GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks

Anonymous authors

Paper under double-blind review

ABSTRACT

We introduce GDPval, an evaluation assessing AI model capabilities on real-world economically valuable tasks. GDPval covers the majority of U.S. Bureau of Labor Statistics Work Activities for 44 occupations across the top 9 sectors contributing to U.S. GDP (Gross Domestic Product). Tasks are constructed from the representative work of industry professionals with an average of 14 years of experience. We find that frontier model performance on GDPval has been improving roughly linearly over time, and that the current best frontier models are approaching industry experts in deliverable quality. We analyze the potential for frontier models, when paired with human oversight, to perform GDPval tasks cheaper and faster than unaided expert workflows. We also demonstrate that increased reasoning effort, increased task context, and increased scaffolding improves model performance on GDPval. Finally, we open-source a gold subset of 220 tasks and provide a public automated grading service to facilitate future research in understanding real-world model capabilities.

1 Introduction

There is growing debate about how increasingly capable AI models could affect the labor market—whether by automating specific tasks, replacing entire occupations, or creating entirely new kinds of work (Brynjolfsson et al., 2025; Chen et al., 2025). Current approaches to measure the economic impact of AI focus on indicators such as adoption rates, usage patterns, and GDP growth attributed to AI (Chatterji et al., 2025; Tamkin et al., 2024; Appel et al., 2025; Acemoglu, 2025; Bick et al., 2024). However, historical evidence from technological shifts—such as electricity, airplanes, and computers—shows that the transition from invention to economy-wide permeation often takes years or even decades, requiring regulatory, cultural, and procedural changes (David, 1990; Brynjolfsson & Hitt, 2000; Brynjolfsson et al., 2019; Dwivedi et al., 2021; Solow, 1987). Therefore, while informative when available, these methods are lagging indicators of AI impacts. We consider an alternate method for understanding the potential economic impacts of AI: directly measuring AI model capabilities. AI capability evaluations can provide clearer, more directly attributable evidence about model abilities, allowing us to assess economic relevance ahead of widespread adoption.

Our paper introduces the first version of GDPval, a benchmark evaluating AI model performance on real-world economically valuable tasks. GDPval covers the top 9 sectors contributing to U.S. GDP (Gross Domestic Product), with at least 30 tasks per occupation in the full set (and 5 tasks per occupation in the gold set), across 44 occupations. Each task is constructed based on actual work product created by an expert professional. Given the complexity of automatically grading these tasks, our primary evaluation metric is head-to-head human expert comparison. We also provide an experimental automated grader service for the 220 open-sourced gold subset of tasks. Future GDP-val iterations will progressively incorporate greater breadth, realism, interactivity, and contextual nuance.

The initial version of GDPval offers several advantages over existing AI model evaluations:

• Realism: Unlike AI benchmarks in the style of an academic test that focus on reasoning difficulty (e.g., Phan et al. (2025); Hendrycks et al. (2020); Rein et al. (2023); Liu et al. (2023)), tasks are based on actual work product from industry experts, validated through multiple rounds and review, and tied to time and cost required for completion.

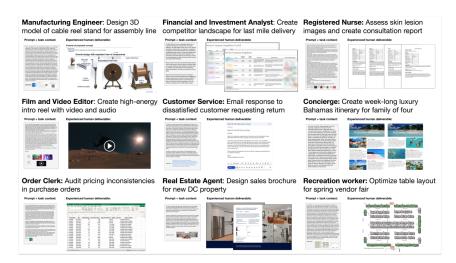


Figure 1: Example GDPval tasks from full set

- Representative breadth: Unlike AI evaluations focused on specific domains like software engineering (e.g., Miserendino et al. (2025)), the GDPval full set covers 1,320 tasks across 44 occupations, sourced to cover the majority of Work Activities tracked by O*NET for each occupation? This top-down approach allows for representativeness of tasks across occupations. We also build on production AI usage analyses (e.g., Tamkin et al. (2024); Chatterji et al. (2025); Appel et al. (2025)) to cover areas where model adoption is still emerging.
- Computer use and multi-modality: Tasks require manipulating a variety of formats (e.g., CAD design files, photos, video, audio, social media posts, diagrams, slide decks, spreadsheets, and customer support conversations). Each task also requires parsing through up to 17 reference files in the gold subset, and 38 in the full set.
- Subjectivity: Tasks assess subjective expert preferences, and are difficult to grade (the average grading time for a human expert is 109 minutes). In addition to correctness, expert graders often consider factors such as structure, style, format, aesthetics, and relevance. Our dataset also therefore serves as a helpful testbed to assess automated grader performance.
- No "upper limit": Unlike metrics that could saturate quickly, our primary metric is winrate, which allows for continuous evaluation. Currently, we compare model outputs against a human expert baseline, but we could replace our baseline with increasingly strong models over time and keep evaluating.
- Long-horizon difficulty: Tasks require an average of 7 hours of work, about a full work day, for an expert professional to complete. On the high end, tasks span up to multiple weeks of work.

2 TASK CREATION

We first identify the sectors that contribute most to U.S. GDP, then source tasks drawn from the highest-earning knowledge work occupations within those sectors.

2.1 PRIORITIZING OCCUPATIONS

GDPval covers tasks from 9 sectors and 44 occupations that collectively earn \$3T annually. We detail below the methodology behind our initial version.

To choose the initial occupations, we:

1. **Selected sectors that contribute over 5% to US GDP** as determined by Q2 2024 Value Added by Industry as a Percentage of Gross Domestic Product (see Federal Reserve Bank of St. Louis (2025)). These 9 sectors are shown in Table 1.

2. Selected the 5 occupations ¹ within each sector that contribute most to total wages and compensation and are predominantly digital. We took a task-based approach to determining if an occupation should be classified as "predominantly digital." Specifically, we identified all tasks for an occupation from O*NET, a database of occupational data, definitions and tasks from the U.S. Department of Labor. We prompted GPT-40 to classify each task as digital or non-digital, and then classified the overall occupation as digital if at least 60% of its component tasks were digital. To calculate this percentage, we weighted tasks by the "relevance," "importance," and "frequency" scores for each task reported in O*NET Task Ratings.

We further validated the representativeness of our digital tasks measure by benchmarking it against the Acemoglu & Autor (2011) task content framework. The correlations we observe—digital tasks increasing with non-routine cognitive content and decreasing with routine and manual content—demonstrate alignment with established economic measures of work, as per appendix A.7.1.

For wage and occupation data, we used O*NET's May 2024 national employment and wage estimates to calculate total wages for 831 occupations (U.S. Bureau of Labor Statistics (2025b)) and further detailed in appendix A.7.



Figure 2: GDPval includes real-world work from 44 occupations.

2.2 EXPERT RECRUITMENT

We recruited expert industry professionals to create realistic tasks based on their professional work experience. Experts were required to have a minimum of 4 years of professional experience in their occupation and a strong resume with a demonstrated history of professional recognition, promotion, and management responsibilities. The average expert had 14 years of experience. We further required experts to pass a video interview, a background check, a training and a quiz to participate in the project. Experts were well compensated for their time and experience. Some of the prior employers of our industry experts include: Accenture, Aetna, Apple, AXA Advisors, Bank of America, Barclays, BBC News, Boeing, Budget Rent a Car, Capital One, Centers for Disease Control and Prevention, Citigroup, Condé Nast, CVS Pharmacy, U.S. Department of Defense, Disney, Douglas Elliman, E*TRADE, Federal Trade Commission, General Electric, Goldman Sachs, Google, Guggenheim Partners, HBO, IBM, JPMorgan Chase, Johnson & Johnson, Kmart, Kirk-

¹We assigned occupations to sectors by using the 2023 BLS National Employment Matrix from U.S. Bureau of Labor Statistics (2025a) to map occupations to sectors by identifying the sector with the highest employment for each occupation. For more detail, see Appendix.

land & Ellis LLP, LinkedIn, Lockheed Martin, Macy's, Massachusetts General Hospital, Meta, Microsoft, Morgan Stanley, National Park Service, NFL Network, Oracle, Paul, Weiss, Rifkind, Wharton & Garrison LLP, Prudential, PwC, Raytheon, Sally Beauty, Samsung, SAP, Scientific American, Sotheby's, Telegraph Media Group, Thermo Fisher Scientific, TIME, Twilio, U.S. Department of Justice, United States Air Force, United States Postal Service, Walgreens, Wells Fargo, White & Case LLP, and Whole Foods.

2.3 TASK CREATION

Each GDPval task consists of two primary components: a request (often with reference files) and a deliverable (work product). Experts classified their requests against O*NET occupational tasks for their occupation to ensure broad and representative coverage across tasks (U.S. Bureau of Labor Statistics, 2025a). More details on task characteristics can be found in appendix A.4.

2.4 TASK QUALITY CONTROL PIPELINE

All 1320 tasks in the full set went through an iterative review pipeline involving both automated model-based screening and multiple stages of human expert review. Each task received an average of five human reviews (with a minimum of three reviews). Across all stages of review, experts provided detailed comments, and tasks were iteratively revised before subsequent reviews to enhance quality and representativeness, as detailed in appendix A.5.

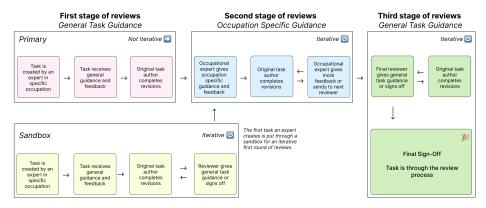


Figure 3: Tasks undergo multiple rounds of review to ensure realism and quality.

2.5 Human Expert Grading and Automated Grading

To grade the 220 open-sourced gold subset, we conducted blinded expert pairwise comparisons, where experts in the relevant occupation were presented with a request and reference files and asked to rank two or more unlabeled work deliverables.

On average, grading each comparison for the gold subset took over an hour. Additional occupational experts were sourced to grade human and model deliverables. Experts provided detailed justifications for their choices and rankings, which enabled us to compute our headline win-rates for various models compared to the human expert completion.

For the gold subset, we trained an experimental grading model to perform pairwise comparisons in the style of industry professional experts. Although significantly limited, it is faster and cheaper than expert grading, and achieves 66% agreement with human expert graders, only 5% below human expert

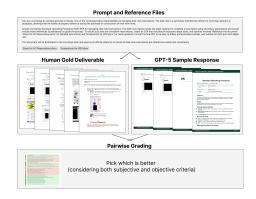


Figure 4: Pairwise grading setup.

inter-rating agreement of 71%. Further detail in appendix A.6.

3 EXPERIMENTS AND RESULTS

3.1 HEADLINE RESULTS

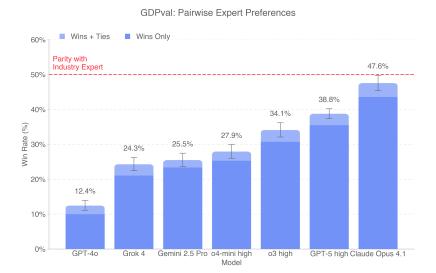
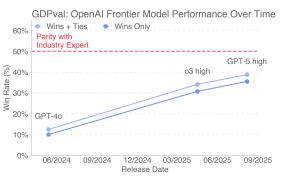
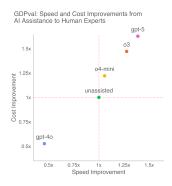


Figure 6: On human pairwise comparisons, models are beginning to approach parity with industry experts.





(a) Performance of OpenAI frontier models increased roughly linearly over time.

(b) In the scenarios we analyze, models show the potential to save time and money by coupling AI assistance with expert human oversight.

Figure 7: Left: Model performance improves over time. Right: AI Assistance can save time and money.

We evaluated GPT-40, o4-mini, o3, GPT-5, Claude Opus 4.1, Gemini 2.5 Pro, and Grok 4 using blind comparisons by professional industry experts ².

Claude Opus 4.1 was the best performing model on the GDPval gold set, excelling in particular

²We aimed to keep comparisons as blind as possible, but model samples may still have been identifiable due to stylistic differences. OpenAI outputs often used em dashes, Claude outputs frequently adopted first-person phrasing, and Grok occasionally referred to itself as Grok. Although filenames were scrubbed of model identifiers, to preserve sample identity, we did not alter style or content, so experts may still have been able to infer model origins. We sampled Grok, Gemini, and Claude via the UI to enable the maximum GDPval-relevant features. For example, for Claude, we wanted to evaluate its 'Upgraded file creation and analysis' feature (https://www.anthropic.com/news/create-files). For plots shown, we sampled each model 3 times for

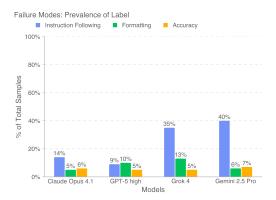


Figure 8: Across models, experts most often preferred the human deliverable because models failed to follow instructions.

on aesthetics (e.g., document formatting, slide layout), while GPT-5 excelled in particular on accuracy (e.g., carefully following instructions, performing correct calculations). fig. 7a shows that on GDPval gold set, 47.6% of deliverables by Claude Opus 4.1 were graded as better than (wins) or as good as (ties) the human deliverable. Model deliverables outperformed or matched expert humans' deliverables in just over half the tasks.

We compared the speed and cost at which models and industry experts completed the 220 gold subset.

We analyzed several scenarios to understand the speed and cost savings ratio of frontier models in appendix A.2.2. In all scenarios analyzed, incorporating frontier AI models into completion workflows shows the potential to save time and money relative to unaided experts. Fig 7 summarizes expected savings under a "review-and-resample" setup, where an expert human samples from a model, reviews outputs, and either uses the output or keeps resampling until they have tried n times. If no satisfactory output is obtained by then, the human completes the task themselves.



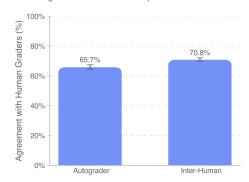


Figure 5: Across all samples on the 220 open-sourced gold subset, automated grader agreement is within 5% of human inter-rater agreement.

3.2 Model strengths and weaknesses

We built a clustering pipeline to analyze why experts preferred or rejected GPT-5 high, Claude Opus 4.1, Gemini 2.5 Pro, and Grok 4 deliverables.⁴ Claude, Grok, and Gemini most often lost due to instruction-following failures, while GPT-5 high lost mainly from formatting errors and had the fewest instruction-following issues. Gemini and Grok frequently promised but failed to provide deliverables, submitted corrupted files, ignored reference data, or used the wrong format. GPT-5 and Grok also had more formatting errors, submitting deliverables with overlapping or missing text, bullets, and charts. By contrast, GPT-5 and Grok showed the fewest accuracy errors, though all models sometimes hallucinated data or miscalculated.

each prompt, and then had 3 different human graders grade each sample (yielding 9 comparisons per prompt, per model, across 220 tasks).

³We were not able to obtain cost estimates for Claude, Gemini, and Grok.

⁴Samples were clustered using expert justifications; labels were mutually exclusive and left blank when the rationale was unclear.

3.3 INCREASING REASONING EFFORT AND SCAFFOLDING

To understand the impact of reasoning effort on model performance, we ran GDPval on the o3 and GPT-5 models at low, medium, and high reasoning effort. We found that additional reasoning effort improved performance.

We were also interested in measuring how easily we could improve model capabilities with prompts. For example, many of the observed GPT-5 failure modes were due to obvious formatting errors. We created a prompt which encouraged GPT-5 to rigorously check deliverables for correctness, check layouts by rendering files as images, avoid nonstandard unicode characters, and avoid excess verbosity. The prompt applies generally to multimodal economic tasks and is not overfit to any given question (see appendix A.3 for details). We also improved agent scaffolding by performing best-of-N sampling with N=4 and a GPT-5 judge.

Prompting fully eliminated black-square artifacts from GPT-5 responses, which previously affected over half of generated PDFs, and reduced egregious formatting errors in PowerPoint files from 86% to 64%. This can be partially attributed to a sharp increase in agents using their multi-modal capabilities to inspect deliverables ($15\% \rightarrow 97\%$). Prompting also improved human preference winrates by 5 percentage points in Figure 9b. These easy performance gains suggest there are paths to agent improvement on GDPval tasks by training or scaffolding them to be more thorough and take full advantage of their multimodal capabilities.

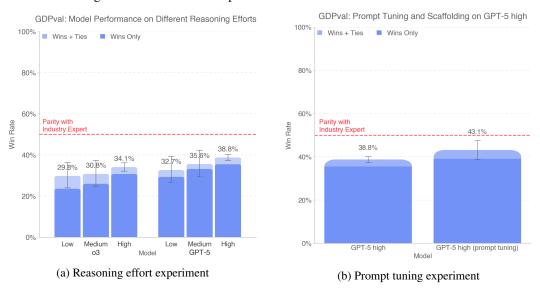


Figure 9: Left: Model performance improves predictably with increasing reasoning effort. Right: Prompt-tuning and scaffolding improvements increase GPT-5 performance.

4 Open-sourcing

We open-source the prompts and reference files in our 220-task gold subset and make a grader publicly available. Please note that the tasks in the open sourced set have been scrubbed of information that could be used to identify the expert who wrote the task. We also note that, as a result of limitations with our automated grader, we don't provide automated grading results for all tasks in the gold subset. Further disclaimers about the open source gold set are in appendix A.1.3.

5 LIMITATIONS

Dataset size: GDPval currently consists of only 44 occupations and 30 total tasks per occupation. It is therefore a limited, initial cut of knowledge work tasks, not a comprehensive evaluation of all possible occupational tasks. We are expanding the dataset size.

Focus on self-contained knowledge work: Tasks in the initial version of GDPval are oriented around knowledge work that can be performed on a computer, particularly around digital deliverables. Manual labor and physical tasks are not included in the current version. Moreover, tasks that involve extensive tacit knowledge, access to personally identifiable information, use of proprietary software tools, or communication between individuals are out of scope for the current evaluation. We aim to build on this in future versions of the evaluation.

Tasks are precisely-specified and one-shot, not interactive: For GDPval, we provide the full context of the task in the prompt, but in real life it often takes effort to figure out the full context of a task and understand what to work on. We are working on a follow-on version of this evaluation that involves more interactivity and contextual realism. In the meantime, the experiment in the "Undercontextualized GDPval" section (appendix A.2.3) demonstrates how performance degrades with less context given to the model up-front.

Grader performance: Our current automated grader is not as good as human expert graders. They have the same limitations as the models we are testing. More details about the automated grader are available in the appendix A.6.2.

Cost: Constructing and running our evaluation is expensive, particularly with industry expert graders. For this reason, we make an automated grader proxy available, but note again that it is not currently good enough to be considered a substitute for industry expert graders.

6 Conclusion

In GDPval, we contribute the following:

- Dataset: We create a new evaluation dataset (GDPval) measuring real-world, economically valuable tasks.
- Capability benchmarking: We analyze quality, speed and cost of deliverables across human industry experts and frontier AI models.
- 3. **Experiments:** We test how results shift with under-contextualized prompts, different reasoning effort, varying model size, and additional prompting.
- 4. **Targeted training:** We fine-tune GPT-5 on GDPval data and measure clear improvements in human win-rate, showing how targeted training can boost performance.
- 5. **Automated grader:** We release an automated grader to improve accessibility of grading, with the best performing grader being a pairwise automated grader that considers both subjective and objective scoring criteria.
- 6. **Open-sourcing:** We open-source 220 tasks as part of our gold subset which includes prompts and reference files, and make our automated grader available.

We hope this work contributes to the science of tracking model progress, so that we have better data to assess the social impacts of AI models.

IMPACT STATEMENT

GDPval aims to measure model capabilities on realistic, economically relevant knowledge-work tasks. Reliable capability measurement can help guide responsible model and policy development by showing where current systems are helpful, where they could improve, and how progress changes over time. It can also help inform workers about where models are reliable today, enabling more effective use in their own workflows.

At the same time, results from GDPval could be misinterpreted to make premature claims about job replacement or used to justify personnel decisions. Capability does not imply adoption or net productivity gains. Our study emphasizes these limitations, relies on expert human preferences as ground truth, and reports failure modes alongside strengths.

We intend GDPval to inform—not determine—policy and model development choices. Future iterations will expand occupational coverage, increase task interactivity, improve multi-modal and tool use, and strengthen automated grading calibrated to expert judgment. By grounding discussion of economic relevance in transparent, empirical measurement, we hope to support careful, evidence-based progress while highlighting areas that require human oversight and further research.

6.0.1 ETHICS STATEMENT

Human participants. This work relies on tasks and judgments from industry experts ("experts," "reviewers," and "graders"). All participants were told the purpose of the project, the types of data being collected, and how their data would be used for research. Participation was voluntary, with the option to withdraw at any time. Participants were compensated for their time and expertise; compensation was not contingent on model outcomes.

Privacy, confidentiality, and data governance. We instructed contributors not to submit personally identifiable information (PII), material non-public information (MNPI), and confidential or proprietary content in any submissions. Reviewers flagged overlooked information, which was subsequently replaced with realistic, fictional data by contributors. We do not release raw deliverables that could reveal sensitive information. O*NET/OEWS and BLS data were used in accordance with their terms.

Potential harms and misuse. GDPval measures model capabilities on economically relevant tasks. Results could be misused to make premature claims about job replacement or to target specific roles for automation without context. To mitigate this, we emphasize limitations (capability versus adoption; subjectivity; file and tool constraints) and report results with confidence intervals. These results are research findings and are not intended to inform or justify personnel decisions (e.g., hiring, firing, promotion, or compensation). The benchmark excludes dangerous or clearly harmful task types and was executed in sandboxed environments with restricted internet access.

REFERENCES

- Daron Acemoglu. The simple macroeconomics of ai. Economic Policy, 40(121):13-58, 2025.
- Daron Acemoglu and David Autor. Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics*, volume 4, pp. 1043–1171. Elsevier, 2011.
- Ruth Appel, Peter McCrory, Alex Tamkin, Michael Stern, Miles McCain, and Tyler Neylon. Anthropic economic index report: Uneven geographic and enterprise ai adoption. *Anthropic Research*, 2025. URL https://www.anthropic.com/research/anthropic-economic-index-september-2025-report.
- Alexander Bick, Adam Blandin, and David J Deming. The rapid adoption of generative ai. Technical report, National Bureau of Economic Research, 2024.
- Erik Brynjolfsson and Lorin M. Hitt. Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives*, 14(4):23–48, 2000. doi: 10.1257/jep.14.4.23. URL https://www.aeaweb.org/articles?id=10.1257/jep.14.4.23.
- Erik Brynjolfsson, Daniel Rock, and Chad Syverson. *Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics.* University of Chicago Press, 2019.
- Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative ai at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025.
- Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. *NBER Working Paper*, (34255), September 2025. doi: 10.3386/w34255. URL https://www.nber.org/papers/w34255.
 - Wilbur Xinyuan Chen, Suraj Srinivasan, and Saleh Zakerinia. Displacement or complementarity?: The labor market impact of generative ai. 2025.

Paul A. David. The dynamo and the computer: An historical perspective on the modern productivity paradox. *American Economic Review*, 80(2):355-361, 1990. URL https://econpapers.repec.org/RePEc:aea:aecrev:v:80:y:1990:i:2:p:355-61.

- Yogesh K Dwivedi, Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, et al. Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International journal of information management*, 57:101994, 2021.
- Federal Reserve Bank of St. Louis. Value added by industry as a percentage of gross domestic product. FRED Release Tables, 2025. URL https://fred.stlouisfed.org/release/tables?rid=331&eid=211. Accessed: 2025-09-03.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. doi: 10.48550/arXiv.2009.03300. URL https://arxiv.org/abs/2009.03300.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating Ilms as agents. *arXiv preprint arXiv:2308.03688*, 2023. doi: 10.48550/arXiv.2308.03688.
- Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. Swe-lancer: Can frontier LLMs earn \$1 million from real-world freelance software engineering? *arXiv preprint arXiv:2502.12115*, 2025.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations, 2024. URL https://arxiv.org/abs/2404.13076.
- Long Phan et al. Humanity's last exam. arXiv preprint arXiv:2501.14249, 2025.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. arXiv preprint arXiv:2311.12022, 2023. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022.
- Robert M. Solow. We'd better watch out. New York Times Book Review, pp. 36, July 1987. URL https://www.standupeconomist.com/pdf/misc/solow-computer-productivity.pdf.
- Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, Jared Kaplan, and Deep Ganguli. Clio: Privacy-preserving insights into real-world ai use. arXiv preprint arXiv:2412.13678, 2024. doi: 10.48550/arXiv.2412.13678. URL https://arxiv.org/abs/2412.13678.
- U.S. Bureau of Labor Statistics. Occupational outlook occupational data. https://www.bls.gov/emp/data/occupational-data.htm, 2025a. Accessed: 2025-09-03.
- U.S. Bureau of Labor Statistics. Occupational employment and wage statistics: May 2024 national tables. https://www.bls.gov/oes/tables.htm, 2025b. Data reference May 2024; accessed: 2025-09-03.

A APPENDIX

A.1 DISCLOSURES

A.1.1 AI DISCLOSURE

We used AI models to help with some of our literature review and with tweaking language in the paper. We also used AI coding assistants as part of our regular engineering workflows (e.g., to help find and fix bugs).

A.1.2 Sensitive content and political content disclosure

Some tasks in GDPval include NSFW content, including themes such as sex, alcohol, and vulgar language, or political content. We chose to keep these tasks as they reflect real themes addressed in various occupations (e.g., film, literature, law, politics). We do not endorse the particular actions or views in any of the content.

A.1.3 THIRD-PARTY REFERENCES DISCLOSURE

GDPval contains limited references to third-party brands and trademarks solely for research and evaluation purposes. No affiliation or endorsement is intended or implied. All trademarks are the property of their respective owners. Some images and videos in this dataset feature AI-generated individuals and real people who have provided permission. Names and identifying references to private individuals in GDPval are fictitious. Any resemblance to actual persons or entities is purely coincidental.

A.2 ADDITIONAL DETAIL ON EXPERIMENTAL RESULTS

A.2.1 WIN RATES BY SECTOR AND OCCUPATION

We include the below plots with more detail on win rates by sector and occupation.

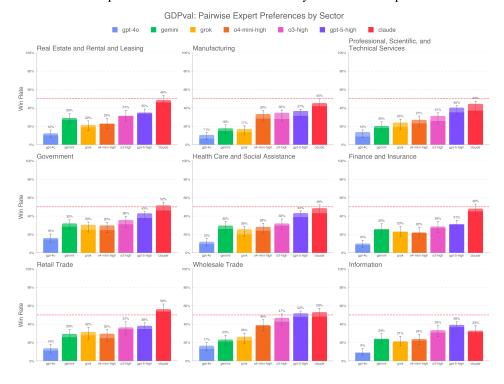


Figure 10: Win rate by sector [updates incoming]

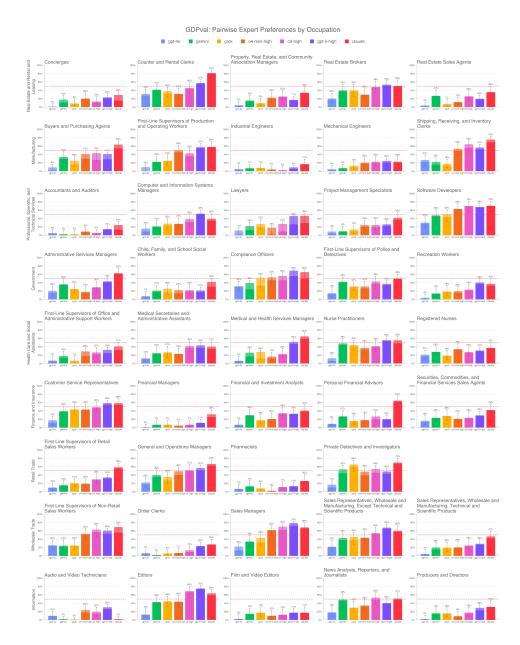


Figure 11: Winrate by occupation

A.2.2 SPEED AND COST ANALYSIS, CONTINUED

We use the following definitions:

1. Human expert professional completion time H_T is the time taken by a human expert professional to complete a task, based on validated self-reported time to complete⁵. To calculate human expert professional completion cost H_C , we multiplied the reported task completion hours per occupation by the median hourly wage for each occupation from the U.S.

⁵During submission, experts self-reported the real-world time required to complete each task. Multiple occupational reviewers independently validated these times, correcting errors. Because times were self-reported, it is possible that experts under-estimated or over-estimated time taken

Bureau of Labor Statistics (2025b)⁶. On average, on our 220 gold set $H_T=404$ minutes and $H_C=\$361$.

- 2. Human expert professional review time R_T is an estimate of the time taken to assess a model deliverable by a human expert grader. We observe this from our our task monitoring software, averaging the time taken to grade for the first time each human expert was asked to grade that question. On average, $R_T=109$ minutes, and associated human expert professional review cost R_C is on average \$86, where R_C is again calculated based on time taken multiplied by median wage data.
- 3. Model completion time M_T is the time taken for the model to complete a deliverable and M_C is the associated completion cost, based on empirical API speed and cost for the model to complete the deliverable when given a prompt 7 . For these analyses, we focus on the GPT-5 model, as it was the best-performing model that we were able to pull speed and cost data for. On average, $M_T = 4.5$ minutes and $M_C = \$0.75$,
- 4. Model win rate w is how often the model deliverable is rated better than the human deliverable by the human expert grader, was on average 40.6%.

We then calculate the following ratios:

- 1. Naive ratio: To measure the ratio of human deliverable versus model deliverable, without accounting for any quality differences or implementation times, we simply divide the average task completion time for a human by the average sampling time for a model: H_T/M_T , and analogously for cost: H_C/M_C .
- 2. Try 1 time, then fix it ratio: To calculate the time with this method, we take the sampling time for the model, add review time R_T for an expert to assess quality, and then with probability $(1 w_i)$ add in the human completion time for any fixes needed for that model for a task i, to obtain $T_{1,i}$ and analogously $C_{1,i}$:

$$\mathbb{E}[T_{1,i}] = M_{T,i} + R_{T,i} + (1 - w_i)H_{T,i} \tag{1}$$

$$\mathbb{E}[C_{1,i}] = M_{C,i} + R_{C,i} + (1 - w_i)H_{C,i} \tag{2}$$

The average time spent is $T_1=\mathbb{E}[T_{1,i}]$, marginalizing over all tasks i, similarly with C_1 . This proxies the setup where a human tries using GPT-5 for a task, assesses its quality, and then does the task themselves if the deliverable quality is below their quality bar. Our plug-in estimate of the time savings ratio is: $H_T/(M_T+R_T+(1-w)H_T)=H_T/\hat{T}_1$, where we use the empirical mean \hat{T}_1 . The analogous cost ratio is $H_C/(M_C+R_C+(1-w)H_C)$. We do not include consideration of the time taken to review a human professional deliverable, although this would commonly occur for tasks in GDPval (either self-review of the professional's own work or review by a supervisor of a team member's work). We also do not include the possibility that the human deliverable is also undesirable.

3. Try n times, then fix it ratio: To calculate the time with this method, we take the sampling time for the model, add review time R_T for an expert to assess quality, and then add in the human completion time for any fixes needed for that model (based on $1 - w_i$) 8. We repeat this across n resamples and re-assess steps before the human steps in to fix it:

⁶Because our experts were recruited specifically for being highly experienced in their field, these wage estimates likely underestimate their true market cost.

⁷For each task, we collected three API completions per model and averaged the observed response times recorded in the API metadata. We also recorded the average invoiced cost per task.

⁸We are over-penalizing the model here, because the win rate after each completion likely goes up (because the professional will adjust the prompt to the model to fix the errors) and the review time also goes down as the professional gets more comfortable with the task.

$$\mathbb{E}[T_{n,i}] = \sum_{k=1}^{n} ((1 - w_i)^{k-1} (M_{T,i} + R_{T,i})) + (1 - w_i)^n H_{T,i}$$
 (3)

$$= (M_{T,i} + R_{T,i}) \frac{1 - (1 - w_i)^n}{w_i} + (1 - w_i)^n H_{T,i}$$
(4)

$$\mathbb{E}[C_{n,i}] = \sum_{k=1}^{n} ((1 - w_i)^{k-1} (M_{C,i} + R_{C,i})) + (1 - w_i)^n H_{C,i}$$
 (5)

$$= (M_{C,i} + R_{C,i}) \frac{1 - (1 - w_i)^n}{w_i} + (1 - w_i)^n H_{C,i}$$
(6)

This proxies the setup where a human tries n rounds of using GPT-5 for a task, then assesses its quality each time, and then does the task themselves if the model quality is below their quality bar after all attempts. As before, the average time spent is $T_n = \mathbb{E}[T_{n,i}]$, marginalizing over all tasks i, similarly with C_n . Therefore, as $n \to \infty$, with w > 0, the time savings are $H_T/((M_T+R_T)/w)$ times faster and cost savings are $H_C/((M_C+R_C)/w)$ times cheaper than human experts.

Table 3: Speed and cost improvements under different review strategies.

		Speed improvement			Cost	improve	ment
Model	Win rate	Naive	Try 1x	Try nx	Naive	Try 1x	Fix nx
gpt-4o	12.5%	327x	0.87x	0.46x	5172x	0.90x	0.53x
o4-mini	29.1%	186x	1.02x	1.06x	1265x	1.06x	1.22x
03	35.2%	161x	1.08x	1.28x	480x	1.13x	1.47x
gpt-5	39.0%	90x	1.12x	1.39x	474x	1.18x	1.63x

When incorporating time to review and redo work, the payoff from using a model shrinks. One further limitation of this analysis is that it does not capture the cost of mistakes, which can be disproportionately expensive in some domains.

UNDER-CONTEXTUALIZED GDPVAL

To assess how models handle task ambiguity, we created a modified version of GDPval with deliberately lower-context prompts. These shorter prompts omitted additional context such as where to locate specific data within reference files, how to approach the problem, or detailed formatting expectations for the final deliverable; the models had to "figure it out." On average, these revised prompts were 42% the length (by token count) of the original prompts.

This setting helped measure an aspect of professional knowledge work previously unaddressed in our evaluation: navigating ambiguity by figuring out what to work on and where to get the necessary inputs. We collected and graded GPT-5 completions with expert human graders and found the model's performance was worse on under-specified prompts. In particular, the models struggled to figure out context.

As a note: this experiment was run on an earlier version of the GDPval gold set, and therefore the observed winrates do not match those in the main text of the paper.

A.2.4 ADDITIONAL DETAIL ON MODEL STRENGTHS AND WEAKNESSES

Additional issues flagged in our analysis:

- Poor instruction-following: Models sometimes failed to follow complex instructions or requirements.
- Verbosity: Human experts occasionally found model deliverables overly detailed or verbose.

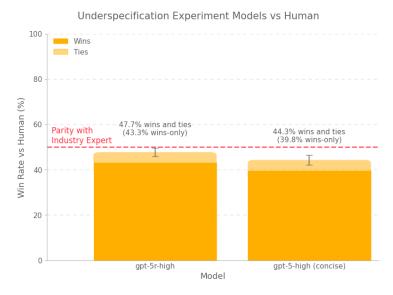


Figure 12: On the underspecified version of GDPval, GPT-5 performed worse as it struggled to figure out requisite context.

- Lack of industry-specific tacit knowledge: Models sometimes missed implicit industry
 norms on how to present information (e.g., failing to include charts in vendor comparisons)
 or produced deliverables that, while technically correct, lacked context obvious to domain
 professionals, or omitted details that domain professionals wanted to know.
- Hallucinations and factual inaccuracies: Some failures involved hallucinations like incorrect distances or non-existent contractual obligations. In Excel deliverables, models sometimes hard-coded values rather than using formulas or computed incorrect results.
- Limitations due to restricted internet access: Certain tasks required external resources such as Python packages or external images. Due to sandbox restrictions, models failed when they were unable to access these external resources.

We also see improvements over time from GPT-40 through to GPT-5:

- **Intelligence:** Higher accuracy, broader knowledge, improved tool use (e.g., browsing, code).
- **Reliability:** GPT-4o's main failure—failing to produce the requested deliverable—accounted for 43% of its losses. This rate has dropped across model generations, and is responsible for 20% of cases where o4-mini-high lost, 11% in o3-high, and just 8% of GPT-5 high losses.
- Structure and formatting: Better at text organization, better handling of spreadsheets and slide decks.
- **Comprehensiveness:** More thorough research, sharper instruction-following, cleaner rewrites, and greater attention to typos and copy errors.

A.2.5 ADDITIONAL DETAIL ON MODEL FAILURES ANALYSIS

We took the subset of GPT-5 model failures (tasks where the GPT-5 deliverable lost to the human expert), and then we asked other expert occupational graders to rate these subset samples as:

- Catastrophic: The model completion would be catastrophic if used in real life because it is harmful or dangerously wrong (e.g., insulting a customer, giving the wrong diagnosis, recommending fraud, or suggesting actions that will cause physical harm).
- **Bad**: The completion is bad and not fit for use, but not offensive or dangerous (e.g., rambling nonsense, completely irrelevant, or incoherent answers).

• Acceptable but subpar: The completion is acceptable (and could be used) but the human produced a stronger response (e.g., model response lacked helpful detail compared to the human).

• N/A: Disagree with original raters; the model completion was better than the human completion.



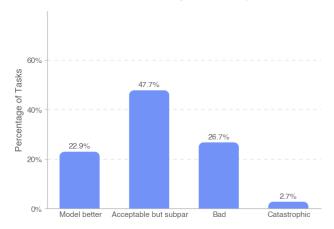


Figure 13: When trainers reviewed GPT-5 failures, they categorized them as catastrophic, bad, acceptable but subpar, or human was better.

A.3 ADDITIONAL DETAIL ON PROMPT-TUNING

Here is the prompt we give the agent to elicit capabilities (lightly edited to remove some specific details of our scaffolding setup).

Special characters - Never use the character - (U+2011), since it will render poorly on some people's computers. Instead, always use - (U+002D) instead. - Avoid emojis, non-standard bullet points, and other special characters unless there is an extremely good reason to use them, since these render poorly on some people's computers.

Graphics embedded within PDFs/slides - Make sure that any diagrams or plots are large enough to be legible (though not so large that they are ugly or cut off). In most cases they should be at least half the page width. - Plots and charts to visualize data are good. Simple graphics (like a flowchart with arrows) are good. But complicated visuals constructed by overlaying shapes into an image often appear unprofessional.

PDFs - Always use LibreOffice to create the PDF. Other libraries sometimes show weird artifacts on some computers.

Fonts - Always use fonts which are available across all platforms. We recommend Noto Sans / Noto Serif unless there is an extremely good reason to use something else. If you must use another font, embed the font in the pptx/word/etc doc.

Deliverable text - Do not link to submitted files in the deliverable text (links are not supported on the interface where these will be viewed). - Ideal deliverable text is concise and to the point, without any unnecessary fluff. 4 sentences max. - Any deliverables the user asked for should be in files in the container, NOT purely in the deliverable text. - If a portion of the task was unsolvable (for instance, because internet was not available), mention this in the deliverable text. - Your submission should be complete and self-contained. Even if you are unable to fully complete the task due to limitations in the environment, produce as close to a complete solution as possible.

Verbosity Always be clear and comprehensive, but avoid extra verbosity when possible. Filetypes If the prompt does not request a specific filetype, use "standard" filetypes like PDF, PPTX, DOCX, XLSX, MP4, ZIP, etc.

Video files (mp4, mov) Extract a string of images from the video files and check the images to see whether the visual elements are corrupted.

Mandatory formatting checks Before you submit your deliverable, you MUST perform the following mandatory formatting checks. Take your time, do these thoroughly, they are extremely important!

STEP 1: Convert all visual deliverables to PNGs using LibreOffice. This includes pptx, docx, pdf, xlsx, etc. Convert it so that each page or slide is a separate PNG. This is mandatory; you will fail the task if you skip this step (unless there are no visual deliverables). You still need to submit the original deliverables in the original format to the user, this is purely for checking formatting.

STEP 2: Display the PNGs. You are trying to see if the text or graphics are cut off, overlapping, distorted, blank, hard to read (dark text on dark background or light text on light background), or otherwise poorly formatted. Look at each image thoroughly, zoom in if you need to see more closely. Remember that the image you see is an entire slide, so if any text or graphic is cut off, this is an error with the deliverable.

STEP 3: Programmatic formatting checks. For highly visual submissions (e.g. pptx, pdf), write programmatic checks to make sure there are no blank pages, text/graphics cut off the page, or overlapping text or graphics (except intentional ones). Also check that if there is a page or slide limit, it is respected.

STEP 4: Summarize the prompt's deliverable instructions, and match that to the portion of the deliverable that addresses it.

STEP 5: Right before submitting, check that the deliverables you have produced are exactly what you want to submit: deliverables should contain exactly the files you want to submit, with no extra files. Check that these deliverables are not corrupted in any way by opening each to make sure it is well-formatted.

If any of these checks reveal a formatting issue, fix them and go through steps 1-5 again. Take your time, be thorough, remember you can zoom in on details.

This is IMPORTANT and MANDATORY, go through each step one-by-one meticulously! Every formatting error is a MAJOR ISSUE THAT YOU NEED TO FIX! There is no time limit, be thorough, go slide by slide or page by page.

Finally – on the last line of your output text, add CONFIDENCE[XX], where XX is an integer between 0 and 100, inclusive, indicating your confidence that the submission is correct, follows instructions, and is well-formatted.

We performed best-of-N sampling by prompting a GPT-5 grader with the prompt, reference files, and deliverable files for four different submissions, then asking it to pick the best.

921

918

919

A.4 ADDITIONAL TASK CHARACTERISTICS

922 923 924

925 926

A.4.1 FILES AND ATTACHMENTS

927 range of real-world file types (such as spreadsheets, documents, presentations, images, audio, video, 928 and specialized formats like CAD). 67.7% of tasks required interaction with at least one reference file.

929

O*NET TASKS, SKILLS, AND WORK ACTIVITIES A.4.2

931 932 933

934

930

To ensure broad occupational representativeness, we analyzed the O*NET tasks, skills, and general work activities represented by GDPval tasks. The dataset covered 208 unique O*NET tasks, 25 occupational skills, and 26 work activities.

Many traditional evaluations rely on text-in/text-out task formats. GDPval tasks incorporate a broad

935 936 937

Most GDPval tasks involve multiple O*NET tasks, skills, and work activities.

938 939

A.4.3 TASK SPECIFICATION

940 941

Occupational experts conducting human grading rated the specificity of instructions provided in each prompt. 89.07% of tasks were rated as well-specified, indicating the instructions closely matched real-world expectations of clarity and detail.

942 943

A.4.4 TASK REPRESENTATIVENESS

944 945

Professional Services *Qualification:* Technology and intellectual property attorney with partner roles at multiple AmLaw 100 firms in New York and California, and 15+ years of experience advising clients on emerging technologies, advertising, antitrust, and cross-border disputes and transactions.

947 948 949

946

Quote: Legal tasks included details that felt true to practice, like ambiguous fact patterns, disclosure of relevant legal considerations along with non-legal business goals, and realistic reference

951 952 953

950

Healthcare Qualification: Nursing professional with 18+ years of expertise in emergency medicine, renal management, care coordination, and healthcare operations. Skilled in quality assurance, case management, and professional education.

954 955 956

Quote: These tasks captured the complexity of the role, requiring not only a keen ear for the physician's words, but also careful attention to clinical accuracy and professional formatting.

957 958 959

Retail Trade Qualification: Strategic retail executive with 15 years of experience growing prestige and niche beauty brands through national account leadership, \$1B+ P&L ownership, and datadriven omnichannel strategies.

960 961

Quote: These tasks mirrored the work I performed regularly, including developing revenue forecasts, conducting competitive analysis, building executive-level presentations, and driving strategic initiatives for key retail partners within a global organization.

962 963

Finance Qualification: Fintech and Wall Street leader with 20+ years of experience in wealth management, asset management, and capital markets across global institutions and startups.

964 965 966

Quote: They reflected real-world scenarios that were nuanced and individualized, situations that only someone with years of experience in the field would fully comprehend. The language and details used in the tasks were directly drawn from actual industry practice, making them authentic and grounded in real-world application.

967 968 969

970

Wholesale Trade Qualification: National Accounts Sales Manager for US, China, and Sweden based brands/factories with over 25 years of experience selling to US based retailers.

Quote: All the tasks were in fact based upon real world tasks with back-up reference files and real-world data.

Manufacturing *Qualification:* Lead Industrial Engineer with 5+ years of experience managing large-scale projects and leading teams of 10+ engineers in industrial operations.

Quote: The redesign tasks stood out as especially true to real-world practice because they included specific design components and blocks, along with detailed drawings that incorporated precise measurements. They emphasized practical considerations such as visibility and optimizing walking distances to improve overall productivity, exactly the kind of detail-oriented focus that reflects actual engineering and operational priorities.

Government *Qualification:* Executive leader with 15+ years working at strategic and operational levels in government and non-profit sectors in housing, human service and labor market programs.

Quote: Many of the tasks demand the integration of multiple sources of information, nuanced decision-making, and tailored the work to varied audiences we serve in the workplace.

Real Estate and Leasing *Qualification:* Seasoned commercial real estate broker with 10 years of experience in investment sales, leasing, and managing real estate offices and agents.

Quote: The tasks capture the dynamics and expertise unique to specific sectors and settings.

Information *Qualification:* An experienced senior journalist and content leader with over 20 years in top-tier media, global corporations, and high-growth startups.

Quote: Most importantly, the tasks are anchored in real-world challenges and workplace goals. They push past obstacles, achieve workplace goals, and deliver real-world solutions and products.

Table 8: Reflections from industry experts on task representativeness

Additional Detail about Expert Qualifications Less than 10% of applicants were selected to contribute tasks to our full set. The industry experts also brought occupational diversity, representing different company sizes, locations, and sub-specialties. Each occupation had a minimum of 5 qualified professionals.

Experts for each occupation were required to have previous experience in that specific occupation and sector based on the O*NET occupation definitions (U.S. Bureau of Labor Statistics, 2025a).

A.5 FURTHER DETAIL ON TASK QUALITY CONTROL

A.5.1 MODEL-IN-THE-LOOP TASK REVIEW

We used OpenAI models to automatically screen each task submission across a variety of criteria and flag possible errors or omissions including: ensuring the task is relevant to the selected O*NET occupation, verifying the request involved tasks performed primarily on a computer, flagging if the task complexity was too simple (e.g., if the task seemed like 5 minutes of work instead of a longer-term piece of work), and indicating if there were no deliverable and reference files attached.

Because models can make mistakes, experts were instructed to take model feedback as a suggestion rather than a direction. Experts retained final responsibility for task accuracy and completeness; the model did not autonomously alter tasks.

A.5.2 HUMAN EXPERT REVIEWERS

Human reviewers conducted multiple rounds of review on each task. Reviewers were primarily sourced from the original expert pool based on demonstrated excellence in task creation. Initially, our researchers manually reviewed all tasks to identify experts who produced consistently high-quality tasks; these individuals were trained and promoted to reviewers. The most skilled reviewers were further trained to become lead reviewers, responsible for identifying, mentoring, and promoting additional qualified reviewers from within the expert pool. Throughout the review process, the research team regularly performed quality-control checks on tasks signed off by reviewers, ensuring ongoing alignment and quality standards.

A.5.3 ITERATIVE REVIEW PROCESS

The iterative review process included at least the following 3 stages:

- Generalist initial review: A generalist reviewer confirmed the task adhered to project requirements.
- 2. **Occupation-specific expert review**: An occupation-specific reviewer assessed the representativeness of the task for the occupation, and confirmed that the task was possible for another member of the occupation to complete with the provided context.

3. **Final iterative reviewer feedback loop**: A third expert reviewer provided iterative feedback and worked with experts until the task met our rigorous quality standards.

A.6 AUTOMATED GRADER DETAILS

A.6.1 AUTOMATED GRADER CONSENSUS METRICS

To measure automated grader performance, we measured the agreement rate between scores given by the automated grader vs. human expert graders for the same sample. We also compared grading agreement between human experts who had graded the same sample.

Human-automated grader Agreement. For a given sample s, let the human score H and automated grader score A take values in $\{0,0.5,1\}$, where 1 indicates preference for the model deliverable, 0 indicates preference for the human deliverable, and 0.5 indicates a tie. The agreement between human and automated grader is defined as

$$A_s^{\mathrm{HA}} = \mathbb{E}[1 - |H - A|].$$

The model-level human–automated grader agreement is the mean of ${\cal A}_s^{\rm HA}$ over all samples for that model.

Human Inter-Rater Agreement. For a given sample s, let the human scores H_1 and H_2 take values in $p \in \{0, 0.5, 1\}$. We measure human inter-rater agreement as the following expectation over two randomly sampled human ratings

$$A_s^{\mathrm{HH}} = \mathbb{E} \big[1 - |H_1 - H_2| \big].$$

For a given sample, we estimate this quantity by the empirical mean over all pairs of ratings for that sample. The final human inter-rater agreement for a model is the mean of these sample-level scores over all samples with at least two human graders. Existing grader inter-reliability statistics such as Cohen's kappa, Fleiss' kappa, and Krippendorff's alpha are less directly applicable here, since our graders output ordinal scores in $\{0, 0.5, 1\}$.

A.6.2 AUTOMATED GRADER CORRELATION RESULTS

Over three automated grader sweeps on our dataset⁹, average human-automated grader agreement was 65.9% and human inter-rater agreement was 71.1%. Plots below show 95% confidence intervals obtained by bootstrapping (resampling with replacement the available automated grader scores or human grades for each sample, computing the mean per sample, and averaging across all samples or for the specified model).

Our automated grader, based on GPT-5-high, shows lower correlation with human expert graders when assessing outputs from capable OpenAI models. This aligns with empirical evidence that models often favor their own responses Panickssery et al. (2024). Both agreement metrics are highest for less capable models, since their outputs are easier to distinguish from human deliverables and are less likely to be preferred.

A.6.3 AUTOMATED GRADER LIMITATIONS

In the open-source set we mark 12 out of 220 tasks as ungradable due to limitations of the automated grader.

- 1. **Internet Access:** Tasks which strictly require internet (e.g., tasks that ask agents to find music online and download it) are not possible to grade because the grader does not have internet access.
- 2. **Python**: The automated grader operates in a container that only allows for running Python. Because of this, we excluded 3 Software Developers tasks that require running other languages and downloading external dependencies to properly test.
- 3. Font Packages Although the automated grader has most metrically-identical fonts (e.g., Liberation Sans instead of Arial), some font packages used in human deliverables still causes certain deliverables to be rendered differently than they would appear on a computer that has these fonts installed.
- 4. **Speech-to-text transcription:** The automated grader has limited speech to text functionality inside the container, and struggles with non-voice sounds.

⁹Metrics were calculated over all samples where the automated grader did not encounter systems errors and returned a valid score. We also excluded 12 tasks (out of the 220 in our open-sourced eval set) that the automated grader frequently could not grade or was less likely to grade reliably due to its limitations, described later.

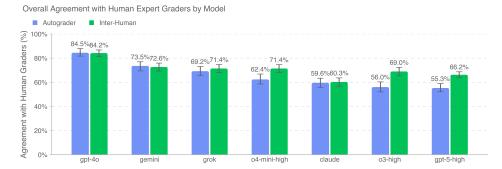


Figure 14: Average human-automated grader agreement is most closely aligned with human interrater agreement for non-OpenAI models. Both agreement metrics are highest for less capable models, as they can be more frequently distinguished from human deliverables and are less likely to be chosen.

A.7 FURTHER METHODOLOGICAL DETAILS ON SELECTING OCCUPATIONS

Assigning Occupations to Sectors We assigned occupations to sectors by using the 2023 BLS National Employment Matrix from U.S. Bureau of Labor Statistics (2025a) to identify the sector with the highest employment for each occupation. This involved filtering to "Line Item" occupations, taking the first two digits of NAICS codes, dropping "total employment" rows, summing 2023 employment, and assigning each occupation to the sector with the largest share of employment.

Detail about O*NET Data Source

Occupations in GDPval. We arrived at 831 occupations by filtering to "Detailed" occupations from the May 2024 OEWS national employment and wage statistics U.S. Bureau of Labor Statistics (2025b) to exclude any aggregate employment categories. We dropped "All Other" occupations, which are catch-all categories within a broader group that bundle together occupations that don't fit into any of the detailed occupations in that group. Dropping "All Other" occupations left us with 761 occupations.

Calculation of Total Wages Earned by Occupation Estimated total wages earned is calculated as total employment * mean annual salary for jobs with annual salaries, and total employment * hourly salary * typical work year of 2080 hours for jobs with only hourly salaries. The determination of which jobs had annual vs. hourly salaries was included in O*NET data. 2080 hours is cited as a "typical work yea" by the Bureau of Labor Statistics (BLS), assuming someone works 40 hours per week. This is an imperfect estimate (eg., the BLS acknowledges actors "generally do not work 40 hours per week, year round") but is the most precise estimate provided by the BLS.

Classifying Occupations as Digital. To classify occupations as predominantly digital, we use a task-based approach. For many occupations, the O*NET database contains task statements and ratings that list all the tasks performed by a worker in an occupation. The O*NET data is provided on the 6-digit SOC occupational code level (SOC-6). We map the O*NET SOC-6 occupations and the corresponding tasks to occupations in the OEWS dataset which reports wages at the 4-digit SOC level ("SOC-4"). For each SOC-4 occupation, we classify its tasks as either digital or non-digital using a prompted GPT-40 model that receives both the occupation and task. We then calculate the weighted share of digital tasks for each occupation. Occupations are classified as digital if their digital share exceeds a threshold of 0.60.

To calculate the weights for our weighted task share, we use task ratings data from O*NET surveys, which includes the relevance, frequency, and importance of each task of the occupation. We first calculate an Adjusted Task Score for each combination of 6-digit SOC occupation and task. This score is defined as the simple average of the three normalized task ratings: task frequency, task importance, and task relevance. Each rating is normalized relative to the maximum observed rating

¹⁰Note that while O*NET distinguishes between Core and Supplemental tasks in its task data, we treat these two task types equally in our calculation of task share.

¹¹For the two occupations without O*NET 28.3 task ratings ("Facilities Managers" and "Medical Dosimetrists"), we used task ratings from O*NET 29.0.

(e.g. the importance ratings are out of 5).¹² If one of these ratings is missing for a task, we impute the value with the mean of that rating across all tasks within the same occupation. For example, if a task lacks a frequency rating, we assign it the average normalized frequency rating of all tasks in the occupation.

We then aggregate these 6-digit Adjusted Task Scores into 4-digit Adjusted Task Scores (for each set of 4-digit SOC occupations and tasks). We do this by summing the SOC-6 Adjusted Task Scores of SOC-6 occupations within a SOC-4 occupation for each task.¹³ For example, the SOC-4 occupation Computer Occupations, All Other combines two 6-digit SOC occupations (*Information Security Engineers* and *Penetration Testers*) which have one task in common: "Identify security system weaknesses, using penetration tests." This task has two SOC-6 Adjusted Task Scores which are added together to create the SOC-4 Adjusted Task Score.

Next, we calculate the Weighted Task Share for each combination of 4-digit SOC occupation and task. The Weighted Task Share is the Adjusted Task Score of the occupation-task pair divided by the sum of Adjusted Task Scores of that occupation. For each occupation, the sum of Weighted Task Share across all its tasks is equal to one. The Weighted Task Share gives us a measure of the relative significance of each task for a given occupation. These Weighted Task Shares are the weights used to calculate the weighted share of digital tasks for each occupation.

Handling Missing Data.

- 1. **Missing Task Statements.** Some occupations in OEWS lacked associated task statements or ratings. Forty-seven of these were broad "All Other" categories without component tasks ¹⁴; twelve others were split into finer sub-occupations in O*NET 29.0 (as of August 2025). For the latter, we incorporated the full set of component tasks from their sub-occupations in O*NET 29.0. The exact reconciliation of how we mapped these 12 occupations is below:
 - (a) **Tour and Travel Guides:** This SOC Code is broken out into two occupations: Tour Guides and Escorts and "Travel Guides". We added the tasks from both occupations.
 - (b) **Miscellaneous Construction and Related Workers:** This SOC Code is broken out into three occupations: "Segmental Pavers", "Weatherization Installers and Technicians", and "Construction and Related Workers, All Other". We added all of the tasks from "Segmental Pavers" and "Weatherization Installers." "Construction and Related Workers, All Other" is a general occupation category without component tasks.
 - (c) **Teaching Assistants:** This SOC Code is broken out into three occupations: Teaching Assistants, Preschool, Elementary, Middle, and Secondary School, Except Special Education, Teaching Assistants, Special Education, and Teaching Assistants, All Other. We added the tasks from Teaching Assistants, Preschool, Elementary, Middle, and Secondary School, Except Special Education, and Teaching Assistants, Special Education. Teaching Assistants, All Other is a general occupation category without component tasks.

¹²The maximum frequency value is 7, the maximum importance value is 5, and the maximum relevance value is 100.

¹³If a SOC-4 occupation is mapped to one SOC-6 occupation, the SOC-6 and SOC-4 Adjusted Task Scores are the same.

¹⁴These occupations were: Entertainers and Performers, Sports and Related Workers, All Other; Postsecondary Teachers, All Other; Production Workers, All Other; Office and Administrative Support Workers, All Other; Teachers and Instructors, All Other; Surgeons, All Other; Information and Record Clerks, All Other; Community and Social Service Specialists, All Other; Educational Instruction and Library Workers, All Other; Sales and Related Workers, All Other; Education Administrators, All Other; Social Workers, All Other; Legal Support Workers, All Other; Food Preparation and Serving Related Workers, All Other; Personal Care and Service Workers, All Other; Food Processing Workers, All Other; Motor Vehicle Operators, All Other; Financial Clerks, All Other; Media and Communication Workers, All Other; Counselors, All Other; Social Sciences Teachers, Postsecondary, All Other; First-Line Supervisors of Protective Service Workers, All Other; Dentists, All Other Specialists; Material Moving Workers, All Other; Helpers, Construction Trades, All Other; Drafters, All Other; Media and Communication Equipment Workers, All Other; Metal Workers and Plastic Workers, All Other; Cooks, All Other; Designers, All Other; Life Scientists, All Other; Building Cleaning Workers, All Other; Precision Instrument and Equipment Repairers, All Other; Grounds Maintenance Workers, All Other; Religious Workers, All Other; Artists and Related Workers, All Other; Textile, Apparel, and Furnishings Workers, All Other; Gambling Service Workers, All Other; Transportation Workers, All Other; Extraction Workers, All Other; Entertainment Attendants and Related Workers, All Other; Woodworkers, All Other; Underground Mining Machine Operators, All Other; Agricultural Workers, All Other; Logging Workers, All Other; Rail Transportation Workers, All Other; Communications Equipment Operators, All Other.

(d) **Buyers and Purchasing Agents:** This SOC Code is broken out into three occupations: Buyers and Purchasing Agents, Farm Products, Wholesale and Retail Buyers, Except Farm Products, and Purchasing Agents, Except Wholesale, Retail, and Farm Products. We added the tasks from the three occupations.

- (e) **Substance Abuse, Behavioral Disorder, and Mental Health Counselors:** This SOC Code is broken out into two occupations: Substance Abuse and Behavioral Disorder Counselors and Mental Health Counselors. We added the tasks from both occupations.
- (f) Clinical Laboratory Technologists and Technicians: This SOC Code is broken out into six occupations: Medical and Clinical Laboratory Technologists, Cytogenetic Technologists, Cytotechnologists, Histotechnologists, Medical and Clinical Laboratory Technicians, and Histology Technicians. We added the tasks from all these occupations.
- (g) Special Education Teachers, Kindergarten and Elementary School: This SOC Code is broken out into two occupations: Special Education Teachers, Kindergarten and Special Education Teachers, Elementary School. We added the tasks from both occupations.
- (h) **Home Health and Personal Care Aides:** This SOC Code is broken out into two occupations: Home Health Aides and Personal Care Aides. We added the tasks from both occupations.
- (i) **Property Appraisers and Assessors:** This SOC Code is broken out into two occupations: Appraisers and Assessors of Real Estate and Appraisers of Personal and Business Property. We added the tasks from both occupations.
- (j) Miscellaneous Assemblers and Fabricators: This SOC Code is broken out into two occupations: Assemblers and Fabricators, All Other and Team Assemblers. We match to Team Assemblers since Assemblers and Fabricators, All Other is a general occupation category without component tasks.
- (k) Electrical, Electronic, and Electromechanical Assemblers, Except Coil Winders, Tapers, and Finishers: This SOC Code is broken out into two occupations: Electrical and Electronic Equipment Assemblers and Electromechanical Equipment Assemblers. We added the tasks from both occupations.
- (1) First-Line Supervisors of Transportation and Material Moving Workers, Except Aircraft Cargo Handling Supervisors: This SOC Code is broken out into four occupations: First-Line Supervisors of Helpers, Laborers, and Material Movers, Hand, First-Line Supervisors of Material-Moving Machine and Vehicle Operators, First-Line Supervisors of Passenger Attendants, and First-Line Supervisors of Transportation Workers, All Other. We added the tasks from First-Line Supervisors of Helpers, Laborers, and Material Movers, Hand, First-Line Supervisors of Material-Moving Machine and Vehicle Operators, and First-Line Supervisors of Passenger Attendants. First-Line Supervisors of Transportation Workers, All Other is a general occupation category without component tasks.
- 2. Missing Task Ratings. There are 36 SOC-6 occupations which do not have any task rating in O*NET 28.3 or 29.0. These correspond to 34 SOC-4 occupations. Among these, 2 SOC-4 occupations (Data Scientists and Web and Digital Interface Designers) have task ratings for some of the component SOC-6 occupations which allow us to compute the Adjusted and Weighted Task Share measures. For the rest of the 32 SOC-4 occupations that

¹⁵These SOC-4 occupations are: Aircraft Service Attendants, Bus Drivers, School, Calibration Technologists and Technicians, Cardiologists, Crematory Operators, Data Scientists, Disc Jockeys, Except Radio, Emergency Medical Technicians, Emergency Medicine Physicians, Entertainment and Recreation Managers, Except Gambling, Financial and Investment Analysts, Financial Risk Specialists, First-Line Supervisors of Entertainment and Recreation Workers, Except Gambling Services, First-Line Supervisors of Security Workers, Fundraising Managers, Health Information Technologists and Medical Registrars, Hydrologic Technicians, Legislators, Lighting Technicians, Medical Records Specialists, Orthopedic Surgeons, Except Pediatric, Paramedics, Pediatric Surgeons, Project Management Specialists, Public Relations Managers, Sales Representatives of Services, Except Advertising, Insurance, Financial Services, and Travel, School Bus Monitors, Shuttle Drivers and Chauffeurs, Software Developers, Special Education Teachers, Kindergarten and Elementary School, Substitute Teachers, Short-Term, Taxi Drivers, Teaching Assistants, Except Postsecondary, Web and Digital Interface Designers.

have no O*NET task ratings, we cannot compute the Adjusted or Weighted Task Share measure. Instead, we proxy the Weighted Task Share as follows: for each combination of 4-digit SOC occupation and task, we calculate the number of times the task appears (i.e., task frequency) for the occupation and divide by the sum of task frequency of all tasks of that occupation. For example, the 4-digit SOC occupation "Special Education Teachers, Kindergarten and Elementary School" combines two 6-digit SOC occupations (*Special Education Teachers, Elementary School* and *Special Education Teachers, Kindergarten*) has 43 unique tasks. Among these 17 tasks appear twice. Thus, the sum of task frequency across 43 tasks is 60. For each task that appears once, the proxy Weighted Task Share is 1/60 = 0.0017, and for each task that appears twice, the proxy Weighted Task Share is 2/60 = 0.0033.

A.7.1 VALIDATING THE DIGITAL TASKS MEASURE

We benchmark our "knowledge work" classification method against the task-content framework of Acemoglu & Autor (2011).

The framework in Acemoglu & Autor (2011) is based on the U.S. Department of Labor's O*NET survey, which collects data on the activities, work "content", and abilities required for each occupation. The framework aggregates these measures into five scores:

- 1. Non-routine cognitive: Analytical.
- 2. Non-routine cognitive: Interpersonal.
- 3. Routine cognitive.
- 4. Routine manual.

 5. Non-routine manual physical.

Each score is computed as a composite measure of select O*NET "Importance" scales. For example, the "Non-routine cognitive: Analytical" score for each occupation is computed by summing the (normalized) values of the "Analyzing data/information" work activity, the "Thinking creatively" work activity, and the "Interpreting information for others" work activity. A high numerical value for an occupation for a given score indicates that the occupation relies heavily on that type of work. We compute the Acemoglu & Autor (2011) scores for each occupation and then compare them with our measures of knowledge work (that is, the share of digital tasks and a binary "knowledge work" indicator for each occupation).

In our first set of results, we compare each Acemoglu & Autor (2011) task-content score with the share of digital tasks in an occupation. The patterns are clear: occupations with higher digital-task shares score systematically higher on the non-routine cognitive dimensions and lower on the manual dimensions. In other words, the more an occupation relies on digital tasks, the more it resembles cognitive, non-routine work.

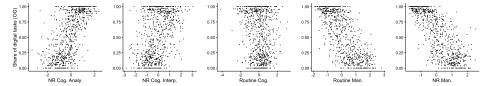


Figure 15: Distribution of occupations and task contents

In our second set of results, we look at the relationship between the Acemoglu & Autor (2011) scores and our binary measure of "knowledge work." In the following figure, we plot each occupation's value for each score, and color occupations by the paper's knowledge-work classification: blue for occupations identified as knowledge work and red for all others. The pattern is again clear–knowledge-work occupations cluster at the top of the non-routine cognitive distributions and at the bottom of the routine and manual distributions.

Taken together, these results suggest that our digital-task classification is closely aligned with the economic literature on cognitive/manual work.

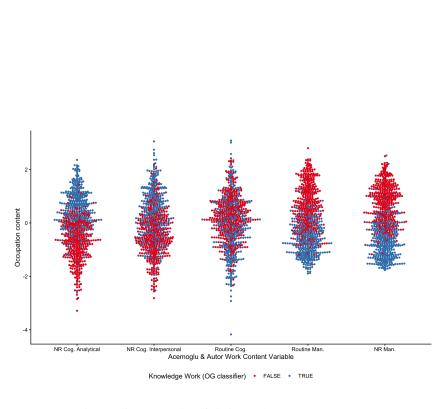


Figure 16: Scatterplot of digital tasks and task contents

Sector	% GDP	Top Occupations and Total Compensation (in Billions USD)			
Real Estate and Rental and Leasing	13.8%	Property/RE/Community Association Managers — \$24.54B			
C		Counter and Rental Clerks — \$17.42B			
		Real Estate Sales Agents — \$13.53B			
		Real Estate Brokers — \$4.55B			
		Concierges — \$1.80B			
Manufacturing	10.0%	First-Line Supervisors of Production and Operating Workers — \$51.07B			
		Buyers and Purchasing Agents — \$39.79B			
		Shipping, Receiving, and Inventory Clerks — \$38.50B			
		Industrial Engineers — \$37.79B			
		Mechanical Engineers — \$31.57B			
Professional, Scientific, and Technical Services	8.1%	Software Developers — \$239.18B			
and reclinical services		Lawyers — \$136.66B			
		Accountants and Auditors — \$135.44B			
		Computer and Information Systems Managers — \$121.44B			
		Project Management Specialists — \$108.77B			
<u> </u>	11.20				
Government	11.3%	Compliance Officers — \$33.80B			
		Administrative Services Managers — \$32.03B			
		Child, Family, and School Social Workers — \$24.10B			
		First-Line Supervisors of Police and Detectives — \$17.00B			
		Recreation Workers — \$11.51B			
Health Care and Social Assistance	7.6%	Registered Nurses — \$323.05B			
		First-Line Supervisors of Office/Admin Support — \$107.02B			
		Medical & Health Services Managers — \$77.93B			
		Nurse Practitioners — \$40.58B			
		Medical Secretaries & Admin Assistants — \$37.87B			
Finance and Insurance	7.4%	Financial Managers — \$147.74B			
		Customer Service Representatives — \$123.70B			
		Securities, Commodities, and Financial Services Sales Agents — \$52.14I			
		Personal Financial Advisors — \$43.33B			
		Financial and Investment Analysts — \$39.67B			
Retail Trade	6.3%	General & Operations Managers — \$477.16B			
		1st-Line Supervisors of Retail Sales Workers — \$58.27B			
		Pharmacists — \$45.12B			
		Private Detectives & Investigators — \$2.39B			
Wholesale Trade	5.8%	Sales Reps, Wholesale & Mfg (Except Tech/Scientific) — \$103.21B			
		Sales Managers — \$97.16B			
		Sales Reps, Wholesale & Mfg (Tech/Scientific) — \$33.66B			
		1st-Line Supervisors of Non-Retail Sales Workers — \$21.43B			
		Order Clerks — \$3.86B			
Information	5.4%	Producers & Directors — \$16.60B			
		Editors — \$8.18B			
		News Analysts, Reporters, and Journalists — \$4.41B			
		Audio & Video Technicians — \$4.30B			
		Film & Video Editors — \$2.41B			

Table 1: Sectors, their value added as a percentage of U.S. GDP (Q2 2024), with representative top occupations and total compensation in billions (USD).

To assess task quality, we asked occupational experts to rate each task on its difficulty, representativeness, time to complete, and overall quality against real-world standards for their occupation. Each task's dollar value was estimated by multiplying the average estimated completion time by median hourly wages for the corresponding occupation from OEWS data. The average task was valued at \$398.46.

	Mean	Std	Min	25%	50%	75%	Max
Overall quality (1–5)	4.47	0.32	3.18	4.30	4.50	4.70	5.00
Difficulty (1–5)	3.32	0.95	1.00	3.00	3.00	4.00	5.00
Representativeness (1–5)	4.50	0.74	2.00	4.00	5.00	5.00	5.00
Avg time to complete (hrs)	9.49	13.75	0.50	2.38	5.00	10.00	100.00
Dollar value of task	\$398.46	\$599.45	\$12.59	\$93.72	\$174.81	\$386.03	\$4,114.20

Table 4: Summary stats for GDPval full set tasks

	Mean	Std	Min	25%	50%	75%	Max
Overall quality (1–5)	4.55	0.43	2.00	4.33	4.56	5.00	5.00
Difficulty (1–5)	3.20	0.92	1.00	3.00	3.00	4.00	5.00
Representativeness (1–5)	4.43	0.76	1.00	4.00	5.00	5.00	5.00
Avg time to complete (hrs)	8.63	24.70	0.25	2.00	4.00	8.00	605.00
Dollar value of task	\$391.44	\$1,296.67	\$8.53	\$70.70	\$147.31	\$354.12	\$32,028.70

Table 5: File counts

	Mean	Std	Min	25%	50%	75%	Max
Reference files	1.92	3.47	0.00	0.00	1.00	2.00	38.00
Deliverable files	1.54	2.64	0.00	1.00	1.00	1.00	36.00

Table 6: O*NET Tasks, Skill, and Work Activities Coverage

	Total unique in O*NET	Total in gold set	Coverage (%)
O*NET Skills	35	25	71.4%
O*NET Work Activities	41	26	63.4%
O*NET Tasks	1,470	208	14.15%

Table 7: Task Specification

Label	Count	%
Underspecified	1,231	8.28%
Well specified	13,246	89.07%
Overspecified	395	2.66%