

Supplementary Information

ESM 2 Operational provenance and declarations

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

S2.1 Scope

This companion describes the provenance of routine operational corrections and the scope of the retrospective analysis supporting EviRoute-DR. It should be read alongside ESM 1 and Online Resource 1. It introduces no additional observations, model fits, statistical tests, or performance claims.

S2.2 Routine operational corrections

The study retrospectively analyses existing operational records. Staff corrected case outcomes during ordinary business operations; the paper subsequently analyses those existing records. The term *reviewer override* denotes this operational correction process, not enrollment in a research experiment. Staff were not recruited, surveyed, or assigned experimental tasks to generate the corrections used in this analysis.

Section 5.2 of the main manuscript retains the final exported outcome flag, including the 31 executed cases with a recorded operational correction. The corrections are not removed or replaced by the rule prior. The analysis concerns case outcomes, not individual employee performance.

Record component	Role in the study	Interpretation boundary
Executed outcome	Factual training and held-out outcome prediction	One observed action outcome per case
Routine staff correction	Part of the retained exported outcome	Ordinary case handling, not a recruited participant study
Rule prior	Input to disagreement correction and reliability fusion	Not treated as factual truth
Non-executed replay annotation	Secondary evaluation of frozen policies	Not an observed intervention outcome or causal estimate

Table S2.1: Distinct uses of operational outcomes, priors, and replay annotations.

S2.3 Ethics and consent declarations

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.

S2.4 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).

S2.5 Relationship to the scientific results

The manuscript analyses 2,100 operational cases using a 600/200/1,300 train/meta-training/test partition and seven candidate actions. Routine corrections remain part of the recorded factual endpoint. The declarations describe the authors' retrospective operational-data classification; they do not assert that a named committee granted an ethics exemption or consent waiver. Factual outcome prediction remains distinct from annotation-dependent counterfactual replay, which does not establish causal intervention effects.