Predicting Where Generative AI Applies and Where It Disappoints: Evidence from 738 U.S. Occupations

Eddie Simmens

Iovine and Young Academy
University of Southern California
Los Angeles, California, USA
simmens@usc.edu

September 2026

Abstract

Generative AI is used across many kinds of work, but its applicability varies sharply by occupation. Most existing measures estimate where AI *could* be useful from expert judgments, task ratings, or assessments of model capabilities. This study instead examines where AI is already being applied, using Microsoft's *Working with AI* dataset of 200,000 Bing Copilot conversations mapped to occupational tasks and outcomes. Combining it with O*NET characteristics for 738 U.S. occupations, I ask whether job characteristics predict how widely generative AI is applied to an occupation's work, and in which occupations workers are less satisfied with AI output even when the AI appears to complete the requested task. Using only non-task O*NET features to avoid target leakage, I evaluate regression models with 10-fold cross-validation. An ensemble of 15 interpretable job-characteristic constructs explains 62% of the variation in AI applicability; a larger ensemble built on 30 framework-grounded constructs, partial-least-squares components, and job-title embeddings raises cross-validated R^2 to 0.72. Communication, computing, creativity, and analytical demands are associated with higher applicability; physical, fine-motor, and mechanical demands with lower. The gap between AI task completion and positive user feedback is substantially larger in physical, mechanical, and hazardous occupations ($d = 0.96$–1.29). Of six dimensions of embodied work, the gap is most strongly associated with continuous monitoring and real-time adjustment ($r = 0.52$), followed by reliance on direct sensory information and variable physical conditions. These occupation-level findings suggest conversational AI struggles most where guidance must keep pace with an evolving physical situation. Supporting such work will require systems that update as conditions change and evaluations that test whether guidance adapts.

Keywords: generative AI, occupational exposure, O*NET, predictive modeling, human–AI interaction, embodied work, workforce.

1 Introduction

Generative AI is becoming part of everyday work, but its usefulness is not evenly distributed across occupations. In some jobs, generative AI appears well suited to the work people frequently perform: drafting, summarizing, explaining, searching, coding, analyzing, or producing written and visual materials. In other jobs, its relevance is narrower, more indirect, or harder to translate into useful assistance. Understanding this variation matters for workers deciding where to invest in new skills, for employers deciding where to pilot AI systems and initiatives, and for researchers seeking to distinguish speculative claims about "AI exposure" from evidence about where generative AI is actually being used today. Much of the existing literature estimates occupational AI exposure from expert judgments, task rubrics, patent–task similarity, or model-rated assessments of what AI systems could plausibly do [12, 18]. These approaches have been useful for mapping potential exposure, but they do not directly observe real-world AI use. The recent *Working with AI* dataset from Microsoft makes a different approach possible: it measures AI applicability at the occupation level from 200,000 Bing Copilot conversations, linking observed AI interactions to O*NET work activities, tasks, and occupations [38]. This makes it possible to ask not only which occupations seem theoretically exposed to generative AI, but which occupational characteristics predict where generative AI is already being applied to work.

In this study, I merge Microsoft's usage-derived AI applicability score with O*NET job-characteristic data for 738 U.S. occupations. The analysis asks two questions. First, can measurable job characteristics predict AI applicability (RQ1)? Second, in which kinds of occupations are workers less satisfied with AI outputs even when the AI appears to complete the task they requested? The second question is examined in two steps: whether the completion–satisfaction gap is larger in some kinds of work than in others (RQ2), and, where it is, which aspects of that work account for it (RQ3). The answers to these questions are valuable because they go beyond a single exposure score. For workers and career guidance professionals, an interpretable model can show which skills and work conditions are associated with greater or lower AI applicability. For employers, the same model can help identify where generative AI is likely to fit existing work and where deployment of novel generative AI initiatives may require more caution. The completion–satisfaction gap is especially important because an AI tool may appear to complete a task while still failing to meet users' needs; knowing where that mismatch concentrates—and which features of the work account for it—is directly useful for deciding where AI fits.

The study uses a deliberately transparent occupation-level design. I merge two public datasets, summarize high-dimensional O*NET features into interpretable job-characteristic constructs, and exclude O*NET task and work-activity statements to avoid target leakage. I then train ensemble models to predict AI applicability, test whether the completion–satisfaction gap widens in hands-on, embodied occupations, and decompose embodied work into narrower constructs to identify which of its aspects account for the gap. The contribution is an interpretable account of where generative AI currently lands in the labor market, what job characteristics help explain that pattern, and which features of the work are most associated with the mismatch between apparent AI task completion and user satisfaction.

2 Related Work

2.1 From automation risk to generative-AI exposure

Early work on occupational automation focused on estimating which jobs were vulnerable to computerization. Frey and Osborne [18] produced one of the most influential occupation-level automation-risk measures by training a classifier on hand-labeled occupations and extrapolating probabilities across detailed O*NET occupations. Their study established occupation-level prediction as a useful way to analyze technological change, but its goal was broad computerization risk rather than generative-AI applicability. Subsequent work shifted from general automation risk toward AI-specific exposure. Felten, Raj, and Seamans [14] proposed a method for linking advances in AI to occupational abilities: the AI Occupational Exposure (AIOE) index [15]. Their approach connects AI applications to O*NET abilities and aggregates those ability-level links to occupations, industries, and geographies. Webb [40] offered a different way to measure exposure by comparing the language of AI patents with O*NET task descriptions, finding that AI exposure was concentrated more heavily in high-skilled, highly educated occupations than prior waves of automation.

The arrival of large language models (LLMs) prompted a new generation of exposure measures focused specifically on generative AI. Eloundou et al. [12] used human raters and GPT-4 to estimate the share of occupational tasks that LLMs could plausibly speed up. Felten, Raj, and Seamans [16], Colombo et al. [8], and Henseke et al. [23] similarly extend task- and ability-based exposure methods into the LLM era using adapted AIOE weights, LLM-generated task ratings, or probabilistic classifications of generative-AI exposure. However, these studies mostly estimate potential exposure rather than observed use. In other words, they ask where generative AI *could* matter, not necessarily where it is already being applied. Tomlinson et al. [38] and Handa et al. [21] mark a major shift by deriving occupational AI measures from real AI conversations. Tomlinson et al. classify Bing Copilot interactions into work activities, tasks, and occupations, producing a usage-derived AI applicability score. Handa et al. perform a related analysis using millions of Claude conversations. These studies make it possible to move beyond expert- or model-rated exposure and study where generative AI is actually showing up in work-related use. This paper builds directly on that shift by treating Microsoft's usage-derived applicability score as the outcome to be explained.

2.2 Predicting usage-derived applicability from job characteristics

The closest concurrent work also uses observed AI interaction data but differs in type of analysis and research goal. Agarwal, Agarwal, and Rana [2] analyze Anthropic Claude usage at the O*NET task level, using a seven-dimension task-trait rubric and applying principal component analysis and K-means clustering to identify task archetypes. Ledingham et al. [26] also work at the task level, using UK Civil Service vacancy data to examine variation in AI exposure across roles and to develop job-redesign scenarios. Alain et al. [3] analyze Claude usage within health care occupations,

mapping interactions to O*NET tasks and computing a per-occupation adoption rate; they find that health care accounts for a small share of Claude conversations, with adoption concentrated in patient-facing roles such as dietitians and nurse practitioners.

This paper differs from those studies in several ways. First, it operates at the occupation level rather than the task level, allowing AI applicability to be modeled as a property of whole occupations rather than isolated task statements. Second, it uses the full non-task O*NET feature space rather than a hand-designed rubric, while deliberately excluding O*NET task and work-activity statements to avoid target leakage. Third, it treats the usage-derived Microsoft score as a predictive target and asks which interpretable job characteristics recover that score. The goal is therefore not only to describe where AI conversations occur, but to explain why some occupations appear more applicable to generative AI than others.

2.3 Productivity, satisfaction, and calibrated reliance

A separate literature studies how generative AI affects task performance. Experimental and workplace studies find that generative AI can improve speed, completion, and output quality in many settings, including customer support, writing, consulting, and software development [6, 9, 10, 33, 34]. At the same time, these gains are uneven. Dell'Acqua et al. [10], for example, emphasize the "jagged frontier" of AI capability: AI can improve performance on some tasks while harming it on others that appear similar but actually fall outside the range of what generative AI can realistically do. This literature is important for understanding AI's potential for enhancing productivity, but it often focuses on specific tasks, companies, or experimental settings rather than asking whether the relationship between AI completion and user satisfaction varies systematically across occupations. The human–automation and human–AI trust literature provides a useful lens for answering that question. Previous work on trust in automation argues that effective human–machine systems require appropriate reliance: users should rely on automation when it is capable but should be skeptical when it is not [27]. More recent work in AI-assisted decision-making shows that explanations, cognitive forcing functions, task difficulty, incentives, and verification costs can all shape whether users over-rely or under-rely on AI systems [7, 39]. However, much of this evidence comes from controlled studies rather than large-scale observations of AI use across occupations.

Recent work has begun to connect AI capability, user preference, and AI adoption more directly. Shao et al. [36] introduce the Human Agency Scale and the WORKBank database, identifying an "automation red-light zone" where expert-assessed AI capability is high but worker desire for automation is low. Henseke [22] similarly shows that occupational exposure predicts, but does not fully determine, generative-AI adoption across European workplaces. These studies suggest that technical capability and user acceptance are related but not identical. This paper explores that idea further by using observed usage-derived measures: it asks whether AI's apparent task-completion rate translates into user satisfaction equally across occupations, whether hands-on, embodied occupations show a larger completion–satisfaction gap, and which aspects of embodied work account for it. In other words, the analysis helps identify where generative AI may be effectively completing work that users still find less satisfactory—and what it is about that work that produces the mismatch.

2.4 Contribution

Taken together, previous work provides strong measures of occupational exposure, emerging usage-derived datasets, evidence of generative-AI productivity effects, and conceptual frameworks for automation and augmentation. What remains less developed is an interpretable, occupation-level account that links observed generative-AI applicability back to measurable job characteristics, while also examining where AI completion and user satisfaction diverge. This paper fills that gap by predicting Microsoft's usage-derived AI applicability score from non-task O*NET characteristics, testing whether the completion–satisfaction gap is larger in hands-on, embodied occupations, and decomposing embodied work into narrower constructs to identify which of its aspects account for the gap.

3 Data and Methods

3.1 Research questions and hypotheses

Table 1 summarizes the research questions and their hypotheses; RQ2 and RQ3 together address the second question posed in the Introduction.

Table 1: Research questions and hypotheses.

	Research question	Hypothesis
RQ1	Can job characteristics predict AI applicability?	**H1:** Models trained on non-task O*NET characteristics recover the applicability score with $R^2 \approx 0.6$ or higher; cognitive and communication features associate positively, physical and manual features negatively.
RQ2	Does AI's task completion translate into user satisfaction equally across occupations?	**H2:** The completion–satisfaction gap is larger in hands-on, embodied occupations—physically demanding, mechanical, and hazardous work.
RQ3	Where the gap appears, which aspects of embodied work account for it?	**H3:** The gap is unevenly distributed across components of embodied work; some components account for substantially more of it than others.

3.2 Datasets

This study combines two public occupation-level datasets: O*NET 30.2 and Microsoft's *Working with AI* dataset (Table 2). Each observation represents an occupation rather than an individual worker, and occupations are identified using Standard Occupational Classification (SOC) codes. O*NET provides the predictor variables describing the characteristics of each occupation, while *Working with AI* provides the generative-AI outcome measures used as prediction targets and comparison variables. O*NET 30.2 [31] is the U.S. Department of Labor's standard database of occupational characteristics. From O*NET, I use eight non-task content domains—categories describing what an occupation is like and what it requires of workers, rather than the specific tasks workers perform: Skills, Knowledge, Abilities, Work Context, Work Styles, Interests, Work Values, and Job Zone. These domains capture required capabilities, working conditions, work-related preferences and tendencies, and levels of occupational preparation. The task and work-activity domains are deliberately excluded because the AI applicability score is itself derived from occupational tasks. Including task-based variables as predictors would allow information used to construct the target to leak into the model inputs. Together, the eight retained domains provide 209 numeric job-characteristic features across 923 detailed O*NET occupations.

Microsoft's *Working with AI* dataset [38] provides occupation-level measures derived from anonymized Bing Copilot conversations between January and September 2024. The dataset covers 785 occupations and includes measures of AI applicability, task completion, user feedback, and impact scope (the scope of AI's involvement in an occupation's work activities). A separate coverage measure, usage breadth, gives the share of an occupation's work activities with any observed Copilot usage; I use it only as a control in the RQ2 robustness checks. The primary outcome for RQ1 is the AI applicability score. In plain terms, this score estimates how much of an occupation's everyday work generative AI can currently assist with, based on how often Copilot conversations involved that occupation's tasks and how successfully the AI performed them. For RQ2, I combine two additional measures: task completion, which estimates how often the AI actually finished the task it was asked to do, and positive user feedback, which reflects how often users signaled they were satisfied with the result. The difference between them is what I call the completion–satisfaction gap: work the AI managed to complete but that the user was nonetheless not satisfied with. For RQ3, I decompose embodied work into six narrower measures built from the same non-task O*NET items, described in Section 3.6.

Figure 1 summarizes the study pipeline. O*NET contributes 209 job-characteristic predictors, Microsoft contributes the usage-derived AI outcome measures, and the merged sample of 738 occupations is used for prediction, the completion–satisfaction analysis, and its decomposition. Figure 2 shows the distribution of AI applicability in the merged sample. The distribution is right-skewed: most occupations have low-to-moderate applicability scores, with a smaller number showing high applicability. This pattern is consistent with generative AI being unevenly applicable across the labor market.

3.3 Preprocessing

I first restrict the O*NET predictors to the eight non-task content domains described in Section 3.2, yielding the 209 numeric job-characteristic features; task and work-activity statements are excluded to avoid target leakage. Because the two datasets report occupations at slightly different levels of detail, I align them to a common SOC occupation level

Table 2: The two source datasets and the merged analysis sample.

	O*NET 30.2	**Microsoft *Working with AI***
Unit	Occupation (SOC)	Occupation (SOC)
Size	923 occupations	785 occupations
Key variables	209 features, 8 domains (Skills, Knowledge, Abilities, Work Context, Styles, Interests, Values, Job Zone)	AI applicability, completion, feedback, impact scope, usage breadth (Bing Copilot, Jan–Sep 2024)
Role	Predictors	Outcome / target
Merged on SOC (inner join, complete O*NET domains): 738 occupations		

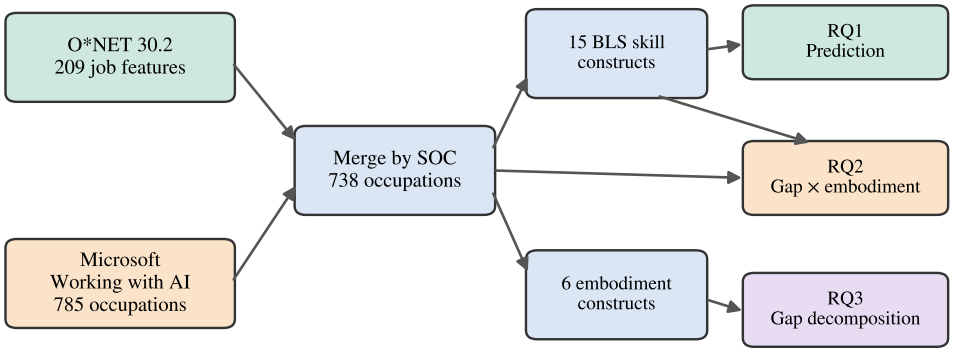


Figure 1: Study pipeline. Two public datasets are merged at the occupation level; the O*NET features are summarized into two interpretable construct sets (15 BLS skill constructs for prediction; six embodiment constructs for decomposing the completion–satisfaction gap); and the research questions are answered on the merged sample. Task and work-activity statements are excluded throughout to avoid target leakage.

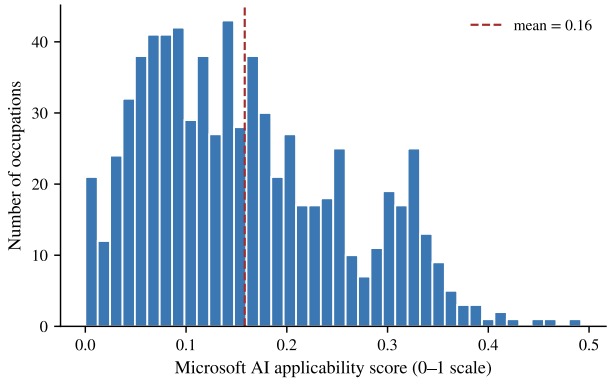


Figure 2: Distribution of the Microsoft AI applicability score across the 738 merged occupations.

before merging. O*NET includes some occupations at the more detailed O*NET-SOC level, while the Microsoft dataset is organized at the SOC level. I therefore truncate O*NET-SOC codes to the shared SOC level, average sub-occupations that map to the same SOC code, and keep only occupations with complete data across all eight domains. I then merge the resulting O*NET occupation records with the Microsoft outcome measures using an inner join on SOC code, producing a final analysis sample of 738 occupations.

The 209 O*NET features are highly intercorrelated, so I also create a smaller set of interpretable constructs for the main predictive models. Following the BLS occupational-skills taxonomy [11], I group 79 O*NET features into 17 theory-grounded skill constructs, including *Communication*, *Computers & IT*, *Mathematics*, *Fine Motor*, and *Physical Strength & Stamina*. Each construct is computed as the mean of its constituent normalized O*NET features. To reduce redundancy, I merge two pairs of highly correlated constructs: speaking/listening with writing/reading ($r = 0.92$), and critical/analytical thinking with problem solving ($r = 0.86$). This yields 15 final skill constructs. These constructs are then standardized to zero mean and unit variance before modeling. Appendix A lists each construct's constituent O*NET items.

3.4 RQ1: Predicting AI applicability

To test whether job characteristics predict generative-AI applicability, I model the Microsoft AI applicability score as a continuous occupation-level outcome. All models are evaluated with 10-fold cross-validation, and predictive performance is reported as the mean cross-validated R^2 across the ten folds. For the strongest model, I also report a 95% confidence interval for R^2, obtained by a percentile bootstrap of the pooled out-of-fold predictions (10,000 resamples of occupations with replacement). Rather than presenting only the most accurate model, I use a modeling strategy that explores both the most interpretable models and less interpretable but higher-performing models. This makes it possible to assess not only how well AI applicability can be predicted, but also how much predictive signal is available from transparent job-characteristic constructs. The modeling proceeds in four stages of increasing complexity. The first stage uses the 15 interpretable skill constructs described above. These construct-level models provide the clearest test of whether broad, named job characteristics—such as *Communication*, *Computers & IT*, *Mathematics*, *Fine Motor*, and *Physical Strength & Stamina*—recover the usage-derived applicability score. I train several regression models on this feature set, including ridge regression, random forest, gradient boosting, and a simple averaging ensemble. The second stage repeats the same modeling approach but incorporates the full set of 209 non-task O*NET features, to test whether the more granular feature space improves prediction beyond the compact construct representation.

The third stage expands the predictors to 30 constructs, each tied to a named labor-economics or occupational-psychology framework, including prior work on task content, automation risk, LLM exposure, and occupational AI applicability [4, 12, 18, 38]; Appendix A lists each construct with its framework anchor and constituent items. I remove redundant and weakly associated features and augment the remaining constructs with a small number of partial-least-squares (PLS) components. The fourth stage adds principal components derived from sentence embeddings of O*NET job titles, which provide a compact text-based signal about occupations without using task statements or Microsoft outcome variables, and combines the strongest predictors using greedy forward ensemble selection; the resulting model uses 40 features in total (the 30 constructs, five PLS components, and five title-embedding components). After fitting the models, I summarize the relationships between job characteristics and AI applicability using Pearson correlations for the interpretable constructs and a cross-validated predicted-versus-actual plot for model fit.

3.5 RQ2: Completion–satisfaction gap

To examine whether AI task completion translates into user satisfaction equally across occupations, I compute a completion–satisfaction gap for each occupation. This gap is defined as the difference between the Microsoft task-completion rate and the positive-feedback rate. The task-completion rate captures how often AI actions associated with an occupation are classified as completed, while the positive-feedback rate serves as a proxy for user satisfaction with those AI outputs. A large gap indicates that AI appears to complete the task, yet users are still unsatisfied with the result.

I then test whether this gap is larger in hands-on, embodied occupations. I use three indicators of embodied work: a physical-demands index, computed as the mean of the standardized *Physical Strength & Stamina* and *Fine Motor* constructs from the RQ1 feature set; the *Mechanical* knowledge construct from the same set; and exposure to hazardous equipment, an O*NET work-context item. For each indicator, occupations are divided into terciles, and I compare the mean completion–satisfaction gap between the high and low terciles using Welch's independent-samples t-test, reporting Cohen's d as the effect size. As robustness checks, I also verify that each relationship holds when the indicator

is treated as a continuous variable, when usage breadth is controlled, and when the analysis is restricted to the half of occupations with above-median usage breadth. This analysis is descriptive rather than causal. It does not show that embodied work causes lower satisfaction with AI. Instead, it tests whether occupations whose work is more physical, mechanical, and hazardous show a systematically larger mismatch between AI task completion and user satisfaction.

3.6 RQ3: Decomposing the gap

Hands-on occupations differ from desk work in many ways at once, so a single "physicality" measure cannot say *which* aspect of embodied work travels with the completion–satisfaction gap. To examine this, I decompose embodied work into six narrower constructs, each computed as the mean of standardized non-task O*NET items (Table 3). *Real-time coupling* captures work that must be continuously monitored and adjusted, often at a pace set by equipment. *Environmental variability* captures work in uncontrolled or changing settings. *Sensory dependence* captures reliance on direct perception—seeing, hearing, and localizing things in the work environment. *Spatial reasoning* captures the need to visualize objects and spatial relationships. *Hands-busy work* captures the degree to which the worker's body is engaged—handling objects, walking, climbing, kneeling, and working in cramped or awkward positions. *Fine manipulation* captures precise manual control.

The six constructs are strongly intercorrelated ($r = 0.53$ to 0.89): occupations that are physically demanding in one respect tend to be demanding in all of them. Coefficients from a single multiple regression are unreliable under this much overlap, so I instead divide the model's explained variance among the constructs by averaging each construct's incremental contribution across all possible orders in which the constructs could enter the model—the Lindeman–Merenda–Gold decomposition [19, 29]. Each construct's share can be read as its average contribution to explaining the gap, with overlapping credit split evenly rather than awarded to whichever construct happens to enter first. Each construct's zero-order correlation with the gap is reported with a 95% confidence interval based on the Fisher z transformation.

Table 3: The six embodiment constructs used in RQ3. Each construct is the mean of the listed O*NET items, standardized before averaging. Parenthetical counts give the number of items. AB = Abilities domain; WC = Work Context domain.

Construct	O*NET items
Real-time coupling (4)	AB: Rate Control; Reaction Time; Response Orientation. WC: Pace Determined by Speed of Equipment.
Environmental variability (5)	WC: Outdoors, Exposed to All Weather Conditions; Outdoors, Under Cover; Indoors, Not Environmentally Controlled; Exposed to Very Hot or Cold Temperatures; Exposed to Contaminants.
Sensory dependence (8)	AB: Far Vision; Glare Sensitivity; Hearing Sensitivity; Night Vision; Peripheral Vision; Sound Localization; Auditory Attention; Visual Color Discrimination.
Spatial reasoning (3)	AB: Visualization; Spatial Orientation; Depth Perception.
Hands-busy work (7)	WC: Spend Time Using Your Hands to Handle, Control, or Feel Objects, Tools, or Controls; Spend Time Walking or Running; Spend Time Climbing Ladders, Scaffolds, or Poles; Spend Time Kneeling, Crouching, Stooping, or Crawling; Spend Time Bending or Twisting Your Body; Spend Time Keeping or Regaining Balance; Exposed to Cramped Work Space, Awkward Positions.
Fine manipulation (4)	AB: Manual Dexterity; Finger Dexterity; Arm-Hand Steadiness; Control Precision.

4 Results

4.1 RQ1: Job characteristics predict AI applicability

Job characteristics explain a substantial share of the variation in generative-AI applicability across occupations. As shown in Figure 3 and Table 4, predictive accuracy increases steadily as the models move from compact, interpretable constructs to the strongest framework-grounded specification. The 15-construct model reaches a 10-fold cross-validated

R^2 of 0.62, meaning that a small set of interpretable skill constructs already explains over 60% of the variation in Microsoft's AI applicability score. Using the full set of 209 non-task O*NET features improves performance to $R^2 = 0.66$. Expanding to framework-grounded constructs with in-fold partial-least-squares components raises the best single interpretable model to $R^2 = 0.68$. The strongest model, which adds job-title sentence embeddings and combines predictors through greedy ensemble selection, reaches $R^2 = 0.72$ under cross-validation.

These results support H1: measurable job characteristics recover much of the usage-derived AI applicability signal. Importantly, most of the predictive value comes from transparent, construct-based inputs rather than from more complex text-based features. The full model improves by about 0.10 R^2 over the simple 15-construct model, but more than half of that gain is achieved before adding job-title embeddings. This suggests that the main structure of AI applicability is already visible in interpretable occupational characteristics. The cross-validated predicted-versus-actual plot for the full-feature model (209 features, $R^2 \approx 0.66$) in Figure 5 shows that it tracks the observed applicability scores well overall, though it slightly compresses predictions for the highest-applicability occupations. As a robustness check, I also tested an automated machine-learning system, AutoGluon [13], on the same full non-task feature set using external five-fold cross-validation. AutoGluon trains and stacks many gradient-boosted-tree and neural-network models, but it reached only $R^2 \approx 0.67$, essentially no better than the simpler full-feature ensemble. This suggests that the reported models are already close to the predictive ceiling available from these O*NET features, and that additional modeling complexity alone does not substantially improve performance.

The relationships between individual constructs and AI applicability are also coherent and interpretable (Figure 4). The strongest positive correlates are *Communication* ($r = +0.59$), *Computers & IT* ($r = +0.57$), *Creativity & Innovation* ($r = +0.42$), *Interpersonal* skill ($r = +0.38$), *Mathematics* ($r = +0.33$), and *Critical Thinking & Problem Solving* ($r = +0.32$). The strongest negative correlates are *Physical Strength & Stamina* ($r = -0.68$), *Fine Motor* skill ($r = -0.65$), and *Mechanical* knowledge ($r = -0.52$). In other words, generative-AI applicability is highest in

Table 4: RQ1: 10-fold cross-validated R^2 for predicting AI applicability, as an escalation from the most interpretable model to the best framework-grounded model. PLS components and job-title embeddings are computed *inside* each fold to avoid leakage; the last row is combined by greedy ensemble selection.

Model	Features	R^2
Interpretable ensemble	15 constructs	0.62
Full-feature ensemble	209 features	0.66
Ridge regression	30 framework constructs + PLS	0.68
Ensemble selection	30 constructs + PLS + title embeddings	**0.72**

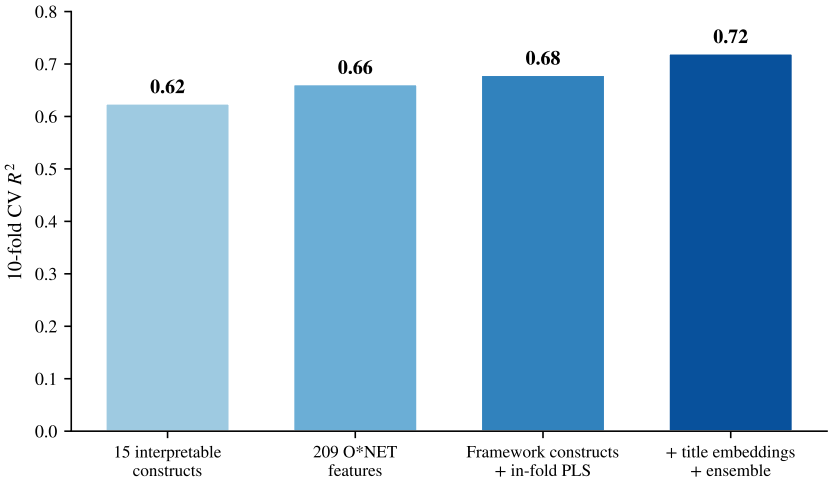


Figure 3: RQ1: 10-fold cross-validated R^2 for predicting AI applicability, escalating from the interpretable 15-construct model to the best framework-grounded model (which adds in-fold PLS components and job-title sentence embeddings, combined by ensemble selection).

occupations centered on communication, computing, analysis, and other cognitive work, and lowest in occupations defined by physical, manual, and mechanical demands.

4.2 RQ2: Hands-on, embodied work shows a larger completion–satisfaction gap

The relationship between AI task completion and user satisfaction varies systematically with how embodied the work is. As shown in Figure 6 and Table 5, hands-on occupations have significantly larger completion–satisfaction gaps across all three embodiment indicators, supporting H2. For the physical-demands index, the mean gap increases from 0.19 in the low tercile to 0.28 in the high tercile ($d = 0.96$, $p < 0.001$). The effect is largest for mechanical knowledge ($d = 1.29$, $p < 0.001$) and similarly large for exposure to hazardous equipment ($d = 1.00$, $p < 0.001$). The pattern is not just a result of dividing occupations into terciles: the same relationships hold when the indicators are treated as continuous variables (mechanical knowledge $r = 0.44$; hazardous equipment $r = 0.48$; physical demands $r = 0.36$). Across the three indicators, the result is consistent: the more physical, mechanical, and hands-on an occupation's work, the wider the mismatch between what AI appears to complete and how satisfied users are with the result.

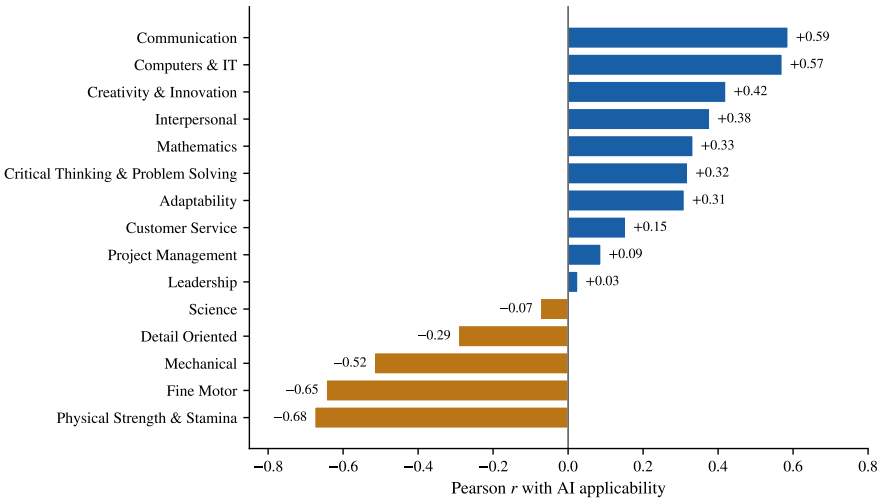


Figure 4: RQ1: Pearson correlation between each of the 15 job-characteristic constructs and AI applicability. Communication- and computing-oriented constructs are most positively associated; physical and manual constructs are most negatively associated.

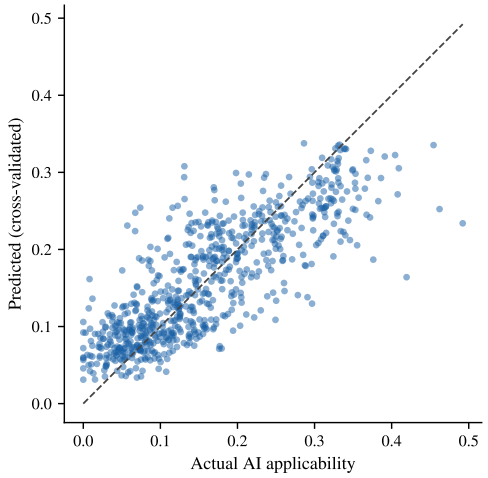


Figure 5: RQ1: cross-validated predicted versus actual AI applicability for the full-feature ensemble (the 209-feature model, $R^2 \approx 0.66$). The model slightly compresses the highest-applicability occupations.

Within these embodied, lower-applicability occupations, AI often records very high task-completion rates, but those completions translate into comparatively low user satisfaction. The relationship is not just a result of sparse usage: it remains significant after controlling for usage breadth, with the hazardous-equipment relationship decreasing from $r = 0.48$ to $r = 0.36$, the mechanical-knowledge relationship from $r = 0.44$ to $r = 0.32$, and the physical-demands relationship from $r = 0.36$ to $r = 0.19$, all still significant at $p < 0.001$. The same pattern holds when the analysis is restricted to the half of occupations with above-median usage breadth ($d = 0.61$, $d = 0.67$, and $d = 0.52$, respectively; all $p < 0.001$). Together, these checks suggest that the gap reflects a substantive mismatch between AI task completion and user satisfaction in hands-on work, rather than merely lower or noisier AI use in those occupations. What H2 does not say is *which* aspect of hands-on work matters: physically demanding occupations differ from desk work in many correlated ways at once. RQ3 addresses this directly.

4.3 RQ3: The gap concentrates in real-time, situated work, not physical effort as such

Decomposing embodied work into the six constructs of Section 3.6 shows that the completion–satisfaction gap is unevenly distributed across them, supporting H3 (Table 6, Figure 7). All six constructs correlate positively with the gap ($r = 0.30$ to 0.52), but they also overlap heavily with one another ($r = 0.53$ to 0.89), so I report each construct's fair share of the jointly explained variance alongside its zero-order correlation. Together, the six constructs explain just under a third of the variance in the gap ($R^2 = 0.31$). Real-time coupling stands out: it is the strongest single correlate ($r = 0.52$, 95% CI $[0.46, 0.57]$) and accounts for 42% of the explained variance—nearly three times the share of any other construct—followed by sensory dependence and environmental variability (15% each). Spatial reasoning (12%), fine manipulation (8%), and hands-busy work (7%) contribute least.

This ordering suggests that the gap is not about physical effort as such. It concentrates where work forms a continuous loop between perception and action—the worker acts, observes the result, and adjusts, often in an environment that differs from site to site and through cues that must be seen, heard, or felt directly. A conversational answer is produced once and does not update, so even an accurate answer arrives as a fixed snapshot where the work itself is a moving process. Notably, the explanation that product intuition would suggest first—that physical workers are dissatisfied because their hands and eyes are occupied and interacting with a text interface is impractical—receives the least support:

Table 5: RQ2: mean completion–satisfaction gap in the low- and high-embodiment terciles, with effect size and significance (t-test, high vs. low).

Embodiment indicator	Low tercile	High tercile	Cohen's d	p
Physical demands (strength + fine motor)	0.194	0.277	0.96	< 0.001
Mechanical knowledge	0.188	0.285	1.29	< 0.001
Exposure to hazardous equipment	0.192	0.279	1.00	< 0.001

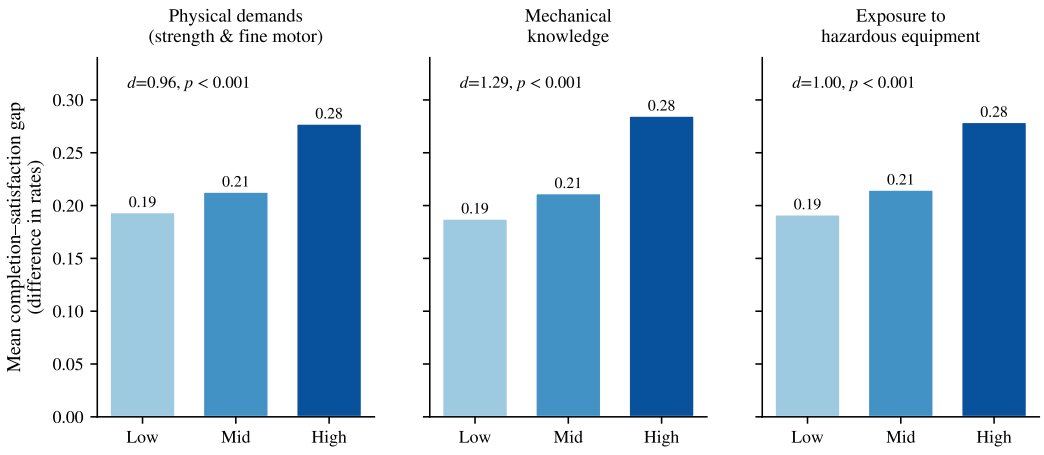


Figure 6: RQ2: mean completion–satisfaction gap by embodiment tercile, for three indicators of hands-on work (a physical-demands index, mechanical knowledge, and exposure to hazardous equipment). The gap grows monotonically in every case.

hands-busy work ranks near the bottom, alongside fine manipulation. Two qualifications apply. First, because the six constructs overlap heavily, the decomposition supports a contrast between the top and bottom of the ranking rather than the exact order of the middle constructs; sensory dependence, environmental variability, and spatial reasoning are within three percentage points of one another and cannot be reliably distinguished. Second, these are occupation-level associations between job characteristics and an aggregate usage measure; they identify where the gap concentrates and which features of the work travel with it, not the mechanism operating within any individual conversation.

5 Discussion

5.1 Summary of findings

Across 738 U.S. occupations, the analysis produces three main findings. First, measurable job characteristics predict usage-derived generative-AI applicability well. A compact model based on 15 interpretable constructs explains approximately 62% of the variance in Microsoft's AI applicability score, the full 209-feature O*NET model explains approximately 66%, and the strongest framework-grounded model reaches $R^2 \approx 0.72$ (pooled out-of-fold $R^2 = 0.73$, 95% bootstrap confidence interval $[0.69, 0.76]$). The relationships are also substantively coherent: occupations involving communication, computing, creativity, and analytical work show higher AI applicability, while occupations involving physical, manual, and mechanical demands show lower applicability.

Second, AI task completion does not translate into user satisfaction equally across occupations. Hands-on, embodied work shows a larger completion–satisfaction gap, with effect sizes between $d = 0.96$ and $d = 1.29$ across a physical-demands index, mechanical knowledge, and exposure to hazardous equipment. This suggests that AI may appear to complete occupationally relevant tasks while still producing outputs that users find less satisfactory, especially where the work itself is physical, mechanical, and hands-on.

Table 6: RQ3: association between each embodiment construct and the completion–satisfaction gap (Pearson r; all $p < 10^{-16}$), and each construct's share of the jointly explained variance (Lindeman–Merenda–Gold decomposition; total $R^2 = 0.31$).

Construct	r	Share of explained variance
Real-time coupling	0.52	42%
Sensory dependence	0.44	15%
Environmental variability	0.43	15%
Spatial reasoning	0.42	12%
Hands-busy work	0.33	7%
Fine manipulation	0.30	8%

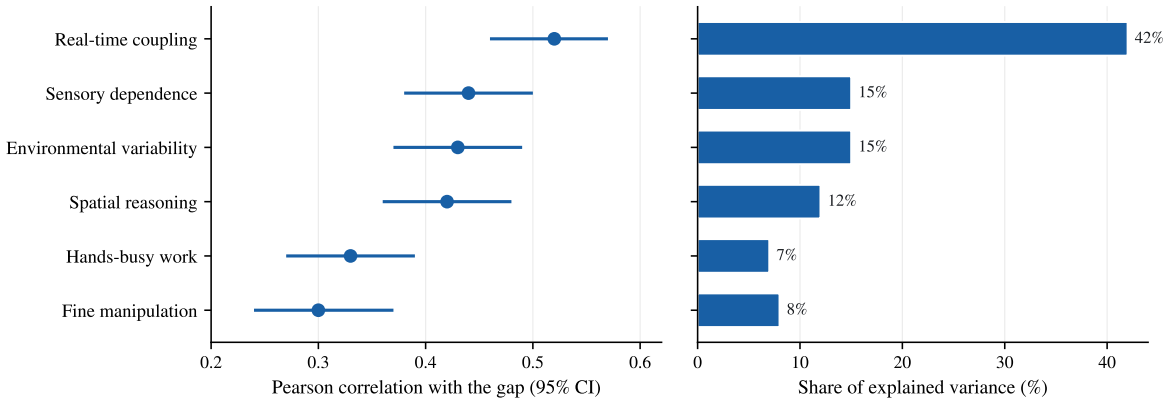


Figure 7: RQ3: each embodiment construct's correlation with the completion–satisfaction gap (left; dots are Pearson r, lines are 95% confidence intervals) and its share of the jointly explained variance (right; Lindeman–Merenda–Gold decomposition, total $R^2 = 0.31$). Real-time coupling—work that is continuously adjusted, often at a pace set by equipment—dominates; hands-busy work and fine manipulation contribute least.

Third, the gap is not about physical effort as such. Decomposing embodied work into six narrower constructs shows that the gap concentrates in work that requires continuous real-time adjustment (often at a pace set by equipment), depends on direct sensory information, and unfolds in variable environments; together these three constructs account for over 70% of the explained variance, with real-time adjustment alone accounting for 42%. Hands-busy work and fine manipulation contribute least. A plausible reading is that a conversational answer, produced once and fixed, cannot participate in work that is a live loop of perceiving, acting, and adjusting: the answer may be correct and still arrive as a snapshot where the work needs something that keeps pace.

5.2 Design and evaluation implications for embodied work

The findings suggest that apparent task completion is an incomplete measure of Bing Copilot's fit for certain types of work. In hands-on occupations, AI task completion translates into user satisfaction less reliably, even after accounting for how broadly AI is used within the occupation. RQ3 helps explain where this mismatch is concentrated. Real-time coupling shows the strongest association with the gap, followed by sensory dependence and environmental variability, while hands-busy work and fine manipulation contribute less. It appears that the main challenge is work that unfolds through a changing cycle of observation, action, and adjustment. This distinction separates access to the system from alignment with the work. Voice input, wearable displays, and other hands-free interfaces may make generative AI easier to use while someone is handling tools or moving through a worksite. However, they do not ensure that the system's understanding remains accurate after the worker acts or conditions change. In other words, a voice assistant can deliver outdated guidance more conveniently. These accessibility features may still be valuable, but the relatively weak contribution of hands-busy work suggests that interface convenience is unlikely to resolve the underlying issue with user satisfaction.

For existing chat-based tools, the findings point toward the less time-sensitive phases surrounding a physical task—preparation, planning, reviewing equipment information, anticipating possible problems, and documenting or explaining completed work—during which workers can pause, provide relevant context, and review output before acting. Assistance during the task itself is a harder design problem. The results suggest that such systems would need to treat the worker's prompts as a snapshot of an evolving situation rather than a complete account of it: distinguishing what the worker has reported from what the system has observed or inferred, tracking which actions have already been taken, and delivering guidance in short, conditional steps that name both the next action and the observation that should confirm it before proceeding. Each step then becomes a point at which the system's understanding can be updated, rather than continuing through a procedure based on conditions that may no longer hold. The sensory-dependence finding similarly argues for multimodal input—images, audio, video, measurements, and equipment data—but as a way to reduce uncertainty about the current situation rather than as a feature in itself. A photograph reveals the visible condition of a component while providing no information about its temperature, pressure, stability, or recent behavior; a system that distinguishes direct observations from user reports, and states explicitly what it cannot determine, is less likely to fill in missing context with assumptions.

The same principle applies to LLM evaluation. A conventional evaluation may judge an LLM response's clarity or domain relevance on its own, but an answer that is reasonable at one moment may become inappropriate after the worker acts. Where information about the changing physical state is available, the model should be evaluated using the complete sequence:

$$\textit{observation} \rightarrow \textit{AI guidance} \rightarrow \textit{worker action} \rightarrow \textit{resulting change} \rightarrow \textit{revised guidance}.$$

An LLM-based judge could evaluate this sequence when provided with a reliable record of the relevant states and actions. It could assess whether the system requested necessary information, made unsupported assumptions, incorporated new evidence, and revised its guidance when earlier assumptions no longer applied. When the changing physical state cannot be observed or reconstructed, the evaluation must make a narrower claim. It cannot determine whether the advice remained correct in the actual environment, but it can test whether the system handled its limited access appropriately. Evaluation cases can deliberately omit important information and measure whether the model identifies what is missing, asks a question that could distinguish among plausible conditions, states its assumptions, and avoids committing prematurely to a long procedure. Paired cases with the same visible symptom but different hidden causes would be especially useful: the desired behavior is not necessarily to infer the correct cause, but to recognize that the available evidence does not support a single confident answer.

The same distinction applies to measuring deployed systems. Task completion alone does not establish that guidance was usable, remained valid, or improved the worker's outcome; evaluations of deployed systems would also need to

capture whether workers could follow a recommendation, how often they corrected the system's understanding, and whether the guidance created rework or interruption. At the occupation level, high generative-AI applicability combined with a small completion–satisfaction gap is consistent with scaling existing conversational tools, whereas a wide gap indicates that use-case selection, interaction design, or evaluation—not model capability alone—deserves attention, even when the AI appears to complete the immediate task. These implications are directions for experimentation rather than proven interventions, subject to the limitations discussed below. Even so, the results support a general principle: AI for embodied work should be assessed not only by the quality of the answer it generates, but by how well it manages uncertainty and remains aligned with the situation as the work unfolds.

5.3 Limitations

Several limitations should be considered when interpreting these results. First, the AI applicability measure comes from a single platform and time period: Bing Copilot conversations from January through September 2024. The findings therefore reflect how generative AI was used in one system during a particular stage of adoption, not the full potential of generative AI across all tools, organizations, or future model capabilities. As AI systems and user practices change, the occupational pattern of applicability may also change. Second, the analysis is conducted at the occupation level. This makes it possible to compare hundreds of jobs using shared O*NET and Microsoft measures, but it also means that within-occupation variation is averaged out. Workers with the same occupational title may perform different tasks, use different tools, work in different organizational contexts, or face different levels of autonomy and risk. As a result, the estimates should be interpreted as occupation-level patterns rather than predictions for every worker in a given job.

Third, the study deliberately excludes O*NET task statements and work-activity descriptions to avoid target leakage (Section 3.2). This improves the validity of the predictive analysis, but it also removes information that would likely be useful for prediction. The reported models therefore estimate how much of AI applicability can be recovered from non-task job characteristics, not the maximum possible accuracy that could be achieved using all occupational text. Finally, the RQ2 and RQ3 findings are descriptive rather than causal. The larger gap in embodied occupations may reflect several factors: only a narrow informational slice of such work can be expressed in a text conversation, answers may be harder to act on away from a desk, and outputs may fit the surrounding conditions of the work less well. The six embodiment constructs also overlap heavily ($r = 0.53$ to 0.89), so the RQ3 decomposition assigns each construct a fair share of the jointly explained variance rather than isolating independent effects, and it is most informative about the contrast between the top and bottom of the ranking. For these reasons, the RQ2 and RQ3 findings should be read as descriptive, exposure-conditional patterns. Future work should test whether the same patterns hold across other AI platforms, later time periods, and more fine-grained worker- or task-level data—and, for RQ3, whether conversation-level evidence confirms the real-time mechanism that the occupation-level pattern suggests.

References

[1] Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives, 33*(2), 3–30. https://doi.org/10.1257/jep.33.2.3

[2] Agarwal, P., Agarwal, H., & Rana, A. (2025). *What work is AI actually doing? Uncovering the drivers of generative AI adoption.* arXiv. https://arxiv.org/abs/2510.23669

[3] Alain, G., Crick, J., Snead, E., Quatman-Yates, C. C., & Quatman, C. E. (2025). Evaluating user interactions and adoption patterns of generative AI in health care occupations using Claude: Cross-sectional study. *Journal of Medical Internet Research, 27*, e73918. https://doi.org/10.2196/73918

[4] Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics, 118*(4), 1279–1333. https://doi.org/10.1162/003355303322552801

[5] Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M

[6] Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

[7] Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188. https://doi.org/10.1145/3449287

[8] Colombo, E., Mercorio, F., Mezzanzanica, M., & Serino, A. (2025). Towards the Terminator economy: Assessing job exposure to AI through LLMs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)* (pp. 9591–9600). https://doi.org/10.24963/ijcai.2025/1066

[9] Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2026). The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers. *Management Science*. Advance online publication. https://doi.org/10.1287/mnsc.2025.00535

[10] Dell'Acqua, F., McFowland, E., III, Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science, 37*(2), 403–423. https://doi.org/10.1287/orsc.2025.21838

[11] Dey, M., Hopson, A., Krutsch, E., Miller, M., & Nguyen, A. (2024). A new data product for occupational skills: Methodology, analysis, and a guide to using the employment projections skills data. *Monthly Labor Review*. U.S. Bureau of Labor Statistics. https://doi.org/10.21916/mlr.2024.19

[12] Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science, 384*(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

[13] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). *AutoGluon-Tabular: Robust and accurate AutoML for structured data.* arXiv. https://arxiv.org/abs/2003.06505

[14] Felten, E. W., Raj, M., & Seamans, R. (2018). A method to link advances in artificial intelligence to occupational abilities. *AEA Papers and Proceedings, 108*, 54–57. https://doi.org/10.1257/pandp.20181021

[15] Felten, E. W., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal, 42*(12), 2195–2217. https://doi.org/10.1002/smj.3286

[16] Felten, E. W., Raj, M., & Seamans, R. (2023). *How will language modelers like ChatGPT affect occupations and industries?* arXiv. https://arxiv.org/abs/2303.01157

[17] Florida, R. (2002). *The rise of the creative class: And how it's transforming work, leisure, community and everyday life.* Basic Books.

[18] Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

[19] Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software, 17*(1), 1–27. https://doi.org/10.18637/jss.v017.i01

[20] Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance, 16*(2), 250–279. https://doi.org/10.1016/0030-5073(76)90016-7

[21] Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., & Ganguli, D. (2025). *Which economic tasks are performed with AI? Evidence from millions of Claude conversations.* arXiv. https://arxiv.org/abs/2503.04761

[22] Henseke, G. (2026). *From exposure to adoption: Generative AI in European workplaces.* arXiv. https://arxiv.org/abs/2604.18849

[23] Henseke, G., Davies, R., Felstead, A., Gallie, D., Green, F., & Zhou, Y. (2025). *How exposed are UK jobs to generative AI? Developing and applying a novel task-based index* (Version 1). arXiv. https://arxiv.org/abs/2507.22748v1

[24] Hochschild, A. R. (1983). *The managed heart: Commercialization of human feeling.* University of California Press.

[25] Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Psychological Assessment Resources.

[26] Ledingham, A., Hollins, M., Lyon, M., Gillespie, D., Yunis-Guerra, U., Siviter, J., Duncan, D., & Hauser, O. P. (2025). *Beyond automation: Redesigning jobs with LLMs to enhance productivity.* arXiv. https://arxiv.org/abs/2512.05659

[27] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[28] Levy, F., & Murnane, R. J. (2004). *The new division of labor: How computers are creating the next job market.* Princeton University Press.

[29] Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.* Scott, Foresman.

[30] Mintzberg, H. (1973). *The nature of managerial work.* Harper & Row.

[31] National Center for O*NET Development. (2026). *O*NET 30.2 database* [Data set]. U.S. Department of Labor. https://www.onetcenter.org/db_releases.html

[32] Newcombe, N. S., & Frick, A. (2010). Early education for spatial intelligence: Why, what, and how. *Mind, Brain, and Education, 4*(3), 102–111. https://doi.org/10.1111/j.1751-228X.2010.01089.x

[33] Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

[34] Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). *The impact of AI on developer productivity: Evidence from GitHub Copilot.* arXiv. https://arxiv.org/abs/2302.06590

[35] Polanyi, M. (1966). *The tacit dimension.* Doubleday.

[36] Shao, Y., Zope, H., Jiang, Y., Pei, J., Nguyen, D., Brynjolfsson, E., & Yang, D. (2025). *Future of work with AI agents: Auditing automation and augmentation potential across the U.S. workforce.* arXiv. https://arxiv.org/abs/2506.06576

[37] Taleb, N. N. (2012). *Antifragile: Things that gain from disorder.* Random House.

[38] Tomlinson, K., Jaffe, S., Wang, W., Counts, S., & Suri, S. (2025). *Working with AI: Measuring the applicability of generative AI to occupations.* arXiv. https://arxiv.org/abs/2507.07935

[39] Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), Article 129. https://doi.org/10.1145/3579605

[40] Webb, M. (2019). *The impact of artificial intelligence on the labor market* (SSRN Working Paper No. 3482150). https://doi.org/10.2139/ssrn.3482150

[41] Weeden, K. A. (2002). Why do some occupations pay more than others? Social closure and earnings inequality in the United States. *American Journal of Sociology, 108*(1), 55–101. https://doi.org/10.1086/344121

A Construct Definitions and Composition

This appendix documents the item-level composition of the two construct sets used in RQ1: the 15 BLS skill constructs (first modeling stage) and the 30 framework constructs of the strongest model (third and fourth stages). The six embodiment constructs used in RQ3 are documented in Table 3 in the main text. Throughout, item prefixes indicate the O*NET domain: AB = Abilities, KN = Knowledge, SK = Skills, WC = Work Context, WS = Work Styles, IN = Interests (RIASEC).

A.1 The 15 BLS skill constructs

Following the BLS occupational-skills taxonomy [11], 79 O*NET items are grouped into 17 constructs. Each item is min–max normalized to $[0,1]$ using the theoretical range of its O*NET scale, and each construct is the unweighted mean of its normalized items. O*NET Work Activities that the BLS taxonomy assigns to these categories are excluded, because the Microsoft applicability score is derived from work activities (Section 3.2). Two construct pairs whose scores correlated above $r = 0.82$ were merged by pooling their item sets and re-averaging: speaking/listening with writing/reading ($r = 0.92$), forming *Communication*, and critical/analytical thinking with problem solving ($r = 0.86$), forming *Critical Thinking & Problem Solving*. Table A1 lists the resulting 15 constructs and their 79 unique items.

Table A1: The 15 BLS skill constructs and their constituent O*NET items. Each construct is the unweighted mean of its min–max-normalized items; parenthetical counts give the number of items.

Construct	O*NET items
Communication (12)	AB: Oral Comprehension; Oral Expression; Speech Clarity; Speech Recognition; Written Comprehension; Written Expression. KN: Communications and Media. SK: Active Listening; Reading Comprehension; Speaking; Writing. WC: Public Speaking.
Critical Thinking & Problem Solving (11)	AB: Deductive Reasoning; Inductive Reasoning. SK: Complex Problem Solving; Critical Thinking; Judgment and Decision Making; Operations Analysis; Systems Analysis; Systems Evaluation. WC: Freedom to Make Decisions; Frequency of Decision Making; Impact of Decisions on Co-workers or Company Results.
Physical Strength & Stamina (10)	AB: Dynamic Strength; Extent Flexibility; Gross Body Coordination; Gross Body Equilibrium; Multilimb Coordination; Stamina; Static Strength; Trunk Strength. WC: Spend Time Bending or Twisting Your Body; Spend Time Walking or Running.
Mechanical (7)	KN: Mechanical. SK: Equipment Maintenance; Equipment Selection; Installation; Operation and Control; Repairing; Troubleshooting.
Adaptability (4)	SK: Active Learning. WS: Adaptability; Self-Control; Stress Tolerance.
Customer Service (4)	KN: Customer and Personal Service. WC: Conflict Situations; Deal With External Customers or the Public in General; Dealing With Unpleasant, Angry, or Discourteous People.
Detail Oriented (4)	AB: Selective Attention. WC: Consequence of Error; Importance of Being Exact or Accurate. WS: Attention to Detail.
Fine Motor (4)	AB: Control Precision; Finger Dexterity; Wrist-Finger Speed. WC: Spend Time Using Your Hands to Handle, Control, or Feel Objects, Tools, or Controls.
Interpersonal (4)	SK: Service Orientation; Social Perceptiveness. WS: Empathy; Social Orientation.
Mathematics (4)	AB: Mathematical Reasoning; Number Facility. KN: Mathematics. SK: Mathematics.
Science (4)	KN: Biology; Chemistry; Physics. SK: Science.
Computers & IT (3)	KN: Computers and Electronics. SK: Programming. WC: E-Mail.
Creativity & Innovation (3)	AB: Originality. KN: Fine Arts. WS: Innovation.
Leadership (3)	SK: Management of Personnel Resources. WC: Work Outcomes and Results of Other Workers. WS: Leadership Orientation.
Project Management (2)	SK: Management of Financial Resources; Management of Material Resources.

A.2 The 30 framework constructs

The 30 constructs of the third and fourth modeling stages were selected in two steps. First, an audit of 66 previously engineered constructs retained those with meaningful association with the target and acceptable redundancy; second, 49 additional constructs anchored in named frameworks were built, and a final screen over the 80 candidates kept the 30 used in the reported models. The strongest model combines these 30 constructs with five in-fold partial-least-squares components and five principal components of job-title sentence embeddings (40 features in total). Table A2 lists the 15 constructs retained from the earlier engineering rounds (each the unweighted mean of min–max-normalized items unless noted); Table A3 lists the 15 newly built constructs (each the unweighted mean of full-sample z-scored items unless noted). Item lists are reported as computed: a small number of specified items that were absent from the leakage-cleaned feature set were dropped at build time, and the professional-closure index's intended apprenticeship component was empty in the source data, leaving four components.

Table A2: The 15 audited constructs retained from earlier feature engineering. Unless noted, each construct is the unweighted mean of its min–max-normalized items.

Construct	Framework anchor	O*NET items / definition
Real-time coordination	Brooks [5]	AB: Reaction Time; Multilimb Coordination; Rate Control; Time Sharing.
Sensory discrimination	Polanyi [35]	AB: Visual Color Discrimination; Auditory Attention; Speech Recognition; Hearing Sensitivity.
Spatial reasoning	Newcombe & Frick [32]	AB: Spatial Orientation; Visualization; Depth Perception.
Tacit knowledge	Polanyi [35]	AB: Problem Sensitivity. SK: Active Learning; Critical Thinking. WC: Frequency of Decision Making.
Hazard exposure	BLS SOII methodology	WC: Exposed to Hazardous Conditions; Exposed to Hazardous Equipment; Exposed to Contaminants; Exposed to Disease or Infections; Exposed to Minor Burns, Cuts, Bites, or Stings; Exposed to Radiation; Exposed to High Places.
Routine cognitive	Autor et al. [4]	SK: Quality Control Analysis. WC: Importance of Repeating Same Tasks.
Autonomy	Hackman & Oldham [20]	WC: Freedom to Make Decisions; Determine Tasks, Priorities and Goals. WS: Initiative.
Creative demand	Florida [17]	AB: Originality; Fluency of Ideas. WS: Innovation.
Customer interaction	Hochschild [24]	WC: Telephone Conversations; Contact With Others.
Quantitative reasoning	Levy & Murnane [28]	AB: Mathematical Reasoning; Number Facility. KN: Mathematics. SK: Mathematics.
Supervision	Mintzberg [30]	SK: Management of Personnel Resources; Coordination; Instructing. WC: Coordinate or Lead Others in Accomplishing Work Activities.
Irreversibility of error	Taleb [37]	AB: Control Precision. WC: Consequence of Error; Importance of Being Exact or Accurate.
Knowledge–physical ratio	Autor et al. [4]	Share of cognitive content, $m_{\text{cog}}/(m_{\text{cog}}+m_{\text{phys}})$, where m_{cog} averages the Document Language, Quantitative Reasoning, and Technology Use construct scores and m_{phys} averages Routine Manual, Hazard Exposure, and Physical Stamina (constituents in the note).
Automation susceptibility	Autor et al. [4]	Share of routine content: routine construct scores (Routine Cognitive, Routine Manual) relative to routine plus nonroutine scores (Nonroutine Analytical, Nonroutine Interpersonal, Nonroutine Manual; constituents in the note).
Professional closure index	Weeden [41]	Mean of z-scores of four occupational-entry requirements from the O*NET Education, Training, and Experience file: expected required education level, related work experience, on-site training, and professional-certification importance.

Note. Constituent constructs used by the two share measures: Document Language = AB: Written Comprehension; Written Expression; Oral Comprehension; Oral Expression. SK: Reading Comprehension; Writing. Technology Use = KN: Computers and Electronics. SK: Programming; Operations Analysis. WC: Spend Time Sitting. Routine Manual = AB: Multilimb Coordination; Manual Dexterity. WC: Spend Time Making Repetitive Motions; Pace Determined by Speed of Equipment. Physical Stamina = AB: Stamina; Trunk Strength; Static Strength; Dynamic Strength. WC: Spend Time Standing; Spend Time Walking or Running. Nonroutine Analytical = AB: Inductive Reasoning; Deductive Reasoning; Problem Sensitivity. SK: Complex Problem Solving; Critical Thinking. Nonroutine Interpersonal = SK: Active Listening; Social Perceptiveness; Service Orientation; Negotiation; Persuasion. Nonroutine Manual = AB: Finger Dexterity; Arm-Hand Steadiness; Control Precision.

Table A3: The 15 newly built framework constructs. Unless noted, each construct is the unweighted mean of its full-sample z-scored items.

Construct	Framework anchor	O*NET items / definition
Routine × manual interaction	Autor et al. [4]	Product of two component scores: routine manual (AB: Finger Dexterity; Manual Dexterity. WC: Spend Time Making Repetitive Motions; Pace Determined by Speed of Equipment) and working-with-things intensity (AB: Manual Dexterity; Finger Dexterity. KN: Mechanical. SK: Equipment Maintenance).
Displacement susceptibility	Acemoglu & Restrepo [1]	Mean of two re-standardized component scores: routine cognitive (SK: Mathematics. WC: Importance of Repeating Same Tasks; Importance of Being Exact or Accurate. WS: Attention to Detail) and routine manual (as above).
Perception and manipulation	Frey & Osborne [18]	AB: Finger Dexterity; Manual Dexterity; Arm-Hand Steadiness. WC: Exposed to Cramped Work Space, Awkward Positions.
Social intelligence	Frey & Osborne [18]	SK: Social Perceptiveness; Negotiation; Persuasion; Service Orientation. WC: Conflict Situations.
AI-exposed perception abilities	Felten et al. [15]	AB: Perceptual Speed; Flexibility of Closure; Selective Attention; Time Sharing.
Code substitution	Eloundou et al. [12]	KN: Computers and Electronics. SK: Programming; Operations Analysis; Technology Design.
Text editing and review	Tomlinson et al. [38]	AB: Written Comprehension. KN: English Language. SK: Writing. WS: Attention to Detail.
Code review and debugging	Tomlinson et al. [38]	SK: Programming; Troubleshooting; Quality Control Analysis. WS: Attention to Detail.
Speech and audio work	Tomlinson et al. [38]	AB: Oral Expression; Speech Clarity; Speech Recognition. KN: Communications and Media.
Technical diagrams and design	Tomlinson et al. [38]	AB: Visualization; Spatial Orientation. KN: Engineering and Technology; Design.
In-person relational work	Tomlinson et al. [38]	SK: Negotiation. WC: Contact With Others; Face-to-Face Discussions with Individuals and Within Teams.
Task significance	Hackman & Oldham [20]	WC: Consequence of Error; Impact of Decisions on Coworkers or Company Results; Health and Safety of Other Workers.
Interest differentiation	Holland [25]	Range (maximum minus minimum) across the six raw RIASEC interest scores (IN: Realistic; Investigative; Artistic; Social; Enterprising; Conventional).
Dominant interest code	Holland [25]	Ordinal index (1–6) of the highest of the six RIASEC interest scores (1 = Realistic, . . . , 6 = Conventional).
Clinical/safety axis	Internal addition (gap analysis)	KN: Medicine and Dentistry; Biology. WC: Exposed to Disease or Infections; Health and Safety of Other Workers.