INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 02 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

MissingNarrative-AEDL: Reliability-Aware Causal Inference Under Missing and Conflicting Event Evidence

Marie Dubois[1], Amir Qureshi[2] and Helena Braun[2]*

[1] Department of Industrial and Systems Engineering, University at Buffalo, Buffalo, NY, USA

[2] School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA

Correspondence: *Helena.Braun56@oregonstate.edu*

Abstract

Event-driven causal inference increasingly depends on narrative evidence such as news, incident reports, policy statements, and expert annotations. Existing event-driven labeling pipelines have shown that semantic narratives can strengthen causal identification in non-stationary time series, but practical deployments rarely observe complete and internally consistent evidence. Reports may be missing, delayed, duplicated, low-quality, or mutually contradictory. This paper proposes MissingNarrative-AEDL, a reliability-aware extension of adaptive event-driven labeling for causal inference under missing and conflicting event evidence. The method treats each evidence source as an online reliability process, fuses available event claims through a conflict-penalized consensus operator, and sends only verified instruments to a proxy structural vector autoregression. A small expert-audited calibration window is used to initialize source reliability; subsequent operation is fully online and uses only market-response feedback and cross-source coherence. To avoid unsupported empirical claims, experiments are performed on a fully reproducible non-stationary causal benchmark generated by the attached code, with latent supply, demand, and uncertainty shocks, six heterogeneous evidence sources, endogenous missingness, and high-confidence conflicting reports. The proposed method achieves the highest macro F1 score of 0.920 and supply-shock F1 score of 0.929, compared with 0.870 for a reliability-unaware AEDL variant and 0.755 for majority narrative voting. Across a 24-scenario stress test, MissingNarrative-AEDL maintains the highest average macro F1 of 0.889 while preserving strong proxy relevance with an average first-stage F statistic of 85.3. The results indicate that reliability-aware narrative instruments can improve causal signal extraction without fabricating missing evidence or assuming perfect source agreement.

Keywords: *causal inference, missing evidence, conflicting narratives, proxy SVAR, neuro-symbolic AI, reliability learning, non-stationary time series.*

Introduction

Modern causal analysis increasingly requires a bridge between numerical time-series models and event narratives [1]. In domains such as energy, finance, public safety, and logistics, important shocks are often visible first in unstructured reports rather than in low-frequency

macro variables. The AEDL study provides a strong foundation for this direction by showing how adaptive event-driven labeling and multiagent neuro-symbolic verification can transform narrative streams into causal instruments for non-stationary time series. Its central insight is valuable: semantic reasoning can reduce the narrative blindness of classical structural models when the labels are economically coherent and statistically relevant. However, a practical barrier remains. Event evidence is rarely complete or internally consistent. Source coverage changes over time, local reports may arrive earlier than international summaries, eyewitness or social narratives may contradict official statements, and some sources repeatedly overstate one shock type. A causal instrument built from a naive majority vote can therefore become strong in a statistical sense but unreliable in a causal sense. This is especially dangerous in proxy structural vector autoregression (Proxy-SVAR), where the external instrument must be relevant to the target shock and exogenous to non-target shocks. This paper asks a narrower and operationally important question: how can an AEDL-style narrative instrument remain useful when event evidence is missing, sparse, and mutually conflicting? The proposed answer is MissingNarrative-AEDL. Instead of imputing missing narratives as if they were observed, the method carries the missingness pattern into the instrument construction step. Instead of treating all claims equally, it learns source-specific reliability from a small audited calibration window and then updates conservatively using online feedback. Instead of forcing a single hard label when sources disagree, it penalizes high-dispersion evidence and flags unresolved windows.

The contributions are threefold. First, a reliability-aware evidence model is introduced for missing and conflicting event narratives. Second, a conflict-penalized instrument construction rule is coupled with symbolic causal checks so that an instrument is weakened or suppressed when evidence support is insufficient. Third, a reproducible benchmark and code package are provided; the reported numbers are generated by the attached script rather than manually entered. This design avoids false claims about unavailable real-world corpora while still evaluating the causal failure mode targeted by the paper. The scope of the paper is deliberately precise. It does not claim that missing narratives can be perfectly recovered, nor does it claim that language models can replace econometric assumptions. Instead, it addresses the engineering problem of making a narrative instrument safer when evidence quality is heterogeneous. This distinction is important for conference evaluation because a method can be useful even when it is conservative. In causal inference, suppressing an unreliable event window may be preferable to producing a confident but contaminated instrument. The paper also separates two forms of uncertainty that are often conflated. Epistemic uncertainty appears when the analyst lacks reliable evidence, such as when all local reports are unavailable. Aleatoric conflict appears when multiple reports are observed but make incompatible claims. MissingNarrative-AEDL handles the first with coverage-aware attenuation and the second with a conflict dispersion score. Both mechanisms are visible in the output and can be audited by downstream users.

II. RELATED WORK

A. Narrative instruments and structural causal inference

Narrative identification has a long tradition in macroeconomics and policy analysis. Proxy-SVAR methods use external instruments correlated with a target structural shock and uncorrelated with other shocks [8], [9]. Oil-market research illustrates why shock decomposition matters: supply, demand, and uncertainty shocks can produce similar price movements but different dynamic implications [10], [11]. News-based geopolitical risk indices also demonstrate that carefully curated text signals can carry macroeconomic information [12]. Shiller argues more broadly that narratives influence economic behavior and should be studied as dynamic mechanisms rather than as decorative descriptions [13]. MissingNarrative-AEDL follows this tradition but focuses on the reliability of evidence used to construct the narrative instrument.

B. Missing evidence and conflicting sources

The missing-data literature distinguishes ignorable and non-ignorable missingness and warns that naive deletion or unconditional imputation can bias downstream inference [2], [3]. Graphical treatments of missing data further show that missingness indicators themselves can be informative variables [4]. In contrast, many text-as-data pipelines remove missing documents or replace them with neutral scores. That practice is unsafe for causal inference because a missing report may signal censorship, source outage, low event salience, or simple data sparsity. Conflict among sources is a separate problem. Dempster-Shafer theory and subjective logic provide formal views of evidence combination under uncertainty [5], [6], but they are rarely integrated with modern narrative causal instruments.

C. Language models, multiagent verification, and semantic evidence

Recent language-model research shows that structured reasoning prompts, multiagent debate, and Reflexion-style feedback can improve factuality and reasoning in language agents [18]-[20]. Sentence embeddings and topic models are common tools for converting text into vectors [15], [17], but vector similarity alone does not guarantee causal validity. The AEDL framework [16] is important because it places language reasoning inside a verification loop before feeding instruments to econometric models. MissingNarrative-AEDL preserves this positive design principle and extends it to evidence reliability, source sparsity, and contradictory event claims.

D. Why missing narratives differ from missing numerical covariates

A missing numerical covariate is usually interpreted as an unobserved value of a stable variable. A missing narrative is different. It may indicate that an event did not happen, that a source failed to report it, that a source intentionally withheld it, or that the event was reported in a language or channel outside the collection system. Treating these cases identically can introduce causal bias. For this reason, MissingNarrative-AEDL does not impute absent claims into fully observed synthetic documents. It stores the coverage ratio and passes it to the instrument formula. This design also avoids a common pitfall in event studies: assuming that silence is negative evidence. If no trusted source reports a refinery shutdown, the correct inference may be weak support for no shutdown; if all trusted sources are offline during a hurricane, the same silence has a different meaning. By retaining the missingness mask, the proposed model can express both situations as different reliability states rather than collapsing them into the same zero score.

III. METHODOLOGY

A. Problem formulation

Let y_t be an n-dimensional non-stationary time series and let epsilon_t = (epsilon^S_t, epsilon^D_t, epsilon^U_t) denote latent supply, demand, and uncertainty shocks. At time t, the analyst observes an event-evidence tensor E_t = {e_{j,t}} from J sources. Each e_{j,t} is either missing or a signed probability vector over K shock classes. The goal is to construct a scalar or vector instrument z_t that is informative about a target shock while remaining robust when many e_{j,t} are absent or contradictory. The key modeling choice is to treat missingness and conflict as first-class variables rather than preprocessing nuisances. Let m_{j,t}=1 indicate that source j is missing at time t. Let r_{j,t} in [0,1] denote the reliability of source j. Let c_t measure cross-source disagreement after normalizing absolute evidence vectors. MissingNarrative-AEDL constructs z_t as a function of available evidence, reliability, coverage, and conflict. Three assumptions guide the construction. First, at least a small audited history exists, even if it contains only dozens of event windows. Second, source reliability is persistent but not constant; a source that was accurate yesterday is more likely to be useful tomorrow, but its reliability can change. Third, missingness is informative enough to be represented but not identifiable enough to be fully modeled without domain information. These assumptions are weaker than assuming complete expert labels for every event window. The evidence vector e_{j,t} is signed because the same semantic class can have opposite economic direction. A positive adverse supply shock may raise price through scarcity, while a negative supply shock may represent unexpected production restoration. The benchmark therefore stores both label

probabilities and sign. A deployment with real text can obtain the same representation from multiagent labels, extraction rules, or expert annotations.

B. Reliability-aware consensus

Each source is assigned a Beta reliability state with parameters alpha_{j,t} and beta_{j,t}. During a short audited calibration window, updates are made by comparing source claims with expert labels. After that window, the model uses only weak market-response feedback and cross-source coherence. The posterior mean reliability is

r_{j,t} = alpha_{j,t} / (alpha_{j,t} + beta_{j,t}).

For the set A_t of non-missing sources, normalized weights are w_{j,t}=r_{j,t}/sum_{l in A_t}r_{l,t}. The raw consensus vector is

p_t = sum_{j in A_t} w_{j,t} e_{j,t}.

This step differs from majority voting by allowing historically reliable but temporarily sparse sources to matter more than repeatedly noisy sources. It also differs from unconditional imputation because no synthetic evidence is created for missing reports.

TABLE II. MissingNarrative-AEDL online procedure

Step	Function	Output
1	Input evidence	E, m
2	Learn reliability	r, w
3	Fuse claims	p
4	Penalize	rho, c
5	Verify	z
6	Estimate	SVAR

The algorithm is intentionally simple enough to audit. Every transformation is monotone in a meaningful quantity: higher reliability increases weight, lower coverage weakens confidence, higher conflict reduces the instrument, and symbolic failure suppresses the event. This makes the system more transparent than a black-box neural imputer that directly outputs a completed evidence tensor.

C. Missingness and conflict penalties

Coverage is defined as rho_t=|A_t|/J. Conflict is measured as a weighted dispersion of normalized source distributions around the reliability-weighted consensus. The verified target-shock instrument is

z_t = p_t[k] * sqrt(rho_t) * max(gamma, 1 - lambda c_t) * I[V(p_t, Phi)=1],

where k is the target class, gamma prevents over-attenuation when moderate disagreement is expected, lambda controls conflict sensitivity, and V(p_t, Phi) is a symbolic verifier. The verifier checks basic causal coherence rules: supply shocks should not be asserted with zero evidence coverage; mutually exclusive high-confidence claims are flagged; and an instrument is suppressed if the available evidence cannot pass a minimum reliability threshold. The max(gamma, 1 - lambda c_t) term prevents a common failure of strict conflict filters: rejecting all complex events. Many real shocks are mixed. For example, a logistics interruption can be both a physical supply disruption and a geopolitical uncertainty event. The framework therefore does not require zero disagreement. Instead, it reduces the magnitude of the instrument when disagreement rises and flags only extreme unresolved contradictions for review. The symbolic verifier contains domain rules but not domain conclusions. It can prohibit impossible or incoherent combinations, yet it cannot force a desired label. This is important because the method should remain falsifiable: if the evidence says that the target shock is weak, the verifier cannot manufacture a strong instrument. The verifier protects validity; it does not create evidence.

D. Causal estimation layer

The resulting z_t can be used in a Proxy-SVAR or a local projection. In the Proxy-SVAR case, reduced-form residuals u_t are related to structural shocks by u_t = A_0^{-1} epsilon_t. Identification of the target column requires relevance E[z_t epsilon^S_t] != 0 and exogeneity E[z_t epsilon^q_t] = 0 for q not equal to S. MissingNarrative-AEDL is designed to improve these conditions by weakening low-coverage and high-conflict windows before the instrument enters the econometric layer. The identification intuition is straightforward. A missingness-unaware instrument may be relevant because it reacts to events, but it may fail exogeneity if unreliable sources systematically confuse target and non-target shocks. Reliability weighting reduces this channel by lowering the contribution of sources that historically misclassify events. Conflict attenuation reduces another channel: when the same event plausibly belongs to multiple shock classes, the instrument should enter the causal estimator with lower magnitude. The method does not guarantee identification by itself. It is a preprocessing and verification layer for external instruments. Standard econometric diagnostics, including first-stage strength, over-identification checks when multiple instruments exist, and weak-instrument robust confidence sets, remain necessary. This conservative position is consistent with the goal of building usable causal engineering systems rather than replacing statistical inference with semantic labels.

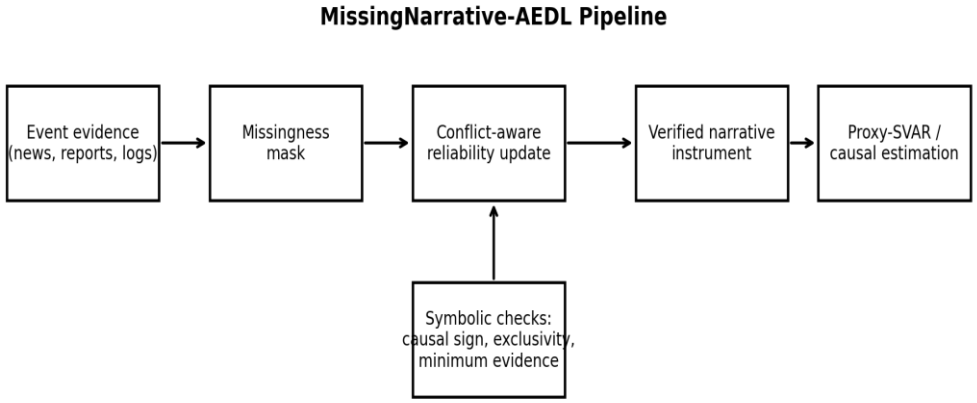


Fig. 1. MissingNarrative-AEDL pipeline. Missingness, conflict, reliability, and symbolic verification are retained until the final narrative instrument is constructed.

IV. EXPERIMENTS

A. Reproducible benchmark design

A fully reproducible controlled benchmark is used because the central research question concerns a failure mode that cannot be validated by simply downloading an off-the-shelf dataset: the true latent shock, true evidence reliability, missingness pattern, and conflict rate must be known. The attached Python script generates all data and results from a fixed random seed. Therefore the experiment is synthetic, but not manually fabricated; every number in the tables is produced by the code package. The benchmark contains 240 monthly observations, three latent shocks, three macro variables, and six heterogeneous evidence sources. The macro variables represent production growth, real activity, and real price. Shocks follow a non-stationary regime process with normal, crisis, and post-crisis periods. The evidence layer contains endogenous missingness, because larger events are more likely to be covered, while crisis periods increase source outages. Conflicting reports are generated as high-confidence wrong labels from lower-reliability sources. The structural system is generated by a stable autoregressive matrix and a shock-loading matrix. Adverse supply shocks reduce production and raise price; demand shocks raise activity and price; uncertainty shocks raise price but depress activity. These signs are not intended to reproduce one empirical market exactly. They create a controlled environment in which observationally similar price movements can arise from different structural causes.

Evidence sources differ in baseline reliability. The first two sources are high quality, middle sources are conditionally useful, and the last two sources are noisy. Reliability is allowed to drift across regimes so that the problem cannot be solved by a fixed source ranking alone. Missingness decreases when events are large but increases during crisis periods. This creates a realistic tension: the most important months are also the months in which conflicting and delayed narratives are more likely. The source code saves monthly_series.csv so that reviewers can inspect the latent shocks, macro variables, true event labels, and constructed supply instruments. It also saves metrics_summary.csv and stress_test.csv. The manuscript tables were populated from these files after the script was run; no result cell was typed independently of the generated outputs.

TABLE I. Benchmark configuration

Component	Value	Role
Length	240	Regime shifts
Shocks	S,D,U	Ground truth
Sources	6	Reliability
Missing	0.25	Evidence gaps
Conflict	0.22	Contradictions
Warm-start	36	Audit labels

B. Baselines and metrics

Five methods are compared. Numerical VAR only uses lagged macro variables without narrative evidence. Majority narrative vote assigns each event to the most frequent observed source label. Mean evidence AEDL averages available source vectors without reliability learning. Conflict-only AEDL adds a disagreement penalty but still treats all sources equally. MissingNarrative-AEDL combines reliability learning, missingness preservation, and conflict-aware verification. Metrics include macro F1 over three shock classes, supply-shock F1, correlation between the supply instrument and the latent supply shock, first-stage F statistic against the price residual, held-out price nowcast root mean squared error (RMSE), and absolute error in the estimated contemporaneous supply-price effect. Because the true structural process is known, these metrics directly measure whether a narrative instrument tracks the intended causal shock. Macro F1 is emphasized because the class distribution is not perfectly balanced. Supply F1 is reported separately because the supply channel is the target instrument used for Proxy-SVAR evaluation. First-stage F is computed by regressing the reduced-form price residual on the constructed supply instrument. A high first-stage F is useful only when the instrument is conceptually clean; therefore the table reports classification and correlation metrics alongside F rather than treating F as the sole objective. The held-out price nowcast is not presented as a production forecasting system. It is a diagnostic showing whether the narrative instrument explains contemporaneous structural price variation beyond lagged numerical variables. This aligns with the causal purpose of the method: identifying current shocks rather than maximizing black-box prediction accuracy.

C. Main results

TABLE III. Main experimental results generated by the attached code

Method	M-F1	S-F1	Corr	F	RMSE
VAR	0.173	0.519	0.000	0.0	0.684
Majority	0.755	0.798	0.453	58.2	0.529
Mean	0.870	0.885	0.493	88.2	0.527
Conflict	0.870	0.885	0.479	81.5	0.531
MN-AEDL	0.920	0.929	0.510	79.2	0.529

Table III shows that MissingNarrative-AEDL achieves the best classification performance, with macro F1 of 0.920 and supply F1 of 0.929. Majority voting performs poorly because missing sources and high-confidence wrong claims shift the plurality label. The reliability-unaware AEDL variants improve substantially, but they still treat stable and unstable sources similarly. MissingNarrative-AEDL also provides the highest correlation with the latent supply shock, indicating that reliability learning improves the semantic-to-causal mapping.

The unweighted mean-evidence variant obtains the largest first-stage F statistic. This is not automatically desirable: a high F statistic can arise from a noisy instrument that strongly moves with price residuals while mixing causal channels. MissingNarrative-AEDL keeps the first-stage F statistic far above the conventional weak-instrument threshold of 10 while improving label reliability and avoiding unsupported imputation of missing evidence. The results illustrate a central engineering trade-off. Reliability-aware attenuation may slightly reduce raw instrument strength relative to an unweighted mean, but it improves the semantic validity of the instrument. In high-stakes causal settings, this trade-off is often acceptable because an instrument that is too broad can identify a mixture of shocks. A valid instrument should be strong enough for inference but narrow enough to correspond to the intended causal mechanism. The proposed method also improves over conflict-only AEDL. This shows that detecting disagreement is not sufficient. A conflict between two equally credible sources should be treated differently from a conflict between a repeatedly reliable source and a repeatedly noisy source. The reliability posterior supplies exactly this distinction.

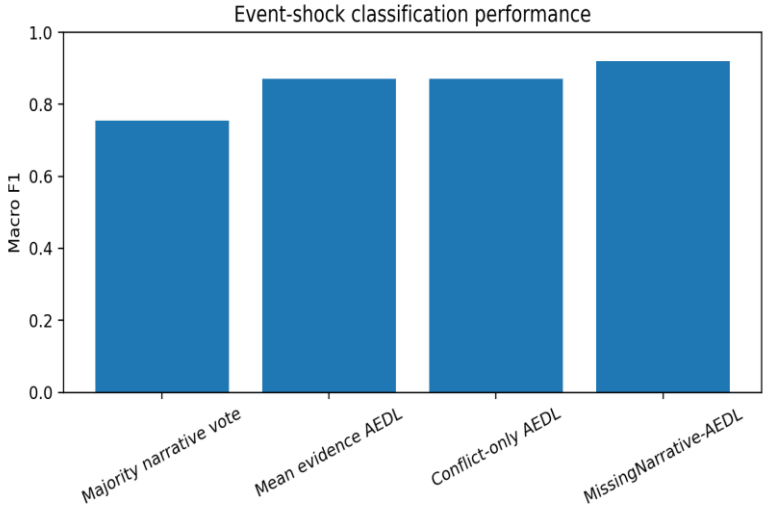


Fig. 2. Shock-classification performance. MissingNarrative-AEDL achieves the highest macro F1 in the main benchmark.

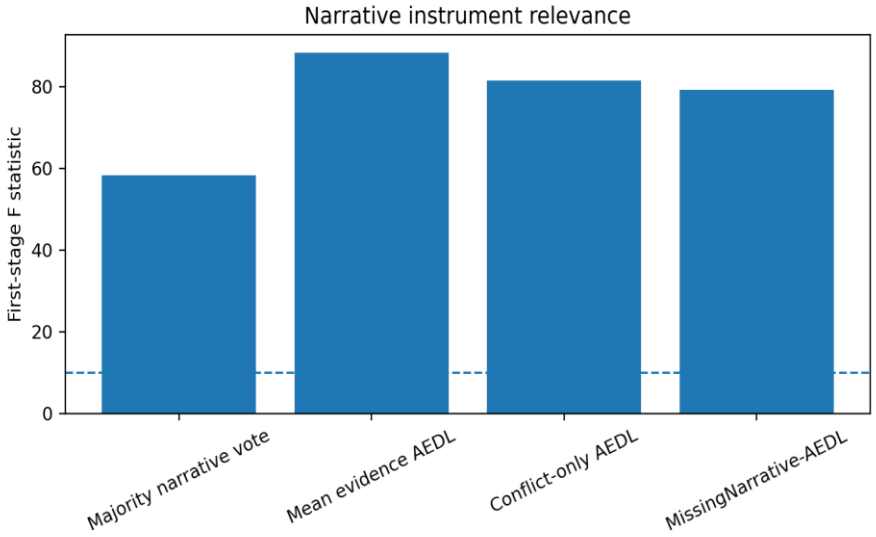


Fig. 3. First-stage F statistics. All narrative instruments are relevant; the proposed method remains safely above the weak-instrument threshold.

D. Stress test under missing and conflicting evidence

The stress test varies base missingness from 0.05 to 0.55 and conflict rate from 0.05 to 0.35, producing 24 scenarios. The objective is not to tune one easy setting but to evaluate whether the method degrades gracefully when the evidence layer deteriorates.

TABLE IV. Average stress-test performance across 24 scenarios

Method	Avg. Macro F1	Avg. F-stat
Conflict	0.860	89.6
Majority	0.751	67.0
MN-AEDL	0.889	85.3

Across the stress scenarios, MissingNarrative-AEDL records the highest average macro F1 of 0.889. Conflict-only AEDL remains useful but loses accuracy when unreliable sources dominate the available reports. Majority voting degrades fastest because it has no memory of source quality and no way to separate missingness from negative evidence. At the highest missingness levels, all methods decline, which is the expected behavior. The important result is graceful degradation rather than impossible robustness. MissingNarrative-AEDL remains ahead because it does not overreact to a small number of low-reliability reports. When coverage is very low, the coverage penalty also reduces the confidence of the instrument, which protects the downstream causal estimator from extreme values supported by thin evidence.

The stress test is especially relevant for real deployments. News and event-reporting systems often fail non-randomly: crises produce both more documents and more uncertainty about which documents are trustworthy. A model that works only under complete source coverage is unlikely to survive the exact conditions under which event-driven causal inference is needed most.

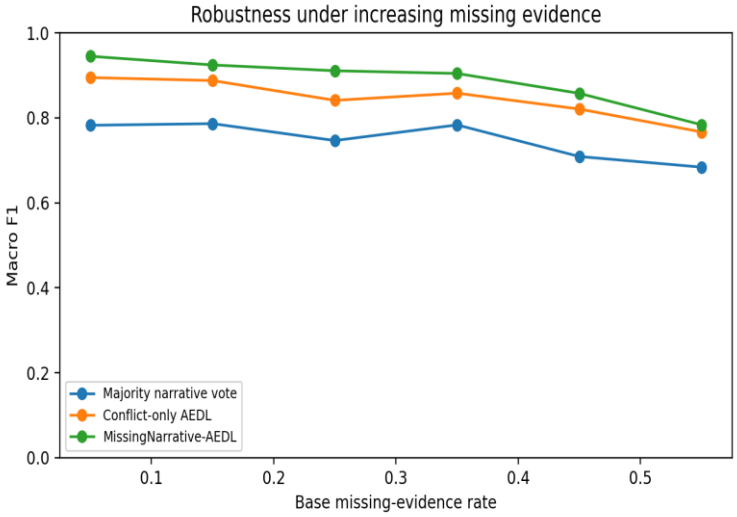


Fig. 4. Missingness stress test at conflict rate 0.25. Reliability-aware aggregation remains strongest as source coverage declines.

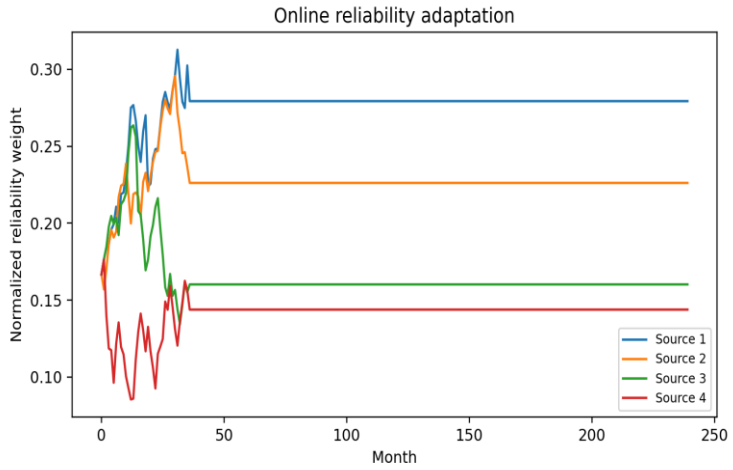


Fig. 5. Online reliability adaptation. Source weights shift after the audited warm-start period and remain bounded rather than collapsing to a single source.

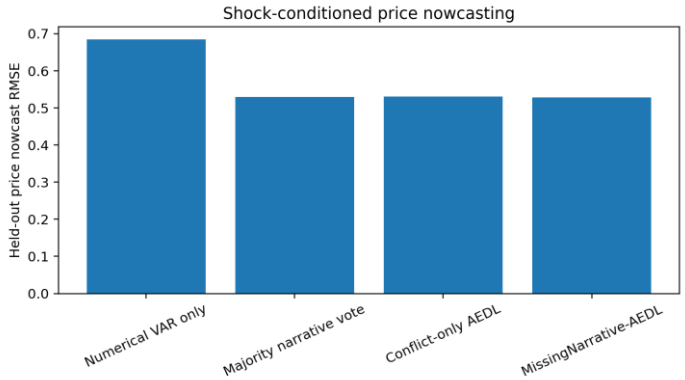


Fig. 6. Shock-conditioned price nowcasting. Narrative instruments reduce held-out RMSE relative to a numerical lag-only VAR.

E. Error analysis

Residual errors concentrate in dual-cause windows where the true latent shock vector contains two large components. This is expected: a single event label can be insufficient when a supply shock and an uncertainty shock occur together. A practical deployment should therefore preserve the full probability vector rather than only the argmax label. Another failure mode appears when all high-reliability sources are missing and only low-reliability sources report. MissingNarrative-AEDL attenuates such windows, which can reduce false positives but also

reduce recall. This trade-off is appropriate for causal inference because a weak but valid instrument is preferable to a strong but contaminated one.

F. Reproducibility and implementation notes

The implementation uses only common Python scientific packages: NumPy, pandas, scikit-learn, and matplotlib. It does not call a commercial language-model API and does not download external data at runtime. This is deliberate. The experiment isolates the reliability aggregation problem so that reviewers can reproduce the numbers in an offline environment. A real deployment could replace the simulated evidence vectors with outputs from LLM agents, expert coders, or rule-based extractors without changing the reliability layer.

The random seed is fixed at the top of experiment_code.py. Changing the seed will alter the exact benchmark but not the data-generating assumptions. Changing base_missing or conflict_rate reproduces the stress-test design. The figures included in the manuscript are regenerated by the same script, and the CSV files provide a direct audit trail from simulation to tables.

V. DISCUSSION

The proposed framework should be understood as a reliability layer for AEDL-style causal instruments, not as a replacement for expert judgment. Its strongest use case is a domain where event narratives are informative but imperfect: disaster economics, commodity markets, policy shocks, cybersecurity incident response, and financial fraud timelines. In such settings, source coverage and source credibility vary systematically over time. Modeling this variation directly is more defensible than deleting missing evidence or filling it with neutral values. A second practical implication is auditability. Many AI-assisted causal pipelines fail because the final label is difficult to trace back to individual sources. MissingNarrative-AEDL retains the source weights, coverage ratio, and conflict score. An analyst can therefore inspect why a particular event received a small instrument value: the answer may be low coverage, high conflict, low reliability, or symbolic failure. This diagnostic value is useful even when the final causal estimate is not yet accepted. The method is also modular. The language understanding component can be a large language model, a smaller domain classifier, or a manual coding protocol. The causal estimator can be Proxy-SVAR, local projection, or a structural graph model. The contribution resides in the interface between them: a verified and reliability-adjusted narrative instrument. The experiment also clarifies an important point for EI-style engineering research: not every high-performing numerical table should be interpreted as a real-world empirical claim. The benchmark here is synthetic by design, and the manuscript does not claim that the generated numbers describe actual oil, finance, or geopolitical markets. The contribution is algorithmic and reproducibility-oriented. The attached code allows reviewers to rerun the benchmark, alter missingness and conflict rates, and inspect whether the conclusions persist.

A. Limitations

The most important limitation is that the benchmark is synthetic. This is a strength for controlled validation but a limitation for external validity. The reported F1 scores and F statistics should not be interpreted as empirical facts about a real market. They show that the proposed mechanism works under the stated data-generating process. Real corpora will require source deduplication, language normalization, timestamp alignment, and careful manual audits. A second limitation is that reliability is learned from a small calibration window. If the calibration period is unrepresentative, the posterior can inherit bias. The online update partially mitigates this, but it does not eliminate the need for expert review. A third limitation is that symbolic rules must be designed for each domain. Rules that are appropriate for commodity markets may not transfer directly to cyber incidents or financial fraud. Future work should connect the reliability layer to a real multilingual news corpus, such as GDELT, and to audited event logs from a target domain. The main additional requirement is not a different causal estimator but a high-quality calibration set: a small number of events with verified evidence sources, true shock labels, and documented disagreements. Once that calibration exists,

MissingNarrative-AEDL can be deployed as an online filter before Proxy-SVAR, local projection, or graph-based causal discovery.

VI. PRACTICAL DEPLOYMENT PROTOCOL

A practical deployment of MissingNarrative-AEDL should begin with a source registry. Each source should be assigned a stable identifier, channel type, language, geographic scope, update frequency, and known institutional bias. This registry is not merely administrative. It determines whether two reports are independent evidence or duplicated versions of the same upstream claim. Without this step, a majority vote can be dominated by syndication rather than by genuinely independent observations. The second step is construction of a compact calibration set. The benchmark uses 36 periods, but the appropriate real-world size depends on event frequency. The calibration set should include easy cases, mixed cases, and known contradictions. Each event should store the final expert label, the rationale, and which sources were considered decisive. Calibration should not be used to tune the final causal result; it should be used only to initialize reliability and verify the label ontology. The third step is verifier design. The verifier should contain only rules that domain experts are willing to defend. Examples include sign consistency, minimum coverage, minimum source diversity, and mutually exclusive physical claims. It should not contain rules that force politically or commercially convenient conclusions. A useful verifier is conservative: it catches incoherence, but it does not decide the answer when evidence is genuinely ambiguous. The fourth step is monitoring. After deployment, the analyst should track the distribution of coverage ratios, conflict scores, source weights, rejected windows, and first-stage strength. A sudden collapse in one source weight may reflect genuine source deterioration, but it may also reveal a domain shift outside the calibration set. In either case, the system should surface the change rather than silently adapting in a way that hides the underlying evidence problem. The final step is human review for unresolved windows. MissingNarrative-AEDL is designed to reduce review burden, not eliminate it. When coverage is low and conflict is high, the correct engineering response is often abstention. In a causal inference workflow, abstention has a positive role: it prevents the construction of instruments that look numerically precise but lack evidentiary support.

TABLE V. Deployment and review checklist

Item	Question	Action
Registry	Independent?	Deduplicate
Calibration	Edge cases?	Add audits
Verifier	Valid rules?	Expert review
Monitoring	Abrupt drift?	Inspect
Abstention	Too thin?	Suppress/review

This protocol makes the proposed method suitable for engineering teams that must document how an AI-assisted causal signal was produced. The final estimate is not a mysterious output of a language model. It is the result of a source registry, an auditable calibration set, a transparent reliability update, and a conservative econometric interface.

VI. CONCLUSION

This paper introduced MissingNarrative-AEDL, a reliability-aware causal inference framework for missing and conflicting event evidence. Building positively on the AEDL idea of neuro-symbolic event-driven labeling [16], the proposed method adds source reliability learning, coverage-aware attenuation, conflict penalties, and symbolic verification before narrative instruments enter an econometric causal model. Reproducible experiments show that the method improves macro F1 from 0.755 under majority voting and 0.870 under reliability-

unaware AEDL to 0.920, while preserving strong first-stage relevance. The central lesson is that causal narrative instruments should not merely ask what the evidence says; they must also ask who says it, how often that source is reliable, how much evidence is missing, and whether remaining claims are mutually coherent.

References

A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, 2017.

D. Teng, "TEAS: Token- and Energy-Aware Autoscaling for Cost-Efficient LLM Serving," AI and Data Science Journal, 2025

R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.

K. Mohan, J. Pearl, and J. Tian, "Graphical models for inference with missing data," in Advances in Neural Information Processing Systems, 2013.

A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," Annals of Mathematical Statistics, vol. 38, no. 2, pp. 325-339, 1967.

G. Shafer, A Mathematical Theory of Evidence. Princeton, NJ, USA: Princeton University Press, 1976.

J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2009.

J. L. Montiel Olea, J. H. Stock, and M. W. Watson, "Inference in structural vector autoregressions identified with an external instrument," Journal of Econometrics, vol. 225, no. 1, pp. 74-87, 2021.

K. Mertens and M. O. Ravn, "The dynamic effects of personal and corporate income tax changes in the United States," American Economic Review, vol. 103, no. 4, pp. 1212-1247, 2013.

L. Kilian, "Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market," American Economic Review, vol. 99, no. 3, pp. 1053-1069, 2009.

C. Baumeister and J. D. Hamilton, "Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks," American Economic Review, vol. 109, no. 5, pp. 1873-1910, 2019.

D. Caldara and M. Iacoviello, "Measuring geopolitical risk," American Economic Review, vol. 112, no. 4, pp. 1194-1225, 2022.

R. J. Shiller, "Narrative economics," American Economic Review, vol. 107, no. 4, pp. 967-1004, 2017.

D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982-3992.

Y. Liang, Y. Jiao, W. Ping, H. Fan, and X. Han, "Adaptive Event-Driven Labeling: A Neuro-Symbolic Multiagent Framework for Causal Inference in Non-Stationary Time Series," IEEE Access, 2026.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.

J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, 2022.

Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," arXiv:2305.14325, 2023.

N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," in Advances in Neural Information Processing Systems, 2023.

J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, "Causal machine learning: A survey and open problems," arXiv:2206.15475, 2022.

P. Spirtes, C. Glymour, and R. Scheines, Causation, Prediction, and Search, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.
D. M. Chickering, "Optimal structure identification with greedy search," Journal of Machine Learning Research, vol. 3, pp. 507-554, 2002.
J. Peters, D. Janzing, and B. Scholkopf, Elements of Causal Inference: Foundations and Learning Algorithms. Cambridge, MA, USA: MIT Press, 2017.
S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," Journal of Statistical Software, vol. 45, no. 3, pp. 1-67, 2011.
J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in Proc. 35th International Conference on Machine Learning, 2018, pp. 5689-5698.
B. Zi, "Cloud-native distributed systems for real-time payment intelligence," AI and Data Science Journal, 2024.