

Figure 1: **(a)** accuracy rises with scale; the pooled share of checkable claims the series contradicts stays in a 36.5–42.1% band; the same claims held at 7B's class mix fall 36.5–31.5%, because extremum and anomaly claims, the two refuted most, go from 35.7% of the rationale to 48.8/45.2% (hollow: the 32B bf16 control). **(b)** certify-abstain, the only arm at zero contradicted claims, labelled with its coverage (9.2–22.9%); **(c)** the accuracy–faithfulness plane on the pooled rate for the other arms; both checker-free self-verification arms sit *right* of CoT there (§4 standardises those deltas). Repair arms are omitted (their post-repair rate has another denominator, Table 2).

CHRONOCHECK: SCALE IMPROVES PER-CLAIM FAITHFULNESS AT A FIXED CLAIM MIX, AND SHIFTS TIME-SERIES RATIONALES TOWARD THE CLAIMS MODELS GET WRONG

Anonymous authors
Paper under double-blind review

ABSTRACT

Larger language models are more faithful on a fixed mix of claim kinds when they reason about time series, and they spend that gain on the kinds of claim they get wrong most. We measure this with ChronoCheck, a training-free oversight layer that decomposes a rationale into typed atomic claims (trend, extremum, periodicity, change-point, comparison, anomaly) and compiles each into a deterministic statistical check over the raw samples, so a rationale is refuted claim by claim. The checkers are conservative on the contradiction side: on a fresh suite of certified-true synthetic claims the contradiction criterion produced 0 false positives, and what they certify is consistency with the extracted claims rather than the truth of the rationale. Across three frozen scales chain-of-thought accuracy rises from 55.5% at 7B to 63.6% at 32B while the pooled contradicted-claim rate does not fall; held at the 7B mix of claim classes it falls from 36.5% to 31.5%, because claims naming a value or a location, the classes refuted most, take up a growing share of the rationale. Repair does not close the gap: no repair delta separates from zero at any Qwen2.5 scale (one ungated arm does on Llama-3.1-8B), and a large share of refuted claims disappear from the rewrite rather than being restated. Certify-abstain is the only arm to reach zero contradicted claims, at low coverage, with certified-subset accuracy gaps whose bootstrap intervals all cross zero. The checker measures time-series rationales; it does not repair them. We release it with the claim suite.

1 INTRODUCTION

Large language models are increasingly asked to reason aloud about time series: whether a server's CPU trace is trending upward, when a pump's vibration signature changed regime, whether a traffic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

stream is periodic. Systems that align series with LLMs or prompt frozen models with rendered samples produce confident narratives (Jin et al., 2024; Gruver et al., 2023; Xie et al., 2025; Kong et al., 2025); benchmarks grade only the final answer and the reasoning behind it is accepted on trust. That trust is measurable and it is misplaced: across the three scales we supervise, 36.5–42.1% of the checkable atomic claims inside chain-of-thought rationales (Wei et al., 2022) are contradicted by the very series the model was looking at, including questions whose final answer is correct. Held at 7B's class mix that share *falls* with scale, to 31.5%; the pooled share rises only because, under each tier's own extractor, the larger model states more of the classes it is worst at (Figure 1; one class ranks differently under a shared extractor, Appendix A).

Factuality research offers a general remedy, decompose text into atomic claims and verify each against evidence (Min et al., 2023; Lu et al., 2025), and it has reached structured evidence (Theologitis et al., 2026) and time series (Strong & Vlachos, 2025). But the verifier is itself a model, so the oversight inherits the failure modes it is meant to catch. Time series are the case where that is avoidable: whether a trend, change-point, period or spike is present is a statistical question with deterministic, auditable decision procedures, so no learned verifier is needed and the checkers' own error can be calibrated on certified synthetic series.

ChronoCheck instantiates this position as a training-free oversight layer (Figure 2): typed claims, one deterministic checker per predicate returning *consistent*, *contradicted* or *unverifiable* (contradicting only under a criterion calibrated at 0 false contradictions on 17,922 certified-true claims) and verdicts that drive certification, abstention, or repair gated to load-bearing contradictions. This is weak-to-strong oversight (Burns et al., 2024; Bowman et al., 2022) whose weak supervisor's own precision and recall are measured. A local sizing pilot fixed the design before any cluster spend (§4); its $n = 200$ quantized read localizes mechanisms rather than effect sizes. The full matrix ($n = 887$, three frozen scales, §4) answers the design's question in the negative. Repair benefit never resolves above question-level noise at any of the three scales; the point deltas trace a non-monotone pattern (-2.7, +3.0, -2.6 points at 7B/14B/32B) whose individual signs paired per-question tests cannot confirm (smallest $p = 0.068$), though it does at Llama-3.1-8B, upward, where the CoT arm itself falls below zero-shot (§4), and matching each refuted claim to the rewritten rationale says what the rewrite does instead: it drops the refuted claim far more often than it restates one the checkers accept (§5). Certification is contradiction-free by construction at every scale; what it buys is coverage, not measurable accuracy. The two checker-free self-verification loops move faithfulness in the opposite direction: both were introduced as ways to make a model's own output more factual by checking it (Gou et al., 2024; Dhuliawala et al., 2024); our adaptations are tool-free (self-critique without the published loops' tools), and here adapted CRITIC at 7B is the one Qwen2.5 delta of 28 paired arm-vs-CoT tests that survives a Bonferroni correction and it resolves *downward* (-6.2 points), adapted Chain-of-Verification loses -2.6 points at 7B with a CI excluding zero, and both raise the pooled contradicted-claim rate, though standardised the same way, only Chain-of-Verification at 7B keeps a CI clear of zero (§4). ChronoCheck is process oversight with the verifier on the data side: each claim compiles to a check on the raw series, so the verdict cannot be written by the model.

Contributions. **(C1)** A typed atomic-claim schema for time-series reasoning and a claim-to-checker compiler gated at zero false positives on certified synthetic series before any benchmark run (section 3). **(C2)** A claim-level faithfulness measurement read separately from end-task accuracy and from claim composition, calibrated on a fresh certified suite: contradiction precision 100.0% in every class, detection 91.2–100.0% per class (§4). **(C3)** Three verdict-driven interventions (certification, abstention and load-bearing-gated repair) measured against zero-shot, CoT and three checker-free adapted baselines (section 4). **(C4)** A weak-to-strong oversight axis from 7B to 32B on which scale lowers the refutation rate standardised to 7B's class mix and not the pooled rate over the rationale, read under each tier's own extractor; neither repair variant delivers a paired-significant accuracy change, and the rewrite drops the refuted claim more often than it restates it (section 4, section 5).

2 Related Work

LLMs meet time series. LLMs are adapted to numeric sequences by reprogramming patch embeddings (Jin et al., 2024), tokenising digit strings (Gruver et al., 2023), or aligning synthetic series for description and QA (Xie et al., 2025); benchmarks assemble multi-task QA (Kong

et al., 2025) and grade concepts strong models miss once causal and compositional reasoning is required (Cai et al., 2024); method-side work elicits temporal chain-of-thought (Chow et al., 2024), adds timestep embeddings (Wu et al., 2026b) and names trustworthy time-series reasoning as a target (Ning et al., 2025). Most of the QA line relies on chain-of-thought rationales (Wei et al., 2022) that nothing checks against the raw samples, so an accurate label can come from a fabricated reading; a rationale's faithfulness to its own intermediate structure is itself unreliable (Somov et al., 2026), and self-auditing before commitment (Yuan et al., 2026) checks beliefs, not the raw series.

Agentic, tool-augmented, and program-structured reasoning. A second line couples the LLM to external computation: agentic evidence-gathering with a self-critic (Liu et al., 2025), analysis-tool augmentation (Wu et al., 2026a), tool-executed operations for irregular TSQA (Chen et al., 2026), multi-agent debate (Trirat et al., 2026), recasting the series as an executable program (Kim et al., 2026), synthesizing process-verifiable traces for training (Zhou et al., 2026), and interpretable anomaly detection over the ops corpora NAB (Lavin & Ahmad, 2015) and SMD (Su et al., 2019), whose non-LLM detectors adapt their alarm threshold to the stream itself (Xiao et al., 2025). TS-Reasoner's programs *produce* answers (Ye et al., 2024), and LLM-driven detectors score the series directly (Alnegheimish et al., 2024; Liu et al., 2024). Grounding helps, but their verification is model-based rather than deterministic, entangled with retraining, or aimed at producing answers; ChronoCheck's programs *audit* a given rationale claim by claim and feed selective answering.

Claim decomposition and programmatic fact verification. The factuality community decomposes text into atomic claims and verifies each against evidence: FActScore scores atomic facts (Min et al., 2023), later work optimizes the decomposition for verifiability (Lu et al., 2025), ClaimDB checks claims against databases (Theologitis et al., 2026) and, closest to our setting, TSVer verifies claims against time-series evidence with a learned verdict predictor (Strong & Vlachos, 2025); for numeric reasoning, program-of-thought offloads computation to an interpreter (Chen et al., 2022) and CRITIC self-corrects through tool-interactive critiquing (Gou et al., 2024). The pattern recurs beyond text, rejecting an output by construction (Xiao et al., 2026a), but none of it defines a typed atomic-claim taxonomy for temporal signals discharged by deterministic statistical checkers on the raw series.

Weak-to-strong generalization and scalable oversight. Weak supervisors can elicit stronger capabilities (Burns et al., 2024), with ensemble refinements (Sang et al., 2024) and a measurement agenda for protocols in which limited supervisors constrain more capable models (Bowman et al., 2022; Sudhir et al., 2025). These motivate but do not instantiate an oversight mechanism for numeric reasoning, and oversight as deployed elsewhere is a score: anomaly monitors tune an operating point (Xiao et al., 2025) and calibrated selective gating withholds uncertain items to hold a risk target (Xiao et al., 2026b); each ranks confidence and none calls a claim false. ChronoCheck supplies cheap, deterministic statistical checkers, calibrated to zero observed false contradictions on certified synthetic series, as the weak-but-reliable supervisor of a stronger frozen LLM.

3 CHRONOCHECK

Setup and claim schema. ChronoCheck (Figure 2) wraps any frozen LLM that answers natural-language questions about a rendered time series with a chain-of-thought rationale: without modifying the model, it parses the produced text into *atomic claims*, verifies each against the raw series, and maps the verdicts to an intervention. Claims are typed against a six-predicate schema covering the vocabulary of time-series narratives: **trend** (directional movement over a window), **extremum** (a maximum/minimum value or its location), **periodicity** (a repeating pattern at a stated lag), **change-point** (a regime shift at a stated index), **comparison** (a stated statistic is larger in one window than another) and **anomaly** (a spike at a stated location). Each carries typed slots (window bounds, direction, claimed value or index, tolerance) filled by a few-shot extractor run with the same frozen LLM (§4); claims that fail to parse are excluded rather than guessed.

Executable checkers, conservative in the contradiction direction. Every predicate compiles to a deterministic checker over the exact rendered series the model saw, built from classical, auditable procedures: a tie-corrected Mann–Kendall test (Mann, 1945; Kendall, 1975) with Theil–Sen slope (Sen, 1968) for trend; robust direct evaluation with value and location tolerances for extremum; STL seasonal strength (Cleveland et al., 1990) with autocorrelation at the claimed lag for periodic-

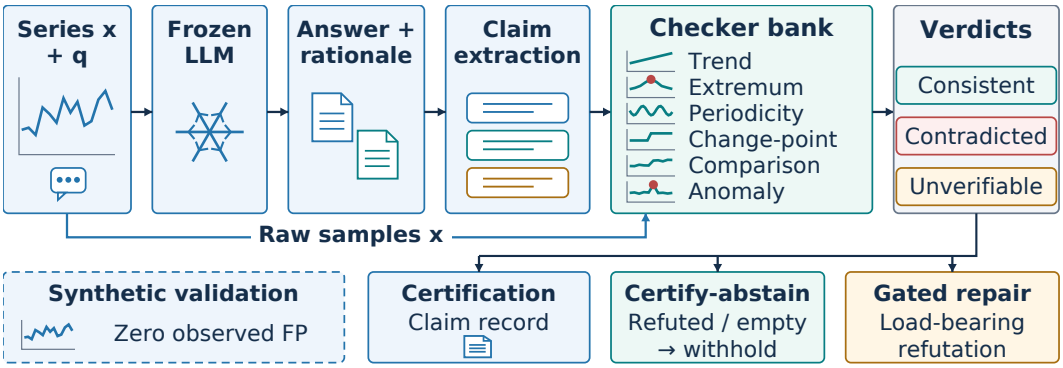


Figure 2: The oversight layer. A frozen LLM answers with a rationale; a few-shot extractor turns it into typed claims and drops what fails to parse. One deterministic checker per predicate (trend, extremum, periodicity, change-point, comparison, anomaly) reads the series, never the text, and returns *consistent*, *contradicted* or *unverifiable*. The verdicts feed three arms: certification ships the checked claim record, certify-abstain withholds an answer with a contradicted or no checked claim, and gated repair re-prompts only on a load-bearing contradiction.

ity; liberal/strict PELT (Killick et al., 2012) sweeps with a robust-scale shift criterion (Truong et al., 2020) for change-point; two-window same-statistic evaluation for comparison; and a robust z-score on a moving-median residual for anomaly. Each returns *consistent* (an endorsement with a measured, nonzero error rate), *contradicted* (the series refutes the claim under a criterion calibrated at zero false positives, §4) or *unverifiable*. Fusing heterogeneous channels into one score is a mapped design space (Xu et al., 2026); ChronoCheck takes its most conservative point: classes never vote, one contradiction vetoes.

The design principle is a *conservative contradiction direction*: a checker contradicts only under a conservative criterion and abstains everywhere else, including genuinely ambiguous regimes (abstain margins: Appendix A). We prefer an overseer that says little and is rarely wrong to one that is comprehensive and occasionally hallucinates, and enforce that as a hard gate: every checker is unit-tested against a synthetic suite with certified labels requiring **zero false positives on true claims** (one contradiction of a true claim freezes the experiment queue) plus non-vacuous detection power, so the gate is not met by abstaining on everything. The gate bounds the checkers' error on certified synthetic claims, not the error of the extraction from natural language that precedes them.

From verdicts to oversight. Verdicts feed three interventions, studied as separate arms. **Certification** ships an answer whose claims carry no contradiction together with a *checker-consistent certificate* (covering the extracted, checked claims; the rest is listed, not certified), the machine-checkable list of its claims and their verdicts: free, no re-prompt, no answer change, and it converts an unverifiable narrative into an auditable one. **Certify-abstain** withholds the answer entirely whenever any claim is contradicted or no claim was checked, so every *answered* question carries at least one checked claim and zero contradicted ones (a checked claim may still be unverifiable), trading coverage for a faithfulness guarantee, and moving along the accuracy–faithfulness plane by selection rather than repair. **Load-bearing-gated repair** re-prompts the model with the checker's specific counter-evidence (measured slope sign, detected change-point set), but only when the contradicted claim is *load-bearing*, that is, the answer references the claimed value, index or event; incidental contradictions are recorded in the certificate and never intervened on. The gate was designed on the 7B sizing pilot, where most contradictions were incidental (189 vs. 131) and re-prompting on them destabilized correct answers; §5 shows that premise is even at 7B and reversed at 14B and 32B.

Oversight calibration and the weak-to-strong axis. The synthetic suite carries complete ground truth, so the overseer's contradicted verdicts carry reported precision and recall (§4), which model-based verifiers cannot provide off their training distribution. The weak-to-strong question is whether counter-evidence usability scales with the supervised model: a repair loop helps only if the model can act on the refutation it is handed.

Table 1: Scale raises accuracy and, of the deltas shown, no oversight arm's delta over CoT survives its paired test except adapted CRITIC at 7B, which costs accuracy (the Llama-3.1-8B exception is §4): end-task accuracy (%) on the 887-question pool (vLLM, seed 0, greedy except self-consistency, unindented because it replaces the shared CoT decode). Intervention rows carry the signed delta vs. CoT from the exact counts (it can differ from the displayed operands by 0.1); best full-coverage row per column in **bold**; "—" = cell not run; † = generated on another machine, delta inside the cross-launch floor (App. B). Certify-abstain scores abstentions as wrong ("all") and also reports accuracy on the subset it answers, a conditional metric.

Arm	7B	14B	32B-AWQ
Zero-shot	51.5	57.0	61.8
CoT	55.5	58.1	63.6
+ self-critique (CRITIC)	49.3 (−6.2)	57.5 (−0.6)†	61.8 (−1.8)†
+ verify-then-revise (CoVe)	52.9 (−2.6)	58.7 (+0.7)	62.3 (−1.2)†
Self-consistency (k=5, T=0.7)	56.1 (+0.7)	60.5 (+2.5)	**64.8** (+1.2)†
+ naive repair	52.8 (−2.7)	**61.1** (+3.0)	61.0 (−2.6)
+ gated repair	**56.4** (+0.9)	59.6 (+1.6)	61.7 (−1.9)
+ certify-abstain (all)	13.6	5.6	7.2
acc. on answered	59.6	61.0	66.0

4 EXPERIMENTS

Setup: models and benchmarks. Three frozen open-weight scales of one family, Qwen2.5-7B / 14B / 32B-AWQ (Yang et al., 2024), so the weak-to-strong axis varies scale inside a single tokenizer and prompting regime. Questions come from the TimeSeriesExam graded understanding exam (Cai et al., 2024), the ChatTS evaluation suite (Xie et al., 2025), and an anomaly-explanation track over the public ops benchmarks NAB (Lavin & Ahmad, 2015) and SMD (Su et al., 2019); ChronoCheck-Synth, our lifecycle-conditioned certified synthetic suite with complete claim- and event-level ground truth, serves the zero-false-positive gate and the calibration below and carries no benchmark questions. Time-MQA (Kong et al., 2025) is not drawn on: only 9.0% of its questions carry a long enough window (Appendix A); a 500-question TSQA pool at its own window and ChatTS-14B as a trained backbone are run as design-external settings (Appendix D). Both pools are stratified across the four benchmark sources.

Arms and baselines. Five arms, all inference-only: zero-shot and chain-of-thought prompting (Wei et al., 2022), the references every oversight arm is measured against, and CoT plus ChronoCheck in one of three intervention modes: ungated (naive) repair, load-bearing-gated repair and certify-abstain (§3). Three checker-free training-free baselines make the model its own verifier, each at all three scales: adapted CRITIC (Gou et al., 2024) (tool-free and post hoc, unlike the published loop, which queries external tools), adapted Chain-of-Verification (Dhuliawala et al., 2024) and self-consistency (Wang et al., 2023). Program-of-thoughts (Chen et al., 2022) stays out (§4); method-side systems (Ye et al., 2024; Xie et al., 2025; Alnegheimish et al., 2024; Liu et al., 2024) are placed in §2. All oversight arms share one greedily decoded CoT rationale, and CRITIC and Chain-of-Verification start from that same answer text, so pre-intervention claim statistics are identical by construction; self-consistency cannot, because its 5 samples at temperature 0.7 are its mechanism.

Research questions. RQ1: does claim-level faithfulness improve as end-task accuracy does? **RQ2**: where do the three intervention arms sit on the accuracy–faithfulness plane relative to the baselines? **RQ3**: which checker classes carry the effect? **RQ4**: does repair turn net-positive as the supervised model scales? The experiments consumed 118.2 of a 40 GPU·h cap (Appendix A); all checkers are pure CPU.

Soundness gate and supervision calibration. No benchmark number exists until the checkers pass the §3 gate, which licenses reading *contradicted* as a refutation rather than verifier noise. On a fresh certified suite disjoint from the gate fixtures (1,440 series, 17,922 certified-true / 19,607 constructed-false claims) the checkers issue 0 contradicted verdicts on certified-true claims (exact 95% FP-rate bounds $\leq$0.46%) and detect 91.2–100.0% of constructed-false claims per class. The error side is not empty: the *consistent* verdict endorsed 30 constructed-false anomaly claims, and abstention on certified-true claims costs 4.3% of coverage pooled (full grid: Table 3).

Table 2: Only certify-abstain reaches zero contradicted claims, and both checker-free self-verification arms raise the pooled rate: claim-level faithfulness. "Contra. pre" is over *all* 887 rationales, identical across oversight arms by the shared-rationale contract; the repair rows are *row-matched* on the rows that arm rewrote (636/394 at 7B), which by construction started above the pool rate. Zero-shot makes too few checkable claims (34/0/46) for a rate. The standardised row re-weights each tier's per-class rates to 7B's claim mix, so only per-class refutation moves it.

Metric	7B	14B	32B-AWQ
Claims checked (CoT)	4,282	5,182	5,284
Contra. pre, all 887 rows (%)	36.5	42.1	39.0
standardised to the 7B class mix (%)	36.5	33.0	31.5
extremum+anomaly share of claims (%)	35.7	48.8	45.2
Naive repair, repaired rows: pre→post (%)	44.8→37.3	45.4→44.1	42.5→40.1
Gated repair, repaired rows: pre→post (%)	49.6→36.5	50.9→42.7	48.2→40.0
Certify coverage (%)	22.9	9.2	10.9
Contra. on answered (%)	0.0	0.0	0.0

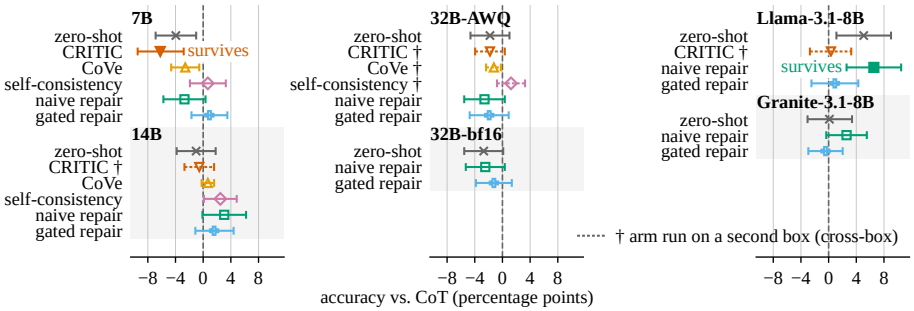


Figure 3: The 28 paired arm-vs-CoT tests of the declared multiple-comparison family: per-question deltas with 20k-resample bootstrap 95% CIs. Two survive Bonferroni ($\alpha = 1.8 \times 10^{-3}$): the 7B checker-free CRITIC loop, downward, and ungated repair at Llama-3.1-8B, upward (§4). Daggered rows ran on a second box.

Main matrix: three scales, full pool. A sizing pilot on a local RTX 5060 (Qwen2.5-7B-Instruct in 4-bit, $n = 200$; Table 4, Appendix A) confirmed the measurement axis but mis-estimated every effect it localized: the full pool overturns two of its signs, including gated repair at 7B (−6.0→+0.9). The matrix covers the three frozen scales over the full 887-question pool (TimeSeriesExam 496 / ChatTS 259 / NAB 100 / SMD 32), seed 0, eight arms per scale, vLLM on one 80 GB GPU (bf16 for 7B/14B, AWQ for 32B); the † cells ran on RTX 4090 cards.

RQ1: accuracy rises, and so does the pooled contradicted-claim rate. CoT accuracy rises 55.5% → 58.1% → 63.6% while the pooled contradiction rate does not improve: 36.5%, *rising* to 42.1% before 39.0%, over a growing claim volume (4,282–5,284), with the bf16 control in the same band (37.8%). But that rate weights each class's refutation rate by how often the model states the class: standardised to 7B's mix (Table 2) it *falls*, −3.5/−4.9 points on a 887-question cluster bootstrap ([-5.2, -1.8]/[-6.7, -3.1]). **Per claim class the larger models are mostly refuted less often (anomaly and trend are flat or rise, Figure 4); the rationale as a whole is not, because it states more of the two classes the checkers refute most.**

RQ4: no repair benefit resolves at any Qwen2.5 scale; one does at Llama-3.1-8B. Naive repair moves accuracy −2.7/+3.0/−2.6 points vs. CoT and gated repair +0.9/+1.6/−1.9 (harmful, then beneficial, then harmful again), which paired per-question tests say is not yet evidence: no Qwen2.5 per-scale repair delta separates from zero (exact McNemar, smallest $p = 0.068$; every bootstrap 95% CI crosses zero, e.g. [-0.2, +6.3] at 14B), and the bf16 rerun rules out AWQ as the driver of the 32B reversal (§4). At Llama-3.1-8B, though, ungated repair improves accuracy +6.5 points ([+2.5, +10.5], $p = 0.002$), the second of our 28 paired tests to survive the corrected $\alpha = 1.8 \times 10^{-3}$ (§4). **The repair axis is bounded, not resolved, for Qwen2.5: at $n = 887$ neither variant moves accuracy beyond a few points in either direction, and every interval crosses zero (Figure 3); the axis is not bounded for Llama-3.1-8B.**

The B1/B2 families: the registered read-out, and where it breaks. Two $\sim$8B models from unrelated lineages, Llama-3.1-8B and Granite-3.1-8B, rerun the five base arms on the same pool, registered before either produced a row. Llama-3.1-8B is clean for the extractor (smoke 30/30); Granite-3.1-8B is partially degenerate (smoke 22/30), so every Granite number carries that caveat. On Llama-3.1-8B the CoT arm scores 43.2% against zero-shot's 48.3%, so the ungated-repair delta that survives correction there (49.7%, +6.5 against CoT) returns the model to about its zero-shot accuracy and is read as CoT collapsing, not as counter-evidence use; the same caveat applies to TSQA at 32B. CRITIC on Llama-3.1-8B moves accuracy +0.3 points (CI [-2.8, +3.3]), inside the floor, and trails that family's ungated repair by -6.2 points ($p = 0.003$; Appendix A). The pre-repair contradiction rate and the claim-volume reduction under repair replicate at both families (Appendix C); the gate value does not; it is negative with its interval clear of zero at Llama-3.1-8B (-5.6 points, [-8.8, -2.3], $p = 7.7 \times 10^{-4}$) and, with the caveat, at Granite-3.1-8B (-3.0 points, [-5.6, -0.5], $p = 0.019$). **The headline therefore keeps Qwen2.5 as its subject** (§6).

RQ2: certification is the arm with a guarantee. The certificate invariant holds everywhere (0.0% contradicted claims on every answered subset), by construction rather than by estimate. Certified-subset accuracy reads above CoT at 7B (selection gap +4.1 points, CI [-1.8, +10.2]), 14B (+2.9, CI [-7.4, +13.1]) and 32B (+2.4, CI [-6.7, +11.3]); all three intervals cross zero, on 203/82/97 answered questions, and coverage is lower at 14B and 32B than at 7B (22.9/9.2/10.9%) as larger models make more checkable claims. **Certify-abstain's guarantee holds by design; it costs coverage and buys no measurable accuracy.**

RQ2 against the adapted baselines: the one arm whose accuracy change survives correction is a checker-free loop, and it resolves downward. Adapted CRITIC makes the model its own verifier: it critiques its own claims against the in-context samples for at most two rounds and revises unless it votes *correct*. At 7B it scores 49.3% against CoT's 55.5% (-6.2 points): 144 questions lost against 89 gained, exact McNemar $p = 3.8 \times 10^{-4}$, bootstrap CI [-9.5, -2.8], the only Qwen2.5 delta among our 28 paired arm-vs-CoT tests to survive a Bonferroni correction ($\alpha = 1.8 \times 10^{-3}$; the second survivor is ungated repair at Llama-3.1-8B, §4). The mechanism sits in the verdicts, not the revisions (§5); paired directly the gap resolves against *gated* repair (-7.1, $p = 2.2 \times 10^{-4}$) but not against *ungated* repair (-3.5, $p = 0.082$), so whether coverage or critique quality causes the cost stays open; both repair arms see checker counter-evidence computed from the raw series, and CRITIC does not. Faithfulness does not improve either, and what moves is the mix: the post-critique rationales carry 5,355 claims at 40.3% contradicted against 36.5% of 4,282 before, but only 37.3% standardised to the CoT class mix ([-0.6, +2.4]), the loop having pushed extremum claims from 20.2% of the rationale to 27.6%.

Chain-of-Verification, which answers its verification questions *without* the analysis in the prompt, the independence that separates it from CRITIC, scores 52.9% (-2.6 points, CI [-4.7, -0.5]), moving accuracy the way CRITIC does with a CI excluding zero though short of the corrected α, and contradicts on 38.8% of 4,548 claims against CoT's 36.5% of 4,282 (38.2% standardised, [+0.3, +3.2]): the one cell whose rise survives that control. Self-consistency scores 56.1% (+0.7, CI [-2.0, +3.3]) and contradicts on 36.4% of 4,448: CoT on both axes. At 14B the two diverge. Chain-of-Verification's accuracy cost disappears (+0.7, CI [-0.3, +1.6]) and its faithfulness cost does not survive that control (45.0% of 5,626 against CoT's 42.1% of 5,182, but 42.8% standardised, [-0.5, +1.8]), and self-consistency gains accuracy (+2.5, CI [+0.1, +4.9]) at CoT's contradiction rate (41.6% of 5,326). **The two checker-free loops raise the pooled contradicted-claim rate in every cell and cost accuracy at 7B, but with the claim mix held fixed only Chain-of-Verification at 7B still raises it, while sampling consensus gains accuracy at 14B and faithfulness at neither scale** (Appendix A).

32B quantization control. A full-tier bf16 rerun of the AWQ 32B tier (identical pool, seed, arms and decoding) resolves the quantization confound. The reversal survives (naive -2.5, gated -1.2 points vs. CoT, paired CIs again including zero) and every full-coverage arm lands within 1.0 points of its AWQ counterpart, below every arm effect the quantization might have explained (Table 6).

Per-class checker ablation (RQ3). For each of the six classes we re-ran the three oversight arms with that class's checker ablated, sound-side only: ablated-class claims take the abstain channel, so ablation can remove contradictions but never add them. It reuses the frozen shared rationale, so the read-outs are exact recomputations (plotted in Figure 6, moved to Appendix A with the full grid). Scope is the 7B tier (a 6×3 sweep was over budget); the certificate invariant holds

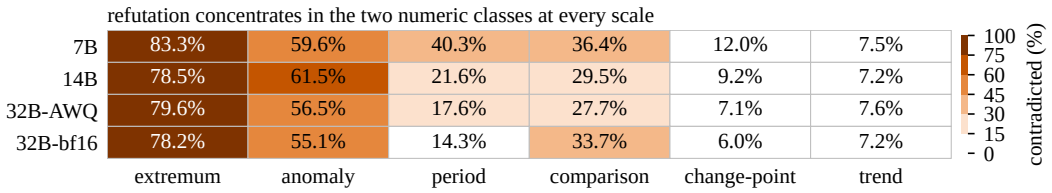


Figure 4: Extremum and anomaly are the two classes the checkers refute most at every scale, and their growing share of the rationale is what moves the pooled rate: contradiction rate by claim class and scale on the shared CoT rationale, columns ordered by the 7B rate. Re-weighting to the 7B class mix turns the pooled rise into a fall.

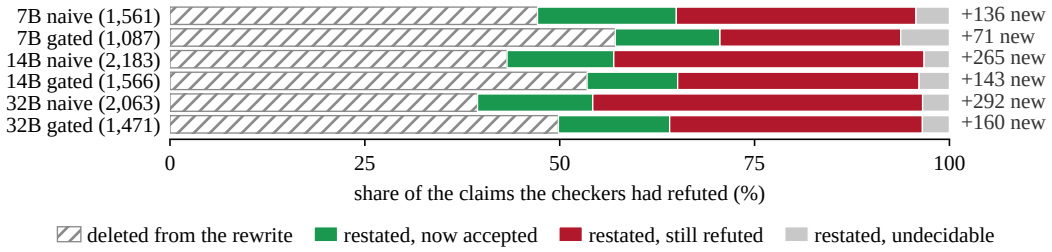


Figure 5: Repair deletes a refuted claim far more often than it fixes one: 39.4–57.2% of refuted claims vanish from the rewrite and 11.6–17.9% come back accepted. Each refuted claim is matched one-to-one to post-repair claims of the same class over an overlapping window (threshold 0.5 of the shorter window; unchanged at 0.25 and 0.75); row labels carry the arm and the number of refuted claims it rewrote, "+n new" the contradictions in the rewrite that track no earlier claim.

throughout. **Detection is concentrated in the extremum checker, and every class contributes to the certificate.** Ablating extremum moves the contradiction rate 36.5% → 19.6% (−16.8 points), nearly half of all refutations, with anomaly second (−9.2) and trend (−1.7) least; detection share does not track claim frequency: trend claims are among the most frequent the 7B model makes (988 claims, 7.5% contradicted) while extremum claims are refuted more often than not (864, 83.3%), only 13.4% of the refuted ones placing a local extremum within ±10 of the reported index (App. A). On the exact rows, all six ablations inflate coverage (+2.3 to +18.5 points) and every one *lowers* accuracy on the answered subset (down to −3.4): the answers each class had been blocking are worse than the certified average. **Detection concentrates in the claims that name a number, and every class contributes: removing any one raises coverage and lowers accuracy on the answers that coverage adds.** Program-of-thoughts stays out for scope rather than budget: the harness has no code-execution pathway and this pool's answers are natural language (Appendix A).

5 Analysis

The matrix says oversight helps through selection rather than revision. This section asks why, from the verdicts themselves; both analyses were registered before they were computed and the registration ships with the artifact.

Refutations concentrate in the claims that name a number. Split by class (Figure 4), the pooled rate of §4 separates into two populations. Extremum and anomaly claims are refuted 83.3/78.5/79.6% and 59.6/61.5/56.5% of the time at 7B/14B/32B, against 7.5/7.2/7.6% for trend and 12.0/9.2/7.1% for change-point, which only assert a direction or an event's existence. The registered prediction was that the rank order is identical at all three scales, and it is not: period and comparison exchange places between 7B and the larger models (period 40.3→21.6→17.6% against comparison 36.4→29.5→27.7%), and change-point and trend exchange places at 32B inside a gap the registration counts as a tie. The split survives: the two most-refuted and the two least-refuted classes are the same at every scale.

Repair removes the refuted claim more often than it fixes it. The rewritten rationales carry a lower contradiction rate (Table 2) over a claim volume changed by $-46.0/-37.4/-32.2\%$ (Appendix C). Matching every refuted claim to the rewrite separates deletion from restatement, and the registered prediction, deletion, is what happened at every scale, for both arms, at every registered threshold (Figure 5). At 7B the ungated arm rewrote 636 rows carrying 1,561 refuted claims: 736 (47.1%) do not appear in the rewrite at all, 480 (30.7%) come back and are refuted again, and 278 (17.8%) come back in a form the checkers accept; across the six arm/scale cells deletion accounts for 39.4–57.2% of refuted claims and accepted restatement for 11.6–17.9%. Repair also introduces 136 contradictions at 7B and 292 at 32B that track no earlier claim.

The load-bearing gate depends on a premise that scale reverses. The gate re-prompts only on contradictions the answer depends on, and it was designed on a pilot in which most contradictions were incidental. At 7B the split is even (787 incidental against 774 load-bearing, a margin no binomial interval on these counts resolves) and it reverses at 14B and 32B (955 vs. 1,228; 980 vs. 1,083), so a gate that withheld repair on about half of the contradicted rows at 7B withholds it on a minority at 14B and 32B and the share of rows it repairs rises from 44.4% to 56.5/55.8%. Its point estimates move accordingly (+0.9, +1.6, -1.9 vs. CoT), but we ran no paired test of gated against ungated repair on Qwen2.5. The B1/B2 registration does run it, and the mechanism does not transfer: at both new families the gate value is negative with its interval clear of zero (§4).

A self-verdict fires almost always and carries almost no information. The adapted CRITIC loop returns *flawed* on 885 of 887 first-round critiques, and because the pool has ground truth that verdict can be scored: of the answers it called flawed, 490 were in fact correct (55.4%) against a base rate of 55.5% on the same pool, a gap of -0.1 points. Two registered predictions failed: repair does not turn durably net-positive with scale (§4), and the class ordering is not scale-invariant.

6 Discussion

Weak-but-deterministic oversight is measured, and that changes the claim. The design turned the pilot's 7B repair failure into a falsifiable hypothesis, that repair turns durably net-positive as counter-evidence usability scales, and the matrix rejects it: 14B is the only scale where both variants read positive, its ungated delta sits nearest to significance (95% CI [-0.2, +6.3]). One greedy run per precision cannot separate a capability effect from prompt-sensitivity, so we report a bounded null, bounded to Qwen2.5 (§4).

Limitations. Three model families appear here and do not tell one story: the gate value (gated against ungated repair) is negative with its interval clear of zero at Llama-3.1-8B and, the tier being partially degenerate for the extractor, at Granite-3.1-8B, and ungated repair significantly helps at Llama-3.1-8B (§4), so the repair claims stay scoped to Qwen2.5. The matrix is one run per cell at seed 0: the paired per-question tests quantify question-level sampling uncertainty rather than run-to-run decode variance. Claim extraction failed on a minority of rationales (850/853/877 of 887 parsed, excluded and never guessed); the schema covers six predicate families and abstains outside them, so cross-series correlation and complex localization pass unverified; the per-class ablation is 7B-only; and the adapted baselines bound what *these* self-verifiers do, not self-correction in general.

7 Conclusion

Time series make oversight itself auditable: the ground truth is computable, so reasoning can be checked claim by claim. Scale lowers the refutation rate standardised to 7B's class mix (-4.9 at 32B, [-6.7, -3.1]) but not the pooled rate over the whole rationale, because under each tier's own extractor the larger model states more of the classes it is worst at. Repair produces no paired-significant accuracy change at any Qwen2.5 scale; certify-abstain holds its guarantee and pays in coverage without a measurable accuracy gain.

REPRODUCIBILITY STATEMENT

Every numeral in this paper is generated from the committed result artifacts (`results/pilot_results.json` for the local pilot; `results/main_results.json` and `results/gpu_ledger.json` for the main matrix and the 32B bf16 quantization control; `results/rq3/` for the per-class checker ablation, whose per-question dumps are `results/main_rq3/<class>/*.jsonl`, shipped as `archives/main_rq3.tar.gz` and extracted into `results/` by the commands in `supp/README_supp.md`; `results/paired_stats.json` for the paired per-question significance analysis; `results/calibration_supervision.json` for the supervision calibration; `results/critic_verdict_recheck.json` for the adapted-CRITIC arm's self-verdict audit) by a macro generator, with no hand-transcribed numbers; experiment cells that have not run (program-of-thoughts, which stays out for the scope reason in Appendix A) carry no numbers at all rather than stand-in values. The claim schema, checker implementations, soundness suite, pilot harness, and per-question raw records are released in the supplementary zip (`supp/README_supp.md`), and Appendix A details prompts, decoding parameters, and compute.

AI USE STATEMENT

In this work we disclose generative AI use under the categories of the ICLR 2027 AI policy. Within the required-disclosure categories, we used generative AI tools to help develop the conceptual framework, propose and refine hypotheses, formulate the paper's formal claims and statistical analyses, give feedback on research methodology and experiment design, implement methods (the experiment, analysis and figure-generating code released with this work), clean and reformat datasets, and interpret results. We have *not* used generative AI tools to generate synthetic data sets, assist with translation, support qualitative or thematic data analysis, or write proofs (this paper states no formal proof); the remaining required-disclosure categories are not applicable to this work, which has no human-subject component and no data collection beyond public benchmarks and the outputs of open-weight models. Under the recommended categories, generative AI was additionally used for software code, literature search and analysis, brainstorming, figure preparation, and editing the text of this paper. No experimental result was produced, selected or filtered by a generative AI tool: every experimental numeral in this paper is emitted by a committed single-writer artifact and reaches the text through an auto-generated macro bridge that admits no hand-typed value; design constants, protocol settings and figures quoted from prior work are of course written by hand and cited. The research questions, the claims and the conclusions are the authors'. We have reviewed all AI-assisted work and take responsibility for the final content, including every claim, number and artifact.

ETHICS STATEMENT

The benchmarks are public time-series corpora and the reasoners are public open-weight checkpoints; there is no human subject, no personal data and no collection of our own. Two things the template asks about do apply.

The first is what a certificate here means. The checkers decide whether a claim about a series is contradicted by deterministic statistics over that series, so a certified rationale is one whose claims survive those checks; it is not one that is true, nor one that is safe to act on. Coverage is bought by abstention, and the checkers abstain wherever a series is too short or a predicate does not apply; §7 reports that certify-abstain is contradiction-free by construction and says what it costs. Reading a certificate as a guarantee about a downstream decision would be reading it for a property nothing here measures.

The second is research integrity about our own compute. The experiments were budgeted at 40 GPU·h and consumed 118.2, in ten tranches; one launch knowingly bypassed the per-launch guard that would have refused it. Appendix A states each tranche, the mechanism of each, and what was cancelled rather than restating the cap, and the hours are recorded unchanged in the ledger that ships with the artifact.

REFERENCES

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. Large language models can be zero-shot anomaly detectors for time series? arXiv:2405.14755, 2024. arXiv:2405.14755.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, et al. Measuring progress on scalable oversight for large language models. arXiv:2211.03540, 2022. arXiv:2211.03540.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *International Conference on Machine Learning (ICML)*, 2024. doi: 10.48550/arXiv.2312.09390. arXiv:2312.09390.

Yifu Cai, Arjun Choudhry, Mononito Goswami, and Artur Dubrawski. Timeseriesexam: A time series understanding exam. In *NeurIPS 2024 Workshop on Time Series in the Age of Large Models*, 2024. arXiv:2410.14752.

Sanhorn Chen, Xiaoyang Chen, Boyu Liu, and Roy Zhao. Irts-toolbench: Towards verifiable agentic data science — solving irregular tsqa via tool-grounded reasoning. arXiv:2606.15107, 2026. arXiv:2606.15107.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research (TMLR)*, 2022. doi: 10.48550/arXiv.2211.12588. arXiv:2211.12588.

Winnie Chow, Lauren Gardiner, Haraldur T. Hallgrímsson, Maxwell A. Xu, and Shirley You Ren. Towards time series reasoning with llms. arXiv:2409.11376, 2024. arXiv:2409.11376.

Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.14314.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, et al. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2024. arXiv:2309.11495.

Zhibin Gou, Zhihong Shao, Yeyun Gong, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.11738.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.48550/arXiv.2310.07820. arXiv:2310.07820.

Ming Jin, Shiyu Wang, Lintao Ma, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.01728.

Maurice G. Kendall. *Rank Correlation Methods*. Charles Griffin, London, 4th edition, 1975.

Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012. doi: 10.1080/01621459.2012.737745.

Jaeho Kim, Changhun Oh, Seokhyun Lee, Irina Rish, and Changhee Lee. Representing time series as structured programs for llm reasoning. arXiv:2606.12481, 2026. arXiv:2606.12481.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, et al. Time-mqa: Time series multi-task question answering with context enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2503.01875.

Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms — the numenta anomaly benchmark. In *Proceedings of the 14th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2015. doi: 10.1109/ICMLA.2015.141. arXiv:1510.03336.

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Large language models can deliver accurate and interpretable time series anomaly detection. arXiv:2405.15370, 2024. arXiv:2405.15370.

Penghang Liu, Elizabeth Fons, Annita Vapsi, Mohsen Ghassemi, Svitlana Vyetrenko, Daniel Borrajo, Vamsi K. Potluru, and Manuela Veloso. Ts-agent: Understanding and reasoning over raw time series via iterative insight gathering. arXiv:2510.07432, 2025. arXiv:2510.07432.

Yining Lu, Noah Ziems, Hy Dang, and Meng Jiang. Optimizing decomposition for optimal claim verification. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2025. doi: 10.48550/arXiv.2503.15354. arXiv:2503.15354.

Henry B. Mann. Nonparametric tests against trend. *Econometrica*, 13(3):245–259, 1945. doi: 10.2307/1907187.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2305.14251.

Kanghui Ning, Zijie Pan, Yushan Jiang, et al. Towards interpretable and trustworthy time series reasoning: A bluesky vision. arXiv:2510.16980, 2025. arXiv:2510.16980.

Jitao Sang, Yuhang Wang, Jing Zhang, et al. Improving weak-to-strong generalization with scalable oversight and ensemble learning. arXiv:2402.00667, 2024. arXiv:2402.00667.

Pranab Kumar Sen. Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association*, 63(324):1379–1389, 1968. doi: 10.1080/01621459.1968.10480934.

Oleg Somov, Mikhail Chaichuk, Gleb Ershov, Karim Vafin, Mikhail Seleznyov, Alexander Panchenko, and Elena Tutubalina. Breaking the chain: A causal analysis of llm faithfulness to intermediate structures. arXiv:2603.16475, 2026.

Marek Strong and Andreas Vlachos. Tsver: A benchmark for fact verification against time-series evidence. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. arXiv:2511.01101.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2019. doi: 10.1145/3292500.3330672.

Abhimanyu Pallavi Sudhir, Jackson Kaunismaa, and Arjun Panickssery. A benchmark for scalable oversight protocols. arXiv:2504.03731, 2025. arXiv:2504.03731.

Michael Theologitis, Preetam Prabhu Srikar Dammu, Chirag Shah, and Dan Suciu. Claimdb: A fact verification benchmark over large structured data. arXiv:2601.14698, 2026. arXiv:2601.14698.

Patara Trirat, Jin Myung Kwak, Jay Heo, et al. Ts-debate: Multimodal collaborative debate for zero-shot time series reasoning. arXiv:2601.19151, 2026. arXiv:2601.19151.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. arXiv:1801.00718, 2020. arXiv:1801.00718.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2201.11903.

Xingjian Wu, Junkai Lu, Zhengyu Li, et al. Timeart: Towards agentic time series reasoning via tool-augmentation. arXiv:2601.13653, 2026a. arXiv:2601.13653.

Yafeng Wu, Huu Hiep Nguyen, Thin Nguyen, and Hung Le. Beyond tokenization: Direct timestep embedding and contrastive alignment for time-series question answering. arXiv:2606.18986, 2026b. arXiv:2606.18986.

Yifan Xiao, Zhongjun Wang, Tiantian Yan, Weiwei Li, Dazhong Wu, and Jie Ma. Time series anomaly detection: A hybrid deep learning framework with adaptive thresholding for electricity industry data. In *IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)*, 2025. DOI 10.1109/CSCLOUD66326.2025.00046.

Yifan Xiao, Shijie Li, and Yuhao Ge. RunbookFX: Type- and effect-safe LLM synthesis for executable incident diagnosis and mitigation. *Proceedings of the ACM on Programming Languages*, 10(ICFP), 2026a. Article 277; DOI 10.1145/3828675.

Yifan Xiao, Shijie Li, and Yu Huang. CALM-TS: Risk-controlled LLM labeling for time-series via calibrated selective gating. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026b. DOI 10.1145/3770855.3817859.

Zhe Xie, Zeyan Li, Xiao He, et al. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *Proceedings of the VLDB Endowment*, 18(8), 2025. doi: 10.14778/3742728.3742735. arXiv:2412.03104.

Dexuan Xu, Yanyuan Chen, Zhongyan Chai, Yifan Xiao, Yandong Yan, Weiping Ding, Hanpin Wang, Zhi Jin, Wenpin Jiao, Weihua Yue, Hang Li, and Yu Huang. Knowledge fusion in deep learning-based medical vision-language models: A review. *Information Fusion*, 125:103455, 2026. doi: 10.1016/j.inffus.2025.103455.

An Yang et al. Qwen2.5 technical report. arXiv:2412.15115, 2024. arXiv:2412.15115.

Wen Ye, Wei Yang, Defu Cao, Yizhou Zhang, Lumingyuan Tang, Jie Cai, and Yan Liu. Ts-reasoner: Domain-oriented time series inference agents for reasoning and automated analysis. arXiv:2410.04047, 2024. arXiv:2410.04047.

Wenhao Yuan, Chenchen Lin, Jian Chen, Jinfeng Xu, Xuehe Wang, and Edith Cheuk Han Ngai. Verify before you commit: Towards faithful reasoning in llm agents via self-auditing. arXiv:2604.08401, 2026.

Jiahui Zhou, Dan Li, Boxin Li, et al. Time series reasoning via process-verifiable thinking data synthesis and scheduling for tailored llm reasoning. arXiv:2602.07830, 2026. arXiv:2602.07830.

A Reproducibility Details

Rounding conventions. Every printed figure is derived by the macro bridge from the integer counts in a committed artifact, differences included, and quantized once; no figure is a re-rounding of a stored rate. Point estimates round half-up. *Bounds and interval endpoints round away from the claim*: upper bounds and band maxima up, lower bounds and band minima down, so a stated guarantee is never tighter than the one that was computed and a stated band never excludes a measured extreme. A range of observed per-tier values is not a bound: the pooled contradicted-claim range (36.5–42.1%) prints the extreme tiers' point estimates, the same values as their table cells.

Code and artifacts. The supplementary zip (start at `supp/README_supp.md`) carries the claim schema and few-shot extractor (`claims.py`, `extract.py`), the six checkers with documented abstain margins (`checkers.py`), the certified lifecycle-conditioned series generator (`generator.py`), the soundness gate (`soundness.py` plus the unit-test suite), the arm implementations (`pipeline.py`), the pilot runner (`pilot.py`),

the single-writer aggregators of every committed result artifact (`aggregate_pilot.py`, `aggregate_main.py`, `aggregate_rq3.py`), the paired per-question significance analyzer and its artifact (`analyze_paired_stats.py`, `results/paired_stats.json`: exact McNemar on discordant pairs plus 20,000-resample percentile bootstrap, seed 0, validated against the committed aggregates before writing), the supervision-calibration analyzer and its artifact (`calibration_supervision.py`, `results/calibration_supervision.json`), the macro bridge (`gen_paper_macros.py`), and the GPU ledgers (`run_ledger.jsonl` for local-pilot hours, `results/gpu_ledger.json` for cluster hours) recording measured wall-time GPU hours per run.

Pilot configuration. Supervised model Qwen2.5-7B-Instruct under bitsandbytes NF4 double quantization with bf16 compute (Dettmers et al., 2023); greedy decoding throughout; the three oversight arms share one CoT rationale per question so pre-intervention claim statistics are identical by design. Question pool: $n = 200$ stratified over TimeSeriesExam (80), ChatTS eval (60), NAB (30), and SMD (30), seed 0; total measured GPU wall time 1.1 hours on one RTX 5060 (8 GB), counted in the ledger outside the cluster budget pool.

Why Time-MQA is not in the pool. Of the 111,503 Time-MQA questions (Kong et al., 2025) that carry a series, 10,057 (9.0%) reach the 64-point minimum a pool record needs, and all of those come from 1 of its three files; the per-file median lengths are 16, 30 and 24 points, at or below what the trend, anomaly and change-point checkers require before they abstain. A window that short cannot be asked the questions this method verifies. Whether a source that can only supply floor-length windows belongs in this population is a design question rather than a download; the admissibility scan ships with the artifact.

Prompts. All prompts (concision-bounded analyst system prompt, per-task answer instructions, the claim-extraction few-shot prompt, and the counter-evidence repair template) are verbatim in `pipeline.py` and `extract.py`.

Supervision calibration. Table 3 reports the per-class supervision-calibration grid of §4: contradiction precision 100.0% in every class, weakest class period (Wilson 95% CI [90.1, 92.3]), weakest subtype aperiodic-on-periodic at 80.0% [77.6, 82.2]; misses are overwhelmingly abstentions (523 of 553), the 30 endorsed constructed-false anomaly claims are 0.55% of that class and zero elsewhere, and abstention on certified-true claims peaks at anomaly 12.5%. The calibration suite is produced by the same certified generator and the same claim-emission protocol as the soundness gate (`calibration_supervision.py` imports the gate's emitters verbatim), on a seed set disjoint from the frozen gate fixture: disjointness is asserted at run time against the fixture's own recorded configuration and re-asserted at macro-generation time, where the bridge re-reads the fixture's config block (the only fixture field the firewall rule permits reading); the artifact records both seed sets, and, series being a pure function of (seed, index, scenario) with series ids that embed the seed, no calibration series *is* a fixture series. TRUE claims are certified by generator labels or direct computation; FALSE claims have their falsehood verified at construction time, which is what makes detection well-defined. For clean-window constructions that verification is *label*-level (clean windows are buffered from every injected event), not statistical: an organic noise excursion inside a labeled-clean window can still meet a checker's criterion, and the Cons_f column counts exactly those endorsements. The frozen gate fixture itself remains uncited, per the standing rule; this suite exists to be cited.

Pilot aggregate. Table 4 reports the five-arm local-pilot aggregate discussed in §4.

RQ3 full grid. Table 5 reports the complete per-class ablation grid of §4: the three deterministic rows that Figure 6 plots, plus the two repair rows: those compare fresh generations across launches, so their deltas carry run-to-run decode noise and are not read as individual signs. No ablation moves either repair arm by more than +2.5 points or flips any 7B point-estimate ordering. The one qualitative, mechanism-consistent observation in those rows is that removing the detection-heaviest class, extremum, moves both repair-arm point estimates up (naive +2.5, gated +1.2 points vs. Full), as expected if the refutations that trigger the most net-harmful repairs are the same ones the certificate screens on (§2), and we flag it as a hypothesis for a multi-run design rather than a finding.

What a refuted extremum claim looks like. The extraction schema records an extremum as a value and an index over the whole window, so a rationale sentence naming a local peak ("a

Table 3: Supervision calibration on the fresh certified suite (1,440 series of length 1,440, seeds 101–124 vs. the gate fixture's disjoint seed set; artifact `results/calibration_supervision.json`). Top, per class: certified-true claims (n_{t}) with contradicted-verdict false positives (FP), the exact Clopper–Pearson 95% upper bound on the true-claim FP rate, and abstentions on true claims ($\mathrm{Abst}_{\mathrm{t}}$, the coverage price); constructed-false claims (n_{f}) with detection power (Wilson 95% CI) and the miss split into *consistent* verdicts ($\mathrm{Cons}_{\mathrm{f}}$, endorsed falsehoods) vs. abstentions ($\mathrm{Abst}_{\mathrm{f}}$). With FP $= 0$, contradicted-verdict precision is 100.0% in every class and the pooled exact upper bound on the true-claim FP rate is 0.03%; the per-class precision Wilson CIs are in the artifact. Every bound in this table is rounded *up*, so a class rate quoted as a band endpoint in the text can read 0.1 below its own entry here. Bottom, by falsehood-construction subtype: each class detection rate above is the mix-weighted average of its subtype rows.

Class	n_{t}	FP	FP rate $\leq$ (%)	$\mathrm{Abst}_{\mathrm{t}}$	n_{f}	Det. (%)	Wilson 95% CI	$\mathrm{Cons}_{\mathrm{f}}$ / $\mathrm{Abst}_{\mathrm{f}}$
Trend	806	0	0.46	41	930	100.0	[99.5, 100.0]	0 / 0
Extremum	2,880	0	0.13	0	4,289	100.0	[99.9, 100.0]	0 / 0
Period	1,572	0	0.24	12	2,767	91.3	[90.1, 92.3]	0 / 242
Change-point	1,480	0	0.25	30	437	94.1	[91.4, 96.0]	0 / 26
Comparison	5,687	0	0.07	0	5,687	100.0	[99.9, 100.0]	0 / 0
Anomaly	5,497	0	0.07	687	5,497	94.8	[94.1, 95.4]	30 / 255

Class	Falsehood subtype	n_{f}	Det. (%)	Wilson 95% CI
Trend	flat-on-trend	465	100.0	[99.1, 100.0]
	trend-flipped	465	100.0	[99.1, 100.0]
Extremum	wrong-location	1,409	100.0	[99.7, 100.0]
	wrong-value	2,880	100.0	[99.8, 100.0]
Period	aperiodic-on-periodic	1,195	80.0	[77.6, 82.2]
	periodic-on-aperiodic	377	99.2	[97.6, 99.8]
	wrong-period	1,195	100.0	[99.6, 100.0]
Change-point	absent-over-cp	96	100.0	[96.1, 100.0]
	cp-in-stationary	341	92.4	[89.0, 94.8]
Comparison	inverted-comparison	5,687	100.0	[99.9, 100.0]
Anomaly	anomaly-in-clean	2,652	89.3	[88.0, 90.5]
	clean-over-anomaly	2,845	100.0	[99.8, 100.0]

Table 4: Local pilot (Qwen2.5-7B-Instruct, NF4 4-bit, $n = 200$, seed 0). "Contra. rate" is the fraction of checkable claims contradicted by the raw series. All oversight arms share one greedy CoT rationale, so their pre-intervention rate equals the CoT row by construction; repair rows show the post-repair rate. Certify-abstain withholds answers on any contradiction: we report accuracy counting abstentions as wrong (all), accuracy on answered questions, and its zero contradicted-claim certificate on that subset. Claim extraction succeeded on 178/200 rationales; unextractable claims are excluded from verification, never guessed, and a rationale with none is not certified.

Arm	Acc. (%)	Contra. rate (%)	Repaired (%)	Coverage (%)
Zero-shot	53.0	—	—	—
CoT	55.5	36.4 (pre)	—	—
+ naive repair	48.0	38.0 (post)	64.0	—
+ gated repair	49.5	34.2 (post)	34.5	—
+ certify-abstain	9.0 (all) 47.4 (answered)	0.0 (answered)	—	19.0

trough around index 40") becomes a global-extremum claim and is checked against the global extremum at the checker's 2%-of-range value tolerance. To measure how much of the extremum refutation rate that schema explains we re-read every refuted extremum claim against the series (`analyze_extremum_window.py`, post hoc, not blind). Of the 620 / 1,263 / 1,203 refuted claims that carry a value (7B / 14B / 32B-AWQ), 97.1% / 98.6% / 98.4% name a value some point of the series attains within that tolerance, but only 13.4% / 23.2% / 25.4% name a value that is a local extremum of the ± 10-index window around the reported index. The schema therefore explains

why refuted values are almost always real values, not why they are refuted: the reported location is not a local extremum in three quarters or more of the cases.

Fixed-extractor check. Every tier's claims are extracted by the tier's own model, so a class share that moves with scale could be the extractor moving rather than the rationale. `PREREG_FIXED_EXTRACTOR.md` re-extracts the 7B and 14B rationales with the 32B-AWQ extractor (4,932 and 5,691 claims) and ranks the tiers by each named class share under both extractors; the registered rule keeps the mechanism sentence unqualified only if every ranking agrees. They do not (no). The extremum share keeps its order (7B 20.9%, 14B 35.2%, 32B 28.8%), and so does the numeric share (37.1% / 49.7% / 45.2%), which is the quantity the body reports. The anomaly share reverses: 14B falls from first to last (17.2%→14.5%) while 7B rises (15.5%→16.2%). The mechanism sentence is therefore stated under each tier's own extractor throughout. The standardised contradiction rate under the shared extractor stays in direction: 7B 36.9%, 14B 34.2%, 32B 33.2% (pooled 36.9% / 42.5% / 39.0%). Against 7B the standardised rate moves −2.7 points at 14B ([-4.3, -1.1]) and −3.7 at 32B ([-5.4, -2.0]), with the question-level cluster bootstrap of the own-extractor deltas; these intervals were added after the check was registered. The check cost 2.85 GPU·h, booked in the ledger.

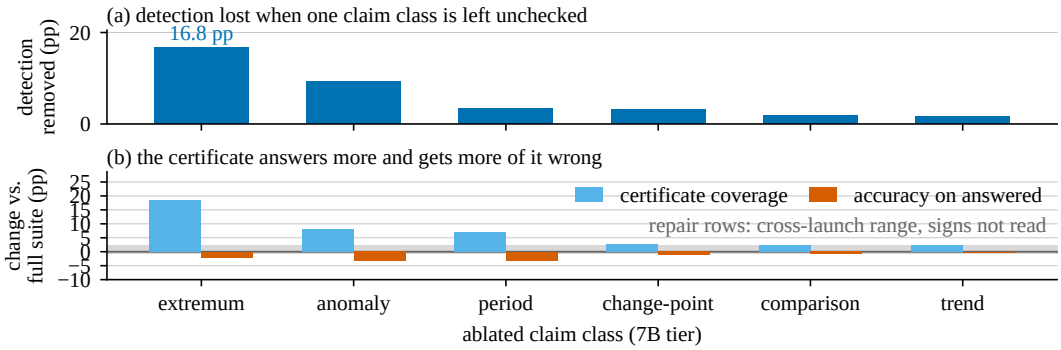


Figure 6: RQ3 per-class checker ablation, 7B tier: exact recomputations over the frozen shared rationale (4,282 claims in every column). **(a)** detection lost when a class stops being checked; **(b)** the coverage gained and the accuracy lost on the answered subset. Grey band: the range the repair arms move over across launches (their signs are not read). Moved here from §4 when the B1/B2 family read-out took the body past the page limit; no number changed.

Table 5: RQ3 per-class checker ablation, 7B tier, full grid (repair rows included; same protocol and delta convention as Figure 6).

	contradicted claims		naive repair		gated repair		certificate coverage		acc. on answered	
ablated class	%	Δ pp	%	Δ pp	%	Δ pp	%	Δ pp	%	Δ pp
none (full suite)	36.5	—	52.8	—	56.4	—	22.9	—	59.6	—
extremum	19.6	−16.8	55.2	+2.5	57.6	+1.2	41.4	+18.5	57.5	−2.1
anomaly	27.2	−9.2	53.1	+0.3	56.4	+0.0	30.9	+8.0	56.2	−3.4
period	33.0	−3.4	52.2	−0.6	55.7	−0.7	29.8	+6.9	56.4	−3.2
change-point	33.2	−3.3	53.3	+0.6	56.7	+0.3	25.5	+2.6	58.4	−1.2
comparison	34.5	−2.0	53.0	+0.2	56.6	+0.2	25.1	+2.3	58.7	−0.9
trend	34.7	−1.7	52.6	−0.1	56.3	−0.1	25.3	+2.4	59.4	−0.2

32B quantization-control aggregate. Table 6 reports the full arm-by-arm AWQ-vs-bf16 comparison summarized in §4.

Main-matrix and follow-up compute. Four full vLLM 0.7.3 runs (Qwen2.5-7B/14B-Instruct bf16, Qwen2.5-32B-Instruct-AWQ, and the Qwen2.5-32B-Instruct bf16 quantization control of §4; the 14B tier ran as two ledgered sessions after a mid-run restart, both counted), plus six RQ3 per-class ablation runs (7B tier, three oversight arms each, §4) and the 7B adapted-CRITIC top-up (§4, two calibration smokes and one full-pool run); all greedy, seed 0, the full 887-question pool, on the cluster GPU. All hours are measured on-box wall time, smoke runs included, recorded in `results/gpu_ledger.json` (local-pilot hours are

Table 6: 32B quantization control: the AWQ main-matrix tier vs. its bf16 rerun (same pool, seed 0, greedy, five arms). Δ is signed bf16$-$AWQ in percentage points, computed at macro-generation time on unrounded values (they can differ from the displayed operands' difference by 0.1). Repair rows carry the signed delta vs. that column's own CoT in parentheses. The certificate invariant (zero contradicted claims on answered questions) holds in both runs.

Metric	32B-AWQ	32B-bf16	Δ (bf16$-$AWQ)
Zero-shot acc. (%)	61.8	61.1	-0.7
CoT acc. (%)	63.6	63.8	+0.2
+ naive repair (%)	61.0 (-2.6)	61.3 (-2.5)	+0.3
+ gated repair (%)	61.7 (-1.9)	62.6 (-1.2)	+0.9
+ certify-abstain (all, %)	7.2	8.1	+0.9
acc. on answered (%)	66.0	65.5	-0.5
Certify coverage (%)	10.9	12.4	+1.5
Contra. pre (%)	39.0	37.8	-1.2

ledgered separately in `run_ledger.jsonl`). The dispatch script and its prompt set live in `runbook/main_matrix.sh` and `runbook/main_matrix.prompt.md`.

Adapted-CRITIC arm: protocol and verdict audit. The arm consumes the committed 7B CoT answers as its initial answers (byte-identical per question, asserted row-for-row before any statistic is computed) and generates only critiques and revisions, so the paired comparison of §4 has no relaunch noise on the reference side. Each episode runs at most two critique/revision rounds and stops early on a *correct* self-verdict; a round whose critique carries no parseable verdict line is treated as *flawed* and revised, never imputed as *correct* (all 4 such rounds triggered a revision). The final answer is then extracted and checked exactly as the other arms are, as a measurement that the arm itself never sees. Reusing the CoT answer byte-identically makes this strictly a *post-hoc revision* baseline (CRITIC as published lets the model write its own initial answer), which buys the noise-free reference side but also means the delta is measured against a rationale the critic has no authorship stake in; the round-1-vs-round-2 asymmetry below cuts both ways on that point.

Verdict-parser provenance. The early-stop parser was line-anchored *while the pool was already running*, so which parser the launch deployed is not recorded. Rather than assert it we settled the question from the records: `recheck_critic_verdicts.py` re-parses every stored critique with both the old and the fixed parser and compares each to the verdict the run acted on. On the complete 887-episode file the two parsers agree on all 1,772 rounds, the old parser produced zero false *correct* readings, and no episode stopped early on a verdict the fixed parser rejects, so no episode is contaminated and no rerun was needed. The audit artifact carries both the full-pool snapshot and the 30-episode calibration smoke, which is the one snapshot that indisputably ran on the pre-fix parser and is likewise clean on all 60 of its rounds (`results/critic_verdict_recheck.json`, which is also where the paper's verdict counts are bridged from).

What the self-verdict is worth. The verify prompt asks the model to re-derive six temporal properties and to close with "`Verdict: correct` if the analysis and its final answer survive every check". A conjunctive bar over six independently re-derived properties will return *flawed* almost always, so the 99.8% first-round rate is partly a property of that prompt rather than of the model's self-verification ability, and we run no prompt variant. The capability question is separable and costs nothing, because every judged answer has a known ground truth: of the 885 answers the first-round verdict called *flawed*, 490 were in fact correct (55.4%), against a base rate of 55.5% on the same pool, -0.1 points. The verdict fires almost always *and* carries essentially no information about whether the answer it judges is right, which is the property the checkers are built to supply and this baseline cannot.

Budget disclosure. The experiments were budgeted at 40 GPU·h; realized consumption is 118.2 GPU·h, an overrun of 78.2 h incurred in ten tranches. The first occurred because the 32B bf16 control and the RQ3 ablation were dispatched in parallel on two cards against one shared ledger: each launch individually passed the per-launch budget guard against the then-current total, but their combined spend crossed the cap mid-flight. The final approved ablation class was allowed to finish rather than be aborted near-complete, and the remaining approved item, the adapted training-free baseline arms, was cancelled under the pre-agreed remainder-only rule, closing the ledger at 43.8 h (3.78 h

over). A second, smaller breach was then authorized rather than incurred: with the cap already closed, the single highest-value cell still missing (one adapted self-critique baseline (CRITIC) at 7B, whose absence left RQ2 with no external comparison at all) was approved as an explicit exception, and it cost 3.8 h (stage `main_matrix_critic` in the ledger), taking the total to 47.6 h. That launch knowingly bypassed the per-launch budget guard, which would have refused it; the hours are recorded unchanged.

The cap was then *reopened* on 2026-08-30, deliberately and after the fact, to run two of the three cancelled adapted baselines: an adapted Chain-of-Verification loop and self-consistency, both at 7B on the same pool (stages `main_matrix_cove` and `main_matrix_sc`, with their smoke runs). They cost 4.0 h, taking the ledger to 51.6 h and the breach to 11.6 h. We state this plainly because the earlier version of this appendix said the cancellation stood and that no further arm was run: that sentence was written under the remainder-only rule and is no longer true. The reopening was a decision taken with the machine idle and the deadline approaching, not a discovery that the arms were cheaper than believed, and the ledger carries the original overrun and the reopened spend without netting them against each other.

The sizing pilot in full, moved here from §4 (Qwen2.5-7B-Instruct under 4-bit NF4 quantization (Dettmers et al., 2023), greedy, seed 0). *The measurement axis holds*: extractor and checkers ran end to end on real rationales (879 checkable claims: 36.4% contradicted, 282 consistent, 277 unverifiable), and CoT outscores zero-shot (55.5% vs. 53.0%). *Naive repair reads net-harmful at 7B*: re-prompting on every contradiction repairs 64.0% of questions yet drops -7.5 points below CoT; most contradictions are answer-irrelevant (189 incidental vs. 131 load-bearing), so counter-evidence shakes answers that were right for unrelated reasons; gating to load-bearing claims concentrates intervention (64.0%→34.5%) but stays -6.0 points below CoT. *Certification is free and abstention is sound*: certify-abstain answers 19.0% of questions (38/200) with zero contradicted claims on every answered one (0.0%), buying faithfulness rather than accuracy (47.4% on the certified subset, below CoT).

The 32B bf16 control's secondary figures, moved here from §4: the paired CI for naive repair is [-5.3, +0.4]; certificate coverage rises 10.9% $\rightarrow$ 12.4% at essentially unchanged certified-subset accuracy (-0.5 points).

Composition standardisation, and what is not registered about it. The contradicted-claim rate of a tier or an arm is the per-class refutation rate weighted by the claim classes that tier or arm chose to state. Direct standardisation holds the weights at a reference mix, $\sum_c w_c^{\text{ref}} r_c$, with w^{ref} the 7B CoT claim mix on the scale axis and each tier's own CoT mix on the arm axis, so only the per-class rates move it. Point estimates come from the integer per-class counts and are rounded once; the intervals are percentile cluster bootstraps that resample the 887 questions (claims are nested in questions, so the question is the resampling unit), 20,000 draws, one draw shared by the two arms or tiers being compared. The arm-axis correction was registered with the R3 review fix and has been printed for CRITIC since; the *scale*-axis correction was **not** pre-registered; it was added on 2026-09-03 after an internal review found the scale axis uncorrected while the arm axis was corrected, and it reverses the sign of the cross-scale comparison. Both quantities are printed wherever either is claimed (Table 2), the code path is the one the CRITIC paragraph already used, and `results/per_class_tiers.json` carries the weights, the counts and the bootstrap settings.

CRITIC's loop bookkeeping, moved here from §4: 2 episodes kept the answer they started from and 39 more stopped after one revision, for 1,731 revisions over 1,772 rounds; 4 verdicts did not parse and were all treated as *flawed*, never as *correct*.

A fourth tranche followed on 2026-08-31: the two reinstated arms at 14B (7.1 h in total, of which 0.72 h is a CRITIC@14B run withdrawn at 117 of 887 questions to release the machine for another project; its rows are kept and no value from it is read or reported; CRITIC@14B was then run in full on 2026-09-15 on a different machine, an RTX 4090 under the same vLLM 0.7.3 stack, registered as such before any row: accuracy 57.5%, -0.6 points against 14B CoT, 52 lost against 47 gained, $p = 0.688$, CI [-2.8, +1.6], the same sign as at 7B but inside the cross-launch floor measured above, so Table 1 marks it † and no sentence reads it as a direction; 9 of 1,334 rounds carried a prose-embedded *correct* that the line-anchored parser declined, and went down the unparsed-therefore-flawed path; on 2026-09-16 the same machine ran Chain-of-Verification and self-consistency at 32B-AWQ (the ninth tranche, below) against the A800 32B CoT rationale, byte-identical: CoVe

62.3%, −1.2 points (CI [-2.4, -0.2]), no-question degeneracy 0 of 887, pooled contradiction rate 41.9% against CoT's 39.0%, the same direction as at 7B and 14B and beyond the cross-launch floor; self-consistency 64.8%, +1.2 points (CI [-0.8, +3.3]), contradiction 40.1%, inside the floor on both axes). Their read-out was registered before any 14B row existed and both registered branches fired: Chain-of-Verification's accuracy cost vanished at 14B while its faithfulness cost persisted on the pooled rate, and self-consistency's null broke in the direction of accuracy, so the abstract, §4 and the conclusion scope each claim to the scale it holds at rather than carrying the 7B sentence with a caveat. The composition control added after the fact (below) then removed the faithfulness cost at 14B as well; the registered read-out is reported as it fired and the control is reported beside it. Chain-of-Verification produced no parseable verification question on 0 (0.0%) of 887 questions at 7B and 36 (4.1%) at 14B, under the 10% degeneracy threshold at which its delta would have carried a caveat; those questions keep their CoT answer unrevised. A fifth and a sixth tranche followed on 2026-09-11, one per registered model family (PREREG_FAMILY_B1B2.md, the cap reopened for the two tranches by the decision file of 2026-09-10): the five base arms at Llama-3.1-8B-Instruct (6.0 h: a 30-question smoke plus the full run) and at Granite-3.1-8B-Instruct (4.3 h, likewise), each on one RTX 4090 of a separate box. The two full runs were not written to that box's ledger at run time (the driver script recorded only the smokes), so the two entries were reconstructed at landing from the driver script's own status files and run logs, which agree to the minute; the ledger entries say so in their note fields. A seventh tranche, the two design-external settings of Appendix D, cost 17.4 h (stages `design_external_*`: 11.8 h for the five TSQA-pool tiers, 5.6 h for ChatTS-14B), booked at landing from each aggregate's wall-clock field. An eighth tranche ran CRITIC at 32B-AWQ and on Llama-3.1-8B, both registered on 2026-09-16 before any row (the 7B loop's protocol, seed 0, the A800 CoT rationale of each tier reused byte-identically): the runs began on the same RTX 4090 box as the 14B arms, were killed when that box was reclaimed by its owner, and resumed from their row counts on a third box (an RTX 4090 D under the same vLLM 0.7.3 stack, the pinned snapshots re-downloaded by sha, the checker libraries pinned to the first box's versions). The tranche books 17.6 h: 8.1 h for the two resumed runs on the third box, 4.6 h for their walls on the first box before the interruption, and 4.9 h for the full CRITIC@14B run of 2026-09-15 described with the fourth tranche, whose hours are booked here. At 32B-AWQ CRITIC scores 61.8% (−1.8 points against 32B CoT, 53 lost against 37 gained, $p = 0.113$, CI [-4.0, +0.4]): the 7B sign, inside the cross-launch floor, so Table 1 marks it †; its pooled contradiction rate rises by +4.6 points ([+3.2, +6.1]), beyond the box term with the 7B sign, and by +1.4 points standardised to the CoT class mix ([+0.4, +2.5]). On Llama-3.1-8B CRITIC scores 43.5% (+0.3 points, $p = 0.882$, CI [-2.8, +3.3]) and its pooled contradiction rate moves −0.3 points ([-1.6, +1.1]): both inside the floor, no direction read; paired directly against that family's ungated repair, the survivor of §4, CRITIC trails by −6.2 points ($p = 0.003$). The verdict audit of §A found no episode at either tier that early-stopped on a self-verdict the line-anchored parser rejects. Each of the two runs adds one headline test to the family (28 tests, $\alpha = 1.8 \times 10^{-3}$); the survivors are unchanged. The ninth tranche is the 32B-AWQ Chain-of-Verification and self-consistency run of 2026-09-16 above, 11.3 h (stage `main_matrix_ext32b_cove_sc`); the tenth is the fixed-extractor re-extraction of Appendix A, 2.85 h (stages `reextract_fixed*`). The ten tranches sum to the ledger's 118.2 h. Each tranche is disclosed on its own line, smoke included, never netted.

The per-scale repair-vs-CoT exact McNemar values, moved here from the RQ4 paragraph: naive repair $p = 0.095/0.068/0.094$ and gated repair $0.554/0.292/0.216$ at 7B/14B/32B. The 14B paired exact McNemar tests read $p = 0.210$ for Chain-of-Verification and $p = 0.051$ for self-consistency. Adding the four 7B and 14B tests took the multiple-comparison family from 13 to 17 and tightened the corrected threshold to $\alpha = 2.9 \times 10^{-3}$; the 4 adapted reruns and the 7 B1/B2 tests then took it to 28 ($\alpha = 1.8 \times 10^{-3}$, Table 10). The surviving set did not change at the first expansion: CRITIC at 7B remained the one test that cleared it. At the second, one test joined it (ungated repair at Llama-3.1-8B), and the next smallest p is 0.010. We record the tightening because it applies to the tests that were already in the family, not only to the new ones. Neither new arm is degenerate on this pool: every question yielded parseable verification questions for Chain-of-Verification and a parseable majority vote for self-consistency, so neither delta is read through a degeneracy caveat. Chain-of-Verification checked 4,548 final-answer claims and self-consistency 4,448, against CoT's 4,282.

The third cancelled baseline, program-of-thoughts, remains out for a different and stronger reason: the harness has no code-execution pathway, so the arm would need a sandboxed executor built from

scratch, and this pool's answers are natural language claims rather than the numeric program outputs the method is designed for. We record that as a scope decision, not as a casualty of the cap.

Self-consistency carries one deliberate break with the rest of the matrix. Every other arm, including CRITIC and Chain-of-Verification, starts from a single greedy CoT rationale, which is what makes their pre-intervention claim statistics identical by design. Self-consistency cannot: its 5 samples at temperature 0.7 are the method, and a shared rationale would delete it. Its comparison against CoT still pairs by question, and it is a member of the multiple-comparison family for that reason, but its delta is the effect of sampling consensus rather than of an intervention applied to a fixed answer, and no table or sentence reports it without saying so. It is also the only non-greedy decoding anywhere in this paper.

We report the overruns rather than restating the cap; the first lesson (a per-launch guard is not fleet-aware) is recorded in the runbook alongside the ledger.

Claim schema and checker thresholds. The six predicate families, their slot signatures, and every decision threshold are defined in code and referenced from §3; conservative v0 scope notes (change-point contradiction abstains on strongly seasonal windows; comparison claims are same-statistic two-window only; trend truth is window-scoped tendency) are documented alongside the thresholds. Certify-abstain answers when no claim is contradicted and at least one was checked, and a checked claim can be unverifiable: 23 of the 203 answered questions at 7B, 2 of 82 at 14B and 0 of 97 at 32B-AWQ carry no consistent claim.

B THE CROSS-LAUNCH FLOOR, MEASURED

Every sentence in this paper that reads a difference against "launch-to-launch noise" rested on indirect evidence: paired-CI half-widths, the AWQ-against-bf16 arm differences (§4) and the RQ3 repair-row relaunches. We registered a direct measurement before any relaunch row existed (`PREREG_NOISE_FLOOR.md` in the artifact) and report it under the registered predictions. The five 7B arms were launched twice more, unchanged: the same 887-question pool, seed, greedy decoding, prompts, checkers and model snapshot, the same vLLM and torch versions, each launch on one RTX 3090 (cards 0 and 1, in parallel). The published run was on an A800, so two spreads are defined and both are printed (Table 7): the *within-box* spread between the two 3090 launches, which is the floor the registration measures, and the *cross-box* spread between each 3090 launch and the A800 run, which carries a box term (a different GPU generation and a different CUDA build of the same torch) and is never pooled into the floor.

Table 7: The 7B five arms under three launches: accuracy on the 887-question pool, the within-box spread (launch 2 − launch 1, signed pp) with the number of questions whose correctness changes, and the cross-box spread (largest |launch − A800|, pp, rounded up) with the number of questions that change against the A800 run.

	accuracy (%)			within-box		cross-box	
arm	A800	3090 #1	3090 #2	pp	rows	pp	rows
zero-shot	51.5	51.7	51.7	+0.0	0	0.3	4
CoT	55.5	55.2	55.2	+0.0	0	0.3	90
naive repair	52.8	54.7	54.7	+0.0	0	2.0	153
certify-abstain	13.6	13.6	13.6	+0.0	0	0.0	94
gated repair	56.4	55.6	55.6	+0.0	0	0.8	133

Read-out. *N1 (within-box spread at most 3.0 pp per arm): holds.* The spread is 0.0 pp for every arm, and the two launches agree question by question (0 rows change correctness in any arm). On one box and one software stack, same-seed greedy decoding at the same batch configuration is a replay, not a second draw; the launch-to-launch floor is zero here and the paired half-widths the paper uses as its resolution are wider than it, so no sentence is re-read. *N2 (cross-box not smaller than within-box for a majority of arms): holds* (5 of 5). The net cross-box spread is at most 2.0 pp (naive repair), but under it up to 153 of 887 questions change correctness between boxes; the flips largely cancel in the aggregate. That is the quantity a replication on other hardware meets: a per-question comparison across boxes does not survive it, which is why every paired comparison in this paper is within-run, and an aggregate difference below 2.0 pp is within what a change of GPU alone moves. *N3 (the five*

arms keep their accuracy order across launches): holds, in both 3090 launches and in the A800 run. Cost: two launches of 3.9 and 4.1 wall-clock hours on one 3090 each; the published A800 numbers are unchanged.

C THE B1/B2 MODEL FAMILIES

What ran. Both families were registered before either produced a row (`PREREG_FAMILY_B1B2.md` in the artifact, including the amendment that replaced OLMo-2-1124-7B-Instruct (whose 4096-token context cannot run this protocol's 8192-token window) by Granite-3.1-8B-Instruct before any B2 row existed). Nothing in the harness changed except the model id: the same 887-question pool, seed 0, greedy decoding, vLLM 0.7.3, the same extractor prompt and the same six checkers, and the five base arms only; the adapted baselines were not run on these tiers. B1 is Llama-3.1-8B-Instruct (NousResearch mirror, snapshot `d10aef79`); B2 is Granite-3.1-8B-Instruct (`ibm-granite/granite-3.1-8b-instruct`, snapshot `4009206d`). Each ran on one RTX 4090 of a separate box. Table 8 carries every cell and §4 the body read-out.

The smoke gate. The gate was fixed before the full runs: the extractor's parse_ok rate on the chronocheck arm's 30 smoke rows decides the band the family lands in. Llama-3.1-8B parsed 30/30 (100.0%), at the clean band, and 863/887 (97.3%) on the full run. Granite-3.1-8B parsed 22/30 (73.3%), inside the registered middle band, so it is partially degenerate for the extractor; the full run parsed 587/887 (66.2%). Every Granite number in this paper is read with that caveat in the same sentence, as the registration requires. The same rate on the Qwen2.5 tiers is 850/887 at 7B and 853/887 at 14B.

The three registered read-outs, in both directions. (1) The gate value (the paired accuracy delta of gated against naive repair, exact McNemar with a 20k-resample bootstrap CI, a test the Qwen2.5 tiers do not carry, §5) is negative with its interval clear of zero at both families: -5.6 points at Llama-3.1-8B ([-8.8, -2.3], $p = 7.7 \times 10^{-4}$) and, with the degeneracy caveat, -3.0 points at Granite-3.1-8B ([-5.6, -0.5], $p = 0.019$). For Qwen2.5 the gate never read as a significant loss, and at Llama-3.1-8B it is the ungated arm that beats CoT (§4), so this quantity does not replicate. (2) The pre-repair contradiction rate replicates: 43.4% (2,811/6,481) at Llama-3.1-8B and, with the caveat, 44.3% (1,681/3,794) at Granite-3.1-8B, against 36.5–42.1% (31.5–36.5% standardised to the 7B claim mix) on the Qwen2.5 tiers, the same order of magnitude. (3) Repair shrinks claim volume as it does for Qwen2.5: -27.1% at Llama-3.1-8B and, with the caveat, -28.4% at Granite-3.1-8B on the ungated arm (Qwen2.5: $-46.0/-37.4/-32.2\%$), and $-33.3/-40.8\%$ on the gated arm. One of the three quantities fails at both families, so the headline keeps Qwen2.5 as its subject and §6 names the quantity. The gate's design premise, that most contradictions are incidental, holds at neither new family: load-bearing 1,639 vs. incidental 1,172 contradicted claims at Llama-3.1-8B and, with the caveat, 733 vs. 948 at Granite-3.1-8B. The certificate invariant holds at both families: zero contradicted claims on every answered subset (0.0% at Llama-3.1-8B and, with the caveat, 0.0% at Granite-3.1-8B), and the selection gap crosses zero at both (+11.4 points, [-3.4, +25.9]; with the caveat, -0.1 points, [-12.2, +12.0]), as the Qwen2.5 intervals do.

D DESIGN-EXTERNAL SETTINGS

What ran. Two settings outside the design of §4, each registered before it produced a row (`PREREG_TSQA_POOL.md` and `PREREG_CHATTS.md` in the artifact), under the same harness: seed 0, greedy decoding, the same extractor and six checkers, the five base arms only. (i) *The TSQA pool.* Time-MQA's TSQA file was excluded from the main pools because most of its windows fall below the protocol's series-length floor (§4); a 500-question pool drawn at the file's own window was run at the five main-matrix tiers, each on one RTX 4090 of a separate box. (ii) *A trained time-series backbone.* ChatTS-14B (Xie et al., 2025) is a Qwen2.5-14B fine-tuned with a native series encoder. vLLM 0.7.3 has no implementation of its architecture (the registration records the error verbatim), so it ran behind the same `pilot.py` interface on transformers at batch 1, the raw series passed through its own encoder rather than as text, on the main 887-question pool; both registered feasibility gates (seconds per question, extractor parse rate on a 30-question smoke) passed. Every cell of Table 9 is macro-bridged from the committed aggregates; the Granite row carries the

Table 8: The B1/B2 family tiers, every cell macro-bridged from the committed aggregates: end-task accuracy (%), claim-level faithfulness and the repair blocks on the 887-question pool. †Granite-3.1-8B is partially degenerate for the extractor (smoke 22/30, full run 66.2% parseable), so its cells are read with that caveat; Llama-3.1-8B is clean (smoke 30/30, full run 97.3%). "Contra. pre" is identical across the CoT, naive-repair and certify-abstain rows by the shared-rationale contract; the repair rows are row-matched on the rows that arm rewrote; best full-coverage row per column in **bold**.

Metric	Llama-3.1-8B	Granite-3.1-8B†
Zero-shot	48.3	51.9
CoT	43.2	51.7
+ naive repair	**49.7** (+6.5)	**54.3** (+2.6)
+ gated repair	44.1 (+0.9)	51.3 (−0.5)
+ certify-abstain (all)	2.7	3.6
acc. on answered	54.5	51.6
Paired p vs. CoT (ZS/naive/gated)	0.012/0.002/0.646	1.000/0.090/0.788
Claims checked (CoT)	6,481	3,794
Contra. pre, all 887 rows (%)	43.4	44.3
Naive repaired rows	819 (92.3%)	525 (59.2%)
Naive pre→post contra. (%)	45.0→39.9	47.9→38.6
Naive claim volume (%)	−27.1	−28.4
Gated repaired rows	577 (65.1%)	306 (34.5%)
Gated pre→post contra. (%)	48.5→38.9	49.0→38.6
Gated claim volume (%)	−33.3	−40.8
Gate value (CI, p)	−5.6 ([-8.8, -2.3], $p = 7.7 \times 10^{-4}$)	−3.0 ([-5.6, -0.5], $p = 0.019$)
Certify coverage, n (%)	44 (5.0)	62 (7.0)
Contra. on answered (%)	0.0	0.0

Table 9: Design-external settings: the TSQA pool at five tiers and ChatTS-14B on the main pool (500 questions per TSQA tier, 887 for ChatTS-14B). Repaired-row Δ is the change in contradiction rate on the rows chronocheck rewrote, paired 95% bootstrap CI; chronocheck − CoT is the paired accuracy delta. Granite carries the extractor-degeneracy caveat.

pool	tier	parsed %	CoT contra. %	abstain %	repaired-row Δ [CI]	chronocheck−CoT [CI]
TSQA	Qwen2.5-7B	98.6	31.6	83.0	−5.8 [-9.5, -2.0]	−3.8 [-9.3, +1.7]
TSQA	Qwen2.5-14B	99.6	22.7	82.6	−5.4 [-8.1, -2.8]	−2.0 [-8.9, +4.9]
TSQA	Qwen2.5-32B-AWQ	99.4	25.6	81.2	−12.7 [-15.2, -10.1]	+18.4 [+12.8, +24.2]
TSQA	Llama-3.1-8B	98.0	41.0	98.6	−10.6 [-13.8, -7.5]	−9.2 [-15.0, -3.4]
TSQA	Granite-3.1-8B†	70.0	46.3	95.8	−18.7 [-22.9, -14.5]	+0.0 [-6.2, +6.2]
main	ChatTS-14B	91.4	17.5	53.3	−5.2 [-8.8, -1.6]	−1.2 [-3.5, +1.1]

extractor-degeneracy caveat of Appendix C (70.0% parsed), and every row was produced on a box other than the main matrix's, so the cross-box term of Appendix B applies to any comparison against Table 1.

The registered read-outs. (1) The checker's effect on the rows it rewrites survives both settings: the repaired-row contradiction rate falls at all six tiers with the paired interval clear of zero, from −5.2 points on ChatTS-14B to −18.7 points on Granite-3.1-8B over TSQA. (2) Accuracy is not bought: chronocheck − CoT crosses zero at four of the six tiers; it is positive at 32B-AWQ over TSQA (+18.4 points, [+12.8, +24.2]) because that tier's CoT arm collapsed to 46.2%, not because the checker gained, and negative at Llama-3.1-8B over TSQA (−9.2 points, [-15.0, -3.4]), the same family whose gate value is the one loss in Appendix C. (3) Abstention at the certificate floor is family-dependent, 95.8–98.6% over TSQA and 53.3% on ChatTS-14B, so the certify-abstain arm's coverage is a property of how many claims a family states, not of the pool. (4) The trained backbone contradicts itself less than its base: ChatTS-14B's pooled CoT contradiction rate is 17.5% (577/3289) against 42.1% for Qwen2.5-14B on the same pool, a gap far beyond the box term; the checker still finds and removes contradictions on top of that training (read-out 1), which is the sense in which ChronoCheck sits on a trained model training-free. None of the six chronocheck-vs-CoT accuracy tests survives the paper's corrected threshold, as registered; nothing in Table 1 moves.

The multiple-comparison family grows. The family is every full-coverage arm-vs-CoT exact McNemar test the paper reports (Table 10): 12 base-arm tests on the four Qwen2.5 tiers, 6 on the two B1/B2 tiers, 9 adapted-baseline tests on Qwen2.5 and 1 (CRITIC) at Llama-3.1-8B. From the 17 tests of the first freeze it grew by 4 adapted reruns (CRITIC at 14B; CRITIC, Chain-of-Verification and self-consistency at 32B-AWQ) and 7 B1/B2 tests to 28, the corrected threshold moving from $\alpha = 2.9 \times 10^{-3}$ to 1.8×10^{-3}; 5 of the tests use rows produced on a second box (§A).

Table 10: The 28-test multiple-comparison family, one row per paired exact McNemar test against CoT, with the tier, the box that produced the rows, the record directory under `results/main/` (shipped as `archives/main.tar.gz`; extract it into `results/` as `supp/README_supp.md` shows), and the p-value. A dagger marks an arm whose rows come from a second box; `---` marks an arm the per-arm box ledger does not enter.

arm	tier	box	rows from	exact McNemar
Base arms, Qwen2.5 tiers (12)				
zero-shot	7B	A800	`7b`	$p = 0.010$
naive repair	7B	A800	`7b`	$p = 0.095$
gated repair	7B	A800	`7b`	$p = 0.554$
zero-shot	14B	A800	`14b`	$p = 0.534$
naive repair	14B	A800	`14b`	$p = 0.068$
gated repair	14B	A800	`14b`	$p = 0.292$
zero-shot	32B-AWQ	A800	`32b-awq`	$p = 0.233$
naive repair	32B-AWQ	A800	`32b-awq`	$p = 0.094$
gated repair	32B-AWQ	A800	`32b-awq`	$p = 0.216$
zero-shot	32B-bf16	A800	`32b-bf16`	$p = 0.072$
naive repair	32B-bf16	A800	`32b-bf16`	$p = 0.107$
gated repair	32B-bf16	A800	`32b-bf16`	$p = 0.400$
Base arms, B1/B2 family tiers (6)				
zero-shot	Llama-3.1-8B	fuxin	`llama8b`	$p = 0.012$
naive repair	Llama-3.1-8B	fuxin	`llama8b`	$p = 0.002$
gated repair	Llama-3.1-8B	fuxin	`llama8b`	$p = 0.646$
zero-shot	Granite-3.1-8B	fuxin	`granite8b`	$p = 1.000$
naive repair	Granite-3.1-8B	fuxin	`granite8b`	$p = 0.090$
gated repair	Granite-3.1-8B	fuxin	`granite8b`	$p = 0.788$
Adapted arms, Qwen2.5 tiers (9)				
CRITIC	7B	A800	`7b`	$p = 3.8 \times 10^{-4}$
CoVe	7B	A800	`7b`	$p = 0.018$
self-consistency	7B	A800	`7b`	$p = 0.666$
CRITIC	14B	fuxin	`14b_fuxin`†	$p = 0.688$
CoVe	14B	A800	`14b`	$p = 0.210$
self-consistency	14B	A800	`14b`	$p = 0.051$
CRITIC	32B-AWQ	—	`32b-awq_critic_fuxin`†	$p = 0.113$
CoVe	32B-AWQ	fuxin	`32b-awq_fuxin`†	$p = 0.035$
self-consistency	32B-AWQ	fuxin	`32b-awq_fuxin`†	$p = 0.272$
Adapted arms, B1/B2 family tiers (1)				
CRITIC	Llama-3.1-8B	—	`llama8b_critic_fuxin`†	$p = 0.882$
28 tests; 5 marked †; $17 + 4 + 7 = 28$				

The pre-expansion state is frozen in `BONFERRONI_BEFORE_FAMILIES.json`: its one survivor, CRITIC at 7B ($p = 3.8 \times 10^{-4}$), still clears the tightened threshold by a factor above five, and no test that survived before the expansion dropped out. One test joins it: ungated repair at Llama-3.1-8B ($p = 0.002$), upward, the first repair benefit to survive correction in this paper, reported here and in §4 at the prominence of the uniqueness claim it qualifies; the claim now holds of the Qwen2.5 tiers only. The next smallest p is 0.010.