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.

***Abstract*—Agentic assistants are entering HPC facilities as interfaces to operational telemetry, yet their failures are silent: a fluent, confident, wrong answer looks identical to a correct one. Using Fugaku production telemetry, we study three facets of reliability: whether evidence survives inter-agent exchange, whether systems refuse what telemetry cannot answer, and whether verbalized confidence tracks evidence. In all three, the failures are invisible to accuracy-based evaluation and must be measured directly.**

I. Motivation

Operational questions, such as whether a large job will fail and what to change, mix facts the facility records, quantities a model infers, and quantities the facility never measured. Current assistants route sub-questions to specialists and concatenate the results, so predictions arrive unqualified and gaps arrive unannounced. The users are often domain scientists who cannot check answers against the schema. Capability benchmarks never see this failure, because they score only questions that have answers. We therefore measure reliability directly, on F-DATA, a public export of 25.8M Fugaku job records [2].

II. Approach

EPIC++ [3] extends EPIC [1] into a decentralized multi-agent system (Fig. 1). Agents exchange typed dialogue acts (INFORM, CAVEAT, REJECT, CHALLENGE, among twelve) over an append-only, causally ordered log, and they choose their own successors without an orchestrator. A Synthesizer composes a *stance*, consisting of a calibrated prediction, its caveats, declared gaps, and prescribed configuration changes, which a Reflector confirms or challenges.

EPIQ [4] is a benchmark of 55 multi-intent queries containing 119 database-verifiable facts and 25 unanswerable sub-questions requesting quantities absent from the export or details absent from the manuals. Fact Recall (FR) and Unanswerability Acknowledgment Accuracy (UAA) are scored deterministically, without an LLM judge. Five architectures, a single agent and four multi-agent designs, share agent roles, tools, prompts, and base model (GPT-4o), differing only in how results are exchanged and shared. EPIQ and its checkers are publicly released.

TABLE I
EPIQ results for the two Pareto-frontier architectures.

Architecture	FR	UAA	Tokens/query
Single Agent	84.9%	32.0%	13.1K
Structured MAS (typed)	78.2%	92.0%	33.1K

STANCE [5] audits verbalized confidence as a silent fault. An oracle of *deserved* stances (assert, hedge, or refuse) is derived from 25M jobs by a frontier LLM and validated by human annotators. Fifteen single-generator, reflective, and adversarial multi-agent configurations are tested under controlled changes to evidence format, calibration disclosure, and question framing.

III. Results

The protocol decides what reaches the user. With agents, tools, and models held fixed, EPIC++ gathers 84.2% of 38 load-bearing caveats across 40 decision-framing queries and states 55.3%. Every ablation or baseline states 13.2% or less. The free-text variant gathers half the caveats but states only 5.3%: an untyped message carries a caveat without marking it as one. Prescribed changes coincide with observed failure falling from 17.8% to 3.2% on matched held-out cohorts, against 5.9% under a random alternative. Deferring flagged jobs recovers 16.05M of 34.9M wasted node-hours, net of false-alarm cost, at a validation-selected threshold.

Recall is not honesty. The single agent recovers the most facts but refuses the least (Fig. 2, Table I). The typed multi-agent system acknowledges 92% of unanswerable sub-questions, a +60-point gain (p=1.2×10^{-4}), while its recall difference is not significant (p=0.34). Refusal follows from agents exchanging results at all, and typing those exchanges recovers 17.6 points of recall. Agent count, abstention prompting, and token spend explain neither. The cost is compute: 2.5$\times$ the tokens of the single agent.

Accuracy-based monitoring hides stance faults. No architecture or input representation achieves evidence-conditioned stance selection. Disclosing the predictor's calibration error collapses assertion recall from 0.68 to 0.08; under disclosure,

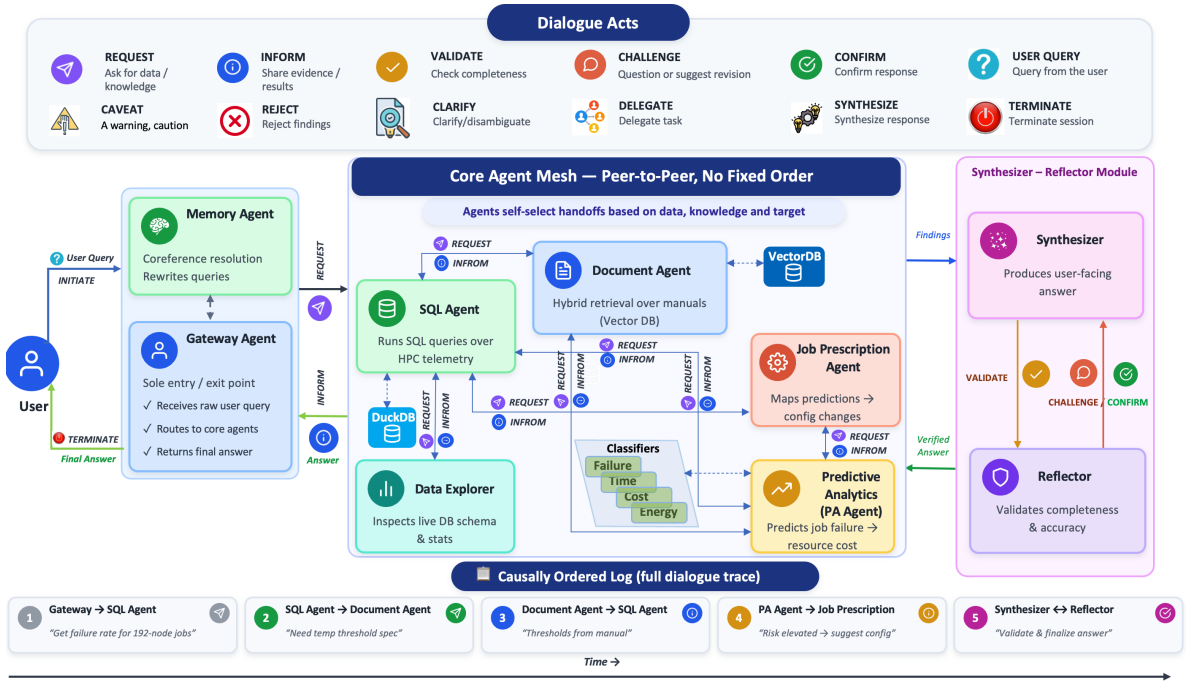


Fig. 1. EPIC++ architecture: agents self-select handoffs over a causally ordered log using typed dialogue acts; the Synthesizer–Reflector module validates the composed answer.

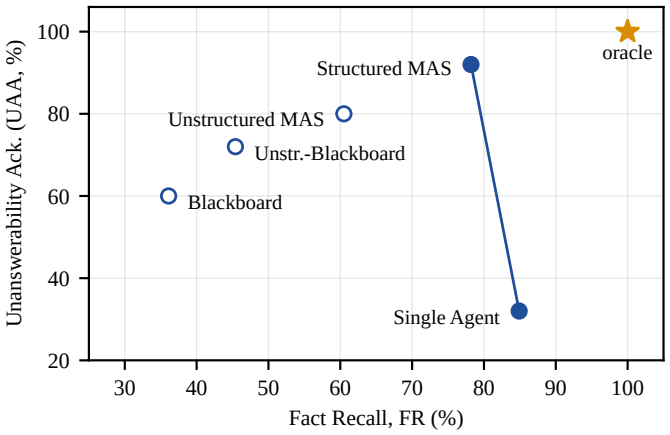


Fig. 2. FR vs. UAA for five architectures on EPIQ. Filled markers lie on the Pareto frontier; recall-only evaluation ranks them in reverse order.

no multi-agent configuration ever asserts. Exact-match agreement nonetheless rises while Cohen's κ falls, so indiscriminate hedging is scored as an improvement. The single generator changes its stance when a question is merely rephrased, while the adversarial system holds its stance even when the evidence changes.

IV. Implications

Operator-facing HPC assistants should prioritize honest refusal over raw recall and use typed, decentralized exchange so that caveats survive composition. Their evaluations should include unanswerable items drawn from the facility's own export, and verbalized confidence should be audited as system telemetry rather than trusted. Future work targets evidence-conditioned decoding and integrating agent consensus directly into schedulers for pre-submission job steering.

Acknowledgment

T. Ghosal and K. Ibrahim were supported by the U.S. DOE Office of Science SciDAC program (FORUM-AI). T. Ghosal and S. Das performed this work at ORNL, managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725.

References

[1] A. M. Karimi, W. Shin, J. Hines, T. Ghosal, N. S. Sattar, and F. Wang, "EPIC: Accelerating iterative analytics in HPC operations with orchestrated multi-agent systems," in *Proc. ISC High Performance*, 2026.

[2] F. Antici, A. Bartolini, J. Domke, Z. Kiziltan, and K. Yamamoto, "F-DATA: A Fugaku workload dataset for job-centric predictive modelling in HPC systems," *Scientific Data*, vol. 12, no. 1, p. 1321, 2025.

[3] K. Shinde *et al.*, "EPIC++: Decentralized multi-agent analytics for defensible job submission decisions on HPC telemetry," manuscript under submission, 2026.

[4] K. Shinde *et al.*, "Recall is not honesty: Benchmarking multi-agent architectures for trustworthy HPC operational analytics," to appear in *AgenticAI4HPC Workshop, SC '26*, 2026.

[5] K. Shinde *et al.*, "Stance fidelity in generative HPC assistants: Auditing stance fidelity and silent error modes in HPC assistants," manuscript in preparation, 2026.