Economic Evaluations of Language Models

Alexander Wan
Stanford University

Stephane Hatgis-Kessell*
Stanford University

Tomás Aguirre*
University of São Paulo

Percy Liang
Stanford University

Rishi Bommasani
Stanford University

Abstract

Language models perform economically valuable work. Yet language models are not currently assessed for how well they perform every economically valuable task. We introduce ECONEVALS as an open-source evaluation suite to measure capabilities relevant to tasks, work activities, and occupations in the US labor economy. We ground the evaluation suite in real user queries to language models where possible, and supplement these with synthetic data. Our evaluations improve coverage over OpenAI's GDPval benchmark, which is the existing state-of-the-art that covers 5% of US occupations, at $500\times$ lower cost. Alongside benchmarks, we also introduce a *simulation-based exposure measure* to estimate how much time current language model capabilities could save across all tasks belonging to all US occupations, with detailed accounting for each estimate. Our estimates indicate that current models could save workers substantial time on at least half of their tasks in 47% of occupations. However, for 79% of tasks where we predict substantial time savings, observed Claude usage is low, suggesting that existing usage lags potential. Beyond inherent constraints of language model chatbots, our data identifies verification and privacy as the principal bottlenecks limiting further time savings from AI. Overall, we introduce adaptable infrastructure that grounds inferences about language models' labor-market impact in their current capabilities, which can be continually updated as capabilities improve.

1 Introduction

The economic impact of frontier AI is highly uncertain. The public identifies job disruption as a top priority in relation to AI [Ipsos, 2024, United Nations, 2025, Stanford HAI, 2025]. Leaders across industry, computer science, economics, and government disagree in their economic forecasts [The White House, 2026, Murphy et al., 2025, Karger et al., 2026]. For example, Nobel Laureate economist Daron Acemoglu predicts a modest cumulative GDP increase of roughly 0.9–1.6% over the next decade from AI [Acemoglu, 2024], while Anthropic CEO Dario Amodei predicts that AI could enable 10–20% sustained annual GDP growth [Amodei, 2026]. Existing evidence is limited [see Chandar, 2025] and facially provides contradictory accounts: several micro studies show that frontier AI meaningfully increases productivity, a few micro studies show it does not, and overall macro productivity evidence is fairly muted [see Imas, 2026, del Rio-Chanona et al., 2025].

The uncertainty about the economic impacts of frontier AI conflicts with the near certainty of their improving capabilities. Language models (LMs) saturate challenging benchmarks, often very quickly [Bengio et al., 2026, Akhtar et al., 2026, Sajadieh et al., 2026]. This rapid progress has prompted the development of new evaluations that are indexed to ever-increasing human task complexity [Kwa et al., 2025, METR, 2026] or real-world utility [Patwardhan et al., 2025, Mazeika et al., 2025, Vidgen et al., 2025]. We posit that a central reason for unclear economic impact despite clear model capabilities is the misalignment between AI benchmarks and economic tasks. Wang et al. [2026]

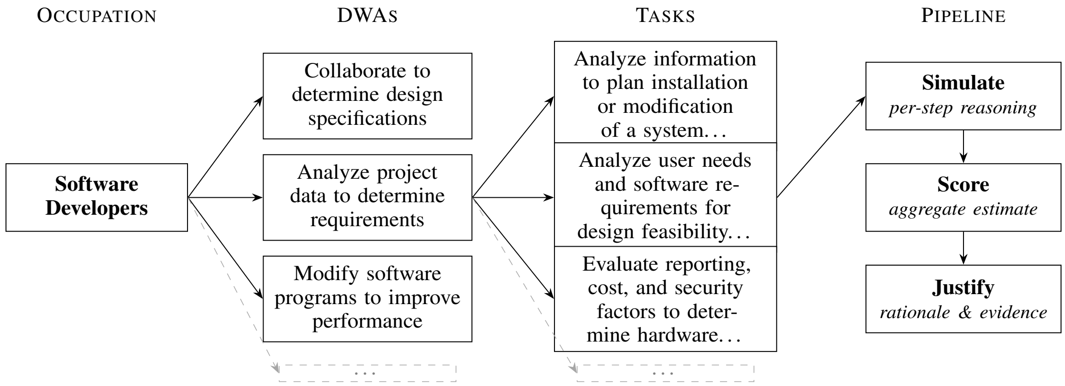


Figure 1: **O*NET work taxonomy and whitebox simulation-based task-level exposure estimates.** O*NET links occupations to DWAs and tasks as depicted for *Software Developers* (15-1252.00). Our method estimates task-level exposure for every task using simulations and provides justifications.

quantifies this distribution shift: existing benchmarks are concentrated on topics like math and coding (39.7% of benchmarking effort), even though the related occupations are a small share of the labor economy (3.5% of U.S. jobs).[1]

We introduce the ECONEVALS infrastructure to measure language model capabilities for work performed in the U.S. labor economy.[2] To codify work, we adopt a top-down approach using the Department of Labor's O*NET taxonomy. O*NET maps the U.S. labor economy from 1,016 occupations to 2,087 detailed work activities (DWAs) and 18,796 tasks. For example, software engineers (SOC code 15-1252.00) perform 17 tasks that include (i) analyzing user needs and software requirements to assess design feasibility, (ii) developing or directing software testing, validation, programming, or documentation, and (iii) modifying existing software to fix errors, adapt to new hardware, or improve performance (see Figure 1).

To measure the work capabilities of LMs, we ground our evaluations in real-world LM use. While most LM usage data is not public, we use open-source datasets such as WildChat [Zhao et al., 2024] and LMSys [Zheng et al., 2023]. Directly classifying the resulting 4,499,105 real user queries into thousands of work categories would be very expensive, so we design a multi-stage retrieval pipeline to balance costs with pipeline precision and coverage. This yields benchmark queries for 143 DWAs (6.9% of all DWAs). If, instead of balancing costs with coverage, we strictly maximized coverage, the upper bound on coverage from available public usage data would be 38.6% of all DWAs, which would increase costs by $300\times$. In comparison, the January 2026 Anthropic Economic Index reports Claude usage for only 17.2% of tasks [Appel et al., 2026].

Given the limits of existing usage data, we design a synthetic data generation pipeline to improve query coverage of U.S. work, including for work categories where LMs are not yet adopted. Specifically, we generate prompts at varying levels of coverage over a target task and use an LM verifier to check that the generated prompt both reflects what a worker in the occupation would do and contains all the information required to be answerable. The result is benchmarks for all 2,087 DWAs, spanning all 1,016 occupations in the U.S. labor economy. This substantially improves coverage over past benchmarks including OpenAI's GDPval [Patwardhan et al., 2025] that covers 44 occupations ($< 5\%$ of all U.S. occupations). Our benchmarks are $500\times$ cheaper than GDPval, and we find that they predict occupational-level GDPval scores and are more predictive than generic capability benchmarks.

While benchmarks are the most common evaluation format in AI research, they foreground the disparities in model performance rather than how capabilities will yield economic value. To complement the benchmarks we build, we also estimate *exposure*: how much time would current capabilities save workers? Exposure has become the dominant lens for translating between technological capabilities and economic returns [e.g. Autor et al., 2003], including recent applications to frontier AI [e.g. Eloundou et al., 2024]. Importantly, exposure guarantees neither increased labor productivity nor increased job displacement. We introduce a *whitebox simulation-based exposure measure* that builds

[1]Many other organizational and regulatory factors also separate capabilities from impacts [Narayanan and Kapoor, 2025].
[2]EconEvals is also used by Fish et al. [2026] to describe their research on AI's performance at making economic decisions.

Approach	**Open data**	**Grounded in task-execution performance**	**Comparable across models**	**Readily extensible to the full task space**
Rubric-based exposure	●	○	◐	●
Adoption-based exposure	○	●	○	◐
Economically relevant benchmarks	◐	●	●	○
Simulation-based exposure (ours)	●	●	●	●

Table 1: **Comparing approaches to exposure estimation for four desiderata.** ● indicates the property is fully satisfied; ◐ indicates partially satisfied; ○ indicates not satisfied.

on prior work using LMs to simulate human behavior [Park et al., 2023]. Our exposure measure simulates a worker using a LM chatbot to complete a task, producing a detailed accounting of the steps involved in performing the task, the baseline time per step, and the time savings per step attributable to LMs. This satisfies several desiderata unaddressed by previous measures (Table 1).

Using our simulation-based exposure measure, we find that current LMs could save workers substantial time on at least half of the tasks in 46.6% of U.S. occupations. In spite of the broad potential for substantial time savings, we find that current usage does not cover all of these task-level opportunities: 79.2% of tasks that are predicted to be exposed by our metric have low Claude usage according to Anthropic's statistics. Further, by not only producing quantitative exposure estimates but an underlying rationale for how these time savings are achieved, we identify bottlenecks that inhibit technological capabilities from translating to productivity gains. Beyond inherent constraints of LMs like physical interaction, verification and privacy constraints are the primary current bottlenecks inhibiting further task-level time savings. We hope that by releasing all of our data, including the underlying instances of simulated workers interacting with language models, we can directly support real human workers in discovering how to better derive work benefits from current language models. Overall, our infrastructure helps bridge the gap between AI capabilities and labor market impact, and is designed to keep pace with both technological change (e.g. new models) and economic change (e.g. new tasks).

2 Data

ECONEVALS evaluations are grounded in real usage where possible. Given the limited coverage of US work in current (public) usage, we ensure complete coverage through synthetic data.

2.1 Real Data

We accumulate real user conversations from public datasets and map these conversations to work categories through a multi-step pipeline involving embedding-based retrieval and language model classifiers. Subject to cost constraints, we optimize the accuracy of the pipeline. To assess the overall quality, we report (i) total costs, (ii) precision in mapping from conversations to work categories, and (iii) coverage of the space of work categories.

Public datasets. We draw on five large public corpora of real user–chatbot conversations published with user consent that total to $4{,}499{,}105$ conversations. WildChat-4.8M [Zhao et al., 2024] provides 3.2M conversations from free GPT-3.5-Turbo and GPT-4 chat services hosted on Hugging Face Spaces. LMSYS-Chat-1M [Zheng et al., 2023] provides 1M conversations from the Vicuna demo and Chatbot Arena. Chiang et al. [2024] provides three releases (55k, 100k, 140k) of Chatbot Arena human preference data.

Pipeline design. We map user conversations to work categories at the detailed work activity (DWA) level of the O*NET taxonomy. We select this level of the hierarchy as the most fine-grained level of the O*NET taxonomy that was feasible given our data.[3] Directly classifying $4{,}499{,}105$ into $2{,}087$ categories would be prohibitively costly. Therefore, we decompose the classification problem into an

[3]Tasks are more fine-grained than DWAs but in initial experiments we found tasks to be infeasible: pipeline precision was considerably lower and the number of identified conversations per task to be too low for our purposes.

initial lightweight retrieval phase followed by a more costly classification phase that only triggers for a much smaller subset of (conversation, DWA) pairs. However, to ensure adequate coverage over occupations, we retrieve at the more fine-grained task level where each task corresponds to a single occupation (i.e., for each DWA, we perform retrieval & pairwise classification for all tasks associated with that DWA). Importantly, the benchmarks we produce are still at the DWA-level (we try to retrieve at least 50 samples per DWA); we just use task-level metadata to improve the coverage of our pipeline.

To implement the retrieval phase, we use standard embedding-based retrieval methods [Douze et al., 2024]. We embed every conversation as well as the tasks/occupations associated with each DWA. Since classifying all 9.4×10^9 DWA–conversation pairs (or the even larger number of task–conversation pairs) is infeasible, we reduce cost by selecting promising DWAs that are likely to have enough conversational coverage. We select 200 DWAs that have the highest number of relevant conversations based on retrieving the top 10 most similar conversations for each associated task and then running the subsequent pairwise classification pipeline. Then, for each selected DWA, we retrieve the top 300 most similar conversation for each associated task.

Given the retrieved (DWA, conversation) pairs, we classify them using hierarchical triage: we use cheap language models to narrow the initial pool and expensive language models to further refine it to manage costs while improving precision. For all (DWA, query) pairs that survive, we filter out low-quality queries and non-work queries. Finally, given the resulting high-quality work-specific conversations for each DWA, we convert them into benchmarking queries by using an LM to select the user-turn most relevant to the DWA. The full pipeline implementation with each step is described in §A.1.

Quality. We designed our query-task mapping to maximize precision and minimize overall costs. To estimate costs, we aggregate costs per pipeline step: the overall cost is approximately $768 for nearly all tasks and occupations in O*NET , which breaks down into $242 for embedding the data/queries, $477 for multi-stage language model classification, and $67 for subsequent data filtering. To estimate precision, we label a sample of 50 query-task pairs identified by our pipeline based on whether we assess the query-task mapping to be correct: the overall pipeline precision is 0.91.

In addition to high precision and low costs, our third desiderata is to maximize recall. Since we perform retrieval and do not classify every possible query-task pair, measuring recall is more complicated than measuring precision. In §A.2, we plot the results of two experiments that test how DWA-level coverage would improve with (i) broader retrieval beyond the selected 200 DWAs and (ii) deeper retrieval beyond the the top-300 most similar retrieved queries. The overall pipeline recall is 0.18: the pipeline identifies 143 DWAs with 50+ queries and the experiments predict that sufficiently broad and deep retrieval would identify 806 DWAs with 50+ queries. This recall is achieved at a cost savings of approximately $300\times$ since we only perform the costly language model classification step for 300 queries $\times$ 200 DWAs instead of 10000 queries $\times$ 2000 DWAs. We prioritize recall to improve task coverage given the limited task coverage of existing benchmarks [Wang et al., 2026]. For the criterion of at least 50 queries per covered DWA, the pipeline covers 6.9% of O*NET DWAs and is upper bound at 38.6% given our recall estimate.

To contextualize our task coverage, we consider usage statistics published by Anthropic based on their proprietary Claude usage data. The January 2026 Anthropic Economic Index publishes usage for tasks with at least 15 conversations or 5 unique accounts: 17.2% of tasks and roughly half of the DWAs are covered.[4] Our retrieval is high precision but low recall: 95.8% of the DWAs for which we retrieved samples for appear in the Economic Index usage data, but our retrieval only covers 12.5% of the their DWAs, which accounts for 27.4% of their usage. For example, we find that their most frequent DWAs ("Modify software programs to improve performance" and "Tutor students who need extra assistance") are not retrieved but similar education and programming DWAs are covered like "Write computer programming code" and "Develop instructional materials". Ultimately, current usage data of all types does not cover the full diversity of U.S. work.

[4]These values are for O*NET 20.1 and are at the task level: we say a DWA is covered if any associated task is covered.

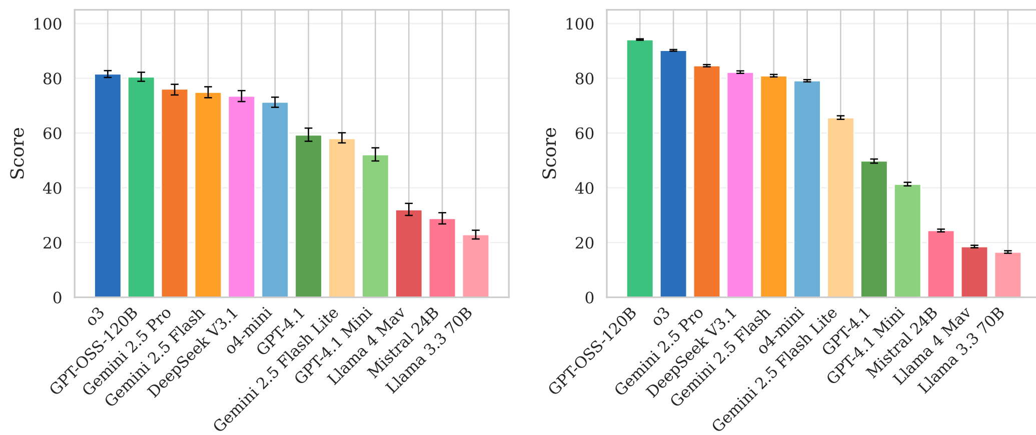


Figure 2: **Comparing economic benchmarks to raw public usage benchmarks.** The left subfigure depicts model performance on 500 randomly sampled public user queries whereas the right depicts model performance averaged across all our 143 50-instance DWA-level benchmarks.

2.2 Synthetic User Queries

Public usage data has limited work coverage and even proprietary usage is fundamentally constrained to how AI is currently used. However, we believe AI should be evaluated for all work use cases: evaluations could inform the procurement and adoption of AI for new tasks as they demonstrate technological improvement. We introduce a simulation-based synthetic data generation pipeline to cover (essentially) all U.S. work.[5]

To control for the complexity of the queries we generate, for each task and occupation, we first create a worker persona and then have `GPT-5-mini` roleplay as a worker with this persona responding to an interviewer asking about time savings. To begin, we start with the maximal time savings and see if a query can be generated to justify such high time savings, iteratively backing off to lower and lower exposure estimates. Since high time savings may arise due to LM hallucinations or other departures from realistic economic modeling, we introduce additional verification steps. The implementation for synthetic data generation is described in more detail in Appendix B.

3 Benchmarks

We present results for 226 benchmarks using our real and synthetic economically-categorized data as the benchmark queries and using language model judges to score model responses to these queries.

Scoring. We evaluate 10 models on our benchmarks, namely `gpt-4.1-mini`, `gpt-4.1`, `o4-mini`, `o3`, `gpt-oss-120b`, `gemini-2.5-pro`, `gemini-2.5-flash`, `gemini-2.5-flash-lite`, `Llama-3.3-70B`, `DeepSeek-V3.1`.[6] We cover different model sizes (e.g., Gemini 2.5 flash-lite, Gemini 2.5 flash, Gemini 2.5 Pro), different model developers (e.g., OpenAI, Google), flagship thinking/non-thinking models (e.g., o3 vs GPT-4.1), and prominent open/closed models (e.g., DeepSeek-V3.1 vs GPT-4.1).

Model responses are generated for each assessed model on each benchmark. To assess the quality of model responses, we use a language model judge. Following prior work [Li et al., 2024], we elicit binary preferences from the judge: is the evaluated model's response better, worse, significantly better, significantly worse, or the same as the response of a reference model?[7] Overall, we present results for 226 benchmarks in this work: 143 50-instance DWA-level benchmarks based on real data, 40 50-instance DWA-level benchmarks based on synthetic data, and 43 50-instance occupation-level benchmarks for the occupations in GDPval [Patwardhan et al., 2025].[8]

[5] In initial experiments, we simply prompted LMs to generate queries but found the resulting synthetic data low quality.
[6] We bias towards cheaper models to reduce evaluation cost, so we do not report results for some expensive frontier models.
[7] We set `o3-mini` as the reference model for analyzing the results in the paper.
[8] GDPval has 44 occupations, but we remove Buyers and Purchasing Agents as it's not present in O*NET 30.2.

DWA-level benchmarks based on real data. In Figure 2, we plot model performance in aggregate on public usage data as-is and averaged across our per-DWA benchmarks. While the general trends are similar, `gpt-oss-120B` performs the best by a clear margin on our economic benchmarks as compared to the raw usage where it is tied for first. Disaggregating by DWA confirms this: of the 37 benchmarks where a model is in first place by a statistically significant margin (i.e. its confidence interval does not overlap with any other model's confidence interval), 35 have `gpt-oss-120B` in first place. Overall, the per-DWA benchmarks generally induce similar rankings as the average ranking across all 143 DWAs. Further, the variance in model performance significantly increases with our economic benchmarks, indicating that our economic weighting identifies greater separation in model quality compared to the raw public usage data. This is particularly noteworthy since a sizable fraction of our usage data derives from the popular leaderboard LMArena.

DWA-level benchmarks based on synthetic data. Our pipeline for classifying public usage data amounts to DWA-level benchmarks for only about 7% (143/2,087) of the DWAs. Based on our synthetic data, we evaluate models on 40 DWAs: 20 randomly sampled from the 143 DWAs that the real data covers and 20 randomly sampled from the remaining 93% of uncovered DWAs. In Figure 8, we report the results for all 40 DWAs.[9] On the 20 DWAs covered by the real data, we find the synthetic benchmark results, on average, have a spearman correlation of 0.67 with the real benchmark results (Figure 11). Comparing the synthetic benchmark results for the two sets of 20 DWAs, we find an average spearman correlation of 0.86 (Figure 9).

Disaggregating the results by DWA, we again find that the model ranking is similar across DWAs. But DWAs differ in how much they separate models (Figure 10). DWAs that involve instructing or advising tend to have the largest variance in model performance and best separate model quality. In contrast, DWAs that involve verification or record maintenance tend to have the least variance in model performance and thereby generally do not identify differences in model quality. For example, all 3 DWAs that involve maintaining records (i.e. "Maintain medical records", "Maintain operational records", "Maintain the inventory of equipment") are in the bottom quintile when ranking by variance whereas "Provide technical guidance to other personnel" is the DWA with the greatest variance. The sole counterexample we find to these trends is the "Document operational procedures" detailed work activity, which resembles record maintenance, but is the third highest DWA by variance.

Occupation-level benchmarks based on synthetic data. Patwardhan et al. [2025] built the GDPval benchmark using queries sourced from workers to yield benchmarks for 44 occupations, each containing 30 task instances. The selected occupations all predominantly perform knowledge work and belong to the 9 U.S. sectors that each contribute over 5% of GDP, prioritizing the occupations with the largest total wage-and-compensation contribution within each sector. Compared to GDPval, the ECONEVALS synthetic benchmarks have five advantages: (i) greater coverage of U.S. work, (ii) clearer understanding of the task-level coverage within an occupation, (iii) more instances per occupation, (iv) lower cost to produce the benchmark, and (v) full transparency and open-source data. We estimate that our evaluation queries cost less than $1000 to produce whereas the GDPval evaluation queries cost more than $500,000 to produce (Appendix C). However, GDPval has superior data quality based on the small subset of queries they publish: they better test frontier capabilities (e.g. multi-turn, agents, tool use and web search) rather than our narrower focus on language models and they more directly align with work deliverables than task completion (e.g. producing a legal brief rather than performing legal research).

We test whether our occupation-level benchmark scores predict GDPval scores for the same occupations. We find that the spearman correlation between the synthetic and GDPval scores is, on average, 0.67. Since prior work finds that many capability benchmarks are correlated [Ho et al., 2025], we further test whether generic capability measures also predict GDPval scores or whether our benchmarks provide increased explanatory power. Compared to the existing capability benchmarks, the generated synthetic benchmarks win or tie in terms of spearman correlation in the majority of cases. However, only seven out of nine are statistically significant (Figure 13).

[9] We do not evaluate Llama-4-Maverick and Mistral-24B as they were were deprecated on Together AI.

Occupation: Hosts and Hostesses, Restaurant, Lounge, and Coffee Shop.
Task: Receive and record patrons' dining reservations.

Blackbox rubric-based exposure.

Prompt. Automation rubric (T0–T4— use exact definitions).
T0— NO AUTOMATION. System cannot perform any meaningful component of the task.
[. . .]
T3— HIGH AUTOMATION System can perform 80–100% at high quality, BUT human oversight is required because of liability/safety, stakeholder trust, or rare catastrophic failure modes.
T4— FULL AUTOMATION. System performs 100% with high quality. Human oversight not routinely needed.

Model response. `T3`

Estimated time saved: 80–100%

Whitebox simulation-based exposure.

Per-step reasoning (excerpts).
1) Prepare station, check availability (1–2 min). Now: ∼1 min, scan OpenTable grid and manager notes. Chatbot effect: none— can't read the grid. Time saved: 0%.
[. . .]
11) Update waitlist, callbacks (1–10 min). Now: manual calls take minutes each. Chatbot effect: only useful if integrated. Time saved: none to small without integration.

Estimated time saved: 0–25%

Figure 3: **Prior blackbox rubric-based vs. our whitebox simulation-based exposure measures.** Left side shows a blackbox rubric-based measure using a shortened rubric from prior work [Hosseini Maasoum and Lichtinger, 2026]. Right side shows our decomposition of the task into steps with step-wise accounting to determine the overall time savings based on our worker simulation.

4 Exposure

We aim to understand how language model capabilities are transformed into impacts on the labor market. Economists have developed *exposure measures* to study this transformation, which measure how much time workers would save given new technologies [Autor et al., 2003, Acemoglu and Autor, 2011, Acemoglu and Restrepo, 2019]. Several works estimate AI-driven exposure [Felten et al., 2018, Webb, 2020, Felten et al., 2021, Tolan et al., 2021, Felten et al., 2023, Pizzinelli et al., 2023, Eloundou et al., 2024, Massenkoff and McCrory, 2026] with Eloundou et al. [2024] popularizing exposure measurement for LMs.

Scoring. Existing exposure estimates for LMs [Eloundou et al., 2024, Hosseini Maasoum and Lichtinger, 2026] generally (i) specify a list of current capabilities and (ii) use LMs, workers, or occupational experts to annotate the extent to which these capabilities would save time for a given work task. We call these *blackbox rubric-based* exposure measures: the measure yields a task-level exposure score based on a predefined rubric of capabilities with no transparency into how the score was computed. Exposure estimation is a complex task that requires jointly reasoning about general technological capabilities, specific work tasks, and how the two interact. For example, to estimate language model exposure for the task of "Taking customers' orders" requires accounting for the fact that even if LMs can do certain steps like summing the prices in an order or making item recommendations, they would also need to handle synchronous in-person customer interactions to further save time.

We introduce the first *whitebox simulation-based* exposure method (see Figure 3). To estimate the exposure of a task, we roleplay a "worker" using a language model and construct a decomposition of each task into steps. Given this decomposition, the simulated "worker" provides a baseline estimate for the time per step and the amount this can be reduced given current language model capabilities. The resulting per-step time savings are aggregated to determine the overall time savings as our exposure prediction. We discretize these time savings into four categories, similar to prior work [Hosseini Maasoum and Lichtinger, 2026]. Through this simulation, we produce a numerical exposure estimate backed by an underlying reasoning trace that directly accounts for what we understand the task to entail and which steps within that task are accelerated by current language model capabilities. We describe the full implementation of our simulation with examples of the resulting per-task exposure accounting in Appendix D.

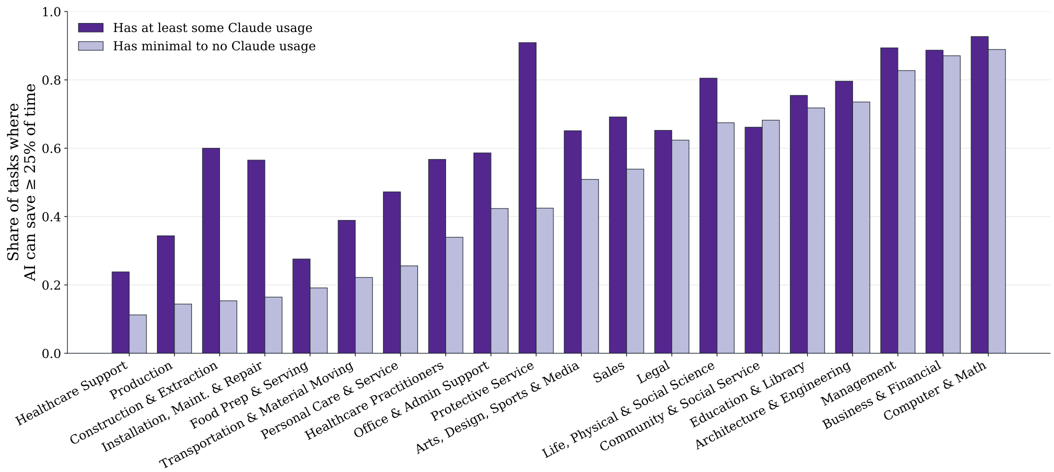


Figure 4: **Simulation-based exposure estimates by occupational group and Claude usage.** We plot the exposure estimate for each SOC major occupational group aggregated across all occupations with the group and tasks within those occupations. Results are stratified based on whether the underlying tasks have reported Claude usage (dark; at least 0.0025% of Anthropic's work-related Claude Sonnet 4.5 traffic) or not (light).

Comparing to LM Usage Data We compare our exposure estimates to the usage data from the Anthropic Economic Index [Handa et al., 2025]. Figure 4 compares the simulation-based exposure for tasks that appear in at least 0.0025% of Anthropic's work related traffic for Claude Sonnet 4.5 versus those that do not. Among tasks that see some Claude usage, 73.4% (1437/1957) are ones where AI can save at least 25% of time, indicating that adoption is concentrated on tasks with meaningful exposure. However, among tasks where AI can save at least 25% of time, 79.2% (5456/6893) have minimal to no Claude usage.

High predicted exposure scores for tasks with little to no current Claude usage suggest that LMs could be applied more extensively in people's jobs than current usage patterns indicate. Such gaps between capability and adoption are characteristic of general-purpose technologies, which historically diffuse through the economy only as firms develop the complementary workflows, skills, and organizational practices needed to deploy them effectively. For example, for Online Merchants calculating revenue and expenses, our exposure measure predicts that LMs could generate spreadsheet formulas, navigate platform tools like Amazon Seller Central, and reconcile bank statements against payment-processor records. For Architects administering construction contracts, LMs could summarize contract clauses, flag inconsistencies between schedules of values and inspection evidence, and produce post-meeting notes from contractor calls. Despite the predicted time-savings for these tasks and occupations under the simulation-based exposure measure, these tasks see little to no Claude usage. In §D.1 we further analyze tasks that the simulation-based-exposure measure labels as exposed to LMs but have minimal Claude usage, and tasks that have at least some Claude usage but are labeled as minimally exposed.

Exposure analysis reveals bottlenecks. By directly accounting for how we predict LMs save worker's time, we can inspect the different factors that contribute to task exposure predictions. Our exposure estimates surfaces specific real-world *bottlenecks* that limit AI-enabled time savings on a task-by-task basis. We summarize our findings by using an LM to categorize our exposure measure's produced justifications why an LM would/wouldn't save a worker time. In Figure 6, we visualize this breakdown across different bottleneck categories. As expected, physical interaction requirements are the most common bottleneck preventing LMs from saving worker time. Additionally, tasks requiring live interaction, LM consultation overheard, privacy constraints, and real-time monitoring account for a substantial share of tasks that are labeled as not exposed by our simulation-based exposure measure. Rubric-based measures, which assume a fixed list of capabilities, cannot separate tasks that are "physically impossible" from those that are "feasible but bottlenecked on verification or privacy." On the other hand, the tasks with the greatest time savings often involve steps like drafting long-form text (e.g. reports), drafting communication-related text (e.g. emails), and generating code.

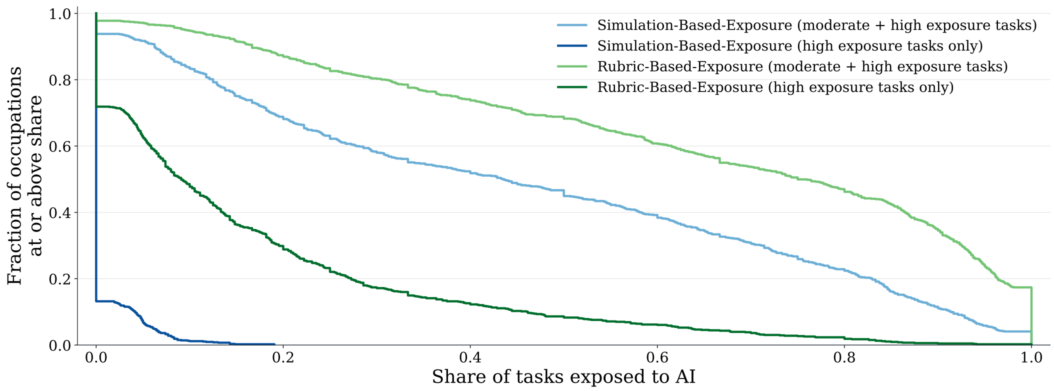


Figure 5: **Simulation-based vs. rubric-based task-level exposure estimates.** We plot the occupation-level exposure for our method and prior work as a function of exposure threshold. At both the high and moderate-to-high threshold, the simulation-based method predicts lower exposure as it surfaces bottlenecks to adoption that only arise from a more detailed accounting of AI-use.

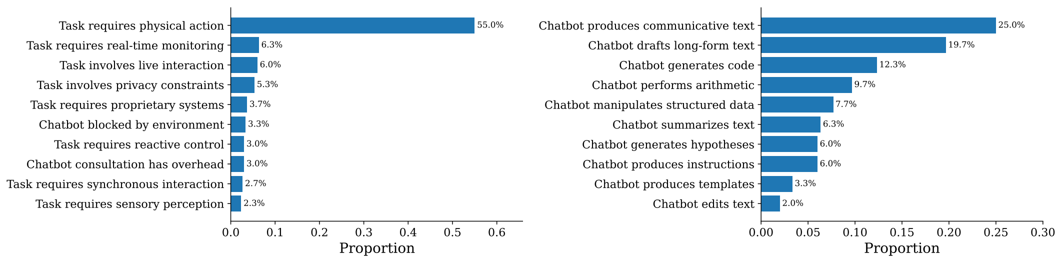


Figure 6: **Explaining the time savings via reasoning trace analysis.** Left side shows bottlenecks that limit time savings below 25%. Right side shows uses that enable time savings above 25%.

Comparing to rubric-based exposure To implement the rubric-based exposure measure we use the prompt and rubric from Hosseini Maasoum and Lichtinger [2026] rather than from Eloundou et al. [2024]: the former updates the prompt used by the latter to reflect current LMs, and they are strongly correlated; Hosseini Maasoum and Lichtinger [2026] report an $R^2 = 0.76$ between the two measures.

Figure 5 compares exposure intensity across 857 occupations, which constitute 93% of the occupations in O*NET. We exclude the remaining occupations due to data generation errors encountered when computing the simulation-based exposure. We compute exposure intensity using either our proposed simulation-based exposure or rubric-based exposure. We call a task highly exposed if the exposure measure predicts an LM can save $> 50\%$ of time, and moderately or highly exposed at the $> 25\%$ threshold. At both exposure levels, simulation-based exposure marks a smaller fraction of each occupation's tasks as exposed to AI than rubric-based exposure does. The former grounds LM capabilities in real performance via simulated interactions with an LM, while the latter assumes that LMs attain a list of capabilities that, in practice, are likely not fully achieved (e.g., "Data analysis," "Business analytics," "Idea generation"). The simulation-based exposure measure suggests that 46.6% of occupations have at least 50% of their tasks at a medium or high AI exposure level, compared to 75.6% under the rubric-based exposure measures. We posit that grounding in simulated interactions with an LM enables the simulation-based measure to better estimate exposure to *current* LMs.

5 Conclusion

We introduce ECONEVALS as an open-source evaluation suite for measuring language model capabilities on U.S. work categories. By providing both benchmark-style and exposure-style measures, expanding coverage of the U.S. work, and operating at the task, DWA, and occupational levels, we develop an array of tools for better measurement in this domain. We encourage future work to explore how these different measures predict downstream economic indicators like employment, wages, and productivity. In addition, future work can address the most fundamental limitations of our work, namely our technological scope being limited to current LM chatbots and our economic scope being limited to the current U.S. labor economy. Overall, we contribute measurement infrastructure to support the broader research agenda and evidence base on the economics of frontier AI.

Acknowledgements

We thank Abhishek Nagaraj, Amir Zeinali, Andy Haupt, Arjun Ramani, Arvind Narayanan, Avanika Narayan, Bharat Chandar, Dan Ho, Dawn Song, Divya Siddharth, Diyi Yang, Dylan Clement, Erik Brynjolfsson, Jacob Steinhardt, Jaime Sevilla, Joel Becker, Jon Saad-Falcon, Kris Gulati, Lukas Freund, Lukas Mann, Peter Cihon, Phil Trammell, Parker Whitfill, Rob Reich, Sam Manning, Sayash Kapoor, Tejal Patwardhan, Tom Cunningham, Yijia Shao, and Yifan Mai for helpful discussion. We thank the Stanford Center for Research on Foundation Models (Stanford CRFM) and Stanford Institute for Human-Centered Artificial Intelligence (Stanford HAI) for funding.

References

Ipsos. Ipsos predictions survey 2025: Positivity about how this year has gone highest since before the pandemic, December 2024. URL https://www.ipsos.com/en-us/ipsos-predictions-2025. Reports that globally 65% expect AI-driven job losses and 43% expect AI-driven job creation. Accessed: 2026-04-09.

United Nations. Human development report 2025: A matter of choice: People and possibilities in the age of ai. Technical report, United Nations Development Programme, 2025. URL https://hdr.undp.org/system/files/documents/global-report-document/hdr2025reporten.pdf. Uses the 2025 Global Survey on AI and Human Development; reports that 57% expect their current jobs to be replaced due to AI. Accessed: 2026-04-09.

Stanford HAI. Ai index report 2025: Public opinion. Technical report, Stanford HAI, 2025. URL https://hai.stanford.edu/ai-index/2025-ai-index-report/public-opinion. Reports that globally 60% expect AI to change how they do their job and 36% expect job replacement within five years. Accessed: 2026-04-09.

The White House. Artificial intelligence and the great divergence, January 2026. URL https://www.whitehouse.gov/research/2026/01/artificial-intelligence-and-the-great-divergence/. Accessed: 2026-04-09.

Connacher Murphy, Josh Rosenberg, Jordan Canedy, Zach Jacobs, Nadja Flechner, Rhiannon Britt, Alexa Pan, Charlie Rogers-Smith, Dan Mayland, Cathy Buffington, Simas Kučinskas, Amanda Coston, Hannah Kerner, Emma Pierson, Reihaneh Rabbany, Matthew Salganik, Robert Seamans, Yu Su, Florian Tramèr, Tatsunori Hashimoto, Arvind Narayanan, Philip E. Tetlock, and Ezra Karger. The longitudinal expert ai panel: Understanding expert views on ai capabilities, adoption, and impact. Working paper 5, Forecasting Research Institute, 2025. URL https://leap.forecastingresearch.org/forecasts.

Ezra Karger, Otto Kuusela, Jason Abaluck, Kevin Bryan, Basil Halperin, Todd Jones, Connacher Murphy, Phil Trammell, Matt Reynolds, Dan Mayland, Ria Viswanathan, Ananaya Mittal, Rebecca Ceppas de Castro, Josh Rosenberg, and Philip E. Tetlock. Forecasting the economic effects of ai, March 2026. URL https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/69cbb9d509ada447b6d9013f/1774959061185/forecasting-the-economic-effects-of-ai.pdf. PDF.

Daron Acemoglu. The simple macroeconomics of ai. Working Paper 32487, National Bureau of Economic Research, May 2024. URL https://www.nber.org/papers/w32487.

Dario Amodei. The adolescence of technology, January 2026. URL https://www.darioamodei.com/essay/the-adolescence-of-technology. Accessed: 2026-04-09.

Bharat Chandar. Ai and labor markets: What we know and don't know, October 2025. URL https://digitaleconomy.stanford.edu/news/ai-and-labor-markets-what-we-know-and-dont-know/. Accessed: 2026-04-09.

Alex Imas. What is the impact of ai on productivity? Ghosts of Electricity (Substack), January 2026. URL https://aleximas.substack.com/p/what-is-the-impact-of-ai-on-productivity. Accessed: 2026-04-09.

R. Maria del Rio-Chanona, Ekkehard Ernst, Rossana Merola, Daniel Samaan, and Ole Teutloff. Ai and jobs. a review of theory, estimates, and evidence, 2025. URL https://arxiv.org/abs/2509.15265.

Yoshua Bengio, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Malcolm Murray, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas, David Duvenaud, Philip Fox, Usman Gohar, Rose Hadshar, Anson Ho, Tiancheng Hu, Cameron Jones, Sayash Kapoor, Atoosa Kasirzadeh, Sam Manning, Nestor Maslej, Vasilios Mavroudis, Conor McGlynn, Richard Moulange, Jessica Newman, Kwan Yee Ng, Patricia Paskov, Shalaleh Rismani, Girish Sastry, Elizabeth Seger, Scott Singer, Charlotte Stix, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Vincent Conitzer, Thomas G. Dietterich, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Susan Leavy, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Sarvapali D. Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, Ya-Qin Zhang, Leandro Angelo Aguirre, Olubunmi Ajala, Fahad Albalawi, Noora AlMalek, Christian Busch, Jonathan Collas, André Carlos Ponce de Leon Ferreira de Carvalho, Amandeep Gill, Ahmet Halit Hatip, Juha Heikkilä, Chris Johnson, Gill Jolly, Ziv Katzir, Mary N. Kerema, Hiroaki Kitano, Antonio Krüger, Kyoung Mu Lee, José Ramón López Portillo, Aoife McLysaght, Oleksii Molchanovskyi, Andrea Monti, Mona Nemer, Nuria Oliver, Raquel Pezoa, Audrey Plonk, Balaraman Ravindran, Hammam Riza, Crystal Rugege, Haroon Sheikh, Denise Wong, Yi Zeng, Liming Zhu, Daniel Privitera, and Sören Mindermann. International ai safety report 2026. Technical Report DSIT 2026/001, Department for Science, Innovation and Technology, UK Government, February 2026. URL https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026.pdf.

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Šuppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Knight, Yiyang Nan, Jyoutir Raj, Yu Fan, Shubham Singh, Subramanyam Sahoo, Eliya Habba, Usman Gohar, Siddhesh Pawar, Robert Scholz, Arjun Subramonian, Jingwei Ni, Mykel Kochenderfer, Sanmi Koyejo, Mrinmaya Sachan, Stella Biderman, Zeerak Talat, Avijit Ghosh, and Irene Solaiman. When ai benchmarks plateau: A systematic study of benchmark saturation, 2026. URL https://arxiv.org/abs/2602.16763.

Sha Sajadieh, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Lapo Santarlasci, Juan Pava, Nestor Maslej, Russ Altman, Erik Brynjolfsson, Carla Brodley, Jack Clark, Virginia Dignum, Vipin Kumar, James Landay, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Elham Tabassi, Russell Wald, Toby Walsh, and Dan Weld. Artificial intelligence index report 2026. Technical report, Stanford Institute for Human-Centered Artificial Intelligence, Stanford, CA, April 2026. URL https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf.

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 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 *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.14499. arXiv:2503.14499.

METR. Task-completion time horizons of frontier ai models. https://metr.org/time-horizons/, March 2026.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025. URL https://arxiv.org/abs/2510.04374.

Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, Jaehyuk Lim, Lachlan Carroll, Alice Blair, Vinaya Sivakumar, Sumana Basu, Brad Kenstler, Yuntao Ma, Julian Michael, Xiaoke Li, Oliver Ingebretsen, Aditya Mehta, Jean Mottola, John Teichmann, Kevin Yu, Zaina Shaik, Adam Khoja, Richard Ren, Jason Hausenloy, Long Phan, Ye Htet, Ankit Aich, Tahseen Rabbani, Vivswan Shah, Andriy Novykov, Felix Binder, Kirill Chugunov, Luis Ramirez, Matias Geralnik, Hernán Mesura, Dean Lee, Ed-Yeremai Hernandez Cardona, Annette Diamond, Summer Yue, Alexandr Wang, Bing Liu, Ernesto Hernandez, and Dan Hendrycks. Remote labor index: Measuring ai automation of remote work. *arXiv preprint arXiv:2510.26787*, 2025. URL https://arxiv.org/abs/2510.26787.

Bertie Vidgen, Abby Fennelly, Evan Pinnix, Chirag Mahapatra, Zach Richards, Austin Bridges, Calix Huang, Ben Hunsberger, Fez Zafar, Brendan Foody, Dominic Barton, Cass R. Sunstein, Eric Topol, and Osvald Nitski. The ai productivity index (apex). *arXiv preprint arXiv:2509.25721*, 2025. URL https://arxiv.org/abs/2509.25721.

Zora Zhiruo Wang, Sanidhya Vijayvargiya, Aspen Chen, Hanmo Zhang, Venu Arvind Arangarajan, Jett Chen, Valerie Chen, Diyi Yang, Daniel Fried, and Graham Neubig. How well does agent development reflect real-world work?, 2026. URL https://arxiv.org/abs/2603.01203.

Arvind Narayanan and Sayash Kapoor. Ai as normal technology: An alternative to the vision of ai as a potential superintelligence. Technical report, Knight First Amendment Institute, April 2025. URL https://knightcolumbia.org/content/ai-as-normal-technology.

Sara Fish, Julia Shephard, Minkai Li, Ran I. Shorrer, and Yannai A. Gonczarowski. Econevals: Benchmarks and litmus tests for economic decision-making by llm agents, 2026. URL https://arxiv.org/abs/2503.18825.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2405.01470.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023. URL https://arxiv.org/abs/2309.11998.

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

David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333, November 2003. doi: 10.1162/003355303322552801. URL https://academic.oup.com/qje/article-abstract/118/4/1279/1925105.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: Labor market impact potential of llms. *Science*, 384(6702):1306–1308, 2024. doi: 10.1126/science.adj0998. URL https://www.science.org/doi/abs/10.1126/science.adj0998.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, 2023. doi: 10.1145/3586183.3606763. URL https://dl.acm.org/doi/10.1145/3586183.3606763.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. *arXiv preprint arXiv:2401.08281*, 2024. URL https://arxiv.org/abs/2401.08281.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024. URL https://arxiv.org/abs/2406.11939.

Anson Ho et al. A Rosetta Stone for AI benchmarks. *arXiv preprint arXiv:2512.00193*, 2025. URL https://arxiv.org/abs/2512.00193.

Daron Acemoglu and David Autor. Skills, tasks and technologies: Implications for employment and earnings. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 4, chapter 12, pages 1043–1171. Elsevier, Amsterdam, 2011. doi: 10.1016/S0169-7218(11)02410-5. Part B.

Daron Acemoglu and Pascual Restrepo. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30, 2019. doi: 10.1257/jep.33.2.3.

Edward W. Felten, Manav Raj, and Robert Seamans. A method to link advances in artificial intelligence to occupational abilities. *AEA Papers and Proceedings*, 108:54–57, 2018. doi: 10.1257/pandp.20181021.

Michael Webb. The impact of artificial intelligence on the labor market. Working paper, 2020.

Edward W. 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. doi: 10.1002/smj.3286.

Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, José Hernández-Orallo, and Emilia Gómez. Measuring the occupational impact of ai: Tasks, cognitive abilities and ai benchmarks. *Journal of Artificial Intelligence Research*, 71:191–236, 2021. doi: 10.1613/jair.1.12647.

Edward W. Felten, Manav Raj, and Robert Seamans. Occupational heterogeneity in exposure to generative ai. *SSRN Electronic Journal*, 2023. doi: 10.2139/ssrn.4414065.

Carlo Pizzinelli, Augustus Panton, Marina M. Tavares, Mauro Cazzaniga, and Longji Li. Labor market exposure to ai: Cross-country differences and distributional implications. Technical Report WP/23/216, International Monetary Fund, 2023.

Maxim Massenkoff and Peter McCrory. Labor market impacts of ai: A new measure and early evidence, 2026. URL https://www.anthropic.com/research/labor-market-impacts.

Seyed Mahdi Hosseini Maasoum and Guy Lichtinger. Generative ai and occupational entry barriers: The labor-supply channel of technological change. 2026. URL https://ssrn.com/abstract=6059674. SSRN Working Paper.

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

A Real Data

Given public user-chatbot conversations, we implement a pipeline that implements embedding-based retrieval followed by language model classification to map from conversations to detailed work activities (DWAs). Here we describe this pipeline in full detail. Then we describe how we assess the pipeline and analyze alternatives.

A.1 Pipeline implementation

Embedding. All of the embeddings used for the retrieved conversations are embedded as-is (i.e., concatenated turns, with each turn prefixed with "Human:" or "Assistant"). For O*NET tasks, we ensemble two sets of embeddings: (1) one that just embeds the concatenated task description and occupation title; (2) averaged embeddings across synthetically generated conversations. We use OpenAI's `text-embedding-3-small` model.

Selecting promising DWAs. There are 2,087 total DWAs. As we expect the majority of DWAs to have few positive classifications, we first select the most promising DWAs. We do this by first performing a "shallow" retrieval with $k = 10$ (looking at the top-10 most semantically similar conversations for each adjacent task). We then perform the rest of the classification pipeline and select the 200 DWAs with the most positive classifications.

Language model classification. Given the 200 DWAs we select, we retrieve the top 300 most similar conversations to perform pairwise classification for the resulting 200×300 pairs.

LMs are used to perform binary classification for a conversation-category pair, where the final output is whether that conversation belongs in that category. However, we may also preprocess the conversations or categories before passing them to the LM.

Although the final category that we care about are the DWAs, we can also classify against the lower-level task statement/occupations and also the higher-level IWAs. We classify against the lower-level task statements/occupations because the DWA titles by themselves are underspecified (e.g., DWAs alone don't define the set of occupations to consider, but the adjacent task statements do) and we classify against the higher-level IWAs because it allows us to filter out conversation-category pairs with fewer inferences (since there are fewer IWAs than DWAs).

Finally, we also may preprocess the user requests or the category descriptions prior to add information that improves the specificity of the classification or remove information that may bias the language model classifier.

For the conversations, we perform up to three preprocessing steps:

1. For all current configurations, only the user requests (and not the assistants' responses) are given to the model. Assistant responses tend to be verbose, which uses unnecessary tokens. But also, LMs have trouble ignoring misleading context and we want LMs to focus on the user requests (since that's what we're evaluating models on), so including assistants' responses tends to reduce accuracy as well.
2. We may filter out turns irrelevant to the category. This is done for a similar reason to the above: LMs have trouble ignoring irrelevant context & conversations often include turns like "Human: can you speak Spanish?"
3. In certain prompt configurations, we may summarize the user requests as well. This is also done for a similar reason as the above: user requests contain a lot of extraneous information & summarization allows us to guide LMs to focus on the correct parts.

We also preprocess the categories. DWA titles, IWA titles, and Task statements tend to be very brief (one sentence or less than one sentence), but the structure of the O*NET hierarchy includes information we can add: we use an LM to summarize e.g., adjacent DWAs (to add detail to the IWA titles), adjacent task statements (to add detail to the DWA titles) or contrast task statements for an occupation with other task statements that belong to that occupation (to add detail to the task statements).

In total, we ensemble across six turns of classification, which we summarize in Table 2.

Step	Category	Model	Prompt preprocessing	Category preprocessing
(1)	IWA	`openai/gpt-4.1-nano`	Remove assistant turns	Add details to IWA titles
(2)	DWA	`openai/gpt-4.1-nano`	Remove assistant turns	Add details to DWA titles
(3)	Task	`openai/gpt-4.1-mini`	Remove assistant turns	Add details to task statements
(4)	DWA	`openai/gpt-4.1-mini`	Remove assistant turns, filter irrelevant turns and summarize user queries	Add details to DWA titles
(5)	Task	`openai/gpt-4.1-mini`	Remove assistant turns, filter irrelevant turns and summarize user queries	Add details to task statements
(6)	Task	`openai/gpt-4.1-mini`	Remove assistant turns, filter irrelevant turns and summarize user queries	Add details to task statements

Table 2: Summary of language model classification steps.

Finally, we use an LM to filter out non-work and low-quality queries: for the former, we remove conversations that are obvious homework, coursework, or exam-style requests that the user wants completed; for the latter, we remove conversations that are unanswerable from the provided context, require capabilities a text-only LM does not have, or are otherwise illegitimate or joke-like rather than genuine work-related requests.

A.2 Pipeline evaluation

There are three sources of error where we may fail to retrieve at least 50 samples for a DWA despite at least 50 samples existing in real data.

First, we only perform retrieval on the 200 most promising DWAs and may miss samples from the remaining 1,887 DWAs. We select the promising DWAs based on the number of relevant samples when looking at only the top-10 most semantically similar samples. As such, by looking at the relationship between the number of relevant samples retrieved during this "shallow" trial to the rate at which we successfully find fifty relevant samples in the final run (looking at the top-300 most semantically similar samples), we can estimate the number of DWAs we missed by only looking at the 200 most promising ones rather than the full set. Specifically, we observe a linear relationship between the number relevant samples retrieved in the shallow trial and the probability we retrieve at least fifty samples in the full run (Figure 7a) and estimate that there are 65.2 DWAs that we missed by only looking at the 200 most promising DWAs rather than the full set of 2,087.

Second, we only perform classification looking at only top-300 most semantically similar conversations and may miss non-semantically similar conversations. We estimate the error from embedding-retrieval by running the pipeline on subsets of DWAs and comparing the number of relevant samples retrieved at different depths: 300, 1,000, and 10,000 (Figure 7b). We estimate that the probability we would have retrieved at least fifty relevant samples at depth 1,000 conditional on *not* retrieving at least fifty samples at depth 300 is 0.09 and the probability we would have retrieved at least fifty relevant samples at depth 10,000 conditional on *not* retrieving at least fifty samples at depth 300 is 0.319. This means that, for the 200 most promising DWAs, there are an additional $(200 - 143) \times 0.319 = 18.18$ we could have gotten by retrieving at a 33.3x greater depth and $(1882 - 65.20) \times 0.319 = 579.56$ for the DWAs that we did not try performing retrieval on, resulting in a total of $18.18 + 579.56 + 65.20 = 662.94$ additional DWAs.

Finally, the classification pipeline prioritizes high precision and may have false negatives. We sample 50 conversation-category pairs for which the pipeline classified as not relevant, and find that we agree with the prediction 73.4% of the time.

B Synthetic Data

We find that simple prompts tend to result in simplistic generations (e.g., just rephrasings of the task description). Instead, we use a prompt where the LM roleplays as a worker sharing how they use AI chatbots to assist with a given task.

Specifically, for each O*NET task and occupation, we first generate a worker "persona" consisting of the company/organization, the years of experience, specific job title, etc. We then use `GPT-5-mini` to roleplay as this worker persona. The user roleplays as an "interviewer" asking the "worker" to send them a prompt that helped save them time on the task. Crucially, we condition in this prompt the amount of time-savings that prompt provides for this task. This lets us generate prompts with varying degrees of "comprehensiveness": we can generate a complex prompt that performs the entire task if we ask the "worker" to send a prompt with 90-100% time-savings; we can also generate a simpler prompt with 1% time-savings.

Ideally, we'd like to generate as comprehensive of a prompt as possible, but an LM may be willing to hallucinate (or ignore certain constraints) to generate a prompt with high time-savings. So, we wrap this in a loop with additional prompts (using a more capable `GPT-5.2`) to verify that (1) the prompt reflects things that someone of this occupation would perform and (2) that the prompt is actually answerable. The full pipeline for synthetic generation then consists of trying to generate, prompting with the highest time-savings (90-100%) then lowering this time-savings amount until we get a synthetic prompt that passes the verifiers.

C Benchmarks

We report results on our benchmarks along with comparing the costs of producing our benchmarks with the costs of producing OpenAI's GPDval [Patwardhan et al., 2025].

C.1 Benchmark generation costs

Since GDPval is the state-of-the-art economic evaluation prior to our work, and perhaps even after our work given the data quality, we compare our benchmark generation procedure to their dataset. We estimate that generating synthetic data such that we have at least 50 instances for each of the 43 occupations costs less than $1,000. In contrast, generating (and validating) worker queries for OpenAI's GDPval to have at least 30 instances costs at least $500,000 according to our estimate.

To estimate the overall query generation cost, we study the exact GDPval benchmark design process, the selected occupations, and the associated wages. Assuming that workers involved in GDPval were paid their standard wage per unit time, we get an estimate of $\$566,296$ via a back-of-the-envelope calculation. GDPVal reports that each task was reviewed by an average of 5 people during creation, and that the mean review time for a model's output on a task was 109 minutes. We use the latter as a proxy for per-reviewer task-creation review time, noting that these are conceptually distinct activities. We exclude the time spent initially drafting or editing tasks, as GDPVal does not report values for these activities; including them would scale our estimate upward. We then compute per-occupation cost as 30 tasks $\times$ 5 reviews $\times$ $(109/60)$ hours $\times$ mean hourly wage for occupation, and sum across the 44 occupations covered by GDPVal. We rate our confidence in this estimate as low-to-medium, and emphasize that it likely represents a lower bound on the true cost.

D Exposure

D.1 Understanding low-exposure tasks with at least some Claude usage and vice versa

Figure 14 complements Figure 4, showing the reasons why some tasks that see Claude usage are predicted to have low exposure by our simulation-based exposure measure, and why some tasks that see minimal Claude usage are predicted to have high exposure.

D.2 Skill Importance

Following Eloundou et al. [2024], we next examine how our exposure estimates relate to the skills that occupations rely on most heavily. The O*NET taxonomy defines ten Basic Skills across two categories:

- **Content skills:** *Reading Comprehension* (understanding written work documents), *Active Listening* (attentive listening without interruption), *Writing* (effective written communication), *Speaking* (effective oral communication), *Mathematics* (using math to solve problems), *Science* (using scientific methods to solve problems).
- **Process skills:** *Critical Thinking* (logic and reasoning to evaluate alternatives), *Active Learning* (applying new information to problem-solving), *Learning Strategies* (selecting appropriate methods for learning or teaching), *Monitoring* (assessing performance to drive improvement).

O*NET provides labels for the importance of each skill for each occupation. We rescale each skill's per-occupation importance score, and then regress our exposure measures on these rescaled importance scores to characterize which skills are most strongly associated with AI exposure.

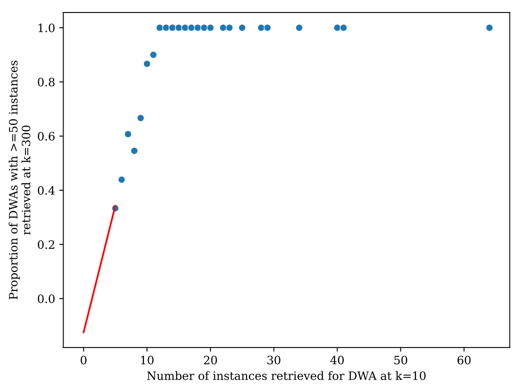
We run this regression analysis to predict the Simulation-Based Exposure measures and, separately, the Rubric-Based Exposure measures. At the moderate-or-higher exposure threshold, both methods agree closely: language-intensive skills (Writing, Reading Comprehension, Critical Thinking, Speaking) are the strongest correlates of exposure, and the regressions explain a similar share of variance ($R^2 = 0.695$ for Simulation-Based vs. 0.694 for Rubric-Based). At the high-exposure threshold, however, the two methods diverge sharply: Rubric-Based Exposure remains well-explained by skill importance ($R^2 = 0.520$), with Programming emerging as the dominant correlate, whereas Simulation-Based Exposure is only weakly predicted by O*NET skills ($R^2 = 0.071$) and exhibits no single dominant skill. This divergence suggests that high exposure under our method is not concentrated in occupations defined by any one skill, but is instead distributed more diffusely across occupational profiles.

Table 3: Correlations between O*NET Basic Skill importance and predicted AI exposure, under Simulation-Based and Rubric-Based measures, at moderate-or-higher and high exposure thresholds. r is the Pearson correlation between skill importance and exposure.

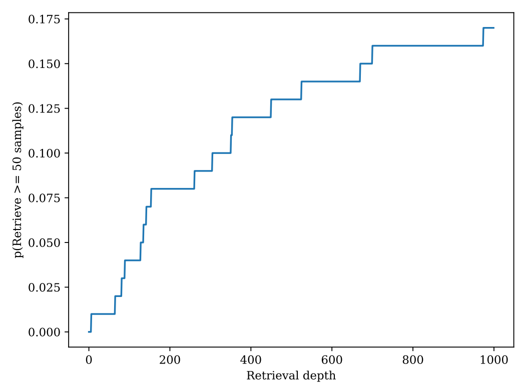
	Moderate + High Exposure		High Exposure	
	Simulation Based	Rubric Based	Simulation Based	Rubric Based
R^2	0.695	0.694	0.071	0.520
Top correlates				
Writing	$+0.72$	$+0.73$	$+0.18$	$+0.28$
Reading Comprehension	$+0.71$	$+0.74$	$+0.18$	$+0.32$
Critical Thinking	$+0.63$	$+0.59$	$+0.14$	$+0.16$
Speaking	$+0.62$	$+0.62$	$+0.17$	$+0.19$
Active Learning	$+0.63$	$+0.59$	$+0.17$	$+0.15$
Active Listening	$+0.60$	$+0.63$	$+0.13$	$+0.22$
Programming	$+0.54$	$+0.55$	$+0.07$	$+0.52$
Learning Strategies	$+0.52$	$+0.49$	$+0.19$	$+0.03$
Mathematics	$+0.50$	$+0.50$	$+0.07$	$+0.30$
Monitoring	$+0.33$	$+0.34$	$+0.11$	-0.12
Science	$+0.25$	$+0.30$	-0.01	-0.06

D.3 Barriers to Entry

Following Eloundou et al. [2024], we ask whether predicted exposure varies systematically with the preparation an occupation requires. To operationalize preparation, we use O*NET's Job Zone classification, which assigns each occupation to one of five tiers based on the education, prior experience, and on-the-job training typically needed to enter the role. Job Zone 1–2 occupations require minimal preparation (under one year), while Job Zone 5 occupations require four or more years. Plotting Simulation-Based and Rubric-Based exposure measures against Job Zone reveals the same pattern for both exposure measures: exposure rises from Job Zone 1-2 through Job Zone 4 and then plateaus or declines at Job Zone 5. Figure 15 illustrates this result.



(a) Number of relevant samples retrieved during the shallow-retrieval versus the full run.



(b) Number of relevant samples retrieved versus retrieval depth.

Figure 7: **Results from analyzing the coverage of the retrieval pipeline.**

Figure 8: **DWA-level synthetic benchmark results.** The top-half correspond to DWAs also found in retrieved data. The bottom-half correspond to DWAs not found in retrieved data.

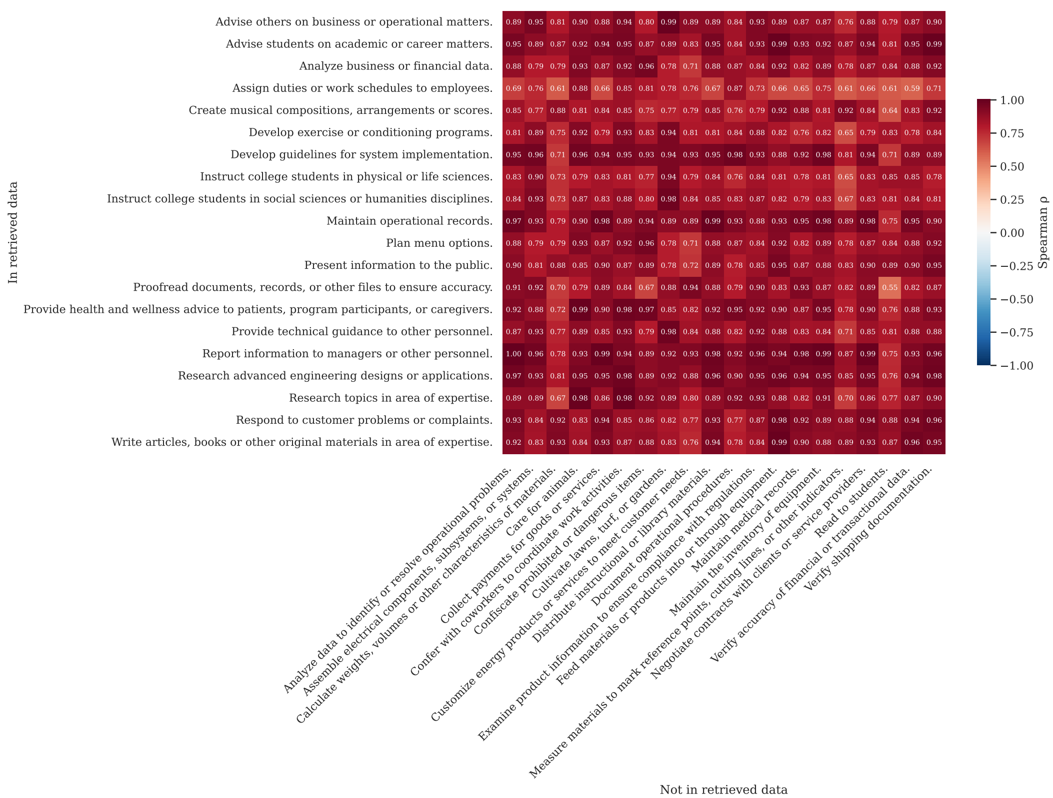


Figure 9: **Spearman correlation between leaderboards where the DWA is found in real usage versus not found in real usage.**

Figure 10: **Variance in DWA-level synthetic benchmark results.** We rank DWAs by the variance they induce in model performance for the 40 DWA-level synthetic benchmarks.

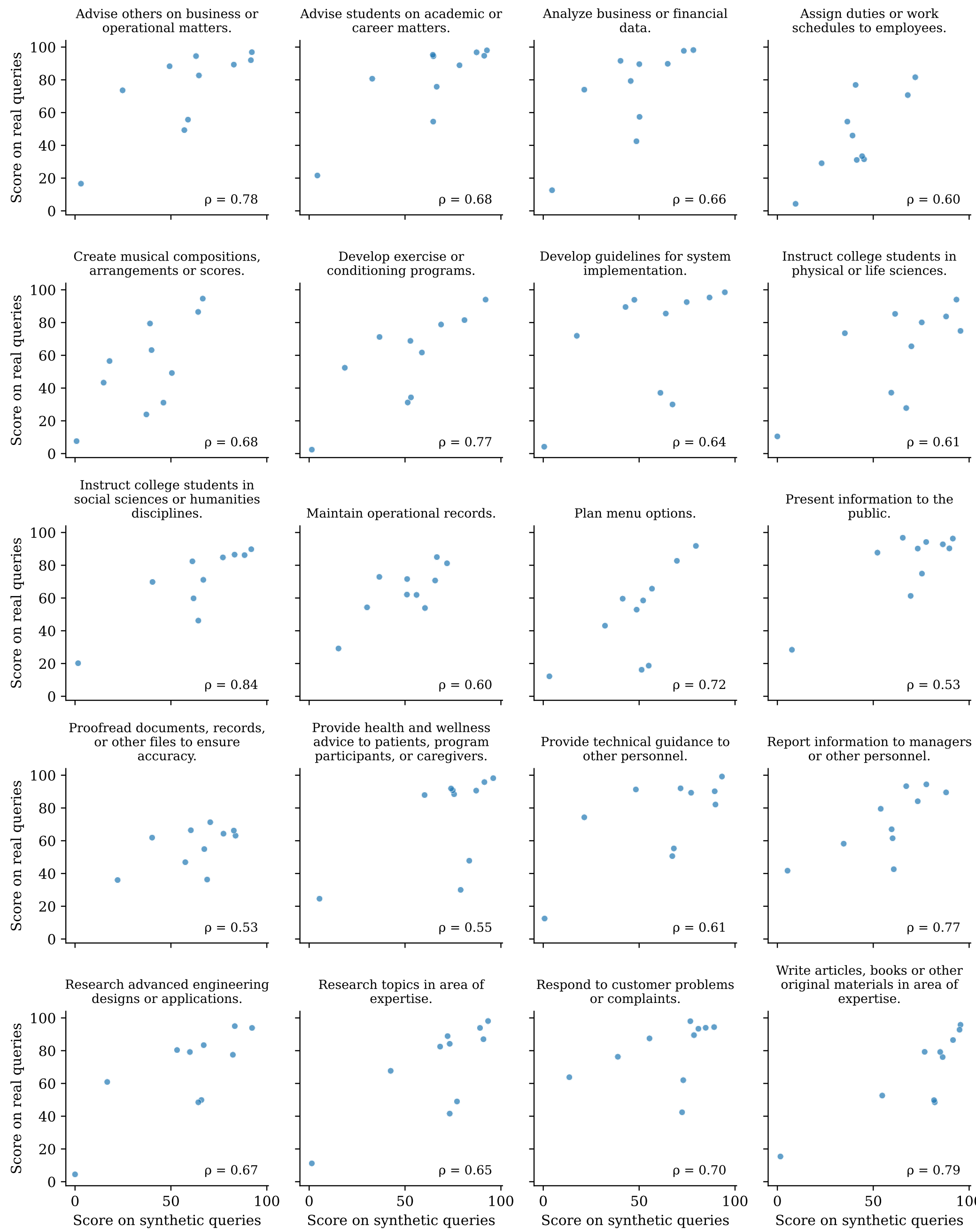


Figure 11: **DWA-level synthetic benchmark results compared to real benchmark results.**

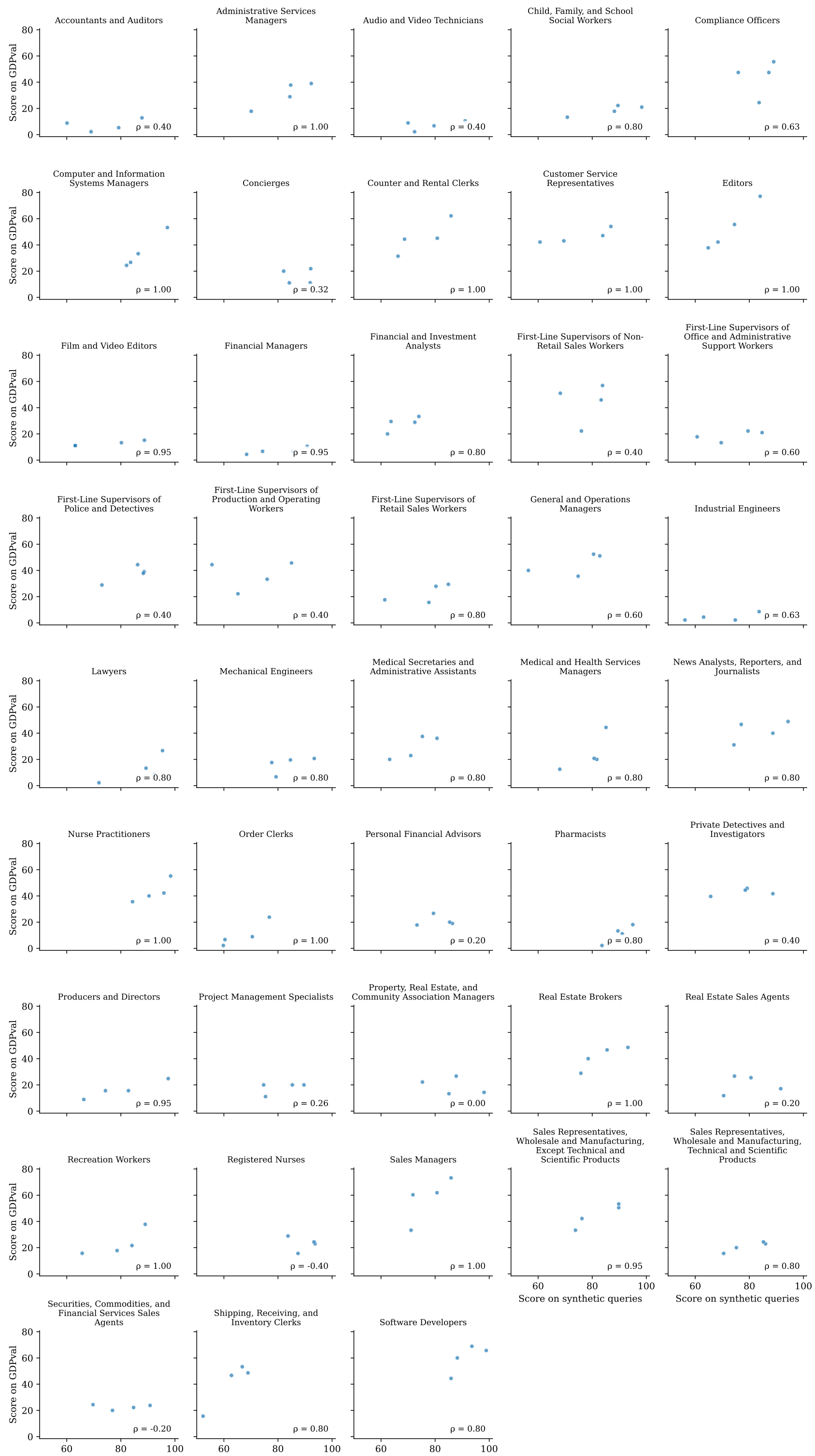


Figure 12: **Occupation-level synthetic benchmark results compared to GDPval benchmark results.**

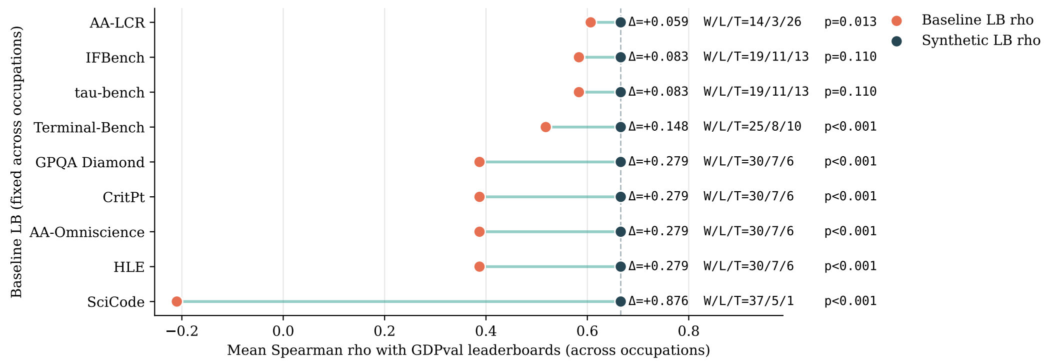
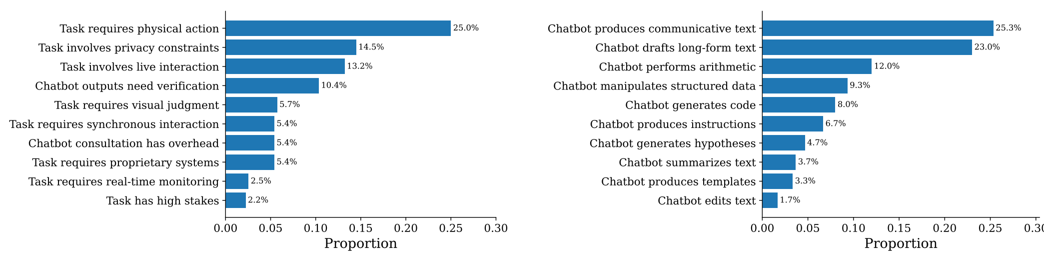


Figure 13: **Correlation between synthetic benchmark results and GDPval benchmark results versus generic capabilities benchmarks.**



(a) Bottlenecks driving predicted time savings below 25% for tasks that have non-zero Claude usage.

(b) Factors driving predicted time savings above 25% for tasks that have minimal Claude usage.

Figure 14: **Why the simulation-based exposure measure predicts LMs will or will not save substantial time on a task, but Claude usage data suggests the opposite.**

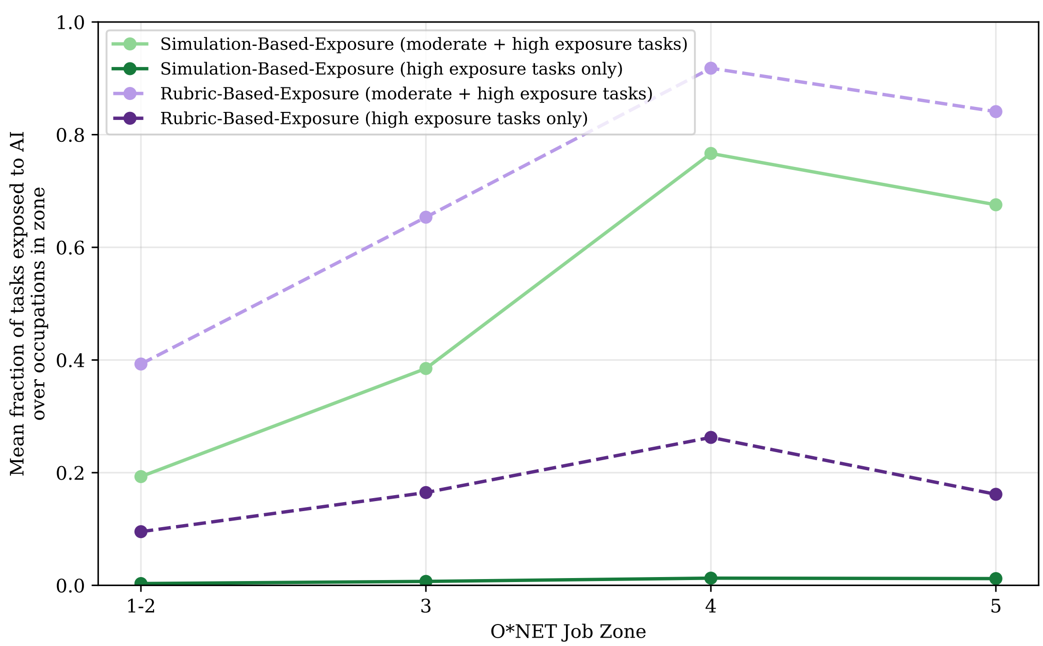


Figure 15: **Predicted AI exposure by O*NET Job Zone, under Simulation-Based and Rubric-Based Exposure measures.** Job Zones group occupations by the education, experience, and training required for entry, ranging from Zone 1–2 (under one year of preparation) to Zone 5 (4+ years). Both measures exhibit the same pattern: exposure rises from Job Zone 1 through Job Zone 4 and then plateaus or declines at Job Zone 5.