WORKBANK: AUDITING AI AGENT READINESS AGAINST WORKER DESIRE ACROSS THE U.S. WORKFORCE

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid rise of AI agents is reshaping the labor market, raising concerns about job displacement, diminished human agency, and overreliance on automation. Yet these debates proceed without answering a fundamental question: which tasks should AI agents automate, and which present opportunities for augmentation? Existing approaches, from agent benchmarks to exposure studies, measure what agents can do but lack the evaluative apparatus to distinguish technical feasibility from social acceptability and what workers want. We address this gap by introducing a task-level auditing framework that contrasts worker desire with technological capability across the automation-augmentation spectrum. Our framework introduces the Human Agency Scale (HAS, H1-H5), a shared evaluation language that moves beyond binary automation framing to capture the full range of agent readiness. Instantiating this framework at scale, we construct WORKBank—comprising data from 1,500 domain workers across 104 occupations and capability assessments from 52 AI experts—as the first evaluation resource to jointly measure worker desire and AI readiness across the U.S. workforce. We demonstrate three evaluative use cases: auditing worker-centered automation demand across sectors, identifying mismatches between the desire-capability landscape and current AI research and investment, and informing collaborative agent design from HAS profiles. Critically, 31.6% of tasks fall in a high-capability but low-desire zone, while 12.7% represent high-desire but low-capability opportunities that current R&D underserves. All data, codebook, and the HAS specification are released publicly to support longitudinal tracking of AI agent readiness across the workforce.

1 INTRODUCTION

Rapid advances in foundation models, such as large language models (LLMs), have catalyzed growing interest in AI agents: goal-directed systems equipped with tool access and multi-step execution capabilities. Their integration into occupational settings is already shaping the labor market (Hoffmann et al., 2024; Demirci et al., 2025). For example, research indicates that around 80% of U.S. workers may see LLMs affect at least 10% of their tasks, with 19% potentially seeing over half impacted (Eloundou et al., 2024). Usage data from Anthropic indicates that in early 2025, at least some workers in 36% of occupations *already* were using AI for at least 25% of their tasks (Handa et al., 2025). While such adoption has shown promise in boosting productivity, it also raises critical debates about the future division of labor between humans and machines, along with concerns about job displacement, reduced human agency, and overreliance on automation.

Yet this debate often proceeds without answering a fundamental question: *which tasks should AI agents automate, and which present opportunities for augmentation?* Answering this requires measuring both what agents can currently do and what workers actually want them to do, across the full diversity of occupational work—a question existing evaluation resources are not equipped to answer. From a coverage perspective, prior research often focuses on a few domains like software engineering (Hoffmann et al., 2024) and customer support (Brynjolfsson et al., 2025), or on occupations that contribute most to GDP (Patwardhan et al., 2026). This scope selection limits our understanding of the real-world complexity of diverse human jobs and the varied nature of open-ended tasks. From a stakeholder perspective, existing studies often emphasize the interests of capital without adequately

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.

considering worker values (Eisfeldt et al., 2023; Baslandze et al., 2026). Furthermore, prior evaluation frameworks treat automation as a threshold question, *i.e.*, exposed or not exposed, automatable or not (Eloundou et al., 2024). This forecloses evaluation of the augmentation spectrum: the range of human-AI collaboration configurations that may be more appropriate, and more desired, than full automation for most tasks (Acemoglu et al., 2026).

Here we introduce a scalable, survey-based auditing framework that addresses these gaps by contrasting worker desire with technological capability at the task level. Our framework leverages an audio-based, semi-structured survey methodology to elicit first-hand insights on how people work and how they envision using AI agents in the United States. To provide a shared evaluation language for the automation-augmentation spectrum, we propose the *Human Agency Scale* (HAS, H1–H5), which quantifies the degree of human involvement required for occupational task completion and quality. Instantiating this framework at scale, we look at the entire workforce that could be impacted by digital AI agents by sourcing occupational tasks from the U.S. Department of Labor's O*NET database (National Center for O*NET Development). Compared to occupation-level evaluations, task-level auditing allows us to better capture the nuanced, open-ended, and contextual nature of real-world work. In total, 1,500 workers across 104 occupations evaluated 844 tasks between January and May 2025. Our auditing framework complements this worker-centric perspective by collecting assessments from 52 AI researchers and practitioners on the same dimensions but focused on current technical feasibility, together forming the AI Agent **W**orker **O**utlook & **R**eadiness **K**nowledge Bank (WORKBank). To the best of our knowledge, this represents the *largest-scale worker-centric* audit of AI agent readiness. Compared to previous research that often focuses narrowly on a few profitable sectors or extrapolates from retrospective usage data, WORKBank provides first-hand data across 13 sectors, enabling us to identify heterogeneous trends across different occupational contexts.

Using WORKBank, we demonstrate three evaluative use cases that yield actionable directions for AI agent development and workforce training that center human agency in the evolving future of work:

1. **Auditing worker-centered automation demand across sectors (Figure 3).** After reflecting on potential job loss concerns and work enjoyment, workers still express positive attitudes toward AI agent automation for 46.1% of tasks, though trends vary significantly by sector.
2. **Identifying mismatches between the desire-capability landscape and current AI research and investment (Figure 4).** Worker desire and technological capability jointly divide tasks into four zones: Automation "Green Light" Zone (high desire and capability), Automation "Red Light" Zone (high capability but low desire), R&D Opportunity Zone (high desire but currently low capability), and Low Priority Zone (low desire and low capability). Notably, 41.0% of Y Combinator company-task mappings are concentrated in the Low Priority Zone and Automation "Red Light" Zone.
3. **Informing human-agent collaboration design through HAS profiles across occupations (Figure 5).** HAS distributions reveal that 45.2% of occupations have H3 (equal partnership) as the dominant worker-desired level, underscoring that augmentation evaluation is at least as important as automation evaluation. To support domain-specific agent design, we release an interactive visualization of HAS profiles across all 104 occupations[1].

2 Related Work

Digital AI Agents Recent advances in foundational models, particularly large language models (LLMs), have sparked a surge in the development of digital AI agents that leverage LLMs to plan actions and interface with external tools (Sumers et al., 2024; Wang et al., 2024). These agents have demonstrated the ability to carry out workflows across diverse domains, including software engineering (Yang et al., 2024; Wang et al., 2025), analytical writing (Shao et al., 2024; Jiang et al., 2024), and customer support (Yao et al., 2025). While many of these agents are designed for full automation, they can also be structured to collaborate with humans (Shao et al., 2026; Luo et al., 2026; Feng et al., 2026; Mozannar et al., 2025). This highlights the broader potential of AI systems to augment, rather than simply replace, human labor (Brynjolfsson, 2022; Acemoglu et al., 2026). The auditing framework proposed in this work systematically examines augmentation versus automation by introducing the Human Agency Scale (HAS).

[1]URL withheld for anonymity; will be included in the camera-ready version.

Table 1: **Comparison of work evaluating AI agent readiness in the workplace.** Prior work measures either technological capability or worker perception, and annotation sources are drawn from a single pool. WORKBank is the first evaluation resource to jointly measure worker desire and expert-assessed capability at the task level across a representative cross-section of the U.S. workforce.

	#Occupations	Perspective	Annotation Source
GDPval (Patwardhan et al., 2026)	44	Capability only	Industry experts
Remote Labor Index (Mazeika et al., 2025)	9	Capability only	Upwork workers
APEX (Vidgen et al., 2025)	4	Capability only	Industry experts
Worker-Seen Automation (Armstrong et al., 2024b)	N/A	Worker perception only	Ipsos Survey
WORKBank (Ours)	104	Capability + Worker desire	Domain workers (Prolific, Upwork, LinkedIn) + 52 AI researchers & practitioners

Evaluating AI Agent Readiness In Workplace Prior work across AI and digital economics has sought to characterize the readiness of AI systems in the workplace. Early work following the launch of ChatGPT established the scale of potential LLM impact: Eloundou et al. (2024) estimated that approximately 80% of the U.S. workforce has at least some tasks exposed to LLM capabilities, while Handa et al. (2025) leveraged real usage data from Claude.ai to identify which economic tasks workers actually perform with AI. Despite their broad coverage, these studies focus primarily on chatbot-style systems and rely on LLM-as-judge or usage-log-based annotations. More recent benchmarks shift toward evaluating agent performance on concrete occupational tasks. Patwardhan et al. (2026) curate tasks across the top 9 GDP-contributing sectors; Mazeika et al. (2025) construct the Remote Labor Index from task types prevalent on the Upwork freelancing platform; and Vidgen et al. (2025) develop APEX, focusing on four economically high-value occupations. While these efforts provide more task-grounded evaluations, they share two key limitations. First, they assess technical capability without accounting for whether workers actually want AI systems to perform these tasks, despite evidence that both worker buy-in and system performance jointly shape adoption outcomes (Van Hoek, 2024; Baslandze et al., 2026). Second, their task selection is biased toward current economic value, overlooking how the distribution of valuable work may evolve as AI capabilities advance.

Work that does incorporate worker perspectives remains limited in scope and granularity. Armstrong et al. (2024b) survey more than 9,000 workers—broadly categorized as office and non-office workers—on their general attitudes toward how technologies like robots and AI will affect their work. While valuable for characterizing broad sentiment, the collected data cannot identify where automation should or should not occur within specific occupational workflows. More fundamentally, prior evaluation frameworks, including all studies above, treat automation as a binary decision, leaving out the possibility of human-agent collaboration configurations that may be more appropriate, and more desired, than full automation in certain scenarios (Acemoglu et al., 2026).

3 Worker-Centric Auditing Framework

To audit AI agent integration into professional work, we capture worker desire and agent readiness at the task level across the automation–augmentation spectrum.

3.1 Audit Granularity and Scope

Our framework focuses on *complex, multi-step tasks* associated with specific occupations (*e.g.*, "Marketing Managers: Compile lists describing product or service offerings"), sourced from the O*NET database. These tasks, unlike isolated, low-level activities (*e.g.*, "track goods or materials" or "translate information"), reflect actual job responsibilities and the kinds of workflows AI agents are poised to impact. Moreover, compared with occupation-level analysis, conducting the audit at the task level enables a more nuanced understanding, as tasks within the same profession can vary significantly and are often highly contextualized.

We scope our audit to computer-compatible tasks, recognizing their susceptibility to *foundation model-powered AI agents*. Drawing from historical and recent literature on agent autonomy (Russell & Norvig, 1995), planning capabilities, and tool use (Mitchell et al., 2025), we define AI agents (excluding physical robots) as "a system or program capable of autonomously performing tasks on behalf of a user or another system by designing its workflow and utilizing available software tools, without the ability to perform physical actions."

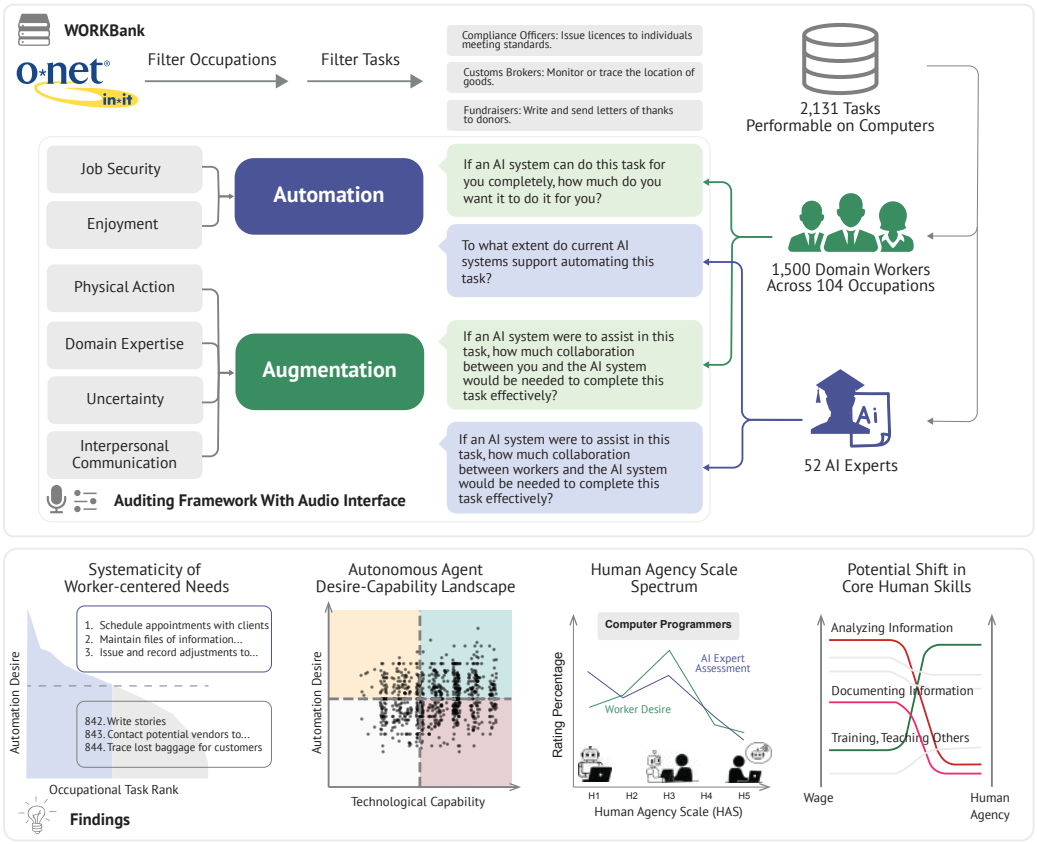

Figure 1: **Overview of the auditing framework and evaluative use cases.** The framework elicits worker desires and expert capability assessments on automation and augmentation through structured prompts and an audio-enhanced interface. We instantiate this framework to build the WORKBank database and demonstrate three evaluative use cases.

3.2 Human Agency Scale

Traditional technology impact studies often ask: *To what degree can this task be automated?* Besides this view of automation, we consider the view of *augmentation*—where technology complements and enhances human capabilities (Autor, 2015), as this new wave of technology holds significant potential to augment human workers through human-agent collaboration. While there are established levels-of-automation taxonomies (Myhre et al., 2019; Parasuraman et al., 2000), there is no established framework for quantifying automation vs. augmentation.

To fill this gap and provide a shared language, we introduce the Human Agency Scale (HAS), a five-level scale from H1 (no human involvement) to H5 (human involvement essential) (See Figure 2). Building on prior work that raises the concept of augmentation (Brynjolfsson, 2022; Handa et al., 2025), we define each level along two dimensions: *team dynamics*, which captures how responsibility for task execution is distributed between humans and AI agents, and *required human involvement*, which captures the degree to which human input is necessary for acceptable task completion. Levels H1-H2 correspond to an automation regime in which AI replaces human effort, while H3-H5 correspond to an augmentation regime in which AI enhances it. Unlike SAE driving automation levels (Committee et al., 2014) that adopt an "AI-first" perspective, HAS provides a human-centered lens for assessing both task properties and appropriate agent development approaches. Importantly, higher HAS levels are not inherently better—different levels suit different AI roles. Tasks at H1-H2 favor automation approaches, while H3-H5 tasks benefit from augmentation strategies. Understanding the ideal level of human involvement is essential both for workers seeking to adapt their skills and for developers aiming to build context-appropriate AI agents. For instance, fully autonomous agents shall be developed for H1 scenarios, while those agents for H3 scenarios must support meaningful coordination and communication with human collaborators.

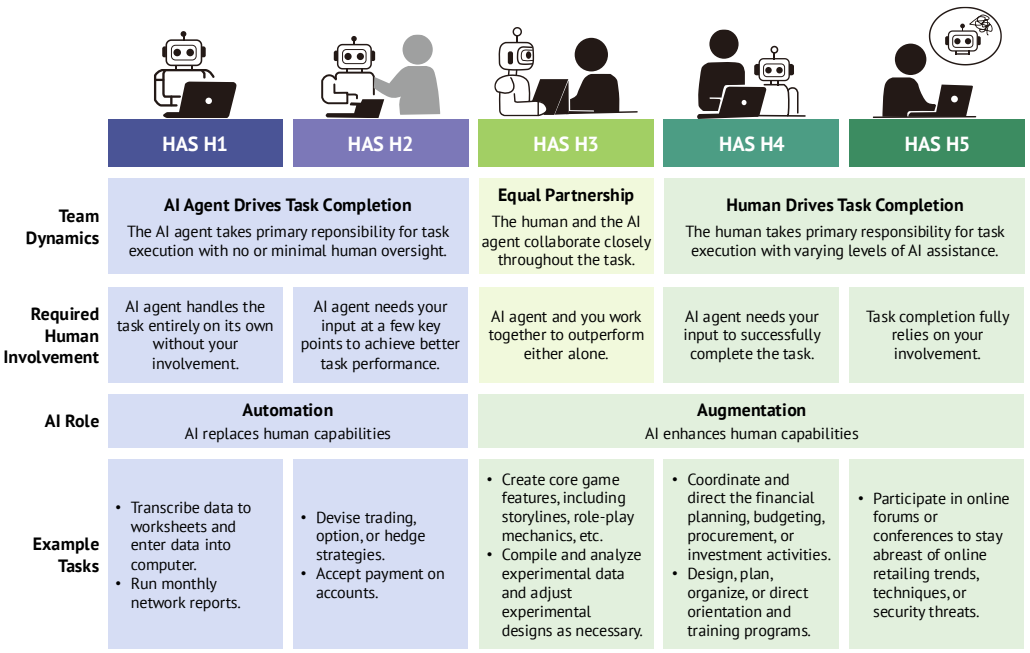

Figure 2: **Levels of Human Agency Scale (HAS).** HAS provides a shared language to quantify automation vs. augmentation, complementing the "AI-first" perspective used in defining levels of automation. Higher HAS levels are not inherently better—different levels suit different AI roles.

3.3 Audit Procedure

Our auditing framework elicits two complementary assessments on each task t: worker desire and expert-assessed capability.

Worker Assessments. Our framework centers on the needs of workers. To support domain workers in providing well-calibrated ratings on *automation desire* $A_w(t)$ and *desired HAS level* $H_w(t)$, we develop an audio-supported survey system and scaffold worker responses through three key designs: (1) **Audio-enhanced Reflection**: The survey begins with an audio-enhanced mini-interview exploring participants' occupational work and AI perspectives. This spoken format enables more natural reflection and helps workers more efficiently contextualize their ratings within their actual work experience, compared to slowly typing their experiences. (2) **Quality Control via Task Familiarity Filtering**: Workers receive only occupation-relevant tasks and must confirm task familiarity before rating, ensuring assessments are grounded in their real experience rather than speculation. (3) **Guided Consideration**: Before answering Likert questions, participants consider factors identified in prior literature—enjoyment and job security concerns for $A_w(t)$ (Armstrong et al., 2024a; Gödöllei & Beck, 2023), and task characteristics like physical actions, domain expertise requirements, uncertainty, and interpersonal elements for $H_w(t)$ (Shah & White, 2025; Parasuraman, 2000; Frank et al., 2019). The full survey questions are included in Appendix A.

AI Expert Assessments. While worker perspectives capture the social demand and acceptance of AI agents, they represent only one side of the integration equation. Domain workers, despite their deep task expertise, may have limited exposure to current AI capabilities and constraints. Thus, we complement workers' perspectives with expert assessments of *current automation capability* $A_e(t)$ and *feasible HAS level* $H_e(t)$ from AI researchers and practitioners, using the same rubrics and drawing on their understanding of existing systems' strengths and limitations. Contrasting $A_w(t)$, $H_w(t)$ and $A_e(t)$, $H_e(t)$ constitutes the core evaluative operation of our audit framework. Alignment between the two perspectives identifies tasks that are well-positioned for near-term deployment. In contrast, divergence reveals critical mismatches: cases where workers desire automation beyond current technical capability point to high-impact directions for further research, while cases where agents are capable but workers resist automation highlight the need to preserve human agency and design for collaborative agents.

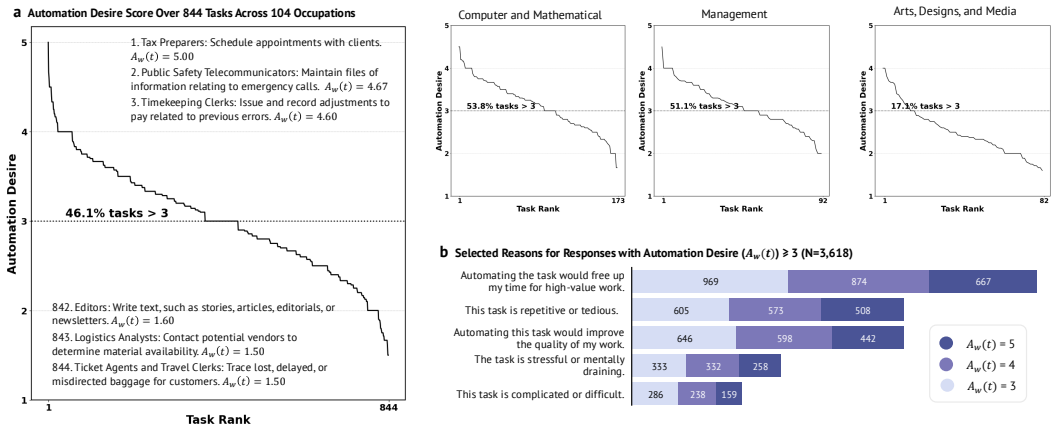

Figure 3: **Our audit reveals structured, heterogeneous worker demand for automation.** **a**, Automation desire scores $A_w(t)$ over 844 occupational tasks, ranked based on WORKBank data, together with sector-specific breakdowns. **b**, Reported reasons for responses with $A_w(t) \geq 3$.

3.4 Instantiating the Audit Framework to Derive WORKBank

We instantiate the audit framework by developing an IRB-approved, self-hosted survey interface and distributing it across Prolific, Upwork, and LinkedIn to ensure a broad domain worker coverage. We source computer-compatible tasks performed at least monthly from the U.S. Department of Labor's O*NET Database (details in Appendix C.1). These tasks reflect complex, multi-step workflows central to our focus. Between January and May 2025, 1,676 workers participated in our audit. A sector-level comparison with Bureau of Labor Statistics workforce data shows that our participant pool captures a broad cross-section of the U.S. workforce (Appendix D.2).

After filtering for adequate representation ($\geq$ 10 participants per occupation), our final sample comprises 1,500 individuals across 104 occupations, from which we compute task-level worker assessments $A_w(t)$ and $H_w(t)$. Across 5,731 task-ratings, all five levels of the Human Agency Scale are substantively populated, with H3 and H4 accounting for 30.13% and 20.31% of ratings respectively. Moreover, within-task rating dispersion at the H3–H4 boundary is comparable to that at other adjacent-level boundaries, indicating no localized confusion.

For expert assessments, we recruited 52 AI experts—PhD researchers and industry practitioners from institutions including Stanford University, MIT, Google, and xAI, *etc.* Each task was independently assessed by at least two experts, with additional reviews ensuring rating consistency (standard deviation $\leq$ 1, see robustness analysis details in Appendix B). Combining these complementary data sources, we construct **W**orker **O**utlook & **R**eadiness **K**nowledge Bank (WORKBank), the first database to capture both worker desires and AI agents readiness for occupational tasks. In §4, we will demonstrate the evaluative use cases of our auditing framework using WORKBank.

4 Evaluation Findings

4.1 Worker-Centered Automation Demand

A fundamental evaluative question for AI agent deployment is not merely whether a task is technically automatable, but whether workers performing that task want it automated. In Figure 3 **a**, we rank tasks by their average worker automation desire scores $A_w(t)$. We find that for 46.1% of tasks, workers currently performing them express a positive attitude (*i.e.*, $A_w(t) > 3$) toward AI agent automation, even after explicitly considering concerns such as job loss and reduced enjoyment, as guided by our auditing framework. As shown in Figure 3 **b**, the dominant motivation "freeing up time for high-value work" was cited in 69.38% of pro-automation responses, followed by task repetitiveness and quality improvement opportunities (each 46.6%). This suggests workers evaluate automation in terms of what it enables, not merely what it replaces. However, the distribution also indicates a mixed attitude, and the results largely differ across sectors.

Where do workers resist AI agent automation? Among our participants expressing negative sentiment to the question "How do you envision using AI in your daily work?" (28%), topic modeling of audio responses identifies three dominant concerns: lack of trust in AI accuracy or reliability (45.0%), fear of job replacement (23.0%), and the perceived absence of human qualities in AI (16.3%) (see Appendix F.1). This sentiment echoes our quantitative findings from the breakdown of automation desire scores across sectors (Figure 3 **a**). In these sector-level breakdowns, the "Arts, Designs, and Media" sector stands out, with only 17.1% of tasks receiving positive desire ratings. Audio responses from participants in this sector reveal nuanced opposition to automating content creation, such as: "*I want it to be used for seamlessly maximizing workflow and, you know, making things less repetitive and tedious and arduous with workflow. No content creation,*" "*I would never use AI to like replace artists. I would be more for personal [project management] use, if anything.*"

4.2 Desire-Capability Landscape for AI Agents in the Workplace

As discussed in §3.3, our auditing framework allows the evaluative operation of contrasting automation desire $A_w(t)$ and expert-assessed technological capability $A_e(t)$, visualized as a desire-capability landscape in Figure 4 **a**. This landscape divides into four zones:

1. **Automation "Green Light" Zone**: Tasks with both high automation desire and high capability. These are prime candidates for AI agent deployment with the potential for broad productivity and societal gains.
2. **Automation "Red Light" Zone**: Tasks with high capability but low desire. Deployment here warrants caution, as it may face worker resistance or pose negative societal implications.
3. **R&D Opportunity Zone**: Tasks with high desire but currently low capability. These represent promising directions for AI research and development.
4. **Low Priority Zone**: Tasks with both low desire and low capability. These are less urgent for AI agent work.

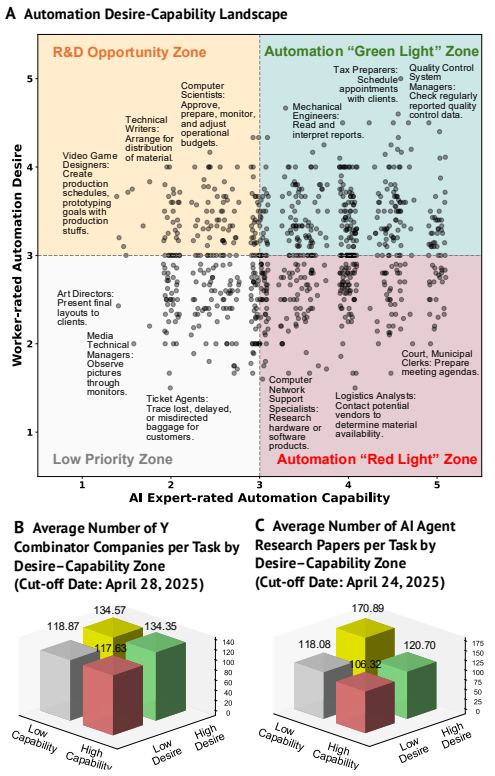

Figure 4: **Our audit reveals the desire-capability landscape, enabling investigations of AI investment and research alignment.**

The desire-capability landscape is robust to the choice of decision threshold. Using the midpoint threshold, 31.64% of tasks fall in the Automation "Red Light" zone and 12.68% in the R&D Opportunity zone. Using more conservative percentile-based thresholds (see robustness analysis details in Appendix B.1), the corresponding proportions vary only modestly (Red Light: 25.95–28.08%; R&D Opportunity: 12.80–16.71%), while the overall qualitative pattern remains unchanged: a substantially larger share of tasks have AI capability exceeding worker desire than the reverse.

Mapping investment to the desire-capability landscape reveals critical mismatches. Identifying the desire-capability landscape allows us to further understand where current investments and research on AI agents are concentrated. We used Y Combinator (YC) companies[2] as a proxy and mapped them to the tasks in our database using `gpt-4.1-mini`. One company could be mapped to multiple tasks, and author annotations reveal that the annotations achieve 92% precision on the final task mapping (see Appendix C.4 for details). As shown in Figure 4 **b**, the company-task mappings are relatively evenly spread across the four zones. Most mapped tasks are concentrated in occupations

[2] `https://www.ycombinator.com/companies`

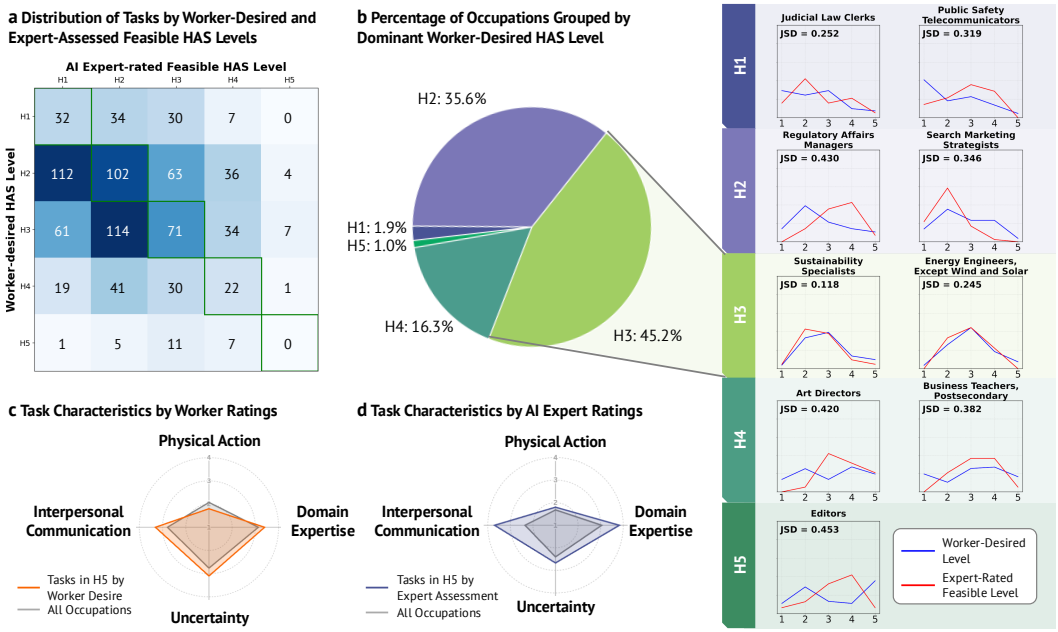

Figure 5: **Human Agency Scale (HAS) distributions expose friction points and underscore opportunities for human–agent collaboration across occupations . a**, Task-level comparison between worker-desired and expert-assessed feasible HAS levels. **b**, Distribution of occupations by their dominant worker-desired HAS level. Full occupation-level results are shown in Figure 8. **c,d**, Task characteristics driving high human agency according to domain workers and AI experts.

related to software development and business analysis. 41.0% of YC companies are mapped to Low Priority and Automation "Red Light" Zone; while many promising tasks within the "Green Light" Zone and Opportunity Zone remain under-addressed by current investments.

AI agent research papers show an emphasis on the R&D Opportunity Zone, but remain concentrated on certain tasks. Following a similar methodology, we collected AI agent papers from arXiv before April 24, 2025, filtering by keyword match on title and abstract (see Appendix C.4 for details), and mapped them to occupational tasks to characterize the distribution of current research efforts. Figure 4 **c** shows that research papers are more concentrated in the R&D Opportunity Zone. While encouraging, the focus remains largely confined to computer science and engineering domains.

4.3 Human Agency Scale (HAS) Spectrum

Beyond automation, AI agents can augment human work. WORKBank's HAS distributions reveal where human-agent collaboration is most viable, most desired, and most at risk of misalignment.

Worker-expert divergence identifies friction points for workplace AI agents. Figure 5 **a** shows workers generally prefer higher levels of human agency than what experts deem technologically necessary, with 47.5% of tasks falling into the lower triangle of the matrix. To quantify this divergence, we compute the Jensen-Shannon Distance (JSD) between the distributions of $H_w(t)$ and $H_e(t)$ at the occupation level. Disagreements are most pronounced in the lower HAS range—five of the ten occupations with the highest JSD scores are also those that experts rate as H1 dominant (see Table 7). This signals potential frictions as AI adoption progresses. Ensuring socially responsible deployment of AI agents and supporting workers currently performing low-HAS tasks warrant further scrutiny. At the other extreme, very few occupations fall in the H5 dominant regime. "Editors" is the only occupation where workers predominantly desire H5. According to AI expert assessments, only "Mathematicians" and "Aerospace Engineers" fall into this category. Representative work descriptions from worker transcripts for these occupations are provided in Appendix F.2. Among the four quantitative dimensions in our auditing framework, worker ratings highlight Interpersonal Communication as a defining feature of H5 tasks (Figure 5 **c**), while expert ratings emphasize both Interpersonal Communication and Domain Expertise (Figure 5 **d**).

Common patterns reveal opportunities for human-agent collaboration. As illustrated in Figure 5 **b** (with full results in Figure 8), many occupations (*e.g.*, "Sustainability Specialists", "Energy Engineers") exhibit an inverted-U shaped distribution for both $H_w(t)$ and $H_e(t)$. While this trend in expert assessments might reflect current technological limitations, it is notable that workers in many domains also prefer a balanced, collaborative partnership with AI. Our analysis of audio transcripts further supports this: the vast majority of workers express either a desire or openness to collaborating with AI to enhance their work. The most common paradigm is "role-based" AI support (described by 23.1% of workers), where individuals anticipate *utilizing AI systems that embody specific roles or personalized functions* (*"[I would like to have an AI agent] trained to automatically analyze the quality control reports of raw sequencing data [...] and suggesting appropriate pre-processing steps."*). Furthermore, 23.0% of workers express a desire for AI systems to function as a supportive assistant for some or all aspects of their workflow (*"[I envision the AI agent] as an assistant who is doing research for me. However, I review every answer because we cannot rely on its accuracy."*), while 16.5% mention pure automation by AI for some aspects of their workflow.

Broader Utilities of WORKBank The evaluative use cases presented above underscore the risk of misalignment in unguided AI agent development and identify worker-centered opportunities for automation and augmentation. Beyond evaluation, WORKBank can support a range of applications across AI agent development, policy, and labor research. For practitioners building AI agents, the dataset could inform benchmark curation by identifying tasks where workers express high automation desire, helping ensure evaluation efforts reflect real workforce needs (Li et al., 2026). On the policy and workforce side, datasets like WORKBank can provide empirical grounding for discussions around meaningful work design (Ranjit et al., 2026), human skill adaptation under AI integration (Liepmann & Hegewisch, 2025), and monitoring AI's broader economic impact (Sajadieh et al., 2026). More broadly, WORKBank's HAS distributions can inform training objectives that optimize for augmentation; for instance, by identifying tasks where models should defer to human judgment or support collaborative decision-making. We hope WORKBank serves as an extensible resource and invite the community to build on it to advance worker-centered AI development.

5 Discussion

We introduce the first large-scale, worker-centric audit of AI agent readiness across the U.S. workforce, contributing both a reusable auditing framework and the WORKBank database. Methodologically, the framework evaluates readiness across the full automation-augmentation spectrum, and contrasts worker-rated desires with AI experts' capability assessments on the same task-level rubrics, enabling three concrete evaluative use cases: auditing worker-centered automation demand across sectors, identifying mismatches in the desire-capability landscape, and informing collaborative agent design from occupational HAS profiles. *For AI developers*, the desire-capability landscape identifies underserved high-desire tasks as priority targets for AI agent development, while HAS profiles provide direct guidance for human-in-the-loop design in agentic systems. From a socio-technical perspective, as AI systems grow increasingly capable, deployment decisions are often guided by technical feasibility alone. Yet it is workers—those most affected by these transitions—who form the foundation of the economy. *For policymakers*, our audit provides task-level empirical grounding for AI regulation and proactive reskilling investment. Our auditing framework offers a timely and practical tool to re-center the evaluation of AI agents on the human side of adoption, complementing performance benchmarks with measurements of social acceptability and appropriate human agency.

Limitations Several limitations should be considered. *First*, our assessments are grounded in O*NET's existing occupational tasks and may not capture new tasks that emerge as AI agents enter the workplace; further analysis of WORKBank's open-ended worker transcripts offers one path to surface such patterns. *Second*, despite prompts to reflect on job-loss and enjoyment concerns, workers may still withhold honest feedback due to job-security concerns; we treat such concerns as critical input for responsible deployment. *Third*, WORKBank covers 104 occupations, a subset of the 287 computer-using occupations identified within the O*NET database. While our findings may not cover the full picture of AI agents for the workplace, our multi-platform recruitment (Upwork, Prolific, LinkedIn) closely matching U.S. workforce sector distributions across the 13 sectors represented in our database. *Finally*, our Y Combinator analysis uses company counts rather than investment amounts (to which we lack access).

AI USE STATEMENT

In this work, LLMs are used to (1) filter computer-using occupations and tasks from O*NET (Appendix C.1), (2) map YC companies and arXiv papers to tasks in our WORKBank database (Appendix C.4), and (3) analyze audio transcripts through topic modeling (Appendix F.1). We manually check all filtered tasks before including them into the final WORKBank database. For the analysis of YC companies, we validate LLM annotations against author annotations on 100 randomly sampled examples, where LLM annotations achieve 92% precision on the final task mapping. LLMs are not used to generate the worker or expert ratings that constitute the core dataset. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Ben Armstrong, Valerie K. Chen, Alex Cuellar, Alexandra Forsey-Smerek, and Julie A. Shah. Automation from the worker's perspective, 2024a. URL https://arxiv.org/abs/2409.20387.

Ben Armstrong, VK Chen, A Cuellar, A Forsey-Smerek, and J Shah. Automation from the worker's perspective: How can new technologies make jobs better, 2024b.

David H Autor. Why are there still so many jobs? the history and future of workplace automation. *Journal of economic perspectives*, 29(3):3–30, 2015.

David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333, 11 2003. ISSN 0033-5533. doi: 10.1162/003355303322552801. URL https://doi.org/10.1162/003355303322552801.

Salomé Baslandze, Zachary Edwards, John Graham, Ty McClure, Brent H Meyer, Michael Sparks, Sonya R Waddell, and Daniel Weitz. Artificial intelligence, productivity, and the workforce: Evidence from corporate executives. Technical report, National Bureau of Economic Research, 2026.

Erik Brynjolfsson. The turing trap: The promise & peril of human-like artificial intelligence. *Daedalus*, 151(2):272–287, 2022.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative ai at work. *The Quarterly Journal of Economics*, pp. qjae044, 2025.

SAE On-Road Automated Vehicle Standards Committee et al. Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. *SAE Standard J*, 3016:1, 2014.

Ozge Demirci, Jonas Hannane, and Xinrong Zhu. Who is ai replacing? the impact of generative ai on online freelancing platforms. *Management Science*, 2025.

Matthew S Dey and Mark A Loewenstein. *On job requirements, skill, and wages*. US Department of Labor, US Bureau of Labor Statistics, Office of Employment . . . , 2019.

Andrea L Eisfeldt, Gregor Schubert, Miao Ben Zhang, and Bledi Taska. The labor impact of generative ai on firm values. *Available at SSRN 4436627*, 2023.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: Labor market impact potential of llms, 2024.

K. J. Kevin Feng, Kevin Pu, Matt Latzke, Tal August, Pao Siangliulue, Jonathan Bragg, Daniel S Weld, Amy X. Zhang, and Joseph Chee Chang. Cocoa: Co-planning and co-execution with ai agents. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791673. URL https://doi.org/10.1145/3772318.3791673.

Morgan R. Frank, David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebrian, David J. Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, Dashun Wang, Hyejin Youn, and Iyad Rahwan. Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539, 2019. doi: 10.1073/pnas.1900949116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1900949116.

Anna F. Gödöllei and James W. Beck. Insecure or optimistic? employees' diverging appraisals of automation, and consequences for job attitudes. *Computers in Human Behavior Reports*, 12: 100342, 2023. ISSN 2451-9588. doi: https://doi.org/10.1016/j.chbr.2023.100342. URL https://www.sciencedirect.com/science/article/pii/S2451958823000751.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*, 2025.

Manuel Hoffmann, Sam Boysel, Frank Nagle, Sida Peng, and Kevin Xu. Generative ai and the nature of work. Technical report, CESifo Working Paper, 2024.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina Semnani, and Monica Lam. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9917–9955, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.554. URL https://aclanthology.org/2024.emnlp-main.554/.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2024.

Yuetai Li, Yichen Feng, Zhangchen Xu, Zixian Ma, Kaiyuan Zheng, Fengqing Jiang, Xinghua Sun, Rulin Shao, Zichen Chen, Yue Huang, et al. Jobbench: Aligning agent work with human will. *arXiv preprint arXiv:2605.26329*, 2026.

Hannah Liepmann and Ariane Hegewisch. Revisiting occupational segregation and the valuation of women's work. 2025.

Hanjun Luo, Chiming Ni, Jiaheng Wen, Zhimu Huang, Bingduo Liao, Sylvia Chung, Yiran Wang, Yingbin Jin, Jialin Li, Xinfeng Li, Wenyuan Xu, XiaoFeng Wang, and Hanan Salam. Centaureval: Benchmarking human-in-the-loop value in agentic coding. In *Forty-third International Conference on Machine Learning*, 2026.

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

Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. Fully autonomous ai agents should not be developed. *arXiv preprint arXiv:2502.02649*, 2025.

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.

Bård Myhre, Are Hellandsvik, and Stig Petersen. A responsibility-centered approach to defining levels of automation. In *Journal of Physics: Conference Series*, volume 1357, pp. 012027. IOP Publishing, 2019.

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

Raja Parasuraman. Designing automation for human use: empirical studies and quantitative models. *Ergonomics*, 43(7):931–951, 2000.

Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297, 2000.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simon Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie Kim, Samuel Miserendino, Gildas Chabot, David Li, Patrick Chao, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. 2026:24005–24040, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/290c2430f91912204f30bbcc990fff1d-Paper-Conference.pdf.

Jaspreet Ranjit, Ke Zhou, Swabha Swayamdipta, and Daniele Quercia. Are we automating the joy out of work? designing ai to augment work, not meaning. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–46, 2026.

S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Series in Artificial Intelligence. Prentice-Hall, Englewood Cliffs, NJ, 1995.

Sha Sajadieh, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Lapo Santarlasci, Juan Pava, Nestor Maslej, Russ Altman, Erik Brynjolfsson, Carla Brodley, Jack Clark, Virginia Dignum, Vipin Kumar, James Landay, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Elham Tabassi, Russell Wald, Toby Walsh, and Dan Weld. The ai index 2026 annual report, April 2026.

Chirag Shah and Ryen W. White. Agents Are Not Enough . *Computer*, 58(10):87–92, October 2025. ISSN 1558-0814. doi: 10.1109/MC.2025.3575829. URL https://doi.ieeecomputersociety.org/10.1109/MC.2025.3575829.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.347. URL https://aclanthology.org/2024.naacl-long.347/.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. 2026:99616–99649, 2026.

Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Suzanne Tsacoumis and Shannon Willison. O* net analyst occupational skill ratings: Procedures. *Alexandria, VA: Human Resources Research Organization*, 2010.

Remko Van Hoek. Insight from industry-early lessons learned about ai adoption in core procurement processes, directions for managers and researchers. *Supply Chain Management: An International Journal*, 29(4):794–803, 2024.

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

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Daniel Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. Openhands: An open platform for ai software developers as generalist agents. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu

(eds.), *International Conference on Learning Representations*, volume 2025, pp. 65882–65919, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a4b6ad6b48850c0c331d1259fc66a69c-Paper-Conference.pdf.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: agent-computer interfaces enable automated software engineering. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, volume 2025, pp. 9965–10017, 2025.

Appendix

Table of Contents

A Audit Protocol **15**

B Robustness Analysis **16**
B.1 Threshold Robustness for the Four-Zone Classification 16
B.2 Annotation Agreement of AI Expert Assessments 16
B.3 Mixed-Effects Model Regression on Worker Responses 16

C Usage of External Data and Resources **17**
C.1 Occupational Information Network (O*NET) 17
C.2 Occupational Employment and Wage Statistics by U.S. Bureau of Labor Statistics 18
C.3 Claude.ai Usage Data 18
C.4 Y Combinator (YC) Company and AI Agent Research Paper Data 18

D WORKBank Statistics **21**
D.1 Full List of Included Occupations 21
D.2 Coverage of the U.S. Workforce 22
D.3 Domain Worker Demographic Information 22
D.4 Recruitment Sources 22

E Additional Results **24**
E.1 Top 20 Tasks Workers Want Automated 24
E.2 Bottom 20 Tasks Workers Want Automated 25
E.3 Full Human Agency Scale Results 26
E.4 Top 10 Occupations By Worker-Expert Discrepancies in HAS Ratings 27
E.5 The Potential Shift of Core Human Skills 27

F Audio Response Analysis **28**
F.1 LLM-based Topic Modeling 28
F.2 Audio Response Examples 30

A AUDIT PROTOCOL

We instantiate our auditing framework (see §3) with an audio-enhanced, semi-structured survey interface to collect first-hand data from domain workers who actually perform those tasks. A detailed codebook documenting all variables, rating scales, and response options is released alongside WORKBank. Our survey is structured as follows:

- A **Mini-interview Section** designed to explore participants' work process and perspectives on the role of AI agents in their work. This section consists of five open-ended questions, allowing participants to share their thoughts freely and edit their audio transcripts in real time:
  - A1 Could you please briefly describe what you do for your work?
  - A2 What tasks do you typically do for your work? Think about this question based on the time you spend on each of them.
  - A3 Please tell us more about the tools or software you use for these tasks. Please try to sort them by usage frequency.
  - A4 For the three tasks that you spend the most of your time on, could you walk us through your process of completing each of them?
  - A5 How do you envision using AI in your daily work?
- A **Task Rating Section** assessing both automation desire ($A_w(t)$) and the desired Human Agency Scale level ($H_w(t)$) for tasks associated with the participant's occupation. For each task, participants respond to a series of structured questions. Items T.I3, T.A1–A3, and T.C1–C5 are all rated on a 5-point Likert scale:

  TASK FAMILIARITY QUESTIONS:

  - T.I1 Have you done this task before?
    (Yes/No; if "No", remaining questions are skipped.)
  - T.I2a With respect to this task, I consider myself a...
    (Novice, Average, Expert)
  - T.I2b How much time do you spend on this task in your daily work schedule?
    (10%, 30%, 50%, 70%, 100%)
  - T.I3 How closely is this task related to your core skills or unique strengths that are essential to your job?

  AUTOMATION DESIRE RATING QUESTIONS:

  - T.A1 If an AI can do this task for you completely, how worried would you be that your job will be replaced?
  - T.A2 Without thinking about salary, how much do you enjoy doing this task?
  - T.A3 If an AI can do this task for you completely, how much do you want an AI to do it for you?
  - T.A4 Why would you like this task to be automated by AI? (Shown if T.A3 ≥ 3; multi-select options):
    - Automating this task would free up my time for higher-value work.
    - This task is repetitive or tedious.
    - Automating this task would improve the quality of my work.
    - The task is stressful or mentally draining.
    - This task is complicated or difficult.
    - Automating this task would help me scale and handle higher output.

  HUMAN AGENCY SCALE RATING QUESTIONS:

  - T.C1 How much does this task require taking physical actions or physical labor?
  - T.C2 How much does this task require dealing with uncertainty or making high-stake decisions?
  - T.C3 How much does this task require specific domain expertise (such as specialized knowledge, unspoken wisdom, or insights gained through experience)?
  - T.C4 How much does this task depend on interpersonal communication or empathy?
  - T.C5 If AI were to assist in this task, how much of your collaboration would be needed to complete this task effectively? (References to H1 to H5 are provided)

T.C6 Why would collaboration be needed for this task? (Shown if T.C5 ≥ 3; multi-select options):
- ∗ This task requires physical actions.
- ∗ This task involves making high-stake decisions which I would like to control.
- ∗ This task requires specific domain knowledge.
- ∗ The task involves nuanced communication or interpersonal skills.
- ∗ The task needs validation or oversight to ensure quality.
- ∗ The task is dynamic and requires adapting to changing circumstances.
- ∗ The task has ethical, sensitive, or subjective aspects.

- A **Demographic Question Section** where we ask about information on age, gender, race, income, education, years of experience in occupation, attitude towards AI, zip code, political orientation, and details about their familiarity with LLMs and how they currently use them (*i.e.*, types of usage and frequency).

B Robustness Analysis

B.1 Threshold Robustness for the Four-Zone Classification

The dual perspectives in WORKBank allow us to categorize the landscape into four zones that surface desire-capability mismatches: 31.6% of tasks fall in a high-capability, low-desire zone, while 12.7% fall in a low-capability, high-desire zone. Our four-zone classification splits at the midpoint (3) on both desire and capability scales. To assess robustness to this threshold choice, we conducted analyses using percentile-based thresholds. At the 70th percentile (capability: 4.00, desire: 3.33), 25.95% of tasks fall in the Automation "Red Light" Zone and 16.71% in the R&D Opportunity Zone. At the 75th percentile (capability: 4.00, desire: 3.43), these shift to 28.08% and 12.80%, respectively. While absolute percentages vary, the key qualitative pattern persists: a substantial proportion of tasks fall into the Red Light zone (capability exceeds desire), while a smaller proportion falls into the Green Light zone (high capability and desire). This confirms that the desire-capability mismatch is not an artifact of threshold choice. The midpoint threshold offers conceptual clarity as it represents tasks exceeding the neutral point.

B.2 Annotation Agreement of AI Expert Assessments

AI experts with practical R&D experience provided ratings for current automation capability $A_e(t)$ and feasible human-agency levels $H_e(t)$. To ensure high-quality annotations, we applied the following controls:

- **Expert qualifications.** Each expert satisfied at least one of:
  1. Current PhD student specializing in NLP, large language models, or AI agents.
  2. PhD in Computer Science with demonstrated expertise in AI, LLMs, or agentic systems.
  3. Industry practitioner (*e.g.*, machine learning engineer or research scientist) with hands-on experience in LLMs and agentic systems.

  In total, 52 experts were recruited from institutions including Stanford University, MIT, Google, and xAI, *etc.*.
- **Assessment protocol.** Every task t was independently assessed by at least two experts, with additional reviews to ensure that $A_e(t)$ and $H_e(t)$ ratings exhibit a standard deviation ≤ 1.

B.3 Mixed-Effects Model Regression on Worker Responses

To assess the extent to which workers' automation desire ratings reflect intrinsic task properties rather than individual demographics, we fitted a linear mixed-effects model of the form

$$y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_k \, X_{k,ij} + u_j + \varepsilon_{ij},$$
$$u_j \sim \mathcal{N}(0, \sigma_u^2), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

Variable	Coef.	Std.Err.	z	P>|z|	Variable	Coef.	Std.Err.	z	P>|z|
Intercept	2.736	0.224	12.208	0.000	llm_usage_by_type__edit[T.Monthly]	0.023	0.065	0.352	0.725
gender[T.Male]	0.070	0.037	1.909	0.056	llm_usage_by_type__edit[T.Never]	-0.185	0.076	-2.447	0.014
gender[T.Other]	-0.210	0.356	-0.591	0.555	llm_usage_by_type__edit[T.Weekly]	-0.094	0.047	-2.009	0.045
gender[T.Prefer not to say]	0.153	0.141	1.084	0.278	llm_usage_by_type__idea_generation[T.Monthly]	-0.096	0.062	-1.545	0.122
education[T.Bachelor's Degree]	0.022	0.081	0.268	0.789	llm_usage_by_type__idea_generation[T.Never]	0.072	0.076	0.952	0.341
education[T.Doctorate (e.g., PhD)]	0.236	0.113	2.086	0.037	llm_usage_by_type__idea_generation[T.Weekly]	-0.014	0.049	-0.279	0.781
education[T.High School]	0.037	0.117	0.319	0.750	llm_usage_by_type__communication[T.Monthly]	-0.042	0.066	-0.641	0.521
education[T.Master's Degree]	0.121	0.084	1.436	0.151	llm_usage_by_type__communication[T.Never]	-0.052	0.067	-0.785	0.432
education[T.Prefer not to say]	0.157	0.221	0.711	0.477	llm_usage_by_type__communication[T.Weekly]	-0.051	0.047	-1.084	0.278
education[T.Professional Degree (e.g., MD, JD)]	-0.087	0.130	-0.670	0.503	llm_usage_by_type__analysis[T.Monthly]	0.078	0.067	1.158	0.247
education[T.Some College, No Degree]	-0.125	0.093	-1.340	0.180	llm_usage_by_type__analysis[T.Never]	-0.013	0.080	-0.165	0.869
experience[T.3-5 years]	0.100	0.060	1.662	0.097	llm_usage_by_type__analysis[T.Weekly]	0.126	0.052	2.415	0.016
experience[T.6-10 years]	0.141	0.065	2.180	0.029	llm_usage_by_type__decision[T.Monthly]	0.022	0.066	0.329	0.742
experience[T.Less than 1 year]	-0.073	0.120	-0.610	0.542	llm_usage_by_type__decision[T.Never]	-0.085	0.077	-1.107	0.268
experience[T.More than 10 years]	0.229	0.067	3.415	0.001	llm_usage_by_type__decision[T.Weekly]	-0.122	0.054	-2.268	0.023
llm_familiarity[T.I have some experience using them.]	-0.266	0.154	-1.726	0.084	llm_usage_by_type__coding[T.Monthly]	-0.021	0.064	-0.328	0.743
llm_familiarity[T.I use them regularly.]	-0.266	0.156	-1.711	0.087	llm_usage_by_type__coding[T.Never]	0.057	0.067	0.850	0.396
llm_familiarity[T.No, I've never heard of them.]	1.370	0.611	2.241	0.025	llm_usage_by_type__coding[T.Weekly]	-0.130	0.059	-2.228	0.026
llm_use_in_work[T.Yes, I use them every day in my work.]	0.077	0.053	1.457	0.145	llm_usage_by_type__system_design[T.Monthly]	-0.171	0.073	-2.334	0.020
llm_use_in_work[T.Yes, I use them every week in my work.]	0.025	0.051	0.493	0.622	llm_usage_by_type__system_design[T.Never]	-0.173	0.073	-2.362	0.018
race[T.Black]	-0.077	0.069	-1.127	0.260	llm_usage_by_type__system_design[T.Weekly]	-0.131	0.070	-1.858	0.063
race[T.Hispanic]	-0.024	0.093	-0.260	0.795	llm_usage_by_type__data_processing[T.Monthly]	-0.026	0.060	-0.430	0.667
race[T.Native American]	-0.209	0.166	-1.257	0.209	llm_usage_by_type__data_processing[T.Never]	0.099	0.061	1.633	0.102
race[T.Other]	0.233	0.088	2.657	0.008	llm_usage_by_type__data_processing[T.Weekly]	0.026	0.052	0.494	0.622
race[T.White]	-0.139	0.062	-2.226	0.026	ai_tedious_work_attitude[T.Somewhat agree]	0.385	0.096	3.991	0.000
income[T.165K-209K]	0.209	0.080	2.621	0.009	ai_tedious_work_attitude[T.Somewhat disagree]	0.538	0.128	4.206	0.000
income[T.209K-529K]	0.212	0.102	2.076	0.038	ai_tedious_work_attitude[T.Strongly agree]	0.685	0.097	7.098	0.000
income[T.30-60K]	0.275	0.063	4.332	0.000	ai_tedious_work_attitude[T.Strongly disagree]	0.439	0.126	3.482	0.000
income[T.529K+]	0.687	0.181	3.792	0.000	ai_job_importance_attitude[T.Somewhat agree]	-0.027	0.050	-0.534	0.593
income[T.60-86K]	0.113	0.066	1.711	0.087	ai_job_importance_attitude[T.Somewhat disagree]	-0.040	0.056	-0.720	0.472
income[T.86K-165K]	0.089	0.064	1.403	0.161	ai_job_importance_attitude[T.Strongly agree]	-0.087	0.065	-1.351	0.177
income[T.Prefer not to say]	-0.035	0.107	-0.329	0.742	ai_job_importance_attitude[T.Strongly disagree]	-0.012	0.071	-0.172	0.864
political_affiliation[T.Green Party]	-0.198	1.157	-0.171	0.864	ai_daily_interest_attitude[T.Somewhat agree]	0.128	0.074	1.724	0.085
political_affiliation[T.Independent]	-0.074	0.053	-1.403	0.161	ai_daily_interest_attitude[T.Somewhat disagree]	0.027	0.103	0.262	0.793
political_affiliation[T.Libertarian]	0.193	0.125	1.539	0.124	ai_daily_interest_attitude[T.Strongly agree]	0.415	0.078	5.308	0.000
political_affiliation[T.No political affliation]	0.063	0.057	1.092	0.275	ai_daily_interest_attitude[T.Strongly disagree]	0.069	0.109	0.633	0.526
political_affiliation[T.Other]	0.208	0.157	1.328	0.184	ai_suffering_attitude[T.Somewhat agree]	-0.299	0.058	-5.191	0.000
political_affiliation[T.Prefer not to answer]	0.136	0.070	1.942	0.052	ai_suffering_attitude[T.Somewhat disagree]	-0.047	0.055	-0.852	0.394
political_affiliation[T.Republican]	0.054	0.048	1.140	0.254	ai_suffering_attitude[T.Strongly agree]	-0.438	0.074	-5.928	0.000
llm_usage_by_type__information_access[T.Monthly]	-0.082	0.065	-1.256	0.209	ai_suffering_attitude[T.Strongly disagree]	-0.023	0.059	-0.391	0.696
llm_usage_by_type__information_access[T.Never]	-0.125	0.077	-1.624	0.104	age	-0.003	0.002	-2.033	0.042
llm_usage_by_type__information_access[T.Weekly]	0.098	0.046	2.149	0.032					

Table 2: Fixed-effects estimates from the mixed-effects regression predicting automation desire ratings.

where y_{ij} is the automation desire rating provided by worker i on task j, $X_{k,ij}$ are the K demographic and attitude covariates (age, gender, education, experience, LLM familiarity/use, income, political affiliation, LLM-usage subtypes, and AI-attitude scales), β_k their fixed effects, u_j a task-specific random intercept, and ε_{ij} the residual error. The model was estimated by REML using the `statsmodels` MixedLM interface in Python.

The fitted model (see Table 2) yielded the following variance-component estimates: $\widehat{\sigma}_u^2 = 0.066, \widehat{\sigma}_\varepsilon^2 = 1.254$, implying an intraclass correlation $\mathrm{ICC} = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + \widehat{\sigma}_\varepsilon^2} = 0.050$, *i.e.*, roughly 5% of total variance in automation desire is attributable to between-task differences. This "small-to-moderate" ICC confirms that task-level properties carry a real signal in workers' automation desire ratings.

We further conduct an additional analysis focusing on tasks where the task-level signal is stronger. We identified "high-signal" tasks based on two criteria: (1) low within-task variance in worker ratings (below the median variance), and (2) strong separate from the overall mean desire rating (above the median absolute deviation). This filtering retained 238 tasks. Fitting on this high-signal subset, the same linear mixed-effects model yields an ICC of 0.405. Critically, the distribution of tasks across the four zones remained remarkably stable: in the high-signal subset, 41.60% of tasks fall in the Automation "Green Light" Zone (vs. 43.13% in the full dataset), 31.09% in the Automation "Red Light" Zone (vs. 31.64%), 15.13% in the Low Priority zone (vs. 12.56%), and 12.18% in the R&D Opportunity zone (vs. 12.68%). Together, these results indicate that, after controlling for a broad array of individual differences, task identity still explains a meaningful fraction of variance in automation desire. In the main paper, we use the average ratings for analysis.

C Usage of External Data and Resources

C.1 Occupational Information Network (O*NET)

We source occupational tasks in this study from O*NET (version 29.2) Task Statements[3]. The O*NET database is a regularly updated database containing information about occupations across the United States. O*NET maps occupations to knowledge, skills, and abilities on different levels of granularity,

[3]https://www.onetcenter.org/dictionary/29.2/excel/task_statements.html

as well as to tasks and detailed work activities. In O*NET, tasks are specific work activities that can be unique for each occupation. In total, there are 18,796 task statements spanning across 923 occupations.

Each task statement is associated with O*NET-SOC Code, Title (*i.e.*, occupation), and Task Type (*i.e.*, "Core"/"Supplementary"). Moreover, O*NET provides annotations of task categories based on the frequency of the task in seven categories ("Yearly or less", " More than yearly", "More than monthly", "More than weekly", "Daily", " Several times daily", " Hourly or more")[4]. As this work focuses on digital AI agents, we filter these task statements based on the following criteria:

1. The occupation mainly involves using computers in its work as judged by `gpt-4o`.
2. The task can be finished on the computer as judged by `gpt-4o`
3. The task does not miss annotation for "Core"/"Supplementary".
4. The task will be done more than monthly.

After filtering, there are 2,131 tasks remaining, spanning across 287 occupations.

Prompt for Filtering Occupations

Does this job mainly involve using computers?
Answer format: "Yes" or "No"
Occupation: {occupation}

Prompt for Filtering Occupational Tasks

For {occupation}, is it possible to finish its work-related task on a computer?
Answer format: "Yes" or "No"
Task: {task}

C.2 Occupational Employment and Wage Statistics by U.S. Bureau of Labor Statistics

We use occupational employment and wage statistics from the U.S. Bureau of Labor Statistics (BLS) to contextualize our findings with economic data. Specifically, we draw on data from the BLS May 2024 Occupational Employment and Wage Statistics Query System[5] to obtain the "Annual Mean Wage" and "Employment" (*i.e.*, number of employees) fields for each occupation in our database. These fields, combined with the collected first-hand data in WORKBank, inform the analysis presented in Figure 9.

C.3 Claude.ai Usage Data

To shed light on the relationship between current large language model (LLM) usage and future worker desires, we compare WORKBank automation desires with existing Claude.ai usage data from the Anthropic Economic Index Handa et al. (2025). The Anthropic dataset reports Claude usage at the task level, following O*NET task definitions. These standardized definitions allow us to directly map tasks from the Anthropic dataset to corresponding tasks in the WORKBank database, enabling a structured comparison across both sources. Our data shows that the top 10 occupations with the highest average automation desire represent only 1.26% of total usage (Figure 3 **c**).

C.4 Y Combinator (YC) Company and AI Agent Research Paper Data

We collect data on Y Combinator (YC) companies and AI agent research papers to assess how current investment and research efforts align with the desire–capability landscape revealed by the WORKBank database (Figure 4). To enable this analysis, we developed an LLM-assisted pipeline that systematically identifies and maps relevant YC startups and academic publications to specific

[4] https://www.onetcenter.org/dictionary/29.2/excel/task_categories.html
[5] `https://www.bls.gov/oes/tables.htm`

occupational tasks in the WORKBank database. We use `gpt-4.1-mini` for annotation, which achieves 92% precision on the final task mapping, validated against author annotations on 100 randomly sampled examples.

YC Company Collection Process The full list of YC companies was retrieved on April 28, 2025, from the official YC website[6]. The initial dataset comprised 5,156 companies. Company descriptions were collected from their respective YC detail pages and filtered using `gpt-4.1-mini`. Each company description was assessed using a binary classification prompt (see Prompt for YC Company Classifier) to determine whether the company is AI-relevant. This process identified 1,723 AI-related companies.

AI Agent Research Paper Collection Process Academic papers were obtained from the arXiv official website[7], with a submission cut-off date of April 24, 2025. Papers were first screened by keyword: their title or abstract must contain "language model" (case-insensitive) and either "agent" or "system." This yielded an initial set of 17,064 papers. This set was refined with `gpt-4.1-mini` using a checklist (see Prompt for AI Agent Paper Classifier) to verify that each paper: (i) describes tasks extending *beyond single-turn raw text completion*, (ii) presents an implemented pipeline, (iii) conducts task-level evaluations, and (iv) involves a realistic task scenario. This process produced a final selection of 1,222 papers. For each paper passing this filter, we performed an additional round of task extraction (see Prompt for Paper Task Extractor); the extracted task statement serves as the paper's representative description in the subsequent mapping step.

Task Mapping Process For each YC company and AI agent paper identified through the above processes, we again employed `gpt-4.1-mini` to perform binary classification of their applicability to each occupational task in the WORKBank database. For YC company-to-task mapping, see Prompt for Company-to-Task Classifier; for agent paper-to-task mapping, see Prompt for Paper-to-Task Classifier.

Prompt for YC Company Classifier

You will be presented with a company description. Your job is to classify if the company is an AI related company or not.
An AI-related company is defined as a company that is involved in the research, development, or application of AI.
Output a boolean value.

The description of the company: {company_description}
The boolean value indicating if the company is an AI-related company:

[6] `https://www.ycombinator.com/companies`
[7] `http://arxiv.org`

Prompt for AI Agent Paper Classifier

You will see a paper TITLE and ABSTRACT. Return True if the paper's main contribution is an LLM-driven AGENT SYSTEM, else False.

Agent-system criteria (must ALL hold):
1. Beyond single turn raw text completion - The LLM's output decides what action or module happens next (planner/controller role), beyond single turn raw text completion.
2. Implemented pipeline - A complete system is actually built and run, not merely proposed.
3. Task-level evaluation - The paper reports results on the entire system performing its task (automatic metrics or user studies).
4. Realistic task - The task matches a plausible real-world workflow or a credible simulated environment.

If ANY criterion is missing, output "False".

———

The title of the paper: {paper_title}
The abstract of the paper: {paper_abstract}
Whether the paper is an LLM-driven agent system:

Prompt for Paper Task Extractor

You will be given the TITLE and ABSTRACT of a research paper describing an agent system. Extract the core task the paper addresses and express it in one concise sentence that begins exactly with: "The paper proposes an agent system to solve the task of...". Your output should be only that sentence, capturing the primary objective of the system.

———

The title of the paper: {paper_title}
The abstract of the paper: {paper_abstract}
The core task the paper addresses, expressed in one concise sentence:

Prompt for Company-to-Task Classifier

You will receive a brief description of a company and its product/service and an indexed list of workflows. For each workflow, decide whether workers involved in that workflow are a primary or explicitly intended user group of the company's offering. If the link is merely incidental, indirect, or speculative, mark it False. When in doubt, default to False.

———

The description of the company: {company_description}
The list of workflows and their descriptions: {workflows}
The dictionary of occupations and whether they are the target users of the company:

Prompt for Paper-to-Task Classifier

You will receive a brief description of a task proposed in the paper and an indexed list of workflows. For each workflow, decide whether the task is related to the workflow and the research is applicable to the workflow. If the link is merely incidental, indirect, or speculative, mark it False. When in doubt, default to False.

———

The description of the task proposed in the paper: {task_description}
The list of workflows and their descriptions: {workflows}
The dictionary of workflows and whether they are related to the task:

D WORKBANK STATISTICS

As detailed in Section 3.4, we retained only occupations with at least ten worker responses from January to May 2025, yielding 1,500 individual assessments across 104 occupations. In this section, we present detailed statistics on occupational coverage, participant demographics, and recruitment sources in the WORKBank database.

D.1 FULL LIST OF INCLUDED OCCUPATIONS

Occupation (O*NET-SOC Title)	N	Occupation (O*NET-SOC Title)	N
Customer Service Representatives	53	Biostatisticians	12
Sales Representatives, Wholesale and Manufacturing, Technical and Scientific Products	35	Quality Control Analysts	12
Accountants and Auditors	31	Advertising and Promotions Managers	12
Clinical Research Coordinators	30	Budget Analysts	11
Medical and Health Services Managers	30	Public Relations Specialists	11
Computer Programmers	28	Financial Managers	11
Web Developers	27	Logistics Analysts	11
Computer and Information Systems Managers	27	Medical Transcriptionists	11
Computer Systems Analysts	25	Radiologists	11
Purchasing Managers	24	Eligibility Interviewers, Government Programs	11
Business Teachers, Postsecondary	24	Court, Municipal, and License Clerks	11
Information Technology Project Managers	24	Securities, Commodities, and Financial Services Sales Agents	11
Financial Quantitative Analysts	23	Sustainability Specialists	11
Insurance Claims and Policy Processing Clerks	21	Web Administrators	11
Computer and Information Research Scientists	20	Geographers	11
Legal Secretaries and Administrative Assistants	19	Management Analysts	11
Secretaries and Administrative Assistants, Except Legal, Medical, and Executive	18	Bioinformatics Scientists	11
Human Resources Specialists	18	News Analysts, Reporters, and Journalists	11
Human Resources Managers	17	Compliance Officers	11
Business Intelligence Analysts	17	Video Game Designers	11
Purchasing Agents, Except Wholesale, Retail, and Farm Products	17	Lawyers	11
Statisticians	16	Proofreaders and Copy Markers	10
Personal Financial Advisors	16	Architects, Except Landscape and Naval	10
Production, Planning, and Expediting Clerks	16	Art Directors	10
Computer User Support Specialists	16	Data Entry Keyers	10
Search Marketing Strategists	16	Public Safety Telecommunicators	10
Architectural and Civil Drafters	15	Producers and Directors	10
Media Technical Directors/Managers	15	Precision Agriculture Technicians	10
Editors	15	Credit Counselors	10
Training and Development Specialists	15	Tax Preparers	10
Transportation Planners	15	Health Informatics Specialists	10
Appraisers and Assessors of Real Estate	14	Aerospace Engineers	10
Fundraisers	14	Medical Secretaries and Administrative Assistants	10
Credit Analysts	14	Technical Writers	10
Writers and Authors	14	Reservation and Transportation Ticket Agents and Travel Clerks	10
Graphic Designers	14	Petroleum Engineers	10
Civil Engineers	14	Molecular and Cellular Biologists	10
Information Security Analysts	14	Social Science Research Assistants	10
Online Merchants	14	Financial Examiners	10
Mathematicians	13	Telemarketers	10
Travel Agents	13	Payroll and Timekeeping Clerks	10
Photographers	13	Quality Control Systems Managers	10
Software Quality Assurance Analysts and Testers	13	Judicial Law Clerks	10
Bookkeeping, Accounting, and Auditing Clerks	13	Database Administrators	10
Desktop Publishers	12	Power Distributors and Dispatchers	10
Cost Estimators	12	Energy Engineers, Except Wind and Solar	10
Regulatory Affairs Managers	12	Network and Computer Systems Administrators	10
Mechanical Engineers	12	Librarians and Media Collections Specialists	10
Supply Chain Managers	12	Treasurers and Controllers	10
Clinical Data Managers	12	Biofuels Production Managers	10
Computer Systems Engineers/Architects	12	Mechanical Engineering Technologists and Technicians	10
Loan Officers	12	Computer Network Support Specialists	10

Table 3: List of occupations covered in the WORKBank database and the number of survey participants from each occupation.

D.2 Coverage of the U.S. Workforce

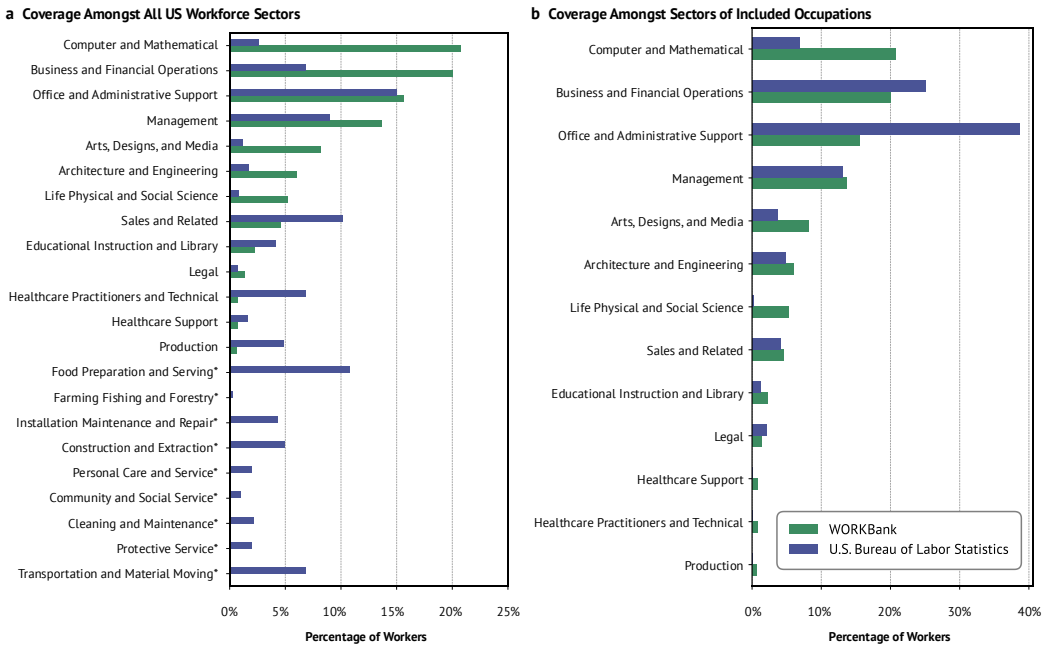

Figure 6: **Sector-level distribution of workers in the WORKBank database compared to U.S. workforce statistics from the Bureau of Labor Statistics. a**, Comparison between WORKBank worker distribution and the U.S. workforce employment statistics across all sectors (sectors not included in WORKBank marked with an asterisk). **b**, Comparison between WORKBank worker distribution and the U.S. workforce employment statistics limited to the 104 occupations included in our database.

We evaluate the representativeness of WORKBank by comparing its sector-level distribution with U.S. workforce data from the Bureau of Labor Statistics[8]. Figure 6 presents a breakdown of workforce coverage by sector, contrasting our database with the full U.S. workforce (**a**) and with the workforce restricted to the 104 occupations included in our database (**b**). The comparisons suggest that our database captures a broad and representative cross-section of the U.S. workforce at the sector level.

D.3 Domain Worker Demographic Information

We compare the demographic profile of domain workers in WORKBank with that of the U.S. workforce. U.S. workforce demographic data are sourced from the 2024 Annual Averages of the Bureau of Labor Statistics (BLS) Current Population Survey[9]. To ensure a fair comparison, we filter the BLS data to include only the 104 occupations represented in our database. Figure 7 shows the breakdown across key demographic dimensions. Our participant pool has a comprehensive demographic coverage, with a tendency toward younger age groups.

D.4 Recruitment Sources

To obtain worker assessments, we mainly recruited domain workers through Prolific, Upwork, and LinkedIn. Table 4 reports the breakdown of recruited domain workers before data filtering. For expert assessments, we recruited 52 AI experts through word of mouth. These AI experts include PhD students (25), postdocs (6), professors (6), and industry practitioners (15) spanning across different career stages.

[8] https://www.bls.gov/oes/tables.htm
[9] https://www.bls.gov/cps/tables.htm

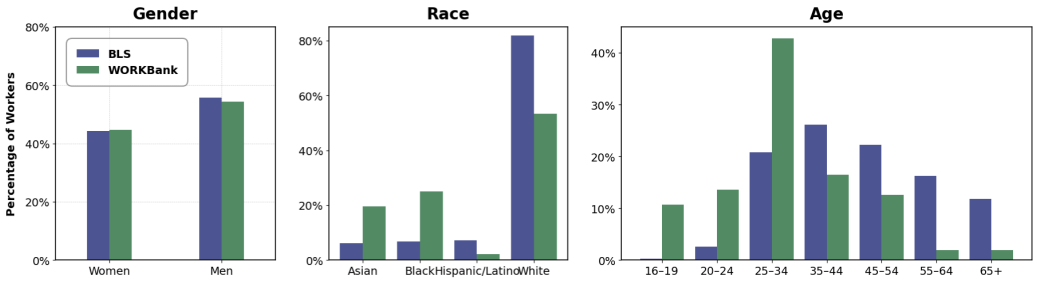

Figure 7: Demographic distribution of worker participants in our study compared to U.S. workforce demographics for the same set of occupations covered in the WORKBank database, based on data from the Bureau of Labor Statistics.

Recruitment Source	Description	# Participants
Prolific	We applied the following prescreen criteria: Employment Status (Full-Time or Part-Time); Nationality (United States); Fluent Languages (English); Method of Work (computer or mobile device used for $\geq$ 25% of work). We also used the "Industry" prescreen filter to ensure participants could only select occupations matching their Prolific profile industry.	1060
Upwork	We required Upwork workers to be located in the United States.	184
Word of Mouth	We distributed the survey link to LinkedIn users located in the United States and distributed flyers within the United States.	434

Table 4: Breakdown of recruitment sources for domain workers in this study. The numbers shown reflect all participants recruited before data filtering. Breakdowns by recruitment source after data filtering are not available, as we did not link participant personal information with data points in WORKBank to preserve privacy.

E ADDITIONAL RESULTS

E.1 TOP 20 TASKS WORKERS WANT AUTOMATED

Task	Average Automation Desire
Tax Preparers: Schedule appointments with clients.	5.00
Public Safety Telecommunicators: Maintain files of information relating to emergency calls, such as personnel rosters and emergency call-out and pager files.	4.67
Payroll and Timekeeping Clerks: Issue and record adjustments to pay related to previous errors or retroactive increases.	4.60
Desktop Publishers: Convert various types of files for printing or for the Internet, using computer software	4.50
Online Merchants: Create or maintain database of customer accounts.	4.50
Quality Control Systems Managers: Direct the tracking of defects, test results, or other regularly reported quality control data.	4.50
Statisticians: Report results of statistical analyses, including information in the form of graphs, charts, and tables.	4.50
Computer User Support Specialists: Maintain records of daily data communication transactions, problems and remedial actions taken, or installation activities.	4.50
Online Merchants: Calculate revenue, sales, and expenses, using financial accounting or spreadsheet software.	4.40
Data Entry Keyers: Store completed documents in appropriate locations.	4.33
Petroleum Engineers: Maintain records of drilling and production operations.	4.33
Logistics Analysts: Apply analytic methods or tools to understand, predict, or control logistics operations or processes.	4.33
Court, Municipal, and License Clerks: Instruct parties about timing of court appearances.	4.33
Data Entry Keyers: Maintain logs of activities and completed work.	4.25
Compliance Officers: Prepare correspondence to inform concerned parties of licensing decisions or appeals processes.	4.25
Web Developers: Back up files from Web sites to local directories for instant recovery in case of problems.	4.20
Web Administrators: Back up or modify applications and related data to provide for disaster recovery.	4.20
Bioinformatics Scientists: Manipulate publicly accessible, commercial, or proprietary genomic, proteomic, or post-genomic databases.	4.17
Network and Computer Systems Administrators: Perform routine network startup and shutdown procedures, and maintain control records.	4.17
Computer and Information Research Scientists: Approve, prepare, monitor, and adjust operational budgets.	4.17

Table 5: Top 20 tasks with highest average automation desire.

E.2 BOTTOM 20 TASKS WORKERS WANT AUTOMATED

Task	Average Automation Desire
Reservation and Transportation Ticket Agents and Travel Clerks: Trace lost, delayed, or misdirected baggage for customers.	1.50
Logistics Analysts: Contact potential vendors to determine material availability.	1.50
Editors: Write text, such as stories, articles, editorials, or newsletters.	1.60
Reservation and Transportation Ticket Agents and Travel Clerks: Contact customers or travel agents to advise them of travel conveyance changes or to confirm reservations.	1.67
Video Game Designers: Provide feedback to designers and other colleagues regarding game design features.	1.67
Librarians and Media Collections Specialists: Code, classify, and catalog books, publications, films, audio-visual aids, and other library materials, based on subject matter or standard library classification systems.	1.67
Editors: Plan the contents of publications according to the publication's style, editorial policy, and publishing requirements.	1.67
Database Administrators: Write and code logical and physical database descriptions and specify identifiers of database to management system, or direct others in coding descriptions.	1.67
Graphic Designers: Key information into computer equipment to create layouts for client or supervisor.	1.67
Mechanical Engineering Technologists and Technicians: Calculate required capacities for equipment of proposed system to obtain specified performance and submit data to engineering personnel for approval.	1.67
Secretaries and Administrative Assistants, Except Legal, Medical, and Executive: Establish work procedures or schedules and keep track of the daily work of clerical staff.	1.67
Graphic Designers: Review final layouts and suggest improvements, as needed.	1.71
Graphic Designers: Prepare illustrations or rough sketches of material, discussing them with clients or supervisors and making necessary changes.	1.71
Mechanical Engineering Technologists and Technicians: Interpret engineering sketches, specifications, or drawings.	1.75
Accountants and Auditors: Prepare, examine, or analyze accounting records, financial statements, or other financial reports to assess accuracy, completeness, and conformance to reporting and procedural standards.	1.75
Editors: Allocate print space for story text, photos, and illustrations according to space parameters and copy significance, using knowledge of layout principles.	1.75
Producers and Directors: Cut and edit film or tape to integrate component parts into desired sequences.	1.75
Graphic Designers: Create designs, concepts, and sample layouts, based on knowledge of layout principles and esthetic design concepts.	1.78
Librarians and Media Collections Specialists: Locate unusual or unique information in response to specific requests.	1.80
Editors: Assign topics, events and stories to individual writers or reporters for coverage.	1.80

Table 6: Bottom 20 tasks with lowest average automation desire.

E.3 Full Human Agency Scale Results

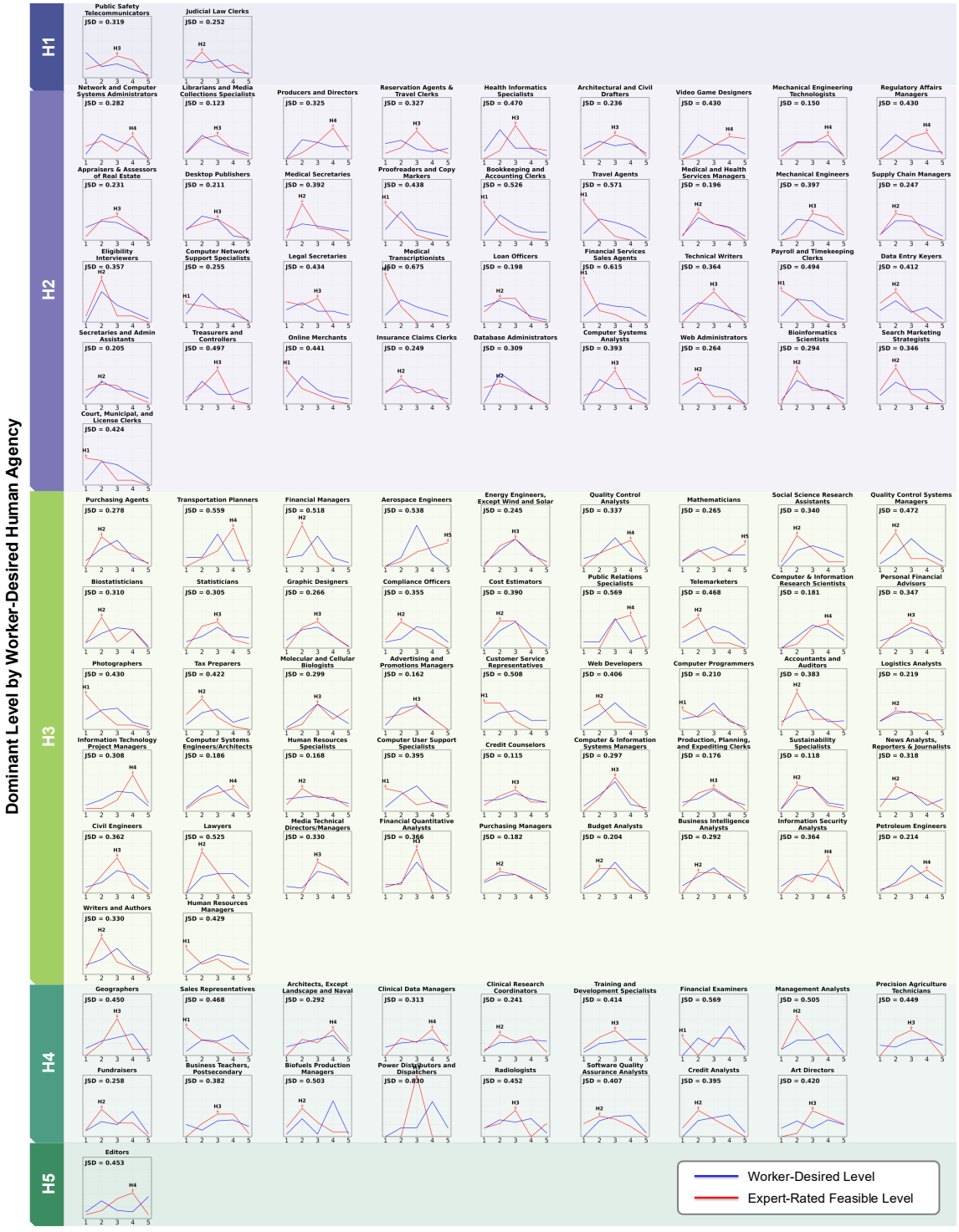

Figure 8: **Distributions of Human Agency Scale (HAS) levels.** The Jensen–Shannon Divergence (JSD) quantifies the divergence between the distribution of worker-desired HAS levels ($H_w(t)$) and expert-assessed feasible HAS levels ($H_e(t)$).

E.4 Top 10 Occupations By Worker-Expert Discrepancies in HAS Ratings

Occupation	Dominant HAS Level By Worker Desire	Dominant HAS Level By AI Expert Assessment	JSD
Power Distributors and Dispatchers	H4	H3	0.830
Medical Transcriptionists	H2	H1	0.675
Securities, Commodities, and Financial Services Sales Agents	H2	H1	0.615
Travel Agents	H2	H1	0.571
Financial Examiners	H4	H1	0.569
Public Relations Specialists	H3	H4	0.569
Transportation Planners	H3	H4	0.559
Aerospace Engineers	H3	H5	0.538
Bookkeeping, Accounting, and Auditing Clerks	H2	H1	0.526
Lawyers	H3	H2	0.525

Table 7: Top 10 occupations with the largest discrepancies between the worker-desired Human Agency Scale levels $H_w(t)$ and AI expert-assessed feasible levels $H_e(t)$. The discrepancy is measured by the Jensen–Shannon divergence (JSD) between these two distributions, where distributions are computed based on all annotated tasks within each occupation.

E.5 The Potential Shift of Core Human Skills

Variation in Human Agency Scale (HAS) across occupations suggests that certain types of human work are more susceptible to automation, while others hold greater potential for augmentation. To this end, we examine the characteristics of tasks that require high human agency to understand how AI agents may shift skill demands.

Translating task-level changes into implications for education and skill training has long been a key lens for analyzing technological transformation, notably pioneered by Autor et al. (2003) in the wave of computers. To operationalize this lens, we use the O*NET database to align each task with its related skills (Generalized Work Activities). In order to create broad groups of common skills across multiple tasks, O*NET defines three levels of work activities: Generalized (GWA), Intermediate (IWA), and Detailed Work Activities (DWA). Each task in the O*NET database is mapped to one or more DWA's, which then corresponds to exactly one GWA. For example, the task "Compile financial data to prepare quarterly budget reports" is mapped to the DWA "Prepare financial documents, reports, or budgets," which is associated with the skill (GWA) "Documenting/Recording Information". As another example, the task "Financial Managers: Approve, reject, or coordinate the approval or rejection of lines of credit or commercial, real estate, or personal loans" is mapped to "Making Decisions and Solving Problems" and "Guiding, Directing, and Motivating Subordinates". Using this mapping, we are able to directly match tasks to their corresponding skills.

We compute the average expert-assessed human agency level $H_e(t)$ for each skill to estimate the degree of human involvement required as AI agents enter the workplace. We also compute the average wage for each skill, using data from the U.S. Bureau of Labor Statistics (Appendix C.2). Wage serves as a proxy for the current economic value of each skill (Dey & Loewenstein, 2019).

As shown in Figure 9, by comparing skill rankings by average wage and required human agency, our analysis reveals three emerging trends:

1. **Shrinking demand for information-processing skills.** Skills related to analyzing data and updating knowledge—while common in today's high-wage occupations (as shown in the left side of Figure 9 in red color)—are less prominent in tasks that demand high human agency.
2. **Greater emphasis on interpersonal and organizational skills.** Skills involving human interaction, coordination, and resource monitoring are more frequently associated with high-HAS tasks (as shown in the left side of Figure 9 in green color), even if they are not currently prioritized in wage-based evaluations.
3. **High-agency skills span diverse aspects.** The top 10 skills with the highest average required human agency encompass a broad range, from interpersonal and organizational abilities to decision-making and quality judgment.



Figure 9: **Comparing skill rankings by average wage and required human agency.** Each line represents a skill (Generalized Work Activity) mapped from O*NET tasks. Based on the skill-task mapping, we compute the average wage using data from the U.S. Bureau of Labor Statistics (May 2024) and the average expert-assessed human agency level $H_e(t)$ to indicate the degree of human involvement required as AI agents enter the workplace. Skills are ranked by average wage (left) and average required human agency (right). The figure highlights the top five skills with the largest upward (green) and downward (red) shifts in rank, suggesting a potential shift in valued workplace skills—from information processing toward interpersonal and organizational competencies.

These findings provide early signals of how AI agent integration may reshape core occupational competencies. As workplace AI agents continue to evolve, longitudinal tracking of task-level changes could yield further insights into how human roles and required skills evolve.

Table 8 provides the full descriptions of the skills shown in Figure 9, ordered by decreasing required human agency (Tsacoumis & Willison, 2010). We exclude physical skills (*i.e.*, "Identifying Objects, Actions, and Events," "Performing General Physical Activities," "Controlling Machines and Processes," "Inspecting Equipment, Structures, or Materials," "Handling and Moving Objects," "Repairing and Maintaining Mechanical Equipment") and "Working with Computers," as the audit focuses on tasks that are likely to be exposed to digital AI agents.

F Audio Response Analysis

F.1 LLM-based Topic Modeling

To analyze our audio transcripts collected from workers, we use LLooM (Lam et al., 2024), an LLM-based topic modeling tool that takes unstructured text to extract high-level concepts.

Using LlooM, we first apply a distillation step with `Claude 3 Opus` on the audio transcripts. This step filters the text to retain only the most relevant quotes based on a seed prompt and performs a

Skill	# Supported Tasks in WORKBank	Average Wage Rank	Average Required Human Agency Rank
Organizing, Planning, and Prioritizing Work: Developing specific goals and plans to prioritize, organize, and accomplish your work.	6	11	1
Training and Teaching Others: Identifying the educational needs of others, developing formal educational or training programs or classes, and teaching or instructing others.	4	21	2
Staffing Organizational Units: Recruiting, interviewing, selecting, hiring, and promoting employees in an organization.	6	6	3
Updating and Using Relevant Knowledge: Keeping up-to-date technically and applying new knowledge to your job.	19	2	4
Developing Objectives and Strategies: Establishing long-range objectives and specifying the strategies and actions to achieve them.	17	3	5
Guiding, Directing, and Motivating Subordinates: Providing guidance and direction to subordinates, including setting performance standards and monitoring performance.	49	4	6
Judging the Qualities of Objects, Services, or People: Assessing the value, importance, or quality of things or people.	47	5	7
Communicating with Supervisors, Peers, or Subordinates: Providing information to supervisors, coworkers, and subordinates by telephone, in written form, e-mail, or in person.	35	12	8
Providing Consultation and Advice to Others: Providing guidance and expert advice to management or other groups on technical, systems-, or process-related topic	56	9	9
Thinking Creatively: Developing, designing, or creating new applications, ideas, relationships, systems, or products, including artistic contributions.	117	7	10
Interpreting the Meaning of Information for Others: Translating or explaining what information means and how it can be used.	16	13	11
Making Decisions and Solving Problems: Analyzing information and evaluating results to choose the best solution and solve problems.	52	10	12
Monitoring Processes, Materials, or Surroundings: Monitoring and reviewing information from materials, events, or the environment, to detect or assess problems.	40	8	13
Assisting and Caring for Others: Providing personal assistance, medical attention, emotional support, or other personal care to others such as coworkers, customers, or patients.	7	26	14
Getting Information: Observing, receiving, and otherwise obtaining information from all relevant sources.	58	15	15
Monitoring and Controlling Resources: Monitoring and controlling resources and overseeing the spending of money.	23	25	16
Analyzing Data or Information: Identifying the underlying principles, reasons, or facts of information by breaking down information or data into separate parts.	89	1	17
Selling or Influencing Others: Convincing others to buy merchandise/goods or to otherwise change their minds or actions.	8	16	18
Documenting/Recording Information: Entering, transcribing, recording, storing, of maintaining information in written or electronic/magnetic form.	177	14	19
Evaluating Information to Determine Compliance with Standards: Using relevant information and individual judgment to determine whether events or processes comply with laws, regulations, or standards.	26	17	20
Communicating with People Outside the Organization: Communicating with people outside the organization, representing the organization to customers, the public, government, and other external sources. The information can be exchanged in person, in writing, or by telephone or e-mail.	8	22	21
Processing Information: Compiling, coding, categorizing, calculating, tabulating, auditing, or verifying information or data.	41	20	22
Estimating the Quantifiable Characteristics of Products, Events, or Information: Estimating sizes, distances, and quantities; or determining time, costs, resources, or materials needed to perform a work activity.	14	23	23
Performing Administrative Activities: Performing day-to-day administrative tasks such as maintaining information files and processing paperwork.	24	18	24
Performing for or Working Directly with the Public: Performing for people or dealing directly with the public. This includes serving customers in restaurants and stores, and receiving clients or guests.	8	24	25
Scheduling Work and Activities: Scheduling events, programs, and activities, as well as the work of others.	10	27	26
Establishing and Maintaining Interpersonal Relationships: Developing constructive and cooperative working relationships with others, and maintaining them over time.	1	19	27

Table 8: Full descriptions of skills covered by the WORKBank database.

Concept	Percentage of Summaries
Lack of trust in accuracy, reliability or capability	45.0%
Fear of job replacement	23.0%
Lack of AI's human qualities or capabilities	16.3%
AI not applicable or useful to specific work	15.6%

Table 9: Identified concepts with the seed prompt "The top most common fears that workers have about AI automation in their work".

summarization of the filtered text. Subsequently, we cluster these summarized quotes using the same LLM to form groups of related concepts. We manually inspect the clusters to ensure the concepts are distinct and merge any overlapping concepts as necessary. Table 9 and Table 10 present the top concept groups identified for two seed prompts, corresponding to the analyses in §4.1 and §4.3, respectively.

Concept	Percentage of Summaries
Role-based support	23.1%
Assistantship	23.0%
Automation	16.5%
Separation of Tasks between AI and Humans	12.0%

Table 10: Identified concepts with the seed prompt "If workers want AI to help, what type of imagined partnership with AI do they prefer".

F.2 Audio Response Examples

F.2.1 Full Transcripts for Direct Quotes in §4.1

Art Director with 6–10 Years of Experience:

So [my work is] basically just first of all looking through all the tasks in the sauna at the start of the day. And then once shoots and, you know, filmings are happening, looking through the footage and pictures, making sure they adhere to brand standards and guidelines and a cohesive voice.

[I spend most of my time] Looking through imagery and making sure that it is consistent and always deliverable, and making selects and things of that nature, just establishing a cohesive tone. [I use] So definitely Bridge, you know, Photo Mechanic, Capture One, Photoshop, Asana for the tasks like I mentioned earlier, you know, Gmail, things like that. [For the detailed procedure,] Yeah, looking at the imagery as it's flowing through during the shoot or, you know, filming if it's video, and then from there going through and selecting and culling and, you know, again, only sharing the best imagery that's cohesive.

[For AI use,] I don't really, unless it's, you know, in some sort of minor way to help the calling process become easier. *I don't want it to be used for content creation. If anything, I want it to be used for seamlessly maximizing workflow and, you know, making things less repetitive and tedious and arduous with workflow. No content creation.*

Art Director with 3-5 Years of Experience:

I manage some anime art projects as part of a company's public relations and community strategy for youth engagement. So I work with artists directly, manage projects and merchandise and tabling at events and all that fun stuff.

I do a lot of internal meetings just to make sure everyone's on the same page. It takes up a lot of my time. I also have to scope out projects, find artist to work with especially those we found on social media, figure out how to get in touch and work with them, work with community groups as well, do this type of stuff. And then I also help, well not directly do, but help assist in getting merchandise produced, including preparing artwork for like production and stuff like that.

I spend a lot of time in Microsoft Word, Microsoft Teams, Outlook, which, you know, my company uses Microsoft Office for everything. But then I also use, to communicate with artists, I use the apps that they use. So sometimes that's LINE, sometimes that's Discord, sometimes that's Twitter. And then I personally have my own notion for project management as well.

I spend a lot of time sculpting out projects, so I generally start brainstorming or collecting all my research, gather all my information into Notion first, and then put it into a Word document. The Word document's a little more formal, but I also like make sure that's still approachable to artists, and then I, you know, export that as a PDF. For communicating with artists, that depends on what they use, but most of them use Discord, so it's just back and forth. They send us a sketch, I send that off to my people for, like, feedback very quickly, and then I, you know, get back to them. I sometimes do have my own autonomy to, like, do the final say on what does work and what doesn't work. And then for at least getting everyone on the same page, I spend a lot of time Microsoft Teams. I obviously have to gather some meeting notes, like, write down some job stuff I want to talk about beforehand, make sure there's no surprises to people, it's just communicating and providing regular updates.

I would never use AI to like replace artists. I would be more for personal [project management] use, if anything, it's to summarize my tasks, for example, or things like improving my writing, using Apple's writing tools, where I can just revise my writing to be a little more concise, but I would never, let it brainstorm on my behalf, just because I find AI to be very poorly performing on those type of tasks.

Graphic Design with More Than 10 Years of Experience:

I do basically architecture presentation, like graphic design work, work like typically. It's like layout design, which is organizing content, image or text for AI storyline, diagram or infographics analysis diagram, render enhancement like poster processing, 3D render for a polished look. I also do topography and color scheme using professional fonts and color that align with the project's theme. I also work with board composition, arranging plan, section, elevation and perspectives. I also work digital and print formatting, which is ensuring high quality output for print or physical brands.

So, I basically do architectural presentations. Graphic design work typically includes layout design, diagram, etc. Later enhancement, like post-printing 3D renders for a polished work. I also give topography color scheme both for precision, visibility, and formality. I basically use drawing for autocad and after that for render I use Photoshop and Illustrator and for 3D render I use Lumion. So basically I spend a lot of time drafting in AutoCAD. The most common tasks just likely include creating floor plans, creating sections, creating elevation, annotation and dimensioning drawings, organizing and managing the layers.

AI can be a game-changer in data architect workflow, helping to improve efficiency, accuracy and even creativity. But I create my design by myself. For research, I use AI.

F.2.2 Transcripts from Occupations Exhibiting High Human Agency (HAS) Levels

As discussed in §4.3, the Human Agency Scale (HAS) spectrum reveals that very few occupations are characterized by a dominant HAS Level 5 (indicating essential human involvement). "Editors" is the only occupation where workers predominantly desire H5. According to AI expert assessments, only "Mathematicians" and "Aerospace Engineers" fall into this category. Below are representative transcripts for these occupations.

Editor with 3-5 Years of Experience:

I proofread and copy-edit marketing materials, mostly in the travel and tourism sector. I also do some copywriting and script writing for different ad clients and some light design work.

I look through flyers, brochures, other marketing materials, and I do several passes for mistakes in grammar, in consistency, in flow and clarity. And I make changes on the document, usually a PDF, and send them back to the client. They fix them, they send me another version, and I do several more passes until we've spent enough time and got it perfect. I mainly use Adobe products, PDFs in Adobe Reader. I also use Microsoft Word and some Adobe Suite products, mostly Illustrator and InDesign.

For copy editing, I read through whatever material the client has sent me. I do a pass for basic grammar. I do a pass for clarity and flow, often changing the copy significantly to make it sound better. For proofreading, I go through the materials. Same thing, but with less of a view toward changing the copy and more toward finding errors in grammar and consistency and even design. For copywriting, I make an outline of my ideas for the project and complete a rough draft. Then I spend some time away from it and revise until I have a polished draft for the client.

I'm resistant to using AI in my daily workflow. If I'm forced to use it, I would use it for basic grammar editing, but I would check each suggestion against my own knowledge very carefully and give it full consideration before adopting it as a change.

Editor With More Than 10 Years of Experience:

So I work in a media company, [masked], and as an editor I make sure that all JavaScript that I'm going to print are formatted correctly, the colors are accurate, and there are no typos.

So a lot of what I do involves sitting at a computer using productivity tools like Adobe Creative Suite, Canva, QuarkXPress, and using the Google Enterprise. So for email, document sharing, I use Google Docs quite a bit for my editing purposes, but I'll also receive files in PDF format. So just working with all the different tools on my computer to get my tasks done every day. So a lot of what I do involves sitting at a computer using productivity tools like Adobe Creative Suite, Canva, QuarkXPress, and using the Google Enterprise. So for email, document sharing, I use Google Docs quite a bit for my editing purposes, but I'll also receive files in PDF format. So just working with all the different tools on my computer to get my tasks done every day.

So one of the most frequent pieces of software I use is Adobe Acrobat, and that is really great for editing PDFs. The next most frequent software I use would be Google Docs. Receiving files through Google Docs, that's a great way to be able to provide updates and edits to annotate the files. And then I would say other tools like Adobe InDesign, Adobe Photoshop, QuarkXPress, Microsoft Publisher. Those are occasionally used, but that's really about all four of the frequently used programs, I would say. So one of the most frequent pieces of software I use is Adobe Acrobat, and that is really great for editing PDFs. The next most frequent software I use would be Google Docs. Receiving files through Google Docs, that's a great way to be able to provide updates and edits to annotate the files. And then I would say other tools like Adobe InDesign, Adobe Photoshop, QuarkXPress, Microsoft Publisher. Those are occasionally used, but that's really about all four of the frequently used programs, I would say.

So, when I'm reviewing PDF documents, I will use the markup tool to add comments and highlight certain sections to make sure that the wording is accurate, or if there's questions regarding the resolution of a photo, I can send that back to mark that up and say, hey, this needs to be a higher resolution photo, it won't print out correctly. So it's just a lot of manual review of every single file before it goes to print to make sure that everything is properly formatted, the colors are accurate, and it will reproduce correctly, just to make sure everything looks good for the customer. So, when I'm reviewing PDF documents, I will use the markup tool to add comments and highlight certain sections to make sure that the wording is accurate, or if there's questions regarding the resolution of a photo, I can send that back to mark that up and say, hey, this needs to be a higher resolution photo, it won't print out correctly. So it's just a lot of manual review of every single file before it goes to print to make sure that everything is properly formatted, the colors are accurate, and it will reproduce correctly, just to make sure everything looks good for the customer.

So, I'm using AI right now when it comes to email, so with the Google suite, there are Google Gemini tools that help with formatting emails. I can take a very simple format for content for email and then using that to expand those topics and make it more of a wordy email. But I would like to be able to use AI more in my proofreading and editing than I am right now, so probably within the next couple months I should be able to do that.

Mathematician With 3-5 Years of Experience:

I do number theory and algebraic geometry, mostly around long-length programs or categorical long-length programs. [In my daily work, I] read papers and write papers.

Solving a math problem, I don't know [whether there is any specific tool to use], just read papers and have an intuition of what the procedure of philosophy should be and work on it.

To be honest, I think [AI is] useless at this moment.

Mathematician With 3-5 Years of Experience:

I am studying geometric representation theory and categorical Langlands program. My work involves coming up the problem to work on and learning math tools to help me think of solutions.

I need to spend a lot of time reading papers and learning math tools. I also need to attend the seminar to find collaborators. Then I work on my problem. [In terms of tools,] I mainly use latex. I spend most of my time studying math. Papers in my field can have hundreds of pages - it takes a long time to understand and try to apply the technique.

At present, I don't think AI has any use for mathematicians, at least for DeepSeek and ChatGPT. One core question I am interested in is whether AI can come up with new stuffs that haven't been proposed before rather than solving problems people craft.

Mathematician With 6-10 Years of Experience:

I used to study number theory, in particular, p-adic Hodge theory in arithmetic geometry. Now, I work on the formalization of p-adic Hodge theory in Lean and also auto-formalization and auto theorem proving.

During formalization, I elaborate, generalize, and fill gaps in mathematical proofs. I design general fomalization frameworks and spend lots of time in Lean coding. Lean coding involves searching theorems, formalizing statements and filling in formalization details in the proof. The last part is the longest part. For auto formalization and formal theorem proving, I spend most of the time coding to establish the LLM's training pipeline and preparing data for the training. I use the interactive theorem prover Lean. I also use LeanSearch and other tools related to Lean to accelerate. I use Python for LLM training and use DeepSeek for coding and debugging.

[Here is a concrete example of my workflow:] I formalized a famous number theory definition, called the period rings of Fontaine. I first wrote down a detailed version of the mathematical statements and proofs I need. Splitting the whole formalization project into several smaller goals. For each smaller goal, I generalize and design suitable definitions and lemmas for formalization. Then I begin actual formalization using Lean. I first write down the definitions and state the theorems in Lean without proof. After this, I fill in the proofs backwards, searching the library for existing theorems to use and design patterns to mimic. During formalization, I revise the natural language proof from time to time.

I think a primary AI tool could help me in filling in searching for theorems and design patterns during formalization. A stronger AI tool would do auto-formalization of theorem statements and provide suggestions in designs. An even stronger AI tool would be able to elaborate and fill gaps in mathematical proofs and autoformalize the human proofs. Additionally, an AI tool strong in coding, debugging, and software engineering would help a lot in coding.

Aerospace Engineer With 3-5 Years of Experience:

I am an aeronautic engineer. I work in the aircraft maintenance, repairs, design aircraft. I work with [masked].

We design aircraft, develop, test, and maintain aircrafts, and the systems that operate within Earth's atmosphere, such as airplane, helicopters, drones, and missiles, though we're not into missiles, though. Our work focuses on making aircraft, machines, very safe and efficient, capable of flight. We use Computer-Aided Design as a tool for Autodex, AutoCAD, Cartier, and Solidwork. And we use Computer Fluid Dynamics. It's ANSYS Fluent, STAR-CNC-MM. What else? We use Finite Element Analysis. It's a tool we use for Nastran.

[In my opinion,] AI is going to be very awesome and it's going to make it very easier for us because most of the time, the main problem we have is detecting where the problem is in the engine, you know, so you have to do a lot of manual jobs and all that. So, but if we have AI, you can possibly tell in the dash cam or whatever, you know, you can possibly tell.

Aerospace Engineer With 1-2 Years of Experience:

I'm a current undergraduate senior and prospective master's student in aerospace engineering, working in guidance, navigation, and controls, so like more simulation, computer programming side of aerospace engineering.

Most of what I do for work has traditionally been programming simulations to evaluate vehicle performance for orbital rockets. And so most of my tasks will be either building out a part of the simulation, programming new features or new testing, or kind of similar types of modeling of different subsystems of the rocket.

In general the tools or software that I use would be Visual Studio Code for the actual programming. The companies I've worked at have used project management tools like JIRA and Confluence. I think also just a lot of internet documentation is useful. And yeah, I very occasionally would use an AI tool like ChatGPT. Generally, my process would be to understand the requirements, which would involve talking to my manager, then kind of going about kind of like pre-reading or other types of information gathering necessary for the task, actual programming, and then like unit testing and other ways of forms of validation for the programming that I completed.

Honestly, I don't use AI too much in my current workflow. I think that the only time that it could come up would be if I'm running into some type of error or bug in my program that I can't find, or kind of a quick piece of code that I could look up how to do, but it's easier to just ask a AI model to generate. But honestly, I use it very, very infrequently.