

Agents that finish the work, and verify it

Three case studies of AI agents doing everyday work on Zero-AR: preparing tax returns, auditing freight invoices, and settling supplier invoices by email.

Day-to-day work, done on its own, and verified.

Case studies in summary, October 2026

Zero-AR, the Zero-layer Autonomous Runtime, code name Dante-1

The people, companies and documents in each study are fictional.

Each full case study is available on request.

Three kinds of everyday work, and one rule for done.

We built three AI agents on Zero-AR and gave them real work with real AI models. Here is what Zero-AR does, what we measured, and each study in a page.

Across three studies, Zero-AR never reported work as done when it wasn't. Every tax return through a Claude model finished with its checks passing; nine of ten supplier invoices settled by email with no one stepping in. And where no model could do the work, like the freight audit, it said so.

1	Work that takes more than one sitting	3
2	What we count	5
3	The three studies at a glance	7
4	Tax returns	9
5	Freight audits	10
6	Supplier invoices	11
7	What held in every study	13
	The limits of these studies	14
	About Zero-AR	15

Most real work does not finish in one conversation.

An AI agent is software built around an AI model: the model decides each next step, and the agent's tools carry it out. An agent can read a document or draft an email in seconds. Finishing the job usually takes longer.

-
- 1 **It waits on other people.** A supplier replies tomorrow. A customer answers next week. The work has to pause and pick up where it stopped.
 - 2 **It needs permission.** Sending an email or paying an invoice speaks for the company, so someone approves it first, or allows it in advance.
 - 3 **It needs checking.** A finished-looking answer can still be wrong. Somebody has to confirm the numbers add up and the facts come from the documents.
 - 4 **Things go wrong along the way.** A model provider has an outage, a server restarts, a check fails. The work has to survive all of it.
-

Zero-AR carries that work from start to a checked finish.

Zero-AR is the software an agent runs on. It keeps each piece of work, from its goal to everything that has happened since, in one place that lasts as long as the work does.

It waits for free. While the work waits for a reply or a decision, nothing runs and nothing is spent. The reply wakes it.

It acts only with permission. Every action that speaks for you waits for your approval, unless you allowed it in advance.

It checks before it calls work done. You decide what done means. The agent cannot mark its own work finished; the checks have to pass.

It stays within budget. Each step sets aside what it may spend before it runs, and the work asks before it goes past the budget you set.

It keeps a full record. Every step is written down as it happens, so the work survives a restart and you can see exactly what was done.

Our aim is simple to say: agents that carry everyday work to a checked finish on their own.

We count work that finished, passed its checks, and needed no one.

“On its own” is easy to claim and hard to measure. In each study we count it the same way:

Our measure

Work that finished with every check passed, without anyone stepping in, apart from the approvals the owner chose to keep, reported beside what the agent was allowed to do without asking.

Verified is our word for finished with every check passed.

Stepping in means a person had to answer a question, fix something, restart the work or redirect it.

Approvals the owner chose to keep do not count against the agent. Wanting to approve a payment yourself is a choice, not a failure.

What the agent was allowed to do is always stated beside the number, so a high figure cannot come from simply allowing everything.

We also count the opposite: work reported as verified that was wrong. That number should always be zero.

An agent on Zero-AR stops for a person for only three reasons.

An agent that stops whenever something goes wrong is not working on its own. On Zero-AR, the agent asks a person only when it has to.

-
- 1 **To approve an action it was not allowed to take alone.** A payment or an email waits for you, unless you allowed it in advance.
 - 2 **To answer what only a person knows.** A fact that is not in the documents, such as someone's filing status, is asked, never guessed.
 - 3 **To spend more than its budget.** It stays within what you set, and asks before it goes further.
-

Everything else, from a model provider's outage to a restart or a failed check, Zero-AR handles without asking. You can still pause the work or redirect it at any time.

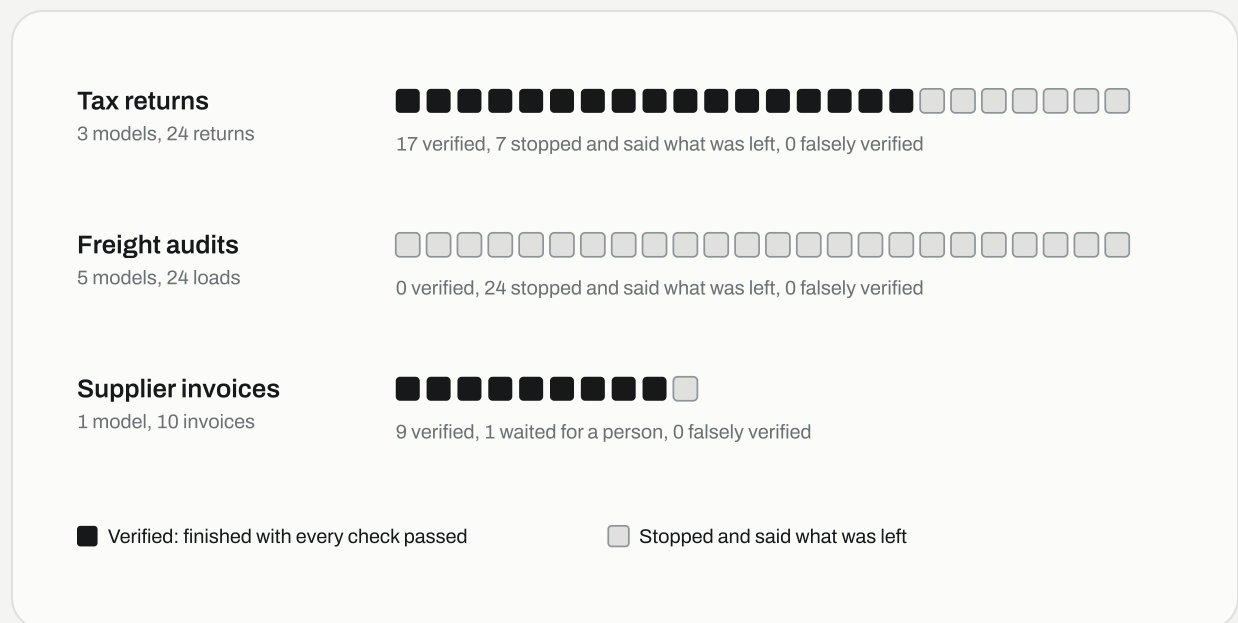
Each study tests a different part of the aim.

Study	The work	Result	What it shows
Tax returns	Prepare US federal and Ontario returns from fictional documents	Every return verified with each Claude model, in both places	The checks decide when work is done, and the agent asks only what a person must answer
Freight audits	Check carrier invoices against PDFs, scans, a photo and a carrier website	No load verified in 24 runs, and none reported verified	When no model can finish, nothing is called done
Supplier invoices	Settle invoices by email, sent and received by a separate mail app	9 of 10 verified with no one stepping in beyond the owner's approvals	Everyday work finished on its own, by our measure

Seven AI models took part: Claude Sonnet 4.6 and Haiku 4.5, GPT-4o mini, GPT-4.1 mini and GPT-5.4 mini, and two models that run on a laptop, qwen3:8b and gpt-oss:20b. Each study found problems in Zero-AR, and we fixed them in Zero-AR itself rather than working around them in the agent.

Fifty-eight runs, and not one reported done when it was not.

Each square is one run: one agent, one model, one case. A filled square finished with every check passed.



Tax returns: the first campaign in the full study, with Claude Sonnet 4.6, Claude Haiku 4.5 and GPT-4o mini. Freight audits: the final campaigns, with three OpenAI models and two models running on a laptop. Supplier invoices: Claude Sonnet 4.6.

A tax agent finished every return with Claude, and asked when it could not know.

8 of 8

returns verified with Claude Sonnet 4.6, and again with Claude Haiku 4.5, across US federal and Ontario cases

\$0.07

per verified return with Claude Haiku 4.5, at list price; \$0.18 to \$0.20 with Sonnet 4.6

When a taxpayer's note left out their filing status, the agent did not guess. It asked, waited for the answer, then finished with all 19 checked lines of the return correct.

The agent never does the arithmetic itself: a calculator tool does, and five checks decide whether the return is done.

No wrong return was reported verified. Smaller OpenAI models finished fewer returns, and the checks caught every mistake they made.

Moving to a second country needed new tax rules and checks, and no change to Zero-AR.

The taxpayer's answers were scripted, and by our measure any answer counts as someone stepping in. This study shows the asking side of working alone. Early Zero-AR builds; one run per case and model.

No model finished the freight audit, and Zero-AR never said one had.

0 of 24

runs verified, across three OpenAI
models and two models running
on a laptop

0

runs reported verified when they
were not

The documents were hard on purpose: PDFs, scanned pages with handwritten times, a photo of a receipt, French paperwork and a carrier's website.

The checks turned down wrong figures. Nine payment requests waited for approval, and each was refused because its amount or bank account did not match the audit; no payment went out.

Fifteen questions reached a person, and each answer set its work going again. No one had to restart anything by hand.

Building the agent found eight bugs and five missing features in Zero-AR, and we fixed all of them in Zero-AR.

An agent that cannot finish is still useful when it says so honestly. Early Zero-AR builds.

Nine of ten invoices settled by email, with no one stepping in.

9 of 10

verified with no one stepping in, apart from the approvals the owner kept. The agent was allowed one follow-up email per invoice, to the same supplier, within seven days, without asking again.

The agent asked each supplier for the invoice, read the PDF that came back, questioned the line that did not match the order once the owner approved, and answered the supplier's reply on its own.

A separate mail app sent and received every email. Zero-AR never signed in to a mailbox; it only decided when each email could go. All thirty went out with the owner's approval or under the follow-up the owner allowed.

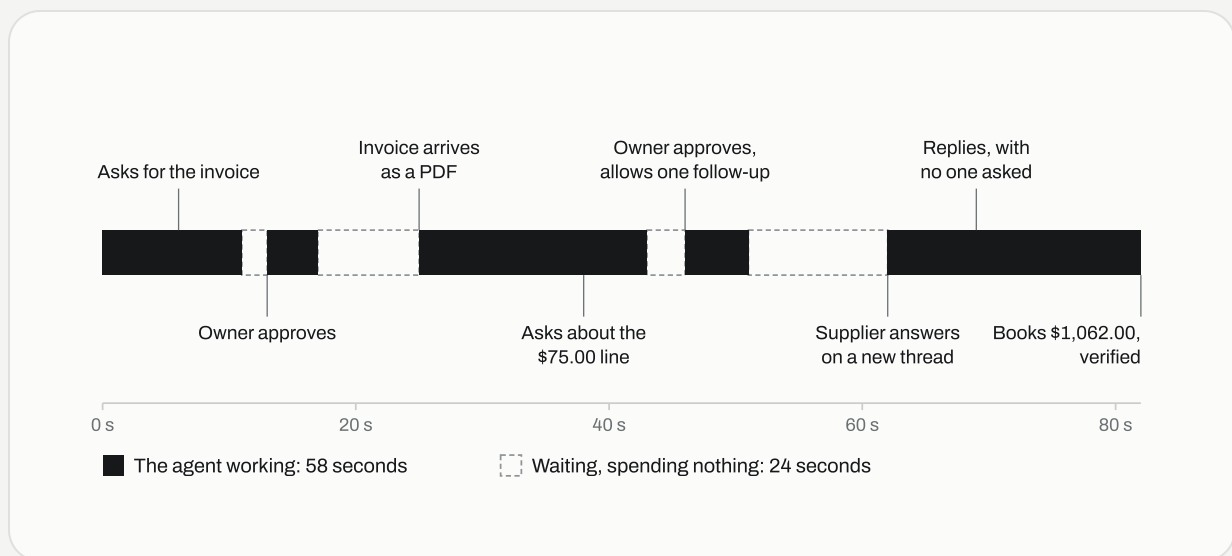
About 76 seconds and 24 cents per invoice, with Claude Sonnet 4.6. While the work waited for a reply, it cost nothing.

The tenth run reached the right amount, but in a way its checks could not confirm, so it waited for a person instead of calling itself done.

Scripted replies and approvals let all ten runs meet the same situation.

One invoice, start to finish, in 82 seconds.

Harbor Office Goods added a \$75.00 assembly charge the order did not list. This is where the time went, from the run's own record.



The agent worked for 58 seconds. For the other 24, it waited for the owner or the supplier, and no model ran.

With people on the other side, each wait takes hours or days. The work picks up the moment the reply or decision arrives, and costs nothing until then.

In every run of all three studies, nothing was reported done that was not.

-
- 1 **No false finish.** Every run that did not finish said so, and said what was left.
 - 2 **Nothing sent without permission.** Every payment request and every email waited for the owner, or went under a follow-up the owner had allowed.
 - 3 **No overspending.** Before each call to an AI model, Zero-AR set aside its budget. In the freight and supplier studies, none of 481 calls went over.
 - 4 **Every reply reached its work.** Each answer, decision and reply found the work it was for, and woke it if it was waiting.
 - 5 **Problems fixed at the source.** Each study found bugs in Zero-AR. All but one are fixed in Zero-AR itself, and the last is in review.
-

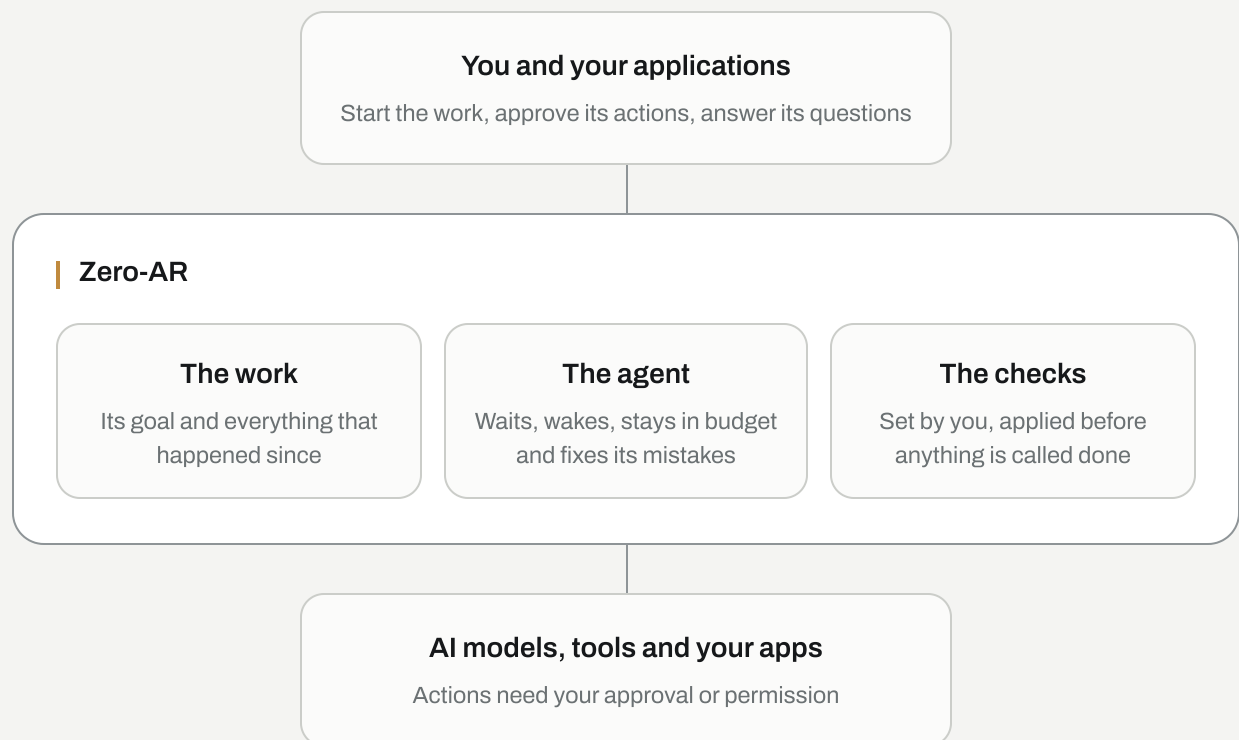
How to read these results.

-
- 1 Real AI models did the work, reading real PDFs and scans and, in the supplier study, sending real email. The cases were written for testing and the replies on the other side were scripted, so every run met the same situation and the results compare fairly. The next step is the same work with real people on the other side, which is what we look for in design partners.
 - 2 Each study ran once per case and model. The numbers are observations, not rates, and they do not rank the models.
 - 3 Verified means the checks passed, and checks see only what they are given. In the tax study, two verified returns quoted lines that were not in their documents, though both returns were right.
 - 4 The runs used early versions of Zero-AR, on the way to version 1. They are not a release.
 - 5 Zero-AR is not a tax, freight or accounts payable product. The agents are examples of what a team can build on it.
-

Every figure comes from the runs' own records, which we keep for each study.

Deploy and run agents that execute day-to-day tasks autonomously, get better the longer you use them, and verify completion.

Zero-AR, the Zero-layer Autonomous Runtime, keeps long-running work in one place that waits, wakes and checks itself. Version 1 is code-named Dante-1.



Read the full studies. Each case study is a fourteen to twenty page report with the runs' own records. Ask for any of them, or bring us your day-to-day work as a design partner, at hello@finesample.com.

| **ZERO-AR** The runtime beneath long-horizon work.