International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Beyond the Aggregate: Certifying Scope, Composition, and Robustness in Analytical Claims

Arindam Chakraborty[1], Saikat Chakraborty[2], Soham Marik[3]

[1,2,3]Independent Researcher, Texas, United States

Abstract

An aggregate that is correct can still be misleading, since its actual scope and the counterexample which would refute it can vary depending on the analysis decisions made. This paper presents an executable release protocol which maintains four questions separate from one another: is the aggregate correct, does it apply to every declared subgroup, what does the composition show and does any counterexample remain? This point is especially important: when dealing with a finite declared family, an exception under each specification is weaker than a single predicate that remains throughout; the protocol is able to distinguish between the two. The certificates include the supporting evidence, the materiality margins, the unresolved checks and the source bindings. When the pipeline is rerun starting from the raw data, all 39 taxi and retail partition diagnostics and the 15 taxi population comparisons are reproduced. The change in the taxi tip ratio ranges from −1.096 to +1.153 percentage points across the three named populations. A new retail matrix examines 58 predicates across three specifications for three measures; persistent counterexamples occur for all three cases under both the original and trimmed means, even though two of the originally chosen examples do not hold. The median checks are given separately because they alter the estimand. The contribution consists of the certificate contract and its implementation, supported by exploratory evidence, not a new decomposition, general LLM reliability, a causal explanation or any advantage over existing verification methods.

Keywords: analytical verification; robust counterexamples; evidence certificates; subgroup scope; specification sensitivity

1. Introduction

Correct arithmetic alone is not sufficient to guarantee that an analytical statement results in the right conclusion. For instance, a falling overall average does not mean that each subgroup had a decrease, nor does it show whether the change took place within the groups or was due to changes in the relative sizes of the groups. Verification procedures should make these two distinctions clear in any statement that is released.

The release protocol in question attaches each claim to a particular population and measure, checks the exact aggregate, examines the declared subgroup family and carries out a composition diagnostic. It is also extended in such a way that it can be determined whether the same counterexample remains when different analysis choices are made or whether each new search results in a different exception. This

difference is important, since in the case of the latter the situation may give an illusion of robustness without in fact providing a persistent explanation.

The idea behind the protocol was based on the taxi comparison, since that study showed sensitivity to both the way the calendar and the population were defined. A later analysis of sensitivity in the retail sector then spotted a similar problem: the two chosen witnesses were dependent on a small number of large sales lines. This observation was turned into a predicate-by-specification certificate and was then evaluated by means of exhaustive computation, all of this being carried out as part of post-review development and not as part of the originally registered design. The main contribution consists in providing a reproducible contract for qualified statements, not in inventing a method for searching for counterexamples or in carrying out a comparative accuracy assessment.

2. Related Work

Many systems have already addressed various aspects of this issue. AggChecker links numerical text summaries to relational data [2, 3]; Scrutinizer enables mixed-initiative checking [4]; CEDAR looks into verification costs [5]; and ClaimDB examines verification with respect to structured databases [1]. There is also another area of research focusing on cherry-picked generalizations, trendlines and Simpson reversals, which looks at aggregate statements that hide opposing subgroups [6, 7, 8, 9], while Slice Finder finds interpretable areas of poor model behaviour [10]. Our protocol draws on these verification and slicing concepts, making the difference between aggregate truth, scope and composition clear in the release artifact, although we have not yet established that this design performs better than any of those systems.

With regard to the accounting aspect, we make use of Kitagawa’s symmetric decomposition of the differences between rates [15], in this case applied to group-weighted means; Lipovetsky and Conklin have also proposed a related approach to the treatment of aggregation and Simpson’s paradox [13]. The arithmetic involved is by no means new: we are not asserting that it is. What is still open for discussion is the question of how a verifier should record and make public its result together with a correct aggregate and any applicable scope exception.

Sensitivity testing is no newer idea either: earlier work in computational fact-checking had already treated claims as parameterized queries and looked at how changes produce alterations in their conclusions [16]; specification curve analysis looks at the results when justified analytical choices are made [17]. The aspect we concentrate on more specifically is the consequence of the order of the quantifiers: a fixed predicate can only be described as persistent if it remains supported throughout the named specifications. We highlight any missing support and keep the changes to the estimator separate from population sensitivity.

There is also a considerable body of work concerning disclosure interventions, for instance, visualization guardrails have been assessed using reader experiments [18]. We have not carried out a similar kind of human evaluation of our certificate implementation; simply providing a more explicit artifact does not in itself prove that readers understand it better.

3. Release Protocol

3.1 Claim Binding and Supported Exceptions

A bound claim includes the population predicate, the measure, the aggregate, the periods, the direction, the allowed dimensions, the materiality margin and the quantifier. The implementation will only

function with structured claims of this type and does not make any effort to reliably extract them from free-text prose. If the direction of the aggregate is incorrect, then the claim is contradicted; if a binding is missing or invalid, it is regarded as insufficient evidence. As for causal language, it needs a design for identification that is outside of this protocol.

For a nonzero overall change Δ and a margin $\varepsilon > 0$, a directional exception is given by the condition $\text{sign}(\Delta) \times \Delta g \leq -\varepsilon$, it being necessary for both periods to each provide at least 1,000 non-null observations for the measure. This threshold is descriptive in character, since it is applied to an exact finite snapshot and not used as a test of statistical significance. The family together with its completed checks is always mentioned. A witness may be able to indicate a broad scope even before the exhaustive search has been completed; however, it is not possible for the lack of a witness in the case of an unfinished search to support a general claim about the family.

3.2 Complete Composition Accounting

Before it is finally released, the composition policy carries out all of the declared one-way partitions (six dimensions in the case of the taxi study and three in the case of retail), keeping every category for accounting purposes, including NULL and undefined values, since these remain as accounting levels rather than as substantive explanations. The share of each group is defined separately for each measure and for each period, based on the count of non-null measures.

The group shares are indicated by $p_0 g$ and $p_1 g$ and the group means by $\mu_0 g$ and $\mu_1 g$. When a group is observed in both periods, its within contribution is given by $W_g = \frac{(p_{0g} + p_{1g})(\mu_{1g} - \mu_{0g})}{2}$; its composition contribution is $M_g = \frac{(\mu_{0g} + \mu_{1g})(p_{1g} - p_{0g})}{2}$ [15]. A group that enters makes a contribution of $p_1 g \mu_1 g$ to a particular turnover term; one that exits makes a contribution of $-p_0 g \mu_0 g$. Never is an absent mean imputed. The total mean change is obtained by adding together the within, composition and turnover contributions.

A partition is classified as having a material composition if $|\Sigma g\ Mg|$ is greater than or equal to ε. The within-term is compared to the overall change in sign only in those cases where each of the contributing groups is present in both periods; small cells remain in the accounting because only scope witnesses need to satisfy the support gate. Components from overlapping partitions are never added together. Since per-category mix terms are uncentered they may cancel, the size of the magnitudes depending on where the outcome arises. What is described here is the accounting, not the unique causes identified.

The total shared-category mix figure itself can be affected by the origin of the outcome when groups begin to enter or leave. An addition of a constant a alters it by a $\Sigma\text{shared}(p_1 g - p_0 g)$, whereas turnover changes by the exact opposite amount, so that the reconstructed total remains the same. That is why we keep the natural origin of each measure rather than asserting that the origin is invariant in the case of incomplete shared support. This point is directly confirmed by a retail stress export that we have included: a shift by 100 modifies some of the values but does not trigger any materiality flag in these data.

3.3 Release States and Certificates

Table 1. Release Rules for a Bound Directional Claim. Composition Disclosure Can Accompany Either Overall Release or Scope Qualification.

Evidence	Release consequence
Invalid binding or incomplete evidence for requested scope	INSUFFICIENT_EVIDENCE; withhold the unsupported statement.

Overall direction is wrong	CONTRADICTED; do not release that direction.
Correct literal overall direction	OVERALL_ONLY; state the declared population and overall scope.
Correct overall direction and material composition	OVERALL_WITH_COMPOSITION; add the partition accounting disclosure.
Broad scope and a supported opposing slice	SCOPE_LIMITED; add WITH_COMPOSITION when applicable.
No supported exception after complete declared-family checking	No supported opposing change reaches the margin within the completed family; this does not prove literal universal direction.

There are three policies that determine release. The overall-only policy issues a literal aggregate and reduces any requested broad scope; the reversal policy includes complete one-way exception checks; and the composition policy includes the accounting disclosure. Regardless of which policy is applied, the resulting certificate contains the input claim, the population, the measure definition, the means, the margin, the support, the completed partitions, the evidence files and the release text. Performing a query that is identical in connection and view provides a consistency replay, but that is all it does: it cannot identify shared semantic errors or take the place of independent validation.

The updated implementation will not grant a positive composition-policy release if the mandatory pass is not complete, even if a scope witness has already been found. It directly rejects invalid policy or scope values as well as nonfinite deltas; it does not support explicitly bound causal requests; and the all-groups equality falls completely outside the scope of the directional verifier. In the case of a claim having the WITHIN_DECLARED_FAMILY label, this simply means that no supported opposing change reaches the margin within that family, not that all groups move in the same direction. With regard to C4, the caller determines completeness by referring to the observed diagnostic count and to the exact coverage of the six declared dimension names, then passes this value explicitly to release(); it is enough to remove just one of the required partitions to prevent the composition-policy release. Nevertheless, the function still relies on the evidence that is binding to the caller; it is not an authenticated proof boundary.

3.4 Counterexample Persistence Across Specifications

Each specification θ includes information on the population, the periods, the measure, the estimator, the support rule and the materiality margin. We define a finite family Θ, keeping the predicate identities s unchanged throughout that family; different populations are considered to be distinct analytical targets in their own right. Since median checks result in a change of the estimand, they are presented in a separate stress contract rather than being silently replaced by a mean claim.

When the total change $\Delta\theta$ is not zero, the opposition gap is given by $v\theta s = -\text{sign}(\Delta\theta)\Delta\theta s - \varepsilon$. A cell is considered to have passed if the gap is nonnegative and each period provides at least 1,000 non-null observations; if a cell is unsupported or the overall direction is not defined, it is marked as UNKNOWN, while if the evaluated gap is negative it is labeled FAIL. The shared-direction account goes an even farther step by requiring that the same nonzero aggregate direction holds throughout Θ.

Let $W\theta$ denote the set of passing predicate identities in the specification θ; all that is needed in order to say that "each specification contains an exception" is for each $W\theta$ to be nonempty, whereas in order to say that a counterexample is persistent it is necessary to have something stronger: namely, that the

intersection of all the Wθ be nonempty. Stated formally, ∀θ ∃s does not imply ∃s ∀θ. Although this distinction is logically elementary, it does affect which statement a certificate is actually able to support.

The implementation lists the entire matrix and, when considering the persistent predicates, chooses the one with the largest worst-case normalized gap, that is, the one with the greatest value of $min_\theta(\frac{v_{\theta s}}{\varepsilon})$; in the case of a tie it selects the predicate lexicographically. Moreover, it examines the distinct coverage patterns in order to find a minimum-cardinality predicate cover of the finite specification family: in other words, it carries out a set-cover calculation at the catalog level, not a minimum explanation or an outcome based on evidence rows. With respect to the current mean contracts based on real data, each of them is found to have a one-predicate cover.

The addition of specifications can only reduce the size of the persistent-witness intersection, never increase it; in order to achieve positive certification it is necessary to have a PASS for the same predicate in each of the required specifications; the absence of a check must never be regarded as a success. All of these properties, however, are subject to the condition that the bindings and measurements are correct from the start; on their own they do not guarantee semantic correctness, statistical confidence or the ability to generalize beyond the declared family.

The robustness certificates include the entire matrix, the original witness as it was fixed, the selected persistent witness, all the specification names, the hash of the source data and the hash of the canonical notebook source. They are provided as separate outputs together with the current composition certificate. However, the implementation has not yet carried out the fusion of all the certificates into a single report written in natural language or the validation of any arbitrary free-form analytical prose.

4. Data and Evaluation

4.1 Taxi Development Data

Our primary data come from the official NYC TLC yellow-taxi dataset for January 2024 and January 2025 [11]: 6,439,850 rows in total. January 1 landed on a Monday in 2024 and a Wednesday in 2025. That mismatch matters, because any weekday-by-hour comparison would otherwise set a holiday against an ordinary working day. We address it in the primary analysis by dropping January 1–2 from both years after validity rules are applied; that leaves 5,923,972 rows. A broader, inclusive valid view of 6,241,527 rows sits aside, reserved for sensitivity checks alone.

Most of this asymmetry sits in nonpositive fares: 38,257 records in 2024 against 145,419 in 2025. Only 28 and 43 records exceeded $500 in each year, respectively. Because this exclusion rule could shape the comparison, we report aggregates both with it applied and without it. Even so, direction stays negative across all five measures. Size is another matter: total amount alone falls by −$0.501 in the primary view, by −$0.582 once we switch to the inclusive valid view and by −$1.111 after the fare rule is dropped entirely.

A fixed sequence governs the validity rules. A pickup gets rejected if it falls outside the file's January; carries a nonpositive duration or one above four hours; sits outside 0–100 miles in distance; falls outside (0, $500] in fare; lands outside $0– $500 in total amount; records a tip-to-fare ratio outside 0– 100%; or shows a non-null passenger count outside 0–8. On its own, a NULL comparison in SQL never triggers these exclusions, because support gets checked separately for each measure. Exact predicates, along with their first-matching exclusion counts, are preserved in the notebook. The airport flag deserves a clarification too: it marks a positive recorded Airport_fee, not an actual geographic airport location.

The calendar exclusion entered after review, added by an analyst rather than flagged automatically by the protocol as a calendar defect. It mitigates a known mismatch, though full calendar comparability is not guaranteed. Every cleaning category stays in the analysis notebook, together with both the calendar and fare-rule sensitivities.

Dated March 18, 2025, the TLC dictionary assigns payment code 0 to Flex Fare and rate code 99 to Null/unknown [12]. For the historical substantive search, it allows payment codes 0–4, rate codes 1–6 and vendor codes 1, 2, 6 and 7; unknown and voided levels drop out, but only from the historical groupings that touch that specific dimension, never from the primary population as a whole. By contrast, the complete one-way pass keeps every level for accounting and category-level checks. Because unknown codes are never read as named operational groups, any taxi exception supported here traces back to a substantive payment or rate level. On its own, though, the dictionary says nothing about historical coding stability.

The margins went up exactly once, right after the first review then sat fixed for 0.5 minutes, $1, 0.25 miles, 1 percentage point and 0.05 passengers. Analysts chose these as materiality thresholds and none of them is an independently validated tolerance. Reclassifying the historical oracle at half, nominal and double margins produces five, five and four claims with opposing slices respectively. As for C4, it is simply the trip-level ratio recorded as 100 times tip_amount divided by fare_amount; because the TLC never records cash tips and it falls short of measuring all tipping.

Table 2. Taxi Primary-Population Means and Fixed Margins, with Changes Computed as Focal Minus Reference; C2 and C4 Depend Strongly on Population Eligibility.

Claim and measure	2024	2025	Change	Margin
C1 Duration (min)	14.854	14.573	-0.281	0.5
C2 Total amount ($)	27.222	26.720	-0.501	1
C3 Distance (mi)	3.184	3.062	-0.122	0.25
C4 Recorded tips (%)	20.461	19.364	-1.096	1
C5 Passengers	1.331	1.286	-0.046	0.05

4.2 Retail Transfer and Post Review Sensitivity

UCI Online Retail [14] supplies a second domain, comparing January–May against June–October 2011. Here the unit of analysis is a sales line, not an invoice or a customer. We retain positive quantity, positive unit price and invoices not marked as cancellations, while missing customer identifiers are left in place. Of 541,909 raw rows, 389,192 fall within the two windows and 9,048 fail these rules, leaving 380,144 lines: 162,474 reference and 217,670 focal. The original analysis applies no upper-value trimming. The windows themselves do not control for seasonality.

Quantity, unit price in pounds sterling and line amount (quantity multiplied by unit price) are evaluated across country, invoice weekday and invoice hour, with each margin set at 5% of its original reference mean. These rules were documented locally with a timestamp and hash prior to inspecting the first retail result, although they were not externally preregistered, as the taxi data had already influenced the protocol by that stage.

The fact that a reviewer had raised concerns about outliers led to an extra sensitivity analysis. For each measure, the original slice that is supported and which has the greatest opposing mean difference is chosen (when there is a tie, the one with the higher dimension and category is selected), then this particular case is then fixed. For each measure separately, any values that exceed the pooled 99th percentile are deleted, the overall direction is recalculated, keeping the original absolute margin is kept. The medians are also compared using the full, unaltered population, where a zero change in the overall median does not count as a directional exception according to the rule in place. In the results that are exported, there is a distinction made between keeping the fixed witness and finding a different witness after reexamining the family. These checks were introduced only after the original results had already been seen.

4.3 Execution and Cost Accounting

The same diagnostics, sensitivity analyses, release policies and consistency checks mentioned above are carried out by the main notebook: specifically covering 30 taxi and nine retail measure-by-partition queries. The accounting fixtures are used to test shared groups, the process of entering and exiting categories, NULL levels and the behaviour in different release states. The 39 partition reconstructions are each reconciled separately against the overall means obtained from independent queries. The timings that are reported include only the SQL components; the exclusions are clearly stated.

A previous exploratory study by JBCS that looked into the matter ranked the grouped searches using a 2% scout and managed to identify all five taxi exceptions after just one exact action, achieving a tie in the uniform median when using a nominal proxy but falling short on the measured query components. The full methods, the reserve ablation, the ranking diagnostic, the trace and the figure from that study are kept in the archived scheduler appendix. Although that search is still optional it cannot take the place of the mandatory one-way diagnostic pass, which is the reason why its stopping budget and 1.3-second query-component cost are not relevant as estimates of the total cost of the full release protocol.

4.4 Post Review Robustness Execution

The extension calculates the three historical taxi populations from the original files (those with a valid primary record, those with a primary record without a payment code 0 and strict documented codes) and then compares all 15 measure-population results with the historical export. In this case, direction envelopes employ a zero sign threshold, which must not be mistaken for the positive margins used in scope exceptions. We do not consider these populations to be interchangeable; the objective is to examine how a more general description relies on eligibility.

For retail, the original mean, the mean calculated after the values exceeding the measure-specific pooled 99th percentile have been removed and the original median are all assessed under three different lenses, with the original absolute measure margins and the support rules remaining unchanged throughout. The original mean and the trimmed mean together make up the primary two-specification contract, whereas the three-specification contract serves as an explicit cross-estimand stress test. The full catalog consists of 58 identities per measure, covering those categories which fail the support rule.

The notebook updated version uses Python 3.12.14, DuckDB 1.4.4, NumPy 2.3.5 and pandas 2.2.3, an environment which is different from the DuckDB 1.5.5 that was required; both the versions used and the checks carried out are recorded. The various additions made here have only been designed following a review of the previous results, so the certificates produced should be regarded as evidence from the development phase and not as the outcome of a test that was carried out without any alterations.

5. Results

5.1 Population Sensitivity and Automatic Composition

When the criteria for population eligibility are altered, two of the five main comparisons undergo a significant change. If the population is restricted to a conservative one, that is, payments 1–4, rates 1–6 and only those with documented vendors, then C4 rises from 21.271% to 22.424% (an increase of 1.153 percentage points), reversing the primary decrease of 1.096 points; at the same time, C2 drops by only $0.020 in this population as against $0.501 in the original population. The primary totals in Table 2 remain accurate for the rows they are associated with; the point that an unqualified statement regarding falling recorded tips and total amounts is found to be dependent on the population is the only thing that changes. All five of the strict-population measures still show opposing trends when their overall directions are recalculated. The supplementary population comparison gives every mean and change for all five measures.

In 2024, Flex Fare records account for 4.53% of the primary population and in 2025 they make up 13.99% as well; the average recorded tip-to-fare ratio is 8.01% and 2.57% respectively, which is not uniformly zero. If you exclude only Flex Fare, the total amount increases by $0.011 and the recorded-tip ratio increases by 1.045 percentage points; this shows the sensitivity associated with that one category alone, while the more stringent population criterion goes a step further by also limiting the rates and vendors. Together, the evidence supports an interpretation that is sensitive to the composition rather than assigning exclusive causality to Flex Fare or providing proof of a past coding change.

In the case of Flex Fare, zero recorded tips make up 62.85% of the non-null tip-ratio observations in 2024 (78,719 out of 125,243) and 88.13% in 2025 (389,345 out of 441,784). The average total amount in this category falls from $25.909 to $23.188 ($-2.721), whereas the average amount for those not in Flex remains almost unchanged, increasing from $27.284 to $27.295 ($+0.011). Although these observed differences help to explain the sensitivity to the choice of population, they do not prove that a coding change has taken place.

The sensitivity analyses which defined the exclusion of both the conservative-population group and Flex Fare were carried out after the review had been completed. Although the initial search had produced a valid exception for vendor-1 at the standard rate for C4, it had not picked up the explanatory composition pattern; that is the reason why the automatic pass being evaluated here was introduced. The fact that this method is able to recover the material accounting terms from these data serves as development evidence and not as independent rediscovery.

The automatic payment partition breaks down the C4 change of −1.096476 percentage points as −0.002473 due to within-categories and −1.094003 as a result of composition, with zero turnover; the composition figure exceeds the one-point limit. The rate partition also identifies composition as being the case (−1.630514 points), this time with a positive within-category amount (+0.534038). For C2, the payment composition accounts for −$0.172198 and the within-category change for −$0.328972; in this instance the composition term remains below its set $1 limit and so the protocol does not treat it as material. In other words, an interesting exclusion sensitivity does not take precedence over the stated threshold.

Table 3. The Two Largest Absolute C4 Payment-Category Mix Terms, with Shares Computed from Non-Null C4 Support; Terms Are Uncentered and Are Not Causal Importance Scores.

Payment category	2024 share	2025 share	Within (pp)	Mix (pp)
1 Credit card	79.71%	73.46%	+0.501373	-1.594458
0 Flex Fare	4.53%	13.99%	-0.504253	+0.500430

It is only by combining both categories that the share transfer becomes apparent: the credit-card share decreases while the Flex Fare share increases; both still appear in the full export together with the other payment codes. The fact that the Flex Fare mix term is positive and not centred does not indicate that its expansion caused the overall figure to rise; rather, the category terms depend on the origin of the chosen outcome and must be read together as part of a single partition, not in isolation.

5.2 Executed C4 Release Certificate

We deliberately construct C4 as an analyst-written statement saying that the overall recorded tip-to-fare ratio had decreased; we do not make the claim that every subgroup saw a decrease. This identical claim is handled by all three of the implemented policies without requiring a parser repair or an additional model call. The example is intentional in that it tests the disclosure branch which the discarded generation pilot had never reached, rather than attempting to determine how frequently a model would make such a claim initially.

Table 4. Executed Policy Comparison, Using the Same Analyst-Authored C4 Input Throughout.

Policy	Returned state	Evidence added to release
Overall-only	OVERALL_ONLY	Correct literal mean and population scope.
Reversal checks	OVERALL_ONLY	No universal scope was requested.
Composition checks	OVERALL_WITH_COMPOSITION	Material payment and rate composition disclosed.

Certificate C4-analyst-overall (executed output)

Status: OVERALL_WITH_COMPOSITION
Origin: analyst-authored demonstration; requested scope: overall
Population: valid-primary trips, January 3–31 in 2024 and 2025
Measure: 100 × recorded tip / fare; margin: 1 pp
Means: 20.460730% → 19.364254%; Δ = -1.096476 pp
Non-null support: 2,765,963 / 3,158,009
Diagnostics: 6/6 one-way partitions complete
Material composition: payment_type and RatecodeID
Consistency replay: passed; same connection and view
Released statement: For the declared population, mean recorded tip/fare ratio fell from 20.460730% to 19.364254%. This is an overall statement; it does not assert a decrease in every subgroup.

Disclosure: In the payment-type partition, the -1.096476 pp change decomposes into -0.002473 pp within-category and -1.094003 pp composition contributions. This is descriptive accounting, not causal attribution.

The machine-readable certificate together with both comparator certificates is contained in the file c4_policy_certificates.json. A supported exception may be enough to meet the scope requirement even if it is not the most informative account of aggregate change. Composition disclosure deals with a completely different question and therefore needs its own full evidence.

5.3 Retail Witnesses and Sensitivity

Table 5. Retail Overall Means, Changes and Original Margins.

Measure	Reference	Focal	Change	Margin
Quantity	10.792	11.058	+0.266	0.540
UnitPrice	3.998	3.767	-0.231	0.200
line_amount	19.948	20.461	+0.513	0.997

Table 6. Selected Original Opposing Mean Witnesses and Their Fixed-Predicate Sensitivities, with Deltas in the Measure's Original Units; Support Is Reference / Focal.

Measure and predicate	Support	Original Δ	Trimmed Δ	Median Δ
Quantity; hour = 10	15,534 / 20,141	-5.278	+0.573	+0.000
Unit price; France	2,573 / 3,532	+1.810	-0.096	-0.100
Line amount; Germany	2,855 / 4,262	-4.504	-2.160	-0.800

Both the witnesses relating to the quantity and the unit price fail when upper-tail trimming is applied; at a quantity cut-off of 100 items and a price cut-off of £16.95, the corresponding fixed-slice changes become +0.573 items and −£0.096. In contrast, the Germany line-amount witness does survive the £185.1596 cut-off, displaying a change of −£2.160 whereas the trimmed overall change is +£0.037 and the original margin is £0.997. The trimmed fixed-slice supports are 15,236/19,808, 2,472/3,410 and 2,816/4,248 respectively, each figure being based on its own trimmed population.

The fact that invoices are concentrated helps to account for this vulnerability. In the reference hour–10 quantity segment, one invoice with a line item of 74,215 items makes up 22.82% of the total quantity in that segment; simply removing that invoice causes the average to drop from 20.937 to 16.161 items. A comparable situation occurs in the focal France price segment, where a single line valued at £4,161.06 represents 21.01% of the total unit prices; when the invoice containing it is deleted the average falls from £5.609 to £4.432. Once again, these are observed concentration diagnostics and not estimates of independent invoice-level uncertainty.

Even after trimming, a new scan still identifies an opposing mean slice for each measure, but this alone does not prove that the original witnesses are robust. On the untrimmed medians, the total number changes by +1 item and there is still another opposing slice, although the change for the fixed hour-10 figure works out to zero. The unit-price medians decrease by £0.02, with no exception supported at the original margin. Both of the overall line-amount medians are £10.08, so the directional reversal cannot

be defined. In the end, none of the three original predicates that were selected qualifies as a median witness.

No single one of the nine original retail partition mix terms achieves the margin it has declared. Since both the country and the hour categories contain groups that were observed in just one window, the accounting keeps their turnover but does not carry out the complete-shared-support sign comparisons. Together, these results indicate that the protocol operates on a different domain and that mean-based scope findings can be very sensitive to large lines, not that operational reversals are robust or that the detection accuracy remains up to standard when compared with independently labelled defects.

5.4 Reconciliation and Mandatory Diagnostic Cost

The 39 partition results are consistent with the overall means when queried separately, using an absolute tolerance of 10^{-9} and a relative tolerance of 10^{-10}. The primary revised run had a maximum internal decomposition residual of 5.44e-15 and took 10.566 seconds to process the 30 taxi diagnostic SQL queries and 0.059 seconds for the nine retail queries. We also carried out an extra comparison on the same machine, running all nine notebook cells fresh under both DuckDB 1.4.4 and 1.5.5, with the same source data, code, Python dependencies and recorded database settings, including the use of four threads throughout. In all 15 of the exported CSV files, the largest absolute numerical difference is 2.90×10^{-11}; there are no categorical or Boolean differences. The content of the C4 and robustness certificates also agrees. This equality check does not include the timing fields and the output hashes; the evidence hashes from each run are verified separately. What this one-run-per-version comparison shows is that the results are numerically reproducible for these specific inputs: it does not indicate performance superiority nor independent semantic validation. It should be noted that internal decomposition residuals measure the reconstruction within each run, so their maximum values can be the same even if there are small differences in the individual components between the two runs.

The cost of the mandatory pass is greater than that of the optional historical search. None of the times associated with loading, Python accounting, sensitivity analyses or certificate construction are included in these SQL totals. The robustness extension keeps its own separate record of the elapsed time. Neither of these measurements represents the complete protocol latency and no efficiency advantage is asserted here.

5.5 Persistent Retail Counterexamples and Taxi Envelopes

All fifteen of the historical comparisons of taxi populations are within tolerance. In the three populations referred to, the C4 values vary between −1.096476 and +1.153184 percentage points and the C2 values range from −$0.501171 to +$0.011177, whereas the other three measures continue to have a negative direction. That slight positive value of C2 is not a substantial increase given that it is subject to a $1 margin; it only breaches the directional envelope which sets the zero threshold.

The retail matrix consists of 522 records, which have been obtained from three measures, three specifications and 58 predicate identities. With regard to each original-versus-trimmed mean contract, 22 predicates satisfy the support criterion in both specifications. There is one instance for quantity, one for unit price and two for line amount as regards the persistent counterexamples. Although the original examples for quantity and unit price do not prove to be persistent, the original Germany line-amount example does.

When it comes to quantity, Germany remains a persistent alternative: its average decreases by 1.019255 items in the original data and by 0.847874 afterwards, even though there are overall increases and a margin of 0.539600 items, with 2,855 out of 4,262 lines originally supporting this and 2,839 out of 4,239

after trimming. As for unit price, in invoice hour 17 the figure goes up by £0.411095 and £0.221754 despite the general trend of declines, with a margin of £0.199895 and support from 7,083 out of 12,207 and 7,008 out of 12,023; the trimmed counterpart, however, only clears the margin by £0.021859.

The line amount for Germany decreases by £4.503658 and £2.160161 when using the two methods, with the original witness being kept unchanged. The most severe normalized excesses above the margins amount to approximately 0.5713 for the Germany quantity, 0.1094 for the hour-17 price and 1.1658 for the Germany line amount, descriptive distances above the thresholds, not confidence levels. It must be noted that the selection from the completed matrix is made afterwards.

Germany remains a persistent example when the median stress specification is added, whereas there is no supported median exception for the unit price at the original margin. Since the equal line-amount medians leave the overall median direction indeterminate, a three-specification directional certificate cannot be issued, although neither of these outcomes contradicts the original arithmetic mean.

The witness pattern which involves only rotation can be shown only through the use of a logical fixture, not through these actual mean contracts: all three of the actual mean contracts allow for exactly one persistent predicate. A total of eleven targeted fixtures are carried out, these covering incomplete evidence, nonfinite results, unsupported equality and checks relating to the order of quantifiers. Another loop also reconstructs the 39 component results from the exported group statistics, thereby independently verifying the arithmetic with respect to the grouping routine, although it still uses the same source bindings and so does not constitute an independent semantic validation.

6. Discussion and Limitations

The certificate draws a distinction between two points. On the one hand, the truth of the aggregate does not determine the scope or the composition; on the other hand, the fact that exceptions are repeatedly found by no means proves that one explanation continues to hold. The original retail sensitivity accurately picked out the fragile cases. The full matrix goes even further by identifying alternatives that survive the named mean checks; it just does not establish that those alternatives would persist in another defensible family or in a future dataset.

The evidence base includes taxi development data and a single retail dataset, together with five and three constructed measure comparisons respectively, all of which are now being used to inform the design. Nothing new has been carried out in the way of an independent test dataset, a naturally occurring claim benchmark, a model-generated-claim experiment or a reader study for the robustness extension. The earlier LLM generations have been kept archived merely as an unsuccessful pilot. The nature of the revised contribution is therefore an implementation and descriptive case study; the extra value it has over simple exhaustive verification has still got to be assessed.

The choice of thresholds, the scales used to determine outcomes, the way families are selected and the criteria for population eligibility all have an impact on release states. The present composition threshold relies on an absolute claim margin and thus may fail to indicate a significant relative change, even though line-level support does not ensure a large number of separate invoices. Additionally, unknown codes and incomplete historical metadata both place limitations on interpretation. As for same-view replay, it checks consistency rather than the independent correctness of the results. Moreover, precise calculations based on finite data do not prove future persistence or statistical significance.

7. Conclusion

The protocol that has been implemented deals with literal aggregate truth, scope exceptions, composition disclosure and counterexample persistence. While the complete retail matrix identifies persistent alternatives for two initially fragile examples, the taxi envelopes show those directions which depend on eligibility - results that provide a reason for adopting a more precise certificate contract, even though they do not involve a new decomposition, a better search algorithm, greater LLM reliability or a causal explanation. What still remains to be done is to have an independently specified evaluation in order to determine whether the certificates in fact improve analytical reports.

8. Data and Code Availability

The executed notebook as revised includes the 39 partition checks, the full component tables, the retail sensitivities, the invoice concentration diagnostics, three C4 certificates, 15 recomputed taxi population comparisons and a retail witness matrix with 522 rows. The robustness_outputs directory contains six contract certificates, direction envelopes, logical fixtures, source bindings and the execution audit. The supplement keeps the original manuscript together with its full scheduler and unsuccessful-generation appendices; the manuscript currently in use has those historical studies outside the main argument. Instead of redistributing them, the raw source files are downloaded and throughout the process pinned inputs and the actual versions of the dependencies are recorded. Finally, the live_engine_comparison supplement contains both the version-specific notebooks that have been freshly executed, the execution records, the outputs and a reproducible comparison script.

9. References

1. M. Theologitis, P. P. S. Dammu, C. Shah, D. Suciu, “ClaimDB: A Fact Verification Benchmark over Large Structured Data,” Proceedings of ACL, pp. 34428–34451, 2026. https://doi.org/10.18653/v1/2026.acl-long.1589
2. S. Jo et al., “Verifying Text Summaries of Relational Data Sets,” Proceedings of ACM SIGMOD, pp. 299–316, 2019.
3. S. Jo et al., “AggChecker: A Fact-Checking System for Text Summaries of Relational Data Sets,” PVLDB 12(12), pp. 1938–1941, 2019.
4. G. Karagiannis, M. Saeed, P. Papotti, I. Trummer, “Scrutinizer: A Mixed-Initiative Approach to Large-Scale, Data-Driven Claim Verification,” PVLDB 13(11), pp. 2508–2521, 2020.
5. T. Jayasekara, I. Trummer, “CEDAR: A System for Cost-Efficient Data-Driven Claim Verification,” PVLDB 18(11), pp. 4492–4504, 2025. https://doi.org/10.14778/3749646.3749708
6. Y. Lin et al., “On Detecting Cherry-Picked Generalizations,” PVLDB 15(1), pp. 59–71, 2022.
7. Y. Lin et al., “OREO: Detection of Cherry-Picked Generalizations,” PVLDB 15(12), pp. 3570–3573, 2022.
8. A. Asudeh et al., “On Detecting Cherry-Picked Trendlines,” PVLDB 13(6), pp. 939–952, 2020.
9. N. Alipourfard, P. G. Fennell, K. Lerman, “Can You Trust the Trend? Discovering Simpson’s Paradoxes in Social Data,” Proceedings of WSDM, 2018.
10. Y. Chung et al., “Slice Finder: Automated Data Slicing for Model Validation,” Proceedings of IEEE ICDE, 2019.
11. New York City Taxi and Limousine Commission, TLC Trip Record Data, accessed September 12, 2026. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

12. New York City Taxi and Limousine Commission, Data Dictionary — Yellow Taxi Trip Records, March 18, 2025, accessed September 12, 2026. https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf
13. S. Lipovetsky, W. M. Conklin, "Data aggregation and Simpson's paradox gauged by index numbers," European Journal of Operational Research 172(1), pp. 334–351, 2006. https://doi.org/10.1016/j.ejor.2004.10.005
14. D. Chen, Online Retail [Dataset], UCI Machine Learning Repository, 2015, License: CC BY 4.0. https://doi.org/10.24432/C5BW33
15. E. M. Kitagawa, "Components of a Difference Between Two Rates," Journal of the American Statistical Association 50(272), pp. 1168–1194, 1955. https://doi.org/10.1080/01621459.1955.10501299
16. Y. Wu, P. K. Agarwal, C. Li, J. Yang, C. Yu, "Toward Computational Fact-Checking," PVLDB 7(7), pp. 589–600, 2014. https://www.vldb.org/pvldb/vol7/p589-wu.pdf
17. U. Simonsohn, J. P. Simmons, L. D. Nelson, "Specification curve analysis," Nature Human Behaviour 4, pp. 1208–1214, 2020. https://doi.org/10.1038/s41562-020-0912-z
18. M. Lisnic, Z. Cutler, M. Kogan, A. Lex, "Visualization Guardrails: Designing Interventions Against Cherry-Picking in Interactive Data Explorers," CHI, 2025. https://doi.org/10.1145/3706598.3713385