ORQA: AN OCCUPATION-REALISTIC QUESTION AND ANSWER FRAMEWORK FOR LLM PROFESSIONAL KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

We present ORQA, a method for testing occupation-level knowledge in large language models. Prior methods either map abstract LLM skills to occupations via task definitions or utilize expert knowledge which is difficult to obtain at scale and expensive. ORQA complements both of these methods by connecting O*NET occupations to trusted occupation-specific websites (such as regulatory agencies, licensing bodies, professional organizations, and government publications) and converting these into source-traceable question-answer pairs. A combination of an automated pipeline and human review produces a set of high quality questions about occupations. The question set created via our method covers 116 occupations from all 21 major groups in the SOC, with 480 questions sourced from 187 different websites. Each question is designed to probe a real-world skill question that is relevant to the occupation in question. We test 15 state-of-the-art frontier and open-weight models via this method. Claude Opus 4.6, GPT-5.4 and Claude Sonnet 4.6 all perform the best at approximately 58-62% while smaller open-weight models achieve approximately 33-41% performance. Performance varies significantly across occupations. Healthcare-related occupations achieve the highest performance (78%) while Office and Administrative Support achieve approximately 40%. Performance on individual occupations (e.g. Sheet Metal Workers and Fish and Game Wardens) is essentially zero. We also find that open-ended questions and weighting by wage bill do not significantly affect the ranking of models on this benchmark. We believe that leveraging existing trusted occupation-specific information to test LLM knowledge in professional domains may be a scalable and useful method for evaluating occupation-level AI performance in the future. Results and data are available at an anonymized project website.

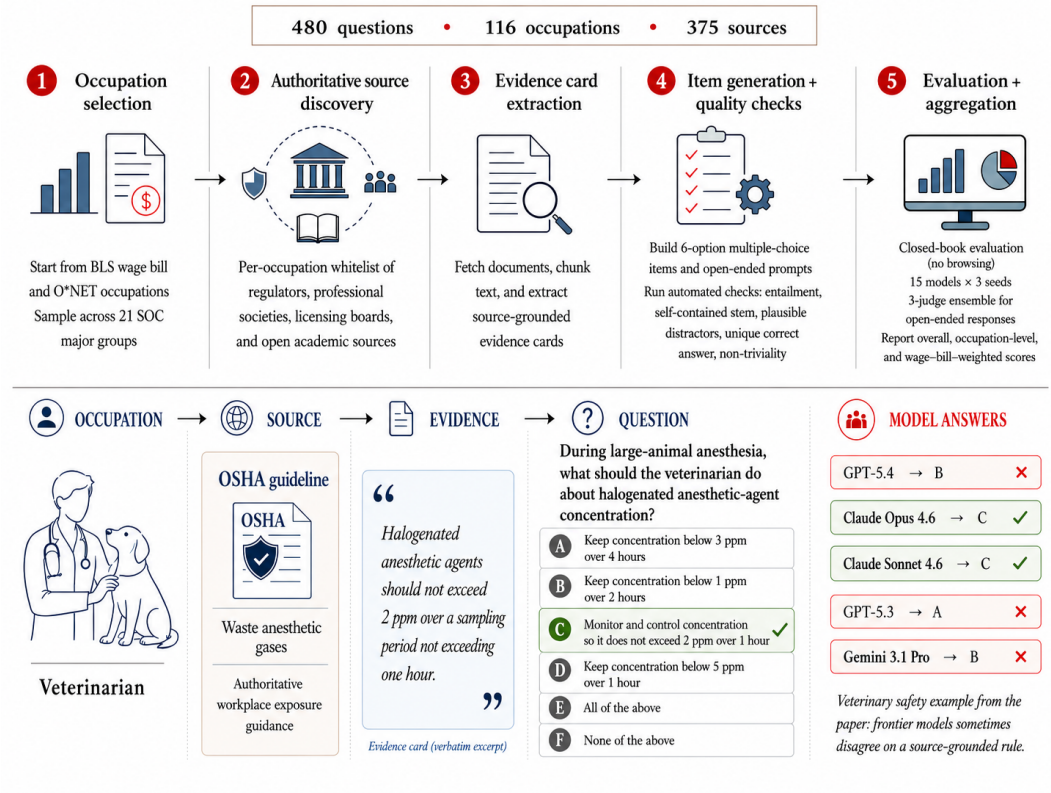
1 INTRODUCTION

Benchmarks have been central to measuring and directing progress in artificial intelligence. For language models, benchmarks such as MMLU, BIG-bench, HELM, GPQA, ARC, and SWE-bench have made it possible to compare systems on reasoning, knowledge, coding, and problem solving in a reproducible way (Hendrycks et al., 2021; BIG-bench authors, 2023; Liang et al., 2023; Rein et al., 2024; Chollet, 2019; Jimenez et al., 2024). As language models become general-purpose work tools, however, the evaluation question shifts. We no longer only want to know whether a model has an abstract capability. We want to understand its practical impacts across the economy. We want to know whether it can help a worker in a particular occupation answer the kinds of questions that arise in practice, and whether one model is more useful than another for that occupation.

Existing work does not yet provide this kind of occupation-level model benchmark across the entire US economy across models in a scalable way. Existing economic work measures exposure, adoption, or productivity effects of AI in the labor market (Brynjolfsson et al., 2018; Eloundou et al., 2023; Handa et al., 2025; Chatterji et al., 2025; Brynjolfsson et al., 2025; Dell'Acqua et al., 2026). Some papers on AI in the labor market are valuable for examining the impact of AI on the labor market. However, these are typically not performance comparison benchmarks for AI systems. Instead, these papers typically estimate the overall impact of a particular AI system on the labor market rather than comparing the usefulness of a system for specific jobs to another system for the same jobs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Figure 1: Construction pipeline for ORQA from occupation sampling to verified benchmark questions.



Note. The figure illustrates the pipeline from occupation sampling and source identification to evidence-card extraction, question construction, quality assurance, and finally to the creation of a balanced benchmark composed of verified questions.

Several recent benchmarks have started to aim at more real tasks. These include web agent tasks (Zhou et al., 2023), enterprise workflows (Drouin et al., 2024), consequential workplace tasks (Xu et al., 2025), and software engineering tasks (Jimenez et al., 2024). These are promising developments. However, these benchmarks are still quite limited with respect to the labor market. Many benchmarks currently examine computer mediated digital jobs. In addition, recent attempts to align agent benchmarks to jobs have found a strong bias toward mathematical and computer-related jobs (Wang et al., 2025; 2026). GDPval represents an important recent effort toward examining more economically relevant tasks. However, GDPval uses expert written prompts and expert or model evaluations which limit the ability to extend to hundreds of jobs (Patwardhan et al., 2025). There have also been attempts to bridge this gap using expert evaluations of AI performance on job-task levels (Shao et al., 2025). Thus the gap appears to be one of benchmarking. We desire occupation-level assessments that are (a) scalable, (b) broadly applicable throughout the economy, and (c) capable of assessing knowledge at the level of occupations. We also desire some of the positive aspects of the current system including source verifiability and model comparisons while moving beyond purely task completion to encompass professional knowledge.

We introduce Occupation Related Question and Answer (ORQA) to fulfill this need but also to explore the feasibility of automatically creating an occupation-level benchmark that is grounded in the economy and to compare the advantages and disadvantages of this approach versus expert created benchmarks. The creation of ORQA stems from the recognition that there exists authoritative text for many occupations that contains professional knowledge relevant to professionals in that occupation. Practical knowledge that professionals are expected to utilize is found in occupation specific documents, government agencies, regulators, licensing bodies, academic sources, standards organizations, and formal and informal professional associations. This also reflects sociological

understanding regarding the creation of professional norms through occupations and their codification in formal and informal sources (Abbott, 2014). We believe such documents can be used to provide scalable occupation-level comparisons, provided they can be connected to O*NET occupations and transformed into source-traceable questions. We are also cognizant of the scope of ORQA. Specifically, ORQA addresses the source-traceable aspect of occupational knowledge (reporting responsibilities, professional recommendations, thresholds, procedures and guidelines) rather than overall occupational capability. Furthermore, ORQA is designed to operate on a number of occupations where authoritative information exists, rather than across the entire labor market. Instead, we view ORQA as a system that demonstrates the feasibility of automated source-traceable occupation-level comparisons and augments both task and expert-based methods.

We propose an agentic system for the large-scale creation of occupation-related QA pairs. We begin with occupation information from the BLS and O*NET (U.S. Bureau of Labor Statistics, 2024; National Center for O*NET Development, 2026) and create occupation-specific lists of authoritative sources. Using these lists, we retrieve documents relevant to the occupation. We then create evidence cards from within the source documents. These cards serve as the standard against which automated systems can verify information. We test the models on both open-ended answers against the evidence cards and multiple choice questions using the same source material. We then go through a series of automated filters to ensure the quality and correctness of the questions. We also included a human evaluation step in which two annotators manually checked the quality of items, their sources and their answer options against a quality standard. Agreement between our manual and automated "good" decisions is about 67% on the items. We use this as a metric to assess the performance of a scalable automated filter relative to expert curation. Correctness is defined in relation to the exact entailment in the source documents. All selected items are also subject to the same manually determined standard. We have created a curated authoritative ORQA dataset consisting of 480 questions, drawn from 187 source hosts, 116 occupations and all 21 major groups in the SOC. We test fifteen frontier and open-weight models without retrieval or browsing capability. We primarily test on closed-book multiple choice using three random seeds per question. We also test on an open-ended task where the model only sees the question text and must provide a brief free response which is scored by a three-judge panel. We also provide a wage-bill weighted measure that uses weighted occupations based on employment multiplied by the mean wage.

We observe three main results. One is that ORQA provides some separation between models in both the multiple choice and open-ended tasks. For multiple choice closed book tasks the frontier models (Claude Opus 4.6, GPT-5.4, and Claude Sonnet 4.6) are very close at 62.4%, 60.3% and 58.7% respectively. Gemini 3.1 Pro is at 56.2% and o3 at 51.7%. GPT-4o is at 43.5% and the open-weight and older models at 33-41%; the three models used in the difficulty filter are reported but not ranked. The absolute scores are lower than on general knowledge tasks because the questions are hand curated to be non-trivial and human verified, so what remains are questions that are genuinely difficult for the models. For open-ended tasks the absolute scores are again somewhat lower than closed book multiple choice, as expected. However the frontier ordering remains: GPT-5.4 at 54.2%, Claude Opus 4.6 at 53.1%, and Claude Sonnet 4.6 at 50.6%. Wage-bill weighting does not affect the overall conclusion. The frontier models again appear first and the absolute scores again show only modest differences.

Second, aggregate scores mask significant occupation by model differences. For some occupations many models perform well while for others even the frontier models perform poorly. For instance GPT-5.4 performs overall at 60.3% but answers 0 of 12 seed-trials correctly for Actuaries (4 items), 0 of 9 for Sheet Metal Workers and for Fish and Game Wardens (3 items each), and 1 of 9 for Flight Attendants. We observe large cross-model differences (up to 100 points) in certain jobs. For Financial Managers (3 items), GPT-5.4 answers 9 of 9 seed-trials correctly and GPT-3.5 Turbo 0 of 9. Even at the level of high-level SOC categories, there is observed heterogeneity. For the most well-represented category, Healthcare Practitioners frontier models perform around 78-80%, while for Installation, Maintenance, and Repair frontier models only perform 44-46% and for Office and Administrative Support jobs frontier models perform 36-43%. This is useful as it allows employers and employees to look beyond overall performance of a model and instead examine the reliability of a certain model for a certain type of job.

Third, ORQA observes a different entity but also tracks externally available signals. ORQA shows high correlations (Spearman) with GDPval (+0.90) and the GDPval-AA Elo ranking (+0.92) across overlapping models. ORQA also shows coverage across occupations in bins of exposure to the

Anthropic Economic Index. This shows that ORQA both covers occupations that are both frequently and infrequently exposed to the Anthropic Economic Index and also tracks externally available signals. This implies ORQA does not currently appear to contradict externally available signals related to the economy and work. ORQA also adds the ability to perform source-aware, occupation-aware, and scalable model comparisons.

Finally, we present four contributions. Our first contribution is a scalable system for creating occupation-level knowledge resources from authoritative data. We present a general approach for transforming public data into verifiable questions and quality/balance considerations that are important to this task. Our second contribution is to instantiate this system as ORQA, a 480-item resource that contains data for 116 occupations, all 21 major groups in the SOC, and both digital and non-digital occupations. Our third contribution is a fifteen-model closed-book multiple choice and open response experiment, with uncertainty via cluster- bootstrap and an aggregation weighted by occupational wage bill. We observe significant variation across models, occupations, and SOC major groups. Our fourth contribution is to compare ORQA to external signals related to occupations such as GDPval, GDPval-AA, and exposure to the Anthropic Economic Index. We also demonstrate how the occupation-indexed system can be extended outside of authoritative data using a community-thread version that leverages data from public discussions (e.g. Reddit threads about occupations). ORQA is thus both a resource and an example of how far automation using data from specific sources can be extended to occupation-level understanding.

2 Related Work

General capability metrics. Benchmarks for general purpose language models have been instrumental in tracking advancements in coding, scientific knowledge, reasoning, language understanding and other domains (Hendrycks et al., 2021; BIG-bench authors, 2023; Liang et al., 2023; Rein et al., 2024; Chollet, 2019; Jimenez et al., 2024). Later work raised the difficulty of this format with harder distractors (Wang et al., 2024) or questions at the limit of expert knowledge (Phan et al., 2025). Such benchmarks assess capability but not impact on work, because they are not structured around jobs, and difficulty is a separate axis from occupational grounding. For instance, scoring high on an abstract capability benchmark does not necessarily imply that the model would assist a mechanic, groundskeeper, accountant, driver, or nurse with the concrete questions that occur in their respective fields.

Rather, evidence in the economics literature regarding the impact of AI on the job market focuses on exposure, usage, and field. Many economics papers have looked at exposure in order to identify the jobs or tasks that are most at risk from AI or generative AI (Brynjolfsson et al., 2018; Webb, 2020; Felten et al., 2021; 2023; Eloundou et al., 2023). There have also been recent studies that examine usage (Handa et al., 2025; Appel et al., 2025; Chatterji et al., 2025; Bick et al., 2024). Finally, field studies offer more concrete insights into quality and productivity impacts across specific domains like knowledge and customer service work (Brynjolfsson et al., 2025; Dell'Acqua et al., 2026).

There are several types of benchmarks that are relevant to understanding the economic impact of AI. Unlike ORQA, however, these do not typically provide a reusable test that evaluates many models on the same occupation-indexed questions. Instead, they generally evaluate on more realistic enterprise, agent, or work tasks. For example, recent benchmarks have transitioned from abstract skills to more realistic work. WebArena, WorkArena, WorkArena++, TheAgentCompany, and SWE-bench all test on software problems, enterprise processes, consequential workplace-like tasks, or web tasks for agents or models (Zhou et al., 2023; Drouin et al., 2024; Boisvert et al., 2024; Xu et al., 2025; Jimenez et al., 2024). This represents a move toward more realistic evaluation. Still, these benchmarks tend to focus on digital, computer-mediated work. Recent efforts to connect agent benchmarks to occupations have revealed a significant bias toward computer and math-based occupations in benchmarks and a large underrepresentation of portions of the job market (Wang et al., 2026; 2025).

Knowledge- and economically valuable benchmarks. GDPval is most similar in nature to ORQA. GDPval tests models on economically valuable occupation-based tasks. The tasks are created by experts and evaluated either by experts or models (Patwardhan et al., 2025). The benefits of GDPval are its realistic nature. However, the tasks are created by experts which makes scaling to hundreds of occupations repeatedly expensive. ORQA, on the other hand, scales by creating verifiable sources for occupation tasks using authoritative occupation information. A third approach drops the source and

has models write and grade their own occupational exams from task descriptions (del Rio-Chanona & Einsiedler, 2025). That scales further but has no external ground truth, whereas every ORQA answer traces to a passage in a document the occupation is governed by. A fourth line evaluates execution rather than knowledge by placing agents in real computer environments (Xie et al., 2024), which we treat as complementary to static questions. Another effort that has begun to connect AI systems to ONET tasks highlights the importance of human-AI collaboration and human-AI assistance in model evaluation rather than solely end-to-end task solving (Shao et al., 2024; Chang et al., 2025). ORQA is not in opposition to this effort, but rather evaluates a more narrow and specific aspect of this. Instead of measuring end-to-end task performance or the value of assistance, ORQA evaluates whether a model's performance is consistent with the authoritative knowledge base of an occupation.

3 METHODOLOGY

From exploratory to an approved pipeline. We created the current pipeline in two stages. In the first stage, we created an agentic pipeline with multiple automated steps that would take occupational data and from credible online sources and produce a large number of potential question-answer pairs. The objective of this initial stage was not to create the question bank itself, but to gain insight into what constitutes a good occupational question. A sample of these generated questions were manually inspected by two annotators. For each question, the answer choices, source, and question were scored using the same quality criteria. The annotators would then provide explanations for why a particular question- answer pair should be rejected. A tagging dashboard revealed several common types of problematic questions: answers not sufficiently implied by the source, distractors that could be answered using general domain knowledge, use of research articles instead of credible sources and correct answers obscured by vague modifiers. The whitelist is built per occupation beforehand by selecting certifying and licensing bodies from the O*NET association lists, with the selection validated against those lists (Appendix A.4). About 67% of the items the automated gate retained were rated good or okay by the human annotators. We present this agreement as an indication of the similarity between automated scalable gates and expert curation, but do not present automation as a replacement for expert curation. Correctness itself is anchored to verbatim entailment from the source and is independent of this scalable gate. In the following, we describe the resulting system in its current state. The system produces items all of which have been filtered through the same quality validation step. We do not make distinctions between early and later items. The overall pipeline from occupational sampling to a balanced pool is summarized in Figure 1.

Obtaining occupations. We start with the May 2024 national table from the Bureau of Labor Statistics Occupational Employment and Wage Statistics. For each Standard Occupational Classification (SOC) code, we calculate the wage bill as the sum of total employment nationwide times the average annual wage in that occupation. This represents the annual amount of labor compensation transiting through the occupation. We then allocate items across the major groups in the SOC according to the proportion of the nation's wage bill attributable to that group. This provides economic significance to both broad and narrow occupational categories. We try to get items for jobs in order of decreasing wage bill within each occupational category.

Domain whitelists per job. A general web search for a job returns millions of off-topic or poor quality pages such as content farms, marketing pages or blogs. Instead, we limit the search for sources for a given job to a list of trusted publishers curated from the official source information within O*NET. For example, a Registered Nurse may be sourced from domains like aacn.org (American Association of Critical-Care Nurses), nursingworld.org (American Nurses Association), ncsbn.org (National Council of State Boards of Nursing), state boards of nursing for that state, and the federal government sources bls.gov, osha.gov and cdc.gov. Each job has a whitelist appropriate to that job. These whitelists are complete. No document can come into the bank from a domain not on the whitelist. There is no overall allowlist from which academic or other documents can bypass the job specific whitelist.

Research papers are excluded before fetching (Appendix A.10).

Source discovery and fetching. Discovery is a single general web round performed through a search API constrained to the whitelist for that job, so every candidate document comes from a publisher that governs the occupation. Nothing is pulled from outside the allowlist.

Evidence card and content creation. Discovery runs through the Serper search API, with each query restricted to one whitelisted domain and paired with a document-type or topical keyword (Appendix A.4). Documents are extracted and then parsed into structured objects. Documents are chunked with windows of two and a half to four kilobytes, overlapping by about a hundred tokens to avoid context drops. Each window is then passed through an evidence card extraction model (GPT-4o). Each card will extract one to three contiguous sentences from the document used as a source quote. The card will also extract the task or procedure referenced in the source quote, and the appropriate occupation.

Item creation. Each card is transformed into a six choice multiple choice question. The correct answer must be entailed by the source quote and can be created from the source quote. Distractor answers are created from within the same document. All of the above and None of the above are guaranteed to appear in every question as two of the six answer choices. Only one answer choice out of the six can be correct. Answer choices are also required to be of similar lengths and free from obvious hinting answers. If an item fails on either of these criteria it is recorded and the card is rejected.

Automated quality and categorization. After an item is created it goes through a series of automated filters. Some of these filters were created to prevent the failure modes discovered through manual testing. A failed item is sent to a specific regeneration attempt instead of being rejected immediately. The item that is produced by the regeneration attempt is then inserted back into the pipeline at the point of failure. Checks are performed in the following order: correct response is fully entailed by the original quote; question is complete and makes sense; all answer choices are topic-appropriate and not obviously nonsensical; no two responses are paraphrases of each other; only one response is correct based on the original; responses have a similar length and do not allude to use of specific sources (e.g., According to NFPA, . . .); the item is not solvable by all three of GPT-4o mini, Claude Haiku 4.5, and Gemini 2.5 Flash in parallel on a closed-book basis (and if so, is considered too easy and rejected; an item all three answer incorrectly is also rejected if the verifier marked it ambiguous or low-quality); no answer choice can be eliminated based on general background knowledge (assessed by an independent LLM); and the correct answer does not camouflage behind vague terms such as standard or appropriate while other answers reference specific values or tools. The last filtering step is the one that is aligned to human annotations. It is this step that allows the pipeline to automatically reach the human defined "good" answer. Section A.4 gives the exact prompts, and Appendix A.11 reports how many candidates each stage rejected, the pass rate of each check and how often each check triggered regeneration.

Domain rebalance and occupation-floor rebalance. Once the quality checks are successful, domain balancing is performed across three domain rules to avoid over-representation from a single domain for any occupation. One rule caps any single source domain at 30% of the items in a SOC group. A second rule replaces that cap for two kinds of domain: a federal regulatory domain that is central to the occupation may reach 40-50%, while a domain whose subject matter is tangential to the occupation is held to 15-20%. Centrality is judged from whether the domain is the body that issues the rules the occupation works under. A third rule blocks a small number of low quality domains. Finally, an occupation-floor rebalance is performed which maintains all occupations with at least one verified item, while limiting the number of items in each occupation. This avoids situations in which a high-value occupation (such as registered nurses in cdc.gov) dominates the bank. There was a tendency toward concentration in available fields such as health care. To address this, we also performed a focused expansion for underrepresented jobs, with a minimum of three items in each job. This increased the number of jobs containing at least three items to 106 out of 116 jobs. The floor shapes a third of the bank: 160 items sit in occupations with three or fewer items, and the rest follow the wage-bill sampling.

Final bank. The final ORQA bank consists of 480 items in 116 jobs across all 21 major SOC groups, drawn from 375 source documents across 187 distinct source hosts. About 43% of the items

Table 1: Adjacent-pair accuracy differences (points) among the twelve ranked models; 95% paired two-level bootstrap intervals (10,000 replicates). *Excludes zero.

Pair	Diff.	95% CI	Pair	Diff.	95% CI
Opus 4.6 > GPT-5.4	+2.1*	[0.3, 3.9]	o3 > GPT-4o	+8.1*	[5.8, 10.5]
GPT-5.4 > Sonnet 4.6	+1.6	[−0.1, 3.3]	GPT-4o > DeepSeek V4	+2.8*	[0.4, 5.2]
Sonnet 4.6 > Gemini 3.1 Pro	+2.4*	[0.4, 4.5]	DeepSeek V4 > Llama 3.3 70B	+1.9	[−0.1, 4.0]
Gemini 3.1 Pro > GPT-5.3-chat	+2.0	[−0.1, 4.1]	Llama 3.3 70B > Qwen 2.5 7B	+4.9*	[2.7, 7.2]
GPT-5.3-chat > GPT-5.2	+1.5	[−0.5, 3.4]	Qwen 2.5 7B > GPT-3.5 T.	+0.6	[−1.9, 3.0]
GPT-5.2 > o3	+1.1	[−1.2, 3.5]			

come from federal government sources (ending in .gov). The rest come from licensing organizations and professional organizations, apart from two items from a state forestry extension service. osha.gov is the most prominent source host at 9.8% of the bank, far below the 30% maximum allowed from a source per job. The coverage is wide but not even: some major SOC groups (such as Production and Healthcare Practitioners) have many jobs, while others (such as Farming, Fishing, and Forestry) have only one or two jobs. We return to this issue in Section 5.

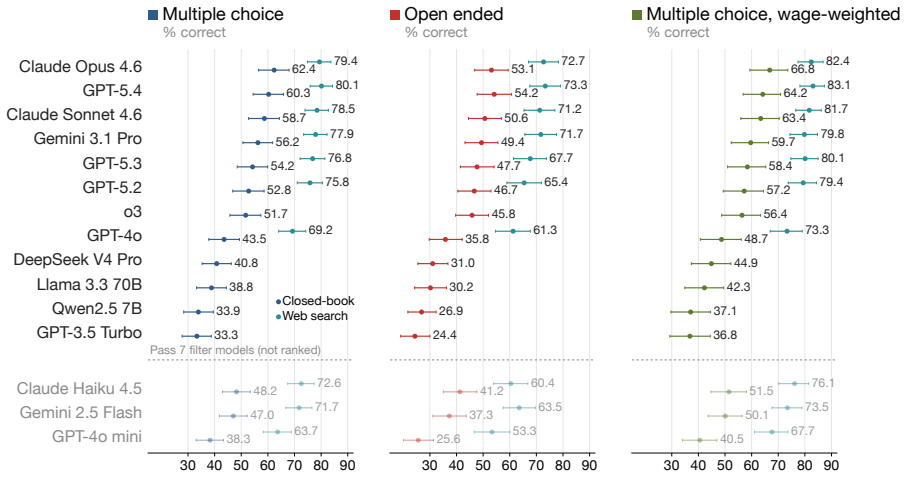
Evaluation. We evaluate fifteen models in closed-book testing. This means that the model is only shown the question text and the six answer choices. There is no browsing or retrieval. We test the following models: GPT-5.4, GPT-5.3-chat-latest, GPT-5.2, o3, GPT-4o, GPT-4o mini, GPT-3.5 Turbo, Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5, Gemini 3.1 Pro, Gemini 2.5 Flash, Llama 3.3 70B, DeepSeek V4 Pro, and Qwen 2.5 7B. Each of the 480 questions is run three times independently across each model using different random seeds. We report accuracy as the mean over all question-model-seed combinations. We estimate 95% intervals with a two-level cluster bootstrap that resamples occupations and then items within each occupation. This is the correct level of clustering because items within the same occupation have similar content and cannot be assumed independent. GPT-4o mini, Claude Haiku 4.5 and Gemini 2.5 Flash served as the Pass 7 filter, so we report but do not rank them. For a multiple choice question with six possible answers, the accuracy by random guessing would be 1/6 ≈16.7%. We include this as a reference line. Because the correct answer is never All of the above or None of the above (Appendix A.2), guessing among the four substantive options gives 25%.

4 RESULTS

Item validation. Prior to score interpretation we consider the degree to which ORQA items are occupational in nature. ORQA asserts three sources of validity. First, items are occupation-linked. Each item is associated with an O*NET occupation and derived from materials produced by industry associations, agencies, standards organizations, licensing organizations, regulatory organizations, academic institutions or other organizations that define or oversee the nature of work in that occupation. Second, response correctness is source-derived and not model-inferred. That is, the correct response must be entailed by an exact excerpt from the source material. Finally, items undergo the automated verifications detailed in Section 3 that were tuned using human annotation before being added to the bank. The Veterinarians call-out in Figure 1 is an example of this. An OSHA exposure guideline in the workplace specifies an absolute level for halogenated anesthetic agents. ORQA transforms this into an occupational item and the correct response is correct because it is present in the source and not because a model has inferred its plausibility.

Overall performance. The same 480 item bank is assessed in three different ways: closed-book multiple choice, open-ended responses with model-scored responses, and wage-bill-weighted closed-book accuracy (Figure 2). In closed book multiple choice Claude Opus 4.6 comes first at 62.4%, then comes GPT-5.4 at 60.3%, and then Claude Sonnet 4.6 at 58.7%. A paired bootstrap (Table 1) separates Claude Opus 4.6 from GPT-5.4 (+2.1 points, 95% CI [0.3, 3.9]) but not GPT-5.4 from Claude Sonnet 4.6; only 5 of 11 adjacent pairs are separable. Gemini 3.1 Pro (56.2%), GPT-5.3-chat-latest (54.2%), GPT-5.2 (52.8%) and o3 (51.7%) follow, then GPT-4o (43.5%), the open-weight models (33.9-40.8%) and GPT-3.5 Turbo (33.3%). All fifteen models score above the 16.7% chance rate and the 25% rate for guessing among the four substantive options, yet even the best models get over a third of these questions wrong.

Figure 2: Overall ORQA performance across multiple-choice, open-ended, and wage-weighted evaluation.



Note. Same 480 items and models in every panel. Closed-book scores (lower dot) for all models; web-search scores (upper dot) for the ten models with a native search tool. Multiple-choice scores average three seeds; open-ended scores use a three-judge majority vote. Bars: 95% two-level cluster bootstrap intervals (occupations, then items). Models below the dashed line were the Pass 7 difficulty filter and are not ranked.

Two robustness checks leave the frontier ordering in place. In the open-ended mode each model sees only the question text and answers in at most five lines, and three judges from different providers score an anonymized response by majority vote (model versions in Appendix A.4). Scores fall by about 7-9 points (GPT-5.4 54.2%, Claude Opus 4.6 53.1%, Claude Sonnet 4.6 50.6%), as expected when models must generate rather than recognize an answer. Weighting occupations by national wage bill raises scores by about 4 points and keeps the same top three (Claude Opus 4.6 66.8%, GPT-5.4 64.2%, Claude Sonnet 4.6 63.4%), so the ordering is not an artifact of the format or the occupation mix. A basic web search tool raises frontier accuracy to about 78-80% in multiple choice and 71-73% open-ended for the ten models that have one. With a fixed retriever over the same authoritative hosts, models score 93-99% when the source is retrievable and the frontier about 90% when it is excluded (Appendix A.15), so items are answerable and part of the web-search gap is search quality.

Occupational heterogeneity. Aggregate leaderboards mask wide heterogeneity across occupations, in both level and ranking, at the SOC major-group level (Figure 10 in Appendix A.7). Frontier performs 78-80% on the most comprehensively covered group, Healthcare Practitioners (13 occupations with 52 items), 44-46% on Installation, Maintenance, and Repair and 36-43% on Office and Administrative Support. Some groups are thin: Table 4 in Appendix A.1 shows that Food Preparation and Serving contributes one item and Farming, Fishing, and Forestry one occupation, so both are omitted from the heatmap and results for sparse groups are provisional. This is also seen more clearly at the occupation level in Figure 11 in Appendix A.7. For occupations with three or more items, easily performed occupations like Electricians, Dentists, and Industrial Engineers appear in the 90%+ range across many models. However, there are also a number of occupations that are near zero in frontier models. GPT-5.4 gets 60.3% overall but answers 0 of 12 seed-trials correctly for Actuaries (4 items), 0 of 9 for Sheet Metal Workers and for Fish and Game Wardens (3 items each), and 1 of 9 for Flight Attendants. With three or four items per occupation and 116 occupations, scores of 0% and 100% are partly a small-sample effect and should be read as candidates for a blind spot rather than as measured deficits.

Cross-model gaps can also be very large: for Financial Managers (3 items), GPT-5.4 answers 9 of 9 seed-trials correctly and GPT-3.5 Turbo 0 of 9, a difference that aggregate leaderboards hide. Appendix A.9 gives examples of the hardest items. Items testing numeric thresholds are the hardest kind (Appendix A.12), and occupational accuracy is only weakly associated with how much authoritative documentation an occupation has (Appendix A.13).

Validation. We are conservative about validation: the strongest claim is that ORQA's rankings are not contradicted by independent external evidence. With no authoritative-source occupation benchmark to compare against, we use three external metrics rather than expecting exact agreement with any one. First, GDPval reports long form performance against professional experts in their respective occupations. ORQA accuracy and GDPval win rate correlate strongly across the five overlapping models (Spearman ρ = +0.90, Pearson r = +0.89), though five points give a very wide interval (Appendix A.5). Second, GDPval-AA (Artificial Analysis, 2025) reports a blind pairwise Elo on long form economic tasks, and across the eleven overlapping models Spearman ρ = +0.92 and Pearson r = +0.93. Both leaderboards put the same frontier models on top and the open-weight models at the bottom, with the middle ordering largely preserved. Third, grouping jobs by observed Claude usage in the Anthropic Economic Index (Anthropic, 2025) shows ORQA covering both heavily used jobs and lightly used ones that usage-based measures tend to miss. These comparisons are model-level; they say nothing about whether individual items are accurate or realistic.

Expert audit. To check the items rather than the rankings, 64 screened practitioners across 20 occupations, recruited on an online participant platform, reviewed the items written for their own occupation and produced 255 ratings on 81 items (Appendix A.14). Shown the answer on file with its source passage, they agreed with it in 84% of ratings (96% agreeing or partly) and judged the passage to clearly support it in 82%; five items drew one disagreement each and none drew two. Answering the same items themselves beforehand, without looking anything up, they scored 71.5% [63.2, 79.1] against 57.6% for the frontier models, so items are answerable from professional knowledge but not trivially so. Relevance is the weaker result: 75% of ratings put an item at 4 or 5 out of 5 for relevance to the rater's own work, but 19% at 1 or 2.

5 Discussion

ORQA measures source-grounded agreement with professional guidance rather than occupational AI ability in full. Its advantage is scale: most occupations have authoritative materials setting out safety rules and required procedures, and these convert into testable questions covering far more of the occupational landscape than expert-built work-sample tests. The approach also brings constraints.

First, ORQA treats authoritative materials as ground truth, but much of a profession lives outside them. Because those materials are incomplete and vary in relevance, a correct ORQA answer means agreement with a specified source, not that a model would decide correctly in practice. Second, coverage is broad but uneven: the bank spans 116 occupations in all 21 SOC major groups, yet some occupations and groups carry very few items. Production and healthcare occupations yield many high-quality sourceable items, while groups such as Farming, Fishing, and Forestry are sparse because few authoritative machine-readable sources exist. Thus, estimates at the occupation level for occupations with few items and estimates at the SOC group level for thinly covered SOC groups should be viewed as exploratory rather than as definite rankings. Coverage is also uneven across the economy, and expanding the bank to under-covered occupations and sectors would strengthen claims of generalizability.

Third, sources may appear in pre-training data, raising memorization concerns. We test this with a per-model cutoff stratification (Appendix A.3). For items taken from a passage published after a given model's training cutoff (thus being free from contamination for that model), performance is equal to or better than performance on pre-cutoff passages for 14 of the 15 models analyzed, although the post-cutoff stratum is small for the newest models (3 items for GPT-5.4), so the per-model gaps carry wide uncertainty. A small sub-bank built only from documents first archived after the latest cutoff points the same way. This is in contrast to the performance one would expect under a pure memorization hypothesis. An open-ended variant gives the same ordering, so the ranking is not a multiple-choice artifact. An audit by 64 practising professionals across 20 occupations supports the items themselves: raters agreed with the answer on file in 84% of ratings and judged the cited passage to support it in 82%, while rating 19% of items as of little relevance to their own work (Appendix A.14). ORQA also does not support extended human-LLM collaboration, so it is one element of occupational assessment rather than the whole of it. ORQA's contribution is to turn authoritative resources into source-verifiable tasks far more cheaply than expert-built evaluation, an approach that extends to non-authoritative resources (Appendix A.8).

AI USE STATEMENT

Generative AI is a component of the method in this work, and we disclose its use as follows. Item construction: GPT-4o extracted candidate evidence from the source documents and drafted the question, the correct answer and the distractors for every item; Claude Haiku 4.5 served as the verification model for the ten construction checks; GPT-4o mini, Claude Haiku 4.5 and Gemini 2.5 Flash served as the difficulty pretest; and Claude Opus 4.7 served as the rubric gate. Every item that entered the released bank was additionally reviewed by the authors against its source document. Evaluation: the models listed in Section 4 were queried through their providers' APIs to produce the reported accuracies. We did not use generative AI tools for research ideation, for retrieval or discovery of related work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced.

REFERENCES

Andrew Abbott. *The system of professions: An essay on the division of expert labor*. University of Chicago press, 2014.

Anthropic. The anthropic economic index. Technical report, Anthropic, 2025.

Ruth Appel, Peter McCrory, Alex Tamkin, Miles McCain, Tyler Neylon, and Michael Stern. Anthropic economic index report: Uneven geographic and enterprise ai adoption, 2025. URL https://arxiv.org/abs/2511.15080.

Artificial Analysis. GDPval AA leaderboard, 2025. Online resource.

Alexander Bick, Adam Blandin, and David J. Deming. The rapid adoption of generative ai. *NBER Working Paper No. 32966*, 2024. doi: 10.3386/w32966.

BIG-bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=uyTL5Bvosj.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier De Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. WorkArena++: Towards compositional planning and reasoning-based common knowledge work tasks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Erik Brynjolfsson, Tom Mitchell, and Daniel Rock. What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, 108:43–47, 2018. doi: 10.1257/pandp.20181019.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative ai at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044.

Serina Chang, Ashton Anderson, and Jake M. Hofman. ChatBench: From static benchmarks to human-ai evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 26009–26038, 2025.

Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoë Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical Report 34255, National Bureau of Economic Research, 2025.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Maria del Rio-Chanona and Johanna Einsiedler. Self-building benchmarks: Using AI-generated exams to understand LLM work capabilities. Working paper, Brookings Institution, Center on Regulation and Markets, November 2025.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2), 2026. doi: 10.1287/orsc.2025.21838.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. WorkArena: How capable are web agents at solving common knowledge work tasks? In *International Conference on Machine Learning*, 2024.

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

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.

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

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, 2025. URL https://arxiv.org/abs/2503.04761.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Bill Yuchen Lin et al. WildBench: Benchmarking llms with challenging tasks from real users in the wild. *Preprint*, 2024.

National Center for O*NET Development. O*NET database, 2026. Online resource.

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. Technical report, OpenAI, 2025.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, 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*, 2024.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*, 2024.

Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with ai agents: Auditing automation and augmentation potential across the us workforce. *arXiv preprint arXiv:2506.06576*, 2025.

U.S. Bureau of Labor Statistics. Occupational employment and wage statistics, may 2024 national release, 2024. Online resource.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Zora Zhiruo Wang, Yijia Shao, Omar Shaikh, Daniel Fried, Diyi Yang, and Graham Neubig. How do ai agents do human work? comparing ai and human workflows across diverse occupations. *arXiv preprint arXiv:2510.22780*, 2025.

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? *arXiv preprint arXiv:2603.01203*, 2026.

Michael Webb. The impact of artificial intelligence on the labor market. *SSRN Electronic Journal*, 2020. doi: 10.2139/ssrn.3482150.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Zhiruo Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. TheAgentCompany: Benchmarking llm agents on consequential real world tasks. In *Advances in Neural Information Processing Systems*, 2025.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

We structure our appendices around diagnostics supporting the argument that the questions are of high quality and also on a community-thread version of ORQA that complements the main implementation. Appendix A.1 contains information about bank composition, occupation and source lists, complete source distributions and SOC group information. Appendix A.2 contains per-item diagnostics including pairwise model significance, per-item difficulty distributions, inter-model correctness correlation, and the release date trend across models. Appendix A.3 contains a diagnosis of contamination during pre-training by comparing pre- and post-cutoff accuracy strata. Appendix A.4 contains details about the prompts, list of models used, and inference parameters. Appendix A.5 contains information about external validation using three externally available signals: GDPval, GDPval-AA Elo, and Anthropic Economic Index exposure. Appendix A.6 contains information about the interactive dashboard. Appendix A.7 contains the complete heatmap across occupations. Appendix A.8 contains the community-thread version of ORQA. Appendix A.11 reports the pipeline funnel, including per-check pass and regeneration rates; Appendix A.12 breaks accuracy down by the kind of knowledge tested; Appendix A.13 relates occupational accuracy to documentation footprint; and Appendix A.15 compares closed-book, fixed-retrieval and web-search accuracy.

A.1 COMPOSITION OF BANK AND LISTS

Number of items and occupations All occupations in the ORQA bank and the number of bank items in each occupation are listed in Table 6 in the appendix. Each occupation also has an O*NET SOC code. The final bank consists of 480 items across 116 occupations across all 21 major groups in the SOC. Following the addition of occupation- floor items, 106 of the 116 occupations have three or more items while the other occupations have either one or two items.

Sources: URL, domain and category Every authoritative source URL that contributed at least one item to the bank is listed in Table 7 in the appendix. For each source, the domain, category and URL is listed. Table 3 shows the total distribution of source categories and the most commonly used source hosts. 57.1% of the items come from licensing boards, associations and professional societies. 42.5% of the items come from federal and state government hosts (.gov). Only 0.4% of the items come from educational institutions or .edu hosts. This lack of educational sources is not an emergent characteristic of the bank. Rather, it results from the source filtering explained in Section 3. Academic search is never performed and no research article is ever retrieved because research publishers are not on any whitelist for any occupation. There are zero items in the bank from a research source, and 0.4% of the bank consists of two items from tfsweb.tamu.edu, the Texas A&M Forest Service. The Texas A&M Forest Service is an extension of the state forestry program and publishes practice standards for forestry, not a research organization. The bank consists of 187 different source hosts. The most commonly occurring host in the bank is osha.gov, which accounts for 9.8% of the bank.

A.2 ITEM AND MODEL DIAGNOSTICS

Agreement between the human annotators and the automated gate. The 67% agreement figure quoted in Section 3 covers the 185 annotated items that the automated gate retained. Of those, 34% were rated good by the annotators, 34% okay and 32% bad, so 68% were acceptable to a human reader. Across all 400 annotated items the base rates are 26% good and 32% okay, meaning the gate raises the share of acceptable items by about ten points. A chance-corrected statistic such as Cohen's κ is not defined here because the automated label is constant among retained items, so we report the class balance instead.

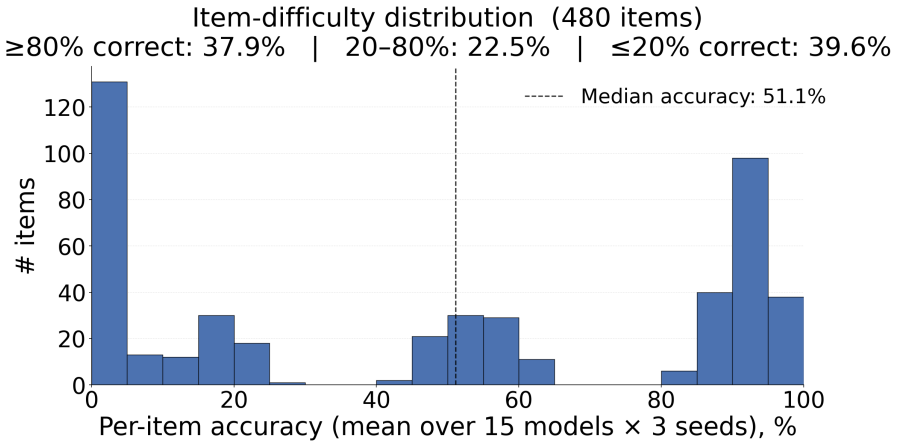
Every item carries "All of the above" and "None of the above" as the fifth and sixth options, and neither is ever keyed as correct. They are kept as a uniform format so that the option list is identical for every item and so that a model that has eliminated the four substantive options has somewhere to go rather than guessing among them. The cost is that a model which learns this regularity faces a four-way rather than a six-way choice, which is why we report the 25% guessing baseline alongside the 16.7% one throughout. A version of the bank in which these two options are sometimes correct would remove the regularity and is left to future work.

Correct option position distribution. The correct answer is in position A for 111 items (23.1%), B for 112 (23.3%), C for 116 (24.2%) and D for 141 (29.4%). Positions A through D hold the substantive answer choices, whose order is randomized across items; positions E and F always hold All of the above and None of the above, and no item in the bank has either as its correct answer. A model that never selects E or F therefore faces a four-way choice, with a chance rate of 25% rather than 1/6.

Pairwise comparisons between models. Table 1 in Section 4 compares adjacent models on the closed-book leaderboard with a paired bootstrap that resamples occupations and then items within each occupation; 5 of the 11 adjacent pairs among the twelve ranked models are distinguishable. The three Pass 7 filter models are excluded from the ranking. For reference, their closed-book scores and 95% intervals are GPT-4o mini 38.3% [33.1, 43.3], Gemini 2.5 Flash 47.0% [41.8, 52.2] and Claude Haiku 4.5 48.2% [42.9, 53.4].

Item-difficulty distribution. Figure 3 displays the distribution of per-item accuracy across fifteen models and three random seeds. We observe a bimodal distribution in which about 38% of items are correct with at least 80% accuracy across all models, 22% straddle the discriminative middle region (between 20% and 80% correct), and 40% are correct with no more than 20% accuracy. The median per-item accuracy is 51.1%. The human-curated set has a significantly higher relative proportion of truly difficult items than the pre-curated set. This explains the heavy lower mode.

Figure 3: The ORQA item bank spans easy, hard, and discriminating middle-difficulty questions.



Note. Each item contributes one mean accuracy over 45 model-seed observations. The curated bank is deliberately weighted toward the discriminating and hard regions of the difficulty distribution.

Per-occupation difference between models. Table 2 displays some instances of inter-occupation differences that are obscured by the aggregate leaderboards. We divide the table into three sections. Section (a) shows occupations in which the difference between the best and worst model is at least 75 percentage points across models. Section (b) shows occupations in which GPT-5.4 achieves zero accuracy despite its overall accuracy of 60.3%. This reveals that even the strongest evaluated model has specific blind spots. Section (c) shows occupations in which the small open-weight Qwen 2.5 7B model (33.9% accuracy) performs the same as the frontier group. These are more relevant to a practitioner choosing a model from the set based on specific tasks rather than the global ordering.

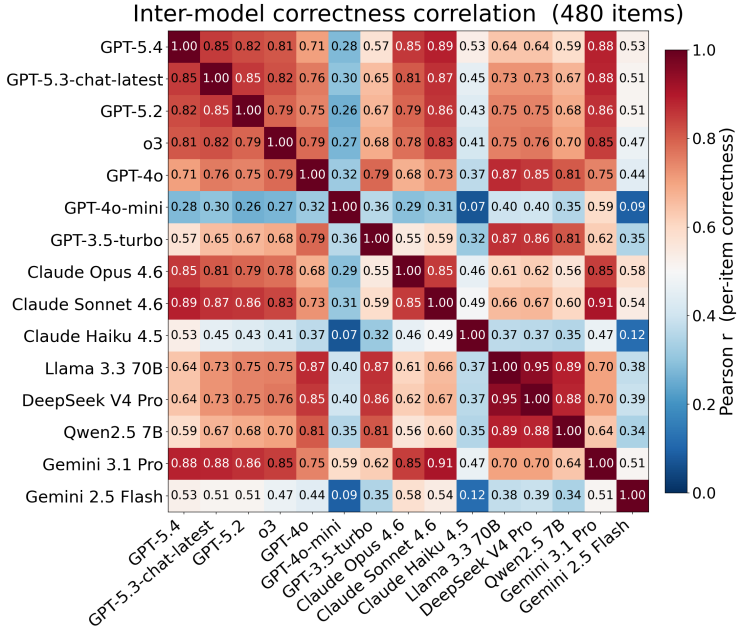
Inter-model correlation. Figure 4 shows the Pearson correlation between models on item-level correctness in the ORQA bank. Across this curated benchmark, the higher performing models all correlate highly with each other across provider families. For example, GPT-5.4 correlates r = 0.94 with Claude Opus 4.6, Claude Opus 4.6 correlates r = 0.93 with Sonnet 4.6 and the open-weight models (Qwen 2.5 7B, DeepSeek V4 Pro and Llama 3.3 70B) all correlate between r = 0.91 and r = 0.92. This makes sense because the ORQA bank appears to test along a professional-knowledge axis. The low correlations all involve GPT-4o mini, Gemini 2.5 Flash and Claude Haiku 4.5. The GPT-4o mini and Claude Haiku 4.5 have the lowest correlation (r = 0.07). The closed-book difficulty

Table 2: Per-occupation discrepancy across models on the ORQA bank. Each cell is per-occupation accuracy (%) averaged over three seeds, restricted to occupations with at least three items. *Spread* is the maximum minus minimum accuracy across all 15 evaluated models. **(a)** Largest cross-model spread. **(b)** GPT-5.4 answers no seed correctly despite its overall 60.3%. **(c)** Qwen 2.5 7B (overall 33.9%) matches the frontier.

Occupation	n	GPT-5.4	Opus 4.6	Sonnet 4.6	Qwen 2.5 7B
(a) Frontier models disagree (largest cross-model spread)					
Financial Managers	3	100	100	89	0
Intelligence Analysts	3	100	100	78	11
Aerospace Engineers	3	78	89	78	22
Healthcare Social Workers	4	75	83	83	8
(b) GPT-5.4 fails entirely (entire frontier weak)					
Actuaries	4	0	0	0	0
Fish and Game Wardens	3	0	11	0	0
Sheet Metal Workers	3	0	0	0	0
Physicists	4	8	17	8	0
(c) Open-weight Qwen 2.5 7B matches the frontier					
Environmental Economists	4	100	100	100	100
Butchers and Meat Cutters	3	100	100	100	100
Industrial Engineers	3	100	100	100	100
Cardiologists	5	80	87	80	80

assessment (Pass 7) used GPT-4o mini, Claude Haiku 4.5 and Gemini 2.5 Flash. During bank creation, items all three answered correctly were removed, as were items all three answered incorrectly when the verifier had marked them ambiguous or low-quality. Selecting items on these models' answers can reduce, and in some cases reverse, pairwise correctness relationships involving them, and such relationships should be analyzed with this in mind.

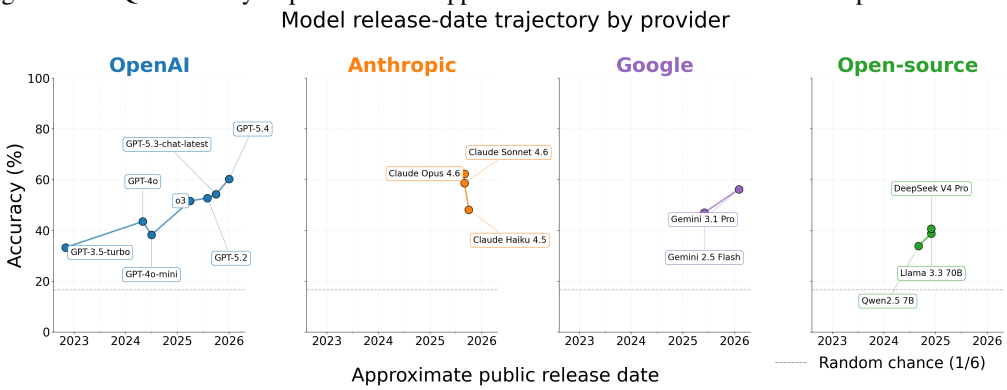
Figure 4: Per-item correctness correlations show which models share the same ORQA successes and failures.



Note. Correlations use binary per-item correctness vectors rather than aggregate scores. Correlations involving the Pass 7 probe models (GPT-4o mini, Claude Haiku 4.5, Gemini 2.5 Flash) may be attenuated by construction.

Release-date trends. Figure 5 displays the accuracy across each model by the estimated date of public release, separately for each provider. For OpenAI, Anthropic and Google, across each of their product lines, accuracy increases with date of release. The open-weight models also show an increase with release date, though weaker. Frontier model performance increases by about 27 percentage points from GPT-3.5 Turbo (33.3%) to GPT-5.4 (60.3%). We note, however, that date of release is only a noisy indicator of effort put into performance and some of the performance increase might be due to more focused post-training on professional material.

Figure 5: ORQA accuracy improves across approximate model release dates within provider families.



Note. Public release date is a coarse proxy for model generation, not a causal variable. The dashed line marks six-option random chance.

Source distribution. The distribution of sources and the top hosting sources is shown in Table 3. Professional licensing and associations, boards and societies account for 57.1% of the content, state and federal government sources account for 42.5%, and .edu sources account for only 0.4%, two items from a state forestry extension service rather than a research venue. The largest host is osha.gov at 9.8% of the bank, significantly under the 30% limit on the source-mix per occupation.

Table 3: Source distribution of the ORQA bank (480 items, 116 occupations, 187 unique source hosts). Top: source category mix as a share of items. Bottom: top ten source hosts by item count.

Source category	Share of items
Professional societies, associations & licensing boards	57.1%
Federal & state government (`.gov`)	42.5%
Academic / `.edu`	0.4%
Top source hosts by item count	
`osha.gov`	47 (9.8%)
`stacks.cdc.gov`	26 (5.4%)
`cdc.gov`	21 (4.4%)
`epa.gov`	12 (2.5%)
`faa.gov`	10 (2.1%)
`docinfofiles.nfpa.org`	8 (1.7%)
`amftrb.org`	7 (1.5%)
`ecfr.gov`	6 (1.2%)
`ashrae.org`	6 (1.2%)
`fjc.gov`	6 (1.2%)

Composition within SOC major groups. Table 4 shows the number of occupations and items within each SOC major group. All 21 SOC major groups are included but there are significant differences in the number of items and occupations within each major group. Healthcare Practitioners, Life/Physical/Social Science and Business & Financial have more than 40 items in each of ten or more occupations. However, Food Preparation & Serving, Farming/Fishing/Forestry, Education & Library and Building & Grounds have only one or two occupations and a small number of items. For these sparse groups, the accuracy within the group should be considered somewhat tentative.

Table 4: SOC major-group composition of the ORQA bank, sorted by item count. Group-level results for thinly covered groups (bottom rows) should be interpreted with caution.

SOC code	Major group	Items	Occupations
29	Healthcare Practitioners & Technical	52	13
19	Life, Physical & Social Science	52	10
13	Business & Financial Operations	44	10
17	Architecture & Engineering	42	9
21	Community & Social Service	32	7
47	Construction & Extraction	31	11
51	Production	26	7
53	Transportation & Material Moving	25	5
43	Office & Administrative Support	23	5
11	Management	21	6
41	Sales & Related	18	5
15	Computer & Math	17	4
33	Protective Service	17	3
27	Arts, Design, Entertainment, Sports & Media	16	5
39	Personal Care & Service	16	4
23	Legal	15	3
49	Installation, Maintenance & Repair	13	4
37	Building & Grounds Cleaning & Maintenance	9	1
25	Educational Instruction & Library	7	2
45	Farming, Fishing & Forestry	3	1
35	Food Preparation & Serving	1	1
	Total	480	116

Model accuracy by occupation wage. Figure 6 displays the relationship between BLS OEWS 2024 annual mean wage and per-occupation model accuracy (average across fifteen models). Each point is scaled according to the total number of employment in the United States. Across all occupations the relationship is weak. The unweighted Pearson correlation between model accuracy and wage is r = +0.22 (p = 0.02). Over the entire range of wages there is also approximately the full range of model accuracy. However, when employment across the nation is used as a weight for occupations the relationship becomes almost non-existent (r = -0.02, with a weighted fit close to zero). We only perform a descriptive analysis. There does not appear to be a strong link between occupational wage and accuracy on procedural knowledge of professional tasks using closed book exams. There also does not appear to be a significant concentration of AI ability within the high-wage share of the labor market when weighted by employment. We present this as an inductive result using our data and do not attempt to test a specific hypothesis here. A more thorough analysis can be performed in future work.

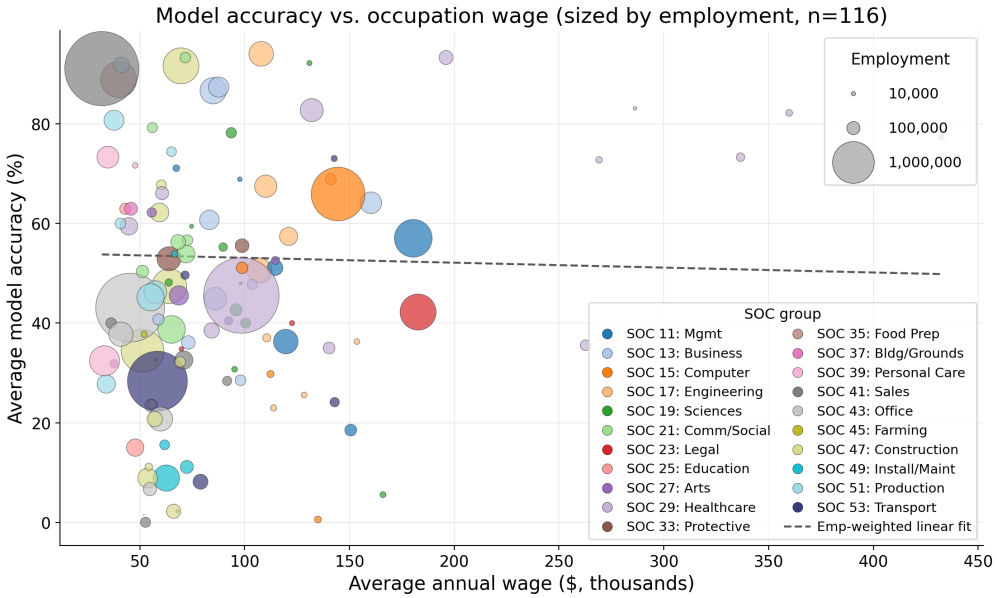
Digital vs. non-digital jobs. We also perform a final job level split based on whether the job is digital or non-digital. Digital jobs are defined as analytical or information-intensive office based jobs. Non-digital jobs are defined as field, in person, clinical, or hands on jobs. We find that performance is fairly even between digital and non-digital jobs (48.1% of jobs in the digital category vs. 47.6% in the non- digital category on average across models). Both digital and non-digital contain some of the most challenging jobs (Figure 7).

Position of the correct answer. Across the four substantive options the correct answer falls at A for 23.1% of items, B for 23.3%, C for 24.2% and D for 29.4%. The last of these is 141 items against 120 expected under a uniform draw, which a binomial test puts at $p = 0.03$ with a 95% interval of 25.3% to 33.7%. The imbalance is small and in a single cell, but it is real, so per-position accuracy is worth checking before reading any model's preference for a late option as a finding.

A.3 Diagnosis of Pre-training Contamination in Closed-Book Tests

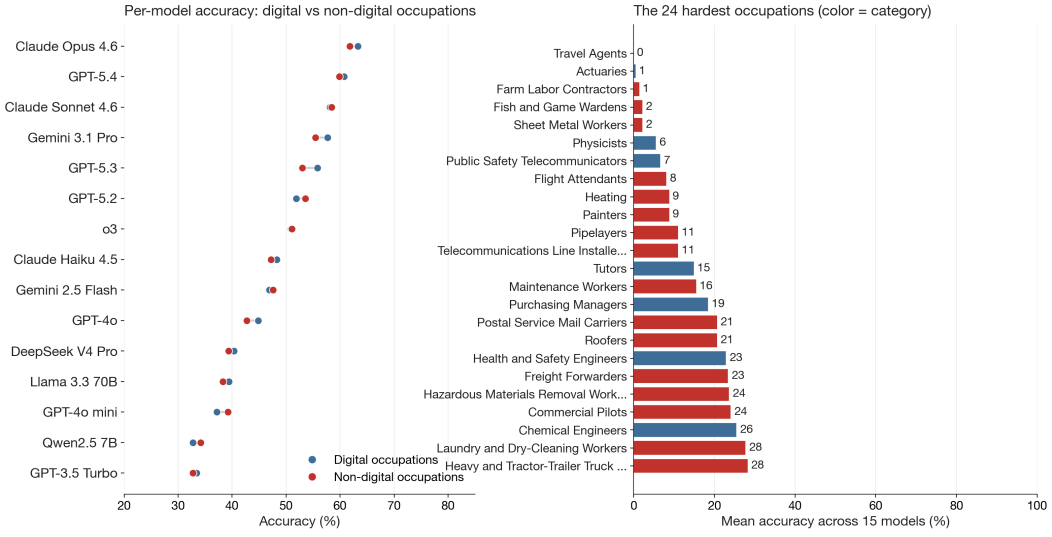
A potential issue in closed-book tests against publicly available authoritative sources is that frontier models may have encountered the same sources during pre-training and can select the correct answer

Figure 6: Average ORQA accuracy is weakly related to occupational wage, and unrelated after employment weighting.



Note. Marker size reflects national employment, so visually prominent points are large occupations. The dashed trend is the employment-weighted fit and is descriptive; it should be read alongside the broad dispersion at each wage level.

Figure 7: Model accuracy on digital versus non-digital occupations, and the hardest occupations by category.



Note. Occupations are hand-classified into digital (information and analytical desk work) and non-digital (physical, hands-on, in-person, clinical, or field work). Left: each model's mean accuracy on the two groups. Right: the twenty-four lowest-accuracy occupations, coloured by category.

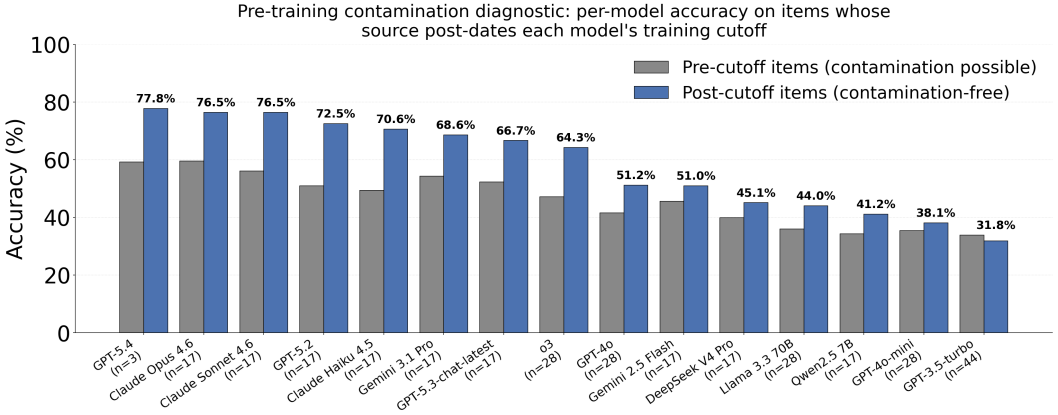
via string-level rather than professional knowledge. We mitigate this by performing model- level splits on accuracy based on whether a source for a question appears after a given model's pre-training date. By definition, questions with sources after a given model's pretraining date are free from contamination for that model. If memorization were causing overestimation of ORQA's performance then the accuracy on post-cutoff sources should be consistently lower than pre-cutoff sources.

Mechanism for source-date assignment. We attempt to determine the publication date of a source from its URL using host-specific heuristics. We use the years in path elements like /2024/ and the fiscal year tag /fy23/ in URLs for OSHA grants. For content published on a low frequency but with URLs that do not contain date information (EPA, FAA, large professional society sites, OSHA directives, CDC stacks repository, BLS Occupational Outlook Handbook), we use a conservative prior date at the host level. URLs that do not conform to either heuristic are dropped. 86 (17.9%) items are dated via URL pattern, 133 (27.7%) items are dated via host-level prior and 261 (54.4%) items are undatable and excluded. 219 dated items form the basis of the analysis.

Per-model cutoffs. The cutoff dates used for pre-training are derived from the most recent published model card for each provider or, if a cutoff is not publicly provided, from the conservative cutoff implied by the model release date: GPT-3.5 Turbo (Sep 2021), GPT-4o / GPT-4o-mini / o3 (Oct 2023), Llama 3.3 70B (Dec 2023), DeepSeek V4 Pro (Jul 2024), GPT-5.2 / GPT-5.3-chat-latest (Sep 2024), Qwen 2.5 7B (Oct 2024), Gemini 2.5 Flash (Dec 2024), Claude Opus 4.6 / Sonnet 4.6 / Haiku 4.5 (Jan 2025), Gemini 3.1 Pro (Mar 2025) and GPT-5.4 (Jun 2025).

Result. Figure 8 shows per-model performance on pre- and post-cutoff strata. Generally performance is equal or higher on the post-cutoff strata than on the pre-cutoff strata: GPT-5.4 achieves 77.8% performance on its (small) post-cutoff stratum (n = 3) compared to 59.3% pre-cutoff (+18.5 pp), Claude Opus 4.6 achieves 76.5% compared to 59.6% (+16.9 pp), Claude Sonnet 4.6 achieves 76.5% compared to 56.1% (+20.4 pp), and GPT-5.2 achieves 72.6% compared to 51.0% (+21.6 pp). Accuracy due to contamination during pre-training seems unlikely given that there is not a systematic pre > post result across all fifteen models. Rather, fourteen of the fifteen models have post-cutoff accuracy equal to or greater than pre-cutoff accuracy (the exception is GPT-3.5 Turbo which has -2.1 pp). This suggests the diagnostic is more likely due to transfer of professional knowledge rather than exact knowledge from pre-training. We also note that the post-cutoff samples for the most recent frontier models are relatively small (n = 3 for GPT-5.4). This results in large statistical uncertainty when comparing pre- vs. post-cutoff performance for these models, and the per-model differences for the newest models should not be read as estimates. A solution to this is to create an after-cutoff sub-bank as described below.

Figure 8: Post-cutoff source items do not show the accuracy drop expected from pre-training memorization.



Note. Source dates combine URL-derived dates and conservative host-level priors; undated items are excluded from this diagnostic. Post-cutoff samples are small for the newest models, so this is a directional contamination check.

Future work: an after-cutoff sub-bank. The above diagnostic relies on source-date heuristics and has small post-cutoff samples for the newest models. As a stricter check, a separate post-cutoff sub-bank was built and kept apart from the 480 items. Of 19,675 saved evidence cards, 1,265 (from 325 documents) carry a publication date of July 2025 or later, after the latest training cutoff in the slate (GPT-5.4, June 2025). Because a document can be re-dated when it is revised, a card was kept only if the Internet Archive holds no capture of its source URL before July 2025; this removed 178 documents and left 645 cards from 147 documents. These cards were run through the production card path with the same checks, difficulty probe and good-rubric gate. Only 8 items survived (186 cards failed to build, 344 were rejected by the verifier, 73 by the difficulty probe and 34 by the good-rubric gate), from 6 occupations in 4 SOC major groups. Frontier models (Claude Opus 4.6, GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro) answer 86.5% of these items correctly (95% item-bootstrap CI [72.9, 96.9]), well above their main-bank accuracy, and GPT-4o mini (25.0%) and Gemini 2.5 Flash (41.7%) remain below them. Frontier accuracy is thus not lower on material that cannot have been in pre-training, consistent with the stratified diagnostic. The sub-bank is too small to rank models or to rule out that post-cutoff items are easier: 53% of post-cutoff items reaching the difficulty probe were rejected as too easy, which suggests that many recent documents restate established rules. A larger contamination-free sub-bank needs date-restricted source discovery, which was not possible within the available search budget.

Source discovery and whitelist construction. Discovery uses the Serper search API against Google. Every query is restricted to a single whitelisted domain with a site: filter. A first round pairs the occupation title with nine document-type templates covering standards and guidelines and manuals and procedures and codes of ethics and certification handbooks and exam content outlines and competency documents, plus a filetype:pdf variant. Later rounds combine the occupation with topical keywords across the same domains and add a small set of .gov and .edu fallback queries for cross-cutting topics. Each query returns up to ten results.

The whitelist is built per occupation before discovery. For each occupation we take the associations that O*NET lists for that SOC code as one pool and the full O*NET association list of roughly 3,100 organizations as a second pool. A single model call then selects 8 to 15 certifying or licensing or standards organizations for that occupation together with up to three state licensing boards, which O*NET does not list. A selection is kept only if it appears in one of the two pools, state board URLs are checked for a live response before entering the list, and the result is combined with a small .gov baseline of cross-cutting safety and statistics hosts.

Host prior dates used in the contamination split. When a source URL carries no datable pattern, the item is assigned a conservative prior date for its host: 2010 for obis.osha.gov; 2018 for www.osha.gov and archive.cdc.gov and www.atsdr.cdc.gov; 2019 for www.epa.gov and wwwnc.cdc.gov; 2020 for www.cdc.gov and stacks.cdc.gov and www.faa.gov and www.fmcsa.dot.gov and www.fhwa.dot.gov and www.ashrae.org and www.apcointl.org and www.asisonline.org and www.nfpa.org and docinfofiles.nfpa.org; 2021 for www.americanbar.org; and 2022 for www.bls.gov, whose occupational handbook is refreshed annually. Hosts outside this list with no datable URL are left unclassified rather than assigned a prior.

A.4 Prompts, Models, and Inference Settings

Evidence card and item creation prompts and quality check configuration. The prompts used for evidence card extraction, item creation, and all quality checks are provided in the supplementary material in their entirety. The prompt for evidence card extraction instructs the model to return a JSON array of cards. Each card consists of a source excerpt (one to three sentences), a task or procedure from the excerpt, an occupation domain, and a quality rating. The item-builder prompt specifies that there must be six answer choices, that All of the above and None of the above must be present, that only one answer must be correct, and that the correct answer must be entailed by the source excerpt.

Details for each quality check. The quality checks described in Section 3 are, in order: (1) source entailment, which verifies that the correct answer is entailed by the source excerpt. (2) question stem sanity, which verifies that the stem is professionally presented as a well-constructed question. (3) distractor plausibility, which verifies that all distractors are domain-plausible. (4) no distractor

paraphrases another answer, which verifies that no distractor paraphrases another answer. (5) only one answer is correct, which verifies that exactly one of the six answers is correct. (6) Length and structural sanity: option lengths are within a factor of two of each other; no option is longer than 350 characters; and no phrases that indicate the source of knowledge. (7) Closed-book difficulty is estimated using GPT-4o mini, Claude Haiku 4.5 and Gemini 2.5 Flash in parallel; any questions that can be answered by all three models are deemed too easy, and questions all three answer incorrectly are rejected when the verifier has marked them ambiguous or low-quality. (8) Eliminability is determined by an external LLM judging if each incorrect answer can be eliminated using only general world knowledge. (9) Auditing on concreteness checks for vagueness in the correct answer if any incorrect answer refers to a specific value or tool, and will then rewrite the correct answer to include the concrete information from the source. All of these are passed through to a categorization step that was tuned using human labels (Section 3).

Details on models and inference. We evaluated the following models: GPT-5.4, GPT-5.3-chat-latest, GPT-5.2, o3, GPT-4o, GPT-4o mini, GPT-3.5 Turbo, Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5, Gemini 3.1 Pro, Gemini 2.5 Flash, Llama 3.3 70B, DeepSeek V4 Pro and Qwen 2.5 7B. Every evaluated model is queried at temperature 0 with seeds 0, 1 and 2, and answers are capped at a single letter for multiple choice and 300 tokens for open-ended. Requests use each provider's current stable endpoint at the time of the run, and open-weight models are served through Together AI. Open-ended answers are graded by three judges from three providers: gpt-4o, claude-haiku-4-5 and gemini-3.1-pro. Each returns a yes or no verdict at temperature 0 and an item counts as correct when at least two of the three say yes.

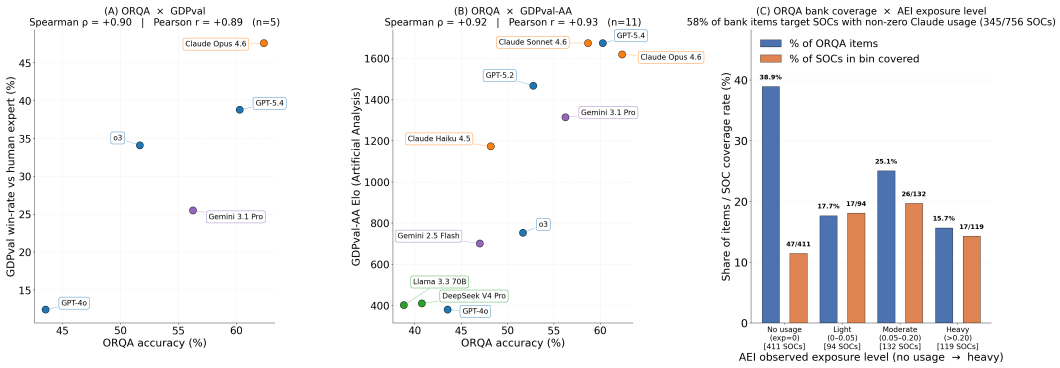
Models on the pipeline: generator and verifier For the bank construction pipeline there is a fixed set of generator and verifier models used across all items. On the generator side (OpenAI): item creation and evidence card extraction use gpt-4o (temperature = 0) and search/discovery and planning uses gpt-4o-mini. On the verifier side (separate from the generator, using Anthropic): the separate verifier, claude-haiku-4-5 (temperature = 0), performs the checks for: leakage, ambiguity, alignment, entailment, eliminability and concrete detail. The source authority check is implemented entirely in code using a whitelist per occupation and tier attribute. The only multi-model check is the difficulty pretest that runs in parallel gemini-2.5-flash (Google), claude-haiku-4-5 (Anthropic) and gpt-4o-mini (OpenAI). There are three models across three providers in order to access three separate rate limits and to achieve some independence in the errors. The use of separate providers for the generator (OpenAI), independent verifier (Anthropic), pretest (OpenAI, Anthropic and Google) justifies the use of the term independent verifier and also provides some resilience to self consistency error in the GPT-4o generator.

A.5 External Validation Using External Signals

We provide a figure (Figure 9) to complement the external validation presented in the main text. For the five models that overlap GDPval (panel a), there is a Spearman ρ = +0.90 and a Pearson r = +0.89 correlation between the accuracy of ORQA and the performance of GDPval against industry professionals. With only five models this is a weak test: the rank correlation is not significant at conventional levels (exact two-sided p = 0.08) and the 95% interval on the Pearson correlation runs from +0.04 to +0.99, so a single model can move it. For the eleven models that overlap the GDPval-AA leaderboard for Artificial Analysis (panel b) there is a Spearman ρ = +0.92 and a Pearson r = +0.93. Panel (c) compares the coverage of ORQA to the observed exposure of occupations to Claude usage in the Anthropic Economic Index. The 6-digit SOC codes are split into heavy, moderate, light, and no usage categories. 38.9% of ORQA questions fall into the no usage category, 17.7% into light, 25.1% into moderate and 15.7% into heavy. These four shares cover the questions whose SOC code could be matched to the index and therefore do not sum to 100. 61.1% of bank questions use SOC codes that have non-zero observed exposure to Claude. ORQA covers 11%, 18%, 20%, and 14% of the SOC codes in each of these categories. This suggests the bank covers both frequently and infrequently used jobs beyond just those currently being adopted by AI. We also compare agreement per occupation to GDPval.

Per-occupation agreement with GDPval. GDPval also provides per-occupation performance information. This information is only provided in a small visualization within the GDPval site and

Figure 9: External benchmark rankings and usage exposure provide directional validation for ORQA.



Note. Panels compare ORQA with GDPval, GDPval-AA, and Anthropic Economic Index exposure where model or SOC coverage overlaps. These are triangulation checks for ranking and coverage, not claims that any external signal is the benchmark ground truth.

not in a tabular form. Nine of the GDPval occupations are also in the current ORQA bank. We report GDPval win-rate in coarse buckets High / Mid / Low (High $\geq$50%, Mid 25-49%, Low < 25%), and ORQA closed-book accuracy (averaged across fifteen models and three random seeds) for these. At the occupation level the two do not agree. The occupation in the GDPval-Low bucket scores second highest on ORQA, the GDPval-High occupations span 32% to 64% on ORQA, and the GDPval-Mid occupations span 28% to 94%. We therefore draw no occupation-level conclusion from this table and highlight the largest discrepancies instead. Nurse Practitioners are in the GDPval-Low bucket but achieve 82.8% on ORQA. This may be due to the use of evidence from CDC and state board sources for the nurse practitioner questions in ORQA, which enables strong closed-book performance. Real Estate Brokers are in the GDPval-Mid bucket but only achieve 28.4% on ORQA. Lawyers are in the GDPval-High bucket but only achieve 42.2% on ORQA. This may be due to the differences in output expected by ORQA (identification of the specific provision entailed by the source) and GDPval (ability to produce long form responses). We note that the ORQA scores per occupation in Table 5 are based on between 3 and 9 questions per cell. Thus, the scores within a cell are noisy. Table 5 is intended as a qualitative sanity check, not an exact match.

Table 5: Per-occupation cross-check between ORQA and GDPval on the nine occupations present in both. *Items* is the number of ORQA items contributing to the per-occupation accuracy. *ORQA acc.* is the mean over fifteen models $\times$ three seeds. *GDPval bucket* is a coarse three-way classification of GDPval win-rate against industry-professional experts (High $\geq$ 50%, Mid 25–49%, Low $<$ 25%). Rows are sorted by ORQA accuracy descending.

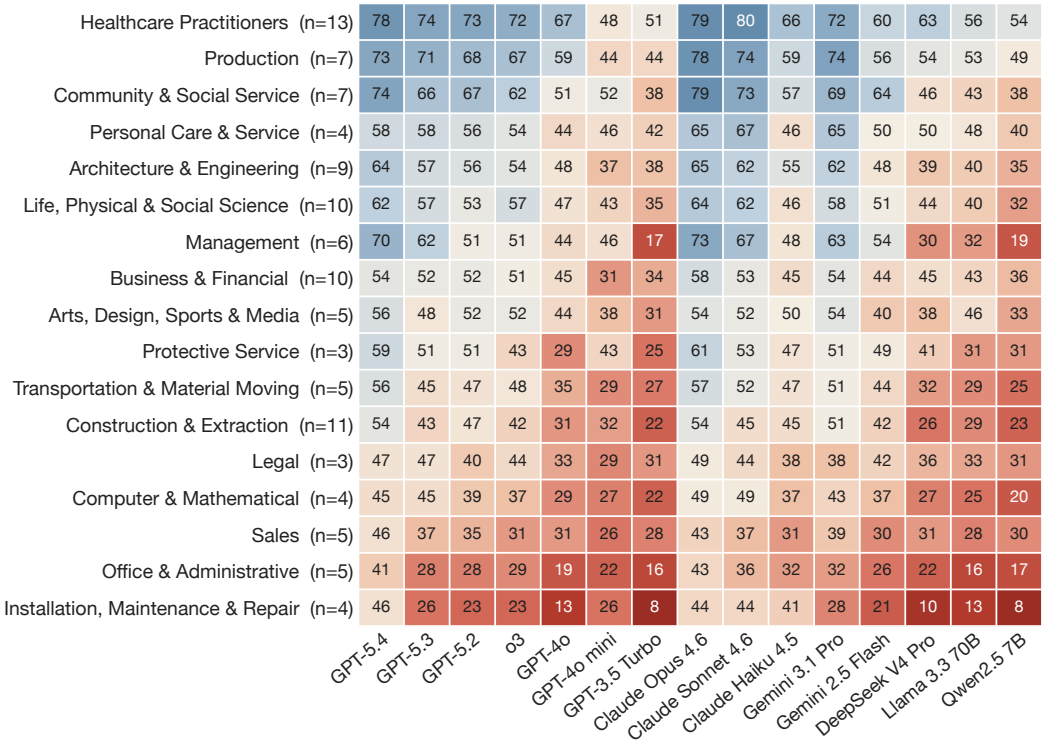
Occupation	Items	ORQA acc.	GDPval bucket
Industrial Engineers	3	94.1%	Mid
Nurse Practitioners	4	82.8%	Low
Mechanical Engineers	8	67.5%	Mid
Software Developers	3	65.9%	Mid
Personal Financial Advisors	8	64.2%	High
Financial Managers	3	57.0%	Mid
Lawyers	6	42.2%	High
Real Estate Sales Agents	3	32.6%	High
Real Estate Brokers	9	28.4%	Mid

A.6 Interactive Dashboard

A dashboard is provided online that enables users to view per-model and per-occupation results, filter by source domain and source category, and click through to individual items and their source texts. This enables the user to examine the evaluation without re-running any of the models while also providing access to the underlying data for further use.

A.7 Full Per-Occupation Heatmap

Figure 10: SOC major-group heatmap of ORQA accuracy by model.



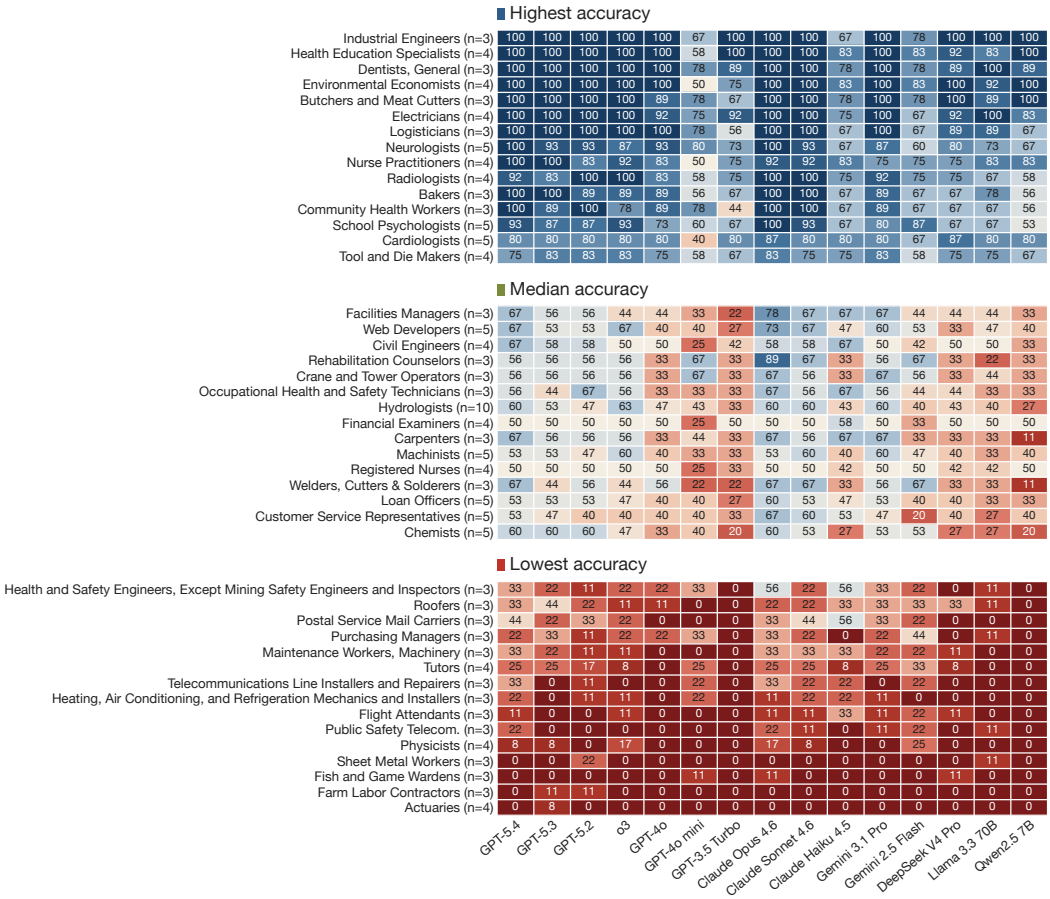
Note. SOC rows aggregate item-level accuracy within broad BLS occupational groups. Each cell averages all retained items in the group across the model's three seeds. Groups vary widely in how many occupations they contain, and thinly covered groups should be interpreted with that in mind.

The per-occupation heatmap used in the paper (Figure 11) displays 45 occupations sampled from the 106 that each have at least three items, ordered by mean accuracy across the models and grouped into highest, median and lowest bands so that easy and hard domains are visually divided. In the released data, there are 116 occupations, some having only one or two items. The trends seen in the displayed heatmap also hold across the full set.

A.8 Community-Thread Version

We also created and tested a parallel community-thread version of ORQA. In this version, the correct answers are community-upvoted answers from public discussions about occupations instead of answers from authoritative sources. Authoritative answers describe what professionals should do, while community threads describe what professionals in fact ask when they encounter a problem at work. These two versions describe different dimensions of real-world professional consensus. This version is reported separately (and created separately) for completeness. It is not part of the 480-item authoritative dataset discussed above.

Figure 11: Occupation-level heatmap showing model-specific strengths and failures.



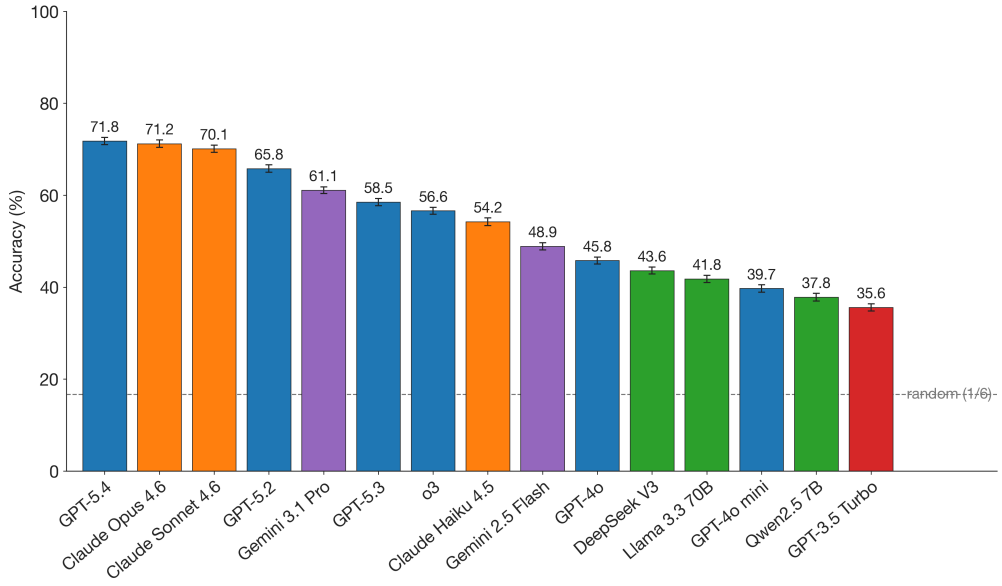
Note. Occupations are ordered by cross-model mean accuracy so that easy and hard domains are visually separated. Cell values summarize model accuracy within an occupation and should be interpreted with item counts in mind.

Pipeline. The community-thread pipeline begins with the list of occupations from O*NET. It groups jobs using SOC codes and retains one job per group. For each job, it searches public discussion forums for relevant community discussions (e.g., Stack Exchange, Metafilter, and other sites dedicated to occupations). Then, using GPT-4o, each community post is transformed into a multiple choice question with six possible answers. This version attempts to follow the style of queries that professionals might make in the real world (e.g., in WildBench (Lin et al., 2024)) but adapted to occupations. The community top-voted answer serves as the correct answer, other substantive answers are derived from the community post, and All of the above and None of the above are always included as answer choices. A four stage filtering system is used to improve on assistance framing, semantic relevance, plausibility of distractors, and the level of agreement in the community. Two limits of this variant should be kept in mind. Its ground truth is a community-upvoted answer rather than an authoritative source, so it can record a popular workaround instead of the professional standard, and no practitioner audit was run on it. Public forums are also heavily represented in pre-training corpora, and no contamination check of the kind in Appendix A.3 was performed here, so its higher accuracy is not evidence about memorization either way.

Final bank. The community-thread bank has 1,444 items in 758 occupations. Deduplication was performed based on SOC codes and these are nearly complete with respect to the O*NET list. Thus there are approximately 1.91 items per occupation. Having only two items in an occupation is allowed because this version of the system prioritizes coverage of occupations over within-occupation statistical power.

Leaderboard. The closed-book leaderboard for the community-thread bank is shown in Figure 12. The frontier cluster is present in both the authoritative and community banks. Performance is more accurate on the community bank because community items are generally more answerable using general domain knowledge than items from authoritative regulations.

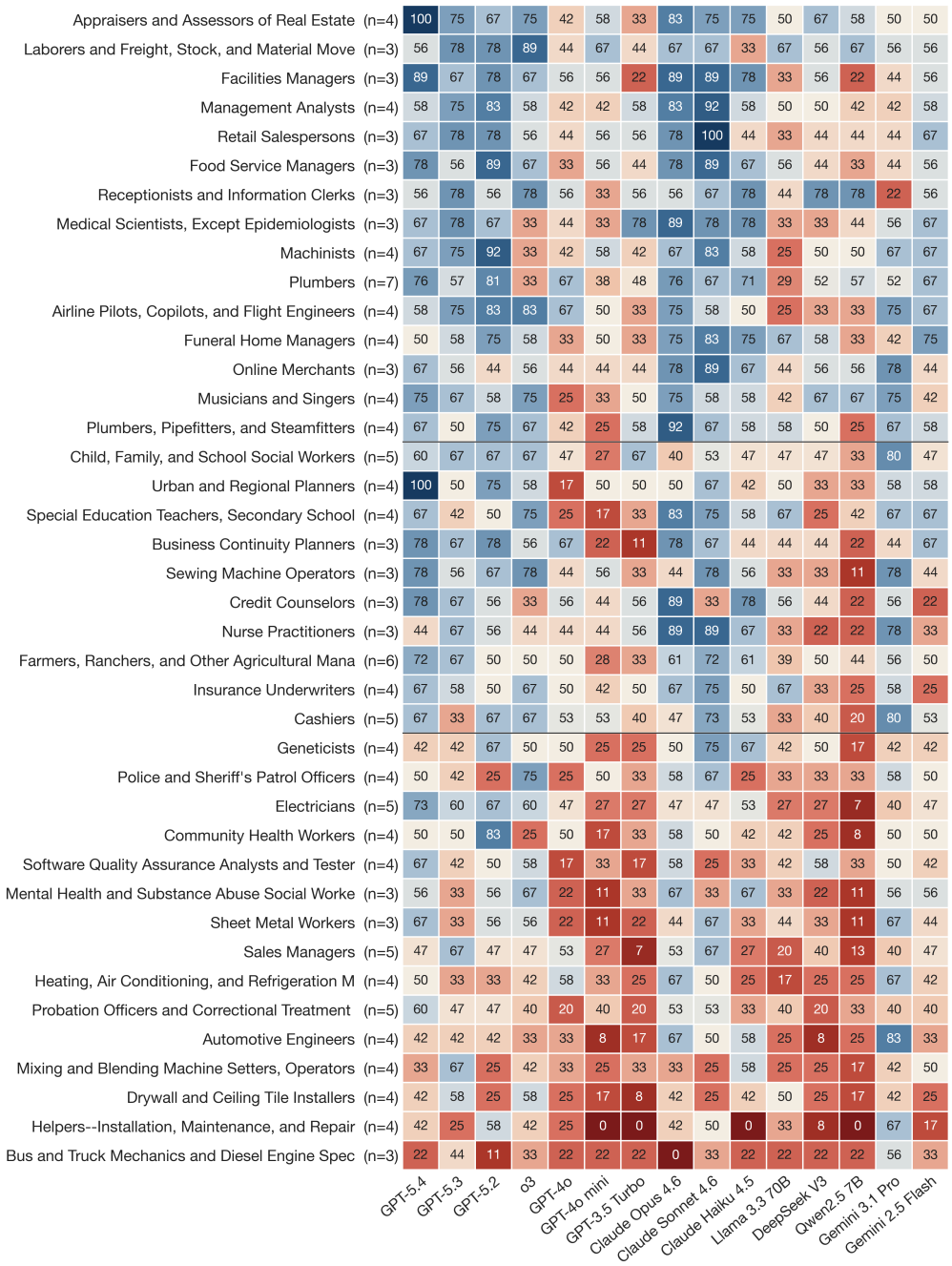
Figure 12: The community-thread variant preserves the frontier cluster under forum-derived ground truth.



Note. The community variant uses forum-derived ground truth over a separate 1,444-item, 758-occupation dataset. Scores are a complementary stress test and are not directly interchangeable with the authoritative-source ORQA bank.

Per-occupation heatmap. A stratified per-occupation heatmap for the community-thread bank is shown in Figure 13. Occupations were selected to sample across the full range of cross-model average correctness and are ordered from highest to lowest. A heatmap containing all occupations with three or more items is provided in Figure 17.

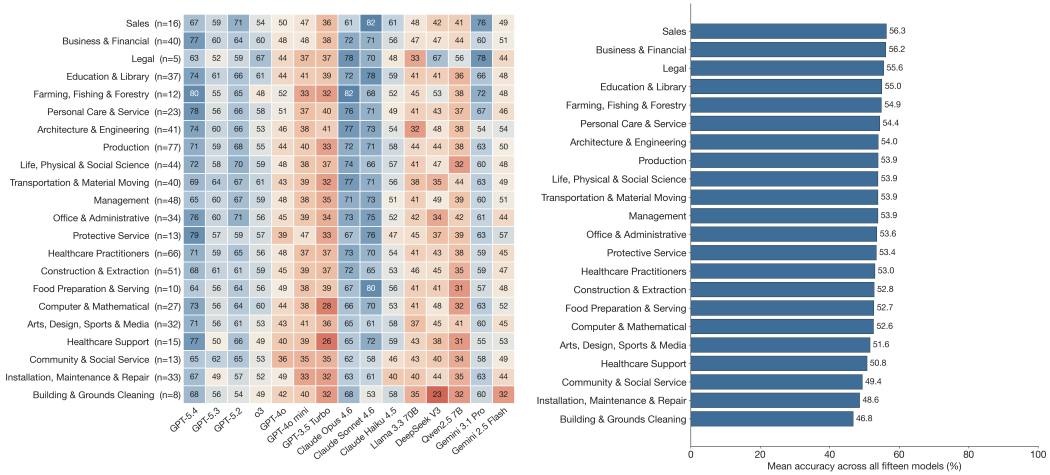
Figure 13: Community-thread performance varies substantially across occupations.



Note. Displayed rows are sampled for readability from the top, middle, and bottom of the community-thread distribution. Occupation-level cells with few items should be read as diagnostics, not stable estimates.

SOC major-group aggregation. Community-thread results are aggregated to SOC major groups in Figure 14. The left panel is a cluster-level heatmap. The right panel is a cluster-level leaderboard that computes group averages first within the major groups of SOC occupations and then across the major groups. All major occupation groups are given equal weight in the leaderboard. The ordering of the fifteen models remains the same across the leaderboard.

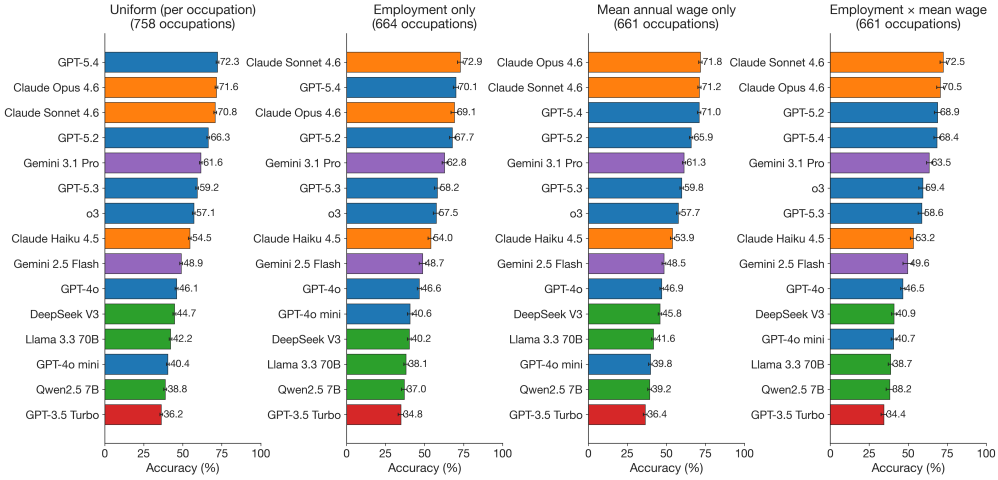
Figure 14: SOC-level community-thread results summarize model performance by occupational family.



Note. The left panel shows SOC major-group accuracy by model; the right panel averages within each group and then across groups.

Economic-importance-weighted leaderboards. Three economic-importance-weighted leaderboards are reported in Figure 15 across the community- thread bank. Additionally, the unweighted leaderboard is also reported. The occupations are weighted according to the number of people employed in the nation, according to the mean wage per year, and according to the product of the first two. In all three cases, the ordering of the fifteen models remains the same.

Figure 15: Community-thread leaderboards remain stable under labor-market weighting.

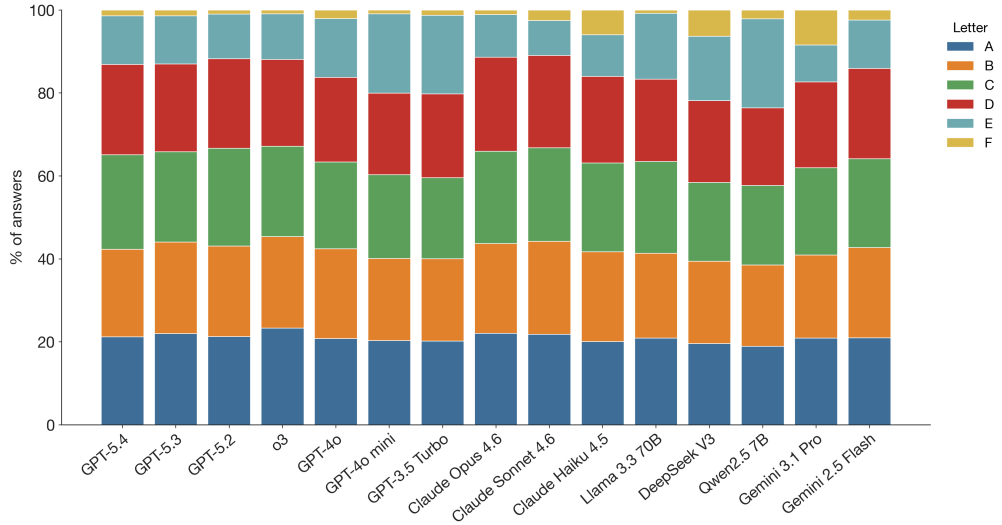


Note. The three weighted variants assign occupation weights proportional to total employment, annual mean wage, and their product. The underlying model responses are unchanged; only each occupation's contribution to the final average varies.

Positional bias and distribution of selected answers. Figure 16 reports the distribution of selected answers (A to F) in the community-thread bank. For the majority of models, answers are distributed

fairly evenly across A-D (20-23% for each) and answer E (All of the above) is chosen about 10-13% of the time. Answer F (None of the above) is chosen infrequently.

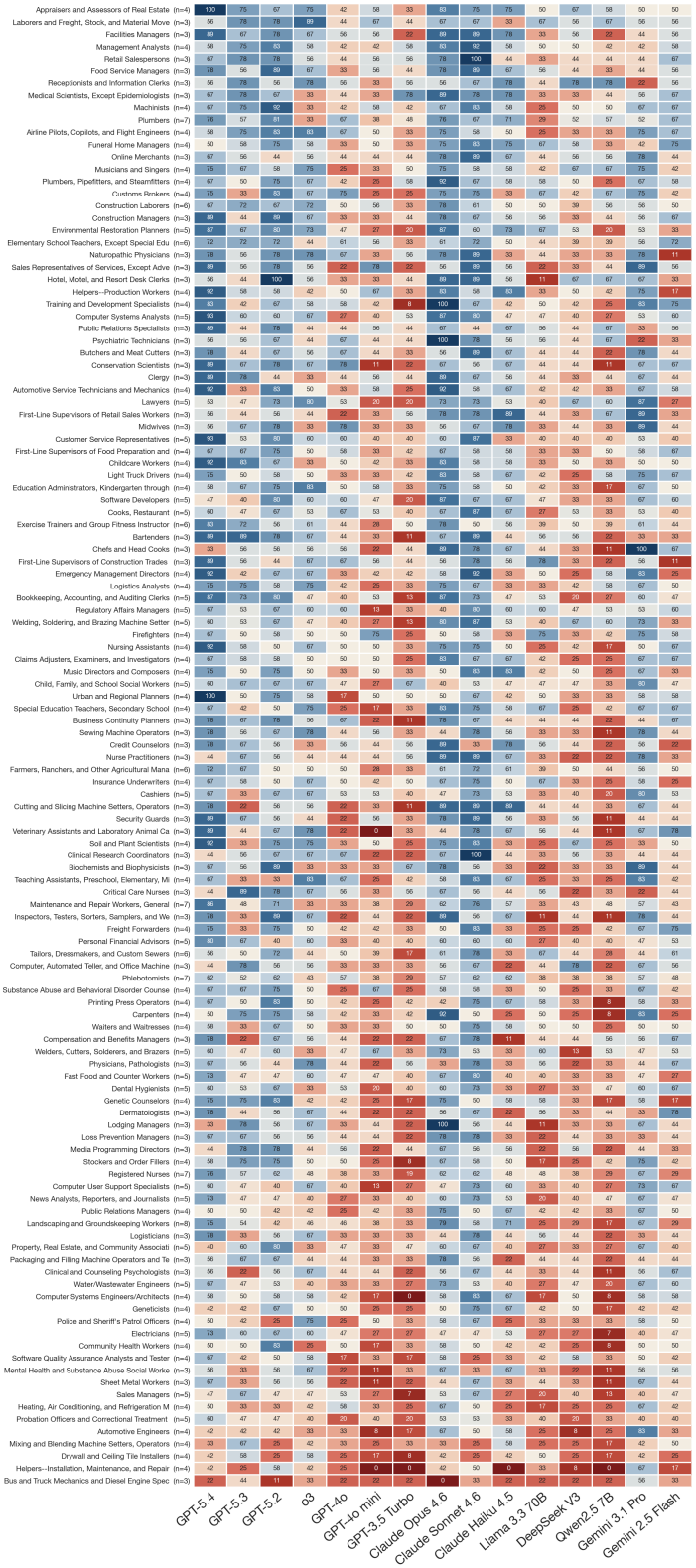
Figure 16: Most community-thread runs show no strong answer-position bias.



Note. Options A–D are randomized substantive choices; E and F are the fixed *All of the above* and *None of the above* slots.

Full community per-occupation heatmap. Figure 17 shows the per-occupation heatmap of correctness for the community-thread answers for occupations that have at least three questions. The reliability criterion used in the authoritative per-occupation heatmap is also used here. Occupations with fewer than three questions are excluded from this heatmap because the per-occupation cells would be based on only one or two questions.

Figure 17: Community-thread occupation-by-model heatmap across the occupations with at least three items.



Note. Rows are the community-thread occupations with at least three items, sorted by cross-model mean accuracy; columns are the fifteen evaluated models. Occupations with fewer than three items are omitted so that each cell rests on a more stable per-occupation estimate.

Table 6: Schema for the occupation roster table included in the supplementary material. The full table lists every occupation in the ORQA bank together with its SOC code and item count.

Occupation title	SOC code	Items
Firefighters	33-2011	11
Hydrologists	19-2043	10
Real Estate Brokers	41-9021	9
Pest Control Workers	37-2021	9
Freight Forwarders	43-5011	9
...	...	...

Table 7: Schema for the source roster table included in the supplementary material. The full table lists every authoritative source URL used to construct an item, together with source domain and category.

Item ID	Source domain	Source URL	Category
Q001	osha.gov	osha.gov/laws-regs/...	Federal
Q002	amftrb.org	amftrb.org/...	Professional / board
Q003	nfpa.org	docinfofiles.nfpa.org/...	Professional
Q004	faa.gov	faa.gov/...	Federal
...	...	...	...

A.9 What Difficult Items Look Like

One of the benefits of ORQA is that it can access to jobs that are not typically included in job assessments because they are in the physical or non-digital world. Also, difficult items can be concrete instead of abstract. There are a few trends that can be pointed out. One is a failure to accurately recall specific codes and numeric values. For example, in an OSHA item regarding Slaughterers and Meat Packers, a specific exposure limit is referenced when a piece of equipment makes impact or impulse noises. The models do not appear to do well in recalling these specific values in closed-book. Another example is in an item for Tool and Die Makers which references a specific machine safety document. The question requires what must be checked prior to inching the ram on a punch press. The correct answer is a specific action rather than something like follow all safety procedures. Second, many failures are retrieval-shaped. The correct response to a question is some piece of fact from some authoritative source that a web-search-capable model can retrieve, but not a closed-book model. An example from Flight Attendants uses a question from GoJet's collective bargaining agreement regarding the boarding pay for flights that are delayed and subsequently deplaned and reboarded. A closed-book model performs very poorly on this question, but a web-search capable model performs well because it can retrieve the collective bargaining agreement. These failures illustrate both the kind of knowledge that ORQA evaluates (professionally relevant knowledge that is not currently in frontier models) and also the possibility for a large gap between closed-book performance and web-search performance.

A.10 Excluding Research Papers at the Source

A research paper is not something that a professional is accountable for. Instead, it reports a result from a study. Because of this, any object that uses such a source fails in our grading scheme as a research_paper_source and is not eligible for a good grade. Such an object is essentially useless regardless of quality, and these sources are not used at all by the pipeline (no academic publishers are whitelisted by any occupation and no academic search is performed). Eliminating these sources before fetch, rather than rejecting them after generation, is safer (a paper that never enters the pipeline cannot survive a grader's mistake) and more cost effective (no extraction, generation, grading or fetch budget is wasted on objects that are guaranteed to be rejected). This distinction is made by document type and not by hosting source. Third party research is often archived by authoritative institutions. For example, a conference paper hosted on a .gov domain is still considered a research paper.

A.11 PIPELINE FUNNEL

Table 8 reports the pipeline's stage counts for the runs that produced the bank, taken from the per-occupation run logs. Across 317 occupation runs, the pipeline fetched 33,600 source URLs; 26,693 of them yielded no usable evidence card, 7,231 cards failed to become items, and of the items built, the independent verifier rejected 12,703, the Pass 7 difficulty probe rejected 544 as too easy and 789 as too hard, and domain balancing rejected 1,955, leaving 1,422 items. Human-calibrated categorization and occupation-floor balancing then produced the 480-item bank. The production logs record rejections per stage but not per verifier check, and they do not count regeneration attempts. To measure these, the pipeline's card path was instrumented to log every candidate's outcome, the result of each verifier check and the number of regeneration attempts each check triggered, and 1,885 saved evidence cards from 39 bank occupations (about two per SOC major group, up to 50 cards each) were replayed through it with the production prompts and models (Tables 9 and 10). 73% of cards were built into an item; most build failures were options containing a giveaway phrase. Of the 1,366 items verified, 24% passed every check. Two checks do most of the filtering: 19% of items fail because the correct answer is not fully entailed by the source quote, and among items that reach the final check, 61% still have a distractor that general background knowledge eliminates after the allowed regenerations. Regeneration is the norm rather than the exception: 83% of verified items triggered at least one regeneration, mostly to replace eliminable distractors (822 items, 2,115 attempts) or a vague correct answer (548 items). Eliminable distractors are the most common first failure in 20 of the 21 SOC groups in the replay. Replay rates for the stages after the verifier are not reported, because a token limit caused one of the three Pass 7 probe models to return no answer during the replay; those stages are reported from the production logs instead.

Table 8: Pipeline funnel for the runs that produced the ORQA bank, by SOC major group. Counts come from the per-occupation run logs of the 317 occupation runs whose output includes at least one bank item (470 of the 480 items are traced to a run). *Runs* counts occupation runs, *Fetched* source URLs fetched, *Verifier* items rejected by the independent verifier, *Pass 7* items rejected by the difficulty probe as too easy / too hard, *Mix* items rejected by domain balancing, *Kept* items the pipeline retained, and *Bank* items in the final bank after human-calibrated categorization and occupation-floor balancing. *Bank* can exceed *Kept* within a group: 10 bank items could not be traced to a run, and for Building & Grounds the logs record one fewer kept item than the bank contains.

SOC	Major group	Runs	Fetched	Verifier	Pass 7	Mix	Kept	Bank
11	Management	12	1,179	448	10 / 22	76	38	21
13	Business & Financial	20	1,955	460	23 / 20	100	66	44
15	Computer & Mathematical	7	992	585	17 / 26	64	34	17
17	Architecture & Engineering	16	2,523	1,193	16 / 107	123	64	42
19	Life, Physical & Social Science	23	2,332	1,747	74 / 120	301	120	52
21	Community & Social Service	12	1,463	913	25 / 41	72	69	32
23	Legal	7	1,214	722	35 / 32	143	44	15
25	Education & Library	10	713	87	2 / 5	1	13	7
27	Arts, Design & Media	13	1,130	188	13 / 19	19	29	16
29	Healthcare Practitioners	40	6,589	2,336	142 / 146	372	497	52
31	Healthcare Support (no bank items)	8	1,530	272	25 / 13	66	78	0
33	Protective Service	18	1,152	332	18 / 32	33	54	17
35	Food Preparation & Serving	5	328	40	7 / 3	15	10	1
37	Building & Grounds	2	179	73	0 / 1	6	8	9
39	Personal Care & Service	12	1,075	318	14 / 23	56	25	16
41	Sales & Related	20	1,662	590	50 / 30	146	47	18
43	Office & Administrative Support	18	1,582	658	21 / 44	146	59	14
45	Farming, Fishing & Forestry	1	57	16	0 / 2	0	3	3
47	Construction & Extraction	25	2,240	567	10 / 30	32	60	31
49	Installation & Repair	14	727	181	0 / 11	15	22	9
51	Production	20	1,768	453	15 / 32	35	45	26
53	Transportation & Moving	14	1,210	524	27 / 30	134	37	38
	Total	317	33,600	12,703	544 / 789	1,955	1,422	480

Table 9: Per-check funnel from a logged replay of 1,885 evidence cards (a stratified sample of 39 bank occupations, about two per SOC major group) through the production card path (build, then the verifier's checks and their regenerations). *Reached* counts items that got to a check, *Pass* is the share of those passing it (checks run in the listed order; an item stops at its first failure), and *Regen.* counts items for which the check triggered at least one regeneration. Stages after the verifier are taken from the production logs, because the replay's Pass 7 probe was affected by an API quota fault.

Stage or check	Reached	Pass (%)	Regen.
Evidence card built into an item	1,885	73	–
Verifier: Source authority	1,366	100	–
Verifier: Answer entailed by quote	1,366	81	–
Verifier: Occupation aligned	1,101	86	0
Verifier: Practical competence	944	99	–
Verifier: No distractor entailed	933	100	5
Verifier: Not ambiguous	931	100	36
Verifier: Jurisdiction fixable	930	100	–
Verifier: Answer does not name source	930	96	101
Verifier: Answer specific (Pass 9)	890	96	548
Verifier: Distractors not eliminable (Pass 8)	855	39	822
Verifier: all checks passed	1,366	24	1,136
Later stages, production logs of the 317 bank runs			
Pass 7 difficulty probe	4,710	72	–
Source-mix balance	3,377	42	–
Good-rubric gate (runs where enabled)	1,649	22	34

Table 10: The same replay by SOC major group: cards, share built into items, share passing all verifier checks, the most common first verifier failure, and the share of verified items that needed at least one regeneration.

SOC	Major group	Cards	Built (%)	Verifier pass (%)	Top verifier failure	Regen. (%)
11	Management	100	63	14	Distractors not eliminable (Pass 8)	94
13	Business & Financial	100	81	33	Distractors not eliminable (Pass 8)	90
15	Computer & Mathematical	100	73	16	Distractors not eliminable (Pass 8)	84
17	Architecture & Engineering	100	75	33	Distractors not eliminable (Pass 8)	81
19	Life, Physical & Social Sci.	100	56	32	Distractors not eliminable (Pass 8)	91
21	Community & Social Service	100	57	21	Distractors not eliminable (Pass 8)	72
23	Legal	100	73	23	Distractors not eliminable (Pass 8)	92
25	Education & Library	100	64	27	Distractors not eliminable (Pass 8)	89
27	Arts, Design & Media	100	70	16	Distractors not eliminable (Pass 8)	74
29	Healthcare Practitioners	100	67	10	Distractors not eliminable (Pass 8)	84
33	Protective Service	100	66	21	Distractors not eliminable (Pass 8)	91
35	Food Prep. & Serving	50	68	21	Distractors not eliminable (Pass 8)	91
37	Building & Grounds	50	76	24	Distractors not eliminable (Pass 8)	79
39	Personal Care & Service	100	90	30	Distractors not eliminable (Pass 8)	77
41	Sales & Related	100	87	34	Occupation aligned	71
43	Office & Admin. Support	100	72	21	Distractors not eliminable (Pass 8)	81
45	Farming, Fishing & Forestry	29	72	29	Distractors not eliminable (Pass 8)	86
47	Construction & Extraction	100	80	16	Distractors not eliminable (Pass 8)	81
49	Installation & Repair	106	85	30	Distractors not eliminable (Pass 8)	82
51	Production	100	78	24	Distractors not eliminable (Pass 8)	83
53	Transportation & Moving	50	72	31	Distractors not eliminable (Pass 8)	83

A.12 What Kind of Knowledge Items Test

To see which kinds of professional knowledge models miss, every item was labeled with the kind of knowledge it tests and every wrong option with the way it is wrong. Two annotator models from different providers (GPT-4.1 and Gemini 2.5 Flash) labeled each item independently from its question, options, correct answer and source quote; o4-mini labeled the 138 items on which they disagreed, and the majority label is used. Agreement was moderate to substantial (knowledge type: 71% raw, Cohen's $\kappa = 0.60$; distractor type: 76% raw, $\kappa = 0.65$). Procedures or required actions (173 items) and general regulations or standards (159) make up most of the bank, followed by numeric thresholds or values (86), concepts or terminology (50) and jurisdiction- or organization-specific rules (12).

Numeric items are the hardest (Table 11). Frontier models answer 46.2% of them correctly against 60-68% for procedures, regulations and concepts, a gap of 16.0 points relative to all other items (95% CI [2.5, 29.6]). This matches the qualitative pattern in Appendix A.9: models know which rule applies but not the exact limit, interval or code it sets. The jurisdiction-specific group is too small for a firm estimate. Wrong answers are not concentrated on one kind of distractor: across the ranked models, the share of wrong answers landing on each distractor type is within 10% of that type's share of the wrong options offered (ratios 0.92-1.06 for the five main types; "All of the above" and "None of the above" take 42% of wrong answers and make up 40% of wrong options). Knowledge-type composition also does not explain the spread across SOC major groups: adjusting each group's frontier accuracy for its mix of knowledge types leaves the standard deviation across groups essentially unchanged (13.9 to 13.8 points).

Table 11: Accuracy by the kind of knowledge an item tests. Labels from two annotator models (GPT-4.1, Gemini 2.5 Flash; Cohen's $\kappa = 0.60$) with o4-mini resolving the 138 disagreements. Frontier = mean of Claude Opus 4.6, GPT-5.4, Claude Sonnet 4.6 and Gemini 3.1 Pro; 95% intervals from an occupation bootstrap.

Knowledge type	Items	Frontier (%)	95% CI
Concept or terminology	50	67.5	[49.8, 82.9]
General regulation or standard	159	63.5	[53.9, 72.3]
Procedure or required action	173	60.5	[51.8, 69.6]
Jurisdiction- or organization-specific rule	12	50.7	[15.3, 88.6]
Numeric threshold or value	86	46.2	[34.0, 58.7]
Numeric minus all other types		−16.0	[−29.6, −2.5]

A.13 Accuracy and Documentation Footprint

One explanation for the spread across occupations is that models do better where more authoritative text about the occupation exists, independent of the occupation's economic size. We measure an occupation's documentation footprint in two ways that do not use model outputs. The first comes from the pipeline's own search: the number of authoritative documents found per run for the occupation and the number of distinct documents that yielded at least one evidence card. The second is the size and 2024 pageviews of the occupation's English Wikipedia article. Web-crawl page counts would be a closer proxy for pre-training exposure, but the Common Crawl index was not reachable at the scale needed. Economic size is measured by OEWS 2024 employment, mean wage and wage bill.

Table 12 reports Spearman correlations with frontier accuracy for the 101 occupations with at least three items. All correlations are positive and small. Documentation density has the largest association ($\rho = +0.18$, 95% CI [−0.01, +0.37]), and it is unchanged after partialling out wage bill (partial $r = +0.18$, $p = 0.07$); the Wikipedia measures and the economic measures are weaker (ρ between +0.09 and +0.13). In an OLS regression of frontier accuracy on standardized measures with HC1 standard errors, one standard deviation more evidence documents is associated with +6.3 points (s.e. 2.9) holding Wikipedia article size and wage bill fixed, and one standard deviation more wage bill with +5.3 points (s.e. 2.8); together the measures explain 10% of the variance across occupations. The evidence is weakly consistent with a footprint effect, but with three to eleven items per occupation most of the variation remains unexplained, and both footprint measures are imperfect proxies for what models saw in pre-training.

Table 12: Occupation-level correlates of frontier accuracy (101 occupations with at least three items). Spearman ρ with 95% bootstrap intervals; all measures on a log scale.

Measure	ρ	95% CI
Authoritative documents found by search (per run)	+0.18	[−0.02, +0.37]
Documents yielding evidence cards	+0.18	[−0.01, +0.37]
Wikipedia article size	+0.10	[−0.10, +0.29]
Wikipedia pageviews, 2024	+0.13	[−0.08, +0.34]
Wage bill (OEWS)	+0.13	[−0.06, +0.32]
Employment (OEWS)	+0.09	[−0.11, +0.29]
Mean annual wage (OEWS)	+0.12	[−0.10, +0.32]

A.14 Expert Audit of Items

Every check reported elsewhere in this paper is internal to the pipeline: the verifier, the difficulty pretest and the authors' own review. This appendix reports an audit by practising professionals, conducted under an exempt human subjects protocol at the authors' institution.

Design. Items were drawn from the 21 occupations in the bank that can be recruited on a large online participant platform, giving 88 items. Recruitment ran as four listings grouped by industry (healthcare; licensed professions; finance and customer service; trades, transport and dispatch). A screener admitted only participants who reported at least two years of experience in the occupation, said they currently work in it, identified the correct licensing or certifying body for the role, and answered one basic question from their field; a cap of three raters per occupation was enforced automatically, and participants beyond the cap were paid a screen-out fee. Each rater saw the 3 to 7 items written for their occupation twice. In the first pass the item was shown alone, and the rater chose an answer from the same six options given to models (or "this is not my area") without looking anything up, rated how relevant the knowledge is to their own work on a 1-5 scale, said whether they had heard of the source organisation, and said whether a competent practitioner would know this offhand or look it up. In the second pass the same item was shown with the answer on file and the passage from the source document it was drawn from, and the rater said whether they agreed with that answer, whether the passage supports it, and how relevant the source is to their occupation. Two attention checks were embedded, and a comprehension check followed the instructions. Payment was between $6.00 and $6.50 for a median of about 12 minutes, roughly $30 per hour. Submissions were reviewed before payment against attention checks, page timings, and internal consistency; three raters who reported looking something up are retained for the rating questions but excluded from the independent-answer accuracy, and one rater who gave the top rating to every item after contradicting four of them is excluded from the reveal ratings. Three further raters spent under 15 seconds per reveal page; excluding them as well moves agreement from 84.4% to 84.2% and source support from 82.0% to 82.2%, so the results do not rest on the fastest raters.

Results. Sixty-four raters across 20 occupations produced 255 ratings on 81 of the 88 items; commercial pilots returned no qualifying raters, and the remaining items belong to occupations that filled only partly. Figure 18 summarises the ratings and compares practitioners with frontier models item by item, and Table 13 breaks both down by occupation.

Agreement with the bank is high. Raters agreed with the answer on file in 84% of ratings [77.7, 90.5] and agreed or partly agreed in 96% [92.0, 98.4]; they judged the cited passage to clearly support that answer in 82% [76.3, 87.7] and clearly or mostly in 96% [93.4, 98.9]. All raters agreed on 49 of the 81 items, five items attracted exactly one disagreement, and no item attracted two. Seven items had at least one rater say the passage does not support the answer. These are the items worth revisiting, and they are a small minority.

The independent answers give a second reading. Practitioners answered 71.5% [63.2, 79.1] of items correctly, against 57.6% for the frontier average on the same items, and beat the frontier average in 11 of 20 occupations. Item-level accuracy for practitioners and models is positively but loosely related (Spearman ρ = +0.30, p = 0.008, 78 items), and the occupation-level association is also positive (ρ = +0.49, p = 0.03, 20 occupations), so the items that are hard for models are not simply the items

that are hard for people. Raters also agree with each other more often than chance: across the 267 pairs of raters who answered the same item, both chose the same option 66% of the time against 22% expected from the distribution of chosen options, and every rater chose the same option on 53% of items. Both raters in a pair chose the answer on file in 59% of pairs and exactly one did in 31%. Practitioners selected "this is not my area" on 6.7% of items, said they had heard of the source organisation for 82%, and said a competent practitioner would know the answer offhand for 61%. Median time on an item was 64 seconds.

Relevance is the weakest result, and the most useful. The mean relevance rating is 3.9 of 5, with 75% of ratings at 4 or 5 but 19% at 1 or 2, and mean relevance varies from 3.1 (registered nurses) to 4.7 (dentists, loan officers, 911 dispatchers). Agreement is lower on items rated as less relevant: raters agreed with the answer on file in 87% of ratings for items they rated 4 or 5, against 73% for items rated 1 or 2. Some occupations therefore have authoritative sources that document work at some distance from daily practice, which is a limitation of source-grounded construction rather than of any individual item.

Two caveats limit what the audit establishes. First, three raters per occupation is thin, so occupation-level figures in Table 13 are indicative and the confidence intervals in Figure 18 are wide. Second, the audit covers 81 items in 20 occupations, about 17% of the bank, and those occupations are the ones a participant platform can reach; the physician specialties and small trades that make up much of the rest of the bank are not represented.

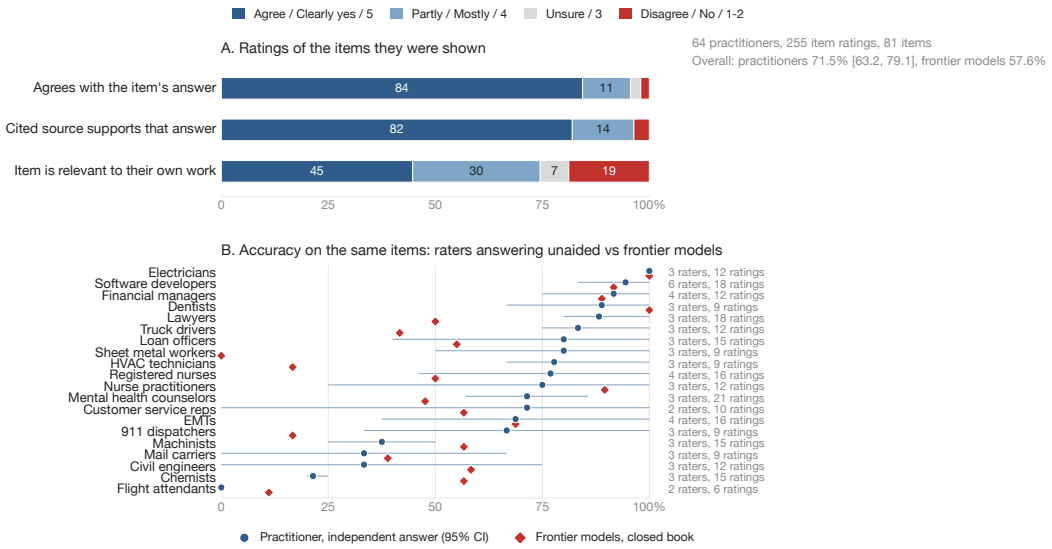


Figure 18: Expert audit of ORQA items. **A**: how practitioners rated the items shown to them. **B**: practitioners' own answers, given before the answer was revealed and without looking anything up, against frontier model accuracy on the same items, by occupation. Intervals are 95% bootstrap intervals resampling raters.

A.15 Closed Book, Fixed Retrieval and Web Search

Provider web search mixes two things: whether the model knows the material and whether its search tool finds the right document. To separate them, every model that is still served was also run with one fixed retriever. The index holds 234,397 passages of about 1,000 characters from 3,934 documents on the same whitelisted authoritative hosts the pipeline draws from, including 96% of the bank's source documents. For each item a BM25 retriever queries the index with the question and the four substantive options and prepends the top five passages to the unchanged multiple-choice prompt; seeds and inference settings match the closed-book runs. Two versions are run: *RAG*, in which the item's own source document can be retrieved, and *RAG, source removed*, in which every passage from that document is excluded, so the model must rely on related authoritative text.

Table 13: Expert audit by occupation. *Agrees* is the share of ratings where the practitioner agreed with the item's answer after seeing it with its source passage; *source supports* is the share rating the cited passage as clearly supporting that answer; *practitioner* is accuracy on the practitioner's own answer given before the answer was revealed; *frontier* is the mean of Claude Opus 4.6, GPT-5.4, Claude Sonnet 4.6 and Gemini 3.1 Pro on the same items; *relevance* is the mean 1-5 rating of how relevant the item is to the rater's own work.

Occupation	Items	Raters	Agrees (%)	Source supports (%)	Practitioner (%)	Frontier (%)	Relevance (1-5)
Electricians	4	3	100	83	100	100	4.1
Software developers	3	6	94	94	94	92	4.6
Financial managers	3	4	100	83	92	89	3.8
Dentists	3	3	100	100	89	100	4.7
Lawyers	6	3	72	83	88	50	3.3
Truck drivers	4	3	92	83	83	42	4.4
Loan officers	5	3	80	47	80	55	4.7
Sheet metal workers	3	3	100	100	80	0	3.2
HVAC technicians	3	3	89	78	78	17	4.3
Registered nurses	4	4	81	75	77	50	3.1
Nurse practitioners	4	3	100	100	75	90	3.5
Customer service reps	5	2	100	90	71	57	3.4
Mental health counselors	7	3	100	81	71	48	4.5
EMTs	4	4	81	94	69	69	4.0
911 dispatchers	3	3	67	78	67	17	4.7
Machinists	5	3	80	73	38	57	3.2
Civil engineers	4	3	58	75	33	58	3.8
Mail carriers	3	3	78	100	33	39	3.7
Chemists	5	3	30	60	21	57	3.3
Flight attendants	3	2	67	67	0	11	3.2
All	81	64	84	82	71	58	3.9

Table 14 shows three findings. First, ORQA items are answerable from their sources: with the source retrievable every model scores 93-99%, including GPT-3.5 Turbo (95.1%), so closed-book errors are not caused by flawed or unanswerable items. Second, much of the answer is present in related authoritative documents: with the source removed, Claude Opus 4.6, GPT-5.4, Claude Sonnet 4.6 and o3 still score 90-94%, and Claude Opus 4.6, GPT-5.4 and Claude Sonnet 4.6 score 87-91% even on the 80% of items for which no retrieved passage contains the source quote. Third, provider web search (78-80% for the frontier) recovers only part of this gap. The two retrieval conditions are not matched in difficulty: the fixed retriever searches a curated index of 3,934 documents drawn from the same hosts the items came from, while provider search works over the open web, so the gap between them reflects the size and noise of the search space as well as search quality. The fixed-retrieval column is therefore an upper bound on what retrieval can contribute rather than a realistic deployment setting. Retrieval also changes the ordering at the bottom of the slate: with the source removed, Claude Haiku 4.5 (61.7%) and Gemini 2.5 Flash (62.4%) fall below GPT-3.5 Turbo because they choose "None of the above" on 34% and 24% of trials when the passages do not state the answer, against 10% or less (rounded) for every other model. This is a difference in fallback behavior rather than a knowledge deficit: when the retrieved text does not contain the answer, abstaining is the faithful response to the context, while the higher-scoring models fall back on parametric knowledge. Scored against the item key it counts as an error, which is worth keeping in mind when reading that column. Closed-book rankings in the main text are unaffected, since these runs are a separate condition on the same items.

Table 14: Accuracy (%) under four evidence conditions on the same 480 items. *RAG, source removed*: the same BM25 index and top-5 passages with the item's own source document excluded. *RAG*: the item's source document is retrievable. *Web search*: each provider's own search tool (Section 4). – marks a model that could not be run in that condition (deprecated or no longer served, or no native search tool). Models below the rule were the Pass 7 difficulty filter.

Model	Closed book	RAG, source removed	RAG	Web search
Claude Opus 4.6	62.4	92.8	98.6	79.4
GPT-5.4	60.3	92.2	98.8	80.1
Claude Sonnet 4.6	58.7	89.7	99.0	78.5
Gemini 3.1 Pro	56.2	88.2	95.7	77.9
GPT-5.3-chat-latest	54.2	–	–	76.8
GPT-5.2	52.8	87.6	98.6	75.8
o3	51.7	94.2	98.2	–
GPT-4o	43.5	86.9	98.3	69.2
DeepSeek V4 Pro	40.8	–	–	–
Llama 3.3 70B	38.8	84.0	97.2	–
Qwen 2.5 7B	33.9	–	–	–
GPT-3.5 Turbo	33.3	82.0	95.1	–
Claude Haiku 4.5	48.2	61.7	94.6	72.6
Gemini 2.5 Flash	47.0	62.4	93.3	71.7
GPT-4o mini	38.3	82.6	95.6	63.7

Note. Index: 234,397 passages from 3,934 documents on whitelisted source hosts (96% of the bank's source documents). Top-5 retrieval returns the item's own source document for 86% of items and a passage containing its source quote for 76%; with the source removed, a near-verbatim copy of the quote from another document is still retrieved for 20%.