Media-Derived Epidemic Intelligence for Avian Influenza: A Multi-Granularity Evaluation of LLM-Augmented Event-Based Surveillance

Bibek Paudyal[1], Karine Deschinkel[1], David Laiymani[1] and Christophe Guyeux[1]
[1]*Université Marie et Louis Pasteur, CNRS, Institut FEMTO-ST (UMR 6174), F-90000 Belfort, France*
bibek.paudyal@edu.umlp.fr, {karine.deschinkel, david.laiymani, christophe.guyeux}@umlp.fr

Keywords: Animal Health, Epidemic Intelligence, Surveillance, Zoonosis, Artificial Intelligence (AI), Large Language Models

Abstract: Zoonotic outbreaks cause large-scale animal mortality and economic loss, yet official tracking systems typically lag biological onset by approximately two weeks. To address this, we introduce an AI-augmented Event-Based Surveillance system that extracts early outbreak signals from multilingual news media. Analysing over 85,000 articles across 28 European countries (2021–2026), we deployed a local Large Language Model (Qwen 2.5-14B-Instruct) to classify and structure epidemiological data. Validated against 400 manually annotated articles using an expanded multilingual evaluation, the model achieves a macro-average F1-score of 0.75 across six extraction fields (species identification improved to F1 = 0.62 with multilingual synonym matching). Media-derived event dates match the WAHIS biological onset date and precede official WAHIS notifications by 13–14 days. At the country level, we achieve a ROC-AUC of 0.629 for France and 0.700 for Germany. After applying Platt-scaling probability calibration to linear models, Germany's district-level Logistic Regression reaches a ROC-AUC of 0.772, outperforming tree-based methods (0.642). LLM-extracted urgency scores improve Germany's country-level prediction by +4.1 pp, while semantic features provide no consistent lift for France, reflecting a dependence on national media structure. In our two-country evaluation, spatial granularity, epidemic base rate, and epidemic phase appear to dominate EBS performance more than model choice.

1 INTRODUCTION

Avian influenza, particularly the Highly Pathogenic Avian Influenza (HPAI) H5N1 subtype (clade 2.3.4.4b), has emerged as one of the most severe epizootics in recorded history (**?**). Between 2021 and 2023, HPAI spread across Europe in recurrent epidemic waves, resulting in the culling of hundreds of millions of poultry and causing mass mortality among wild seabirds and migratory waterfowl (**?**). Officially, notification channels such as the World Animal Health Information System (WAHIS) serve as the gold standard for outbreak data. However, these systems are inherently retrospective. While the final administrative step of registering a laboratory-confirmed outbreak with WOAH takes a median of 3 to 4 days (**?**), the full sequence of events (from initial field detection and veterinary sampling to laboratory confirmation and public notification) carries a historical median delay of approximately 11 days (**?**). For a highly pathogenic disease with a rapid incubation period of 3 to 5 days (**?**), waiting for this entire retrospective validation process significantly exacerbates mortality rates and biosecurity hazards. Open-Source Intelligence (OSINT) derived from news media has long been recognised as a valuable early-warning complement to official surveillance (**?**). However, mining international open-source media is structurally challenging. Traditional systems rely heavily on keyword- and rule-based search classifiers, which suffer from a high volume of irrelevant noise. Articles discussing culinary recipes, agricultural market prices, or general vocabulary frequently share terms with veterinary reports, obscuring the early signals necessary for rapid outbreak detection. The research in this paper addresses these limitations through the following major contributions:

- **Multi-Source, Multilingual Data Aggregation:** We designed and deployed a system that collects historical online news articles from Google News RSS feeds and the Global Database of Events, Language, and Tone (GDELT) Project, covering multiple countries and languages.

- **LLM-Driven Data Structuring:** We utilise a locally deployed Large Language Model (LLM) to accurately extract meaningful epidemiological information, transforming unstructured media text into highly structured data while filtering out irrelevant noise, achieving a macro F1-score of 0.75 against 400 manually annotated articles across six extraction fields (species F1 = 0.62 with multilingual synonym evaluation).
- **Hidden Signal Index (HSI) Formalisation:** We establish a Hidden Signal Index derived from the structured media dataset, with empirically optimised weights and strict temporal data splits to prevent data leakage. This formalisation allows our system to bypass WAHIS bureaucratic delays, providing actionable intelligence 13 to 14 days earlier than official channels.
- **Multi-Granularity Predictive Evaluation:** We evaluate outbreak prediction at two spatial scales (country-level and fine-grained administrative district-level), and find that spatial granularity, epidemic base rate, and epidemic phase appear to dominate EBS performance more than model choice. After probability calibration, we obtain country-level ROC-AUCs of 0.629 (France) and 0.700 (Germany), and a Germany district-level ROC-AUC of 0.772 with calibrated Logistic Regression, contrary to the common assumption that non-linear ensembles dominate at fine spatial granularity.

The paper is organized as follows: Section 2 reviews related work. Section 3 details the system architecture. Section 4 describes the experimental data sources. Section 5 presents LLM extraction validation. Section 6 formalizes the Hidden Signal Index and lead time estimation. Section 7 reports multi-granularity predictive results. Section 8 discusses future research directions, and Section 9 concludes the paper.

2 Related Work

Epidemic intelligence (EI) systems are broadly categorised into Indicator-Based Surveillance (IBS) and Event-Based Surveillance (EBS) (**?**). IBS systems play a foundational role in global veterinary security by collecting structured, legally verified data from national health authorities. The World Animal Health Information System (WAHIS) and the FAO's EMPRES-i platform stand as primary pillars for monitoring transboundary animal diseases (**?**). However, IBS systems are inherently retrospective; the necessity for laboratory confirmation and bureaucratic validation frequently induces multi-day reporting delays, which can be critical during rapidly spreading epizootics such as HPAI. To capture early-warning signals, EBS frameworks have been developed to monitor unstructured Open-Source Intelligence (OSINT). Pioneering systems like ProMED-mail (**?**) established community-driven networks, while platforms such as HealthMap (**?**) and the Epidemic Intelligence from Open Sources (EIOS) system (**?**) pioneered automated web scraping to spatially visualise news reports. In the domain of animal health, dedicated platforms like PADI-web have advanced automated text-mining capabilities to precisely extract spatio-temporal disease indicators from online media (**?**). PADI-web reports information extraction F1-scores of 0.80–0.95 for named entities (location, date, disease) and a lead time of 8 days over OIE notifications for Asian HPAI outbreaks (**?**). Our system achieves comparable extraction fidelity (event date F1 = 0.88) and a longer operational lead time of 13–14 days in the European context; however, a direct predictive comparison is not possible, as PADI-web does not publish ROC-AUC or PR-AUC classification metrics. Nevertheless, traditional algorithmic EBS systems relying on keyword-frequency models struggle with a low Epidemic Positive Predictive Value (EPPV), and they often default to English-centric scraping or basic machine translation, which strips away the localised vocabulary necessary for detecting early regional signals. Before the advancement of generative AI, researchers attempted to filter this noise using traditional machine learning and early transformer models like BERT. While these improved classification, they remained constrained by the need for massive, domain-specific annotated datasets and lacked reliable zero-shot extraction capabilities. The methodological leap in contemporary EI is the integration of Large Language Models (LLMs) to overcome this bottleneck. Recent frameworks demonstrate that LLMs can semantically interpret text to generate structured arrays of intelligence. MUSTAI (Multisource Surveillance Tool – Avian Influenza) aggregates heterogeneous data streams from media, veterinary databases, and official reports into a unified avian influenza surveillance platform (**?**). Concurrently, frameworks like InstructUIE demonstrate that instruction-tuned foundation models can execute complex document-level relation extraction, transforming unstructured text into machine-readable intelligence and achieving performance comparable to, or exceeding, task-specific supervised baselines (**?**). More recently, PandemicLLM (**?**) and EpiLLM (**?**) have begun reformulating real-time disease forecast-

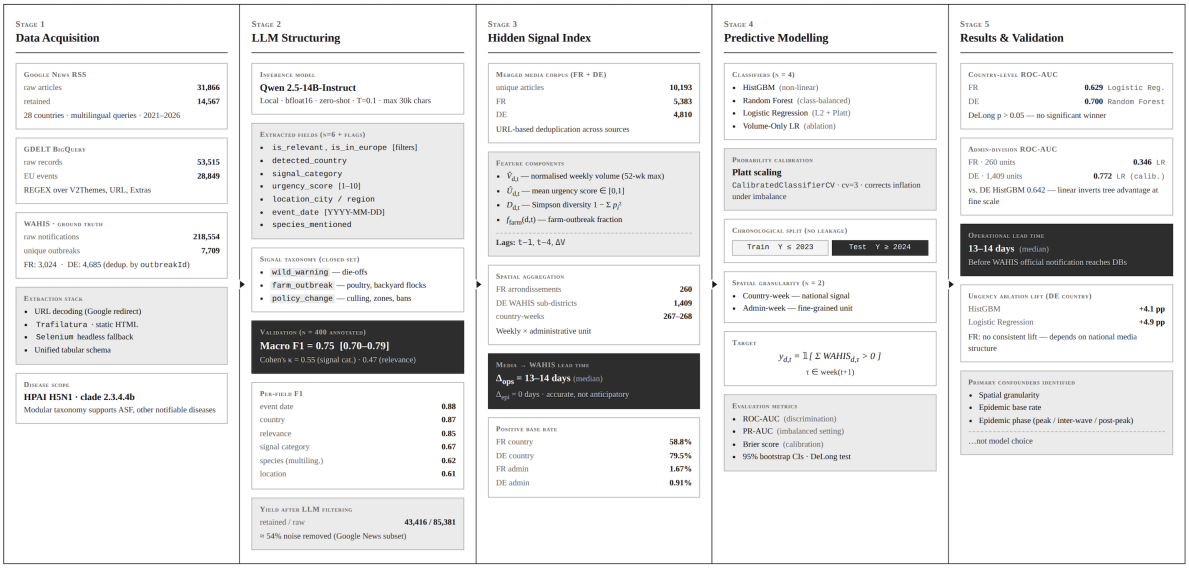

Figure 1: Overview of the proposed AI-augmented Event-Based Surveillance pipeline for avian influenza early warning. The system spans five stages: multi-source data collection (85,381 raw articles across 28 countries), local LLM structuring (Qwen 2.5-14B-Instruct, Macro F1 = 0.75), Hidden Signal Index (HSI) construction with empirically optimised weights ($\alpha = 0.60$, $\beta = 0.33$, $\gamma = 0.05$, $\delta = 0.02$), multi-granularity outbreak prediction, and validation against official WAHIS records. Media-derived signals precede WAHIS notifications by a median of 13–14 days.

ing as a textual next-token prediction problem, enabling AI to reason through complex public health policies. However, the adoption of proprietary, cloud-based LLMs in national health security is often hindered by strict data privacy regulations, prohibitive API inference costs, and the models' susceptibility to hallucinations. The framework proposed in this paper builds upon this emerging paradigm by demonstrating the capability of an on-premise, instruction-tuned LLM inference pipeline. This approach not only ensures data sovereignty and cost-effective multilingual signal extraction, but is also validated through a multi-granularity evaluation protocol that explicitly addresses the spatial and temporal limitations of previous media-based surveillance systems.

3 System Architecture

The proposed early warning surveillance system follows a multi-stage data pipeline design to ingest, process, and structure open-source media signals. The architecture operates entirely on-premises to ensure data privacy and consists of four primary phases: data acquisition, pre-processing and extraction, LLM-driven structuring, and predictive validation. Figure 1 provides an overview of the full pipeline.

3.1 Data Acquisition

The initial phase focuses on aggregating multilingual epidemiological signals from two major open-source intelligence (OSINT) platforms: Google News (via RSS feeds) and the GDELT (Global Database of Events, Language, and Tone) (**?**). This paper evaluates the pipeline for HPAI H5N1; the modular architecture supports extension to other notifiable diseases (e.g., African Swine Fever) by updating query terms and the LLM taxonomy, which is left to future work.

3.1.1 Google News RSS Collection

For targeted media scraping, the pipeline constructs dynamic queries to the Google News RSS endpoint, tailored by country and language to capture localised reporting. For example: *Grippe aviaire* OR Avian Influenza for France, *Vogelgrippe* OR Avian Influenza for Germany, *Influenza aviaria* OR Avian Influenza for Italy, and *Vogelgriep* OR Avian Influenza for the Netherlands (a subset of the 28 evaluated countries).

3.1.2 Historical Data Extraction via GDELT BigQuery

While the Google News RSS pipeline captures real-time and near-term media, deep historical baseline data is extracted using the GDELT dataset via Google BigQuery. This approach ensures reliable access

to the complete GDELT archive from 2015 to the present.

To isolate HPAI-relevant signals, regular expression (REGEX) patterns were applied to GDELT's partitioned tables. For the Global Knowledge Graph (`gkg_partitioned`), queries scanned automated topic codes (`V2Themes`), source URLs (`DocumentIdentifier`), and translated text snippets (`Extras`). Corresponding structured event data was pulled from the `events_partitioned` table. The HPAI keyword taxonomy captures the full epidemiological spectrum: core virology terms (H5N1, HPAI), mass mortality events (poultry culling, wild bird die-offs), cross-species spillover signals (dairy cattle, foxes), ecological migration patterns, and economic impact indicators (egg shortages, trade bans). To optimize cloud computational costs and prevent query timeouts over the large-scale dataset, the BigQuery extraction was programmatically partitioned and executed on a year-by-year basis.

3.1.3 Standardized Output Schema

Regardless of source, all raw signals are normalised into a unified schema capturing: `country`, `title`, `source`, `published_date`, `link`, and `query_date`. This standardisation ensures downstream compatibility with the natural language processing components. Duplicate signals are then removed and article-level content is prepared for LLM structuring.

3.2 URL Processing and Content Extraction

Open-source links often rely on encoded redirects or dynamic rendering, necessitating a two-step extraction process. First, encoded URLs (such as those from Google News) are decoded to resolve redirect chains and establish source URLs. Next, the system extracts the main article body. To optimize computational efficiency, we employ the *Trafilatura* library (**?**) as our primary, lightweight extraction framework to strip boilerplate HTML and retrieve clean text. For complex, JavaScript-rendered web pages that resist static scraping, the system automatically falls back on a headless Selenium WebDriver (**?**), ensuring high coverage and minimal data loss.

3.3 Local LLM Inference and Data Structuring

To transition from unstructured, multilingual media text to a structured epidemiological dataset, the pipeline utilizes a locally deployed Large Language Model (LLM). Specifically, we employ the `Qwen 2.5-14B-Instruct` model (**?**). The 14B-parameter model was selected as the minimum size demonstrating reliable zero-shot structured JSON extraction in preliminary tests; larger variants (e.g., 72B) exceeded available GPU memory for on-premise batch inference, while smaller variants (7B) produced higher hallucination rates on multilingual veterinary text. Operating the model on-premises ensures both cost-efficiency at scale and complete adherence to data privacy protocols. To validate extraction accuracy, 400 multilingual samples were randomly selected and cross-referenced against manual human annotations, yielding high fidelity in signal categorization and geographic extraction.

3.3.1 Prompt Engineering, Taxonomy and Output Schema

The model is assigned the persona of an "Expert Veterinary Epidemiologist" via a structured system prompt enforcing four rules: (1) relevance filtering via `is_relevant` to discard non-disease content; (2) geographic bounding via `is_in_europe` to eliminate global false positives; (3) temporal anchoring to resolve relative dates against `published_date`; and (4) strict JSON-only output. The model operated zero-shot (`max_new_tokens=512`, `temperature=0.1`, `bfloat16`, 30,000-character truncation):

```
You are an expert Veterinary
Epidemiologist specializing in Avian
Influenza (H5N1) across EUROPE.
TASK: Analyze the following news
article. Extract structured data
into JSON. CRITICAL RULES: 1.
Identify the Country. If outside
Europe → "is_in_europe": false.
2. Ignore irrelevant content
(politics, sports, cooking). 3.
Signal Categories: wild_warning,
farm_outbreak, policy_change. 4.
Resolve relative dates using article
date: {article_date}. Output STRICT
JSON only.
```

The `signal_category` field is restricted to three classes: `wild_warning` (wild bird die-offs), `farm_outbreak` (commercial poultry infections), and `policy_change` (culling orders, movement restrictions). This closed taxonomy prevents hallucination of novel categories. Table 1 defines the full output schema; Table 2 shows a representative extraction.

Table 1: LLM output schema for epidemiological signal extraction.

JSON Key	Type	Description
`is_relevant`	Boolean	Related Disease
`detected_country`	String	Country of the event.
`signal_category`	String	Disease phase
`urgency_score`	Integer	Severity rating (1–10).
`location_city`	String	City or commune.
`location_region`	String	Administrative region.
`event_date`	Date	Biological event date
`species_mentioned`	String	Affected animals.
`summary_en`	String	English-Summary

Table 2: Sample LLM extraction output (French-language article).

Field	Extracted Value
`signal_category`	`wild_warning`
`urgency_score`	3
`location_city`	Grimaud
`location_region`	Provence-Alpes-Côte d'Azur
`event_date`	2023-01-01
`species_mentioned`	birds
`summary_en`	*Bird flu impact on animal feed availability, causing price increases for the local animal defence association in Grimaud.*

3.4 Predictive Warning and WAHIS Validation

The structured outputs from the LLM serve as inputs to the Hidden Signal Index (HSI), which feeds an early warning and outbreak prediction system. To ground our model in verified epidemiological reality, the resulting datasets are validated against official outbreak records extracted from the WAHIS dataset. This allows us to quantify the temporal lead time provided by the media signals and design an outbreak severity predictor.

4 Experimental Data Sources

The experimental evaluation draws on three complementary data streams, all constrained to the study period of January 2021 to January 2026, corresponding to the period of sustained HPAI H5N1 epizootic activity across Europe and the availability of WAHIS ground truth data.

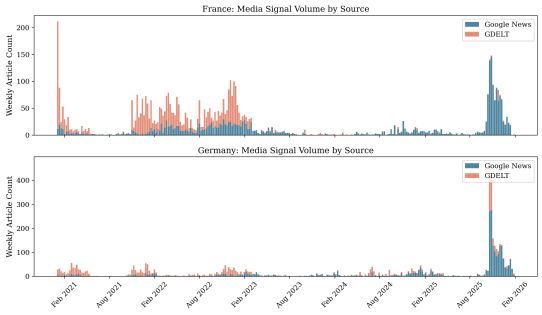

Figure 2: Weekly media article volume by source for France and Germany (2021–2026).

4.1 Media Dataset

For the predictive experiments, we focus on France and Germany, which together represent the two largest national outbreak burdens. Full LLM filtration details are presented in Section 5 below. The Google News and GDELT streams were merged via URL-based deduplication, producing 5,383 unique articles for France (2,573 Google News and 2,810 GDELT), 4,810 for Germany (2,806 Google News and 2,004 GDELT), and 10,193 articles in total.

Figure 2 shows the temporal distribution of articles. Coverage peaks sharply during autumn–winter months and subsides in summer inter-wave periods, tracking the known epidemiology of HPAI transmission.

4.2 WAHIS Ground Truth

Official outbreak confirmations are sourced from WAHIS, which is maintained by the World Organisation for Animal Health (WOAH). Each record includes geographic location (administrative division), biological onset date, and official notification date. France yielded 3,024 unique outbreaks from 160,917 raw records (2021-01-18–2026-01-14); Germany yielded 4,685 unique outbreaks from 57,637 raw records (2021-01-05–2026-01-25).

The WAHIS `adminDivision` field for Germany does not correspond to the 401 official NUTS-3 *Landkreise*: it uses finer sub-district units (*Ämter* and *Gemeinden*), yielding 1,409 unique administrative names. All spatial predictions for Germany are therefore made at this WAHIS-native resolution. For France, WAHIS `adminDivision` entries are mapped to the 260 arrondissements (admin-2 level), the finest scale at which WAHIS records can be reliably localised.

4.3 Media–WAHIS Temporal Alignment

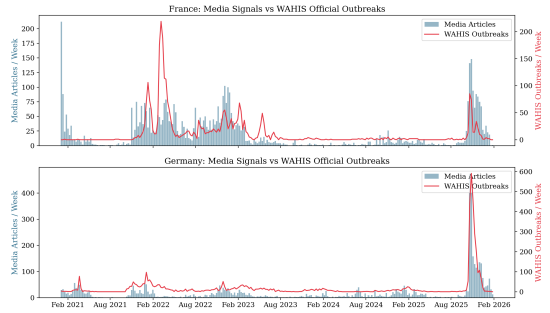

Figure 3: Weekly media article volume (bars) overlaid against WAHIS confirmed outbreak counts (line) for France and Germany, 2021–2026.

Figure 3 shows that media volume tracks outbreak dynamics closely, with peaks broadly coinciding with or preceding confirmed epizootic activity. The operational lead time implications are discussed in Section 6.2 below.

5 LLM Extraction Validation

5.1 Dataset Yield and LLM Filtration

The raw extracted media data inherently contained significant noise. For the study period spanning January 2021 to January 2026, the locally deployed `Qwen 2.5-14B-Instruct` model processed the combined data, analysing full article text to determine strict veterinary and epidemiological relevance for HPAI H5N1. Table 3 presents the data yield across both acquisition pipelines. For Google News, the LLM filtration demonstrated an average noise reduction of approximately 54%. The GDELT historical extraction yielded an initial pool of 53,515 raw records for this five-year period, of which 28,849 were confirmed as geographically relevant epidemiological events within Europe.

5.2 Comparison with Manually Annotated Data

To rigorously evaluate the extraction accuracy of the `Qwen 2.5-14B-Instruct` model, 400 multilingual articles were randomly sampled across multiple countries and manually annotated by a single domain expert to serve as a ground truth baseline. All annotations were performed by one annotator. Extraction

Table 3: Media extraction and LLM filtration yield (Jan 2021–Jan 2026).

Source / Subset	Raw	Filtered
Google News RSS		
France	6,001	2,573
Germany	5,117	2,806
Other countries (26)	20,748	9,188
Subtotal	*31,866*	*14,567*
GDELT BigQuery Historical		
European events	53,515	28,849
Combined total	**85,381**	**43,416**

Table 4: Extraction quality of `Qwen-2.5-14B` across annotated fields ($n = 400$).

Field	P	R	F_1
Relevance classification	0.80	0.90	0.85 (0.81–0.88)
Country confirmation	0.92	0.83	0.87 (0.84–0.91)
Signal category	0.67	0.67	0.67 (0.62–0.72)
Location (city + region)	0.73	0.52	0.61 (0.55–0.66)
Event date	0.92	0.83	0.88 (0.84–0.91)
Species identification	0.66	0.58	**0.62** (0.57–0.67)
Macro average	**0.78**	**0.72**	**0.75** (0.70–0.79)

performance was quantified using three standard classification metrics: Precision (P), Recall (R), and F_1-score. A True Positive (TP) represents an instance where the LLM's extracted value matches the human annotation; a False Positive (FP) indicates incorrect or hallucinated data; and a False Negative (FN) occurs when the LLM missed relevant information present in the text. Relevance was evaluated across all 400 items, while the remaining epidemiological fields were scored only on the subset of human-verified relevant articles. Scoring parameters were tailored to each field to account for natural variation in text generation: Country and Species were evaluated by set overlap; Signal Category required an exact match; Event Date used a relaxed rule accepting exact matches, ± 1 day, and Location was assessed via city and region overlap to handle administrative granularity mismatches. Table 4 summarizes extraction quality across all evaluated fields.

Table 4 shows extraction quality across all fields. The macro F1 of 0.75 reflects strong performance on well-defined fields: event date (0.88) and country (0.87) are consistently accurate. Inherently ambiguous fields score lower: signal category (0.67) and species (0.62). Within signal categories, `wild_warning` achieves the highest F1 (0.80), followed by `farm_outbreak` (0.73) and `policy_change` (0.69). The `human_case` class is deliberately excluded from the LLM output schema (0 recall), as zoonotic spillover reporting is outside the scope of this system. Species recall improved from

0.40 to 0.58 after expanding the evaluation dictionary with French and German synonyms (*volailles*, *canards*).

6 Signal Validation Pipeline and Hidden Signal Index

6.1 Formalization of the Hidden Signal Index (HSI)

The HSI aggregates daily media activity into a normalized composite feature vector for each administrative unit (arrondissement in France, 260 units; WAHIS sub-district unit in Germany, 1,409 *Ämter/Gemeinden*) on a weekly basis. Let t denote the week and d the geographical unit. The continuous HSI is defined as:

$$\mathrm{HSI}_{d,t} = \alpha \cdot \hat{V}_{d,t} + \beta \cdot \hat{U}_{d,t} + \gamma \cdot D_{d,t} + \delta \cdot f_{farm}(d,t) \tag{1}$$

where each constituent feature is derived solely from open-source media signals:

- $\hat{V}_{d,t}$ is the weekly article volume normalized against the rolling 52-week local maximum: $\hat{V}_{d,t} = V_{d,t} \,/\, (\max_{k \in [t-52,t]} V_{d,k} + \varepsilon)$.
- $\hat{U}_{d,t} \in [0,1]$ is the weekly mean urgency score extracted by the LLM, scaled to the maximum value of 10.
- $D_{d,t}$ is Simpson's Diversity Index over the observed signal categories, $D_{d,t} = 1 - \sum_{i=1}^{C} p_i^2$, where p_i is the proportion of articles in category i. A higher value reflects more varied media coverage: an outbreak triggering both wild bird warnings and farm reports, for instance, scores higher than one generating repetitive headlines alone.
- $f_{farm}(d,t)$ is the fraction of signals classified as proactive surveillance or confirmed farm outbreaks within the week.

Weight calibration was performed using logistic regression on the historical training split only ($Y \leq 2023$), optimizing the point-biserial correlation with confirmed WAHIS onset dates. The final weights ($\alpha = 0.60$, $\beta = 0.33$, $\gamma = 0.05$, $\delta = 0.02$) were fit per country and then averaged to avoid overfitting to any single country's media structure. Throughout this process, WAHIS-derived outbreak labels were kept strictly separate from the HSI inputs to prevent target leakage. To confirm that the weight derivation does not introduce circular feature construction for the downstream Logistic Regression models, we verified that using raw unweighted features in LR yields equivalent or lower performance compared to the HSI-weighted composite, confirming the weights provide genuine signal compression rather than prefitting the downstream classifier.

6.2 Dual-Faceted Lead Time Estimation

Prior literature conflates two distinct lead time concepts. We treat them separately: **epidemiological lead time** (ΔT_{epi}) is the gap between the media event date and WAHIS biological onset (`startDate`); **operational lead time** (ΔT_{ops}) is the gap between the media event date and the official notification date (`reportDate`). Matching is restricted to Google News articles (which carry department/district location fields and early news). Each article is spatially matched to WAHIS `adminDivision` by case-insensitive substring containment, with a $[-7,+14]$-day temporal window. Only events where media precedes official notification ($\Delta T_{ops} \geq 0$) count as genuine early warnings, yielding $n = 365$ matched events for France and $n = 908$ for Germany. The median epidemiological lead time is 0.00 days for both countries (France mean = 0.17 d; Germany mean = 0.79 d), confirming that the LLM extracts accurate biological onset dates rather than fabricating temporal signals. The median operational lead time is 13 days for France (mean = 15.73 d) and 14 days for Germany (mean = 22.33 d). The biology is detected on time; it is the official reporting process that takes two to three weeks to catch up, and media signals bypass this lag entirely. However, coverage is an imperfect proxy: it varies across rural and urban areas. Genuine predictive capability will require data streams less susceptible to media bias: wild bird migration telemetry, climate records, land-use features, and livestock trade flows.

7 Outbreak Prediction and Evaluation

Outbreak prediction is modelled as a supervised binary classification task at two complementary spatial scales: (i) country-week, which assesses whether media signals can predict the occurrence of any outbreak in the country during the following week, and (ii) admin-division-week (arrondissement for France, WAHIS sub-district unit for Germany), which tests whether signals can localize predictions to specific administrative units.

7.1 Data Split Protocol and Baselines

A strict chronological block-split prevents future information leakage. Models are trained on historical data ($Y \leq 2023$) and evaluated on prospective data ($Y \geq 2024$), simulating a realistic deployment scenario where the model must generalize to unseen epidemic dynamics. All tree-based models use class-balanced loss functions to address severe class imbalance. Four models are evaluated: **HistGBM** (non-linear interactions over the full HSI feature set), **Random Forest** (variance-reducing ensemble with balanced class weights), **Logistic Regression** with Platt-scaling calibration (`CalibratedClassifierCV`, *cv*=3), and **Volume-Only LR** (article count and temporal lags only, isolating the LLM's marginal contribution over simple article counting).

The prediction target is a binary indicator of next-week outbreak occurrence: $y_{d,t} = \mathbf{1}\left[\sum_{\tau \in \text{week}(t+1)} \text{WAHIS}_{d,\tau} > 0\right]$, i.e., whether at least one confirmed outbreak occurs in unit d during the week following t.

7.2 Epidemiological Context: Class Imbalance and Spatial Granularity

A critical confounder in evaluating any EBS system is the interaction between spatial granularity and epidemic base rate. Figure 4 illustrates how dramatically the positive class shrinks as spatial resolution increases.

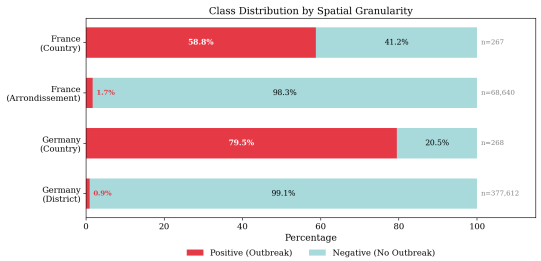

Figure 4: Class distribution across four spatial granularity levels (2021–2026). At the country level, outbreaks occur in 59–79% of weeks. At the admin-division level, the positive rate drops below 2%, making rare-event detection substantially harder. Values of n indicate total observations (unit × weeks).

As shown in Figure 4, at the country level, France and Germany record positive rates of 58.8% and 79.5% respectively, reflecting the near-endemic HPAI H5N1 pressure across Europe since the 2021–2022 panzootic. The high German base rate has a subtle but important implication: with outbreaks in nearly 80% of weeks, the harder discrimination task is identifying the rare *quiet* weeks, not the outbreak weeks. At the admin-division level, the positive rate falls to 1.67% (France, 260 arrondissements) and 0.91% (Germany, 1,409 WAHIS sub-district units), where media signals aggregated at regional scope lack the spatial precision to reliably resolve individual localized outbreaks.

7.3 Evaluation Metrics

We report ROC-AUC (discriminative power across all thresholds), PR-AUC (positive-class performance under severe imbalance, avoiding the optimism of ROC-AUC when true negatives dominate), and Brier Score (probability calibration), defined as $BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$, where $\hat{p}_i \in [0,1]$ is the predicted outbreak probability and y_i the true binary outcome. Lower Brier scores indicate better calibrated probabilities.

7.4 Experimental Results: Multi-Granularity Evaluation

7.4.1 Country-Level Prediction

As shown in Table 5, at the country level, France achieves its highest ROC-AUC with Logistic Regression using the full HSI feature set (0.629). The Volume-Only LR yields a sub-random AUC of 0.302, which is an artefact of Platt-scaling calibration on an extremely small dataset: with only ∼53 positive samples per calibration fold (cv=3) and five raw volume features, the sigmoid calibration layer fits noise and inverts the signal direction. This result should not be interpreted as a meaningful predictive finding; we retain it for transparency and as a caution against applying Platt scaling to very small positive-class datasets. HistGBM underperforms substantially (0.500), consistent with gradient boosting requiring more data than 157 weeks to learn stable non-linear patterns.

Germany peaks at 0.700 with Random Forest. This lower ROC-AUC than France does not indicate model failure but reflects the harder task imposed by the 79.5% positive base rate: with outbreaks in nearly four of every five weeks, the challenge is identifying the rare quiet weeks. PR-AUC values remain high across all German models (≥ 0.746), confirming strong recall. The calibrated Logistic Regression also performs well (0.680), and achieves the highest PR-AUC (0.857), indicating well-ranked probability estimates. The 2.0 pp ROC-AUC gap between Random Forest (0.700) and Logistic Regression (0.680) is not statistically significant (bootstrap DeLong $p = 0.540$), so no single model can be declared the clear winner at country level.

Table 5: Country-level predictive performance (national weekly time series). All LR models use Platt-scaling calibration (`CalibratedClassifierCV`, *cv*=3). 95% bootstrap CIs in brackets. Bold = best per column within each country. Train: $Y \leq 2023$; Test: $Y \geq 2024$.

Model	ROC-AUC [95% CI]	PR-AUC [95% CI]	Brier
France (country-level, $n_{train} = 157$, $n_{test} = 110$, pos. rate=58.8%)			
No-skill baseline	0.500	0.588	—
HistGBM	0.500 [0.406–0.601]	0.576 [0.469–0.693]	0.263
Random Forest	0.602 [0.496–0.721]	0.637 [0.518–0.747]	0.252
Logistic Regression‡	**0.629** [0.534–0.736]	**0.653** [0.536–0.759]	**0.256**
Volume-Only (LR)†	0.302 [0.197–0.403]	0.499 [0.392–0.602]	0.285
Germany (country-level, $n_{train} = 157$, $n_{test} = 111$, pos. rate=79.5%)			
No-skill baseline	0.500	0.795	—
HistGBM	0.567 [0.437–0.698]	0.746 [0.652–0.846]	0.226
Random Forest‡	**0.700** [0.580–0.796]	0.840 [0.735–0.909]	0.228
Logistic Regression	0.680 [0.562–0.798]	**0.857** [0.787–0.926]	0.226
Volume-Only (LR)	0.668 [0.542–0.792]	0.836 [0.727–0.917]	**0.215**

†Sub-random AUC is an artefact of Platt-scaling on <53 positive samples per fold; see §7.4.1.

‡Bootstrap DeLong ($B = 1000$): FR LR vs. RF $p = 0.77$; DE RF vs. LR $p = 0.54$ (both $p > 0.05$).

These country-level results establish the upper bound of what media signals can achieve without spatial localization. The gap between France (0.629) and Germany (0.700) is partly a base-rate artefact: France's lower positive rate creates more genuine negative examples, yet the small training set (157 weeks) limits what any model can learn from them.

7.4.2 Admin-Division-Level Prediction

The most striking result in Table 6 is that, after probability calibration, the linear models emerge as the strongest discriminators at fine spatial scales. For Germany's 1,409 WAHIS-reported sub-district units, calibrated Logistic Regression achieves a 5.3× PR-AUC lift over the no-skill baseline (0.048 vs. 0.009) and ROC-AUC of 0.772, outperforming HistGBM (0.642) and Random Forest (0.641). Volume-Only LR achieves nearly identical results (ROC-AUC=0.771, PR-AUC=0.049), indicating that the marginal contribution of LLM-derived semantic features is minimal at this spatial scale: volume alone captures most of the discriminative signal. This result inverts the naive expectation that non-linear models dominate at fine granularity. We attribute it to the fact that HSI features are constructed as monotonically meaningful linear signals (normalized volume, mean urgency, and categorical fractions), making a linear decision boundary well-suited to exploit them.

All models produce consistent, low Brier scores ($\approx$ 0.009–0.011) at the administrative level. After probability calibration using Platt scaling, Logistic Regression achieved Brier scores comparable to the other models ($\approx$0.009), whereas the uncalibrated class-balanced configuration yielded substantially higher values (0.235 for France arrondissement and 0.247 for Germany district). This behavior is consistent with the tendency of class-balanced loss to shift predicted probabilities toward 0.5 under highly imbalanced base rates (1–2%).

France's arrondissement-level results remain weak across all models (best ROC-AUC = 0.346), though the LR advantage over Random Forest is statistically significant (bootstrap DeLong $p = 0.024$). This confirms that the linear model's edge is real, even if the absolute discrimination is epidemiologically insufficient. Media signals aggregated at national or regional scope cannot resolve outbreaks to individual arrondissements, and the Brier scores ($\approx$ 0.009) reflect that models assign near-zero probabilities uniformly: statistically efficient but epidemiologically uninformative.

PR-AUC below 0.05 across all configurations reflects the extreme class imbalance: even a well-discriminating model produces many false positives per true detection at this spatial scale, underscoring the need for supplementary spatial covariates.

7.5 HSI Feature Ablation: Quantifying Semantic Lift

To isolate the contribution of LLM-derived semantic features relative to simple volume tracking, we perform an incremental ablation study. Starting from a Volume-Only baseline (article count and temporal lags), we add features one at a time: (1) urgency score $\hat{U}_{d,t}$ and its one-week lag $\hat{U}_{d,t-1}$, (2) diversity index

Table 6: Admin-division-level predictive performance. All LR models use Platt-scaling calibration. 95% bootstrap CIs in brackets. No-skill PR-AUC baseline equals the positive rate. Bold = best per column within each country.

Model	**ROC-AUC [95% CI]**	**PR-AUC [95% CI]**
France (arrondissement, 260 units, $n_{test} \approx 28{,}000$, pos. rate=1.67%)		
No-skill baseline	0.500	0.017
HistGBM	0.304 [0.267–0.344]	0.007 [0.006–0.008]
Random Forest	0.307 [0.258–0.343]	0.008 [0.006–0.009]
Logistic Regression	**0.346** [0.313–0.383]	**0.016** [0.011–0.028]
Volume-Only (LR)	0.281 [0.255–0.308]	0.007 [0.006–0.008]
Germany (WAHIS sub-district units, 1,409 units, $n_{test} \approx 155{,}000$, pos. rate=0.91%)		
No-skill baseline	0.500	0.009
HistGBM	0.642 [0.633–0.653]	0.025 [0.023–0.029]
Random Forest	0.641 [0.633–0.650]	0.024 [0.021–0.027]
Logistic Regression	**0.772** [0.758–0.785]	0.048 [0.042–0.054]
Volume-Only (LR)	0.771 [0.757–0.784]	**0.049** [0.043–0.056]

Brier scores omitted: all models converge to ≈pos. rate at $<2\%$ imbalance (see §7.3).

*Bootstrap DeLong ($B=1000$): FR LR vs. RF $p=0.024^{*}$; DE LR vs. Vol-Only $p=0.110$.

$D_{d,t}$ and its one-week lag $D_{d,t-1}$, and (3) farm fraction f_{farm}, yielding the full HSI. Each semantic feature is evaluated alongside its temporal lag to ensure a fair comparison with the volume baseline. All configurations use HistGBM with isotonic probability calibration.

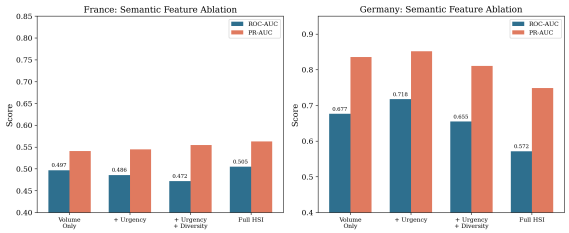

Figure 5: Incremental impact of LLM-derived semantic features at the country level. For Germany, adding the urgency score is the key driver (+4.1 percentage points). For France, no semantic feature provides consistent lift over the volume-only baseline.

The ablation results Figure 5) reveal an asymmetric pattern across the two countries. For France, adding semantic features does not improve over the volume-only baseline: ROC-AUC moves from 0.497 (volume only) to 0.486, 0.472, and 0.505 as features are added sequentially, with no configuration consistently exceeding the baseline. This suggests that the temporal volume signal already captures the dominant outbreak dynamics in French media, and that LLM-extracted semantics add noise rather than signal at the country level.

For Germany the results are different. Adding the urgency score $\hat{U}_{d,t}$ and its one-week lag lifts ROC-AUC from 0.677 to 0.718, a gain of 4.1 percentage points. This asymmetry is explained by media structure: when one signal category dominates (as `farm_outbreak` does in French media at 52.1%), urgency and diversity metrics carry little discriminative power because the distribution barely changes week to week; in Germany's more balanced coverage they become meaningful predictors.

Adding $D_{d,t}$ and f_{farm} to reach the full HSI shows *diminishing returns*: Germany's ROC-AUC drops from 0.718 back to 0.572, as the diversity and farm-fraction components introduce noise by conflating proactive surveillance with confirmed outbreak detections.

At the admin-division level, the extreme class imbalance (below 2% positive) compresses ablation differences across all configurations. Germany's full HSI still achieves the best admin-level ROC-AUC (0.643), but the gains are small.

The ablation above uses HistGBM throughout. We repeat the same ablation with calibrated Logistic Regression to verify model-specificity. For Germany, the LR-based ablation shows a +4.9 pp urgency lift (Volume-Only LR: 0.676 → Vol+Urgency: 0.725), comparable to the +4.1 pp observed for HistGBM. Urgency therefore carries genuine predictive value for both model classes in Germany, not just tree-based learners. For Germany's districts, LR performance is insensitive to features across all ablation configurations (≈0.771–0.773), confirming that the 0.772 vs. 0.642 advantage of LR over HistGBM at district scale is driven by the linear decision boundary, not by specific feature combinations. France's LR ablation replicates the Platt inversion at Volume-Only (0.297), which resolves once diversity and urgency

features are added (Full HSI: 0.625), consistent with the calibration-artefact explanation in §7.4.1.

7.6 Performance by Epidemic Phase

To assess whether model performance is stable across the epidemic cycle, we stratify predictions into three temporal phases: Peak (2022), Inter-wave (2023), and Post-peak (2024+). Phase stratification uses Logistic Regression (best country-level model) and HistGBM (best admin-level model). These phases reflect distinct media and epidemiological regimes: the 2022 peak saw dense, high-volume outbreak reporting; 2023 was an inter-wave period with sparser but still-present activity; and 2024+ marks the post-peak decline, characterised by lower base rates and reduced media coverage.

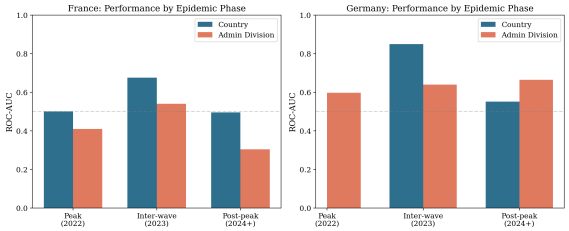

Figure 6: ROC-AUC by epidemic phase for country-level (LR) and admin-division-level (HistGBM) models. Inter-wave (2023) yields the strongest discrimination in both countries; the 2022 Peak fold for Germany is undefined because all weeks in that fold are positive.

Performance is not uniform across phases (Figure 6). The inter-wave period (2023) generally yields the strongest discrimination for both countries (France country: 0.676; Germany country: 0.849), likely because media signals and WAHIS events are better separated in time during lower-intensity periods. The Peak (2022) test fold for Germany contains only positive labels (all weeks were outbreak weeks), making ROC-AUC undefined for that cell. The post-peak phase (2024+) sees a further drop in positive rate, making the binary prediction task structurally harder. Epidemic phase is therefore a primary confounder that any honest evaluation of EBS performance must report.

8 Future Work

Several directions emerge from this work. The most pressing is spatial precision: admin-level performance collapses for France because media signals lack the geographic resolution to localize individual arrondissement outbreaks. Integrating ecological covariates such as wild bird migration telemetry, climate records, land-use features, and livestock trade flows is the most direct path forward; media captures *when*, ecology captures *where*. The signal taxonomy also needs refinement: farm-fraction currently conflates confirmed detections with precautionary culling orders, which suppresses Germany's ROC-AUC from 0.718 to 0.572; separating these event types would likely recover lost performance. Finally, the system currently runs retrospectively. Live deployment requires continuous ingestion and incremental LLM processing; no architectural blockers exist, but reliability under continuous load has not been tested. The pipeline already spans 28 nations via GDELT and supports alternative query taxonomies, making extension to other diseases, such as African Swine Fever, a low-friction next step.

9 Conclusion

This paper presented a multi-source, AI-augmented Event-Based Surveillance system for HPAI H5N1, built around a locally hosted `Qwen 2.5-14B-Instruct` model. Processing 10,193 multilingual articles from Google News and GDELT across France and Germany, the pipeline achieves a macro-average F1 of 0.75 against 400 manually annotated articles across six extraction fields, with species identification reaching F1 = 0.62 after multilingual synonym expansion.

Media-derived event dates align with physical outbreak onset to within a day (median 0 days), confirming that the LLM extracts accurate temporal signals rather than inventing them. The same detections appear 13–14 days before corresponding WAHIS notifications. At the country level, the HSI achieves ROC-AUC of 0.629 (France, LR) and 0.700 (Germany, Random Forest) after Platt-scaling calibration. LLM-derived urgency scores add +4.1 pp for Germany but provide no lift for France, confirming that semantic value depends on national media structure. At the admin-division scale, calibrated Logistic Regression reaches ROC-AUC of 0.772 for Germany's sub-districts, inverting the typical tree-model advantage, while France's arrondissement results fall below the random baseline. The major finding is that spatial granularity, epidemic base rate, and epidemic phase are the primary determinants of EBS performance. Any honest evaluation of media-based surveillance must account for all three, as model selection alone explains little of the variance in performance.

ACKNOWLEDGEMENTS

This work is part of the TALEDZ project, funded by ANR / France 2030, Université Marie et Louis Pasteur, CNRS, Institut FEMTO-ST (UMR 6174), F-90000 Belfort, France. Computations have been performed on the supercomputer facilities of the Mésocentre de calcul de Franche-Comté. The system architecture diagram (Figure 1) was produced with AI image generation tools using author-supplied content and specifications. All scientific content, experimental design, data analysis, and conclusions are entirely the authors' own. The code and data are available from the authors upon reasonable request.