

Disagreement reliability routing for multimodal large language model evidence pipelines with imperfect counterfactual priors

Nataraj Agaram Sundar

natsundar@ebay.com

eBay (United States)

Tejas Morabia

eBay (United States)

Research Article

Keywords: multimodal LLMs, evidence routing, counterfactual reliability, cost-sensitive learning, probabilistic prediction, selective decision systems, production telemetry, reliability

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10963637/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. The authors are employees of eBay Inc. The authors declare no other competing financial or non-financial interests.

Disagreement reliability routing for multimodal large language model evidence pipelines with imperfect counterfactual priors

Nataraj Agaram Sundar and Tejas Morabia

eBay Inc., San Jose, CA 95125, USA

Correspondence: natsundar@ebay.com

Research article

Abstract

Multimodal large-language-model (LLM) evidence pipelines often attach rule-derived outcomes to candidate remediation actions even though only the executed action yields a factual production outcome. Treating those counterfactual rules as ground truth can miscalibrate routing, while discarding them wastes structured operational knowledge. We introduce EviRoute-DR, a disagreement-reliability routing method that learns when rule priors fail and combines corrected-rule experts with direct factual outcome experts through a reliability-conditioned meta-learner. The study uses 2,100 production cases, seven candidate actions per case (14,700 case-action records), and exactly one observed production outcome per case. Base models are trained on 600 factual cases, combined on 200, and evaluated on 1,300 held-out factual cases spanning main, temporal-labelled, hard, and workflow-shift cohorts. On factual outcomes, EviRoute-DR achieves AUC 0.926 and Brier score 0.0910, compared with 0.921/0.0971 for direct logistic regression and 0.924/0.0922 for a random-forest baseline. The paired bootstrap Brier improvement over logistic regression is 0.0060 (95% CI 0.0021–0.0092). A separate production-derived counterfactual replay shows a stronger cost-resolution frontier at moderate operating costs, reaching approximately 0.815 resolution near 12 cost units. Because the logged action is always the argmax of the production score vector, we do not interpret those scores as stochastic propensities and make no inverse-propensity, doubly robust causal, or production-effect claim. The results support disagreement-aware reliability fusion as a practical methodology for evidence routing under partial factual feedback.

Keywords: multimodal LLMs; evidence routing; counterfactual reliability; cost-sensitive learning; probabilistic prediction; selective decision systems; production telemetry; reliability

1 Introduction

Evidence-producing LLM systems increasingly combine retrieval, optical character recognition (OCR), generation, policy checks, and human review. Operational failures are not homogeneous: evidence can be missing, retrieval channels can conflict, OCR confidence can degrade, and corrective actions can have sharply different costs. A routing layer must therefore decide not only whether a case is uncertain, but also which intervention is justified.

Existing LLM routing work primarily studies selection among models or experts under performance-cost trade-offs, often using preference labels, full-information supervision, or bandit feedback [1, 2]. Offline policy evaluation provides principled tools when the historical behavior policy and outcome support are known [3, 2, 4, 5]. These assumptions are not automatically satisfied in evidence-remediation systems: production logs commonly observe only one executed action, while the remaining actions are represented by rule-derived counterfactual outcomes.

This paper studies that asymmetry directly. The central premise is that a rule-derived action outcome can be useful without being treated as factual truth. EviRoute-DR learns the error of the rule prior on executed production actions, converts that error model into a corrected action-success probability, and then combines corrected-rule and direct factual experts with reliability features. Here “DR” denotes disagreement-reliability; it does not denote the classical doubly robust estimator.

The methodological contributions are:

1. a formal partial-feedback routing problem in which one factual action outcome coexists with imperfect rule-derived counterfactual priors;
2. a disagreement correction $q_\theta(x, a)$ that estimates where a rule prior is likely to be wrong, and a closed-form corrected success probability;
3. a reliability-conditioned stacker that combines direct factual experts, rule-correction experts, missingness, conflict, logged policy score, and action cost;
4. a partition-disjoint production evaluation separating factual held-out prediction from a secondary production counterfactual replay; and
5. explicit abstention from causal OPE when the logs do not establish stochastic selection from the score vector.

2 Related work

2.1 Routing and selective decision systems

RouteLLM learns preference-based routers to trade response quality against model cost [1]. The unpublished preprint by Wei et al., *Learning to Route LLMs from Bandit Feedback*:

One Policy, Many Trade-offs (2025; arXiv:2510.07429), emphasizes that deployment exposes only the selected model outcome. It is identified here as a preprint rather than included in the published-reference list. Selective prediction similarly studies systems that abstain under uncertainty rather than force a low-confidence prediction [6]. EviRoute-DR differs in target: the action is a remediation step in an evidence pipeline, and the central supervision problem is disagreement between an imperfect operational prior and the single factual executed-action outcome.

2.2 Calibration and uncertainty

Probability calibration is essential when predictions feed downstream decisions [7]. Deep ensembles and related uncertainty methods capture predictive variability but do not by themselves specify how to use a structured operational counterfactual prior [8]. EviRoute-DR treats missingness and signal conflict as reliability variables in a meta-level combination of factual and corrected-rule experts.

2.3 Offline policy evaluation with imperfect counterfactuals

Under identification and overlap assumptions, classical doubly robust evaluation combines a reward model with behavior-policy weighting and is consistent if either nuisance component is correctly specified [3]. Counterfactual risk minimization learns from logged bandit feedback under propensity assumptions [2]. CANDOR studies imperfect counterfactual annotations inside doubly robust OPE [4], and Carranza and Athey study robust policy learning under heterogeneous observational sources [5]. These methods motivate the partial-feedback framing, but the present production log always records the maximum-score action; consequently, we do not reinterpret its scores as randomized propensities.

3 Problem formulation

For case i , let x_i denote pre-decision telemetry and let $\mathcal{A}$ be the seven-action set. The supplied production export provides a rule prior $r_i(a) \in \{0, 1\}$ and action cost $c_i(a)$ for every action. A separate replay annotation $\tilde{Y}_i(a)$ is distinct from that rule prior; it is an evaluation endpoint, not a training label. Exactly one action A_i is executed, yielding factual resolution outcome $Y_i(A_i) \in \{0, 1\}$. For $a \neq A_i$, the available outcome is a production-derived counterfactual annotation and is not used as a factual training label.

The factual disagreement target is

$$D_i = \mathbf{1}\{Y_i(A_i) \neq r_i(A_i)\}. \quad (1)$$

A disagreement expert estimates

$$q_\theta(x_i, a) = P(D_i = 1 \mid x_i, a). \quad (2)$$

The rule-corrected success probability is then

$$\hat{p}_\theta^R(x, a) = r(x, a)[1 - q_\theta(x, a)] + [1 - r(x, a)]q_\theta(x, a). \quad (3)$$

Equation 3 has an interpretable limiting behavior: when disagreement probability tends to zero, the rule is trusted; when it tends to one, its label is reversed; at $q = 0.5$, the rule contributes no directional information.

4 EviRoute-DR

4.1 Base experts

Two direct factual experts estimate $P(Y = 1 \mid x, a)$ from executed-action rows: regularized logistic regression and random forest. Three disagreement experts estimate q_θ : pooled logistic correction, regime-conditioned logistic correction, and random-forest correction. The regime-conditioned model includes interactions between action, telemetry missingness, and a training-defined high-conflict indicator.

4.2 Reliability-conditioned fusion

Let $z(x, a)$ contain the direct factual probabilities, corrected-rule probabilities, raw rule prior, missing-channel count, conflict score, logged production action score, and candidate action cost. A regularized logistic meta-learner is fit only on the 200-case validation partition:

$$\hat{p}_{\text{ER}}(x, a) = \sigma(\beta_0 + \beta^\top z(x, a)). \quad (4)$$

The separation between base-model training and meta-learning prevents the meta-learner from fitting its own in-sample base predictions. All meta-learner inputs are standardized on the 200 meta-training cases, with logistic inverse regularization $C = 0.1$. This partition is not an independent calibration test. As shown in Fig. 1, the deployed score combines the two expert branches and reliability inputs; it is not an upper-confidence bound.

4.3 Routing score

For a cost-aversion parameter $\lambda \geq 0$, EviRoute-DR selects

$$\pi_\lambda(x) = \arg \max_{a \in \mathcal{A}} [\hat{p}_{\text{ER}}(x, a) - \lambda c(x, a)/20]. \quad (5)$$

We retain the archived grid $\Lambda = \{0\} \cup \text{logspace}(-4, 2, 160)$, the divisor 20, and all fitted primary models. The divisor sets the scale of the cost-aversion parameter; it is not a currency conversion. All 161 operating points are released, and the plotted upper envelope is a descriptive, outcome-dependent replay summary, not a prospectively selected deployment policy.

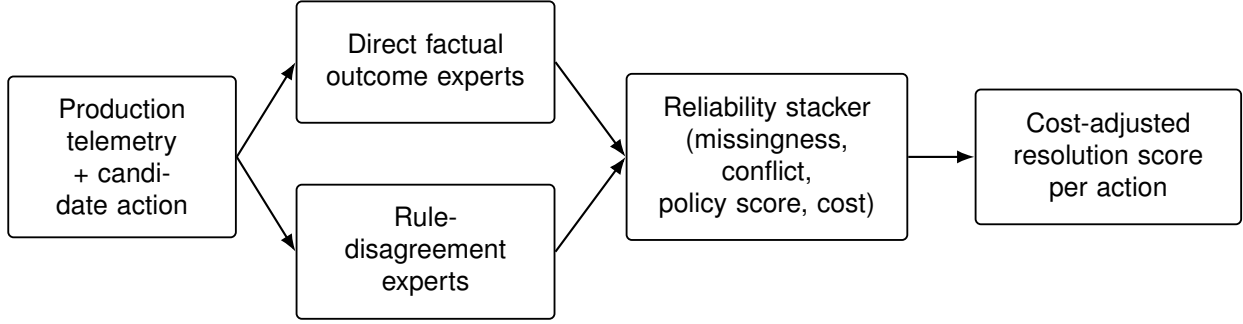


Fig. 1 EviRoute-DR architecture: direct factual experts and rule-disagreement experts feed a reliability-conditioned stacker and a cost-adjusted resolution score. Each training case contributes one executed-action outcome. Non-executed replay outcomes are reserved for evaluation, while the separate rule prior is an input

5 Production data and evaluation protocol

5.1 Data

The dataset contains 2,100 production cases and seven candidate actions per case, for 14,700 case-action rows. Exactly one action per case is marked as factually observed. The action set comprises no action, retrieval retry, OCR retry, regeneration, policy review, human review, and abstention. Pre-decision reliability variables include evidence-channel missingness and signal conflict; the production decision log also provides a seven-action score vector.

Using the latest production outcome file, the factual rule disagreement rate is 17.38% (365/2,100) over executed actions. The score vectors sum to one within numerical tolerance, but the executed action is the score argmax in 100% of cases. This observation alone does not identify the action-generation mechanism. Without documentation of stochastic selection, we use the values only as production policy scores, not behavior propensities. Inverse propensity scoring (IPS), self-normalized IPS (SNIPS), classical doubly robust off-policy evaluation (OPE), and causal policy-value claims are outside the scope of this paper. Reconstructing a deterministic rule would not create support for actions it never selects.

5.2 Recorded endpoint, cost, and annotation provenance

The factual endpoint is the binary `resolution_success` field from the latest `production_potential_outcomes.csv`, restricted to rows marked `potential_outcome_observed=1`. Each recorded observation end is exactly 24 hours after its decision timestamp. We retain the final exported success flag, including the 31 executed cases whose recorded outcomes were corrected by staff through routine operational review (recorded as reviewer overrides), rather than reconstructing success from the rule prior. These corrections were made during

ordinary business operations and were not elicited through research-specific recruitment, surveys, questionnaires, or assigned experimental tasks. The retrospective analysis uses recorded case outcomes rather than evaluating individual staff performance. The export does not supply the internal success-adjudication predicate, exact time of resolution, or an override-policy specification; accordingly, “resolution” denotes this recorded 24-hour endpoint, not a separately validated measure of case closure or monetary recovery. ESM 2 specifies the source hierarchy and these boundaries.

The routing cost is the panel’s `action_cost`, which matches `cost_v1.total_cost` exactly on all 14,700 case–action keys. In dimensionless utility units,

$$c_i(a) = L_i(a) + R_i(a) + T_i(a) + W_i(a) + U_i(a) + A_i(a), \quad (6)$$

where the stored components represent failure loss, review effort, latency, incorrect routing, unnecessary intervention, and abstention, respectively. The alternative `action_cost` column in the outcome file is all zero and is not used. These components are already weighted; we do not apply a second set of coefficients or interpret them as dollars. ESM 2 gives their ranges and action-specific terms. They are fixed throughout this analysis, but the exports do not document when the operational weights were originally chosen.

The rule priors and non-executed replay labels have different roles. The former supply operational structure to the predictors; the latter score frozen policies only in secondary replay. The available metadata identifies contextual annotation using diagnosis, missingness, and conflict, but does not include its executable construction procedure or independent validation. The analysis therefore tests performance against the supplied annotation resource, not counterfactual ground truth. An analytical annotation-error sensitivity in ESM 2 quantifies the dependence on that resource without altering any factual outcome.

5.3 Splits and leakage controls

The supplied DR development protocol defines 600 factual training cases, 200 validation cases, and 1,300 held-out test cases. The test set contains a 200-case main test slice and three additional labelled cohorts: temporal (550), hard (250), and workflow shift (300). No adjudicated failure class, factual outcome, downstream loss, final-resolution field, or post-decision result is exposed as a prediction-time feature. The primary factual test uses only the single observed outcome for each test case. The cohort names are supplied labels, not evidence of an independently verified forward-time split: the logged decision dates overlap across partitions. The inputs are supplied as decision-time features; their availability is an assumption not independently established by per-feature capture timestamps. This distinction is important because the exported cost vectors are constant within diagnostic class. The additional ablations, calibration checks, and annotation sensitivity are post-hoc diagnostics on the same held-out cases, not a new untouched test.

5.4 Baselines and metrics

We compare the raw rule prior, direct logistic regression, direct random forest, rule-correction logistic regression, regime-conditioned correction, rule-correction random forest, and full EviRoute-DR. Primary metrics are area under the receiver operating characteristic curve (AUC), Brier score, and log loss. Brier score and log loss measure probabilistic accuracy; neither isolates calibration. Separate diagnostics use 10 equal-width probability bins, expected calibration error (ECE), and a logistic calibration intercept and slope, without recalibrating the test predictions. Paired case bootstraps quantify differences against the direct baselines. The secondary replay evaluates the full seven-action production counterfactual ledger only after model fitting and is labelled separately from factual evaluation.

6 Results

6.1 Factual held-out prediction

Table 1 reports the primary 1,300-case factual test. EviRoute-DR has the best point estimate on all three metrics. The improvement over direct logistic regression is clearest in probabilistic accuracy: Brier score improves by 0.0060 with a paired bootstrap 95% CI of 0.0021–0.0092, and log loss improves by 0.0171 (95% CI 0.0059–0.0273). Against the random forest, EviRoute-DR improves the point estimates but the paired bootstrap intervals include zero; we therefore report a modest point-estimate improvement, not established superiority or equivalence.

Fig. 2 visualizes Brier scores for all primary methods; Table 1 also provides their AUC and log loss.

The disagreement model itself is learnable: pooled disagreement prediction reaches AUC 0.823 and Brier 0.0986 on factual test rows, while the regime-conditioned variant reaches AUC 0.814 and Brier 0.0979. This supports predictable rule-error structure conditional on the supplied context; it does not establish intervention effectiveness for unexecuted actions.

Across labelled held-out slices, EviRoute-DR has Brier scores of 0.0896 on hard cases, 0.0964 on the main test, 0.0815 on the temporal-labelled cohort, and 0.1061 on workflow shift. It is strongest relative to both direct baselines on the workflow-shift Brier score and remains competitive on the other slices; detailed values are provided in Supplementary Information (ESM 2).

6.2 Calibration diagnostics

The 10-bin ECE is 0.0427 for EviRoute-DR, 0.0452 for direct logistic regression, and 0.0393 for random forest. Joint logistic calibration intercept/slope estimates are 0.240/1.213,

Table 1: Factual held-out prediction on 1,300 executed production actions. Lower Brier score and log loss are better.

Method	AUC	Brier	Log loss
Rule prior	0.7875	0.1692	2.3380
Direct logistic	0.9213	0.0971	0.3301
Direct random forest	0.9235	0.0922	0.3185
Rule-correction logistic	0.9074	0.0986	0.3362
Regime-conditioned correction	0.9006	0.0979	0.3361
Rule-correction random forest	0.8968	0.1336	0.4239
EviRoute-DR	0.9262	0.0910	0.3131

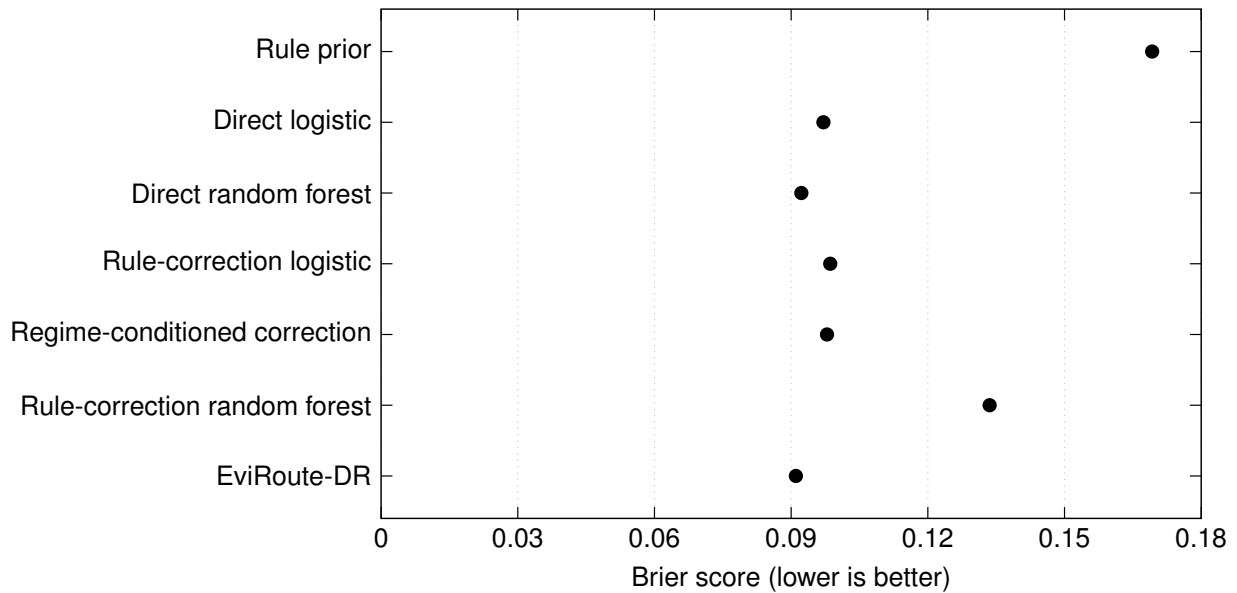


Fig. 2 Primary factual probability quality on 1,300 executed actions. Dots show Brier scores, for which lower is better. The full model improves the point estimate over both direct baselines; paired intervals support the logistic comparison but not superiority over random forest

0.472/1.013, and $-0.328/1.120$, respectively (ideal: 0/1). These descriptive metrics do not uniformly favor EviRoute-DR; the lower Brier score should therefore be described as improved probabilistic accuracy, not uniformly best calibration. ESM 2 includes the exact definitions, bin counts, and reliability diagram.

6.3 Production counterfactual replay

The separate replay uses the seven-action production counterfactual ledger as an evaluation benchmark. These non-executed alternatives are not reclassified as factual outcomes and are not used to train the primary models. Fig. 3 shows the upper envelope of the archived cost-aversion sweep; Online Resource 1 retains all operating points, including those omitted from that envelope.

At mean cost near 10 units, the interpolated EviRoute-DR replay frontier reaches approximately 0.802 resolution, compared with approximately 0.783 for the rule-prior frontier. Near 12 units, EviRoute-DR reaches approximately 0.815 versus 0.786 for the rule-prior frontier. Fixed human review reaches 0.805 resolution at mean action cost 19.5, whereas fixed abstention reaches 0.756 at mean cost 2.48. These results identify intermediate operating points in the supplied replay, not proof of superior intervention decisions. Fixed abstention is a particularly important comparator: its 0.756 replay success rate is the supplied endpoint annotation, not a convention that abstention automatically resolves a case. They are descriptive replay results, not causal estimates of deployment impact.

7 Ablations and interpretation

Table 2 isolates the inputs to the meta-learner rather than comparing only standalone experts. Each reduced model is refit on the same 200 factual meta-training cases using $C = 0.1$, then scored on the same 1,300 test cases. “Direct” denotes the logistic and random-forest factual experts; “corrected” denotes the three disagreement-corrected experts; “reliability” denotes missingness, conflict, production score, and cost. The “rule branch” includes all three corrected experts and the raw rule prior. Base expert parameters are held fixed. These are post-hoc component diagnostics, not additional independent experiments.

The direct-only stack has Brier score 0.0963; adding reliability and then the raw rule gives 0.0957 and 0.0950. The full model reaches 0.09104, consistent with a contribution from explicit disagreement correction beyond generic direct-expert stacking. However, the rule branch plus reliability reaches 0.09101, marginally lower than the full model, while the full model has the higher AUC and lower log loss. Thus the data do not show that every component is necessary or that full fusion dominates every reduced model on every metric. They support a narrower conclusion: correction and reliability supply most of the probability-quality gain, while the direct branch improves the full model’s discrimination and log-loss point estimates. No additional component-significance claim is made from these

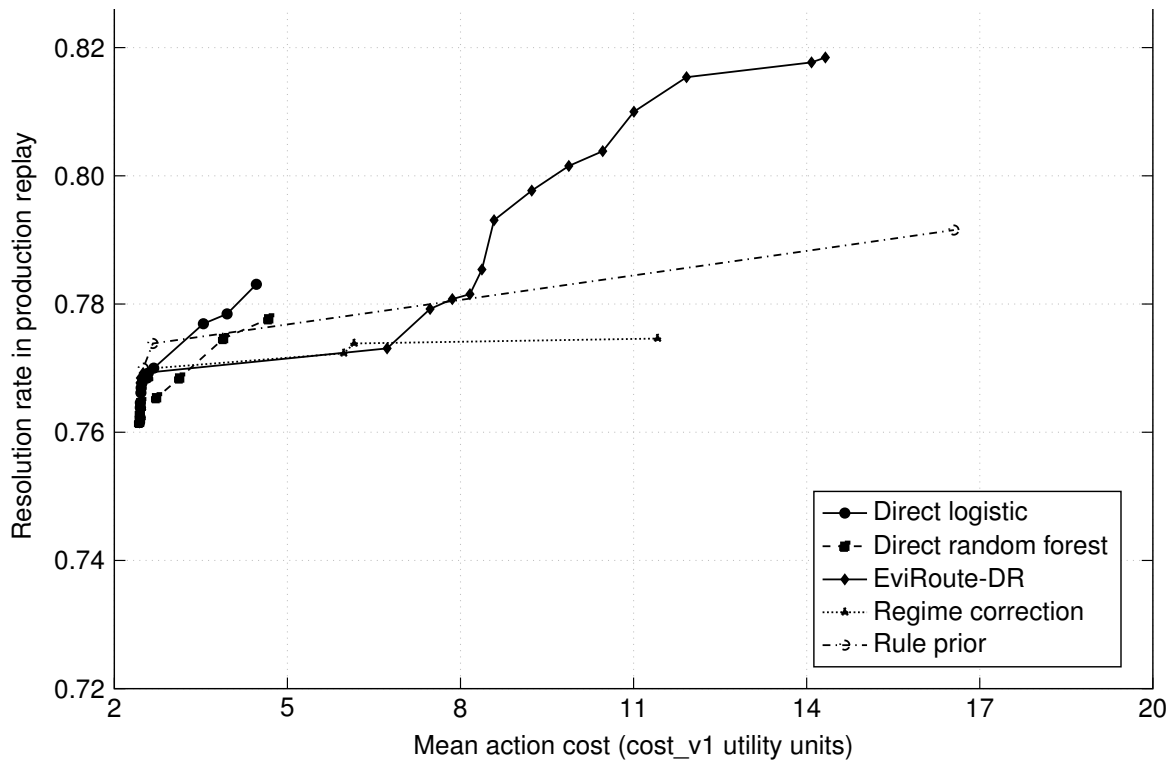


Fig. 3 Descriptive production-derived replay envelope over the archived cost-aversion sweep. Distinct line styles and markers identify the methods. Connected segments describe interpolated mixtures, not additional executed policies. Replay is secondary, annotation-dependent evidence and is not causal OPE

point estimates.

Table 2: Matched meta-learner ablations on the same 1,300 factual test cases. All rows use the same base fits, meta-training cases, standardization, and logistic regularization.

Meta-learner inputs	AUC	Brier	Log loss
Direct-only stack	0.9240	0.0963	0.3344
Direct + reliability	0.9250	0.0957	0.3270
Direct + raw rule + reliability	0.9255	0.0950	0.3249
Direct + corrected experts	0.9175	0.0916	0.3200
Rule branch + reliability	0.9242	0.0910	0.3140
Full EviRoute-DR	0.9262	0.0910	0.3131

A separate sensitivity keeps the routing models fixed and perturbs only assumptions about unexecuted replay annotations. At a common 12-unit expected cost, cost-bracket mixtures of archived EviRoute-DR and rule-prior policies differ by 0.0322 in replay resolution. Under uniform 10% binary-label error confined to unexecuted annotations, the expected gap is 0.0417. Under targeted error affecting at most 1% of the 7,800 unexecuted test annotations, the worst-case gap can instead reach -0.0207 . These are analytical perturbation scenarios, not measured error rates, new observations, or confidence intervals. They show that the replay gain is not robust to arbitrary concentrated annotation error; ESM 2 gives the construction and full sensitivity table.

The factual comparison remains deliberately narrow. EviRoute-DR’s AUC difference versus random forest is +0.0027, with a paired interval spanning zero. The empirically supported advance is improved probabilistic accuracy relative to direct logistic learning, small but inconclusive point-estimate gains versus random forest, structured disagreement prediction, and a secondary annotation-dependent replay trade-off.

8 Discussion

EviRoute-DR addresses a common operational mismatch: a system may possess dense action rules but sparse factual outcomes. Naively training on all rule labels confuses procedural knowledge with observed truth; discarding them entirely ignores useful structure. Disagreement-reliability learning treats the rule as a fallible prior whose trustworthiness can be estimated from executed production outcomes.

The method is particularly relevant to multimodal evidence systems because missing channels and conflicting signals are not generic nuisance variables; they affect whether a remediation rule is trustworthy. The workflow-shift result is encouraging in this respect, but independent replication on another deployment is still needed.

The production score log also illustrates why causal terminology must be disciplined. Although the seven scores resemble probabilities, every logged action is their argmax and the export does not establish randomized selection. Importance weighting requires

the actual behavior probabilities and support for the target policy; nonzero model scores alone do not provide that support. Relabelling a deterministic model score as a propensity would create an invalid OPE analysis. We instead report factual prediction and a separate counterfactual replay.

9 Limitations

First, this is a retrospective single-system study. The test cohorts broaden operating conditions but do not constitute an external organizational replication. Second, only one action outcome per case is factual; the complete seven-action replay ledger is a production-derived counterfactual evaluation resource and cannot establish causal intervention effects. Third, the logs do not establish valid stochastic propensities or target-policy overlap, so standard IPS/DR policy values are not reported. Fourth, EviRoute-DR's advantage over a strong random forest is modest on factual prediction and not statistically decisive in the paired bootstrap. Fifth, production costs are utility specifications; different organizations may assign different action-cost trade-offs. Sixth, the internal resolution-adjudication rule, feature-capture timing, and independent validation of counterfactual annotations are not supplied; the sensitivity analysis cannot replace those records. Finally, access to raw production records is restricted, so public reproducibility relies on released code, aggregate outputs, and protocol descriptions rather than public case-level data.

10 Conclusion

EviRoute-DR reframes evidence remediation as learning under asymmetric feedback: one factual action outcome, dense but imperfect rule priors, and reliability signals that indicate when those priors may fail. On 1,300 held-out factual production cases, the method improves probabilistic accuracy over direct logistic learning and achieves the best primary-model point estimates, while a reduced rule-branch stack narrowly improves its Brier score. A separate production counterfactual replay shows a stronger intermediate cost-resolution frontier. The results support disagreement-aware reliability fusion as a practical methodological advance while preserving a strict boundary between factual prediction, counterfactual replay, and causal policy evaluation.

Supplementary materials

ESM 2: Supplementary Information. Endpoint and cost definitions, source hierarchy, complete factual and slice results, component ablations, calibration diagnostics, and counterfactual-annotation sensitivity.

Online Resource 1: Code and aggregate results. Executable analysis and figure

sources, aggregate CSV tables, input hashes, and reproduction checks. No raw case-level production records are included.

ESM 2: Supplementary operational clarification. Retrospective operational-record provenance, routine staff corrections, and matching ethics, consent, and data-availability statements. This companion adds no observations or statistical results.

Declarations

Author contributions. Nataraj Agaram Sundar: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review and editing, Project administration. Tejas Morabia: Conceptualization, Methodology, Investigation, Validation, Supervision, Project administration, Writing – review and editing.

Funding. No external funding was received for this work.

Competing interests. The authors are employees of eBay Inc. The authors declare no other competing interests.

Ethics approval

Not applicable. This study is a retrospective computational analysis of de-identified operational records generated during routine business operations. Staff corrections recorded as reviewer overrides were part of ordinary case handling and were not collected through a research intervention or participant study. No individuals were recruited, enrolled, surveyed, or assigned research-specific tasks for this study, and the analysis does not evaluate individual staff performance. No customer-identifying information or reviewer identities are reported.

Consent to participate

Not applicable. No human participants were recruited or enrolled in this retrospective operational-data study, and no research-specific interaction or intervention with employees was conducted. The reviewer-override records arose from routine operational work rather than study-specific participation.

Consent for publication

Not applicable. The article contains no identifiable photographs of individuals or other identifying personal information.

Data availability

The case-level datasets analysed during the current study are not publicly available because they contain proprietary operational telemetry and decision outcomes. Requests for access to de-identified records may be directed to the corresponding author at natsundar@ebay.com on reasonable request. Access is subject to organizational authorization, privacy and legal review, and applicable data-use restrictions. De-identified aggregate result tables, analysis code, and the evaluation protocol are provided in the accompanying supplementary materials (ESM 1 and Online Resource 1).

Code availability. Analysis code and aggregate outputs are included in Online Resource 1. The code expects the production input tables described in ESM 2; raw production records are not included in the public resource.

Use of artificial intelligence tools. Generative AI tools were used to assist with manuscript organization. The authors reviewed the analyses, checked reported outputs against the supplied production files, determined the scientific claims, and retain full responsibility for the manuscript.

References

- [1] Ong I, Almahairi A, Wu V, Chiang W-L, Wu T, Gonzalez JE, Kadous MW, Stoica I. RouteLLM: Learning to Route LLMs from Preference Data. In: International Conference on Learning Representations; 2025. <https://openreview.net/forum?id=8sSqNtaMr>.
- [2] Swaminathan A, Joachims T. Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In: Proceedings of the 32nd International Conference on Machine Learning. PMLR. 2015;37:814–823. <https://proceedings.mlr.press/v37/swaminathan15.html>.
- [3] Dudík M, Langford J, Li L. Doubly Robust Policy Evaluation and Learning. In: Proceedings of the 28th International Conference on Machine Learning; 2011. p. 1097–1104. <https://icml.cc/Conferences/2011/papers.php.html>.
- [4] Mandyam A, Tang S, Yao J, Wiens J, Engelhardt BE. CANDOR: Counterfactual ANnotated DOubly Robust Off-Policy Evaluation. In: Proceedings of the 7th Conference on Health, Inference, and Learning. PMLR. 2026;333:644–687. <https://proceedings.mlr.press/v333/mandyam26a.html>.
- [5] Carranza AG, Athey S. Robust Offline Policy Learning with Observational Data from Multiple Sources. In: Proceedings of the 28th International Conference on Artificial Intelligence and Statistics. PMLR. 2025;258:4897–4905. <https://proceedings.mlr.press/v258/carranza25a.html>.
- [6] Geifman Y, El-Yaniv R. Selective Classification for Deep Neural Networks. In: Advances in Neural Information Processing Systems. 2017;30. <https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html>.

- [7] Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. In: Proceedings of the 34th International Conference on Machine Learning. PMLR. 2017;70:1321–1330. <https://proceedings.mlr.press/v70/guo17a.html>.
- [8] Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In: Advances in Neural Information Processing Systems. 2017;30. https://papers.nips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM2SupplementaryClarification.pdf](#)
- [ESM2SupplementarySource.zip](#)