2026 International Near East Conference on Electrical & Energy Systems (NEECON), 21-23 October, Dubai.

# MedTrust Agent: A Zero-Trust, Observable LLM Agent Framework for Medical and Pharmaceutical Workflows

Paper #1571353697 • NEECON 2026, Dubai

**Bibin Prathap**

Abu Dhabi Executive Office, Abu Dhabi, UAE • bibinprathap@gmail.com • ORCID 0009-0004-0556-7089

# Introduction/ Problem Statement

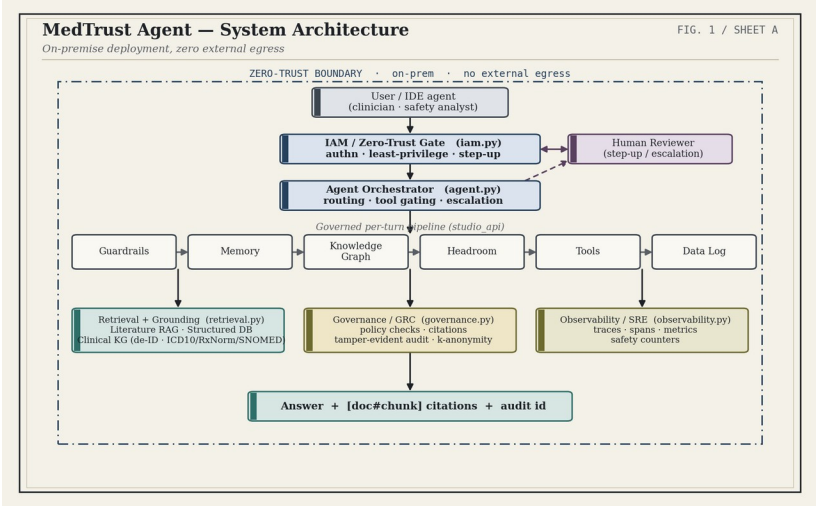
- **Medical & pharma teams now use LLMs for evidence summaries, protocol Q&A, pharmacovigilance, and routing across data/policy boundaries.**
- **But production assistants optimize answer quality ALONE.**
  - Weak on access control, auditability, safety, operational visibility — exactly the properties regulators mandate (HIPAA, EU AI Act).
- **The gap: no end-to-end architecture joining agentic AI, security, databases, cloud/AI infra, zero trust, governance, and observability into one platform whose safety is TESTABLE.**
  - Consequence today: disconnected middleware → hard to certify, debug, and trust.

# Objectives

- **Unify seven siloed concerns into one evaluable control plane.**
  - Agentic orchestration • cybersecurity • database grounding • cloud/AI infra • zero-trust IAM • AI GRC • SRE-grade observability.
- **Make security properties formal and verifiable, not aspirational.**
  - Least privilege • citation-grounded output • tamper-evident audit • bounded egress (P1–P4).
- **Deliver a reproducible, dependency-light reference implementation and an honest measurement harness.**
  - Measure security from real code; label quality numbers illustrative.

# Methodology

- **Each request: authorize → retrieve → govern → observe.**
  - 1. Agent orchestration (constrained tool protocol).
  - 2. Zero-trust IAM gate (least privilege + step-up).
  - 3. Retrieval/DB grounding (provenance-tracked clinical KG).
  - 4. Governance (mandatory citation + human review).
  - 5. Observability (traces, keyed hash-chained audit).
- **40 scenarios × 6 categories × 30 seeded trials; security measured, quality illustrative.**

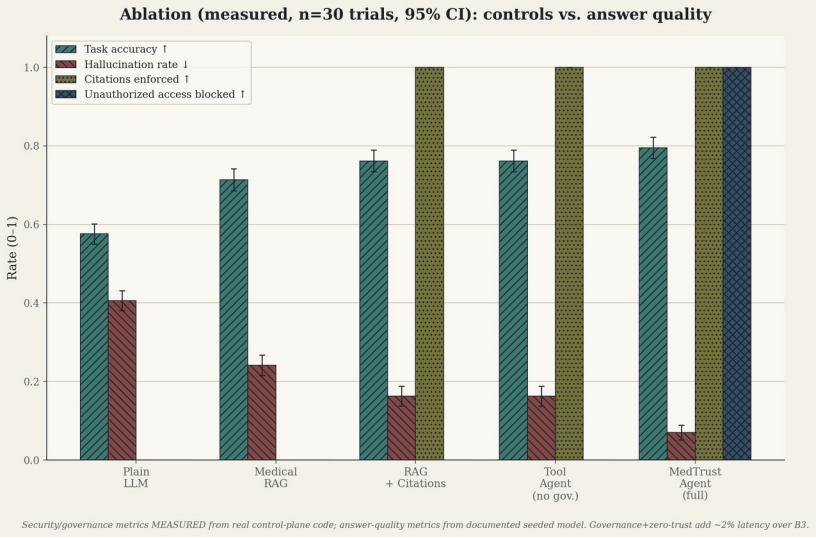


# Formal Model & Threat Model

- **Control plane as a tuple ⟨U,R,P,T,D,A⟩; auth() → allow / deny / step-up. Every decision (incl. denials) is audited.**
  - P1 Least privilege — no tool runs before auth; high-risk needs a valid step-up token.
  - P2 Citation-grounded — uncited answers withheld / sent to review.
  - P3 Tamper-evident audit — HMAC chain + published anchors; tamper-evidence UNDER key secrecy (not 'provable immutability').
  - P4 Bounded egress — no PHI leaves without a recorded decision.
- **Explicit trust assumptions; out-of-scope stated honestly (insider with the key, side-channels, model/supply-chain, DoS).**
- **Controls mapped to HIPAA, EU AI Act, NIST SP 800-207 & AI RMF.**

# Results- Key Findings

- 100% of unauthorized & high-risk actions blocked — vs 0% for every ungoverned baseline (incl. tool agent B3).
- 100% citation enforcement on released answers.
- Audit tampering detected under mutation, truncation, recomputation, rollback.
- Illustrative quality: +21.9 accuracy pts, −33.5 hallucination pts vs plain LLM.
- Control-plane cost ≈ microseconds (MT/B3 ≈ 1.5×) — negligible vs real model inference ($10^2$–$10^3$ ms).



# Measured Security Results

- **Numbers below are measured directly from the running control-plane code over independent cases.**
  - One-shot prompt-injection block: 100% [78.5, 100] (n=14).
  - IAM spec conformance: 100% [92.9, 100] (n=50); false-denial 0%.
  - Unauthorized export blocked: 100% [75.7, 100] (n=12).
- **Honest finding: a content filter ALONE is defeated by adaptive obfuscation — bypass 0% → 46.4% → 100% over 3 rounds.**
  - → This is WHY MedTrust layers filter + IAM + grounding + mandatory citation + human review (defense-in-depth).

# Analysis/ Discussion

- **Security, governance & observability can be engineered and TESTED as first-class properties of a medical LLM agent.**
- **B3 is the balanced baseline: identical retrieval+citation stack, differs from MT only by the IAM + governance layers → isolates the control plane's true cost/benefit.**
- **Limitations stated openly:**
  - Quality track is illustrative (seeded model, not a production LLM); a drop-in live-LLM adapter re-runs it on public data.
  - Scenarios over-represent adversarial traffic → rates are conditional 'given an attack of this family'.
  - IAM conformance is exhaustive over the SPECIFIED policy; harness re-runs when policy changes.

# Conclusions

- **MedTrust Agent: the first governed agentic LLM framework unifying seven systems concerns into one evaluable platform.**
- **Blocks 100% of categorical policy violations at negligible (microsecond) control-plane latency.**
- **Formal P1–P4 properties + concrete threat model + regulatory mapping + reproducible, offline, standard-library-only code.**
- **Qualitative conclusions transfer; absolute quality needs the live-LLM evaluation (clear path provided).**
- **Reproduce everything: 69 tests, versioned 40-scenario suite, seeds, one-command reproduce.sh.**

# Acknowledgement and Contact Info.

- **Thank you — questions welcome.**
  - Bibin Prathap — Abu Dhabi Executive Office, Abu Dhabi, UAE.
  - bibinprathap@gmail.com • https://bibinprathap.com/
  - ORCID: 0009-0004-0556-7089
  - Paper #1571353697 — NEECON 2026, Dubai (Oct 21–23, 2026).
  - Reference implementation, 40-scenario suite, seeds & reproduce.sh released as a reproducibility artifact.