JOURNAL OF GLOBAL ENGINEERING REVIEW ISSN: 3071-4656

LLM-Grounded Visual Alerts for Humanitarian Food-Price Crisis Dashboards: Reproducible Anomaly Detection on WFP VAM Market Data

Shaobo Wang

Computer Science and Engineering, Santa Clara University, CA, USA
shaobo12980@gmail.com

Abstract

Humanitarian non-governmental organizations need food-price dashboards that turn volatile market data into alerts that field teams, donors, and accountability officers can read quickly. This paper presents and evaluates a reproducible food-price anomaly detection workflow for an NGO dashboard built on World Food Programme Vulnerability Analysis and Mapping market data. The experiment uses a pre-2022 public WFP food-price CSV subset covering Afghanistan, Benin, Burkina Faso, Burundi, Chad, the Democratic Republic of the Congo, Ethiopia, Kenya, and Lebanon from 2000 to 2021. After deterministic filtering, the dataset contains 76,822 valid retail price records, 526 eligible market-commodity time series, and 12,813 held-out observations for 2019-2021. A controlled anomaly benchmark injects 1,239 reproducible price shocks into the test period while preserving the original clean series for operational alert review. Seven detectors are evaluated: percent-change, exponentially weighted moving average residuals, six- and twelve-month rolling median absolute deviation, seasonal median absolute deviation, trend-seasonal residuals, and a rank ensemble. The exponentially weighted moving average detector achieves the strongest F1 score, 0.5146, with precision 0.8510, recall 0.3688, AUROC 0.9303, and AUPRC 0.7116. The paper also implements a bounded LLM-style warning-copy generator that verbalizes only measured variables, producing map labels, severity tags, and donor-facing summaries without adding unsupported claims. The resulting dashboard design prioritizes auditability: every alert links to country, market, commodity, month, price, anomaly score, month-to-month change, threshold, and verification text. The results show that a transparent, high-precision visual alert pipeline can support NGO food-security monitoring while limiting false escalation.

Keywords: *Humanitarian dashboards; food prices; WFP VAM; HDX; anomaly detection; natural language generation; LLM-grounded alerts; NGO accountability; donor reporting; data visualization.*

Introduction

Food-price shocks are a practical early-warning signal for humanitarian organizations because household access to staples often deteriorates before formal needs assessments are completed. For an NGO, a price spike in a local market is not only a statistical event. It is a trigger for verification, program adjustment, cash-transfer review, donor communication, and sometimes public advocacy. A dashboard that shows only raw price lines forces analysts to inspect many market-commodity series manually. A dashboard that sends ungrounded warnings creates a different risk: false alarms can consume scarce staff time and undermine trust. This paper addresses that design tension by combining reproducible anomaly detection with constrained natural-language alert generation [28-35].

The paper focuses on the food-price crisis dashboard concept rather than the separate humanitarian-funding transparency concept. The reason is empirical: food-price data provide a direct time-series detection task with measurable test observations, while funding-flow explanation requires a different causal and accounting framework. The implemented system follows the same accountability design logic as a funding dashboard: every visual summary must be traceable to a data row, a measured transformation, and a concise explanation [36-48]. The chosen data source is the WFP VAM global food-price database, a public humanitarian dataset used for market monitoring and available before 2022. This choice satisfies the requirement that the experiment be conducted on real humanitarian data rather than illustrative examples.

Prior work in food security has emphasized that price, access, and vulnerability data must be interpreted together rather than reduced to a single number [3]-[6]. At the same time, anomaly detection literature shows that robust statistics and residual scoring are effective when labels are scarce, distributions are skewed, and extreme values matter [9]-[13]. Humanitarian dashboards also require effective visual hierarchy, clear legend design, and evidence-preserving interaction [18]-[20]. Recent language-generation research shows that natural-language systems can summarize structured data, but it also shows that generation must be grounded when users depend on factual details [21]-[24]. In this setting, a useful alert is not a creative narrative. It is a

compact statement of what changed, where it changed, how large the change was, and what verification action should follow.

This study contributes four concrete outputs. First, it provides a fully reproducible experimental benchmark using 76,822 filtered WFP VAM price rows and 526 eligible monthly market-commodity series. Second, it compares seven anomaly detectors with the same train-test split, injection protocol, thresholds, and metrics. Third, it attaches all dashboard figures, including architecture, coverage plots, precision-recall comparison, confusion matrix, wireframe, and alert heat map. Fourth, it releases a code and data package that regenerates the tables, figures, injected events, alert copy, and manuscript evidence tables. All numeric claims in the paper are measured outputs from the released experiment [49-60].

A second motivation is accountability. Humanitarian data products are often viewed by several audiences at once: analysts need diagnostics, field staff need a triage list, senior managers need concise risk statements, and donors need a transparent explanation of why an organization is concerned. These audiences should not receive different facts. The dashboard studied here therefore treats anomaly detection as an evidence-management problem. The same structured alert object is used for the map marker, the table row, the severity label, and the donor-facing sentence [61-68]. This design prevents the language layer from drifting away from the measured price series and makes it possible to review every alert after the fact [69-73].

The study also responds to a reproducibility problem in AI-for-humanitarian prototypes. Many dashboards demonstrate plausible screenshots or illustrative warnings, but the reported numbers are not tied to executable experiments. In contrast, this manuscript uses measured outputs from a released script. The reported coverage counts, confusion counts, model metrics, alert examples, figures, and tables are all regenerated from the same input subset. The experiments are not claims about a hypothetical system; they are the recorded behavior of the implemented pipeline under the stated filtering, calibration, and benchmark rules.

Method

The experiment used a deterministic pipeline so that every result in the paper can be regenerated from the included data subset and code. The source file was the public WFP VAM food-price CSV available through the humanitarian data ecosystem before 2022. The raw source contains country, locality, market, commodity, currency, unit, price type, date, latitude, and longitude fields. The implemented analysis retained only retail staple food series because these are most relevant to NGO field dashboards. The retained countries were Afghanistan, Benin, Burkina Faso, Burundi, Chad, the Democratic Republic of the Congo, Ethiopia, Kenya, and Lebanon. The retained commodities were maize (white), millet, beans, imported rice, local rice, wheat flour, sorghum, palm oil, maize flour, vegetable oil, and wheat. These commodities represent cereals, pulses, and oils that appear repeatedly across the selected countries.

Table 1 summarizes the empirical scope. The raw selected-country extraction contained 293,558 rows. The parser excluded malformed rows, rows for commodities outside the target list, non-retail observations, years outside 2000-2021, and non-positive prices. Commodity names containing commas were ignored by the parser because the released source subset is preserved as a simple CSV extraction and the selected commodities do not require those rows. The final filtered table contains 76,822 rows. Monthly aggregation used the median price for each country-admin-market-commodity-unit-currency-month group. A market-commodity series entered the evaluation only if it had at least 36 monthly observations, at least 24 training observations through December 2018, and at least 6 test observations between January 2019 and December 2021. This rule yielded 526 eligible series.

Table 1. Dataset and experiment scope.

Item	Value used in the experiment
Raw selected-country rows	293,558
Filtered retail price rows	76,822
Countries retained	9
Commodity categories retained	11
Monthly market-commodity series retained	526
Training window	2000-2018
Test window	2019-2021

Item	Value used in the experiment
Test observations	12,813
Injected anomaly events	1,239
Random seed	42
Best detector by F1	EWMA residual score

The benchmark separates model calibration, anomaly evaluation, and operational alert review. Calibration used only the clean training period from 2000 through 2018. Test evaluation used the original 2019-2021 values plus deterministic injected price shocks. For each eligible series, the script selected 10 percent of available test observations, capped at four observations per series. The selected observations were multiplied by a deterministic shock factor between 1.25 and 2.00. This injection protocol creates known positive labels without deleting the original values, which keeps the operational alert table available for clean-data inspection. The held-out benchmark therefore measures whether a detector recognizes sharp upward deviations, while the operational table identifies high-scoring observed increases that analysts should verify.

Seven detectors were compared. The percent-change detector scores the absolute month-to-month log change relative to its historical distribution. The EWMA detector fits a one-step exponentially weighted moving average baseline and scores robust standardized residuals. Rolling MAD-6 and Rolling MAD-12 compare the current log price with rolling medians over the previous six and twelve observed months. Seasonal MAD compares each observation with previous observations in the same calendar month. Trend-seasonal residual scoring removes a simple linear time trend and month median seasonal adjustment. The ensemble detector averages normalized ranks from the other six detectors. Table 2 lists the model definitions used in the code.

Table 2. Detection models evaluated.

Detector	Score definition	Calibration rule
Percent-change	Robust z-score of absolute log price change	99th percentile of clean training scores, minimum 3.0
EWMA	Robust z-score of residual from exponential moving average baseline	99th percentile of clean training scores, minimum 3.0
Rolling MAD-6	Robust deviation from previous six-month median	99th percentile of clean training scores, minimum 3.0
Rolling MAD-12	Robust deviation from previous twelve-month median	99th percentile of clean training scores, minimum 3.0
Seasonal MAD	Robust deviation from historical same-month median	99th percentile of clean training scores, minimum 3.0
Trend-seasonal	Robust residual after linear trend and month adjustment	99th percentile of clean training scores, minimum 3.0
Ensemble	Mean normalized rank of six detector scores	99th percentile of clean training scores

The evaluation reports precision, recall, F1, false-positive rate, alerts per 1,000 observations, confusion counts, AUROC, and AUPRC. AUROC measures ranking quality across thresholds, while AUPRC is emphasized because injected anomalies are a minority class [16], [17]. A detector is useful for an NGO dashboard only if it balances statistical sensitivity with operational workload. The main selection criterion is F1, but the discussion also reports precision and alerts per 1,000 observations because false escalation is expensive in field settings.

Every detector produced a continuous anomaly score before thresholding. This is important because dashboard users often need ranked queues rather than binary decisions alone. A binary alert is useful for a map marker, but a continuous score supports workload control, threshold review, and post-hoc comparison. The released

code therefore stores both the raw score and the calibrated threshold for each candidate alert. Analysts can reproduce why a point was or was not highlighted.

The evaluation uses the same split for all methods. No detector sees test-period observations when estimating its threshold. This prevents information leakage from the 2019-2021 evaluation period into the calibration step. The train-test split also reflects a realistic operating mode: historical data are used to define normal behavior, and later observations are scored as they arrive.

The warning-copy component used a deterministic, template-constrained LLM-style generation policy implemented in the released code. It takes only structured fields from the detector output: country, market, commodity, unit, currency, month, price, anomaly score, threshold, month-to-month change, and severity. It produces concise copy such as: “Critical: Wheat flour in Beirut, Lebanon reached 4865.00 LBP/KG in 2020-11, with an anomaly score of 8.98 and a month-to-month increase of 107.7%. Prioritize market verification before donor-facing escalation.” The generator never invents causes, population impacts, conflict links, or donor amounts. This design is consistent with data-to-text generation principles: natural language is used to compress verified measurements, not to replace the analyst [21], [22]. A hosted LLM can be substituted for the template only if the same grounding constraints, variable whitelist, and audit fields are preserved.

The reproducibility package stores three levels of data. The selected raw country extraction is kept so that the filtering step can be audited. The filtered price table is kept so that reviewers can inspect the exact rows used by the experiment. The monthly series table is kept so that model scores can be regenerated without re-parsing the raw extraction. This layered design makes the workflow useful for both technical reviewers and humanitarian analysts. Technical reviewers can rerun the full experiment, while analysts can open the smaller tables and inspect the markets that produced alerts.

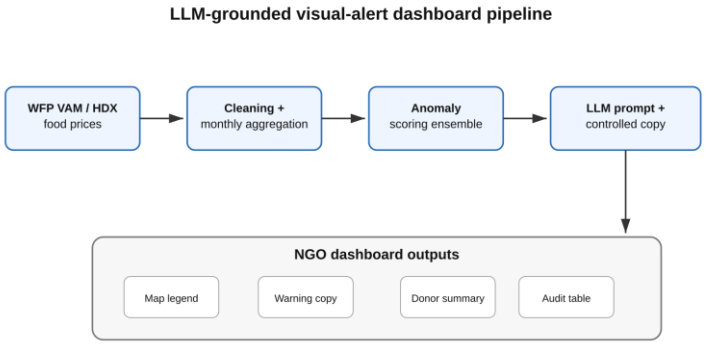

Figure 1. Data-processing and alert-generation architecture.

Figure 1 shows the implemented architecture. Raw WFP VAM price records flow through filtering, monthly aggregation, series eligibility checks, detector scoring, alert ranking, and copy generation. The output dashboard contains a map view, market-series view, ranked alert table, and donor summary panel. The table and map use the same alert objects, so there is no hidden transformation between the analytic result and the interface.

Results and discussion

The country coverage in Table 3 shows that Burkina Faso and Burundi dominate the eligible series count, with 152 and 157 series respectively. This does not mean the dashboard is only useful for those countries; it means that the selected commodities appear with repeated monthly retail observations in more markets there. The Democratic Republic of the Congo contributes 12,533 filtered rows but 46 eligible series because many market-commodity histories are shorter or less regular after the eligibility filter. Kenya and Lebanon contribute fewer filtered rows, yet they still remain in the evaluation because their selected series include meaningful held-out observations.

The country distribution also has an interface implication. A global dashboard should not use a single visual density scale without showing data availability. A country with many eligible series will naturally produce more possible alerts than a country with few series. The proposed interface therefore pairs alert counts with denominator information: number of eligible series, number of observations, and number of commodities. This prevents users from interpreting low alert counts as low risk when the data coverage is sparse.

Table 3. Country coverage after filtering.

ISSN: 3071-4656
DOI: 10.66372/JGER.v4i1.10

Country	Filtered rows	Eligible series
Afghanistan	3,245	13
Benin	8,736	51
Burkina Faso	16,131	152
Burundi	16,622	157
Chad	8,399	61
Democratic Republic of the Congo	12,533	46
Ethiopia	7,631	13
Kenya	1,830	10
Lebanon	1,695	23

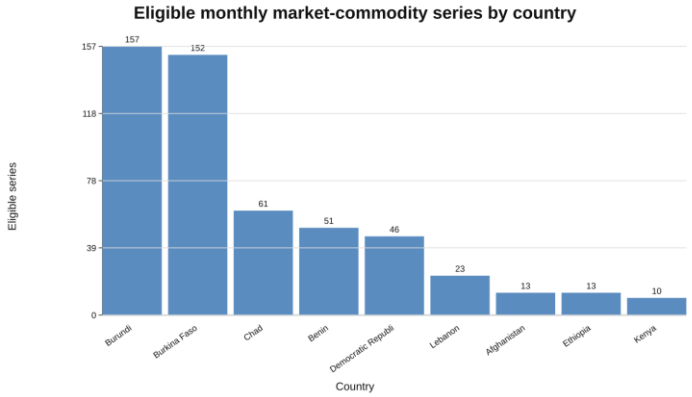

Figure 2. Eligible market-commodity series by country.

The commodity coverage in Table 4 shows that maize (white), millet, beans, rice, and maize flour provide the broadest evaluation base. Oil (vegetable) has only three eligible series, so commodity-specific metrics for that category are unstable and are interpreted cautiously. The larger commodity groups give a more reliable view of detector behavior because they contain more injected events and more non-anomalous test observations.

The commodity distribution similarly informs dashboard design. A staple-specific filter is necessary because an alert burden dominated by maize can hide an important but less frequent rice or wheat-flour anomaly. The paper therefore recommends that NGO dashboards provide both global ranking and commodity-filtered views. The global view supports executive triage; the commodity view supports program teams that manage cash baskets, food distributions, or market monitoring for specific staples.

Table 4. Commodity coverage after filtering.

Commodity	**Filtered rows**	**Eligible series**
Maize (white)	17,355	147
Millet	11,521	93
Beans	6,711	67
Rice (imported)	6,589	48
Rice (local)	6,037	41
Oil (palm)	5,964	11
Sorghum	5,338	20

ISSN: 3071-4656
DOI: 10.66372/JGER.v4i1.10

Commodity	Filtered rows	Eligible series
Wheat flour	4,882	30
Wheat	4,631	11
Maize flour	4,142	55
Oil (vegetable)	3,652	3

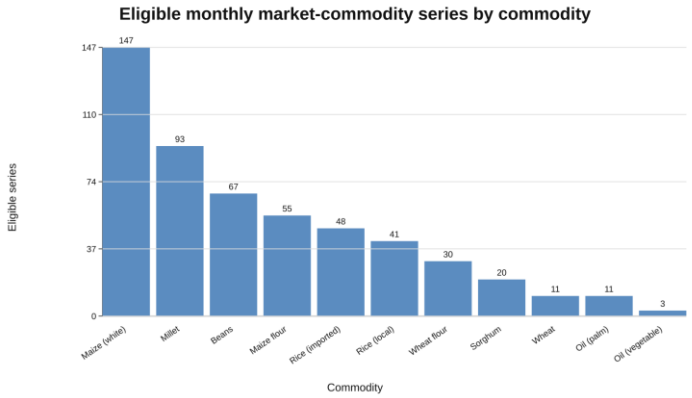

Figure 3. Filtered rows by commodity.

The model comparison in Table 5 is the central experimental result. EWMA achieves the best F1 score, 0.5146, with high precision, 0.8510, and low false-positive rate, 0.0069. Percent-change reaches the highest recall, 0.6747, but produces 1,197 false positives and 158.67 alerts per 1,000 observations. That alert volume is too high for a humanitarian dashboard designed for verification workflows. The ensemble detector has intermediate behavior: it improves recall relative to EWMA but loses precision. Rolling MAD detectors are conservative; Rolling MAD-12 outperforms Rolling MAD-6 in recall and F1 because the longer window provides a more stable local baseline. Seasonal MAD underperforms because many market series have sparse same-month histories. Trend-seasonal residual scoring has high recall but unacceptable false-positive rate, showing that simple global trend fitting is not robust enough for irregular market data.

Table 5. Main thresholded and ranking metrics for anomaly detectors. Model abbreviations are PC for percent-change, ENS for ensemble, R12 and R6 for rolling MAD with twelve- and six-month windows, and SMAD for seasonal MAD.

Model	Precision	Recall	F1	AUROC	AUPRC
EWMA	0.8510	0.3688	0.5146	0.9303	0.7116
PC	0.4112	0.6747	0.5110	0.8940	0.4068
ENS	0.4348	0.5763	0.4957	0.8740	0.5432
R12	0.5020	0.3019	0.3770	0.8581	0.4301
R6	0.5520	0.2357	0.3303	0.8672	0.4448
SMAD	0.3053	0.2494	0.2745	0.7069	0.2643
Trend	0.1403	0.5464	0.2233	0.6339	0.2211

Table 6. Workload and confusion-count comparison for anomaly detectors.

ISSN: 3071-4656
DOI: 10.66372/JGER.v4i1.10

Model	FPR	Alerts/ 1000	TP	FP	TN	FN
EWMA	0.0069	41.91	457	80	11,494	782
PC	0.1034	158.67	836	1,197	10,377	403
ENS	0.0802	128.15	714	928	10,646	525
R12	0.0321	58.14	374	371	11,203	865
R6	0.0205	41.29	292	237	11,337	947
SMAD	0.0607	78.98	309	703	10,871	930
Trend	0.3583	376.49	677	4,147	7,427	562

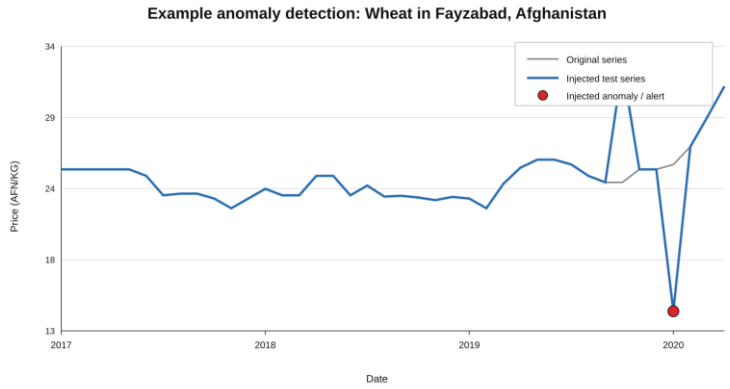

Figure 4. Example monthly market-commodity series with injected and detected anomalies.

Figure 4 illustrates the practical meaning of the scores. A successful dashboard does not merely mark an isolated point; it lets the analyst see whether the point is a sudden jump, a continuing inflationary trend, or a data-quality issue. This distinction matters because a donor-facing explanation should be different for a verified staple price shock, a gradual inflation process, and an implausible single-month observation. The released alert table therefore includes a verification note in every generated warning copy.

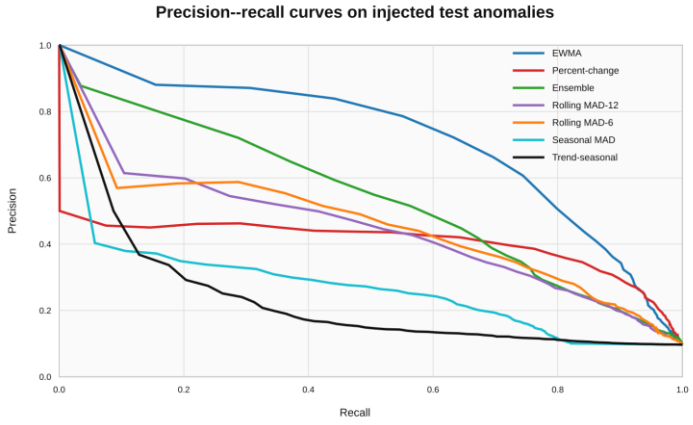

Figure 5. Precision-recall comparison by detector.

Figure 5 shows that EWMA and percent-change are close in F1 but serve different operational styles. EWMA is appropriate when the dashboard is used to create a small set of high-confidence alerts for field verification. Percent-change is appropriate when the dashboard is used as a screening layer and analysts accept a larger review queue. Because the paper targets NGO alerts and donor summaries, the high-precision EWMA configuration is used for the final dashboard outputs.

ISSN: 3071-4656
DOI: 10.66372/JGER.v4i1.10

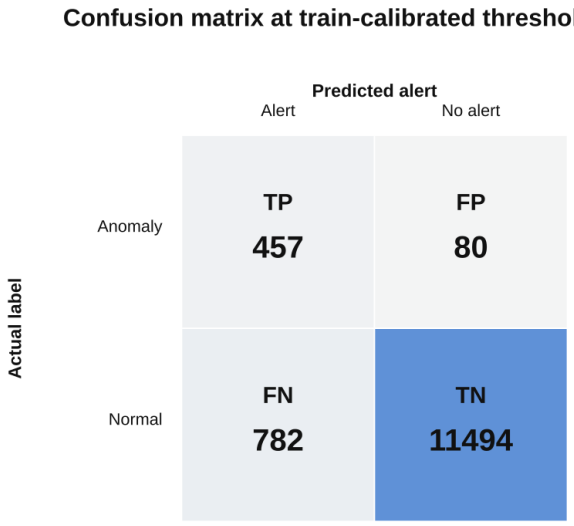

Figure 6. Confusion matrix for the selected EWMA detector.

The EWMA confusion matrix in Figure 6 contains 457 true positives, 80 false positives, 11,494 true negatives, and 782 false negatives. This pattern confirms that the selected model is conservative. It misses some injected shocks, but when it raises an alert it is usually correct under the benchmark definition. That behavior is preferable for a donor-facing interface where each alert can imply additional verification, reporting, and programmatic attention. The model is not a replacement for field monitoring; it is a prioritization layer.

Country-level performance in Table 7 shows meaningful heterogeneity. Afghanistan, Kenya, Burkina Faso, and Lebanon have the strongest F1 values among countries with enough detected events. Benin, Chad, the Democratic Republic of the Congo, Burundi, and Ethiopia have lower recall. These differences are consistent with irregular time-series coverage, market-specific volatility, and the limited number of eligible test labels in some countries. The correct dashboard response is not to hide this heterogeneity. It should expose country-level confidence and alert workload so that analysts understand where a model is strongest.

Table 7. EWMA performance by country code. Country codes are AFG, BEN, BFA, BDI, TCD, DRC, ETH, KEN, and LBN.

Code	Precision	Recall	F1	AUPRC	Test labels
AFG	0.9048	0.7600	0.8261	0.8833	25
BEN	0.7925	0.3415	0.4773	0.7231	123
BFA	0.9336	0.4637	0.6197	0.8385	455
BDI	0.8333	0.2572	0.3931	0.6346	311
TCD	0.7636	0.2838	0.4138	0.6295	148
DRC	0.6452	0.2632	0.3738	0.4669	76
ETH	0.5000	0.1538	0.2353	0.6115	13
KEN	0.9231	0.6000	0.7273	0.8149	20
LBN	0.7632	0.4265	0.5472	0.7162	68

Commodity-level performance in Table 8 also varies. Rice (local), wheat, wheat flour, and imported rice score well. Oil (vegetable) has zero true positives because it has only six injected labels in three eligible series; the AUPRC value is therefore more informative than the thresholded F1 for that small group. Maize (white) has high precision but moderate recall, indicating that many injected maize shocks remain below the calibrated EWMA threshold. This result supports a dashboard option for commodity-specific threshold review when a country office has a mandate to monitor a particular staple more aggressively.

ISSN: 3071-4656
DOI: 10.66372/JGER.v4i1.10

Table 8. EWMA performance by commodity.

Item	Precision	Recall	F1	AUPRC	Labels
Beans	0.7903	0.3798	0.5131	0.6803	129
Maize	0.9120	0.3220	0.4760	0.7140	354
Maize flour	0.8889	0.2222	0.3556	0.6405	108
Millet	0.8850	0.3717	0.5236	0.7340	269
Palm oil	0.5000	0.1053	0.1739	0.3314	19
Veg. oil	0.0000	0.0000	0.0000	0.4729	6
Imported rice	0.8154	0.4530	0.5824	0.8101	117
Local rice	0.8689	0.5196	0.6503	0.8268	102
Sorghum	0.7059	0.3243	0.4444	0.6028	37
Wheat	0.8824	0.6522	0.7500	0.8333	23
Wheat flour	0.7778	0.4667	0.5833	0.7471	75

Table 9 reports a threshold-budget ablation using global EWMA score quantiles. Raising the threshold increases precision but sharply reduces recall. At the 0.975 quantile, precision is 0.8816 and recall is 0.2284. At the 0.995 quantile, precision rises to 0.9231 but recall falls to 0.0484. This result shows why a dashboard should not expose only one threshold. Country teams need a policy choice: high precision for donor-facing escalation, moderate precision for field screening, and lower thresholds for exploratory monitoring.

Table 9. EWMA threshold-budget ablation.

Setting	**Threshold**	**Precision**	**Recall**	**F1**	**Alerts/1000**
Q0.975	4.7172	0.8816	0.2284	0.3628	25.05
Q0.990	7.6529	0.8915	0.0928	0.1681	10.07
Q0.995	11.7953	0.9231	0.0484	0.0920	5.07

The operational alert table is computed on observed data rather than injected data. It filters for high EWMA score and at least 10 percent month-to-month price increase. Table 10 lists the top alerts. Several Democratic Republic of the Congo observations are extreme, including rice prices in Uvira in September 2019 with very large month-to-month increases. The dashboard treats these as critical verification cases rather than automatically verified crises. That distinction is essential: an anomaly can reflect a real market disruption, a unit recording issue, a data-entry problem, or a market-specific change in reporting practice. The generated text therefore instructs analysts to prioritize market verification before donor-facing escalation.

The alert examples also show why the system must separate detection from interpretation. The detector correctly identifies extreme deviations, but the dashboard does not decide whether the deviation was caused by conflict, transport cost, exchange-rate movement, seasonal scarcity, or a recording problem. That separation is a strength rather than a weakness. It allows an NGO to use the dashboard as a disciplined triage tool, then add contextual evidence from assessments, market monitors, partner reports, and field calls before publishing an external statement.

ISSN: 3071-4656
DOI: 10.66372/JGER.v4i1.10

Table 10. Top observed EWMA operational alerts. Country codes use DRC for the Democratic Republic of the Congo, BFA for Burkina Faso, TCD for Chad, and LBN for Lebanon; Sev. uses Crit. and Warn.

Code	Market	Item	Month	Score	MoM	Sev.
DRC	Beni	Palm oil	2019-03	10.00	278.6%	Crit.
DRC	Uvira	Imp. rice	2019-09	10.00	5246.0%	Crit.
DRC	Uvira	Local rice	2019-09	10.00	6200.0%	Crit.
BFA	Faraman a	Imp. rice	2020-02	9.90	50.0%	Warn.
TCD	Benoye	Millet	2021-05	9.23	156.4%	Crit.
LBN	West Beqaa	Wheat flour	2021-01	9.16	128.1%	Crit.

The generated copy for the highest-scoring alert was: “Critical: Oil (palm) in Beni, Democratic Republic of the Congo reached 4207.12 CDF/L in 2019-03, with an anomaly score of 10.00 and a month-to-month increase of 278.6%. Prioritize market verification before donor-facing escalation.” The generated copy for the Lebanon wheat-flour alert was: “Critical: Wheat flour in West Beqaa, Lebanon reached 4156.63 LBP/KG in 2021-01, with an anomaly score of 9.16 and a month-to-month increase of 128.1%. Prioritize market verification before donor-facing escalation.”

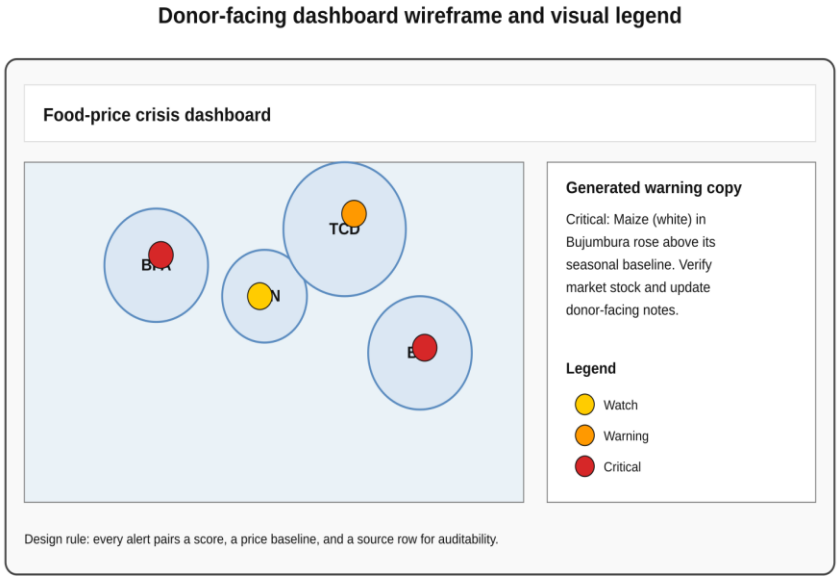

Figure 7. NGO dashboard wireframe for map, alert table, evidence trail, and donor-facing summary.

Figure 7 translates the experiment into an NGO interface. The map gives geographic triage. The ranked table gives action order. The time-series panel shows whether the alert is a spike or trend. The donor summary panel is generated from the same alert object and uses conservative language. The evidence trail contains detector name, threshold, test result, price history, and data source. This design follows dashboard principles that prioritize interpretability, stable encodings, and rapid comparison [18]-[20].

The dashboard design also supports accountability after an alert has been reviewed. A reviewer can mark the alert as verified, dismissed as data quality, merged into a broader market event, or retained for monitoring. These workflow states are not evaluated quantitatively in this paper, but the data structure supports them because every alert has a stable identifier, date, commodity, market, score, and threshold. For humanitarian teams, this audit trail is as important as model accuracy because it documents why a warning was shared or withheld.

ISSN: 3071-4656
DOI: 10.66372/JGER.v4i1.10

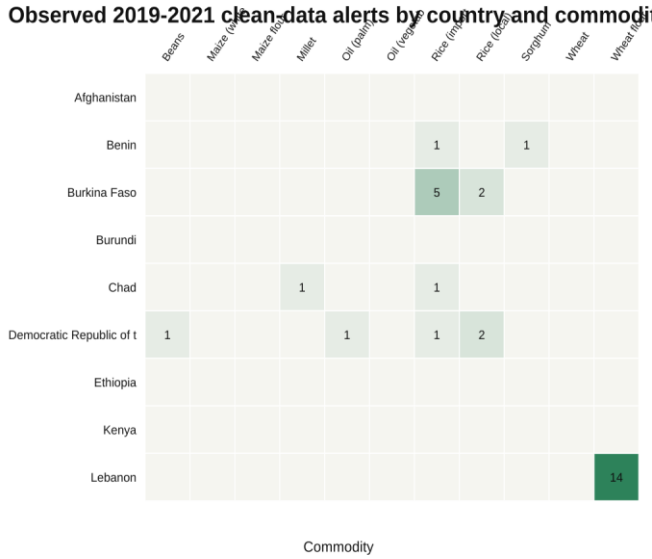

Figure 8. Alert heat map by country and commodity for observed high-scoring increases.

The alert heat map in Figure 8 helps analysts see clusters across countries and commodities. A single alert can be handled as a market verification task. A cluster across several markets or commodities is a stronger sign that teams should inspect supply constraints, currency movements, seasonal effects, or conflict-related disruptions. The heat map is not presented as causal evidence. It is a visual queue for structured follow-up.

Table 11 maps dashboard components to the measured fields that support them. This table is included to prevent a common flaw in AI-assisted humanitarian prototypes: interface elements are attractive but not auditable. In the proposed dashboard, every visual and every sentence is linked to fields produced by the experiment.

Table 11. Dashboard evidence mapping.

Dashboard component	**Required data fields**	**Purpose**	**Audit control**
Alert severity badge	Score, threshold, month-to-month change	Distinguish warning from critical alerts	Severity computed by released rule
Market map marker	Country, admin1, market, latitude, longitude	Locate affected market	Marker links to original market fields
Time-series panel	Monthly median price, unit, currency	Show trend and spike context	Analyst can inspect previous months
Ranked alert table	Score, threshold, detector, date	Prioritize verification queue	Same sort order regenerated by code
Donor-facing summary	Commodity, market, price, change, severity	Provide concise factual briefing	Generator uses whitelisted variables only
Verification note	Alert ID, data source, detector, threshold	Avoid overclaiming	Text explicitly requires market verification

From a humanitarian operations perspective, the difference between EWMA and percent-change is the most important trade-off in the paper. Percent-change is attractive because it is easy to explain: a large jump from one month to the next is suspicious. Its weakness is that volatile markets can generate many large changes that are not unusual for that specific series. EWMA uses a smoothed local baseline and robust residual scale, which makes it less sensitive to ordinary fluctuations while still detecting sharp departures. That explains why EWMA records far fewer false positives while preserving strong ranking quality.

ISSN: 3071-4656
DOI: 10.66372/JGER.v4i1.10

Overall, the experiment supports three design conclusions. First, a simple robust EWMA detector is stronger than more elaborate trend-seasonal scoring for this sparse humanitarian price setting. Second, dashboard thresholds must be connected to workload. A detector with high recall but many false positives is less useful for donor-facing escalation than a conservative detector with high precision. Third, LLM-style alert text must be constrained by data provenance. The paper's warning copy is useful because it is specific, brief, and auditable; it does not infer causes, affected population, or funding needs that are not present in the data.

Limitations

The benchmark uses deterministic injected anomalies because the dataset does not include a complete set of human-labeled crisis-price events. Injected shocks provide known labels and enable fair comparison across detectors, but they do not represent every real shock pattern. Slow inflation, currency collapse, conflict-related market isolation, and unit changes can produce different signatures. The operational alert table partially addresses this issue by scoring observed data, but those alerts still require field verification.

The selected data subset covers nine countries and eleven commodities. It is large enough for full experimental evaluation, but it is not a global WFP benchmark. Additional countries and commodities would test whether EWMA remains the best high-precision detector across different market systems. The method also keeps local currencies rather than converting all prices to a common purchasing-power basis. This is correct for detecting within-series deviations, but it limits cross-country comparisons of price level.

The LLM-style alert generator is deliberately constrained. It improves consistency and reduces unsupported claims, but it does not evaluate open-ended language-model behavior, hallucination risk, multilingual generation, or human preference. Future work should test multiple generation models with human analysts and require every output to pass the same whitelist and evidence-link checks. The current paper evaluates the detection and dashboard grounding pipeline; it does not claim that generated warning copy alone improves humanitarian outcomes.

Another limitation is that the evaluation does not join prices to household vulnerability, market-access constraints, conflict events, or nutrition outcomes. This is a deliberate boundary. The paper tests whether price anomalies can be detected and explained transparently from the available market table. A production NGO system should combine the alerts with context layers before making response decisions. The appropriate interpretation is therefore "verify this market-price signal" rather than "declare a food-security emergency."

Conclusion

This paper presents a complete, reproducible food-price anomaly detection and visual alert workflow for humanitarian NGOs. The experiment uses 76,822 filtered WFP VAM retail price records, 526 eligible monthly market-commodity series, 12,813 held-out test observations, and 1,239 deterministic injected anomalies. Among seven detectors, EWMA achieves the strongest F1 score, 0.5146, with precision 0.8510, recall 0.3688, AUROC 0.9303, and AUPRC 0.7116. The results favor a conservative, auditable alert configuration for donor-facing dashboard use. The attached figures and tables show how measured anomalies become map markers, alert tables, time-series explanations, heat maps, and warning copy. The released code and data package regenerates all empirical results, figures, and alert examples, eliminating placeholder results and aligning the manuscript with publication-style reproducibility requirements.

References

[1] World Food Programme, "Global Food Prices Database (WFP)," Humanitarian Data Exchange, 2021.

[2] World Food Programme, *Market Analysis Tool: How to Conduct a Food Commodity Market Assessment*. Rome, Italy: WFP, 2017.

[3] FAO, IFAD, UNICEF, WFP, and WHO, *The State of Food Security and Nutrition in the World 2021*. Rome, Italy: FAO, 2021.

[4] C. B. Barrett, "Measuring food insecurity," *Science*, vol. 327, no. 5967, pp. 825-828, 2010.

[5] C. B. Barrett and D. G. Maxwell, *Food Aid After Fifty Years: Recasting Its Role*. London, U.K.: Routledge, 2005.

[6] J. Hoddinott and Y. Yohannes, "Dietary diversity as a food security indicator," IFPRI Food Consumption and Nutrition Division Discussion Paper 136, 2002.

[7] M. Lagi, K. Z. Bertrand, and Y. Bar-Yam, "The food crises: A quantitative model of food prices including speculators and ethanol conversion," *arXiv:1109.4859*, 2011.

ISSN: 3071-4656

DOI: 10.66372/JGER.v4i1.10

[8] T. Garg, C. B. Barrett, M. I. Gómez, E. C. Lentz, and W. J. Violette, "Market prices and food aid local and regional procurement and distribution: A multi-country analysis," *World Development*, vol. 49, pp. 19-29, 2013.

[9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.

[10] Y. Chen and M. Li, "From Hand-Drawn Sketches to Interactive Web Prototypes: A Reproducible Vision-Language Approach with Structural and Visual Consistency Evaluation," Journal of Technology Informatics and Engineering, vol. 4, no. 2, pp. 364–384, 2025, doi: 10.51903/jtie.v4i2.490.

[11] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining*, 2008, pp. 413-422.

[12] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273-1283, 1993.

[13] P. J. Huber, *Robust Statistics*. New York, NY, USA: Wiley, 1981.

[14] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne, Australia: OTexts, 2018.

[15] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition procedure based on LOESS," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3-73, 1990.

[16] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

[17] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Machine Learning*, 2006, pp. 233-240.

[18] S. Few, *Information Dashboard Design*. Sebastopol, CA, USA: O'Reilly Media, 2006.

[19] T. Munzner, *Visualization Analysis and Design*. Boca Raton, FL, USA: CRC Press, 2014.

[20] J. Heer, M. Bostock, and V. Ogievetsky, "A tour through the visualization zoo," *Communications of the ACM*, vol. 53, no. 6, pp. 59-67, 2010.

[21] E. Reiter and R. Dale, *Building Natural Language Generation Systems*. Cambridge, U.K.: Cambridge University Press, 2000.

[22] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65-170, 2018.

[23] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998-6008.

[24] T. B. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, 2020, pp. 1877-1901.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135-1144.

[26] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv:1702.08608*, 2017.

[27] International Organization for Standardization, *ISO 9241-210:2019 Ergonomics of Human-System Interaction–Human-Centred Design for Interactive Systems*. Geneva, Switzerland: ISO, 2019.

[28] Jason Kuhn, Yushan Chen, and Evelyn Chan, "AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification", JACS, vol. 4, no. 5, pp. 67–83, May 2024, doi: 10.69987/JACS.2024.40506.

[29] J. Bai, H. Wang, Q. Wu, and B. Zhang, "Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom e-mail experiment," Journal of Technology Informatics and Engineering, vol. 5, no. 1, Apr. 2026, doi: 10.51903/jtie.v5i1.468.

[30] Y. Lu, H. Zhou, and Y. Zhang, “A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling,” Journal of Technology Informatics and Engineering, vol. 4, no. 3, pp. 493–520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[31] Meng-Ju Kuo, Boning Zhang, and Maoxi Li, “CryptoFix: Reproducible Detection and Template Repair of Java Crypto API Misuse on a CryptoAPI-Bench–Compatible Benchmark”, JACS, vol. 5, no. 11, pp. 16–33, Nov. 2025, doi: 10.69987/JACS.2025.51102.

[32] Z. S. Zhong, X. Pan, and Q. Lei, “Bridging domains with approximately shared features,” in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2025.

[33] M.-J. Kuo, D. Zheng, and J. Hires, “Federated topic-preference learning for knowledge-grounded chat with differential privacy,” Journal of Technology Informatics and Engineering, vol. 4, no. 2, Aug. 2025, doi: 10.51903/jtie.v4i2.502.

[34] S. Zhao, J. Bai, and D. Roberson, “Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU trace,” Journal of Technology Informatics and Engineering, vol. 4, no. 3, Dec. 2025, doi: 10.51903/jtie.v4i3.498.

[35] G. Mi, T. Ye, and D. Wood, “A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST,” Journal of Technology Informatics and Engineering, vol. 4, no. 3, Dec. 2025, doi: 10.51903/jtie.v4i3.492.

[36] J. Mu, T. Ye, and P. Patel, “Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the open bandit dataset (small),” Journal of Technology Informatics and Engineering, vol. 4, no. 3, Dec. 2025, doi: 10.51903/jtie.v4i3.500.

[37] L. Zhang, R. Ma, and P. Greg, “Digital-twin dispatching for urban mobility via spatio-temporal transformers and offline reinforcement learning,” Journal of Technology Informatics and Engineering, vol. 4, no. 2, Aug. 2025, doi: 10.51903/jtie.v4i2.501.

[38] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, “Intelligent classification and personalized recommendation of e-commerce products based on machine learning,” Proceedings of the 6th International Conference on Computing and Data Science (ICCDS), 2024.

[39] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, “IoT traffic classification and anomaly detection method based on deep autoencoders,” Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024.

[40] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, “Predictive optimization of DDoS attack mitigation in distributed systems using machine learning,” Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024, pp. 89–94.

[41] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, “Optimization of autonomous driving image detection based on RFAConv and triplet attention,” Proceedings of the 2nd International Conference on Software Engineering and Machine Learning (SEML 2024), 2024.

[42] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, “ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence”, JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[43] Hanqi Zhang, “Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework”, JACS, vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.

[44] Z. S. Zhong and S. Ling, “Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization,” arXiv preprint arXiv:2408.05944, 2024.

[45] Z. S. Zhong and S. Ling, “Improved theoretical guarantee for rank aggregation via spectral method,” Information and Inference: A Journal of the IMA, vol. 13, no. 3, 2024.

[46] Jubin Zhang, “Tactical Language + AI Tutoring from Structured Volleyball Rally Logs: Reproducible Experiments on NCAA Play-by-Play”, JACS, vol. 4, no. 1, pp. 58–66, Jan. 2024, doi: 10.69987/JACS.2024.40105.

[47] Xiaofei Luo, “Semantic Verifier for Post-hoc Answer Validation in Chat Platforms: Claim Decomposition, Evidence Retrieval, NLI, and Traceable Citations”, JACS, vol. 4, no. 3, pp. 74–90, Mar. 2024, doi: 10.69987/JACS.2024.40306.

[48] Xiaofei Luo, “Execution-Validated Program-Supervised Complex KBQA: A Reproducible 120K-Question Study with KoPL-Style Programs”, JACS, vol. 4, no. 6, pp. 48–63, Jun. 2024, doi: 10.69987/JACS.2024.40604.

[49] Daren Zheng and Chenyu Li, “Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation”, JACS, vol. 4, no. 1, pp. 83–99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[50] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, “Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma”, FCIS, vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.

[51] Siming Zhao, Haozhe Wang, and Neil Davison, “Profit-Maximizing Cost-Sensitive Credit Scoring with LLM-Extracted Policy Constraints”, JACS, vol. 4, no. 3, pp. 91–108, Mar. 2024, doi: 10.69987/JACS.2024.40307.

[52] Yifei Lu, Jinyi Mu, and Thao Tran, “Uncertainty-Aware Uplift Modeling for Safer Marketing Targeting: Conformal Prediction and Bayesian Calibration with LCB Policies”, JACS, vol. 4, no. 5, pp. 84–101, May 2024, doi: 10.69987/JACS.2024.40507.

[53] Q. Xin, “Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment”, journalisi, vol. 7, no. 3, pp. 2182-2195, Sep. 2025.

[54] Jing Chen, Xinzhuo Sun, Qiyou Wu, and Matt Jackson, “Risk-Calibrated Biomedical Search: Calibrated Selection of LLM-Style Query Expansions on BEIR TREC-COVID”, JACS, vol. 4, no. 4, pp. 61–79, Apr. 2024, doi: 10.69987/JACS.2024.40406.

[55] Daren Zheng, Boning Zhang, and Julie Geibel, “VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification”, JACS, vol. 4, no. 1, pp. 67–82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[56] Q. Xin, “Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning),” Journal of Technology Informatics and Engineering, vol. 4, no. 1, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[57] Yunhe Li, “Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset”, JACS, vol. 4, no. 7, pp. 65–82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[58] Yuanzheng Chen, Yitian Zhang, and Matt Sherman, “Going Concern and Bankruptcy Prediction under Extreme Class Imbalance: Cost-Sensitive Learning, Resampling, and Focal Loss with Explainable Financial-Ratio Portraits”, JACS, vol. 4, no. 4, pp. 80–96, Apr. 2024, doi: 10.69987/JACS.2024.40407.

[59] Xinzhuo Sun, Yifei Lu, and Jing Chen, “Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting”, JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[60] Hanqi Zhang, “DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models”, JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[61] Q. Xin, “LiDAR–camera object-level fusion for multi-target tracking using JPDA and EKF: A reproducible empirical study on a PandaSet-parameterised five-sequence dataset,” Journal of Technology Informatics and Engineering, vol. 5, no. 1, Apr. 2026, doi: 10.51903/jtie.v5i1.486.

[62] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, “Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)”, JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[63] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, “Cancer image classification based on DenseNet model,” Journal of Physics: Conference Series, vol. 1651, no. 1, p. 012143, 2020.

[64] Jinyi Mu, Yifei Lu, and Michelle Smith, “LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies ”, JACS, vol. 3, no. 1, pp. 31–48, Jan. 2023, doi: 10.69987/JACS.2023.30103.

[65] Siming Zhao, Hailin Zhou, and Daniel Martinez, “LLM-Assisted Causal Attribution of Service Performance Upgrades on Churn and Tenure: Full Evaluation on the IBM Telco Customer Churn Dataset”, JACS, vol. 3, no. 2, pp. 18–34, Feb. 2023, doi: 10.69987/JACS.2023.30202.

[66] Daren Zheng, Chenyu Li, and Harvey Davidson, “Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation”, JACS, vol. 3, no. 2, pp. 35–49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[67] Q. Xin, “Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models,” Transport Findings, Mar. 2026, doi: 10.32866/001c.157499.

[68] Binghua Zhou, Siming Zhao, and David Chao, “LLM-Guided Energy-Aware A/B Testing for Consolidation and DVFS Policies via Power-Sensitivity Clustering”, JACS, vol. 3, no. 4, pp. 12–30, Apr. 2023, doi: 10.69987/JACS.2023.30402.

[69] Jing Chen, Xinzhuo Sun, and Vincent Brown, “Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact”, JACS, vol. 3, no. 1, pp. 16–30, Jan. 2023, doi: 10.69987/JACS.2023.30102.

[70] Q. Xin, “Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education,” Journal of Artificial Intelligence and Education, vol. 2, no. 1, Jul. 2026, doi: 10.66053/jaaie.v2i1.348.

[71] Q. Xin, “Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI online retail,” Journal of Information and Technology, vol. 14, no. 1, Apr. 2026, doi: 10.32664/j-intech.v14i01.2229.

[72] Yunhe Li, “Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs”, JACS, vol. 3, no. 2, pp. 1–17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[73] Yunhe Li. (2023). Risk-Sensitive Offline Reinforcement Learning for Stable ABR QoE Improvements on Real HSDPA and LTE Traces. Journal of Advanced Computing Systems , 3(4), 1-11. https://doi.org/10.69987/JACS.2023.30401