Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Structured Data Verification: When Facts Need a Query

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-31 · structured data verification · database claim verification · fact verification · claimdb · text-to-sql evaluation · data provenance · abstention · reproducibility · data verification vs validation

Original article: https://reliablecontext.com/articles/structured-data-verification-evaluation



In this article

1. Executive Summary
2. Introduction and Background
3. The ClaimDB Evidence Regime
4. A Worked Synthetic Witness
5. Boundary Changes: A Perturbation Matrix
6. Separating Four Quantities
7. Comparison with Passage, Table and Time-Series Benchmarks
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

Structured data verification is the task of deciding whether a claim about a database is entailed, contradicted, or undecidable from that database. The central methodological point of this report is that the evidence object is not a passage. It is a bundle: a frozen snapshot, a schema, an executable query, the result with its intermediate counts, and a verdict mapped from that result. A prose passage cannot establish an aggregate over millions of rows, and a verdict without the bundle cannot be independently checked.

The most direct test bed is **ClaimDB** (Theologitis, Dammu, Shah and Suciu, 2026). It pairs **53,368** claims with databases from BIRD (aclanthology.org), covers **80** unique real-life databases (aclanthology.org), and averages 11.3 tables and 4.6 million records per claim (arxiv.org). The authors convert that evidence to roughly 110 million tokens per claim (aclanthology.org), so reading it is not an option. In their SQL-tool agent setting on the public test set, **17 of 30** models stay below 55% accuracy and macro-F1 (aclanthology.org), and both closed and open models struggle with abstention (aclanthology.org). These are paper-reported results, not independent reproductions.

Three scope limits shape what those numbers justify. The labels hold only for a fixed snapshot (arxiv.org). Claims are written by a language model from a question and a JSON answer, and the SQL is then discarded (aclanthology.org), so the released records do not carry the executable witness. Upstream annotation errors in BIRD propagate to ClaimDB (aclanthology.org), so counts can change.

A synthetic worked example (labeled throughout, built in SQLite) shows how fragile an aggregate verdict is. A claim that at least 60% of eligible orders exceeded $100 is contradicted at **82 of 137** (59.85%), one qualifying row short of the threshold. Of eight non-equivalent query variants, **eight** changed the numerator or denominator but only **two** flipped the verdict, while three equivalent rewrites changed nothing. Verdict agreement hides most scope errors; counts expose them.

The practical recommendation is to score four things separately: answer accuracy, abstention, query validity, and evidence reproducibility, the last with an explicit numerator, denominator and Wilson interval (doi.org). In a synthetic audit, 34 of 40 witnesses reproducing gives a 95% interval of 70.9% to 92.9%. Executable evidence

is valid only under the declared snapshot and semantics, and it says nothing about whether the database is complete or true.

Introduction and Background

Fact verification research grew up on passages. The FEVER (Fact Extraction and VERification) dataset contains **185,445** claims generated by altering Wikipedia sentences (aclanthology.org), and its evidence is a sentence a reader can inspect in seconds. Claims over databases break that assumption. The evidence for "most orders in the northern region exceeded $100 in 2025" is the set of rows satisfying an eligibility predicate, the join that attaches region to order, the treatment of missing amounts, and the arithmetic that turns counts into a proportion.

This report is aimed at fact-verification and knowledge-fusion researchers who must decide whether a benchmark or verifier can justify a claim computed from large, joined tables, and when it must abstain. It is a research-methods article. It does not evaluate database products.

Four kinds of database-grounded claim

Each kind demands different evidence.

- **Row lookup:** one entity and one attribute ("order 1042 was paid"). Evidence is a single row and a key.
- **Join:** attributes held in different tables ("customers in the North region placed order 1042"). An inner join emits one output row for each matching row in the other table (www.postgresql.org), so join semantics matter.
- **Aggregation:** a quantifier over a set ("most", "at least 60%", "the average exceeded"). Evidence is the full eligible set, the aggregate function, and its treatment of nulls and duplicates.
- **Out-of-schema question:** a concept the schema does not represent ("most orders were paid by credit card" with no payment column). The correct verdict is abstention.

ClaimDB makes the last category explicit. Its not-enough-info claims are based on concepts not represented anywhere in the schema (arxiv.org), express opinions or value judgments that cannot be objectively verified, or are counterfactual, and all are meant to be definitely unanswerable from the database at hand (arxiv.org).

Verification versus validation

The terms are used loosely, so precision helps. The NIST (National Institute of Standards and Technology) glossary defines verification as confirmation, through objective evidence, that specified requirements have been fulfilled (csrc.nist.gov), and validation as confirmation that the requirements for a specific intended use are fulfilled (csrc.nist.gov). US Environmental Protection Agency guidance applies the same line to data: verification evaluates completeness and correctness against stated requirements (www.epa.gov), while validation extends the evaluation to the analytical quality of the data set (www.epa.gov). The guidance cautions that organizations define the terms differently (www.epa.gov).

For claim checking, verifying a claim asks whether the stated requirement, the claim, is met by the data as it stands. Validating the database asks whether the data are fit for use. An executable query can do the first. It cannot do the second, and a verdict never certifies the database itself.

The ClaimDB Evidence Regime

Construction

ClaimDB starts from BIRD, which supplies about 11,000 natural-language-to-SQL examples. After filtering by query structure and result size, roughly 6,500 pairs remain. Each SQL query is executed for an exact answer, and the paper states that the SQL is then no longer needed and can be discarded (aclanthology.org). A language model (gpt-5) is prompted to write claims entailed or contradicted by the answer (arxiv.org). The totals are 12,855 entailed, 16,529 contradicted and 23,984 not-enough-info claims (arxiv.org).

Two consequences follow for anyone treating a ClaimDB label as a witness. First, a label derives from a question-answer pair produced by a gold SQL query, not from a recorded query attached to the claim; the dataset card lists claim identifier, BIRD identifier, database name, label, claim, extra information and category as fields, and no query field (huggingface.co). Second, a model-phrased claim may drift from the semantics of the gold query, for example on ties or nulls. The available documents do not publish a measured drift rate.

Scope: what "80 databases" covers

The abstract states that ClaimDB consists of 80 unique databases. That count describes the benchmark as a whole, not the test set. The paper divides the data into three disjoint splits (arxiv.org): training, a public test set and a private test set whose labels are hidden. Both test sets contain approximately 1,000 examples (aclanthology.org), fully balanced with roughly 333 examples per label, and the test claims come from the dev split of BIRD. The training split keeps the full spectrum of not-enough-info claims (arxiv.org). The dataset card lists 999 claims for testing with hidden labels (huggingface.co), an archive of 80 databases (training) (huggingface.co) and an archive of 11 databases (test) (huggingface.co). The documents consulted do not state how the 11 test databases relate to the 80, so the public test should never be described as "over 80 databases".

Every claim is paired with exactly one large, multi-table database (arxiv.org). The paper's introduction and conclusion round the averages to 11 tables and 4.5 million records, slightly different from the 11.3 tables and 4.6 million of Section 3. Version 1 of the preprint also reported that none of the models exceeded 83% accuracy (arxiv.org), a statement the abstract of the revised version no longer carries. These are small differences, but they show why the version of every cited statistic should be recorded: the arXiv record shows the paper was last revised in April 2026.

Model conditions and reported results

The headline experiment gives each agent a single tool, the ability to execute arbitrary SQL queries (arxiv.org), with a limit of 20 tool calls per claim. The study evaluates 30 proprietary and open-source models. The exact condition behind the often-quoted figure is that 17 of 30 stay below 55% accuracy and macro-F1 (arxiv.org). The best model is gpt-5-mini.

A no-tool control is informative. Without database access, the best model is near random, with a macro-F1 of 0.253 and accuracy of 0.367 (arxiv.org), which supports reading the benchmark as rewarding computation rather than memorized knowledge. On abstention, the failure modes differ by model. The two strongest

proprietary models almost never predict not-enough-info for entailed or contradicted claims (arxiv.org), whereas one open model family predicts it roughly half the time (arxiv.org). A single accuracy figure cannot show that split.

What the reported results justify

The authors state their own limits. Labels are assessed against the fixed database snapshot (arxiv.org). The evaluation uses agents equipped with a single tool, the ability to execute SQL queries, so results partly reflect SQL skill. Annotation errors in BIRD propagate to ClaimDB, as noted above. The dataset is released under the same license as BIRD (arxiv.org); the license terms should be re-read before any redistribution, because this report did not audit them beyond the authors' statement.

Three bounded conclusions follow. A reported accuracy supports a statement about performance under that tool, prompt and snapshot. It does not support a statement about organizational data with different schemas, null conventions or duplicate rates. And it says nothing about whether the labels would survive re-derivation by a second team, which is the question the next sections formalize.

A Worked Synthetic Witness

All data in this section are **synthetic**: generated with a fixed random seed and executed in SQLite to illustrate method, not to measure any real system. The database has three tables: **customers** (200 rows), **orders** (1,203 rows) and **contacts** (353 rows, one to three per customer).

Step 1: normalize the claim

Claim C1: "At least 60% of paid orders placed by North-region customers in 2025 exceeded $100." Normalization makes every quantifier and implicit scope explicit.

- **Quantifier:** "at least 60%" becomes successes divided by eligible rows is at least 0.60.
- **Population:** orders joined to customers where region equals North.
- **Eligibility predicate:** status paid; timestamp on or after 2025-01-01 and strictly before 2026-01-01; amount not null.
- **Success predicate:** amount strictly greater than 100, in recorded US dollars.

The null rule is a declared choice. SQL comparisons with a null operand yield null rather than true or false (www.postgresql.org), and SQL uses three-valued logic (www.postgresql.org), so a filter on a nullable amount silently drops those rows. Excluding them from the denominator is a semantic decision the witness must state.

Step 2: freeze the snapshot

Serialize the database file and record its checksum. For an on-disk SQLite database, serialization is just a copy of the disk file (www.sqlite.org). In a server database, the equivalent is a consistent export, and PostgreSQL's dump tool makes consistent exports even while the database is in concurrent use (www.postgresql.org). Isolation matters too: under the default Read Committed mode, two commands in one

transaction can see different data (www.postgresql.org), so a witness that runs several queries must pin one snapshot. SQLite in write-ahead-log mode gives a reader an unchanging snapshot of the database from when its transaction started (www.sqlite.org).

Step 3: execute and inspect intermediate counts

The normalized query (whitespace collapsed) hashes with SHA-256 to a value beginning 076ddfcfe737. Intermediate counts, which any reviewer can recompute:

- North-region customers: **50**
- Their orders, all statuses and years: **307**; of those, paid: **222**; of those, placed in 2025: **151**
- Of the 151, amount null (excluded): **14**
- Eligible denominator: **137**; eligible orders over $100: **82**
- Proportion: 82 / 137 = **59.85%**

The threshold needs ceil(0.60 × 137) = 83 successes. The claim misses by exactly one order, so CONTRADICTED is correct under the declared semantics, but a one-row scope error could flip it.

Step 4: map the result to the three-way verdict

- **Empty eligible set:** SQLite notes that the SQL standard requires sum() of no rows to be NULL (www.sqlite.org), so an empty population needs an explicit rule (here, NOT ENOUGH INFO) rather than a silent zero.
- **ENTAILED:** the query is valid, the eligible set is non-empty, and the proportion meets the rule.
- **CONTRADICTED:** the query is valid, the eligible set is non-empty, and the proportion fails the rule.
- **NOT ENOUGH INFO:** the claim concerns a concept the schema lacks, is subjective or counterfactual, or cannot be expressed as a valid query over the snapshot.

Table 1 records three synthetic witnesses in a claim-to-query format. C2 is out of schema and C3 is subjective.

Claim	Source database / version	Tables and rows used	Eligibility predicate	Query hash	Output	Verdict	Reviewer
C1: at least 60% of paid 2025 North orders exceeded $100	synthetic_orders, snapshot S1 (file checksum recorded)	customers (50 North of 200), orders (151 North, paid, 2025)	status paid; half-open 2025 date range; amount not null	076ddfcfe737	82 of 137 = 59.85 %	CONTRADICTED	Synthetic reviewer R1
C2: most North orders were paid by credit card	synthetic_orders, snapshot S1	none: no payment-method column	not expressible	none (no valid query)	not applicable	NOT ENOUGH INFO (out of schema)	Synthetic reviewer R1
C3: North customers are happier than South customers	synthetic_orders, snapshot S1	none: no satisfaction measure	not expressible	none (subjective)	not applicable	NOT ENOUGH INFO (subjective)	Synthetic reviewer R2

Two of the three rows have no query at all. For those rows, the evidence object is the schema plus a recorded reason the claim cannot be expressed. A verifier that returns NOT ENOUGH INFO without a reason, or invents a query for a column that does not exist, has not produced evidence. This parallels FEVER's pairing of label and evidence, in which the best pipeline reaches only 31.87% accuracy even when handed the correct evidence (aclanthology.org).

Boundary Changes: A Perturbation Matrix

A witness is only as good as its resistance to small, plausible changes in scope. *Table 2* applies eleven synthetic changes to the base query for C1 (threshold 60%, base 82 of 137).

ID	Change (synthetic)	Equivalent?	Numerator / denominator	Share	Verdict	Counts differ?
Base	Declared semantics	n/a	82 / 137	59.85%	CONTRADICTED	n/a
P1	Null amounts kept in denominator	No	82 / 151	54.30%	CONTRADICTED	Yes
P2	Extra join to contacts (fan-out)	No	124 / 226	54.87%	CONTRADICTED	Yes
P3	Date cutoff written as <= '2025-12-31'	No	81 / 136	59.56%	CONTRADICTED	Yes
P4	Status filter dropped	No	112 / 186	60.22%	**ENTAILED**	Yes
P5	Region filter dropped	No	286 / 567	50.44%	CONTRADICTED	Yes
P6	Region via IN subquery	Yes	82 / 137	59.85%	CONTRADICTED	No
P7	Region via EXISTS, year via substring	Yes	82 / 137	59.85%	CONTRADICTED	No
P8	Threshold written as >= 100	Yes (no amount equals 100)	82 / 137	59.85%	CONTRADICTED	No
P9	Five exact duplicate over-$100 orders added	No	87 / 142	61.27%	**ENTAILED**	Yes
P9b	Five duplicate at-or-under-$100 orders added	No	82 / 142	57.75%	CONTRADICTED	Yes
P10	Entity split: three of six qualifying orders of one customer moved to a second identifier in another region	No	80 / 134	59.70%	CONTRADICTED	Yes

Read the matrix by column. Of the eight non-equivalent changes (P1 to P5, P9, P9b, P10), **eight** altered the counts but only **two** (P4 and P9) altered the verdict. A reviewer comparing verdicts alone would have accepted six wrong witnesses. The three equivalent rewrites (P6 to P8) left counts and verdict unchanged, as they should, although P8 is equivalent only because no synthetic amount equals exactly 100.00, a data property a witness should state.

Why each change matters

- **Nulls (P1):** count(*) counts input rows, while count(expression) counts only rows where the value is not null (www.postgresql.org), and avg() averages only non-null inputs (www.postgresql.org). Different forms give different denominators from the same data.
- **Joins (P2):** a one-to-many join repeats the "one" side, and listing several tables forms their Cartesian product before conditions apply (www.postgresql.org). The denominator inflated from 137 to 226 with no change in the claim, because each qualifying order is repeated once per matching contact row.
- **Date cutoffs (P3):** the table holds an order at 23:30 on 31 December 2025 and one at 00:15 on 1 January 2026. In this SQLite database the timestamps are stored as text, so the cutoff <= '2025-12-31' is a string comparison that sorts '2025-12-31 23:30' after '2025-12-31' and drops that order. A related hazard in PostgreSQL: it silently ignores a time zone indication on a literal of type timestamp without time zone (www.postgresql.org), so a local-time cutoff is not a stable specification there.
- **Duplicates (P9, P9b):** SELECT keeps duplicates by default (www.postgresql.org), and SQLite's count(DISTINCT X) counts distinct values instead of the total number of non-null values (www.sqlite.org).
- **Entity identifiers (P10):** a split identifier moves rows between groups. OWL does not assume that different names denote different individuals (www.w3.org), whereas databases usually behave as if keys were unique.
- **Schema omissions:** under the closed-world assumption, the negation of an unprovable literal is assumed true (www.cs.ubc.ca), and databases usually treat an absent fact as false (www.w3.org). Applied to an out-of-schema claim, that reading produces an answer where abstention is correct.
- **Alternate plans (P6, P7):** the PostgreSQL planner is trusted to give every correct plan the same row set (www.postgresql.org). But equivalence of two arbitrary SQL queries is undecidable (arxiv.org), and an equivalence-checking system notes that, because of this, some cases yield neither proof nor counterexample (www.cidrdb.org). Plan choice can also vary because planner statistics are random samples rather than exact (www.postgresql.org). Rewrites are therefore tested by execution, which is evidence for those inputs only.

Row order and numeric type matter as well. Without ORDER BY, rows are returned in whatever order the system finds fastest (www.postgresql.org), so a claim about "the top customer" needs a declared tie-break rule, and comparing two floating-point values for equality might not always work as expected (www.postgresql.org), which matters for threshold claims such as P8. The execution-based text-to-SQL literature met the same issue. Zhong and colleagues found that the Spider metric had a 2.5% false negative rate on average and 8.1% in the worst case (aclanthology.org), and proposed comparing denotations on a set of databases to reduce false positives (aclanthology.org). The equivalent practice for fact verification is to run a witness on perturbed copies of the snapshot, as this section does by hand.

Separating Four Quantities

One accuracy number conflates at least four properties. Each needs its own denominator.

Answer accuracy

Accuracy is the share of claims whose predicted verdict equals the gold verdict. With three labels, macro-averaging takes an unweighted mean over labels and does not take label imbalance into account (scikit-learn.org). Balanced accuracy, the macro-average of recall, avoids inflated estimates on imbalanced data (scikit-learn.org).

Abstention behavior

Selective classification, also known as the reject option (proceedings.neurips.cc), lets a user set a desired risk level (proceedings.neurips.cc). For language models, abstention is increasingly recognized for its potential to mitigate hallucinations and enhance safety (direct.mit.edu). ClaimDB's own split, between models that rarely abstain and a model that abstains about half the time on answerable claims, fits that picture. Abstention should be reported as coverage, selective risk, abstention precision and abstention recall, not folded into accuracy.

Query validity

Validity asks whether the program that produced the evidence is correct under the declared semantics. In text-to-SQL, execution accuracy counts a prediction as correct when the executed results of the predicted and ground-truth queries are identical (arxiv.org), which is agreement on one database state. The perturbation matrix shows that agreement on one state can mask a different eligibility rule. Validity needs the declared predicate, an intermediate-count audit, and perturbation tests, and it needs the gold query to be right.

Evidence reproducibility

ACM's artifact badging defines reproducibility as an independent group obtaining the same result using the author's own artifacts (www.acm.org), and replicability as obtaining it with independently developed artifacts (www.acm.org). For a witness, the artifacts are the snapshot, the query and the semantics note. The Research Data Alliance recommendations for dynamic data citation provide a template: query timestamping (www.rd-alliance.org), a persistent identifier for the query (www.rd-alliance.org), and result-set verification by checksum (www.rd-alliance.org). FAIR principle R1.2 adds that data should be associated with detailed provenance (www.nature.com).

Provenance theory supplies the vocabulary. "Why" provenance is the source data that influenced the existence of the output (dl.acm.org), and "where" provenance is the location in the source databases from which data was extracted (dl.acm.org). A witness is a subset of records sufficient to produce an output record (homepages.inf.ed.ac.uk). For an aggregate claim, the why-provenance is the eligible set of 137 rows, not the single number 82. The W3C PROV model defines provenance as information about the entities, activities and people involved in producing a piece of data (www.w3.org), and a derivation as a transformation of an entity into another (www.w3.org). Here the snapshot and result are entities, the execution is an activity, and the reviewer is an agent.

Comparison with Passage, Table and Time-Series Benchmarks

Benchmarks differ in what they let a second researcher re-check. *Table 3* places ClaimDB beside its neighbors.

Regime	Evidence object	Label set	Scale (as reported)	What a second researcher can re-check
FEVER (2018)	Wikipedia sentences	Supported, Refuted, NotEnoughInfo	185,445 claims (aclanthology.org)	Read the cited sentences; label and evidence scored together
FEVEROUS (2021)	Sentences and/or table cells (datasets-benchmarks-proceedings.neurips.cc)	Supports, refutes, not enough information	87,026 verified claims (datasets-benchmarks-proceedings.neurips.cc)	Cell-level evidence; a baseline gets both evidence and verdict right for 18% of claims (datasets-benchmarks-proceedings.neurips.cc)
TabFact (2020)	One Wikipedia table per statement (openreview.net)	Entailed or refuted only (openreview.net)	118k statements	Small tables; programs can be executed against them (openreview.net)
TSVer (2025)	Time frames over time series (aclanthology.org)	Verdict labels plus a separate justification (aclanthology.org)	304 claims (aclanthology.org)	Chosen date range can change the verdict
ClaimDB (2026)	A whole multi-table database, queried by SQL	Entailed, contradicted, not enough info	53,368 claims; released with an 80-database training archive and a separate 11-database test archive (overlap with the 80 not stated)	Rerun a query on the snapshot; released records do not include the query

Three observations follow. First, passage evidence is inspectable by reading, and FEVER's scoring couples label with evidence. Second, TSVer is the closest structured relative, but its evidence is time frames over curated time series, and the paper's baseline loads the specified slices rather than executing a query. Third, ClaimDB shifts the burden from reading to computing, which is what makes the query the evidence.

Transfer limits

Program-aided table verification shows how far computation can carry a verdict on small tables: Binder is a training-free neural-symbolic framework that maps the task input to a program (openreview.net). Those results concern tables small enough to show in a prompt, and they do not transfer to millions of rows, joins and null conventions without the checks above. Table-grounded verification is hard even before scale: on SciTab, all models except GPT-4 performed barely above random guessing (aclanthology.org). At the other end, AVeriTeC annotates claims with question-answer pairs supported by evidence available online (proceedings.neurips.cc).

Open-world questions remain unsolved. A snapshot is a closed world by construction, whereas a real claim ("the company never reported a loss") may concern facts outside it. In truth discovery, the core difficulty is that conflicts usually exist among collected multi-source information (dl.acm.org), and knowledge-based trust treats

a source with few false facts as trustworthy (www.vldb.org). Those fusion methods quantify uncertainty across sources. A single-database verdict does not, so ENTAILED implicitly assumes the snapshot is the whole truth.

Data Analysis and Evidence

Literature-reported figures

Reported figures most relevant to a methods decision, with their conditions attached:

- **Scale and labels:** 53,368 claims; 12,855 entailed, 16,529 contradicted and 23,984 not-enough-info.
- **Test design:** two test sets of about 1,000 claims each, fully balanced (aclanthology.org); the hidden-label set has 999 claims (huggingface.co).
- **Accuracy:** 17 of 30 models below 55% accuracy and macro-F1 in the SQL-tool setting (arxiv.org).
- **No-tool control:** accuracy 0.367 and macro-F1 0.253 for the best model (arxiv.org).
- **Evidence size:** roughly 110 million tokens per claim.

Two practices from evaluation statistics apply. Clustered, non-independent questions make a score look more precise than it is, and ClaimDB generates between one and three claims per question-answer pair (arxiv.org), with many claims sharing a database, so independence is doubtful. Card and colleagues warn that underpowered experiments make it difficult to discern statistical noise from real differences (aclanthology.org). Rankings of nearby models on a 1,000-claim test set deserve interval estimates, and the papers consulted do not make that comparison.

Synthetic audit: verdict and abstention matrices

Table 4 shows a **synthetic** audit of 120 claims (40 per gold label) with invented counts that illustrate the arithmetic. Rows are gold labels, columns are predictions. E is entailed, C contradicted, N not enough info.

Gold \ Predicted	E	C	N	Row total
E	30	4	6	40
C	5	28	7	40
N	9	8	23	40
Column total	44	40	36	120

Collapsing to answered versus abstained gives the abstention view: 84 predicted answers (E or C), of which 58 are correct, and 36 abstentions, of which 23 fall on truly unanswerable claims. The derived metrics are below.

Metric (synthetic)	Value	Computation
Answer accuracy	67.5%	81 / 120; Wilson 95% interval 58.7% to 75.2%
Macro-F1	0.673	mean of F1 for E (0.714), C (0.700), N (0.605)

Metric (synthetic)	Value	Computation
Coverage	70.0%	84 answered / 120
Selective accuracy (answered only)	69.0%	58 / 84; selective risk 31.0%
Abstention precision	63.9%	23 / 36 abstentions were correct
Abstention recall	57.5%	23 / 40 unanswerable claims; Wilson 95% interval 42.2% to 71.5%
Unanswerable claims answered anyway	42.5%	17 / 40

The matrix shows why accuracy alone is insufficient: a verifier at 67.5% accuracy still answers 17 of 40 unanswerable claims. Macro-F1 (0.673) and accuracy (0.675) nearly coincide only because the synthetic classes are balanced, as in ClaimDB's test sets; on the training split, where not-enough-info claims are more common, they would diverge. Human adjudicators of edge cases should be assessed with chance-corrected agreement, because some agreement is due to chance.

Synthetic reproducibility sample

Suppose an independent reviewer re-executes **n = 40** sampled witnesses from the released bundle and matches the recorded numerator, denominator and verdict in **k = 34**. The point estimate is 85.0%, and the Wilson score interval is 70.9% to 92.9%. The Wilson interval is preferable to the textbook normal approximation: the NIST handbook notes that the Wilson lower limit cannot be negative (www.itl.nist.gov), and that normal-approximation limits may not be accurate enough with few failures (www.itl.nist.gov). Brown, Cai and DasGupta recommend the Wilson or equal-tailed Jeffreys intervals for small samples (doi.org). Had all 40 reproduced, the Wilson interval would run from 91.2% to 100%, and the rule of three, under which 3/n bounds the proportion when zero events occur in n trials (rti.org), gives 7.5% as a rough upper bound on the failure rate. Forty witnesses can therefore support "at least about 71% reproduce", not "the benchmark is reproducible".

Implications and Future Directions

For benchmark authors

- **Release the witness:** publish the query (or its text and hash), the snapshot checksum, the eligibility predicate and intermediate counts. ClaimDB discards the SQL by design, which suits model evaluation but leaves a gap for audit.
- **Record the semantics note:** null handling, duplicate handling, tie-breaks and cutoff conventions as fields, not folklore.
- **Version everything:** the paper's abstract changed between preprint versions (see above), so a claim count with no version is a weak citation.
- **Report four quantities:** accuracy, abstention, query validity, and reproducibility with an interval.
- **Perturb the snapshot:** run each witness on perturbed copies, as in test-suite evaluation.

For verifiers and their evaluators

- **Treat NOT ENOUGH INFO as an output with a reason:** schema omission, subjectivity, counterfactual, or invalid query.
- **Return the bundle with the verdict:** snapshot, query, counts and semantics.
- **Test tool-free and tool-equipped conditions:** the near-random no-tool control shows whether computation is doing the work.

How to evaluate data verification tools

Ask whether the tool exposes the executable query and the snapshot it ran against; whether it reports the eligible-row count; whether it can abstain with a stated reason; and whether a second person can rerun it and obtain the same counts. Ask for a perturbation report rather than one accuracy figure. Because sources vary in quality, it is essential to provide provenance (homepages.inf.ed.ac.uk), and the same reasoning applies to a verifier.

A note on systems of record

A **system of record** keeps one canonical record per real thing and a verbatim, cited source and observation date behind every value. Used with a verifier, it addresses a different layer than a benchmark: it governs where values came from and which past state of the record a verdict refers to. Reliable Context, the publisher of this report, is a vendor in this category; this is the vendor's own description of the category, and no product was evaluated for this report. It is not a benchmark, a query-evaluation harness or a text-to-SQL engine, and readers who need a data catalog, a master data management suite, a vector database or a chatbot should choose those instead. Whether a given system of record meets the witness requirements above is a question to test with the checklist, not to assume.

Open problems

First, claim semantics: a model-written claim may diverge from the gold query, and no public measurement of that divergence for ClaimDB was found. Second, snapshot validity: the database may be incomplete or wrong, and an executable verdict cannot say so. Third, review cost: human reproduction of witnesses is expensive, so sampling with intervals is the realistic method, and sample sizes should be disclosed.

Frequently Asked Questions (FAQs)

How do you verify structured data accuracy?

Separate verification from validation (csrc.nist.gov). For a specific claim, freeze a snapshot, state the eligibility predicate, execute a query, inspect intermediate counts, record the result and semantics, then test the boundary cases in the perturbation matrix.

What is database fact checking?

It decides a claim by computing over rows rather than reading text. FEVER evidence is a Wikipedia sentence, whereas ClaimDB evidence is a multi-table database averaging roughly 110 million tokens per claim.

What does verifying a fact against a database query involve?

A normalized claim, a declared population and predicate, a valid query, a result with numerator and denominator, and a rule mapping the result to ENTAILED, CONTRADICTED or NOT ENOUGH INFO. It also requires decisions on nulls, duplicates and joins.

How are data provenance and lineage used for verification?

Why-provenance identifies the source rows that influenced an output (dl.acm.org), and where-provenance the source location (dl.acm.org). For an aggregate verdict, the eligible rows are the why-provenance, and a derivation links the result to the snapshot.

What is the difference between data verification and data validation?

Verification confirms that specified requirements are fulfilled (csrc.nist.gov); validation confirms that requirements for a specific intended use are fulfilled (csrc.nist.gov).

What is a system of record for data provenance?

In this report, a store that keeps one canonical record per real thing, with a cited source and observation date behind each value, and can reproduce past states; this is the report's working definition, not a universal property of the category, and Reliable Context, the publisher, is a vendor in this category. It fixes which state a verifier examined; it does not replace the query, snapshot or semantics note.

How should structured data fact verification for AI be evaluated?

Score accuracy, abstention, query validity and reproducibility separately. ClaimDB reports that more than half of 30 models score below 55% accuracy (aclanthology.org), but that finding concerns one tool setting and one benchmark; intervals and perturbation tests are needed before transfer to other schemas.

Conclusion

A benchmark or verifier can justify a claim computed from large, joined tables when its verdict ships with its evidence object: the snapshot, the schema, the executable query, the eligible-row count and result, the semantics for nulls, duplicates, joins and cutoffs, and the rule that maps result to label. Under those conditions another researcher can rerun the computation, and a sampled reproducibility rate can be reported with an interval rather than asserted.

ClaimDB moves the field toward that standard by making computation unavoidable. Its paper-reported results, such as 17 of 30 models below 55% in the SQL-tool setting, are informative about model behavior in that setting. They do not establish that individual labels are independently reproducible, because the released records do not carry the query and the labels hold only for a fixed snapshot. The synthetic example makes the same point at small scale: a one-row difference flips a verdict, and six of eight scope errors would pass a verdict-only comparison.

Abstention deserves first-class treatment as a decision with a denominator, a coverage cost and a reason. Future work should publish witnesses, version counts, perturb snapshots, report intervals, and keep the distinction between what a database says and what is true.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://aclanthology.org/2026.acl-long.1589.pdf
2. https://aclanthology.org/2026.acl-long.1589
3. https://arxiv.org/html/2601.14698v2
4. https://doi.org/10.1214/SS/1009213286
5. https://aclanthology.org/N18-1074
6. https://www.postgresql.org/docs/current/queries-table-expressions.html
7. https://csrc.nist.gov/glossary/term/verification
8. https://csrc.nist.gov/glossary/term/validation
9. https://www.epa.gov/sites/default/files/2015-06/documents/g8-final.pdf
10. https://huggingface.co/datasets/michaeltheologitis/claimdb
11. https://arxiv.org/html/2601.14698v1
12. https://www.postgresql.org/docs/current/functions-comparison.html
13. https://www.postgresql.org/docs/current/functions-logical.html
14. https://www.sqlite.org/c3ref/serialize.html
15. https://www.postgresql.org/docs/current/app-pgdump.html
16. https://www.postgresql.org/docs/current/transaction-iso.html
17. https://www.sqlite.org/isolation.html
18. https://www.sqlite.org/lang_aggfunc.html
19. https://www.postgresql.org/docs/current/functions-aggregate.html
20. https://www.postgresql.org/docs/current/datatype-datetime.html
21. https://www.postgresql.org/docs/current/sql-select.html
22. https://www.w3.org/TR/owl2-primer
23. https://www.cs.ubc.ca/sites/default/files/tr/1977/TR-77-16.pdf
24. https://www.postgresql.org/docs/current/using-explain.html
25. https://arxiv.org/pdf/2412.05561
26. https://www.cidrdb.org/cidr2017/papers/p51-chu-cidr17.pdf
27. https://www.postgresql.org/docs/current/datatype-numeric.html
28. https://aclanthology.org/2020.emnlp-main.29
29. https://aclanthology.org/2020.emnlp-main.29.pdf
30. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

31. https://scikit-learn.org/stable/modules/model_evaluation.html
32. https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html
33. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00754
34. https://arxiv.org/pdf/2305.03111
35. https://www.acm.org/publications/policies/artifact-review-and-badging-current
36. https://www.rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf
37. https://www.nature.com/articles/sdata201618
38. https://dl.acm.org/doi/10.5555/645504.656274
39. https://homepages.inf.ed.ac.uk/jcheney/publications/provdbsurvey.pdf
40. https://www.w3.org/TR/prov-dm
41. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/68d30a9594728bc39aa24be94b319d21-Abstract-round1.html
42. https://openreview.net/forum?id=rkeJRhNYDH
43. https://aclanthology.org/2025.emnlp-main.1519
44. https://openreview.net/forum?id=lH1PV42cbF
45. https://aclanthology.org/2023.emnlp-main.483
46. https://proceedings.neurips.cc/paper_files/paper/2023/hash/cd86a30526cd1aff61d6f89f107634e4-Abstract.html
47. https://dl.acm.org/doi/10.1145/2897350.2897352
48. http://www.vldb.org/pvldb/vol8/p938-dong.pdf
49. https://aclanthology.org/2020.emnlp-main.745
50. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
51. https://rti.org/publication/look-rule-three

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.