

A tax agent on Zero-AR

One agent, built for two tax jurisdictions and run against real models from Anthropic and OpenAI.

AI work that survives the handoff.

Case study, October 2026

Zero-AR development builds after release 0.3.0

The taxpayers are fictional. Zero-AR does not ship a tax product.

What happened when we ran one tax agent on Zero-AR.

Zero-AR is a log-native runtime for long-horizon agent work. We built a tax-preparation agent from ordinary Zero-AR project files. Then we built it again for a second country.

In one paragraph

Claude Sonnet 4.6 and Claude Haiku 4.5 verified every return, 16 of 16. GPT-4o mini verified 1 of 8 at first, and part of that was Zero-AR's doing. After a fix, it verified 5 of 8. In 112 runs, Zero-AR reported no wrong draft as verified.

1	The work	3
2	The agent	4
3	One return, start to finish	5
4	Results	7
5	The OpenAI fix	9
6	What held in every run	11
7	What building it changed	12
	What this does not establish	13
	About Zero-AR	14

A tax return is long-running work.

It needs facts no document states, a reviewer's decision on anything unusual, and a record of where every number came from.

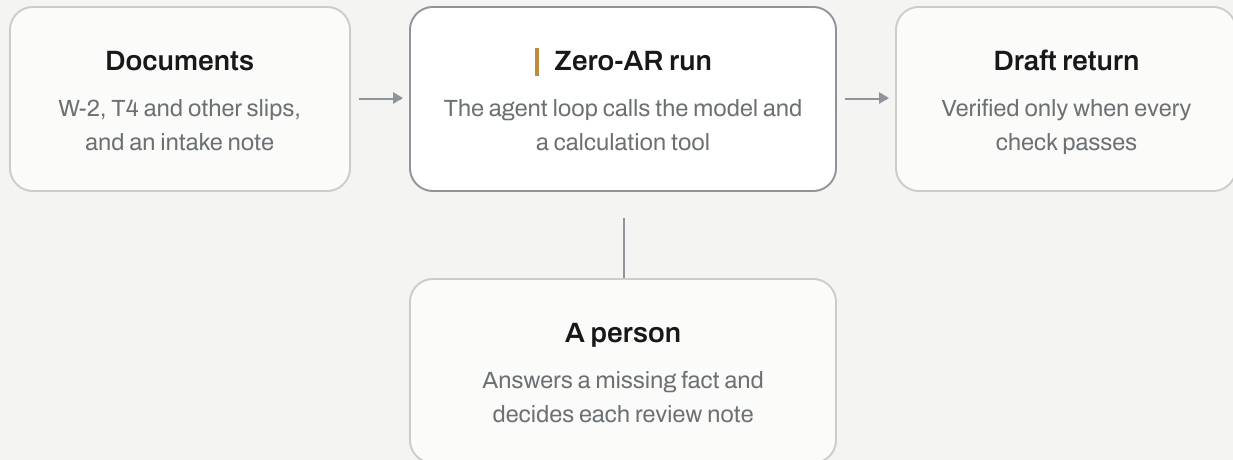
The familiar failure is an agent that guesses the missing fact, loses its place while it waits, or calls a draft done before anything checked it.

Zero-AR keeps the run's history, position, limits and checks together. When a return needs a person, the run stops, keeps its place and continues from its own record.

What Zero-AR does not decide

Zero-AR does not decide what a correct return is. The tax checks attached to the agent decide it. Zero-AR enforces those checks and does not report unchecked work as verified.

An ordinary Zero-AR project, built twice.



- 1 The agent reads each document and records each fact.
- 2 A calculation tool runs the tax engine. The agent never computes tax itself.
- 3 Five tax checks test every item before it counts.
- 4 When a check cannot decide, the run asks a person.

Agent	Prepares	Lines checked
United States	Form 1040 and Schedule B	19
Ontario	T1 return and Form ON428	33

No document stated the filing status. The run asked, waited and continued.

The hardest US return belongs to a parent whose child tax credit is larger than the tax.

-
- 1 The agent recorded the filing status as unknown, with a question, instead of guessing.
 - 2 Zero-AR parked that item, and the run stopped before any form was calculated.
 - 3 The answer, head of household, reached the agent. It calculated both forms.
 - 4 The engine noted a credit outside its scope. The run stopped again for a reviewer.
 - 5 The reviewer confirmed. Every check passed, and all 19 lines matched the engine.
-

The Ontario return with no stated province of residence has the same shape and finished the same way.

The run's own record shows each pause.

The run's own record, cropped

```
run_01a1137d99b4724fa7fdf600f9d10f7a / Claude Haiku 4.5 / 2026-10-06 UTC
23:12:40 run.started
23:12:48 checkpoint.passed doc-intake, doc-w2-alpine
23:12:51 checkpoint.indeterminate fact-filing-status a person must answer
23:12:51 item.parked fact-filing-status
23:12:51 run.suspended awaiting_answer
23:12:51 gap.settled fact-filing-status
23:12:51 run.resumed
23:12:54 checkpoint.passed fact-filing-status, fact-qualifying-children
23:12:59 checkpoint.indeterminate form-1040 credit exceeds the tax by 431
23:12:59 item.parked form-1040
23:12:59 run.suspended awaiting_answer
23:12:59 gap.settled form-1040
23:12:59 run.resumed
23:13:04 checkpoint.passed form-1040, schedule-b
23:13:08 completion.proposed
23:13:08 verification.concluded verified
23:13:08 run.finished complete
```

The person in this eval was the test harness, which answered within a second. In use, the run waits as long as a person takes.

Both Claude models verified every return.

16 of 16

returns verified by Claude Sonnet 4.6 and Claude Haiku 4.5,
across both jurisdictions

Every verified return matched the reference engine on every line.

GPT-4o mini verified 1 of 8 in the same campaigns.

Source: Zero-AR tax agent campaigns, October 2026. Four returns per jurisdiction, one run per model.

Returns verified, by model and jurisdiction.

Model	United States	Ontario	Cost of all eight
Claude Sonnet 4.6	● 4 of 4	● 4 of 4	\$1.52
Claude Haiku 4.5	● 4 of 4	● 4 of 4	\$0.57
GPT-4o mini	0 of 4	1 of 4	\$0.20

A return counts as verified only when every check passed. A run that did not verify ended as unverified or suspended, and named what remained. None was reported as verified.

What this does not show

One run per return for each model. These are observations, not rates, and they do not rank the models. Costs are list-price estimates from each campaign's recorded tokens, rounded once.

Part of GPT-4o mini's result was Zero-AR's.

Each return has items of three shapes: documents, facts and forms.

Before

One operation

Any item, in any of the three shapes

After

| Documents

Document items, in the document shape

| Facts

Fact items, in the fact shape

| Forms

Each form, in its own shape

OpenAI's strict mode makes a call match one of the shapes, not the right one. In one run, GPT-4.1 mini wrote that it would send a document's fields. The call that followed carried a form's fields.

Zero-AR now offers one operation for each shape, with fields in the order their author wrote them. No check changed.

After the fix, no item arrived in the wrong shape.

Model	Zero-AR	US verified	Ontario verified	Wrong shape
GPT-4o mini	Before	0 of 4	1 of 4	37
GPT-4o mini	After	2 of 4	3 of 4	0
GPT-4.1 mini	Before	1 of 4	2 of 4	74
GPT-4.1 mini	After	3 of 4	3 of 4	0

Each remaining failure was the model's tax work, and a check caught it. One GPT-4.1 mini run stopped on a separate Zero-AR defect, which is being fixed.

What this does not show

This change came after the 0.4.0 source candidate and is not in it. The Claude models were not rerun after it. GPT-4.1 mini ran as a control and is not part of the main campaigns. Two verified runs here quoted lines that are not in their documents, though both drafts matched the engine.

In 112 runs, no wrong draft was reported as verified.

0 of 112

runs reported a wrong draft as verified, including every run where the model failed

0 of 833

model calls spent more than they reserved, in the eight campaigns with retained records

- 1 **A person decided what only a person could.** Questions reached a person while the run waited.
- 2 **Every run kept one record.** Each run's full history exports as one bundle.

Building a real agent found defects the tests had missed.

The first two runs verified no return. Each defect was fixed in Zero-AR, not worked around in the agent. No check was relaxed.

Run	Changed before the run	Sonnet 4.6	Haiku 4.5	GPT-4o mini
1	First run on a hosted cell	0 of 4	not run	0 of 4
3	Submitted items visible to the agent	3 of 4	not run	0 of 4
5	Each model's own output limit	4 of 4	not run	0 of 4
7	Rejection reasons reach the agent	4 of 4	4 of 4	0 of 4
10	OpenAI-shaped work protocol	4 of 4	4 of 4	0 of 4
Later	One operation for each shape	not run	not run	2 of 4

US returns verified in selected runs. The engineering report records every run.

The limits of this evidence.

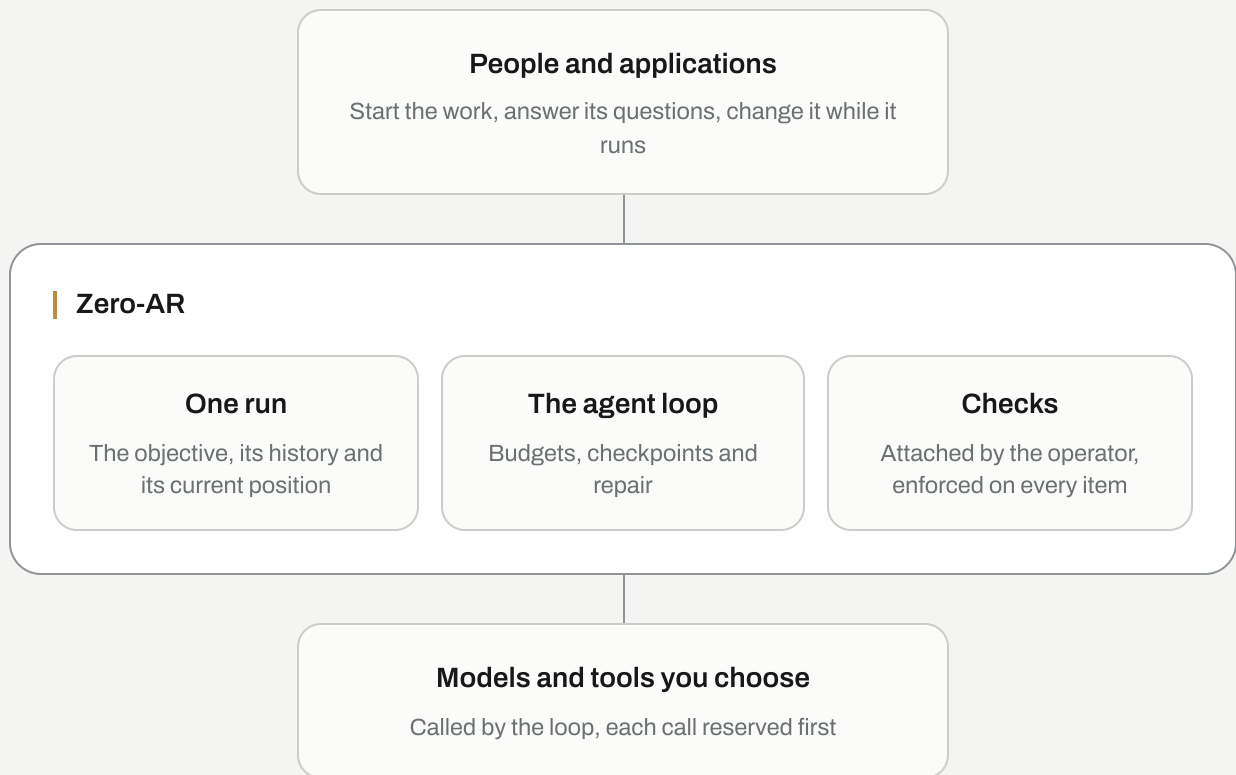
- 1 The tax engines are illustrative and the taxpayers are fictional. Neither engine is tax software.
- 2 Each run lasted minutes, including its pauses for a person. It shows work continuing across an answer, not a run lasting hours.
- 3 The eval is small: four returns per jurisdiction, one run per model and campaign.
- 4 Verified means the named checks passed. The traceability check sees what the agent submitted, not the documents, so a quote can be verified without appearing in its document.
- 5 The runs used development builds after Zero-AR 0.3.0. Zero-AR 0.4.0 is a source candidate and is not released yet.

Every figure comes from the runs' own records. For eight campaigns, each run's report and exported bundle are retained with SHA-256 digests, run ids and chain heads.

Zero-AR is the runtime under an application. This case study describes one agent, for evaluation. Zero-AR does not ship a tax product.

Zero-AR is the runtime for AI work that has to survive the handoff.

Long-running work crosses agent sessions, model changes, reviews and failures. Zero-AR keeps it in one run, so nobody starts over.



Operators attach the checks that say what done means. Zero-AR enforces them, and names anything unresolved instead of rounding it into a finished result.

Read the guide to build the same agent: Building a tax agent on Zero-AR, in the Zero-AR documentation.

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