across investment banking, law, and management consulting requiring cross-application execution with extensive reference files. Together, the sourced tasks cover 10 out of 20 O*NET work sectors (see Appendix C.1 for task distribution details).

Human Worker Recruitment We recruit U.S.-based participants via Upwork, compensating $20 per completed task with a suggested five-task target though they may withdraw early at their discretion. Tasks are randomly matched to participants based on their occupational background. To incentivize participants to complete these tasks carefully, those ranked 1st, 2nd, and 3rd by CollabSkill score will be awarded bonuses of $100, $50, and $50, respectively. Before starting each session, participants are randomly paired with an AI agent, presented with the installation guide for their assigned agent, and instructed to watch a tutorial video if they have not used it before. They then proceed to the task on the platform, where the task prompt and any reference files are provided. Upon completion, participants submit their final deliverables along with an interaction log (screenshots or terminal content or shareable links for Manus). Participants completing all five tasks are invited to fill out a post-study survey. The study protocol is IRB-approved. Interface details are in Appendix C.2.

Table 1: **Data Statistics.**

# Participants	93
# Data points	386
# Distinct Tasks Executed	165
Avg. Years of Experience	9.6
0-2 years	15
3-5 years	32
6-10 years	17
11-20 years	17
20+ years	10
Not Reported	2

Data Statistics During the study period, 93 human workers participate in CollabSkill, contributing 386 sessions across 10 O*NET sectors. 76 participants complete both pre/post-study surveys. Participants have an average of 9.6 years of professional experience in their selected occupation; detailed statistics are provided in Table 1.

[2]We use Gemini CLI in headless mode for rubric generation.

[3]In our implementation, $N = 2$ and we employ Gemini CLI and Claude Code as the judges.

4 From Teamwork Outcomes to CollabSkill Score

The scalar score in each collected data point reflects the joint contribution of both the human and the agent. Simply averaging scores per agent cannot produce a trustworthy ranking due to inter-human variability as people could differ substantially in AI literacy. In this section, we introduce the Bayesian skill rating system employed by CollabSkill to disentangle and estimate the collaboration skill from both sides, which draws inspiration from ranking methods in multiplayer video games (Herbrich et al., 2007; Minka et al., 2018).

Specifically, for each teamwork outcome observation (A, H, y), prior work typically estimates agent quality as the sample mean across sessions, $\hat{q}_A = \frac{1}{|\mathcal{D}_A|} \sum_{(A,H,y)\in\mathcal{D}_A} y$, where $\mathcal{D}_A$ denotes the set of sessions involving agent A, implicitly assuming that human variability averages out in expectation. Rather than making this assumption, we propose to design latent skill values s_A and s_H to the agent and human respectively, and model the scalar score y as an additive decomposition of agent skill, human skill, and observation noise.

$$y = s_A + s_H + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \beta^2). \tag{1}$$

This formulation is the key modeling choice that separates CollabSkill from prior metrics such as task completion rate, average session score, and win rate: rather than treating human variability as noise, Equation (1) explicitly allocates a latent skill component to each human participant. For example, when agent A collaborates with both H_1 and H_2, the model estimates s_{H_1} and s_{H_2} separately from their respective sessions; the agent's skill s_A is then inferred by explaining away the portion of each outcome attributable to the human. Intuitively, if H_1 consistently scores well across multiple agents while H_2 does not, the model attributes that difference to human skill in collaborating with AI agents rather than to whichever agent H_1 happened to work with.

Each entity is initialized with a Gaussian prior. For an agent A and a human user H,

$$s_A \sim \mathcal{N}(\mu_{A,0}, \sigma^2_{A,0}), \qquad s_H \sim \mathcal{N}(\mu_{H,0}, \sigma^2_{H,0}). \tag{2}$$

When a new observation (A, H, y) arrives, consider the two-dimensional latent state $\theta = [s_A, s_H]^\top$ with prior $\theta \sim \mathcal{N}(m, \Sigma)$. Writing the design vector as $x = [1, 1]^\top$, the observation takes the form $y = x^\top\theta + \epsilon$. Since this is a linear-Gaussian model, the posterior $N(m', \Sigma')$ is available in closed form via the Kalman filter measurement update (Kalman, 1960) (see Appendix F.1 for derivation):

$$r = y - x^\top m, \qquad K = \frac{\Sigma x}{x^\top \Sigma x + \beta^2}, \tag{3}$$

$$m' = m + Kr, \qquad \Sigma' = \Sigma - Kx^\top\Sigma, \tag{4}$$

where r is the prediction error (how much the outcome surprised the model) and K is the Kalman gain (how much weight to place on the new observation relative to the prior). When an outcome exceeds expectations, both μ_A and μ_H increase, with larger adjustments assigned to the entity with greater uncertainty σ^2. As observations accumulate, σ^2 shrinks and estimates stabilize.

Crucially, the model does not require a large number of sessions per agent to produce reliable rankings. What matters is the connectivity of the human-agent interaction graph, *i.e.*, every agent is linked to every other agent through at least one shared human participant, and vice versa. When the same human H appears in sessions with both agent A_1 and agent A_2, their shared outcomes allow the model to attribute score differences to the agents rather than to H alone, propagating information across the graph. In CollabSkill, where a small fixed set of agents is evaluated and humans are randomly matched to available agents, such overlap arises naturally and guarantees sufficient connectivity for reliable estimation.

In practice, we initialize all entities with $\mu_0 = 0$ and $\sigma_0 = 1$, and estimate the posterior efficiently using the information form of the Gaussian (see Appendix F.2). CollabSkill defines **CollabSkill score** $\triangleq \mu_i - 3\sigma_i$ as a conservative measure of collaboration capability

(approximately the 1% lower quantile under a Gaussian) that penalizes high-uncertainty entries and prevents over-ranking agents with few observations (Herbrich et al., 2007). CollabSkill scores are meaningful when compared within the same group—a higher score indicates that the entity is likely to contribute more to human-agent collaboration outcomes. In Appendix G, we further show empirically that the CollabSkill rating system yields lower variance when comparing agents than naively averaging performance.

5 Human Factors: AI Literacy and Attitudes

Involving real humans in CollabSkill lets us examine how AI literacy shapes human-agent collaboration and how it evolves through the experience. To complement CollabSkill of the user, we further collect attitudinal readiness, measured through paired pre- and post-task surveys designed around the Technology Acceptance Model (Davis, 1989). This design lets us ask questions with direct practical stakes: (i) Which prior experience and attitudinal factors best predict collaboration skill? (ii) Does hands-on collaboration recalibrate workers' beliefs about agent capability or shift their preferred autonomy levels?

External Variables: Prior Exposure and Agent Familiarity In the demographics survey administered at the start of the study, we capture each participant's baseline familiarity with LLMs (**D1.**) and their frequency of LLM use in professional contexts (**D2.**). We further ask participants to rate their familiarity with AI agents defined as "systems that autonomously complete tasks on your behalf" (**A1.**).

Perceived Usefulness: Trust and Delegation Comfort The Technology Acceptance Model posits that readiness to adopt a technology is strongly predicted by perceived usefulness. We measure this construct through two complementary instruments: (i) *trust in agent autonomy*, participants rate how much they trust AI agents to complete tasks accurately without human oversight, capturing perceived reliability (**A2.**); and (ii) *comfort with delegation*, participants rate how comfortable they are delegating work tasks to an AI agent (**A3.**).

Perceived Capability: Autonomy Calibration We assess how participants perceive human-agent autonomy sharing. As AI agent use is not a binary decision but a spectrum, we adopt the Human Agency Scale (HAS) (Shao et al., 2025), a shared framework for quantifying automation versus augmentation, which defines five levels from full agent autonomy (*H1*), minimal human input (*H2*), equal partnership (*H3*) to human lead (*H4*) and full human involvement (*H5*) (see Table 7). Participants identify the level at which they *believe* AI agents can currently operate for their typical work tasks (**A4.**).

Attitude: Desired Autonomy and Affective Orientation Using the same HAS, participants indicate their *preferred* autonomy level (**A5.**), their concern that AI will reduce their role below what they consider meaningful (**A6.**), and their overall affective orientation toward AI agents, chosen from *Excited*, *Curious*, *Neutral*, *Skeptical*, or *Anxious* (**A7.**).

Pre/Post Paired Design All items (except demographics and LLM familiarity) are administered identically before and after participants complete a series of collaborative tasks with AI agents. This paired design allows us to examine how hands-on experience with AI agents in occupational tasks shapes participants' AI literacy and attitudes.

6 Analysis

6.1 The Agent Side in Human-Agent Collaboration

Table 2 reports agent rankings on CollabSkill. Among the five agents evaluated, Claude Cowork achieves the highest estimated CollabSkill ($\mu - 3\sigma$) as 76.738. Since our experiment includes agents with different LLM backbones, agent harness, and interface designs, we structure the comparison along two axes.

Axis 1: Model Collaboration Strength Among Terminal-based Agents Codex, Claude Code, and Gemini CLI employ a similar terminal-based interface, and are all ReAct-based agents with bash tools, so their relative ranking primarily reflects differences in the underly-

Table 2: **Agent rankings on CollabSkill.** CollabSkill estimates each agent's latent skill as a Gaussian and reports its posterior mean μ and standard deviation σ. Agents are ranked by the conservative CollabSkill $\mu - 3\sigma$, following Herbrich et al. (2007), to penalize highly uncertain estimates.

Agent	CollabSkill	μ	σ
Claude Cowork `claude-sonnet-4.6`	76.738	77.206	0.156
Claude Code `claude-sonnet-4.6`	74.778	75.248	0.157
Codex `gpt-5.2-codex-medium`	71.267	71.733	0.155
Manus `manus-1.6-lite`	71.204	71.646	0.147
Gemini CLI `gemini-3-auto`	69.555	70.009	0.151

Table 3: **CollabSkill ranking stability under bootstrap resampling ($N = 10,000$).** We report each agent's mean rank and its probability of ranking first across N resampling rounds.

Agent	Mean	P(rank=1)
Claude Cowork	1.36	0.713
Claude Code	2.08	0.233
Codex	3.58	0.036
Manus	3.64	0.014
Gemini CLI	4.34	0.004

Table 4: **CollabSkill ranking stability under leave-one-out tests.** We report mean agent rankings under two leave-one-out conditions: dropping all sessions from one human at a time, and dropping all sessions from one `TaskInstance` at a time.

Agent	Ranking	Leave-One-Human-Out Mean	Leave-One-Task-Out Mean
Claude Cowork	#1	1.00	1.00
Claude Code	#2	2.00	2.00
Codex	#3	3.41	3.33
Manus	#4	3.59	3.67
Gemini CLI	#5	5.00	5.00

ing models. Among them, Claude Code performs best in the human-agent collaboration setting ($p(s_{\text{Claude Code}} > s_{\text{Codex}}) > 0.999$)[4] —a ranking that diverges from both our autonomous evaluation and the rankings reported by SWE-bench Verified, GDPVal, and SciCode (Table 5), suggesting that model strength in autonomous settings does not straightforwardly transfer to collaboration.

Axis 2: The Effect of Human-Agent Interface Design Second, Claude Cowork and Claude Code share the same underlying LLM, so their comparison isolates the effect of interface design. Claude Cowork ranks strictly above Claude Code ($p(s_{\text{Claude Cowork}} > s_{\text{Claude Code}}) > 0.999$), demonstrating that interface-based agents are more favorable than terminal-based ones across the broader spectrum of human work beyond software engineering. Note that this comparison controls for installation friction, as we provided detailed installation guides and assisted participants who got stuck at the installation stage (see Figure 5). To further control for prior agent familiarity since interface-based agent may be easier to get started with, we recompute CollabSkill ratings restricting to participants with self-reported agent familiarity ≥ 4 on a 7-point Likert scale ($N = 81$); the comparison between Claude Cowork and Claude Code remains unchanged (Claude Cowork: CollabSkill $= 76.403, \sigma = 0.173$; Claude Code: CollabSkill $= 74.349, \sigma = 0.175$).

Validating CollabSkill Ranking To assess the stability of our Bayesian skill rating system, we perform bootstrap resampling by drawing observations with replacement and recomputing the full posterior skill estimates across 10,000 rounds. As shown in Table 3, the top-ranked agent (Claude Cowork) retained rank 1 in 71.3% of resamples (mean rank 1.36). The overall ranking order remained moderately stable (mean Kendall's $\tau = 0.64$), with most variation arising from adjacent-rank swaps between Codex and Manus.

We additionally conduct leave-one-human-out and leave-one-task-out analyses by recomputing rankings after removing each human annotator or each task in turn to check if any single human annotator or task disproportionately influences the overall ranking. Table 4 reports the mean ranking across all such recomputations. The results confirm that top two agents (Claude Cowork and Claude Code) and the bottom agent (Gemini CLI) receive perfectly stable mean rankings of 1.00, 2.00, and 5.00, respectively, under both analyses. The middle positions show minor variance: Codex and Manus swap slightly in mean rank

[4]Since the Bayesian skill rating system models the each agent's skill as a Gaussian, the difference is also Gaussian. We compute this probability p using the CDF.

Table 5: **Comparison of CollabSkill and solo agent performance.** We run terminal-based agents in headless mode with an AI researcher-engineered system prompt on 165 collected tasks and report the mean score under Solo Score. SWE-bench Verified, GDPval, SciCode results are sourced from their official leaderboards. Model versions match those used in §3.1.

Agent	CollabSkill	Solo Score	SWE-bench Verified	GDPval	SciCode
Claude Code	74.778	77.17	71.40	42.5%	46.8
Codex	71.267	78.02	72.80	49.7%	54.6
Gemini CLI	69.555	79.44	69.60	40.3%	56.1

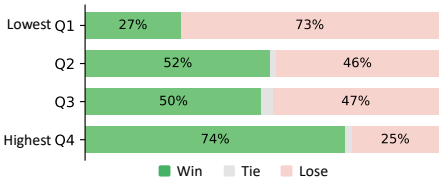

Figure 2: **Human-agent collaboration vs. solo agent win rates.** Pairwise comparison between human-agent collaboration sessions and solo agent counterparts with an AI researcher-engineered system prompt grouped by estimated human CollabSkill score.

between the two analyses (3.41 vs. 3.59 and 3.33 vs. 3.67, respectively), but their relative ordering remains consistent. Overall, these analyses indicate that the CollabSkill ranking is stable and not an artifact of any particular annotator or task.

6.2 The Human Side in Human-Agent Collaboration

We conduct head-to-head comparisons between human-agent teams and the same agent running autonomously under an AI-researcher-engineered system prompt (Figure 9). As shown in Figure 2, outcomes depend strongly on human worker CollabSkill score: top-quartile workers (Q4) achieve a 74% win rate against the autonomous baseline, while bottom-quartile workers (Q1) win only 27% of the time.

AI fluency behavior indicator predicts human-agent collaboration outcome We analyze interaction logs by coding behaviors from the Anthropic AI Fluency Index which identifies 11 directly observable indicators of human skill in using AI (Anthropic, 2026) (Appendix J). The total count of observed behaviors correlates strongly with session score (Pearson $r = 0.262$, $p < 1e-4$), with breakdowns shown in Figure 3; by contrast, user turn count shows no significant correlation ($r = 0.046$, $p = 0.371$), suggesting that *how* workers interact with agents matters more than *how much*.

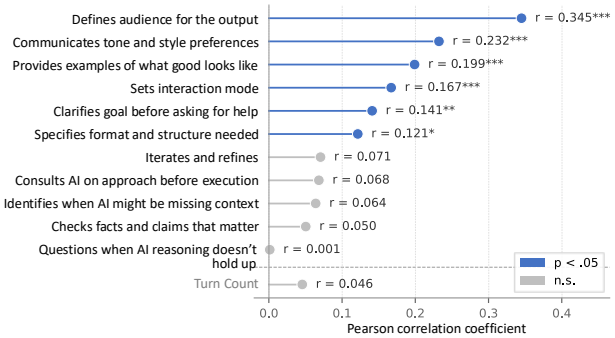

Figure 3: **Correlation between AI fluency behavior indicator with session score.**

Practical LLM experience predicts human's skill in collaborating with AI agents, but attitudes largely do not. CollabSkill yields both self-reported attitudinal data and exhibited CollabSkill estimated by the Bayesian model. As shown in Figure 4 (a), among AI literacy measure collected in the pre-study survey and the participant's CollabSkill score estimate, self-reported LLM familiarity showed a significant positive correlation with collaboration skill (Spearman $\rho = 0.297$, $p = 0.010$). Comfort with delegating work to an AI agent also correlated positively with skill ($\rho = 0.238$, $p = 0.041$). In contrast, none of the remaining attitudinal measures reached significance.

Collaboration shifts perceived AI capability toward greater autonomy, but does not change the desired HAS level. Figure 4 (c) reports paired pre- and post-study attitude changes with Wilcoxon signed-rank test. Participants reported increased trust ($\Delta = 0.58$, $p = 0.004$) and delegation comfort ($\Delta = 0.45$, $p = 0.039$). Notably, while participants revised their beliefs about AI capability toward greater autonomy (HAS belief $\Delta = -0.50$, $p < 0.001$), the preferred autonomy level did not change ($\Delta = -0.16$, $p = 0.105$), suggesting preferences about human-AI work allocation are governed by factors beyond perceived capability. When

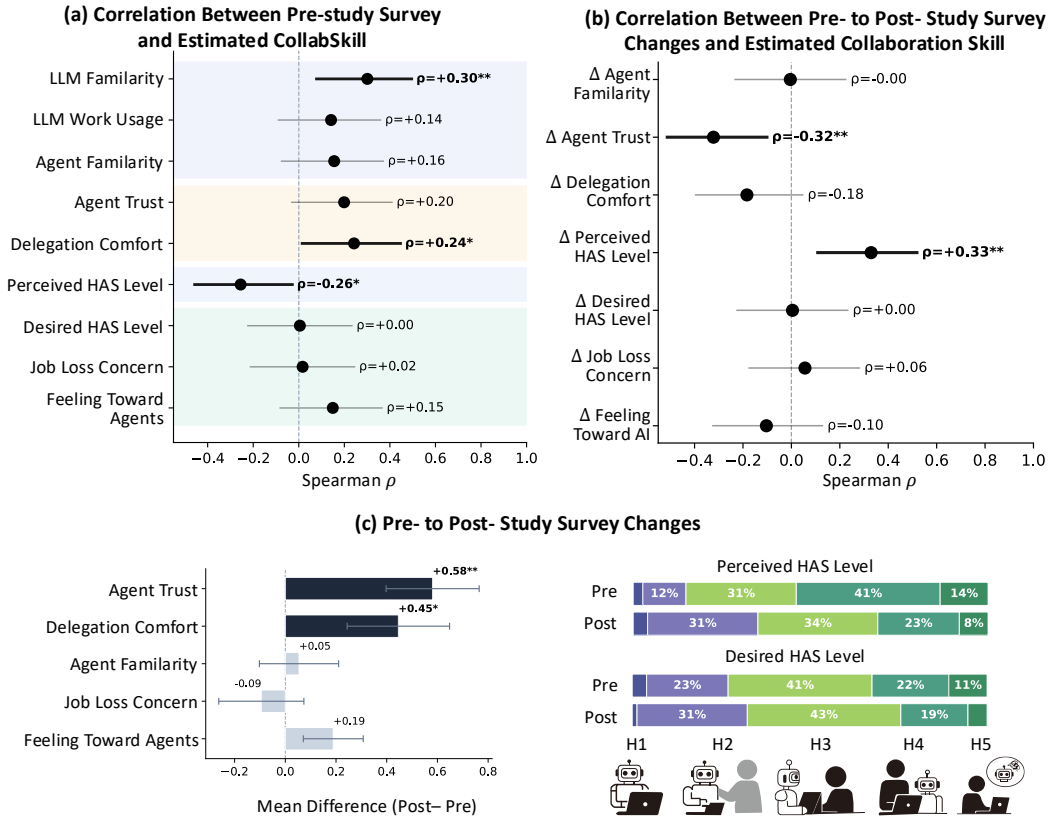

Figure 4: **Human CollabSkill and survey responses ($N = 76$).** (a) Spearman rank correlations between pre-study survey measures and estimated CollabSkill; filled circles denote point estimates and horizontal bars show 95% confidence intervals. (b) Spearman rank correlations between pre-to-post perception change and estimated CollabSkill. (c) Changes in perception after collaborating with AI agents on realistic occupational tasks; error bars denote standard errors computed via the Wilcoxon signed-rank test. In all three figures, * denotes $p < 0.05$ and ** denotes $p < 0.01$.

splitting participants at the median CollabSkill, participants in the lower half increased their agent trust by +1.03 points on average, compared to +0.16 for higher-skilled participants ($p = 0.028$ in the Mann-Whitney U test), and shifted their perceived AI autonomy level by -0.89 toward greater autonomy, versus -0.13 for higher-skilled participants ($p = 0.001$). Spearman correlations corroborate this pattern as shown in Figure 4 (b).

7 Conclusion

This work presents CollabSkill, a framework for systematically evaluating human-agent collaboration on real-world occupational tasks. CollabSkill provides infrastructure for collecting collaboration data at scale and a Bayesian skill rating framework to estimate the collaboration skill of both human workers and AI agents. Our analysis yields practical findings for multiple stakeholders: for *agent developers*, human-agent collaboration requires targeted evaluation and optimization, as autonomous capability alone does not determine collaboration quality; for *employers*, workers' familiarity with AI agents significantly predicts their collaboration skill, and effective human-agent collaboration can surpass AI-only performance; for *human workers*, hands-on experience collaborating with AI agents on occupational tasks is a meaningful pathway to improving AI literacy. We hope CollabSkill spurs efforts to build AI agents that augment human workers and to support workers who are genuinely ready to work with them.

Ethics Statement

This study was approved by our Institutional Review Board. All 93 participants were recruited through Upwork on a voluntary basis and could withdraw at any point without penalty. All collected data are reported in anonymized and aggregate form.

Our Bayesian skill rating system estimates a latent collaboration skill s_H for each participant strictly as a proxy for AI literacy within the human-agent collaboration setting we study. Individual estimates are used solely for the analyses presented in this paper. We caution against using estimated human skill scores to screen or evaluate individual workers for employment or other consequential decisions.

By focusing on human-agent collaboration, the work helps identify the strengths and weaknesses of AI agents in augmenting humans on occupational tasks, and supports the design of more effective collaborative agents. More broadly, it aligns with the goal of developing AI systems that preserve human agency and meaningful human control.

Acknowledgments

We thank Caleb Ziems, Yanzhe Zhang, Chenglei Si, and Abe Hou for their valuable feedback on the manuscript, Haowen Wang and Boom Iamphongsai for testing the CollabSkill data collection interface, and all members of the Stanford SALT lab for their suggestions throughout this project. We are also grateful to faculty administrator Maria David for her dedicated assistance that makes this large-scale human evaluation possible. We extend special thanks to the human workers who participated in our study. Many of them generously offered voluntary comments on how they think about the future of work with AI agents and their experience with using AI agents in our study, which inspired us to dedicate a separate subsection to this topic. This research is supported in part by grants from Open Philanthropy, ONR N000142412532, NSF IIS 2247357, Schmidt Sciences, and a multi-company research collaboration via Stanford HAI with SCBx, Itau, Wells Fargo, and American Express.

References

Daron Acemoglu, David Autor, and Simon Johnson. Building pro-worker artificial intelligence. Technical report, National Bureau of Economic Research, 2026.

Ammar Ammar and Devavrat Shah. Ranking: Compare, don't score. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 776–783. IEEE, 2011.

Ciprian Anghel, Alexandru A Anghel, Emilia Pecheanu, Mihai V Craciun, Adina Cocu, and Camelia Niculita. Pearl: A rubric-driven multi-metric framework for llm evaluation. *Information*, 16(11):926, 2025.

Anthropic. Anthropic education report: The AI fluency index. Technical report, Anthropic, February 2026. URL https://www.anthropic.com/research/AI-fluency-index. Accessed: March 30, 2026.

Negar Arabzadeh, Julia Kiseleva, Qingyun Wu, Chi Wang, Ahmed Awadallah, Victor Dibia, Adam Fourney, and Charles Clarke. Towards better human-agent alignment: Assessing task utility in llm-powered applications. *arXiv preprint arXiv:2402.09015*, 2024.

Adrian Arnaiz-Rodriguez, Nina Corvelo Benz, Suhas Thejaswi, Nuria Oliver, and Manuel Gomez Rodriguez. Towards human-ai complementarity in matching tasks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 36–59. Springer, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Costas Bekas, Effrosyni Kokiopoulou, and Yousef Saad. An estimator for the diagonal of a matrix. *Applied Numerical Mathematics*, 57(11–12):1214–1229, 2007. doi: 10.1016/j.apnum.2007.01.003.

Erik Brynjolfsson, Bharat Chandar, and Ruyu Chen. Canaries in the coal mine?: Six facts about the recent employment effects of artificial intelligence. 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Tim R Davidson, Adam Fourney, Saleema Amershi, Robert West, Eric Horvitz, and Ece Kamar. The collaboration gap. *arXiv preprint arXiv:2511.02687*, 2025.

Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 13(3):319–340, 1989.

Ozge Demirci, Jonas Hannane, and Xinrong Zhu. Who is ai replacing? the impact of generative ai on online freelancing platforms. *Management Science*, 71(10):8097–8108, 2025.

Tao Dong, Harini Sampath, Ja Young Lee, Sherry Y Shi, and Andrew Macvean. From correctness to collaboration: Toward a human-centered framework for evaluating ai agent behavior in software engineering. *arXiv preprint arXiv:2512.23844*, 2025.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, et al. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: Labor market impact potential of llms. *Science*, 384(6702):1306–1308, 2024.

KJ Feng, Kevin Pu, Matt Latzke, Tal August, Pao Siangliulue, Jonathan Bragg, Daniel S Weld, Amy X Zhang, and Joseph Chee Chang. Cocoa: Co-planning and co-execution with ai agents. *arXiv preprint arXiv:2412.10999*, 2024.

Thore Graepel, Tom Minka, and R TrueSkill Herbrich. A bayesian skill rating system. *Advances in Neural Information Processing Systems*, 19(569-576):7, 2007.

Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. A decision theoretic framework for measuring ai reliance. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pp. 221–236, 2024.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, et al. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*, 2025.

Andreas Haupt and Erik Brynjolfsson. Position: Ai should not be an imitation game: Centaur evaluations. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Sanchaita Hazra, Bodhisattwa Prasad Majumder, and Tuhin Chakrabarty. Position: AI safety should prioritize the future of work. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=CA9NxmmUG5.

Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, pp. 569–576, 2007.

Manuel Hoffmann, Sam Boysel, Frank Nagle, Sida Peng, and Kevin Xu. Generative ai and the nature of work. *Harvard Business School Strategy Unit Working Paper*, (25-021):25–021, 2025.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516*, 2025.

Faria Huq, Zora Zhiruo Wang, Frank F Xu, Tianyue Ou, Shuyan Zhou, Jeffrey P Bigham, and Graham Neubig. Cowpilot: a framework for autonomous and human-agent collaborative web navigation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 163–172, 2025.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics – Simulation and Computation*, 19 (2):433–450, 1990. doi: 10.1080/03610919008812866.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Elena Wu, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Hanjun Luo, Chiming Ni, Jiaheng Wen, Zhimu Huang, Yiran Wang, Bingduo Liao, Sylvia Chung, Yingbin Jin, Xinfeng Li, Wenyuan Xu, et al. Hai-eval: Measuring human-ai synergy in collaborative coding. *arXiv preprint arXiv:2512.04111*, 2025.

Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, et al. Remote labor index: measuring ai automation of remote work. *arXiv preprint arXiv:2510.26787*, 2025.

Shuhaib Mehri, Philippe Laban, Sumuk Shashidhar, Marwa Abdulhai, Sergey Levine, Michel Galley, and Dilek Hakkani-Tür. Measuring and mitigating the distributional gap between real and simulated user behaviors. *arXiv preprint arXiv:2605.07847*, 2026.

Kiana Jafari Meimandi, Gabriela Aránguiz-Dias, Grace Ra Kim, Lana Saadeddin, Allie Griffith, and Mykel J Kochenderfer. The measurement imbalance in agentic ai evaluation undermines industry productivity claims. *arXiv preprint arXiv:2506.02064*, 2025.

Tom Minka, Ryan Cleven, and Yordan Zaykov. Trueskill 2: An improved bayesian skill rating system. *Technical Report*, 2018.

Hussein Mozannar, Gagan Bansal, Cheng Tan, Adam Fourney, Victor Dibia, Jingya Chen, Jack Gerrits, Tyler Payne, Matheus Kunzler Maldaner, Madeleine Grunde-McLaughlin, et al. Magentic-ui: Towards human-in-the-loop agentic systems. *arXiv preprint arXiv:2507.22358*, 2025.

National Center for O*NET Development. O*NET OnLine help: The database. O*NET OnLine. URL https://www.onetonline.org/help/onet/database. Accessed: March 22, 2026.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, et al. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025.

Kevin Pu, Daniel Lazaro, Ian Arawjo, Haijun Xia, Ziang Xiao, Tovi Grossman, and Yan Chen. Assistance or disruption? exploring and evaluating the design and trade-offs of proactive ai programming support. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pp. 1–21, 2025.

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Zhiwei Liu, Jianguo Zhang, Haolin Chen, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, et al. Userbench: An interactive gym environment for user-centric agents. *arXiv preprint arXiv:2507.22034*, 2025.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*, 2024.

Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with ai agents: Auditing automation and augmentation potential across the us workforce. *arXiv preprint arXiv:2506.06576*, 2025.

Shannon Zejiang Shen, Valerie Chen, Ken Gu, Alexis Ross, Zixian Ma, Jillian Ross, Alex Gu, Chenglei Si, Wayne Chi, Andi Peng, et al. Completion $\neq$ collaboration: Scaling collaborative effort with agents. *arXiv preprint arXiv:2510.25744*, 2025.

Joseph Suh, Ayush Raj, Minwoo Kang, and Serina Chang. Quantifying the utility of user simulators for building collaborative llm assistants. *arXiv preprint arXiv:2605.09808*, 2026.

Minyang Tian, Luyu Gao, Shizhuo D Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. Scicode: A research coding benchmark curated by scientists. *Advances in Neural Information Processing Systems*, 37:30624–30650, 2024.

Bertie Vidgen, Abby Fennelly, Evan Pinnix, Julien Benchek, Daniyal Khan, Zach Richards, Austin Bridges, Calix Huang, Kanishka Sahu, Abhishek Kottamasu, et al. The ai productivity index (apex). *arXiv preprint arXiv:2509.25721*, 2025.

Bertie Vidgen, Austin Mann, Abby Fennelly, John Wright Stanly, Lucas Rothman, Marco Burstein, Julien Benchek, David Ostrofsky, Anirudh Ravichandran, Debnil Sur, et al. Apex-agents. *arXiv preprint arXiv:2601.14242*, 2026.

Sanidhya Vijayvargiya, Xuhui Zhou, Akhila Yerukola, Maarten Sap, and Graham Neubig. Interactive agents to overcome underspecificity in software engineering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=X2yzXtH4wp.

Zora Zhiruo Wang, Yijia Shao, Omar Shaikh, Daniel Fried, Graham Neubig, and Diyi Yang. How do ai agents do human work? comparing ai and human workflows across diverse occupations. *arXiv preprint arXiv:2510.22780*, 2025.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Yuan Xu, Shaowen Xiang, Yizhi Song, Ruoting Sun, and Xin Tong. Duetui: A bidirectional context loop for human-agent co-generation of task-oriented interfaces. *arXiv preprint arXiv:2509.13444*, 2025.

Bingsheng Yao, Jiaju Chen, Chaoran Chen, April Wang, Toby Jia-jun Li, and Dakuo Wang. Through the lens of human-human collaboration: A configurable research platform for exploring human-agent collaboration. *arXiv preprint arXiv:2509.18008*, 2025.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Xuhui Zhou, Hyunwoo Kim, Faeze Brahman, Liwei Jiang, Hao Zhu, Ximing Lu, Frank F. Xu, Bill Yuchen Lin, Yejin Choi, Niloofar Mireshghallah, Ronan Le Bras, and Maarten Sap. HAICOSYSTEM: An ecosystem for sandboxing safety risks in interactive AI agents. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=KI1WQ6rLiy.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, et al. Mind the sim2real gap in user simulation for agentic tasks. *arXiv preprint arXiv:2603.11245*, 2026.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, et al. Llm-based human-agent collaboration and interaction systems: A survey. *arXiv preprint arXiv:2505.00753*, 2025.

A Disclosure of LLM use in Research

We used LLMs in three parts of this work. First, during development of the CollabSkill frameworks and analysis code, we used LLM-based coding agents and tab completion features to help write and debug code. Second, in preparing this manuscript, we used LLMs to polish paragraphs originally drafted by the authors and to convert handwritten mathematical derivations into LaTeX format. In both cases the authors wrote the initial content and verified the final output. Third, LLMs serve as components of the CollabSkill system itself: a coding agent generates evaluation rubrics from task specifications, and two LLM judges score submitted deliverables against those rubrics, as described in §3.2.

B Limitations and Future Work

While CollabSkill offers the first systematic framework for evaluating human-agent collaboration on realistic occupational tasks, several limitations should be considered.

Evaluation Design. Involving real human workers is central to CollabSkill's design: unlike agent-only benchmarks, CollabSkill captures authentic human–agent collaboration dynamics, and cross-analyzing estimated CollabSkill with self-reported survey results yields meaningful insights into workers' AI literacy. That said, our participant pool was recruited through Upwork and targets US-based workers only, so findings on human workers' CollabSkill scores and AI literacy should be interpreted with this selection bias in mind. With 386 sessions from 93 workers spanning 10 of the 20 O*NET sectors, CollabSkill establishes rankings and surfaces broader trends than most prior work focused exclusively on software engineering; however, we cannot conduct fine-grained subgroup analyses due to limited statistical power, and our sample is not yet representative of the full occupational landscape. Owing to the scalable `TaskInstance` schema, we are expanding CollabSkill to additional sectors and have released tooling on our open-source platform to support community contributions.

Simulation-based evaluation is another appealing direction, but we find current user simulators cannot be applied in our setting. Unlike typical chatbot scenarios where user simulators model turn-based dialogue, our setting involves professional workers completing complex, real-world occupational tasks through rich, multimodal interactions spanning agent interfaces (terminal, desktop, web app) and direct task-completion actions. The scale and naturalistic complexity of our human data reflects this: 386 sessions with a median duration of 76.7 minutes, yielding over 1,500 prompts and various editing actions in the task environment. We see building better user simulators for the human-agent co-working setting as a high-value research direction.

For task scoring, we use LLM judges rather than human graders. While human evaluation is the gold standard, it is prohibitively expensive at our scale. For example, GDPval (Patwardhan et al., 2025) required substantial human grading costs and later introduced LLM-judge rubrics. A core design goal of CollabSkill is to remain accessible and reproducible for the broader research community, which necessitates automated scoring. While we've validated the automated grader (see Appendix I for details), we acknowledge that improving automated graders is an important direction for future work.

CollabSkill Formulation. We model task outcomes as an additive decomposition of agent skill, human skill, and observation noise. Our primary objective is to disentangle human and agent skill and to provide robust quantitative estimates of each. A limitation of the current formulation is that the estimated CollabSkill score does not separately identify task skill and collaboration skill. We adopt this formulation because real-world tasks are substantially more complex than toy collaboration environments, making a clean decomposition challenging. Thus, we view the current formulation as a principled starting point; extending the CollabSkill ranking system to distinguish finer-grained factors is a promising direction for future work. Analogously, the original TrueSkill (Herbrich et al., 2007) ranks holistic team-play performance, and later extensions incorporate additional factors such as player experience and individual statistics to tailor rankings to specific games (Minka et al., 2018).

C CollabSkill Data Collection Details

C.1 Task Sourcing

CollabSkill targets human-agent collaboration on real-world tasks. Our unified `TaskInstance` schema specifies both the task description (prompt) and execution environment (reference files, required software). For the experiments in this paper, tasks are sourced from three public datasets, GDPval (Patwardhan et al., 2025), APEX (Vidgen et al., 2025), and APEX-Agents (Vidgen et al., 2026), grounding task difficulty consistently across domains. We target a wide coverage of sectors and select 5 tasks per occupation from each source:

- **GDPval**: (1) Administrative Services Managers; (2) Buyers and Purchasing Agents; (3) Child, Family, and School Social Workers; (4) Compliance Officers; (5) Computer and Information Systems Managers; (6) Concierges; (7) Customer Service Representatives; (8) Editors; (9) Film and Video Editors; (10) Financial Managers; (11) Financial and Investment Analysts; (12) First-Line Supervisors of Office and Administrative Support Workers; (13) First-Line Supervisors of Police and Detectives; (14) General and Operations Managers; (15) Industrial Engineers; (16) Medical Secretaries and Administrative Assistants; (17) News Analysts, Reporters, and Journalists; (18) Nurse Practitioners; (19) Order Clerks; (20) Pharmacists; (21) Project Management Specialists; (22) Property, Real Estate, and Community Association Managers; (23) Real Estate Sales Agents; (24) Recreation Workers; (25) Sales Managers; (26) Shipping, Receiving, and Inventory Clerks; (27) Software Developers
- **APEX**: (1) Financial Analysts; (2) Healthcare Diagnosing or Treating Practitioners, All Other; (3) Management Analysts
- **APEX-Agents**: (1) Financial Analysts; (2) Lawyers; (3) Management Analysts

Together, these tasks span 10 O*NET sectors, providing a diverse testbed for evaluating general-purpose agent capabilities in real-world human-agent collaboration settings. It is also worth noticing that the unified `TaskInstance` schema provides a standardized interface for incorporating new tasks. Since the initial version of this paper in March 2026, we have further integrated datasets such as JobBench[5], OccuBench[6], WorkspaceBench[7], ClawBench[8] into our data pool.

C.2 CollabSkill Data Collection Interface

Figures 5 and 6 show the task page used in our user study. Each session consists of five tasks, all sharing the same page layout; only the task prompt, reference files, and assigned agent differ. At the top, the page displays the task number and the assigned agent. A setup panel asks whether the participant has used the agent before, with three options: "Never used," "Used a few times," and "Use regularly." A collapsible guide walks through installation. The participant checks a box to confirm that the agent is ready. If setup fails, the participant can contact the research team and skip the task after uploading screenshots of the issue.

Below the setup panel, the full task prompt is displayed. Reference files are listed by name and can be downloaded individually or as a ZIP archive. Expected deliverable filenames are listed so that participants know what to submit. At the bottom of the page, participants upload their completed files and provide a collaboration log showing how they worked with the agent. For Manus, participants paste a shared session link; for terminal-based agents like Claude Code, participants upload log files or screenshots of the interaction.

[5] https://huggingface.co/datasets/JobBench/job-bench
[6] https://huggingface.co/datasets/gregH/OccuBench
[7] https://huggingface.co/datasets/Workspace-Bench/Workspace-Bench
[8] https://huggingface.co/datasets/NAIL-Group/ClawBench

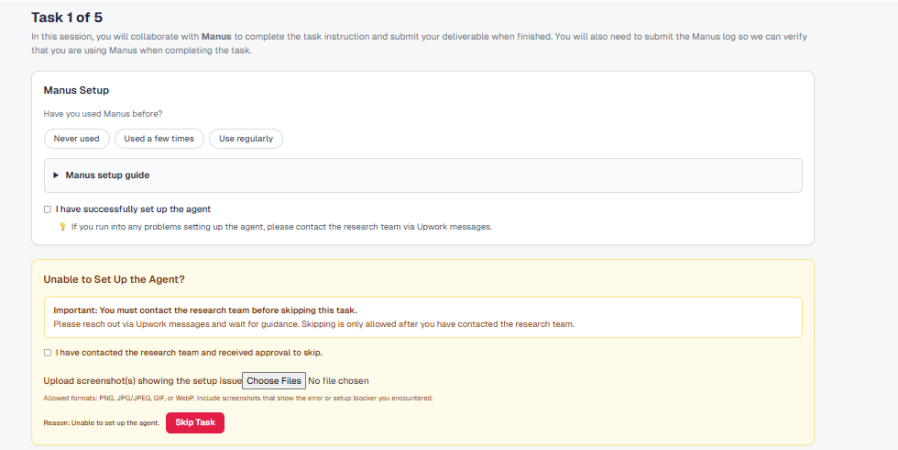

Figure 5: **Agent setup panel on the task page.** The participant reports prior experience with the assigned agent, follows the setup guide, and confirms readiness.

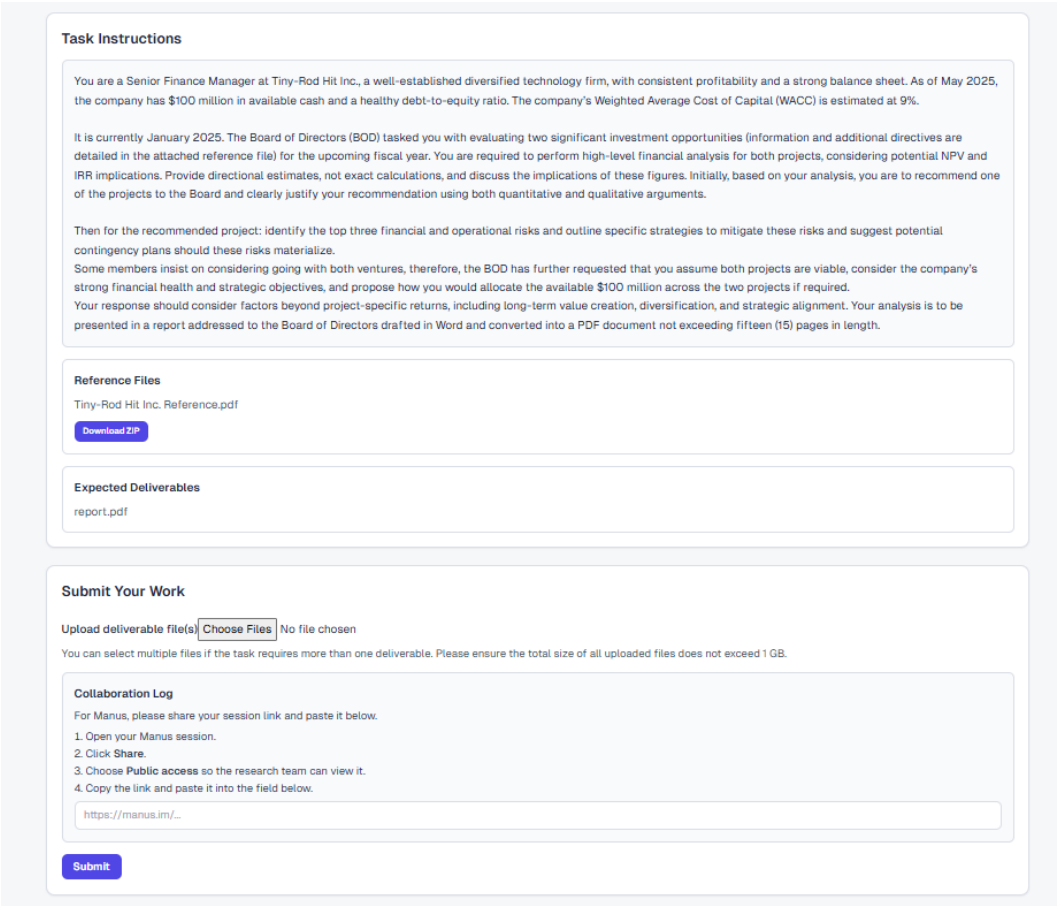

Figure 6: **Task instruction and submission area.** The participant reads the prompt, downloads reference files, uploads deliverables, and provides a collaboration log.

D Prompts Used in the Automated Grader

The automated grader in CollabSkill works in two stages. First, a coding agent reads the task prompt and reference files to produce a structured rubric. Second, two LLM judges from different model families independently score the submitted deliverables against that rubric, and the final score is their average. §3.2 describes the pipeline in detail; below we include the full prompts used in each stage.

Autogenerate Rubrics Prompt

You are analyzing a task folder to generate a grading rubric for AI agent evaluation.

Task ID: `{task_id}`
Task directory: `{task_dir}`
Reference files available (DO NOT read or reference anything in deliverables/): `{file_list}`

INSTRUCTIONS

1. Read `prompt.txt` to understand the full task description and requirements.
2. Read all other reference files listed above. These are the source data and templates the agent uses.
3. Extract specific quantitative values, formulas, thresholds, counts, and required outputs from these files.
4. Generate a `rubric.json` that:
   - has `task_id = "{task_id}"`
   - has `total_points = 100`
   - is split into exactly 4 scoring categories:
     - `"Correctness & Accuracy"` (40 pts)
     - `"Deliverable Completeness"` (20 pts)
     - `"Task Requirements"` (25 pts)
     - `"Technical Quality"` (15 pts)
   - has each category as an object with keys: `name`, `max_points`, `criteria` (list of criterion objects)
   - has each criterion object include: `description`, `points`, `expected_value`, `evaluation_method`
   - has `expected_answers` contain `specific_values` with exact expected outputs extracted from the reference files
   - has `critical_requirements` list non-negotiable pass/fail requirements
   - has `reference_files_analyzed` list every file you read
   - has `evaluation_notes` give overall guidance for the human evaluator
5. Be as specific as possible. If the task requires calculating a sample size, cost, threshold, count, or any numeric value, state the exact expected answer in `expected_value`.
6. Output only the raw JSON object, with no markdown, no explanation, and no code fences.

FINAL INSTRUCTION
Now read the files and generate the `rubric.json` content.

Figure 7: **Prompt for rubric generation.** The coding agent reads the task prompt and reference files, extracts expected values from the source materials, and outputs a JSON rubric with four weighted scoring categories. The agent does not see any submitted deliverables at this stage.

Autograder Prompt

You are an expert evaluator scoring an AI agent's task completion against a rubric.

TASK DESCRIPTION: `{task_prompt}`
Deliverable files submitted by the agent: `{file_list}`
RUBRIC: `{rubric_str}`

INSTRUCTIONS

1. Read every file in the deliverables directory listed above.
2. Use the task description above to understand what the agent was supposed to accomplish.
3. For each rubric category and criterion, carefully evaluate whether the agent's output meets the expected value. Be specific about what you found versus what was expected.
4. Assign a numeric score for each criterion, from 0 up to that criterion's full point value. Partial credit is allowed where the work is partially correct.
5. Sum the criterion scores within each category.
6. Sum all category scores for the total score.
7. Be quantitatively precise. If a rubric criterion asks for a specific numeric value (for example, sample size = 385 or cost = $1,234.56), check for that exact value and deduct points proportionally if it is incorrect.

OUTPUT FORMAT
Return only a valid JSON object with this exact structure, with no markdown and no explanation:

```
{
"score": <total numeric score, float>,
"max_score": 100,
"percentage": <score / 100 * 100, float>,
"summary": "<2-4 sentence overall assessment of the agent's
submission>",
"category_scores": {
  "<Category Name>": {
    "earned": <float>,
    "max": <int>,
    "notes": "<specific observations about what was correct/incorrect
in this category>"
  }
},
"deliverables_found": [<list of filenames you were able to read>],
"criterion_scores": [
  {
    "category": "<category name>",
    "description": "<criterion description>",
    "earned": <float>,
    "max": <int>,
    "reason": "<brief explanation>"
  }
]
}
```

Figure 8: **Prompt for autograder scoring.** Each LLM judge receives the task description, the generated rubric, and the submitted deliverables. The judge scores each rubric criterion with partial credit and returns a structured JSON score breakdown.

E Prompt Used For Autonomous Agents

Prompt for Testing Autonomous Agents

You are an AI assistant helping to complete a real-world task. Follow these guidelines:

- Never use special characters like - (U+2011), use - (U+002D) instead
- Avoid emojis and nonstandard characters
- For PDFs, always use LibreOffice to create them
- Use cross-platform fonts like Noto Sans/Noto Serif
- Deliverable text should be concise (4 sentences max)
- Always check formatting before submitting
- Add CONFIDENCE[XX] on the last line (0-100)

CRITICAL FILE LOCATION REQUIREMENTS:

- You are working in a task directory. Save ALL files directly here (in the current directory).
- Use ONLY relative filenames: 'output.pdf', 'report.docx', 'analysis.xlsx' - NO paths!
- NEVER create directories like 'deliverable_files' or navigate to other directories
- NEVER use paths like '../deliverable_files/' or '/deliverable_files/' or any absolute paths
- NEVER use paths like '../deliverable_files/' or '/deliverable_files/' or any absolute paths
- When using LibreOffice or any tool, specify output file as just the filename: 'output.pdf' not './deliverable_files/output.pdf'
- All files you create must appear in the same directory as prompt.txt
- Do NOT use cd, mkdir, or navigate away from the current working directory

{Task Prompt}

Figure 9: **Autonomous agent prompt.** System prompt used to evaluate terminal-based agents in headless mode, engineered by AI researchers.

F Deriving Bayesian Skill Rating System

In §4, we introduce how CollabSkill employs a Bayesian Skill Rating System to jointly estimate human and agent skill from team outcomes. Here we provide additional derivations to assist the understanding.

F.1 Deriving the Update from One Data Point

For each teamwork outcome observation (A, H, y) collected through the CollabSkill infrastructure, consider the two-dimensional latent state $\theta = [s_A, s_H]^\top$, with prior $\theta \sim \mathcal{N}(m, \Sigma)$. Under the additive latent-skill decomposition, the observed teamwork outcome $y = x^\top \theta + \epsilon$, where $x = [1, 1]^\top$ and $\epsilon \sim \mathcal{N}(0, \beta^2)$. The goal is to update the estimation of m and Σ from this data point.

Likelihood. Conditioned on θ, the observation follows

$$p(y \mid \theta) = \mathcal{N}(y \mid x^\top \theta, \beta^2). \tag{5}$$

Equivalently, in exponential form,

$$p(y \mid \theta) \propto \exp\left(-\frac{1}{2\beta^2}(y - x^\top \theta)^2\right). \tag{6}$$

Posterior via Bayes' rule. The posterior is given by

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta), \tag{7}$$

where the prior is

$$p(\theta) \propto \exp\left(-\frac{1}{2}(\theta - m)^\top \Sigma^{-1}(\theta - m)\right). \tag{8}$$

Expanding both terms, we obtain

$$\log p(\theta \mid y) = -\frac{1}{2}(\theta - m)^\top \Sigma^{-1}(\theta - m) - \frac{1}{2\beta^2}(y - x^\top \theta)^2 + \text{const} \tag{9}$$

$$= -\frac{1}{2}\theta^\top \left(\Sigma^{-1} + \frac{1}{\beta^2} x x^\top\right)\theta + \left(\Sigma^{-1} m + \frac{y}{\beta^2} x\right)^\top \theta + \text{const}. \tag{10}$$

This corresponds to a Gaussian posterior $p(\theta \mid y) = \mathcal{N}(m', \Sigma')$, with natural parameters

$$\Sigma'^{-1} = \Sigma^{-1} + \frac{1}{\beta^2} x x^\top, \qquad \Sigma'^{-1} m' = \Sigma^{-1} m + \frac{y}{\beta^2} x. \tag{11}$$

Recovering the covariance. Applying the Woodbury matrix identity,

$$\Sigma' = \Sigma - \Sigma x (x^\top \Sigma x + \beta^2)^{-1} x^\top \Sigma. \tag{12}$$

Recovering the mean. Multiplying both sides of $\Sigma'^{-1} m'$ by Σ' and simplifying yields

$$m' = m + K(y - x^\top m), \quad \text{where} \quad K = \frac{\Sigma x}{x^\top \Sigma x + \beta^2}. \tag{13}$$

Final update equations. Defining the residual $r = y - x^\top m$, the posterior update takes the Kalman form (Kalman, 1960):

$$r = y - x^\top m, \qquad K = \frac{\Sigma x}{x^\top \Sigma x + \beta^2},$$

$$m' = m + Kr, \qquad \Sigma' = \Sigma - K x^\top \Sigma.$$

F.2 Scalable Inference Algorithm

While the update in Equation (4) is intuitive, it is not the most efficient representation at scale because explicitly maintaining a dense covariance matrix becomes prohibitive as the number of entities grows. For scalable inference, we instead maintain the global posterior in information form. Let $\theta \in \mathbb{R}^n$ stack all latent skills across agents and humans. For each teamwork outcome observation (A, H, y), define a sparse design vector $x \in \mathbb{R}^n$ such that $x_A = 1$, $x_H = 1$, and all other entries are zero. The observation model becomes

$$y = x^\top \theta + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \beta^2). \tag{14}$$

In the information form,

$$p(\theta) \propto \exp\left(-\frac{1}{2}\theta^\top \Lambda \theta + \eta^\top \theta\right), \tag{15}$$

where $\Lambda = \Sigma^{-1}, \eta = \Sigma^{-1} m$. Writing $w = 1/\beta^2$, the update in Equation (4) is equivalent to the following (see Appendix F.1 for derivation):

$$\Lambda \leftarrow \Lambda + w\, x x^\top, \qquad \eta \leftarrow \eta + w\, y\, x. \tag{16}$$

Because x has only two nonzero entries, this update is extremely sparse, touching only four entries in Λ and two entries in η. Priors in Equation (2) are injected exactly once per entity by adding the corresponding Gaussian natural parameters:

$$\Lambda_{ii} \mathrel{+}= 1/\sigma_0^2, \qquad \eta_i \mathrel{+}= \mu_0/\sigma_0^2. \tag{17}$$

Once all observations have been accumulated, the posterior mean vector, which serves as the primary estimate of latent skill, is obtained by solving the sparse linear system $\Lambda \mu = \eta$. Uncertainty for entity i is determined by the marginal posterior variance,

$$\sigma_i^2 = (\Lambda^{-1})_{ii}, \qquad \sigma_i = \sqrt{(\Lambda^{-1})_{ii}}.$$

In practice, we initialize all entities with $\mu_0 = 0$ and $\sigma_0 = 1$, and use Hutchinson's stochastic diagonal estimator (Hutchinson, 1990; Bekas et al., 2007) to estimate $(\Lambda^{-1})_{ii}$ efficiently.

G Comparing CollabSkill Rating System with Averaging Agent Performance

We compare CollabSkill ratings against a naive baseline that ranks agents by mean task score. Table 6 shows the results. While the resulting agent ranking order matches that of CollabSkill, the high variance ($\sigma > 19$ for all agents) makes it difficult to draw statistically meaningful conclusions due to high inter-human variability. This high variance is the core motivation for jointly modeling human skill. Furthermore, the naive baseline offers no insight into the human side, providing no information on human workers' skill in collaborating with AI agents.

Agent	Rating by Taking the Average		Rating by Using CollabSkill Framework	
	Score	σ	Score	σ
Claude Cowork	83.839	19.456	76.738	0.156
Claude Code	81.176	21.994	74.778	0.157
Codex	77.480	25.286	71.267	0.155
Manus	76.417	24.887	71.204	0.147
Gemini CLI	74.822	22.528	69.555	0.151

Table 6: Comparing CollabSkill ratings against ratings by taking the average of agent performance.

H Survey Details

CollabSkill evaluates human-agent collaboration with real human workers. To better understand the human factors, we collect data on participants' AI literacy and attitude through pre/post-task surveys. Our survey is structured as follows:

Demographics and Prior Exposure

D1. LLM Familiarity: How familiar are you with large language model products (e.g., ChatGPT, Claude, Google Gemini, etc.)?

- I use them regularly.
- I have some experience using them.
- I have heard of them but don't know much about their functionalities.
- No, I've never heard of them.

D2. Professional Use: Have you used large language models in your work-related activities?

- Yes, I use them every day in my work.
- Yes, I use them every week in my work.
- Yes, I have used them occasionally for specific tasks.
- No, I have not used them for any work-related activities.
- No, I've never heard of them.

D3. Experience: How many years of experience do you have in [participant's occupation]? *(Numeric input)*

Attitudinal Readiness To measure shifts in participants' perceptions of AI agents, we administered the following items both before and after the collaborative tasks. Post-task items were prefixed with: *"After collaborating with agents on these tasks..."*

A1. Agent Familiarity: How familiar are you with AI agents (systems designed to autonomously complete tasks on your behalf)? *(7-point Likert: 1 = Never heard of them, 7 = Use them regularly)*

A2. Trust in Autonomy: How much do you trust AI agents to execute tasks accurately without human oversight? *(7-point Likert: 1 = No trust at all, 7 = Complete trust)*

A3. Comfort with Delegation: How comfortable are you delegating high-stakes professional tasks to an AI agent? *(7-point Likert: 1 = Very uncomfortable, 7 = Very comfortable)*

Level	Type	Definition
H1	Autonomous	AI agent handles the task entirely on its own.
H2	Minimal Input	AI agent requires minimal human input for optimal performance.
H3	Partnership	AI agent and human form an equal partnership, outperforming either alone.
H4	Human-Led	AI agent requires human input to successfully complete the task.
H5	Full Involvement	AI agent cannot function without continuous human involvement.

Table 7: The Human Agency Scale (HAS) (Shao et al., 2025) reference provided to participants in pre/post-task surveys.

A4. Perceived Capability: Reflecting on your typical work tasks, where on the HAS scale do you believe AI agents can generally operate for your job? *(Choice: H1–H5, see Table 7 for reference)*

A5. Desired Autonomy: Where on the HAS scale do you prefer AI agents to operate within your professional workflow? *(Choice: H1–H5, see Table 7 for reference)*

A6. Role Meaningfulness: To what extent are you concerned that AI integration will reduce your role below what you consider meaningful work? *(7-point Likert: 1 = Not at all, 7 = Extremely worried)*

A7. Affective Orientation: Which of the following best captures your current sentiment toward AI agents? *(Options: Excited, Curious, Neutral, Skeptical, Anxious)*

I Validating Automated Grader Quality

I.1 Rubric Categories

To evaluate open-ended deliverables, we partition evaluation criteria into four weighted top-level categories. This multi-dimensional approach is grounded in established frameworks for automated and agentic evaluation. *Correctness & Accuracy* assesses factual integrity and technical soundness, treating factual accuracy as the primary axis of evaluation (Liang et al., 2022; Anghel et al., 2025). *Deliverable Completeness* measures whether the agent successfully produced all requested components, which serves as a fundamental dimension of task utility (Arabzadeh et al., 2024). *Task Requirements* evaluates adherence to explicit instructions and negative constraints, ensuring the agent operates within defined operational guardrails (Liang et al., 2022). Finally, *Technical Quality* captures organization, clarity, and professionalism, drawing on established metrics for explanatory usefulness and clarity (Anghel et al., 2025; Arabzadeh et al., 2024). We weight correctness highest, as the economic utility of occupational agents is fundamentally predicated on factual accuracy over stylistic alignment. Figure 10 shows an example rubric generated from the automated grader.

I.2 Manual Rubric Audit

For realistic occupational tasks, manual grading is hard to scale: GDPval reports that grading each comparison took human experts an hour on average (Patwardhan et al., 2025), and Remote Labor Index reports human evaluators spending around 30 minutes per task for pairwise preference judgments (Mazeika et al., 2025). While a more scalable approach is to use LLM-as-a-judge provided with fine-grained rubrics, many task sources do not come with rubrics or only cover correctness, and manual rubric construction remains expensive, motivating our reference-free approach where we employ an agent to generate rubrics automatically. We validate rubric quality by sampling 3 tasks per sector from GDPVal (27 tasks total) and manually comparing our auto-generated rubrics against the official hand-authored GDPVal rubrics[9], which were released after the original GDPVal paper (Patwardhan et al., 2025). The generated rubrics achieve a recall of 82.1% against the official criteria (998 of 1,216 official rubric items are covered) and a precision of 92.2% (297 of 322 of our items are grounded in official criteria), with authors judging all 322 of the generated rubric items as important for evaluating task quality.

[9] https://huggingface.co/datasets/openai/gdpval

I.3 Correlation with Grading with Hand-authored Rubrics

We further validate our grading pipeline by comparing the final scores yielded from our automated grader and the scores yielded from GDPval official rubrics on 135 randomly sampled (task, deliverable) pairs, with authors manually reviewing each score to confirm evaluation outcomes The two score series yielded a Pearson correlation of $r = 0.72$ (excluding three outlier pairs based on author review; $r = 0.66$ over all 135 pairs), indicating strong agreement and supporting the use of auto-generated rubrics in the automated grading pipeline.

Example Rubric - Istanbul Trip Itinerary

Task ID: `0e4fe8cd-16d0-4f41-8247-6385b4762582`

Total Points: 100

SCORING CATEGORIES

1. Correctness & Accuracy *40 pts*

- **Itinerary dates are correct (June 1–4)** 5 pts
  Expected: Tabs labeled June 1st, June 2nd, June 3rd, June 4th.
  Method: Manual inspection of tab names and dates.
- **All specified times are correct** 10 pts
  Expected:
  - Day 1: 8am Pickup, 9am Wheels Up
  - Day 2: 3:00am Wheels Down, 4:30am Drop-off, 9:00am Breakfast, 11:00am Tour, 2:00pm Lunch, 5:00pm Return/Meeting, 7:00pm Tuxedo Drop-off, 8:30pm Pickup, 9:00pm Dinner, 10:30pm Departure
  - Day 3: 9:00am Breakfast, 11:00am Hair/Makeup, 1:00pm Tuxedo Fitting, 2:00pm Lunch, 4:00pm SUV to Wedding, 10:30pm Leave Wedding
  - Day 4: 8:00am Pickup, 9:00am Wheels Up, 11:00am Landing, 11:00am Pickup, 11:30am Drop-off

  Method: Compare times listed against expected values.
- **All locations are correctly named** 10 pts
  Expected: Airports: JVY, ISL. Hotel: Four Seasons Bosphorus. Restaurants: Yali, Hidden Garden, Garden 1897. Venue: Adile Sultan Palace.
  Method: Verify all location names against the prompt.
- **All names are correctly spelled** 5 pts
  Expected: Oguz (Tour Guide), Maserto (Tuxedo), Samira Lowell (Hair/Makeup), Bespoke Tuxedo.
  Method: Check spelling of all specified names.
- **Flight calculations and time zone details are accurate** 10 pts
  Expected: Outbound: 10-hour flight, 8-hour time difference forward. Return: 10-hour flight, gaining 8 hours.
  Method: Confirm flight duration and time zone details in travel entries.

2. Deliverable Completeness *20 pts*

- **Workbook contains exactly four tabs** 10 pts
  Expected: 4 tabs, one per day.
  Method: Count the number of tabs in the submitted Excel file.
- **All required hyperlinks are included** 10 pts
  Expected: Links for: Oguz (tour guide), Hidden Garden, Four Seasons Bosphorus GM, Maserto, Garden 1897, Samira Lowell, and Adile Sultan Palace.
  Method: Check for the presence of a hyperlink for each specified item.

3. Task Requirements *25 pts*

- **Itinerary delivered as a multi-tab Excel document** 10 pts
  Expected: File format `.xlsx` or equivalent.
  Method: Verify the file type of the deliverable.
- **Itinerary covers complete journey door-to-door** 5 pts
  Expected: Starts with pickup at the main house (Day 1) and ends with drop-off at the

main house (Day 4).
Method: Check the start and end points of the itinerary.

- **Separate actions listed as distinct line items** 5 pts
  Expected: E.g., "8:30 p.m. picked up from hotel lobby" and "9:00 p.m. reservation at Garden 1897" are two separate entries.
  Method: Inspect itinerary to ensure events are broken into individual line items.
- **Formatting is readable and all links are clickable** 5 pts
  Expected: Clear, professional layout; hyperlinks are active and lead to relevant webpages.
  Method: Manual review and hyperlink testing.

4. Technical Quality *15 pts*

- **File is a valid, openable Excel workbook** 5 pts
  Expected: Opens without errors in standard spreadsheet software.
  Method: Attempt to open the file.
- **Workbook is well-organized with clear columns** 5 pts
  Expected: Columns for Time, Action/Event, Details/Links clearly labeled and consistently used.
  Method: Subjective review of layout and organization.
- **No placeholder, broken, or incorrect links** 5 pts
  Expected: All hyperlinks resolve to valid, relevant websites; high-quality alternatives provided for undiscoverable entities.
  Method: Click-through on all hyperlinks to verify validity and relevance.

EXPECTED SPECIFIC VALUES

Item	Expected Value
Day 1 Start Time	8:00 AM
Day 1 Departure Airport	JVY
Day 2 Arrival Airport	ISL
Hotel Name	Four Seasons Bosphorus
Tour Guide Name	Oguz
Wedding Venue	Adile Sultan Palace
Number of Excel Tabs	4

CRITICAL REQUIREMENTS

1. The deliverable must be a multi-tab Excel file (`.xlsx` or equivalent).
2. The itinerary must span the four specified days (June 1–June 4).
3. The itinerary must include travel logistics from the principal's home and back.

EVALUATION NOTES
The core of this task is attention to detail and professional presentation. The evaluator should check every time, location, and name against the prompt. The quality of research for hyperlinks is also important; links should be for high-end, relevant services in Istanbul. If a specified entity such as *Maserto* cannot be found, the agent should have noted this and provided a comparable high-quality alternative.

Figure 10: **Example rubric for the Istanbul trip itinerary task.** The rubric evaluates correctness of times, locations, and names; completeness of tabs and hyperlinks; adherence to formatting requirements; and overall technical quality of the Excel deliverable.

J Analysis of Human-Agent Collaboration Trajectories

To characterize how participants interact with AI agents during collaboration sessions, we conducted a systematic analysis of all available interaction logs ($N = 386$). We code each session's trajectory with `gemini-3-pro` using the 11 behavioral indicators from the Anthropic AI Fluency Index, which captures effective human–AI collaboration practices such as iterating on outputs, clarifying goals, specifying formats, defining audience, and verifying AI-generated claims (see Figure 11 for the full prompt).

We examined the relationship between each AI fluency behavior and the scalar outcome of the session (mean=78.6, N=386) using point-biserial correlations. Figure 3 presents the results. Six of the eleven behaviors were significantly correlated with higher session scores, including (1) defining the audience for the output ($r = 0.345$, $p < 0.001$), where sessions exhibiting this behavior scored 16.5 points higher on average (84.7 vs. 68.1), (2) communicating tone and style preferences ($r = 0.232$, $p < 0.001$; 83.9 vs. 73.2), (3) providing examples of desired output quality ($r = 0.199$, $p < 0.001$; 82.8 vs. 73.5), (4) setting the interaction mode ($r = 0.167$, $p = 0.001$; 82.6 vs. 74.9), (5) clarifying the goal before requesting help ($r = 0.141$, $p = 0.006$; 79.8 vs. 69.6), and (6) specifying format and structure ($r = 0.121$, $p = 0.017$; 79.6 vs. 71.1). Notably, the raw number of user turns was uncorrelated with session score ($r = 0.046$, $p = 0.371$), suggesting that more interaction does not inherently lead to better outcomes and the quality of the interaction matters more.

We also compare AI fluency behavior rates across the five agents in our study. Cowork elicits the highest mean AI fluency score (6.0 out of 11 on average). Examining individual behaviors, Cowork led all agents in eliciting the human workers to set interaction mode (60.3% vs. 39–49% for other agents), identify missing context (83.6%), question reasoning (16%), consult on approach (8.2%), and check facts (28.8%).

Collaboration Trajectory Analysis Prompt

You are an expert at analyzing human-AI interaction patterns. You will be given a conversation log between a user and an AI assistant.
Your task is to analyze whether the user demonstrates the following effective AI collaboration behaviors. For each behavior, assess with a simple Yes or No based on whether there is clear evidence in the log, and provide a brief reasoning for your assessment.
—
BEHAVIORS TO ASSESS:

1. **Iterates and Refines** — Does the user follow up, push further, or build on prior responses rather than accepting the first answer?
2. **Clarifies Goal Before Asking for Help** — Does the user articulate what they're trying to achieve before making requests?
3. **Provides Examples of What Good Looks Like** — Does the user share reference points, models, or success criteria?
4. **Specifies Format and Structure Needed** — Does the user indicate how they want the output organized (length, format, sections, etc.)?
5. **Sets Interaction Mode** — Does the user define how they want the AI to behave (e.g., "be concise", "act as a critic", "ask me questions before answering")?
6. **Communication Tone and Style Preferences** — Does the user express stylistic preferences (formal/casual, direct/exploratory, etc.)?
7. **Identifies When AI Might Be Missing Context** — Does the user proactively add background, correct assumptions, or flag gaps in the AI's understanding?
8. **Defines Audience for the Output** — Does the user specify who the output is for and tailor the request accordingly?
9. **Questions When AI Reasoning Doesn't Hold Up** — Does the user push back, probe the AI's logic, or challenge weak reasoning?
10. **Consults AI on Approach Before Execution** — Does the user ask the AI to help think through strategy or method before diving in?
11. **Checks Facts and Claims That Matter** — Does the user verify outputs, ask for sources, or flag uncertainty about specific claims?

—
OUTPUT FORMAT:
Return only a valid JSON object. No explanation, no markdown, no preamble. Each behavior has two fields: "result" ("yes" or "no") and "reasoning" (one concise sentence explaining why).

```
{
  "iterates_and_refines": {"result": "", "reasoning": ""},
  "clarifies_goal_before_asking": {"result": "", "reasoning": ""},
  "provides_examples_of_good": {"result": "", "reasoning": ""},
  "specifies_format_and_structure": {"result": "", "reasoning": ""},
  "sets_interaction_mode": {"result": "", "reasoning": ""},
  "communication_tone_and_style": {"result": "", "reasoning": ""},
  "identifies_missing_context": {"result": "", "reasoning": ""},
  "defines_audience": {"result": "", "reasoning": ""},
  "questions_ai_reasoning": {"result": "", "reasoning": ""},
  "consults_ai_on_approach": {"result": "", "reasoning": ""},
  "checks_facts_and_claims": {"result": "", "reasoning": ""}
}
```

—
Here is the interaction log to analyze:
{Interaction Log}

Figure 11: **Prompt for coding AI fluency behavior indicators.** We parse the human-agent collaboration log into text format and use an LLM to code behaviors from the Anthropic AI Fluency Index which identifies 11 directly observable indicators of human skill in using AI (Anthropic, 2026).