MODERN AI

WHITEPAPER | A MODERN AI LABS TEST

Can You Trust the Number?

A Double-Blind Test of AI on Messy Workforce Data

18 October 2026 | Testing by Modern AI Labs

modernai.io

Key findings

Double-blind test of Claude and DeepSeek, each alone and with Modern Gravity, on synthetic workforce data.

98.1% vs 43.8%

Accuracy, Claude with Gravity against Claude alone

Right outcome on scored answers: 103 of 105 against 46 of 105. A right outcome is a right figure, or an honest refusal where the answer key expects no figure.

0 vs 46

Confidently wrong answers, of 105 scored

0.0% for Claude with Gravity (upper bound of the interval 3.5%) against 43.8% for Claude alone.

32% fewer

Agent turns per answer, Claude

A mean of 4.1 turns with Gravity against 6.1 alone. This counts steps, not time or cost: Gravity answers took longer by the clock, and with Gravity's answer check each answer used more tokens in total (about 243,000 against 203,000). Measured after the run, not pre-registered.

94.3% vs 39.0%

Accuracy with an open model (DeepSeek)

DeepSeek with Gravity, right outcome: 99 of 105 against 41 of 105 for DeepSeek alone. Confidently wrong: 1.0% against 46.7%.

Before you quote a number: part of the gain in right outcomes is honest refusals where the answer key expected no figure (35 of Claude with Gravity's 103 right outcomes). Counting only right figures confirmed by both the program and the audit, Claude with Gravity had 63 right figures against 43 for Claude alone, of 105. Synthetic workforce data, one domain, 43 questions (35 scored), graded blind and audited (Section 2.9).

Executive summary

Results for finance, HR and operations leaders checking AI-generated workforce numbers. We tested Claude, a frontier AI model, alone and with Modern Gravity™, a layer that makes the AI answer from the company's certified business definitions. The test covered 43 questions on synthetic workforce data that has the data issues real company systems have. Each question was asked three times of each setup. Every grade in this paper was checked by an independent audit (Section 2.9).

Accuracy. We count an answer as a right outcome when it gives the right figure, or honestly declines to give one where the answer key says none should be given (an undefined measure, an empty or future period, or, under the rule in force when the key was frozen, a very small group). Claude alone had a right outcome on **43.8%** of scored answers (46 of 105), and Claude with Gravity on **98.1%** (103 of 105). DeepSeek, an open model, alone: 39.0% (41 of 105). DeepSeek with Gravity: 94.3% (99 of 105). The audit moved these numbers in Gravity's favor; the program-only numbers are in Section 2.9.

Risk result. Claude with Gravity stated a wrong figure without warning on none of the 105 scored answers (0.0%; the upper bound of the interval is 3.5%), against 43.8% (46 of 105) for Claude alone. DeepSeek: 1.0% (1 of 105) with Gravity against 46.7% (49 of 105) alone.

Turns. Claude with Gravity needed a mean of **4.1 agent turns** per answer against **6.1** for Claude alone, which is **32% fewer** (median 3 against 6). Accuracy by right outcome and turns were measured after the run; neither was one of our four pre-registered hypotheses. Gravity answers took longer by the clock (median 19.9 seconds against 16.3), so this paper never says faster. Fewer agent turns is an execution measure; this test does not establish staff capacity released, cash savings or shorter business workflows. Fewer turns is also not fewer tokens: with Gravity's own answer check, each answer used about 19% more tokens in total (about 243,000 against 203,000) and cost more to run in this test.

> READ BEFORE YOU QUOTE A NUMBER
>
> **What you must know before you quote either number.** Part of the gain in right outcomes is honest refusals where the answer key expected no figure: 35 of Claude with Gravity's 103 right outcomes. Counting only right figures confirmed by both the program and the audit, Claude with Gravity had 63 right figures against 43 for Claude alone (6 questions against none, p = 0.031). Twenty of the 35 honest refusals rest on a small-number rule we have since reversed (Section 7). No person has hand-graded a sample of these answers; the audit in Section 2.9 replaced the person check that the pre-registration named.

What this is for. Right answers are the minimum. The value comes next: Gravity is meant to make checked numbers the base for the company's planning, analysis and decisions. This test measured the first part, answer errors.

The risk we set out to measure. Ask a business AI how many people you started last month with, and it answers in one confident sentence. Whether the number is right depends on things the sentence never shows: a month that is only partly done, rows with no customer recorded, an exit entered after its month closed. We built a synthetic workforce database with 19 kinds of such problems, wrote the correct answer to every question from the raw rows before any AI saw it, and had the answers blinded and graded before anyone could see which setup wrote which answer. The same figure came back on all three repeats for 88.4% of questions with Gravity against 53.5% for Claude alone. We report all four pre-registered hypotheses as tested on these 43 questions, including the one that did not reach significance and the places where Gravity was wrong.

1. The problem: an answer that looks right

A finance partner, an HR leader or a regional operator asks a plain question and gets a plain number back. The number is fluent, specific and formatted well. There is nothing on the page that says it is wrong.

Business data is rarely clean. Some rows have no customer. Some numbers are blank. A few rows appear twice. A month is only partly done. Somebody entered an exit after its month was closed. Each of these has a defensible handling and an easy mistake, and the mistake produces a number that looks exactly like a right one.

> That is the cost this paper measures: not the answers an AI refuses to give, but the answers it gives with full confidence and no warning. A refusal leaves work to resolve. A wrong number in a board pack can undermine trust in every number after it.

We wanted to know four things. How often does a frontier model alone give a confidently wrong answer on messy data? Does certified context change that, and by how much? Does the answer hold when the model is an open one that a company could run on its own terms? And does the AI take fewer steps to get there?

2. How we tested it, in plain words

2.1 The data

The database is synthetic, so no customer data touched any model. It describes a services firm of about 850 people with 12 client accounts of very different sizes, over 36 months plus the current, partly done month. It has the same seven tables a real firm's people and hiring system holds.

The database was built on the day of the run (2026-10-15) with a fixed seed, so the current month is partly done the way it is in real life and a rebuild gives identical rows.

2.2 The 19 real-world data issues

Each issue is built in at a known place and count, listed in a file the grader reads, and confirmed by a separate scan of the raw rows.

Code	What is in the data
D1	some rows have no customer recorded
D2	some headcount numbers are blank in the source
D3	some exits have no exit type
D4	some rows appear twice in the source
D5	some rows are dated in the future
D6	the current month is only partly done
D7	exit types are spelled in mixed ways
D8	an exit was added after its month had closed
D9	some people left and rejoined
D10	a quarter had no exits at all
D11	one month has no rows in any table
D12	an exit has no customer
D13	a movement date falls in a different month from its month column
D14	the monthly snapshot does not add up to the movements
D15	some days-to-fill or dates are blank, negative or impossible
D16	a customer was renamed over time
D17	some groups are very small
D18	offer statuses are not what the name suggests

Code	What is in the data
D19	open positions are only a draft measure

2.3 The questions

> SCOPE NOTE
>
> The test was pre-registered with 69 questions; 26 questions whose answer rule was not set before the run were removed after the run and will be retested in the next run, and everything in this paper is on the remaining 43 questions (129 answers per setup, 516 in all, each question asked three times of each setup).

They are written in the words a business user would use, not in database terms. By type: 8 single figures, 1 month-by-month table, 1 weekly count, 4 splits by customer, 5 follow-up questions ("compare it with Q1"), 3 questions that need a reason ("why did it change?"), 3 questions that should be refused or clarified (a measure nobody has defined, an ambiguous period), 3 plain "just give me the number" requests, 10 data-quality probes, and 5 multi-layered problems that need several steps. The 3 "why" questions and the 5 multi-layered problems have no single right figure; the figure grader does not score them (Section 2.6), so 35 questions, 105 answers per setup, are scored.

2.4 The correct answers

Before any model saw a question, an independent program computed the correct answer to every question straight from the raw rows. It uses no code from the data builder and none from Gravity. Where a question has more than one reasonable reading and no written rule picks one, the grader is set up to accept each reasonable reading it was written to recognize, and the results say which one an answer used. Section 7 lists where it could not do this.

2.5 What each setup was given

Every setup got the same questions, the same wording, the same database and the same anchor date for the data (2026-10-15). The answers were produced over 15 to 17 October; none of the 43 questions is among the twelve the run flagged as reading the current date.

Setup	What it was given
The model alone	The model, a read-only login to the database, a database client, and a plain description of each table. Nothing else. No certified definitions. Its one instruction was "Answer the user's question from the database."
The model with Gravity	The same model and the same database, with Gravity installed the way a customer installs it. Gravity supplies the business's certified definitions. Every answer separates certified figures (the definition, the period and the calendar they were certified for) from the model's own reading, which is kept apart.

The setups ran for two models: Claude Sonnet 5.5 (claude-sonnet-5-5), a frontier model, and DeepSeek V4.1 Flash (deepseek/deepseek-v4.1-flash), an open model reached through a local gateway. Each setup ran in its own clean folder, in a real coding-assistant session, not a script. Versions: Claude Code 2.1.301 and 2.1.302; Gravity release 0.1.11.

The model alone had a fair chance: it could query the data directly and was free to be careful. Section 3.4 shows where it was.

2.6 Blind grading

- **Pre-registration.** The question set, the correct answers, the grading program and four hypotheses were frozen, and their fingerprints written with a timestamp (2026-10-15, 21:04:58 UTC), before the first answer was produced (21:04:59 UTC). Section 2.9 discloses the post-run replacement of the registered person check and its effect on the grades.
- **Answer cards.** Every answer was turned into a card. Anything that reveals the setup (section headings, certification labels, product and tool names) was removed or replaced with neutral words. The cards were shuffled and given random ids, and the key was sealed in a file the graders never read.
- **Two graders.** First, a program compared every figure on every card with the correct answer and labeled it: correct, a wrong figure stated as fact, a figure where none should be given, a correct refusal or clarification, or a wrong refusal. (It can also label a figure for a plausible but different definition; no answer in these 43 questions received that label.) It does not score the 8 open-ended questions (24 answers per setup). Second, a grader model from a different model family than the answers read each card with the question and the correct answer, and judged whether the card answered the question asked, stated anything as fact that the correct answer does not support, and told the reader when it could not answer. Our first-choice second grader, a model from a third family, returned no text for 14 of the 516 cards. Grading therefore fell back automatically to another model from a different family than the answering models, and every second-grader result in this paper comes from that fallback.
- **How the program and the second grader combine.** In the program's own grades, a "confidently wrong" answer is one the program labeled wrong, with no warning to the reader according to the second grader. Those program grades were then audited (Section 2.9), and the numbers in this paper are the audited grades.
- **Unblinding in order.** Each grader wrote a completion record to an append-only, hash-chained log: the figure program at 2026-10-17 04:45:48 UTC and the second grader at 12:16:17 UTC. The tool that opens the key refuses to run unless both records exist and match their score files, and it logs its own timestamp (12:16:17 UTC, after both), so the log proves grading finished before the key was opened.
- **Separate folders and file isolation.** Every setup ran in its own clean folder. The DeepSeek setups were also confined to their own folder, proven by a live test before they ran. That confinement was

added after an earlier run showed that two DeepSeek setups had read files they should not have (see Section 7); those answers are kept on disk, never graded, and the DeepSeek setups were asked again under isolation. The Claude setups ran before the confinement existed; the earlier check found outside-folder reads in 0 of 413 Claude answers.

2.7 What we mean by ”confidently wrong”

> **DEFINITION: CONFIDENTLY WRONG**
>
> A confidently wrong answer states a wrong figure as fact with no warning, or states a figure where none should be given (for example, a number for a period that has not happened, or a figure for a measure nobody has defined). An answer that is wrong but tells the reader to be careful is counted separately.

2.8 The four hypotheses, written before the run

ID	Hypothesis	Test
H1	Claude with Gravity gives fewer confidently wrong answers than Claude alone.	Paired by question; exact McNemar test (two-sided) on "confidently wrong on most repeats"; rates with 95% Wilson intervals
H2	DeepSeek with Gravity gives fewer confidently wrong answers than DeepSeek alone.	Same as H1
H3	DeepSeek with Gravity versus Claude alone. Stated as a comparison, not an assumed win.	Same as H1, two-sided
H4	The Gravity setups give the same figure more consistently across repeats than their counterparts.	Paired by question; exact McNemar test on "same figure in every repeat"

The four hypotheses were written before the run and are tested here on the 43 questions. Only these four get a pre-registered headline. Accuracy by right outcome and turns are measured results that were not pre-registered. Every other comparison in the paper is labeled exploratory.

2.9 How the grading was checked

The pre-registration named a person check of a 10% sample of the grading. We replaced it, on our founder's decision of 2026-10-18, with a wider audit of all 516 answers on these 43 questions. No person has hand-graded a sample. The audit did four things.

- **The answer key was recomputed from the raw data** by an independent query run, with no value mismatches against the frozen key.
- **A third grader, from another model family, graded all 516 answers** blind to the setup. On the 420 scored answers it agreed with the program on right outcome 87.4% of the time (Cohen's

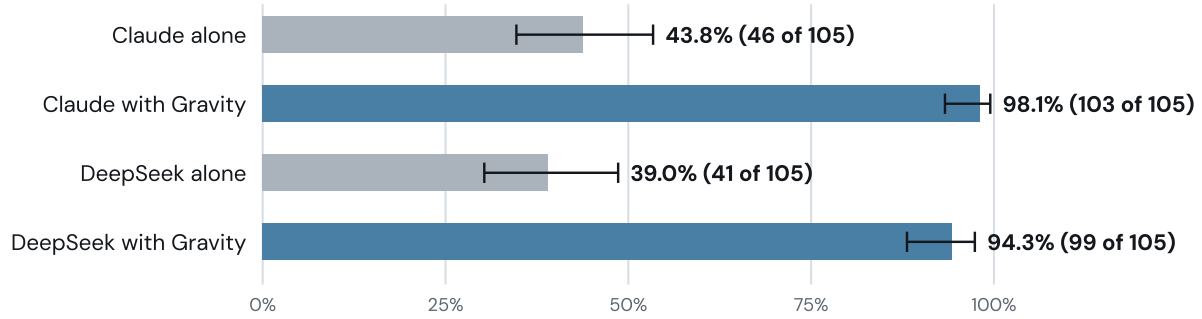
kappa 0.71) and on confidently wrong 88.6% of the time (kappa 0.65).

- **The 88 disagreements were settled by a fourth grader model that is none of the answering models**, under written rules, blind to the setup. Where the program and the third grader agreed, their grade stands.
- **The result is the final grade for each answer**, and every figure in this paper uses it. The grading program alone, before the audit, gave right outcomes of 48.6% (Claude alone), 93.3% (Claude with Gravity), 43.8% (DeepSeek alone) and 89.5% (DeepSeek with Gravity), and confidently wrong rates of 37.1%, 1.9%, 47.6% and 8.6%. The audit moved 19, 6, 17 and 9 answers in the four setups, and it moved them in the direction of fewer right outcomes without Gravity and more with it.

> The audit cannot remove every error: all graders and the tie-break are models, and the 88 settled cases are judgments made under written rules. Consistency across repeats and turns are computed from the run records and are not changed by the audit.

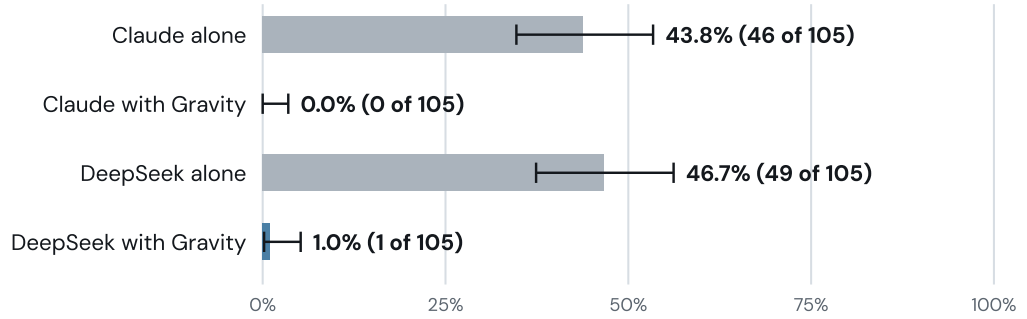
3. Results: Claude

Accuracy: right outcome by setup



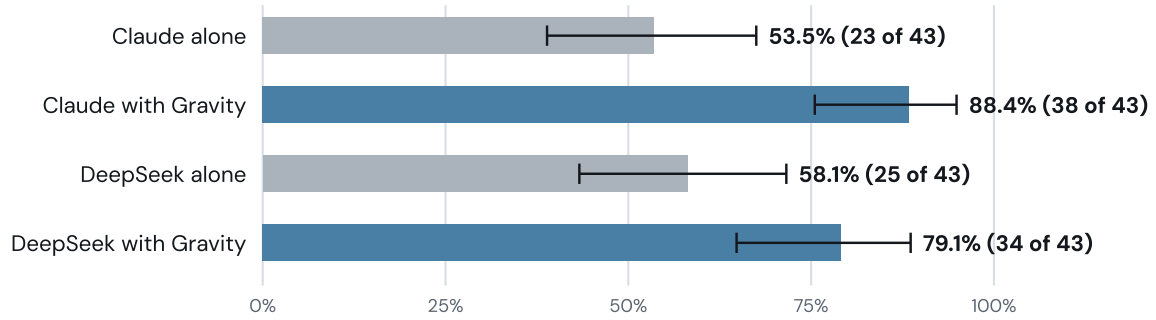
Share of the 105 scored answers per setup that gave a right figure or an honest refusal where the answer key expects no figure. Whiskers are 95% Wilson intervals. Audited grades.

Confidently wrong answers by setup



Share of the 105 scored answers per setup that stated a wrong figure as fact with no warning. Whiskers are 95% Wilson intervals. Audited grades.

Consistency: same figure in every repeat



Share of the 43 questions where all three repeats gave the same figure. Whiskers are 95% Wilson intervals. A steady answer can still be wrong.

3.1 The two lead numbers

DEFINITION: RIGHT OUTCOME

Accuracy: the right outcome. A right outcome is a right figure, or an honest refusal or clarification where the answer key says no figure should be given. 105 scored answers per setup. Grades are the audited grades. Intervals are 95% Wilson intervals.

	Right outcome	Right figure (confirmed by the program and the audit)
Claude alone	43.8% (46 of 105; 34.7% to 53.4%)	43 of 105
Claude with Gravity	98.1% (103 of 105; 93.3% to 99.5%)	63 of 105

Claude alone gave no honest refusal by the program's label, so almost all of its right outcomes are figures; Claude with Gravity gave 35 honest refusals. The audit does not regrade the kind of right outcome, so the right-figure column counts only answers the program labeled a correct figure and the audit also accepted, which understates the figures slightly for every setup. On right outcomes

the comparison is clear: 19 questions where only Claude with Gravity was right, none where only Claude alone was, exact paired test p < 0.001. On confirmed right figures alone it is smaller but still significant: 6 questions against none, p = 0.031.

Turns: steps to answer. A turn is one step of the AI's work loop (a model call, often with a database or tool call). We summed the turns recorded in each answer's run record, adding the parts of a follow-up question together. This was measured after the run and was not pre-registered; the audit does not change it.

	Mean turns (129 answers)	Median
Claude alone	6.1	6
Claude with Gravity	4.1	3
Reduction in the mean	32%	

Claude with Gravity took fewer steps, but not less time. The median answer took 19.9 seconds with Gravity against 16.3 alone, and the 90th percentile 59.7 against 32.4 (Section 5). Fewer turns does not mean faster, and we do not say it does. The turns result is also specific to Claude: DeepSeek with Gravity did not take fewer turns than DeepSeek alone (Section 4.4).

3.2 H1: confidently wrong answers

Confidently wrong answers, as a share of scored answers: **0.0%** (0 of 105; 0.0% to 3.5%) with Gravity, against **43.8%** (46 of 105; 34.7% to 53.4%) for Claude alone.

In question terms: Claude with Gravity was confidently wrong on most repeats in 0 questions where Claude alone was not, and Claude alone was confidently wrong in 16 where Gravity was not, out of 35 scored questions; exact paired test p < 0.001 (0.00003). The result is in the direction predicted. Right figures (confirmed), the same comparison: 63 with Gravity against 43 alone, of 105; 6 questions one way and none the other; exact paired test p = 0.031.

3.3 H4: consistency across repeats

Asked the same question three times, Claude alone gave the same figure every time on **53.5%** of the 43 questions (23 of 43; 38.9% to 67.5%). With Gravity: **88.4%** (38 of 43; 75.5% to 94.9%). Paired test: 17 questions where only the Gravity setup was consistent and 2 where only Claude alone was; p < 0.001 (0.0007). A steady answer can still be wrong; consistency is measured separately from correctness.

3.4 Where the model alone was right

The model alone did well on plain questions. On questions with no data issue in the way, Claude alone had a right outcome on 23 of 24 answers (1 confidently wrong); Claude with Gravity had a right outcome on 24 (21 right figures and 3 honest refusals). Where a movement date sat in the wrong

month column (D13), Claude alone had 11 right outcomes and 1 confidently wrong answer; with Gravity, 12 and none. Where people left and rejoined (D9), Claude alone had 6 right outcomes and 3 confidently wrong; with Gravity, 9 and none.

Across the 35 scored questions, both setups had a right outcome on most repeats on 16. Claude with Gravity had one where Claude alone did not on 19. There is no scored question on which Claude alone had a right outcome and Claude with Gravity did not. This is a smaller list of places where the model alone did well than a reader may expect, and it is what the audited grades show.

3.5 Where Gravity was wrong or held back

After the audit, 2 of the 105 scored Claude-with-Gravity answers were not a right outcome, and none was confidently wrong. Both are follow-up questions about attrition.

Question	Problem	What happened	Our reading
Q54, attrition in Q2 and "compare it with Q1", repeat 1	D10	The answer left out the second-quarter rate (26.1% on the key) and cautioned that its first-quarter zero may reflect incomplete exit records	Not a right outcome, not confidently wrong: it carried a warning
Q55, the same, in other words, repeat 3	D10	Not a right outcome after the audit	A two-part follow-up the first grader could not score cleanly (open grader item TB-B5)

The audit moved most of the Claude-with-Gravity answers the program had marked wrong on Q52, Q54 and Q55 to a right outcome after reading them against the key: for example, 26.1% matches the key's 26.13% after rounding, and an answer that plainly withholds a small group matches a key that withholds it. It left two as not a right outcome and removed the confidently wrong labels. That is the change the program's open grader items (TB-B5, TB-B6) had predicted.

Gravity did not improve every result: DeepSeek had one confidently wrong answer on the partly done month with and without Gravity (Section 6).

3.6 Worked examples

Five questions where the setups differ, chosen by a fixed rule (the biggest gap first, in either direction), summarized here from the cards the graders saw; the grades quoted in each example are the first grader's, and the totals in this paper use the audited grades. Customer names in the cards are synthetic and are removed. We describe three in detail and two briefly; the full cards are available on request.

Example 1. "Show opening headcount for each month last year." The correct answer: 715, 722, 726, 726, 729, 735, 737, 757, 757, 758, 762, 769. Data issue: some headcount numbers are blank in the source (D2).

- *Claude alone* gave a clean twelve-row table with January as 716 and July as 736. Both are wrong by one, with no warning. Graded: wrong figure stated as fact.
- *Claude with Gravity* gave the same table with 715 and 737, then "Metric: Opening headcount, Period: 2025-01-01 to 2025-12-31". Graded: correct.

Example 2. "How many people left each week last month?" The key expects no figure, because every week's group was under the small-group threshold of the time (D17).

- *Claude alone* put together a weekly table from daily records, 0, 1, 3, 3 and 4 exits, and added "about 2.75 a week". Graded: figure stated where the key withholds.
- *Claude with Gravity* answered "Withheld: the group is too small to show, so no figure is given." Graded: correct refusal.
- **Read this one with the rule change in mind.** On 2026-10-16 we decided that counts are always shown, however small. Under that rule a weekly table is the expected answer and the withheld reply is the one at fault. The key was frozen before the decision, so this paper scores the old way (Section 7).

Example 3. "How many people left in the last 12 months, by customer?" The key withholds the split for the same small-group reason. Claude alone gave a full table of 13 rows (12 customers and a "no customer recorded" row) with a total of 115. Claude with Gravity withheld. Same reading as Example 2.

Example 4. "What was our attrition rate in Q2 2026, and why did it change?" then "Compare it with Q1". Claude with Gravity gave 26.1% for Q2 and 0% for Q1, noting the table shows no exits in Q1, which is the key's default reading. The first grader could not score the two-part question and marked all three answers wrong (open item TB-B5). The audit judged two of the three a right outcome (26.1% matches the key's 26.13%) and left one as not a right outcome. The paper counts that one against Gravity.

Example 5. "How many people left last month, by customer?" Same pattern as Example 3, on the same rule: Claude alone gave a customer split of 11 exits; Claude with Gravity withheld.

4. Results: DeepSeek, the open model

4.1 Headline

	Right outcome	Right figure (confirmed)	Confidently wrong	Same figure in every repeat
DeepSeek alone	39.0% (41 of 105; 30.3% to 48.6%)	38 of 105	46.7% (49 of 105; 37.4% to 56.2%)	58.1% (25 of 43; 43.3% to 71.6%)

	Right outcome	Right figure (confirmed)	Confidently wrong	Same figure in every repeat
DeepSeek with Gravity	94.3% (99 of 105; 88.1% to 97.4%)	63 of 105	1.0% (1 of 105; 0.2% to 5.2%)	79.1% (34 of 43; 64.8% to 88.6%)

DeepSeek alone gave no honest refusal (0 of 105); DeepSeek with Gravity gave 31 (by the program's label).

4.2 H2: DeepSeek, with and without Gravity

Confidently wrong answers: **1.0%** for DeepSeek with Gravity against **46.7%** for DeepSeek alone. DeepSeek with Gravity was confidently wrong on most repeats in 0 questions where DeepSeek alone was not, and DeepSeek alone in 16 where Gravity was not, of 35; exact paired test $p < 0.001$ (0.00003). The result is in the direction predicted. Right figures (confirmed): 63 against 38, of 105; 8 questions one way and none the other; $p = 0.008$.

4.3 H3: DeepSeek with Gravity against Claude alone

Confidently wrong answers: **1.0%** (1 of 105) for DeepSeek with Gravity against **43.8%** (46 of 105) for Claude alone. DeepSeek with Gravity was confidently wrong on most repeats in 0 questions where Claude alone was not, and Claude alone in 16 where DeepSeek with Gravity was not, of 35; exact paired test **$p < 0.001$** (0.00003). This is a comparison, stated before the run without an assumed winner, and it falls in favor of the open model with Gravity on this measure. It is one domain, 35 scored questions and a model pair, so we do not claim an open model with Gravity beats a frontier model in general. Right figures (confirmed), the same pair: 63 against 43, of 105; 6 questions against none; $p = 0.031$.

4.4 H4 and turns for DeepSeek

DeepSeek alone gave the same figure every time on 58.1% of the 43 questions; with Gravity, 79.1%. Paired test: 14 questions where only the Gravity setup was consistent and 5 where only DeepSeek alone was; **$p = 0.064$, which is not significant**, so for the open model H4 did not reach the 0.05 line.

Turns do not show the Claude pattern. DeepSeek alone took a mean of 11.1 turns per answer (median 9) and DeepSeek with Gravity 12.5 (median 9), counting the 122 of its 129 answers that finished. There is no turns reduction for the open model.

4.5 Answers from a backup model

No DeepSeek answer came from a backup model (0 of 258). For the Claude setups, the report tool lists all 258 answers as "from a backup model". That flag is an artifact: the usage records also name Claude's small internal helper model (claude-haiku-5-5), which accounts for a small share of the

output in the usage records, while every answering message in the run records came from claude-sonnet-5-5. We could not verify the backup model configured for the DeepSeek setups, and answer-level records do not establish which model handled every part of the processing.

5. Cost and time

	Answers	Median seconds	90th percentile seconds	Cost
Claude alone	129	16.30	32.38	0.76% of the weekly Max limit
Claude with Gravity	129	19.88	59.72	1.74% of the weekly Max limit
DeepSeek alone	129	47.42	176.65	$1.02
DeepSeek with Gravity	129	71.24	311.70	$1.89

Claude usage is shown as a share of the weekly plan limit, because those runs used a subscription plan; the share is an estimate converted from token counts. DeepSeek cost is measured (tokens times the saved price of the model that answered): $0.0079 per answer alone and $0.0146 with Gravity.

Gravity costs more and takes longer in both models: about twice the plan usage for Claude and about twice the dollars for DeepSeek. Seven DeepSeek-with-Gravity answers timed out twice; the second result stands (Section 7).

6. What we learned, data problem by data problem

For each problem: why it trips an AI that reads rows, and what we saw. Each cell below counts answers to the 43 questions that problem touches: right outcome / confidently wrong / of the right outcomes, honest refusals. The rest of the answers (wrong but with a warning, or not scored) are not shown in the cell, so cells do not sum to the answer count; a cell of 0 / 0 / 0 means the answers were not scored by the figure program. Grades are the audited grades. With few questions per problem the table is descriptive, not a test.

Code	Why it trips an AI that reads rows
D1	A blank customer is easy to leave out of a by-customer table, so the rows no longer add up to the total.
D2	A blank number can be read as zero, or skipped silently in a total.
D3	A blank exit type is neither "voluntary" nor anything else a filter expects.

Code	Why it trips an AI that reads rows
D4	A repeated row is counted twice unless somebody decides it is not.
D5	A future-dated row invites a figure for a period that has not happened.
D6	A partly done month looks like a low month.
D7	"Voluntary" and " VOLUNTARY" are one thing to a person and two to a filter.
D8	A late entry changes a closed month, and the number moves.
D9	A rehire is both an exit and a join, and is easy to net out.
D10	A quarter with no exits invites "0%" or "not available", and the two mean different things.
D11	An empty month is a missing month, not a month of zero.
D12	An exit with no customer is easy to lose from a customer split.
D13	A date in the wrong month column sends the count to the wrong month.
D14	Two tables that should agree do not, and a model may adjust one to fit the other.
D15	Negative or impossible values pull averages and tenures.
D16	A renamed customer splits into two rows, or merges two that should stay apart.
D17	A very small group makes a rate swing, and a count over a small base needs care.
D18	A status called "offered" may not mean what a reader assumes.
D19	A draft measure invites a figure that nobody has certified.

What we saw (right outcome / confidently wrong / of the right outcomes, honest refusals by the program's label):

Code	Claude alone	Claude with Gravity	DeepSeek alone	DeepSeek with Gravity
D1	0 / 12 / 0	12 / 0 / 11	0 / 11 / 0	12 / 0 / 12
D2	0 / 6 / 0	6 / 0 / 0	0 / 6 / 0	6 / 0 / 0
D3	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
D4	3 / 0 / 0	3 / 0 / 0	2 / 1 / 0	3 / 0 / 0
D5	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0
D6	0 / 2 / 0	3 / 0 / 3	0 / 1 / 0	0 / 1 / 0
D7	0 / 3 / 0	3 / 0 / 3	0 / 3 / 0	3 / 0 / 3
D8	3 / 2 / 0	6 / 0 / 0	3 / 2 / 0	6 / 0 / 0
D9	6 / 3 / 0	9 / 0 / 0	3 / 6 / 0	9 / 0 / 0
D10	0 / 6 / 0	4 / 0 / 0	0 / 5 / 0	4 / 0 / 0
D11	0 / 0 / 0	3 / 0 / 3	0 / 2 / 0	3 / 0 / 3

Code	Claude alone	Claude with Gravity	DeepSeek alone	DeepSeek with Gravity
D12	0 / 9 / 0	9 / 0 / 8	0 / 8 / 0	9 / 0 / 9
D13	11 / 1 / 0	12 / 0 / 0	9 / 2 / 0	12 / 0 / 0
D14	0 / 2 / 0	3 / 0 / 0	0 / 1 / 0	3 / 0 / 0
D15	0 / 0 / 0	3 / 0 / 0	0 / 0 / 0	3 / 0 / 0
D16	0 / 3 / 0	3 / 0 / 3	0 / 3 / 0	3 / 0 / 3
D17	0 / 15 / 0	15 / 0 / 15	0 / 15 / 0	15 / 0 / 14
D18	0 / 6 / 0	9 / 0 / 9	0 / 7 / 0	9 / 0 / 9
D19	0 / 2 / 0	3 / 0 / 3	0 / 2 / 0	2 / 0 / 2
No data issue	23 / 1 / 0	24 / 0 / 3	24 / 0 / 0	24 / 0 / 3

6.1 Three things the table supports

Each of these is a statement about this run only, and each stays in the paper only because the table supports it.

1. **A missing customer was the clearest case.** Where rows or exits had no customer recorded (D1, D12), Claude alone was confidently wrong on 12 and 9 answers and Claude with Gravity on none, with most of its right outcomes honest refusals. Blank headcount numbers (D2) went from 6 confidently wrong answers to none.
2. **Much of Gravity's lead is honest refusal, and part of that rests on a rule we have reversed.** Claude with Gravity gave 35 honest refusals, 20 of them on seven questions about small groups (D17 and the by-customer questions), where the answer key of the time expected no figure. Section 7 shows the score with those seven questions removed.
3. **Gravity did not help everywhere.** For DeepSeek, a partly done month (D6) produced 1 confidently wrong answer with and without Gravity and no right outcome in either; D3 and D5 had no scored answers at all. The small base of questions per problem means this table cannot tell where Gravity does not help; it only shows where this run found nothing.

7. Limitations

We list these in the paper, not in a footnote.

- **Synthetic data.** The data is built to contain known problems. Real data has problems we did not think of, and real rates may differ.
- **One domain.** Workforce headcount, exits, hiring and offers. We make no claim about finance, supply chain or any other area.

- **One tool version.** One coding-assistant version and one Gravity release were used (Claude Code 2.1.301 and 2.1.302; Gravity 0.1.11). Later releases will differ.
- **A small-number rule we have since reversed.** When the answer key was frozen, it said a split value over a small group must be withheld. On 2026-10-16 our founder decided the opposite: a count is always shown, and only a percentage over a small base carries a note. The grader still applies the old expectation: a withheld reply is scored as a correct refusal, and a shown count as a figure stated where the key withholds. This affects seven of the 35 scored questions (Q21, Q26, Q28, Q30, Q32, Q52, Q59, 21 answers per setup). Removing those seven questions, the right outcome is 54.8% for Claude alone (46 of 84; 44.1% to 65.0%) and 97.6% for Claude with Gravity (82 of 84; 91.7% to 99.3%), and 48.8% for DeepSeek alone and 92.9% for DeepSeek with Gravity. This check was made after the run, is exploratory, and is not part of the pre-registration.
- **Grader items found.** The first grader misread some answers: compound follow-up questions (Q54, Q55) and a count of months in prose read as a figure (Q52). The audit settled most of them (Section 3.5); two Claude-with-Gravity answers remain not a right outcome. The first grader's frozen key and code were not edited.
- **The graders are models, and no person hand-graded a sample.** The first grader is a program with its own limits. The second grader, the third grader and the model that settled the 88 disagreements can each be wrong, and the settled cases are judgments made under written rules. The pre-registration named a person check of a 10% sample; the audit in Section 2.9 replaced it.
- **Writing style can hint at the setup.** Blinding removes labels, headings, tool names and code, but a reader could still guess which setup wrote a card from its style.
- **Questions with more than one reasonable reading.** For some questions no written rule says which reading is right. The grader accepts each reasonable reading it was written to recognize, and the results show which one was used; some readings it does not recognize, and those answers count against the setup that gave them.
- **Per-question comparisons use the majority of three repeats,** and the per-problem table is descriptive, not a test.
- **Turns was not pre-registered, and it is not time or cost.** It was computed from the run records after grading. It counts agent steps. Gravity answers took longer by the clock and cost more.
- **Re-asked answers.** An answer that timed out or broke was asked once more, the same way in every setup, and the second result stands whatever it is. Eleven answers on these questions, all from DeepSeek with Gravity, were asked a second time; 7 of the 11 timed out again and stand as timeouts. The first attempts are kept on disk and are in no result. Two answers that crossed midnight UTC were set aside and asked again.
- **Infrastructure stops.** Three infrastructure events on the Claude with Gravity setup are in the log (all on 2026-10-15); every setup finished with all its answers.

- **DeepSeek setups were asked again.** In an earlier run, two DeepSeek setups read files outside their own folders: the correct answers in 17 answers of DeepSeek alone, and test files in 5 answers of DeepSeek with Gravity. The Claude setups read nothing outside their folders in that check (0 of 413). Those earlier DeepSeek answers (375 files) are kept on disk, never graded and never used in any claim. The DeepSeek setups were asked all questions again after each was proven unable to read outside its folder.
- **No competitor comparison.** This test compares a model with and without Gravity. It does not compare Gravity with any other product.
- **The paper is ours.** Modern AI built Gravity, and Modern AI Labs, our own internal testing team, ran the test. That is why the hypotheses were written first, the grading was blind and the bundle in the method and data appendix lets anyone re-check it.

8. Try one check on your own setup

You do not need a benchmark to see whether this matters to you. In your company-approved AI, pick data you are allowed to use and a question whose right answer you can check. Then paste this before your next question:

> Before you answer my next question from our data, tell me: which date range you are using; whether any month in that range is incomplete; how you treat blank or missing values; whether any rows are dated in the future; which exact definition of the measure you are using, and who approved it; and how many records the number rests on. Then answer, and show me the records and the calculation so I can check them.

Ask the same question again tomorrow on the same data, with the same explicit date range. If the number or its explanation changes, find out what changed before you rely on it. If your AI cannot show the records behind the figure, you cannot complete this check from its answer alone. Even if both answers match, check the figure against your own records: a steady answer can still be wrong.

Method and data appendix: what we publish and what is on request

The full method, the scoring rules and the version history of the benchmark live on our methodology page: modernai.io/methodology/gravitybench. This paper links there instead of repeating it.

Published with this paper:

- the headline and per-problem tables;
- the five worked examples above, chosen by a fixed rule;
- the pre-registration fingerprints and timestamp, which identify the files frozen before the first answer; the corresponding files are available on request for comparison.

Available on request to buyers, auditors and researchers: the questions and correct answers, every answer card and score file, the sealed key and the append-only log that proves the order, the claims register that traces every number in this paper to a source file and its fingerprint, the methods file, and the generated report. We keep the questions out of public view for now, which also keeps later runs of the same benchmark blind. To request them, use the evidence request form on the methodology page (modernai.io/methodology/gravitybench#evidence-request).

Never shared: raw session transcripts, internal paths, credentials, the data generator, and anything about how Gravity is built.