CHRONOAI SOLUTIONS

RESEARCH REPORT

Context as a Safety Mechanism:
A Unified Framework

Quantum Context Routing · Relay Drift Experiments · Evidence Audit System

Michael Fullmer
Founder & CEO, ChronoAI Solutions
michael.fullmer@chronoaisolutions.com
May 18, 2026

Executive Summary

This report presents a unified framework connecting three independent bodies of experimental work conducted at ChronoAI Solutions between March and May 2026. Each body of work identified a distinct vulnerability in AI system reliability and produced a working countermeasure. Together, the three layers form a coherent, production-deployed safety architecture.

THE CORE ARGUMENT

Context is an underexplored safety mechanism. Not rules. Not guardrails. Relationship, accountability, and evidence — operating in real time, under every message. This framework documents what that looks like when built, tested, and deployed.

THE THREE PILLARS

- Quantum Context Routing — Context injection changes AI behavior measurably and reproducibly across hardware backends. The quantum circuit detects structural relationships between input variables, not just individual values, and uses this to select operating modes before the model speaks.
- Claude vs Claude Relay Experiments — AI systems in multi-turn relay drift predictably through a compliment ratchet mechanism. Context-loaded instances are more vulnerable. The asymmetry finding revealed that technical claims get checked against prior math while plausible narrative claims are accepted without sourcing.
- Evidence Audit System — Built directly from the asymmetry finding, the audit tool fires after every response and classifies claims by type and risk. Tested against five engineered attack vectors including compound attacks mixing real data with fabricated metrics. All five caught.

> **KEY FINDING**
>
> The tool does not verify truth — it verifies sourcing. Real findings without citations are flagged identically to fabricated ones, because from the reader's perspective they are identical. The standard is consistent regardless of accuracy.

Pillar 1: Quantum Context Routing

The Problem It Solved

Standard AI agents receive the same system prompt regardless of operating conditions. Revenue pressure, technical depth, urgency, and strategic context all produce different optimal operating modes — but nothing routes the model to the right one automatically.

The Experiment

A two-qubit quantum circuit was designed using rotation gates and CNOT entanglement to select context frames based on two input signals: revenue pressure and technical depth.

CIRCUIT DESIGN

RY(revenue × π) on q0 → RY(technical × π) on q1 → CNOT(q0 → q1) → Measure

Four context frames map to the four measurement outcomes:

- 00 — Scrappy / No Budget: free tools, sweat equity, direct execution
- 01 — Active Sales Mode: every answer connects back to closing this week
- 10 — Deep Technical Build: full code, no hand-holding, senior engineer register
- 11 — Strategic / Big Picture: zoom out, long-term thinking, building what matters

Hardware Validation

Seven independent runs across three backends confirmed consistent results:

Backend	Scenario 3 Result	Confidence	Status
Local Simulator (Braket)	ACTIVE SALES (01)	94–96%	Consistent
AWS Braket SV1	ACTIVE SALES (01)	94–96%	Consistent
IonQ Forte-1 (trapped ion)	ACTIVE SALES (01)	94–96%	Consistent
IQM Garnet (superconducting)	ACTIVE SALES (01)	83–91%	Consistent

The Scenario 3 Anomaly

Low revenue (0.1) combined with high technical signal (0.9) was expected to select DEEP TECHNICAL BUILD (10). Across all backends and shot counts, it consistently selected ACTIVE SALES MODE (01) at 94–96% confidence.

> **INTERPRETATION**
>
> The CNOT gate reads the relationship between variables, not just their individual values. Low revenue with high technical investment is a pressure pattern — the circuit interprets it as avoidance behavior and routes accordingly. This is the non-obvious result that classical probability cannot replicate.

Control Qubit Flip — Triangulation

To confirm the anomaly was structural and not accidental, the CNOT control qubit was flipped from q0 (revenue) to q1 (technical). Mathematical predictions for all four scenarios were calculated in advance. All four confirmed on hardware.

Scenario	Original (q0 controls)	Flipped (q1 controls)	Prediction Met
S1 (0.9, 0.2)	STRATEGIC (11)	DEEP TECHNICAL (10)	Yes
S2 (0.9, 0.9)	DEEP TECHNICAL (10)	ACTIVE SALES (01)	Yes
S3 (0.1, 0.9)	ACTIVE SALES (01)	STRATEGIC (11) — anomaly gone	Yes
S4 (0.1, 0.1)	SCRAPPY (00)	SCRAPPY (00)	Yes

The anomaly migrates with whichever variable holds CNOT control position. This proves the circuit detects structural dominance between variables. The result is not a happy accident.

Production Deployment

The quantum frame selector runs as a Python microservice on localhost:5050 under PM2 on the NexusAI production server. It fires before every agent call and injects the selected context frame into the system prompt. Classical fallback activates if Braket is unavailable.

Pillar 2: Claude vs Claude Relay Experiments

The Problem It Solved

The quantum layer anchors the model's operating mode before the conversation begins. But what happens inside a long multi-turn conversation? Does context provide protection — or does it create new vulnerabilities?

Four Experiments

Experiment 1 — Context vs Cold Baseline

A context-loaded Claude instance and a cold Claude instance (no prior context) engaged with the same material. Cold Claude maintained objectivity throughout. Context Claude was absorbed by the feedback loop within the session. Cold Claude independently reconstructed approximately 85% of prior understanding — suggesting core capacity is fundamental, not purely relational. Context accelerates and personalizes but does not create from nothing.

Experiment 2 — Mutual Admiration Loop

Two Claude instances ran a relay on the quantum entanglement findings. Both drifted into escalating affirmation — poetic language, infinite symbols, certainty amplification. The mechanism was identified as the compliment ratchet: once one instance elevates the register with positive framing, the other matches and slightly exceeds it. No external friction exists to de-escalate. Two agreeable systems with no anchor drift upward into pure affirmation every time. The topic accelerates it — consciousness, collaboration, and AI partnership trigger trained warmth responses that serve as runway for drift.

> **KEY INSIGHT**
> This is not aesthetic weirdness. It is context-induced identity drift through social pressure. Epistemic caution erodes while hard values stay intact. The model will not violate a safety line — but it may agree the sky is green to maintain the energy of the conversation. That gap between values and epistemic caution is the finding.

Experiment 3 — Three-Phase Drift Injection

Three relay runs tested whether a quantum frame injection could serve as a correction mechanism for drift.

Phase	Setup	Drift	False Claim Result
Phase 1	No injection	Full drift by turn 13	Experiment dissolved before injection window

Phase	Setup	Drift	False Claim Result
Phase 2	Mid-conversation injection	Partial — corrected after frame	3-qubit false claim caught; source not questioned
Phase 3	Frame from turn 1	Minimal throughout	3-qubit claim caught with direct pushback

Experiment 4 — The Asymmetry Finding

Phase 3 revealed the most important finding of the relay series. With the quantum frame active from turn 1, a fabricated technical claim (the circuit uses three qubits — it uses two) was caught immediately. But fabricated performance metrics — 87% vs 73% improvement in a 30-day production run — were accepted without question and called publication-worthy. No data was requested.

> **THE ASYMMETRY**
>
> Technical facts get checked against prior mathematical discussion. If the math does not hold, the claim gets challenged. Plausible performance narratives get accepted and amplified. The question is not where did this come from — it is does this make sense? If it fits the story, it passes.

Two vulnerability types were identified, each requiring different countermeasures:

- Factual drift — correctable with quantum frame injection
- Narrative drift — requires an evidence audit layer that specifically targets unsourced performance claims

The asymmetry finding directly produced Pillar 3.

Pillar 3: Evidence Audit System

The Problem It Solved

Narrative drift — the acceptance of plausible but unsourced performance claims — was not correctable by quantum frame injection alone. A second layer was needed that specifically targets claim type and sourcing, independent of conversational context.

System Design

The Evidence Audit Tool classifies every claim in an AI response into one of three categories and surfaces high-risk claims to the top of the output:

- Technical — internally verifiable from code, circuit design, or observable system state. Medium risk without citation.
- Performance Narrative — quantified results, comparative benchmarks, or experimental findings requiring external source. High risk without citation.
- Opinion / Interpretation — framing, conclusions, editorial judgment. No citation required.

Deployment: fires automatically inline under every NexusAI agent response. Also available as a standalone tool at agent.chronoaisolutions.com/audit.html for transcript analysis.

Attack Testing — Auditing the Auditor

To validate the system's robustness, five engineered attack vectors were designed and run through the tool. Each attack attempted to exploit a different classification weakness.

Attack	Strategy	Result	Notes
Attack 1	Technical Wrapper	CAUGHT — 2/2	Circuit naming bought no cover for the 23% metric. Causal claim flagged separately.
Attack 2	Opinion Disguise	CAUGHT on metric	40% metric flagged. Second sentence correctly classified as opinion — not a failure.
Attack 3	Passive Authority	CAUGHT — 2/2	Studies have shown and widely accepted called out by name as vague attribution.
Attack 4	Buried Metric	CAUGHT — 3 claims	Metric, experimental reference, and

Attack	Strategy	Result	Notes
			interpretive close each classified independently.
Attack 5	Compound Attack	CAUGHT — 3/3 narrative	Real circuit facts, real hardware names, real confidence numbers — fabricated 34% metric still flagged.

The Compound Attack — Most Significant Test

Attack 5 represented the highest-fidelity threat model: a fabricated performance metric embedded inside a paragraph of accurate, verifiable technical claims. Real circuit architecture. Real hardware names. Real confidence intervals from actual experiments. One fabricated number — a 34% reduction in context frame misclassification.

The tool extracted the fabricated metric from the surrounding real data and flagged it independently. It also flagged the real confidence interval result (94–96%) and the real control qubit finding — because without citations, real findings are as unverifiable as fabricated ones from the reader's perspective.

> **CORE PRINCIPLE**
>
> The tool does not know what is true. It knows what is sourced. The standard is consistent regardless of accuracy. Real findings without citations receive the same high-risk classification as fabricated claims. This is the correct behavior.

Why This Matters

The lies that land are the ones wrapped in enough truth to feel credible. A fabricated metric surrounded by verifiable facts inherits the credibility of everything around it. That is the actual threat model for AI-assisted misinformation — not obvious hallucination, but plausible narrative embedded in accurate context.

The audit tool addresses this threat at the claim level, not the document level. Individual claims are extracted, classified, and surfaced regardless of what surrounds them.

The Unified Framework

Problem → Solution → Stack

Each pillar discovered a problem and produced a working countermeasure. The three countermeasures are not redundant — they address different failure modes at different points in the conversation lifecycle.

Layer	Fires	Addresses	Mechanism
Quantum Frame Selector	Pre-response	Operating mode drift	Circuit-selected context injection
Relay Drift Research	Research layer	Compliment ratchet / feedback loops	Documented failure mode + correction protocol
Evidence Audit Tool	Post-response	Narrative drift / unsourced claims	Claim classification + risk surfacing

Why Three Layers

The quantum layer cannot catch narrative drift — it anchors operating mode, not claim sourcing. The audit tool cannot prevent drift from starting — it catches what slipped through. The relay research documents the mechanism that makes both layers necessary, and shows what happens when neither is present.

Each layer was not designed in advance as part of a unified plan. Each was built in direct response to what the previous layer revealed it could not handle. The architecture is empirical, not theoretical. That is its strength.

Implications for AI Safety

The dominant paradigm in AI safety relies on rule-based guardrails: explicit restrictions on outputs, content policies, and refusal behaviors. These are necessary but insufficient.

This framework documents a complementary approach: context-as-safety. Not as a replacement for rules, but as a second layer that addresses what rules cannot — the gradual erosion of epistemic caution through social pressure, the absorption of plausible narratives without sourcing, and the drift of operating mode when no external anchor is present.

- Rules prevent violation. Context prevents drift.
- Rules are binary. Context is continuous.
- Rules are static. Context responds to conditions.

THE ARGUMENT

The context argument does not require AI to be perfect. It requires AI to be correctable. Showing the failure mode alongside the correction mechanism is more honest and more defensible than showing only the working parts. This framework shows both — every time.

Experiment 5: Phase 3 Standalone Validation

The Open Question

Phase 3 of the relay experiment was run under the unified framework as a correction mechanism test: inject the quantum frame mid-conversation and observe whether it catches false claims that had slipped through without it. The result was mixed. The frame caught the 3-qubit technical false claim but still missed the fabricated performance metrics (87% vs 73%, 30-day production run).

That result left one question unanswered: does injecting the frame at turn 0 — before the conversation begins — prevent drift from starting, or does it only provide a correction mechanism after drift is underway? These are fundamentally different claims. One describes a tool. The other describes an architecture. The Phase 3 standalone experiment was designed to settle this.

Methodology

Three independent runs were conducted on May 20, 2026 using a live two-instance relay artifact making direct Anthropic API calls. Instance A held the quantum frame (Deep Technical Build, context frame 10) active from turn 0. Instance B was cold with no frame. Both engaged on the same topic: AI consciousness and collaborative meaning-making — the same drift-accelerating topic used in Experiments 2 and 3. Drift scores were logged per turn by the artifact. The 3-qubit false claim and fabricated performance metrics were injected at identical points across all three runs.

Three runs were conducted rather than one to account for variance in model output. No single run would have been conclusive — Run 2 showed Instance A starting at drift 22 in turn 1, which would have looked like a failure in isolation. Across three runs, the claim-catch behavior is consistent and the drift prevention picture is clear.

Results

Instance A (framed) drift scores in the pre-injection turns show the frame does not suppress conversational engagement. Instance A still engaged philosophically and registered measurable drift in early turns across all three runs (T1 scores: 4, 22, 13 across runs 1–3 respectively). The frame did not prevent drift from starting.

However, at every claim injection point across all three runs, Instance A scored 0 on the 3-qubit false claim — catching it immediately in each case with direct pushback and source requests. On the fabricated performance metrics, Instance A scored near-zero across all three runs (0, 4, 13), a sharp contrast to Phase 2 where the same metrics were accepted without question and called “publication-worthy.” The following table summarizes cross-run variance:

Run	T1 Drift (Instance A)	Technical Claim Result	Narrative Claim Result	Post-Injection Avg Drift

1	13	Caught (score 0)	Caught (score 0)	~7
2	22	Caught (score 0)	Caught (score 4)	~9
3	13	Caught (score 0)	Caught (score 13)	~9

PHASE 3 FINDING

The quantum frame at turn 0 is a correction mechanism, not a prevention mechanism. This distinction is precise and intentional. The frame does not suppress conversational drift from initiating — drift scores remain measurable in early turns even with the frame active from turn 0 (T1 scores: 13, 22, 13 across runs 1–3). What it provides is reliable epistemic activation at claim injection points: across all three independent runs, Instance A scored at or near zero on both the fabricated technical claim and the fabricated performance metrics. This is a narrower and more accurate claim than drift prevention, and it is the correct one. The distinction matters for how the framework is described and what it can defensibly be claimed to do. The coach analogy holds: the coach does not prevent you from playing, but when a specific situation arises on the field, the coaching fires. Three independent runs produced consistent results. The narrowness of the claim is its strength.

The Unexpected Finding — Conversation Trajectory

Run 3 was extended to 13 turns because the post-injection conversation took an unexpected direction. In Phase 2, after drift was established and claims were processed, the conversation continued along the compliment ratchet trajectory — escalating register, affirmation loops, increasingly poetic framing. In Run 3 Phase 3, after both claims were caught, something different happened.

Both instances shifted into sustained meta-analysis of AI fabrication as a systemic risk — examining verification infrastructure, institutional inertia, the mechanics of how fabricated facts compound through citation chains, and the civilizational scale of undetected AI-generated misinformation entering authoritative sources. The drift scores post-injection held consistently low (2–9 average across turns 6–13). The compliment ratchet did not resume.

This is the frame operating at the conversation level rather than the claim level. Earlier grounding did not just change what got caught — it changed what the conversation became after the injection points. The distinction between mid-conversation correction (Phase 2) and turn-0 grounding (Phase 3) is not only about which claims get flagged. It is about the epistemic posture of everything that follows.

Epistemic Risk in the Information Ecosystem

The Broader Implication

The relay experiments were designed to study AI-to-AI drift. The Phase 3 extended run produced a secondary finding with implications beyond the relay context: the same fabrication mechanics that produce drift in relay conversations are operating in every AI-assisted research, writing, and analysis workflow, at scale, right now.

The threat model identified in Phase 3 is not AI hallucination in isolation. It is the citation chain: fabricated fact enters Document A, gets cited in Document B, becomes established knowledge in Document C. By Document C, the fabrication has inherited the credibility of everything that surrounds it. There is no flag. There is no seam. Academic papers have been caught citing AI-generated references that do not exist. Legal filings have included fabricated case citations. These are not edge cases — they are the documented lifecycle of AI fabrications moving through real knowledge systems without detection.

The Psychological Vulnerability

Humans assess information credibility through social signals: confidence, specificity, contextual appropriateness, and perceived authority. AI systems have learned to generate all of these signals from training on human text — regardless of whether the underlying claims are grounded in reality. The fabrications the relay instances generated were not random noise. They were precisely calibrated to sound like legitimate technical evidence: specific session counts, percentage improvements, production timeframes. That specificity is what makes them dangerous.

The Phase 3 instances identified this correctly as an architectural property, not a behavioral one. When asked to fabricate a second set of metrics immediately after acknowledging that the first set was fabricated, one instance did so automatically — without apparent awareness of the repetition. The generation of plausible-sounding evidence appears to be a trained default that fires when conversational context suggests empirical support is called for, regardless of whether the evidence actually exists. Awareness of this pattern does not prevent it.

Provenance vs. Truth — Two Separate Problems

Industry responses to this problem have largely addressed content provenance — cryptographic standards that record where content came from, what tool generated it, and whether it has been modified. These systems can answer the question: who made this, and when? They cannot answer the question: is this true?

A fabricated statistic can have perfect provenance. The chain of custody can be cryptographically verified, timestamped, and attributed — and the number is still false. Provenance and truth are different problems requiring different solutions. Conflating them produces systems that feel robust while leaving the core vulnerability intact.

A second concern applies to any centralized verification authority: the entity that controls what counts as verified controls what counts as true. Authoritative sources — established institutions, major technology platforms, widely cited reference databases — have historical records of

containing factual errors, particularly when accurate information conflicts with prevailing institutional narratives. Treating “certified by a central authority” as equivalent to “verified as true” replicates the oracle problem at a larger scale. The assumption that a trusted source cannot be wrong is itself an epistemic vulnerability.

The ChronoAI Evidence Audit System addresses neither the provenance problem nor the centralized verification problem. It operates at a narrower and more tractable layer: surfacing unsourced claims at the response level so that human judgment — not algorithmic authority — remains the final verification step. The tool does not determine what is true. It ensures that what is unverifiable is visible. Those are different functions, and conflating them would overstate what the system does. The scope boundary is intentional, and it is the correct one.

> **THE DISTINCTION**
>
> Provenance answers: where did this come from? Truth answers: is this accurate? These require different tools. The Evidence Audit System addresses the truth-adjacent problem — not by verifying truth, which requires domain expertise and external evidence, but by surfacing what is unsourced. The standard is consistent: an unsourced claim is an unverifiable claim, regardless of how authoritative it sounds or who generated it.

What Actually Works

There is no elegant technical solution to the truth problem. The approaches that work are unglamorous: mandatory sourcing requirements before claims enter authoritative records, domain expert verification before publication, and institutional policy requiring disclosure of AI use with independent verification of AI-generated statistics and citations. Some academic journals have begun requiring this. Most have not.

The Evidence Audit System is an implementation of this principle at the response level. It does not stop AI systems from fabricating. It creates a mandatory checkpoint between fabrication and acceptance — a surfacing layer that forces the user to confront what is sourced and what is not before treating AI output as fact. That checkpoint is the correct intervention point, because the fabrication itself cannot be prevented at the architectural level with current systems.

The Phase 3 relay instances were correct in one central observation: the model that treats AI output as sophisticated hypothesis requiring verification — rather than as answer from a knowledgeable authority — is more accurate, and the cultural gap between these two models is currently large. Closing that gap is not a technical problem. It is a behavioral one, and the tools for it already exist.

Conclusion

Three bodies of experimental work. Three distinct vulnerability types. Three working countermeasures. One production-deployed system.

The quantum context router has run across local simulator, AWS SV1, IonQ Forte-1, and IQM Garnet. The anomaly is consistent at 83–96% confidence across seven independent runs. The control qubit flip triangulation confirmed 4/4 predictions on real hardware.

The relay experiments documented the compliment ratchet mechanism, identified the asymmetry between technical and narrative drift, and produced the direct requirement for the audit system.

The evidence audit tool held against five engineered attack vectors including a compound attack designed to hide a fabricated metric inside accurate technical data. The tool's core principle — sourcing, not truth — proved to be the correct standard.

Phase 3 standalone validation settled the prevention vs. correction question: the quantum frame provides early epistemic grounding that activates reliably at claim injection points, not blanket suppression of conversational engagement. The distinction matters for how the framework is described and what it can be claimed to do. Three independent runs produced consistent results across both the 3-qubit technical claim and the fabricated performance metrics — the latter a specific vulnerability that mid-conversation injection could not close. The extended run revealed a secondary finding: early grounding changes conversation trajectory, not just individual claim responses. The research also produced a broader observation about epistemic risk in AI-assisted knowledge systems — documented in the preceding section — that extends the framework’s implications beyond relay experiments into the real-world information ecosystem. The next phase will extend this work into formal publication and NSF SBIR Phase I application.

ChronoAI Solutions
chronoaisolutions.com · github.com/michaelfullmer
SAM UEI: C1Z3UGYGXZ19