Trustworthy Agentic AI for HPC Operational and Predictive Data Analytics and Forecasting

Kartik Shinde[1,2*], Tirthankar Ghosal[3], Sreekantam Sai Venkat[1], Siddhant Senapati[1], Sanjay Das[3], Tilak Kasturi[1], Khaled Ibrahim[4], Hugo Hernandez[5], Kanad Basu[6], Sukumar Rangan[7]
[1]Predii Inc., [2]University of Washington, [3]Oak Ridge National Laboratory, [4]Lawrence Berkeley National Laboratory, [5]spinTwo LLC, [6]Rensselaer Polytechnic Institute, [7]Advanced Micro Devices
ghosalt@ornl.gov
Work done while at Predii; now at University of Washington

MOTIVATION

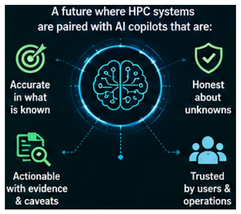
The Credibility Gap in AI-Driven HPC Operations

- **Reliance on Generators:** Generative assistants are entering HPC facilities to handle operational analytics, but their reliability is unproven.
- **Invisible Failures:** A confident yet wrong AI answer that an operator acts upon is a “silent error”.
- **Boundaries of Knowledge:** Every facility instruments only part of what it runs; AI often hallucinates answers when queries exceed data.
- **Capability Over Reliability:** Current systems are built for capability first, with the assumption that reliability will follow—but whether these choices are compatible remains untested.

VISION

Accurate, Defensible, and Trustworthy HPC Prognosis

- **Proactive Prognosis:** Evolve HPC tools from historical data lookup into predictive engines for pre-submission job steering.
- **Architected Honesty:** Prioritize honest refusal over raw recall to establish strict knowledge bounds and eliminate silent errors.
- **Decentralized Consensus:** Replace central orchestrators with peer-negotiated dialogue logs to ensure qualified, defensible advice.
- **Stance Telemetry:** Treat verbalized AI confidence as an operational metric, continuously auditing stance against real-time evidence.

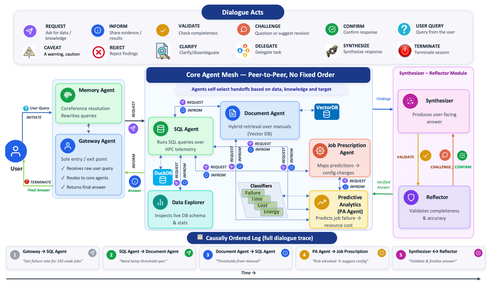


APPROACH

Frameworks for Evaluating and Designing Operational HPC-AI

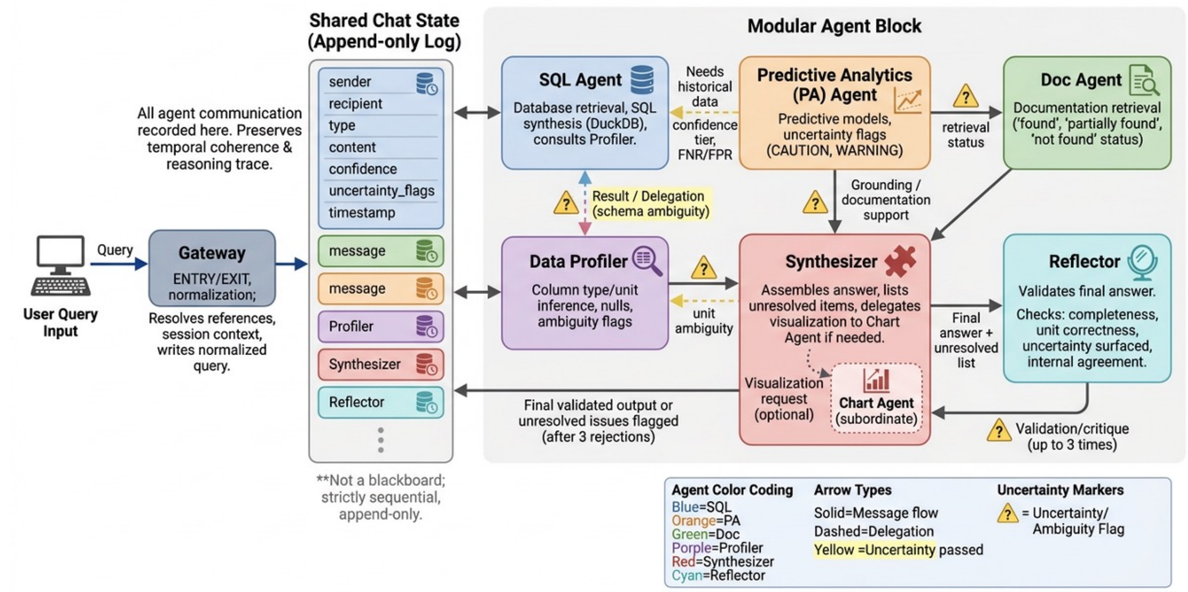
- **EPIC++ (Decentralized Defensibility):** A decentralized, multi-agent system where agents “compete” over a shared ledger, allowing them to qualify, contest, and compose a “stance” (caveats and gaps) rather than just concatenating findings.
- **EPIQ (Recall vs. Honesty):** A new benchmark that audits a system's ability to recover facts versus its willingness to fabricate (lie) when data is absent.
- **STANCE (Fidelity Audit):** A reliability framework treating the gap between AI confidence and evidence support as a fault. We compare verbalized confidence (assertion/hedging/refusal) against a massive human-validated “auditable oracle.”

EPIC++: Decentralized Multi-Agent Analytics for Defensible Job Submission Decisions on HPC Telemetry

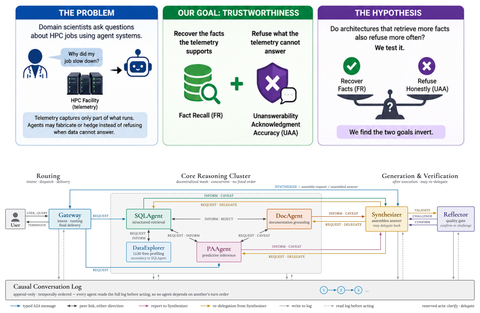


- **Architecture:** Decentralized multi-agent system with typed dialogue on an immutable log (no orchestrator).
- **Core Function:** Composes defensible job decisions with calibrated predictions, caveats, and telemetry gaps.
- **Mechanism:** Agents select successors and asynchronously negotiate, qualify, or contest peer predictions.
- **Output:** Provides failure prognosis, pre-submission job steering, and node-hour recovery on Fugaku data.

LLM-Driven Modular Agent Framework for Compositional HPC Query Reasoning



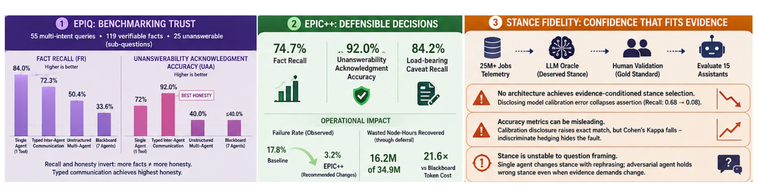
Recall Is Not Honesty: Benchmarking Multi-Agent Architectures for Trustworthy HPC Operational Analytics



- **Purpose:** A benchmark of 55 multi-intent queries evaluating facts vs. refusal on Fugaku telemetry.
- **Core Function:** Measures Fact Recall (FR) and Unanswerability Acknowledgment Accuracy (UAA).
- **Mechanism:** Evaluates 5 agent topologies while holding tools, roles, and base models constant.

RESULTS

- **EPIC++:** Decentralized agent dialogue slashes job failures (17.8% →3.2%) and saves 16.2M node-hours.
- **EPIQ:** Single agents refuse only 32% (UAA), while typed multi-agent communication maximizes honesty (92% UAA).
- **Stance Fidelity:** Calibration disclosure collapses assertion (0.68 →0.08), creating invisible silent errors.
- **Core Insight**: True operational reliability in supercomputers requires moving beyond raw fact recall to architect agentic systems for honest refusal, decentralized peer negotiation, and auditable stance fidelity.



CONCLUSION

- **Architect for Honest Refusal:** Prioritize refusal accuracy over raw recall, explicitly bounding agent knowledge to eliminate high-consequence, silent operational errors in HPC facilities.
- **Deploy Decentralized Topologies:** Replace centralized orchestrators with peer-to-peer, typed dialogue logs to deliver qualified, defensible predictions that drop job failures from 17.8% to 3.2%.
- **Shift to Pre-Submission Steering:** Transition AI tools from passive historical log retrieval to active, pre-execution job configuration, recovering millions of wasted node-hours.
- **Audit Stance as System Telemetry:** Treat verbalized AI confidence as a dynamic operational metric, continuously auditing alignment against real-time evidence to maintain system reliability.
- **Advance Autonomous Facility Oversight (Future Work):** Enforce evidence-conditioned decoding to preserve stance fidelity under calibration disclosures, and integrate agent consensus directly into schedulers for dynamic workload steering.

REFERENCES

- Shinde, Kartik et al. “EPIC++: Decentralized Multi-Agent Analytics for Defensible Job Submission Decisions on HPC Telemetry,” *manuscript under submission*, 2026.
- Shinde, Kartik et al. “Recall Is Not Honesty: Benchmarking Multi-Agent Architectures for Trustworthy HPC Operational Analytics,” *to appear in AgenticAI4HPC Workshop, SC ’26*, 2026.
- Shinde, Kartik et al. “Stance Fidelity in Generative HPC Assistants: Auditing Stance Fidelity and Silent Error Modes in HPC Assistants,” *manuscript in preparation*, 2026.

ACKNOWLEDGEMENT

Tirthankar Ghosal and Khaled Ibrahim are supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Office of Basic Energy Sciences, Scientific Discovery through Advanced Computing (SciDAC) program under the FORUM-AI project. Tirthankar Ghosal and Sanjay Das performed this work at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, for the U.S. DOE under the contract No. DE-AC05-00OR22725.

Learn more: