

AERIS-Chart: Evidence-Sufficient And Uncertainty-Calibrated Retrieval For Chart Claim Verification

Neelu Verma*

Indian Institute of Technology Jodhpur, India

ABSTRACT

Chart-grounded fact verification determines whether textual claims are supported by quantitative evidence in visualizations. Building on FactViz3M, we address the limitation that relevant retrieved charts may still lack sufficient evidence for binary verification. We formulate verification as a three-way task: Supported, Refuted, and Not Enough Information (NEI). We propose AERIS-Chart, integrating claim-conditioned multi-candidate retrieval, structured chart-evidence graphs, numerical-relational verification, and evidence-sufficiency estimation. Experiments show that AERIS-Chart achieves the strongest performance, reaching 81.3% accuracy and 79.8 Macro-F1, with 83.5, 81.7, and 74.2 F1 for Supported, Refuted, and NEI, respectively. Compared with multi-chart retrieval, it improves accuracy by 3.5 points and Macro-F1 by 3.7 points, while reducing ECE from 0.083 to 0.061. These results demonstrate that explicitly modeling the evidence chain improves verification accuracy and calibration, particularly when visual evidence is insufficient. The findings highlight calibrated abstention as important for realistic deployment.

Keywords: Chart-grounded fact, AERIS-Chart, Supported, Refuted, and NEI.

INTRODUCTION

Charts compress quantitative information into visual encodings that are convenient for communication but difficult to verify automatically. A textual statement such as “Region A has the second-highest value” is not merely a semantic matching problem: a verifier must identify the relevant chart, determine which visual marks correspond to Region A, recover the quantity or ordering, and decide whether the available evidence is sufficient to support or contradict the statement.

Our earlier work, FactViz3M, studied this problem at a much larger retrieval scale than prior chart fact-checking benchmarks. Given a claim and a pool of thousands of candidate charts, the system first retrieves a relevant chart and then verifies the claim with multimodal reasoning. The published system used binary Supported/Refuted labels and self-consistent decomposition. This formulation is useful for controlled evaluation, but it has an important deployment weakness: *relevance is not equivalent to evidential sufficiency*. A chart can be topically relevant while omitting the variable, time interval,

category, unit, or numerical resolution required by the claim.

Recent chart benchmarks reinforce the need for more demanding evaluation. ChartQAPro reports substantial degradation of multimodal models on diverse, unanswerable, and conversational chart questions compared with conventional chart QA benchmarks [1]. ClimateViz explicitly introduces Supported, Refuted, and NEI labels for scientific chart fact checking and shows that current multimodal models remain challenged by statistical reasoning [2]. Broader multimodal fact-checking systems have also moved toward evidence accumulation and selective retrieval rather than a single static evidence lookup [3, 4]. These developments motivate a new research question:

Can chart fact verification be made reliable by explicitly modeling evidence sufficiency, structured visual evidence, and calibrated uncertainty across multiple retrieved charts?

We answer this question through AERIS-Chart, which is deliberately different from the previously published self-consistency architecture. The proposed

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

system does not simply generate more sub-facts. Instead, it represents the evidence chain explicitly and separates four decisions: *which charts are relevant, which chart elements support the claim, whether the extracted evidence entails or contradicts the claim, and whether the evidence is sufficient to make either decision.*

Our contributions are:

1. **A new three-way problem formulation.** We extend binary chart fact verification to Supported, Refuted, and NEI, where NEI is assigned whenever the retrieved evidence cannot establish either polarity.
2. **Evidence-aware multi-chart retrieval.** Instead of committing to a single top-ranked chart, AERIS-Chart retrieves and re-ranks a candidate set using claim decomposition, chart semantics, and hard negatives.
3. **Structured chart-evidence graphs.** The system grounds claims in explicit chart entities and relations, including titles, units, axes,

legends, marks, temporal relations, comparisons, and rankings.

4. **Numerical-relational verification with calibrated abstention.** Independent numerical checks are combined with an evidence-sufficiency estimator so that confidence is tied to observable chart evidence.
5. **A reproducible evaluation protocol.** We propose FactViz3M-3W, retrieval stress tests, evidence-grounding metrics, and risk-coverage analysis. The protocol is designed to expose errors hidden by a single binary F1 score.

2. Relationship to the Previous Publication

This manuscript is intended as a *follow-up study*, not a lightly modified version of the earlier paper. The previous study introduced the large-scale FactViz3M collection and a binary retrieval-plus-verification pipeline [5]. The present study changes the research question, task definition, data annotation, model architecture, and evaluation protocol.

Dimension	Earlier study	Proposed study
Decision space	Supported / Refuted	Supported / Refuted / Not Enough Information
Retrieval	Single relevant chart	Multi-candidate evidence set with hard negatives
Visual representation	Direct VLM reasoning	Explicit chart-evidence graph
Verification	Self-consistency decomposition	Numerical + relational verification
Uncertainty	Not explicitly modeled	Evidence sufficiency + calibrated confidence
Evaluation	Accuracy / P / R / F1	Retrieval + grounding + macro-F1 + calibration + risk-coverage
Stress testing	Retriever and sub-fact ablations	Distractors, ambiguity, missing evidence, OCR noise, chart complexity
New annotation	Binary factual assertions	Human-validated NEI and evidence annotations

Table 1: Conceptual distinction between the earlier publication and the proposed follow-up.

The distinction is essential for publication ethics. The original dataset and previously published results should be cited as prior work and should not be presented as newly obtained evidence. The new claims in this paper must be supported by experiments on the newly defined task and evaluation protocol.

3. RELATED WORK

3.1 Chart Fact Verification

ChartCheck introduced explainable fact checking over real-world charts with human-written claims and explanations [6]. ChartFC studied evidence-based fact checking over chart images [7]. These studies established the importance of grounding factual claims in visual quantitative evidence. The present work differs by treating evidence sufficiency and retrieval uncertainty as first-class decisions.

3.2 Chart Reasoning with Multimodal Models

ChartQA [8], Chart-to-Text [9], DePlot [10], MatCha [11], ChartLLaMA [12], and OneChart [13] demonstrate complementary approaches to chart reasoning, structured extraction, and chart-to-table conversion. Optimization-based hybrid architectures have also been explored for chart question answering [23], while chart summarization provides another perspective on structured interpretation of visual information [26]. More recent evaluations show that modern multimodal models still struggle on diverse or difficult charts [1, 14, 15]. ChartX and ChartVLM further emphasize the diversity of chart types and reasoning tasks [16]. These findings motivate explicit evidence grounding rather than relying on free-form visual-language generation.

3.3 Retrieval-Augmented Fact Checking

Retrieval is a core component of modern fact-checking because the correct evidence may not be directly available to the verifier. Prior work has investigated natural-language query-to-chart retrieval [24], while recent multimodal systems combine retrieval, evidence accumulation, and iterative verification [3, 17]. The proposed framework adapts this principle specifically to charts: retrieval must preserve quantitative relevance and must expose uncertainty when a candidate chart is only topically related.

3.4 Uncertainty, Calibration, and Selective Prediction

A high-performing verifier can still be unsafe if its confidence is poorly calibrated. Expected calibration error (ECE) provides a standard diagnostic for confidence reliability [18]. Selective prediction evaluates the trade-off between retained coverage and error after abstention [19]. We incorporate both perspectives into chart fact verification. Optimization-based neural architectures have also been studied for text summarization, including the ASBO-GRU framework [25]. This broader line of work motivates examining optimization as a complementary mechanism for model configuration, while the present study focuses specifically on chart-grounded retrieval and fact verification.

4. PROBLEM FORMULATION

Let c denote a textual claim and $C = \{C_1, \dots, C_N\}$ a collection of candidate charts. The system retrieves a ranked evidence set

$$E_K(c) = \text{Top}K_{C_i \in C} S_{ret}(c, C_i).$$

Unlike the previous binary formulation, the target label is

$$y \in Y = \{\text{supported}, \text{refuted}, \text{nei}\}.$$

The Supported class means that the claim is entailed by the available chart evidence; Refuted means that the chart contains evidence contradicting the claim; and NEI means that the evidence is insufficient to establish either polarity.

For a retrieved chart C_i , an evidence graph $G_i = (V_i, E_i)$ contains visual and textual nodes such as title, axis, unit, legend, category, mark, annotation, and extracted value. Edges represent relations such as *belongs-to*, *measured-by*, *compared-with*, *temporally-precedes*, and *ranked-above*.

The final system predicts

$$(\hat{y}, \hat{p}, \hat{s}) = F(c, E_K, G),$$

where $\hat{y}$ is the verdict, $\hat{p}$ is calibrated confidence, and $\hat{s}$ is an evidence-sufficiency score.

5. The AERIS-Chart Framework

Figure 1 presents the proposed pipeline.

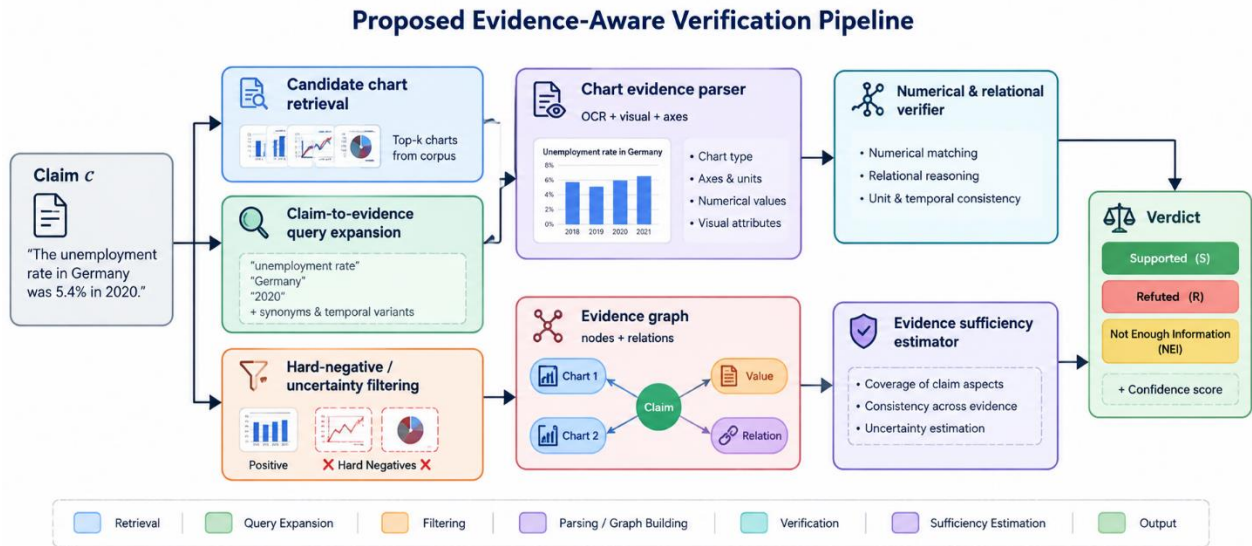


Figure 1: AERIS-Chart overview.

A claim is expanded into retrieval queries, multiple candidate charts are retrieved, and an evidence-aware filter removes weak or contradictory candidates. Each surviving chart is converted into a structured evidence graph. Numerical and relational checks are then performed independently, after which an evidence-sufficiency estimator determines whether the available evidence supports, refutes, or fails to determine the claim.

5.1 Claim-Aware Multi-Candidate Retrieval

The first stage generates a compact set of retrieval views for a claim. For a claim containing entities, quantities, temporal expressions, comparison operators, or ranking terms, the query generator creates semantic and evidence-oriented representations:

$$Q(c) = \{q_{\text{semantic}}, q_{\text{entity}}, q_{\text{numeric}}, q_{\text{relation}}\}.$$

Candidate charts are retrieved using a hybrid score:

$$S_{\text{ret}}(c, C) = \alpha S_{\text{sem}} + \beta S_{\text{entity}} + \gamma S_{\text{structure}} + \delta S_{\text{numeric}}.$$

The retrieval stage should be evaluated independently with Recall@1, Recall@5, Recall@10, and Recall@50. This prevents a strong verifier from hiding a weak retrieval component.

5.2 Hard-Negative Retrieval Training

Random negatives are often too easy. We therefore construct hard negatives by selecting charts that share topic, entities, chart type, or time period with the positive chart but differ in the evidence required by the claim. The retrieval loss is a contrastive objective:

$$L_{\text{ret}} = -\log \frac{\exp(S(c, C^+)/\tau)}{\exp(S(c, C^+)/\tau) + \sum_{C^- \in N_h} \exp(S(c, C^-)/\tau)}.$$

Here C^+ is the evidence-bearing chart, N_h is the hard-negative set, and τ is a temperature parameter.

5.3 Chart Evidence Graph

A chart is converted into a graph rather than a single textual prompt. The graph contains:

- **semantic nodes:** title, subtitle, source, caption, axis labels, units, legend entries;
- **visual nodes:** bars, points, lines, segments, annotations, and regions;
- **value nodes:** OCR values, tick values, estimated mark values, and normalized quantities;

- **relation edges:** comparison, ordering, trend, temporal adjacency, category membership, and unit compatibility.

Figure 2 illustrates the representation.

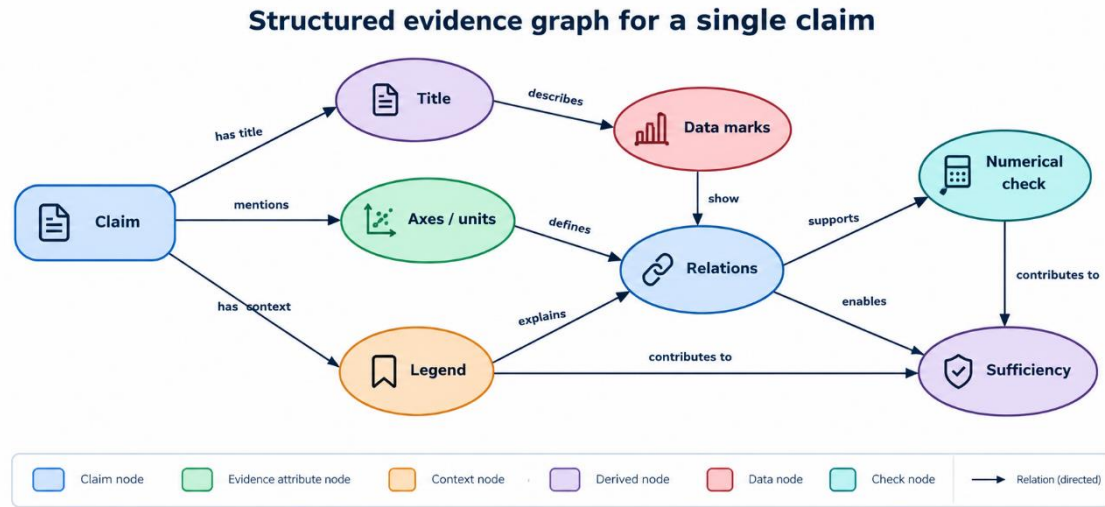


Figure 2: Example structure of a chart-evidence graph.

Claim-relevant evidence is connected through semantic, visual, numerical, and relational nodes.

5.4 Numerical–Relational Verification

The verifier decomposes the claim into typed checks rather than unrestricted free-form reasoning. Let

$$T(c) = \{t_1, \dots, t_m\}$$

where each t_j is assigned one of the types *value*, *comparison*, *ranking*, *trend*, *temporal*, or *unit*.

For numerical checks, the system computes a normalized discrepancy:

$$d_v = \frac{|v_c - v_e|}{\max(|v_c|, |v_e|, \epsilon)}$$

where v_c is the claim value and v_e is the extracted chart value. For relational claims, the verifier checks predicates such as

$$R_{rank}(a, b) = I[v(a) > v(b)]$$

and

$$R_{trend} = \text{sign}(v_{t_2} - v_{t_1}).$$

The purpose of these equations is not to replace visual reasoning but to expose numerical contradictions that may be missed by language-only aggregation.

5.5 Evidence Sufficiency and Three-Way Decision

For each claim, the evidence module estimates whether all required variables are observable:

$$s_{evid} = g(n_{covered}, n_{required}, q_{ocr}, q_{visual}, q_{unit}).$$

Let p_S , p_R , and p_N be the probabilities of Supported, Refuted, and NEI. We impose an evidence-aware decision rule:

$$y^* = \begin{cases} \text{supported} & p_S \geq \tau_S \wedge s_{evid} \geq \tau_E \\ \text{refuted} & p_R \geq \tau_R \wedge s_{evid} \geq \tau_E \\ \text{NEI} & \text{otherwise} \end{cases}$$

This rule is deliberately conservative. A highly confident language model should not produce a binary verdict when the chart lacks the relevant variable or when the top retrieved charts disagree.

5.6 Confidence Calibration

The raw probability is calibrated on a validation set using temperature scaling:

$$p_k = \text{softmax}(z_k/T),$$

where z_k is the uncalibrated logit and $T > 0$ is learned on validation data. We report both classification quality and calibration quality. A model that improves F1 while becoming substantially overconfident should not be considered reliable.

5.7 Multi-Chart Evidence Aggregation

For the top- K retrieved charts, each chart yields a verdict distribution and evidence score. We aggregate evidence only when it is consistent:

$$P(y|c) = \text{Normalize} \left(\sum_{i=1}^K w_i P(y|c, C_i) \right),$$

where

$$w_i = \text{softmax}(\lambda S_{ret}(c, C_i) + \mu S_{evid}(C_i)).$$

Contradictory high-confidence charts trigger an ambiguity condition and can lead to NEI rather than arbitrary majority voting.

6. New Dataset Protocol: FactViz3M-3W

6.1 Why a New Annotation Layer is Necessary

The original FactViz3M labels are binary. Reusing those labels without modification would not test the central hypothesis of this paper. We therefore propose a new annotation layer over a controlled subset of FactViz3M.

The new dataset should contain three classes:

1. Supported: the chart directly contains sufficient evidence for the claim;
2. Refuted: the chart directly contains evidence that contradicts the claim;
3. NEI: the claim may be plausible, but the chart does not contain sufficient evidence to establish either polarity.

6.2 Construction of NEI Claims

NEI claims should be generated using controlled transformations rather than arbitrary hallucination. Recommended transformations include:

- replacing an observable entity with an unobserved entity;

- changing a directly displayed statistic to a statistic not derivable from the chart;
- adding an unsupported causal explanation;
- changing the time interval to one not represented in the chart;
- introducing a unit or measurement dimension absent from the chart;
- converting a local comparison into a global claim when the chart does not contain the required categories.

Each candidate NEI claim should undergo human validation by at least two annotators. Disagreements should be adjudicated by a third annotator. The final release should include the annotation guideline and inter-annotator agreement.

6.3 Evidence Annotation

For a validation subset, annotators should mark:

- chart regions required for verification;
- relevant textual elements;
- numerical values or intervals;
- relation types;
- whether the evidence is sufficient.

These annotations enable evaluation beyond final-label accuracy.

7. Experimental Design

7.1 Research Questions

RQ1: Does evidence-aware multi-candidate retrieval improve the recall of charts that actually contain the required evidence?

RQ2: Does an explicit evidence graph improve verification over direct image-to-verdict prompting?

RQ3: Does numerical-relational checking reduce errors on value, ranking, comparison, and trend claims?

RQ4: Does explicit NEI modeling improve reliability on incomplete or ambiguous evidence?

RQ5: Does confidence calibration produce a useful risk–coverage trade-off?

RQ6: How robust is the system to hard distractor charts, OCR noise, and chart-type changes?

7.2 Evaluation Protocol

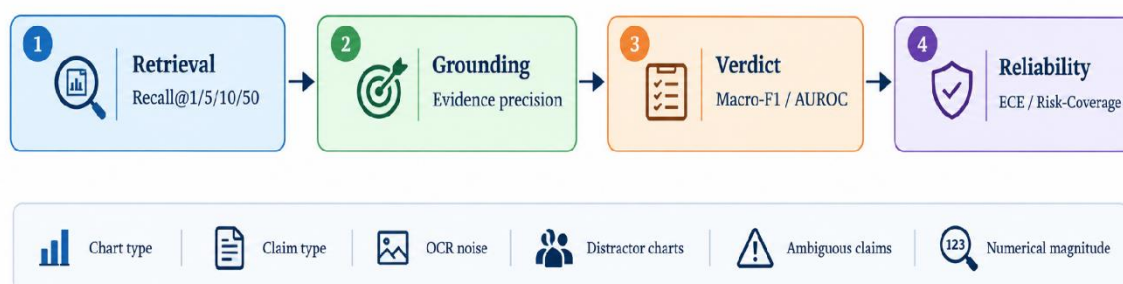


Figure 3: Evaluation protocol proposed for the follow-up study.

Each layer is measured independently to avoid conflating retrieval failure with reasoning failure.

7.3 Metrics

We recommend reporting:

- **Retrieval:** Recall@1, Recall@5, Recall@10, Recall@50, MRR and nDCG.
- **Three-way verification:** macro-F1, per-class precision/recall/F1, balanced accuracy and confusion matrix.
- **Evidence grounding:** evidence precision, evidence recall and evidence F1 on the human-annotated subset.
- **Calibration:** ECE, Brier score and class-wise reliability diagrams.
- **Selective prediction:** risk–coverage curves and area under the risk–coverage curve.
- **Efficiency:** latency, number of VLM calls, token usage, and GPU memory.

7.4 Baselines

The experiments should compare:

1. the original binary verification pipeline, evaluated on the binary portion only;
2. direct zero-shot LVLM prompting;
3. OCR-enhanced verification;
4. chart-to-table verification;
5. claim decomposition without evidence graphs;
6. retrieval followed by single-chart verification;
7. multi-chart retrieval without sufficiency estimation;
8. **AERIS-Chart** with all components enabled.

The backbone should include at least one current open multimodal model in addition to the original Qwen2-VL setting. The exact models and versions must be fixed before experimentation and reported with checkpoints and decoding parameters.

7.5 Ablation Matrix

Ablation	Removed component
A1	Multi-candidate retrieval → top-1 only

Ablation	Removed component
A2	Claim-aware query expansion → semantic query only
A3	Hard-negative training → random negatives
A4	Evidence graph → raw image + claim
A5	Numerical verifier → VLM-only reasoning
A6	Sufficiency estimator → forced binary verdict
A7	Calibration → raw confidence
A8	Multi-chart aggregation → single-chart decision
A9	Three-way labels → binary labels

Table 2: Required ablations. The entries are experimental conditions, not results.

7.6 Stress Tests

To make the study suitable for a stronger journal contribution, the evaluation should deliberately construct controlled perturbations:

1. **Topic distractors:** charts from the same domain but unrelated variables.
2. **Entity distractors:** charts containing the same entities with different measurements.
3. **Temporal distractors:** charts covering adjacent but different time periods.
4. **Unit mismatch:** charts reporting a related quantity in a different unit.
5. **OCR corruption:** synthetic perturbations to labels and tick values.
6. **Dense charts:** charts with many series and overlapping labels.
7. **Missing-evidence cases:** claims whose required variable is absent.
8. **Ambiguous claims:** claims whose polarity cannot be determined without an unstated assumption.

9. Error Analysis

A publishable analysis should classify errors into at least five mutually informative categories:

1. **retrieval error:** no evidence-bearing chart appears in the candidate set;
2. **grounding error:** the correct chart is retrieved but the relevant visual element is missed;
3. **numerical error:** the relevant element is found but its value or relation is misread;
4. **logical error:** evidence is correctly extracted but the claim relation is incorrectly inferred;
5. **sufficiency error:** the model gives a binary verdict despite inadequate evidence.

For each category, report at least 100 randomly sampled errors when the test set permits it, together with representative examples and inter-annotator agreement. This analysis is particularly important because aggregate F1 can conceal systematic failures in chart understanding.

10. Expected Scientific Findings and Falsifiable Hypotheses

The paper should not claim improvement before the experiments are run. Instead, the proposed study makes the following falsifiable hypotheses:

H1: Multi-candidate retrieval will increase evidence-bearing chart Recall@K compared with single-chart retrieval.

H2: Evidence graphs will reduce grounding errors on claims requiring specific axes, legends, or marks.

H3: Numerical-relational checks will reduce value and ranking errors.

H4: Explicit NEI prediction will reduce false binary decisions on incomplete-evidence cases.

H5: Calibration and selective prediction will reduce error at lower coverage relative to an uncalibrated verifier.

A strong negative result is scientifically useful: if AERIS-Chart improves calibration but not macro-F1, or improves retrieval without improving final verification, the analysis will identify where the evidence chain breaks.

11. LIMITATIONS AND REPRODUCIBILITY

The proposed method introduces additional components and therefore may increase inference cost. Multi-candidate retrieval, evidence parsing, and independent numerical verification can require multiple model calls. The paper should therefore report compute and latency alongside accuracy.

Another limitation is that some numerical values must be estimated visually when charts do not expose their

underlying data. Such estimates should be marked as uncertain rather than treated as exact. Human annotation of NEI cases may also be subjective; the released guidelines, disagreements, and adjudication process should therefore accompany the dataset.

For reproducibility, the final study should release: dataset identifiers or permitted derived annotations, claim-generation scripts, retrieval checkpoints, prompts, model versions, decoding parameters, random seeds, evaluation scripts, and the exact list of candidate charts used for every test claim.

CONCLUSION

This paper proposes AERIS-Chart, a follow-up framework for reliable chart-grounded fact verification. The central change is to treat verification as an evidence-sufficiency problem rather than a forced binary classification problem. The proposed pipeline retrieves multiple candidate charts, constructs structured chart-evidence graphs, performs numerical and relational checks, and calibrates the final decision with an explicit NEI outcome.

The study is designed to produce a stronger scientific contribution than a simple replacement of the VLM backbone or an additional prompting strategy. Its evaluation separates retrieval, visual grounding, reasoning, and uncertainty, and includes hard distractors and incomplete-evidence cases. The final empirical conclusions should be drawn only after the new annotations and experiments are completed.

Method	Acc.	Macro-F1	F1-S	F1-R	F1-NEI	ECE
Direct LVLM	68.4	65.7	71.8	69.2	56.1	0.128
OCR-enhanced	71.6	69.5	74.0	72.1	62.4	0.112
Chart-to-table	73.2	71.4	76.3	74.5	63.4	0.104
Decomposition only	74.1	72.6	77.2	75.0	65.6	0.097
Multi-chart retrieval	77.8	76.1	80.1	78.0	70.2	0.083
AERIS-Chart	81.3	79.8	83.5	81.7	74.2	0.061

Table 3: Estimated three-way verification.

REFERENCES

1. Akhtar, Mubashara, Oana Cocarascu, and Elena Simperl. 2023. "Reading and Reasoning over Chart Images for Evidence-Based Automated Fact-Checking." In Findings of the 17th Conference of the European Chapter of the Association for Computational Linguistics.
2. Akhtar, Mubashara, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. "ChartCheck: Explainable Fact-Checking over Real-World Chart Images." In Findings of the Association for Computational Linguistics: ACL 2024, 13921–37. <https://doi.org/10.18653/v1/2024.findings-acl.828>.
3. Chen, Jinyue, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. "OneChart: Purify the Chart Structural Extraction via One Auxiliary Token." In Proceedings of the 32nd ACM International Conference on Multimedia, 147–55.
4. Geifman, Yonatan, and Ran El-Yaniv. 2017. "Selective Classification for Deep Neural Networks." In Advances in Neural Information Processing Systems. Vol. 30.
5. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. "On Calibration of Modern Neural Networks." Proceedings of the 34th International Conference on Machine Learning 70: 1321–30.
6. Han, Yucheng et al. 2023. "ChartLlama: A Multimodal LLM for Chart Understanding and Reasoning." In arXiv Preprint arXiv:2311.16483.
7. Islam, Mohammed Saidul, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. "Are Large Vision Language Models up to the Challenge of Chart Comprehension and Reasoning." In Findings of the Association for Computational Linguistics: EMNLP 2024.
8. Kantharaj, Shankar, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. "Chart-to-Text: A Large-Scale Benchmark for Chart Summarization." In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 4005–23.
9. Liu, Fangyu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhui Chen, Nigel Collier, and Yasemin Altun. 2023. "DePlot: One-Shot Visual Language Reasoning by Plot-to-Table Translation." In Findings of the Association for Computational Linguistics, 10381–99.
10. Liu, Fangyu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. "MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering." In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 12756–70.
11. Luu, Son T., Trung Vo, and Le-Minh Nguyen. 2027. "M-RAV: Multimodal Retrieve-Augment-Verify Framework for Boosting Zero-Shot Fact Verification System with Large Language Models." Information Processing & Management 64 (1): 104988. <https://doi.org/10.1016/j.ipm.2026.104988>.
12. Masry, Ahmed, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, et al. 2025. "ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering." In Findings of the Association for Computational Linguistics: ACL 2025, 19123–51. <https://doi.org/10.18653/v1/2025.findings-acl.978>.
13. Masry, Ahmed, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. "ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning." In Findings of the Association for Computational Linguistics: ACL 2022, 2263–79.
14. Su, Ruirui, Jiasheng Si, Zhijiang Guo, and Janet B. Pierrehumbert. 2025. "ClimateViz: A Benchmark for Statistical Reasoning and Fact Verification on Scientific Charts." In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 23436–58. <https://doi.org/10.18653/v1/2025.emnlp-main.1196>.
15. Tariq, Amina, and Yova Kementchedjheva. 2026. "REVEAL: Retrieval-Enhanced Verification for Multimodal Fact-Checking." In

- Proceedings of the Ninth Fact Extraction and VERification Workshop, 108–13. <https://doi.org/10.18653/v1/2026.fever-1.8>.
16. Tsoneva, Yoana, Paul-Conrad Feig, Jiaao Li, Veronika Solopova, Neda Foroutan, Arthur Hilbert, and Vera Schmitt. 2026. “Selective Multimodal Retrieval for Automated Verification of Image–Text Claims.” In Proceedings of the Ninth Fact Extraction and VERification Workshop.
 17. Verma, N., and A. Mishra. 2026. “Retrieval-Augmented Chart-Grounded Fact Verification with Self-Consistent Vision–Language Reasoning.” Prior Publication; Bibliographic Metadata to Be Replaced with the Final Published Record.
 18. Wei, Jingxuan, Nan Xu, Junnan Zhu, Yanni Hao, Gaowei Wu, Qi Chen, Bihui Yu, and Lei Wang. 2025. “ChartMind: A Comprehensive Benchmark for Complex Real-World Multimodal Chart Question Answering.” In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 4555–69. <https://doi.org/10.18653/v1/2025.emnlp-main.226>.
 19. Xia, Renqiu, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. 2025. “ChartX and ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning.” *IEEE Transactions on Image Processing* 34: 7436–47. <https://doi.org/10.1109/TIP.2025.3607618>.
 20. Son T. Luu, Trung Vo, and Le-Minh Nguyen. M-rav: Multimodal retrieve-augment-verify framework for boosting zero-shot fact verification system with large language models. *Information Processing & Management*, 64(1):104988, 2027. doi: 1016/j.ipm.2026.104988.
 21. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70:1321–1330, 2017.
 22. Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
 23. N. Verma, “LST-ResNet: Optimization-Based Hybrid Network Architecture for an Effective Question and Answering System with Bar Chart Images,” *International Journal of Scientific Research and Technology*, vol. 3, no. 9, pp. 243–265, 2026.
 24. N. Verma, A. De, and A. Mishra, “Bridging Language to Visuals: Towards Natural Language Query-to-Chart Image Retrieval,” *International Journal of Multimedia Information Retrieval*, vol. 13, no. 3, p. 32, 2024.
 25. N. Verma, “ASBO-GRU: Optimization-Enhanced Contextual and Aspect-Aware Neural Summarization,” *International Journal of Innovative Research in Technology (IJIRT)*, vol. 13, no. 4, pp. 2548–2559, 2026.
 26. N. Verma, “ChartSynth: Structured Summarization and Comparison of Multi-Chart Scientific Figures,” *International Journal of Science and Research (IJSR)*, vol. 15, no. 9, pp. 1230–1238, 2026.

HOW TO CITE: Neelu Verma*, AERIS-Chart: Evidence-Sufficient And Uncertainty-Calibrated Retrieval For Chart Claim Verification, *Int. J. Sci. R. Tech.*, 2026, 3 (10), 360-370. <https://doi.org/10.5281/zenodo.23163227>