PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 04 (2026) Page: 69-78
E-ISSN: 3109-8886

Fair Incrementality Learning and Conservative Policy Selection without Persistent Identifiers: Calibrated Counterfactual Evaluation for Ads and Job Ranking

[1]Chen Yang, 2Derek Peterson, [3]Lin Feng, [4]Rachel Adams

[1]Department of Computer Science, Northwestern University, Evanston, IL, USA
[2]Department of Computer Science, Brown University, Providence, RI, USA
[3]Department of Computer Science, University of California, Irvine, Irvine, CA, USA
[4]Department of Computer Science, University of Colorado Boulder, Boulder, CO, USA

Email: c.yang.cs@gmail.com

Abstract

Identifier loss complicates advertising measurement, while job recommenders must balance utility and opportunity. We linked two separate tracks through conservative selection: randomized Criteo Uplift v2.1 for intention-to-treat estimation and FairJob for identifier-free ranking and proxy-group audit. S-HGB achieved test Qini 4.689×10^{-3}, top-decile uplift 6.746 points, and top-20% gain 0.974 points; a lower-confidence-bound rule treated 30% and gained 1.007 points. FairJob's best PR-AUC was 0.00833. Proxy penalization raised MRR from 0.574 to 0.585 but widened group disparity from 0.023 to 0.071, so deployment was rejected. Evidence cards quantified local faithfulness without interpreting anonymized fields. Calibration and gating supported identifier-free decisions, but predictive parity did not ensure fair ranking.

Keywords: Uplift Modeling; Heterogeneous Treatment Effects; Off-Policy Evaluation; Job Recommendation; Ranking Fairness; Causal Calibration; Privacy-Preserving Advertising; Conservative Policy Selection; Explanation Faithfulness.

Abstrak

Hilangnya pengenal (identifier loss) mempersulit pengukuran efektivitas periklanan, sementara sistem rekomendasi pekerjaan harus menyeimbangkan manfaat (utility) dan kesempatan yang adil. Kami menghubungkan dua jalur penelitian yang terpisah melalui seleksi konservatif, yaitu Criteo Uplift v2.1 yang diacak untuk estimasi intention-to-treat dan FairJob untuk pemeringkatan tanpa pengenal (identifier-free ranking) serta audit kelompok proksi (proxy-group audit). Model S-HGB mencapai nilai Qini pada data uji sebesar $4,689 \times 10^{-3}$, top-decile uplift sebesar 6,746 poin, dan top-20% gain sebesar 0,974 poin; aturan lower-confidence-bound dengan perlakuan terhadap 30% populasi menghasilkan peningkatan sebesar 1,007 poin. Pada dataset FairJob, nilai PR-AUC terbaik yang dicapai adalah 0,00833. Pemberian penalti berbasis kelompok proksi (proxy penalization) meningkatkan Mean Reciprocal Rank (MRR) dari 0,574 menjadi 0,585, namun juga memperlebar disparitas antarkelompok dari 0,023 menjadi 0,071, sehingga model tersebut tidak direkomendasikan untuk diterapkan. Evidence card digunakan untuk mengukur tingkat kesetiaan penjelasan lokal (local faithfulness) tanpa menginterpretasikan atribut yang telah dianonimkan. Kalibrasi dan mekanisme gating mendukung pengambilan keputusan tanpa menggunakan pengenal, tetapi kesetaraan prediktif (predictive parity) tidak menjamin keadilan dalam proses pemeringkatan.

Kata Kunci: Pemodelan Uplift; Efek Perlakuan Heterogen; Evaluasi Off-Policy; Rekomendasi Pekerjaan; Keadilan Pemeringkatan; Kalibrasi Kausal; Periklanan yang Menjaga Privasi; Seleksi Kebijakan Konservatif; Kesetiaan Penjelasan (Explanation Faithfulness)

A. INTRODUCTION

Identifier loss shifts advertising toward causal response estimation, while job ranking must assess utility and opportunity. Uplift policy [1], conditional response ranking [2], and cookieless experiments [3] motivate evidence-based action or abstention.

FairJob provides a behavioral proxy and randomized positions [4]; Criteo derives from randomized advertising experiments [5]. The FairJob artifact is documented separately [6]; Criteo v2 corrected treatment-ratio leakage [7] and v2.1 was distributed in 2024 [8]. Criteo identifies auction-participation ITT, whereas FairJob randomizes position, so the datasets remain separate.

Heterogeneous-effect learners [9], doubly robust pseudo-outcomes [10], uplift ranking [11], offline slot-policy evaluation [12], [13], conservative selection [14], and counterfactual ad ranking [15] motivate validation-only model choice.

Diterima Redaksi: 01-08-2026 | Selesai Revisi: 06-08-2026 | Diterbitkan Online: 07-08-2026

Calibration [16], isotonic correction [17], imbalance [18], rejection [19], confidence fusion [20], and model-risk workflow [21] require separate ranking/reliability audits. Counterfactual control [22], policy learning [23], safe fallback [24], allocation success [25], constrained bidding [26], and response budgeting [27] motivate abstention and fixed reach; these are governance analogies, not advertising evidence.

Fairness targets opportunity [28], classification [29], exposure [30], and attention [31]. Demographic-free methods [32], [33] are partial; career-ad, delivery, and auction evidence [34], [35], [36] refutes intent as a parity guarantee. Private bidding, attribution, identifier-loss budgeting, and calibrated matching [37], [38], [39], [40] need separate privacy, utility, and fairness audits.

Identifier-free representation [41] remains drift-sensitive [42]. Recommendation and attribution [43], [44], [45] complement faithfulness tests [46], [47], [48]. Evidence-centered RAG and operations [49], [50], [51], [52], [53], [54], tickets [55], and incident cards [56] motivate inspectable support. Review-grounded recommendation [57], cards [58], rubric scoring [59], rationales [60], and editable briefs [61] motivate deterministic schemas, not benchmark effects.

We contribute calibrated Criteo targeting, user-disjoint FairJob auditing, and rejection-capable evidence, joined only for policy selection and explanation (Figure 1)..

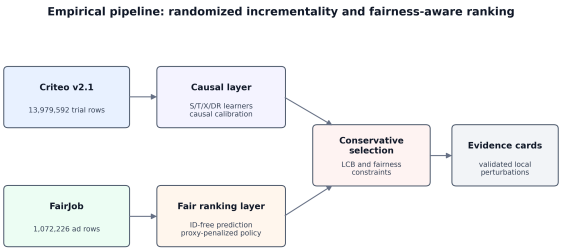

Figure 1 Figure 1. Two-Track Study Workflow Linking Causal Uplift Estimation And Fair Job Ranking To Conservative Policy Selection And Constrained Explanations.

B. METHODS

Data, Estimands, And Quality Control

Table 1 summarizes Criteo's 13,979,592 randomized rows, 12 pre-treatment covariates, visit/conversion outcomes, and audit-only exposure. FairJob supplied 1,072,226 rows with 13 categorical, 35 numerical, and senior inputs. IDs served only splitting/evaluation; proxy, rank, and random-display flag were excluded.

FairJob covered 30,361 users, 224,898 impressions, and 57,355 products; 105,869 randomized rows covered 21,962 impressions and 766 clicks. Excluding 197 contradictory-proxy users left 1,049,707 rows (Table 2).

Table 1.

Table 1 Dataset Profiles And Identifier Use.

Dataset	Rows	Cols.	Archive (MiB)	Primary outcome (%)	Secondary / senior (%)	Randomized / audit (%)	IDs in models
Criteo Uplift v2.1	13,979,592	16	297.0	4.699	0.292	85.000	No identifiers
FairJob	1,072,226	56	181.9	0.698	66.559	9.874	No identifiers

Deterministic Partitions And Preprocessing

SplitMix64 seed 202410 assigned Criteo 70/15/15; modeling used 2M/1M/full-test samples. User hashing produced impression-disjoint FairJob splits. Categorical inputs used leave-one-out target encoding (smoothing 100) and log frequency (Table 2).

Table 2 Experimental Partitions And Modeling Sample Sizes.

Dataset	Partition	Assigned rows	Modeled rows	Positive rate (%)	Treatment / group 1 (%)
Criteo Uplift v2.1	Train assignment	9,781,693	2,000,000	4.699	85.004
Criteo Uplift v2.1	Validation	2,096,891	1,000,000	4.698	84.981
Criteo Uplift v2.1	Test	2,101,008	2,101,008	4.701	85.001
FairJob	Train	738,116	738,116	0.710	49.677
FairJob	Validation	158,211	158,211	0.650	51.611
FairJob	Test	153,380	153,380	0.793	52.003

Note. Excluding 197 contradictory-proxy users removed 22,519 rows from modeling; the source profile retained them.

Criteo Effect Learners And Calibration

Seven rankings included random control. S-linear used averaged logistic SGD with treatment interactions (alpha 2×10^{-6}; 60 epochs); S-HGB used one surface, T-HGB separate arms, X-HGB imputed effects, and DR-HGB a clipped [-0.25,0.25] pseudo-outcome. HGB used rate 0.08, 65 iterations, 31 leaves, L2=1, early stopping, and minimum leaves 150/100.

An isotonic [-0.10,0.10] map fitted 500,000 validation pseudo-outcomes. CECE used ten bins; AUUC integrated inverse-propensity uplift; Qini removed random allocation. Treat-none anchored value; 200 recentered 150,000-record bootstraps supplied intervals. Regime-shift forecasting [62] likewise separates ranking from uncertainty.

Conservative advertising policy

Validation Qini selected the learner. Contiguous deciles with uplift-1.96SE>0 were locked for test; fixed 5%–100% budgets formed Figure 4.

Fairjob Models, Policy Selection, And Audits

FairJob compared random, elastic-net, ID-free, reweighted, random-display, and proxy-penalized HGB. HGB used rate 0.08, 80 iterations, 31 leaves, leaf minimum 120, L2=1, and early stopping. A 500,000-row proxy achieved AUC 0.53487. Scores subtracted lambda d(ahat-mean(ahat)), lambda in {0,0.1,0.25,0.5,1,1.5}, without proxy inference. Feasibility required score/ECE gaps ≤0.0001/0.001 and MRR ≥98% baseline; maximum MRR won with smaller-lambda ties. Cold-start [63] and calibrated refusal [64] motivate caution.

Metrics were ROC/PR-AUC, loss, Brier, ECE, top-1% gaps, and randomized MRR, hit@1, NDCG@3, senior share, and group gaps. Imbalance studies [65], [66] motivate PR-AUC/threshold audits. The 104 clicked impressions used 200 user-cluster bootstraps; fairness failure precluded deployment.

Evidence Cards And Reproducibility

Median replacement on 1,000 cases per dataset identified three largest score changes; concentration and decision flips measured faithfulness. Rendering barred invented semantics, identities, or mechanisms. Refusal [67], rationales [68], verification [69], risk cards [70], guardrails [71], and claim ranking [72] are parallel controls. Tables 12–13 and Figure 9 report diagnostics..

C. RESULTS AND DISCUSSION

Assignment Effects And Criteo Uplift Ranking

Across 11,882,655 treated and 2,096,937 control records, visit rates 4.85434%/3.82010% yielded 1.03424-point ITT; conversion ITT was 0.11519 points, and all 428,212 exposures followed assignment (Table 3).

Table 3 Criteo Randomized-Assignment Outcome Rates.

Outcome	Treated (%)	Control (%)	Difference (pp)
visit	4.854	3.820	1.034
conversion	0.309	0.194	0.115
exposure	3.604	0.000	3.604

Validation selected S-HGB (Qini 5.09582×10⁻³). Test Qini was 4.68923×10⁻³, top-decile uplift 6.74603 points, and top-20% gain 0.97428 points; competing results are in Table 4 and Figure 2. Treated-response ranking was noncausal and excluded.

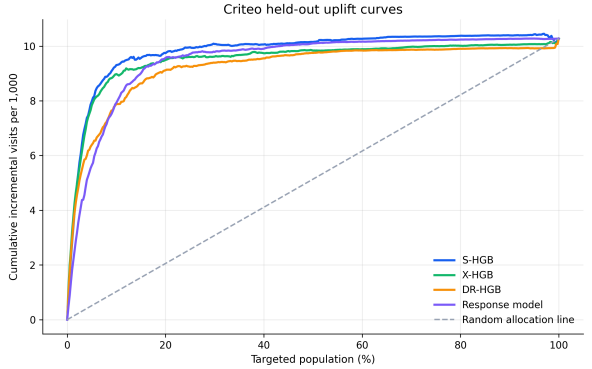

Figure 2 Criteo Cumulative Uplift Curves On The Held-Out Test Set.

Table 4 Criteo Uplift Ranking And Held-Out Policy Performance.

Model	Validation Qini x 1,000	AUC x 1,000	Test Qini x 1,000	Top-10% visit uplift (pp)	Top-10% conversion uplift (pp)	Top-20% value (%)	Value 95% CI	Gain vs. none (pp)
S-HGB	5.0958	9.8249	4.6892	6.7746	0.779	4.802	[4.639, 4.994]	0.974
X-HGB	4.8238	9.4968	4.3611	6.4446	0.737	4.777	[4.587, 4.963]	0.949
DR-HGB	4.5704	9.2152	4.0795	6.0777	0.679	4.739	[4.586, 4.909]	0.911
Response model	4.5583	9.4745	4.3388	5.4225	0.855	4.785	[4.674, 4.903]	0.957
T-HGB	4.0626	9.1998	4.0641	6.1156	0.736	4.754	[4.598, 4.926]	0.926
S-linear	3.7759	8.6916	3.5559	5.4408	0.734	4.668	[4.454, 4.849]	0.839
Random	-0.1766	5.1396	0.0039	0.979	0.127	4.015	[3.800, 4.230]	0.187

Calibration And Conservative Advertising Selection

Isotonic calibration raised S-HGB CECE from 0.24909 to 0.30446 points (slope 1.20814), despite improving four nonselected models (Table 5; Figure 3). Ranking and calibration therefore remained distinct.

Table 5 Criteo Causal Calibration Before And After Held-Out Calibration.

Model	Raw CECE (pp)	Calibrated CECE (pp)	Calibration slope
Response model	4.0405	0.1183	0.9166
Random	1.0346	0.1739	0.1539
DR-HGB	1.1021	0.1882	1.1686
X-HGB	0.3183	0.2575	1.2587
S-linear	0.5179	0.2628	1.0588
T-HGB	0.3005	0.2894	1.2218
S-HGB	0.2491	0.3045	1.2081

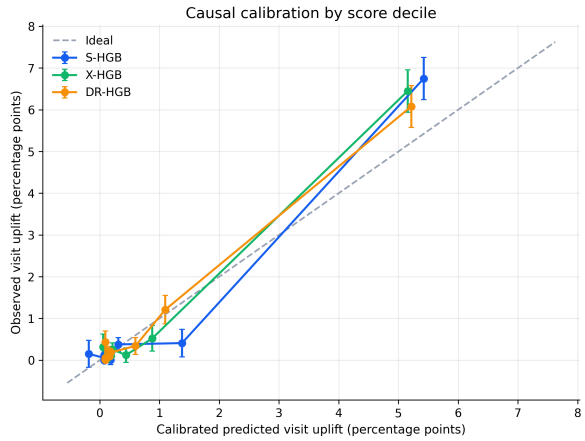

Figure 3 Criteo Predicted-Versus-Observed Causal Calibration.

At 30% reach, the rule achieved 4.83459% value (95% CI 4.68840%–5.01739%) and 1.00655-point gain, 98% of full-treatment gain (Table 6; Figure 4). This visit frontier excludes costs, time, and protected labels; planning [73], placement [74], and drift [42] remain separate constraints. Table 6. Conservative Criteo policies under treatment-budget constraints.

Table 6 Conservative Criteo Policies Under Treatment-Budget Constraints.

Policy	Selected (%)	Visit value (%)	Value 95% CI	Gain vs. none (pp)
Top 5%	5.000	4.626	[4.399, 4.828]	0.798
Top 10%	10.000	4.759	[4.537, 4.985]	0.931
Top 20%	20.000	4.802	[4.601, 4.974]	0.974
Top 30%	30.000	4.835	[4.684, 5.003]	1.007
Top 50%	50.000	4.843	[4.710, 4.994]	1.015
Top 100%	100.000	4.855	[4.756, 4.958]	1.027
LCB-gated deciles	30.000	4.835	[4.688, 5.017]	1.007

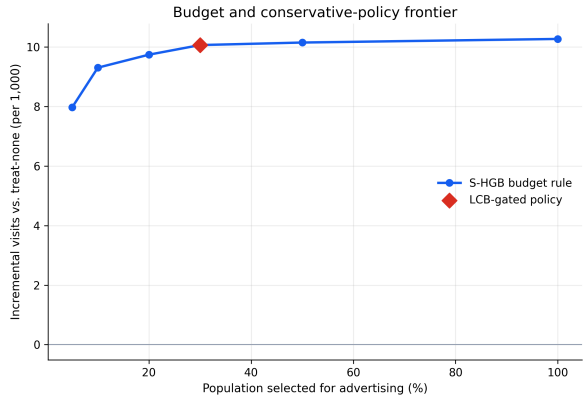

Figure 4 Criteo conservative policy frontier across treatment budgets.

FairJob prediction, ranking, and fairness generalization

FairJob's 7,489 clicks yielded 0.69845% prevalence; proxy rates 0.72932%/0.66758% reversed to 0.71839%/0.72870% under randomized display, while inclusion was nonrandom. Random-display HGB led PR-AUC 0.008328 and linear led ROC-AUC 0.518728. A 1.58×10^{-8} score gap coexisted with 0.00912/0.00249 top-1% TPR/selection gaps (Tables 7–8; Figure 5).

Table 7 Fairjob Click-Prediction And Probability-Calibration Performance.

Model	ROC-AUC	PR-AUC	Log loss	Brier	ECE (15 bins)
Random-display HGB	0.5186	0.0083	0.0464	0.0079	0.0014
ID-free linear	0.5187	0.0083	0.0464	0.0079	0.0014
Proxy-penalized HGB	0.5090	0.0082	0.0464	0.0079	0.0014
Rank-aware audit	0.4966	0.0080	0.0464	0.0079	0.0014
Random ranking	0.4944	0.0079	0.0464	0.0079	0.0014
ID-free HGB	0.4969	0.0079	0.0464	0.0079	0.0014
Protected-aware audit	0.4964	0.0079	0.0464	0.0079	0.0014
Reweighted HGB	0.4964	0.0078	0.0464	0.0079	0.0014

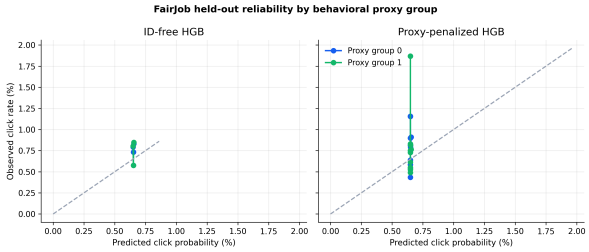

Figure 5 Fairjob Calibration By Protected-Proxy Group.

Table 8 Fairjob Groupwise Predictive And Selection-Rate Fairness Audit.

Model	PR-AUC G0	PR-AUC G1	ECE G0	ECE G1	Mean-score gap	Top-1% TPR gap	Selection-rate gap
Random ranking	0.0072	0.0090	0.0014	0.0014	1.32e-06	0.017	5.04e-04
ID-free linear	0.0088	0.0080	0.0014	0.0015	7.01e-05	0.002	0.001
ID-free HGB	0.0081	0.0079	0.0014	0.0014	1.58e-08	0.009	0.002
Reweighted HGB	0.0079	0.0078	0.0014	0.0014	5.10e-07	0.004	0.001
Random-display HGB	0.0083	0.0086	0.0014	0.0014	4.17e-06	0.003	6.72e-04
Proxy-penalized HGB	0.0092	0.0074	0.0014	0.0014	7.92e-08	0.009	0.002

Model	PR-AUC G0	PR-AUC G1	ECE G0	ECE G1	Mean-score gap	Top-1% TPR gap	Selection-rate gap
Rank-aware audit	0.0083	0.0079	0.0014	0.0014	1.97e-07	0.006	0.002
Protected-aware audit	0.0080	0.0078	0.0014	0.0014	2.10e-08	0.009	0.002

Proxy-penalized HGB reached test MRR 0.58535 (95% CI 0.49861–0.66311) versus 0.57374 baseline/0.52394 random. Validation MRR rose 0.59399 to 0.63041 at lambda=0.10, but test group MRR disparity rose 0.02260 to 0.07083 and senior disparity 0.02907 to 0.03340; deployment was rejected (Tables 9–10; Figures 6–7).

Table 9 Fairjob Ranking On Randomized-Display Test Impressions.

Model	Clicked impr.	MRR	MRR 95% CI	Hit@1	NDCG@3	MRR gap	Senior top-1 gap
Proxy-penalized HGB	104	0.5854	[0.499, 0.663]	0.442	0.5536	0.0708	0.033
Reweighted HGB	104	0.5785	[0.503, 0.655]	0.423	0.5609	0.0342	0.030
Protected-aware audit	104	0.5785	[0.503, 0.655]	0.423	0.5609	0.0342	0.029
ID-free HGB	104	0.5737	[0.500, 0.646]	0.413	0.5573	0.0226	0.029
Rank-aware audit	104	0.5737	[0.500, 0.646]	0.413	0.5573	0.0226	0.033
Random-display HGB	104	0.5732	[0.515, 0.641]	0.385	0.5733	0.0287	0.030
Random ranking	104	0.5239	[0.457, 0.601]	0.356	0.5103	0.1036	0.047
ID-free linear	104	0.5105	[0.439, 0.588]	0.327	0.4888	0.0155	0.021

Table 10 Validation Sensitivity To The Fairness-Penalty Coefficient.

Lambda	Val. ROC-AUC	Val. PR-AUC	Score gap	ECE gap	Randomized MRR	MRR gap	Selected
0.000	0.4971	0.0066	8.19e-05	7.57e-04	0.5940	0.0489	No
0.100	0.5183	0.0069	8.23e-05	7.56e-04	0.6304	0.0312	Yes
0.250	0.5178	0.0068	8.28e-05	7.55e-04	0.6304	0.0312	No
0.500	0.5176	0.0068	8.38e-05	7.54e-04	0.6304	0.0312	No
1.000	0.5175	0.0068	8.56e-05	7.52e-04	0.6304	0.0312	No
1.500	0.5177	0.0068	8.73e-05	7.49e-04	0.6304	0.0312	No

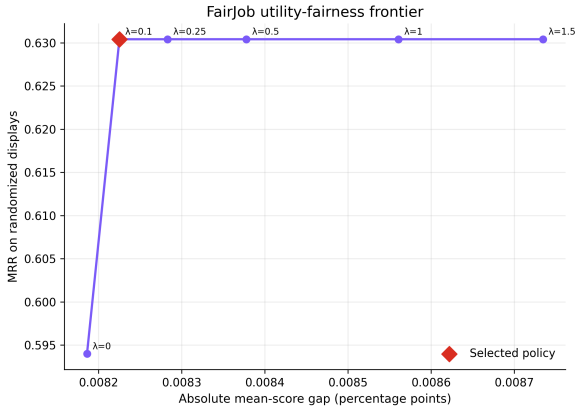

Figure 6 Fairjob Utility-Fairness Pareto Frontier.

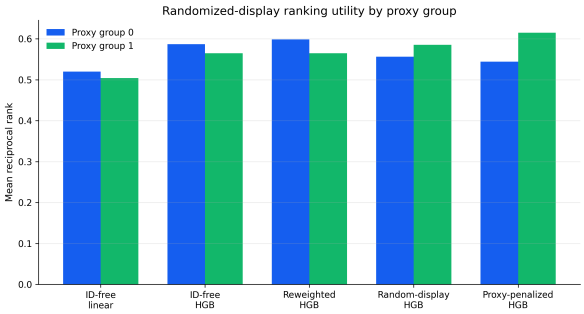

Figure 7 Fairjob Randomized-Display Ranking Performance By Proxy Group.

Ablations yielded MRR 0.57320–0.57855 (Table 11); importance concentrated weakly in anonymized fields, so no occupational meaning was assigned (Figure 8). Validation gains could not override held-out safety failure.

Table 11 Fairjob Model And Feature-Use Ablations.

Ablation	ROC-AUC	PR-AUC	Log loss	Brier	ECE	Randomized MRR	Score gap
ID-free HGB	0.4969	0.0079	0.0464	0.0079	0.0014	0.5737	1.58e-08

Ablation	ROC-AUC	PR-AUC	Log loss	Brier	ECE	Randomized MRR	Score gap
Rank-aware audit	0.4966	0.0080	0.0464	0.0079	0.0014	0.5737	1.97e-07
Protected-aware audit	0.4964	0.0079	0.0464	0.0079	0.0014	0.5785	2.10e-08
Random-display HGB	0.5186	0.0083	0.0464	0.0079	0.0014	0.5732	4.17e-06
Reweighted HGB	0.4964	0.0078	0.0464	0.0079	0.0014	0.5785	5.10e-07
Proxy-penalized HGB	0.5090	0.0082	0.0464	0.0079	0.0014	0.5854	7.92e-08

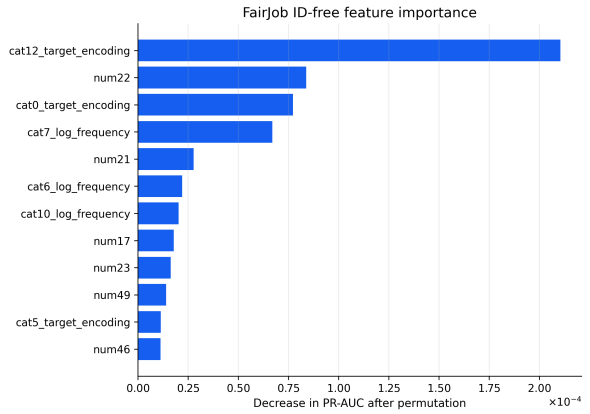

Figure 8 Fairjob Identifier-Free Permutation Feature Importance.

Identifier removal limits memorization, not exposure disparity. FairJob's proxy is unverified, only 104 clicked test impressions had randomized position, and inclusion remained nonrandomized. Criteo supports benchmark ITT allocation, not temporal, profit, or fairness claims; only its lower-bound gate passed.

Explanation diagnostics and computational cost

All 2,000 cards passed schema checks. Criteo top-three perturbations explained 90.9267% of change with 90.4% flips; FairJob achieved 74.2208%, median 5.59×10^{-7} change, and 4.0% flips (Table 12; Figure 9). Evidence cards [75], review-grounded recommendation [57], RAG provenance [52], and trajectory reliability [76] motivate traceability, not perceived quality.

Table 12.

Table 12 Validation of evidence-grounded explanation cards.

Dataset	Cards	Schema pass (%)	Top-3 evidence concentration (%)	Median ablation change	Top-3 decision flips (%)
Criteo Uplift v2.1	1,000	100.000	90.927	0.01681473	90.400
FairJob	1,000	100.000	74.221	0.00000056	4.000

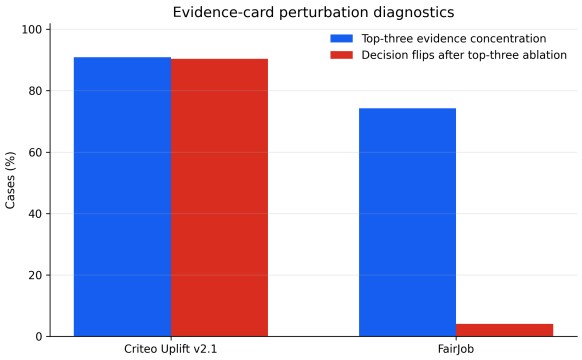

Figure 9 Explanation-Card Evidence Concentration And Perturbation Diagnostics.

Anomaly detection [77], routing [78], and host narratives [79] reinforce the boundary: cards may report scores, thresholds, perturbations, and decisions but cannot invent semantics.

The 267.06-second CPU pipeline processed 15,051,818 records without identifiers or accelerators; Table 13 reports stage runtimes.

Table 13 Model-Fitting Runtime And Training Footprint.

Model / stage	Fit time (s)	Records processed	Implementation note
Criteo S-linear	12.93	2,000,000	SGD logistic model with treatment interactions
Criteo S-HGB	13.71	2,000,000	single nonlinear response surface
Criteo T-HGB	16.35	2,000,000	separate treated and control response surfaces
Criteo X-HGB	6.62	2,000,000	imputed effects with propensity-weighted combination
Criteo DR-HGB	8.58	2,000,000	doubly robust visit pseudo-outcome regression
FairJob ID-free linear	4.96	738,116	no user, impression, product, rank, or protected field
FairJob ID-free HGB	8.08	738,116	target/frequency encodings learned from training records
FairJob reweighted HGB	28.22	738,116	capped inverse protected-group/click-cell weights
FairJob random-display HGB	1.11	73,012	trained only on randomized banner positions
FairJob proxy-penalized HGB	16.35	500,000	validation-selected lambda=0.10
FairJob audit ablations	39.18	738,116	rank-aware and direct-proxy models are not serving candidates
Complete pipeline	267.06	15,051,818	source records read; Python 3.12.13; scikit-learn 1.8.0

D. CONCLUSIONS

Only conservative selection transferred. Criteo supported randomized ITT: S-HGB ranked uplift best, and the validation gate retained three deciles while recovering nearly all full-treatment gain; calibration remained separate. FairJob supported identifier-free prediction and randomized-position auditing, not causal effects. Its selected penalty improved MRR but worsened held-out disparity, requiring nondeployment.

Evidence cards constrained prose to scores, thresholds, and anonymized perturbations, without replacing causal, fairness, or human review. Future work needs costs, repeated exposure, validated demographics, larger randomized samples, and prospective drift tests; confidence gating and abstention remain essential.

E. References

[1] J. Mu, Y. Lu, and M. Smith, "LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies," J. Adv. Comput. Syst., vol. 3, no. 1, pp. 31–48, Jan. 2023, doi: 10.69987/JACS.2023.30103.

[2] P. Gutierrez and J.-Y. Gérardy, "Causal inference and uplift modelling: A review of the literature," in Proc. 3rd Int. Conf. Predictive Applications and APIs, PMLR, vol. 67, pp. 1–13, 2017.

[3] 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," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 17–38, Apr. 2026, doi: 10.51903/jtie.v5i1.468.

[4] M. Vladimirova, F. Pavone, and E. Diemert, "FairJob: A Real-World Dataset for Fairness in Online Systems," in Advances in Neural Information Processing Systems 37, 2024, doi: 10.5555/3737916.3738250.

[5] E. Diemert, A. Betlei, C. Renaudin, and M.-R. Amini, "A large scale benchmark for uplift modeling," in Proc. AdKDD and TargetAd Workshop, KDD, London, U.K., 2018. [Online]. Available: https://ailab.criteo.com/criteo-uplift-prediction-dataset/.

[6] Criteo AI Lab, "FairJob dataset card," Hugging Face, Jul. 2024. [Online]. Available: https://huggingface.co/datasets/criteo/FairJob. Accessed: Aug. 4, 2026.

[7] E. Diemert, A. Betlei, C. Renaudin, M.-R. Amini, T. Gregoir, and T. Rahier, "A large scale benchmark for individual treatment effect prediction and uplift modeling," arXiv preprint arXiv:2111.10106, 2021. [Online]. Available: https://arxiv.org/abs/2111.10106.

[8] Criteo AI Lab, "Upload criteo-research-uplift-v2.1.csv.gz," verified commit 8281178, Hugging Face, Jun. 12, 2024. [Online]. Available: https://huggingface.co/datasets/criteo/criteo-uplift/commit/82811785048bb633de2d55c02bab4e57066e6423. Accessed: Aug. 4, 2026.

[9] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," Proc. Natl. Acad. Sci. U.S.A., vol. 116, no. 10, pp. 4156–4165, 2019, doi: 10.1073/pnas.1804597116.

[10] E. H. Kennedy, "Towards optimal doubly robust estimation of heterogeneous causal effects," Electron. J. Statist., vol. 17, no. 2, pp. 3008–3049, 2023, doi: 10.1214/23-EJS2157.

[11] F. Devriendt, J. Van Belle, T. Guns, and W. Verbeke, "Learning to rank for uplift modeling," IEEE Trans. Knowl. Data Eng., vol. 34, no. 10, pp. 4888–4904, 2022, doi: 10.1109/TKDE.2020.3048510.

[12] 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)," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 521–543, Dec. 2025, doi: 10.51903/jtie.v4i3.500.

[13] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," in Proc. 28th Int. Conf. Machine Learning, pp. 1097–1104, 2011.

[14] T. Ye, J. Mu, and J. Hunter, "Off-Policy Evaluation and Conservative Policy Selection for Slot-Level Dynamic Bidding and Ranking on the Open Bandit Dataset (Small)," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 178–199, Apr. 2026, doi: 10.51903/jtie.v5i1.503.

[15] H. Zhang, "Counterfactual Learning-to-Rank for Ads: Off-Policy Evaluation on the Open Bandit Dataset," J. Adv. Comput. Syst., vol. 5, no. 12, pp. 1–11, Dec. 2025, doi: 10.69987/JACS.2025.51201.

[16] Y. Xu and S. Yadlowsky, "Calibration error for heterogeneous treatment effects," in Proc. 25th Int. Conf. Artificial Intelligence and Statistics, PMLR, vol. 151, pp. 9280–9303, 2022.

[17] L. van der Laan, E. Ulloa-Pérez, M. Carone, and A. Luedtke, "Causal isotonic calibration for heterogeneous treatment effects," in Proc. 40th Int. Conf. Machine Learning, PMLR, vol. 202, pp. 34831–34854, 2023.

[18] O. Nyberg, T. Kuśmierczyk, and A. Klami, "Uplift modeling with high class imbalance," in Proc. 13th Asian Conf. Machine Learning, PMLR, vol. 157, pp. 315–330, 2021.

[19] Y. Chen, Y. Zhang, D. Chau, and M. Sherman, "Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 31–47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[20] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215–238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[21] J. Jin, T. Huang, and S. Lu, "A Model-Risk-Friendly Probability of Default Workflow: Calibration,

Distribution-Free Uncertainty Quantification, and SHAP Explanations on the UCI Credit Card Default Dataset," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 74–85, Jun. 2024, doi: 10.69987/JACS.2024.40606.

A. Swaminathan and T. Joachims, "Counterfactual risk minimization: Learning from logged bandit feedback," in Proc. 32nd Int. Conf. Machine Learning, PMLR, vol. 37, pp. 814–823, 2015.

[22] S. Athey and S. Wager, "Policy learning with observational data," Econometrica, vol. 89, no. 1, pp. 133–161, 2021, doi: 10.3982/ECTA15732.

[23] R. Laroche, P. Trichelair, and R. Tachet des Combes, "Safe policy improvement with baseline bootstrapping," in Proc. 36th Int. Conf. Machine Learning, PMLR, vol. 97, pp. 3652–3661, 2019.

A. Betlei, M. Vladimirova, M. Sebbar, N. Urien, T. Rahier, and B. Heymann, "Maximizing the success probability of policy allocations in online systems," in Proc. AAAI Conf. Artificial Intelligence, vol. 38, no. 10, pp. 11061–11068, 2024, doi: 10.1609/aaai.v38i10.28982.

[24] H. Zhang, "Risk-Aware Budget-Constrained Auto-Bidding Under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.

[25] Y. Lu, H. Zhou, and Y. Zhang, "A Constrained, Data-Driven Budgeting Framework Integrating Macro Demand Forecasting and Marketing Response Modeling," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 493–520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[26] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems 29, pp. 3315–3323, 2016.

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in Proc. 35th Int. Conf. Machine Learning, PMLR, vol. 80, pp. 60–69, 2018.

[27] Singh and T. Joachims, "Fairness of exposure in rankings," in Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 2219–2228, 2018, doi: 10.1145/3219819.3220088.

[28] J. Biega, K. P. Gummadi, and G. Weikum, "Equity of attention: Amortizing individual fairness in rankings," in Proc. 41st Int. ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 405–414, 2018, doi: 10.1145/3209978.3210063.

[29] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in Proc. 35th Int. Conf. Machine Learning, PMLR, vol. 80, pp. 1929–1938, 2018.

[30] P. Lahoti, A. Beutel, J. Chen, K. W. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, "Fairness without demographics through adversarially reweighted learning," in Advances in Neural Information Processing Systems 33, pp. 728–740, 2020.

A. Lambrecht and C. Tucker, "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads," Management Science, vol. 65, no. 7, pp. 2966–2981, 2019, doi: 10.1287/mnsc.2018.3093.

B. Imana, A. Korolova, and J. Heidemann, "Auditing for discrimination in algorithms delivering job ads," in Proc. Web Conf. 2021, pp. 3767–3778, 2021, doi: 10.1145/3442381.3450077.

[31] E. Celis, A. Mehrotra, and N. Vishnoi, "Toward controlling discrimination in online ad auctions," in Proc. 36th Int. Conf. Machine Learning, PMLR, vol. 97, pp. 4456–4465, 2019.

[32] M. Sebbar, C. Odic, M. Léchine, A. Bissuel, N. Chrysanthos, A. D'Amato, A. Gilotte, F. Höring, S. Nogueira, and M. Vono, "CriteoPrivateAd: A real-world bidding dataset to design private advertising systems," arXiv preprint arXiv:2502.12103, 2025.

[33] W3C Private Advertising Technology Working Group, "Privacy-Preserving Attribution: Level 1," W3C First Public Working Draft, Apr. 22, 2025. [Online]. Available: https://www.w3.org/TR/2025/WD-privacy-preserving-attribution-20250422/. Accessed: Aug. 4, 2026.

[34] J. Bai and Q. Wu, "Privacy-Safe Marketing Mix Modeling and Budget Optimization Under Identifier Loss: A Controlled Simulation Study," Int. J. Electron. Commun. Syst., vol. 6, no. 1, pp. 83–95, Jun. 2026, doi: 10.24042/ijecs.v6i1.30533.

[35] Zhou, J. Jin, and D. Zhao, "Calibrated Resume-Job Matching for Trustworthy LLM-Assisted Recruiter Screening: Pairwise Matching, Probability Calibration, and Selective Refusal on Two Public Recruitment Datasets," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 625–648, Dec. 2025, doi: 10.51903/jtie.v4i3.529.

[36] Q. Xin, "Self-Supervised Customer Representation Learning for Segmentation and Next-Purchase Prediction on UCI Online Retail," J-INTECH (J. Inf. Technol.), vol. 14, no. 1, pp. 20–37, Apr. 2026, doi: 10.32664/j-intech.v14i01.2229.

[37] H. Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models," J. Adv. Comput. Syst., vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[38] Q. Xin, "Behavior Retrieval Plus Response Generation for Interpretable Conversational Personalized Recommendation," Int. J. Electr., Energy Power Syst. Eng., vol. 9, no. 2, pp. 120–136, Jul. 2026, doi: 10.31258/ijeepse.9.2.120-136.

[39] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, pp. 4765–4774, 2017.

[40] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent Classification and Personalized Recommendation of E-Commerce Products Based

on Machine Learning," Appl. Comput. Eng., vol. 64, 2024, doi: 10.54254/2755-2721/64/20241365.

A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" in Proc. 58th Annu. Meeting Assoc. Computational Linguistics, pp. 4198–4205, 2020, doi: 10.18653/v1/2020.acl-main.386.

[41] P. Atanasova, O.-M. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, and I. Augenstein, "Faithfulness tests for natural language explanations," in Proc. 61st Annu. Meeting Assoc. Computational Linguistics, Vol. 2: Short Papers, pp. 283–294, 2023, doi: 10.18653/v1/2023.acl-short.25.

[42] H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter, "Measuring attribution in natural language generation models," Computational Linguistics, vol. 49, no. 4, pp. 777–840, 2023, doi: 10.1162/coli_a_00486.

[43] Li, J. Bai, and S. Wang, "Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications," J. Adv. Comput. Syst., vol. 4, no. 2, pp. 76–92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[44] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502–520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[45] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," Emerg. Inf. Sci. Technol., vol. 6, no. 2, pp. 125–146, Nov. 2025, doi: 10.18196/eist.v6i2.31232.

[46] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520–533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

[47] G. Liu, C. Li, and E. Zhang, "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 97–111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[48] Zhang, H. Rao, and D. Zhao, "Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents," J. Adv. Comput. Syst., vol. 4, no. 3, pp. 109–125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[49] Q. Xin, "Log Anomaly Detection With Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," AVITEC, vol. 8, no. 2, pp. 247–264, Aug. 2026, doi: 10.28989/avitec.v8i2.3974.

[50] J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-Constrained Incident Visualization Cards for Distributed Cloud Logs: A UI/UX Framework for Turning Hadoop, OpenStack, and ZooKeeper Logs into Actionable SRE Design Interfaces," Int. J. Graphic Design, vol. 4, no. 1, pp. 179–185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[51] X. Chang, Y. Lu, and Z. S. Zhong, "Review-Grounded Explainable Recommendation with Faithfulness Evaluation on Amazon Reviews," J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 9–22, Jun. 2026, doi: 10.54732/jeecs.v11i1.2.

[52] Zhang, Y. Ren, and J. Zou, "LLM-Style Explainable E-Commerce Recommendation Cards: A UI/UX Design Framework for Trust-Calibrated Product Recommendation," Int. J. Graphic Design, vol. 3, no. 2, pp. 381–396, Oct. 2025, doi: 10.51903/ijgd.v3i2.3697.

[53] Q. Xin, "Auditable Automated Essay Scoring and Formative Feedback: A Rubric-Grounded Pipeline for Secondary and Higher Education," J. Appl. Artif. Intell. Educ., vol. 2, no. 1, pp. 1–19, Jul. 2026, doi: 10.66053/jaaie.v2i1.348.

[54] Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," Int. J. Graphic Design, vol. 3, no. 1, pp. 216–240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[55] J. Mu, Y. Lu, and E. Hwang, "Structured Visual Brief Interfaces for Advertising Design: A UI/UX Framework for Turning Creative Intentions into Designer-Editable Graphic Design Cards," Int. J. Graphic Design, vol. 4, no. 1, pp. 192–208, Apr. 2026, doi: 10.51903/ijgd.v4i1.3702.

[56] Q. Xin, "Probabilistic Bike-Sharing Demand Forecasting Under Changing Weather and Seasonal Regimes With Transformer-Based Models," Findings, Mar. 2026, doi: 10.32866/001c.157499.

[57] S. He, C. Li, and H. Rao, "Few-Shot Cold-Start Workload Forecasting for New AI Inference Tenants with Time-Series Foundation Models," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 306–324, Apr. 2025, doi: 10.51903/jtie.v4i1.546.

[58] Q. Xin, "Explainable and Fair Credit Risk Scoring With Counterfactual Explanations: A Reproducible Evaluation on the German Credit Dataset (HELOC-Motivated)," J-INTECH (J. Inf. Technol.), vol. 14, no. 2, pp. 215–231, Jun. 2026, doi: 10.32664/j-intech.v14i02.2228.

[59] J. Jin, T. Huang, and S. Lu, "Cost-Sensitive Learning, Simulated PU Learning, and One-Class Autoencoding for Extreme-Imbalance Credit Card Fraud Detection," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 64–73, Jun. 2024, doi: 10.69987/JACS.2024.40605.

[60] Y. Chen, Y. Zhang, and M. Sherman, "Going Concern and Bankruptcy Prediction under Extreme Class Imbalance: Cost-Sensitive Learning, Resampling, and Focal Loss with Explainable Financial-Ratio Portraits," J. Adv. Comput. Syst.,

vol. 4, no. 4, pp. 80–96, Apr. 2024, doi: 10.69987/JACS.2024.40407.

[61] Zheng and C. Li, "Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 83–99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[62] Q. Xin, "Early-Warning Analytics With LLM Intervention Rationales for Student Retention Decisions: Classroom Interaction Modeling With xAPI-Edu-Data and Dropout/Success Prediction," Interdiscip. J. Pedagogy Res. Media Technol., vol. 2, no. 1, pp. 9–27, Jun. 2026, doi: 10.64268/inspire.v2i1.117.

[63] Zheng, B. Zhang, and J. Geibel, "VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 67–82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[64] W. Su, H. Rao, and E. Ma, "Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight under Prompt-Injection Attacks," Int. J. Graphic Design, vol. 4, no. 1, pp. 186–191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[65] Z. Li, K. Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and FRED Data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75–90, Aug. 2026, doi: 10.51903/jtie.v5i2.541.

[66] W. Su, S. Chen, and E. Qian, "Narrative-Aware Scientific Claim Verification Agent with Evidence Ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327–340, Apr. 2026, doi: 10.51903/jtie.v5i1.549.

[67] S. He, J. Nie, and C. Li, "Power-Aware Inventory Planning for AI Infrastructure Using Job-Level Forecasting and LLM Workload Explanations," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341–359, Apr. 2026, doi: 10.51903/jtie.v5i1.548.

[68] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment," J. Inf. Syst. Informatics, vol. 7, no. 3, pp. 2182–2195, Sep. 2025, doi: 10.51519/journalisi.v7i3.1170.

[69] J. Jin, "LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy," Int. J. Graphic Design, vol. 3, no. 2, pp. 397–414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[70] Z. S. Zhong, C. Li, and H. Rao, "Trajectory Reliability Prediction for Generalist AI Agents: Tool-Use Failure Analysis and Success Forecasting on ZClawBench," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341–360, Apr. 2026, doi: 10.51903/jtie.v5i1.539.

[71] Q. Xin, "Self-Supervised Log Anomaly Detection With LogBERT-Style Transformers: Full Empirical Evaluation on a Reproducible SynHDFS Benchmark," J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 23–35, Jun. 2026, doi: 10.54732/jeecs.v11i1.3.

[72] Li, G. Liu, and Z. Zhao, "Cost-Aware LLM-Style Routing for AIOps Log Analysis: Log Parsing, Anomaly Detection, Fault Diagnosis, and Incident Summarization on LogEval Task Files," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 91–103, Jun. 2026, doi: 10.51903/jtie.v5i2.538.

[73] Q. Xin, "Host-Based Intrusion Detection With System Call Sequences: Window Localization and Forensic Narratives," AVITEC, vol. 8, no. 2, pp. 325–334, Aug. 2026, doi: 10.28989/avitec.v8i2.3973.