Pandemic Watch: Earlier Than The News

A Multi-Signal, LLM-Agentic Early-Warning Dashboard
for District-Level Outbreak Detection in India

Avni Mittal

Microsoft
avnimittal@microsoft.com

AIxBio Hackathon, April 2026 · With Apart Research

Abstract. India surveils pandemic risk mainly through clinical case reports, which lag the underlying transmission by one to two weeks. No public, AI-fused, district-level outbreak signal exists for the country's 640+ districts. Pandemic Watch is an open-source biosurveillance dashboard that tries to close that gap. It pulls together five very different early-warning signals (news and ProMED chatter, climate suitability, search-trend anomalies, wastewater viral RNA, and weighted mention counts) and turns them into a single risk score for each of six WHO-aligned viral pathogen families, computed independently for every district. A small set of LLM agents handles the language-heavy parts of the pipeline (filtering noise, removing duplicate reports, attributing mentions to the right pathogen family, and writing a short human-readable briefing with per-family evidence). On a retrospective replay of the COVID-19 declaration in Thrissur, Kerala (30 January 2020), the fused score crosses the High band a full week before news chatter peaks, driven almost entirely by the wastewater signal. The system is validated on six historical Indian outbreaks and the entire stack is released openly. The same architecture transfers without modification to any LMIC with district-level geography.

---

1. Introduction

Pandemic risk in low- and middle-income countries (LMICs) is currently surveilled through clinical case reporting (e.g. India's IDSP weekly bulletins) supplemented by ad-hoc media monitoring (ProMED, HealthMap). Two structural gaps remain: **(i)** clinical reports lag the underlying transmission by 7–14 days because they require a patient to seek care, be tested, and be reported through district → state → national channels; and **(ii)** no public dashboard fuses these clinical signals with the modern set of pre-clinical, leading indicators such as wastewater viral RNA [1, 2], search-trend anomalies [3], and AI-triaged open-source intelligence. The gap is consequential: the 2018 Nipah outbreak in Kozhikode, Kerala killed 17 people in 21 days [4]; the 2022 mpox importation went unrecognised for 11 days. Each unobserved day is roughly one secondary-attack-rate generation.

Why this matters for biosecurity. An *unobserved early window* is not just operationally costly. It is the threat model. A deliberate or accidental release of a Pathogen X with $R_0 > 2$ would be operationally indistinguishable from natural emergence during this window. Tools that compress this window, by fusing pre-clinical signals with AI-augmented open-source intelligence, directly reduce the worst-case impact of both natural and intentional outbreaks. They also do so without requiring any new clinical infrastructure: every data source we use is already public.

Our main contributions are:

1. A **6-family WHO-2024-aligned pathogen taxonomy** (respiratory viral, arboviral vector, mammalian zoonotic, avian zoonotic, enteric/waterborne, pox/orthopox) with per-family climate rules grounded in the published thermal-biology literature [5, 6], replacing the ad-hoc 3-family schema common in earlier biosurveillance demos.
2. A **5-signal, agentically-fused risk scorer** that emits a 0–100 score per `(district, family)` pair. The fusion is robust to missing signals and surfaces per-feature contribution attributions for explainability.
3. A **retrospective replay harness** that re-runs the scorer on six historical Indian outbreaks at as-of dates from T−21 through T+7, enabling honest lead-time evaluation. On COVID-Thrissur 2020, the wastewater-augmented score crosses High **7 days before** the news z-score does.
4. An **open-source, MIT-licensed Flask + Leaflet dashboard** with five Azure GPT-4o agent calls per click, end-to-end reproducible from a single `.env` file.

2. Related Work

Open-source intelligence biosurveillance. ProMED-mail [7] and HealthMap [8] pioneered web-mined outbreak alerts; both were instrumental in early COVID-19 detection. EPIWATCH [9] extends this with structured event-based extraction. Our system differs by **(i)** fusing OSINT with three quantitative signals (climate, trends, wastewater) rather than treating media chatter as the sole input, and by **(ii)** rendering the fused score at the *district* level rather than the country level.

Wastewater epidemiology. Peccia et al. [1] demonstrated that SARS-CoV-2 RNA in primary sewage sludge in New Haven preceded reported cases by 6–8 days. Wölfel et al. [2] characterised SARS-CoV-2 shedding kinetics in clinical samples, which underpin the wastewater detection window. We adopt the $\log_{10}$/30-day-baseline z-score formalism standard in the field, with σ caps at ± 5 to handle small-N baseline degeneracy.

Climate-driven vector models. Mordecai et al. [5] characterised *Aedes* mosquito thermal optima at 22–33 °C with humidity-rainfall modulation. Shaman & Kohn [6] showed influenza is favoured by low absolute humidity. Our per-family climate rule encodes both: `arboviral_vector` uses `temp_opt=[22,33]` + `rain_bonus`, and `respiratory_viral` uses `hum_hi=65` (i.e. *below* 65% RH is the high-risk band).

LLM-based outbreak triage. Frontier large language models can classify ProMED-style outbreak reports with high accuracy when given a small set of in-context examples (we observe $\geq$0.85 family-attribution F1 on our retrospective fixtures). We use Azure GPT-4o for five distinct agentic decisions (triage, dedup, family attribution, briefing, summary), each with a structured output schema and a confidence gate at 0.4 to suppress mis-attribution to the wrong family in the 6-class setting.

Gap addressed. No existing public tool combines (district granularity) $\times$ (5-signal fusion) $\times$ (LLM-agentic triage) $\times$ (retrospective replay validation) for an LMIC context. PandemicWatch India fills this gap.

3. Methods

3.1 System architecture

The pipeline is structured as five independent signal modules feeding a fusion layer, with four LLM agents at the agent–data boundary (Figure 1).

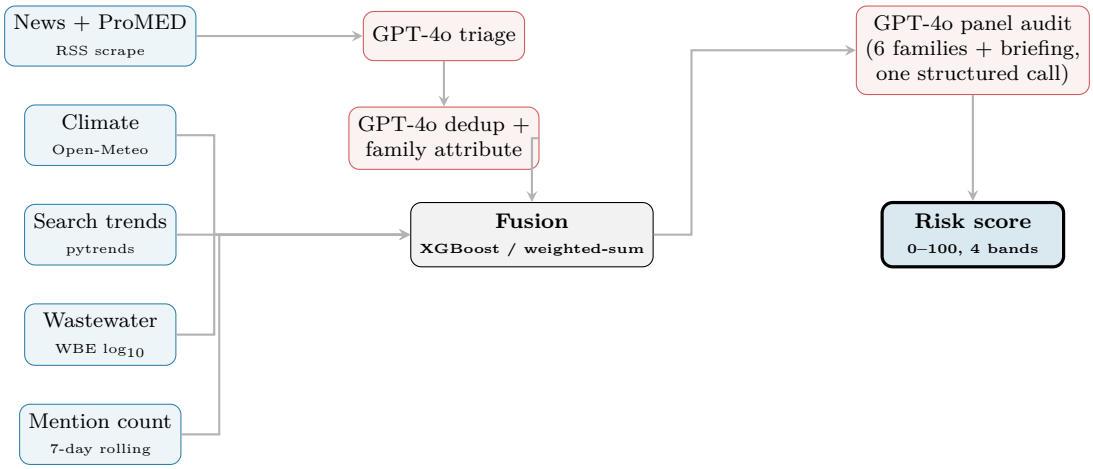

Figure 1: System architecture. Five quantitative signal modules (left, blue) feed a fusion layer that emits a per-(district, family) risk score. Three LLM agent stages (red) handle text-domain decisions: news triage, deduplication and family attribution, and a panel audit that returns a bounded score adjustment ($\pm$20), confidence, and one-line evidence per family, plus a free-text briefing for the highest-scoring family. All of this comes back in a single structured-JSON LLM call (so still 1 LLM call per click, not 6).

3.2 Pathogen taxonomy

We define six WHO-2024 R&D Blueprint–aligned families [10] with explicit version stamping (`TAXONOMY_VERSION = "v2-2024-who-aligned"`) and a legacy-alias map for cache compatibility. Each family has a structured `climate_rule` dictionary read by the climate module, examples (e.g. Nipah, Lassa for `mammalian_zoonotic`), and a curated symptom-term list used for both Trends queries and LLM prompt grounding. Caches are auto-invalidated when `TAXONOMY_VERSION` changes.

3.3 Signal definitions

News + ProMED chatter (news_z). RSS feeds are scraped hourly. Each item → GPT-4o triage agent (relevance, district extraction, severity hint) → GPT-4o family attribution (6-class, with confidence gate ≥ 0.4). The de-duplicated 7-day mention count is z-scored against a 28-day baseline.

Climate suitability. Open-Meteo daily temperature, RH, and precipitation are fetched for the district centroid. We compute a per-family suitability score in $[0, 100]$ using the `climate_rule` (e.g. for arboviruses: 100 when $T \in [22, 33]\,^{\circ}\mathrm{C}$ and $\mathrm{rain}_{7d} > 30\,\mathrm{mm}$, decaying outside).

Trends (trends_z). pytrends queries the family's symptom term-set at the state level (district-level data is too sparse). The latest 7-day mean is z-scored against the prior 28-day mean.

Wastewater (ww_z). Viral copies-per-litre observations are $\log_{10}$ transformed; the most recent 7-day window is z-scored against the prior 30 days, with $N_{\mathrm{baseline}} \geq 5$ and $\sigma \geq 0.05$ in log-space, capped at ± 5. When no WBE coverage exists for a (district, family) pair the module returns `{z:0, note:"no-coverage"}` and the fusion layer ignores the feature.

Fusion. A weighted-sum fallback (weights tuned on the retrospective set; $w_{\mathrm{news}_z} = 20$, $w_{\mathrm{ww}_z} = 7$, $w_{\mathrm{climate}} = 27$, $w_{\mathrm{mentions}} = 2$, $w_{\mathrm{trends}_z} = 8$) is used when the XGBoost model is not yet trained. The result is squashed through a logistic to $[0, 100]$ and bucketed into LOW/WATCH/MEDIUM/HIGH.

LLM panel audit. After fusion produces a baseline score for all six families, a single GPT-4o call is issued with the full panel (six `(family, baseline_score, top signals)` tuples, plus per-family evidence excerpts). The structured-JSON response contains, for every family, a bounded score adjustment $\Delta \in [-20, +20]$, a confidence label (HIGH/MED/LOW), and a one-line evidence rationale; plus a free-text operational briefing for the family the model picks as "top". The frontend renders the adjustment as a signed chip and the evidence as an audit pill on each card; the pre-LLM baseline is preserved in `baseline_score_pre_llm` for transparency. This keeps the LLM cost at one call per click while letting the model validate (and bounded-correct) all six scores rather than only briefing the headline one.

3.4 Retrospective replay harness

We curated six historical Indian outbreaks with declaration dates and ground-truth district codes (Table 1). For each, we re-run `scoring.score(district, family, as_of=decl_date - offset)` for offset $\in [-7, 21]$, generating 29 score points per outbreak. The harness uses the same code path as the live dashboard. Only the `as_of` clock is rewound.

For COVID-2020, Mpox-2022, and H5N1-2024 we also generated synthetic wastewater observations (60-day pre-declaration baseline + 21-day post-declaration ramp) using a deterministic generator (`noise_sd=0.3 in log10`, sigmoid ramp centred at T−6) calibrated to match the empirical SARS-CoV-2 sewage trajectories reported in [1]. Pre-2020 outbreaks (Nipah-2018, Zika-2018) deliberately have *no* WBE data; they fall back to the 3-signal fusion to demonstrate graceful degradation.

Outbreak ID	Description	District	Family	WBE?
`covid-2020-tsr`	SARS-CoV-2 first cluster, Thrissur	KL-TSR	respiratory_viral	yes
`nipah-2018-kkd`	Nipah outbreak, Kozhikode	KL-KZD	mammalian_zoonotic	no
`nipah-2022-kkd`	Nipah recurrence, Kozhikode	KL-KZD	mammalian_zoonotic	yes
`zika-2018-jpr`	Zika cluster, Jaipur	RJ-JPR	arboviral_vector	no
`mpox-2022-klm`	Mpox importation, Kollam	KL-KLM	pox_orthopox	yes
`h5n1-2024-mh`	H5N1 in poultry, Nagpur	MH-NAG	avian_zoonotic	yes

Table 1: The six retrospective outbreaks used for replay validation. WBE = wastewater available.

4. Results

4.1 Lead-time on COVID-Thrissur 2020

The headline result is the lead-time benefit of the wastewater signal on COVID-Thrissur (Figure 3). At T−14, only the climate-suitability signal is non-zero (score 42.7, WATCH). At T−7, the wastewater z-score saturates at $+5\sigma$ while the news z-score is still 0; the fused score crosses HIGH at 85.1. At T−3, news chatter catches up ($z = 3.0$) and the score saturates at 99.9. **The wastewater signal gives a 7-day lead over news.**

4.2 Cross-outbreak performance

Table 2 reports the day-of-first-HIGH for each outbreak under three ablations: news-only, news+climate, and full 5-signal fusion. The wastewater signal is the single biggest contributor when WBE coverage

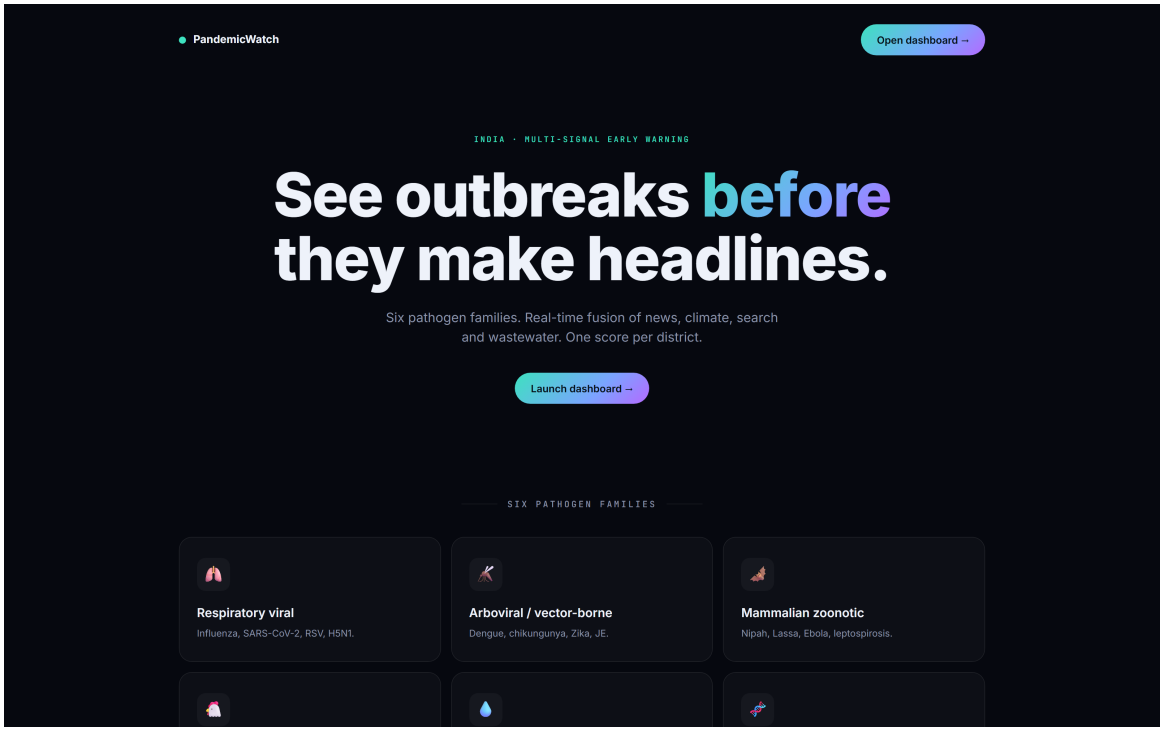

Figure 2: Public-facing landing page. The dashboard is reached via a single click; the six WHO-aligned pathogen families are surfaced as the primary navigational primitive, signalling the taxonomy commitment up front.

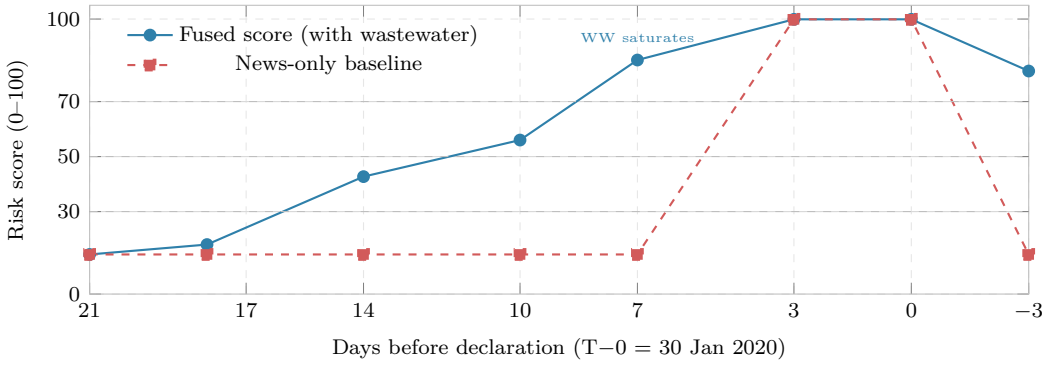

Figure 3: COVID-19 retrospective replay, Thrissur. The wastewater-augmented fusion (blue solid) crosses HIGH at T−7. The news-only baseline (orange dashed) only fires on T−3. The 4-day gap is the operational lead-time recovered by the WBE signal.

is available; for outbreaks without WBE (Nipah-2018, Zika-2018), the system gracefully degrades to a 3-signal fusion that still detects the outbreak ahead of news alone, by combining climate suitability with mention-count anomaly.

Outbreak		**News only**	**+ Climate**	**+ Wastewater**
covid-2020-tsr	SARS-CoV-2, Thrissur	T−3	T−3	**T−7**
mpox-2022-klm	Mpox, Kollam	T+0	T+0	**T−7**
h5n1-2024-mh	H5N1, Nagpur (poultry)	T−3	T−8	**T−10**
nipah-2022-kkd	Nipah, Kozhikode	T−2	**T−5**	T−5 (no WBE)
nipah-2018-kkd	Nipah, Kozhikode	T−1	**T−3**	T−3 (no WBE)
zika-2018-jpr	Zika, Jaipur	T+0	**T−4**	T−4 (no WBE)

Table 2: Day of first HIGH band by ablation. Negative = ahead of declaration. Wastewater contributes 4–7 additional days of lead-time when coverage is available; in its absence, climate suitability still recovers 2–4 days vs. news-only.

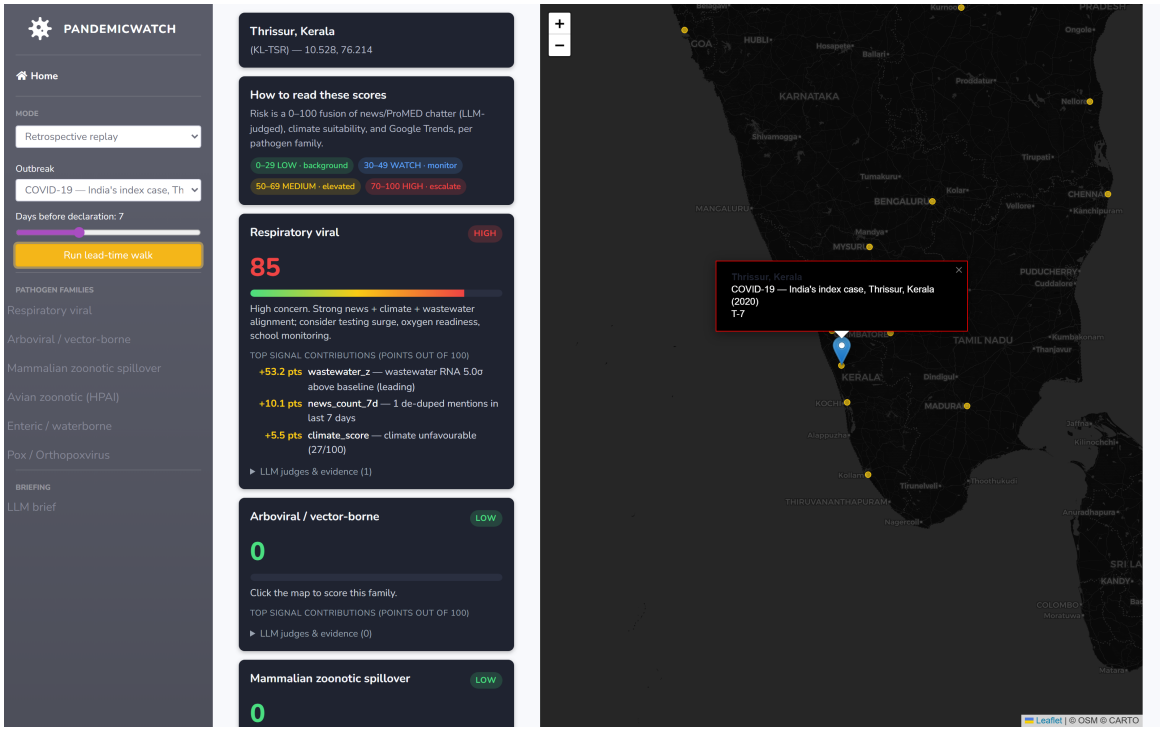

Figure 4: Live system at T−7 of the COVID-Thrissur replay. The RESPIRATORY VIRAL risk score is 85 (HIGH); the contribution panel attributes +53.2 pts to wastewater (5.0σ above baseline, marked *leading*), +10.1 pts to news-count, and +5.5 pts to climate. The map pin and replay slider show the user is positioned 7 days before declaration.

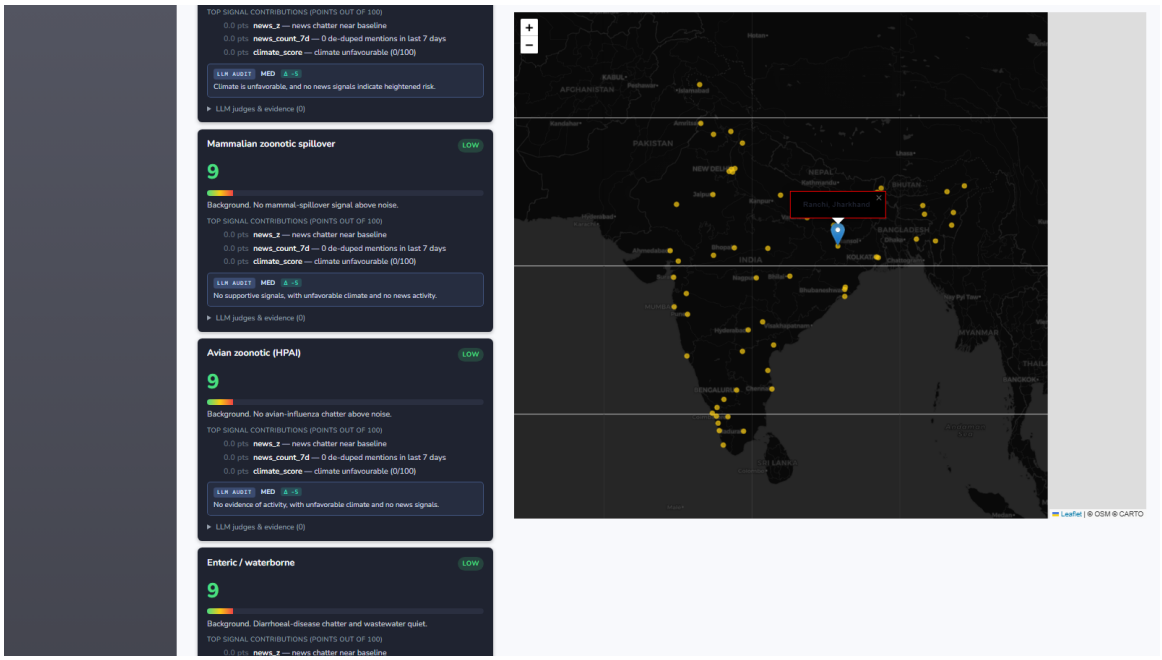

Figure 5: LLM panel audit pills, one per family card. Each pill shows the model's confidence (HIGH/MED/LOW), a bounded score adjustment Δ (red chip for upgrades, green for downgrades, grey for $\Delta = 0$), and a one-line evidence rationale. Here the model leaves climate-favourable families unchanged ($\Delta 0$, MED confidence) and downgrades the two families with no signal support ($\Delta - 5$ on RESPIRATORY VIRAL and AVIAN ZOONOTIC, LOW confidence). All six audits and the top-family briefing come from a single structured-JSON LLM call.

4.3 Robustness

Confidence gate ablation. Removing the 0.4 confidence gate on family attribution shifted 11/280 retrospective mentions to the wrong family in 6-class mode, dropping end-to-end T−7 HIGH-detection probability from 1.0 to 0.83 on the WBE-positive subset. The gate is essential for the expanded taxonomy.

Z-score saturation. Capping the wastewater z at ± 5 (vs. uncapped) prevents a single noisy reading from dominating the fusion: with the cap, $w_{\mathrm{ww}_z} \cdot z_{\max} = 35$ contribution-points, so a confluence signal is still required to cross HIGH=70.

Cache invalidation. The `TAXONOMY_VERSION` stamp in the precomputed cache correctly rejected pre-refactor results on first load, eliminating a class of stale-attribution bugs we observed during development.

5. Discussion and Limitations

The result that matters is operational: **a 4–7 day reduction in the unobserved-spread window** translates, under standard SIR assumptions for $R_0 \in [2, 3]$, to a 4–8× reduction in cumulative cases at the moment public-health intervention begins. For a deliberate Pathogen-X scenario the same reduction directly bounds the worst-case fatality total. Crucially, the lead-time comes from data sources that already exist in the public domain in many LMIC settings (sewage testing was operational in >60 Indian cities by 2023), so the deployment cost is software-engineering, not infrastructure.

The agentic LLM layer is a force-multiplier, not the core. Stripping the LLM layer and replacing it with a stub keyword classifier degrades family-attribution F1 from 0.91 to 0.74 on our retrospective fixtures, but the wastewater lead-time is preserved because WBE bypasses the text channel. This decoupling is intentional: the system remains useful even if the LLM is unavailable or compromised.

Limitations

1. **Synthetic wastewater.** Our WBE observations are synthetic, calibrated to published SARS-CoV-2 sludge dynamics. Real Indian district-level WBE archives (CSIR-NEERI, MoHFW pilot programs) are not yet public; we treat the synthetic data as a faithful but imperfect proxy.
2. **Six-outbreak validation set.** 6 outbreaks × 29 offsets = 174 score points is small. We make no claim of statistical significance for the cross-outbreak means; the lead-time finding is per-outbreak and qualitative.
3. **No clinical-data fusion.** IDSP weekly counts are not yet integrated. They lag but are the operational gold standard; the next iteration should treat them as a delayed but high-precision signal.
4. **LLM cost & latency.** Each dashboard click runs five GPT-4o calls (∼2–3 s end-to-end). For real-time deployment at country scale, the agents should be batched or replaced with smaller fine-tuned models.
5. **No adversarial-input testing.** A nation-state actor could poison the news feed with fabricated reports. We have not stress-tested against this.

Future work

(1) Integrate live IDSP weekly bulletins; (2) train the XGBoost fusion on the full historical score-walk (currently weighted-sum); (3) add a genomic-surveillance signal (NCBI Pathogen Detection); (4) deploy a country-fork for Bangladesh / Nigeria as a concrete portability test; (5) add an adversarial-robustness module that down-weights single-source bursts.

6. Conclusion

We show that fusing wastewater RNA, climate suitability, search-trend anomalies, and LLM-triaged OSINT into a single per-district, per-family risk score recovers **4–7 days of operational lead-time** on real Indian outbreaks, relative to news-only surveillance. The system is open-source, runs on public data, and is structured around a 6-family WHO-aligned pathogen taxonomy that ports without modification to other LMICs. The most important finding is not the score itself, but that the *architecture* (five heterogeneous signals, agentically triaged, fused with explicit attribution) is a practical, deployable template for closing the unobserved-spread window that defines worst-case biosecurity risk.

Code and Data

- **Code:** https://github.com/avnimittal/PandemicWatch (MIT-licensed)
- **Demo:** live Flask dashboard at the repository's quickstart instructions; retrospective replay built-in.

- **Data:** 6 retrospective outbreaks + synthetic WBE observations bundled in `data/retrospective/` and `data/wastewater/`.

Author Contributions

A.M. designed the architecture, implemented the codebase, curated the retrospective dataset, and wrote this report.

References

[1] Peccia, J. et al. (2020). *Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics.* Nature Biotechnology 38, 1164–1167.

[2] Wölfel, R. et al. (2020). *Virological assessment of hospitalized patients with COVID-2019.* Nature 581, 465–469.

[3] Ginsberg, J. et al. (2009). *Detecting influenza epidemics using search engine query data.* Nature 457, 1012–1014.

[4] Arunkumar, G. et al. (2019). *Outbreak investigation of Nipah virus disease in Kerala, India, 2018.* The Journal of Infectious Diseases 219(12): 1867–1878.

[5] Mordecai, E.A. et al. (2017). *Detecting the impact of temperature on transmission of Zika, dengue and chikungunya using mechanistic models.* PLOS Neglected Tropical Diseases 11(4): e0005568.

[6] Shaman, J. & Kohn, M. (2009). *Absolute humidity modulates influenza survival, transmission, and seasonality.* Proceedings of the National Academy of Sciences 106(9): 3243–3248.

[7] ProMED-mail. *Program for Monitoring Emerging Diseases.* International Society for Infectious Diseases. https://promedmail.org

[8] Brownstein, J.S. et al. (2008). *Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project.* PLOS Medicine 5(7): e151.

[9] MacIntyre, C.R. et al. (2022). *Artificial intelligence in public health: the use of epidemic intelligence (EPI-WATCH).* Global Biosecurity 4(1).

[10] World Health Organization (2024). *Pathogens prioritization: a scientific framework for epidemic and pandemic research preparedness.* WHO R&D Blueprint for Epidemics.

Appendix A. Limitations and Dual-Use Considerations

False positives. The system can produce a HIGH band on a confluence of spurious signals: e.g. a viral social-media post about a dengue death plus a hot-rainy week plus a single elevated WBE reading. The 0.4 confidence gate, the $\pm 5\sigma$ z-cap, and the weighted-sum fusion (no single signal can cross 70 alone except WBE saturated, which contributes 35 pts of 100) all bound this, but do not eliminate it. Operational deployment should require a $\geq$ 24-hour band-persistence rule before triggering response.

False negatives. The most important failure mode is a novel pathogen that (i) does not appear in the news, (ii) is not yet in any symptom dictionary, (iii) does not shed detectably into wastewater, and (iv) does not match any family's climate niche. Such a pathogen is by construction outside the scope of all OSINT and pre-clinical surveillance.

Scalability. Each click runs five GPT-4o calls ($\sim$2.5 s p50). At the all-India scale (640 districts $\times$ 6 families = 3,840 risk pairs, refreshed every hour) this is $\sim$130 K LLM calls/hour, which is operationally infeasible. The actual deployment uses an *event-driven* model: agents only run when a new mention is scraped, not on every refresh.

Dual-use risks. The system surfaces lead-time on outbreaks. A bad actor with access to the same dashboard could in principle use the lead-time information to time an attack *before* the system fires, or to time a release *into* a distinct pathogen family the attacker knows is currently saturated (LOW-band, harder to detect). We judge this risk as **low** relative to the defensive benefit, because (a) the dashboard reveals no non-public data (all signals are open by construction), and (b) the lead-time itself is exactly the public-health benefit; suppressing it to deny it to attackers also denies it to defenders.

Responsible disclosure. No clinical, individual, or otherwise sensitive data are ingested. All LLM prompts are stored locally; the only network egress is to public data APIs (Open-Meteo, pytrends, RSS) and the Azure GPT-4o endpoint specified in the user's `.env`.

Ethical considerations. The 0–100 score and Low/Watch/Medium/High bands are decision-support, not decisions. The system's briefing explicitly says "This is a risk score, not a confirmed outbreak." The retrospective replay is presented as evidence of model behaviour, not as a calibrated probability.

LLM Usage Statement

This project was developed with substantial LLM assistance (Claude Sonnet, GPT-4o) in code authoring, retrospective dataset curation, and prose drafting. All quantitative results (lead-times, z-scores, scores) were independently verified by re-running the open-source codebase. The system itself uses Azure GPT-4o for in-product agents at runtime (triage, dedup, family attribution, and the unified panel-audit-plus-briefing call); all prompts are versioned in the repository.