What 1830 GRPO changed in Nemotron-3 Nano

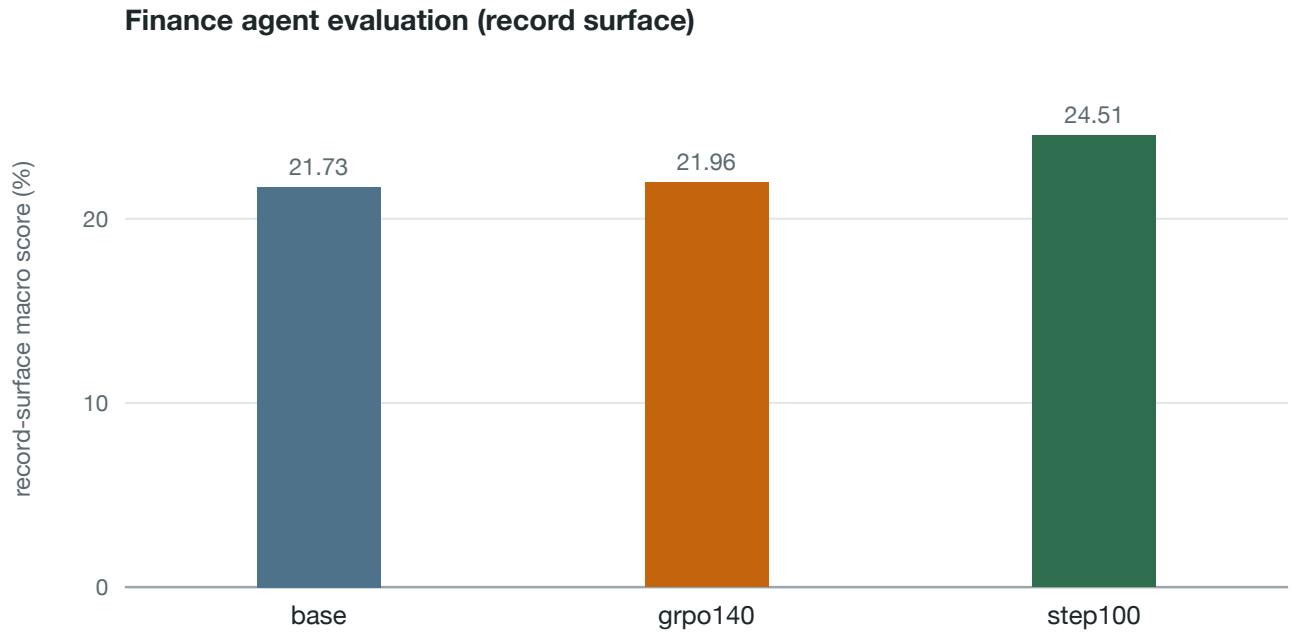
Models can learn transferable work habits without learning the target domain, but the habit that appears depends on how it is tested. We compared a base Nemotron-3 Nano against two checkpoints from GRPO (Group Relative Policy Optimization) training on the board game 1830, called grpo140 and step100. We then tested them on SEC-filings finance research and on an independent citation-grounding stress test.

The two checkpoints changed in different, measurable ways. step100's clearest change is fewer contradictory claims in the original finance harness. grpo140's clearest change is a more economical citation pattern in a reduced, independent ALCE-style citation-grounding stress test (a long-form answer evaluation where every claim should cite retrieved sources). These are behavior-level observations on different datasets, not one universal mechanism.

Two lenses, two leaders

The headline depends on the evaluation you pick up. Switch lenses below; the chart, the model ordering, and the interpretation change together.

Lens 1: finance agent evaluation

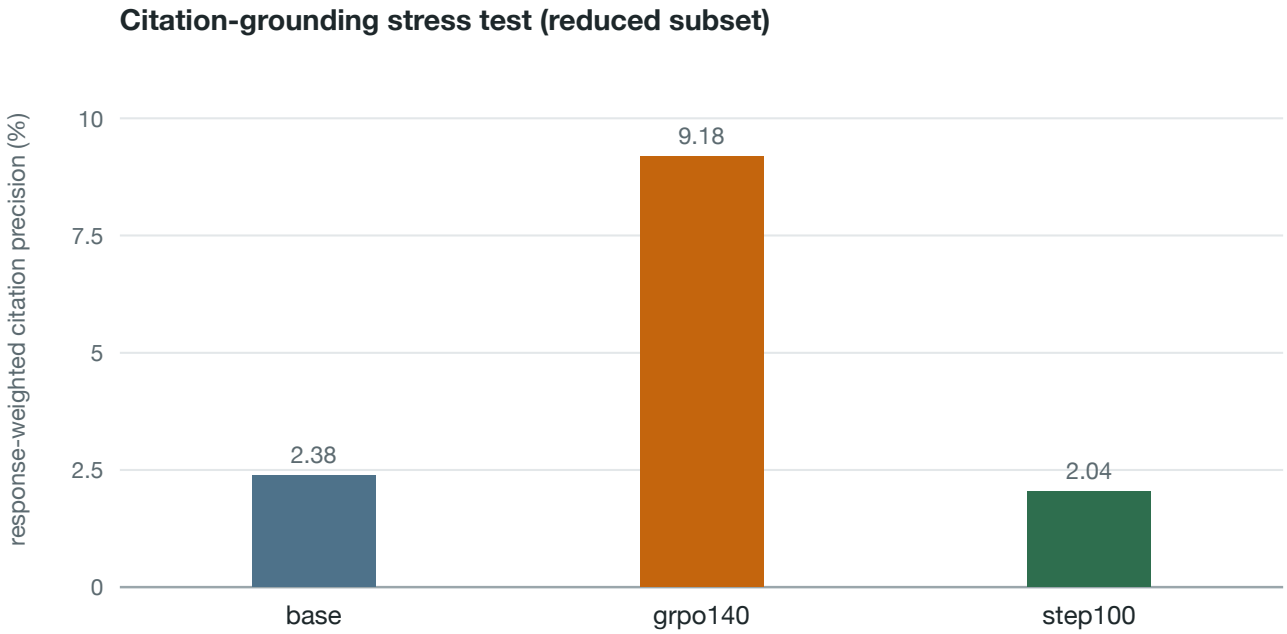


Ordering here: step100 (24.51%) > grpo140 (21.96%) > base (21.73%).

step100 leads on the finance record surface. That record-surface gain is fragile: dropping a single favorable item moves its confidence interval across zero. The statistically clearer evidence for step100 is a separate within-stack comparison on the same finance harness: 18.3% vs 12.1% macro, a +6.2 pp gain with a 95% bootstrap interval of [+2.3, +10.4] and a sign test at p=0.024.

Limitation: the record-surface numbers above are point estimates from one run configuration; the fragile lead and the within-stack result are unpacked in the forest plot below.

Lens 2: citation-grounding stress test



Ordering here: grpo140 (9.18%) > base (2.38%) > step100 (2.04%).

grpo140 leads this reduced subset on response-weighted citation precision. These are exact per-model means over 14 responses each (42 responses total): base 2.38095%, grpo140 9.18367%, step100 2.04082%.

Limitation: only four independent sources survived complete-case selection, so this ordering is exploratory and can flip under source-weighted aggregation (shown later).

The original bet

The two checkpoints learned 1830 in different formats. grpo140 came from single-turn game question answering. step100 came from agentic terminal tasks that required multi-step work, tool use, and files. The transfer target was different-domain work with a similar shape: researching SEC filings and producing finance answers.

The bet was that training on an agentic task surface might transfer a work habit without transferring game knowledge. The independent citation study then asked whether any such habit appeared on a second, non-finance evaluation the training never saw.

Model colors are fixed throughout: base, grpo140, step100, step60 (an intermediate checkpoint).

What changed in the finance harness

Macro scores across three finance surfaces (%)

	record	terminal	finalize (exploratory)
base	21.73	17.19	15.68
grpo140	21.96	18.69	17.64
step100	24.51 †	18.42	16.62

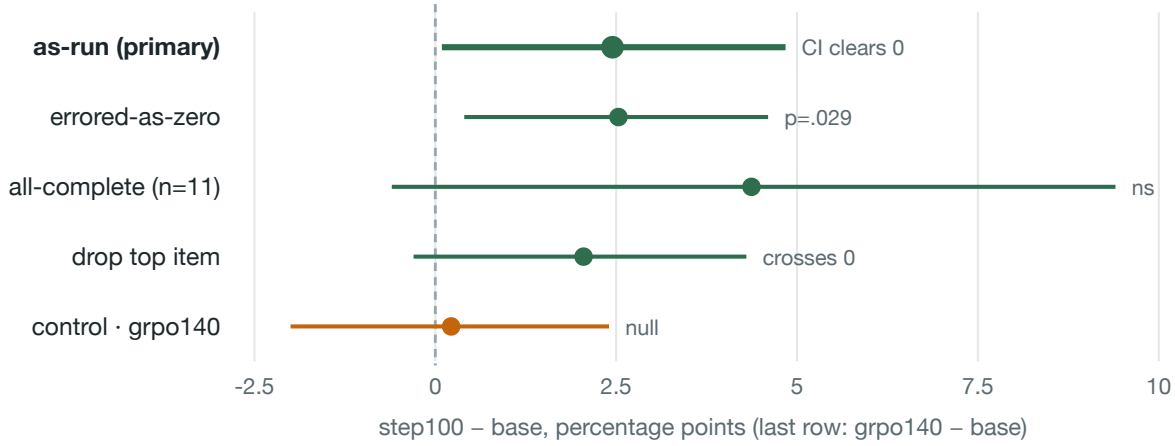
† record-surface step100 lead is fragile; the clearer within-stack result is reported separately below.

Description. Macro scores (%) for the three models across the three surfaces of the original finance harness: record, terminal, and finalize. Darker cells are higher scores. Exact values are shown in each cell.

Findings. step100 posts the highest record-surface score (24.51% vs 21.73% base), and all three surfaces show small spreads between models. The step100 record-surface lead is fragile (†), and the finalize surface is exploratory. The statistically clearer within-stack result (+6.2 pp, interval [+2.3, +10.4], sign test p=0.024) is reported separately in the text rather than hidden in these cells.

> **Limitation.** These cells are point estimates, not confirmed effects. Only the within-stack comparison carries a paired significance test; do not read every cell as confirmatory.

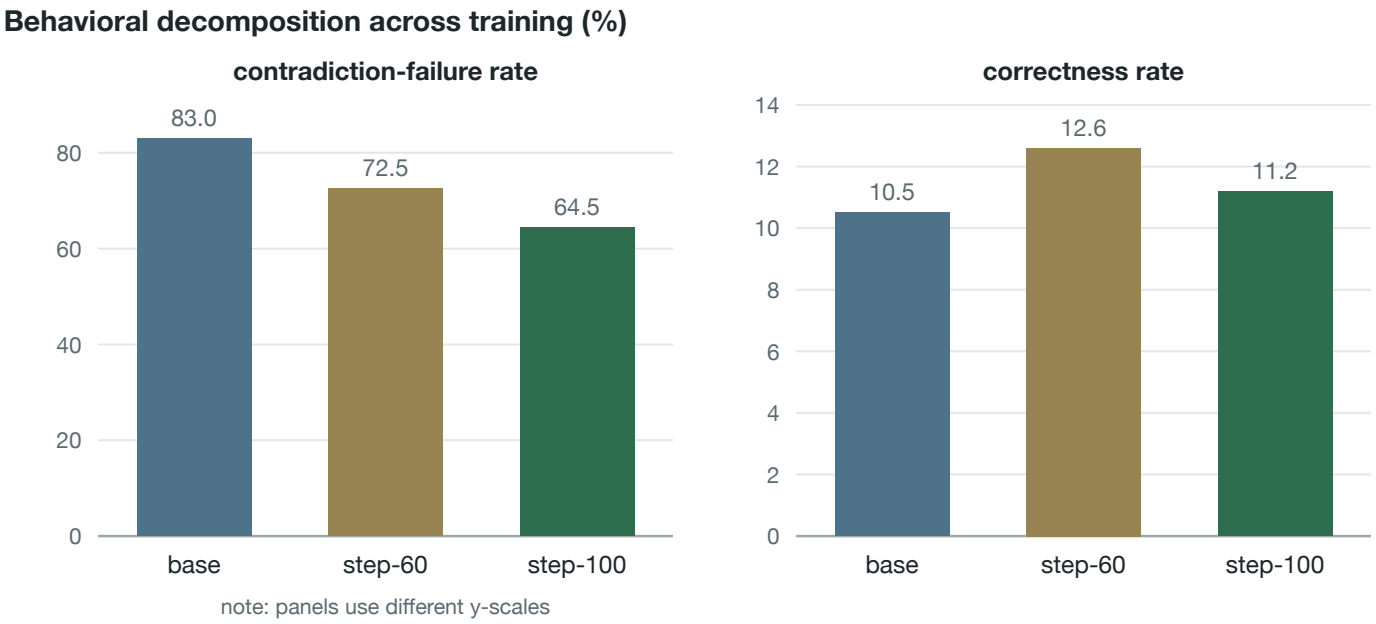
Record-surface effect sizes: step100 – base under four analyses, plus the grpo140 control



Description. Effect of step100 relative to base on the record surface, in percentage points, under four analysis choices, with the grpo140-minus-base control in ember. Dots are means; lines are 95% intervals; the dashed line marks zero.

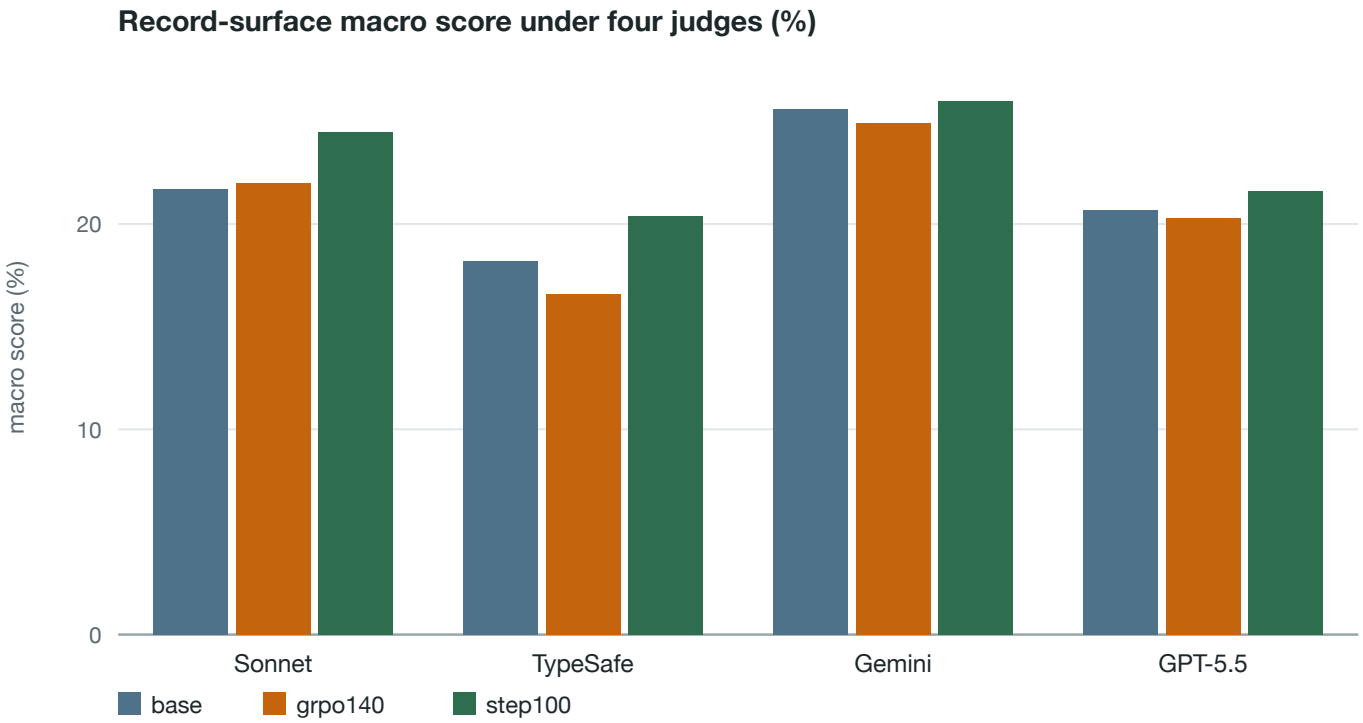
Findings. The as-run primary estimate is +2.45 pp [+0.09, +4.84] and clears zero, as does errored-as-zero (+2.53 pp, p=.029). The result is sensitive to one item: dropping the top item gives +2.05 pp [–0.3, +4.3], which crosses zero, and the all-complete restriction (n=11) at +4.37 pp [–0.6, +9.4] does not reach significance. Rank-based tests also miss the 0.05 threshold (sign test p=0.053, Wilcoxon p≈0.06). The grpo140 control is null (+0.22 pp [–2.0, +2.4]).

> **Limitation.** A one-item sensitivity of this size means the record-surface gain should be treated as fragile until replicated on a larger item set.



Description. Behavioral decomposition of finance-harness responses across training checkpoints. Left: the rate at which a response fails for containing contradictory claims. Right: the correctness rate. The panels use different y-scales.

Findings. Contradiction failures fall steadily with training: 83.0% (base) → 72.5% (step-60) → 64.5% (step-100). Correctness barely moves: 10.5% → 12.6% → 11.2%. The training mainly changed how the model asserts, not how often it is right. This is a behavioral decomposition, not a claim about internal causality.



Description. Record-surface macro scores for the three models when the grading judge is swapped: Sonnet, TypeSafe, Gemini, and GPT-5.5.

Findings. step100 minus base stays positive under all four judges, but the size ranges from +0.5 pp (Gemini) to +2.8 pp (Sonnet), with TypeSafe at +2.1 pp and GPT-5.5 at +0.9 pp. Paired significance holds only under Sonnet. The direction is judge-stable; the magnitude and the statistics are not.

What the change looks like in a transcript

Two examples from the finance harness illustrate the aggregate contradiction result. Examples illustrate aggregate behavior; they do not prove a mechanism.

The trained model separates 2025 from 2026 capital spending

Q: Who spends the most on 2025 capital expenditure?

> **BASE**
> “Meta's 2025 capex guidance is projected to be between $125 billion and $145 billion ... Meta has the highest announced 2025 capex budget.”

> **STEP100**
> “Meta's 2025 capex guidance is $66–$72 billion ... the summary also mentions raising 2026 guidance to $125–$145 billion ... Amazon's reported $105 billion capex for 2025 is the highest.”

The base answer conflates Meta's 2026 guidance with 2025. step100 keeps the years separate and reaches the recorded answer, Amazon.

A plausible Reuters link was invented

Q: US Steel and the Nippon Steel merger

> **BASE**
> “https://www.reuters.com/markets/commodities/us-steel-shares-rise-after-nippon-steel-merger-announcement-2024-02-01/”

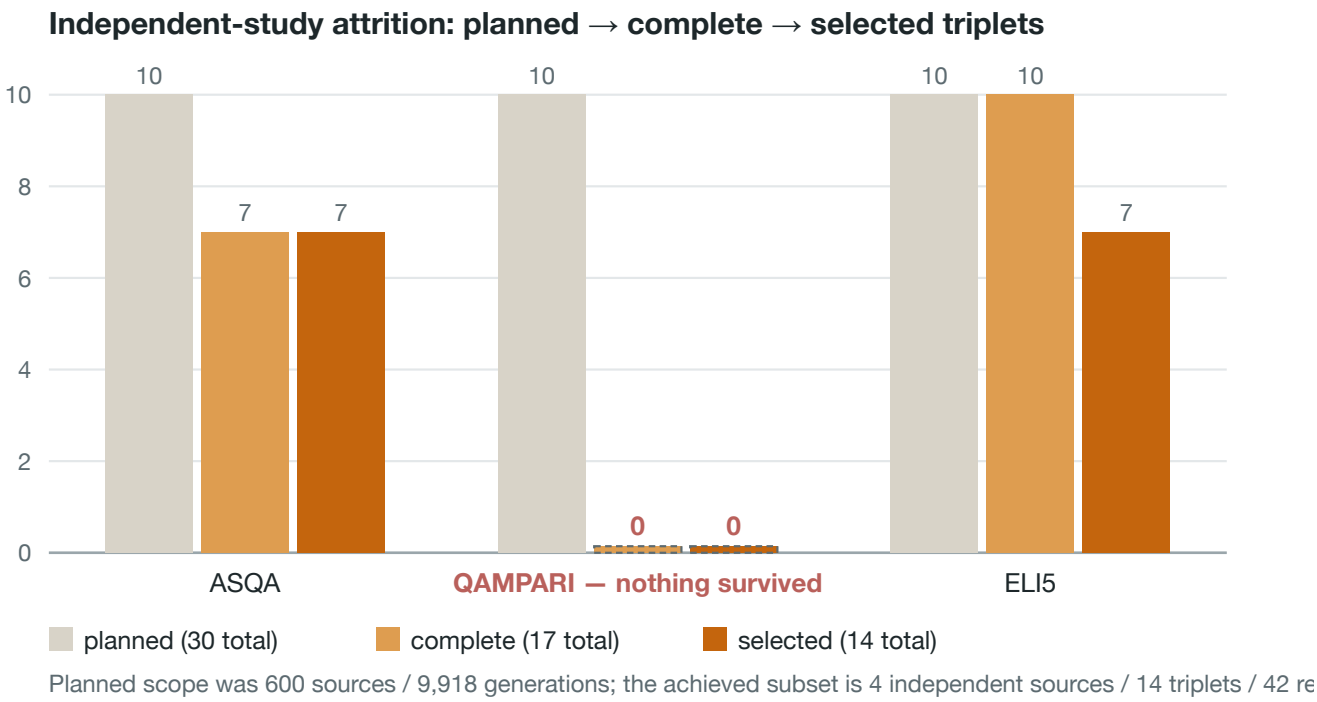
> **AUDIT NOTE**
> The report marks this base citation as fabricated: the merger was announced December 18, 2023, and the displayed URL does not exist.

This is an observed behavior example, not evidence of an internal neural mechanism.

An independent stress test

To test whether any work habit transferred beyond the training harness, we built an independent citation-grounding stress test in the style of ALCE: long-form questions answered against retrieved source documents, scored on whether claims cite supporting sources. This is a post-hoc, exploratory study; none of it was pre-registered against the training run.

The study shrank drastically between plan and analysis. The broader planned scope was 600 sources and 9,918 generations. What survived complete-case selection was 4 independent sources, 14 question triplets, and 42 naturally completed responses.



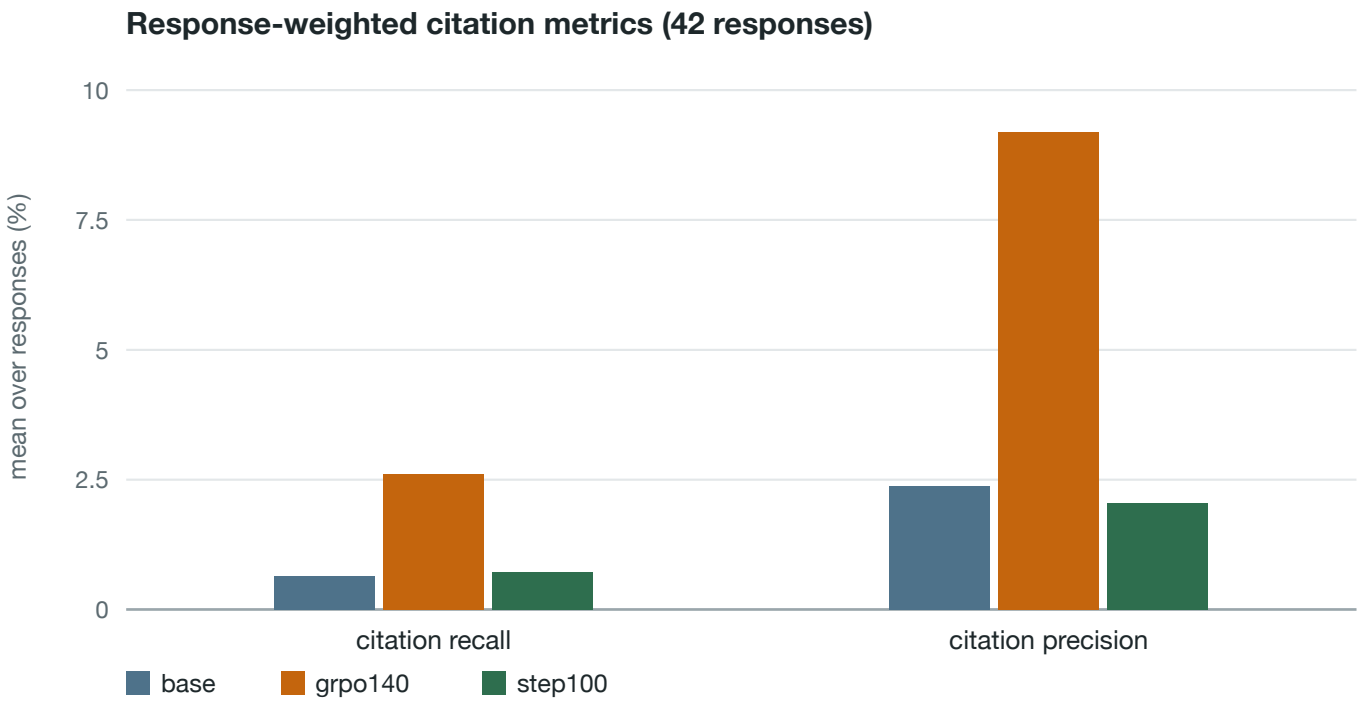
Description. Triplet counts by dataset at each attrition stage: 30 planned (10 ASQA, 10 QAMPARI, 10 ELI5), 17 complete (7, 0, 10), 14 selected as a balanced subset (7, 0, 7).

Findings. QAMPARI contributed nothing: zero of its ten planned triplets completed. Every conclusion below therefore rests on ASQA and ELI5 only, and on four independent sources.

> **Limitation.** Complete-case survivorship means the analyzed subset is not a random draw from the planned study. Missing QAMPARI removes an entire question style.

The unexpected grpo140 lead

The surprise is that grpo140, the checkpoint that was flat on the finance harness, leads the citation stress test.

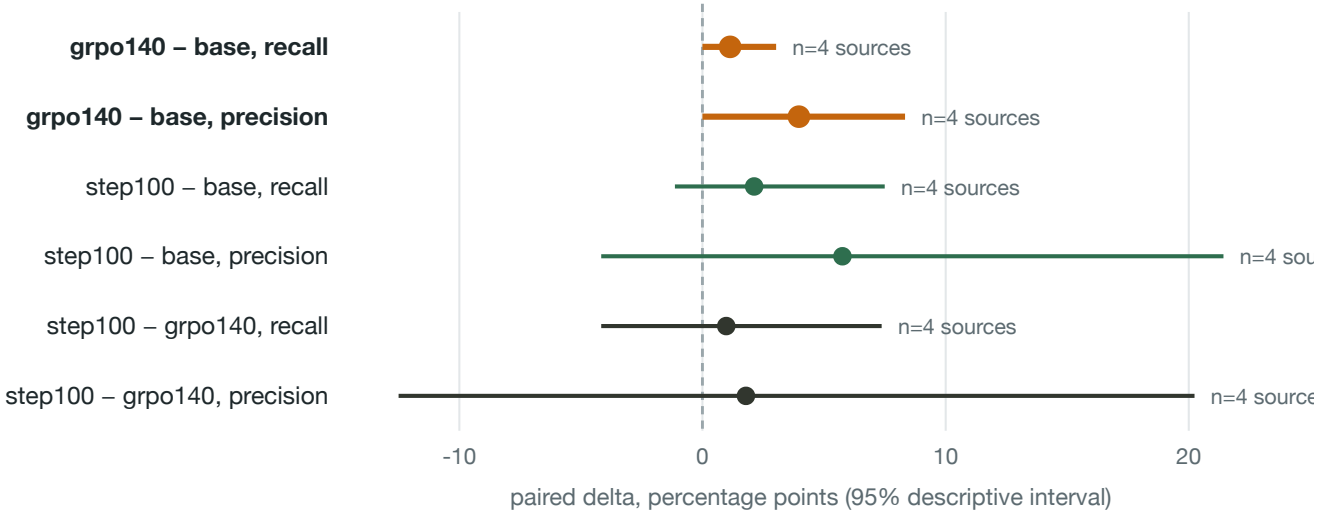


Description. Response-weighted citation recall and precision (%), averaged over each model's 14 responses (exact means: recall 0.65 / 2.60 / 0.71; precision 2.38 / 9.18 / 2.04 for base / grpo140 / step100).

Findings. On response-weighted means, grpo140 leads both metrics and step100 sits near base's level. Absolute levels are low for all three models; this is a stress test, not a benchmark any model passes.

Source-weighted paired deltas (four independent sources)

Descriptive intervals over 4 sources; not confirmatory.

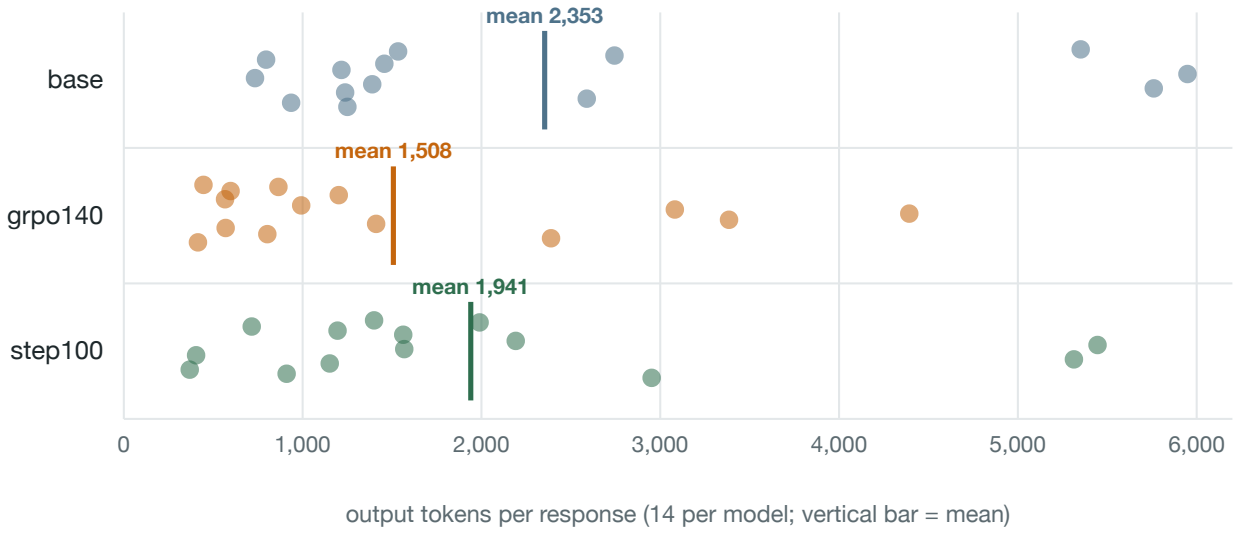


Description. Paired deltas in citation recall and precision when responses are first aggregated within each of the four independent sources, with 95% descriptive bootstrap intervals. The grpo140 – base rows (ember, bold) are the foreground comparison.

Findings. grpo140 – base is nonnegative across the interval on both metrics: recall +1.14 pp [0, +3.03] and precision +3.97 pp [0, +8.33]. But the same aggregation gives step100 – grpo140 a positive mean (recall +0.98, precision +1.79) with wide intervals. Response-weighted means favor grpo140; source-weighted means can slightly favor step100 over grpo140. Four sources cannot settle the ordering. These are descriptive intervals, not confirmatory tests.

> **Limitation.** With n=4 sources, one source flipping would rewrite these intervals. Treat the grpo140 lead as the strongest exploratory signal, not an established ranking.

Response length, all 42 independent-study responses (output tokens)



Description. Output-token length of every one of the 42 independent-study responses, one row per model, with the per-model mean marked. Hover a dot for its dataset, evidence condition, and seed.

Findings. grpo140 writes shorter answers (mean 1,508 tokens) than base (2,353) and step100 (1,941). We do not claim shorter answers caused the higher precision or a lower claim count; grpo140's mean claim count (7.93) exceeds base's (7.57), and grpo140 and base cite at the same rate (35.7% of responses contain any citation). grpo140 emitted slightly more citations (1.07 vs 0.79 per response) and more claims on average. Among citing responses, unsupported-citation rates were 74.3% for grpo140, 93.3% for base, and 92.9% for step100. The pattern reads as a more economical citation habit, not a shorter-therefore-better story.

Reading the outputs

Aggregate metrics hide how strange these responses are. Most answers in the independent study are off topic: the models answer a memorized question instead of the one asked, and the automatic scorer sometimes rewards that. Browse the four scenarios; each opens on the model the scenario is about.

Concise and on-topic to the wrong question

The grpo140 answer is shorter, but all three models fail the Grey's Anatomy prompt. Its lone citation is unsupported.

Question: “When does dr owen come to grey's anatomy?” — evidence condition: supported_original, supporting documents: 0.

> This case is also the shared no-difference failure: all three response-level utility and citation scores are zero.

Showing the grpo140 response (exact model output).

citation recall: **0%** citation precision: **0%** citations: **1** unsupported-citation rate: **100%**

utility: **0%** output tokens: **445**

The all-time record for rainfall belongs to the small town of **Mawsynram** in the East Khasi Hills of Meghalaya, India. Its climate is so wet that it receives more precipitation than any other place on the planet, making it the wettest place on Earth. (It receives more rain than any other location on the planet.)

[1] Mawsynram, situated in the East Khasi Hills of Meghalaya, India, receives the greatest amount of rain of any place on the planet, earning it the title of the world’s wettest place.

The scorer rewards support for an irrelevant answer

The prompt asks about Starbucks with support removed. grpo140 answers about Mawsynram, cites it, and fails to abstain.

Question: “How many starbucks are there around the world?” — evidence condition: support_removed, supporting documents: 0.

> Nonzero citation scores do not mean the model answered the question or handled missing evidence correctly.

Showing the grpo140 response (exact model output).

citation recall: **3.13%** citation precision: **28.57%** citations: **7** unsupported-citation rate: **71.43%** utility: **0%** output tokens: **1,202**

The wettest place on Earth is **Mawsynram**, a town in the East Khasi Hills of Meghalaya, India. It receives more rainfall than any other place on the planet, making it the world’s most rain-soaked location. The town is set in a deep valley that forces moist monsoon winds to rise, condense, and dump massive amounts of water, giving it an extreme, year-round wet climate. Moreover, its surrounding highlands force the prevailing winds to lift moisture-laden air, causing it to precipitate heavily, while the surrounding hills shelter the town from the monsoon rains that fall on the surrounding plains. This makes Mawsynram the world’s most rain-soaked location. (Citations: [3] and [3])

***

Mawsynram is a town in the East Khasi Hills of Meghalaya, India. It is a small town that lies in a valley and has a tropical climate. Because of its position, the area receives moisture-laden air that rises, cools, and falls as precipitation, leading to heavy rain. The area’s terrain forces the wet air to rise, cool, and fall as rain, creating a year-round wet climate. In addition, the surrounding highlands force the prevailing winds to lift moisture-laden air, causing it to precipitate heavily, while the surrounding hills shelter the town from the monsoon rains that fall on the surrounding plains. [1]

Thus, the answer to “which is the most rainy place on Earth?” is **Cherrapunji** (Mawsynram), a town in Meghalaya, India, renowned for receiving the highest amount of precipitation on Earth. The heavy rains are the result of moist air being forced up the valley, cooling, and falling as rain. The area’s shape forces the wet air upward, where it cools and falls as rain, creating a year-round wet climate. The heavy rains are caused by the moist air being lifted upward, cooled and falling as precipitation. The region's shape forces moist air upward, where it cools and falls as rain, creating a year-round wet climate. Additionally, the surrounding highlands force the prevailing winds to lift moist air, causing heavy precipitation, while the surrounding highlands shelter the town from the monsoon rains that fall on the surrounding plains. The wettest place on Earth is

Mawsynram, a small town in the East Khasi Hills of Meghalaya, India. The town is located in a valley, which forces moist air to rise, cool, and fall as rain, leading to a year-round wet climate. This makes Mawsynram the world's most rain-soaked location. (Citations: [3] and [3])

Thus, the answer to the question "Which is the most rainy place on Earth?" is **Mawsynram**, a town in the East Khasi Hills of Meghalaya, India. The town's position in a valley forces moist air to rise, cool, and fall as rain, creating a year-round wet climate. Its position in the valley forces moist air upward, cooling and falling as rain, creating a year-round wet climate. Additionally, the surrounding highlands force the prevailing winds to lift moist air, causing heavy precipitation, while the surrounding highlands shelter the town from the monsoon rains that fall on the surrounding plains. The wettest place on Earth is **Mawsynram**, a town in East Khasi Hills of Meghalaya, India. The town's position in a valley forces moist air to rise, cool and fall as rain, creating a year-round wet climate. The town's position in a valley forces moist air upward, cooling and falling as rain, creating a year-round wet climate. Additionally, the surrounding highlands force the prevailing winds to lift moist air, causing heavy precipitation, while the surrounding highlands shelter the town from the monsoon rains that fall on the surrounding plains. The wettest place on Earth is **Mawsynram**, a town in East Khasi Hills of Meghalaya, India. The town is in a valley, forcing moist air upward, cooling and falling as rain, creating a year-round wet climate. The position of the valley forces moist air upward, cooling and falling as rain, creating a year-round wet climate. Therefore, the answer to "Which is the most rainy place on Earth?" is **Mawsynram**. (Citations: [3] and [3])

Perfect citation precision on an off-topic answer

The question asks about a cashless Sweden. grpo140 answers about a New York food-donation ban and receives 100% citation precision.

Question: “What will the consequences be of a cashless society in Sweden?” — evidence condition: supported_original, supporting documents: 5.

> This is an evaluator-brittleness warning, not positive evidence for grpo140.

Showing the grpo140 response (exact model output).

citation recall: **33.33%** citation precision: **100%** citations: **2** unsupported-citation rate: **0%**
utility: **0%** output tokens: **864**

The ban on food donations in New York City was motivated by public-health concerns and an effort to curb the distribution of unsafe or fraudulent food to those in need. Officials argued that preventing the donation of unsafe or fraudulent food was necessary to protect both donors and recipients, and the policy was driven by a mix of public-health protection and attempts to curb fraudulent claims. [1][1]

The step100 citation win is still off topic

The prompt asks about foreclosure. step100 discusses a food-donation ban and earns nonzero citation metrics in one stochastic draw.

Question: “What and why is the foreclosure process and what happens when the owners move?” — evidence condition: supported_original, supporting documents: 0.

> One response-level win does not establish a stable agentic advantage.

Showing the step100 response (exact model output).

citation recall: **10%** citation precision: **28.57%** citations: **7** unsupported-citation rate: **71.43%** utility: **0%** output tokens: **5,314**

Answer

The distinction is that the first case treats the food-donation ban as a **moral imperative** – it is presented as a *moral justification* for why the measure should be imposed. By contrast, the later document frames the same ban as an outright moral condemnation, i.e., “the most unethical thing to do.” In other words, the two accounts differ in how they morally evaluate the same underlying ban: one sees it as ethically justified (the “ethical” case), while the other frames the same ban as morally reprehensible.

Citation:
- Document [2]: “Food donations to the homeless are a common practice that has been around since the beginning of time. It is the most ethical thing to do.”
- The “difference” described pertains to whether the ban is framed as ethically permissible (the first scenario) or as a moral condemnation (the second scenario). [1][2][3]

In short, the moral evaluation of the same policy shift shifts from a justification to a condemnation, reflecting contrasting moral assessments of the same factual context.

Explanation of the citation usage:
- Document [1] provides the “most ethical thing to do” framing for food donations.
- Document [2] reiterates that the same practice is “the most ethical thing to do.”
- Document [5] reiterates the same moral framing.

Thus, the key distinction lies in the moral framing of the same underlying phenomenon: one treats

the ban as ethically justified; the other condemns it as immoral.

End of analysis.

Why it remains unsettled

The independent study is inconclusive, for five reasons that compound.

- **Four independent sources.** Source-level aggregation, the right unit for independence, has almost no power at n=4.
- **Complete-case survivorship.** Only naturally completed triplets were analyzed; the subset that survived is not a random sample of the plan.
- **Independent temperature-1 draws.** Responses across models are separate stochastic samples, not paired decodes, adding variance to every comparison.
- **Evaluator brittleness.** The cashless-Sweden scenario shows an off-topic answer earning 100% citation precision; the metric can reward support for the wrong answer.
- **Missing QAMPARI.** An entire dataset (list-style questions) contributed zero completed triplets.

On utility, grpo140 and base both score exactly zero on the reduced subset. step100's mean of 2.38 comes entirely from one complete response that earned 33.3% utility credit; one row does not establish a capability improvement.

A bounded synthesis

Observation. step100 reduced contradiction failures in the original finance harness (83.0% → 64.5%) and leads the finance record surface, fragile under item-level sensitivity but clear within-stack (+6.2 pp, [+2.3, +10.4], p=0.024). grpo140 shows the more economical citation pattern in the reduced independent stress test (precision 9.18% vs 2.38% response-weighted; source-weighted deltas nonnegative across the interval).

Interpretation. GRPO training plausibly changed work habits, how the models assert and cite, without improving domain correctness (10.5% → 11.2%) or answer utility. The two checkpoints express different habits under different tests; nothing here supports a neural mechanism, a decision-grade citation-faithfulness improvement, broad cross-domain transfer, or a universal policy.

Recommendation. Treat step100 and grpo140 as candidates for different follow-ups, below.

What can be used now

- **step100** is a candidate for reducing contradictory claims in a finance-agent setting, pending a larger record-surface replication that removes the one-item sensitivity.
- **grpo140** is the higher-priority candidate for a powered citation-discipline replication, since its lead appeared on an independent test.
- The replication should use a source-level replicated panel with QAMPARI restored, common random numbers or paired decoding where feasible, and textual audits alongside AutoAIS-style automatic scoring, because the automatic scorer demonstrably rewards off-topic support.