

# Human Baselines for Benchmarks: AI Now Outperforms Junior Accountants

A human baselining study of 12 licensed CPAs on tasks adapted from APEX-Accounting

---

**AUTHOR**

Aden Barton

**PUBLISHED**

October 2026

**CONTACT**

[adenbarton@mercor.com](mailto:adenbarton@mercor.com)

**WEB**

[mercor.com](https://mercor.com)

Benchmark scores have soared over the last few years, but AI adoption seems to be contributing only slightly, if at all, to higher productivity growth in U.S. economic statistics. It's a trillion-dollar paradox.

That tension is hard to resolve without understanding how benchmarks relate to real work, and how humans themselves would do that work. Human baselining – measuring how well humans perform on benchmark tasks – is a key tool for gaining this understanding. But as benchmark tasks have grown longer and more complex, human baselining has become less common because of the added expense and logistical burden.

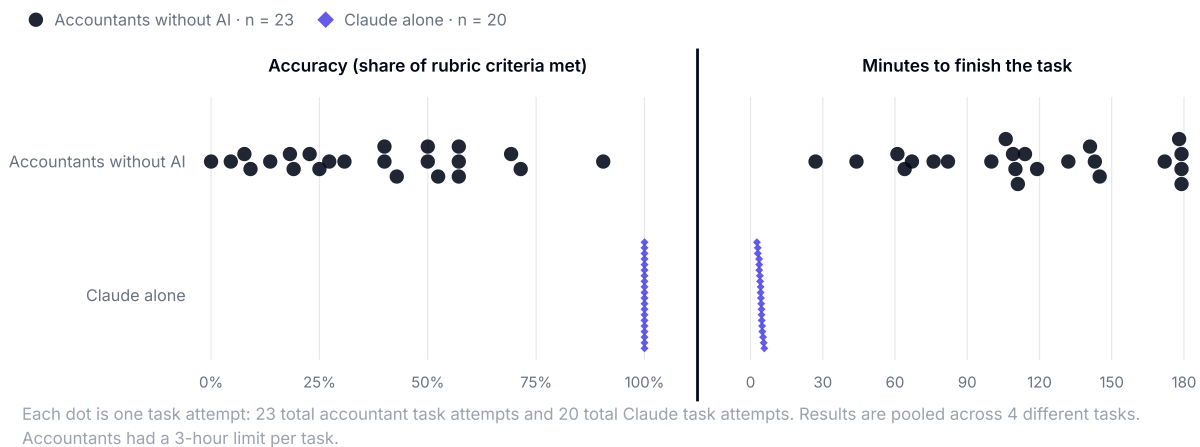
To help bridge that gap, we hired 12 junior accountants to complete tasks adapted from our recent [APEX-Accounting benchmark](#), completing some tasks with assistance from Claude and some without. We also compare accountants' performance against unaided AI models.

We find that on medium-length, well-defined accounting tasks, frontier AI models are now faster and more accurate than junior accountants. On both dimensions, AI outperformed not just the average accountant but the best one in our sample.

FIGURE 1

### Claude is faster and more accurate than accountants on medium-length, detail-oriented accounting tasks

Accuracy and time for accountants and Claude Opus 5 on timed, realistic accounting tasks

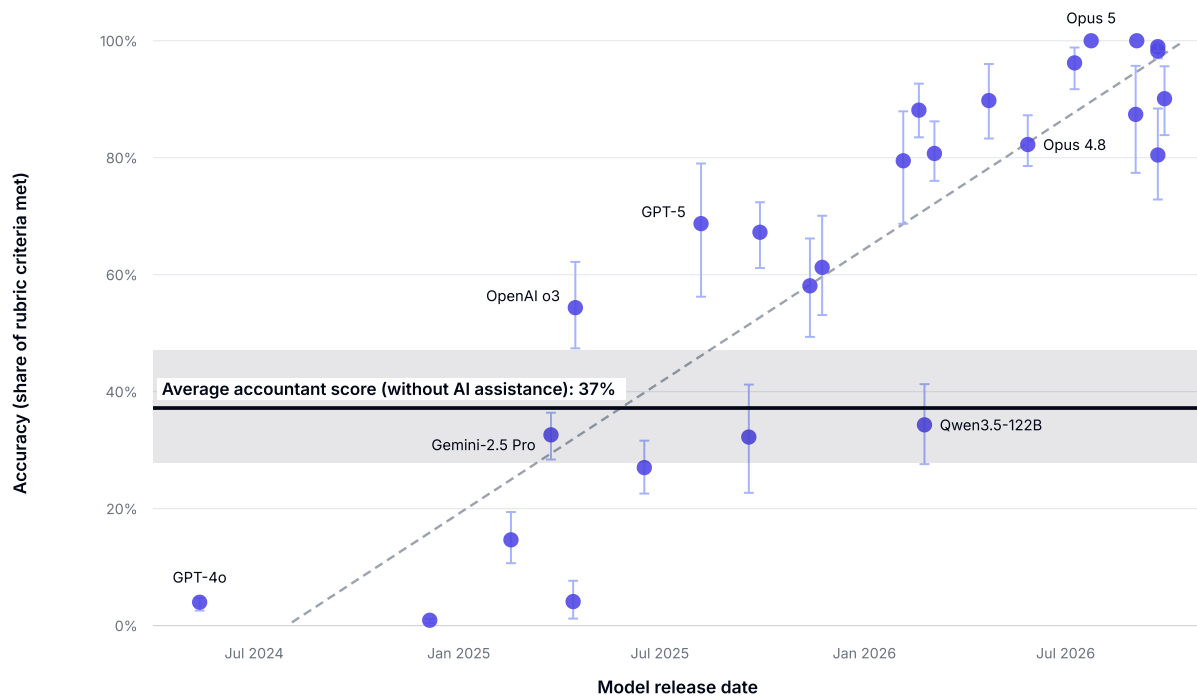


What matters even more than that gap is how quickly AI took the lead, since the pace may be a signal of where capabilities are headed. As shown in the graph below, just a year and a half ago, the best AI models still fell short of the average accountant's score of around 37%. Today, models ace those same tasks.

FIGURE 2

### AI models passed accountants on these tasks in 2025 and now score at or near 100%

Accuracy of AI models on the study's accounting tasks by model release date, compared with accountants without AI



Every model tested is shown, including budget and open-weight variants, which is why several recent releases sit below the accountant line. Scores are averaged across the 4 study tasks. Error bars and the shaded band are 95% bootstrap intervals across runs. Dashed line is a linear fit across all models, drawn between 0% and 100%.

But these results do not imply accountants are replaceable – far from it, for two reasons.

First, our tasks tested only a small set of skills that happen to be what models are best at: being detail-oriented, searching through files, and following instructions very closely.

Because the tasks were originally designed to stump models, they evolved to include hard-to-spot (but realistic) accounting requirements that compounded on one another, so a single overlooked number could produce a very low score. These are important skills in accounting (and knowledge work more generally), but they are not the whole job. Given this difficulty, our task authors expected average scores to be quite low: 30% for a junior and 55% for a mid-level accountant.

Second, much of an accountant's job was absent from our setting. There were no coworkers to ask questions, and participants came to the tasks cold, without the context one would normally have at a job. That made the tasks much harder. It also meant we did not measure skills like client

communication, asking the right questions, and collecting tacit context over time. These skills will likely become a larger part of accountants' work as models take on more of the detail-oriented tasks.

---

## 01 Study Design

### Task Selection and Iteration

To start, we selected accounting tasks from our recent benchmark, [APEX-Accounting](#), which were designed by real accounting or finance professionals. To make them more tractable for the study (i.e. to reduce estimated human completion time), we simplified and pared down the tasks. Additionally, to reduce pure search time, we gave the accountants a file map describing at a high level what each file contained.

Then, we piloted the simplified tasks with two initial waves of participants from Mercor's network of experts. Based on their feedback, we revised the tasks further to remove any outstanding ambiguities and potentially unfair items before the full study.

The four tasks were all month-end close scenarios that accountants would realistically have to complete in their line of work. Each participant was given a simulated company's working files and had to dig through those files to find the right numbers, figure out which calculation approach the company's rules called for, do the math, and then deliver a completed summary table of the results.

The submissions were graded on rubric items also written by accountants, ranging from 13 to 22 total items depending on the task. One item on the hotel task, for example, required the response to identify that \$4,000 of parking revenue had been miscoded as room revenue. The grades were produced by an LLM judge. To ensure they were calibrated, we had 8 task submissions hand-graded – 2 for each of the 4 tasks – by the task authors; their grades showed a 98% agreement with our LLM judge.

For the AI-only runs, we tested the models as tool-using ReAct agents over an API gateway and capped them at 250 steps. We matched everything (grader prompts, rubrics, file system, etc.) in the task set-up and grading between the human alone and AI alone tasks.

### Study Flow

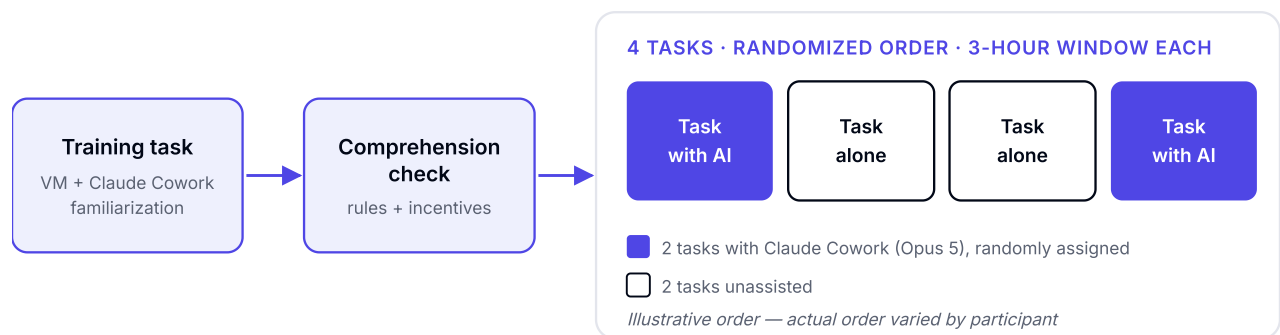
For the full study, we administered these tasks on a virtual machine to 12 junior accountants, with about five and a half years of average accounting experience. See Appendix A2 for more specific details on the participants' backgrounds. The virtual machine was meant to replicate the feel of logging onto one's desktop at work, with all files living locally on the desktop.

Before they took the study tasks, accountants were required to complete a training task to familiarize themselves with the virtual machine and Claude Cowork. We also administered a comprehension check to ensure participants understood the rules and incentives before taking the study.

Each participant had a three-hour window per task and was paid for the full three hours no matter how much time they took, so there was no incentive for the participants to artificially prolong their time spent on each task. We also offered a bonus for each task criterion they satisfied.

We randomized the order of the four tasks, as well as which two tasks they completed with AI assistance (Claude Cowork with Opus 5). We used Claude as the collaboration model because Anthropic was at the top of the APEX-Accounting leaderboard when the study began. The graphic below depicts the study flow.

FIGURE 3



## 02 Results

### Augmentation

The study was initially designed to measure the uplift from AI, i.e. the extra help a person gets from using AI to accomplish work. We expected that professional accountants would add domain knowledge that would make their work with AI better than either the AI or the expert alone. Instead, our findings show that humans with AI are slower and slightly less accurate than AI alone.

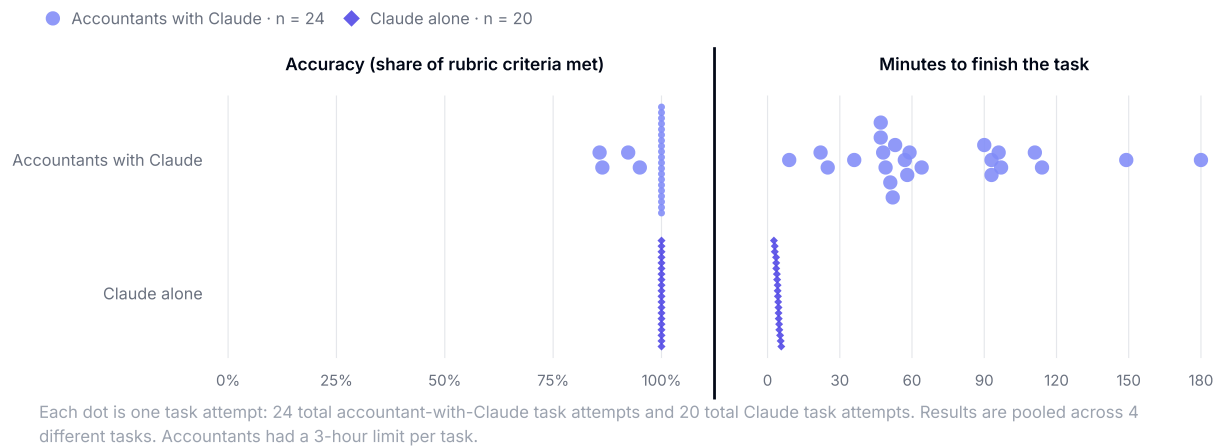
Of course, that conclusion follows from AI alone getting 100% on the tasks. If Claude can already get a perfect score, the rubric will not measure any increase in quality from human assistance. As a result, any human time spent reviewing the output is time the model would not have spent alone.

Further, accountants working with Claude took about 15X longer per task than Claude working alone. And, interestingly, in three of the four AI-assisted sessions that did not receive a perfect score, Claude had the correct figure at some point, but it didn't make it into the final submission. In two, Claude revised away from it and the participant deferred to the new answer; in the third, the participant overrode Claude's initial correct advice.

FIGURE 4

### Claude alone is faster and slightly more accurate than accountants working with Claude

Accuracy and time for accountants using Claude and for Claude Opus 5 alone on timed, realistic accounting tasks



## Automation

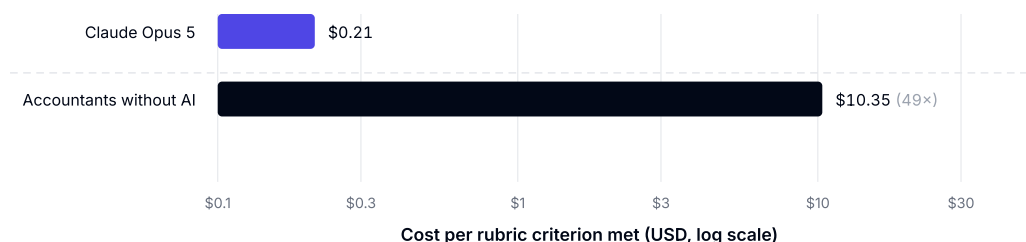
Because AI alone scores perfectly, we are mostly measuring automation. As a result, we focus on the results of unassisted humans against AI alone. As shown in the very first chart, participants performed substantially worse than models released over the last year and a half. Models are also less expensive: comparing cost per task criterion, frontier models are more than an order of magnitude cheaper than humans.

Importantly, most study participants submitted before their 3-hour limit, suggesting time was not the main constraint.

FIGURE 5

### Claude costs far less than an accountant per correct rubric criterion

Total cost of the model, or of accountant time at the US median accountant wage, divided by the average number of rubric criteria met



Claude Opus 5 is the model used throughout this study. Its cost is estimated by multiplying input and output tokens by their respective token prices, at list rates with no caching discount, so it is an upper bound. Accountant wage is the BLS median, not the rate paid in the study.

## 03 Why Did Models Outperform Humans?

When we saw the results, we went back to the accountants who helped make these tasks and triple-checked the tasks' fairness. We concluded that these tasks were fair, meaning the right answer was findable from the starting files (shown – at least partially – by the models' success).

We built the tasks – ironically – to fail models, so they are full of hard-to-spot accounting conditions and difficult edge cases. For example, in one task a block of hotel room stays at the end of December was not billed until January 3, after the month ended, so it was missing from December revenue even though it belonged there. In another, the company background said the company does a lease reclassification step every month, even though many companies only do it when preparing financial statements. Nearly every participant working without AI skipped it.

But while these accounting requirements were difficult, none of them were unreasonable as tests of accounting knowledge. On this question, one of the task authors said the following:

"The errors [to be corrected in the task] themselves are not rare. Something coded to the wrong account, a bill that lands after the month ends, a commission that never made it into the books, that is a normal close, and an accountant sees them constantly."

However, they continued, "what is unrealistic is the setting and the scoring," meaning model outperformance came not from the task difficulty per se but from its interaction with the setting and the scoring. As the task author concluded, "the traps are realistic, and the conditions and the

scoring are what drove the results."

On the setting, there are many ways that the task-taking diverged from reality, but to name three:

- **The lack of context.** In a real job, people would have months or years of context building up to a task like this.
- **No ability to ask questions** or clarify ambiguities.
- **Lack of 1:1 expertise fit with the task.** Although the tasks could all be solved with foundational accounting knowledge, they tested specific concepts that may not come up regularly depending on an accountant's specialization, e.g. how a nonprofit splits shared costs.

On the scoring, mistakes cascade. Missing one of these hard-to-spot errors could result in a very low score. Consistent with this, the task authors estimated that even mid-level accountants would likely only get around 50% to 60%, depending on the task.

Overall, then, the experts had lots of difficult things to catch, little context, and no ability to ask questions. These all compounded the tasks' difficulty.

Participants' descriptions of task difficulty confirm this view (Appendix A3). Many cited the volume of material, e.g., "there was enough info to do almost any task, but that means I had to sort through all of it." Other accountants noted the inability to ask questions, e.g., "In reality I can usually ask my client questions as I work."

The result was that we ended up testing precisely what the models are best at: being detail-oriented, searching over the whole file set, and following instructions very closely, i.e. not defaulting to heuristics that work 80% of the time. Humans, by contrast, often rely on context and colleagues not only to catch what they miss but also to brainstorm with.

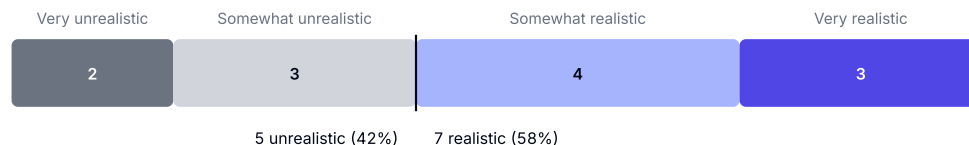
One task author observed that "with a lot of files to work through, people tended to settle early on an approach that looked reasonable and carry it through without checking it back against the instructions." In the event revenue task, for instance, two participants split revenue by accounting category rather than by event type, which the prompt asked for.

The realism gap around the setting shows up in the 5 of 12 participants who said the tasks were not realistic. These participants mostly had objections not to the tasks themselves, but to the lack of interaction or colleagues. For example, one expert said "real work has colleagues [sic] and supervisors you can brainstorm with and ultimately managers there that can help figure things out with you when you get stuck."

FIGURE 6

### 7 of 12 accountants rated the tasks as realistic

Participant responses to "How realistic were these tasks compared with work you actually do in your job?"



All 12 participants answered. The two "very unrealistic" respondents later clarified that their objections were to the setting (e.g. no coworkers), not the tasks themselves.

However, on the tasks themselves, experts tended to be positive, although not universally. One said "nothing felt unrealistic with the tasks." In fact, when we followed up with the two participants who said the tasks were "very unrealistic," they clarified that their objections were with the setting (e.g. lack of coworkers), not with the Excel tasks themselves. See Appendix A3 for the full free-form responses on task realism.

## 04 Interpreting the Results

These results don't imply that accountants are fully replaceable, but they are informative about the future of work.

The most important takeaway is that models are getting much better very quickly. If we had run the study a year and a half ago, the average participant would have beaten the models. Now even the best participant loses badly. Extrapolating this progress – even acknowledging the contrivance of these tasks – suggests model capabilities surpassing humans in many knowledge work tasks over the next several years.

And even if model progress were frozen today, these results suggest that current AI capability would still provide large productivity gains to accountants. Setting aside the harder criteria and just looking at the items that humans tend to get right (defined as greater than 60% pass rate on average), models alone tend to provide a 34x speedup relative to unassisted humans.

Beyond being faster, models today already seem more meticulous than junior accountants. That's extremely valuable in a field where small errors can be extraordinarily costly.

As a result, accountants' work will likely shift toward the parts of the job where speed and meticulousness matter less. This probably means things like handling client relationships, asking the right questions of colleagues, or operating in the presence of uncertainty or lack of clear rules.

---

## 05 A Reflection on Benchmarking and Uplift Studies

Finally, these results raise some open questions about benchmarking and about studies measuring how AI augments human productivity, often called uplift studies.

There are tasks that only humans can do, like attending an in-person meeting, and there are tasks that only AI can do, like reading 100 Excel files in under a minute. Traditionally, benchmarks focus on the overlap of human and AI work: tasks that both humans and AI can do, like creating a slide deck over several hours. As a result, benchmarks typically represent real-world tasks that people perform in their jobs, and uplift studies accordingly measure how much AI speeds up humans on those tasks.

But much of the benchmarking field has moved towards technically difficult tasks designed to fail AI, many of which humans could not do on their own and therefore are not representative of everyday work.

Importantly, those benchmarks likely do not set out to create non-human tasks but rather start with everyday tasks and make them harder to induce failures. Think of exam questions built to challenge top students.

These study results are a prime example of that dynamic. On the parts of the tasks we would expect humans to complete well, models are as good as the best accountants, just much faster. But because the tasks were iterated on to stump models, the density of errors pushed beyond what we would expect even a strong accountant to catch at work, and that is where models exceeded even the best study participants.

Current benchmarks, then, increasingly look something like the right-hand Venn diagram, with some tasks able to be performed by humans and some not.

FIGURE 7



These new AI-only tasks are likely some of the most economically valuable tasks. You wouldn't ask a human to run 60 miles an hour, but the ability to transport people at that speed is precisely what allows cars to increase economic productivity.

That shifts the goal of frontier benchmarking from measuring how well AI does existing human work to measuring tasks that were formerly impossible or prohibitively costly. Identifying those new tasks, of course, is its own challenge.

And if benchmarks are sampling from tasks only AI can perform, it becomes much harder to estimate the economic value of rising benchmark scores. With human tasks, we at least have existing statistics, like wages, that can serve as a proxy. But tasks only AI can do aren't performed in the economy yet (or only recently started to be performed).

At the very least, then, benchmark releases should try to measure how their tasks split between those only AI can do and those humans also perform (in part by human baselining) to help gauge the economic significance of the results.

*Interested in working on similar problems or have feedback for us? Please email [adenbarton@mercor.com](mailto:adenbarton@mercor.com).*

*We thank Whitney Zhang for her help in designing the study and for giving feedback on the write-up drafts.*

*Additionally, we thank the accounting experts who took part in the study as well as those who helped run and create the tasks and study infrastructure.*

## A Appendix

### A1. Task Descriptions

**Task 1.** A nonprofit runs four programs, and each program's December 2024 spending was compared to its budget, program by program, in dollars and as a percent. Some costs, like payroll taxes, sat in a shared pool rather than under any one program. Those had to be spread across all four programs and the two support functions first. Any line where a program ran \$14,000 or more over budget got flagged with a note on what drove it.

**Task 2.** A hotel's booking system and its accounting books disagreed about December room revenue, which also threw off the commissions owed to its online travel agencies for the month. Both gaps had to be tracked down item by item, with the amount and the account or platform for each, then fixed with correcting journal entries. Then the net effect of both fixes on December income.

**Task 3.** A venue's events in December brought in more revenue than budgeted, but the average per event came in under plan. Actual and budget, with the gap between them, were needed on the number of events, the average revenue per event, food and beverage minimums, and the mix of event types by count. Last came an explanation of the difference, backed by revenue and event count for each type of event.

**Task 4.** Three companies in one group each rent a property. The holding company has an office, the trucking division a terminal yard, and the logistics division a warehouse. Under the U.S. lease accounting rule, each lease sits on the books two ways. There's a lease liability for what's still owed and a right-of-use asset for the value of using the space, and both come down over time. For December, each lease schedule was rolled forward a month, working out how much of the asset got used up and the drop in the liability. The numbers went into the December entries for each company's QuickBooks file.

### A2. Participant Background (n = 12)

- **Location:** All 12 based in the US.
- **Credentials:** All 12 report an active US CPA license.
- **Education:** All hold an accounting-specific degree. 8 of 12 hold a master's degree.
- **Experience:** Accounting-specific experience averages 5.4 years (median 5.3, range 3.8–8.0).
- **Sector:** The reported current role is split between 6 in public practice and 6 in industry.
- **Public accounting background:** 11 of 12 have public accounting experience, and half (6) have worked at a Big Four firm.

- **Seniority:** 2 are staff/associate level (17%), 7 are senior accountant or senior auditor (58%), and 3 are manager or above (25%). None are partner, director, or executive level.
- **Specialization:** 5 in general ledger, close, and reporting; 4 in audit/assurance; 2 in accounting management or controllership; 1 in tax.
- **Hands-on work:** All 12 were currently practicing and still doing hands-on preparation work. 7 of them also review others' work.
- **Prior AI exposure:** 4 had previously worked on a Mercor AI-data project; 2 describe using AI tools in their current job.

### A3. Participant Free-Form Responses

Because participant task order was randomized, the task number referred to in the following comments may not match the task number given in Appendix A1.

| If anything felt unrealistic compared with your real work, please describe it.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | What, if anything, was the most difficult part of these tasks?                                                                                                                                                                                                                                           |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>1 Based on my experience, the tasks given from my manager are much more straightforward and somewhat a subset of the task that I did during the study. For instance, in one of the cases, there seems to be multiple accounting concepts overlapped with each other, including revenue, COGS, leases, etc. In corporate setting, it is usually segregated by multiple departments (Revenue would have their own, leases have their own, etc.). The whole purpose of this is that they master a certain field and would be less likely to make an error due to the complexity nature. Even when they have to make management judgment, they have enough experience to provide inputs based on historical practice.</p> | <p>The most difficult part of the tasks has to be the amount of data that one has to pull and to digest before performing the tasks. Some workpapers may be irrelevant and are just noise, but we wouldn't be sure unless I open it or ask Claude if this would be relevant for the particular task.</p> |
| <p>2 —</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <p>The most difficult part was the first task where it was a non profit client where there were allocation entries that needed to be made.</p>                                                                                                                                                           |
| <p>3 In my day to day work, I would probably be given less direction, but this is more a product of doing the same sort of work frequently.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <p>Some of the work was a little bit outside my normal area of expertise. For instance, I routinely prepare journal entries, but not necessarily for a nonprofit organization. The AI assistant was very helpful in bridging these gaps though.</p>                                                      |
| <p>4 nothing felt unrealistic with the tasks</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | <p>Making sure I understood the full scope or if I was providing enough information. In reality I can usually ask my client questions as I work so that is the only difference</p>                                                                                                                       |
| <p>5 —</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <p>Understand the questions, what is needed for the task</p>                                                                                                                                                                                                                                             |

| If anything felt unrealistic compared with your real work, please describe it.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | What, if anything, was the most difficult part of these tasks?                                                                                                                                                                                                                                                                                                                                                                                                                        |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>6 I generally perform tasks along these lines in my every day work, but the tasks here felt very open concept and more "explore and figure it out" whereas tasks at my job I have never done I would walk through or be shown examples before just going forward to complete them.</p>                                                                                                                                                                                                                                                                                                           | <p>I found it difficult trying to navigate through the amount of information and files provided. It seemed like there was enough info to do almost any task, but that means I had to sort through all of it and remember which items I used if I had to go back to something.</p>                                                                                                                                                                                                     |
| <p>7 For Task 4, I would have the actual PMS to pull the data myself for validity check as well as being able to talk through the requirement more thoroughly with a supervisor.</p>                                                                                                                                                                                                                                                                                                                                                                                                                | <p>checking AI answers and sources against the raw data to make sure the data is being pulled correctly</p>                                                                                                                                                                                                                                                                                                                                                                           |
| <p>8 Usually there would be someone to ask questions to. When you are involved with a company for a few months/years, you already know where to find files/data and how to manipulate it. And you usually have accounting software to help you build reports or analysis rather than having to do something from scratch</p>                                                                                                                                                                                                                                                                        | <p>for me, it was finding the best way to present the data. i.e. creat live formulas, make any tables look well organized, use sources etc.</p>                                                                                                                                                                                                                                                                                                                                       |
| <p>9 —</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <p>Task #4 because the events had a lot of things going on there wasnt just one budget report and then one event report there were a lot of other items like food &amp; beverga ethat shouldnt be included in the events total but was charge to the clients together and I found the totals of the events minus the F&amp;B didnt add up to the other report where I added everything together (venue, service charge and other two items that made up the total per the budget)</p> |
| <p>10 It could be seen either way. Realistic in that if you had all of this information this cleanly in the real world, I could see how these tasks could be completed within 3 hours. But in the real world, you don't have all of this information. You go back-and-forth with so many stakeholders (internal and external). Tasks of this magnitude takes days and are worked on a few days before close, then several days throughout month end close (first 4 days of the month). I am in month end close right now and I could not imagine getting these tasks done in less than 3 hours.</p> | <p>Time. One non-AI task felt like there was so much information and so much to do that I didn't get to conclude, cleanly wrap up before running out of time. On the opposite end of the spectrum, the AI assisted tasks felt incomplete because I completed the tasks well before the 3 hour mark.</p>                                                                                                                                                                               |
| <p>11 real work has colleagues and supervisors you can brainstorm with and ultimately managers there that can help figure things out with you when you get stuck.</p>                                                                                                                                                                                                                                                                                                                                                                                                                               | <p>The most difficult part was having to work on stuff that I don't have hands on experience with in a work environment. It was also having an expectation to complete a task in 3 hours with our without AI and that is difficult given how much AI can expedite the solution.</p>                                                                                                                                                                                                   |
| <p>12 —</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | <p>It was difficult being unfamiliar with the industry and the files. There were some nuances that took some time to learn.</p>                                                                                                                                                                                                                                                                                                                                                       |

### Follow-up with the two experts who rated the tasks as very unrealistic

"Hi Aden! Yes, the excel tasks themselves are very much realistic."

"Hi Aden! Sorry for the delayed response, yes the tasks themselves were extremely realistic!"