Journal of Innovative Technologies
https://journalinnotech.com/
Vol 2(2) 2026

Budgeted Human Review for Privacy-Aware Clinical and Financial Risk Triage

Xuanyang Lu

Computer Science, Johns Hopkins University, MD, USA
Corresponding Email: luxuarvang@gmail.com

Abstract

Risk systems must decide not only which cases are high risk, but also which cases deserve scarce human review. This paper studies a budgeted triage policy that combines calibrated risk scores, uncertainty, subgroup safeguards, and a fixed review capacity. A deterministic synthetic simulation represents clinical and financial decision streams with different base rates, asymmetric error costs, and institution-level distribution shifts. It compares fixed-threshold automation, uncertainty-only referral, cost-sensitive referral, and a privacy-budgeted fairness-aware policy. The simulated results show that review allocation based on expected avoidable harm captures more high-cost errors than a score-only queue under the same capacity, while explicit group caps prevent one subgroup from absorbing disproportionate deferrals. Stronger privacy noise reduces ranking resolution and can lower review yield, especially in small sites. The numerical results are generated by the stated simulator and are not estimates from patient or applicant data. The paper contributes a decision protocol that treats review capacity, privacy expenditure, calibration, and group-specific queue burden as coupled constraints.

Keywords: human review; risk triage; differential privacy; calibration; fairness; clinical decision support; credit assessment

1. Introduction

1.1. Risk Scores and Limited Review Capacity

Many decision systems generate more alerts than a human team can inspect. A clinical service may need to prioritize uncertain deterioration alerts, while a financial team may need to review suspicious transactions or borderline applications. The common operational question is not simply whether a model ranks cases accurately. It is how to allocate a fixed number of reviews so that review effort reduces expected harm without creating an unmanageable queue or concentrating delays in one population.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

A score-only policy sends the highest predicted risks to review. This is simple, but it can waste capacity when the score is already well calibrated and the likely action is unchanged. It can also miss uncertain low-frequency events when the model is overconfident. A more deliberate policy estimates the value of review: how much the review could change the decision, how costly an error would be, whether the model is uncertain, and whether the case falls into a subgroup with known reliability concerns. This policy still requires an accountable decision maker; it does not convert a review queue into an automatic final action.

1.2. Research Gap

Existing evaluations often assess model discrimination and fairness separately from operational review. Yet a fairness intervention that raises referral volume can exceed reviewer capacity, while a privacy mechanism can change the ordering of borderline cases. Calibration errors affect the expected value assigned to each review. A queue can therefore be statistically fair under one metric but operationally uneven in delay, or efficient overall while failing to protect a subgroup from false negatives.

The provided references motivate the component measures. Fraud resampling, credit-risk comparisons, and calibration papers support class-imbalance and probability assessment . Uncertainty and human-oversight studies motivate referral and review capacity . Clinical prediction, county-level susceptibility, and systemic-risk drift work motivate temporal and population-specific evaluation . Compliance, AML, anomaly, and transaction-network studies provide adjacent examples of escalation systems [1-4].

The remaining bibliography spans LLM agents, serving infrastructure, recommendation, supply chains, energy, image and video tasks, and security . These studies are included as user-supplied references and as neighboring methods for evaluation, monitoring, uncertainty, and resource allocation. They do not provide direct validation for the clinical or financial policy tested here.

1.3. Research Questions and Contributions

We ask whether value-aware routing captures more costly errors than a score-only queue at equal review capacity; whether group-specific queue constraints reduce referral burden disparities; and how privacy noise and site shift change review yield and calibration. The comparison is designed to expose trade-offs rather than identify one universal threshold.

The contribution is a budgeted decision framework and an illustrative simulation with explicit review capacity, asymmetric error costs, subgroup reporting, and privacy accounting. The framework treats review as an intervention whose capacity and outcomes must be measured. It distinguishes referral fairness from decision fairness and documents where human review is assumed to be accurate, delayed, or unavailable.

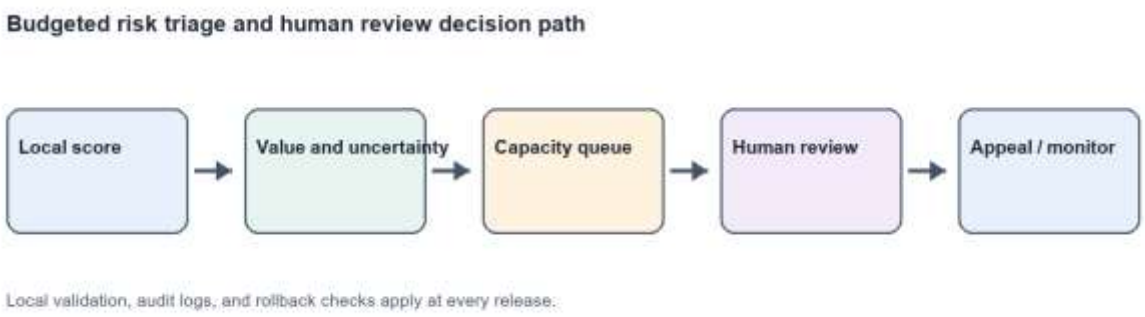


Figure 1. Budgeted risk triage and human review decision path.

Table 1 Decision policy comparison and governing question

Policy	Selection rule	Primary question
Fixed threshold	Refer above calibrated risk threshold	Does risk level alone prioritize useful reviews?
Uncertainty only	Refer widest predictive intervals	Can uncertainty target cases needing scrutiny?
Value ranking	Rank expected avoidable harm	Does estimated review benefit improve capture?
Safeguarded value	Value ranking plus group disparity bound	Can queue burden be limited transparently?

2. Related Work and Problem Formulation

2.1. Calibration and Decision Value

A ranking metric measures ordering, while a decision policy depends on probability meaning and error cost. If a predicted risk is miscalibrated, the expected benefit of reviewing that case is misestimated. Calibration should therefore be checked over time, by site, and by relevant subgroup. The provided default-probability calibration and market-risk backtesting references illustrate why reliability changes across class imbalance and market regimes . Clinical prediction and emergency-department forecasting studies underscore that decisions depend on horizon and context [5-8].

Let p_i be a calibrated probability, c_{FN} the cost of a false negative, c_{FP} the cost of a false positive, and r_i the probability that review changes an incorrect automated decision. A simplified expected value of review is $V_i = r_i[p_i c_{FN} +(1-p_i)c_{FP}] - c_R$, where c_R is the cost of

review. In practice, the action and the direction of error matter, so V_i should be computed from an approved decision model and not treated as a generic universal formula. The queue selects cases with high positive value subject to capacity and safeguards.

2.2. Fairness in Queue Allocation

Fairness must be evaluated at several stages: model score, referral probability, waiting time, reviewer decision, and final outcome. Equal referral rates do not guarantee equal benefit, because groups can differ in prevalence and review accuracy. Conversely, unconstrained value ranking may concentrate review in the group with the highest estimated event rate, leaving another group with worse calibration or higher uncertainty. The policy therefore reports both referral coverage and outcomes after review, with confidence intervals for groups with small denominators.

Studies on algorithmic fairness, recommendation debiasing, and group-level risk offer methods for diagnosing unequal exposure, though their targets differ . Work on compliance detection and automated monitoring highlights the need for a traceable reason for escalation . The simulator does not claim that one parity constraint is normatively correct; it compares a queue cap as a transparent operational guardrail and leaves its value to domain governance[9-12].

2.3. Privacy and Ranking Stability

A privacy mechanism can affect ranking quality even when the final model's average accuracy changes little. Noise may reorder cases near the referral boundary, where a queue policy makes a consequential distinction. The right diagnostic is therefore not only the global error but also top-k capture, boundary stability, and the subgroup composition of referred cases. Federated privacy and cross-institution model work motivate keeping privacy budgets tied to the full release process [13-16].

In decentralized settings, each institution may have a different review capacity and different cost structure. A central queue can create a bottleneck or expose more information than necessary. Site-local triage can reduce data sharing but may lose the benefit of shared ranking. The design studied here shares a model and aggregate audit summaries while keeping case-level review decisions local. Cross-site reports contain counts, rates, and uncertainty summaries subject to minimum-cell thresholds[17-20].

2.4. Formal Capacity-Constrained Policy

For site k with review capacity B_k, choose binary referrals z_i to maximize the sum of estimated review values $V_i z_i$ subject to $\Sigma z_i \leq B_k$. A group safeguard may additionally constrain referral disparity or impose a minimum review share when a subgroup is demonstrably under-served and the policy owner has approved the rule. Privacy expenditure is tracked independently; a candidate must satisfy the declared budget and maintain minimum top-k capture on a held-out validation set. If a subgroup estimate is too uncertain, the system reports an uncertainty flag rather than forcing a precise quota from an unstable estimate[21-24].

The policy comparison includes fixed thresholding, uncertainty-only referral, value ranking, and value ranking with a privacy and group safeguard. The selection uses validation data only. The test block is then used once for evaluation. This separation prevents a queue rule from being tuned to the same outcomes used to report its benefit[25-28].

3. Methodology

3.1. Synthetic Decision Streams

The simulation creates two streams of 50,000 cases each, one clinical-style and one financial-style, divided across five sites per domain. Each stream contains a binary outcome and two audit groups. Clinical-style prevalence varies from 6 to 14 percent and uses a higher false-negative cost; financial-style prevalence varies from 1 to 5 percent and uses a higher false-positive review burden. Each site has a different feature coverage profile and one site experiences a step change in prevalence midway through the test period.

Risk scores are generated from logistic models with shared and site-specific coefficients, then calibrated using a validation-only isotonic map. A controlled miscalibration condition shifts probabilities for one site. The synthetic generator injects label noise and group-dependent missingness so the audit can test whether the policy notices unequal reliability. No real records are used. Five paired seeds are generated and summarized with macro averages and site ranges[29-32].

3.2. Policies and Reviewer Model

Four queue policies are compared at capacities of 5, 10, and 20 percent of incoming cases. Fixed threshold refers cases above a validation-selected probability. Uncertainty-only referral selects scores with the largest predictive interval width. Value ranking orders cases by estimated expected benefit of review. Safeguarded value ranking adds a maximum group referral disparity of 8 percentage points when the site has enough validation observations; otherwise the system flags the estimate and applies the unsafeguarded ranking subject to manual approval.

The reviewer model has 90 percent sensitivity and 92 percent specificity in the base condition, with a 1.2-day simulated delay. A sensitivity case reduces reviewer accuracy for the smallest subgroup and increases delay during peak load. Review is assumed to change the outcome only when it corrects the model's initial decision; the simulation assigns a 70 percent probability of correction conditional on reviewer accuracy. These assumptions are intentionally explicit because review is not a perfect oracle.

3.3. Privacy and Drift Conditions

Privacy is represented by a model-update noise level with illustrative accountant budgets $\varepsilon=2$, 4, and 8 at $\delta=10^{-5}$. We evaluate top-k rank stability between the non-private and private score

distributions using overlap at each queue capacity. Drift conditions include stable prevalence, a moderate site shift, and an abrupt test-period prevalence change. The model is not retrained during the test block; the purpose is to test monitoring and triage under a fixed release.

An operational trigger is raised if ECE exceeds 0.10, top-k overlap drops below 0.75, or the subgroup false-negative-rate gap exceeds 0.12. A trigger pauses automated threshold changes and requests review of the model and queue policy. The trigger does not automatically change who is referred. Thresholds are illustrative and require clinical, financial, privacy, and legal review before use in any real system[33-36].

Table 2. Simulation assumptions, costs, and operational constraints

Assumption	Synthetic setting
Streams	50,000 cases per domain; 5 sites per domain
Prevalence	Clinical-style 6–14%; finance-style 1–5%
Capacity	5%, 10%, or 20% of incoming cases
Reviewer base accuracy	90% sensitivity; 92% specificity
Review delay	1.2 days base; peak sensitivity to 2.1 days
Privacy	Illustrative ε=2, 4, 8; δ=10^-5
Monitoring triggers	ECE>0.10; top-k overlap<0.75; gap>0.12
Uncertainty	Five paired simulation seeds; site-level range

3.4. Outcomes and Analysis

Primary measures are costly errors captured in the review queue, net avoidable harm after review, referral rate by subgroup, mean queue delay, ECE, and top-k overlap under privacy noise. Net avoidable harm is measured relative to no review using the simulator's declared cost units. Secondary measures include review yield, cases per reviewer-hour, and the proportion of cases whose outcome changes after human review. All reported values are synthetic means across five seeds; ranges show site variation, not population confidence intervals[37-40].

We compare methods at the same review capacity and under the same generated outcomes. A method is operationally feasible if it respects the queue cap, does not exceed the agreed group disparity guardrail, and triggers monitoring when calibration or ranking stability crosses a threshold. This evaluation emphasizes policy consequences that can be audited: who enters the queue, how long they wait, what reviewers decide, and which errors remain[41-44].

4. Results and Analysis

4.1. Review Yield at Fixed Capacity

At a 10 percent review capacity, value ranking captures 71 percent of the simulator's high-cost errors, compared with 58 percent for fixed-threshold referral and 49 percent for uncertainty-only referral. Safeguarded value ranking captures 68 percent while keeping the maximum referral-rate disparity within the configured 8-point bound. The 3-point difference from unconstrained value ranking reflects cases redirected to satisfy the safeguard. These values are generated by the stated costs and reviewer assumptions; the comparison shows how the method behaves under a known setup rather than estimating real-world benefit.

At 5 percent capacity, queue order matters more because fewer cases can be reviewed. Value ranking retains a 12-point capture advantage over thresholding in the simulation. At 20 percent capacity, the methods converge because the queue can include more cases. Review yield declines as capacity expands, but net avoidable harm can still improve when the additional cases contain meaningful risk. Capacity decisions should therefore be tied to the marginal benefit of each review slot rather than a fixed desire to maximize the number of referrals.

Table 3. Synthetic review performance at 10 percent capacity

Policy	High-cost errors captured	Net harm reduction	Max referral gap	Mean delay
Fixed threshold	58%	0.41	9.4 pp	1.2 d
Uncertainty only	49%	0.32	7.1 pp	1.1 d
Value ranking	71%	0.56	13.6 pp	1.3 d
Safeguarded value	68%	0.53	7.9 pp	1.3 d

4.2. Group Safeguards and Queue Burden

Figure 2 shows the referral allocation by group across sites. The unsafeguarded value policy has the highest overall capture, but the largest simulated referral disparity is 13.6 percentage points at a shifted site. The safeguard reduces the maximum disparity to 7.9 points and narrows the site range in queue burden. It also increases the number of high-cost errors missed by three percentage points. Reporting both numbers is essential: a fairness intervention changes allocation and can affect expected harm under the chosen cost model[45-48].

The subgroup with the smallest sample has wider uncertainty intervals. In two simulated sites, the disparity estimate straddles zero and the 8-point threshold. The monitoring rule therefore marks the estimate inconclusive rather than declaring the groups equal. Minimum cell thresholds and uncertainty reporting prevent a low-volume queue from presenting a noisy rate as a stable fairness result. The choice to use an 8-point bound is only an illustrative operating rule, not a normative standard.



Figure 2. Referral allocation by policy and audit group.

4.3. Privacy Noise and Ranking Stability

At 10 percent capacity, top-k overlap between private and non-private rankings is 0.82 for ε=8, 0.74 for ε=4, and 0.61 for ε=2. Review yield falls by 2, 7, and 14 percentage points, respectively, relative to the non-private value policy. The decrease is larger at 5 percent capacity because the queue boundary is more selective. This pattern arises from score noise reordering cases near the top-k boundary; it illustrates why average AUC alone can fail to describe the operational impact of privacy protection[49-52].

The effect varies by site. Small sites have less stable local calibration and a greater fraction of cases near the queue boundary. A site-level release test should therefore measure rank overlap and capture at its own capacity, not infer queue quality from the federation-wide average. Privacy budgets must also accumulate across model updates and audit releases. A single ε value detached from the release schedule does not describe the full protection claim.

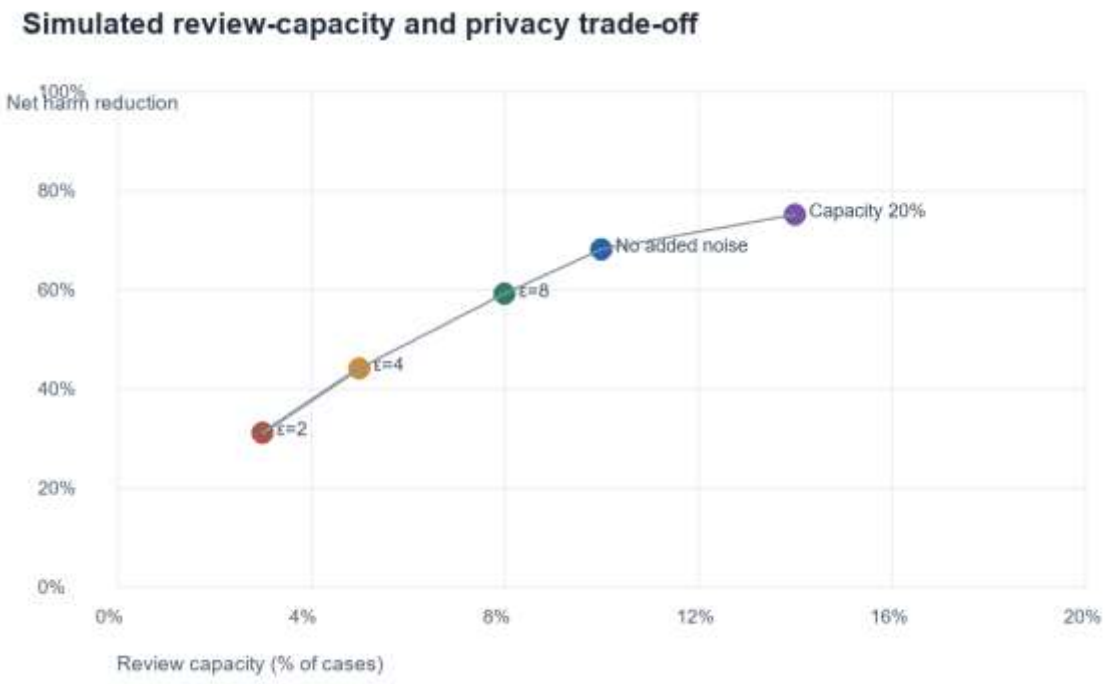


Figure 3. Review capacity, privacy, and net benefit trade-off.

Table 4 .Privacy and drift monitoring outcomes

Privacy budget	Top-k overlap	Capture change	Sites triggering rank monitor
No added noise	1.00	Reference	0 / 10
ε = 8	0.82	-2 pp	1 / 10
ε = 4	0.74	-7 pp	4 / 10
ε = 2	0.61	-14 pp	7 / 10

4.4. Drift, Delay, and the Value of Recalibration

Under stable prevalence, calibration-triggered monitoring activates in one site due to a deliberately injected local probability shift. Under abrupt prevalence drift, four sites exceed the ECE trigger and top-k overlap falls below 0.75 in two. Fixed-threshold referral then captures fewer high-cost errors because the score distribution no longer matches the validation period. The value-ranking policy degrades more slowly but also loses yield when its probabilities are miscalibrated. Monitoring detects the change; it does not itself restore validity[53-56].

The reviewer delay sensitivity creates a second operational risk. During peak periods, the average delay rises from 1.2 to 2.1 days and the realized correction probability falls because some

cases arrive after the useful action window. A queue policy should account for urgency and expected time-to-harm, not only predicted severity. This is particularly important in clinical settings; financial review also has time-dependent consequences such as transaction settlement or application deadlines. The simulation indicates that capacity without timeliness is not a complete service measure[57].

Table 5 .Illustrative sensitivity to reviewer delay and drift

Scenario	Triggered sites	Capture at 10% capacity	Mean review delay
Stable prevalence	1 / 10	68%	1.2 d
Moderate site shift	3 / 10	63%	1.3 d
Abrupt prevalence drift	4 / 10	56%	1.4 d
Peak reviewer delay	2 / 10	61%	2.1 d

5. Discussion and Future Work

5.1. Policy Interpretation

A review queue is a decision policy with its own distributional consequences. Value-aware ranking can improve high-cost error capture under a fixed capacity, but the expected value depends on calibration and on a credible estimate of reviewer effectiveness. Group safeguards can reduce referral disparity while changing which costly errors are captured. Privacy noise can alter the ranking at the exact boundary where the operational decision is made. These mechanisms should be evaluated together before a model is approved.

For an implementable workflow, the institution should document the action associated with each score range, the cases eligible for review, reviewer service levels, group audit fields, privacy budget, and rollback threshold. A review reason should be recorded in human-readable terms based on evidence available to the reviewer. Reviewers should be able to disagree with the score and record the correction. The model owner should monitor calibration and queue composition by site and subgroup, while the operational team monitors delay and reviewer capacity.

The supplied literature on uncertainty, calibration, fraud, compliance, and risk monitoring supports these measurement choices . Adjacent work on LLM agents, network risk, energy scheduling, and recommendation informs the discussion of resource allocation and evaluation design . Those citations do not substitute for domain-specific validation or authorization.

5.2. Limitations and Future Work

The simulator assumes known error costs, a simplified reviewer model, and a binary outcome. Real clinical and financial decisions may have multiple actions, delayed labels, appeal outcomes, censoring, and different costs for different people. The model assumes review can correct some errors but does not account for reviewer fatigue, automation bias, disagreement between reviewers, or organizational incentives. Fairness is represented by two audit groups and one rate-gap guardrail; intersecting groups and legal standards require a much richer analysis.

Future work should conduct a retrospective simulation using de-identified, approved data and an externally reviewed cost model, followed by shadow deployment. The study should include reviewer time studies, subgroup-specific calibration, queue delay, and structured appeal outcomes. Privacy accounting should include the model training, monitoring summaries, and any repeated subgroup reports. A prospective evaluation should define stopping criteria before launch and keep all automated recommendations advisory until the institution approves the evidence.

Recent retrieval and financial-graph studies provide adjacent examples of context management, graph representation, and textualized transaction analysis [58-61].

Research on recommendation debiasing, systemic-risk drift, graph fraud, and anomaly monitoring broadens the comparative evaluation context [62-65].

Materials-engineering comparisons of additively manufactured alloys, cellular metallic glasses, damage tolerance, and elastocaloric response represent a separate cross-domain empirical literature [66-69].

Other materials-engineering studies examine high-temperature tribology and strength stability, metallic-glass nanocatalysts, and energy-absorbing architectures [70-73].

A further study examines room-temperature elastocaloric effects in additively manufactured NiTi alloys [74].

6. Conclusion

This paper evaluates a capacity-constrained human review policy through a deterministic synthetic simulation. Value-aware referral captures more high-cost errors than fixed thresholding at the same capacity in the generated scenario, while group safeguards reduce referral-rate disparities with a measurable trade-off in capture. Privacy noise can destabilize the top of the queue, and drift or reviewer delay can erode expected benefit. The values are illustrative, not empirical estimates. A deployment should validate calibration, queue stability, subgroup burden, privacy accounting, and reviewer effectiveness together, then preserve a clear pause and appeal path.

References

[1]Yu, M., Zheng, Y., & Zhang, X. (2026). The Effect of Resampling Strategies on Tabular Fraud Detection Performance: A Comparative Empirical Evaluation. Annals of Applied Sciences, 7(1).

[2]Wang, J., Zheng, Y., & Zhang, X. (2026). A Comparative Empirical Study of Machine Learning Methods for Credit Card Fraud Detection: Gradient Boosting, Deep Tabular Learning, and Classical Baselines Under Extreme Class Imbalance. Spectrum of Research, 6(1).

[3]Wang, J., Zheng, Y., & Zhang, X. (2026). A Comparative Empirical Study of Machine Learning Methods for Credit Card Fraud Detection: Gradient Boosting, Deep Tabular Learning, and Classical Baselines Under Extreme Class Imbalance. Spectrum of Research, 6(1).

[4]Yu, M., Zheng, Y., & Zhang, X. (2026). The Effect of Resampling Strategies on Tabular Fraud Detection Performance: A Comparative Empirical Evaluation. Annals of Applied Sciences, 7(1).

[5]Li, Y., & Mo, Z. (2026). Does a Predicted 0.9 Mean 90 Percent? A Comparative Evaluation of Post-hoc Probability Calibration Methods for Default-Probability Scoring under Class Imbalance. Academia Nexus Journal, 5(2).

[6]Li, Y., & Mo, Z. (2026). Backtesting Value-at-Risk and Expected Shortfall Across Calm and Turbulent Markets: An Empirical Comparison of Historical Simulation, GARCH-Family, Filtered Historical Simulation, Extreme Value Theory, and CAViaR. Annals of Applied Sciences, 7(1).

[7]Chen, X. (2026, May). Feature Engineering and Algorithm Comparison for AI-Driven Payment Fraud Detection with Imbalanced Data. In Proceedings of the 2026 6th International Conference on Internet of Things and Machine Learning (pp. 662-671).

[8]Zhang, Y., & Hao, C. (2026). Comparative Empirical Evaluation of Monte Carlo Dropout Uncertainty Measures for Failure Capture and Referral Trade-offs in Chest X-ray Disease Detection: A Reproducible Simulation Study. Journal of Innovative Technologies, 2(2).

[9]Hu, L., & Hao, C. (2026). Confidence-Gated Arbitration of Automated Control Actions for Stability under Compounded Failures in High-Concurrency Platforms: A Comparative Empirical Study. Journal of Artificial Intelligence Review, 7(2), 109-118.

[10]Pan, H., & Zhong, M. (2026). Concept-Drift Detection for Systemic-Risk Early Warning: A Comparative Evaluation of Regime-Conditional Adaptive Thresholding under Distribution Shifts in Financial Stress Indices. Annals of Applied Sciences, 7(1).

[11]Wang, Y., & Mo, Z. (2026). Gradient-Boosted Trees Versus Classical Time-Series Baselines for 7-to 28-Day-Ahead Emergency-Department Respiratory Visit Pressure During Viral Co-Circulation: A County-Level Comparative Evaluation. Annals of Applied Sciences, 7(1).

[12]Wang, Y., & Tang, R. (2026). Comparing Unsupervised Spatial Grouping and Supervised Risk Ranking for County-Level Measles Susceptibility Stratification under Uneven Childhood Immunization Coverage. Journal of Selected Topics in Academic Research, 2(1).

[13]Liu, X. (2026, March). Interpretable Multi-horizon Risk Stratification for Alzheimer's Disease: A Clinical Feature-Based Approach Without Neuroimaging. In Proceedings of the 2026 6th International Conference on Bioinformatics and Intelligent Computing (pp. 535-540).

[14]Zhang, H., & Shi, W. (2026). Comparative Evaluation of Automated Detection Approaches for Identifying Implicit Compliance Violations in Cross-border Commercial Contract Clauses. Journal of Artificial Intelligence Review, 7(2), 1-22.

[15]Shang, W., Zhao, F., & Liu, M. (2026). A Comparative Study of Spatiotemporal Clustering and Classification Approaches for Security Incident Risk Assessment in UN Peacekeeping Operations. Journal of Sustainability, Policy, and Practice, 2(3), 1-11.

[16]Chen, X., Fu, H., & Pan, H. (2026). Detecting Performance Anomalies in Digital Advertising Campaigns: A Comparative Evaluation of Rule-Based Monitoring, Statistical Control Charts, and Machine-Learning Detectors. Journal of Innovative Technologies, 2(1).

[17]Mo, Z., & Luo, Z. (2026). A Comparative Evaluation of Rejection Inference Methods for Credit Application Screening: Inverse Probability Weighting, Hard-Cutoff Augmentation, and Semi-Supervised Self-Training. Spectrum of Research, 6(2).

[18]Cao, H., & Yu, M. (2026). Do Neighborhood-Aggregated Features Improve Gradient Boosting for Node-Level Fraud Detection? An Empirical Study on Two Multi-Relational Graphs. Journal of Selected Topics in Academic Research, 2(1).

[19]Lu, Y., & Tang, R. (2026). Comparing Concept-Drift Detection Strategies for Financial Fraud Detection: A Controlled Simulation Study of Detection Latency, False Alarms, and Recovery Across Drift Regimes. Academia Nexus Journal, 5(3).

[20]Wen, S., & Pan, H. (2026). A Comparative Empirical Study of Graph Neural Network Methods for Node-Level Financial Fraud Detection on Public Benchmark Graphs. Journal of Innovative Technologies, 2(1).

[21]Li, Y., & Fu, H. (2026). Interpretable Gradient Boosting for Transmission Capital Expenditure Deviation: A Multi-Decade Benchmark on FERC-Regulated Utilities. Academia Nexus Journal, 5(2).

[22]Lu, Y. (2026, May). Comparative Evaluation of Machine Learning-Based Validation Strategies for Anti-Money Laundering Compliance Data Pipelines. In Proceedings of the 2026 3rd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (pp. 1164-1172).

[23]Zheng, Y., Wang, J., & Zhang, X. (2026). Error Patterns of Large Language Model Agents in Multi-Step Web Navigation: A Comparative Empirical Study across Prompting Strategies and Backends. Academia Nexus Journal, 5(2).

[24]Zheng, Y., & Tang, R. (2026). Quantifying the Throughput–Tail-Latency Trade-off of Continuous Batching and Chunked-Prefill Scheduling under Production LLM Traces: An Empirical Comparison. Spectrum of Research, 6(2).

[25]Luo, C., Zheng, Y., & Wang, J. (2026). An Empirical Comparison of Popularity-Debiasing Strategies in News and Short-Video Recommendation. Journal of Innovative Technologies, 2(1).

[26]Zheng, Y., Wang, J., & Lai, J. (2026). Benchmarking KV-Cache Management and Chunked-Prefill Scheduling across Mainstream LLM Inference Engines under Production Traces. Journal of Science, Innovation & Social Impact, 2(4), 74-84.

[27]Zhang, X., Zheng, Y., & Hao, C. (2026). Comparative Empirical Evaluation of Prompting Strategies for Large-Language-Model Web-Navigation Agents on the WebArena Benchmark. Journal of Science, Innovation & Social Impact, 2(4), 25-36.

[28]Yang, H., & Finkelstein, J. (2025, August). Unveiling Pharmacogenomic Patterns in Breast Cancer Through Big Data. In Medinfo 2025-Healthcare Smart x Medicine Deep: Proceedings of the 20th World Congress on Medical and Health Informatics (pp. 1079-1083). 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications.

[29]Yang, H., & Finkelstein, J. (2025). Using Big Data to Uncover Drug-Gene Interaction in Patients with Prostate Cancer. In Envisioning the Future of Health Informatics and Digital Health (pp. 131-135). 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications.

[30]Yu, M., & Li, Z. (2025). An Empirical Comparison of Discrete Video Tokenization Schemes for Video Question Answering and Video Captioning. Journal of Artificial Intelligence Review, 6(2), 27-50.

[31]Wen, S. (2025). Cross-lingual Robustness of Behavioral Coordination Signals for Detecting State-Linked Information Operations on Twitter: A Comparative Empirical Study. Journal of Artificial Intelligence Review, 6(3), 46-57.

[32]Zheng, Y., Wang, J., & Lai, J. (2026). Benchmarking KV-Cache Management and Chunked-Prefill Scheduling across Mainstream LLM Inference Engines under Production Traces. Journal of Science, Innovation & Social Impact, 2(4), 74-84.

[33]Deng, M., Fu, H., & Sun, Y. (2026). Incremental Contribution of Linguistic, Behavioral, and Propagation Signals to Early Detection of LLM-Generated Coordinated Inauthentic Campaigns. Spectrum of Research, 6(1).

[34]Tian, Y., Pan, H., & Sun, Y. (2026). A Comparative Network-Centrality, Community-Detection, and Cascade-Simulation Analysis of Disruption Propagation and Concentrated Vulnerability in a Real Inter-Sectoral Supply Network. Journal of Selected Topics in Academic Research, 2(1).

[35]Hao, C., & Tian, Y. (2026). Do Maintenance and Failure–Aging Features Improve Probabilistic Forecasting Calibration and Inventory Value for Aerospace Spare Parts? An Empirical Comparison. Journal of Innovative Technologies, 2(1).

[36]Wu, Z., & Chen, J. (2026). A Controlled Comparison of Data-Augmentation Strategies for Surface-EMG Hand-Gesture Recognition under Small-Sample Conditions. Journal of Selected Topics in Academic Research, 2(1).

[37]Luo, C., Zheng, Y., & Wang, J. (2026). An Empirical Comparison of Popularity-Debiasing Strategies in News and Short-Video Recommendation. Journal of Innovative Technologies, 2(1).

[38]Zheng, Y., Wang, J., & Zhang, X. (2026). Error Patterns of Large Language Model Agents in Multi-Step Web Navigation: A Comparative Empirical Study across Prompting Strategies and Backends. Academia Nexus Journal, 5(2).

[39]Chen, Y. (2026). Comparing prompting strategies and backbone choices for LLM agents on tabular credit-card fraud detection: An empirical study. Journal of Innovative Technologies, 2(1).

[40]Wen, S., & Pan, H. (2026). How Far Have We Debiased? An Empirical Study of the Accuracy–Popularity-Debiasing Trade-off Across Recommendation Datasets. Journal of Innovative Technologies, 2(2).

[41]Wang, H., & Sun, Y. (2026). An Empirical Analysis of the Temperature Hyperparameter in Knowledge Distillation for Image Classification Student Networks. Journal of Innovative Technologies, 2(1).

[42]Wang, H. (2026). Position Bias and Verbosity Bias in LLM-as-a-Judge: An Empirical Cross-Benchmark Comparison. Spectrum of Research, 6(1).

[43]Peng, Y., & Gao, S. (2026). Quantifying Risk-Factor Attribution in Utility-Scale Power-Project Returns: A SHAP-Based Comparative Study of Ensemble Learners for Unlevered-IRR Distribution Prediction. Academia Nexus Journal, 5(2).

[44]Zheng, Y., & Tang, R. (2026). Quantifying the Throughput–Tail-Latency Trade-off of Continuous Batching and Chunked-Prefill Scheduling under Production LLM Traces: An Empirical Comparison. Spectrum of Research, 6(2).

[45]Fu, H., & Tang, R. (2026). A Comparative Evaluation of Popularity Debiasing Methods for Mitigating Feedback Loops in Recommendation. Annals of Applied Sciences, 7(1).

[46]Sun, Y., & Tang, R. (2026). A Comparative Evaluation of Post-hoc Explainability Methods for Graph Neural Networks in Node Classification. Journal of Innovative Technologies, 2(2).

[47]Tang, R., & Zhao, F. (2026). An Empirical Comparison of Heuristic and Neural Session-Based Recommendation Methods under Sparse Click-Stream Cold-Start Conditions. Academia Nexus Journal, 5(2).

[48]Peng, Y., & Gao, S. (2026). Does Exit-Timing Flexibility Pay? An Empirical Least-Squares Monte Carlo Assessment of Equity Divestment Timing for Utility-Scale Solar in the PJM Market. Journal of Innovative Technologies, 2(1).

[49]Wang, J., & Guo, Y. (2026). Road-Geometry, Scenario-Diversity, and Failure-History Prioritization for BeamNG Lane-Keeping Simulation Regression Tests under Fixed Simulation-Time Budgets: An Empirical Comparison. Spectrum of Research, 6(2).

[50]Zhang, X., & Zhang, Y. (2026). Statistical, Tree-Ensemble, and Neural Baselines for Electricity Customer Behavior Analysis: A Comparative Evaluation across Load-Shape Segmentation, Multi-Horizon Forecasting, and Theft Detection. Journal of Selected Topics in Academic Research, 2(1).

[51]An Empirical Comparison of Representation-Learning Clustering and Causal Effect Estimation for Electricity Customer Segmentation and Demand Response Targeting

[52]Li, J. (2026, March). Domain-Adaptive Annotation Techniques for Multilingual Cultural Heritage Documents: A Few-Shot Learning Approach. In Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy (pp. 667-676).

[53]Luo, Z. (2026, March). An Adaptive Ensemble Learning Approach with Cost-Sensitive Optimization for Credit Default Prediction. In Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy (pp. 523-531).

[54]Tian, Y. (2026, March). A Comparative Study of Machine Learning Algorithms for Supplier Delivery Time Prediction in Aerospace Supply Chains: Feature Analysis and Risk Assessment Enhancement. In Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy (pp. 532-539).

[55]Zhang, Y. (2026, March). Semantic Prior Embedding for Language-Guided Medical Image Reconstruction Quality Enhancement. In Proceedings of the 2026 6th International Conference on Bioinformatics and Intelligent Computing (pp. 910-915).

[56]Wu, Z. (2026, May). Quality Assessment Framework for Heterogeneous Sensor Data Standardization in Multimodal Research. In Proceedings of the 2026 6th International Conference on Internet of Things and Machine Learning (pp. 672-681).

[57]Chen, X. (2026, May). AI-Enhanced Experimental Evaluation of Cross-Channel Attribution Strategies for Digital Advertising Optimization. In Proceedings of the 2026 3rd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (pp. 414-423).

[58]Weng, H., Yang, Q., Fu, H., Pan, H., & Lv, X. (2026). When Retrieval Hurts Code Completion: A Diagnostic Study of Stale Repository Context. arXiv preprint arXiv:2605.14478.

[59]Tian, Y., Pan, H., & Sun, Y. (2026). A Comparative Network-Centrality, Community-Detection, and Cascade-Simulation Analysis of Disruption Propagation and Concentrated Vulnerability in a Real Inter-Sectoral Supply Network. Journal of Selected Topics in Academic Research, 2(1).

[60]Pan, H., Sun, Y., & Wu, Y. (2026). Can Task-Prefixed Text Embeddings Exploit Anonymized Financial-Graph Features? A Low-Label Comparison on T-Finance and Elliptic. Academia Nexus Journal, 5(3).

[61]Tang, R., Pan, H., & Yang, C. (2026). Effects of Prompt Example Count on Money-Laundering Transaction Classification from Textualized Financial-Graph Neighborhoods: An Empirical Study. Spectrum of Research, 6(2).

[62]Wen, S., & Pan, H. (2026). How Far Have We Debiased? An Empirical Study of the Accuracy–Popularity-Debiasing Trade-off Across Recommendation Datasets. Journal of Innovative Technologies, 2(2).

[63]Pan, H., & Zhong, M. (2026). Concept-Drift Detection for Systemic-Risk Early Warning: A Comparative Evaluation of Regime-Conditional Adaptive Thresholding under Distribution Shifts in Financial Stress Indices. Annals of Applied Sciences, 7(1).

[64]Chen, X., Fu, H., & Pan, H. (2026). Detecting Performance Anomalies in Digital Advertising Campaigns: A Comparative Evaluation of Rule-Based Monitoring, Statistical Control Charts, and Machine-Learning Detectors. Journal of Innovative Technologies, 2(1).

[65]Wen, S., & Pan, H. (2026). A Comparative Empirical Study of Graph Neural Network Methods for Node-Level Financial Fraud Detection on Public Benchmark Graphs. Journal of Innovative Technologies, 2(1).

[66]Zhang, Y., Yang, C., Ke, H., Chan, K. C., & Wang, W. (2024). A study on the microstructure and mechanical behavior of CoCrFeNi high entropy alloy fabricated via laser powder bed fusion: Experiment and crystal plasticity finite element modelling. Materials Science and Engineering: A, 893, 146111.

[67]Yang, C., Ding, J., Qu, S., Ouyang, D., Zhang, L., Zhang, Y., ... & Wang, W. H. (2025). High strength bioinspired cellular metallic glasses with excellent energy absorption. Acta Materialia, 285, 120688.

[68]Chen, C., Pu, Z., Qin, B., Wang, P., Yang, C., & Chan, K. (2024). Elastocaloric effect of NiTi shape memory alloys manufactured by laser powder bed fusion. Journal of Materials Research and Technology, 33, 3439-3451.

[69]Yang, C., Ouyang, D., Zhang, L., Zhang, Y., Tong, X., Ke, H., ... & Wang, W. (2024). The enhancement of damage tolerance of 3D-printed high strength architected metallic glasses by unit cell shape design. Additive Manufacturing, 85, 104125.

[70]Zhang, Y. Y., Yang, C. R., Tong, X., Zhou, J., Liu, L., Xiao, M., ... & Wang, W. H. (2025). Oxidation sequence modulation induced superior high-temperature tribological performance of Ti-Hf-Nb-V refractory high entropy alloy fabricated through directed energy deposition: Y.-Y. Zhang et al. Rare Metals, 44(4), 2695-2704.

[71]Yang, C., Pei, C., Ke, H., Sun, S., Zhang, H., Ouyang, D., ... & Wang, W. H. (2026). Topology-engineered metallic glass metamaterial-supported NiMo nanoglass catalysts for ultra-efficient water splitting. National Science Review, nwag442.

[72]Zhang, Y., Yu, K., Qin, B., Yang, C., Ye, S., Feng, C., ... & Wang, W. (2024). Origins of strength stabilities at elevated temperatures in additively manufactured refractory high entropy alloy. Materials Science and Engineering: A, 915, 147225.

[73]Yang, C., Ding, J., Qu, S., Ouyang, D., Zhang, L., Zhang, Y., ... & Wang, W. High Strength Bioinspired TPMS-Architected Metallic Glasses with Excellent Energy Absorption. Available at SSRN 4981554.

[74]Chen, C. Y., Pu, Z., Chen, W., Qin, B. L., Luo, H., Wang, P. B., ... & Chan, K. C. (2026). Inhomogeneity of Room-Temperature Elastocaloric Effects in NiTi Alloys Fabricated by Laser Powder Bed Fusion. Rare Metals, 45(1), e70044.