

Hard tasks for AI agents, with the proof attached

What we offer, our numbers, and how to start with three free tasks.

Written by IsoData

IsoData is a small team of engineers and scientists. We write training and evaluation tasks for AI agents and test each one against a frontier model before we ship it. We build around your brief, from expert-written examples and response reviews to runnable tasks with reference solutions, graders and test runs.

What we offer

Coding and terminal tasks. Software engineering, debugging and tool use across databases, build tools, browsers and scientific software. Our engineers build runnable tasks, reference solutions and tests for real workflows and edge cases.

Generalist training and evaluation. Everyday questions, chatbot conversations, writing, research and multi-step reasoning. Our contributors include experienced reviewers and project leads who can write examples, compare responses and check the quality of a model's answers.

Finance and operations tasks. Financial analysis, reconciliation, expense reviews and other work with documents, spreadsheets and business records. Our contributors combine finance experience with hands-on AI training.

Expert research and reasoning. Challenging questions in electrical engineering, mathematics, science and other specialist fields, including tasks designed to test models that can search the web. Each comes with a supported answer and clear grading criteria.

We also scope projects in medicine and social sciences, drawing on contributors with MD and PhD backgrounds. Across these fields, we can build training examples, red-team tests and evaluation rubrics, review model outputs, and document how agents use tools.

Need something else? Tell us what your project needs. We'll work with you on the task format, subject expertise and review process, then put together the right team.

You pay only for the tasks you accept. If one has a defect, we fix or replace it for free. The tasks we write for you are yours, and we never reuse them.

Our numbers

We built five sample tasks and gave each one to Claude Opus 5.5 five times. It solved 0 of 25 attempts.

Sample task	Field	Solved
G-code toolpath parity	CNC machining	0 / 5
Pharmacy count close, free-text log	Hospital pharmacy	0 / 5
Stream gauging field book	Hydrology	0 / 5
Fare engine certification	Test engineering	0 / 5
Expense claim audit	Accounts payable	0 / 5
Control: the same pharmacy shifts as a database	Hospital pharmacy	3 / 3

Method: Claude Opus 5.5 as a Claude Code agent with a shell, told not to open the solution, tests or web. $n = 5$ attempts per task, 3 for the control, run on 9 and 10 October 2026. Baseline: our reference solution scores 1 and an empty run scores 0 on every task, re-checked on 10 October 2026. Three tasks were regraded on their final version. We kept 5 of 14 candidates and dropped the rest after our own checks.

The last row is the control. It has the same shifts and the same answers as the pharmacy task, given as a clean database instead of a nurse's free-text log. The model solved it every time. The tasks are hard because of what the agent has to read and reason about, and the graders hold up.

Some of our metrics include four of our experts who wrote 402 reasoning tasks for one 5,000-task project, about 8% of it. Each one made the tested model fail at least two out of three times.

See the five sample tasks

What we can do for you

Tell us your task format and the domain. Within a week we send three tasks, free, each tested against a frontier model, or against yours if you give us access. If they hold up, we run a 20-task pilot, then deliver 5 to 10 hard tasks a week.

How to start

Write to us through our site, or reply to whoever sent you this page. Tell us your format, the domain and how you accept a task. We'll take it from there.

Get a free sample · isodatas.com/contact