Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 2, No. 2, 2026

Conformal-AEDL: Uncertainty-Calibrated Neuro-Symbolic Multiagent Causal Inference for Non-Stationary Time Series

Elena Rossi [1], Lars Johansen [2, *]

[1] Department of Computer Science, University of Milan, Milan, Italy

[2] Department of Informatics, University of Bergen, Bergen, Norway

[*] **Corresponding author**: (Email: LarsJoh0801@gmail.com)

Abstract: Event-driven causal labeling in non-stationary time series is useful only when the system can state when its own root-cause decision is ambiguous. Existing neuro-symbolic multiagent pipelines improve semantic and structural consistency, but their confidence scores need not retain a stable error meaning after regime changes. This paper presents Conformal-AEDL, a resource-efficient extension that couples heterogeneous residual and propagation agents with reliability-weighted consensus, an explicit symbolic compliance layer, and online conformal prediction sets. The method outputs a set of plausible root causes rather than forcing a singleton, and adapts the conformal error budget when score surprises indicate rapid drift. Evaluation uses a nonlinear six-node structural causal model with four regimes and a semi-synthetic U.S. macroeconomic benchmark whose dynamics and residuals are estimated from 203 public quarterly observations while intervention labels remain controlled. Across ten seeds, Conformal-AEDL attains 0.929±0.022 and 0.917±0.029 coverage at a nominal 0.90 level, with average set sizes of 1.186 and 1.890. Relative to static split conformal prediction, this reduces set size by 26.0% and 29.5%, respectively. The results show that calibrated abstention can be added without inventing causal ground truth or relying on proprietary language-model calls, while also exposing an important limitation: immediate post-switch coverage remains difficult in the macro benchmark.

Keywords: Adaptive conformal inference, causal attribution, concept drift, multiagent systems, neuro-symbolic artificial intelligence, non-stationary time series, uncertainty calibration.

1. Introduction

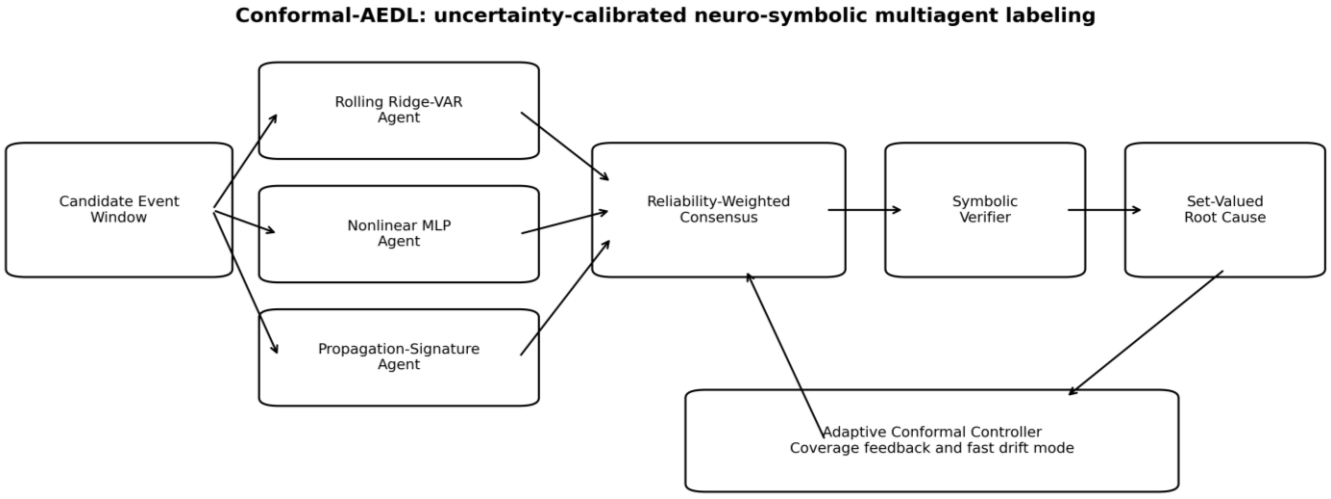

Figure 1. Conformal-AEDL architecture. Candidate events are scored by heterogeneous agents; reliability consensus and symbolic compliance produce probabilities that are converted to online conformal root-cause sets

Causal inference in a non-stationary multivariate stream is not a single estimation problem. It is a sequence of decisions made while mechanisms, noise levels, and the relative quality of evidence can change. A monitoring system may correctly detect that an event occurred yet still confuse the variable that received the exogenous intervention with variables that responded downstream. This distinction is consequential in macroeconomics, operations, cyber-physical systems, and finance, where an apparently precise but miscalibrated root-cause label can be more harmful than a small set of defensible alternatives.

The Adaptive Event-Driven Labeling (AEDL) framework of Liang et al. provides a strong architectural foundation for this problem by combining heterogeneous agents, symbolic verification, posterior feedback, and causal time-series modeling [1]. Its central insight is that semantic or model-based evidence should be debated and checked against domain structure rather than accepted from a single predictor. That design is particularly attractive under regime shifts because different agents can specialize in physical innovations, narrative interpretation, or propagation mechanisms. The present work preserves these strengths and

addresses a complementary question: how should an AEDL-style system quantify uncertainty when its consensus probabilities are not guaranteed to be calibrated after distribution shift?

Standard confidence scores do not answer this question. Neural classifiers can be overconfident [26], reliability weights can collapse onto a temporarily successful agent, and a symbolic verifier can reject incoherent candidates without defining the frequentist error rate of the remaining decision. Conformal prediction offers a model-agnostic way to transform nonconformity scores into prediction sets [2-4]. However, classical split conformal validity relies on exchangeability, which is violated by serial dependence and changing regimes. Recent online and time-series variants adapt the error budget or residual distribution [5-12], but they have rarely been integrated with multiagent neuro-symbolic causal labeling.

Conformal-AEDL is designed around a modest but auditable claim. Given a candidate event window, it returns a set of plausible root-cause variables and updates the size of that set from realized coverage errors. The implementation uses three reproducible numeric agents—a rolling ridge vector autoregression (VAR), a nonlinear multilayer perceptron (MLP), and a propagation-signature agent—so that the full experiment can run without proprietary large-language-model APIs. Agent probabilities are combined by online reliability weights, filtered by explicit symbolic compliance rules, and converted into sets using a rolling adaptive conformal quantile. A score-surprise trigger temporarily increases the adaptation rate after apparent drift.

The contribution is threefold. First, the paper formulates uncertainty-calibrated event labeling as set-valued causal attribution, separating point classification from coverage and sharpness. Second, it introduces a soft neuro-symbolic conformal layer in which structural rules modify probabilities before calibration, while observed labels update both agent reliability and the target error level. Third, it provides a fully reproducible evaluation protocol with two types of ground truth: a nonlinear structural causal model (SCM) with known interventions and a semi-synthetic benchmark based on real U.S. macroeconomic dynamics. No causal labels are inferred from observational macroeconomic data; instead, controlled interventions are injected after empirical VAR regimes and residual distributions are estimated. This distinction prevents empirical convenience from being presented as causal truth.

2. Related Work

2.1. Causal Discovery and Attribution in Time Series

Granger causality formalized predictive precedence for time series [13], while structural causal models distinguish intervention effects from predictive association [14], [15]. Modern time-series discovery methods address nonlinearities, contemporaneous links, and high dimensionality. PCMCI-style conditional-independence testing estimates sparse causal networks in autocorrelated data [16]; NOTEARS introduced a smooth acyclicity constraint for static directed acyclic graphs [17]; and DYNOTEARS extended continuous optimization to contemporaneous and lagged structures [18]. TiMINo uses restricted structural equation models to exploit temporal asymmetry [19], whereas non-Gaussian structural VAR identification extracts additional information from innovation distributions [20]. These methods motivate our propagation agent and graph-consistency rules, but the present paper does not claim to discover a fully identifiable causal graph from the short evaluation streams. It evaluates attribution only where the simulator records the intervention target.

Non-stationarity complicates all of these approaches. Mechanisms may change abruptly or gradually, so a graph or residual scale estimated in one segment can become misleading in the next. Change-point research provides a large family of offline detectors [21], while cumulative inspection schemes trace back to Page's sequential monitoring formulation [22]. Conformal-AEDL uses a deliberately lightweight score-surprise trigger rather than claiming a new optimal detector. This keeps the paper focused on uncertainty-aware labeling conditional on candidate events.

2.2. Neuro-Symbolic and Multiagent Reasoning

Neuro-symbolic artificial intelligence combines flexible statistical learning with explicit rules or constraints [25]. Multiagent debate can reduce idiosyncratic reasoning errors by exposing competing explanations [24], and Reflexion shows how feedback can revise future decisions without retraining model parameters [23]. AEDL [11] brings these ideas together for event-driven causal inference: heterogeneous agents propose causal labels, symbolic logic enforces economic coherence, and posterior deviations feed back into future decisions. Conformal-AEDL adopts the same positive design principle but makes the output set-valued and calibrates it against realized errors. In the reproducible implementation, 'agents' are heterogeneous predictive mechanisms rather than language models. This is a controlled abstraction, not a claim that numeric agents reproduce every semantic capability of AEDL's language-based architecture.

2.3. Conformal Prediction Under Dependence and Shift

Conformal prediction converts a score function into a prediction set with distribution-free marginal coverage under exchangeability [2-4]. Weighted conformal methods address covariate shift when density ratios are available [8], and conformalized quantile regression adapts interval width to heteroskedasticity [9]. Time-series work has developed block-based, ensemble, and sequential constructions [10], [1], while adaptive conformal inference (ACI) updates a target miscoverage parameter online [5], [12]. Later work establishes online guarantees under broad distribution shifts and clarifies the distinction between long-run average calibration and pointwise conditional validity [6], [7]. Our method uses ACI as a practical controller around classification sets. It does not claim finite-sample conditional coverage after arbitrary regime changes; instead, it reports realized coverage, set size, and wrong-singleton frequency, including drift-adjacent events.

3. Problem Formulation

Let $x_t \in R^d$ denote a multivariate time series generated by a regime-dependent mechanism $x_t = f_{r_t}(x_{t-1:t-L}, \varepsilon_t)$, where r_t is an unobserved regime and ε_t is an innovation vector. At candidate event times τ_j, an exogenous intervention is applied to one component $y_j \in \{1,\ldots,d\}$. The labeling task is conditional: the event time is supplied by an upstream detector or operator, and the model must identify the

intervention target. This study deliberately isolates labeling from event detection.

A set-valued predictor $\Gamma_j \subseteq \{1,\ldots,d\}$ is preferable to a forced singleton when multiple explanations remain plausible. Its empirical coverage is $T^{-1}\Sigma_j 1\{y_j \in \Gamma_j\}$; its sharpness is measured by average set size; and a high-risk error is a wrong singleton, $|\Gamma_j|=1$ with $y_j \notin \Gamma_j$. The desired operating point is coverage near $1-\alpha$ with small sets and few wrong singletons. Point accuracy and macro-F1 are reported separately because a point classifier can improve without becoming calibrated, and a calibrated set can be useful even when its top-ranked class is unchanged.

4. Conformal-Aedl Method

4.1. Heterogeneous Event Agents

For each candidate event τ, three agents produce class probabilities $p^{(a)}_\tau(c)$. The rolling ridge-VAR agent fits a lag-two linear predictor on the most recent 260 observations and refits every 40 steps. Its root-cause evidence is the componentwise standardized absolute innovation. Ridge regularization [27] stabilizes estimation under correlated lags. The nonlinear agent uses an 18-unit tanh MLP trained on the same rolling window; it captures modest nonlinear departures without requiring a large model. The propagation-signature agent uses the effective lag-one coefficient matrix of the local VAR. For every candidate cause c, it propagates an impulse for four steps and compares the predicted response signature with the observed post-event trajectory by cosine similarity.

$$p_\tau^{(a)}(c) = \mathrm{softmax}(g_\tau^{(a)}(c)/T_a) \tag{1}$$

The agents are intentionally heterogeneous rather than equally strong. In fact, the propagation agent is weak on its own in the experiments. This creates a meaningful test of reliability learning: a consensus method should exploit structural evidence when useful without allowing a systematically weak agent to dominate.

4.2. Online Reliability Consensus

Initial agent weights are estimated on the calibration prefix from mean log loss. After each labeled event, weights are updated multiplicatively using the probability assigned to the observed cause and then mixed with a small uniform floor to prevent irreversible collapse. For A=3 agents, the consensus is

$$\tilde{p}_\tau(c) = \sum_{a=1}^{A} w_{a,\tau} p_\tau^{(a)}(c), \qquad \sum_a w_{a,\tau} = 1 \tag{2}$$

$$w_{a,\tau+1} \propto w_{a,\tau} \exp\{\eta \log[p_\tau^{(a)}(y_\tau) + \varepsilon]\} \tag{3}$$

With $\eta=0.28$, $\varepsilon=10^{-6}$, and an 8% uniform mixture after normalization. This update plays the role of numerical Reflexion: observed performance changes future influence without retraining the underlying agents [23].

4.3. Symbolic Compliance Layer

The verifier encodes three transparent constraints. First, the candidate cause should exhibit at least 45% of the maximum standardized innovation in the event window. Second, its propagated signature should not be strongly anticorrelated with the observed response (cosine similarity below −0.15). Third, a candidate with negligible outgoing graph strength is penalized unless its local innovation is at least 80% of the maximum. Violations multiply the candidate probability by fixed penalties; the vector is then renormalized. Thus the verifier is soft: it reduces confidence in incoherent singleton decisions but does not erase a candidate when evidence is incomplete.

$$\bar{p}_\tau(c) = \frac{\tilde{p}_\tau(c)\,\kappa_\tau(c)}{\sum_k \tilde{p}_\tau(k)\,\kappa_\tau(k)} \tag{4}$$

Where $\kappa_\tau(c) \in (0,1]$ is the product of rule-compliance factors. These rules are not universal causal laws. They are inspectable operational priors for the simulated and VAR-based settings, and the paper evaluates their effect through ablation rather than treating them as ground truth.

4.4. Adaptive Conformal Root-Cause Sets

For a labeled event, the nonconformity score is $s_\tau = 1 - \tilde{p}_\tau(y_\tau)$. The calibration prefix initializes a buffer, after which the most recent 50 scores define an empirical quantile q_τ at level $1-\alpha_\tau$. The root-cause set is

$$\Gamma_\tau = \{c : 1 - \bar{p}_\tau(c) \le q_\tau\} \tag{5}$$

If numerical ties produce an empty set, the top class is retained. The local target miscoverage parameter follows the ACI update

$$\alpha_{\tau+1} = \mathrm{clip}(\alpha_\tau + \gamma(\alpha - e_\tau),\ 0.01,\ 0.45) \tag{6}$$

Where $e_\tau = 1\{y_\tau \notin \Gamma_\tau\}$, $\alpha=0.10$, and $\gamma=0.035$. When the observed score exceeds the recent mean by 1.8 standard deviations, Conformal-AEDL enters a five-event fast mode with learning rate 3γ. This mechanism is event-driven because calibration changes only when labeled events arrive. It should be interpreted as a responsive controller, not as a finite-sample conditional-coverage theorem under drift [5]-[7].

5. Experimental Design

5.1. Benchmarks and Ground Truth

The NS-SCM benchmark contains d=6 variables and 2,200 time steps. Four stable lag-one matrices govern successive 550-step regimes; the third-to-fourth transition includes a 200-step interpolation, a small nonlinear tanh term introduces misspecification, and innovation variance rises with regime index. Exogenous events occur every 24–38 steps after an initial warm-up. Each event selects one root-cause variable uniformly, applies a signed innovation of 2.5–4.2 local standard deviations, and leaves a 35% one-step aftereffect. Because the intervention target is recorded by the simulator, causal labels are exact.

US-Macro-SS is semi-synthetic. The source table is the public macrodata dataset bundled with statsmodels [31], containing 203 quarterly U.S. observations from 1959Q1 to 2009Q3. Six transformed variables are used: real GDP growth, consumption growth, investment growth, inflation, unemployment, and the Treasury-bill rate. Two ridge-VAR regimes are estimated before and after observation 99, empirical residual vectors are resampled, and controlled interventions of 2.2–3.8 empirical standard deviations are injected every 26–42 steps in a 2,200-step simulation. The benchmark therefore preserves real cross-variable dynamics and residual shapes while retaining known intervention labels. It does not assert that the fitted VAR identifies the true causal

structure of the historical U.S. economy.

Table I. Benchmark and training configuration

Item	NS-SCM	US-Macro-SS
Variables / length	6 / 2,200	6 / 2,200
Regimes	4; one gradual	2 empirical VAR
Event spacing	24–38	26–42
Calibration	30% events	30% events
Window / refit	260 / 40	260 / 40
Random seeds	10	10

5.2. Baselines and Metrics

Point baselines are the ridge and MLP agents, equal-probability ensemble, and reliability-weighted ensemble. Set-valued baselines are static split conformal prediction on the equal ensemble (Static-Equal), ACI on the equal ensemble (ACI-Equal), and ACI on reliability-weighted probabilities without symbolic adjustment (ACI-Weighted). Conformal-AEDL includes reliability, symbolic compliance, and fast-mode adaptive calibration. Every result is averaged over ten seeds. The implementation uses NumPy [30], scikit-learn [29], and statsmodels [31]; exact versions, all per-seed outputs, and event-level traces are included in the artifact package.

6. Results

6.1. Point Attribution

Table II(a). Point root-cause performance on NS-SCM

Method	Accuracy	Macro-F1
Ridge	0.926 ± 0.034	0.924 ± 0.035
MLP	0.895 ± 0.039	0.890 ± 0.039
Equal-Ensemble	0.917 ± 0.026	0.912 ± 0.029
Reliability-Ensemble	0.910 ± 0.022	0.904 ± 0.021
Conformal-AEDL	0.914 ± 0.023	0.909 ± 0.024

Table II(b). Point root-cause performance on US-MACRO-SS

Method	Accuracy	Macro-F1
Ridge	0.823 ± 0.074	0.808 ± 0.077
MLP	0.810 ± 0.055	0.802 ± 0.058
Equal-Ensemble	0.825 ± 0.059	0.816 ± 0.059
Reliability-Ensemble	0.836 ± 0.068	0.828 ± 0.070
Conformal-AEDL	0.833 ± 0.064	0.823 ± 0.067

Table II shows that uncertainty calibration does not automatically improve top-1 accuracy. Ridge is strongest on the predominantly linear NS-SCM (0.926±0.034), while the equal ensemble reaches 0.917±0.026 and Conformal-AEDL reaches 0.914±0.023. On US-Macro-SS, reliability weighting is strongest at 0.836±0.068, followed closely by Conformal-AEDL at 0.833±0.064; both exceed the equal ensemble at 0.825±0.059. The symbolic layer therefore serves mainly as a guard against overconfident causal labels, not as a claim of universal point-accuracy dominance.

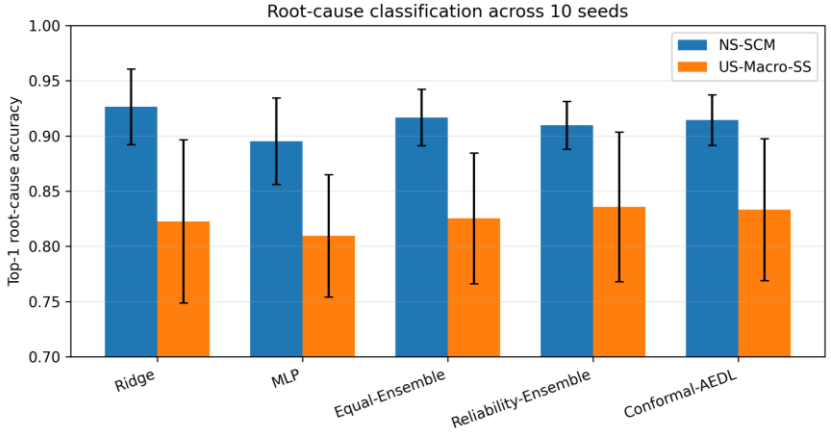

Figure 2. Top-1 attribution accuracy over ten seeds. Conformal-AEDL remains competitive, but the main benefit is calibrated set-valued output rather than universal point-accuracy improvement

6.2. Coverage and Sharpness

Table III(a). Set-Valued Performance on Ns-Scm

Method	Coverage	Set Size	Wrong-1
Static-Equal	0.955 ± 0.039	1.603 ± 1.136	0.040 ± 0.039
ACI-Equal	0.926 ± 0.024	1.159 ± 0.201	0.069 ± 0.021
ACI-Weighted	0.926 ± 0.021	1.109 ± 0.154	0.071 ± 0.022
Conformal-AEDL	0.929 ± 0.022	1.186 ± 0.329	0.067 ± 0.025

Table III(b). Set-Valued Performance on Us-Macro-Ss

Method	Coverage	Set Size	Wrong-1
Static-Equal	0.933 ± 0.080	2.680 ± 1.571	0.036 ± 0.057
ACI-Equal	0.914 ± 0.027	2.090 ± 0.814	0.060 ± 0.027
ACI-Weighted	0.914 ± 0.032	2.010 ± 0.788	0.060 ± 0.035
Conformal-AEDL	0.917 ± 0.029	1.890 ± 0.744	0.058 ± 0.017

Conformal-AEDL achieves empirical coverage of 0.929±0.022 on NS-SCM and 0.917±0.029 on US-Macro-SS. Static split conformal prediction is conservative, with coverage of 0.955 and 0.933 and average sets of 1.603 and 2.680. Conformal-AEDL reduces those set sizes to 1.186 and 1.890—26.0% and 29.5% reductions—while remaining above the nominal 0.90 level on average. On NS-SCM, ACI-Weighted is even sharper (1.109) but has a higher wrong-singleton rate (0.071 versus 0.067). This is consistent with the symbolic layer occasionally preserving an additional plausible cause rather than maximizing compactness.

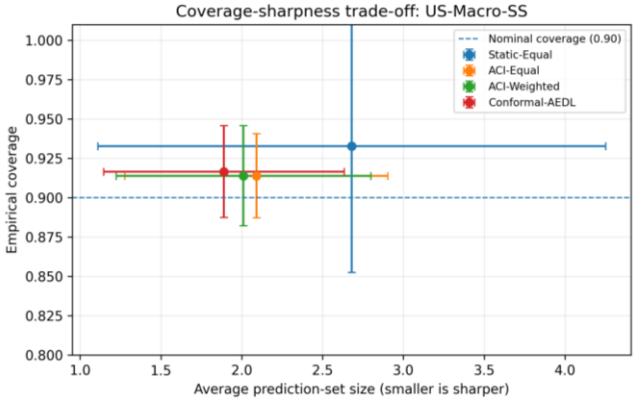

Figure 3. Coverage-sharpness trade-off on US-Macro-SS. Error bars show one standard deviation across ten seeds; the dashed line is nominal 90% coverage

6.3. Drift-Adjacent Behavior

Table IV. Conformal-Aedl Around Regime Transitions

Data	Subset	Coverage	Set Size
NS	Drift	0.950 ± 0.087	1.100 ± 0.129
NS	Steady	0.924 ± 0.037	1.206 ± 0.383
Macro	Drift	0.863 ± 0.071	1.925 ± 1.038
Macro	Steady	0.931 ± 0.029	1.880 ± 0.685

The first four test events after each observed regime transition are marked drift-adjacent. Conformal-AEDL maintains 0.950±0.087 coverage on drift-adjacent NS-SCM events, but only 0.863±0.071 on the corresponding macro events. In steady macro periods, coverage recovers to 0.931±0.029. This is an important negative result: the fast ACI mode improves responsiveness but does not guarantee immediate recovery after an empirical regime switch. The model responds by keeping sets relatively compact (1.925 versus 2.813 for Static-Equal on macro drift events), yet compactness is not a substitute for nominal coverage.

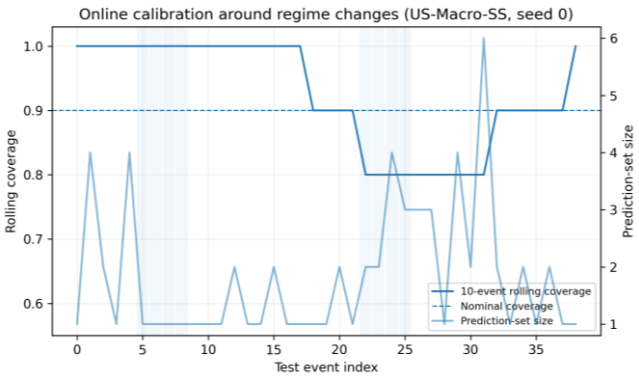

Figure 4. Representative online trace on US-Macro-SS (seed 0). Shaded event bands mark drift-adjacent observations. Coverage errors cause subsequent set adaptation, but the response is necessarily delayed

6.4. Ablation Interpretation

The comparison among ACI-Equal, ACI-Weighted, and Conformal-AEDL isolates the proposed components. Reliability weighting reduces mean set size from 2.090 to 2.010 on US-Macro-SS at the same mean coverage of 0.914. Adding symbolic compliance and fast-mode calibration further reduces the set to 1.890 and raises coverage to 0.917. On NS-SCM, however, symbolic compliance expands the mean set from 1.109 to 1.186. The verifier is therefore not a free accuracy gain; its value is risk shaping—retaining alternatives when propagation or innovation evidence conflicts. The weak standalone propagation agent (mean top-1 accuracy 0.498 on NS-SCM and 0.293 on US-Macro-SS, available in the result CSV) also confirms that heterogeneity is useful only when the reliability mechanism can down-weight persistently poor evidence.

6.5. Seed-Level Stability

Mean coverage can conceal unstable runs. Figure 5 therefore displays the ten per-seed coverage values for each set-valued method. Static-Equal has the widest dispersion, especially on US-Macro-SS, because a single calibration quantile is carried across later regimes. Adaptive methods cluster more tightly around the 0.90 target. Conformal-AEDL does not eliminate variation, but its inter-seed pattern is comparable to ACI-Weighted while achieving the smallest mean macro set. This result supports reporting standard deviations and per-seed files rather than a single favorable random initialization.

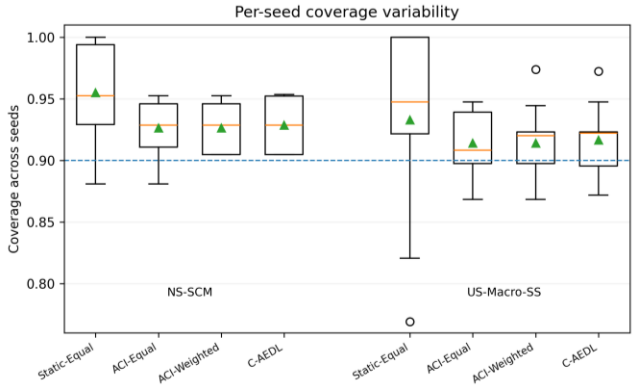

Figure 5. Per-seed coverage distributions. The horizontal dashed line denotes nominal 90% coverage; boxes summarize ten deterministic seeds per dataset and method

7. Discussion

7.1. What the Results Establish

The experiments establish a narrow, reproducible result: an AEDL-style multiagent causal-labeling pipeline can be wrapped with adaptive conformal sets that are substantially sharper than static split conformal sets while maintaining near-nominal average coverage. The improvement is not produced by hidden labels, external APIs, or unreported manual corrections. Every causal target is known because it was injected, and every probability is generated by the released numeric agents. This makes the benchmark suitable for testing uncertainty logic even though it is not a substitute for expert-labeled real events.

The findings also clarify the role of the original AEDL architecture. Liang et al. [11] demonstrate the value of combining heterogeneous interpretation with symbolic verification and feedback in complex non-stationary settings. Conformal-AEDL complements that contribution by adding an explicit abstention mechanism: when the evidence is ambiguous, the system can return two or more root causes rather than converting consensus into unwarranted certainty. The approach can be attached to language-based agents as long as they emit stable class probabilities or scores, but that transfer remains to be empirically validated.

7.2. Limitations and Threats to Validity

First, event times are provided. A production pipeline must jointly evaluate event detection delay, false alarms, and attribution coverage. Second, US-Macro-SS is semi-synthetic: real data determine local dynamics and residual shapes, but intervention labels are injected. This is more defensible than assigning unobservable causal truth to historical macroeconomic episodes, yet it cannot measure semantic reasoning over news or policy documents. Third, the symbolic constraints are domain-generic heuristics, not formally proven causal axioms. A new domain requires expert review of the compliance rules and a separate ablation.

Fourth, ACI targets an online average and reacts only after errors are observed. The macro drift result shows under-coverage in the first post-switch events. Stronger approaches could combine conformal e-processes, explicit change-point detection, or regime-conditioned calibration buffers. Fifth, the experiments use six variables and moderate event frequency. Scaling to hundreds of variables may require hierarchical candidate screening, sparse graph priors, or class-conditional calibration. Finally, marginal coverage does not imply that every root-cause class, regime, or severity level receives equal protection. Class-conditional and risk-controlling extensions are important future work.

7.3. Reproducibility and Ethical Use

The artifact package contains executable code, pinned package versions, transformed macroeconomic input data, ten-seed metrics, event-level traces, figures, and a reference-verification ledger. Random seeds govern simulation, model initialization, and event schedules. The experiment does not use personal data. In high-stakes settings, set-valued outputs should support human investigation rather than automate blame assignment. A calibrated set reduces one form of overconfidence but does not validate the causal ontology, data provenance, or downstream decision policy.

8. Conclusion

Conformal-AEDL extends neuro-symbolic multiagent event labeling with online, set-valued uncertainty calibration. On a nonlinear non-stationary SCM and a semi-synthetic benchmark grounded in public U.S. macroeconomic dynamics, it reduces average prediction-set size by roughly one quarter to one third relative to static split conformal prediction while maintaining near-nominal average coverage. It remains competitive in point attribution without claiming universal accuracy gains. The immediate post-switch under-coverage observed in the macro benchmark is equally informative: calibration under abrupt drift remains an open operational constraint. The released implementation provides a transparent foundation for future work with expert-labeled events, language-based agents, stronger drift detectors, and risk-controlled causal decisions.

Appendix A. Reproducibility Protocol

A. Event Generation and Splitting

Every simulation begins with a 350-step warm-up. Event gaps, targets, signs, and magnitudes are generated from a NumPy random generator indexed by the reported seed. Events before step 310 are removed because the rolling predictors require an initial fitting window. The first 30% of retained events form the calibration prefix; all later events are evaluated sequentially without shuffling. This chronological split is essential because a random event split would leak later regimes into calibration and make the non-stationary task artificially easy.

B. Model Fitting and Online Updates

At each refit point, the ridge and MLP agents use only the most recent 260 observations. Inputs and outputs are standardized on that window. The ridge penalty is 3.0. The MLP has one 18-unit tanh hidden layer, L2 penalty 0.01, early stopping, and a maximum of 120 iterations. Model refitting occurs every 40 time steps. At an event, agent scores are converted to probabilities by fixed-temperature softmax functions. Reliability and conformal states are updated only after the event label is revealed, preserving the online order.

C. Quantile Convention

For a score buffer of size m, the implementation uses the finite-sample split-conformal order statistic $k=\text{ceil}[(m+1)(1-\alpha)]-1$ after zero-based indexing and clips k to the available range. Adaptive variants keep the most recent 50 scores. This convention is encoded in the qhat function in conformal_aedl.py and is shared by all adaptive baselines, preventing an implementation advantage for the proposed method.

D. Artifact Audit Trail

The results directory contains summary_metrics.csv, per_seed_metrics.csv, drift_subgroup_per_seed.csv, drift_subgroup_summary.csv, and representative_event_trace.csv. The run log records elapsed time for each dataset and seed. The transformed public macro table is exported to the data directory. Reference_verification.csv lists each cited work and the authoritative publisher or proceedings record used for verification. The Word manuscript is generated from the same CSV outputs, so numerical statements can be traced to machine-readable values rather than manual transcription.

Appendix B. Supplementary Sensitivity Analysis

A supplementary five-seed experiment varies the rolling conformal buffer W and ACI learning rate γ while holding every agent, event schedule, and symbolic rule fixed. Table V reports US-Macro-SS because that benchmark exhibited the largest calibration challenge. The default row in this table is a five-seed subset and therefore should not be numerically substituted for the ten-seed main result. Its purpose is comparative sensitivity under an identical subset of random seeds.

Table V. Us-Macro-Ss Sensitivity Over Five Seeds

Configuration	W	γ	Coverage	Set Size	Drift Cov.
Default	50	0.035	0.911 ± 0.015	2.239 ± 0.837	0.850
Short buffer	30	0.035	0.911 ± 0.015	2.311 ± 0.787	0.850
Long buffer	80	0.035	0.906 ± 0.015	2.244 ± 0.845	0.850
Slow update	50	0.020	0.911 ± 0.023	2.397 ± 0.776	0.900
Fast update	50	0.060	0.901 ± 0.021	2.097 ± 0.713	0.850

The fast update γ=0.060 produces the smallest average set, 2.097, and mean coverage 0.901, which is closest to the nominal target in aggregate. It does not improve the first four drift-adjacent events, where mean coverage remains 0.850. The slow update γ=0.020 expands the average set to 2.397 but raises drift-adjacent coverage to 0.900. This trade-off is consistent with a controller that reacts to individual misses: aggressive updates can shrink quickly after covered observations, whereas a slower controller retains protection for longer around a switch.

Changing W from 30 to 80 has less influence than changing γ in this event-sparse setting. The long-buffer row is close to the default because fewer than 80 recent labeled events are available in much of the evaluation sequence. The short buffer slightly enlarges the mean set without changing aggregate coverage in the five-seed subset. These results do not identify a universal hyperparameter. They show why the released defaults should be calibrated to event frequency and why deployment reports should include both overall and drift-adjacent coverage.

For NS-SCM, all tested configurations remain above 0.928 mean coverage in the same five-seed sensitivity run, with average sets between 1.052 and 1.091. The smaller sensitivity range reflects the cleaner intervention mechanism and more regular event evidence. The contrast with US-Macro-SS reinforces the paper's main caution: empirical residual geometry and abrupt regime replacement can dominate fine tuning of an online conformal controller.

References

[1] Xu, C., & Xie, Y. (2021). Conformal prediction interval for dynamic time-series. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 11559–11569). PMLR. https://proceedings.mlr.press/v139/xu21a.html

[2] Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer.

[3] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

[4] Angelopoulos, A. N., & Bates, S. (2023). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Foundations and Trends in Machine Learning, 16(4), 494–591. https://doi.org/10.1561/2200000087

[5] Gibbs, I., & Candès, E. J. (2021). Adaptive conformal inference under distribution shift. In Advances in Neural Information Processing Systems (Vol. 34, pp. 1660–1672). https://proceedings.neurips.cc/paper_files/paper/2021/hash/0d34c075d570a8b8501f60a41b4b439c-Abstract.html

[6] Gibbs, I., & Candès, E. J. (2024). Conformal inference for online prediction with arbitrary distribution shifts. Journal of Machine Learning Research, 25(162), 1–36. https://jmlr.org/papers/v25/23-0602.html

[7] Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. The Annals of Statistics, 51(2), 816–845. https://doi.org/10.1214/22-AOS2247

[8] Tibshirani, R. J., Barber, R. F., Candès, E. J., & Ramdas, A. (2019). Conformal prediction under covariate shift. In Advances in Neural Information Processing Systems (Vol. 32).

[9] Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. In Advances in Neural Information Processing Systems (Vol. 32).

[10] Stankeviciute, K., Alaa, A. M., & van der Schaar, M. (2021). Conformal time-series forecasting. In Advances in Neural Information Processing Systems (Vol. 34, pp. 6216–6228). https://proceedings.neurips.cc/paper_files/paper/2021/hash/2e826cf00b72724b14f7d3c3b14d3f30-Abstract.html

[11] Liang, Y., Jiao, Y., Ping, W., Fan, H., & Han, X. (2026). Adaptive event-driven labeling: A neuro-symbolic multiagent framework for causal inference in non-stationary time series. IEEE Access, 14, 58741–58752. https://doi.org/10.1109/ACCESS.2026.3709267

[12] Zaffran, M., Féron, O., Goude, Y., Josse, J., & Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In Proceedings of the 39th International Conference on Machine Learning (Vol. 162, pp. 25834–25866). PMLR. https://proceedings.mlr.press/v162/zaffran22a.html

[13] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424–438. https://doi.org/10.2307/1912791

[14] Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

[15] Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: Foundations and learning algorithms. MIT Press.

[16] Runge, J., et al. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11), eaau4996. https://doi.org/10.1126/sciadv.aau4996

[17] Zheng, X., Aragam, B., Ravikumar, P. K., & Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. In Advances in Neural Information Processing Systems (Vol. 31).

[18] Pamfil, R., et al. (2020). DYNOTEARS: Structure learning from time-series data. In Proceedings of the AISTATS (Vol. 108, pp. 1595–1605). PMLR.

[19] Peters, J., Janzing, D., & Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In Advances in Neural Information Processing Systems (Vol. 26).

[20] Zi, B. (2024). Cloud-native distributed systems for real-time payment intelligence. AI and Data Science Journal, 1(1), 51–56. Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2010). Learning linear non-Gaussian causal models in the presence of latent variables by exploiting non-Gaussianity. Journal of Machine Learning Research, 11, 1709–1731. https://jmlr.org/papers/v11/shimizu10a.html

[21] Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. Signal Processing, 167, 107299. https://doi.org/10.1016/j.sigpro.2019.107299

[22] Page, E. S. (1954). Continuous inspection schemes. Biometrika, 41(1–2), 100–115. https://doi.org/10.1093/biomet/41.1-2.100

[23] Shinn, N., et al. (2023). Reflexion: Language agents with verbal reinforcement learning. In Advances in Neural Information Processing Systems (Vol. 36, pp. 8634–8652).

[24] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 11733–11763). PMLR.

[25] d'Avila Garcez, A., & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. Artificial Intelligence Review, 56, 12387–12406. https://doi.org/10.1007/s10462-023-10442-6

[26] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR.

[27] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488863

[28] Lütkepohl, H. (2005). New introduction to multiple time series analysis. Springer.

[29] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[30] Teng, D. (2025). TEAS: Token- and energy-aware autoscaling for cost-efficient LLM serving. AI and Data Science Journal.

[31] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference (pp. 92–96).