OpenAccountants

RESEARCH REPORT · OA-RR-2026-02 · 13 AUGUST 2026

How accurate is AI on tax?

In April 2026 we asked AI to write the tax law of the world — six topics, every country. Then we had accountants mark it, and turned their marking into a benchmark that any model can sit. This is what came back, including the parts that do not flatter us.

Guides drafted 1,928 Facts produced 66,253 Facts judged by accountants 2,713

Benchmark questions 60 Model runs scored 120

Summary

The experiment, and the five findings

> We gave AI agents one instruction — write the tax law of every country on earth, six topics each, every claim carrying its source — and let them run for three months. Out came **1,928 guides across 186 countries, containing 66,253 individually checkable facts**. Fourteen accountants then marked 2,713 of those facts in five languages. Section 2 sets out exactly how. Everything below rests on that, plus a benchmark we built from the marked answers so the same questions can be put to any model.

1. **A frontier model with no internet answered 97% of tax questions put to it, and got 77% right.** It almost never refuses. On questions where we handed it the country, the exact item, the governing statute and the tax year, it declined twice out of sixty.

2. **Giving it web search raised that to 88%, fixing seven answers and breaking none.** The improvement is real and statistically significant. Retrieval is not optional decoration; it is the difference between one error in four and one in eight.

3. **But it reached an actual tax authority for only 1 question in 4.** Government sites blocked it — canada.ca returned 403, Peru's gov.pe returned 418, Saudi Arabia's ZATCA 404'd. Forty-three percent of all its answers trace to a single website: PwC's Worldwide Tax Summaries.

4 **Retrieval fixes facts that went stale. It does not fix rules with conditions.** Every single error that survived web search was the same shape: the model gave the headline rate and dropped the taxpayer class that changes it — filer versus non-filer, resident versus non-resident, commercial versus private.

5 **Searching made the model more confident faster than it made it more accurate.** High-confidence answers tripled from 14 to 45, while the error rate among high-confidence answers did not improve. Confidently wrong answers went from one to four.

What this is for

Accountants are being asked whether they can rely on an AI's tax answer. Until now the honest reply was a shrug. This report replaces the shrug with a number, a method anyone can repeat, and a clear statement of what the number does not cover.

Section 1

What we measured, and what we cannot

> Two separate things are measured here and they must never be quoted as one. The first is **how good our own drafts were**, judged by the accountants who reviewed them. The second is **how good AI is at tax generally**, measured by sitting a model down in front of a benchmark. The first is a floor. The second is a ceiling. Neither is "the accuracy of AI on tax", and anyone offering you that single number is selling something.

Why our own number is a floor. It counts how often an accountant *flagged* one of our drafted facts. That is not how often the draft was wrong. An error nobody noticed looks exactly like a fact that was right. The true error rate is that number or higher, never lower.

That distinction is not theoretical, and the fairness point cuts toward the reviewers. They were told, in writing: *"Do I need to read every section? No. Review the sections you are confident about. Flag or skip what you are unsure of."* and *"You're not the last line of defence."* A reviewer who confirmed everything was following the instructions we gave them. Where this report finds thin review, that is a finding about our instrument, not about anyone's diligence.

Why the benchmark number is a ceiling. We made the questions easy on purpose, and we drew them from a pool that is biased in the model's favour. Section 15 sets both problems out in full. The short version: treat 77% and 88% as the best case, not the expected case.

Section 2

What we actually did

> In April 2026 we asked AI to write the tax law of the world. Six topics for every country on earth. Then we asked qualified accountants to mark it. That is the experiment this report is built on, and as far as we can tell nobody has run it at this scale before, so the method matters as much as the result.

Step 1 — the instruction. One web-researching agent per country, each asked to produce the same six guides:

TOPIC	WHAT IT HAS TO COVER
Tax Overview	the shape of the system, who administers it
Personal Income Tax	bands, rates, allowances, filing
Corporate Income Tax	rates, small-company regimes, reliefs
VAT / GST	rates, registration thresholds, exemptions
Payroll & Social Contributions	employer and employee rates, ceilings, deadlines
Company Formation & Entity Choice	entity types, capital, registration

Step 2 — the required output shape. The agents could not write prose and stop. Every claim had to come back as a structured fact with its own fields — the item, the value, the unit, any qualifier, what kind of thing it is (`rate`, `threshold`, `rule`, `formula`, `definition`), the legal reference, whether that reference had actually been researched, a source URL, and an effective date.

That constraint is the reason this report is possible. A guide written as an essay cannot be marked fact by fact; 66,253 individually addressable facts can. It is also what let us later discover that **100% of drafted facts carried a legal reference and 72.7% carried a live source URL at the moment they were written** — the agents really were reading sources, not reciting.

Step 3 — what came out.

1,928
guides

186
countries and territories

66,253
individual facts

3
months, April to June

MONTH	GUIDES WRITTEN	JURISDICTIONS TOUCHED
April 2026	378	180
May 2026	552	104
June 2026	945	173
July–August 2026	53	28

April went wide — 180 jurisdictions in one pass, a thin layer over most of the world. May and June went deep, adding topics to countries that already had one.

Step 4 — the marking. We sent accountants a workbook: one row per fact, with the drafted value, a verdict column, a corrected-value column and a notes column. Twenty-five workbooks came back, in five different layouts and five languages — English, Portuguese, Spanish, Georgian and Indonesian — because the reviewers were working accountants in their own countries, not annotators we trained. Each verdict was loaded back against the exact fact it judged, with the workbook's SHA-256 recorded so any row can be traced to the file it came from.

What the reviewers were told matters for reading every number that follows: *"Review the sections you are confident about. Flag or skip what you are unsure of."* They were not asked to audit exhaustively, and they were told explicitly they were not the last line of defence.

> **Why this is worth publishing rather than keeping**
>
> The result is a corpus of tax facts where a named, credentialled professional has personally ruled on each one — 2,713 of them so far, across 14 countries and 5 languages. Public AI benchmarks in tax are single-country and synthetic. This is multi-country and adjudicated by practitioners in the jurisdictions concerned. It is the only material we know of that can answer "is the AI right about tax" with a human signature behind each answer.

Section 3

The benchmark

As of August 2026 we could find no public benchmark measuring whether AI-stated tax rules and thresholds are correct across multiple jurisdictions. Every tax benchmark we located is single-jurisdiction: TaxCalcBench and SARA are US, TaxPraBen is China, FiscalQA Pro is France, SteuerEx is Germany, RO-FIN-LLM is Romania. The two cross-jurisdictional legal benchmarks we found, Multi-Legal-Bench and CrossLex, exclude tax entirely. One multi-jurisdiction tax benchmark has been announced — TaxOS's TaxBench v1, due Q1 2026 — and has not been released.

That is a claim about a search rather than about the world, and we would rather be shown a counter-example than assert a first. But it is why we built our own, from material we have and others do not: tax facts a named, credentialled accountant has personally adjudicated.

OA-TAXFACTS-60. Sixty questions drawn from facts reviewed by accountants between May and July 2026, capped at six per country so no single reviewer's workbook could define the result. Fifteen jurisdictions: India, Nigeria, Venezuela, Cameroon, Pakistan, Cyprus, Tanzania, Peru, Nepal, Portugal, South Africa, Indonesia, Canada, Saudi Arabia and Brazil.

Each question gives the model the jurisdiction, the guide and section the fact sits in, the item being asked about, the governing statute, and the period the answer must be in force for. That is far more help than a real user gives. It is deliberate: we wanted to know what the model can do at its best, because a failure under generous conditions is a much stronger finding than a failure under hard ones.

The controls, because a benchmark you cannot trust is worse than none.

- **The answer key was quarantined.** Questions and answers were written to separate directories. The model runs read only the questions.
- **The searching arm was barred from us.** It could not use openaccountants.com or any of our tools. Our site publishes the very answers being scored; letting it read them would have graded our data against itself.
- **Grading was blind and doubled.** Two independent graders marked the same sheet with the arm labels stripped and the rows shuffled, so neither could tell which answer came from the searching run. They agreed on 52 of 54 judgement calls (96%). The two disagreements were both "partially right" versus "right", and we took the stricter mark in each.
- **Exact matches never reached a grader.** 66 of 120 answers were settled mechanically. Nobody exercises judgement over whether "15%" equals "15.0%".
- **Wrong year counts as wrong.** Being right about 2024 when asked about 2026 is the failure this study exists to measure.

Section 4

The scoreboard

No internet

76.7%

46 of 60 · 95% CI 65–86%

Answering from what the model already knows. Declined 2 questions.

With web search

88.3%

53 of 60 · 95% CI 78–94%

Same model, same questions, allowed to go and look. Declined 1.

Because both arms answered the identical questions, we can compare them question by question rather than only in aggregate — a much stronger test than two separate scores.

46	7	0	7
both arms right	search fixed	search broke	neither got right

Seven fixed and none broken. On a paired test that is significant (McNemar exact, p = 0.016), which in plain terms means it is unlikely to be luck. **Search does not trade one error for another. It removes a specific class of error and adds nothing back.**

The confidence intervals overlap, and at sixty questions they were always going to. The paired comparison is what carries the finding; the two headline percentages are the context around it.

Section 5

What search fixes: facts that went stale

All seven repairs are the same story. The model knew a figure that used to be right.

> **South Africa — voluntary VAT registration threshold**
>
> From memory: **R50,000**
> Correct: **R120,000** from 1 April 2026

> **Peru — the UIT, the indexing unit half the tax code is denominated in**
>
> From memory: **PEN 5,600**
> Correct: **PEN 5,500** — and searching found the decree that set it, Decreto Supremo N° 301-2025-EF of 17 December 2025

This is the failure mode nobody can engineer away. A model's knowledge is fixed at training; tax law is not. Every budget, every finance act, every indexation order moves figures the model still believes. The Peru case is instructive twice over: the UIT is a multiplier, so one stale number silently corrupts every threshold, penalty and allowance derived from it.

It is also worth naming what this means for anyone comparing AI vendors. **A model's tax accuracy decays continuously after its training cutoff, and the decay is invisible from the outside.** A benchmark score published in January describes a different product by December.

Section 6

What search does not fix: rules with conditions

This is the most important finding in the report, and the one that most directly answers what an accountant is actually for.

Seven questions defeated both arms. Searching the entire public internet fixed none of them. Here is every one:

JURISDICTION	ITEM	WHAT BOTH ARMS SAID	WHAT IS ACTUALLY TRUE
Pakistan	§153(1)(a) withholding, goods, to a company	5%	Filer 5% / non-filer 10%
Pakistan	§151 profit on debt	15–20%	20% bank profit / 15% others, doubled for non-filers
India	Motor vehicles	15%	15%, but 30% commercial
Venezuela	Non-resident professional activities	34%	34% on 90% of gross income
Tanzania	Rent, other assets	10%	0% resident / 10% non-resident
India	The valid GST rate set	9 or 11 rates	13 rates, including 0.1, 0.25, 1.5, 6 and 14
Pakistan	Securities held 1–2 years	12.5%	15%

Six of the seven are one error wearing different clothes: **the model states the headline rate and drops the class of taxpayer that changes it.** Filer or non-filer. Resident or non-resident. Commercial or private. Bank or other. The answer is not fabricated, it is not stale, and it will survive a citation check — the statute really does say 5%, for one kind of taxpayer.

> **Why this is the dangerous class**
>
> A stale figure is detectable: check the source and the mismatch is obvious. A dropped condition is not. It reads as correct, cites correctly, and is wrong only for the client who happens to fall on the other side of the line. It is exactly the error a non-specialist cannot catch, and exactly the one a practitioner catches on sight — because the practitioner is not recalling a rate, they are asking who the taxpayer is.

The seventh case is different and worth its own note. Pakistan's securities rate is recorded in our own data as the bare string `0.15`, with no unit. Both arms read the question and answered 12.5%. Our stored value means 15%, but a machine reading `0.15` has every reason to render it as 0.15%. That is our bug, not the model's: 159 facts across the corpus are bare numbers with no unit attached, and 3 are Excel date serials that were never converted. Small in number, but each one is a fact that will be read wrongly by the very audience we built the data for.

Section 7

The confidence trap

We asked both arms to say how sure they were. The closed-book results are, frankly, better than we expected:

STATED CONFIDENCE	NO INTERNET		WITH WEB SEARCH	
	ANSWERS	ERROR RATE	ANSWERS	ERROR RATE
High	14	7%	45	9%
Medium	34	18%	13	8%
Low	10	50%	1	100%
Declined	2	—	1	—

Closed-book, the model is well calibrated: when it says it is sure it is wrong 7% of the time, when it hedges it is wrong 18%, and when it says it is unsure it is wrong half the time. That is a genuinely useful signal, and it deserves saying plainly because the popular account of language models says the opposite.

Then search changes the picture in a way nobody should ignore. High-confidence answers went from 14 to 45. The error rate among them did not improve — it went from 7% to 9%. Having found *a* source, the model became sure; but the sources it found were often a professional summary that carries the same simplification the model already had. The count of confidently-wrong answers went from one to four.

So retrieval improves the average answer while degrading the usefulness of the model's own hedging. The user gets a better answer and a worse warning signal at the same time. Anyone building "AI checks its sources, so it must be right" into a workflow should sit with that.

One more number for the reliance question: closed-book, the model volunteered that a figure might have changed since training on 40% of answers, and flagged 48% as worth verifying. That is honest behaviour. It still answered 97% of the questions.

Section 8

The source problem: the authorities are closed

We recorded what the searching arm actually cited, and this is the finding with the widest implications beyond us.

WHAT BACKED THE ANSWER	QUESTIONS	SHARE
Professional summary (PwC, Deloitte, and similar)	39	65%
Official tax authority	10	17%
Statute text	5	8%
Other secondary sources	5	8%
Nothing found	1	2%

A primary source was reached for 15 of 60 questions — one in four. And the substitution is concentrated to a degree we did not anticipate: **26 of the 60 answers, 43% of the whole benchmark, cite the same website** — PwC's Worldwide Tax Summaries.

It was not for want of trying. The run recorded being turned away, by name, from:

AUTHORITY	WHAT HAPPENED
canada.ca (CRA)	HTTP 403 — refused, on two separate pages
gob.pe (Peru)	HTTP 418
zatca.gov.sa (Saudi Arabia)	HTTP 404 on the VAT Law PDF and the announcement
pwc.pt · seg-social.pt (Portugal)	HTTP 403 / unreadable
epfindia.gov.in · einvoice1.gst.gov.in (India)	blocked or 404
seniat.gob.ve (Venezuela) · tra.go.tz (Tanzania)	blocked

> **What this means for everyone building AI tax tools, not only us**
>
> When an AI answers a tax question about most of the world, it is very probably not reading the law or the tax authority. It is reading a Big Four summary of them. Those summaries are good, written by professionals, and deliberately simplified — which is precisely how the dropped-condition errors in Section 5 get in and stay in. The industry has quietly converged on a single secondary source, and a rule that PwC summarises in one line is a rule every AI will state in one line.

The practical consequence for a reader deciding what to trust: **"the AI cited a source" and "the AI read the law" are not the same claim**, and today they differ three times out of four.

Section 9

Where this sits against everything else published

Our two central findings were arrived at independently, on our own data. Both turn out to be well supported by peer-reviewed work we found only afterwards, which is the outcome we wanted — a result nobody else has seen is usually a result that is wrong.

Finding: search fixes staleness and little else. FreshQA (Findings of ACL 2024) tested GPT-4 on the same day with and without search. On *fast-changing* facts it went from 26.0% to 67.7%, a gain of nearly 42 points. On *never-changing* facts it went from 92.7% to 96.0% — a gain of 3. Retrieval is transformative exactly where facts move and close to irrelevant where they do not. Our tax result is the same shape at a smaller magnitude.

The same paper records something that should end the "just make the model bigger" argument for tax: *"there are flat scaling curves on questions that involve fast-changing knowledge: simply increasing the model size does not lead to reliable performance gains."*

Finding: search does not stop confident error. Magesh et al. (*Journal of Empirical Legal Studies*, 2025) preregistered 202 queries against the commercial legal-research tools that are the closest existing analogue to what we are building:

SYSTEM	ACCURATE	INCOMPLETE	HALLUCINATED
Lexis+ AI (retrieval-backed)	65%	18%	17%
Westlaw AI-Assisted Research (retrieval-backed)	42%	25%	33%
Ask Practical Law AI (retrieval-backed)	20%	63%	17%
GPT-4, no retrieval	49%	8%	43%

Their conclusion, verbatim: *"RAG systems are no panacea."* Note the uncomfortable detail — two of the three retrieval-backed commercial products were *less* accurate overall than plain GPT-4, partly by refusing more often. Anyone quoting this study as "retrieval fixes hallucination" has not read the table.

The closest tax comparator, and why our number is not on its scale. TaxCalcBench (Column Tax) has models compute complete US tax returns, graded deterministically. Its live leaderboard shows the same retrieval effect we measured, in the same direction, on the same vendor's models:

MODEL	WHOLE RETURNS CORRECT	INDIVIDUAL LINES CORRECT
GPT-5.6 Sol with web search	58%	85.6%
Claude Opus 5 with web search	40%	81.9%
Claude Opus 5, no search	18%	70.7%
Claude Sonnet 5	6%	60.7%

> **Do not put our 88% next to their 40%**
>
> Those two numbers measure different things and the comparison would flatter us by about forty points. A TaxCalcBench "correct return" requires every evaluated line of a Form 1040 to be right at once — one wrong line fails the whole return. Ours is per-fact. The honest comparator for our number is their *per-line* column, 60–86%, and on that scale our results are unremarkable rather than exceptional. We are stating this because it is exactly the misreading this report would otherwise invite.

Two further reference points for what closed-book models do on legal and factual recall. Dahl et al. (*Journal of Legal Analysis*, 2024, over 800,000 queries) found legal hallucination rates *"between 58% of the time with ChatGPT 4 and 88% with Llama 2"*. And on LegalBench's

`sara_numeric` task — given the statutes, state the tax owed — GPT-4 scored **8.3%**, while on `sara_entailment`, deciding whether a rule applies at all, it scored **86.8%**. That 78-point gap between reasoning about a rule and producing the right figure is the single most striking number in the literature, and it points the same way our Section 6 does.

One paper worth quoting to any accountant weighing this up. Cheng et al. (COLM 2024, Outstanding Paper) open with tax as their motivating case: *"imagine a layperson using an LLM for tax advice, without realizing that the effective cutoff of the tax code is 2022 and thus outdated — despite the fact that the reported cutoff is advertised as 2023."* A companion paper found LLaMA-2's knowledge peaks around 2019 against a stated 2022 cutoff. A model's advertised knowledge cutoff is not the date its tax knowledge actually stops.

And a warning about maintained data, which cuts against a lazy reading of our own conclusion. The HoH benchmark (ACL 2025) found that *"even when current information is successfully retrieved, the mere presence of outdated information in the context leads to at least 20% performance drop"* — outdated context is worse than no context. That is an argument for a corpus that is dated and maintained, not for one that is merely large. It is also a direct warning to us: our 66,253 facts include many nobody has looked at since they were written.

> **On vendor accuracy claims, including the ones we will be tempted to make**
>
> We checked the published figures from tax-AI vendors. Thomson Reuters' 90–99% pass rates are self-graded by their own attorneys on a private dataset, and are for legal rather than tax skills. Blue J's headline is a thumbs-down rate, not an error rate, and its "reduction in error rate" figures never disclose the baseline. TaxGPT's "99%" is *time saved*, not accuracy — its own FAQ on research accuracy gives no number. Intuit's "100% accurate calculation guarantee" is a reimbursement promise. Not one of these is a measurement of whether the answer was right. We are naming this because the pressure to publish a flattering number is the same pressure we are under, and the only defence is stating the denominator.

Section 10

How our own drafts scored, judged by accountants

Separately from the benchmark, fourteen accountants reviewed 2,713 of our AI-drafted facts and recorded a verdict on each.

74.3%	25.7%	467	231
clean first time — the reviewer changed nothing (2,015 of 2,713)	needed a change of some kind (698 of 2,713)	had the wrong figure	figure right, source or context wrong

About three in four facts were right first time; one in four needed an accountant to change something. Two thirds of the changes were to the figure itself. A second and more generous reading — counting a citation-only fix as "the number was right" — gives **82.8%**. Both are true, they answer different questions, and neither should be quoted without saying which.

The average hides the actual finding. Value accuracy by country ranged from 99.1% to 50.5%:

COUNTRY	REVIEWER	FACTS	VALUE ACCURACY
Cameroon	Nkinyam Courage Ndasi	113	99.1%
Venezuela	Jose Padilla	287	96.2%
Cyprus	Christos Thoma	336	92.9%
Canada	Edgar Lautsyus	227	89.4%
Peru	Maria Clemencia Valverde Rios	144	86.0%
India	Mayur Deokar	227	85.8%
Portugal	Mário Vale	97	82.5%
Illinois (US)	Amir Pelinkovic	125	76.0%
Indonesia	Rilia Putri	198	68.0%
Brazil	Ariane Marrocos	125	56.6%
Pakistan	Ibrar Ali	170	53.5%
Nepal	Ashish Bista	65	50.8%
South Africa	Werner Britz	487	50.6%
United States	Christopher Aryee · Amir Pelinkovic	95	50.5%

The spread is not reviewer strictness. It tracks how much of a country's tax law is published in a form a machine can read, which is the same finding Section 8 arrives at from the other direction.

Section 11

What the AI actually gets wrong

Pooling the reviewers' 698 corrections with the benchmark failures gives a consistent taxonomy. In descending order of how often it happens and how hard it is to catch:

1. **The dropped condition.** The right rule, minus the exception or class that changes it. *Peru:* "no reduced rate" — there is one, for small hospitality businesses. *Pakistan:* one surcharge rate where salaried and non-salaried differ. The commonest and the most dangerous, because it survives every automated check.

2. **The component that was left out.** *Cameroon:* 10% where the answer is 11%, because the additional council tax rides on top. Arithmetically close, practically wrong.

3. **The right rule applied to the wrong taxpayer.** *Illinois:* a net-loss provision that applies to corporations was applied to individuals, moving a cap from $100k to $500k.

4. **The national answer to a local question.** *India:* a single e-way-bill threshold, where each state sets its own.

5. **The stale figure.** *Cyprus:* "5.0–5.25% in recent years" against an actual 3.50% for 2026. Note the hedge — it was not confidently wrong, it was vaguely wrong, which is harder to flag and just as unusable.

6. **The invented attribution.** The figure is right; the authority cited for it is not. Rare in review, but dominant in a separate instrument we run: of 91 recorded deviations where a model was given a reviewed guide and asked to answer from it, 65 were invented attributions.

That last instrument gives a third data point worth stating, because it isolates the model from the retrieval problem entirely. Given a correct, accountant-reviewed guide directly in context, three models still deviated from it:

MODEL	RUNS	DEVIATED FROM THE GUIDE
Claude Haiku 4.5	62	14.5%
Claude Sonnet 5	62	17.7%
Claude Opus 4.8	62	24.2%

Same 31 questions, same guides, two runs each. Two cautions on reading it: at 62 runs per model these differences are inside the margin of error and should not be read as a model ranking, and the bigger model doing worse is interesting rather than established. The finding that *is* solid is the floor: **even handed the correct answer in context, roughly one answer in six drifts from it.** Supplying good data is necessary and not sufficient.

Section 12

Which model wrote it? We can bound it, not name it

We expected to name the model in this report. We cannot name the version, and the honest account of why is more useful than a guess would have been.

No record exists. The drafting was done by agent sessions doing web research, not by a script calling an API with a logged model parameter, and the output files were never committed. We checked every layer that could have captured it: the audit log for all 620 authoring events records `{slug, facts, topic, source, jurisdiction}` and no model; no provenance column exists on the facts, the skills or the change sets; the signed ingest receipts record who *verified*, never who *generated*; the one model column that exists anywhere in the schema is an empty table.

And the working assumption was wrong. We believed the drafts were Claude Opus 4.8. The chronology rules that out for most of the corpus: the first mention of that model anywhere in our codebase is 9 June 2026, by which date 941 of 1,928 guides already existed. The step that produced the facts the accountants actually reviewed is pinned in code to a different model, `claude-sonnet-4-6`, and has never been changed. The only artifact in the entire company that names anything is one Korean review workbook referring to "Claude's proposed values" — which establishes the family and no version.

What we can do instead is bound it by dates. Setting the public release dates of each model against our own guide-creation dates rules out most of the possibilities:

DATE	EVENT	GUIDES WRITTEN
5 Feb 2026	Claude Opus 4.6 released	–
17 Feb 2026	Claude Sonnet 4.6 released	–
9–14 Apr 2026	**our first drafting run**	378
16 Apr 2026	Claude Opus 4.7 released	–
20–27 May 2026	second drafting run	523
28 May 2026	**Claude Opus 4.8 released**	–
29 May – 8 Jun 2026	third drafting run	321

The consequence is unambiguous. **901 of the 1,230 guides drafted before review — 73% — were written before Claude Opus 4.8 existed.** The largest single batch, 350 guides on 13 April, predates even Opus 4.7 by three days; the newest model available that day was Opus 4.6. Only the 321 guides from 29 May onwards could have used 4.8 at all, and nothing records whether they did.

So the correct statement is: **drafted between April and June 2026 by Claude-family agents doing live web research, on models available at the time — Opus 4.6 or Sonnet 4.6 for the bulk of the corpus; exact versions not recorded and not reconstructable.** We have said this rather than the version we assumed, because a report about other people's unverified claims cannot rest on one of our own.

Model release dates above are from public secondary sources rather than a vendor changelog, and are stated to the day only where those sources agree.

One thing the same investigation did establish, and it is good news: the drafting was genuinely grounded. Of the surviving draft artifacts, 100% carried a legal reference and 72.7% carried a live source URL *at the moment of drafting*, and they cite documents that could not have come from a model's memory — including a District of Columbia payment booklet whose filename is date-stamped weeks before the draft was written.

> **A bug this uncovered**
>
> Those source URLs were captured and then lost. Tracing one country end to end: 74 of 74 facts had a URL when drafted, all 74 survived the first load, and only 9 of 73 still have one today. Corpus-wide, 3% of facts carry a source URL against 47% carrying a legal reference. **We did not fail to collect the links. We collected them and dropped them in a later rebuild.** That is a fixable pipeline defect, and it is now the highest-value repair we know of.

Section 13

The number that reframes all the others

Everything above concerns facts somebody checked. Here is the whole corpus:

66,253	3,359	5.1%	3.0%
facts published across 186 jurisdictions	have ever been reviewed by an accountant	of the corpus has any human verdict at all	carry a source URL a machine can follow

The other 94.9% does not have an accuracy rate of 74%. It has an accuracy rate of *unknown*. Applying a reviewed subset's number to an unreviewed corpus is the single most tempting mistake available to us, and we are not going to make it in our own report. Kenya is the sharp example: fourteen published guides, 156 facts, zero reviews by anyone, ever.

This is why the benchmark in Sections 3 to 7 matters more than our own 74.3%. It measures the thing that scales.

Section 14

What we got wrong about our own data

An earlier draft of this report contained a finding that the United Kingdom showed 182 facts confirmed, zero corrections and zero annotations, and drew the obvious conclusion about the reviewer.

That conclusion was false, and the fault was ours. The UK facts were imported with a flag that stamped every row "correct" without a human ever seeing them, from a workbook whose every row read "Pending". Our own records show the reviewer, James Power, returned two genuine

corrections — the two the founder remembered independently. Our import overwrote his actual work with a fabricated hundred percent and then displayed it on a dashboard.

> **Still live**
>
> Those 182 fabricated rows remain in production today. A fix script exists and has never been run. Until it is, our own dashboard reports a verification that did not happen — the precise failure this company exists to prevent, sitting inside the company.

We are recording this in the report rather than quietly fixing it first, because a research report from a company whose product is provenance should show what happened when its own provenance failed. We came within one edit of publishing a named professional's supposed negligence that was in fact our software's.

Section 15

Limits of this study

Stated plainly, because every one of them would be found by a sceptical reader anyway.

1. **Sixty questions is a small benchmark.** The confidence intervals are wide (±10 points) and they overlap between arms. The paired comparison carries the finding; the absolute percentages are indicative.

2. **The benchmark pool is biased toward facts our AI already got right.** 55 of the 60 answers are ones an accountant confirmed rather than corrected — that is what short numeric facts look like after review. Both arms are therefore being tested on material selected for answerability. This inflates both scores, and it is the single biggest reason to read 77% and 88% as ceilings.

3. **The ground truth is an accountant's verdict, not the law itself.** Where a reviewer confirmed a fact without checking it, an error passes into the answer key. Two questions in this very set exposed defects in our own stored values.

4. **Grading was done by a model, not a human panel.** Two independent graders agreed 96% of the time and disagreements were resolved toward the stricter mark, but this is not the same as accountant adjudication.

5. **One model family, one point in time.** These runs are Claude models in August 2026. We have not tested GPT, Gemini or open-weight models on this benchmark, so nothing here supports a claim about which vendor is best. The published TaxCalcBench leaderboard suggests GPT-5.6 currently leads on US returns, which is a reason to be sceptical of reading our single-family result as a statement about AI in general.

6. **Our numbers are not on the same scale as the public benchmarks.** Ours are per-fact; TaxCalcBench's headline is per-whole-return. Section 9 sets out the correct comparator and why the flattering comparison is the wrong one.

7. **Four countries are excluded and the exclusion is unresolved.** The UK, Tanzania and Nigeria workbooks recorded no corrections at all, and one Saudi workbook is a duplicate. Excluding them makes our headline worse, not better: 25.7% flagged rather than 22.4%. They are held out pending the re-examination described below.

Section 16

What is still open

1. **Run the UK fix.** 182 rows in production assert a verification that never happened. This is the only item here that is actively misleading someone today.

2. **Finish the miss-rate check.** Two of a planned twenty independent re-checks are done. Both found drift rather than reviewer error — figures correct when reviewed, superseded since. Completing this is what converts "at least a quarter were wrong" into a real estimate of what review missed.

3. **Re-examine the four held-out countries** and either restore them with a rigour caveat or state why they cannot be counted.

4. **Restore the lost source URLs.** 72.7% of facts had one when drafted; 3% have one now. Recovering them from the draft artifacts is mechanical and would raise machine-checkability by an order of magnitude.

5. **Close the 119 open review issues.** Every issue raised by reviewers in July is still marked open, though the underlying corrections were applied. The tracker is dead and the record cannot say which fix answered which issue.

6. **Record the model from now on.** Every future draft should carry the generating model, the date and the retrieval mode, so the next edition of this report can name what this one could not.

Section 17

What this says, plainly

A frontier model is a competent tax researcher and an unreliable tax authority. Asked a question about a rate or a threshold, it will answer nearly always, and on generous terms it will be right roughly three times in four from memory and closer to seven in eight if it can search. For a great many purposes that is genuinely useful, and it is better than the popular account of these systems suggests.

It is not what reliance requires. The residual errors are not random noise that more compute will sand away — they are structural, and they cluster in one place. **The model is good at the rate and bad at the exception.** Searching fixes what has gone out of date; it does not fix a rule that has a condition attached, because the summaries it can actually reach have already dropped the condition. And having searched, it becomes markedly more confident without becoming correspondingly more correct.

The gap that remains is therefore not a knowledge gap. It is the difference between recalling a rate and asking who the taxpayer is. That is the accountant's question, and nothing we measured closes it.

Two things follow for what we build. First, the binding constraint is not more guides; it is machine-reachable primary sources, because the world's tax authorities are closed to agents and the industry has quietly standardised on one Big Four summary as a substitute. Second, the review that matters is not "is this figure current" — search handles that — but "which taxpayer does this figure apply to". That is where a named professional adds something a model demonstrably does not have, and it is where our reviewers' corrections concentrated too.

There is a third implication we did not expect to find, and it is a caution rather than a strategy. A large corpus of unmaintained facts is not obviously better than no corpus: the published evidence is that stale material sitting in a model's context is worse than an empty context, because the model reconciles the two and loses. We have 66,253 facts and human eyes on 5.1% of them. The work that matters is not the next ten thousand facts. It is dates, sources and review on the ones we already have.

> **Reproducing this**
>
> The benchmark builder and scorer are committed as `scripts/build-tax-fact-benchmark.mjs` and `scripts/score-tax-fact-benchmark.mjs`. The sample is drawn with a fixed seed, so the same 60 questions come back on every run. Question set, both arms' raw answers, both graders' marks and the tally are retained.

OpenAccountants Research Report OA-RR-2026-02 · 13 August 2026 · Figures verified against the production database on the date of publication. Reviewer names are published with the reviewers' work; corrections and disputes to the research team.