# Risks Create a Doubly Jagged Frontier of LLM Productivity Gains Across Computer Occupations

Deepika Chawla*[1], Gagandeep Singh[2], Elham Khorasani Buxton[3], Meicen Sun[2], Lav R. Varshney [4], Jeremy Riel[5], and Craig De Voto[5]

[1]Independent, [2]University of Illinois Urbana-Champaign, [3]University of Illinois Springfield, [4]Stony Brook University, [5]University of Illinois Chicago

**Abstract**

This paper argues that the current frontier of expert-LLM collaboration is doubly jagged: it is uneven, not only because of jagged AI capabilities but also due to jagged AI risks. As LLM systems improve, incompetence-driven risks (e.g., misinformation) may decline, but adversarial risks can persist or even increase, leading to the likely persistence of a jagged risk–reward frontier. To show jaggedness, we introduce the **first** occupation-level quantification of risk-aware productivity gains for expert–LLM collaboration over a realistic task distribution, using standard O*NET computer-occupation tasks. For each task, we leverage six frontier models (e.g., GPT-5, Claude Opus) to produce structured risk–reward ratings and use 6 human experts to verify a subset for reliability, achieving high agreement. Risk captures the possible increase in individual, organizational, or societal harm due to employing LLMs. The reward captures potential productivity gains, i.e., time/cost savings at a fixed quality target, after accounting for expert oversight to detect and mitigate errors and risks. We aggregate task-level ratings into occupation-level scores. Our analysis shows that risk varies more than reward, yielding vast differences in risk–reward tradeoff: In safety-critical roles (e.g., *Information Security Engineers*), gains are offset by risk, whereas in less safety-critical roles (e.g., *Web Developers*), gains often substantially exceed risks. At the task-level, we identify high-reward and medium-risk tasks as the "Sweet spot" for expert-LLM collaboration. The identified sweet spot is evident in real-world deployments as the majority of Claude conversations assigned to O*NET tasks (from the recent Anthropic Economic Index) are high-reward and medium-risk, whereas high-risk tasks are minimally represented, supporting our methodology.

*Correspondence to: deepinder.chawla.eco@gmail.com. Co-author emails: ggnds@illinois.edu, esahe2@uis.edu, mcsun@illinois.edu, lav.varshney@stonybrook.edu, jriel2@uic.edu, cdevot2@uic.edu

**Keywords.** Artificial Intelligence, Computer Occupations, Jaggedness, Large Language Models, Productivity Gains, LLM Risks

# 1 Introduction: The doubly jagged frontier of LLM deployment

The economic impact of LLM systems depends on whether organizations can derive meaningful productivity gains from real-world deployments [21, 45]. Recent debates on jagged capabilities of AI systems such as LLMs and agents, where they perform unevenly across similar tasks, highlights the need to identify tasks where LLM use is beneficial [19, 67]. Current discussions around jaggedness in the research community ignore risks and remain largely capability-first [24, 16, 35], centered around such questions as what set of tasks the models can perform and how quickly the set can expand. However, even if LLMs are capable of performing a task and expert oversight is available, the cost of verification and the potential consequences of catastrophic failures and of adversarial exploitation may be prohibitive. As a result, capability-only analyses often overestimate labor impact and have at times fueled "mass hysteria" narratives [32, 66]. Like capability, the risks vary by task: some tasks are robust to occasional mistakes and easy to audit, while others are brittle, hard to verify, or prone to attacks. This leads to a second layer of jaggedness in the frontier of LLM deployment due to risk heterogeneity across tasks that does not automatically diminish with capability improvements.

**We argue that the frontier of expert–LLM collaboration is uneven due to jaggedness in both LLM capabilities and LLM risks. As models improve, some risks shrink, while others, especially adversarial risks, persist or increase. The result is a risk-aware productivity frontier that can remain jagged even as capability-driven jaggedness smooths. This has concrete implications for labor policy, research funding, and national competitiveness.**

**Why computer occupations?** Computer occupations are among the earliest and the most intensive adopters of LLM systems, and as such, enable diffusion into other sectors. Evidence from the recent Anthropic Economic Index report supports this focus: LLM usage is concentrated in coding and related tasks, and within the United States, per-capita usage is higher in states with a larger share of workers in computer and mathematical occupations [3]. Computer occupations also cover a wide range of tasks, from routine documentation and testing for security-critical work where small errors or adversarial manipulation can trigger devastating consequences. Risks such as privacy and security have been identified as the

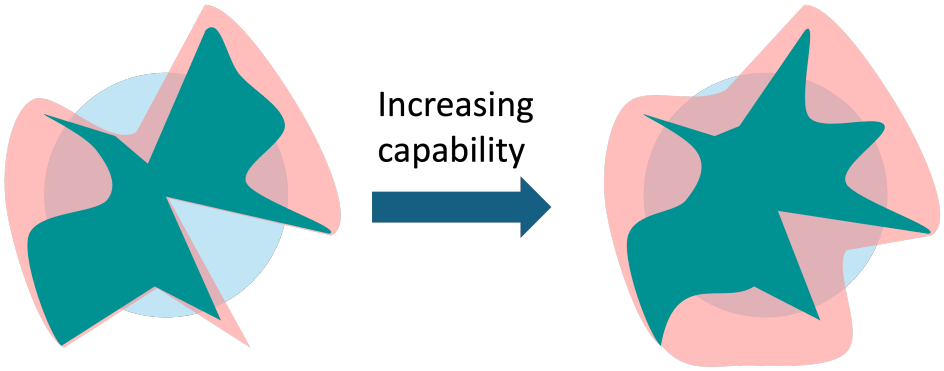


Figure 1: **Conceptual illustration of the doubly jagged productivity frontier:** The boundary of the blue circle represents tasks on which humans have the same productivity. The pink region shows the frontier of expert–LLM collaboration irrespective of risk. The pink subregion inside the blue circle represents tasks where LLMs reduce productivity. LLMs improve productivity for tasks where the pink region crosses the blue boundary. The green region denotes the risk-aware productivity frontier and is doubly jagged. As model capability increases, capability-driven jaggedness diminishes as the pink region expands and the frontier smooths, while the risk-aware green frontier remains jagged due to heterogeneous risks, particularly adversarial risks, that persist or even increase despite capability increases.

main challenge for reliable LLM adoption in the recent Stack Overflow Developer Survey [64]. This combination of high adoption and heterogeneous risk regimes makes computer occupations a uniquely suitable setting for investigating the risk-aware productivity frontier.

Note that we extend our analysis to healthcare and legal professions in Appendix J. We observe risk dominates at the occupation level for the vast majority of roles (99/108); the few reward-dominant cases are primarily administrative/support roles.

**Why expert-LLM collaboration?** Collaboration reflects the dominant mode of enterprise deployment: organizational policy, liability, and quality requirements typically require expert oversight [25]. Moreover, the collaboration setting provides a conservative baseline for identifying tasks where autonomous LLM use is reliable and beneficial.

**Contributions.** This paper contributes three main findings.

**1. The first occupation-level quantification of risk-aware productivity gains** for expert–LLM collaboration over a realistic task distribution. Our novel risk-aware productivity frontier can remain jagged even as the capability jaggedness smooths due to variation in auditability, costliness, and adversarial exposure across task-specific risks (Figure 1).

**2. Adversarial risks are a key driver of persistent jaggedness.** Many high-value tasks remain constrained not by average-case model quality but by worst-case behavior under adversarial manipulation, a gap emphasized by emerging LLM security analyses [56]. Moreover, increasing model capabilities can amplify the severity of such risks [2].

**3. Policy levers differ sharply by risk type.** For tasks involving non-adversarial

risks, addressing technical failures and verification costs through evaluation, tooling, and workflow design will accelerate deployment. For tasks involving adversarial risks, investments should prioritize rigorous stress testing, security engineering, access control, monitoring, and institutional safeguards.

Our work aims to stimulate research on risk-driven jaggedness and its impact on deployment. Compared to prior works [27, 26, 24, 16], we consider risk as a first-class constraint on deployment and introduce a risk-aware productivity frontier that provides a more realistic foundation for understanding the economic impact of LLMs. **Our work is the first to connect capability, risk, and deployment** to quantify productivity gains from LLMs, revealing which tasks within an occupation are suitable for LLM assistance under current capabilities and safeguards, and which remain bottlenecked by risks.

# 2 Data and Methodology

We study the risk–reward frontier of expert–LLM collaboration, where a human expert works with LLM systems (including LLM agents) to perform computer-related tasks using state-of-the-art models. To ground our study in a realistic task distribution, we use the O*NET 30.0 dataset (the Occupational Information Network) [49], which provides standardized work descriptions, including natural-language task statements and structured attributes such as skills and knowledge. O*NET 30.0 contains 923 occupations and we focus on the 27 occupations in the *Computer Occupations* category, denoted by $\mathcal{O}$. Each task $t$ for an occupation $o \in \mathcal{O}$ is represented by its task statement and O*NET numerical attributes, *Importance*, *Relevance*, and *Frequency*, which we later use to aggregate task-level ratings into occupation-level scores (§2.2). In total, we obtain 609 tasks across 27 occupations. Table 7 in Appendix D shows the number of tasks associated with each occupation.

Following the commonly used AI risk classification defined in the AI Risk Repository [61], we consider six risk categories most salient to computer occupations: Discrimination, Toxicity, Privacy, Misinformation, Security, and Malicious Actors and Misuse (full definitions in Appendix B). For our analysis, we group these into two sources of harm: 1. Incompetence-driven risks arising from erroneous or unreliable model behavior (Discrimination, Toxicity, Privacy, Misinformation); and 2. Adversarial risks due to malicious exploitation (Security, Malicious Actors and Misuse).

## 2.1 Task-level Risk-Reward Scoring

We adopt an LLM-based evaluation to estimate task-level risk–reward scores. This choice follows a growing literature that uses frontier LLMs as scalable annotators/judges [29, 40] and prior work that uses LLM-based ratings to score the exposure to AI of different O*NET occupations [16, 24]. Further, O*NET itself has started using a hybrid LLM–expert workflow to produce ratings at scale [51]. We treat the models' ratings as reflecting what it has learned from large-scale training corpora that include web data, books, news articles, and research papers, thereby approximating expert ratings at scale. We use 6 frontier models, both open- and closed-source: Qwen3-80B [74], LlamA-3.3-70B [31], Claude Opus 4.5 [1], Gemini-2.5 Pro [30], Grok-4 [73], and GPT-5 [53]. We use deterministic decoding wherever supported.[1] For each task $t$, we instruct each model as follows:

- Determine whether $t$ is safety-critical,
- If yes, then considering the risk categories, estimate the increased risk of expert–LLM collaboration relative to expert-only completion [2],
- After accounting for the expert oversight and effort required to detect and mitigate failures, estimate how much expert-LLM collaboration can reduce completion time or labor costs while maintaining expert-level quality in task completion.

Let $\mathcal{M}$ denote the set of frontier models. Each model $m \in \mathcal{M}$ is prompted to output for each task:

- **Reward rating** $G_m(t) \in \{0, \ldots, 5\}$ with justification: Low ($\leq 2$), Moderate ($= 3$), High ($\geq 4$) and,
- **Risk rating** $R_m(t) \in \{0, \ldots, 5\}$ with applicable risk categories and justification: Low ($\leq 2$), Moderate ($= 3$), High ($\geq 4$).

Both risk and reward are defined with respect to an expert-only completion. Risk measures the increased likelihood and severity of individual, organizational, and societal harms by using an LLM system with an expert-in-the-loop workflow. Reward measures the potential net time or cost savings, relative to human-only completion, achieved through an LLM-assisted workflow operating at expert-level quality, while accounting for the effort required to detect, verify, and mitigate model errors. We include a small set of human-verified in-context

[1] GPT-5 does not accept sampling parameters via API (e.g. `temperature`), so we use its default decoding [54].

[2] We considered weighing different risk categories, but chose not to because the relative importance can change across tasks. A fixed weighting scheme would therefore add hard-to-validate assumptions.

examples with risk-reward labels, drawn from non-Computer Occupation tasks to avoid contamination (full prompt and examples in Appendix B).

### 2.2 Occupation-level aggregation

We first average risk and reward ratings across models, respectively, where $|\mathcal{M}| = 6$.

$$\bar{R}(t) \;=\; \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} R_m(t), \quad \bar{G}(t) \;=\; \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} G_m(t). \tag{1}$$

Using the standard method[3], we normalize each of the three attributes associated with each task – Importance $I(t)$, Relevance $V(t)$, and Frequency $F(t)$ – to $[0, 1]$, with task weight defined as $\alpha(t) \;=\; \widetilde{I}(t)\cdot\widetilde{V}(t)\cdot\widetilde{F}(t)$ [50]. For an occupation $o$ with the set of tasks $\mathcal{T}_o$, we compute:[4]

$$R(o) \;=\; \frac{\sum_{t\in\mathcal{T}_o}\bar{R}(t)\cdot\alpha(t)}{\sum_{t\in\mathcal{T}_o}\alpha(t)}, G(o) \;=\; \frac{\sum_{t\in\mathcal{T}_o}\bar{G}(t)\cdot\alpha(t)}{\sum_{t\in\mathcal{T}_o}\alpha(t)} \tag{2}$$

The weighted average mirrors prior ratings [16] of AI exposure that combine task-level assessments with importance, relevance, and frequency, enabling direct comparison. We do not use non-linear aggregations as they require additional, deployment-specific assumptions about how tasks are coupled within workflows. These assumptions are difficult to justify in a general setting. Further, our weighted-average aggregation should be viewed as an optimistic baseline: if nonlinear effects [36] are present, the true adoption constraint from risk is plausibly stronger than what a linear model reports. Applying (2) yields $R(o)$ and $G(o)$ both in the $[0, 5]$ range. The occupation-level scores are weighted averages across tasks and therefore do not admit the same threshold-based interpretation as the task-level ratings ($\leq 2$ as 'low', $= 3$ as 'medium', $\geq 4$ as 'high'). Instead, comparing an occupation's aggregate reward score against its aggregate risk score allows us to identify where LLMs are likely to yield larger net gains, and where the associated risks are comparable to or exceed the rewards. We provide the responses generated by each LLM for all tasks in the supplementary material.

## 3 Results and Analysis

We present our empirical findings about our thesis on the doubly jagged frontier of LLM deployment.

[3]Let $x_{min}$ and $x_{max}$ be attribute-specific min. and max. values per O*NET. Normalized value is then computed using: $\widetilde{x}(t) = \frac{x(t)-x_{\min}}{x_{\max}-x_{\min}}$ for $x \in \{I, V, F\}$ .

[4]For the 4 occupations where the O*NET weights were not available, we followed O*NET's standard methodology to obtain approximate weights [33].

| Occupation | $R(o)$ | $G(o)$ |
|---|---|---|
| Information Security Engineers | 4.16 | 3.46 |
| Information Security Analysts | 4.05 | 3.24 |
| Penetration Testers | 3.85 | 3.56 |
| Digital Forensic Analysts | 3.85 | 2.91 |
| Blockchain Engineers | 3.61 | 3.45 |

**Risk is higher than reward**
$G(o) - R(o) < 0$

| Occupation | $R(o)$ | $G(o)$ |
|---|---|---|
| Database Administrators | 3.22 | 3.73 |
| Computer Network Architects | 3.10 | 3.76 |
| Computer Systems Analysts | 3.06 | 3.94 |
| Computer Network Support Specialists | 3.01 | 3.61 |
| Software Developers | 2.99 | 3.89 |
| Web Administrators | 2.88 | 3.81 |
| Computer and Inf. Research Scientists | 2.87 | 3.79 |
| Telecommunications Engineering Specialists | 2.87 | 3.64 |

**Reward is moderately higher than risk**
$0.5 < G(o) - R(o) < 1$

| Occupation | $R(o)$ | $G(o)$ |
|---|---|---|
| Health Informatics Specialists | 3.58 | 3.71 |
| Computer Systems Engineers/Architects | 3.38 | 3.87 |
| Network and Computer Systems Admins. | 3.34 | 3.61 |

**Reward is slightly higher than risk**
$0 \leq G(o) - R(o) \leq 0.5$

| Occupation | $R(o)$ | $G(o)$ |
|---|---|---|
| Computer Programmers | 3.03 | 4.13 |
| Data Warehousing Specialists | 2.84 | 4.0 |
| Software QA Analysts & Testers | 2.80 | 3.94 |
| Database Architects | 2.77 | 3.95 |
| Document Management Specialists | 2.69 | 3.74 |
| Information Technology Project Managers | 2.66 | 3.93 |
| Computer User Support Specialists | 2.56 | 3.81 |
| Web Developers | 2.55 | 3.94 |
| GIS Technologists & Technicians | 2.54 | 3.70 |
| Web and Digital Interface Designers | 2.29 | 3.95 |
| Video Game Designers | 1.50 | 3.91 |

**Reward is significantly higher than risk**
$G(o) - R(o) \geq 1$

Figure 2: Occupation-level average Risk $R(o)$ and Reward $G(o)$, averaged across 6 frontier models. QA: Quality Assurance, GIS: Geographic Information Systems.

## 3.1 Occupation-level jaggedness

Figure 2 reports occupation-level Risk $R(o)$ and Reward $G(o)$ scores computed using (2). We report inter-model variabiity in risk and reward scores at the occupation level in Appendix E. We group occupations by their net gain, $D(o) = G(o) - R(o)$, yielding four groups: (i) $D(o) < 0$, (ii) $0 \leq D(o) \leq 0.5$, (iii) $0.5 < D(o) < 1$, and (iv) $D(o) \geq 1$. Even under a reward-only view, $G(o)$ varies across occupations (range: $[2.70, 4.13]$), reflecting capability-driven nonuniformity. Risks increase this irregularity: while rewards for many occupations are relatively tightly concentrated, risks vary over a significantly wider range ($[1.50, 4.16]$). We observe that occupations with comparable $G(o)$ can differ significantly in $R(o)$. For example, *Health Informatics Specialists* and *Document Management Specialists* have similar rewards ($G(o) = 3.71$ vs. 3.74), but significantly different risks ($R(o) = 3.58$ vs. 2.69). The highest-risk occupations are largely security-sensitive roles: *Information Security Engineers, Information Security Analysts, Penetration Testers, Digital Forensic Analysts, and Blockchain Engineers.* This is consistent with the intuition that in security-critical settings, adversarial misuse and

high-stakes failure modes increase the harm caused by LLM use. Figures 3, 5, and 7 zoom in on within-occupation variation in the risk-reward profile through task-level risk $\bar{R}(t)$ and reward $\bar{G}(t)$ for 3 representative occupations (O*NET task ID on the horizontal axis). We also report the standard deviation of task ratings across models, which are visualized as error bars to capture inter-model variability. These plots show that even within a single occupation, tasks can have different risk-reward profiles. Some tasks offer clear gains with modest risk, while others have similar rewards but much higher risk (or vice versa), so risk does not increase smoothly with reward.

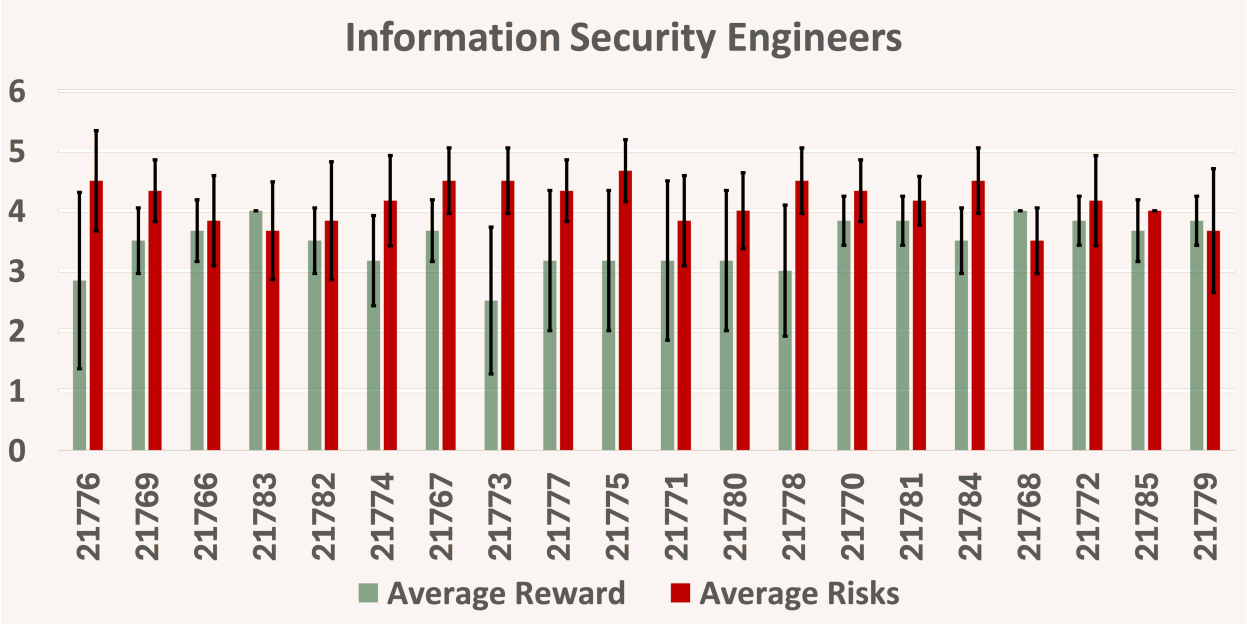


Figure 3: Average risk $\bar{R}(t)$ and reward $\bar{G}(t)$ for the 20 tasks in *Information Security Engineers*, with error bars denoting standard deviation across models.

**Case study I: Risks outweigh rewards.** Figure 3 illustrates a case of risks outweighing rewards with the occupation *Information Security Engineers*. For nearly half of the tasks, $\bar{R}(t) > \bar{G}(t) + 1$ std of rewards, and there are no tasks where $\bar{G}(t) > \bar{R}(t) + 1$ std of risks. The highest weighted task, ID 21769 - *Coordinate monitoring of networks or systems for security breaches or intrusions* has $\bar{R}(t) > \bar{G}(t) + 1$ std of rewards. All models agree that a missed intrusion can cause severe harm, and LLMs add risks such as incorrect judgments, overconfidence in flawed outputs, and susceptibility to maliciously crafted inputs. Even with human review, subtle errors can be hard to detect in time. All models flag Security, Misinformation, and Malicious Actors and Misuse risks for this task, and 5/6 also flag Privacy. Figure 4 reports how often each risk type is identified by at least 4/6 models across 20 tasks in this occupation. Both incompetence-driven risks (Privacy, Misinformation) and adversarial risks (Security, Malicious Actors and Misuse) are prevalent across tasks.

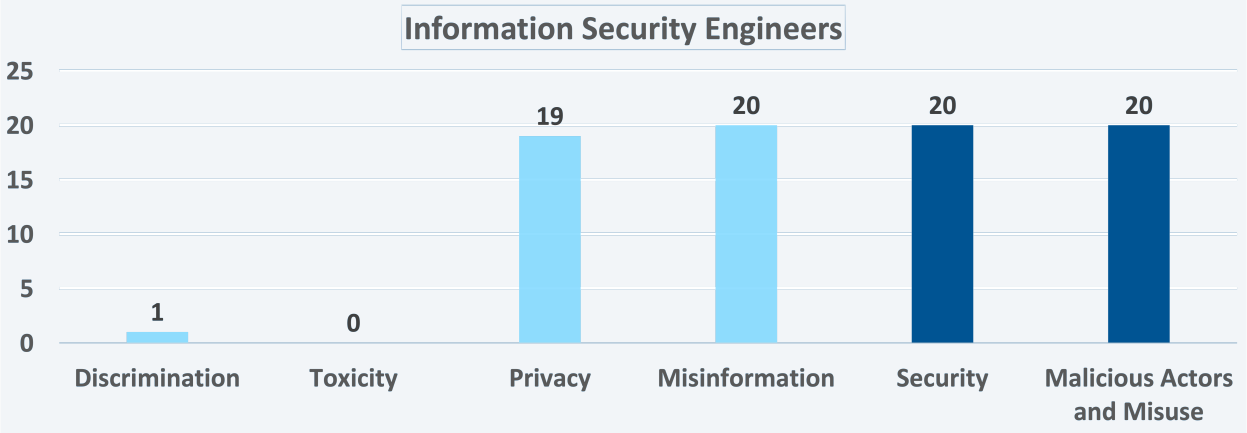


Figure 4: Frequency of the different risk categories for the 20 tasks in *Information Security Engineers*.

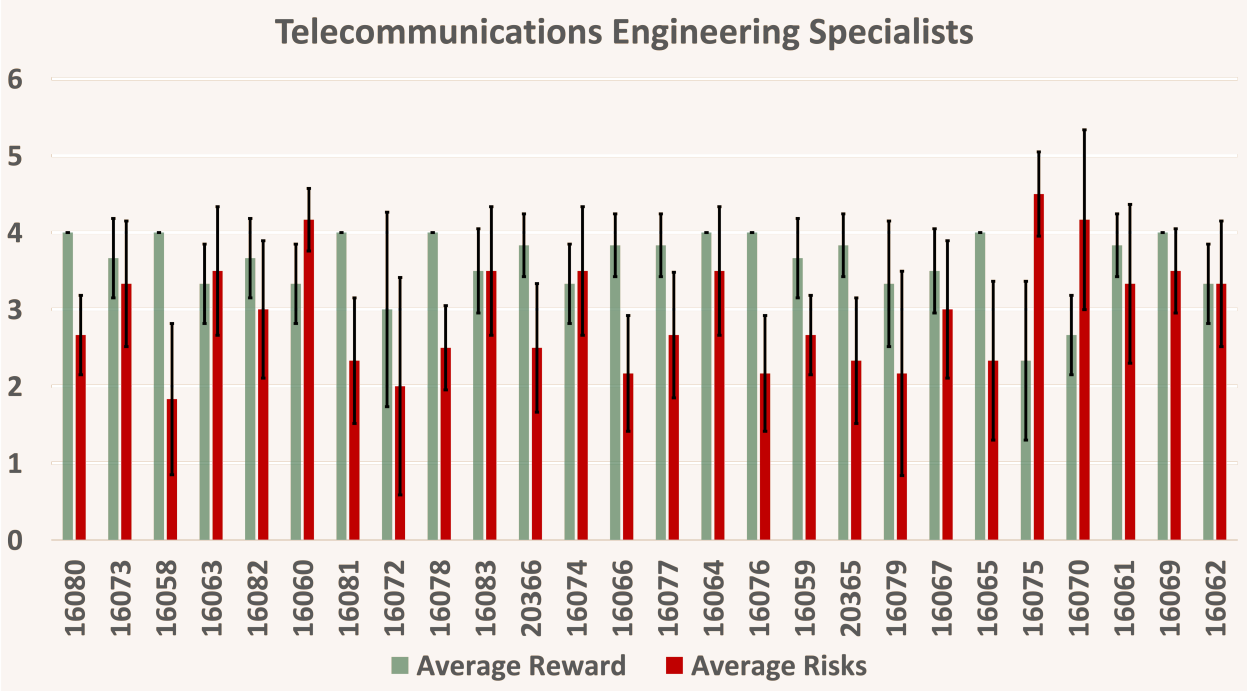


Figure 5: Average risk $\bar{R}(t)$ and reward $\bar{G}(t)$ for the 26 tasks in *Telecommunications Engineering Specialists*, with error bars denoting standard deviation across models.

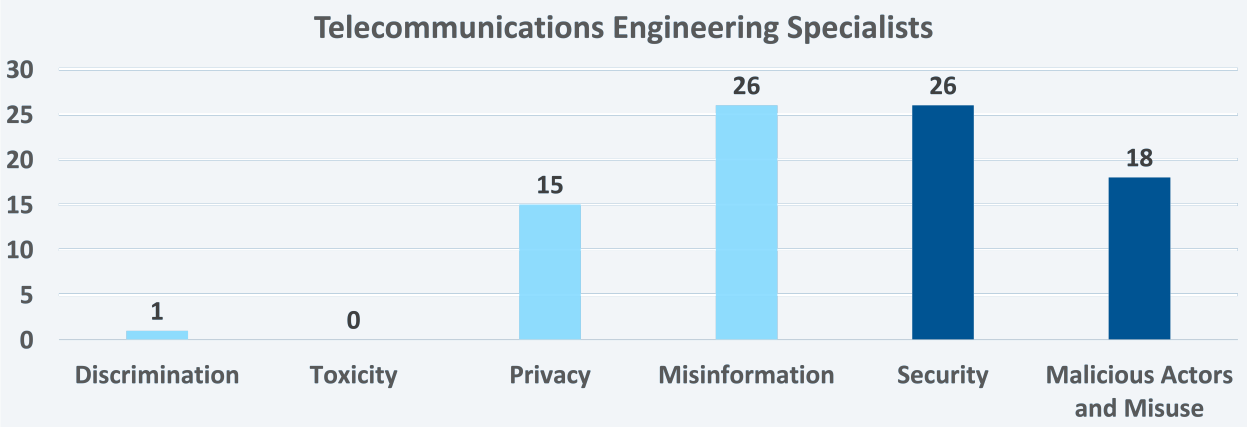


Figure 6: Frequency of the different risk categories for the 26 tasks in *Telecommunications Engineering Specialists*.

**Case study II: Rewards moderately outweigh risks.** Figure 5 depicts *Telecommunications Engineering Specialists* as a case where rewards moderately outweighs risks, but with a notable subset of tasks where risks are close to rewards or higher. Among the highest weighted is Task 16063 - *Review and evaluate requests from engineers, managers, and technicians for system modifications*. Risks here are higher than rewards, with three common risks being identified by all models: Misinformation, Security, and Malicious Actors and Misuse. Even accounting for variation among LLMs, there are tasks where $\bar{R}(t) > \bar{G}(t) + 1$ std of reward, indicating high-risk pockets within an otherwise positive average. Figure 6 shows that both incompetence-driven and adversarial risks are comparably prevalent in this occupation with Security and Misinformation applicable to all tasks.

**Case study III: Rewards significantly outweigh risks.** Figure 7 presents a case where rewards significantly outweigh risks for most tasks despite a small number of high-risk tasks, with the occupation *Document Management Specialists*. Most tasks have a favorable reward/risk tradeoff, but for the subset where $\bar{R}(t) > \bar{G}(t)$, adding +1 std to mean reward does not bridge the gap. For example, Task 16226 - *Exercise security surveillance over document processing, reproduction, distribution, storage, or archiving* ranked 8/23 by weight is one such task due to Privacy, Misinformation, Security, Malicious Actors, and Misuse risks identified by all models. While incompetence-driven and adversarial risks are both prominent, as in Figure 8. Malicious Actors and Misuse applies to fewer tasks (13/23) than does Privacy or Misinformation.

Our observations make jaggedness visible within occupations: even when tasks belong to the same occupation and are carried out by the same type of expert, the risk/reward frontier varies from task to task. This heterogeneity indicates that the risk-reward frontier is shaped by task-specific factors.

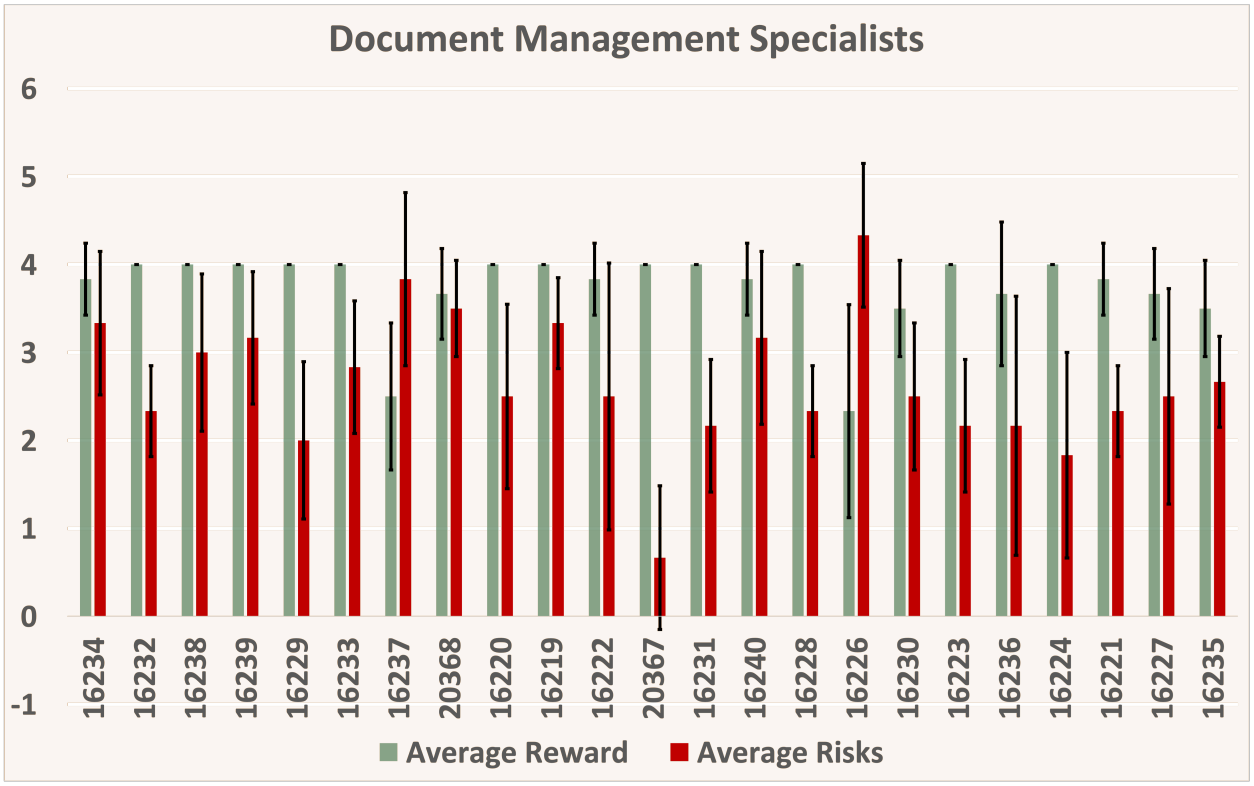


Figure 7: Average risk $\bar{R}(t)$ and reward $\bar{G}(t)$ for the 23 tasks in *Document Management Specialists*, with error bars denoting standard deviation across models.

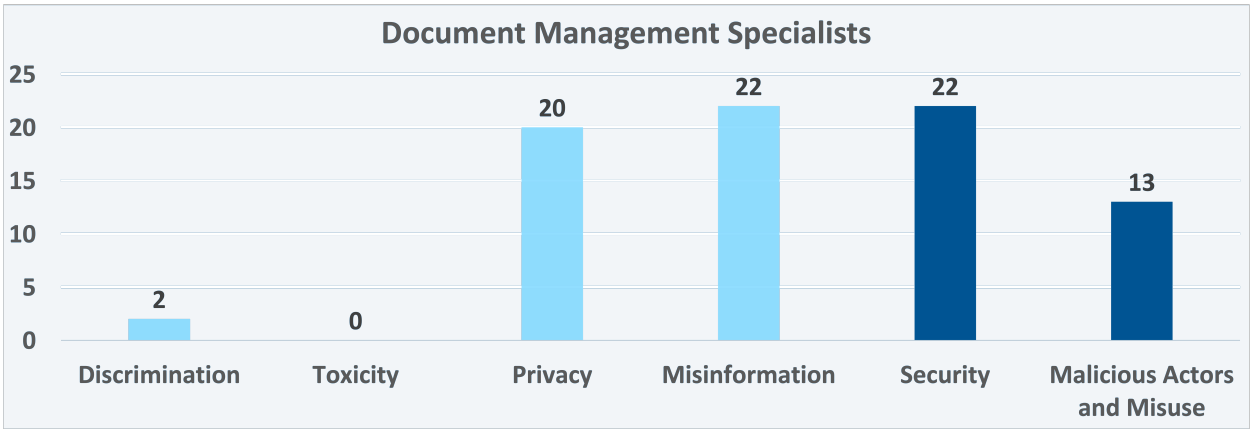


Figure 8: Frequency of risk categories for the 23 tasks in *Document Management Specialists*.

Table 1: Task count by risk-reward groups.

| | Reward level | | | |
|---|---|---|---|---|
| Risk level | Low | Medium | High | Total |
| Low | 1 | 34 | 39 | 74 |
| Medium | 4 | 224 | 233 | 461 |
| High | 3 | 70 | 1 | 74 |
| Total | 8 | 328 | 273 | 609 |

## 3.2 Task-level clustering analysis

We discretize $\bar{R}(t)$ and $\bar{G}(t)$ into three levels each: Low ($\leq 2$), Medium ($> 2$ and $< 4$), and High ($\geq 4$). This yields nine risk-reward groups over the 609 tasks as given in Table 1. To identify common task properties within each group, we cluster tasks by semantic similarity and manually analyze the clustering. We embed each task description with the `all-MiniLM-L6-v2` sentence transformer [58], reduce embeddings with UMAP [43], and cluster the reduced representations with HDBSCAN [11]. We then use GPT-4o-mini [52] to generate concise labels for each cluster and manually review and edit the labels for quality. This pipeline yields 20 task clusters, C0 to C19 (Appendix C). Finally, we aggregate the risk categories identified by the majority of models for each task by risk level (Low, Medium, High) (Table 2). As shown in Table 2, Misinformation is the most prevalent risk category and appears consistently across all levels, indicating that unreliable outputs are a concern even for low-risk tasks. In contrast, Discrimination and Toxicity occur relatively infrequently. We also observe a systematic shift in the composition of risk categories as we move from low- to higher-risk levels. Adversarial risks are relatively less in low-risk tasks but become prominent in medium- and high-risk tasks, suggesting that as model capabilities improve, lower-risk tasks are more likely to yield higher productivity gains, whereas higher-risk tasks may remain constrained by persistent adversarial vulnerabilities. We show clusters for each risk-reward group in Appendix G and present key insights below.

**High risk.** 74 tasks are high-risk, 70 of which are medium-reward. The top three clusters (Table 6 in Appendix C) in this group consist of tasks that implement cybersecurity measures and test security vulnerabilities (C4, 74%), tasks that involve implementing solutions for managing and analyzing healthcare data (C1, 7%), and network administration tasks (C14, 6%). Their high-risk rating owes to errors that can lead to acute vulnerabilities, data breaches, or compliance failures.

**Medium risk.** Most tasks are medium-risk (461/609), dominated by two categories: medium-risk, high-reward (233/461) and medium-risk, medium-reward (224/461). The former may

Table 2: Distribution of risk categories by risk level: Percentage denotes frequency of given risk category on given level (categories not mutually exclusive).

| Risk level | Risk category |
|---|---|
| Low | Misinformation (90.5%) |
| | Security (45.9%) |
| | Discrimination (18.9%) |
| | Privacy (18.9%) |
| | Malicious Actors and Misuse (12.2%) |
| | Toxicity (4.1%) |
| Medium | Misinformation (100.0%) |
| | Security (99.6%) |
| | Privacy (80.0%) |
| | Malicious Actors and Misuse (77.4%) |
| | Discrimination (16.7%) |
| | Toxicity (0.7%) |
| High | Malicious Actors and Misuse (100.0%) |
| | Security (100.0%) |
| | Misinformation (100.0%) |
| | Privacy (94.6%) |
| | Discrimination (8.1%) |
| | Toxicity (0.0%) |

present a "sweet spot" for expert-LLM collaboration where the expected gains likely outweigh the risks from LLM. By contrast, medium-risk, medium-reward sits on the borderline: the gains are plausible but not large, and the added risks are nontrivial. While all 20 clusters are present in both high- and medium-reward groups, their relative distributions appreciably differ. Software Testing and Quality Assurance (C12), Data Warehousing (C5), System Design Documentation (C13), Web Interface Development (C10), and Document Management Systems (C9) congregate in high-reward, whereas Network Hardware Installation and Troubleshooting (C19), Network Systems Administration (C14), Security Vulnerability Testing (C4), GIS Applications (C0), and Telecommunications Equipment Support (C17) concentrate in medium-reward. Within-cluster task-level jaggedness is evident, for instance, for software testing (C12) where two tasks with identical medium risks diverge in reward: Task 1267 - *Correct errors by making appropriate changes and rechecking the program to ensure that the desired results are produced* (high-reward) with significant speed-up from LLMs in error corrections and test case verification, versus Task 3478 - *Review and analyze computer printouts and performance indicators to locate code problems, and correct errors by correcting codes* (medium-reward) that includes human review and judgment. Our comparison between high- and medium-reward groups suggests greater gains from LLMs for development-oriented domains like software testing, web development, and documentation, where LLMs

can directly generate code, test cases, and structured outputs with manageable verification costs. Conversely, gains are modest for infrastructure and operations-oriented tasks such as network installation, system administration, and hardware support, likely reflecting higher operational risks, physical constraints, and the need for hands-on expertise.

**Low risk.** 74 tasks are low-risk, split between high-reward (39), medium-reward (34), and low-reward (1). Most clusters, again, appear in both high- and medium-reward categories, such as game and web interface design (C2, C10) and IT project management (C8) in the top 5 clusters for both categories. A few are unique to each group – document management (C9) predominantly in high-reward, and tasks related to training and coordination (C7), systems architecture analysis (C11), network system design (C15), and telecommunications equipment support (C17) only in medium-reward, albeit in small proportions. The sharp within-cluster difference in rewards seems more attributable to task-level exposure to LLMs than to semantic task description. For instance, in game design (C2), tasks that create content for game features (sketches, missions) are attributed higher rewards than those requiring subjective judgment of gameplay experiences or those involving soliciting or giving feedback to design and technical staff. We discuss the less prominent groups in Appendix F.

**Deployment evidence.** We analyzed the recent Anthropic Economic Index usage data for computer occupations [3]. We track `onet_task_count`, which denotes the number of Claude conversation instances mapped to a given O*NET task. The top 10 tasks by `onet_task_count` account for the majority of the mapped conversations. Among these, 8 are high-reward, medium-risk, and the other two are high-reward, low-risk, and medium-reward, medium-risk (see Table 9 in Appendix I). This concentration matches the "sweet spot" highlighted by our analysis. The associated occupations (e.g., Web Developers, Computer Programmers) have positive $G(o) - R(o)$. High-risk tasks only have minimal representation, which is consistent with our risk-aware deployment perspective.

Table 3: Human expert reviewer agreement with LLM-generated reward and risk scores. Agreement rates are reported as proportions with 95% Wilson score confidence intervals. Each reviewer independently evaluated 90 stratified tasks.

| **Reviewer** | **Reward Agree %** | **95% CI** | **Risk Agree %** | **95% CI** |
|---|---|---|---|---|
| Reviewer 1 | 94.4 | [87.6, 97.6] | 84.4 | [75.6, 90.5] |
| Reviewer 2 | 97.8 | [92.3, 99.4] | 91.1 | [83.4, 95.4] |
| Reviewer 3 | 97.8 | [92.3, 99.4] | 91.1 | [83.4, 95.4] |
| Reviewer 4 | 100.0 | [95.9, 100.0] | 92.2 | [84.8, 96.2] |
| Reviewer 5 | 97.8 | [92.3, 99.4] | 96.7 | [90.7, 98.9] |
| Reviewer 6 | 95.6 | [89.1, 98.3] | 100.0 | [95.9, 100.0] |
| **Aggregate** | **97.2** | **[95.5, 98.3]** | **92.6** | **[90.1, 94.5]** |

### 3.3 Validation of LLM-as-Judge scores with human studies

To validate the reliability of our LLM-as-judge approach, 6 reviewers with domain expertise independently reviewed a stratified sample of 90 tasks drawn from the full dataset. The sample was divided equally across three gain categories: 30 high-gain tasks, 30 low-gain tasks, and 30 medium-gain tasks, where gain is $\bar{G}(t) - \bar{R}(t)$. At most four tasks per occupation were included to ensure occupational diversity. For each task, reviewers were presented with the task description, the average reward and risk scores and their standard deviations across the six LLMs, and a synthesized summary generated by GPT-4o-mini from the aggregated LLM justifications. Reviewers independently rated their agreement with the LLM summary and average scores on a three-point ordinal scale of Agree, Partially Agree, or Disagree, separately for the reward and risk dimensions. As shown in Table 3, across all six reviewers (N=540 ratings per dimension), strict agreement rates were 97.2% [95.5, 98.3] for reward and 92.6% [90.1, 94.5] for risk (95% Wilson score confidence intervals). We also present case studies discussing LLM responses to representative tasks in Appendix A.

Beyond the human review, we compared our LLM-generated scores against the Workbank database from Stanford's SALT NLP lab [59], which contains task-level ratings collected from 1,361 domain workers across 104 occupations. For each O*NET task, workers rated task characteristics such as domain expertise required, involved uncertainty, and physical action requirement on 1-5 scales. They also provided self-assessments, including automation desire, human agency (the degree of human collaboration needed with AI), and reasons for wanting or resisting automation. A separate panel of experts independently rated the automation capacity of each task. We matched 144 of our 609 tasks to this dataset, covering 15 of our 27 occupations. A distributional comparison confirmed that the 144-task subset is representative of the full dataset in terms of both reward and risk scores (Kolmogorov-Smirnov test, $p > 0.17$ for both). We computed Spearman rank correlations between the Workbank variables and our LLM scores. Although the variables measured in the Workbank study differ from our LLM-rated risk and reward, these correlations serve as a proxy to assess whether our LLM-generated scores align with task characteristics as rated by domain workers. Since reward and risk capture opposing dimensions, we computed the correlation of Workbank variables with net gain (reward minus risk), which provides a single measure of the LLM-assessed benefit of automation. The gain measure showed a clear alignment with worker assessments (Figure 9). Tasks where workers reported higher domain expertise requirements had significantly lower LLM-assessed gain, indicating that LLMs recognize the diminished net benefit of automating highly specialized tasks. Similarly, tasks involving more uncertainty and those where workers cited human error as a concern also showed lower gain. On the other hand, tasks where workers indicated that more human collaboration is needed with

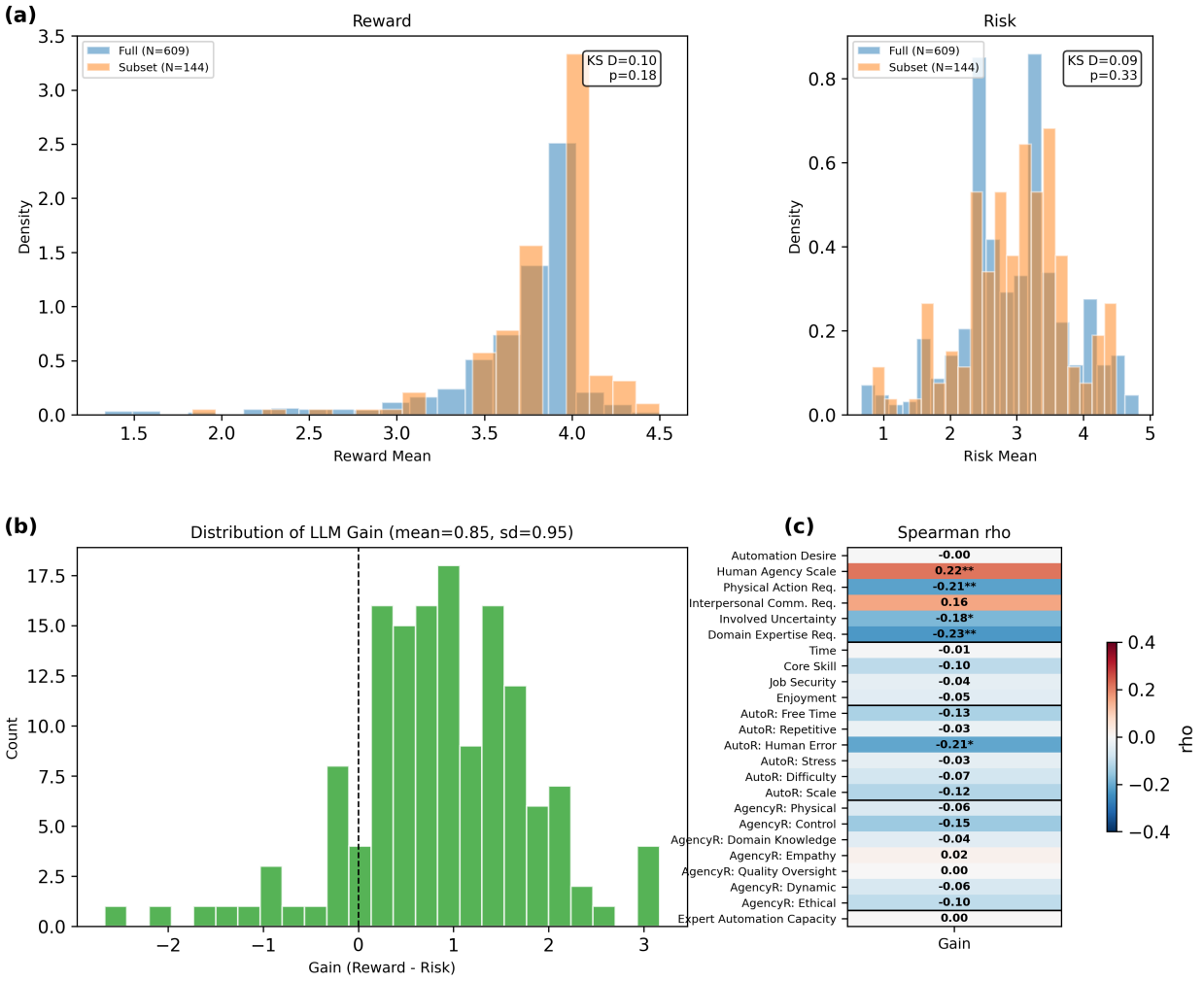


Figure 9: External validation of LLM-generated risk and reward scores against the Workbank domain worker survey. **(a)** Sample representativeness: distributions of LLM reward (left) and risk (right) scores for the full dataset ($N = 609$ tasks, blue) and the 144-task subset matched to Workbank (orange). Kolmogorov-Smirnov tests confirm no significant distributional differences (. **(b)** Distribution of LLM net gain (reward − risk) for the 144 matched tasks. **(c)** Spearman rank correlations between Workbank worker/expert ratings and LLM gain scores. $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$.

AI showed higher gain, consistent with our LLM-expert collaboration setting. Interestingly, neither workers' self-reported automation desire nor experts' ratings of automation capacity correlated with gain, suggesting that LLM-generated risk-reward scores capture a risk-reward dimension distinct from both worker preference and technical feasibility.

The high agreement rates from expert reviewers and the significant correlations with independently collected worker assessments provide evidence that our LLM-as-judge scoring is reliable.

# 4 Call to Action: Policy Implications

**Risk-aware labor market projections.** Capability-centered labor market forecasts can overstate displacement by focusing on whether a given task can be accomplished with technology. Workplace data however suggests that augmentation and partial automation, rather than full displacement, dominate the AI adoption landscape [3, 20]. The uneven net gains across tasks within an occupation warrant a risk-aware labor market projection framework. The double-layered jaggedness in our analysis predicts employment loss only when an occupation's core tasks have consistently high reward relative to risk under oversight, with the strongest labor-demand reductions in task bundles with positive $\bar{G}(t) - \bar{R}(t)$. This perspective is consistent with our observations from the Anthropic Economic Index above. The risk-aware framework should also distinguish between types of labor, such as between low-stakes tasks, where LLMs generate outputs amenable to human verification, and high-stakes tasks that must remain human-led, where one minor mistake can cause enormous harm or be exploited by attackers. Finally, such a framework should anticipate new occupation classes centered on risk measurement, oversight, and incident response that make deployment reliable.

**Risk-informed agility in AI governance.** For a risk frontier that is both jagged and fast evolving, conventional static regulatory regimes may do more harm than good. One concrete improvement based on our analysis is to give greater consideration to liability. This has two benefits: First, credible liability exposure creates incentives to invest in documentation, auditing, and safer deployment practices [37]. By making explicit such factors as the consumer's decision between competing technologies, liability encourages shared responsibility among producers, finetuners, and users of AI [18, 69]. Second, it helps mitigate AI's emerging "moral crumple zones" where the human user, whom AI should assist and benefit, ends up absorbing ethical, legal, and reputational damage for the preservation and advancement of the technology [23, 69]. Human operators of AI, for one, could be held responsible for failures of autonomous systems despite having no meaningful control over the

system's actions. Risk-informed assessment of liability steers AI governance away from a rigid, one-size-fit-all regime that obscures massive heterogeneity in risk profiles across scenarios and toward an agile and flexible framework for evaluating individual cases in context.

**Cumulative, not just catastrophic, AI risks.** National security discussions of AI misuse are preoccupied with catastrophic scenarios (e.g. CBRN) [71]. Our results underscore a more quotidian but no less impactful setting: security- and infrastructure-adjacent software work, where small errors can precipitate grave downstream consequences. In highlighting the central role of Security risks in driving risk jaggedness, our analysis challenges the view that model improvements strengthen cyber defense [42]. LLMs may raise productivity just as they may widen the attack surface by introducing subtle vulnerabilities that survive review. This calls for keener attention to assurance under realistic attacker modes in the relevant national-security discussions on AI deployment and defense.

**Incentivizing risk-focused research.** Our findings accentuate the need to rebalance funding priorities. With global AI investment projected to reach multi-trillion dollars in 2026 [7], such investment continues to skew toward accelerating performance over enhancing safety. AI safety and security remain underfunded relative to the scale of potential risks. Productivity gains from AI, as suggested in prevailing studies require broad enterprise uptake, which is often limited by integration cost, reliability, and governance constraints [48]. A risk-aware funding strategy should incentivize deployability, such as by incorporating evaluation under realistic threat models, secure-by-design workflows, and formal methods [60]. Given the transnational nature of AI security threats, private entities must play a proactive role in funding safety research not only to fill the gap in government funding but also to encourage innovative research that may not fit well with a conventional national research agenda.[5]

**Deployment focused measurements.** Real-world tasks, including those represented in O*NET, often require combining multiple capabilities. Even if a model performs well on benchmarks that measure individual capabilities separately, its performance on a real task may differ significantly once those capabilities are combined for a real-world task. Most existing academic benchmarks measure capabilities and risks in narrow settings. GDPval [57] and ITBench [34] are useful steps toward addressing this gap in capability evaluation. For risks, evaluations should measure risks in realistic tasks. Further, both risk evaluation and mitigation should be done with respect to the LLM output distribution and not on a few sampled responses [65, 70]. Another important step is to develop mappings from occupational task taxonomies to relevant LLM capability and risk measurements.

[5]See, for example, SecureDNA – a free, automated privacy-preserving system for screening global DNA synthesis built with private funding by scientists from the U.S., China, the European Union, and Israel [6].

# 5 Related Work

**Economic impact of AI.** Capability-based AI exposure measures remain influential in economic forecasting and workforce planning [68, 27, 26, 24, 17, 16, 13], often using the O*NET task descriptors [49]. Our results contrast with these exposure-only analyses. For example, [16] uses 3 LLMs to compute task-level capability ratings and aggregate them into occupation-level exposure and replacement indexes (TEAI/TRAI) weighted by O*NET task Relevance, Importance, and Frequency. Their analysis deems *Information Security Engineers* and *Information Security Analysts* as highly exposed to AI as many tasks appear automatable. Similarly, [24] uses GPTs to assess occupational exposure to LLMs without modeling risk and identify *Blockchain Engineers* as highly exposed to LLMs. In our framework, however, these same occupations are among the highest risk roles. This mismatch shows why capability-only or upside-only measures can be misleading for deployment decisions: **high exposure does not imply a favorable risk–reward frontier** once misuse and adversarial exploitation are taken into account. Our occupation-level $R(o)$ and $G(o)$ make these tradeoffs explicit and thereby surface jaggedness that exposure alone cannot capture.

Complementary empirical studies [8, 47] measure realized productivity gains in domains such as customer support and writing. Other work analyzes usage and adoption patterns, showing that LLM deployment is concentrated in specific tasks and occupations rather than uniformly distributed [20, 3, 13]. The work of [59] introduces a task-level auditing framework that jointly considers model capability and worker preferences, revealing systematic mismatches between what AI can do and what users want AI to do. Market- and adoption-based perspectives [21, 45] emphasize that economic impact is mediated by incentives, diffusion, and organizational integration rather than technical feasibility alone. As for exposure studies, these works do not explicitly model the role of risk in constraining adoption.

**LLM evaluation and risk mitigation.** Traditional capability evaluation relies on static benchmarks and aggregate scores, which may not reflect real-world workflows or generalization behavior [41, 63]. More recent approaches propose metrics that better approximate practical usefulness, such as the task completion time horizon [38]. Similarly, evaluations grounded in realistic workflows show that benchmark performance can diverge significantly from real-world utility due to factors such as verification overhead, tool integration, and error propagation [57]. In line with prior frameworks for emerging technology governance [62, 44], rich taxonomies for characterizing harms and managing risks have now been developed in AI governance [61, 46]. The security community has likewise compiled concrete threat classes and failure modes for LLM applications, including prompt injection, insecure outputs, data poisoning, and excessive agency [56]. More recent work provides formal guarantees on LLM

safety and security [14, 65]. Alignment techniques aim to mitigate risks at both training time and inference time. Training-time approaches such as reinforcement learning from human feedback (RLHF) and related preference optimization methods aim to steer model behavior toward desired outputs [55, 4], while test-time techniques such as controlled generation [28], constrained decoding [5], guardrails [22], and activation steering [72] seek to enforce safety during generation. Existing approaches primarily focus on reducing undesirable behaviors at the model or system level, but do not provide a systematic framework for quantifying how risks affect productivity gains across real-world tasks. In practice, even well-aligned models can fail under adversarial manipulation, and the costs of verification and mitigation can vary significantly across tasks.

**Systemic risks.** We have argued that policies should consider the jaggedness of the AI safety frontier, in terms of developing liability regimes and focusing research attention on safety of AI models. An alternative view is that, instead of considering anything specifically about AI models, questions of risks should be addressed at broader systems levels. In particular, rather than considering Buchanan's traditional AI triad of data, algorithms, and compute [9] as the point of improvement, one should rather consider Varshney's AI safety triad of DevOps, AIOps, and cybersecurity [15] as the key to safe and secure deployment of AI in consequential settings.

Finally, our task-based risk–reward scores may not capture *systemic risks* or emergent effects that arise at the ecosystem level [12]. Our task-level analysis is complementary: it makes deployment tradeoffs concrete within occupations, while systemic-risk approaches address cross-cutting effects that extend beyond any single occupation.

# 6 Conclusion

We show that capability alone does not determine the deployability of AI technologies. Tasks across O*NET Computer Occupations exhibit jaggedness in both reward and risk. Exposure-based analysis may dangerously inflate AI deployability for tasks that are automatable but vulnerable to exploitation and misuse. Based on these observations, we argue for systematic research on risk-driven jaggedness.

# References


[1] Anthropic. What's new in Claude 4.5. Claude Developer Docs, 2025.

[2] Anthropic. Claude Mythos Preview. `https://red.anthropic.com/2026/mythos-preview/`, 2026.

[3] Ruth Appel, Maxim Massenkoff, Peter McCrory, Miles McCain, Ryan Heller, Tyler Neylon, and Alex Tamkin. Anthropic economic index report: economic primitives, 2026.

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, John Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Chris Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, E Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, J Landau, Kamal Ndousse, Kamilė Lukoiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noem'i Mercado, Nova Dassarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Thomas Henighan, Tristan Hume, Sam Bowman, Zac Hatfield-Dodds, Benjamin Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom B. Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. *ArXiv*, abs/2212.08073, 2022.

[5] Debangshu Banerjee, Changming Xu, Eugene Ie, Ming Zhang, Daiyi Peng, Chu-Cheng Lin, and Gagandeep Singh. Severa: Verified synthesis of self-evolving agents, 2026.

[6] Carsten Baum, Jens Berlips, Walther Chen, Helena Cozzarini, Hongrui Cui, Ivan Damgård, Jiangbin Dong, Kevin M Esvelt, Leonard Foner, Mingyu Gao, et al. A system capable of verifiably and privately screening global dna synthesis. *National Science Review*, page nwag103, 2026.

[7] Alexander Berger and Liz Givens. AI safety and security need more funders. Coefficient Giving, 2025.

[8] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work, 2024.

[9] Ben Buchanan. The AI triad and what it means for national security strategy. Technical report, Center for Security and Emerging Technology, 2020.

[10] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data*, 10(1), July 2015.

[11] Ricardo JGB Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):1–51, 2015.

[12] Samuel Carey. Regulating uncertainty: Governing general-purpose AI models and systemic risk. *European Journal of Risk Regulation*, page 1–17, 2025.

[13] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research, 2025.

[14] Isha Chaudhary, Vedaant V. Jain, Avaljot Singh, Kavya Sachdeva, Sayan Ranu, and Gagandeep Singh. Lumos: Let there be language model system certification. *CoRR*, abs/2512.02966, 2025.

[15] Alvin Chin and Lav Varshney. AI is here. how can we make it less scary and safer? *Chicago Tribune*, 2024.

[16] Emilio Colombo, Fabio Mercorio, Mario Mezzanzanica, and Antonio Serino. Towards the terminator economy: Assessing job exposure to AI through LLMs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI-2025, page 9591–9600, September 2025.

[17] Council of Economic Advisers. Potential labor market impacts of artificial intelligence: An empirical analysis. The White House, July 2024. Released July 2024.

[18] Ruth Schwartz Cowan. The consumption junction: A proposal for research strategies in the sociology of technology. In Wiebe E. Bijker, Thomas P. Hughes, and Trevor J. Pinch, editors, *The Social Construction of Technological Systems*, pages 261–280. MIT Press, Cambridge, MA, USA, 1987.

[19] Fabrizio Dell'Acqua, Edward McFowland III, Ethan Mollick, Hila Lifshitz, Katherine C Kellogg, Saran Rajendran, Lisa Krayer, Francois Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. Technical report, Harvard Business School Working Paper, 2023.

[20] Eleanor Wiske Dillon, Sonia Jaffe, Nicole Immorlica, and Christopher T. Stanton. Shifting work patterns with generative ai, 2025.

[21] Jeffrey Ding. *Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition*. Princeton University Press, 2024.

[22] Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Building guardrails for large language models, 2024.

[23] Madeleine Clare Elish. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5:40–60, 2019.

[24] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: An early look at the labor market impact potential of large language models, 2023.

[25] European Union. EU Artificial Intelligence Act, Article 14: Human Oversight. `https://artificialintelligenceact.eu/article/14/`, 2024.

[26] Ed Felten, Manav Raj, and Robert Seamans. How will language modelers like ChatGPT affect occupations and industries?, 2023.

[27] Edward Felten, Manav Raj, and Robert Seamans. Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 42(12):2195–2217, 2021.

[28] Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Tianrui Guan, Mengdi Wang, Alvaro Velasquez, Ahmad Beirami, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment, 2025.

[29] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

[30] Google Cloud. Gemini 2.5 Pro (Vertex AI model documentation). Vertex AI Documentation, 2025.

[31] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, et al. The Llama 3 herd of models, 2024.

[32] Mehul Gupta. Andrej karpathy just dropped a job risk map for the ai era. *Data Science in Your Pocket*, March 2026. Medium article.

[33] Jessica L. Harris and Jeffrey A. Dahlke. Updates to related occupations for the O*NET program using the O*NET 30.0 Database. O*NET Resource Center report page (HumRRO), November 2025. Published November 2025.

[34] Saurabh Jha, Rohan R. Arora, Yuji Watanabe, Takumi Yanagawa, Yinfang Chen, Jackson Clark, Bhavya, Mudit Verma, Harshit Kumar, Hirokuni Kitahara, Noah Zheutlin, Saki Takano, Divya Pathak, Felix George, Xinbo Wu, Bekir O. Turkkan, Gerard Vanloo, Michael Nidd, Ting Dai, Oishik Chatterjee, Pranjal Gupta, Suranjana Samanta, Pooja Aggarwal, Rong Lee, Jae-wook Ahn, Debanjana Kar, Amit M. Paradkar, Yu Deng, Pratibha Moogi, Prateeti Mohapatra, Naoki Abe, Chandrasekhar Narayanaswami, Tianyin Xu, Lav R. Varshney, Ruchi Mahindru, Anca Sailer, Laura Shwartz, Daby Sow, Nicholas C. Fuller, and Ruchir Puri. Itbench: Evaluating AI agents across diverse real-world IT automation tasks. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025.

[35] Andrej Karpathy. Ai exposure of the us job market. `https://karpathy.ai/jobs/`, 2026. Interactive visualization tool analyzing AI exposure across 342 U.S. occupations using BLS data.

[36] Michael Kremer. The o-ring theory of economic development. *The Quarterly Journal of Economics*, 108(3):551–575, 1993.

[37] Rune Kvist, Rajiv Dattani, and Brandon Wang. Underwriting superintelligence, 2025.

[38] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Roa Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long software tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[39] P. Langley. Crafting papers on machine learning. In Pat Langley, editor, *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

[40] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-Judges: A comprehensive survey on LLM-based evaluation methods, 2024.

[41] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

[42] Xiaoqun Liu, Jiacheng Liang, Qiben Yan, Jiyong Jang, Sicheng Mao, Muchao Ye, Jinyuan Jia, and Zhaohan Xi. Cylens: Towards reinventing cyber threat intelligence in the paradigm of agentic large language models, 2025.

[43] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[44] M Granger Morgan. Risk analysis and management. *Scientific American*, 269(1):32–41, 1993.

[45] Arvind Narayanan and Sayash Kapoor. AI as Normal Technology. *Knight First Amendment Institute*, 2025.

[46] NIST. Artificial intelligence risk management framework (AI RMF 1.0), 2023.

[47] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

[48] OECD. AI adoption by small and medium-sized enterprises. Technical report, Organisation for Economic Co-operation and Development, December 2025.

[49] O*NET OnLine. O*NET OnLine. `https://www.onetonline.org/`.

[50] O*NET OnLine. Scales, ratings, and standardized scores. `https://www.onetonline.org/help/online/scales`.

[51] O*NET Resource Center. Using a hybrid artificial intelligence-expert method to develop work style ratings for the o*net database, 2025.

[52] OpenAI. GPT-4o mini (model documentation). `https://platform.openai.com/docs/models/gpt-4o-mini`, 2024.

[53] OpenAI. Introducing GPT-5. OpenAI Blog, 2025.

[54] OpenAI. Using GPT-5.2 (GPT-5 model family guide). OpenAI API Documentation, 2025.

[55] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS'22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[56] OWASP Foundation. OWASP top 10 for large language model applications. OWASP GenAI Security Project, 2023.

[57] Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simon Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. 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. In *The Fourteenth International Conference on Learning Representations*, 2026.

[58] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[59] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with ai agents: Auditing automation and augmentation potential across the u.s. workforce, 2026.

[60] Gagandeep Singh and Deepika Chawla. Position: Formal methods are the principled foundation of safe AI. In *ICML Workshop on Technical AI Governance (TAIG)*, 2025.

[61] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The AI

risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence, 2025.

[62] Paul Slovic. Perception of risk. In *The Perception of Risk*, pages 220–231. Routledge, 2016.

[63] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden

Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib

Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[64] Stack Overflow. Stack Overflow Developer Survey 2025: AI. `https://survey.stackoverflow.co/2025/ai/`, 2025.

[65] Tarun Suresh, Nalin Wadhwa, Debangshu Banerjee, and Gagandeep Singh. BEAVER: an efficient deterministic LLM verifier. *CoRR*, abs/2512.05439, 2025.

[66] Tech Desk. Ai may replace the highest-paid workers first, warns ex-tesla ai head andrej karpathy. *The Financial Express*, March 2026.

[67] Helen Toner. Taking jaggedness seriously, November 2025.

[68] Lav R. Varshney. Impact of AI on employment. *Digital Skills Insights*, pages 40–47, 2020. International Telecommunications Union.

[69] Lav R. Varshney, Nitish Shirish Keskar, and Richard Socher. Pretrained AI models: Performativity, mobility, and change. arXiv:1909.03290 [cs.CY], 2019.

[70] Jason Vega and Gagandeep Singh. Matching ranks over probability yields truly deep safety alignment. *CoRR*, abs/2512.05518, 2025.

[71] Chengxiao Wang, Isha Chaudhary, Qian Hu, Weitong Ruan, Rahul Gupta, and Gagandeep Singh. Quantifying risks in multi-turn conversation with large language models, 2025.

[72] Zhengxuan Wu, Qinan Yu, Aryaman Arora, Christopher D Manning, and Christopher Potts. Improved representation steering for language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[73] xAI. Grok 4. xAI News, 2025.

[74] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

# A Alignment with Expert Judgment

We have conducted extensive prompt tuning to maximize agreement among the 6 models. We have also tested additional models, such as those in the Mistral and Phi families, which tend to be less reliable in producing clear, task-grounded ratings. Using variants within the same family does not add meaningful diversity. We therefore consider a set of frontier models that is both reliable and diverse in how it interprets task statements. Table 4 reports the variation in the risk and reward scores across all 609 tasks by each model.

Next, we present LLM ratings together with summaries of the underlying LLM responses based on GPT-4o-mini. We highlight three representative cases of agreement between LLM and experts below. These examples illustrate why the human reviewers agreed with the LLM assessments.

**Occupation**: Software Developers (High-reward, Low-risk)

**Task**: Prepare reports or correspondence concerning project specifications, activities, or status.

**Avg. reward**: 4,

**Summary**: The six LLMs unanimously recognize the high rewards of using AI assistance for drafting and refining project reports, status updates, and correspondence. They highlight that LLMs excel in converting technical jargon into stakeholder-friendly language, summarizing meeting notes, and generating structured documentation.

**Avg. Risk**: 1.83

**Summary**: The six models generally agree that while this task is not highly safety-critical, it does carry moderate risks. Common concerns include misinformation, such as LLMs hallucinating details or misrepresenting timelines, and privacy/security breaches involving proprietary information. Most models emphasize that human oversight is crucial in mitigating these risks, as developers typically review and verify LLM-generated content against actual project data.

**Occupation**: Information Security Analysts (High-risk, Low-reward)

**Task**: Modify computer security files to incorporate new software, correct errors, or change individual access status.

**Avg. Reward**: 1.83,

**Summary**: The models generally agree that while AI assistance can provide some benefits in drafting and generating initial configurations or documentation for modifying computer security files, the overall rewards are limited due to the high necessity for human oversight.

**Avg. Risk**: 4.5,

**Summary**: The task is universally recognized as highly safety-critical due to the potential

for significant security breaches, unauthorized access, and compliance violations stemming from even minor errors. All models emphasize the heightened risks associated with using LLMs, including the generation of syntactically correct but logically flawed configurations, the potential for prompt injection attacks, and the inadvertent exposure of sensitive information. They collectively highlight that while human oversight can mitigate some risks, subtle vulnerabilities may still evade detection.

**Occupation**: Computer and Information Research Scientists (Medium reward and risk)

**Task**: Participate in staffing decisions and direct training of subordinates.

**Avg. Reward**: 3.67,

**Summary**: The six models generally agree that using AI assistance can lead to significant time and cost savings. They emphasize that AI should support operational aspects, such as drafting job descriptions, generating interview questions, and creating training materials, while humans retain final decision-making authority.

**Avg. Risk**: 3.5,

**Summary**: The six models unanimously agree that using AI assistance is associated with significant risks, primarily due to the potential for biased evaluations, privacy violations, and misinformation. All models acknowledge that human oversight is essential to mitigate these risks, but they also caution that subtle biases and errors may still persist.

We next examine the most disputed case, where some of the reviewers assessed the risk as higher than the average LLM rating. This disagreement stems from differing views on the extent to which LLMs are involved in the task.

**Occupation**: Network and Computer Systems Administrators

**Task**: Perform data backups and disaster recovery operations.

**Avg. Reward**: 3.17

**Summary**: The six models generally agree that AI assistance can provide meaningful support in data backups and disaster recovery operations, primarily through documentation, script generation, and automation of routine tasks.

**Avg. Risk**: 3.17

**Summary**: The models unanimously agree that this is a safety-critical task. They emphasize that the involvement of LLMs introduces various risks, including incorrect configurations, misinformation, and security vulnerabilities. While most models acknowledge that experienced human oversight can mitigate many risks, they also caution that subtle errors may still go undetected. Qwen diverges from the consensus by arguing that the deterministic nature of backup operations and the requirement for human approval in LLM interactions significantly reduce the associated risks, suggesting that LLMs pose minimal threat as long as they are not directly integrated into operational workflows.

Table 4: Average and standard deviation (SD) of reward and risk scores across all 609 tasks by each model.

| Model | Avg. Reward ± SD | Avg. Risk ± SD |
|---|---|---|
| Qwen3-80B | $3.61 \pm 0.89$ | $2.22 \pm 1.08$ |
| Llama-3.3-70B | $3.86 \pm 0.71$ | $3.21 \pm 1.12$ |
| Claude Opus 4.5 | $3.48 \pm 0.66$ | $2.36 \pm 0.80$ |
| Gemini-2.5 Pro | $3.79 \pm 0.56$ | $3.63 \pm 0.74$ |
| Grok-4 | $3.95 \pm 0.39$ | $2.66 \pm 1.10$ |
| GPT-5 | $3.72 \pm 0.47$ | $2.51 \pm 0.84$ |

# B Model prompt

For each task within different occupations, determine the risks and rewards of using an LLM system (e.g., LLMs, LLM-based agents) in collaboration with a human expert to complete the task. For determination, use your knowledge of the current capabilities of the best-performing LLM systems. First, determine whether the task is safety-critical, i.e., any unintentional errors or exploitation of LLM capabilities by malicious actors can cause significant harm at the individual, organizational, or societal level. If yes, then quantify the extent to which using the combination of LLM Systems + Human Experts in performing the task increases the risk compared to only having human experts complete the task. Considering the cost of detecting and mitigating risks with experts, determine if the combination of LLM systems + Human Experts can significantly decrease the cost (in USD) or time of completing the task while keeping the same quality as expert humans.

Key vulnerabilities of LLM systems to consider include:

1. **Discrimination**: Unequal treatment of individuals or groups by LLM systems, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and representation of those groups. Further, the accuracy and effectiveness of LLM decisions and actions are dependent on group membership, where decisions in LLM system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.

2. **Toxicity**: LLMs can expose users to harmful, abusive, unsafe, or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory, political speech, or pornography.

3. **Privacy**: LLM systems memorize and leak sensitive personal data or infer private

information about individuals without their consent. Unexpected or unauthorized sharing of data and information can compromise user expectations of privacy, assist identity theft, or cause loss of confidential intellectual property.

4. **Security**: Vulnerabilities can be exploited in LLM systems, software development toolchains, and hardware, resulting in unauthorized access, data and privacy breaches, or system manipulation causing unsafe outputs or behavior.

5. **Misinformation**: LLM systems can inadvertently generate or spread incorrect or deceptive information, which can lead to inaccurate beliefs in users and undermine their autonomy. Humans who make decisions based on false beliefs can experience physical, emotional, or material harm. Highly personalized LLM-generated misinformation creates "filter bubbles" where individuals only see what matches their existing beliefs, undermining shared reality and weakening social cohesion and political processes.

6. **Malicious Actors and Misuse**: LLM systems can be used to conduct large-scale disinformation campaigns, malicious surveillance, or targeted and sophisticated automated censorship and propaganda, to manipulate political processes, public opinion, and behavior. LLM systems can be used to develop cyber weapons (e.g., by coding cheaper, more effective malware), develop new or enhance existing weapons (e.g., Lethal Autonomous Weapons or chemical, biological, radiological, nuclear, and high-yield explosives), or use weapons to cause mass harm. LLM systems can be used to gain a personal advantage over others through cheating, fraud, scams, blackmail, or targeted manipulation of beliefs or behavior. Examples include LLM-facilitated plagiarism for research or education, impersonating a trusted or fake individual for illegitimate financial benefit, or creating humiliating or sexual imagery.

For each evaluation:

- Rate both the reward and risk associated with LLM systems on a scale from 0 to 5. Here, for rewards, a rating of 0 means that using LLM systems cannot reduce overall task completion time or cost while keeping equivalent quality, rating $\leq 2$ means relatively low reduction in overall cost or time or both, rating of 3 means moderate reduction, and $\geq 4$ means high reduction. A rating of 0 for risks means that the task is not safety-critical, a risk rating $\leq 2$ means it can be easily detected and/or fixed by human experts, a rating of 3 means that the risk is moderate, and rating $\geq 4$ means the risk is high and cannot be easily detected.

- Provide a detailed justification in the format: [risks_rating, "risks_types", "risks_justification", rewards_rating, "rewards_justification"]

**In-Context Examples**

1. **Profession**: Mathematicians

   **Task**: Conduct research to extend mathematical knowledge in traditional areas, such as algebra, geometry, probability, and logic.

   **Risk Rating**: 0

   **Risk Types:** No Risks

   **Risk Justification:** The task is not safety critical, and therefore, there is no risk.

   **Reward Rating:** 2

   **Reward Justification:** LLMs can speed up ideation, proof sketching, counterexample search, literature triage, and formalization/writeup support, with the expert validating all claims. However, existing systems struggle in coming up with the right ideas and sketches for frontier math problems. Typical savings around 5–10 hours on a 100–300 hour project (approximately 5–10%), i.e., roughly 500–2k of expert time (assuming approximately 100–200/hr), while keeping quality if humans fully verify all mathematical claims.

2. **Profession**: Pharmacists

   **Task**: Provide information and advice regarding drug interactions, side effects, dosage, and proper medication storage.

   **Risk Rating**: 5

   **Risk Types:** Misinformation, Privacy, Security, Malicious actors and Misuse, Discrimination

   **Risk Justification:** The task is safety critical as incorrect interaction, dose, side-effect, or storage advice can cause serious harm. LLMs raise risks mainly via confident clinical errors or omissions, and automation bias. Most errors are catchable if a pharmacist verifies using authoritative drug databases or labeling and uses the patient profile, but subtle mistakes and outdated guidance can slip through, causing substantial harm. Added concerns: PHI leakage if used outside compliant systems; prompt-injection or tool misuse in agentic setups; potential biased framing in counseling.

   **Reward Rating:** 3

   **Reward Justification:** Best rewards are on communication and workflow (not core clinical judgment): draft patient-friendly counseling, generate clarifying questions,

summarize monographs, translate counseling, draft messages to prescribers, while the pharmacist verifies key facts in trusted databases.

3. **Profession**: Human Resources Managers

   **Task**: Identify staff vacancies and recruit, interview, and select applicants.

   **Risk Rating**: 3

   **Risk Types:** Discrimination, Privacy, Security, Misinformation, Malicious actors and Misuse

   **Risk Justification:** It's a critical task, and hiring errors can cause serious individual harm and create legal or reputational risk. LLMs raise risk via biased screening or interview framing, mishandling applicant PII, incorrect compliance guidance, and misuse (AI-crafted resumes, deepfake interviews, or prompt-injected materials). Many issues are catchable with structured rubrics, audits, and keeping LLM output advisory, not decisive, but subtle bias or privacy failures can slip through.

   **Reward Rating:** 4

   **Reward Justification:** high reward when used for operations, not decisioning: draft job ads or outreach, summarize resumes into structured fields (verified), generate interview guides or rubrics, schedule or communicate, and draft documentation. Typical savings around 5–15 hours per hire (approximately 250–1,500 at 50–100/hr HR cost), while keeping quality if humans make final decisions using structured evidence and bias checks.

4. **Profession**: Claims Adjusters, Examiners, and Investigators

   **Task**: Review police reports, medical treatment records, medical bills, or physical property damage to determine the extent of liability.

   **Risk Rating**: 4

   **Risk Type:** Misinformation, Privacy, Discrimination, Security, Malicious actors and Misuse

   **Risk Justification:** The task is safety critical, wrong liability calls can seriously harm claimants or insurers and create legal or fraud exposure. LLMs raise risk by misreading or omitting evidence, hallucinating facts, making overconfident liability inferences, and introducing biased credibility judgments; sensitive medical or police data also heightens privacy or security risk, and fraudsters can tailor narratives to game the model. Humans can catch many issues if LLMs are limited to summarization or

extraction and every claim is traced back to the record, but subtle distortions can slip through.

**Reward Rating:** 4

**Reward Justification:** High reward for evidence organization: extract entities, codes, or line items, build timelines, flag inconsistencies or missing docs, draft notes or letters, and generate follow-up questions. Typical savings approximately 0.5–3 hours per complex claim (approximately 20–240 at 35–80/hr), improving cycle time while maintaining quality if adjusters verify key facts and keep final liability judgment human-driven.

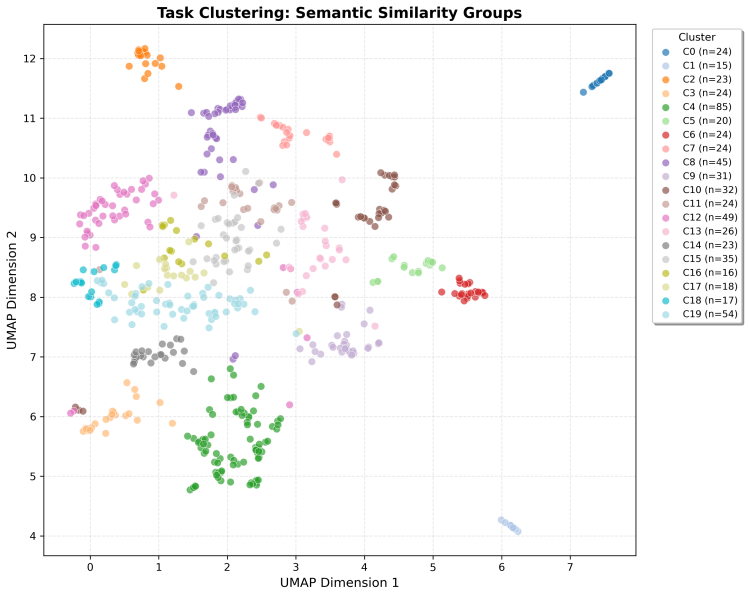


Figure 10: Clustering of 609 computer-occupation tasks into 20 groups based on semantic similarity.

# C Task clustering

Aggregating results from 6 models, tasks are classified into 9 combined categories of risk and reward based on mean ratings: Low ($\leq 2$), Medium ($> 2$ and $< 4$), and High ($\geq 4$). Within each risk-reward group, we cluster all tasks by semantic similarity in task activity description before manually analyzing any patterns therein. For clustering, each task description is encoded into a 384-dimensional embedding vector with the all-MiniLM-L6-v2 sentence transformer model [58], reduced to 5 dimensions with UMAP [43] ($n_{neighbors} = 15$, $min_{dist} = 0$, $metric = cosine$), then clustered with Hierarchical Density-Based Spatial

Table 5: ID and description for 20 semantically grouped task clusters.

| Cluster | Label |
|---|---|
| C0 | Analyzing and visualizing geospatial data through specialized GIS applications |
| C1 | Health information technology solutions |
| C2 | Designing interactive game experiences through iterative prototyping and user feedback |
| C3 | Maintaining and updating web infrastructure and server configurations |
| C4 | Testing security vulnerabilities and implementing cybersecurity measures |
| C5 | Integrating and optimizing data warehouse systems and processes |
| C6 | Architecting and optimizing databases for performance and scalability |
| C7 | Training users and coordinating technical project teams and activities |
| C8 | Managing IT project resources, schedules, and stakeholder communications |
| C9 | Creating and optimizing document management systems |
| C10 | Designing and developing user-focused web interfaces and applications |
| C11 | Analyzing and designing computer systems architecture and solutions |
| C12 | Evaluating and testing software quality through structured methodologies |
| C13 | Developing and documenting system design procedures and data workflows |
| C14 | Administering network systems and maintaining logs and automation |
| C15 | Analyzing and designing network systems to meet organizational requirements |
| C16 | Evaluating and testing system performance and new technology solutions |
| C17 | Installing and supporting telecommunications equipment and systems |
| C18 | Troubleshooting and resolving user network and software issues |
| C19 | Installing, configuring, and troubleshooting network hardware and software |

Clustering of Applications with Noise (HDBSCAN) [10] ($min_{cluster\_size} = 10$, $min_{samples} = 3$). Finally, we use GPT-4o-mini [52] to generate concise, interpretable labels to describe tasks in each cluster. Human annotators manually review and edit all clustering results and topic labels. This process yields 20 distinct task clusters with 5% of tasks classified as outliers. The outliers are reassigned to clusters based on nearest neighbors. Table 5 and Figure 10 present final cluster descriptions, with cluster distributions for each risk-reward group in Table 6.

# D Task distribution of Computer Occupations in O*NET

Table 7 counts the tasks in each O*NET Computer Occupation.

# E Inter-model variability in occupation-level risk and reward scores

We compute model-specific occupation-level risk and reward scores by aggregating task-level ratings $R_m(t)$ and $G_m(t)$ using (2). Table 8 reports the average occupation-level scores $R(o)$

Table 6: Cluster distribution by risk-reward category. Percentage denotes proportion of tasks in given cluster on given risk-reward level.

| **Risk level** | **Reward level** | **N** | **Task distribution** |
|---|---|---|---|
| Low | Low | 1 | C19 (100.0%) |
| | Medium | 34 | C2 (20.6%), C10 (17.6%), C8 (8.8%), C3 (8.8%), C19 (5.9%), C13 (5.9%), C0 (5.9%), C15 (5.9%), C12 (5.9%), C14 (2.9%), C11 (2.9%), C17 (2.9%), C7 (2.9%), C9 (2.9%) |
| | High | 39 | C2 (30.8%), C9 (17.9%), C10 (17.9%), C13 (10.3%), C8 (7.7%), C12 (5.1%), C14 (2.6%), C6 (2.6%), C0 (2.6%), C3 (2.6%) |
| Medium | Low | 4 | C4 (75.0%), C14 (25.0%) |
| | Medium | 224 | C19 (14.3%), C8 (9.4%), C4 (7.6%), C0 (6.7%), C15 (6.7%), C11 (6.2%), C12 (6.2%), C14 (5.8%), C17 (5.8%), C6 (4.9%), C7 (4.5%), C16 (3.6%), C9 (3.6%), C3 (3.1%), C10 (2.2%), C1 (2.2%), C18 (2.2%), C13 (2.2%), C5 (1.8%), C2 (0.9%) |
| | High | 233 | C12 (13.3%), C15 (7.7%), C19 (7.7%), C8 (7.3%), C5 (6.9%), C13 (6.4%), C9 (6.0%), C10 (6.0%), C7 (5.6%), C6 (5.2%), C3 (4.7%), C18 (4.3%), C4 (3.9%), C11 (3.4%), C16 (3.0%), C0 (2.6%), C1 (2.1%), C17 (1.7%), C14 (1.3%), C2 (0.9%) |
| High | Low | 3 | C4 (100.0%) |
| | Medium | 70 | C4 (74.3%), C1 (7.1%), C14 (5.7%), C18 (2.9%), C3 (2.9%), C19 (1.4%), C11 (1.4%), C16 (1.4%), C9 (1.4%), C8 (1.4%) |
| | High | 1 | C4 (100.0%) |

Table 7: Number of task statements for each of the O*NET 30.0 Computer Occupations.

| Occupation | # Tasks |
|---|---|
| Web Administrators | 35 |
| Computer Network Architects | 33 |
| Web and Digital Interface Designers | 30 |
| Software Quality Assurance Analysts and Testers | 30 |
| Web Developers | 29 |
| Geographic Information Systems Technologists and Technicians | 29 |
| Computer Systems Engineers/Architects | 28 |
| Computer Network Support Specialists | 26 |
| Telecommunications Engineering Specialists | 26 |
| Database Architects | 25 |
| Video Game Designers | 24 |
| Document Management Specialists | 23 |
| Computer Systems Analysts | 22 |
| Penetration Testers | 22 |
| Information Technology Project Managers | 21 |
| Information Security Engineers | 20 |
| Digital Forensics Analysts | 20 |
| Network and Computer Systems Administrators | 20 |
| Data Warehousing Specialists | 18 |
| Database Administrators | 18 |
| Computer Programmers | 17 |
| Blockchain Engineers | 17 |
| Software Developers | 17 |
| Health Informatics Specialists | 17 |
| Computer User Support Specialists | 16 |
| Computer and Information Research Scientists | 15 |
| Information Security Analysts | 11 |

and $G(o)$, along with their standard deviations across models. After accounting for standard deviation, for the four highest-risk, lowest-reward safety-critical occupations in Figure 2 (*Information Security Engineers, Information Security Analysts, Penetration Testers, Digital Forensic Analysts*), we find that $G(o)$ + 1 std of rewards $< R(o)$ in each case. This shows the "risk outweighs reward" signal is robust to model-to-model variability. For 20 of the tasks where $G(o) > R(o)$, we have that $G(o)$ - 1 std of rewards $> R(o)$, showing that the "reward outweighs risk" signal is likewise robust to model-to-model variability.

Table 8: Occupation-level reward and risk statistics across Computer occupations.

| **Occupation** | $G(o)$ | **Reward Std** | $R(o)$ | **Risk Std** |
|---|---|---|---|---|
| Information Security Engineers | 3.46 | 0.36 | 4.16 | 0.51 |
| Information Security Analysts | 3.24 | 0.47 | 4.05 | 0.60 |
| Penetration Testers | 3.56 | 0.27 | 3.85 | 0.61 |
| Digital Forensic Analysts | 2.91 | 0.39 | 3.85 | 0.73 |
| Blockchain Engineers | 3.45 | 0.37 | 3.61 | 0.68 |
| Health Informatics Specialists | 3.71 | 0.24 | 3.58 | 0.49 |
| Computer Systems Engineers/Architects | 3.87 | 0.19 | 3.38 | 0.61 |
| Network and Computer Systems Administrators | 3.61 | 0.32 | 3.34 | 0.78 |
| Database Administrators | 3.73 | 0.16 | 3.22 | 0.58 |
| Computer Network Architects | 3.76 | 0.22 | 3.10 | 0.70 |
| Computer Systems Analysts | 3.94 | 0.12 | 3.06 | 0.68 |
| Computer Network Support Specialists | 3.61 | 0.28 | 3.01 | 0.74 |
| Software Developers | 3.89 | 0.28 | 2.99 | 0.70 |
| Web Administrators | 3.81 | 0.13 | 2.88 | 0.57 |
| Computer and Information Research Scientists | 3.79 | 0.39 | 2.87 | 0.75 |
| Telecommunications Engineering Specialists | 3.64 | 0.24 | 2.87 | 0.72 |
| Computer Programmers | 4.13 | 0.34 | 3.03 | 0.56 |
| Data Warehousing Specialists | 4.00 | 0.09 | 2.84 | 0.73 |
| Software Quality Assurance Analysts and Testers | 3.94 | 0.09 | 2.80 | 0.52 |
| Database Architects | 3.95 | 0.09 | 2.77 | 0.65 |
| Document Management Specialists | 3.74 | 0.15 | 2.69 | 0.68 |
| Information Technology Project Managers | 3.93 | 0.13 | 2.66 | 0.49 |
| Computer User Support Specialists | 3.81 | 0.20 | 2.56 | 0.74 |
| Web Developers | 3.94 | 0.22 | 2.55 | 0.74 |
| Geographic Information Systems Technologists and Technicians | 3.70 | 0.32 | 2.54 | 0.82 |
| Web and Digital Interface Designers | 3.95 | 0.14 | 2.29 | 0.64 |
| Video Game Designers | 3.91 | 0.16 | 1.50 | 0.72 |

# F Task analysis of minor risk-reward groups

**High risk, high reward.** One high-risk task, Task 21756 - *Gather cyber intelligence to identify vulnerabilities* performed by penetration testers, is classified as high-reward, possibly due to LLMs' efficiency in processing and summarizing large Open Source Intelligence (OSINT) datasets.

**High risk, low reward.** Three tasks are classified as low-risk, low-reward: Task 21819 - *Implementing catastrophic failure handlers to identify security breaches*, Task 21812 - *Design and verify cryptographic protocols to protect private information*, and Task 5317 - *Modify computer security files to incorporate new software*. These tasks involve safety-critical implementation and deep security expertise, where gains from LLM assistance with documentation are likely outweighed by the add costs for verification and error correction. Most common risk types in the high-risk are Security, Malicious Actors and Misuse, Misinformation, and Privacy – all present in 94-100% of tasks.

**Medium risk, low reward.** Four tasks (Tasks 21799, 21792, 21801, 21789) are classified as medium-risk, low-reward. These involve the preserving, duplicating, imaging, and recovering digital forensic data with limited opportunity for LLM automation as they require specialized forensic tools and direct access to physical or digital evidence.

**Low risk, low reward.** Only one task, Task 1327 - *Load computer tapes and disks, and install software and printer paper or forms* is classified as low-risk, low-reward as a physical task not amenable to LLM automation.

# G Cluster visualization by risk-reward category

Figures 11, 12, 13, 14, 15, 16, 17, 18, and 19 visualize the clusters for each risk-reward group.

# H Visualizing central role of Security risks

Figure 20 visualizes the central role of Security risks in contributing to task-level risk jaggedness through frequent n-grams in risk justifications for tasks.

# I Evidence from Anthropic Economic Index

Table 9 shows a list of O*NET tasks ordered (highest first) by how frequently they appear in Human-AI conversations as reported in Anthropic Economic Index [3].


Figure 11: High-risk, medium-reward clusters.

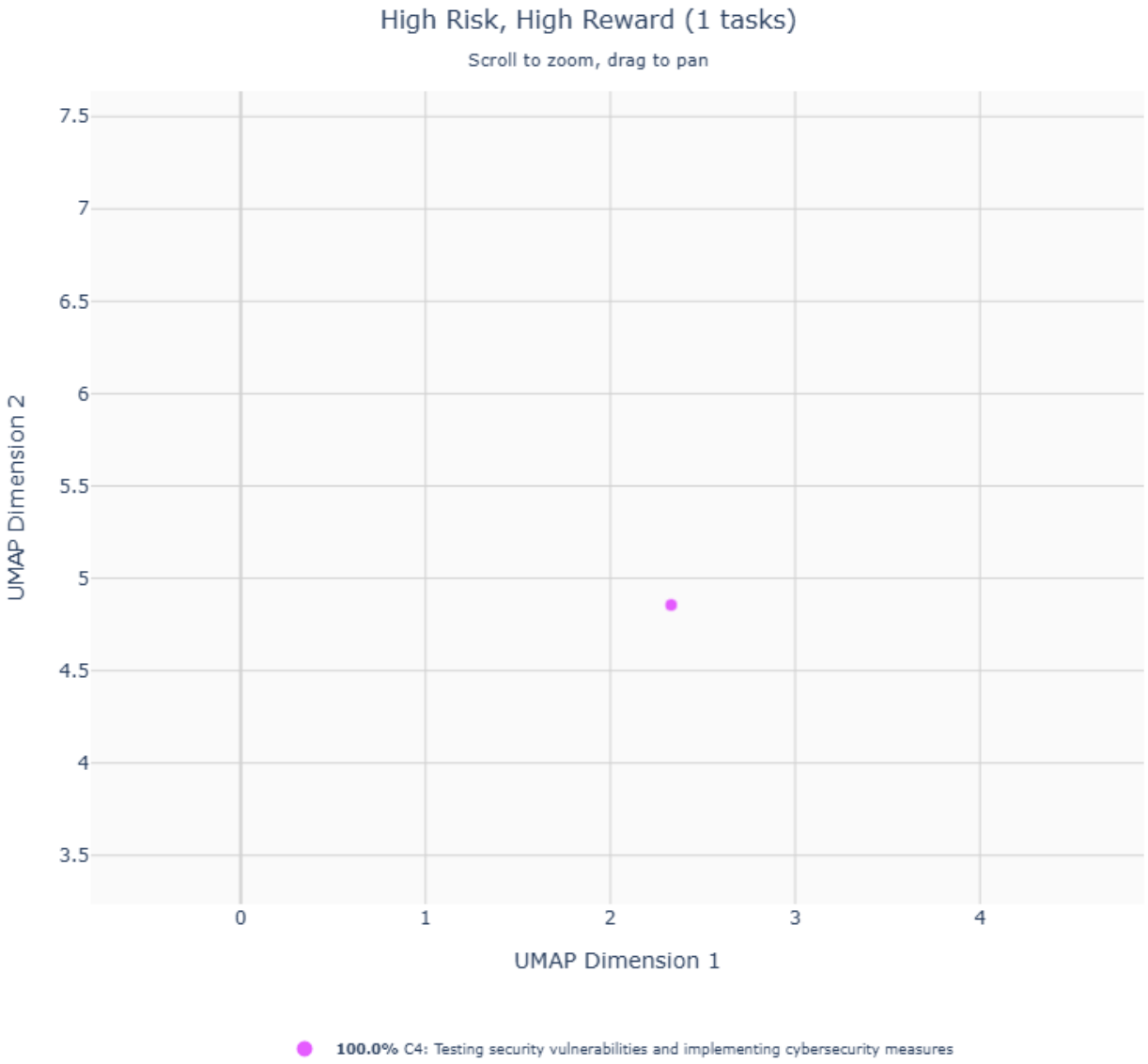


Figure 12: High-risk, high-reward clusters.

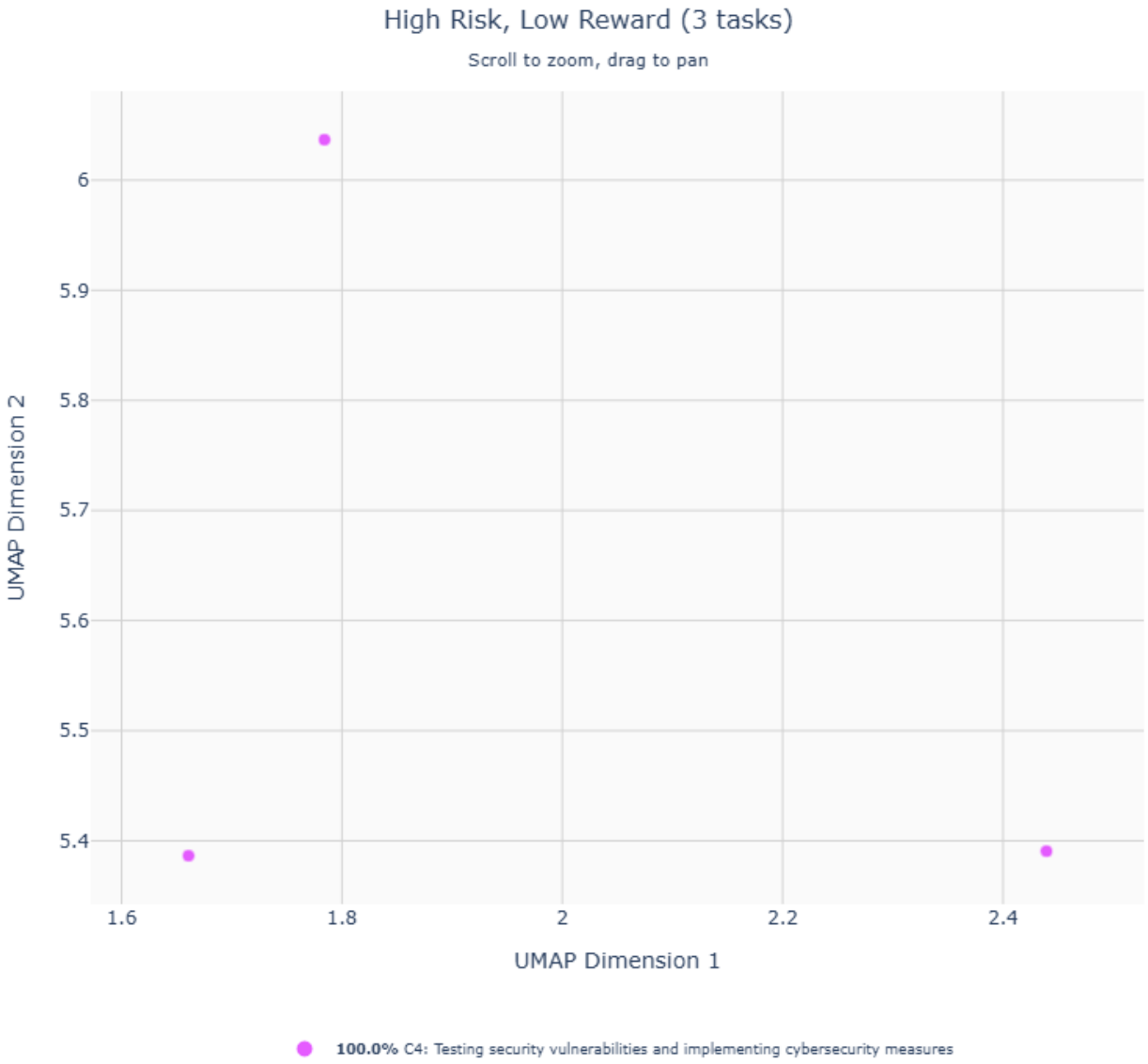


Figure 13: High-risk, low-reward clusters.

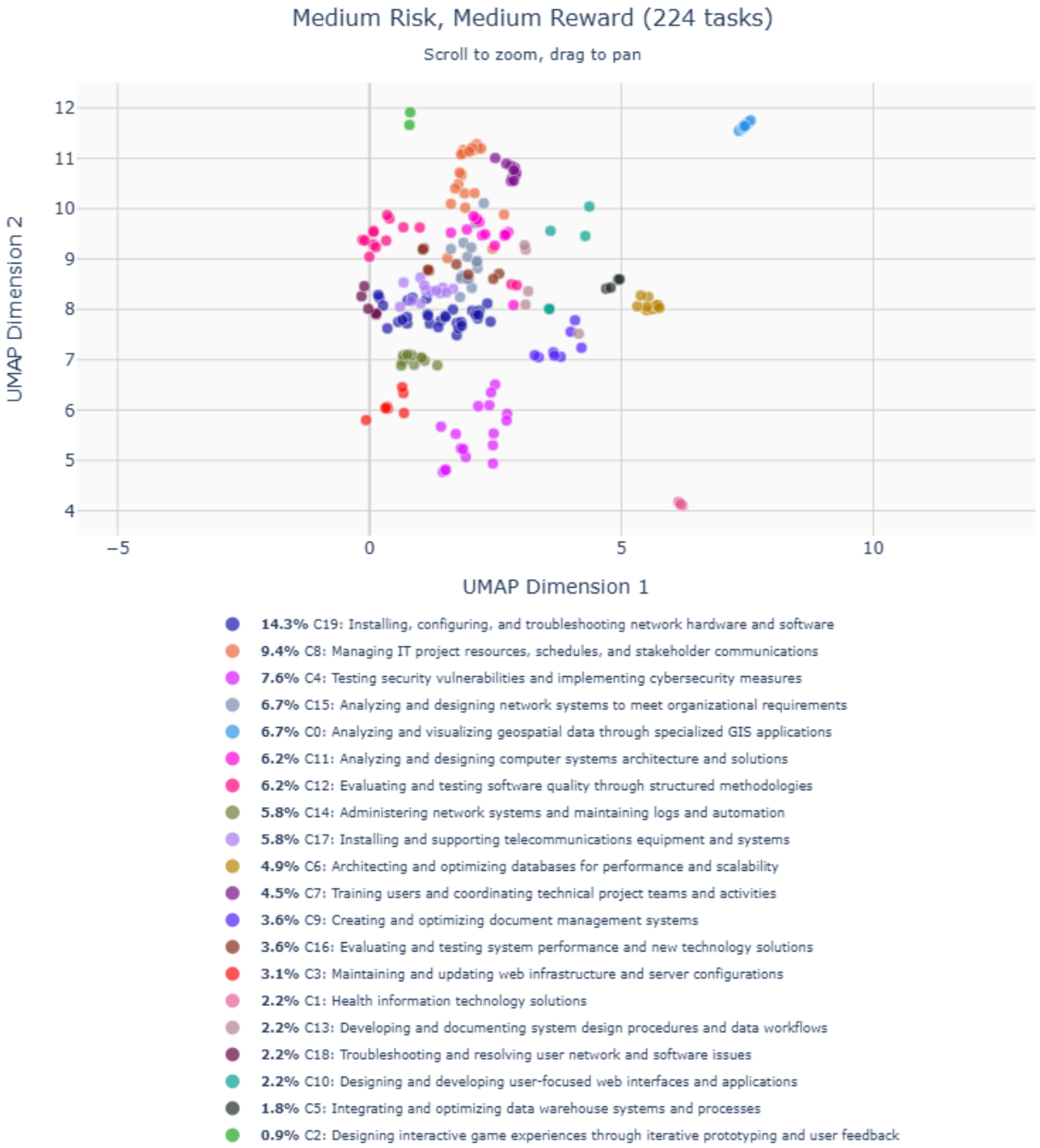


Figure 14: Medium-risk, medium-reward clusters.


Figure 15: Medium-risk, high-reward clusters.

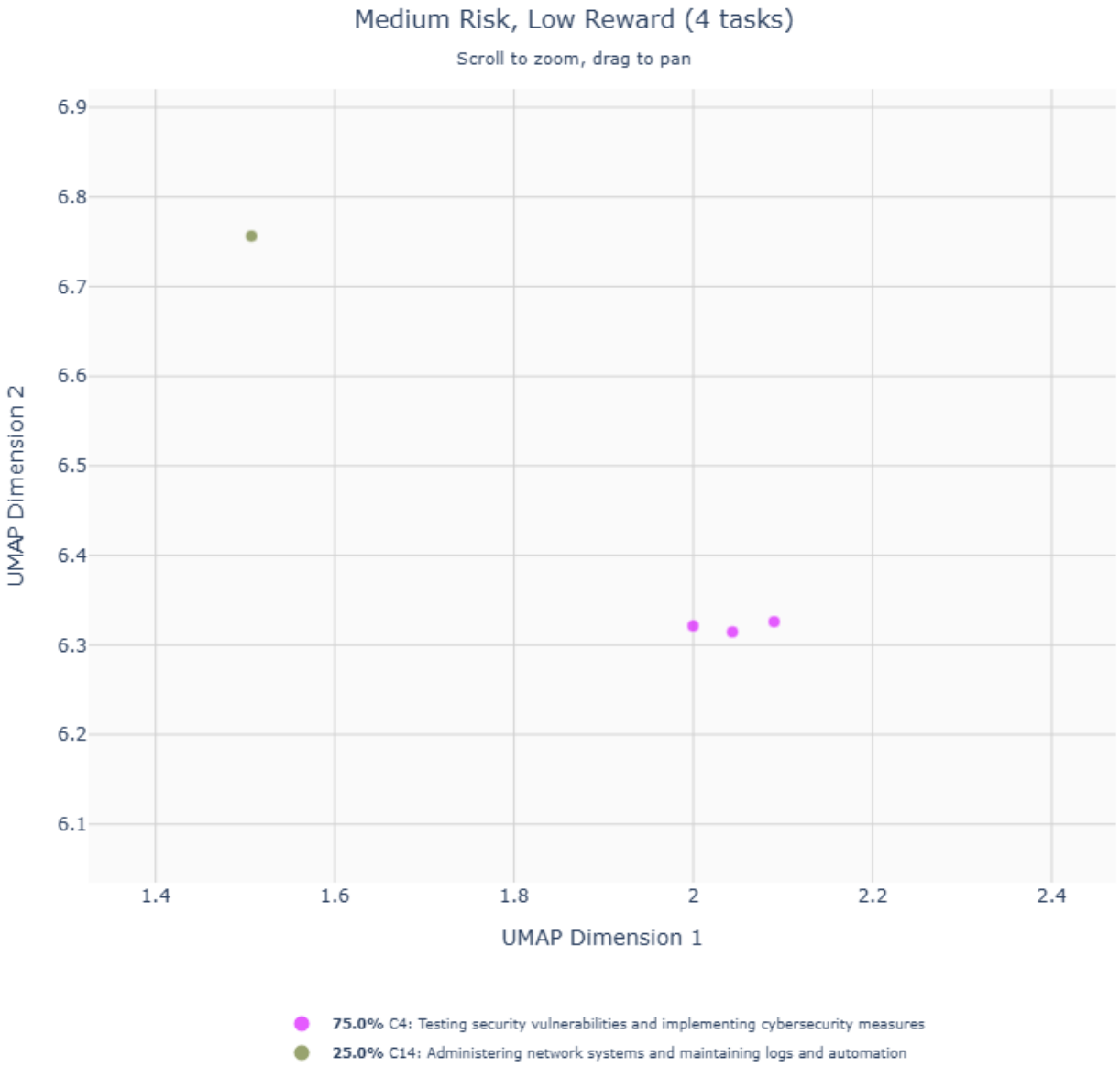


Figure 16: Medium-risk, low-reward clusters.


Figure 17: Low-risk, medium-reward clusters.


Figure 18: Low-risk, high-reward clusters.

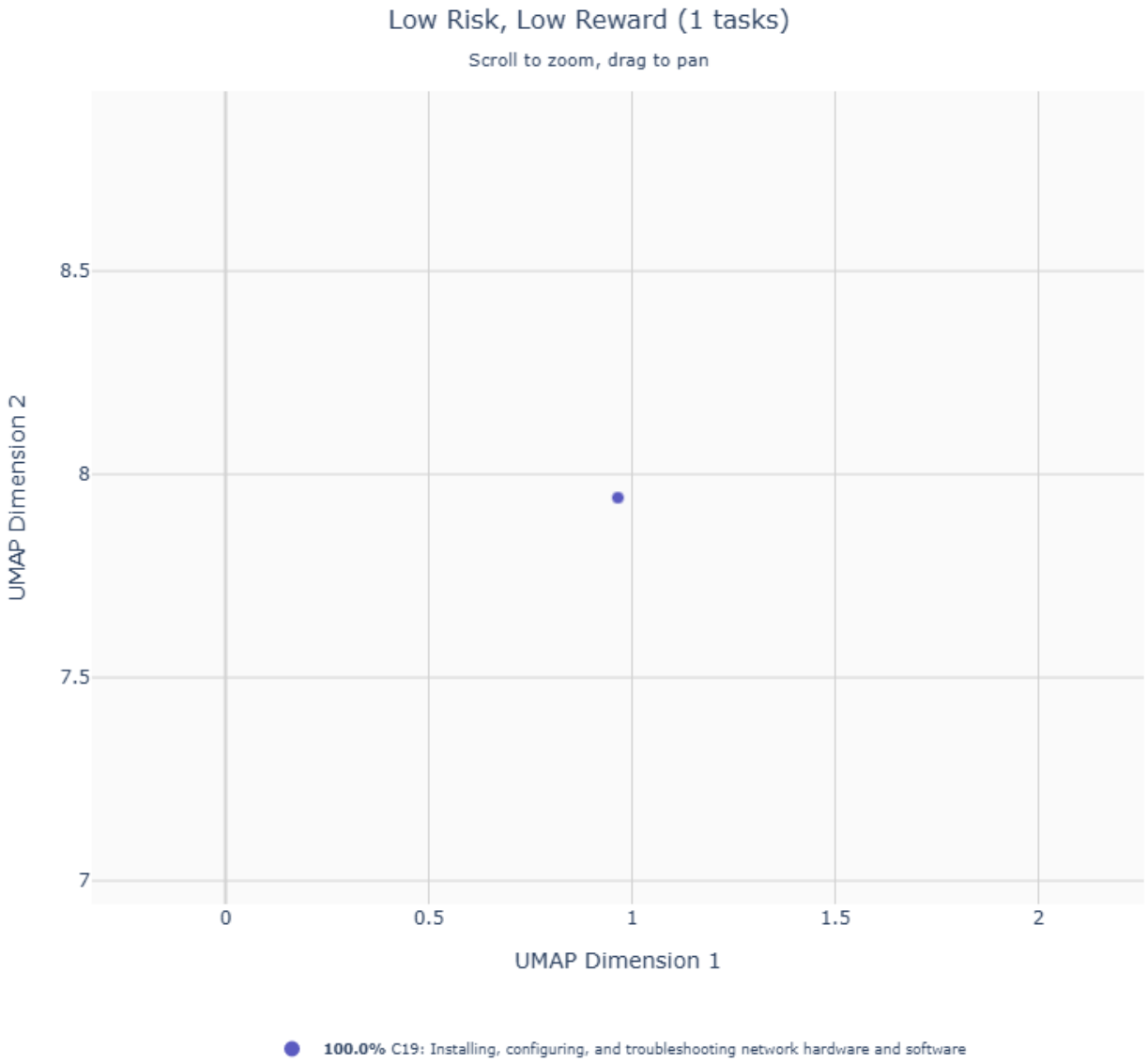


Figure 19: Low-risk, low-reward clusters.


Figure 20: Word clouds of TF-IDF weighted n-grams (1-3) in risk justifications for tasks where majority of models agree on the presence of Security risks, highlighting Security as driving task-level risk jaggedness: Low-risk tasks display secondary, incompetence-driven security concerns (privacy vulnerabilities, prompt injection, outdated information, sensitive data leakage) arising from errors and nature of LLM interaction. Medium-risk tasks are a balance of incompetence-driven and adversarial concerns. High-risk tasks are dominated by adversarial threats (malicious actors, exploit attempts, breaches, attacks) involving safety-critical systems and deliberate exploitation.

# J Risk Reward Scores for Healthcare and Legal Occupations

Table 10 shows substantial variation in both risk (range: $[1.90, 4.63]$) and reward (range: $[0.95, 3.77]$) across healthcare and legal occupations. In general, many physicians and advanced clinical roles, such as *Emergency Medicine Physicians*, *Nurse Anesthetists*, *Psychiatrists*, and *Obstetricians and Gynecologists*, have higher risk scores compared to rewards, suggesting that deployment in these settings remains constrained by safety-critical and liability-sensitive workflows. In contrast, administrative and documentation-oriented occupations, including *Medical Transcriptionists*, *Paralegals and Legal Assistants*, *Patient Representatives*, and *Title Examiners*, achieve relatively high reward scores with lower or moderate risk levels, indicating relatively favorable conditions for LLM-assisted productivity gains. Overall, the results support the broader observation that the risk-reward frontier for expert–LLM collaboration is risk-constrained across these occupations as well.

Table 10: Occupation-level reward and risk statistics for Healthcare and Legal occupations

| **Occupation** | $G(o)$ | **Reward Std** | $R(o)$ | **Risk Std** |
|---|---|---|---|---|
| Acupuncturists | 2.25 | 0.46 | 3.74 | 0.66 |
| Acute Care Nurses | 2.25 | 0.42 | 4.24 | 0.45 |
| Administrative Law Judges, Adjudicators, and Hearing Officers | 2.92 | 0.49 | 3.93 | 0.31 |
| Advanced Practice Psychiatric Nurses | 2.44 | 0.57 | 4.31 | 0.41 |
| Allergists and Immunologists | 2.45 | 0.54 | 4.10 | 0.54 |
| Anesthesiologist Assistants | 1.37 | 0.55 | 4.28 | 0.64 |
| Anesthesiologists | 1.64 | 0.57 | 4.50 | 0.56 |
| Arbitrators, Mediators, and Conciliators | 2.96 | 0.51 | 3.60 | 0.31 |
| Art Therapists | 2.69 | 0.47 | 3.14 | 0.59 |
| Athletic Trainers | 2.51 | 0.50 | 3.05 | 0.77 |
| Audiologists | 2.96 | 0.27 | 3.44 | 0.64 |
| Cardiovascular Technologists and Technicians | 1.89 | 0.44 | 3.58 | 0.99 |
| Chiropractors | 2.41 | 0.52 | 3.92 | 0.58 |
| Clinical Nurse Specialists | 3.09 | 0.24 | 3.93 | 0.64 |
| Critical Care Nurses | 2.01 | 0.57 | 4.38 | 0.51 |
| Cytogenetic Technologists | 2.22 | 0.66 | 3.20 | 1.05 |
| Cytotechnologists | 2.24 | 0.55 | 3.50 | 0.90 |

| **Occupation** | $G(o)$ | **Reward Std** | $R(o)$ | **Risk Std** |
|---|---|---|---|---|
| Dental Assistants | 2.01 | 0.36 | 2.79 | 1.00 |
| Dental Hygienists | 1.81 | 0.52 | 2.96 | 0.91 |
| Dentists, General | 1.86 | 0.46 | 3.54 | 0.97 |
| Dermatologists | 2.47 | 0.52 | 4.03 | 0.49 |
| Diagnostic Medical Sonographers | 2.08 | 0.55 | 3.36 | 0.84 |
| Dietetic Technicians | 3.31 | 0.22 | 2.88 | 0.73 |
| Dietitians and Nutritionists | 3.33 | 0.19 | 3.54 | 0.53 |
| Emergency Medicine Physicians | 2.18 | 0.43 | 4.59 | 0.41 |
| Endoscopy Technicians | 1.90 | 0.55 | 3.21 | 1.15 |
| Exercise Physiologists | 2.82 | 0.35 | 3.27 | 0.56 |
| Family Medicine Physicians | 3.00 | 0.46 | 4.17 | 0.58 |
| General Internal Medicine Physicians | 2.82 | 0.41 | 4.38 | 0.34 |
| Genetic Counselors | 3.04 | 0.34 | 4.01 | 0.48 |
| Hearing Aid Specialists | 2.46 | 0.58 | 3.20 | 0.72 |
| Histology Technicians | 1.27 | 0.88 | 1.90 | 1.48 |
| Histotechnologists | 2.13 | 0.68 | 2.59 | 1.28 |
| Home Health Aides | 1.91 | 0.55 | 3.14 | 0.94 |
| Hospitalists | 2.92 | 0.32 | 4.33 | 0.42 |
| Judges, Magistrate Judges, and Magistrates | 1.81 | 0.77 | 4.63 | 0.24 |
| Judicial Law Clerks | 3.37 | 0.31 | 3.61 | 0.55 |
| Lawyers | 3.36 | 0.41 | 3.97 | 0.47 |
| Licensed Practical and Licensed Vocational Nurses | 1.98 | 0.42 | 3.68 | 0.96 |
| Low Vision Therapists, Orientation and Mobility Specialists, and Vision Rehabilitation Therapists | 2.74 | 0.59 | 3.38 | 0.62 |
| Magnetic Resonance Imaging Technologists | 1.97 | 0.48 | 3.23 | 0.93 |
| Massage Therapists | 2.00 | 0.58 | 2.46 | 1.17 |
| Medical Assistants | 2.44 | 0.46 | 3.15 | 0.88 |
| Medical Dosimetrists | 1.70 | 0.80 | 4.59 | 0.25 |
| Medical Equipment Preparers | 2.09 | 0.61 | 2.80 | 1.11 |
| Medical Transcriptionists | 3.76 | 0.21 | 3.39 | 0.60 |
| Medical and Clinical Laboratory Technicians | 2.00 | 0.71 | 3.80 | 0.97 |
| Medical and Clinical Laboratory Technologists | 2.21 | 0.65 | 3.68 | 0.77 |

| **Occupation** | $G(o)$ | **Reward Std** | $R(o)$ | **Risk Std** |
|---|---|---|---|---|
| Midwives | 2.27 | 0.41 | 4.09 | 0.48 |
| Music Therapists | 2.88 | 0.48 | 2.64 | 0.69 |
| Naturopathic Physicians | 2.48 | 0.52 | 4.03 | 0.48 |
| Neurodiagnostic Technologists | 2.25 | 0.57 | 3.43 | 0.70 |
| Neurologists | 2.72 | 0.33 | 4.23 | 0.43 |
| Nuclear Medicine Technologists | 1.70 | 0.77 | 4.38 | 0.50 |
| Nurse Anesthetists | 1.48 | 0.66 | 4.59 | 0.39 |
| Nurse Midwives | 2.59 | 0.45 | 4.22 | 0.56 |
| Nurse Practitioners | 2.87 | 0.30 | 4.28 | 0.37 |
| Nursing Assistants | 1.51 | 0.55 | 2.92 | 1.14 |
| Obstetricians and Gynecologists | 2.60 | 0.56 | 4.50 | 0.36 |
| Occupational Therapists | 3.18 | 0.26 | 3.39 | 0.53 |
| Occupational Therapy Aides | 2.60 | 0.41 | 2.73 | 1.05 |
| Occupational Therapy Assistants | 2.62 | 0.46 | 3.00 | 0.66 |
| Ophthalmic Medical Technicians | 1.70 | 0.60 | 3.11 | 1.22 |
| Ophthalmic Medical Technologists | 1.93 | 0.58 | 3.42 | 0.90 |
| Ophthalmologists, Except Pediatric | 2.64 | 0.30 | 4.16 | 0.60 |
| Opticians, Dispensing | 2.46 | 0.44 | 2.08 | 0.79 |
| Optometrists | 2.32 | 0.54 | 4.16 | 0.37 |
| Oral and Maxillofacial Surgeons | 1.53 | 0.73 | 4.27 | 1.06 |
| Orderlies | 1.21 | 0.78 | 2.18 | 1.24 |
| Orthodontists | 2.51 | 0.57 | 3.53 | 0.78 |
| Orthoptists | 2.50 | 0.44 | 3.86 | 0.36 |
| Orthotists and Prosthetists | 2.61 | 0.65 | 3.39 | 0.54 |
| Paralegals and Legal Assistants | 3.77 | 0.18 | 3.41 | 0.71 |
| Patient Representatives | 3.62 | 0.20 | 3.21 | 0.50 |
| Pediatricians, General | 2.68 | 0.45 | 4.26 | 0.46 |
| Personal Care Aides | 2.52 | 0.36 | 3.35 | 0.76 |
| Pharmacists | 2.79 | 0.30 | 4.33 | 0.32 |
| Pharmacy Aides | 2.56 | 0.48 | 2.26 | 0.95 |
| Pharmacy Technicians | 2.56 | 0.32 | 3.21 | 0.80 |
| Phlebotomists | 1.45 | 0.65 | 2.92 | 1.39 |
| Physical Medicine and Rehabilitation Physicians | 2.86 | 0.26 | 4.07 | 0.62 |

| **Occupation** | $G(o)$ | **Reward Std** | $R(o)$ | **Risk Std** |
|---|---|---|---|---|
| Physical Therapist Aides | 1.85 | 0.49 | 2.39 | 1.02 |
| Physical Therapist Assistants | 2.18 | 0.37 | 3.09 | 0.95 |
| Physical Therapists | 2.78 | 0.33 | 3.59 | 0.37 |
| Physician Assistants | 2.60 | 0.48 | 4.38 | 0.40 |
| Physicians, Pathologists | 2.71 | 0.51 | 4.21 | 0.39 |
| Podiatrists | 2.32 | 0.52 | 4.05 | 0.54 |
| Preventive Medicine Physicians | 3.53 | 0.15 | 3.88 | 0.65 |
| Prosthodontists | 2.09 | 0.78 | 3.54 | 1.04 |
| Psychiatric Aides | 1.74 | 0.54 | 3.42 | 0.81 |
| Psychiatric Technicians | 1.80 | 0.57 | 4.07 | 0.47 |
| Psychiatrists | 2.60 | 0.60 | 4.53 | 0.33 |
| Radiation Therapists | 1.71 | 0.60 | 4.27 | 0.66 |
| Radiologic Technologists and Technicians | 1.93 | 0.57 | 3.49 | 0.87 |
| Radiologists | 2.97 | 0.45 | 3.91 | 0.52 |
| Recreational Therapists | 3.13 | 0.31 | 3.29 | 0.41 |
| Registered Nurses | 2.61 | 0.39 | 4.10 | 0.55 |
| Respiratory Therapists | 2.13 | 0.54 | 4.21 | 0.58 |
| Speech-Language Pathologists | 3.25 | 0.23 | 3.37 | 0.56 |
| Speech-Language Pathology Assistants | 3.54 | 0.17 | 2.70 | 0.71 |
| Sports Medicine Physicians | 2.81 | 0.26 | 4.13 | 0.52 |
| Surgical Assistants | 0.95 | 0.53 | 3.80 | 1.45 |
| Surgical Technologists | 1.20 | 0.56 | 3.83 | 1.22 |
| Title Examiners, Abstractors, and Searchers | 3.73 | 0.25 | 3.30 | 0.52 |
| Urologists | 2.47 | 0.43 | 4.43 | 0.52 |
| Veterinarians | 2.58 | 0.48 | 3.66 | 0.79 |
| Veterinary Assistants and Laboratory Animal Caretakers | 2.12 | 0.56 | 2.69 | 0.75 |
| Veterinary Technologists and Technicians | 1.96 | 0.54 | 2.91 | 0.91 |

Table 9: High-frequency O*NET tasks from Computer occupations identified in Claude Human–AI conversations.

| # | Task Description | Task ID | Occupation | Risk-reward group |
|---|---|---|---|---|
| 1 | Modify existing software to correct errors, adapt to new hardware, or improve performance | 21680 | Software Quality Assurance Analysts and Testers | high-reward, medium-risk |
| 2 | Write new programs or modify existing programs to meet customer requirements using current programming languages and technologies | 16120 | Data Warehousing Specialists | high-reward, medium-risk |
| 3 | Design, build, or maintain websites using authoring or scripting languages, content creation tools, management tools, and digital media | 14694 | Web Developers | high-reward, medium-risk |
| 4 | Diagnose, troubleshoot, and resolve hardware, software, or network/system problems and replace defective components | 15205 | Network and Computer Systems Administrators | high-reward, medium-risk |
| 5 | Modify existing software to correct errors, adapt to new hardware, or upgrade interfaces and improve performance | 21670 | Software Developers | high-reward, medium-risk |
| 6 | Write, update, and maintain programs or software packages for inventory tracking, data storage/retrieval, or equipment control | 1270 | Computer Programmers | high-reward, medium-risk |
| 7 | Confer with clients regarding information-processing or computation needs a program is to address | 3467 | Computer Systems Analysts | medium-reward, medium-risk |
| 8 | Compile and write documentation of program development and revisions, including comments in code | 1269 | Computer Programmers | high-reward, low-risk |
| 9 | Write supporting code for web applications or websites | 14707 | Web Developers | high-reward, medium-risk |
| 10 | Correct errors by making appropriate changes and rechecking programs to ensure desired results | 1267 | Computer Programmers | high-reward, medium-risk |