Integrating Data and Causal Reasoning to Automate Root Cause Analysis

D. Mandelli [a*], C. Wang [a]

[a] Idaho National Laboratory, 2525 N. Fremont Ave., 83415 Idaho Falls (ID), USA

Abstract: This paper presents a structured, data-driven framework for automating root cause analysis (RCA) in complex nuclear power plant environments. Traditional RCA depends on manual interpretation of heterogeneous data sources (i.e., telemetry, maintenance records, and system documentation), making the process time-consuming, expertise-dependent, and vulnerable to inconsistencies. These challenges are amplified by system interdependencies, incomplete or ambiguous data, and unclear temporal relationships between causes and effects. The proposed framework integrates plant telemetry, historical records, and system architecture into a unified "hypothesize-test-compare-decide" workflow. It combines three complementary reasoning dimensions: structural plausibility, temporal coherence, and historical evidence. A knowledge graph derived from model-based systems engineering constrains hypotheses to feasible components and failure modes, while a vector-based retrieval system links candidate explanations to relevant documentation and supporting information. The method generates, evaluates, and ranks multiple causal hypotheses, explicitly representing uncertainty and alternative possibilities. Evidence is categorized and used to refine rankings, ensuring traceability and engineering interpretability throughout the RCA process. A representative case study is included to demonstrate how jointly analyzing timing, system structure, and supporting evidence improves diagnostic rigor, consistency, and transparency. The results illustrate how integrated causal reasoning can enhance RCA efficiency while still preserving expert oversight.

Keywords: root cause analysis, artificial intelligence, temporal patterns, causal reasoning

1. INTRODUCTION

Root cause analysis (RCA) is an important task at nuclear power plants. In particular, 10 CFR 50, Appendix B, Criterion XVI, requires measures to ensure that conditions adverse to quality are promptly identified and corrected. For significant conditions adverse to quality, the cause must be determined, corrective action taken to preclude repetition, and the condition, cause, and corrective action documented and reported [1].

Motivated by issues discussed in RCA guidance and observed in manual investigation workflows [2, 3], this paper targets three recurring process vulnerabilities. The first is fixation on the most salient symptom: analysts converge on the most recent alarm or most visible damage, without systematically testing whether that signal is a cause or a downstream consequence. Because equipment failures cascade, downstream effects are frequently louder than upstream initiating conditions, and an analyst focused on the loudest signal may document a symptom but leave the root cause unexamined. The second vulnerability is incomplete causal coverage: investigations can close as soon as a plausible proximate cause is identified, without evaluating the contributing and root-cause categories covered in a thorough investigation. The third vulnerability is inconsistent retrieval of plant history: the condition reports and operating experience (OE) records most relevant to a current event are rarely retrieved in a structured way (e.g., an analyst may not locate a prior RCA documenting an identical failure mode on the same component family, even when that record exists in the corrective action database).

With this in mind, this paper presents DACKAR [1], a structured diagnostic pipeline for equipment

* Corresponding author, diego.mandelli@inl.gov

[1] Digital Analytics, Causal Knowledge Acquisition and Reasoning (DACKAR) GitHub repository: https://github.com/idaholab/DACKAR

failure RCA at nuclear power plants. It is designed to support equipment-related RCA within that context by organizing candidate causes into proximate, contributing, and root-cause layers commonly used in RCA practice [2]. DACKAR ensures that all defined causal categories are evaluated, that every conclusion is linked to a specific evidence source, and that coverage gaps cannot be silently omitted. The generated output is a structured assessment card that an analyst can hand to supervision and that a resident inspector can audit. The novelty of DACKAR lies not in automated fault diagnosis but in deterministic enforcement of causal coverage completeness and traceable evidence accountability, converting investigative rigor from a human discipline issue into a process-verifiable property.

2. THE DACKAR FRAMEWORK

This section describes the DACKAR analysis workflow (i.e., how evidence is ingested, hypotheses are scored, and causal coverage is enforced). A full description of the system architecture, integration interfaces, and deployment infrastructure is provided in [4].

DACKAR operates in four sequential stages:

1. Evidence ingestion: telemetry, plant documents, the plant model, and CMMS data are loaded and indexed
2. Multi-line causal assessment: five parallel investigation lines evaluate each active hypothesis against the available evidence
3. Coverage enforcement: the pipeline verifies that all 12 DACKAR causal categories have been evaluated, and it flags any that have not
4. Structured output generation: a machine-readable RCA assessment card and run manifest are produced for analyst review.

3. DESIGN OBJECTIVES

DACKAR is designed around three operational objectives that correspond directly to the three failure modes defined in Section 1. Against investigative fixation, the pipeline evaluates all plausible causal hypotheses in parallel rather than sequentially, and applies a scoring framework that weights evidence by quality and source authority rather than by prominence or recency.

Against coverage gaps, the pipeline enforces completion of all 12 DACKAR causal categories before producing an output. Any category that is applicable but not evaluated generates a structured attention flag that blocks investigation closure until the analyst resolves it.

Against inconsistent history use, the pipeline systematically searches historical condition reports, work orders, past RCA records, and fleet OE databases, applying source authority weights and returning structured evidence citations rather than raw document retrieval results.

4. INPUTS AND OUTPUTS

The developed workflow is such that it accepts four input categories, all produced during normal plant operations:

- Telemetry and historian data provide the time-stamped record of system parameters before, during, and after the event.
- Plant documents include condition reports, work orders, maintenance procedures, surveillance test records, and completed RCA records for prior similar events.
- The plant model is a structured representation of component physical and functional connections and known failure modes, as drawn from design documentation, FMEA records, and industry failure data.
- Operational context from the CMMS indicates the current maintenance status of affected equip-

ment: namely, whether PM tasks are current or overdue, whether open work orders exist, and whether the equipment has been recently returned from maintenance.

Plants without a formal plant model for the affected system can configure DACKAR so as to include a partial model covering the failed component and its immediate neighbors. This minimum viable configuration covers physical plausibility and timing analysis while also flagging the incomplete model as a data quality attention item, communicating the limitation explicitly rather than silently degrading the output.

The pipeline output is a structured RCA assessment card (i.e., the *rca_card*) containing the ranked hypothesis list with composite scores, evidence citations linking each conclusion to a specific source record, all active attention flags, the recommended corrective action category, and a full run manifest that links every score and weight to its source data and calculation.

5. ANALYTICAL ENGINES

The pipeline relies on four specialized data analysis engines that convert raw plant records into structured artifacts consumed by the scoring layer: named-entity recognition (NER), PM compliance, event similarity, and equipment similarity. These engines handle data extraction and structuring: they do not perform causal reasoning but produce the typed, scored inputs upon which the five investigation lines operate.

The NER module is designed to identify and extract specific entities (e.g., component IDs, failure modes, degradation phenomena) from documents such as condition reports and work orders, enabling structured retrieval beyond keyword matching. A causal relation extraction module identifies directed cause-effect triples from text by employing two extractors: one for formal engineering documents (with full dependency-parse handling of passive voice and embedded clauses) and one for informal CMMS entries (lighter-weight). Both extractors apply conjecture detection at the relation level, filtering modal-qualified statements from asserted causal links, and they normalize health status to a standard vocabulary (acceptable / degraded / failed / unknown) so that the evidence scores remain comparable across heterogeneous document sources.

The PM compliance engine assesses whether PM was current on the affected equipment at the time of the event. For each PM task linked to the failed component, it compares the last completed execution date against the scheduled interval in order to determine status (pass / fail / unknown) and compute the number of days overdue. A scope analysis maps each task to the failure mode identifiers it is designed to detect or prevent; a failure mode with no covering task generates a scope gap flag that feeds directly into Category J (inspection and testing program inadequacy) scoring. An effectiveness analysis complements the schedule check by computing the fraction of historical PM executions that recorded a “degraded as-found” condition, addressing whether prior inspections were actually detecting degradation prior to the functional failure. The summary fields (compliance rate, overall compliance status, maintenance-induced risk, and the primary-failure-mode scope gap flag) are consumed directly by the governance scoring dimension without requiring the causality engine to re-parse individual work order records.

The Event Similarity engine addresses the recurring RCA vulnerability of inconsistent use of plant history by searching internal records, fleet databases, and industry sources for any historical events that match the current one in regard to asset type, component identifier, failure mode, event type, and operating conditions. Retrieved events are ranked by similarity and tagged with a three-tier confidence multiplier (plant 1.00, fleet 0.80, industry 0.60) that scales their contribution to evidence scoring. A recurrence characterization is produced for each match: whether the current event represents a first occurrence, a known recurring pattern, or a recurrence after a prior corrective action was applied and closed. Recurrence after a supposedly effective prior corrective action serves as a direct input to the

Category L (systemic organizational weakness) assessment.

The Equipment Similarity engine extends the evidence retrieval scope beyond the specific failing component and to nominally identical sister equipment at other plants or in other trains, matching on component type, service conditions, and manufacturer or model. This extension is particularly relevant to design deficiency (Category H) and supply chain defect (Category K) hypotheses, which by nature affect every unit from the same design class or production batch. Sister equipment documents are added to the evidence retrieval query plan and weighted with the same three-tier confidence hierarchy as direct OE.

Temporal sequencing is evaluated using Allen interval algebra [5], which classifies each anomaly window relative to the failure event into one of five canonical relations: PRECEDES, OVERLAPS, CONTAINS, DURING, and FOLLOWS. Each relation carries a base causal support score ranging from 0.90 (PRECEDES, strongest causal support) to 0.10 (FOLLOWS, inconsistent with causation). A simultaneous-event epsilon window prevents spurious temporal contradictions arising from measurement uncertainty and accommodates fast-transient events. The engine also validates observed signal latency against failure mode and effect analysis (FMEA)-specified expected degradation windows, meaning that temporal scores reflect not only whether a signal preceded the failure, but also whether it did so at a physically plausible time.

The pipeline output stage uses a large language model (LLM)-based synthesizer to generate the natural-language narrative sections of the rca_card from the structured scoring results, evidence citations, and attention flags. A deterministic template fallback is always available: if the LLM is unavailable or returns output that fails validation, the synthesizer substitutes a structured template that populates all required rca_card fields from the run manifest directly. This ensures that the governed, machine-readable content of the rca_card (i.e., scores, citations, flags, and the run manifest) is always produced regardless of LLM availability, and that the determinism property applies to the assessment record even when the narrative prose varies between runs.

6. LINES OF INVESTIGATION

The pipeline applies five parallel lines of investigation to evaluate the current event evidence against the set of active causal hypotheses.

The *physical plausibility* line queries the plant model to establish which components are physically or functionally connected to the reported failure location, thus bounding the causal search space to plausible candidates. Only components appearing in the model as direct neighbors or known functional dependencies are considered, preventing spurious hypotheses and ensuring that non-obvious connections are not overlooked.

The *timing analysis* line examines time-stamped historian data to determine, one by one, whether each anomalous signal preceded the failure, was coincident with it, or appeared only afterward. A signal that first deviated from normal only after failure had occurred cannot be a cause; any hypotheses requiring that signal as a precursor are penalized accordingly. Anomalies appearing hours or days before the event are weighted as potential precursor indicators supporting a gradual-degradation hypotheses.

The *documentary evidence* line searches the full plant record (condition reports, work orders, past RCA records, and surveillance tests) by utilizing the equipment identifier, system classification, failure mode description, and component type, so that the relevant records are retrieved even when terminology differs. Contradictory records attenuate the hypothesis score rather than eliminating the hypothesis, raising uncertainty rather than forcing a binary exclusion.

The *PM compliance* line retrieves PM program records for the failed component and evaluates whether scheduled tasks were current at the time of failure. For overdue tasks, it assesses whether the specific overdue work corresponds to the failure mode being investigated and whether the PM intervals

are consistent with industry recommendations, flagging potential frequency nonconformances for contributing cause evaluation.

The *fleet and industry experience* line retrieves OE from the plant's internal history, fleet records from INPO IRIS, and industry databases, applying source authority weights: internal plant records (1.0), fleet INPO records (0.80), and broader industry records (0.60). Fleet records documenting the same failure mode on similar equipment serve as independent corroboration (or, in cases when the postulated mechanism was not observed under comparable circumstances, as attenuating evidence).

7. SCORING AND COVERAGE ENFORCEMENT

For each active causal hypothesis h, the pipeline computes a composite score from the following five dimensions:

$$S_h = w_{\text{struct}} \cdot S_{\text{struct}} + w_{\text{temp}} \cdot S_{\text{temp}} + w_{\text{tel}} \cdot S_{\text{tel}} + w_{\text{evid}} \cdot S_{\text{evid}} + w_{\text{gov}} \cdot S_{\text{gov}} \tag{1}$$

where weights $(w_{\text{struct}}, w_{\text{temp}}, w_{\text{tel}}, w_{\text{evid}}, w_{\text{gov}})$ sum to 1.0 and are category-specific:

- equipment-origin hypotheses (categories A–F) use $(0.30, 0.20, 0.20, 0.20, 0.10)$
- human-performance hypotheses (category G) use $(0.05, 0.10, 0.05, 0.65, 0.15)$
- inspection-gap hypotheses (category J) use $(0.05, 0.05, 0.05, 0.55, 0.30)$.

The documentary evidence component is updated for each retrieved record per:

$$E_{\text{doc}} = \text{clamp}\big(0.30\,E_{\text{prior}} + 0.55\,S_{\text{support}}\,w_{\text{auth}} + 0.15\,E_{\text{context}} - 0.45\,C_{\text{contradict}},\ [0,1]\big) \tag{2}$$

where w_{auth} is drawn from six source-tier values $\{1.00, 0.80, 0.70, 0.50, 0.40, 0.30\}$, corresponding to plant-instance records, plant procedures, plant FMEA records, plant-family data, INPO IRIS reports, and U.S. Nuclear Regulatory Commission ADAMS documents, respectively, and $C_{\text{contradict}}$ is the weighted contradiction score from opposing records. The negative contradiction term ensures that contradictory evidence attenuates the hypothesis score instead of being silently discarded. Hypotheses are ranked by S_h: the highest-scoring hypothesis is designated the primary finding only if no near-tie flag is active, and all alternatives are reported with their scores in the output regardless.

Coverage enforcement runs independently of the scoring process. Before the pipeline produces a final output, it verifies that each of the 12 DACKAR causal categories has been addressed. A category is considered addressed if it produced at least one scored hypothesis (regardless of whether that hypothesis scored high or low) or if the physical plausibility analysis explicitly excluded it as inapplicable to the current failure location and event type.

8. GOVERNANCE AND TRACEABILITY CONTROLS

Producing a ranked hypothesis list is necessary, but on its own this list is insufficient for a defensible RCA. Two additional controls are required: coverage enforcement, which ensures that no causal category is silently skipped, and traceability, which ensures that every conclusion is linked to a specific evidence source in a form verifiable by an auditor. This section describes how DACKAR implements both.

8.1. Twelve-Category Causal Taxonomy

The 12 causal categories in DACKAR are organized across three depth layers consistent with common RCA terminology: proximate causes (categories A through F), contributing causes (categories G through K), and root cause (category L). Table 1 presents the full taxonomy along with plant-specific examples.

The three-layer structure is aligned with the AP-913 equipment reliability framework. Proximate causes (A-F) describe what physically happened at the failed component. Contributing causes (G-K) describe

Table 1: DACKAR 12-category causal taxonomy for equipment RCA

Code	Category	Depth Layer	Example in Nuclear Plant Context
A	Equipment Degradation	Proximate	Bearing wear from lubricant breakdown; valve seat erosion from particulate flow; insulation resistance degradation from thermal cycling
B	Support System Loss	Proximate	Cooling water flow reduction to motor cooler; instrument air pressure drop affecting valve actuator; DC bus voltage sag affecting control logic
C	Upstream System Influence	Proximate	Cavitation-inducing conditions propagated from upstream pump; water hammer transient from upstream valve actuation; pressure fluctuation exceeding design band
D	Downstream System Influence	Proximate	Backpressure from downstream isolation reducing flow below minimum; thermal feedback from downstream load change affecting operating point
E	Operating Context	Proximate	Equipment operated outside design basis conditions during outage evolution; transient loads imposed during surveillance test cycling
F	External Hazard	Proximate	Vibration from adjacent construction activity; environmental temperature exceedance in equipment room; flooding or moisture intrusion
G	Human Execution Error	Contributing	Incorrect reassembly following maintenance; lubrication specification applied to wrong fitting; procedure step performed out of sequence
H	Design Deficiency	Contributing	Design margin inadequate for actual service conditions; material specification not appropriate for fluid chemistry; component rating insufficient for actual duty cycle
I	Configuration Baseline Deviation	Contributing	As-built configuration differs from design basis documentation; temporary modification not reflected in controlled drawings; software version did not match qualified configuration
J	Inspection Program Gap	Contributing	Surveillance test interval insufficient to detect degradation before functional failure; acceptance criteria not commensurate with known failure mode; inspection technique not qualified for defect type
K	Supply Chain Defect	Contributing	Non-conforming replacement part introduced during procurement; counterfeit or substandard material installed; certificate of conformance not verified before installation
L	Systemic Organizational / Programmatic Weakness	Root Cause	CAP threshold too high to capture precursor conditions; work management process does not enforce PM schedule compliance; engineering review process does not require FMEA update when PM interval is changed

Table 2: Attention flag types generated by DACKAR

Flag Type	Trigger Condition	Resolution Required
Coverage gap	A required causal category has neither a scored hypothesis nor a recorded exclusion basis	Analyst evaluates the category or documents a technical basis for exclusion; silence is not permitted
Data quality	A required input is absent or below the minimum quality threshold (historian gap, incomplete plant model, unindexed condition report corpus)	Analyst provides the missing data or acknowledges the limitation and describes how it was addressed in the investigation record
Contradictory evidence	Documentary or fleet experience records significantly conflict with the primary hypothesis	Analyst evaluates whether the conflict is resolvable or reflects genuine uncertainty to be reflected in the primary finding confidence rating
Near tie	Composite score gap between first- and second-ranked hypotheses is below the discrimination threshold (0.05)	Analyst conducts additional investigation to discriminate between competing hypotheses before a primary finding is recorded

why the enabling conditions were present. The root cause (L) describes the management system weakness whose correction prevents the same class of failure from recurring across the plant.

As a DACKAR process requirement, all 12 categories must be evaluated for any event that meets the significance threshold. A category may be explicitly excluded (an external hazard category can be ruled out if the event occurred in a controlled room with no adjacent activities and no environmental exceedances), but the exclusion must be stated and its basis recorded. The pipeline enforces this by tracking every category's disposition and generating an attention flag for any category lacking either a scored hypothesis or a recorded exclusion basis.

8.2. Attention Flags and the Analyst Interface

An attention flag is a structured output item that names a specific condition requiring analyst resolution before investigation closure. It remains open in the rca_card until a documented response is provided. Table 2 summarizes the four flag types.

8.3. Audit Trail Structure

Figure 2 illustrates the full hypothesis lifecycle and governance chain. For every score in the rca_card, the manifest records the source data item that produced it, the specific calculation applied, and the result. An auditor can trace any score back to the underlying evidence record without reconstructing the investigation or interviewing the analyst. This traceability is generated automatically as a byproduct of pipeline execution: no additional documentation effort is required beyond providing the input data. Analyst judgment, as applied in resolving attention flags and endorsing or modifying hypotheses, is recorded through the analyst's documented response to each flag.

9. EXAMPLE: SERVICE WATER CHECK VALVE LEAK

9.1. Event Description

The following event is a synthetic representative composite designed to exercise all five lines of investigation simultaneously. It was not derived from any specific plant record. All identifiers, records, intervals, retrieval results, and scores in this demonstration are synthetic.

Event EVT-U1B-2025-0312 was initiated when a Unit 1B service water surveillance identified a

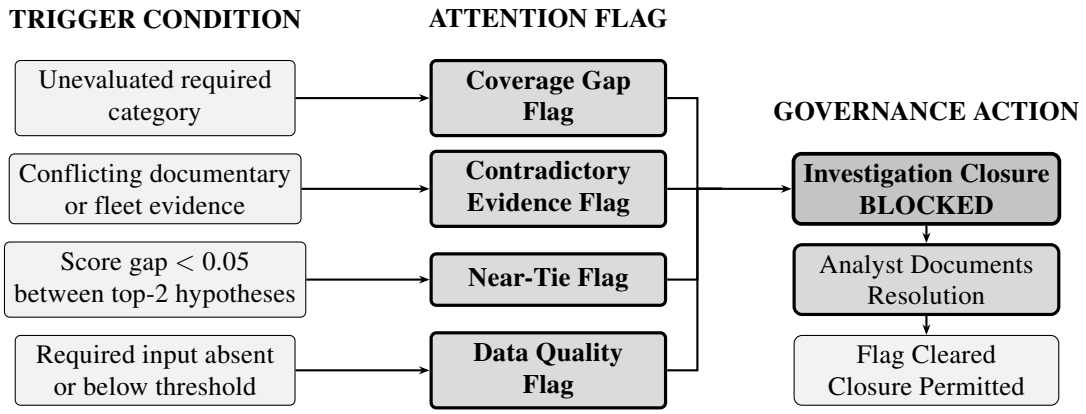

Figure 1: Attention flag governance loop. Each trigger condition generates a structured attention flag that blocks investigation closure. All four flag types converge at the same governance gate regardless of type. Closure is permitted only after the analyst has provided a documented resolution for every active flag; coverage gap flags cannot be satisfied by silence, near-tie flags require discriminating evidence, and data quality flags require either the missing data or a documented acknowledgment of the limitation.

leak on a check valve supplying safety-related heat exchangers at a rate exceeding the technical specification limit. The condition was entered into the CAP as a significant condition adverse to quality, initiating cause evaluation under the plant CAP. The valve had been in service for approximately 6 years; no automatic alarms actuated before discovery and no pre-failure anomalies were detected in the continuous monitoring signals (i.e., the condition was identified entirely through the scheduled surveillance test). Work order records indicated that the scheduled PM inspection for this valve was overdue by approximately 178 days at the time of the event.

9.2. RCA Assessment

DACKAR evaluated the event against six causal hypotheses that the physical plausibility analysis identified as being consistent with the plant model for the service water check valve and its connected systems:

- CHK-SEAT-EROSION (category A): progressive seat erosion from high-cycle wear, potentially accelerated by particulate matter in the service water system exceeding the design specification for the seat material.
- CHK-DISC-DAMAGE (category A): check valve disc damage from repeated high-frequency disc slam during service water flow cycling.
- PM-FREQ-NONCONF (category J): PM frequency nonconformance—the plant's 18-month inspection interval exceeded the 12-month vendor-specified maximum for valves in high-cycle service.
- PM-CFG-CTRL-GAP (category I): configuration control gap; a 2021 PM interval revision was implemented without an engineering evaluation of the deviation from the vendor-specified frequency.
- VND-BATCH-TRACE (category K): vendor supply chain; the installed valve falls within a production lot range identified in a fleet operating experience record as being associated with premature seat fatigue failure.
- OE-SCREEN-MISS (category L): systemic organizational weakness; the relevant fleet OE record was screened as non-applicable in 2023 and not incorporated into the plant PM program or

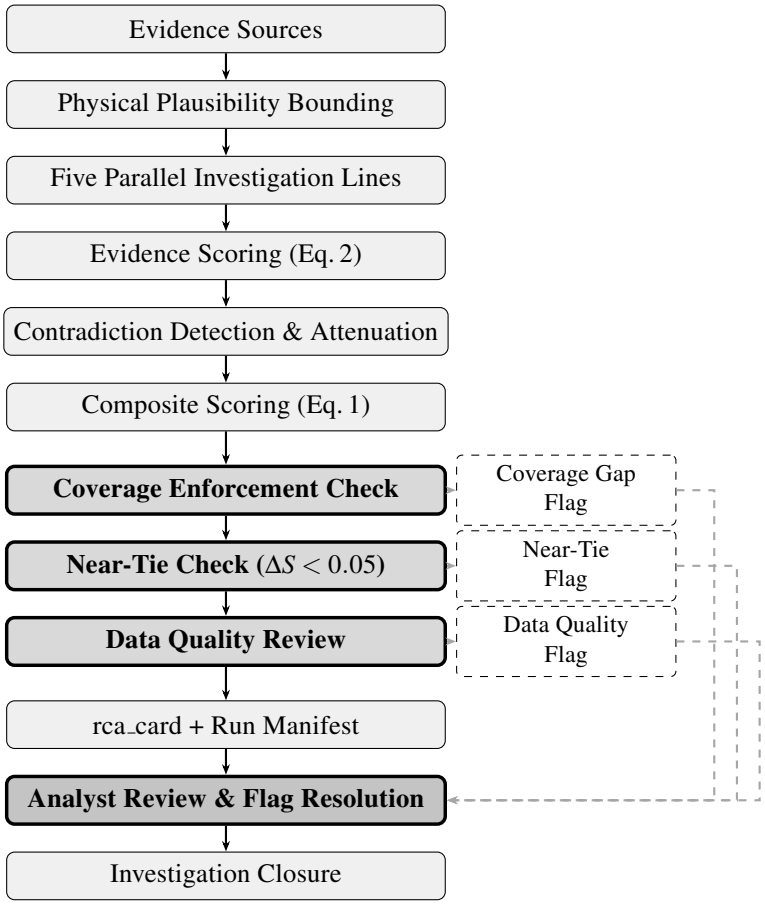

Figure 2: Hypothesis lifecycle through the DACKAR governance chain. The lightly shaded nodes are processing steps that transform evidence into scored hypotheses. The bold-border, more heavily shaded nodes are governance enforcement checks that generate structured attention flags whenever their conditions are not satisfied (dashed arrows). All flags converge at the bold-border heavily shaded analyst review gate (solid return arrows); investigation closure cannot proceed until every flag carries a documented analyst resolution.

engineering evaluations.

The timing analysis applied to the historian record found no pre-failure telemetry anomalies: both monitored signals (pump discharge pressure and heat exchanger flow) remained within normal operating bounds throughout the event window, and no operator alarms were generated prior to the surveillance test. The failure was discovered entirely through scheduled surveillance, not continuous monitoring. With no pre-failure telemetry signal available, the discriminating weight for this event falls entirely on the documentary evidence and governance streams.

The documentary evidence line retrieved seven evidence snippets from plant and fleet sources. Two plant condition reports from prior leak events on the same valve model supported CHK-SEAT-EROSION. An INPO IRIS fleet record (IRIS-OE-2023-SW-0047) documenting four failures of the same valve model at three separate pressurized-water reactor units, attributed to production lot range GWB-2020-L05 through GWB-2020-L09, supported VND-BATCH-TRACE. This record carried a source authority weight of 0.80 (fleet INPO tier). A 2021 plant work order documenting the PM interval revision and the original vendor specification both supported PM-FREQ-NONCONF by confirming that the plant's 18-month interval deviated from the vendor-specified 12-month maximum for high-cycle service; the same work order supported PM-CFG-CTRL-GAP by showing that the revision was implemented without a recorded engineering evaluation of the vendor specification deviation. A 2023 OE screening decision log, in which IRIS-OE-2023-SW-0047 had been classified as non-applicable for this valve class, provided documentary evidence for the OE-SCREEN-MISS root cause hypothesis.

The most analytically significant evidence interaction in this example involves VND-BATCH-TRACE. A 2025 teardown report retrieved by the documentary evidence line found the seat insert dimensions to be within the acceptance criteria, appearing to contradict the fleet OE attribution of failures to a manufacturing lot defect. The NER engine resolved this apparent contradiction by extracting the installed lot number from the teardown report text (GWB-2020-L07) and cross-referencing it against the affected range in IRIS-OE-2023-SW-0047 (GWB-2020-L05 through GWB-2020-L09): the installed valve was found to fall within that range. The IRIS-OE record further states that the dimensional defect manifests as fatigue failure after 12-24 months of cycling service, not at initial acceptance inspection (making the teardown finding consistent with, rather than contradictory to, the fleet OE attribution). The contradiction weight was modulated rather than zeroed, VND-BATCH-TRACE was retained as a live candidate, and an open item for vendor lot traceability was generated in the rca_card.

The PM compliance line confirmed that the most recent PM inspection was overdue by 178 days at the time of the event, directly supporting categorization under both PM-FREQ-NONCONF (category J) and PM-CFG-CTRL-GAP (category I). The Event Similarity engine identified this as the third leak on the same valve model within a span of 18 months; both prior CAP entries had been closed at proximate cause and without recurrence evaluation, despite an explicit supervisor observation of the pattern, a finding that fed directly into OE-SCREEN-MISS (category L) as evidence of a systemic failure to act on available fleet experience.

Table 3 presents the post-refinement composite scores. The pre-refinement leader was CHK-SEAT-EROSION (category A); after evidence retrieval, PM-FREQ-NONCONF overtook it at 0.456 vs. 0.454: a rank inversion driven by the `surveillance` weight profile placing $w_{\text{evid}} = 0.55$ on category J vs. $w_{\text{evid}} = 0.20$ on category A, so evidence improvements carry $2.75\times$ the composite impact for the PM hypothesis. CHK-DISC-DAMAGE was filtered following evidence retrieval. The rca_card returns `depth_complete: true`, spanning all three causal layers: proximate (category A), contributing (categories I, J, and K), and root cause (category L).

9.3. Attention Flag: Missing FMEA Failure-Mode Latency Data

The plant model included failure mode designations for the check valve but no failure mode latency estimates (the expected time from degradation onset to functional failure). Without latency data, the

Table 3: Post-refinement composite scores for event EVT-U1B-2025-0312 (synthetic representative event; see [4] for full component-level scores and run manifest). All scores are post-multiplier values (quality multiplier = 0.682; see Section 9.3). Contradiction penalties were absorbed into the evidence component per Eq. 2. The score gap between Rank 1 and Rank 2 is $0.002 < 0.05$ threshold, triggering the near-tie flag.

Hypothesis	Cat.	Composite	Notes
PM-FREQ-NONCONF	J	**0.456**	Rank inversion from pre-refinement
CHK-SEAT-EROSION	A	**0.454**	Fleet OE contradiction modulated
PM-CFG-CTRL-GAP	I	0.391	
VND-BATCH-TRACE	K	0.360	Lot cross-ref. contradiction resolved
OE-SCREEN-MISS	L	0.347	Root cause; analyst determination required
CHK-DISC-DAMAGE	A	—	Filtered at evidence threshold

timing analysis could not evaluate whether the absence of pre-failure signals was consistent with the expected degradation rate for this failure mode. A data quality flag was generated and a quality multiplier of 0.682 was applied uniformly to all hypothesis scores to reflect the incomplete model, with the analyst being required to either supply FMEA latency data or document the limitation in the investigation record. Scores reported in Table 3 are post-multiplier values.

9.4. Near-Tie Result and the Safety Argument

The scoring process produced the following results for the two highest-ranked hypotheses:

- Rank 1: PM-FREQ-NONCONF (PM frequency nonconformance, category J): composite score 0.4560
- Rank 2: CHK-SEAT-EROSION (check valve seat erosion, category A): composite score 0.4540

.

The score gap of 0.002 is below the pipeline's near-tie discrimination threshold of 0.05. The pipeline therefore did not designate PM-FREQ-NONCONF as the primary finding. Instead, it generated a near-tie attention flag and issued a mandatory tie-break requirement: the analyst must conduct additional investigation to discriminate between these two hypotheses before the investigation can be closed with a single primary finding.

This behavior (i.e., refusing to manufacture a definitive conclusion when the evidence does not support one) is a deliberate design feature rather than a limitation. The two hypotheses are not equivalent from a corrective action perspective: seat erosion implies a component-level corrective action (replacement, seat material upgrade, water chemistry monitoring), while PM frequency nonconformance implies a program-level corrective action (restoring the deferred task, reviewing PM intervals for similar valves in the same system). Forcing one conclusion when the evidence equally supports both would produce a corrective action that addresses only part of the actual problem.

A structured assessment tool that explicitly surfaces the limits of its own conclusions is more compatible with a questioning attitude and the conservative decision-making required by that safety culture than one that silently selects a winner. Resolving the J/A near-tie requires additional discriminating investigation: a quantitative seat condition assessment based on the valve inspection data, along with a review of the PM extension authorization in order to confirm whether adequate technical justification was documented. Separately, two additional open items were generated for candidates that scored above the minimum threshold but below the pipeline's write-back confidence floor: vendor lot traceability confirmation for VND-BATCH-TRACE (category K), and re-evaluation of the 2023 OE screening decision that classified IRIS-OE-2023-SW-0047 as being non-applicable for OE-SCREEN-MISS

(category L). All four items were identified systematically by the pipeline rather than left to the analyst's recollection.

9.5. Determinism and Repeatability

Here, "determinism" refers to the reproducibility of the evidence aggregation process: given identical inputs, the pipeline produces identical scored outputs. In this respect, multiple independent runs of the DACKAR pipeline on the EVT-U1B-2025-0312 event produced identical composite scores and identical hypothesis rankings on every run. The only variation between runs was in the natural-language narrative text generated to explain the findings; the explanatory prose varied in phrasing while the underlying assessment was identical. Recall that DACKAR is in fact not a probabilistic model asserting a posterior probability on a root cause, but rather a reproducible evidence-weighting procedure whose output uncertainty is expressed through scored alternatives and attention flags. This deterministic behavior is operationally significant for two reasons: first, the plant organization can rely on the assessment as a stable record, and second, differences between two rca_cards for different events reflect differences in the evidence basis, not random variation in the pipeline (an analyst comparing findings across events can attribute score differences to the underlying evidence rather than to assessment-process noise).

10. DATA PREREQUISITES

Deploying DACKAR for a specific equipment class requires three data prerequisites: 1) a plant model covering the equipment type, with at least one failure mode entry per component type; 2) a condition report and work order corpus indexed by equipment identifier such that the documentary evidence line can retrieve records by tag or functional location; and 3) historian data at a resolution sufficient to establish signal precedence. A minimum viable configuration with these three inputs produces partial but useful output, with data quality flags communicating which lines of investigation are degraded and to what degree. Fleet OE integration via INPO IRIS improves evidence line quality but is not required for minimum viable operation; plants without IRIS access operate in an internal-records-only configuration, with this lack of access being noted in the rca_card.

11. CONCLUSIONS

DACKAR enforces multi-level causal accountability aligned with the AP-913 equipment reliability framework and produces a traceable evidence record consistent with the documentation requirements of 10 CFR 50, Appendix B, Criterion XVI, applied systematically to each investigation processed by DACKAR. What the pipeline enforces is completeness: all 12 DACKAR causal categories are evaluated, every conclusion is linked to a specific evidence source, and coverage gaps surface as structured attention flags rather than as silent omissions. Every rca_card reflects the analyst's documented review and endorsement of the structured assessment, preserving investigative authority while eliminating the structural gaps that manual practice leaves open.

The service water check valve example shows the safety argument for structured coverage enforcement. The pipeline identified a near-tie between two competing hypotheses, with a score gap of 0.002, and declined to force a primary finding. This result is operationally significant: the two hypotheses imply different corrective actions, and selecting one arbitrarily would produce a corrective action that addresses only part of the actual problem. The pipeline's explicit identification of the near-tie and its mandatory tie-break requirement reflect the same conservative decision-making standard that nuclear safety culture requires of human investigators. A tool that surfaces ambiguity rather than manufacturing false confidence under time pressure is more compatible with a questioning attitude than is a tool that always returns a definitive answer.

The deterministic behavior of the developed pipeline (i.e., five runs producing identical scores and rankings) supports the use of rca_card outputs as stable investigation records. The assessment does not

change between reviews, and differences between rca_cards for different events reflect differences in the evidence basis rather than pipeline variation.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy Office of Nuclear Energy, under the Light Water Reactor Sustainability (LWRS) program. Artificial intelligence tools assisted with the development of the presented framework and the revision of this paper. All technical analyzes, designs, results, and conclusions were developed and verified by the authors.

REFERENCES

[1] U.S. Nuclear Regulatory Commission, "Title 10, code of federal regulations, part 50, appendix b: Quality assurance criteria for nuclear power plants and fuel reprocessing plants," tech. rep., U.S. Nuclear Regulatory Commission, Washington, DC.

[2] International Atomic Energy Agency, "Root cause analysis following an event at a nuclear installation: Reference manual," Tech. Rep. IAEA-TECDOC-1756, International Atomic Energy Agency, Vienna, Austria, 2014.

[3] S. Sklet, "Comparison of some selected methods for accident investigation," *Journal of Hazardous Materials*, vol. 111, no. 1-3, pp. 29–37, 2004.

[4] C. Wang, D. Mandelli, C. M. Godbole, V. Agarwal, M. Movassat, D. Liang, and E. Nur, "Open-source release of digital analytics, causal knowledge acquisition and reasoning (dackar)," Tech. Rep. INL/RPT-25-85560, Idaho National Laboratory, Idaho Falls, ID, 2025.

[5] J. F. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, 1983.