BioWatch Brief[1]

Kevin Fickert
University of Pennsylvania

Jessica Anderson
University of Pennsylvania

David Nai
University of Pennsylvania

With

Apart Research

Abstract

Biosecurity analysts face a growing asymmetry: the volume of global outbreak reporting has expanded substantially over the past two decades, while the number of trained personnel capable of synthesising that information in real time has not. When a novel pathogen finding or unusual outbreak is reported, an analyst typically spends hours triaging WHO Disease Outbreak News, ProMED-mail alerts, and policy documents to construct an initial risk picture—a manual workflow that scales poorly under pressure.

We present **BioWatch Brief**, a rapid pathogen risk assessment tool that compresses this intake stage into minutes. The system implements a three-stage LLM pipeline—structured extraction, retrieval against a curated corpus of historical outbreaks and biosecurity policy frameworks, and grounded analysis—to produce a structured risk card from an arbitrary input report. The architecture is deliberate: by separating fact extraction from retrieval and synthesis, we constrain LLM outputs at each stage and surface uncertainty rather than mask it.

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

1. Introduction

Biosecurity practice operates under chronic information overload. Public-health practitioners and biosecurity analysts must monitor a continuous flow of outbreak reports, laboratory findings, and policy advisories drawn from sources including WHO Disease Outbreak News, ProMED-mail, national CDCs, peer-reviewed literature, and gray literature. When a novel signal appears—an unusual cluster, an unexpected geography, a new pathogen finding—the analyst must rapidly form a risk picture: what is this, how serious is it, what historical events does it resemble, what policy and regulatory frameworks apply, and what are the immediate next steps? In current practice, this triage takes hours of manual effort per signal, and the volume of incoming reports vastly exceeds available analytical capacity.

This gap matters for AI safety and biosecurity in two distinct ways. First, undetected or under-prioritised outbreak signals carry direct catastrophic risk: the cost of missing an early warning is asymmetrically high. Second, AI systems are now plausible candidates to assist with this work, but naïve deployment of large language models for high-stakes risk assessment introduces its own failure modes—hallucinated facts, over-confident classifications, and silent omission of uncertainty. A useful tool in this domain must be designed to mitigate these failure modes rather than amplify them.

From a technical standpoint, the core challenge is taking an unstructured biosecurity report and transforming it into actionable intelligence. To that end, we developed a Large Language Model (LLM) pipeline that takes raw report text as input and processes it in three distinct stages—extraction, retrieval, and analysis—to produce a structured risk assessment.

Our main contributions are:

1. A staged pipeline architecture for biosecurity risk assessment. By decomposing the task into structured extraction, retrieval, and analysis, we improve reliability, interpretability, and consistency relative to monolithic prompting approaches.
2. A category-separated keyword retrieval system that explicitly distinguishes historical outbreak analogues from policy and regulatory context, ensuring both classes of evidence are surfaced rather than allowing one to dominate.
3. A curated open corpus of 21 historical outbreak and biosecurity policy entries with normalised fields for pathogen, location, transmission, response history, and source provenance.

2. Related Work

Existing systems for pathogen surveillance and outbreak intelligence broadly fall into three categories.

Aggregation and surveillance platforms. ProMED-mail (Carrion & Madoff, 2017; Burki, 2023) and WHO Disease Outbreak News are the dominant manual-curation sources for global outbreak intelligence. HealthMap and the FAO's EMPRES-i platform provide automated aggregation of news and animal-health reports. These systems excel at capturing what is happening but provide minimal analytical synthesis: an analyst must still read, contextualise, and assess each item independently.

Commercial AI-driven epidemic intelligence. Platforms such as BlueDot and Metabiota apply machine learning to structured surveillance feeds, primarily for travel-risk and pandemic-forecast use cases. These systems are closed-source and oriented toward institutional clients; they are not accessible to individual public-health analysts and do not provide transparent reasoning over their outputs.

LLM-based information triage in adjacent domains. Recent work in clinical decision support and threat intelligence has demonstrated the use of large language models for structured extraction and retrieval-augmented generation over expert-curated corpora. To our knowledge, no published open system has applied this architecture specifically to the biosecurity practitioner intake workflow.

Our work occupies the gap between aggregation tools (which present raw reports) and commercial intelligence platforms (which are inaccessible and opaque). BioWatch Brief is open, single-analyst-scale, and explicitly designed for transparency: every retrieved fact is traceable to a corpus entry, and every LLM judgment is constrained by a published schema.

3. Methods

Both LLM stages use OpenAI's gpt-4.1-mini accessed via the Responses API. The corpus consists of 21 entries: 16 historical outbreaks and 5 policy or framework documents, curated from WHO Disease Outbreak News, US CDC sources, and peer-reviewed literature, normalised to a shared schema with fields for pathogen, location, transmission, response history, lessons learned, and source URLs. The frontend is a React single-page application; the backend is a FastAPI server exposing a single /analyze_report endpoint.

Our LLM Pipeline Architecture

The system was designed with three stages: (1) structured extraction, (2) context retrieval, and (3) analysis and synthesis.

The first stage, structured extraction, is focused on taking the input from the user (raw text from a report) and uses the LLM to break it down into data request from the data layer. The request includes fields such as pathogen name, location, transmission modes, and contextual tags (short keywords that capture important context that isn't captured by the other fields). The output from this stage is constrained to a predefined JSON structure. We utilize a safe parsing function to ensure it is valid.

In the second stage, we use a simple retrieval-augmented generation (RAG) method to identify relevant context from the curated corpus. We implemented this using a keyword-based scoring system. To that end, we developed two scoring functions:

- Outbreak scoring: prioritizes matching historical events based on pathogen, location, and transmission similarities.

- Policy scoring: prioritizes matching policy and framework corpus entries by looking at key words present in the documents.

This separation was an intentional design choice to ensure our intelligence report includes both policy guidance and historical events when relevant. To that end, each entry in the corpus is scored independently and ranked in its category. The system selects the top-k results for each category. In our case we used top three historical events and top 2 policy entries.

In the third and final stage, the retrieved context and structured extraction are passed to the LLM, which then generates the intelligence report/risk assessment. Like the structured extraction, the output is constrained to a predefined JSON structure to ensure the front end receives consistent output to reliably display the results.

4. Results

We evaluated the system using a curated scenario library of input reports spanning four conditions: clearly high-risk outbreaks with known historical analogues (e.g., a fictional Nipah-like cluster in South Asia), ambiguous early-stage reports with limited information, clearly low-risk findings, and one deliberately out-of-distribution case where no close corpus match exists. Test reports were author-written to control variables and to ensure ground-truth labels for risk classification.

Across these scenarios, the pipeline produced syntactically valid JSON output in every run, and the staged architecture surfaced both historical analogues and policy context for every query that had matching corpus entries. For a fictional Nipah-like cluster scenario, the system **correctly identified the Kerala 2018 outbreak as the top historical analogue** (score 7), produced a calibrated risk classification of 4/high, and surfaced three operationally relevant uncertainty flags including pending laboratory confirmation. This was emulated in other scenarios with clear historical analogues in the corpus. For the deliberately out-of-distribution scenario, the retrieval layer correctly returned its placeholder "No Historical Comparisons available" entry rather than producing a forced low-quality match, and the synthesis stage propagated this absence into an explicit uncertainty flag in the final output card.

A qualitative observation from the evaluation: the system's uncertainty flags—generated by the analysis stage when input data was incomplete—correlated with the cases where ground-truth assessment was itself ambiguous to the human authors. This suggests the LLM is detecting genuine ambiguity in the input rather than producing boilerplate hedging. We did not observe cases where the model invented historical events or fabricated source URLs not present in the retrieved context, which we attribute to the architectural separation between retrieval and synthesis.

We note significant limitations on the strength of these claims, addressed in Section 5. The scenario library is small, author-constructed, and was used during development; we have not run the system against held-out adversarial inputs or evaluated inter-rater reliability of the risk classifications. Results here should be read as evidence that the architecture functions as designed, not as evidence of practitioner-grade accuracy.

5. Discussion and Limitations

An observation from this project is that LLMs on their own do not consistently preserve structured inputs without explicit constraints. During development, we saw that the LLM would omit low-confidence or placeholder retrieval results unless explicitly instructed to include them. This shows that LLMs optimize for plausibility and coherence rather than strict fidelity to input data.

The broader implication for AI safety is that for critical systems, it is not sufficient to rely on LLM outputs alone. The system must include explicit control mechanisms that enforce consistency and correctness.

This is not unique to our pipeline; it reflects a general property of how transformer-based LLMs are trained to produce fluent continuations. The practical implication is that any safety-critical LLM deployment which depends on faithful preservation of structured inputs—clinical decision support, legal document review, intelligence analysis—should treat structural fidelity as a property to be enforced through architectural constraints (schemas, validators, retrieval boundaries), not a property to be requested through prompts.

Limitations

Our retrieval system utilizes keyword-based scoring, which is limited in capturing semantic relationships between the data queries and documents in the corpus. This means relevance may be underestimated or overestimated in some cases. In cases where similar concepts are expressed using different terminology, some entries in the corpus may be scored lower than they should be.

Scoring thresholds for relevance and other rules of thumb were hardcoded which may not be appropriate for other report types.

Future Work

The natural next step would be to change the retrieval method from keyword-scoring based to embedding-based to better capture semantic matching. Additionally, the corpus could be expanded.

6. Conclusion

We present BioWatch Brief, a tool designed to compress the intake stage of biosecurity outbreak triage from hours to minutes. The system's architectural commitments—staged decomposition, explicit retrieval over a curated corpus, and structured constraints on LLM outputs—are not merely engineering conveniences; they reflect a position about how LLMs should be deployed in safety-critical settings. Coherent prose is cheap; reliable behaviour under partial information is not, and we found it must be designed in rather than prompted for.

The most natural extension of this work is a richer retrieval layer—dense embeddings over an expanded corpus, integration with live RSS feeds from CDC and WHO surveillance bodies—and a structured feedback loop that allows domain experts to correct individual outputs over time. We are also interested in studying how practitioners actually use the tool: a fast assessment is only valuable if it is trustworthy enough to act on, and that trust must be earned through calibration evidence at a scale we have not yet generated.

Code and Data

Include links if applicable.

- ***Code repository**: https://github.com/jessanderson222/aparthackathon-aixbio/*
- ***Data/Datasets**: [Small JSON dataset included with submission.]*

Author Contributions

K.F. led the project and contributed the data layer and most of the main logic. J.A. implemented the frontend. D.N. researched the corpus, served as pathology/infectious disease expertise and wrote most of the manuscript, which all authors finally reviewed.

References

AMA citation format.

1. Carrion M, Madoff LC. ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. *Int Health*. 2017;9(3):177-183. doi:10.1093/inthealth/ihx014
2. Burki T. "Unfettered flow": how ProMED-mail keeps the world alert. *Lancet*. 2023;401(10373):259-260. doi:10.1016/S0140-6736(23)00167-8

Appendix A

Input for a fictional Nipah virus-like outbreak in Kerala, India:


```
{
  "report_text": "Health authorities in Kerala, India have reported a cluster of 4 suspected encephalitis cases among healthcare workers at a hospital in Kozhikode district. Initial laboratory testing suggests Nipah virus. Three of the four affected workers had contact with a recently deceased patient who presented with similar symptoms before death. The patient had reportedly consumed fruit from trees frequented by fruit bats. No travel history outside the district has been identified for any case. The state health ministry has activated outbreak response protocols and notified WHO."
}
```


Response from OpenAI (after LLM round-trips x2 to OpenAI (extraction, then synthesis) plus the corpus scoring in between).


```
{
  "brief_title": "Suspected Nipah Virus Cluster Among Healthcare Workers in Kerala, India",
  "summary": "A cluster of 4 suspected Nipah virus encephalitis cases has been detected among healthcare workers in Kozhikode district, Kerala, linked to a deceased patient with probable zoonotic exposure. Human-to-human transmission is suspected given healthcare worker involvement. Outbreak protocols have been activated and WHO notified.",
  "risk_level": 4,
  "risk_label": "high",
  "risk_rationale": [
    "Nipah virus is associated with high fatality and has known nosocomial transmission risks.",
    "Cluster involves healthcare workers, indicating potential human-to-human spread.",
    "Zoonotic exposure identified via contact with fruit bats, consistent with prior outbreaks in Kerala.",
    "No cases reported outside the district, limiting geographic spread currently.",
```

```
    "State health authorities are responding promptly with outbreak protocols
and WHO notification."
  ],
  "historical_comparisons": [
    {
      "title": "Nipah virus",
      "summary": "Healthcare workers and community members in Kerala affected
by Nipah virus with very high fatality rate, linked to fruit bat reservoir
and hospital transmission.",
      "relevance": "Matched based on pathogen, location, transmission modes,
and tags.",
      "source": [
        "https://www.who.int/southeastasia/outbreaks-and-
emergencies/surveillance-and-alert/surveillance-and-risk-assessment/nipah-
virus-outbreak-in-kerala"
      ],
      "score": 7
    }
  ],
  "policy_context": [
    {
      "title": "No Policy Guidance available.",
      "summary": "No relevant policy guidance was found that matches the
request.",
      "relevance": "No sufficiently relevant policy guidance was found.",
      "source": [],
      "score": 0
    }
  ],
  "live_signal_assessment": "Current evidence signals early cluster with
nosocomial transmission potential; human-to-human spread likely but no large-
scale community spread yet observed.",
  "recommended_next_steps": [
    "Enhance infection prevention and control (IPC) measures in healthcare
settings.",
    "Conduct thorough contact tracing and active surveillance in Kozhikode
district.",
    "Implement community awareness on avoiding fruit bat exposure.",
    "Perform confirmatory laboratory testing and genomic analysis to confirm
Nipah virus and track spread.",
    "Maintain rapid notification to WHO and coordinate with national health
authorities."
  ],
  "uncertainty_flags": [
    "Initial laboratory testing suggests Nipah virus but confirmation
pending.",
    "Limited case number (4 cases) restricts understanding of full outbreak
scope.",
```

```
    "No data on secondary transmission outside healthcare workers currently
available."
  ],
  "sources_used": [
    "Original report provided",
    "Structured extraction provided",
    "WHO Nipah virus outbreak Kerala historical data:
https://www.who.int/southeastasia/outbreaks-and-emergencies/surveillance-and-
alert/surveillance-and-risk-assessment/nipah-virus-outbreak-in-kerala"
  ]
}
```


LLM Usage Statement

LLMs were used as core components of the system for the structured extraction and final analysis. Anthropic Claude's assistance was also used to brainstorm approaches and draft portions of the implementation. All system behavior and outputs were manually reviewed and validated by the authors.