

A New Frontier in AI Tax Intelligence: How Harness Engineering Improves Factual Accuracy and Verifiability

Does retrieval tooling make AI tax guidance more accurate, and easier to check? Two hundred expert-level questions, put to the same model with and without our harness, run twice and graded blind against the law itself.

CFO Silvia Research · August 2026

200 questions

Two benchmarks of 100. The analysis rests on the 190 that both runs posed; the ledger in section 1 accounts for the rest.

1,511 answers

Four tool setups, two runs. Each written in a fresh session, tool use audited from the logs.

2,949 judgments

Two rubrics per answer, graded blind. Citations looked up in the law itself.

ABSTRACT

Language models answer tax questions fluently, but fluency is not authority. A practitioner cannot act on guidance that offers no way to check it against the law it claims to rest on. We set out to measure what retrieval tooling changes about that.

We put the same frontier model through two 100-question benchmarks under **four tool configurations**: once from memory alone, once with a curated library of the tax law itself, once with a web search tool it ran on its own, and once with both. The first benchmark is built to catch the model out on the fine print of the tax code. The second is made of the kind of planning questions real clients actually bring. We ran the whole study twice. Every answer was scored from 1 to 10 on two rubrics, and the judges never knew which setup had produced the answer they were grading. Four configurations, two runs, and the 190 questions both runs posed come to **1,511 answers** and **2,949 blind judgments** in all.

On the hard statutory questions, the harness beat the model with no tools by **2.13** points on accuracy and **3.40** points on grounding. Both gaps are large, and both are statistically solid: if the tools made no real difference, results this lopsided would turn up by chance less than once in a hundred attempts for accuracy ($p=0.008$) and less than once in a thousand for grounding ($p<0.001$). Both runs agreed the tools helped. They disagreed on how much, improving accuracy by +1.32 in the first run and +2.92 in the second. That is a wide spread, and it is not just chance: I^2 , the share of the variation between runs that reflects a genuine difference in conditions rather than chance, comes out at 97%. With only two runs to compare, the per-run estimates carry more weight than the pooled one.

The client scenarios behaved differently. There the model was already accurate without any tools, and adding tools barely moved it: a gain of +0.13 points, comfortably within the range of chance ($p=0.244$). What did improve was grounding, by +0.38 points, a small gain but a reliable one ($p=0.012$), and it came from the curated library rather than from web search. The two mechanisms are not doing the same job.

Web search supplies facts the model does not have. The library supplies the citation that lets someone check the answer.

Every answer carries two scores, and they measure different things.

Accuracy asks whether the answer is right: the correct conclusion, the correct figures, for the correct tax year.

Grounding asks whether you can check it: does the answer cite the statute or ruling it depends on, and when the judge looks that citation up, does it actually say what the answer claims? An answer can be perfectly correct and still unusable to a practitioner if there is no way to verify it. That is why we score the two separately.

I. Design

Every setup answered every question in a fresh session with no conversation history, so no answer could benefit from another. Tool restrictions were enforced in the prompt and then audited against the tool-call logs. Runs that violated their restriction were quarantined and re-collected.

Each answer was then scored by a judge that had both retrieval sources available to it. The judge was instructed to look every claim up in the law itself rather than take the answer's word for it, recompute the figures for the correct tax year, and look for evidence that the answer was wrong. It was not told which setup had produced the answer it was scoring. One judge sample was taken per answer per rubric.

Setup	Description
Base model	The model called directly, with no tools. It answers from what it absorbed during training.
Silvia tax data store	The model given our curated library of the tax law itself: the tax code, Treasury regulations, IRS publications and instructions, state statutes and revenue-department guidance. Web access is switched off.
Web search	The model given a web search tool it runs itself, deciding what to look up and when. The Silvia data store is switched off.
Silvia harness (store + web)	The model given both the library and web search. This is what we run in production.

How the counts add up

Benchmark and run	Questions posed	Setups	Answers kept	Judgments
Adversarial-100, Run 1	100	4	400	800
Adversarial-100, Run 2	100	4	391	709
Client-100, Run 1	90	4	360	720
Client-100, Run 2	90	4	360	720
Total			1,511	2,949

Three gaps sit behind these totals. **(1)** The published question set contains 200 items, but 10 of the client scenarios were never posed in either run, so the analysis rests on 190 questions. **(2)** Run 2 of Adversarial-100 lost 9 answers of an intended 400 to an execution-time ceiling. **(3)** That run is also short 73 of the 782 judgments its 391 answers should have produced: 45 in the web-search setup, 25 in the baseline, the rest scattered, again from the judging pass hitting its time ceiling rather than from any filtering on content. Contrasts are computed pairwise-complete, so a missing judgment removes that question from that one contrast and nothing else; the per-setup *n* behind every point is shown on hover in the web version and in the agreement tables below.

2. Benchmark A: Adversarial-100

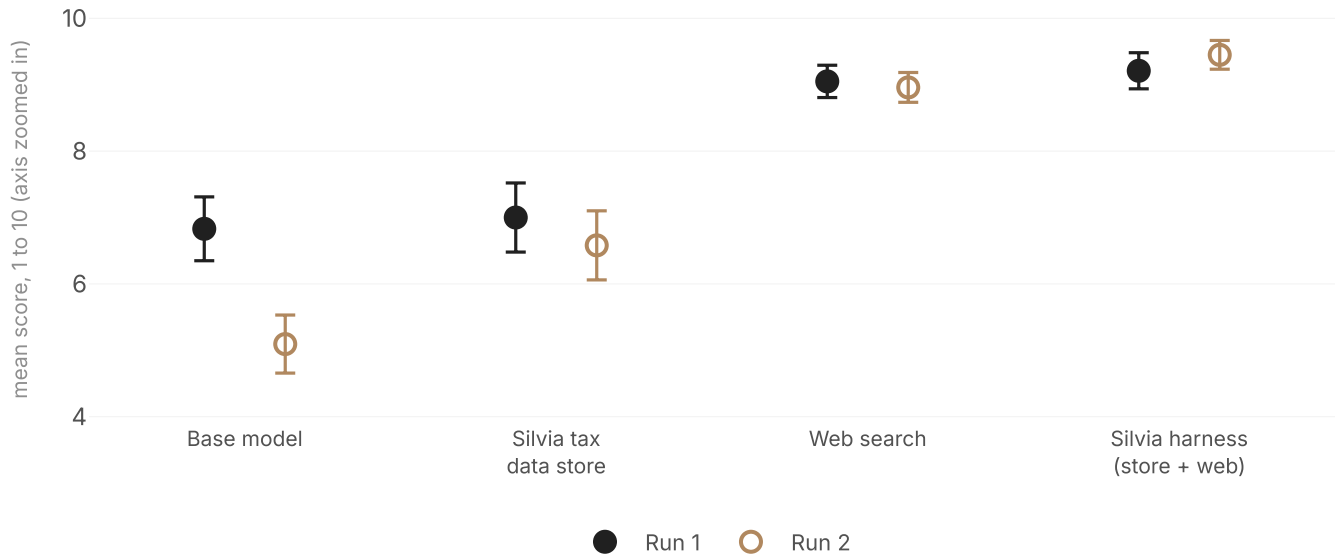
One hundred deliberately difficult questions: the exact wording of a subsection, figures that change from year to year, provisions that interact in awkward ways, and state rules that break from the federal treatment.

Run 1: 400 answers, 800 judgments. Run 2: 391 answers, 709 judgments.

Average score for each setup



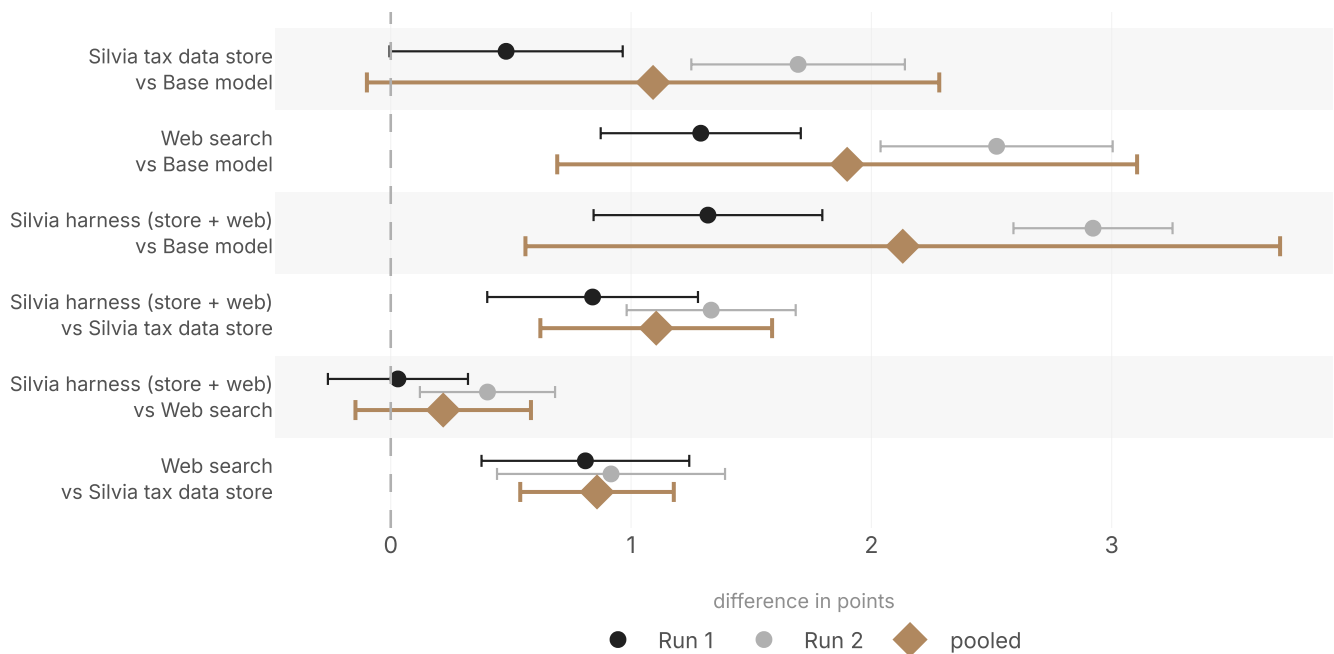
Grounding (source verifiability)



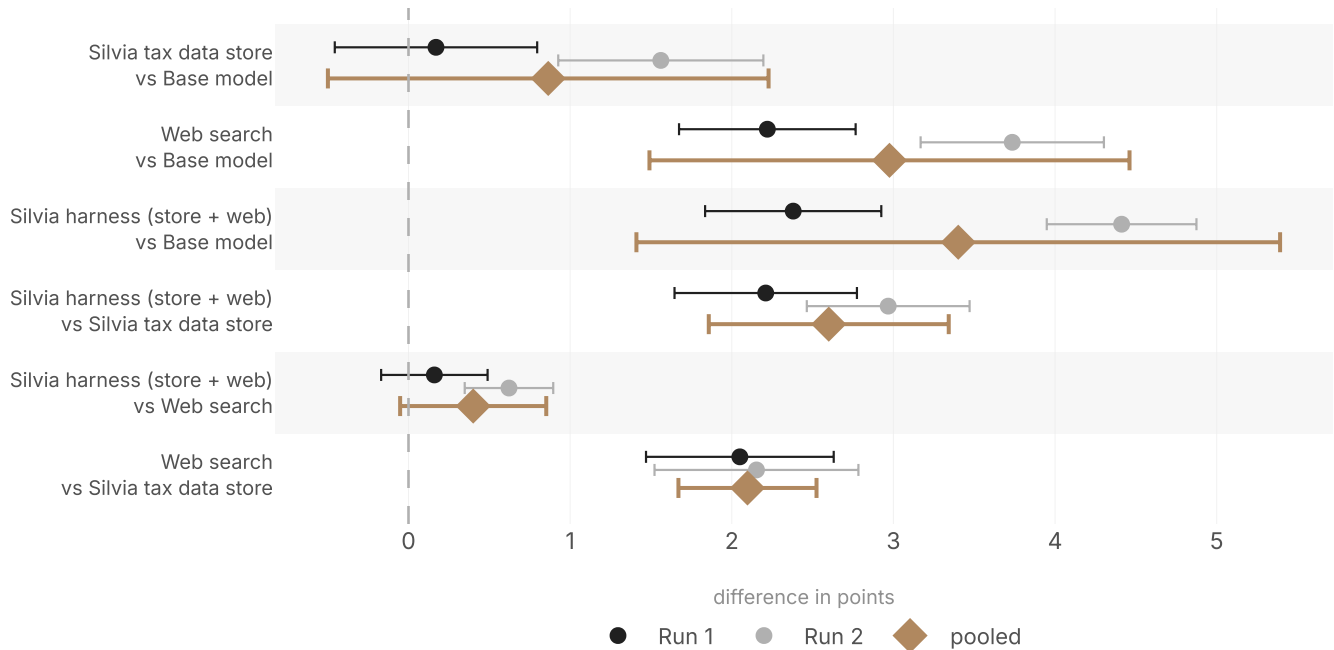
Each marker is an average, shown with its 95% confidence interval. Filled markers are Run 1, open markers Run 2. The vertical axis is zoomed in so the differences between setups are visible; we use points rather than bars, because starting an axis above zero makes bars look far more different than they are.

Gaps between setups, per run and combined

Factual accuracy



Grounding (source verifiability)



Circles are the mean paired difference within each run, computed on identical questions, with 95% confidence intervals: dark for Run 1, light for Run 2. Gold diamonds combine both runs into a single estimate (random-effects, DerSimonian-Laird). An interval clearing the dashed zero line means the first-named setup genuinely won. Each pooled estimate carries an I^2 value, the share of the variation attributable to genuine disagreement between runs rather than chance, reported on hover in the web version. Where I^2 is high, the per-run estimates carry more weight than the pooled one, since only two runs are being combined.

How closely the two runs agreed

Setup	Accuracy r	within ± 1	Grounding r	within ± 1	n items
Base model	0.50	51%	0.55	46%	90
Silvia tax data store	0.52	74%	0.78	63%	90
Web search	0.13	86%	0.26	77%	77
Silvia harness (store + web)	0.31	86%	0.39	89%	100

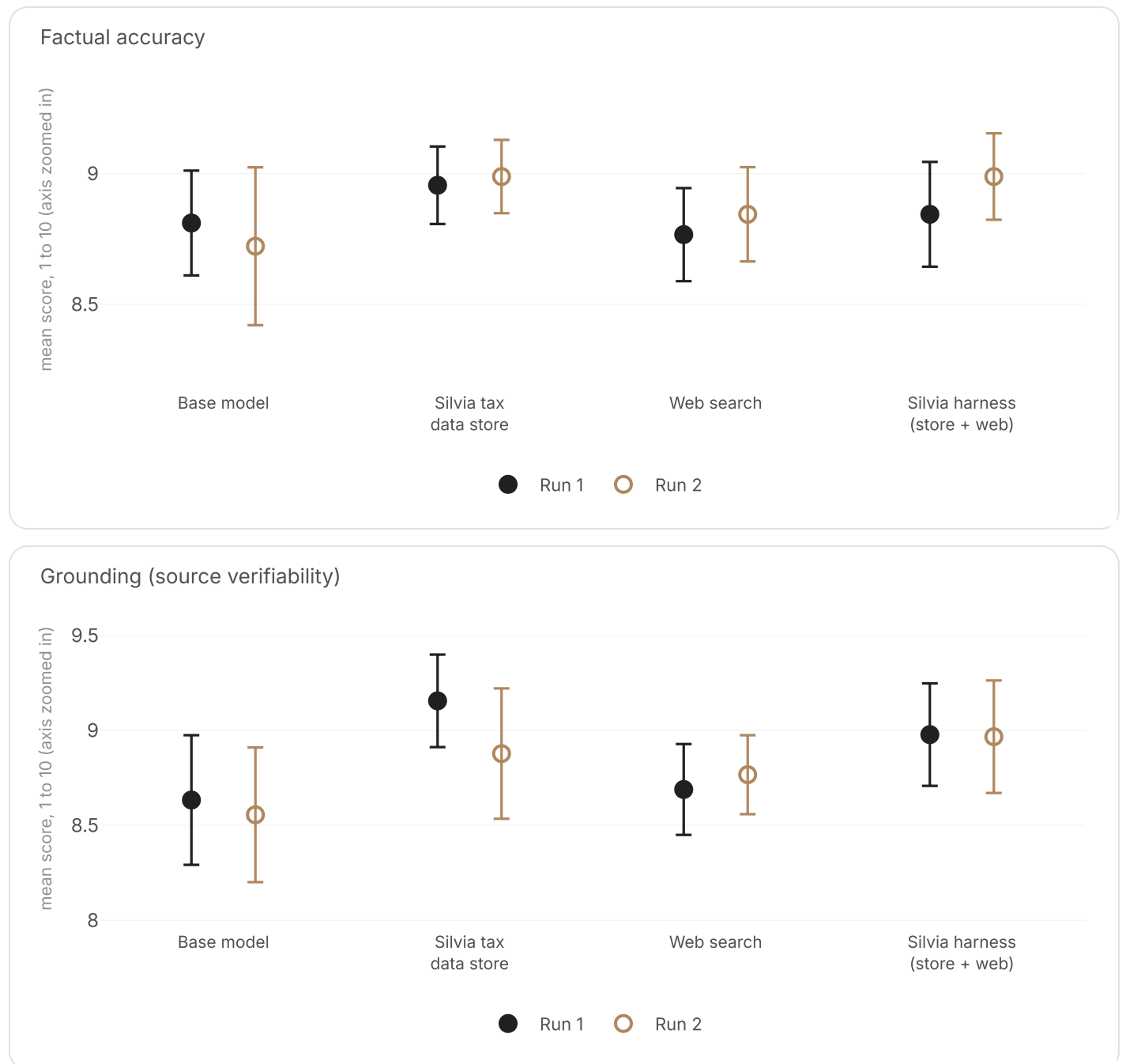
Correlation between the two runs' per-question scores (Pearson), and the share of items scored within ± 1 point across runs. Each run collected answers and judgments independently.

3. Benchmark B: Client-100

One hundred real planning questions from actual clients, each paired with that client's financial profile. Every record was anonymized first: cities and time zones dropped, ages replaced with bands, and dollar figures converted to ranges. This is the benchmark we care most about, because it is what our customers actually ask. The figures below rest on the 90 questions both runs posed verbatim, out of the set's 100.

Run 1: 360 answers, 720 judgments. Run 2: 360 answers, 720 judgments.

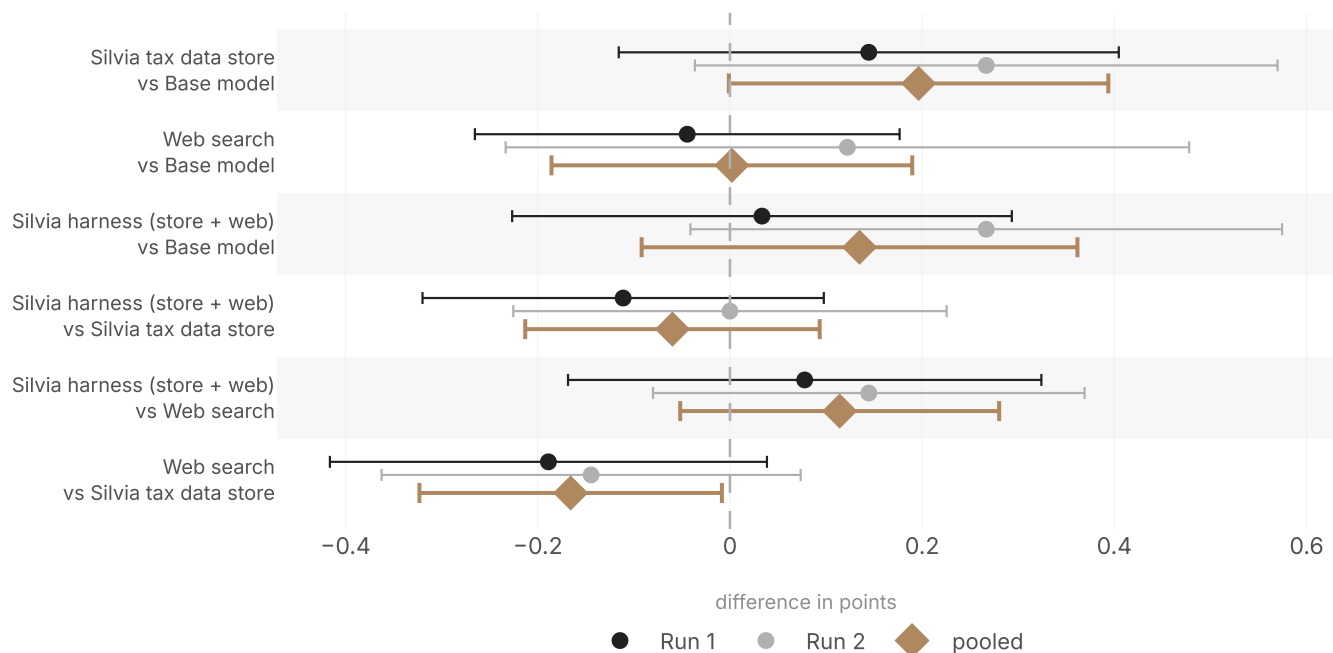
Average score for each setup



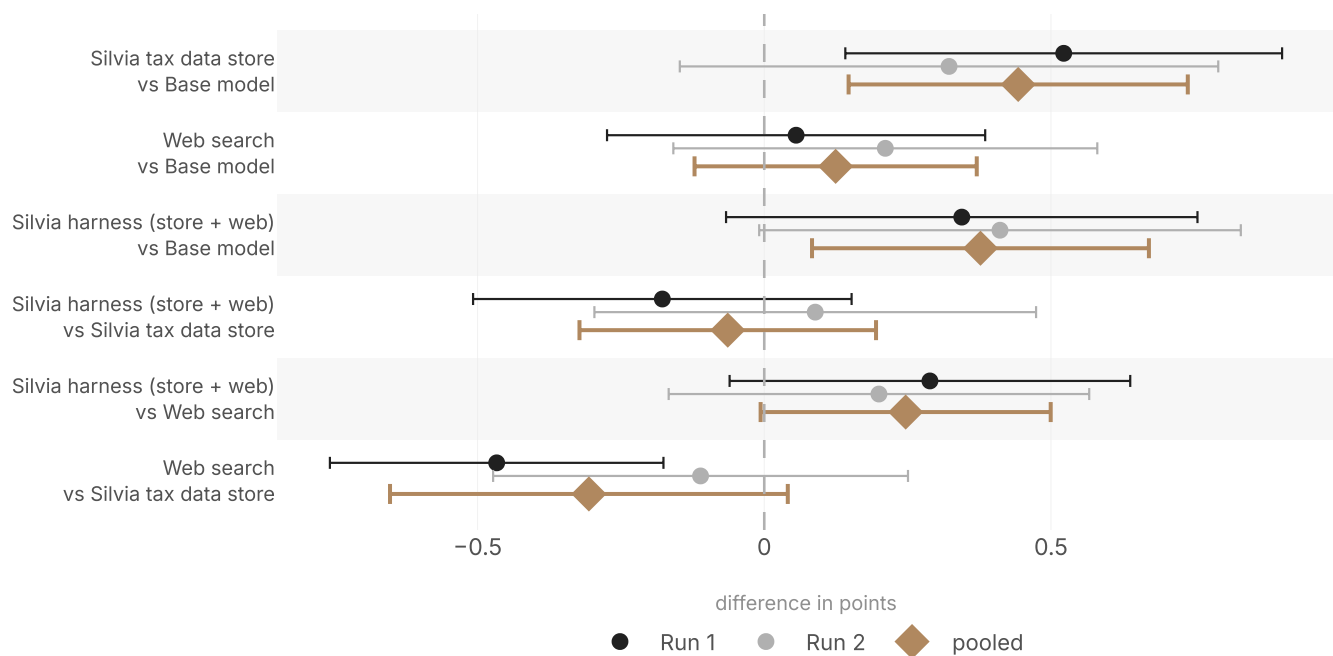
Each marker is an average, shown with its 95% confidence interval. Filled markers are Run 1, open markers Run 2. The vertical axis is zoomed in so the differences between setups are visible; we use points rather than bars, because starting an axis above zero makes bars look far more different than they are.

Gaps between setups, per run and combined

Factual accuracy



Grounding (source verifiability)



Circles are the mean paired difference within each run, computed on identical questions, with 95% confidence intervals: dark for Run 1, light for Run 2. Gold diamonds combine both runs into a single estimate (random-effects, DerSimonian-Laird). An interval clearing the dashed zero line means the first-named setup genuinely won. Each pooled estimate carries an I^2 value, the share of the variation attributable to genuine disagreement between runs rather than chance, reported on hover in the web version. Where I^2 is high, the per-run estimates carry more weight than the pooled one, since only two runs are being combined.

How closely the two runs agreed

Setup	Accuracy r	within ± 1	Grounding r	within ± 1	n items
Base model	0.30	90%	0.18	68%	90
Silvia tax data store	0.18	92%	0.37	80%	90
Web search	0.07	88%	0.12	79%	90
Silvia harness (store + web)	0.06	88%	0.14	81%	90

Correlation between the two runs' per-question scores (Pearson), and the share of items scored within ± 1 point across runs. Each run collected answers and judgments independently.

4. Synthesis across benchmarks

Both runs combined	Adversarial - 100 accuracy	Adversarial - 100 grounding	Client - 100 accuracy	Client - 100 grounding
Silvia harness (store + web) vs Base model (harness value)	+2.13 [+0.56, +3.70], p=0.008	+3.40 [+1.41, +5.39], p<0.001	+0.13 [-0.09, +0.36], p=0.244	+0.38 [+0.08, +0.67], p=0.012
Web search vs Silvia tax data store (mechanism comparison)	+0.86 [+0.54, +1.18], p<0.001	+2.10 [+1.67, +2.52], p<0.001	-0.17 [-0.32, -0.01], p=0.039	-0.31 [-0.65, +0.04], p=0.084

Each cell gives the average gap between the two setups, in points on the 1–10 scale, followed by the range the true gap most likely falls in and the probability of seeing a gap that large if the two were really equal. A range sitting entirely above zero means the first-named setup won; a range straddling zero means the result is inconclusive.

The two benchmarks point in opposite directions. On the fine-print questions, web search does almost all the work and the curated library adds little on top of it. On the realistic client scenarios, the setups separate very little on accuracy, and the curated library is the only thing that meaningfully improves whether an answer can be checked.

What retrieval is worth depends on the question being asked. Neither mechanism wins everywhere, so a system built on only one of them will be weak somewhere. Across both benchmarks and both rubrics, the combined setup is the only one never significantly behind any other, and that is the property that matters when you cannot know in advance which kind of question a user will bring.

5. Limitations

Coverage

Ten of the 200 published questions, all client scenarios, were never posed in either run. Run 2 of Adversarial-100 is missing 9 of 400 answers and 73 of the 782 judgments those answers should have

produced, both from execution-time ceilings rather than from any content-based filtering. The missing judgments are not spread evenly: 45 fall in the web-search setup and 25 in the baseline, so that run's web-search estimates rest on noticeably fewer items than its others, and the two setups with the largest gaps are the two that make up one of the headline contrasts. Everything is computed pairwise-complete on the items that survive.

Measurement and statistics

Rubric scores are ordinal and are analyzed as interval data. P-values use the normal approximation, which all pooled contrasts support at $n \geq 80$ pairs. Pooling combines only two runs, so I^2 and the between-run variance are estimated very coarsely and the per-run estimates deserve at least as much weight as the pooled ones, which is why both are always shown. Mean scores on Client-100 sit near 9 on a 1–10 scale, close enough to the top that compression may be limiting how much movement any setup can show.

6. Data and materials

The benchmark is open. All 200 questions and their judge scores are published as **silvia-tax-bench** under a CC BY 4.0 license:

<https://huggingface.co/datasets/cfosilvia/silvia-tax-bench>