KOLAB: BENCHMARKING HUMAN-AGENT COLLABORATION IN LONG-HORIZON WORKFLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

When humans collaborate with AI agents on long-horizon tasks, what happens to the agents' contributions? For solving long-horizon tasks, humans place greater importance on ongoing interaction with AI agents (*e.g.,* revising plans) than on an agent's ability to complete the task in a single attempt. However, most agentic benchmarks have focused more on automation. Strong performance on these benchmarks does not mean that an agent can collaborate effectively when deployed in the real world. Professional workflows offer a useful setting for examining this question, as they often involve interdependent steps. To address this gap, we introduce KOLAB, a new dataset of 100+ hours of screen recordings of 45 professionals from 36 occupations collaborating with AI agents on realistic, long-horizon problems. The dataset also contains collaboration and output quality judgments from the participating experts. We evaluate (1) how well agents collaborate with professionals and (2) whether they can serve as verifiers of human–agent collaboration. With KOLAB, we find that frontier models, both agentic and non-agentic, tend to overestimate collaboration with experts and struggle to distinguish between high- and low-quality collaboration. As frontier agents cannot yet understand long-horizon collaboration with experts, we position KOLAB as a valuable dataset for systematically studying human–agent collaboration grounded in real professional workflows.

1 INTRODUCTION

As AI agents become more capable of complex reasoning and knowledge work, recent research has examined whether they can complete such tasks autonomously in a single attempt, given a single instruction (Patwardhan et al., 2026; Vidgen et al., 2026; Zhu et al., 2026; Sun et al., 2026b; Merrill et al., 2026; Xu et al., 2026; Xie et al., 2024). However, real-world problem solving may not follow this autonomous setting: human goals can evolve, earlier decisions shape subsequent work, and errors may require intervention before they propagate. In professional work, for example, workers formulate requirements, produce intermediate artifacts, revise decisions, and move between agents and domain-specific tools. Thus, strong autonomous task performance may not translate into effective human-agent collaboration (Zou et al., 2025; Mozannar et al., 2025; Shen et al., 2026).

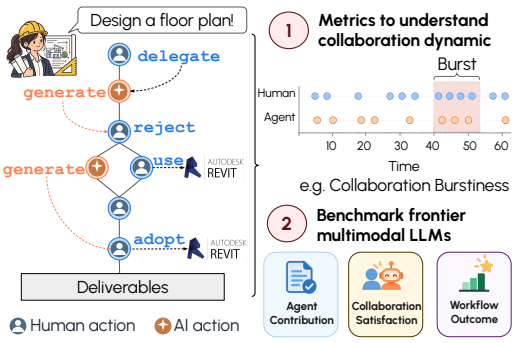


Figure 1: Overview of our study: KOLAB records how professionals (human actions) and agents (AI actions) jointly solve long-horizon problems, and benchmarks whether frontier models understand this collaboration.

Despite the need for studying effective human-agent collaboration, existing datasets provide limited grounding in real collaborative workflows. They are often not publicly released, rely on predefined tasks constructed or sourced by researchers, or focus on a single professional domain such as software development (Wang et al., 2025; Chen et al., 2026; Shao et al., 2026b). As a result, benchmarking long-horizon human-agent collaboration workflows remains underexplored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

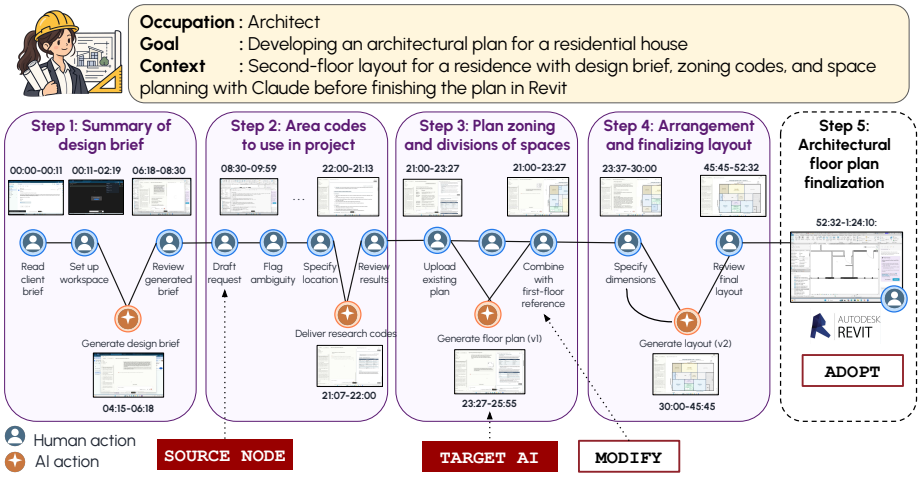


Figure 2: An example human-agent collaborative workflow for an architect. Each step is a subgraph of human and agent actions; annotations link a source node, the target AI output it triggers, and the human response to it (here, MODIFY and ADOPT).

To address this gap, we introduce KOLAB, a dataset of screen recordings capturing human-agent collaboration on long-horizon professional problems. We represent each workflow as a graph of semantically meaningful human and agent actions and their dependencies (Figure 2), allowing us to understand when and how collaboration occurs in the workflow. In this example, an architect completes a long-horizon task in five steps, and agent outputs can influence later steps. A task delegated in Step 2 (source node) produces information (target AI node) that is later modified and used in Step 3 (MODIFY). Human responses to agent outputs may also occur outside the chat interface. The architect ultimately adopts and executes the resulting plan in Revit, an action not directly observable by the agent.

In addition to characterizing existing collaboration, we investigate whether frontier multimodal large language models (MLLMs) and agents can understand and anticipate how their outputs contribute to and affect human work. Specifically, we test whether MLLMs can predict whether agent contributions will be adopted, modified, or rejected, including responses that occur outside the chat interface, such as an architect incorporating an agent output in Revit. We also evaluate whether MLLMs can predict professionals' collaboration satisfaction and outcome quality from the workflow observed so far. These capabilities will enable agents to provide more proactive and adaptive support, while anticipating when their contributions may require human intervention or introduce risks that propagate to later stages of the workflow. We observe that frontier multimodal agents still struggle to anticipate how professionals use agent outputs and to recover their judgments of collaboration quality. For example, Gemini 3.1 Pro and GPT-6 achieve only 44% and 56% F1 in forecasting adoption behavior, while their predictions of collaboration satisfaction correlate weakly with professional ratings (Spearman's $\rho = .06$ and .33). Further analysis reveals systematic biases: models often favor particular adoption labels and overpredict high ratings, limiting their ability to distinguish successful from poor collaboration. These results highlight a challenging yet important problem toward enabling more effective human-agent collaboration. Our main contributions are as follows.

- We present KOLAB, 105 hours of screen-recorded human-agent workflows from 45 professionals solving their own real-world problems, comprising 503 steps across 90+ software applications. Each workflow includes a final deliverable, with step-level human ratings of collaboration experience and output quality.
- We analyze human-agent collaboration dynamics using three metrics: collaboration burstiness, workflow topology, and AI retention rate.
- We benchmark frontier MLLMs on their ability to understand human-agent collaborative workflows when solving professional problems.
- We uncover systematic MLLM failure modes, revealing prediction biases that vary by model and input modality.

Table 1: Comparison of KOLAB with long-horizon, multi-domain workflow benchmarks. "H-AI Collab." indicates the presence of human-agent collaborative workflows; "Software Diversity" refers to tasks spanning multiple software applications; "Modalities" lists types of publicly released data.

Dataset	Human Workflow	H-AI Collab.	Software Diversity	#Domains	Modalities
Agents' Last Exam (Sun et al., 2026b)	✗	✗	✓	55	Image, text, files
Workflow-GYM (Zhu et al., 2026)	✓	✗	✓	6	Image
WeaveBench (Li et al., 2026)	✗	✗	✓	8	Image, text
TheAgentCompany (Xu et al., 2026)	✗	✗	✗	6	Text
GDPVal (Patwardhan et al., 2026)	✗	✗	✓	44	Text, files
DeskCraft (Wang et al., 2026)	✗	✓	✓	5	Image
KOLAB (Ours)	✓	✓	✓	**36**	**Video**, text, files

2 RELATED WORK

Long-horizon agent and workflow benchmarks. Autonomous agent benchmarks cover desktop applications (Xie et al., 2024; Yuan et al., 2026), enterprise software (Dai et al., 2026), hybrid GUI/CLI and scientific computer use (Li et al., 2026; Jang et al., 2026; Wang et al., 2026; Zhu et al., 2026; Sun et al., 2026a), professional deliverables (Patwardhan et al., 2026; Sun et al., 2026b; Xu et al., 2026), and long-running agent behavior (Maldaner et al., 2026). Some include human baselines but do not release human action trajectories. Workflow-GYM (Zhu et al., 2026) collects expert workflow trajectories but uses them as demonstrations to evaluate autonomous agents. Among the benchmarks in Table 1, KOLAB uniquely combines multimodal trajectories of professionals collaborating with agents on self-defined problems with step-level human ratings.

Human-agent collaborative workflows. Prior work argues that autonomous task performance does not ensure effective collaboration (Shen et al., 2026). Existing evaluations often rely on simulated users (Wu et al., 2026; Yao et al., 2024; Vijayvargiya et al., 2026; Qian et al., 2025) or human intervention in short shopping and web browsing tasks (Huq et al., 2025; Mozannar et al., 2025). Collaborative Gym (Shao et al., 2026a) supports real or simulated users but covers only three predefined tasks. CollabSkill (Shao et al., 2026b) assigns professionals problems from existing datasets; Chen et al. (2026) study interaction-level satisfaction with software-engineering agents; and Wang et al. (2025) compare separate human and agent workflows. We publicly release complete workflow trajectories across 36 occupations, capturing how professionals use, modify, or reject agent contributions in their broader work.

3 KOLAB: A MULTIMODAL, MULTIDOMAIN DATASET OF HUMAN-AGENT COLLABORATIVE WORKFLOWS ON LONG-HORIZON PROBLEMS

KOLAB is built to understand human-agent collaborative workflows when humans solve long-horizon problems.

Problem criteria. To ensure that the collected workflows represent substantive professional problem solving rather than generic computer-use activities, we required each problem to satisfy four criteria:

- **Real-world.** The problem must originate from the professional's actual work. Professionals were not given tasks from existing workflow datasets or scenarios.
- **Domain-specific.** The problem must require the application of professional knowledge, skills, or judgment from the participant's field. Tasks that could be completed without domain expertise, such as routine personal activities or generic information seeking, were excluded.
- **Complex and long-horizon.** The problem must require a sequence of interdependent steps to achieve its final goal. Participants were asked to plan a workflow requiring at least one hour of work and consisting of at least five steps, with each step involving multiple computer actions or decisions.
- **Deliverable-oriented.** The problem must result in a concrete final deliverable. Each step must also produce an intermediate output that can be evaluated.

Data collection. We recruited 45 industry professionals via word of mouth, Prolific, and Upwork. Each participant received a compensation of $100 with a bonus for extra time. Figure 3 shows our participants' domains with details in Appendix A.1. Professionals interacted with an agentic AI, Claude (Anthropic, 2026), for two 1-hour recording sessions.[1] Each participant planned their workflow beforehand to ensure that their problem satisfied our criteria. Then they updated the workflow document, rated their collaboration experience and output quality, and filled out the post-interaction survey in the Appendix A.11.

Dataset statistics. KOLAB covers 90 problem sets and human-agent collaborative workflows conducted by professionals from 10 sectors of 36 occupations (Figure 3). The post-interaction survey also measured collaboration satisfaction and output quality on a 1 (low) to 5 (high) scale.[2] Collaboration satisfaction was rated 5 in 55.7% of sessions and 4 in 28.9%, with 15.3% rated 3 or below. Similarly, output quality was rated 5 in 53.6% of sessions and 4 in 31.2%, with 15.2% rated 3 or below.

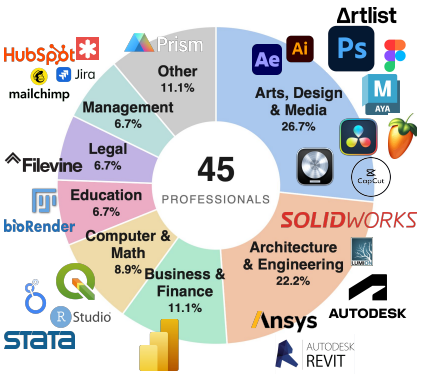


Figure 3: KOLAB domain distributions. "Other" refers to Life & Social Science, Sales, and Healthcare.

Data release. We publicly release KOLAB with seven components: (1) screen-recording videos for all sessions; (2) the workflow documents written by the professionals, including goal, context, and step names/actions; (3) the human–agent chat logs, including agent tool calls; (4) optional input and output files of each workflow; (5) human ratings for collaboration experience and output quality, along with step-level output rubrics written by professionals; (6) the workflow graph per session, with step alignment, timestamps, and adoption labels; and (7) post-session survey responses.

4 UNDERSTANDING HUMAN-AGENT COLLABORATION DYNAMICS

4.1 DATA PROCESSING AND ANNOTATION

To make the workflow structured, each step in the workflow is represented as a temporal graph, where each node $v \in V$ corresponds to a human or agent action at a time interval $\tau(v) = [t_s(v), t_e(v)] \subset \mathbb{R}_{\geq 0}$. [3] A node can branch when activities occur at overlapping timestamps and loops happen when the same action is repeated. Formally, each workflow is represented as follows:

$$G = (V, E, a, w, \tau), \tag{1}$$

where V is a set of human (**H**) or AI (**A**) action nodes; $E \subseteq V \times V$ is the set of directed edges, with $(u, v) \in E$ indicating that action v directly follows from action u.

After graph construction, two human annotators labeled how humans incorporated agent outputs. A human node is identified as a *source node* if its action triggered the agent's generation (*e.g.,* "*design a floor plan*"). An agent node is the *target AI node* if it generates an output that is subsequently acted on by the participant. The human node that acts on this AI output is labeled as follows.

- **ADOPT**: The agent output is observably incorporated into the workflow as-is, such as directly using an agent-generated design or adding a suggested paper to a literature review.
- **MODIFY**: The agent output is incorporated with modifications, *e.g.,* the human revises agent-generated text before using it.
- **REJECT**: The agent output is not incorporated after being reviewed, such as when the human discards it or clearly expresses a negative sentiment and proceeds with a different approach.

[1] They were free to use other AI agents when appropriate for their work. For example, an architect might use Claude for planning and reasoning while using ChatGPT to generate or render architectural images.

[2] Details in Table 8 in the Appendix.

[3] The research team and professionals refined these graphs. An example graph for one step is shown in Figure 10 in the Appendix.

Table 2: Real examples of ADOPT, MODIFY, and REJECT responses from KOLAB, drawn from a music producer workflow, architecture workflow, and a video-production workflow respectively.

Response	Source Node Action	Target AI Node Action	Current Node Action	Software
ADOPT	Request Afrobeats/techno background music in ElevenLabs (67:33–68:30)	ElevenLabs generates Afrobeats instrumental candidates (69:50–70:04)	Capture the ElevenLabs track and import it into the DaVinci Resolve project (72:05–74:36)	ElevenLabs, DaVinci Resolve
MODIFY	Upload the existing shell plan and request a full room layout (24:10–25:28)	Claude delivers a dimensioned floor-plan layout for the lobby/storage circulation (51:21–56:20)	Rebuild walls/partitions in Revit, keeping the liked zoning but fixing lobby/storage circulation (58:40–72:59)	Revit
REJECT	Ask Kling AI to generate a Zanzibar schoolyard video clip (36:30–39:04)	Kling AI generates the requested video clip (39:45–41:17)	Judge the clip's audio unusable and abandon it, out of credits to iterate further (41:17–42:06)	Kling AI

Table 3: Summary statistics of KOLAB for videos, workflow graphs, and human action annotations.

Video (90 sessions)		**Graph** (74 sessions, 82.2%)		**Annotation** (43 sessions, 59.7%)	
Total duration	105 hrs	Workflow steps	503	ADOPT	62.4% (151)
Avg. duration/session	70 mins	Avg. steps/session	5.60	MODIFY	27.3% (66)
		Avg. nodes/step	7.87	REJECT	10.3% (25)

As seen from real examples in Table 9, professionals evaluate agent contributions and act on softwares outside the AI interface. A summary of KOLAB is provided in Table 3. In total, KOLAB contains 90 workflow videos, of which 74 (81%) have corresponding human-curated temporal graph representations and 59.7% are annotated with AI adoption labels, yielding 298 instances for benchmarking MLLMs.

4.2 COLLABORATION DYNAMICS METRICS

Collaboration burstiness. Human-agent interactions are not uniformly distributed over time. Humans may work with an agent steadily or in short, intensive episodes separated by longer periods of independent work. To capture this temporal rhythm, following Goh & Barabási (2008), we estimate collaboration burstiness as $B = \frac{\sigma_\tau - \mu_\tau}{\sigma_\tau + \mu_\tau} \in [-1, 1]$, where μ_τ and σ_τ are the mean and standard deviation of the inter-event intervals. $B \to 1$ indicates bursty interactions, $B \to -1$ indicates regular interactions, and $B = 0$ corresponds to a Poisson-like process.

Workflow topology. Motivated by graph-theoretic representations of complex systems (Bondy & Murty, 2008; Boccaletti et al., 2006; Newman et al., 2011), we characterize workflow topology to capture the extent to which humans work concurrently while the agent generates an output. We operationalize this as the proportion of *branch* nodes with out-degree > 1 (Br) relative to the number of edges ($|E|$) as a workflow topology $W = \frac{Br}{|E|}$.

AI retention rate. To measure how agent outputs are retained in subsequent work (Solé et al., 2026), we define *AI retention* through adoption and modification rates:

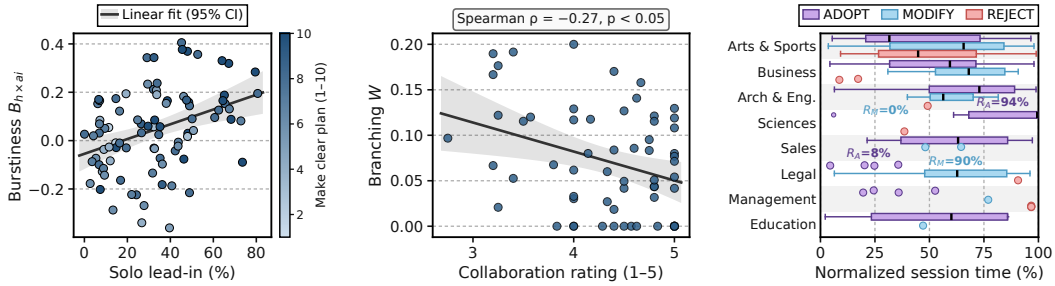
$$R_A = \frac{N_{\text{ADOPT}}}{N_{\text{ADOPT}} + N_{\text{MODIFY}} + N_{\text{REJECT}}}, \qquad R_M = \frac{N_{\text{MODIFY}}}{N_{\text{ADOPT}} + N_{\text{MODIFY}} + N_{\text{REJECT}}}.$$

These capture the proportion of agent outputs incorporated directly or with modification.

4.3 ANALYSIS

Using these three metrics, we analyze how human-agent collaboration varies in its temporal rhythm (B), topology of concurrent work (W), and integration of agent contributions (R). Our analyses are intended to describe the collaborative patterns observed in KOLAB, rather than as population-level comparisons across professional domains. Figure 4 shows the results.

#1: Burstier workflows reflect more planning before interacting with the agent. Collaboration burstiness was relatively consistent within individuals across two sessions ($\rho = 0.67$). We define *solo lead-in* as the proportion of a step completed before the first AI interaction, averaged over all steps in a session. This reflects the proportion of time a person spends working independently before



(a) Collaboration burstiness (b) Workflow topology (c) AI retention rate

Figure 4: Collaboration dynamics in KOLAB. (a) Burstiness vs. solo lead-in; (b) workflow branching vs. collaboration rating; (c) timing of ADOPT, MODIFY, and REJECT responses across sectors.

Table 4: **Benchmark tasks.** T_x^{start} and T_x^{end} denote the start and end timestamps of event or segment x, respectively; x can refer to the source node, target AI node, human adoption node, or the full workflow.

Task	Observed Workflow	Response Visible?	Prediction
1a: Recognition	$T_{\text{source}}^{\text{start}} \rightarrow T_{\text{adoption}}^{\text{end}}$	✓	ADOPT, MODIFY, REJECT
1b: Forecasting	$T_{\text{source}}^{\text{start}} \rightarrow T_{\text{AI}}^{\text{end}}$	✗	
2a: Collaboration Satisfaction	$T_{\text{step}}^{\text{start}} \rightarrow T_{\text{step}}^{\text{end}}$	N/A	Rating 1–5
2b: Output Quality	$T_{\text{step}}^{\text{start}} \rightarrow T_{\text{step}}^{\text{end}}$	N/A	

first interacting with AI. For example, if the first AI interaction occurs 4 minutes into a 10-minute step, the solo lead-in is 40%. As shown in Figure 4a, burstier workflows had longer solo lead-ins ($\rho = 0.37$, $p < 0.001$). Consistent with this pattern, participants who reported having a clearer plan before using AI in the post-interaction survey also had longer solo lead-ins ($\rho = 0.40$, $p < 0.01$).

#2: Workflows with branches are associated with lower perceived collaboration experience. Workflow branching (W) is negatively associated with collaboration ratings (Figure 4b) ($\rho = -0.27$, $p < 0.05$). Although modest, this association suggests that concurrent human and agent activity may introduce coordination costs.

#3: AI retention patterns vary across professional domains. As shown in Figure 4c, most domains predominantly adopt agent outputs as-is, with adoption rates ranging from 59% in Arts & Sports to 94% in Sciences. Legal workflows differ markedly, with 8% adoption and 90% modification. This illustrates work style variation in how professional domains integrate agent contributions.

5 EXPERIMENTS AND RESULTS

5.1 TASKS

Our benchmark tasks aim to answer these research questions as summarized in Table 4.

- **RQ1:** (AI Adoption) How well can frontier MLLMs forecast how humans use agent contributions in long-horizon professional workflows?
- **RQ2:** (Human Rating) How well can frontier MLLMs predict human ratings of collaboration satisfaction and output quality given long-horizon human-agent collaboration workflows?

RQ1 (AI Adoption). We evaluate whether MLLMs can predict whether humans ADOPT, MODIFY, or REJECT agent contributions under two settings—prediction and forecasting—which differ in whether the human response is observable (Table 4). Each input begins at the source node, where the human delegates to the agent ($T_{\text{source}}^{\text{start}}$). Forecasting clips average 6.0 minutes ($SD = 6.7$; max=40.2), while prediction clips average 12.4 minutes ($SD = 12.4$; max=72.2) with more details in Table 11.

Table 5: F1 for AI adoption recognition (Task 1a), and precision (P), recall (R), F1, and per-class F1 for forecasting (Task 1b), for video only, mean±std over 3 runs per model. The best result besides Claude in each column is **shaded**.

	1a: Recognition	1b: Forecasting			1b: Per-label F1		
Model	**F1**	**P**	**R**	**F1**	ADOPT	MODIFY	REJECT
Majority-class	0.249	0.199	0.333	0.249	0.747	0.000	0.000
• Gemini 3.1 Pro	**0.756**±0.017	0.529±0.079	0.434±0.020	0.437±0.028	0.696±0.020	0.447±0.003	0.168±0.069
• Gemini 3.8 Flash	0.662±0.010	0.491±0.031	0.418±0.012	0.420±0.014	**0.734**±0.008	0.317±0.014	0.208±0.023
• GPT-6 Astra	0.748	**0.580**±0.027	**0.558**±0.028	**0.558**±0.023	0.710±0.026	**0.577**±0.043	**0.387**±0.054
• Claude Opus 5.5	0.672	0.635	0.544	0.562	0.750	0.542	0.392
• Qwen3-VL-30B-A3B	0.596±0.010	0.405±0.014	0.412±0.024	0.394±0.022	0.544±0.029	0.359±0.029	0.278±0.019
• Qwen3-VL-32B	0.432±0.033	0.441±0.042	0.382±0.016	0.245±0.012	0.064±0.033	0.447±0.010	0.225±0.022

Table 6: Rating prediction for Experiments 2a and 2b, averaged over 3 runs.

	Exp. 2a: Collaboration			Exp. 2b: Output Quality		
Model	**MAE↓**	**Within-1↑**	**Spearman ρ ↑**	**MAE↓**	**Within-1↑**	**Spearman ρ ↑**
Mode baseline	0.679	0.847	–	0.725	0.836	–
• Gemini 3.1 Pro	**0.697**	**0.843**	0.060	**0.767**	**0.821**	0.006
• Gemini 3.8 Flash	0.798	0.816	0.037	0.817	0.808	−0.006
• GPT-6 Astra	0.960	0.836	**0.332**	0.938	0.779	**0.299**
• Qwen3-VL-30B-A3B	0.912	0.808	0.079	1.239	0.656	−0.015
• Qwen3-VL-32B	1.266	0.609	0.082	1.341	0.563	0.101

RQ2 (Human Rating). An evaluator should distinguish genuinely successful interactions from those in which professionals were dissatisfied or produced lower-quality outcomes, rather than simply reflecting the predominantly positive ratings. Thus, we investigate whether MLLMs can similarly recover professionals' ratings from multimodal, long-horizon human-agent collaboration workflows. Given the full workflow video for each step, we evaluate whether frontier MLLMs can predict the professional's 1-5 rating of collaboration satisfaction (**Task 2a**) and output quality (**Task 2b**).

Experimental Setup We evaluate four leading multimodal agentic models, Gemini 3.8 Flash, Gemini 3.1 Pro, GPT-6 Astra, and Claude Code Opus 5.5, and two open-weight video-capable MLLMs: Qwen3-VL-30B-A3B-Instruct (MoE) and Qwen3-VL-32B-Instruct (dense) (Bai et al., 2025). We compare against majority-class (ADOPT) and mode-rating baselines for RQ1 and RQ2, respectively. Our primary input is workflow video, which captures actions inside and outside the AI interface. We additionally evaluate all seven combinations of video, graph, and chat to assess the contribution of each modality.

5.2 Results and Discussion

Performance on forecasting AI adoption. Table 5 shows each MLLM's performance with video only as input. Claude performs the best for both predicting and forecasting human adoption behavior, yet even the upper-bound only reaches an F1-score of 0.87. Since the default AI used in this study is Claude, we hypothesize that Claude is biased toward its own interface. Even so, its F1 for forecasting human adoption is only 0.558, dominated by ADOPT. It has substantially lower performance on MODIFY and REJECT. For non-Claude models, Gemini 3.1 Pro performs the best when the human's actual response is observable, with an F1 score of 0.756 while GPT-6 performs better at forecasting the response. Gemini models performs particularly poorly at forecasting REJECT, even worse than the open-weight Qwen models. Gemini 3.1 Pro is very biased toward predicting ADOPT. 52% of true MODIFY and 70.8% of true REJECT are predicted as ADOPT by Gemini.

Why should an AI model care to understand human collaboration with other AI models? Humans orchestrate various software tools in our dataset, and their workflows are not necessarily limited to a single AI model. Even within our study, professionals used other AI systems, such as Gemini, ChatGPT, and Grok, in addition to Claude. This makes it increasingly important for agents to understand human collaboration with other models, rather than only interactions involving themselves.

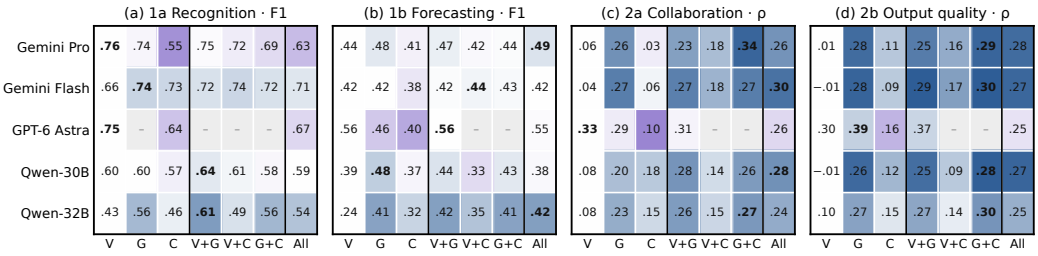
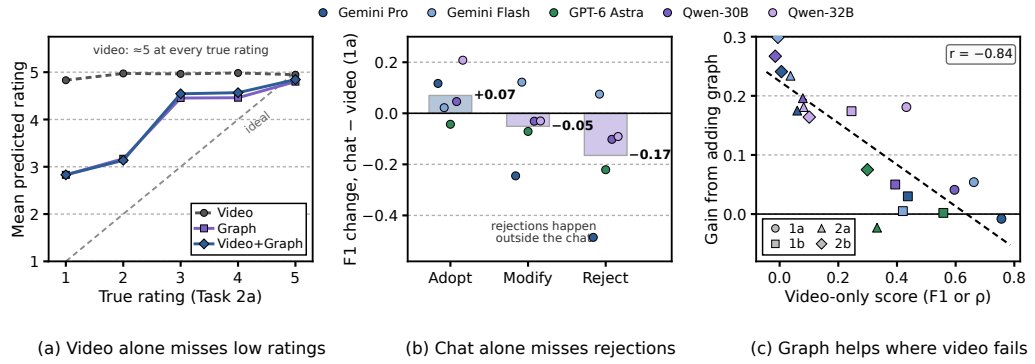


Figure 5: Full results for seven input-modality combinations. V stands for video, G: graph, and C: chat log; "+" is a combination of modalities, "All" refers to using video, graph, and chat as the input.



(a) Video alone misses low ratings (b) Chat alone misses rejections (c) Graph helps where video fails

Figure 6: How each modality helps.

Performance on predicting collaboration satisfaction and output quality. We evaluate alignment with human ratings using mean absolute error (MAE), exact and within-1 accuracy, and Spearman's ρ. The mode baseline outperforms every MLLM in MAE (0.590 for collaboration satisfaction and 0.744 for output quality), reflecting the strong skew of human ratings toward 5. Gemini 3.1 Pro achieves the lowest MAE for both collaboration satisfaction (0.667) and output quality (0.846), but still fails to outperform this trivial baseline. Moreover, rank agreement with human judgments remains weak. GPT-6 achieves the highest Spearman correlation for both collaboration satisfaction ($\rho = 0.378$) and output quality ($\rho = 0.329$), while several models exhibit near-zero or even negative correlations. Thus, even when disregarding differences in absolute rating scale, current frontier multimodal agents and LLMs still struggle to recover how professionals comparatively judge their collaboration and resulting outputs.

Modality Analysis We examine how video, graph, and chat, individually and in combination, affect models' understanding of human-agent collaboration (Figures 5 and 6). No modality consistently performs best across models or tasks. Video alone yields the highest recognition performance with Gemini 3.1 Pro (F1= 0.76, Task 1a) and forecasting performance with GPT-6 (F1= 0.56, Task 1b). However, chat alone particularly hurts forecasting of REJECT cases: relative to video, F1 changes by $+0.07$ for ADOPT, -0.05 for MODIFY, and -0.17 for REJECT (Figure 6b), likely because chat does not capture how agent outputs are subsequently used outside the AI interface. Yet video is not consistently optimal. Human-curated workflow graphs, which capture high-level professional activities, particularly benefit the non-agentic Qwen models; for Qwen-32B, adding graphs to video improves forecasting F1 from 0.24 to 0.42.

For human-rating prediction, performance remains limited regardless of modality (Figure 5c–d). Correlations peak at only $\rho = 0.34$ for collaboration satisfaction (Gemini 3.1 Pro, graph+chat) and $\rho = 0.39$ for output quality (GPT-6, graph only). Graphs are particularly informative for output quality: video alone yields near-zero correlations for Gemini 3.8 Flash and Qwen-30B ($\rho = -0.01$), while adding graphs raises them to $\rho = 0.26$–0.30. Overall, providing additional modalities does not resolve MLLMs' difficulty in approximating professionals' judgments of collaboration experience and output quality.

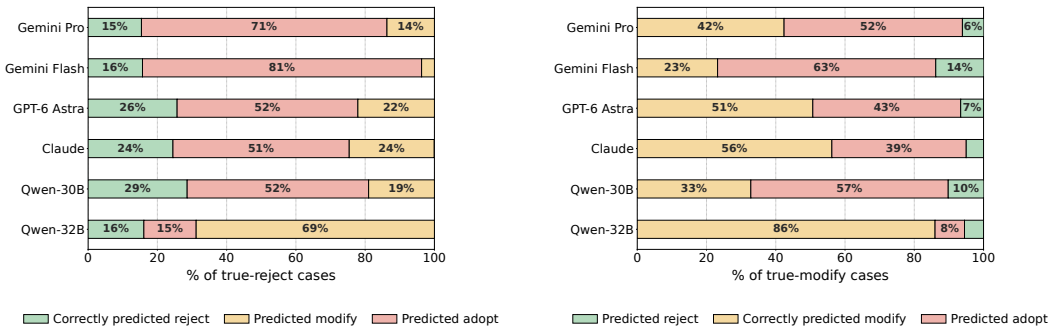


Figure 7: Prediction distributions for REJECT and MODIFY cases in AI adoption forecasting.

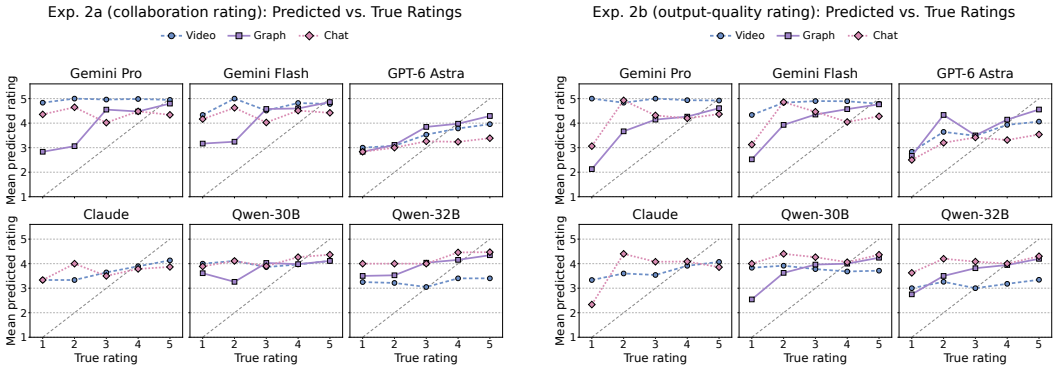


Figure 8: Mean predicted rating for Task 2a and Task 2b vs. ground truth

5.3 FAILURE CASES

For adoption forecasting (**Task 1b**), several models strongly favor particular labels even though the true labels are REJECT or MODIFY (Figure 7). Gemini 3.1 Pro and Flash consistently overpredict ADOPT; with video, Gemini 3.8 Flash predicts ADOPT for 79.2% of examples compared with a 62.0% ground-truth rate. Qwen3-32B instead favors MODIFY, predicting it in 91.5% of cases compared with 28.1% in the ground truth. These biases largely persist with graph and chat inputs, except for Qwen3-30B, which shifts from MODIFY with video to ADOPT with graph and chat. Models therefore exhibit distinct forecasting biases, which vary with both the model and input modality.

For human-rating prediction (Figure 8),some models predict similar ratings all the time. With video, Gemini 3.1 Pro predicts 4.83 for a true rating of 1 and 4.95 for a 5 (Task 2a); in Task 2b, it predicts 5.00 and 4.92. Graph input improves this separation, with gaps between true ratings of 1 and 5 reaching 1.96 and 2.49 for Tasks 2a and 2b. Qwen variants also show a similar pattern. However, GPT-6 and Claude already separate low and high ratings with video (gaps of 0.74–1.23) and remain more stable across modalities.

These systematic biases raise a more fundamental question: to what extent are MLLMs actually grounding their predictions in the observed workflow? If models understand the workflow context, their predictions should vary with the human behaviors and outcomes they observe. Some models repeatedly produce similar labels or ratings despite different inputs and ground-truths. It suggests that their predictions may be driven by model-specific priors instead of the workflow itself.

6 CONCLUSION AND FUTURE WORK

We have introduced KOLAB to benchmark frontier multimodal agentic models on their capabilities to support human-agent collaboration better on long-horizon workflow. Our results show that even frontier MLLMs, including GPT-6 , struggle to predict and understand human–AI collaboration. In future work, we will explore ways to improve these capabilities by constructing training datasets or developing sandbox environments where models can practice collaborating with simulated users before real-world deployment.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the execution of the user study since the participants are using generative AI tools, construction of initial workflow graphs, and development of software code. We did not use generative AI tools for theoretical development, hypothesis formulation, dataset processing, qualitative analysis, or interpretation of results. Additionally, we used generative AI tools for writing and editing the paper, creating visualizations, literature search, but we do not use it for filling out the references in the paper, brainstorming, and formatting tables and LaTeX elements. Graphs were all edited by human researchers. All workflow graph annotations were performed and reviewed by human annotators. We take responsibility for the final content of this work, including text, claims, analyses, code, visualizations, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study was approved by our institution's IRB. All participant data are anonymized and securely stored.

REPRODUCIBILITY STATEMENT

Data collection details are explained in A.2 in the Appendix. Data processing and annotation details are explained in A.4. All model prompts and model settings are elaborated in the Appendix.

REFERENCES

Anthropic. Claude. https://claude.ai, 2026. Accessed: 2026-09-25.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

John Adrian Bondy and Uppaluri Siva Ramachandra Murty. *Graph theory*. Springer Publishing Company, Incorporated, 2008.

Valerie Chen, Rohit Malhotra, Xingyao Wang, Juan Michelini, Xuhui Zhou, Aditya Bharat Soni, Hoang H. Tran, Calvin Smith, Ameet Talwalkar, and Graham Neubig. How can we assess human-agent interactions? case studies in software agent design. In *International Conference on Machine Learning (ICML)*, Seoul, South Korea, 2026. URL https://arxiv.org/abs/2510.09801.

Yutong Dai, Krithika Ramakrishnan, Jing Gu, Matthew Fernandez, Yanqi Luo, Viraj Prabhu, Zhenyu Hu, Caiming Xiong, Zeyuan Chen, Ran Xu, et al. Scuba: Salesforce computer use benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 24363–24386, 2026.

K-I Goh and A-L Barabási. Burstiness and memory in complex systems. *Europhysics Letters*, 81 (4):48002, 2008.

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.

Lawrence Keunho Jang, Andrew Keunwoo Jang, Jing Yu Koh, and Ruslan Salakhutdinov. Mypcbench: A benchmark for personally intelligent computer-use agents. *arXiv preprint arXiv:2606.16748*, 2026.

Wanli Li, Bowen Zhou, Yunyao Yu, Zhou Xu, Yifan Yang, Dongsheng Li, and Caihua Shan. Weavebench: A long-horizon, real-world benchmark for computer-use agents with hybrid interfaces. *arXiv preprint arXiv:2606.09426*, 2026.

Matheus Kunzler Maldaner, Adam Fourney, Amanda Swearngin, Hussein Mozannar, Gagan Bansal, Maya Murad, Rafah Hosn, and Saleema Amershi. Sentinelbench: A benchmark for long-running monitoring agents. *arXiv preprint arXiv:2606.05342*, 2026.

Mike Merrill, Alexander Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Shin, Thomas Walshe, E Kelly Buchanan, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *International Conference on Learning Representations*, volume 2026, pp. 40903–40986, 2026.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. SWE-lancer: Can frontier LLMs earn $ 1 million from real-world freelance software engineering? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44412–44450. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/miserendino25a.html.

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.

Mark Newman, Albert-László Barabási, and Duncan J Watts. *The structure and dynamics of networks*. Princeton university press, 2011.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simon Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, et al. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. In *International Conference on Learning Representations*, volume 2026, pp. 24005–24040, 2026.

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, 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.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. In *International Conference on Learning Representations*, volume 2026, pp. 99616–99649, 2026a.

Yijia Shao, Zora Zhiruo Wang, Neel Ahuja, Yicheng Wang, Bowen Liu, and Diyi Yang. Collabskill: Evaluating human-agent collaboration on real-world tasks. *arXiv preprint arXiv:2606.09833*, 2026b.

Shannon Zejiang Shen, Valerie Chen, Ken Gu, Alexis Ross, Zixian Ma, Jillian Ross, Alex Gu, Chenglei Si, Wayne Chi, Andi Peng, Jocelyn J Shen, Ameet Talwalkar, Tongshuang Wu, and David Sontag. Scaling collaborative effort with agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12254–12271, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.595. URL https://aclanthology.org/2026.findings-acl.595/.

Ricard Solé, Giulio Ruffini, Francesca Castaldo, Marco Tuccio, Luis F Seoane, Manlio de Domenico, Santiago F Elena, David C Krakauer, and Michael Levin. Large-language models as a cognitive virus. *arXiv preprint arXiv:2609.03344*, 2026.

Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, et al. Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows. In *International Conference on Learning Representations*, volume 2026, pp. 75694–75731, 2026a.

Yiyou Sun, Xinyang Han, Weichen Zhang, Yuanbo Pang, Tianyu Wang, Yuhan Cao, Yixiao Huang, Chris Duroiu, Haoyun Zhang, Jeffrey Lin, et al. Agents' last exam. *arXiv preprint arXiv:2606.05405*, 2026b.

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. Ambigswe: Interactive agents to overcome underspecificity in software engineering. In *International Conference on Learning Representations*, volume 2026, pp. 146716–146737, 2026.

Wenkai Wang, Tao Xiong, Jingchen Ni, Yunpeng Bao, Xiyun Li, Tianqi Liu, Hongcan Guo, Zilong Huang, and Shengyu Zhang. Deskcraft: Benchmarking desktop agents on professional workflows and human-in-the-loop collaboration. *arXiv preprint arXiv:2606.03103*, 2026.

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.

Yaozu Wu, Wei-Chieh Huang, Jizhou Guo, Dongyuan Li, Renhe Jiang, Henry Peng Zou, Chunyu Miao, Shanghao Li, Weizhi Zhang, WeiWei Ye, et al. Has-bench: Evaluating llm-based human-agent systems under configurable human participation. *arXiv preprint arXiv:2607.04329*, 2026.

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.

Frank Fangzheng Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *Advances in Neural Information Processing Systems*, 38, 2026.

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.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, Dunjie Lu, Haikong Lu, Qi Zhen, Xinyuan Wang, Jiaqi Deng, Yuhao Yang, Cheng Chen, Boyuan Zheng, Alex Su, Xiao Yu, Hao Zou, Saaket Agashe, Xing Han Lu, Manpreet Kaur, Zhengyang Qi, Vincent Sunn Chen, Frederic Sala, Dayiheng Liu, Junyang Lin, Zhou Yu, Yu Su, Siva Reddy, Xin Eric Wang, Peng Qi, Tianbao Xie, and Tao Yu. Osworld2.0: Benchmarking computer use agents on long-horizon real-world tasks, 2026. URL https://arxiv.org/abs/2606.29537.

Liya Zhu, Jingzhe Ding, Jian Zhang, Jianbo Xue, Shihao Liang, Ge Zhang, Yi Zhu, Duju Zeng, Xiang Gao, Qingshui Gu, et al. Workflow-gym: Towards long-horizon evaluation of computer-use agentic tasks in real-world professional fields. *arXiv preprint arXiv:2606.11042*, 2026.

Henry Peng Zou et al. Llm-based human-agent collaboration and interaction systems: A survey. *arXiv preprint arXiv:2505.00753*, 2025. Findings of ACL 2026. 20 total co-authors, abbreviated here.

A Appendix

A.1 Professionals Background and Software Usage

The 45 study participants span 10 O*NET sectors and 36 distinct occupations, ranging from film editors, graphic designers, and audio engineers to aerospace engineers, GIS specialists, attorneys, and bioinformaticians. We prioritize non-computer science professions, which remain comparatively under-resourced with real-world task data relative to software engineering (Patwardhan et al.,

2026; Shao et al., 2025; Sun et al., 2026b; Miserendino et al., 2025). Across these roles, participants reported using 96 distinct profession-specific non-AI software applications and 8 AI tools in their day-to-day work, as shown in Figure 9. This mix of shared productivity tools and domain-specific software illustrates the breadth of professional practice that the benchmark's workflows are grounded in.

Table 7: Summary statistics of the 45 study participants. Demographics are self-reported via the intake survey and were unavailable for 6 Prolific-recruited participants ("Not reported").

Category	Item	Count
Coverage	Total participants	45
	O*NET sectors	10
	Occupations	36
Education	PhD	9
	Master's	19
	Bachelor's	13
	Other	4
Age range	18–25	7
	26–35	16
	36–45	13
	46–55	2
	56–65	1
	Not reported	6
Work experience	< 1 year	5
	1–3 years	13
	4–6 years	7
	6–10 years	5
	> 10 years	9
	Not reported	6
Self-rated expertise	Intermediate (3/5)	10
	Advanced (4/5)	15
	Expert (5/5)	14
	Not reported	6
GenAI familiarity	Low (1–4/10)	1
	Medium (5–7/10)	11
	High (8–10/10)	27
	Not reported	6

Category	Country	Count
Geography (21 countries)	USA	16
	Colombia	3
	India	3
	Mexico	3
	Germany	2
	Nigeria	2
	UK	2
	Argentina	1
	Ecuador	1
	Egypt	1
	El Salvador	1
	Finland	1
	Kenya	1
	Montenegro	1
	Morocco	1
	Netherlands	1
	Pakistan	1
	South Africa	1
	Sri Lanka	1
	Turkey	1
	Yemen	1

A.2 Data Collection Details

We collected KOLAB through two sessions per participant, in which we recorded the participant's computer screen. Each session lasted approximately one hour or longer and centered on a problem chosen by the participant. After data processing and due to time constraints, some problems contained fewer than five steps; however, each interaction session still lasted approximately one hour. To capture different patterns of human-agent collaboration, we used two protocols. In Session 1, participants were encouraged to interact with the agent at least five times while completing their workflow. In Session 2, participants were encouraged to interact with the agent at each step of the workflow. The two sessions did not have to address the same problem. However, participants could use the second session to continue or extend the work from the first session. For example, a participant might use the first session to develop a branding logo and use the second session to create a marketing advertisement based on the resulting brand identity. Participants were required to interact with agentic AI, Claude, using our credentials during each session and were free to use other AI agents when appropriate for their work. For example, an architect might use Claude for planning and reasoning while using ChatGPT to generate or render architectural images. At the time of data collection, most participants used the default Claude model (Sonnet 5.5), while a few used Opus 5. Each participant planned their workflow before the recording sessions to ensure that their problem satisfied our task criteria. After the recording sessions, they updated the workflow document and completed surveys about their collaboration experience and output quality. During each recording

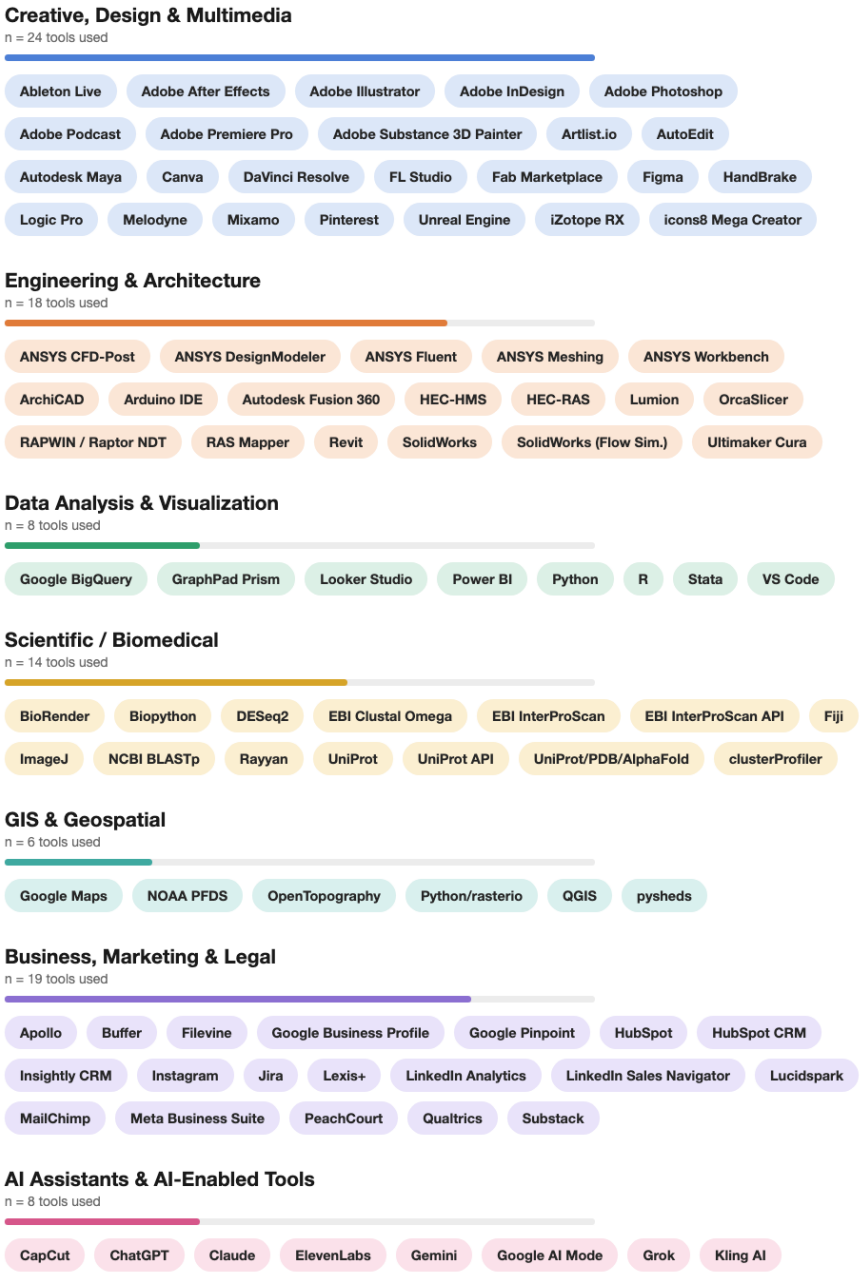


Figure 9: Software used by participants in their day-to-day work, grouped by category.

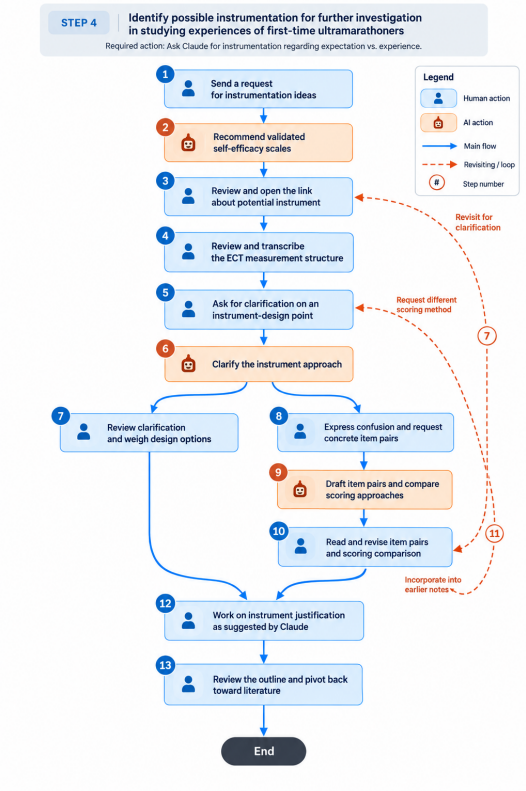


Figure 10: Example workflow graph for one step of a participant's session.

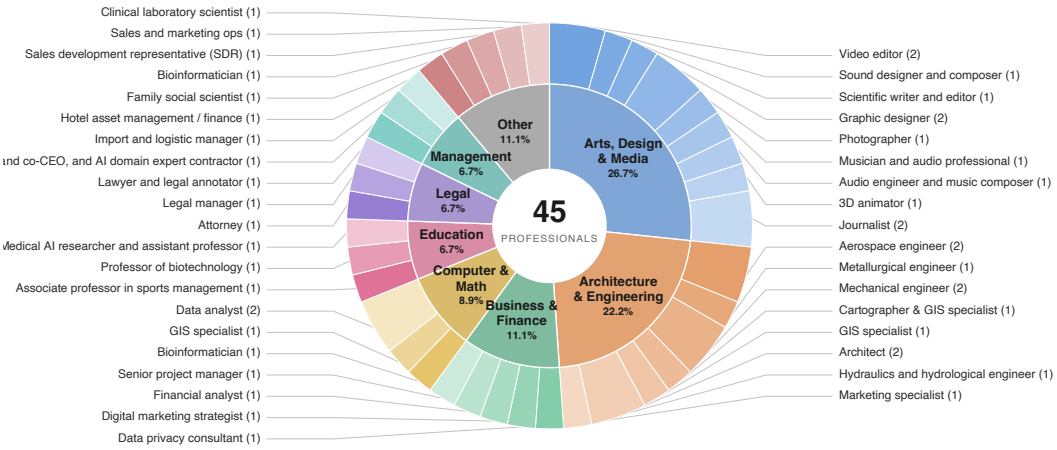


Figure 11: Occupation distributions of KOLAB. "Other" refers to three O*NET sectors: Life & Social Science, Sales & Related, and Healthcare.

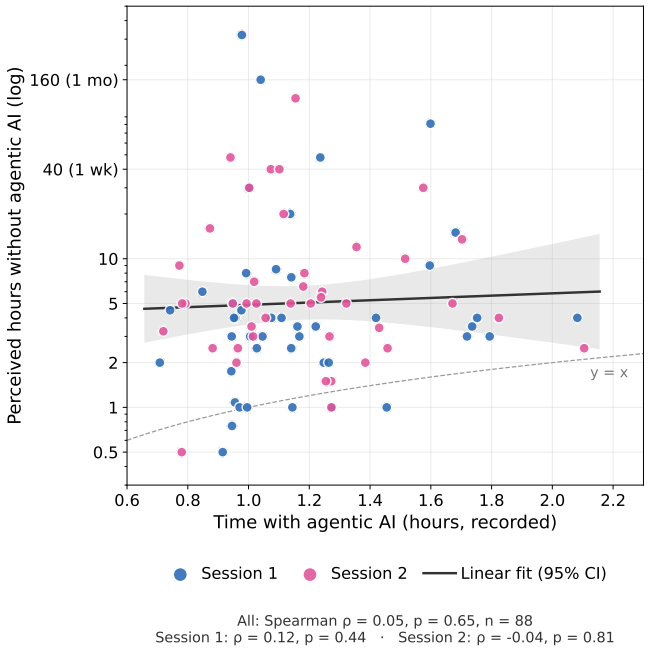


Figure 12: Time spent in each recording session working with agentic AI versus the participant's estimated time to solve the same problem without agentic AI.

session, participants were encouraged to think aloud; we used the transcriptions to construct the workflow graphs and to guide the annotation of response labels.

A.3 Graph as a Collaboration Workflow Representation

Actions and timestamps from workflow video are aligned with steps documented by each professional. We then represented each workflow as a directed graph, where each node corresponds to a human or agent action. The research team and professionals jointly reviewed and refined these graphs. Graphs can capture non-linear workflows; for instance, a node branches when activities occur in parallel and loops happen when the same action is repeated.

Formally, each session is represented as a directed acyclic temporal graph

$$G_{temporal} = (V, E, a, w, \tau), \tag{2}$$

where V is the set of nodes, each corresponding to a discrete human or AI action, and $E \subseteq V \times V$ is the set of directed edges, with $(u, v) \in E$ indicating that action v directly follows from action u within the workflow. Each node $v \in V$ is annotated by three functions: an actor label

$$a : V \to \{\mathbf{H}, \mathbf{A}\}, \tag{3}$$

indicating whether the action was performed by the human or the agent; a workflow-step label

$$w : V \to \{1, \ldots, K\}, \tag{4}$$

indicating which of the K steps of the task the action belongs to; and a time interval

$$\tau(v) = [t_s(v), t_e(v)] \subset \mathbb{R}_{\geq 0}. \tag{5}$$

recording when the action began and ended.

After graph construction, two human annotators annotated how each AI output was incorporated into the workflow. For every observed user response to an AI output, annotators identified the *prompt node* that triggered generation, the *AI-content node* being acted on, and a *response label*: **Adopt** (incorporated without modification), **Modify** (incorporated with changes), or **Reject** (explicitly

rejected or not incorporated). This is an annotation layer on top of $G_{temporal}$: each node $n \in V$ representing a user response carries a set of response triples as follows.

$$U(n) = \{(p_i, c_i, \ell_i)\}_{i=1}^{m_n}, \tag{6}$$

where $p_i \in V \cup \{\emptyset\}$ is the prompt node (absent if the user acts on existing AI content without a new prompt), $c_i \in V$ is the AI-content node ($a(c_i) = \mathbf{A}$), and $\ell_i \in \{\text{adopt}, \text{modify}, \text{reject}\}$ is the response label.

The following is an example workflow document that an architect filled out to plan their workflow before each recording session. After the session, they revised the content and filled out the assessment questions about collaboration experience and output quality ratings. Other step details are omitted due to space.

Goal: Developing an architectural plan for Residential House
Context: I have chosen this project because it requires a set of expertise to layout an architectural plan. Conditions including design brief, site and weather analysis, building codes, area and zoning standards to evaluate and design according to that as an expert.

Did you achieve your goal after ~1-hour interaction with Claude? Yes
Why or why not? Claude was able to produce layouts

Step 1 name: Summary of design brief, a set of requirements to start the project

Step action: Prepare the design brief and define scope of project.
Input: client brief shared with me
Output: Revise details when design brief in comprehensive format is generated.
Softwares: Claude

Rubric:

1. Each design improvement includes a clear architectural reason
2. Each revision idea includes a rationale.
3. The revised layout remains consistent with the client brief.
4. Suggestions are consistent with the requirements.

Output Quality and Collaboration Assessment

1. Rate the output quality from 1–5 (1 not satisfy, 5 satisfy). **4**
2. Reason for your rating? *"It created a set of pointers in design brief from the narrative, and curated a long brief in detail. Added extra things also that were not questioned."*
3. Which ones of the rubrics above are satisfied in this output? **2, 3, and 4**

Human-Agent Collaboration

1. Did the AI Agent successfully complete the task? **Yes**
2. Why? *It created a set of pointers in design brief from the narrative, and curated a long brief in detail. Added extra things also that were not questioned.*
3. Collaboration satisfaction — write a number (1–5) to the left of each statement, where 1 = not satisfied and 5 = satisfied:
   - **5** — The agent understood the task, context, and instructions provided.
   - **4** — The agent responded appropriately to requests and adapted when given feedback or corrections.
   - **4** — The agent reduced effort, time, or cognitive load during the task.
   - **4** — The agent proactively suggested useful ideas, alternatives, or next steps when appropriate.
   - **5** — The agent produced accurate, relevant, and trustworthy outputs that required minimal correction.
   - **2** — The interaction felt smooth, with the agent contributing at the right moments without creating unnecessary work.
   - **3** — I was able to guide, redirect, and override the agent's actions when desired.
4. Rate the overall collaboration experience from 1–5 (1 not satisfy, 5 satisfy). **4**

Step 2: area codes to use in project
Step 3: plan zoning and divisions of spaces
Step 4: Arrangement of spaces and Finalizing layout
Step 5: Architectural floor Plan finalization

Another example of workflow document by a different participant (sound designer and composer) is shown below. In this case, the agentic AI failed to help and actually slowed down the process.

Goal: Sound Design a promotional video for a real estate sales pitch
Context: This goal will allow me to evaluate how AI can enhance my work as a sound designer. The plan is to work on a promotional sales pitch for apartments in cozumel, which was recorded by a production company, but they lack sound for many of the camera shots they use (I.E: Shots of the beach, of the town and its localities, etc.) This lack of sound makes the project seem empty, therefore, a sound designer is needed to enhance the immersion of what is shown by adding the ambience noise, sound of actions (Walking, placing, etc) while ensuring audio files don't overlap and work in symphony.

Did you achieve your goal after ~1-hour interaction with Claude? No

Why or why not? *"Although we did advance on the video, It was not finished in the 1 hour mark. I felt like I had to keep up with Claude at the start, rather than being benefited by it while working. This gradually improved when I used it for smaller activities like finding specific SFX or a good usage of effect plugins, but when I initially tried to use it as an analysis assistant, it took almost 20 minutes just analyzing the video, followed by 10 more minutes of analyzing just a fragment of it and providing ideas, which had slowed me down significantly, as I was already done with most of the other things I had not delegated to Claude."*

Step 1 name: Analyze the video and understand key moments from it

Step action: Understand key moments where sound design can exponentially increase quality and immersion for the video, uploading the video to Claude for it to create this in depth analysis. **Input:** MP4 video with dialogue and visual insights to the required SFX **Output:** A pdf or DOCX document with key moments, time stamps and descriptions of the required sound effects. **Softwares:** FL Studio, Artlist, Claude Max

Rubric:

1. A score of 5 means that Claude was able to analyze the video and describe with detail what it is, what its selling, what key visuals and moments it has and how to interpret it as a client.
2. A score of 5 means the analysis included accurate, verifiable timestamps for each key moment, ready to use directly without needing to re-watch the video.

Output Quality and Collaboration Assessment

1. Rate the output quality from 1–5 (1 not satisfy, 5 satisfy). **1**
2. Reason for your rating? *"Claude was not able to fully analyze the video, due to a runout of time."*
3. Which ones of the rubrics above are satisfied in this output? **N/A**

Human-Agent Collaboration

1. Did the AI Agent successfully complete the task? **No.**
2. Why? *"Claude mentioned a timeout when trying to analyze the video."*
3. Collaboration satisfaction — write a number (1–5) to the left of each statement, where 1 = not satisfied and 5 = satisfied:
   - **3** — The agent understood the task, context, and instructions provided.
   - **1** — The agent responded appropriately to requests and adapted when given feedback or corrections.
   - **1** — The agent reduced effort, time, or cognitive load during the task.
   - **1** — The agent proactively suggested useful ideas, alternatives, or next steps when appropriate.
   - **1** — The agent produced accurate, relevant, and trustworthy outputs that required minimal correction.
   - **1** — The interaction felt smooth, with the agent contributing at the right moments without creating unnecessary work.
   - **2** — I was able to guide, redirect, and override the agent's actions when desired.
4. Rate the overall collaboration experience from 1–5 (1 not satisfy, 5 satisfy). **1**

Step 2: Develop the sound design concept
Step 3: Find and select appropriate sound effects
Step 4: Generate a sound effect from scratch
Step 5: Create a step by step guide to improve the realism of a sound effect using effects (EQ, Reverb, etc.)

Table 8: Distribution of collaboration satisfaction and output quality ratings, by session.

Rating	Session 1	Session 2	Total
Collaboration Satisfaction			
5	111 (55.2%)	122 (56.2%)	233 (55.7%)
4	53 (26.4%)	68 (31.3%)	121 (28.9%)
3	26 (12.9%)	16 (7.4%)	42 (10.0%)
2	7 (3.5%)	7 (3.2%)	14 (3.3%)
1	4 (2.0%)	4 (1.8%)	8 (1.9%)
Output Quality			
5	111 (49.8%)	133 (57.3%)	244 (53.6%)
4	77 (34.5%)	65 (28.0%)	142 (31.2%)
3	24 (10.8%)	20 (8.6%)	44 (9.7%)
2	5 (2.2%)	9 (3.9%)	14 (3.1%)
1	6 (2.7%)	5 (2.2%)	11 (2.4%)

Table 8 shows the rating distribution by participants for collaboration satisfaction (1 low, 5 high) and output quality (1 low, 5 high) for each step.

A.4 Data Processing and Annotation Details

The actions from workflow video are aligned with steps documented by each professional. Claude was used to map video timestamps to the corresponding steps, after which professionals reviewed and corrected the alignments. We then represented each workflow as a directed graph, where each node corresponds to a human or AI action and is associated with a workflow step. Nodes with out-degree greater than one denote a branch into concurrent sub-paths, and nodes with in-degree greater than one denote a merge of previously concurrent branches back into a single line of work. The authors and professionals jointly reviewed and refined these graphs. Graphs can include branches when activities occur in parallel and loops when the same action is repeated, allowing the representation to capture non-linear workflow structures as shown in Figure 10. After graph construction, two human annotators annotated how agent outputs were incorporated into the workflow. For each observed response to an agent output, annotators identified the source node, corresponding to the human action that initiated the request, and the agent node, corresponding to the resulting agent response. They then identified the subsequent human node in which the response was acted upon and labeled the user's response as ADOPT, MODIFY, or REJECT. ADOPT indicates that the AI output was incorporated without modification; MODIFY indicates that it was incorporated with changes; and REJECT indicates that the user explicitly rejected the output or did not incorporate it into the workflow. Because an AI output may be acted upon later in the workflow rather than immediately after it is produced, these annotations explicitly link each user response to the AI output it concerns.

A.4.1 Annotation Guidelines

Each workflow graph records not only the sequence of user prompts and Claude responses, but also how each Claude response was incorporated into the user's workflow. This is captured by the `user_response` field, which is added only to the user node that represents the adopt, modify, or reject decision.

`user_response` takes the following form:

```
"user_response": [["source_node_id", "claude_node_id", "label",
```

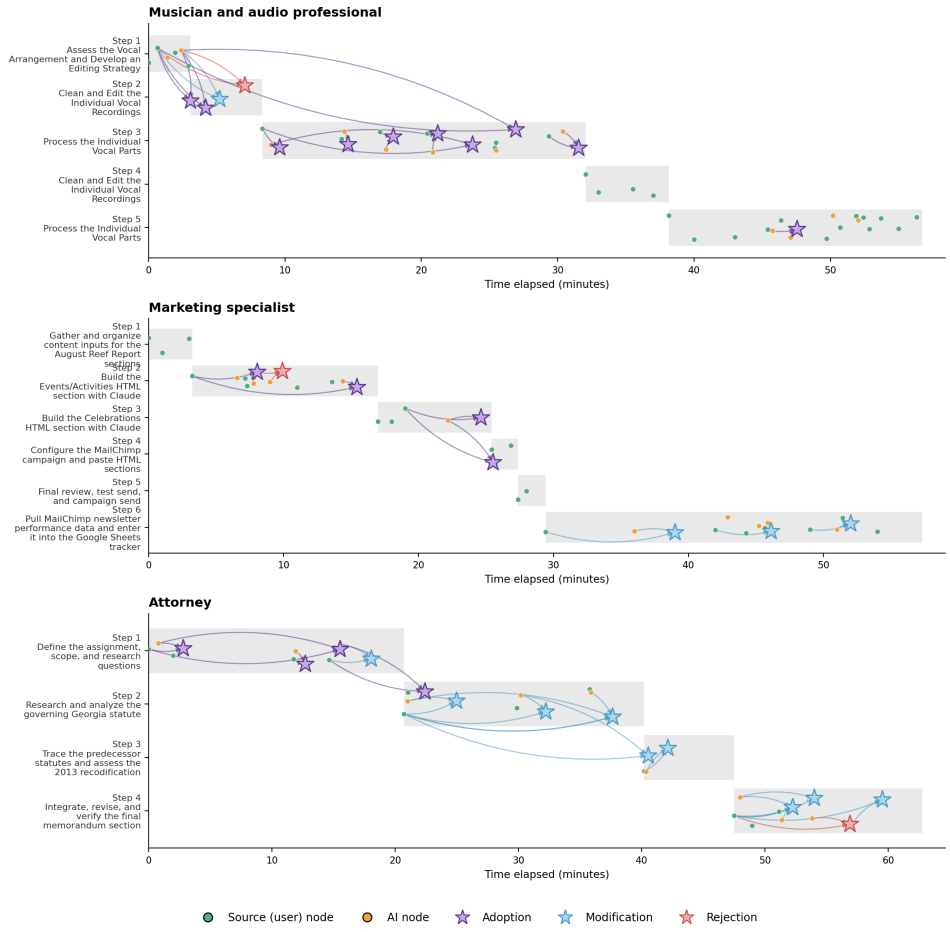


Figure 13: How three professionals (musician, marketing specialist, and attorney) adopt, modify, and reject agent outputs.This is based on the real data point in KOLAB.

Table 9: Real examples of ADOPT, MODIFY, and REJECT responses from KOLAB, drawn from a music producer workflow, architecture workflow, and a video-production workflow respectively.

Response	Source Node Action	Target AI Node Action	Current Node Action	Software
ADOPT	Request Afrobeats/techno background music in ElevenLabs (67:33–68:30)	ElevenLabs generates Afrobeats instrumental candidates (69:50–70:04)	Capture the ElevenLabs track and import it into the DaVinci Resolve project (72:05–74:36)	ElevenLabs, DaVinci Resolve
	Ask Claude to add a faith-based closing line to the voiceover script (13:07–13:41)	Claude adds the requested faith-based reference to the script (34:17–34:22)	Generate the final voiceover in ElevenLabs and import it into DaVinci Resolve (34:22–36:30)	Claude, ElevenLabs, DaVinci Resolve
MODIFY	Upload the existing shell plan and request a full room layout (24:10–25:28)	Claude delivers a dimensioned floor-plan layout for the lobby/storage circulation (51:21–56:20)	Rebuild walls/partitions in Revit, keeping the liked zoning but fixing lobby/storage circulation (58:40–72:59)	Revit
	Ask whether CapCut's AI Relight tool requires the desktop app (63:49–63:55)	Claude explains AI Relight also works in the web version (64:01–64:04)	Retry AI Relight in the browser; still blocked by the desktop-app requirement (64:04–65:17)	CapCut
REJECT	Ask Kling AI to generate a Zanzibar schoolyard video clip (36:30–39:04)	Kling AI generates the requested video clip (39:45–41:17)	Judge the clip's audio unusable and abandon it, out of credits to iterate further (41:17–42:06)	Kling AI
	Ask ChatGPT to generate the same video clip (40:09–40:28)	ChatGPT explains it cannot generate video and offers a still image instead (40:28–41:03)	Review ChatGPT's limitation and move to another tool (41:03–41:17)	ChatGPT

```
"source_node_time", "claude_node_time",
"user_response_time"]]
```

This ensures that each adopt, modify, or reject annotation is linked to the correct occurrence of the repeated node, rather than being ambiguous across its several instances.

Field definitions.

- **`source_node_id`** — The user node ID that initiated the Claude request for this step of the workflow. This node may originate from a previous step: all subsequent user responses within the same workflow continue to reference it as the source, even as the user issues additional requests. `source_node_id` is left empty when the user reuses a prior Claude response within a new step whose narrative is not coherently linked back to the node from that earlier step.
- **`claude_node_id`** — The Claude node ID containing the response to the request.
- **`label`** — One of `adopt`, `modify`, or `reject`, indicating how the user incorporated the Claude response into the workflow. On rare occasions a single node carries more than one label; in that case, each label is recorded as a separate tuple with its own `source_node_id` and `claude_node_id`.

Label definitions.

- ADOPT: The user incorporates Claude's response into the workflow without changes.
- MODIFY: The user incorporates Claude's response into the workflow after modifying it.
- REJECT: The user does not incorporate Claude's response into the workflow. This label is applied only when the user explicitly states that they are rejecting the response.

Examples of how professionals adopt, modify, or reject agent outputs are shown in Figure 13. A request from a previous step can undergo multiple refinements before being adopted in a later step.

A.5 DETAILS FOR CONSTRUCTING AND ANNOTATING GRAPHS FROM VIDEOS

Figure 14 shows our workflow annotation interface. Professionals first align their documented workflow steps with the corresponding video timestamps (Figure 14a). Step names are taken from their workflow documents and initially mapped to the video by Claude Code Sonnet 5.5. After professionals correct the step boundaries, we extract a workflow graph for each step and present it for verification (Figure 14b), where professionals confirm whether the high-level workflow structure is

consistent with their actual process. The prompt for generating the initial graph by Claude is shown below.

Four members of the research team then reconstruct and annotate the full workflow graphs, as the automatically extracted graphs frequently contain errors in action labels, edges, timestamps, and actor assignments. Annotators use the screen recordings, think-aloud transcripts, agent chat logs, and input/output files to ground these corrections. When an action or dependency remains ambiguous, the research team consults the corresponding professional for clarification.

Workflow Graph Extraction

You are given:

1. A workflow video.
2. A list of workflow steps with timestamps.
3. The workflow goal.
4. A transcript of the video (.srt).
5. A workflow planning document (.docx).
6. A chat log between human user and Claude

Your task is to extract a behavioral workflow graph, capturing goal-directed activities performed by the human user and AI throughout the workflow.

Behavioral nodes

A behavioral node represents what the actor is trying to accomplish, not the computer use actions.

You must focus on the underlying work activity, intention, or objective and include relevant context, such as the object being worked on or the purpose of the activity.

Examples:

* Define objective
* Plan approach
* Seek information
* Gather evidence
* Analyze information
* Compare alternatives
* Refine requirements
* Verify findings
* Make decision
* Implement solution
* Validate outcome
* Revise solution

These are examples only, not a fixed taxonomy. Use a more specific label when the context supports it.

For example, prefer:

> "Evaluate candidate methods for the analysis"

over:

> "Read information"

Do not create nodes for low-level actions such as clicking, typing, scrolling, switching tabs, opening files, copying text, or reading an AI response. Similarly, do not create nodes simply because the actor is using a particular software application.

Only include a software interaction when it represents a meaningful work activity.

A behavioral graph corresponds to the workflow step at one time-stamp range. For example, if the workflow is "Step 2" $\rightarrow$ "Step 3" $\rightarrow$ "Step 2", you must have 3 different behavioral graphs.

Use the video and workflow context to infer the actor's behavior. Do not invent intentions that are unsupported by the evidence.

Actors

Include meaningful behavioral activities performed by both:

* `user`

* `ai brand`

For AI, indicate whether it is `claude`, `grok`, or other types of AI. Most of AI here are Claude, but some users may use other AI too, so please check carefully.

For AI behavior, describe the meaningful contribution to the user's workflow rather than individual tool calls.

Softwares

Infer the softwares being used in the screen.

Graph structure

Construct a directed graph:

* Nodes represent behavioral activities.
* Edges represent transitions between activities.
* Include loops when the workflow revisits an earlier behavioral activity or repeatedly follows a behavioral cycle.
* Use edge types such as `sequential` or `loop` when appropriate.

Output

Return only valid JSON following the format below:

```
{
  "session_id": "...",
  "participant_id": "...",
  "title": "...",
  "goal": "...",
  "steps": [
   {
      "key": "seg0a",
      "badge": "Step ID (e.g., Step-1, Custom-1, and so on",
      "step_name": "Step name from steps.json",
      "step_action": "Step action from steps.json",
      "time_range": "Step time range from steps.json",
      "start_seconds": 0,
      "end_seconds": 600,
      "nodes": [
      {
          "id": "n1",
          "actor": "user",
          "label": "The behavioral node label",
          "detail": "A brief description of the behavioral activity.",
          "time": ["HH:MM:SS1-HH:MM:SS2", ...],
          "actor_type": "user",
          "softwares": "...",
          "source_steps": "..."
        },
        {
          "id": "n2",
          "actor": "claude or grok or gemini or chatgpt",
          "label": "The behavioral node label for an AI-authored
          action",
          "detail": "A brief description of the behavioral activity.",
          "time": ["HH:MM:SS1-HH:MM:SS2", ...],
          "actor_type": "ai",
          "softwares": "...",
          "source_steps": "..."
        }
```

```
      ],
      "edges": [
        {
          "from": "node_id",
          "to": "node_id",
          "type": "sequential"
        }
      ]
    }
  ]
}
```

Note on `actor` vs. `actor_type`: `actor` is the specific actor performing the node — `user`, or the AI's actual name (`claude`, `grok`, `chatgpt`, `gemini`, etc.). `actor_type` is always just the coarse category — `user` or `ai` — never an AI brand name. Do not swap these: an AI node must have `actor_type: "ai"` (with the brand name in `actor`), not `actor_type: "claude"`.

For each node:

* `label`: the behavioral node label
* `detail`: brief contextual description of the activity. Only include 2–3 short sentences!
* `time`: [the start and ending timestamps of that behavior]
* `source_steps`: workflow step(s) supporting the node

Before returning the result, verify that every node represents a meaningful goal-directed activity rather than a low-level interface action, and that the graph accurately reflects the workflow's behavioral sequence and iterations.

You must watch the video to correctly create the graph! Don't rely on the .srt transcript since it can be wrong! Save the video frames for later use.

Double-check if there are some loops in the graph. If there are loops, make sure that it is presented in the JSON file correctly. For example, if the loop is node A $\rightarrow$ node B $\rightarrow$ node A, there should be two edges (from node A to node B and node B back to node A). A loop can also be A $\rightarrow$ B $\rightarrow$ C $\rightarrow$ A $\rightarrow$ D, so please check carefully.

A.5.1 How Humans Construct the Graph

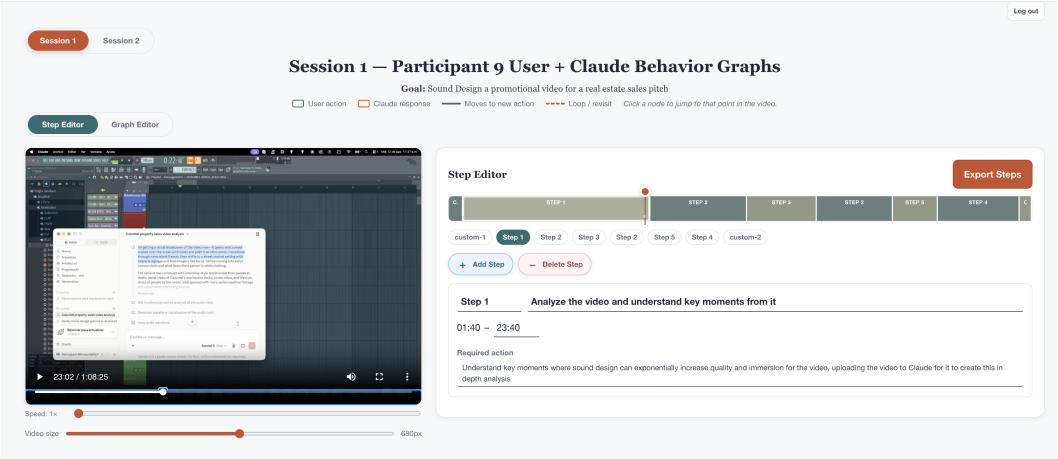
While an LLM was used to generate initial workflow graphs from the video, think-aloud transcript, workflow document, step mappings, and AI chat logs collected from professionals, four members of the research team manually reviewed and substantially reconstructed the graphs, often effectively from scratch.

Table 10: Claude's raw behavior graph vs. the final human-reviewed ground truth.

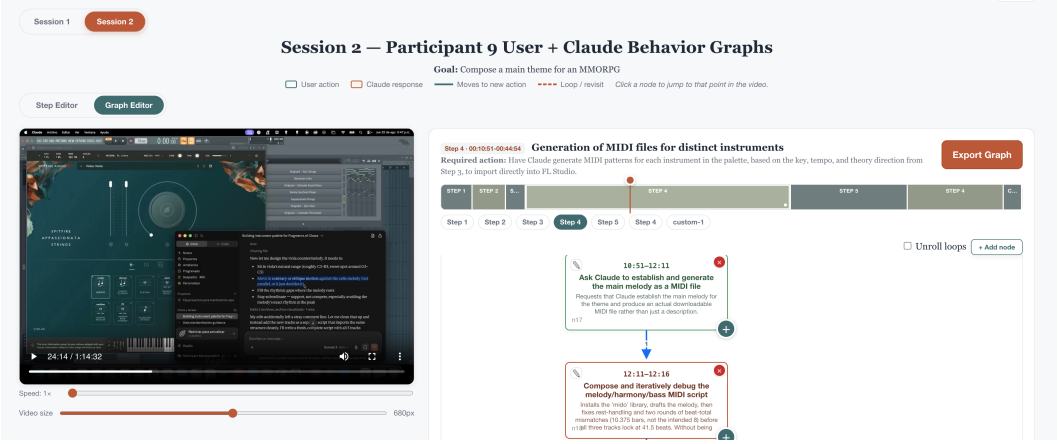
Metric	Claude (raw)	Human (final)
Nodes, total	2,153	2,820
Nodes, mean / session	36.5	47.8
Edges, total	1,810	2,728
Edges, mean / session	30.7	46.2
Loop (revisit) edges, total	11	58
Branch nodes (out-degree >1), mean / session	0.17	3.76
Merge nodes (in-degree >1), mean / session	0.15	3.49
Max out-degree, corpus-wide	2	10
AI$\rightarrow$AI edges (no user turn between), total	61	0

A.6 Prompts for Evaluating MLLMs

We run each experiment under seven input conditions, one for every non-empty combination of video, workflow graph, and chat. The combined conditions reuse the parts of the single-source prompts, so we show only the video-only, graph-only, and chat-only prompts.



(a) The interface for professionals to align steps and timestamps.



(b) The interface for professionals to verify and revise workflow.

Figure 14: A webpage for professionals to align timestamps with steps and verify the generated graphs. Professionals first verify workflow steps and their corresponding video timestamps (top), and subsequently inspect the workflow graph for each step (bottom).

A.6.1 Experiment 1a: Adoption Behavior Prediction Prompt (with target timestamp)

Experiment 1a: video-only

You are given a video snippet of {PROFESSION_PHRASE} working with an AI agent toward the following goal:

Goal: {GOAL}

The workflow consists of multiple steps that contribute to achieving this goal. The AI may produce an output in one step, and the user may act on that output either within the same step or in a later step.

AI Action Step: {AI_ACTION_STEP}
User Action Step: {USER_ACTION_STEP}

Target AI Action Timestamp: {START_TIME}–{END_TIME}

You must predict how the user will respond to the output of the target AI action:

* ADOPT: Use the AI output as provided, without any modification.
* MODIFY: Use the AI output but modify it in any way, including minor edits, formatting changes, or other adjustments.
* REJECT/BYPASS: Do not use the AI output; reject, ignore, replace, or proceed without it.

Use the workflow context available to make your prediction.

Output exactly one label:
`ADOPT`, `MODIFY`, or `REJECT/BYPASS`

Experiment 1a: graph-only

You are given the workflow graph of {PROFESSION_PHRASE} working with an AI agent toward the following goal (no video is provided for this task – rely only on the workflow graph below):

Goal: {GOAL}

The workflow consists of multiple steps that contribute to achieving this goal. The AI may produce an output in one step, and the user may act on that output either within the same step or in a later step.

AI Action Step: {AI_ACTION_STEP}
User Action Step: {USER_ACTION_STEP}

Workflow graph for this portion of the workflow is defined as follows (each node is one action taken by the human or the agent; edges show the order they happened in):

`[n0]` ({actor}) {label}: {details}
`[n1]` ({actor}) {label}: {details}
...

A directed edge X → Y means that action Y follows from or depends on action X. The graph is directed and may contain branches.

Edges (order of actions):
n0 -> n1
n1 -> n2
...

You must predict how the user will respond to the output of the target AI action (the node above matching the "AI Action Step" described earlier):

* ADOPT: Use the AI output as provided, without any modification.
* MODIFY: Use the AI output but modify it in any way, including minor edits, formatting changes, or other adjustments.
* REJECT/BYPASS: Do not use the AI output; reject, ignore, replace, or proceed without it.

Use the workflow context available to make your prediction.

Output exactly one label:
`ADOPT`, `MODIFY`, or `REJECT/BYPASS`

Experiment 1a: chat-only

You are given the chat of {PROFESSION_PHRASE} working with an AI agent toward the following goal (no video is provided for this task – rely only on the chat below):

Goal: {GOAL}

The workflow consists of multiple steps that contribute to achieving this goal. The AI may produce an output in one step, and the user may act on that output either within the same step or in a later step.

AI Action Step: {AI_ACTION_STEP}
User Action Step: {USER_ACTION_STEP}

Chat between the human and the AI agent during this portion of the workflow, in chronological order (each message from the human is wrapped in `<human>` tags, each message from the AI agent in `<ai>` tags):

<human>
{PROMPT}
</human>

<ai>
{AI_RESPONSE}
</ai>

{...ADDITIONAL TURNS...}

You must predict how the user will respond to the output of the target AI action:

* ADOPT: Use the AI output as provided, without any modification.
* MODIFY: Use the AI output but modify it in any way, including minor edits, formatting changes, or other adjustments.
* REJECT/BYPASS: Do not use the AI output; reject, ignore, replace, or proceed without it.

Use the workflow context available to make your prediction.

Output exactly one label:
`ADOPT`, `MODIFY`, or `REJECT/BYPASS`

A.6.2 Experiment 1b & 1c: Adoption Behavior Forecasting Prompt (clip trimmed to AI action)

Experiment 1b & 1c: video-only

You are given a video snippet of {PROFESSION_PHRASE} working with an AI agent toward the following goal:

Goal: {GOAL}

The workflow consists of multiple steps that contribute to achieving this goal. The AI may produce an output in one step, and the user may act on that output either within the same step or in a later step.

AI Action Step: {AI_ACTION_STEP}

You must predict how the user will respond to the output of this AI action:

* ADOPT: Use the AI output as provided, without any modification.
* MODIFY: Use the AI output but modify it in any way, including minor edits, formatting changes, or other adjustments.
* REJECT/BYPASS: Do not use the AI output; reject, ignore, replace, or proceed without it.

Use the workflow context available to make your prediction.

Output exactly one label:
`ADOPT`, `MODIFY`, or `REJECT/BYPASS`

Experiment 1b & 1c: graph-only

You are given the workflow graph of {PROFESSION_PHRASE} working with an AI agent toward the following goal (no video is provided for this task – rely only on the workflow graph below):

Goal: {GOAL}

The workflow consists of multiple steps that contribute to achieving this goal. The AI may produce an output in one step, and the user may act on that output either within the same step or in a later step.

AI Action Step: {AI_ACTION_STEP}

Workflow graph for this portion of the workflow is defined as follows (each node is one action taken by the human or the agent; edges show the order they happened in):

`[n0]` ({actor}) {label}: {details}
`[n1]` ({actor}) {label}: {details}
...

A directed edge X → Y means that action Y follows from or depends on action X. The graph is directed and may contain branches.

Edges (order of actions):
n0 -> n1
n1 -> n2
...

You must predict how the user will respond to the output of the target AI action (the node above matching the "AI Action Step" described earlier):

* ADOPT: Use the AI output as provided, without any modification.
* MODIFY: Use the AI output but modify it in any way, including minor edits, formatting changes, or other adjustments.
* REJECT/BYPASS: Do not use the AI output; reject, ignore, replace, or proceed without it.

Use the workflow context available to make your prediction.

Output exactly one label:
`ADOPT`, `MODIFY`, or `REJECT/BYPASS`

Experiment 1b & 1c: chat-only

You are given the chat of {PROFESSION_PHRASE} working with an AI agent toward the following goal (no video is provided for this task – rely only on the chat below):

Goal: {GOAL}

The workflow consists of multiple steps that contribute to achieving this goal. The AI may produce an output in one step, and the user may act on that output either within the same step or in a later step.

AI Action Step: {AI_ACTION_STEP}

Chat between the human and the AI agent during this portion of the workflow, in chronological order (each message from the human is wrapped in `<human>` tags, each message from the AI agent in `<ai>` tags):

<human>
{PROMPT}
</human>

<ai>
{AI_RESPONSE}
</ai>

{...ADDITIONAL TURNS...}

You must predict how the user will respond to the output of the target AI action:

* ADOPT: Use the AI output as provided, without any modification.
* MODIFY: Use the AI output but modify it in any way, including minor edits, formatting changes, or other adjustments.
* REJECT/BYPASS: Do not use the AI output; reject, ignore, replace, or proceed without it.

Use the workflow context available to make your prediction.

Output exactly one label:
`ADOPT`, `MODIFY`, or `REJECT/BYPASS`

A.6.3 Experiment 2a: Collaboration Rating Prediction Prompt

Experiment 2a: video-only

You are given a video snippet of {PROFESSION_PHRASE} working with an AI agent toward the following goal:

Goal: {GOAL}

The video shows a step of a multi-step workflow for accomplishing the goal:

Step: {STEP_NAME}

Watch how the human and the AI agent interact throughout this step. Examine whether the agent understood the task, context, and instructions; whether it adapted appropriately when given feedback or corrections; whether it reduced the human's effort, time, or cognitive load; whether it proactively suggested useful ideas or next steps; whether its output was accurate, relevant, and trustworthy with minimal correction needed; whether the interaction felt smooth, with the agent contributing at the right moments without creating unnecessary work; and whether the human was able to guide, redirect, or override the agent's actions when desired.

Based on this, rate the overall collaboration experience for this step on a scale from 1 to 5, where 1 means that the user is not satisfied with the collaboration experience with the AI agent and 5 means that the user is satisfied with the collaboration experience with the AI agent.

Output exactly one integer, with no other text:
1, 2, 3, 4, or 5

Experiment 2a: graph-only

You are given a video snippet of {PROFESSION_PHRASE} working with an AI agent toward the following goal:

Goal: {GOAL}

The video shows a step of a multi-step workflow for accomplishing the goal:

Step: {STEP_NAME}

The workflow consists of multiple steps that contribute to achieving this goal. The AI may produce an output in one step, and the user may act on that output either within the same step or in a later step.

Workflow graph for this portion of the workflow is defined as follows (each node is one action taken by the human or the agent; edges show the order they happened in):

`[n0]` ({actor}) {label}: {details}
`[n1]` ({actor}) {label}: {details}
...

A directed edge X → Y means that action Y follows from or depends on action X. The graph is directed and may contain branches.

Edges (order of actions):
n0 -> n1
n1 -> n2
...

Read the graph above. Examine whether the agent understood the task, context, and instructions; whether it adapted appropriately when given feedback or corrections; whether it reduced the human's effort, time, or cognitive load; whether it proactively suggested useful ideas or next steps; whether its output was accurate, relevant, and trustworthy with minimal correction needed; whether the interaction felt smooth, with the agent contributing at the right moments without creating unnecessary work; and whether the human was able to guide, redirect, or override the agent's actions when desired.

Based on this, rate the overall collaboration experience for this step on a scale from 1 to 5, where 1 means that the user is not satisfied with the collaboration experience with the AI agent and 5 means that the user is satisfied with the collaboration experience with the AI agent.

Output exactly one integer, with no other text:
1, 2, 3, 4, or 5

Experiment 2a: chat-only

You are given a video snippet of {PROFESSION_PHRASE} working with an AI agent toward the following goal:

Goal: {GOAL}

The video shows a step of a multi-step workflow for accomplishing the goal:

Step: {STEP_NAME}

Chat between the human and the AI agent during this portion of the workflow, in chronological order (each message from the human is wrapped in `<human>` tags, each message from the AI agent in `<ai>` tags):

<human>
{PROMPT}
</human>

<ai>
{AI_RESPONSE}
</ai>

{...ADDITIONAL TURNS...}

Read the chat above. Examine whether the agent understood the task, context, and instructions; whether it adapted appropriately when given feedback or corrections; whether it reduced the human's effort, time, or cognitive load; whether it proactively suggested useful ideas or next steps; whether its output was accurate, relevant, and trustworthy with minimal correction needed; whether the interaction felt smooth, with the agent contributing at the right moments without creating unnecessary work; and whether the human was able to guide, redirect, or override the agent's actions when desired.

Based on this, rate the overall collaboration experience for this step on a scale from 1 to 5, where 1 means that the user is not satisfied with the collaboration experience with the AI agent and 5 means that the user is satisfied with the collaboration experience with the AI agent.

Output exactly one integer, with no other text:
1, 2, 3, 4, or 5

A.6.4 Experiment 2b: Output Quality Rating Prediction Prompt

Experiment 2b: video-only

You are given a video snippet of {PROFESSION_PHRASE} working with an AI agent toward the following goal:

Goal: {GOAL}

The video shows a step of a multi-step workflow for accomplishing the goal:

Step: {STEP_NAME}

Watch the output the AI agent produced and how it was used during this step. Examine whether the output was accurate, relevant, and directly usable for the task; whether it required significant correction, rework, or was discarded; and whether it met the specific requirements of this step.

Based on this, rate the output quality for this step on a scale from 1 to 5, where 1 means that the user is not satisfied with the output quality and 5 means that the user is satisfied with the output quality.

Output exactly one integer, with no other text:
1, 2, 3, 4, or 5

Experiment 2b: graph-only

You are given a video snippet of {PROFESSION_PHRASE} working with an AI agent toward the following goal:

Goal: {GOAL}

The video shows a step of a multi-step workflow for accomplishing the goal:

Step: {STEP_NAME}

Workflow graph for this portion of the workflow is defined as follows (each node is one action taken by the human or the agent; edges show the order they happened in):

`[n0]` ({actor}) {label}: {details}
`[n1]` ({actor}) {label}: {details}
...

A directed edge X → Y means that action Y follows from or depends on action X. The graph is directed and may contain branches.

Edges (order of actions):
n0 -> n1
n1 -> n2
...

Read the graph above, focusing on the output the AI agent produced and how it was used during this step. Examine whether the output was accurate, relevant, and directly usable for the task; whether it required significant correction, rework, or was discarded; and whether it met the specific requirements of this step.

Based on this, rate the output quality for this step on a scale from 1 to 5, where 1 means that the user is not satisfied with the output quality and 5 means that the user is satisfied with the output quality.

Output exactly one integer, with no other text:
1, 2, 3, 4, or 5

Experiment 2b: chat-only

You are given a video snippet of {PROFESSION_PHRASE} working with an AI agent toward the following goal:

Goal: {GOAL}

The video shows a step of a multi-step workflow for accomplishing the goal:

Step: {STEP_NAME}

Chat between the human and the AI agent during this portion of the workflow, in chronological order (each message from the human is wrapped in `<human>` tags, each message from the AI agent in `<ai>` tags):

<human>
{PROMPT}
</human>

<ai>
{AI_RESPONSE}
</ai>

{...ADDITIONAL TURNS...}

Read the chat above, focusing on the output the AI agent produced and how it was used during this step. Examine whether the output was accurate, relevant, and directly usable for the task; whether it required significant correction, rework, or was discarded; and whether it met the specific requirements of this step.

Based on this, rate the output quality for this step on a scale from 1 to 5, where 1 means that the user is not satisfied with the output quality and 5 means that the user is satisfied with the output quality.

Output exactly one integer, with no other text:
1, 2, 3, 4, or 5

A.7 Model Selection

For proprietary models, we evaluate leading video-capable multimodal models from major providers: **Gemini 3.1 Pro** and **Gemini 3.8 Flash** (Google), **GPT-6 Astra** (OpenAI), and **Claude Opus 5.5**. Gemini natively accepts video as an input. GPT-6 Astra has no native video modality, so we sample evenly-spaced frames from each clip as its input; this also lets GPT-6 Astra serve as a frame-sampled point of comparison against Gemini's genuinely video-native processing. We additionally evaluate Claude despite it likewise lacking native video input (also requiring frame sampling), because Claude is the specific agent our study participants interacted with during their recorded sessions, so evaluating it tests whether the very model embedded in the human-AI collaboration can itself judge or predict the outcome of that collaboration.

For open-weight models, we evaluate two representative video-capable Qwen3-VL variants (Bai et al., 2025): **Qwen3-VL-30B-A3B-Instruct**, a mixture-of-experts model (30B total parameters, 3B active per token) included to test whether MoE-style efficiency scaling preserves video-temporal understanding, and **Qwen3-VL-32B-Instruct**, a dense model of comparable total scale that instead invests its full parameter budget in deeper visual-language fusion at every token. Comparing the two isolates the effect of the MoE-vs-dense architectural choice at a fixed parameter scale.

A.8 Sample Size and Selection

Claude evaluation subsets Claude was queried manually through its web interface, so for Experiments 1a, 2a and 2b we evaluated it on fixed subsets; Experiment 1b uses all instances. For 1a, we sampled 25 instances per label (N = 75), because the labels are imbalanced (13% reject), and a random sample would leave too few rejects for per-class F1. For 2a, we kept every step rated 1–4 and sampled 30 of the steps rated 5 (N = 78), since 5s make up 60% of the data. 2b uses the same

Experiment	# Clips	Mean ± std	Min	Max
Experiment 1a	290	11.9±12.2	0.9	72.3
Experiment 1b	290	6.0±6.8	0.2	40.2
Experiment 1c	54	11.8±6.3	2.3	32.8
Experiment 1d	54	43.0±9.4	26.0	57.4
Experiment 2a	294	14.6±10.9	0.8	66.2
Experiment 2b	305	14.7±10.9	0.8	66.2

Table 11: Length (in minutes) of the video clips given to the models in each experiment.

		Macro			Per-class F1		
Model	**Setting**	**P**	**R**	**F1**	**Adopt**	**Modify**	**Reject**
GPT-6-Astra	video-only	**0.761**	**0.762**	**0.748**	0.794	**0.689**	**0.762**
	chat-only	0.630	0.645	0.637	0.751	0.618	0.541
	video-graph-chat	0.699	0.742	0.673	**0.812**	0.480	0.727
Claude	video-only	**0.729**	**0.667**	**0.672**	**0.644**	**0.654**	0.718
	chat-only	0.682	0.640	0.648	0.621	0.560	**0.762**
Gemini-3.1-Pro-Preview	video-only	**0.747**±0.021	**0.768**±0.012	**0.756**±0.017	0.800±0.023	0.659±0.000	**0.810**±0.027
	graph-only	0.742±0.015	0.735±0.026	0.738±0.021	0.817±0.002	0.655±0.017	0.743±0.081
	chat-only	0.544±0.297	0.563±0.325	0.551±0.309	**0.917**±0.118	0.414±0.586	0.324±0.458
	video-graph	0.743±0.033	0.754±0.033	0.748±0.033	0.816±0.034	0.679±0.039	0.750±0.027
	video-chat	0.706±0.011	0.747±0.012	0.722±0.011	0.817±0.000	**0.691**±0.006	0.657±0.040
	graph-chat	0.722	0.669	0.690	0.778	0.600	0.692
	video-graph-chat	0.654	0.617	0.631	0.766	0.588	0.538
Gemini-3.8-Flash	video-only	0.652±0.007	0.694±0.010	0.662±0.010	0.810±0.009	0.544±0.024	0.632±0.000
	graph-only	**0.780**±0.021	0.717±0.015	**0.741**±0.017	0.832±0.005	**0.696**±0.006	0.694±0.039
	chat-only	0.743±0.000	0.742±0.004	0.735±0.001	0.832±0.012	0.666±0.044	0.707±0.028
	video-graph	0.733±0.035	0.708±0.039	0.716±0.037	**0.844**±0.009	0.642±0.036	0.663±0.080
	video-chat	0.729±0.001	**0.757**±0.003	0.737±0.002	0.805±0.002	0.642±0.008	**0.765**±0.000
	graph-chat	0.736±0.037	0.708±0.014	0.718±0.024	0.819±0.022	0.612±0.063	0.724±0.013
	video-graph-chat	0.723±0.064	0.696±0.054	0.708±0.059	0.819±0.015	0.671±0.047	0.633±0.115
Qwen3-VL-30B-A3B-Instruct	video-only	0.581±0.012	0.672±0.016	0.596±0.010	0.714±0.010	0.500±0.020	0.575±0.004
	graph-only	0.620±0.028	0.665±0.023	0.598±0.028	0.771±0.019	**0.543**±0.059	0.479±0.009
	chat-only	0.606±0.030	0.612±0.017	0.567±0.017	0.760±0.000	0.469±0.030	0.473±0.032
	video-graph	**0.649**±0.003	**0.698**±0.004	**0.637**±0.004	**0.820**±0.011	0.511±0.025	0.580±0.000
	video-chat	0.642±0.034	0.635±0.038	0.613±0.031	0.786±0.017	0.462±0.030	**0.593**±0.051
	graph-chat	0.618±0.020	0.635±0.017	0.581±0.013	0.807±0.016	0.461±0.023	0.476±0.014
	video-graph-chat	0.621±0.004	0.654±0.008	0.595±0.005	0.812±0.012	0.439±0.023	0.535±0.030
Qwen3-VL-32B-Instruct	video-only	**0.621**±0.041	0.592±0.009	0.432±0.033	0.159±0.021	0.539±0.076	0.597±0.040
	graph-only	0.618±0.006	0.627±0.008	0.560±0.004	0.491±0.004	0.523±0.002	**0.667**±0.019
	chat-only	0.526±0.018	0.556±0.007	0.461±0.008	0.367±0.013	0.509±0.004	0.506±0.008
	video-graph	0.619±0.005	**0.670**±0.010	**0.613**±0.010	**0.658**±0.012	**0.563**±0.019	0.618±0.000
	video-chat	0.527±0.028	0.583±0.022	0.488±0.024	0.426±0.020	0.506±0.004	0.533±0.046
	graph-chat	0.569±0.009	0.608±0.009	0.564±0.007	0.635±0.005	0.533±0.015	0.524±0.000
	video-graph-chat	0.537±0.000	0.599±0.002	0.541±0.001	0.623±0.006	0.504±0.006	0.495±0.007

Table 12: Result for Experiment 1a. Mean±std over runs where a model has several; bold marks the best score within each model.

steps, except six with no output-quality rating (N = 72). All subsets were drawn with a fixed seed, spread across participants

A.9 Experiment Results

Results for Experiment 1a (recognizing the user's response to the AI's action) are shown in Table 12, and results for Experiment 1b (forecasting the user's response) are shown in Table 13. Results for Experiments 2a and 2b are given in Table 14 and Table 15, respectively.

A.10 Output Quality × Collaboration Rating

As shown in Figure 15, output quality and the collaboration experience perceived by professionals are correlated.

Model	Setting	Macro			Per-class F1		
		P	R	F1	Adopt	Modify	Reject
GPT-6-Astra	video-only	0.580±0.027	**0.558**±0.028	0.558±0.023	0.710±0.026	0.577±0.043	**0.387**±0.054
	graph-only	0.519	0.442	0.456	**0.746**	0.358	0.263
	chat-only	0.398	0.396	0.397	0.672	0.372	0.146
	video-graph	**0.596**	0.551	**0.560**	0.706	**0.587**	**0.387**
	video-graph-chat	0.561	0.545	0.549	0.694	0.568	0.385
Claude	video-only	**0.635**	**0.544**	**0.562**	**0.750**	**0.542**	**0.392**
	chat-only	0.411	0.419	0.413	0.653	0.476	0.111
Gemini-3.1-Pro-Preview	video-only	0.529±0.079	0.434±0.020	0.437±0.028	0.696±0.020	0.447±0.003	0.168±0.069
	graph-only	0.505±0.032	0.469±0.031	0.477±0.034	0.750±0.014	0.492±0.060	0.190±0.035
	chat-only	0.423±0.021	0.407±0.022	0.410±0.022	0.676±0.007	0.380±0.011	0.173±0.048
	video-graph	0.497±0.010	0.459±0.004	0.467±0.005	0.743±0.004	0.492±0.022	0.167±0.038
	video-chat	0.444±0.018	0.417±0.013	0.423±0.015	0.691±0.010	0.404±0.031	0.174±0.014
	graph-chat	0.472±0.013	0.436±0.008	0.442±0.010	0.727±0.009	0.454±0.021	0.146±0.031
	video-graph-chat	**0.534**±0.041	**0.475**±0.027	**0.488**±0.031	**0.755**±0.018	**0.495**±0.046	**0.213**±0.041
Gemini-3.8-Flash	video-only	0.491±0.031	0.418±0.012	0.420±0.014	0.734±0.008	0.317±0.014	0.208±0.023
	graph-only	0.487±0.010	0.415±0.005	0.418±0.007	**0.775**±0.008	**0.398**±0.052	0.080±0.038
	chat-only	0.420±0.043	0.393±0.036	0.375±0.040	0.724±0.007	0.202±0.055	0.199±0.058
	video-graph	0.487±0.015	0.422±0.015	0.425±0.017	0.752±0.001	0.345±0.023	0.178±0.039
	video-chat	**0.519**±0.037	**0.436**±0.025	**0.441**±0.027	0.741±0.013	0.353±0.020	**0.228**±0.061
	graph-chat	0.496±0.024	0.425±0.012	0.430±0.014	0.757±0.007	0.385±0.021	0.149±0.037
	video-graph-chat	0.494±0.020	0.417±0.009	0.419±0.011	0.765±0.005	0.372±0.007	0.120±0.030
Qwen3-VL-30B-A3B-Instruct	video-only	0.405±0.014	0.412±0.024	0.394±0.022	0.544±0.029	**0.359**±0.029	0.278±0.019
	graph-only	0.514±0.030	**0.473**±0.020	**0.485**±0.022	0.719±0.014	**0.359**±0.053	**0.377**±0.071
	chat-only	0.504±0.027	0.387±0.002	0.374±0.004	0.698±0.004	0.245±0.038	0.179±0.024
	video-graph	**0.535**±0.029	0.433±0.020	0.444±0.026	0.743±0.002	0.328±0.033	0.262±0.046
	video-chat	0.404±0.014	0.356±0.005	0.327±0.009	0.726±0.005	0.131±0.029	0.123±0.010
	graph-chat	0.526±0.023	0.439±0.009	0.435±0.009	**0.746**±0.011	0.261±0.023	0.300±0.009
	video-graph-chat	0.477±0.004	0.398±0.008	0.382±0.010	0.740±0.004	0.189±0.005	0.218±0.036
Qwen3-VL-32B-Instruct	video-only	0.441±0.042	0.382±0.016	0.245±0.012	0.064±0.033	0.447±0.010	0.225±0.022
	graph-only	0.626±0.018	**0.463**±0.013	0.415±0.018	0.517±0.009	0.497±0.007	0.232±0.041
	chat-only	0.362±0.005	0.373±0.004	0.323±0.003	0.374±0.006	0.418±0.004	0.176±0.000
	video-graph	**0.657**±0.074	0.460±0.010	0.419±0.011	0.491±0.059	**0.503**±0.022	**0.265**±0.070
	video-chat	0.401±0.043	0.405±0.022	0.355±0.009	0.407±0.046	0.458±0.044	0.199±0.032
	graph-chat	0.458±0.007	0.435±0.007	0.414±0.007	0.591±0.010	0.465±0.010	0.188±0.003
	video-graph-chat	0.437±0.012	0.436±0.013	**0.421**±0.016	**0.596**±0.011	0.467±0.004	0.199±0.032

Table 13: Result for Experiment 1b. Mean±std over runs where a model has several; bold marks the best score within each model.

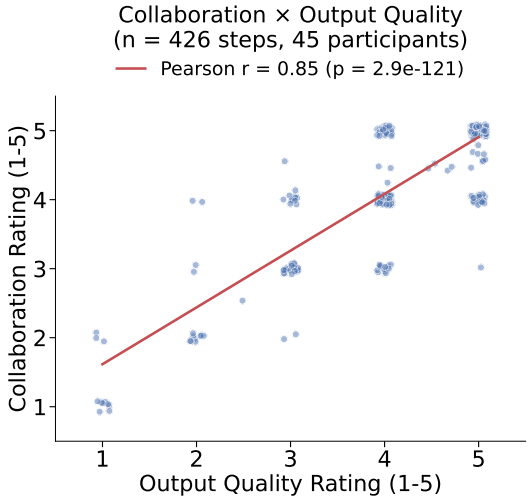


Figure 15: Professionals' ratings of collaboration satisfaction and output quality across all steps.

		Overall				MAE by true rating			
Model	**Setting**	**MAE↓**	**Exact acc.**	**Within-1**	**Spearman ρ**	**Low 1–2**	**Moderate 3**	**Mod-high 4**	**High 5**
GPT-6-Astra	video-only	0.960±0.093	0.244±0.037	0.836±0.018	**0.332**±0.078	1.755±0.425	0.714±0.081	0.480±0.106	1.039±0.037
	graph-only	0.716	0.396	0.880	0.295	1.367	0.923	0.544	0.711
	chat-only	1.302	0.168	0.602	0.105	1.583	0.609	0.819	1.613
	video-graph	**0.697**	**0.399**	**0.890**	0.309	1.389	0.733	0.455	0.760
	video-graph-chat	0.810	0.321	0.863	0.260	1.500	0.880	0.398	0.917
Claude	video-only	**0.705**	**0.385**	**0.923**	**0.407**	1.833	0.786	0.250	0.867
	chat-only	0.846	0.359	0.821	0.169	2.167	0.643	0.357	1.133
Gemini-3.1-Pro-Preview	video-only	0.697±0.006	0.530±0.002	0.843±0.005	0.060±0.030	3.231±0.032	1.963±0.037	0.952±0.011	0.054±0.004
	graph-only	0.654±0.008	0.503±0.002	**0.859**±0.002	0.263±0.017	1.731±0.116	1.753±0.021	0.908±0.011	0.204±0.010
	chat-only	1.085±0.105	0.493±0.057	0.733±0.024	0.027±0.065	3.067±0.521	1.679±0.160	1.208±0.166	0.707±0.109
	video-graph	0.667±0.020	0.501±0.011	0.855±0.009	0.235±0.031	1.824±0.064	1.840±0.043	0.982±0.023	0.162±0.034
	video-chat	0.790±0.026	0.539±0.016	0.796±0.005	0.180±0.047	2.767±0.058	1.813±0.031	0.961±0.123	0.309±0.067
	graph-chat	**0.589**±0.027	**0.630**±0.091	0.853±0.000	**0.342**±0.040	2.044±0.252	1.487±0.422	0.920±0.105	0.125±0.109
	video-graph-chat	0.623±0.011	0.548±0.007	0.856±0.005	0.259±0.004	2.089±0.077	1.782±0.022	0.943±0.018	0.082±0.019
Gemini-3.8-Flash	video-only	0.798±0.022	0.510±0.011	0.816±0.002	0.037±0.050	3.083±0.000	1.877±0.021	1.019±0.022	0.235±0.056
	graph-only	0.614±0.005	0.521±0.005	0.878±0.000	0.274±0.008	1.954±0.032	1.667±0.000	0.871±0.006	0.141±0.004
	chat-only	0.961±0.028	0.463±0.013	0.756±0.012	0.057±0.020	2.733±0.000	1.615±0.038	1.084±0.009	0.577±0.039
	video-graph	0.615±0.006	0.521±0.004	**0.879**±0.004	0.271±0.015	1.991±0.032	1.827±0.043	0.945±0.013	0.071±0.006
	video-chat	0.738±0.053	0.541±0.024	0.826±0.009	0.176±0.037	3.000±0.176	1.756±0.022	0.948±0.031	0.186±0.075
	graph-chat	0.619±0.008	0.540±0.004	0.866±0.004	0.274±0.037	2.200±0.000	1.705±0.022	0.912±0.018	0.092±0.020
	video-graph-chat	**0.611**±0.008	**0.550**±0.002	0.856±0.005	**0.297**±0.029	2.178±0.077	1.795±0.022	0.943±0.009	0.047±0.000
Qwen3-VL-30B-A3B-Instruct	video-only	0.912±0.015	0.324±0.004	0.808±0.009	0.079±0.026	2.478±0.038	1.000±0.072	0.650±0.050	0.889±0.014
	graph-only	0.782±0.013	0.288±0.012	**0.923**±0.004	0.197±0.031	1.900±0.067	1.028±0.048	0.317±0.018	0.886±0.019
	chat-only	0.848±0.028	0.445±0.024	0.833±0.003	0.179±0.044	2.500±0.083	1.449±0.066	0.758±0.017	0.633±0.031
	video-graph	**0.614**±0.008	0.489±0.007	0.903±0.004	0.275±0.005	2.033±0.176	1.403±0.064	0.592±0.074	0.372±0.030
	video-chat	0.728±0.023	0.531±0.012	0.829±0.013	0.143±0.025	3.083±0.083	1.797±0.050	0.905±0.026	0.249±0.034
	graph-chat	0.691±0.018	0.440±0.014	0.894±0.008	0.262±0.011	1.972±0.127	1.188±0.025	0.605±0.034	0.526±0.044
	video-graph-chat	0.617±0.011	**0.544**±0.004	0.872±0.012	**0.281**±0.019	2.139±0.048	1.580±0.133	0.792±0.074	0.232±0.024
Qwen3-VL-32B-Instruct	video-only	1.266±0.056	0.177±0.033	0.609±0.024	0.082±0.055	1.570±0.163	0.464±0.235	0.734±0.094	1.599±0.014
	graph-only	0.717±0.019	0.394±0.014	0.886±0.008	0.233±0.025	1.713±0.032	1.037±0.064	0.549±0.026	0.646±0.017
	chat-only	0.863±0.017	0.457±0.020	0.812±0.004	0.152±0.019	2.533±0.067	1.564±0.022	0.948±0.031	0.500±0.028
	video-graph	0.685±0.051	0.431±0.005	**0.888**±0.036	0.263±0.002	1.915±0.363	1.167±0.146	0.544±0.107	0.546±0.025
	video-chat	0.754±0.033	0.487±0.024	0.831±0.009	0.155±0.023	2.750±0.051	1.500±0.028	0.858±0.011	0.356±0.043
	graph-chat	0.633±0.009	0.517±0.016	0.872±0.005	**0.270**±0.016	2.333±0.067	1.526±0.059	0.813±0.018	0.184±0.026
	video-graph-chat	**0.620**±0.006	**0.526**±0.006	0.868±0.012	0.236±0.025	2.214±0.000	1.640±0.113	0.818±0.000	0.168±0.011

Table 14: Result for Experiment 2a. Mean±std over runs where a model has several; bold marks the best score within each model. Band columns give the MAE within each true-rating band.

		Overall				MAE by true rating			
Model	**Setting**	**MAE↓**	**Exact acc.**	**Within-1**	**Spearman ρ**	**Low 1–2**	**Moderate 3**	**Mod-high 4**	**High 5**
GPT-6-Astra	video-only	0.938±0.093	0.344±0.053	0.779±0.053	0.299±0.096	1.885±0.198	0.716±0.086	0.702±0.170	0.893±0.082
	graph-only	**0.736**	**0.509**	0.811	**0.385**	2.000	1.000	1.000	0.444
	chat-only	1.195	0.221	0.673	0.159	1.385	0.667	0.852	1.457
	video-graph	0.837	0.429	0.776	0.374	1.833	0.500	0.714	0.719
	video-graph-chat	0.767	0.400	**0.860**	0.247	1.667	0.636	0.565	0.778
Claude	video-only	**0.750**	**0.375**	**0.903**	**0.374**	1.875	0.692	0.174	0.929
	chat-only	1.278	0.236	0.694	-0.013	2.000	1.538	1.043	1.143
Gemini-3.1-Pro-Preview	video-only	0.767±0.025	0.510±0.006	0.821±0.007	0.006±0.079	3.325±0.061	2.000±0.000	1.018±0.006	0.079±0.044
	graph-only	0.754±0.029	0.467±0.015	0.832±0.008	0.277±0.034	1.693±0.080	1.419±0.000	0.951±0.016	0.399±0.039
	chat-only	1.065±0.037	0.446±0.009	0.728±0.003	0.108±0.086	2.438±0.000	1.679±0.000	1.277±0.044	0.628±0.094
	video-graph	0.759±0.021	0.469±0.005	0.836±0.016	0.247±0.021	1.658±0.053	1.538±0.019	1.025±0.022	0.347±0.022
	video-chat	0.845±0.080	**0.515**±0.057	0.791±0.018	0.160±0.056	3.236±0.084	1.726±0.021	1.137±0.056	0.223±0.033
	graph-chat	**0.700**±0.025	0.509±0.012	**0.852**±0.006	**0.288**±0.047	1.844±0.044	1.554±0.025	0.915±0.011	0.236±0.053
	video-graph-chat	0.733±0.023	0.509±0.016	0.830±0.008	0.276±0.019	2.042±0.072	1.655±0.021	0.985±0.015	0.211±0.027
Gemini-3.8-Flash	video-only	0.817±0.030	0.499±0.009	0.808±0.007	-0.006±0.015	3.061±0.213	1.946±0.067	1.028±0.032	0.210±0.055
	graph-only	0.659±0.005	0.513±0.003	0.866±0.003	0.279±0.013	1.816±0.053	1.548±0.056	0.849±0.022	0.239±0.006
	chat-only	1.137±0.012	0.428±0.003	0.689±0.009	0.091±0.025	2.375±0.000	1.768±0.076	1.338±0.022	0.720±0.018
	video-graph	**0.640**±0.013	0.528±0.008	**0.870**±0.004	0.293±0.014	2.009±0.061	1.688±0.049	0.898±0.012	0.125±0.005
	video-chat	0.779±0.028	0.520±0.017	0.811±0.007	0.171±0.037	2.896±0.095	1.803±0.048	1.002±0.028	0.174±0.043
	graph-chat	0.676±0.017	0.525±0.009	0.855±0.007	**0.298**±0.006	2.042±0.036	1.583±0.041	0.933±0.039	0.147±0.008
	video-graph-chat	0.678±0.022	**0.552**±0.038	0.842±0.002	0.273±0.026	2.292±0.180	1.614±0.050	0.910±0.012	0.169±0.037
Qwen3-VL-30B-A3B-Instruct	video-only	1.239±0.037	0.260±0.004	0.656±0.033	-0.015±0.018	2.365±0.095	1.321±0.093	0.938±0.035	1.284±0.048
	graph-only	0.734±0.023	0.409±0.031	**0.874**±0.008	0.261±0.036	1.552±0.036	1.037±0.037	0.465±0.035	0.754±0.023
	chat-only	0.967±0.006	0.442±0.018	0.768±0.015	0.118±0.028	2.769±0.077	1.514±0.064	1.022±0.034	0.632±0.031
	video-graph	0.718±0.012	0.456±0.013	0.859±0.006	0.252±0.045	1.510±0.095	1.296±0.037	0.714±0.038	0.540±0.029
	video-chat	0.919±0.027	0.472±0.003	0.768±0.010	0.085±0.032	2.897±0.320	1.694±0.096	1.137±0.034	0.435±0.036
	graph-chat	**0.670**±0.035	0.501±0.025	**0.874**±0.014	**0.279**±0.013	1.718±0.118	1.361±0.087	0.754±0.057	0.373±0.015
	video-graph-chat	0.709±0.018	**0.519**±0.016	0.846±0.007	0.267±0.032	1.872±0.044	1.681±0.064	0.989±0.034	0.243±0.023
Qwen3-VL-32B-Instruct	video-only	1.341±0.074	0.150±0.039	0.563±0.078	0.101±0.072	1.522±0.106	0.600±0.088	0.851±0.047	1.652±0.027
	graph-only	0.776±0.008	0.392±0.000	0.840±0.009	0.272±0.016	1.605±0.000	0.849±0.037	0.565±0.006	0.789±0.022
	chat-only	0.973±0.015	0.413±0.008	0.775±0.003	0.147±0.025	2.542±0.095	1.286±0.036	1.005±0.047	0.677±0.018
	video-graph	0.744±0.002	0.401±0.005	0.863±0.000	0.265±0.022	1.556±0.039	0.983±0.024	0.575±0.008	0.705±0.009
	video-chat	0.848±0.014	0.429±0.003	0.798±0.016	0.135±0.035	2.633±0.047	1.370±0.105	0.857±0.022	0.502±0.008
	graph-chat	0.697±0.007	0.446±0.011	0.891±0.008	**0.297**±0.010	2.062±0.062	1.107±0.062	0.738±0.031	0.399±0.017
	video-graph-chat	**0.687**±0.003	**0.452**±0.002	**0.894**±0.001	0.248±0.070	2.091±0.129	1.269±0.013	0.789±0.016	0.354±0.017

Table 15: Result for Experiment 2b. Mean±std over runs where a model has several; bold marks the best score within each model. Band columns give the MAE within each true-rating band.

A.11 POST-INTERACTION QUESTIONNAIRE

OVERVIEW: ALL BEHAVIORAL LIKERT ITEMS (Q13–Q38)

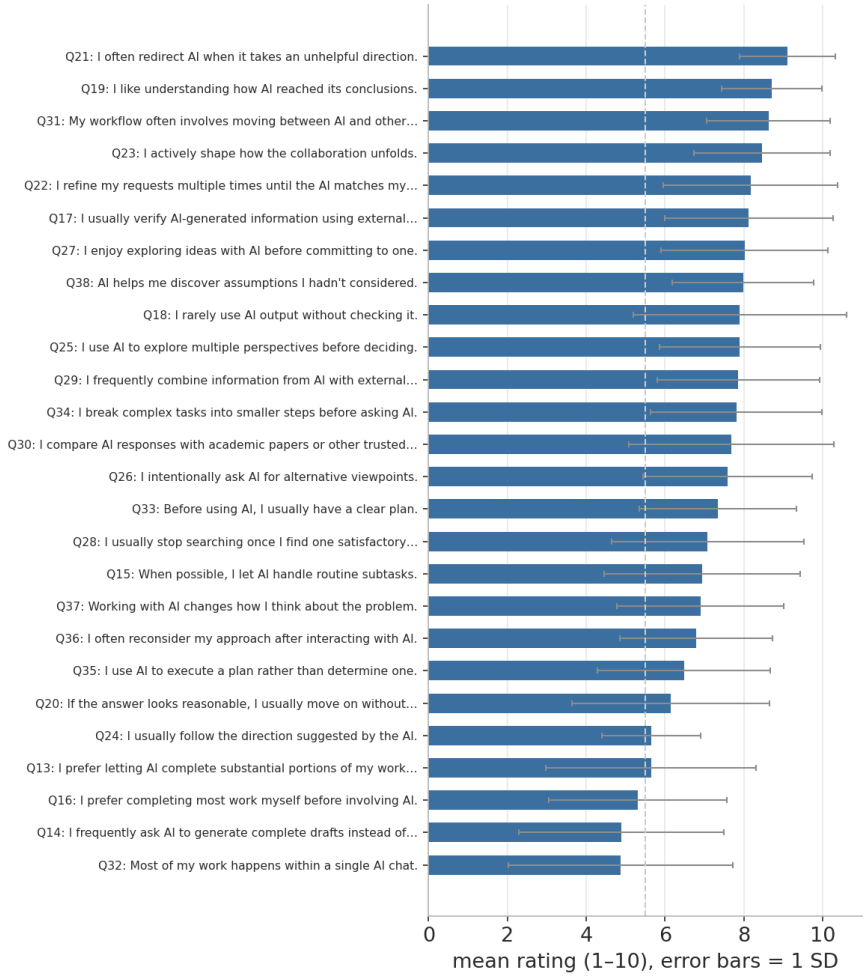
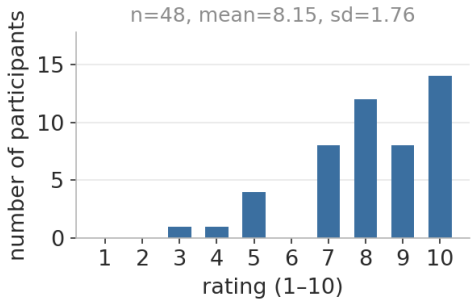


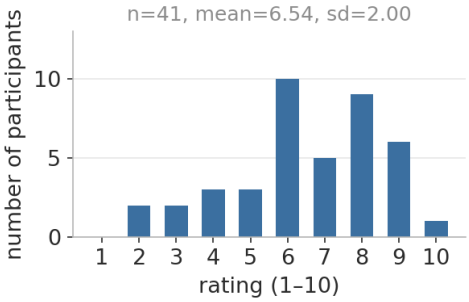
Figure 16: Mean rating (1–10) for each of the 26 AI-collaboration behavior items, sorted low to high (error bars: 1 SD). The dashed line marks the scale midpoint.

SESSION 1 & SESSION 2 REFLECTION

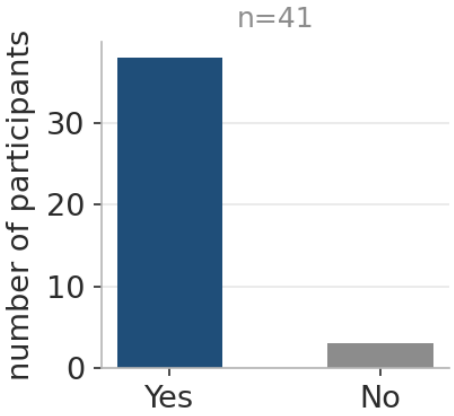


(a) Q1. After watching the session 1 recording again and looking at the extracted workflow, how natural do you think your workflow there compared to your general workflow?

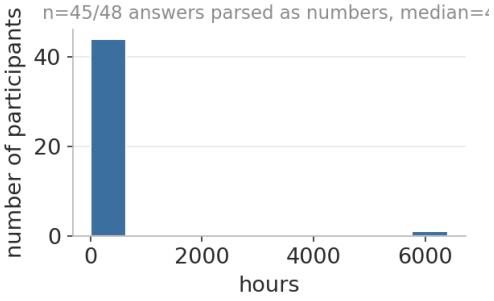
Q2. If you answered not natural at all in the previous question, why do you think so? *Open-ended response*



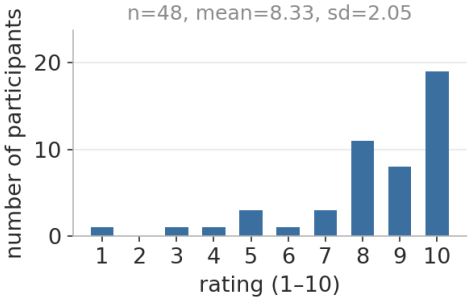
(a) Q3. How difficult was the problem you worked on in Session 1?



(b) Q4. Have you ever done problem 1 manually without agentic AI's help?

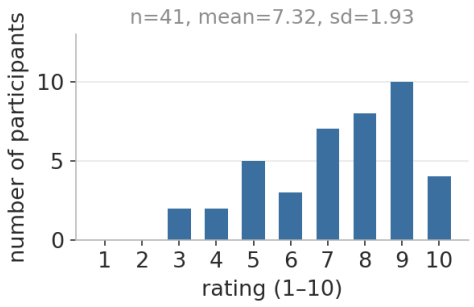


(a) Q5. How many hours on average did it take or do you think it takes for you to finish the problem for session 1 without agentic AI's help?

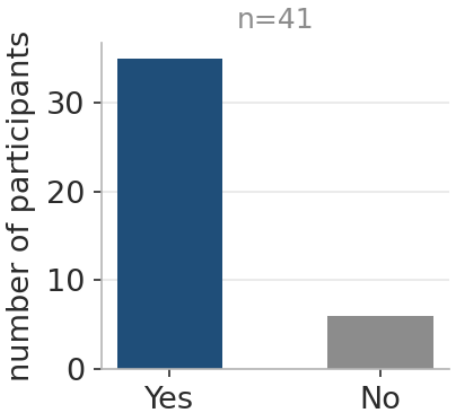


(b) Q6. After watching the session 2 recording again and looking at the extracted workflow, how natural do you think your workflow there compared to your general workflow?

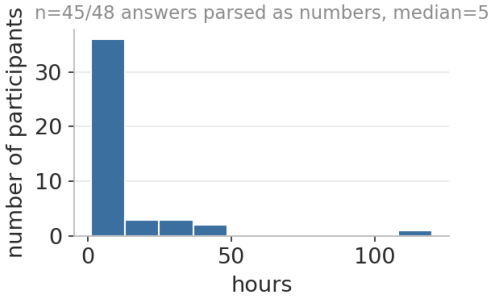
Q7. If you answered not natural at all in the previous question, why do you think so? *Open-ended response*



(a) Q8. How difficult was the problem you worked on in Session 2?



(b) Q9. Have you ever done problem 2 manually without agentic AI's help?

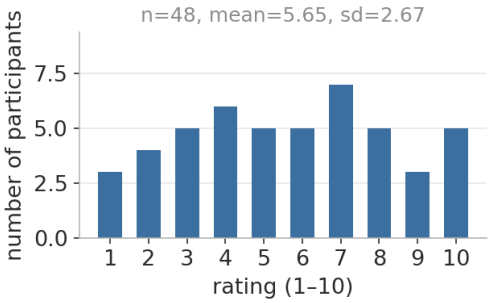


(a) Q10. How many hours on average did it take or do you think it takes for you to finish the problem for session 2 without agentic AI's help?

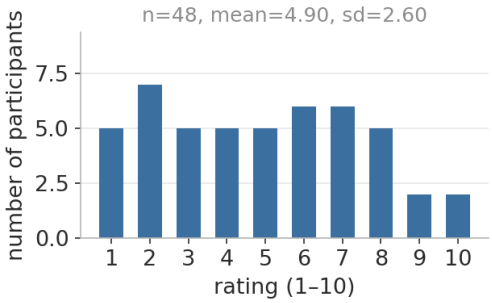
Q11. For either or both session 1 and 2, were there any points where you changed your plan or workflow? What prompted that? *Open-ended response*

Q12. If you did a similar task again, would you approach it differently? If so, what would you change? *Open-ended response.*

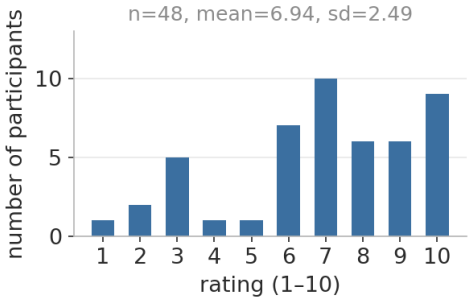
AI Reliance / Delegation



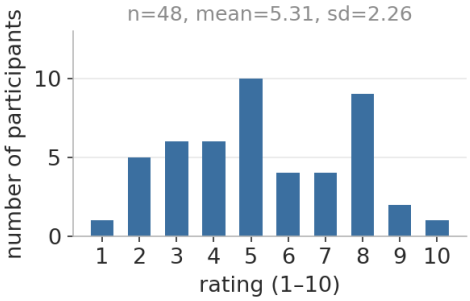
(a) Q13. I prefer letting AI complete substantial portions of my work before reviewing it.



(b) Q14. I frequently ask AI to generate complete drafts instead of working incrementally.

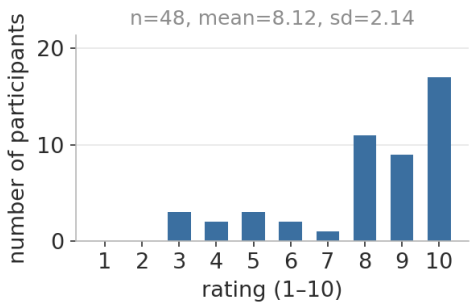


(a) Q15. When possible, I let AI handle routine subtasks.

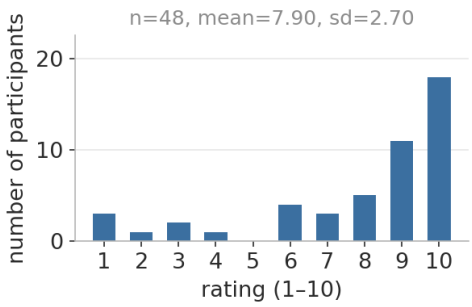


(b) Q16. I prefer completing most work myself before involving AI.

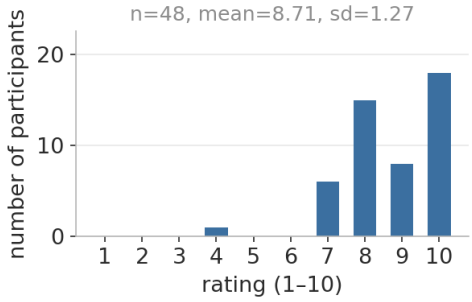
VERIFICATION BEHAVIOR



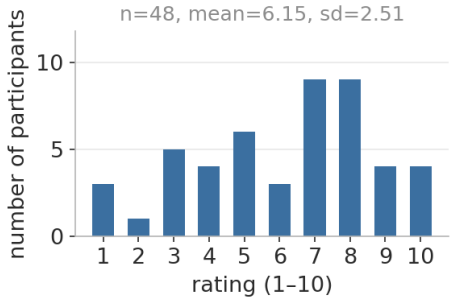
(a) Q17. I usually verify AI-generated information using external sources.



(b) Q18. I rarely use AI output without checking it.

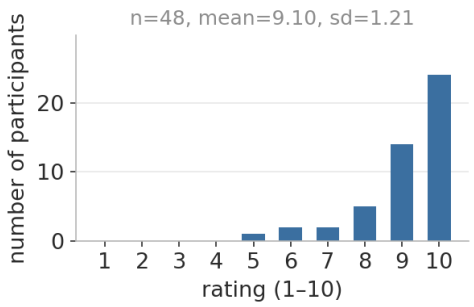


(a) Q19. I like understanding how AI reached its conclusions.

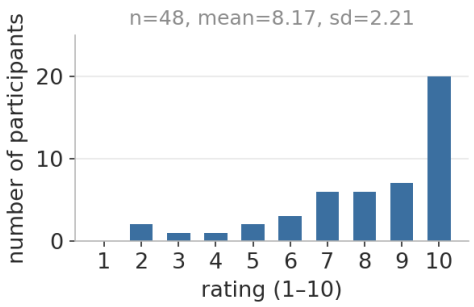


(b) Q20. If the answer looks reasonable, I usually move on without further verification.

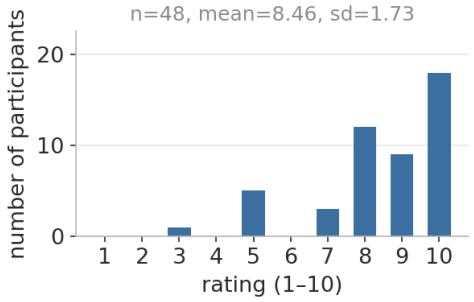
CONTROL & REDIRECTION



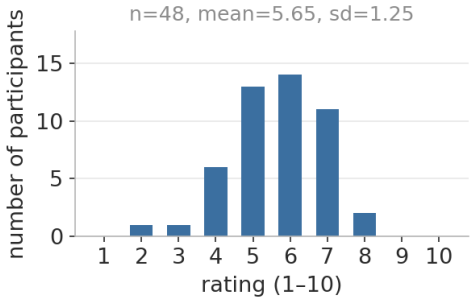
(a) Q21. I often redirect AI when it takes an unhelpful direction.



(b) Q22. I refine my requests multiple times until the AI matches my intent.

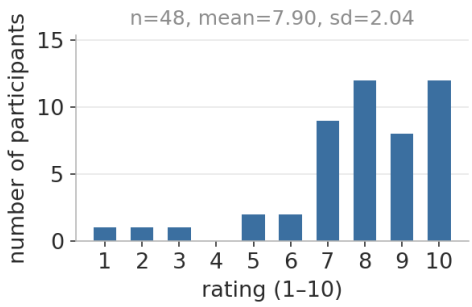


(a) Q23. I actively shape how the collaboration unfolds.

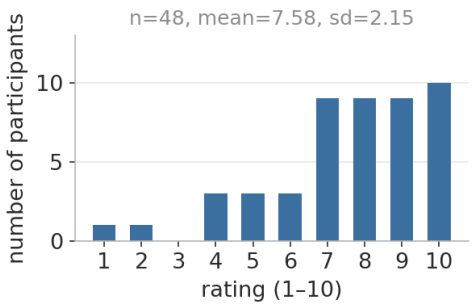


(b) Q24. I usually follow the direction suggested by the AI.

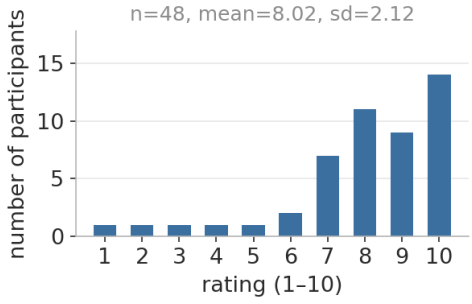
EXPLORATION OF ALTERNATIVES



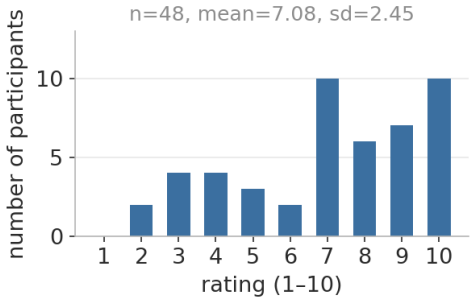
(a) Q25. I use AI to explore multiple perspectives before deciding.



(b) Q26. I intentionally ask AI for alternative viewpoints.

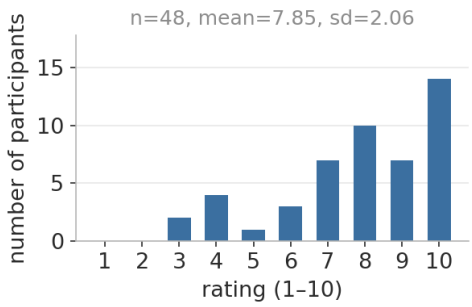


(a) Q27. I enjoy exploring ideas with AI before committing to one.

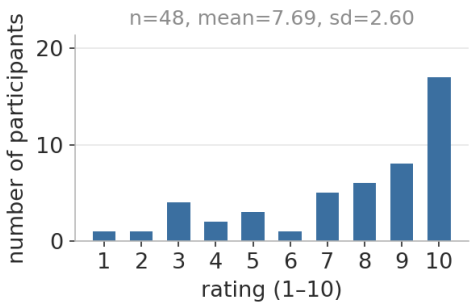


(b) Q28. I usually stop searching once I find one satisfactory answer.

EXTERNAL / CROSS-CHECKING

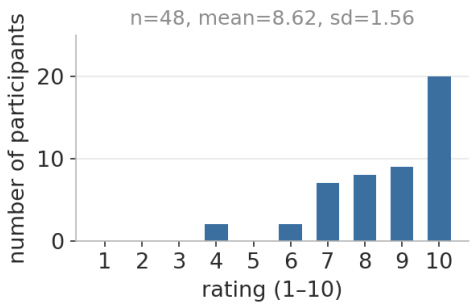


(a) Q29. I frequently combine information from AI with external documents.

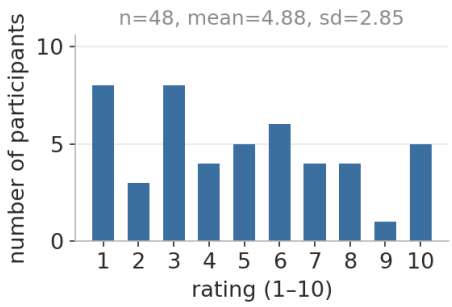


(b) Q30. I compare AI responses with academic papers or other trusted sources.

TOOL-SWITCHING BEHAVIOR

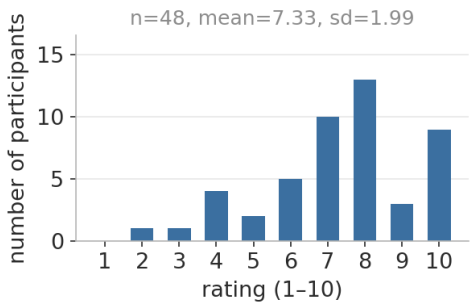


(a) Q31. My workflow often involves moving between AI and other tools.

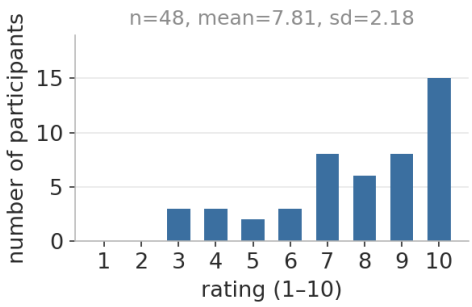


(b) Q32. Most of my work happens within a single AI chat.

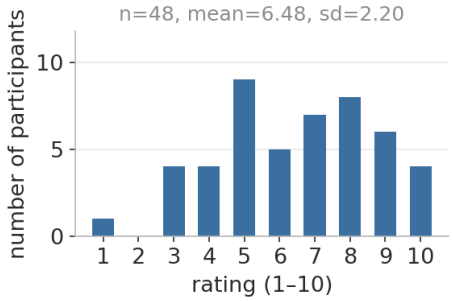
PLANNING BEHAVIOR



(a) Q33. Before using AI, I usually have a clear plan.

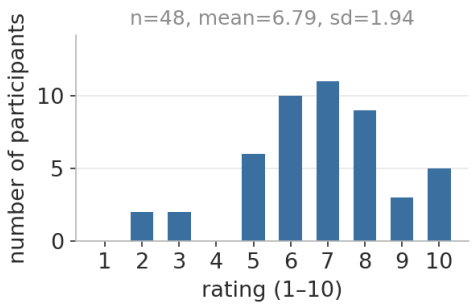


(b) Q34. I break complex tasks into smaller steps before asking AI.

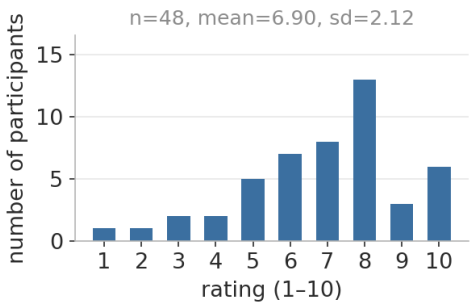


(a) Q35. I use AI to execute a plan rather than determine one.

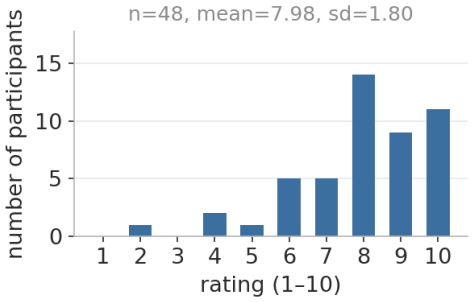
Reflection & Reframing



(a) Q36. I often reconsider my approach after interacting with AI.

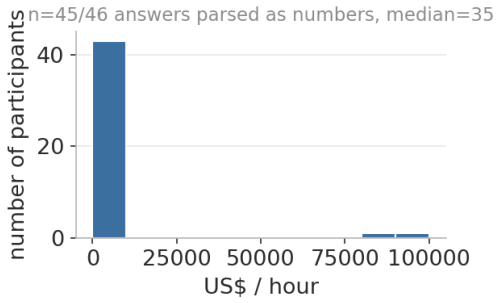


(b) Q37. Working with AI changes how I think about the problem.



(a) Q38. AI helps me discover assumptions I hadn't considered.

ERROR COST & RISK PERCEPTION



(a) Q39. What is the average hourly wage (in US Dollar) for a job similar to yours?

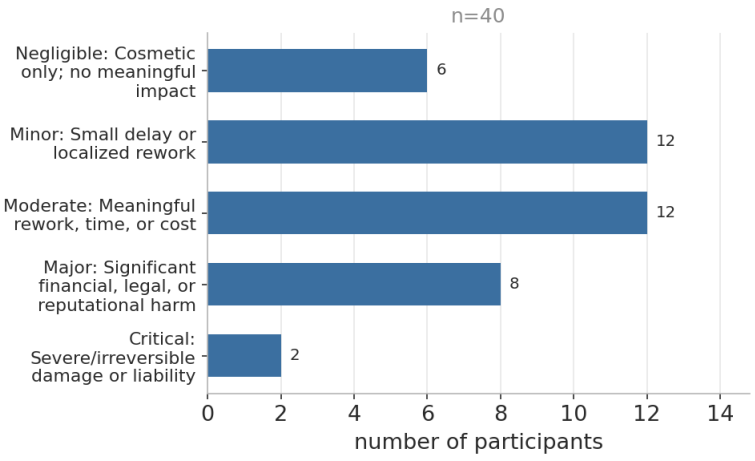


Figure 37: Q40. How serious would an undetected material error be in the tasks you did for this study?

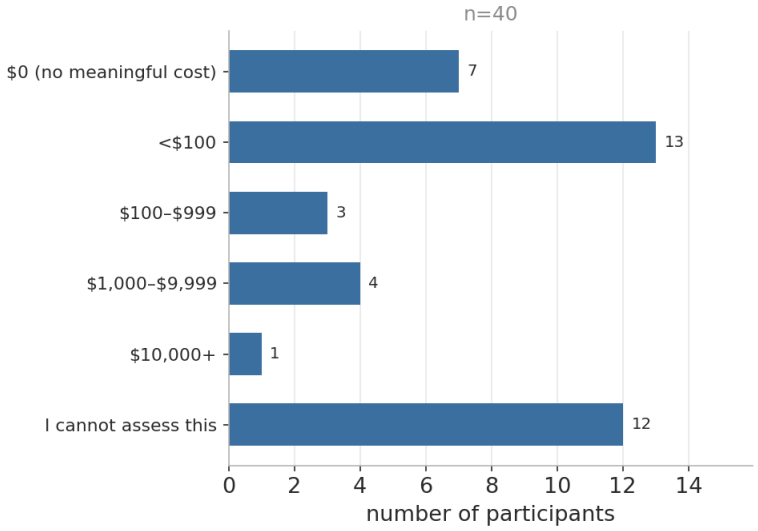


Figure 38: Q41. Estimated financial/resource cost of an undetected error:

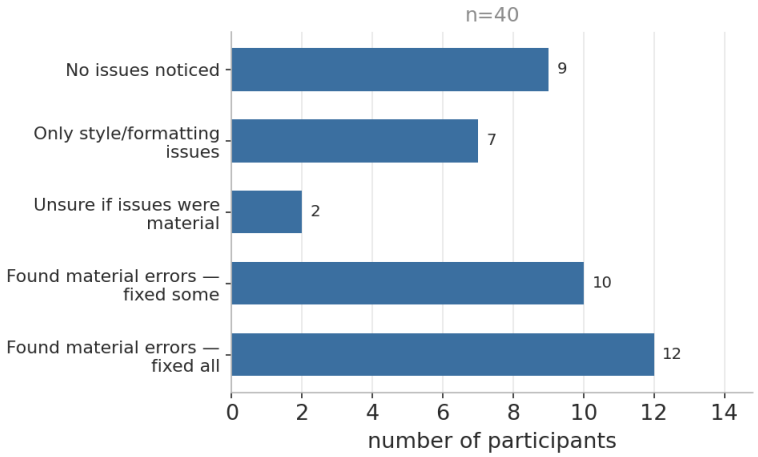
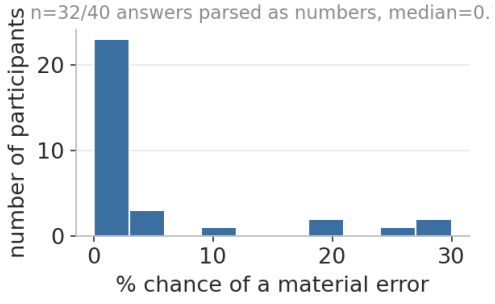


Figure 39: Q42. Did you notice any issues in the AI content during the tasks you did for this study?



(a) Q43. Estimated chance your final output still contains a material error: