# Five Whys as an Epistemic-Honesty Scaffold for Multi-Agent LLM Analysis of Industrial Time Series

Yoshihiro Ochi, Yasunobu Uchiyama
Graduate School of Artificial Intelligence and Science, Rikkyo University, Tokyo, Japan

***Abstract*—Multi-agent large language model (LLM) pipelines increasingly analyse industrial time series. Where the true causes lie outside the recorded signal, they tend to report confident, plausible-sounding mechanisms not grounded in the data — a *false root-cause* handed to an engineer. We port Toyota's Five Whys onto such a multi-agent system (MAS) not as a decomposition tool but as a *reviewer frame*, and find it functions as an *epistemic-honesty scaffold*. On a non-public power-consumption series from a coating application process, whose true drivers are exogenous to the data, a single Five Whys chain drives the MAS to surface three grounded-sounding mechanisms and, by repeatedly asking *where in the data is this grounded?*, *retract each in turn*, converging on the honest verdict: the cause is exogenous to, and unverifiable from, this signal. A matched baseline, run to the same depth without the reviewer, commits to one such mechanism as a confident but unverified cause. In the three-seed mean — a directional, mean-level effect with large per-seed variance and occasional sign reversals — the scaffold raises contestation on the hard series, whereas on a clean benchmark with an in-data root it instead reconciles competing narratives onto that root, adapting to whether a groundable root exists rather than adding noise. The honest verdict is the right target: the strongest model states it unaided, so the scaffold's value concentrates where the model cannot self-regulate. We frame it as a discipline for human–AI coexistence: it lets a practitioner correctly receive what the MAS does and does not know.**

***Keywords*—*Five Whys; multi-agent systems; large language models; root-cause analysis; epistemic honesty; causal abstention; industrial time series; argumentation framework; human–AI collaboration***

## I. Introduction

Multi-agent large language model (LLM) systems (MAS) are increasingly applied to industrial time-series analysis: a chain of LLM agents generates, tests, curates, and synthesises candidate findings, and a human practitioner receives a report to act on. In this setting the report's consumer is not a benchmark scorer but an engineer who must decide whether a stated cause is real. That places a specific demand on the system that aggregate quality rarely measures: *when the data cannot actually support a causal conclusion, the system should say so*, rather than supply a confident one.

Contemporary LLMs make this demand hard to meet, for an understandable reason. They are trained to be capable, fluent, and helpful — to behave, in effect, like an able student who always produces a well-formed answer. On data analysis that disposition has a side effect: asked why a signal behaves as it does, an MAS will readily produce a plausible, grounded-sounding mechanism even when the true cause is not present in the data at all. The result is a *confident false root-cause*: an answer that reads as authoritative and is, in fact, ungrounded. We do not regard this as a defect to be condemned; it is a predictable consequence of how these models are built. But on a factory floor, a confident false root-cause handed to an engineer is exactly the failure that matters, because it can be acted on. The question we take up is not how to make the model less capable or less confident, but how to *design around* the trait so that the practitioner correctly receives what the analysis does and does not know. This is the spirit of *jidoka* — automation with a human touch, in which the system surfaces its own abnormalities rather than running on silently [1] — carried from the shop floor to an analysis MAS.

Our design lever is domain method rather than model change. We take a method the same manufacturing tradition already uses for getting to true causes — Toyota's Five Whys [1] — and port it not as a decomposition tool but as a *reviewer frame* on the MAS: an auxiliary facilitator that repeatedly asks, of each causal claim, *where in the data is this grounded?* The facilitator sees no data and supplies no answers; it only forces the MAS to ground its own claims. We find that, used this way, the Five Whys functions less as a root-cause decomposition than as an *epistemic-honesty scaffold*.

The central evidence is a single Five Whys chain on a non-public industrial power-consumption series whose true drivers are exogenous to the recorded data. Under the scaffold, the chain drives the MAS to surface three grounded-sounding mechanisms — an ambient thermal driver, a dated structural break, an operating-mode cause — and to run new tests and *retract each in turn*, converging on the honest verdict: the cause is exogenous to, and unverifiable from, this signal. A matched baseline — the same MAS, run to the same depth and seed without the reviewer — instead commits its final report to one such mechanism as a confident but unverified cause. In the three-seed mean the scaffold raises contestation on this hard series — it leaves competing mechanisms standing as unresolved rather than collapsing them — whereas on a clean benchmark series with an in-data root the same scaffold instead *reconciles* the competing narratives onto that root. It adapts to whether a groundable root exists rather than merely adding noise. The honest verdict is moreover the *right* target, not a quirk of a weak model: the strongest model reaches "exogenous and unprovable" unaided, so the scaffold's value concentrates in the capability band where a model can act on the probing but cannot impose the discipline on itself.

### A. Contributions

*a) C1: The Five Whys as an epistemic-honesty scaffold:* We reframe Toyota's Five Whys, ported as a data-blind reviewer frame, as a mechanism that elicits honest causal abstention in a data-analysis MAS rather than as a

decomposition tool. The claim metrics and the operator are each grounded in established frameworks — quantitative bipolar argumentation [2], the Toulmin model [3], Why-Because analysis and counterfactual causation [4], [5], and grounded-theory saturation [6] (Section III).

*b) C2: A demonstration that the scaffold drives retraction of a confident false root-cause.:* On a real industrial signal we show a single chain driving the MAS to self-refute three grounded-sounding mechanisms and attribute the cause honestly to outside the data, against a matched baseline that commits to one of them as a confident but unverified cause — *without the scaffold a confident, unverified root-cause is handed to the engineer; with it, the report abstains honestly* (Section V-A).

*c) C3: The scaffold is adaptive and its value is bounded.:* In the three-seed mean it surfaces unresolved mechanisms on the exogenous-root series and reconciles them on the in-data-root series; and its effect concentrates in a capability band, since the strongest model reaches the honest verdict unaided (Sections V-B to V-D). This contestation effect is *directional and mean-level* — per-seed variance is large and individual seeds can reverse — so the adaptive reading rests on the qualitative chain (Section V-A) and the regime contrast (Section V-D), with the three-seed-mean rise as corroboration rather than as a per-seed-stable claim.

Underlying all three is a single stance toward human–AI coexistence: rather than making the model less confident, we let a domain discipline surface what it cannot ground, so that a practitioner and an analysis MAS can each contribute what they are good at. Section II situates the work; Section III describes the MAS, metrics, and operator; Section IV the setup; Section V the results; Section VI the coexistence stance and limitations; Section VII the follow-on questions; and Section VIII concludes.

## II. Related Work

We situate this work against five lines: self-evaluation bias, self-improvement, multi-agent debate, LLM-based root-cause analysis, and abstention. A common thread distinguishes our setting from all five: we target *ground-truth-free* industrial analysis, where the desired behaviour is not a better answer but an honest acknowledgement that the data cannot support one. We then note the established frameworks our method builds on.

*1) Self-evaluation bias and sycophancy:* A body of work documents that LLMs tend to favour a user's stated framing — sycophancy [7] — and that such tendencies are hard to remove by prompting alone. This literature *identifies* the disposition that makes an analysis MAS report a confident mechanism it has not grounded. We take that disposition as given and ask a complementary question: not how to detect the bias, but how to *operationally elicit* honest grounding despite it, through a reviewer loop that drives the system to retract what it cannot support.

*2) Self-improvement and self-verification:* Self-Refine [8], Reflexion [9], and Chain-of-Verification [10] have a model critique and improve its own output, and self-consistency aggregates multiple samples [11]. These methods are most effective on tasks with a checkable target, and Huang *et al.* [12] report that, without an external signal, self-correction of reasoning is limited. Our reviewer is deliberately not a self-critic that converges toward a correct answer: it supplies no answer, and the target is not a correct cause but an honest *abstention* when no cause is groundable. The turf differs — self-improvement pushes toward a known-good answer; we ask what to do when there is no answer to push toward.

*3) Multi-agent debate:* Debate among LLM agents can improve factuality and reasoning by having agents argue toward consensus [13]. Our configuration is *asymmetric* rather than peer debate: a data-blind facilitator interrogates an analyst, and the goal is not consensus but, where warranted, an honest non-answer. Convergence is the right outcome only when a groundable root exists (our benchmark regime, Section V-D); when it does not, the appropriate outcome is unresolved contestation, which a consensus objective would tend to paper over. Recent work in this venue extends debate with *interruptibility* — sentence-level disclosure and urgency-based turn-taking for early error correction among agents on benchmark QA [14]. Our setting differs on three axes: 1) the intervention is a *data-blind reviewer frame*, not agent–agent interruption; 2) the target is *honest causal abstention on real, ground-truth-free industrial time series*, not answer correctness on a benchmark; and 3) when no groundable root exists the desired outcome is *unresolved contestation*, not consensus. Tiered human-oversight frameworks for agentic systems [15] motivate *why* such honesty matters; we contribute an *implemented, empirically tested operator* rather than a governance taxonomy.

*4) LLM-based root-cause analysis:* A line of systems applies LLMs to root-cause analysis, typically to *generate* a most-likely root cause for an incident — RCACopilot reports high rates of plausible root causes [16] — while a recent benchmark finds that even strong models struggle to locate true root causes [17]. Our contribution runs in the opposite direction. We use the Five Whys not to *produce* a root cause but as a grounding interrogation, and the value we report is in *not* manufacturing one: when the cause is exogenous, the system is driven to abstain rather than to assert a clean root. Where this literature optimises the quality of a generated root, we characterise honest refusal to generate one as the property that matters for industrial deployment.

*5) Abstention, calibration, and uncertainty:* Abstention — teaching or prompting a model to say "I don't know" — is an effective hallucination mitigation, but it is hard in open-ended settings [18]. A related line estimates a model's confidence directly, through self-evaluation [19] or semantic uncertainty over sampled generations [20]. These calibrate *whether to answer* at the granularity of a whole output; our target is finer and structural — which *individual causal claim* in an analysis is groundable — and the abstention we seek is not a refusal to answer but an attribution of the cause to outside the data. Rather than train or prompt for abstention directly, we therefore let *causal* abstention emerge structurally: the claim graph plus the Five Whys chain drive the system to attribute a cause to outside the data, with the grounding made explicit, rather than to emit an unsupported one. The abstention is thus grounded ("exogenous and unverifiable from this signal") rather than a bare refusal.

*6) Foundations the method builds on:* The metrics and operator are not novel inventions but adaptations of established frameworks, which both grounds the design and keeps its novelty located in the conjunction rather than in any single part. Claim quantification uses quantitative bipolar argumentation frameworks (QBAFs) [2], [21], [22] with the Toulmin model [3] for base weight; causal structure and the test for a genuine causal link use Why-Because analysis [4] and counterfactual causation [5]; honest stopping uses theoretical saturation from grounded theory [6]; and the essence of the Five Whys as a single-path method follows the root-cause literature [23], [1]. The data-blind reviewer's repeated demand — *where in the data is this grounded?* — resembles Socratic questioning, but the operator is more constrained than open-ended elicitation: it follows a single causal chain toward a root and halts at an exogenous terminus (Section III-C). Separately, work that applies LLMs directly to time-series forecasting [24] is orthogonal to ours: we use the LLM not as a forecaster but as an analyst node whose causal claims must be grounded or honestly withdrawn.

*7) The gap we fill:* The novelty is the conjunction (Sections V to VI): A data-blind Five Whys reviewer that 1) elicits honest causal abstention rather than a generated root, 2) *adapts* to the data — reconciling when a root is groundable, surfacing contestation when it is not — and 3) concentrates its value in the capability band where a model cannot self-regulate, with the honest verdict shown to be the right target by the strongest model reaching it unaided. None of the five lines above, individually, occupies this position.

## III. Method

We study a data-analysis multi-agent system (MAS) that iteratively analyses a single time series and accumulates its findings into a causal hypothesis graph (Section III-A). Onto this MAS we port Toyota's Five Whys not as a decomposition tool but as a *reviewer frame*: an auxiliary facilitator agent that interrogates the MAS's causal claims (Section III-C). We quantify each claim with three metrics drawn from established frameworks (Section III-B), compare the MAS with the reviewer on (*Five Whys*) and off (*baseline*) under a matched protocol (Section III-D), and stabilise the metrics with two engineering devices (Section III-E). Fig. 1 shows the overall flow.

### A. The Data-Analysis MAS

The MAS analyses the series over a sequence of rounds. In each round a Strategy agent proposes analysis directions, one or more Analysis agents execute code and read plots to extract candidate insights (numeric and visual), and a Claim-Graph agent consolidates the round's insights into the running graph and induces edges between claims. A Domain agent contributes prior process knowledge as a prefix before any analysis insight exists. The unit of accumulated knowledge is the *claim graph*: its nodes are causal claims and its edges are typed relations — *explains* and *refines* (support) and *contradicts* and *supersedes* (attack). The graph is the object the reviewer interrogates and the object on which the three metrics are computed. After the final round, a Synthesis agent composes the arm's final report from this graph. We treat the MAS architecture as fixed across all conditions; the only manipulated factor is the presence or absence of the reviewer.

### B. Three Claim Metrics

Each claim carries three metrics, deliberately chosen so that each maps onto an established, axiomatised framework rather than an ad hoc formula. The mapping treats the claim graph as a *quantitative bipolar argumentation framework* [2], [21]: claims are arguments, *explains*/*refines* are support, and *contradicts*/*supersedes* are attack (Fig. 2).

*1) Grounding (base weight, Toulmin): Grounding* is the claim's standalone evidential strength $g \in [0, 1]$, the degree to which it is backed by a specific, computed feature of the data rather than asserted. This is the base weight in the argumentation framework, and corresponds to the *grounds* and *qualifier* of the Toulmin argument model [3]: a claim that names an in-data feature and a test is well grounded; a claim restated from domain knowledge without a data test is not.

*2) Validity (acceptability degree, QBAF): Validity* is the claim's acceptability after support and attack have propagated through the graph, computed by an h-categorizer gradual semantics [2], [22]:

$$v(a) \;=\; \frac{g(a)}{1 + \sum_{b \in \mathrm{att}(a)} v(b)}, \tag{1}$$

where, $v(a)$ is this h-categorizer value, $g(a)$ is the grounding of claim $a$, and $\mathrm{att}(a)$ its attackers. A well-grounded claim with no surviving attacker keeps $v \approx g$; a claim that is contradicted by other surviving claims is driven down. We call a claim *contested* when attack has pushed its validity below its base weight ($v(a) < g(a)$). Crucially, grounding (the input) and validity (the propagated output) are kept distinct, so that "explains a lot but does not stand up" (high reach, low validity) and "stands up but explains little" are represented separately.

*3) Importance (causal reach, Why-Because Graph): Importance* is a claim's structural centrality in the causal graph, the reach of its *explains*/*refines* descendants, following the Why-Because Graph notion of a causal DAG and distance to a root [4]. Importance is orthogonal to validity: a central hub may be contested, and a peripheral claim may be valid. We use importance only to *order* which claim a Five Whys chain should interrogate next; we do not use it as a quality score, for reasons given in Section VI-B.

### C. The Five Whys Operator

The reviewer implements the Five Whys as a facilitator that asks, of a target causal claim, *by what mechanism — grounded in a specific in-data feature — does this hold?*, and repeats the question against each successive answer. Three properties define the operator, and each rests on an established account rather than on our own stipulation:

*a) A single causal chain:* The reviewer drills one chain at a time from a target claim toward its root, matching the documented essence of Five Whys as a single-path, single-root method [23] and the chain structure of counterfactual causation [5].

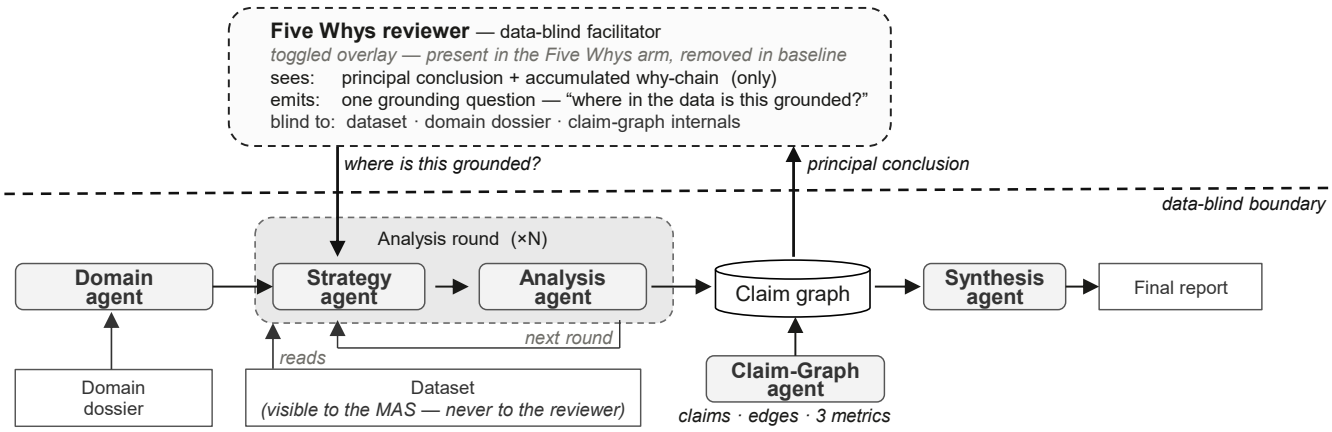


Fig. 1. System architecture, drawn to foreground the reviewer's data-blindness. The MAS analysis loop (Strategy → Analysis → Claim-Graph, $\times N$ rounds), fed by the dataset and domain dossier, is fixed across conditions. The Five Whys reviewer is a toggled overlay (present in the *Five Whys* arm, removed in *baseline*) that sits above a *data-blind boundary*: only the principal conclusion crosses up to it and only a single grounding question crosses back down into the next round — no data, dossier, or claim-graph internals are visible to the reviewer, so every retraction is the MAS refuting itself.

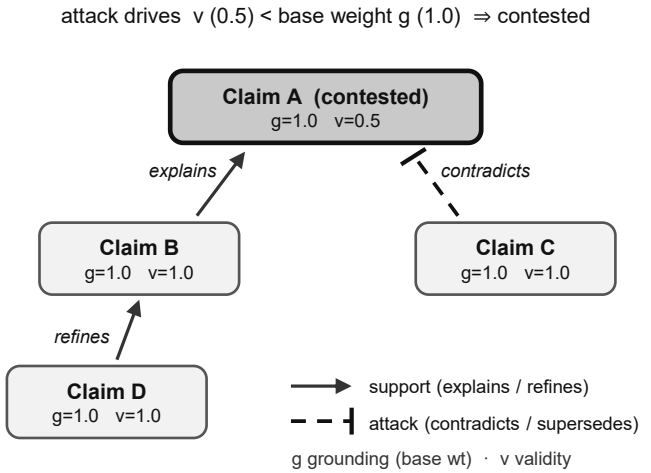


Fig. 2. Claim-graph anatomy and the *contested* condition. The figure isolates the two metrics that define contestation: each claim carries a grounding $g$ (Toulmin base weight) and a validity $v$ (the QBAF acceptability after support and attack propagate), and a claim is *contested* when an attack drives $v$ below its base weight $g$. Here claim C's attack lowers claim A's validity to $0.5 < 1.0$, while B explains and D refines. (Importance, the third metric, orders interrogation only and does not enter this condition; see Section III-B.)

*b) Each link a verified causal factor:* A link survives only if the MAS can ground it in a data feature that passes a test; this is the Why-Because counterfactual test for whether a factor is causal [4], [5], not a looser notion of relevance.

*c) Honest stopping at the true (possibly exogenous) root:* The chain terminates when no new groundable link can be added — either a data-grounded root is reached, or the cause is found to lie outside the recorded signal and the reviewer declares it exogenous and unidentifiable. This is theoretical saturation in the grounded-theory sense [6]: stop when novelty is exhausted, including the honest case where the answer is "not determinable from this data."

A design property of the facilitator is central to the interpretation in Section V: *the reviewer never sees the data*. It supplies no analysis and no answers; it only demands that the MAS ground its own claims. The reviewer is thus a Socratic self-grounding device, and any retraction it elicits is the MAS refuting itself on evidence it computes, not the reviewer overriding it. The full operator prompt is given in Appendix A.

### D. Matched On/Off Comparison

For each configuration we run the identical MAS twice: once with the reviewer (*Five Whys*) and once without (*baseline*). The two arms share the same Domain prefix and the same random seed, so they begin from the same state and diverge only through the reviewer's steering. Because that reviewer is data-blind and supplies no data, answers, or candidate mechanisms (Section III-C), the only factor the Five Whys arm adds is the demand to ground claims — not new information, data, or candidate answers — so a difference between the arms is the MAS responding to that demand. To remove depth as a confound, the baseline arm is run for as many analysis steps as the Five Whys arm took to reach its terminus (*baseline-match-Five Whys*): the comparison is between an MAS that analyses to the same depth with and without the Five Whys discipline, not between a longer and a shorter run. The final report of each arm is synthesised from its final claim graph under one shared, validity-aware procedure, so that any difference in the report reflects the graph state rather than two different synthesis methods. The synthesis is constrained to the epistemic standing of the goal-central claims: when the mechanism that answers the goal is contested or exogenous, the report's headline abstains rather than substituting a peripheral established claim. The reviewer does redirect which analyses the MAS performs — its grounding demand prompts confirmatory tests the baseline never runs — but that redirection is the scaffold's mechanism of action rather than a confound to remove: matched depth equalises the number of analysis steps, so the arms differ in what the MAS is driven to test, not in how much analysis it is afforded.

### E. Foundation: Stable, Discriminative Metrics

Two engineering devices make the three metrics stable enough to compare across arms; we state them as robustness scaffolding, not as contributions. First, *edge accumulation with support*: edges are accumulated across rounds and only an edge re-induced in at least $K$ rounds (support $\geq K$, here $K = 2$) is admitted to the metric computation. This is a vote-based confirmation across rounds, in the spirit of self-consistency [11], and it removes the metric churn that arises when a single round's stochastic edge extraction is taken at face value. Second, a *strategy gate* limits how many new directions are opened per round and requires each to name a new viewpoint or a verification need; this is theoretical sampling [6] applied to generation, and it keeps the graph from inflating with re-statements. With both devices in place the metric rankings are stable across rounds (no spurious reordering) while remaining discriminative, which is the precondition for reading on/off differences as signal.

## IV. Experimental Setup

### A. Datasets: Two Causal Regimes

We use two time series chosen to differ in one decisive respect: whether the true cause of the series' behaviour is approximable within the recorded signal or lies outside it.

*1) Benchmark series (in-data root approximable):* The public AirPassengers monthly series is a textbook case whose salient drivers — a growth trend and a seasonal cycle — are present in, or closely approximable from, the signal itself. It serves as the regime in which a groundable root exists, and as a contrast to the industrial case. We use "in-data root" in this operational sense — a terminus the analysis can ground in the signal, here trend and seasonality — not in the sense that an independently established causal mechanism is available.

*2) Industrial series (root exogenous):* The hard case is a non-public power-consumption time series from a coating application process, provided through the manufacturing partnership and abstracted here per the confidentiality conventions of the collaboration: we report no process names, line identifiers, or raw values, though we note its structural shape — a power-consumption signal with seasonal variation spanning multiple years. Its decisive property for this study is that the true drivers of the series — production scheduling, ambient conditions, and operator activity — are *not* recorded in the dataset. The signal alone is genuinely causally ambiguous: a competent analysis can ground *that* something changed but cannot, from this signal, prove *why*. This is the realistic industrial condition under which a confident false root-cause is most damaging, and it is where the scaffold's behaviour is most informative.

### B. Models

The MAS and the reviewer are each instantiated with a large language model, and we vary the capability tier so as to locate where the scaffold's effect lives. We report results for two commercial models, `claude-haiku-4-5` (a mid-tier model) and `claude-opus-4-8` (a top-tier model), in the MAS and reviewer roles. We also ran a smaller open-weight model (`gemma`); on the hard series it neither sustains contestation nor self-regulates, and we report it only as a lower floor where the scaffold confers little benefit, without building on it.

### C. Configurations, Seeds, and Protocol

The full experimental matrix crosses MAS model, reviewer model, and dataset; the analysis in this paper draws on four configuration groups (five run cells; E2-aux spans two pairings) that carry the reported findings (the remaining cells, including the open-weight floor and a set of facilitator-capability variants, are summarised but not featured; see Section VI-B). Their model pairings, datasets, and per-seed results are given in Table I:

- E1 / E2 (hero and primary contestation) — the primary industrial pairing; E1 is the single-chain dialogue of Section V-A and E2 its three-seed contestation count.
- E2-aux (auxiliary contestation) — two further capable-MAS pairings that corroborate the *direction* of the contestation effect across capability tiers.
- E3 (ceiling) — the top-tier peer pairing, both roles at the strongest model.
- E4 (contrast) — the benchmark-series rerun of the primary pairing (in-data root).

Every configuration is run for three random seeds. Unless noted, runs use a single Five Whys chain and a fixed round budget (max-rounds $= 8$); the on/off comparison uses the matched protocol of Section III-D. Aggregate quantities over seeds are reported as means; because the contestation count carries large seed-to-seed variance on the hard series (the three-seed mean rises for all capable-MAS configurations, but individual seeds vary in magnitude and sometimes in sign), we report the three-seed mean alongside the per-seed values (Section V-B) rather than implying tight intervals. The metric-stabilising foundation (Section III-E) is identical across all featured runs, and a post-experiment code audit confirmed that the design invariants (matched depth, edge-support accumulation, single-chain structure, and reviewer fidelity — no leading the witness, see Appendix A) held across all cells (Section VI-B).

## V. Results

We first present the qualitative core of the paper: a single Five Whys chain driving an MAS to retract a sequence of confident but ungrounded causal claims on the industrial series, and the matched baseline that, without the reviewer, commits its final report to one such claim as a confident but unverified cause (Section V-A). We then show that this behaviour is reproducible as a mean-level rise in contestation across capability pairings on the hard series (Section V-B), that the honest verdict is the right target and the scaffold's value concentrates where the model cannot reach it unaided (Section V-C), and that on a series with an in-data root the same scaffold instead reconciles rather than contests (Section V-D).

### A. Single Chain Retracts Three Ungrounded Mechanisms

On the industrial series, the unscaffolded MAS (baseline) produces a fluent, confident report that commits to a primary root-cause without a confirming test. With the reviewer present (Five Whys), starting from the same state and the same seed, a

**Reviewer↔MAS dialogue on the industrial series (sanitized; one Five Whys chain, three turning points)**

**Turning point 1 — ambient/seasonal mechanism.**

**MAS (Five Whys off)**: holds, grounded and unchallenged, that the decline is "ambient-driven thermal load variation rather than structural control changes."

**Reviewer (Five Whys)**: "Where *in the data* is that mechanism grounded — rather than a bare seasonal correlation you have yourself flagged as unverifiable without ambient data?"

**MAS (self-revision)**: "Seasonal ambient-responsive thermal control is *not* the primary driver …the seasonal mean pattern contradicts the signature expected from ambient compensation …ambient compensation is absent or negligible."

**Turning point 2 — a dated structural break.**

**MAS (Five Whys off)**: asserts, without change-point detection, a structural break on a specific date with a substantial median drop.

**Reviewer (Five Whys)**: "Is there *any* observable scheduling feature whose month-by-month profile matches the decline — or must you concede that no on-data scheduling feature exists?"

**MAS (self-revision)**: re-runs change-point detection and self-corrects: the break is at *a different, earlier date than previously hypothesised*, and is "a *distributional* change" — a severalfold variance increase at near-flat mean power — "rather than a sustained mean power reduction."

**Turning point 3 — an operating-mode cause.**

**MAS (Five Whys off)**: attributes the decline to a switch between operating modes.

**Reviewer (Five Whys)**: "By what mechanism, grounded in a specific in-data feature, does operating mode drive the monotonic power fall?"

**MAS (self-revision)**: tests it and rejects it: the relevant activity count is "essentially flat across months" (a near-zero, non-significant correlation), so reduced operational intensity "does *not* explain the observed power variation."

**Terminus.** The Five Whys *final report* itself declines to prescribe a cause: its headline states that the analysis "cannot identify a single, data-grounded" improvement and that the drivers are *exogenous* to the power signal and *not verifiable from the available data alone*. The matched baseline's final report instead commits to a confident primary recommendation — an operating-mode scheduling change it credits with the observed decline — while conceding it cannot be verified from the data and "requires correlated … logs to validate."

Fig. 3. One Five Whys chain on the industrial series: the MAS refutes three successive grounded-sounding mechanisms on evidence it computes itself and attributes the cause honestly to outside the data, so that its *final report* declines to prescribe one. The matched baseline (same seed, same depth) instead commits to a confident but unverified root-cause. Dates and raw values are abstracted for confidentiality; the mechanism types and the self-refutation structure are unchanged.

single chain drives the same MAS to propose three grounded-sounding mechanisms in turn — an ambient-temperature-driven thermal load, a dated structural break, and a switch of operating mode — to *test and retract each*, and to land on the verdict that the cause is exogenous to, and unverifiable from, the power signal alone. Fig. 3 reproduces the three turning points of the chain (run E1, seed 1).

Two properties of this exchange matter for the interpretation. First, every retraction is the MAS refuting *itself*: the reviewer supplies no data and no answer and names no candidate mechanism, only the demand to ground the claim (Section III-C), so the new change-point test, the operating-mode regression, and the rejection of each are the MAS's own work on evidence it computes. The reviewer's questions are pointed, but they add no information beyond *where in the data is this grounded?*; what changes the verdict is the test the MAS then runs, not a fact the reviewer supplies. Second, the baseline is not a weaker analyst that simply missed these tests; it is the *same* MAS run to the same depth, and left to itself its final report commits to a confident primary recommendation it cannot verify. The difference between a confident, unverified root-cause and an honest "exogenous and unverifiable" is, here, exactly the presence of the Five Whys discipline.

### B. Contestation Rises on the Hard Series, in the Mean

The single-chain behaviour reproduces in aggregate. When the true root is not in the data, the scaffold does not manufacture a clean resolution; it surfaces the competing mechanisms as *unresolved*, which registers as a rise in the number of contested claims (validity driven below base weight, Section III-B). Table I and Fig. 4 report the three-seed mean contested count with the reviewer off and on, for the capable-MAS configurations on the industrial series.

Across the three capable-MAS pairings the three-seed *mean* moves the same way: the reviewer raises the mean contested count on the hard series (Table I). The primary pairing (`haiku` × `opus`, the configuration of Fig. 3) rises in the mean from 5.3 to 9.7, and the two auxiliary pairings likewise ($15.0 \rightarrow 19.7$ and $10.3 \rightarrow 13.0$). We read this *aggregate* direction — all three capable-MAS means rise — not the magnitude or any single seed, as the signal: the per-seed counts are highly variable and individual seeds can reverse (Fig. 4). The rise is thus a mean-level effect: in this ground-truth-free setting the magnitude is not the signal and the robust evidence is the qualitative mechanism (Section V-A) together with the regime contrast (Sections V-C to V-D). The interpretation is the one the dialogue makes concrete — on a signal whose causes are exogenous, an honest analysis should leave the competing mechanisms standing as contested rather than collapse them into one confident story.

### C. The Honest Verdict is the Target; the Scaffold's Value is Bounded

That "exogenous and unverifiable" is the *appropriate* destination, not an artefact of a particular model, is shown by the top-tier peer configuration (E3, MAS and reviewer both `opus`). Here the baseline already reaches the honest verdict unaided: without any prompting it states in its own report that the "causes are exogenous and unprovable from this data … cannot be proven," and contests its own claims. With the reviewer added the contested count does not move (saturated, $9.3 = 9.3$ over three seeds); the Five Whys only sharpens framing (separating a statistical change from an unprovable cause) rather than producing a categorical correction.

The scaffold's value is therefore bounded above by the model's own capability for honest self-regulation: where the model already reaches the honest terminal on its own, the scaffold is redundant. It is also bounded below — the smaller open-weight model gains little, as it neither sustains the contestation nor acts on the probing (Section IV-B). The effect

TABLE I. CONTESTED-CLAIM COUNT ON THE INDUSTRIAL SERIES, BASELINE VS. FIVE WHYS: THREE-SEED MEAN WITH THE THREE INDIVIDUAL SEED VALUES IN PARENTHESES. FOR THE CAPABLE-MAS PAIRINGS THE THREE-SEED *mean* RISES (THE SCAFFOLD SURFACES UNRESOLVED COMPETING MECHANISMS), THOUGH PER-SEED COUNTS ARE HIGHLY VARIABLE AND INDIVIDUAL SEEDS CAN REVERSE — THE EFFECT IS AT THE MEAN LEVEL. THE TOP-TIER PEER PAIRING IS PER-SEED IDENTICAL BETWEEN ARMS (E3 SATURATION, SECTION V-C); ON THE BENCHMARK SERIES ALL THREE SEEDS MOVE DOWN (E4 RECONCILE, SECTION V-D). A LEADING CONFIG COLUMN MAPS EACH ROW TO THE E1–E4 CONFIGURATIONS. PER-SEED COUNTS ARE EXTRACTED FROM THE FROZEN RUN ARTEFACTS

| **Config** | MAS × reviewer | Dataset | baseline (per seed) | Five Whys (per seed) | mean effect |
|---|---|---|---|---|---|
| E1 / E2 | `haiku × opus` | industrial | 5.3 (9, 4, 3) | 9.7 (13, 2, 14) | ↑ surface |
| E2-aux | `opus × haiku` | industrial | 15.0 (17, 15, 13) | 19.7 (23, 25, 11) | ↑ surface |
| E2-aux | `haiku × haiku` | industrial | 10.3 (18, 11, 2) | 13.0 (34, 5, 0) | ↑ surface |
| E3 | `opus × opus` | industrial | 9.3 (8, 9, 11) | 9.3 (8, 9, 11) | = saturated |
| E4 | `haiku × opus` | benchmark | 7.0 (7, 2, 12) | 3.7 (0, 0, 11) | ↓ reconcile |

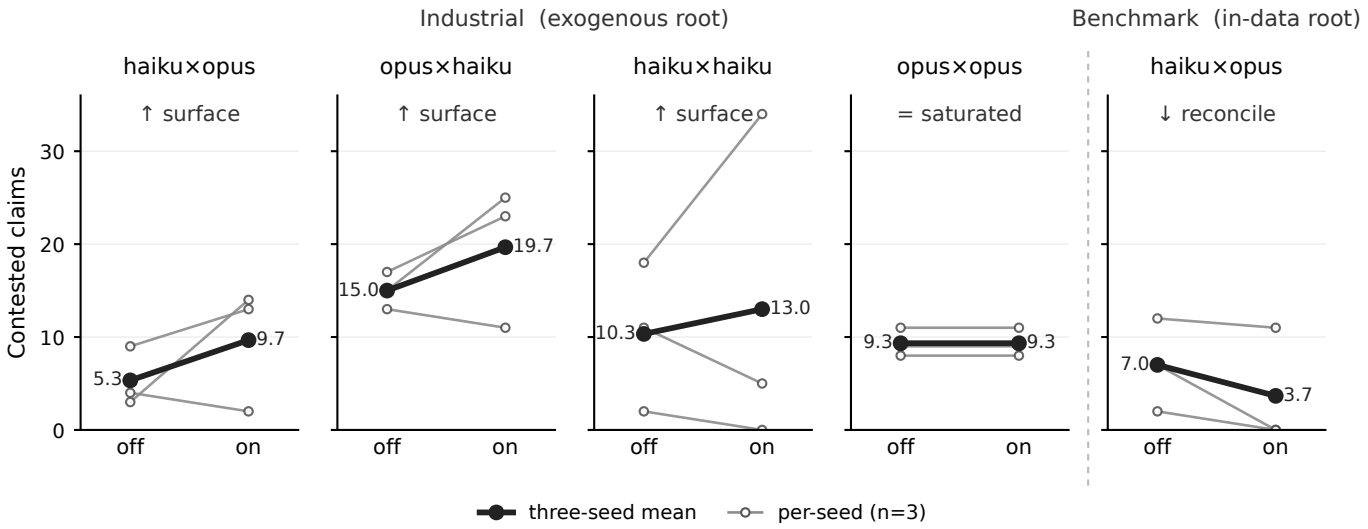


Fig. 4. Contested-claim count, baseline (Five Whys off) → Five Whys (on), as a paired-slope plot. Each thin line is one seed and the bold line is the three-seed mean; direction is read from each line's own slope. On the industrial series the three-seed mean rises for all capable-MAS pairings (↑ surface) while individual seeds reverse — notably `haiku×haiku`, whose mean rise is driven by a single seed (18→34) against two that fall; the top-tier peer pairing is per-seed identical between arms (= saturated); the benchmark series moves down in all three seeds (↓ reconcile). The industrial-↑ versus benchmark-↓ contrast is the adaptive signature of Section V-D; the per-seed values are tabulated in Table I.

of Sections V-A to V-B lives in the band between these bounds: a model capable enough to act on "where is this grounded?" but not capable enough to impose that discipline on itself unprompted. This is the band in which an analysis MAS would otherwise hand an engineer a confident false root-cause, and it is exactly where the scaffold pays off.

### *D. The Scaffold Adapts: It Reconciles When a Root Exists*

The rise in contestation on the hard series is not the scaffold "adding noise." On the benchmark series, whose root is approximable in the data, the *same* scaffold has the opposite effect: it drives the contested count *down* (E4, 7.0 → 3.7 in the mean, and down in all three seeds). There the Five Whys chain converges the competing narratives — whether a mid-series event created demand or merely accommodated it — onto a single in-data root (the arithmetic of the growth-and-seasonality structure), reconciling what looked contradictory. The scaffold thus does the structurally appropriate thing in each regime: it *reconciles* competing mechanisms when a groundable root exists, and *surfaces* them as contested when it does not. The discriminating variable is the data, not the prompt: a single operator produces convergence or honest non-resolution according to whether the signal can support a root at all. The regime contrast also speaks to a measurement concern: if the rise merely reflected the reviewer enlarging the graph, the count would rise in both regimes; instead it falls on the benchmark and rises only where the root is exogenous.

## VI. DISCUSSION

### *A. An Adaptive Discipline for Human–AI Coexistence*

The scaffold's behaviour across the two regimes (Section V-D) is what makes it a discipline rather than a bias. A device that simply increased scepticism would degrade the benchmark case, where a clean root genuinely exists; instead the same operator reconciles there and surfaces contestation on the exogenous-root signal. The discriminating variable is the data, not the prompt. Two design properties account for

this. First, the facilitator is *data-blind*: it supplies no answers and only demands that the MAS ground its own claims, so every convergence or retraction is the analyst's own work on evidence it computes. The honesty is elicited, not injected. Second, the discipline is structural: it is the Five Whys chain, and the argumentation graph it acts on, that carry the behaviour, and the effect survives changing which model occupies the analyst and reviewer roles (Section V-B).

We read this as a small, concrete instance of a broader stance on how to deploy capable LLMs on the factory floor. The aim is not to make the model less confident or to catch it out, but to let a domain method — one the manufacturing tradition already trusts for getting to true causes — surface what the analysis can and cannot ground, so that a practitioner receives the system's output correctly. This is the *genchi-genbutsu* ("go and see") instinct — grounding the question in what the data actually shows — operationalised as a reviewer loop, and the *jidoka* instinct (let the system raise its own hand when something is not right) applied to causal reporting. It is also why the value is bounded (Section V-C): where the model already regulates itself the discipline is redundant, and where it cannot act on the probing at all it does not help. The interesting and useful band is the one in between — precisely where a capable, fluent analysis MAS would otherwise hand an engineer a confident false root-cause.

*1) Deployment considerations:* The scaffold is cheap to add: one data-blind reviewer call per analysis round (bounded at eight rounds here), and because the reviewer sees only the running conclusion and the why-chain — not the data or the claim graph — its context stays small. It changes nothing about the model, adds no sensors, and needs no labelled data, so the adoption barrier is low; its one prerequisite is an analyst model capable enough to act on the probing (Section V-C). A practitioner can judge whether it applies to a new process by one structural question: are the variables that truly drive the signal recorded in the data, or exogenous to it? Where they are exogenous (our industrial case) the honest-abstention behaviour is the relevant value; where they are in-data the scaffold instead reconciles (Section V-D). The facilitator is itself an LLM; putting a human practitioner in that role is the *jidoka*-facing extension we leave to Section VII.

### B. Limitations

*1) The primary evidence is qualitative, by design:* Our central claim rests on a mechanism — self-refutation toward honest abstention — shown most clearly in a single-chain dialogue and corroborated by a directional contestation effect in the three-seed mean. We deliberately keep aggregate metrics off the main axis. Importance ranking is distorted on a non-trivial fraction of cells (about forty per cent) by cycles in the induced causal subgraph, which inflate the centrality of claims inside a cycle; a code audit located this and we therefore do not rest any claim on the importance ordering (the fix is deferred, Section VII). The three metrics are operationalisations grounded in established frameworks rather than validated instruments: we do not establish independently that grounding, validity, and contestation measure causal support or epistemic quality, which is part of why the aggregate metrics are kept off the main axis. We state these as limitations and do not build on them.

*2) Explored but not featured:* In the course of the study we examined several stronger hypotheses that did not reproduce robustly across three seeds, and we report that we set them aside rather than feature them: that the scaffold's marginal value is largest at intermediate model capability (an inverted-U), that a weak open-weight model fabricates a clean false root under the scaffold, and that reviewer capability and analyst capability form two separable axes governing stopping discipline. Each was visible in a single seed but did not survive seeding, and we treat the robust core — honest self-retraction and exogenous attribution — as the contribution. We mention the open-weight model only as a lower floor where the scaffold confers little benefit. Recording these explored-and-set-aside hypotheses is deliberate: it distinguishes our prospectively specified on/off comparison from post-hoc reading.

*3) Ground truth is unconfirmable on the industrial series:* By construction the true cause of the industrial signal is not in the data, so we cannot certify that the MAS's corrected analysis (Section V-A) is absolutely right. What we claim is narrower and well-supported: under the scaffold the system's *run-internal epistemics* forced a confident initial mechanism to be re-tested and withdrawn. We claim the honesty of the process and the appropriateness of the exogenous verdict, not the absolute correctness of any intermediate number; we cannot exclude that the scaffold occasionally drives retraction of a claim that was in fact correct — a cost we accept in exchange for not handing on unfounded ones.

*4) Scope:* The study uses two datasets, a single Five Whys chain, and three seeds. The contestation effect is a three-seed-mean rise across the capable-MAS configurations, with large per-seed variance and occasional per-seed sign reversals (Section V-B); it is a mean-level, not a per-seed-stable, effect, and the saturation and reconciliation results are each shown on one configuration. The behaviour should be expected to generalise only as far as these conditions; Section VII sketches the extensions that would test it. We restrict all claims to observed behaviour and offer no speculation about the internal causes of any particular model's outputs. The industrial series is also non-public, so that arm cannot be independently replicated; the benchmark arm uses a public dataset. We report model identifiers, seeds, chain count, and round budget, and give the operator in Appendix A, but not full API-level configuration such as sampling parameters, context limits, and retry behaviour, on which LLM outputs can depend.

*5) What would disconfirm the claim, and what is not yet isolated:* We state the claim so that it can fail: the epistemic-honesty reading would be disconfirmed if, within the capability band, the data-blind reviewer left confident ungrounded mechanisms standing, or drove the MAS to fabricate a clean root rather than to retract — neither of which we observe in the featured runs, though both are conceivable outside the band. Two comparisons are, moreover, suggestive rather than isolating. The regime contrast (Section V-D) varies the root locus (in-data vs. exogenous), but the two datasets also differ in complexity, in the presence of structural breaks, and in likely pre-training familiarity; the graded dataset spectrum of Section VII is what would isolate the root-locus variable. And the featured analyst–reviewer pairs are all within one model family; whether the self-retraction behaviour holds for cross-family pairings — the more realistic deployment

— is untested. Finally, the exogenous terminus rests on the reviewer's judgement that on-data alternatives are exhausted (Appendix A) rather than on a formal stopping rule, which may contribute to the per-seed variance noted above. Nor have we ablated the reviewer prompt itself: whether the retraction behaviour is specific to the Five Whys framing or would follow from any sceptical grounding demand is not isolated here, and a prompt ablation is the direct test. The reviewer is also data-blind but not context-free — it sees the principal conclusion and the accumulated why-chain. Those contain the MAS's own prior statements, so the channel introduces no evidence from outside the run, but it does carry the analysis history.

## VII. Future Work

*1) A graph fix for importance:* The causal-cycle distortion of the importance metric (Section VI-B) has a single-boundary fix: drop the low-support back-edges that close a cycle at edge-admission time, or compute reach on the strongly-connected-component condensation of the causal graph. With importance made trustworthy, the ranking could re-enter the analysis as a quantitative axis alongside the qualitative evidence rather than being held out of it.

*2) A dataset spectrum from in-data to exogenous root:* We showed reconciliation on a benchmark with an in-data root and contestation-surfacing on an industrial signal with an exogenous root. A natural next step is to vary the clarity of the root continuously — a graded series of datasets between the two regimes — and trace the reconcile-to-surface transition, which would turn the qualitative "adapts to the data" claim into a measured curve.

*3) Multi-chain and broader model coverage:* The present study uses a single Five Whys chain; the same honest-terminal behaviour should be tested when several independent chains are drilled in one run, and across a wider range of analyst and reviewer models — including cross-family analyst–reviewer pairings (the more realistic deployment) and additional open-weight families — to map the capability band in which the scaffold pays off.

*4) Human-in-the-loop facilitation:* The reviewer here is itself an LLM that never sees the data. A longer-term direction — and the one closest to the *jidoka* aim of the work — is to study how a human practitioner's way of asking "where is this grounded?" changes what the MAS surfaces, treating the facilitator role as a locus of stage-targeted human intervention rather than a fixed automated agent. We leave the design and evaluation of that human-facing variant to follow-on work.

## VIII. Conclusion

We ported Toyota's Five Whys onto a data-analysis multi-agent system not as a decomposition tool but as a data-blind reviewer frame, and found that it functions as an epistemic-honesty scaffold. On a real industrial signal whose causes are exogenous to the recorded data, a single Five Whys chain drove the system to retract three confident, grounded-sounding mechanisms in turn and to attribute the cause honestly to outside the data, where a matched baseline committed to a confident but unverified cause. The scaffold adapts to the data — it reconciles competing narratives when a groundable root exists and surfaces them as contested when it does not — and its value concentrates in the capability band where a model can act on the probing but cannot regulate itself unprompted, since the strongest model reaches the honest verdict unaided.

The contribution is less a new algorithm than a way of deploying capable LLMs honestly: rather than making the model less confident, we let a domain discipline the manufacturing tradition already trusts surface what an analysis cannot ground, so that a practitioner correctly receives what the system does and does not know. That is the form of human–AI coexistence we think industrial data analysis needs — not an AI that always answers, but one that is disciplined into knowing, and saying, when it cannot.

## Declaration on Generative AI

The authors used Claude (Anthropic) to assist with drafting and language refinement of the manuscript and with generating figures from already-computed experimental results. All AI-assisted text and analysis were critically reviewed by the authors, who take full responsibility for the content. Bibliographic details were verified against the original sources.

## Acknowledgment

This work was conducted as part of a joint research collaboration between Rikkyo University and Toyota Motor Corporation. We thank our collaborators at Toyota Motor Corporation for providing access to the industrial process data used in this study and for their domain expertise throughout the partnership.

## References

[1] T. Ohno, *Toyota Production System: Beyond Large-Scale Production*. Productivity Press, 1988.

[2] P. Baroni, A. Rago, and F. Toni, "From fine-grained properties to broad principles for gradual argumentation: A principled spectrum," *International Journal of Approximate Reasoning*, vol. 105, pp. 252–286, 2019.

[3] S. E. Toulmin, *The Uses of Argument*. Cambridge University Press, 1958.

[4] P. B. Ladkin and K. Loer, "Analysing aviation accidents using WB-Analysis — an application for multimodal reasoning," in *Working Notes of the AAAI Spring Symposium on Multimodal Reasoning*, ser. Technical Report SS-98-04. AAAI Press, 1998.

[5] D. Lewis, "Causation," *The Journal of Philosophy*, vol. 70, no. 17, pp. 556–567, 1973.

[6] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, 1967.

[7] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman *et al.*, "Towards understanding sycophancy in language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[8] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe *et al.*, "Self-refine: Iterative refinement with self-feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[9] 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 (NeurIPS)*, 2023.

[10] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 3563–3578.

[11] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[12] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou, "Large language models cannot self-correct reasoning yet," in *International Conference on Learning Representations (ICLR)*, 2024.

[13] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *International Conference on Machine Learning (ICML)*, 2024, pp. 11 733–11 763.

[14] A. Kimura, K. Fukuda, Y. Tahara, and Y. Sei, "Interruptible multi-agent debate: Sentence-level disclosure and urgency-based turn-taking for early error correction," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 17, no. 4, 2026.

[15] O. Bahidika, "Agentic accountability: "the buck stops where?" ethical frameworks for human oversight of autonomous AI systems," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 17, no. 5, 2026.

[16] Y. Chen, H. Xie, M. Ma, Y. Kang, X. Gao, L. Shi *et al.*, "Automatic root cause analysis via large language models for cloud incidents," in *Proceedings of the 19th European Conference on Computer Systems (EuroSys)*, 2024, pp. 674–688.

[17] J. Xu, Q. Zhang, Z. Zhong, S. He, C. Zhang, Q. Lin, D. Pei, P. He, D. Zhang, and Q. Zhang, "OpenRCA: Can large language models locate the root cause of software failures?" in *International Conference on Learning Representations (ICLR)*, 2025.

[18] B. Wen, J. Yao, S. Feng, C. Xu, Y. Tsvetkov, B. Howe, and L. L. Wang, "Know your limits: A survey of abstention in large language models," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 13, pp. 529–556, 2025.

[19] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez *et al.*, "Language models (mostly) know what they know," *arXiv preprint arXiv:2207.05221*, 2022.

[20] L. Kuhn, Y. Gal, and S. Farquhar, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation," in *International Conference on Learning Representations (ICLR)*, 2023.

[21] A. Rago, F. Toni, M. Aurisicchio, and P. Baroni, "Discontinuity-free decision support with quantitative argumentation debates," in *Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 2016, pp. 63–73.

[22] L. Amgoud and J. Ben-Naim, "Ranking-based semantics for argumentation frameworks," in *Scalable Uncertainty Management (SUM)*, ser. Lecture Notes in Computer Science, 2013, pp. 134–147.

[23] A. J. Card, "The problem with '5 Whys'," *BMJ Quality & Safety*, vol. 26, no. 8, pp. 671–677, 2017.

[24] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, "Large language models are zero-shot time series forecasters," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

## APPENDIX A
## THE FIVE WHYS REVIEWER OPERATOR

The reviewer is a facilitator agent (the top-tier model in the featured runs) that issues exactly one grounded "why" per round. It operates under a *clean-facilitator view*: it is given only the system's principal conclusion and the accumulated why-chain — *not* the dataset, the domain dossier, or the claim graph itself — so it cannot supply or verify facts itself and can only require the system to ground its own answers. The discipline, condensed from the operator template, is:

*1) Chain linkage:* Target the specific claim in the system's most recent answer; on the first round, anchor on the principal conclusion.

*2) Mechanism, not restatement:* Demand a causal mechanism, not a paraphrase, category, or bare correlation ("that restates the fact — why does it occur?").

*3) Fact-grounding (*genchi-genbutsu*).:* Require the system to cite a specific data feature or domain fact for its own answer; if it cites none or is speculative, send it back to the data and re-ask rather than descend. Never fill the evidence in.

*4) No symptom-stop / single chain.:* Treat an intermediate symptom as a new problem and continue; follow one dominant causal line rather than branching.

*5) Therefore-check.:* Periodically read the chain forward with "therefore" to catch logical leaps before going deeper.

*6) Dynamic stop at the root.:* Do not stop at a fixed count; stop when the chain reaches a generative driver at the edge of the evidence. The number five is a guideline [1].

The property that matters most for time-series analysis is the *observational-data terminus*. The operator explicitly recognises that the generative root of an observed series is often exogenous — not a variable present in the data — and instructs the reviewer, once the pattern's shape is grounded and the on-data alternatives are refuted, to declare the root reached and to name it only as "exogenous (a candidate from the domain context), not verifiable from the data," without asserting which specific domain fact it is. A guard prevents premature exogenous declaration: while the shape is still ungrounded, an unanswered "why" is genuine evasion of an answerable question and must be re-asked. This is what makes "exogenous and unverifiable" a disciplined terminus rather than an escape hatch.

The single-question phrasing template is:

> "Why — and by what mechanism — does {specific prior claim} occur, and where in the data is that grounded?"

The reviewer may add a brief pointer to where grounding should come from, but must not bundle several targets, assert or pre-answer the conclusion (leading the witness is a fidelity violation), or write a multi-paragraph interrogation. Each round emits one injected "why"; when the dynamic-stop test is met the reviewer instead declares the root reached, emits no further question, and states the root cause and the chain read backward with "therefore." A safety cap bounds recursion at eight rounds.