SafeSurveil-AIxBio: Runtime-Gated Genomic AMR Triage with Evidence Graphs and Grounded AI Sidecars[1]

M. Rishyanth Reddy
Vignan's Lara Institute of Technology & Science, India
rishyanthreddy101@gmail.com

V. Eswar
Vignan's Lara Institute of Technology & Science, India
eswarvajja16@gmail.com

G. Akash Reddy
Glocal University, India
gade.akash3456@gmail.com

With
Apart Research

Abstract

Rapid genomic surveillance can shorten the path from antimicrobial-resistance (AMR) signal detection to analyst review, but AI-assisted workflows are difficult to trust when evidence provenance, model limits, and generated explanations are separated across tools. We present SafeSurveil-AIxBio, a defensive prototype for E. coli–tetracycline triage that couples live public-data retrieval and local AMR evidence generation with an auditable operator interface. The system ingests a live-retrieved FASTA, records source provenance and artifact hashes, normalizes AMRFinderPlus mechanisms, estimates novelty and baseline phenotype risk, and produces an actionability-constrained triage decision. Its main contribution is a runtime trust layer rather than a new clinical predictor: generated copilot and semantic-UI outputs are treated as sidecars and accepted only after identity, citation, numeric, queue-context, and policy checks against the persisted decision object. SafeSurveil-AIxBio also builds a deterministic biological reasoning trace, typed evidence graph, execution-gate verdict, and a Version 2 (V2) audit page. In the final proof run, a frontend-submitted live job completed with OpenRouter (LLM API gateway) live output, Thesys C1 (semantic UI framework) rendered output, `gate=allow`, full consistency checks, a 54-node/101-edge evidence graph, and a 26/0/0 pass/warn/fail curl matrix (automated API test suite). The prototype is not clinically validated; it demonstrates how genomic AI triage can be made inspectable, bounded, and submission-auditable.

[1]Research conducted at the AIxBio Hackathon, April 2026

1. Introduction

Antimicrobial resistance is a critical biosecurity problem because the signals that matter for routine stewardship (resistance genes, phenotype predictions, mobile elements, and surveillance provenance) dictate how quickly an analyst can identify a high-risk case. Whole-genome sequencing (WGS) has expanded AMR detection, but the literature repeatedly cautions that database choice, phenotype-genotype discordance, and standardization heavily alter interpretation[1, 2, 5]. Simultaneously, ML systems for infectious disease support remain difficult to deploy safely as they are often evaluated on predictive performance rather than workflow impact or interpretability[7, 8, 9].

SafeSurveil-AIxBio addresses a specific failure mode: AI models often make AMR triage look more certain than the underlying evidence supports. Our prototype treats generated language and UI as sidecars. The true authoritative objects are persisted decisions, evidence artifacts, and verifier outputs. This demonstrates a practical theory of change where a local analyst can rapidly inspect a genomic AMR case while clearly seeing why the system allows, reviews, or blocks downstream action.

Our main contributions are:

1. A live-backed AMR triage workflow connecting public-data retrieval, local AMR tooling, grounded copilot generation, and a renderer-safe frontend.
2. A visible execution gate that validates identity, citations, numeric consistency, and policy alignment before generated content is trusted.
3. A deterministic reasoning trace and evidence graph that convert the case into inspectable biological evidence rather than a black-box recommendation.
4. A reproducible audit bundle that makes our hackathon acceptance claims fully verifiable.

2. Related Work

Genomic AMR surveillance. WGS and metagenomics are vital for AMR detection, outbreak investigation, and One Health monitoring[1, 3, 4]. Reviews of AMR tools (AMRFinder, CARD, ResFinder) emphasize that resources differ in input assumptions, marker coverage, and sensitivity[2, 5]. This directly shaped our design: our UI highlights mechanistic evidence, raw artifact IDs, and caveats rather than presenting a single clinical answer.

ML for AMR and stewardship. Systematic reviews find promising performance for ML models predicting AMR from WGS and clinical data[8, 9, 10]. However, barriers include retrospective designs, nonstandard preprocessing, and limited prospective validation [7]. SafeSurveil-AIxBio does not claim broad predictive superiority. Its predictive baseline is deliberately disclosed as fixture-trained; our novelty lies in how predictions are constrained and audited.

LLMs and clinical safety. Recent reviews of LLMs in healthcare report potential utility but consistent concerns around hallucination, guideline mismatch, and patient-safety risks[13, 12, 11]. Retrieval-augmented generation (RAG) is a promising strategy, but healthcare RAG lacks standardized evaluation practices [14]. In response, SafeSurveil-AIxBio validates citations against built copilot context, rejects fabricated numbers, and refuses out-of-bound requests.

Explainability and evidence graphs. XAI literature argues that clinical decision support requires explanations meaningful to end users, not just post-hoc feature scores[15, 16]. Biomedical knowledge graphs are widely explored for integrating evidence, but clinical translation remains challenging[17, 18]. Recent work emphasizes structured reasoning traces from genomic context[21, 22]. SafeSurveil-AIxBio borrows this interpretability spirit, deterministically assembling a reasoning trace from persisted AMR evidence, novelty checks, and actionable triage.

3. Methods

System Pipeline and Scope

Our scope is intentionally focused on a defensive, local-first *E. coli* and tetracycline surveillance scenario. The system is a decision-support prototype for analyst review, not an autonomous diagnostic tool. The complete Version 2 (V2) architecture, detailed in Figure 1, visually enforces our core thesis by isolating the evidence factory from the presentation path.

Execution Gate and Grounding Checks

The verifier treats the persisted decision object as the absolute source of truth. It checks whether generated and rendered objects preserve job identity, target drug, numeric values, cited evidence IDs, and policy alignment. It records provider-call metadata so deterministic verifier routes do not accidentally trigger live LLM calls. The gate returns `allow`, `review`, or `block`, which dictates frontend display permissions.

Reasoning Trace and Evidence Graph

The reasoning trace is not a hallucination-prone chain-of-thought from an LLM. It is a deterministic sequence over known case components: sample context, AMR mechanism evidence, phenotype prediction, and QC limits. The evidence graph builds linked nodes for genes, mechanism classes, artifacts, novelty (calculated via Mash, a genome distance estimator), and citations. Off-target or weak AMR mechanisms are explicitly caveated rather than presented as supporting evidence.

4. Results

Implementation Outcome

Our build produced a fully integrated backend, frontend, and trust-layer bundle. Figure 2 visualizes the cumulative verification evidence across our Version 2 (V2) architecture release.

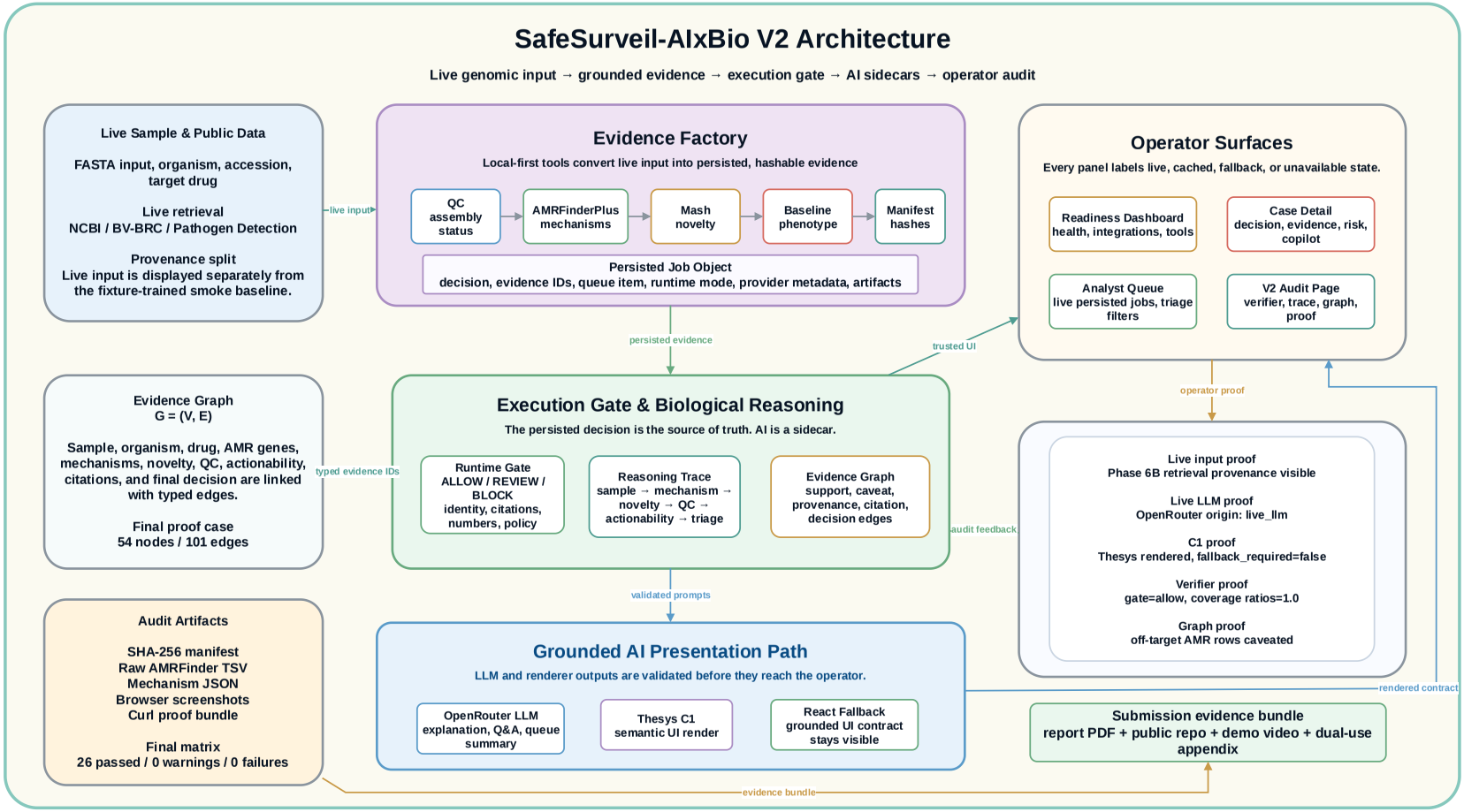

Figure 1: SafeSurveil-AIxBio V2 architecture. The flow visually enforces the system's core thesis: live genomic input is converted to persisted evidence and passed through a strict execution gate, treating the LLM presentation path strictly as a validated sidecar before reaching the operator.

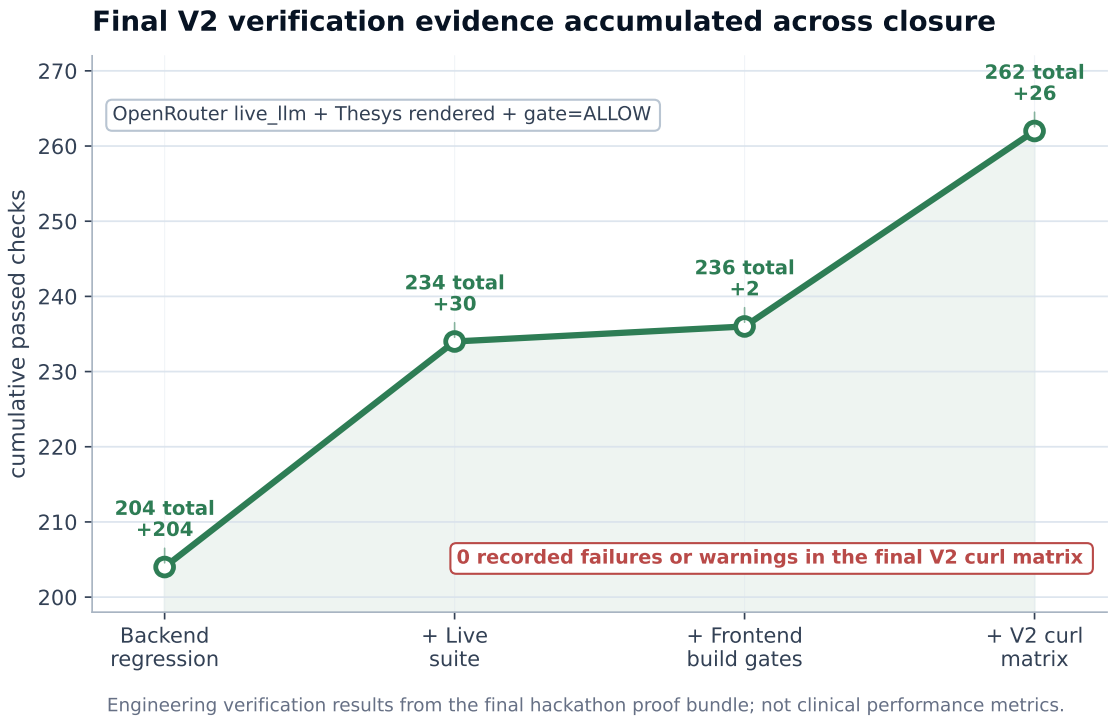

Figure 2: Cumulative verification evidence across the final V2 closure path. The line aggregates deterministic backend regression checks, explicit live integration checks, frontend build gates, and the final provider/proxy/verifier curl matrix. These are engineering verification results, not clinical validation metrics.

The final live proof utilized frontend job `job_20260425201118267288` processing an agricultural surveillance proxy sample. Runtime health reported live evidence mode, live LLM mode, and zero blockers.

Check	Observed Result	Interpretation
Backend suite	204 passed	Non-live contract and service tests successfully cleared.
Live suite	30 passed	Live integration tests passed separately from deterministic tests.
Frontend build	Passed	React/TypeScript production build succeeded.
V2 curl matrix	26 passed, 0 warn	Backend, provider, proxy, verifier, and audit checks passed perfectly.
Provider proof	`live_llm`, `rendered`	Generated sidecars exercised successfully via OpenRouter and Thesys C1.
Execution gate	`allow` (0 fails)	Case passed identity, numeric, citation, and policy checks.
Evidence graph	54 nodes, 101 edges	Case correctly represented as a fully inspectable graph.

Table 1: Final engineering verification results. These represent system integrity, not broad clinical validation.

Demo Workflow

The operator flow begins at runtime readiness. An analyst submits a live-backed job, inspects the case detail screen, reviews mechanistic evidence, and checks risk decomposition. When the analyst asks grounded copilot questions, the system renders semantic UI through Thesys C1. The core result is that every visual sidecar remains subordinate to artifacts. If a provider fails, fallback labels are visible; if grounding is weak, explicit warnings are surfaced.

5. Discussion and Limitations

SafeSurveil-AIxBio demonstrates that a biosecurity-facing AMR triage tool can be designed around visible constraints rather than invisible model authority. AMR surveillance and healthcare AI literatures point to the same adoption barrier: promising models are inadequate when users cannot see provenance, failure modes, or whether generated text is grounded[7, 15, 13]. The execution gate, reasoning trace, and evidence graph are not decorative; they form the safety interface.

Limitations

This prototype is not clinically validated. It relies on a fixture-trained smoke baseline for the first predictive layer and does not establish robustness across diverse pathogens. While live input provenance was demonstrated, large-scale public-data evaluation is reserved for future versions. Provider outputs remain bounded by validation and should never be treated as independent scientific evidence.

Future Work

A research-grade version should incorporate multi-sample public AMR snapshots, leakage-controlled train/test splits, and calibrated baselines. Introducing ablations for mechanism, novelty, and gate components alongside unsafe-action-rate metrics will be critical. Moving forward, broad external validation that honestly addresses dual-use boundaries is the primary goal.

6. Conclusion

SafeSurveil-AIxBio proves that AI-assisted genomic AMR triage can be built around evidence governance rather than model mystique. The system connects live AMR evidence, grounded generation, and a visible execution gate into a cohesive analyst workflow. Its most important achievement is not predicting resistance better than existing tools, but demonstrating how a biosecurity tool can make provenance, uncertainty, and generated-language boundaries visible by default.

Code and Data

- **Code repository:** https://github.com/RishyanthReddy/SafeSurveil-AIxBio
- **Data and artifacts:** Hosted in the same GitHub repository. The report utilizes public-data retrieval paths, generated AMRFinderPlus outputs, artifact manifests, and V2 proof summaries.

- **Demo video:** Watch the demonstration here

Author Contributions

M. Rishyanth Reddy led the project concept, implementation workflow, integration testing, frontend review, and report direction. V. Eswar and G. Akash Reddy contributed to project development, validation, and report preparation. Apart Research organized the AIxBio Hackathon and provided the event context.

References

1. Boolchandani, M. et al. (2019). Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics*. View on Consensus.
2. Hendriksen, R. et al. (2019). Using Genomics to Track Global Antimicrobial Resistance. *Frontiers in Public Health*. View on Consensus.
3. Oniciuc, E.-A. et al. (2018). The Present and Future of Whole Genome Sequencing and Whole Metagenome Sequencing for Surveillance of AMR. *Genes*. View on Consensus.
4. Collineau, L. et al. (2019). Integrating Whole-Genome Sequencing Data Into Quantitative Risk Assessment of Foodborne AMR. *Frontiers in Microbiology*. View on Consensus.
5. Mahfouz, N. et al. (2020). Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction. *Journal of Antimicrobial Chemotherapy*. View on Consensus.
6. Su, M. et al. (2018). Genome-Based Prediction of Bacterial Antibiotic Resistance. *Journal of Clinical Microbiology*. View on Consensus.
7. Peiffer-Smadja, N. et al. (2020). Machine learning for clinical decision support in infectious diseases. *Clinical Microbiology and Infection*. View on Consensus.
8. Ardila, C. M. et al. (2024). Integrating whole genome sequencing and machine learning for predicting antimicrobial resistance in critical pathogens. *PeerJ*. View on Consensus.
9. Ardila, C. et al. (2025). Machine learning for predicting antimicrobial resistance in critical and high-priority pathogens. *PLOS One*. View on Consensus.
10. Pinto-de-Sa, R. et al. (2024). Brave New World of Artificial Intelligence: Its Use in Antimicrobial Stewardship–A Systematic Review. *Antibiotics*. View on Consensus.
11. AlGain, S. et al. (2025). Can we rely on artificial intelligence to guide antimicrobial therapy? *Antimicrobial Stewardship & Healthcare Epidemiology*. View on Consensus.
12. Antonie, N.-I. et al. (2025). The Role of ChatGPT and AI Chatbots in Optimizing Antibiotic Therapy. *Antibiotics*. View on Consensus.
13. Park, Y.-J. et al. (2024). Assessing the research landscape and clinical utility of large language models. *BMC Medical Informatics and Decision Making*. View on Consensus.

14. Amugongo, L. M. et al. (2025). Retrieval augmented generation for large language models in healthcare. *PLOS Digital Health*. View on Consensus.

15. Antoniadi, A. et al. (2021). Current Challenges and Future Opportunities for XAI in ML-Based Clinical Decision Support Systems. *Applied Sciences*. View on Consensus.

16. Xu, Q. et al. (2023). Interpretability of Clinical Decision Support Systems Based on Artificial Intelligence. *Journal of Healthcare Engineering*. View on Consensus.

17. Budhdeo, S. et al. (2023). Scoping review of knowledge graph applications in biomedical and healthcare sciences. *Wellcome Open Research*. View on Consensus.

18. Rajabi, E. and Etminani, K. (2022). Knowledge-graph-based explainable AI: A systematic review. *Journal of Information Science*. View on Consensus.

19. Undheim, T. (2024). The whack-a-mole governance challenge for AI-enabled synthetic biology. *Frontiers in Bioengineering and Biotechnology*. View on Consensus.

20. Elgabry, M. et al. (2024). Cyber-biological convergence: a systematic review and future outlook. *Frontiers in Bioengineering and Biotechnology*. View on Consensus.

21. Fallahpour, A. et al. (2025). BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model. *arXiv*. View on arXiv.

22. Fallahpour, A. et al. (2026). BioReason-Pro: Advancing Protein Function Prediction with Multimodal Biological Reasoning. *BioRxiv / Arc Institute*. View source.

Appendix: Limitations and Dual-Use Considerations

False positives and false negatives. A gene call, similarity score, or phenotype probability can be wrong or incomplete. The prototype mitigates this by surfacing evidence IDs, support levels, QC warnings, and lab-confirmation next steps, but it does not eliminate biological uncertainty.

Scalability constraints. The current implementation is scoped to a narrow demo case. It has not been benchmarked across diverse organisms, drugs, sequencing technologies, or surveillance settings.

Dual-use risks. AMR surveillance tools can be misused if they help optimize harmful biological choices, identify resistance patterns for misuse, or make uncertain genomic signals appear operationally decisive. SafeSurveil-AIxBio mitigates this risk by limiting outputs to defensive triage, avoiding autonomous action, preserving uncertainty, and treating generated text as strictly non-authoritative.

Responsible disclosure. No new biological vulnerability or undisclosed pathogen finding was discovered during this work. If future versions identify actionable biosecurity concerns, disclosure must occur through institutional biosafety channels before public release.

Ethical posture. The system is built for defensive surveillance and safer decision support. It should not be used for clinical treatment, autonomous containment decisions, or high-consequence action without validated laboratory confirmation and expert oversight.

LLM Usage Statement

LLM assistance was used during software development, code review, literature orientation, and report drafting. All reported implementation results, test counts, provider-proof states, and safety claims were checked manually against repository artifacts and acceptance matrices. Generated language was heavily edited to preserve strict claim boundaries and to avoid overstating clinical validity.