

# **DETAILED ERROR ANALYSIS**

Point-by-Point Mapping of Reviewer Comments to Paper Issues

February 6, 2026

## **Contents**

<b>1</b>	<b>REVIEWER 1: Methodological Concerns</b>	<b>2</b>
<b>2</b>	<b>REVIEWER 2</b>	<b>4</b>
<b>3</b>	<b>REVIEWER 3: Structure &amp; Clarity</b>	<b>5</b>
<b>4</b>	<b>REVIEWER 4: Linguistic &amp; Methodological Quality</b>	<b>7</b>
<b>5</b>	<b>REVIEWER 5: Experimental Completeness</b>	<b>9</b>
<b>6</b>	<b>REVIEWER 6: Technical Details</b>	<b>12</b>
<b>7</b>	<b>REVIEWER 7: Rigor &amp; Reproducibility</b>	<b>14</b>
<b>8</b>	<b>REVIEWER 8: Deep Technical Critique</b>	<b>18</b>
<b>9</b>	<b>SUMMARY TABLE: Common Errors Across All Reviewers</b>	<b>23</b>
<b>10</b>	<b>CRITICAL PATH SUMMARY</b>	<b>28</b>
10.1	6 FATAL Issues - Must Fix . . . . .	28
10.2	11 MAJOR Issues - Should Fix . . . . .	28
10.3	3 MINOR Issues - Nice to Have . . . . .	28

# 1 REVIEWER 1: Methodological Concerns

Table 1: Reviewer 1 - Detailed Issue Mapping

R1.#	Reviewer Comment	Location in Paper	Error Type & Severity
R1.1	<b>Novelty unclear:</b> "Provide a clearer positioning of what is novel beyond 'a unified evaluation pipeline."	Abstract lines 8-12 Introduction Section 1, paragraph 3 Contributions list (lines 91-93)	<b>MAJOR - Conceptual Gap</b> Paper states "unified pipeline" but reviewers see this as procedural/engineering contribution, NOT methodological novelty. Current contribution bullets are vague.
R1.2	<b>COCOMO II unfair:</b> "Add experiments with recalibrated COCOMO II for a fairer comparison."	Section 2.1 (COCOMO Recap) Section 5.1 Table 1 (baseline results) Section 4 (no calibration mentioned)	<b>FATAL - Invalid Baseline</b> Paper uses COCOMO II with UNCALIBRATED parameters (likely A=2.94, B=0.91 defaults). Comparing RF against uncalibrated COCOMO creates "straw man" argument. MMRE=2.790 is suspiciously high. Must fit A, B on training data per schema.
R1.3	<b>Modern datasets missing:</b> "Include modern datasets (GitHub, Jira-based effort logs, DevOps metrics) to improve relevance."	Section 3.1 (Data Sources) Datasets are 1993-2022, no GitHub/Jira	<b>MAJOR - Generalization Concern</b> All datasets are historical/legacy. No modern DevOps metrics, CI/CD telemetry, or Agile story points. Limits external validity for contemporary projects.

R1.#	Reviewer Comment	Location in Paper	Error Type & Severity
R1.4	<b>Additional metrics:</b> "Report additional error metrics such as MAPE, MdMRE, or RAE."	Section 2.3 (Evaluation Metrics) Section 5 (Results tables)	<b>MINOR - Incomplete Metrics</b> Only reports MMRE, PRED(25), MAE, RMSE, $R^2$ . Missing MdMRE (median-based, more robust), MAPE, RAE. Easy to add.
R1.5	<b>Confidence intervals:</b> "Provide confidence intervals for all reported metrics."	Section 5.1 Table 1 All result tables	<b>MINOR - Uncertainty Reporting</b> Results show mean values only (e.g., MMRE=0.647). Should report "Mean [95% CI]" using bootstrap or standard error from 10 seeds.
R1.6	<b>Length reduction:</b> "Reduce length by moving some methodological details to appendices or supplementary material."	Entire document	<b>MINOR - Formatting</b> Paper may be verbose in preprocessing details (Section 3.2-3.4). Can move IQR formulas, detailed harmonization rules to Supplementary.
R1.7	<b>Reproducibility:</b> "Release the harmonized dataset and scripts for reproducibility."	Section 9 (Data Availability) No GitHub link provided	<b>MAJOR - Reproducibility Gap</b> Paper says "upon reasonable request" but no actual repository link. Reviewers cannot verify claims. Should upload to Zenodo/Figshare with DOI.

## 2 REVIEWER 2

**Note:** Reviewer 2 provided an attachment. Based on the email, this was added separately. The main concerns likely overlap with Reviewer 8's detailed technical review.

*Action: Read Reviewer 2 attachment carefully and map to specific sections.*

### 3 REVIEWER 3: Structure & Clarity

Table 2: Reviewer 3 - Detailed Issue Mapping

R3.#	Reviewer Comment	Location in Paper	Error Type & Severity
R3.1	<b>Introduction structure:</b> "The Introduction should make a compelling case... What is already known? What is missing? What needs to be done?"	Section 1, paragraphs 1-3	<b>MAJOR - Structure Weakness</b> Introduction jumps to "unified framework" without clearly establishing (1) state of art, (2) research gap, (3) specific research questions. Need explicit subsections: 1.1 Motivation, 1.2 Research Gap, 1.3 Contributions.
R3.2	<b>Related Work insufficient:</b> "Compare the references... draw the paper's motivation. Cite: aisy.202300706, patcog.112890, ACCESS.3480205, engappai.111655"	Section 7 (Related Work) NO comparison table	<b>MAJOR - Missing SOTA Comparison</b> Section 7 describes evolution of SEE but does NOT compare specific papers in a table. Must create comparison: Study — Year — Approach — Schemas — Statistical Tests — MMRE — Our Advantage. Must cite 4 DOI papers suggested.
R3.3	<b>Assumptions &amp; limitations:</b> "Highlight all assumptions and limitations of your work."	Section 6 (Threats to Validity) NO explicit Assumptions section	<b>MAJOR - Missing Critical Section</b> Paper has "Threats to Validity" but no explicit "Assumptions" (e.g., linear cost-effort, 160h/month, no team dynamics). Need NEW Section 3.6 or subsection in Methods: "Assumptions and Limitations" (2 pages).

R3.#	Reviewer Comment	Location in Paper	Error Type & Severity
R3.4	<b>Figure 1 description:</b> "Describe clearly Figure 1 within the text."	Figure 1 caption (line 130) No detailed explanation in text	<b>MINOR - Figure Caption Weak</b> Figure 1 shows CO-COMO pipeline vs ML framework but caption is 1 sentence. Need 3-4 sentence caption explaining: (a) left side is COCOMO, (b) right side is ML, (c) key differences are preprocessing + model flexibility.
R3.5	<b>Conclusion structure:</b> "Consider: (i) Strengths and weaknesses, (ii) Assessment and implications, (iii) Projection/recommendations"	Section 8 (Conclusion)	<b>MINOR - Conclusion Enhancement</b> Current conclusion summarizes findings but lacks explicit "Strengths/Weaknesses" subsection and "Practical Recommendations" for PM. Can restructure as: 8.1 Summary, 8.2 Strengths & Limitations, 8.3 Recommendations.

## 4 REVIEWER 4: Linguistic & Methodological Quality

Table 3: Reviewer 4 - Detailed Issue Mapping

R4.#	Reviewer Comment	Location in Paper	Error Type & Severity
R4.1	<b>Introduction too short:</b> "The introduction is too short, the limitations of their research related to this paper should be pointed out"	Section 1 (Introduction) Only 3 paragraphs	<b>MAJOR - Insufficient Context</b> Introduction is only 600 words. Missing: (1) detailed problem scoping, (2) explicit research questions, (3) scope/limitations preview. Should expand to 4-5 paragraphs or add subsections.
R4.2	<b>Related Work lacks comparison:</b> "Detailed explanations for advantage and drawback of each related method. Cite: TSMC.2025.3580086, TFUZZ.2025.3569741, TETCI.2025.3647653"	Section 7 (Related Work) Figure 7 (related work diagram)	<b>MAJOR - Missing SOTA Models</b> Paper mentions ML evolution but does NOT discuss: (1) XGBoost/LightGBM/CatBoost, (2) Deep Learning approaches, (3) Recent 2024-2025 papers. Must add comparison table + cite 3 DOI papers.
R4.3	<b>Newer models missing:</b> "There are some newer model can be as candidate algorithm for solving this problem."	Section 2.2 (Multi-Schema Framework) Section 4.2 (Models: LR, DT, RF, GB)	<b>MAJOR - Outdated Model Selection</b> Paper only tests LR, DT, RF, GB. Missing: XGBoost, LightGBM, CatBoost, Neural Networks (MLP), LSTM for sequential effort. These are 2020+ SOTA. Must add XGBoost at minimum.

R4.#	Reviewer Comment	Location in Paper	Error Type & Severity
R4.4	<b>Post-hoc tests:</b> "Post hoc statistical tests can be used to discuss the results."	Section 4.4 (Significance Testing) Wilcoxon + Holm-Bonferroni mentioned	<b>MINOR - Statistical Enhancement</b> Paper uses Wilcoxon + Cliff's Delta, which is good. Could add Friedman test + Nemenyi post-hoc for comparing 5+ models across multiple datasets (more robust than pairwise Wilcoxon).
R4.5	<b>Language quality:</b> "Linguistic quality needs improvement. Grammatical errors affect quality."	Entire document e.g., "it is worth noting" appears multiple times	<b>MAJOR - AI-Generated Tone</b> Paper uses formulaic phrases: "it is worth noting," "captures the nuances," "substantially outperforming." Sounds templated. Need manual rewrite to sound natural. Run Grammarly + human editing.

## 5 REVIEWER 5: Experimental Completeness

Table 4: Reviewer 5 - Detailed Issue Mapping

R5.#	Reviewer Comment	Location in Paper	Error Type & Severity
R5.1	<b>Generalization:</b> "Would be interesting to see if these models hold up across different methodologies. Add more datasets."	Section 3.1 (Datasets) Section 6 (External Validity)	<b>MAJOR - Limited Generalization</b> All datasets are historical waterfall projects. No Agile, no Scrum, no DevOps. Cannot claim generalization to modern SDLC. Need: (1) Agile story point datasets, (2) CI/CD telemetry, OR (3) explicit limitation statement.
R5.2	<b>Paper structure:</b> "At the end of the introduction, incorporate the structure of the paper."	Section 1, end of Introduction	<b>MINOR - Roadmap Missing</b> Introduction ends with contributions list but no roadmap. Should add: "The remainder of this paper is organized as follows: Section 2 presents background, Section 3 describes datasets..."
R5.3	<b>Figure quality:</b> "Figures 1 and 2 are suboptimal. Enhance quality."	Figure 1 (COCOMO vs ML) Figure 2 (Unit conversions) ALL figures have NO CAPTIONS	<b>FATAL - Missing Captions &amp; Low Resolution</b> Reviewer 7 also flags this: "None of the figures contain captions." LaTeX code shows figures but captions may be formatted incorrectly or missing. ALL figures must have: (1) High-res export (600dpi PDF), (2) Proper caption with label.

R5.#	Reviewer Comment	Location in Paper	Error Type & Severity
R5.4	<b>Ablation study:</b> "Incorporate ablation study to evaluate each part of the proposed method."	NO ablation study in paper	<b>MAJOR - Missing Validation</b> Paper claims pre-processing (unit harmonization, log transform, IQR capping) improves results but provides NO ablation: Raw vs +Log vs +Log+IQR vs Full. Must add NEW Table: "Ablation Study - RF Performance with Progressive Preprocessing Steps."
R5.5	<b>Limitations detail:</b> "Incorporate the limitation of the proposed method in more detail."	Section 6 (Threats to Validity)	<b>MINOR - Expand Limitations</b> Threats to validity is generic. Need explicit limitations: (1) FP n=24 too small, (2) No team dynamics features, (3) Assumes 160h/month uniform, (4) Historical data bias.
R5.6	<b>Figure numbering:</b> "The numbering of the figures should be added to the manuscript."	All figures	<b>MINOR - Formatting Issue</b> Figures may be numbered correctly in LaTeX but reviewer cannot see them (PDF rendering issue?). Ensure figure labels are visible: Figure 1:, Figure 2:, etc.
R5.7	<b>Section structure:</b> "Some sections are disorder. Integrate brief 1-2 sentence subsections."	Section 3 (multiple small subsections)	<b>MINOR - Organization</b> Section 3 has many small subsections (3.1, 3.2, 3.3, 3.4, 3.5). Some are only 1 paragraph. Can merge: 3.2 + 3.3 into "Data Cleaning," 3.4 into existing section.

R5.#	Reviewer Comment	Location in Paper	Error Type & Severity
R5.8	<b>Cite additional papers:</b> "Consider: 10.1007/s44248-024-00016-0, 10.21203/rs.3.rs-7556543/v1"	Section 7 (Related Work)	<b>MINOR - Missing Citations</b> Must cite 2 additional papers in Related Work + explain how they relate to this study.
R5.9	<b>Linear Regression justification:</b> "If relationship is non-linear, LR might not work well, limiting framework performance."	Section 4.2 (Models) Table 1 shows LR performs badly	<b>MAJOR - Model Justification</b> Paper includes LR as baseline but results show MMRE=4.5 (worst). Should either: (1) Remove LR from comparison, OR (2) Justify why it's included: "LR serves as a sanity check for linearity assumptions; poor results confirm non-linear relationships."

## 6 REVIEWER 6: Technical Details

Table 5: Reviewer 6 - Detailed Issue Mapping

R6.#	Reviewer Comment	Location in Paper	Error Type & Severity
R6.1	<b>"Overall" aggregation unclear:</b> "Abstract mentions MMRE0.647 but does not clarify if averaged across schemas or from specific one. Given schema differences, specify to avoid misinterpretation."	Abstract line 10 Section 5.1 Table 1 ("Overall")	<b>FATAL - Ambiguous Metrics</b> Table 1 title: "Overall test performance" but HOW is "overall" computed? (1) Pooling all LOC+FP+UCP predictions? (2) Macro-average (unweighted mean of 3 schemas)? (3) Micro-average (weighted by sample size)? LOC n=947 dominates FP n=24. MUST define explicitly.
R6.2	<b>Equation labels duplicate:</b> "Section 2.1 equation references undefined. Equation for Time presented twice with nearly identical wording."	Section 2.1, Equations 1-2 Lines 120-130	<b>MAJOR - Duplicate/Formatting Error</b> LaTeX shows: $\text{Time} = C \times E^D$ appears TWICE (Eq 2 and again after line 130). Second instance is redundant. DELETE second copy. Also fix equation labels: use <code>\label{eq:cocomo-time}</code> properly.
R6.3	<b>FP sample size protocol:</b> "FP n=24 is very small. May limit statistical reliability. Discuss limitation and impact on conclusions for FP."	Section 3.1 (FP schema n=24) Section 4.1 (Train-test protocol) Section 5.2 (FP results)	<b>FATAL - Statistical Power Issue</b> 80/20 split on n=24 gives 19 train / 5 test. Grid search with 5-fold CV on 19 training samples is HIGHLY UNSTABLE. Should use: (1) Leave-One-Out CV for FP, (2) Bootstrap confidence intervals, (3) Label FP results "exploratory."

R6.#	Reviewer Comment	Location in Paper	Error Type & Severity
R6.4	<b>R<sup>2</sup> column shows “–”:</b> ”Table 1 shows ‘–’ for R <sup>2</sup> for all models. If computed, report. Otherwise remove or explain.”	Section 5.1 Table 1 All result tables	<b>MAJOR - Missing Metric</b> LaTeX code shows: $R^2 \uparrow --$ $-- - - - -$ . Either: (1) R <sup>2</sup> was computed but not filled in (typo), OR (2) R <sup>2</sup> not computed. Section 2.3 defines R <sup>2</sup> formula so SHOULD be reported. Must compute and fill OR explain ”R <sup>2</sup> not applicable for COCOMO II due to negative values.”

## 7 REVIEWER 7: Rigor & Reproducibility

Table 6: Reviewer 7 - Detailed Issue Mapping

R7.#	Reviewer Comment	Location in Paper	Error Type & Severity
R7.1	<b>Formatting catastrophic:</b> "None of the figures or tables contain captions. Figures are low resolution with unreadable text. Paper lacks page/line numbers."	ALL figures (1-8) ALL tables LaTeX preamble (no lineno package)	<b>FATAL - Submission Standards Violation</b> This is CRITICAL. LaTeX code shows figures with includegraphics but captions may be missing or incorrectly formatted. MUST: (1) Add caption{...} for EVERY figure, (2) Export figures as vector PDF 600dpi, (3) Add usepackage{lineno} + linenumbers in preamble for easy review.
R7.2	<b>AI-generated tone:</b> "Language feels unnatural and formulaic, as if heavily templated or generated. Authors need to manually revise for natural academic tone."	Entire document Phrases: "it is worth noting," "captures nuances," "substantially"	<b>MAJOR - Writing Quality</b> Multiple reviewers (R4, R7) flag this. Paper overuses generic phrases. Must: (1) Run Grammarly, (2) Manual human rewrite removing clichés, (3) Vary sentence structure.

R7.#	Reviewer Comment	Location in Paper	Error Type & Severity
R7.3	<b>COCOMO II straw man:</b> "COCOMO error rates suspiciously high. Not stated if A, B calibrated or defaults used. Using uncalibrated creates straw man, invalidates comparison."	Section 2.1 (COCOMO) Section 4.2 (no CO-COMO calibration mentioned) Table (MMRE=2.790) 1	<b>FATAL - Invalid Baseline</b> SAME as R1.2 and R8.X. Paper does NOT describe CO-COMO implementation. MMRE=2.790 suggests default parameters on heterogeneous data. MUST: (1) Fit A, B using scipy.optimize on training set per schema, (2) Report COCOMO (original) vs COCOMO (calibrated) vs RF, (3) Explain limitations for FP/UCP (may need FP-to-LOC conversion).
R7.4	<b>SOTA models missing:</b> "Model selection outdated. Fails to include XGBoost, LightGBM, CatBoost. No deep learning or LLMs."	Section 4.2 (Models: LR, DT, RF, GB) Section 7 (Related Work)	<b>MAJOR - Incomplete Comparison</b> SAME as R4.3. Paper only uses scikit-learn basic models. Modern SOTA for tabular data: XGBoost, LightGBM, CatBoost. Must add XGBoost at minimum. Deep learning (MLP) optional but should mention in limitations: "Neural networks not explored due to small FP/UCP sample sizes."

R7.#	Reviewer Comment	Location in Paper	Error Type & Severity
R7.5	<b>Interpretability unsupported:</b> "Authors claim RF provides interpretability but results focus only on error metrics. Must include feature importance (Gini or SHAP)."	Section 6 (Discussion) NO feature importance plot	<b>MAJOR - Missing Analysis</b> Paper says RF is interpretable (Section 6, paragraph 1) but provides NO feature importance analysis. MUST add: (1) NEW Figure: RF feature importance (Gini impurity), (2) NEW paragraph explaining top 3 features (e.g., Size, Time, Developers contribute X%).
R7.6	<b>Ablation study missing:</b> "Complex pipeline but no validation of individual components. Need ablation to verify if gains come from framework or just log/IQR."	NO ablation study	<b>MAJOR - Missing Validation</b> SAME as R5.4. Must create NEW Table: "Ablation Study - RF MMRE with Progressive Preprocessing" Raw → +Log → +Log+IQR → +Log+IQR+Harmonization Show that each step contributes to performance.
R7.7	<b>Sample sizes unclear:</b> "Data reporting vague. FP n=24 means test set 5 data points, statistically insufficient. Must list sample sizes for all splits."	Section 3.1 (Data sources) NO explicit train/test counts per schema	<b>FATAL - Missing Critical Info</b> Paper says "n947 LOC, n=24 FP, n=71 UCP" but does NOT report: (1) Train vs test split counts per schema, (2) Per-source counts before/after dedup. MUST create: NEW Table 1 in Section 3.1: "Dataset Manifest" with columns: Source — Link/DOI — Schema — Raw Count — After Dedup — Train — Test.

R7.#	Reviewer Comment	Location in Paper	Error Type & Severity
R7.8	<b>Generalization unclear:</b> "Main contribution is engineering pipeline, not methodological. Need to show approach generalizes to unseen datasets or different organizations."	Section 6 (External Validity) NO cross-organization validation	<b>MAJOR - Limited Generalization</b> Paper tests on random holdouts from SAME historical pool. Does NOT test: (1) Leave-one-source-out (LOSO) CV per dataset origin, (2) Transfer to new organization. Must either: (1) Add LOSO experiment, OR (2) Acknowledge limitation: "Cross-organizational validation is future work."
R7.9	<b>Figure anomalies:</b> "Discussion figures appear strange. LOC error curve has few points. LR error DECREASES as size increases (contradicts theory). FP ground truth is smooth curve, not scattered (simulations?)."	Figure in Section 6 (Error Profiles) May be Figure 5 or later	<b>MAJOR - Questionable Visualizations</b> Reviewer suspects: (1) Figures may show TRAINING data not TEST, OR (2) Ground truth line interpolated incorrectly. MUST: (1) Verify all plots use TEST SET only, (2) Plot actual scatter points, not interpolated curves, (3) Add figure caption explaining what each line/point represents.

## 8 REVIEWER 8: Deep Technical Critique

Table 7: Reviewer 8 - Detailed Issue Mapping

R8.#	Reviewer Comment	Location in Paper	Error Type & Severity
R8.1	<b>Limited novelty:</b> "Main findings (RF & COCOMO) well established. Unified framework is procedural, not methodological. Contribution incremental."	Abstract Introduction (contributions)	<b>MAJOR - Conceptual Weakness</b> Reviewer 8 (most technical) says paper is good empirical study but does NOT introduce new modeling paradigm. Must: (1) Reframe contributions to emphasize REPRODUCIBLE BENCHMARK + HARMONIZATION PROTOCOL as main value, NOT just "RF wins," (2) Downplay novelty claims.
R8.2	<b>No cross-schema learning:</b> "Models trained independently for LOC/FP/UCP. No cross-schema generalization, transfer learning, or shared representation. Doesn't address fragmentation."	Section 2.2 (Framework) Section 4.1 (Train-test)	<b>FATAL - Conceptual Gap</b> Paper claims "unified multi-schema" but trains 3 SEPARATE models (one per schema). No model trained on LOC+FP+UCP jointly. Current approach is "3 parallel pipelines" NOT "unified." Must: (1) Clarify in Abstract/Intro, OR (2) Add cross-schema transfer experiment (train LOC, test FP).

R8.#	Reviewer Comment	Location in Paper	Error Type & Severity
R8.3	<b>Data imbalance not addressed:</b> "Effort datasets highly skewed. Standard loss functions biased toward majority ranges. May inflate performance while masking poor behavior on large projects."	Section 3.3 (Outlier handling) Section 4.2 (Models use MSE loss)	<b>FATAL - Imbalance Ignored</b> Paper does IQR capping (mitigates outliers) but does NOT address CLASS IMBALANCE (many small projects, few large). Models minimize MSE which gives equal weight to all samples. Large projects (high impact) may be systematically under-predicted. MUST: (1) Analyze error distribution by project size quantile, (2) Consider weighted loss or focal loss for regression (cite: 10.1038/s41598-025-22853-y).
R8.4	<b>Imbalance-aware learning opportunity:</b> "Study would be strengthened by incorporating focal loss variants for regression. Recent work (10.1038/s41598-025-22853-y) shows focal loss improves robustness on long-tailed targets."	Section 4.2 (Models) Section 8 (Future Work)	<b>MAJOR - Missed Methodological Opportunity</b> Reviewer suggests NOVEL contribution: adapt focal loss from classification to regression for imbalanced effort data. This would add genuine novelty. Can: (1) Implement focal MSE loss for RF/GB, OR (2) Cite paper and add to Future Work: "Focal loss for imbalanced regression datasets."

R8.#	Reviewer Comment	Location in Paper	Error Type & Severity
R8.5	<b>Dataset provenance missing:</b> "Data aggregation mentioned but not auditible. Need source table: dataset name, year, link/DOI, schema, raw count, dedup count, final count."	Section 3.1 (Data sources) Section 9 (Data Availability)	<b>MAJOR - Reproducibility Gap</b> SAME as R7.7. Paper lists GitHub links but no structured manifest. MUST create: NEW Table 1 (Dataset Provenance) with 6 columns: Source — Year — Link/DOI — Schema — Raw # — After Dedup — Final #. Example rows: DASE 2023 — github — LOC — 1200 — -150 — 1050.
R8.6	<b>Deduplication leakage risk:</b> "Dedup criteria: project_no, title, size, effort. Titles/IDs inconsistent across corpora. May still have near-duplicates causing train-test leakage."	Section 3.1 (Deduplication) Lines: "exact duplicates matched on..."	<b>FATAL - Data Leakage Risk</b> Paper says duplicates removed by matching project_no, title, size, effort but: (1) Project titles often differ slightly ("ProjectA" vs "Project A v2"), (2) Same project in 2 sources may have slightly different effort values (rounding). Risk: SAME project in train+test = LEAKAGE. MUST: (1) Document exact dedup algorithm (case insensitive? fuzzy match?), (2) Verify no overlap between train/test via project name analysis.

R8.#	Reviewer Comment	Location in Paper	Error Type & Severity
R8.7	<b>Target leakage: Developers feature:</b> "If Developers derived from Effort/Time, using it as feature creates target leakage. Only use Developers if in raw dataset."	Section 3.2 (Unit Harmonization) Line: "Developer count inferred as $\text{ceil}(\text{Effort}/\text{Time})$ "	<b>FATAL - Feature Leakage</b> LaTeX shows: "Developer count is inferred as $\text{ceil}(\text{Effort}/\text{Time})$ " (line 180). This is LEAKAGE: Developers = $f(\text{Effort}) \rightarrow$ using Developers as feature = using target to predict target. MUST: (1) REMOVE Developers from features IF it's inferred from Effort, (2) Only use Developers if present in raw data BEFORE effort is known.
R8.8	<b>Hyperparameter search on FP overfits:</b> "FP n=24: 80/20 split gives 19 train. Grid search with 5-fold CV on 19 samples can overfit to idiosyncrasies."	Section 4.2 (Hyperparameter tuning) FP schema	<b>MAJOR - Statistical Power Issue</b> SAME as R6.3. 5-fold CV on n=19 means each fold has 4 samples. Grid search may select parameters that work on these 4 but fail on new data. MUST: (1) For FP: use LOOCV (Leave-One-Out) instead of 5-fold, (2) Reduce hyperparameter search space for FP (fewer configurations), (3) Report wider confidence intervals.

R8.#	Reviewer Comment	Location in Paper	Error Type & Severity
R8.9	<b>Class imbalance acknowledgement:</b> "Effort is continuous regression, not classification, so no class imbalance. But should mention focal loss paper for future classification work."	Section 4.2 (Models) Section 8 (Future Work)	<b>MAJOR - Clarification Needed</b> Reviewer notes effort is CONTINUOUS (regression) so "class imbalance" doesn't apply. But should: (1) Explain effort is continuous, no classes, (2) Acknowledge SIZE IMBALANCE (many small projects, few large), (3) Cite focal loss paper (10.1038/s41598-025-22853-y) for future classification tasks (e.g., project risk categories).

## 9 SUMMARY TABLE: Common Errors Across All Reviewers

Table 8: Common Critical Errors Identified by Multiple Reviewers

Error ID	Issue Description	Reviewers	Severity & Action Required
E1	<b>COCOMO II baseline un-calibrated / unfair comparison</b> Using default A, B parameters creates "straw man." Must fit COCOMO on training data per schema.	R1.2, R7.3, R8 (implicit)	<b>FATAL - MUST FIX</b> 1. Implement scipy.optimize to fit A, B on train set 2. Report COCOMO (original) vs (calibrated) vs RF 3. Explain FP/UCP COCOMO limitations <b>Location:</b> Section 2.1, 4.2, Table 1
E2	<b>"Overall" aggregation undefined</b> Table 1 shows "overall" metrics but method unclear: pooled? macro-avg? micro-avg? LOC n=947 dominates FP n=24.	R6.1, R8.2	<b>FATAL - MUST DEFINE</b> 1. Add subsection defining aggregation: "macro-average (unweighted mean) across 3 schemas" 2. Create NEW table: per-schema results <b>Location:</b> Abstract line 10, Section 5.1
E3	<b>Figures missing captions / low resolution</b> ALL figures lack captions or have formatting issues. Cannot understand plots.	R5.3, R7.1	<b>FATAL - MUST FIX</b> 1. Add caption... for EVERY figure (8+ figures) 2. Export figures as vector PDF 600dpi 3. Add line numbers: usepackagelineno <b>Location:</b> All figures 1-8
E4	<b>FP n=24 protocol inappropriate</b> 80/20 split gives 5 test samples. Grid search on 19 training unstable.	R6.3, R7.7, R8.8	<b>FATAL - MUST CHANGE</b> 1. For FP: use LOOCV (Leave-One-Out CV) 2. Report bootstrap 95% CI 3. Label FP results "exploratory" <b>Location:</b> Section 4.1, 5.2

Error ID	Issue Description	Reviewers	Severity & Action Required
E5	<b>Dataset manifest / provenance missing</b> Cannot audit data sources. Dedup criteria may allow train-test leakage.	R7.7, R8.5, R8.6	<b>FATAL - MUST CREATE</b> 1. NEW Table 1: Dataset Provenance Columns: Source — Link/DOI — Schema — Raw# — Removed# — Final# 2. Document dedup algorithm (exact vs fuzzy match) <b>Location:</b> Section 3.1
E6	<b>Target leakage: Developers feature</b> "Developers = ceil(Effort/Time)" uses target to create feature.	R8.7	<b>FATAL - MUST REMOVE</b> 1. DELETE Developers from features if inferred from Effort 2. Only use Developers if present in raw dataset BEFORE effort known <b>Location:</b> Section 3.2 line 180
E7	<b>Novelty unclear / contribution incremental</b> Reviewers see "RF & CO-COMO" as known result. "Unified pipeline" is procedural, not methodological novelty.	R1.1, R3.1, R4.1, R8.1	<b>MAJOR - REFRAME</b> 1. Rewrite Abstract/Intro emphasizing: REPRODUCIBLE BENCHMARK + HARMONIZATION PROTOCOL 2. Downplay "novel model" claims 3. Reframe as empirical validation study <b>Location:</b> Abstract, Section 1
E8	<b>Related Work lacks SOTA comparison</b> No comparison table with recent papers. Missing discussion of XG-Boost/LightGBM/DL. Must cite DOI papers.	R3.2, R4.2, R5.8	<b>MAJOR - MUST ADD</b> 1. Create comparison table: Study — Year — Approach — Schemas — MMRE 2. Cite 4+ DOI papers suggested 3. Discuss advantages/limitations vs prior work <b>Location:</b> Section 7

Error ID	Issue Description	Reviewers	Severity & Action Required
E9	<b>SOTA models missing (XGBoost, LightGBM)</b> Only tests LR, DT, RF, GB (2000s models). Missing 2020+ SOTA: XGBoost, CatBoost.	R4.3, R7.4	<b>MAJOR - SHOULD ADD</b> 1. Add XGBoost as 5th model (scikit-learn compatible) 2. Update all result tables 3. Justify why DL not used (small sample sizes) <b>Location:</b> Section 4.2, Table 1
E10	<b>Ablation study missing</b> Pipeline has multiple components (harmonization, log, IQR) but no validation of individual contributions.	R5.4, R7.6	<b>MAJOR - MUST ADD</b> 1. NEW Table: Ablation Study RF MMRE: (raw) — (+log) — (+log+IQR) — (full) 2. Show each step contributes <b>Location:</b> NEW Section 5.3 or 5.4
E11	<b>Interpretability claim unsupported</b> Paper says RF is interpretable but provides NO feature importance analysis.	R7.5	<b>MAJOR - MUST ADD</b> 1. NEW Figure: RF feature importance (Gini or SHAP) 2. NEW paragraph explaining top 3 features <b>Location:</b> NEW Section 5.3
E12	<b>R<sup>2</sup> column shows "—" (missing or unexplained)</b> Table 1 has R <sup>2</sup> column but all entries are "—". If computed, report. If not, remove or explain.	R6.4	<b>MAJOR - MUST FIX</b> 1. Compute R <sup>2</sup> for all models (formula in Section 2.3) 2. Fill in table OR 3. Remove column + explain "R <sup>2</sup> negative for some models" <b>Location:</b> Table 1, all result tables
E13	<b>Duplicate equation (Time = C × E<sup>D</sup>)</b> Section 2.1 presents Time equation twice with nearly identical wording.	R6.2	<b>MAJOR - DELETE</b> 1. Remove second instance of Time equation 2. Fix equation labels/refs <b>Location:</b> Section 2.1 lines 120-130

Error ID	Issue Description	Reviewers	Severity & Action Required
E14	<p><b>Assumptions &amp; Limitations section missing</b></p> <p>No explicit "Assumptions" (e.g., 160h/month, linear cost-effort, no team dynamics). Limitations vague.</p>	R3.3, R5.5	<p><b>MAJOR - MUST ADD</b></p> <ol style="list-style-type: none"> <li>1. NEW Section 3.6: Assumptions and Limitations (2 pages)</li> <li>2. List: (1) 160h/month uniform, (2) Historical data bias, (3) FP n=24 exploratory, (4) No team dynamics</li> </ol> <p><b>Location:</b> After Section 3.5</p>
E15	<p><b>Generalization unclear / no cross-org validation</b></p> <p>Tests on random holdouts from SAME pool. No leave-one-source-out (LOSO). No modern datasets (GitHub/Jira).</p>	R1.3, R5.1, R7.8	<p><b>MAJOR - ACKNOWLEDGE OR ADD</b></p> <ol style="list-style-type: none"> <li>1. Add LOSO CV experiment (train on dataset A, test on B) OR</li> <li>2. Acknowledge limitation: "Cross-org validation is future work"</li> <li>3. Add modern datasets OR cite limitation</li> </ol> <p><b>Location:</b> Section 4.1, Section 6</p>
E16	<p><b>Language quality / AI-generated tone</b></p> <p>Formulaic phrases: "it is worth noting," "captures nuances," "substantially." Sounds templated.</p>	R4.5, R7.2	<p><b>MAJOR - MANUAL REWRITE</b></p> <ol style="list-style-type: none"> <li>1. Run Grammarly</li> <li>2. Human editing to remove clichés</li> <li>3. Vary sentence structure</li> </ol> <p><b>Location:</b> Entire document</p>
E17	<p><b>Data imbalance not addressed</b></p> <p>Many small projects, few large. MSE loss gives equal weight. Large projects may be under-predicted.</p>	R8.3, R8.4	<p><b>MAJOR - ANALYZE OR ACKNOWLEDGE</b></p> <ol style="list-style-type: none"> <li>1. Analyze error by project size quantile (small/medium/large)</li> <li>2. Consider weighted loss or focal loss (cite: 10.1038/s41598-025-22853-y)</li> <li>3. Add to limitations: "Imbalance toward small projects"</li> </ol> <p><b>Location:</b> Section 3.3, Section 4.2</p>

Error ID	Issue Description	Reviewers	Severity & Action Required
E18	<p><b>Additional metrics missing (MAPE, MdMRE, RAE)</b></p> <p>Only reports MMRE, PRED(25), MAE, RMSE, <math>R^2</math>. Missing median-based metrics.</p>	R1.4	<p><b>MINOR - EASY ADD</b></p> <ol style="list-style-type: none"> <li>1. Add MdMRE, MAPE, RAE to Section 2.3</li> <li>2. Update all result tables</li> </ol> <p><b>Location:</b> Section 2.3, Table 1</p>
E19	<p><b>Confidence intervals missing</b></p> <p>Results show mean only (<math>MMRE=0.647</math>). Should report "Mean [95% CI]" from 10 seeds or bootstrap.</p>	R1.5	<p><b>MINOR - EASY ADD</b></p> <ol style="list-style-type: none"> <li>1. Compute bootstrap 95% CI for all metrics</li> <li>2. Change format: "0.647 [0.589, 0.712]"</li> </ol> <p><b>Location:</b> All result tables</p>
E20	<p><b>Paper structure: roadmap missing</b></p> <p>Introduction ends with contributions but no roadmap ("Section 2 presents..., Section 3...").</p>	R5.2	<p><b>MINOR - EASY ADD</b></p> <ol style="list-style-type: none"> <li>1. Add paragraph at end of Section 1: "The remainder of this paper is organized as follows..."</li> </ol> <p><b>Location:</b> Section 1, end</p>

## 10 CRITICAL PATH SUMMARY

Based on analysis of all 8 reviewers, the following issues are **BLOCKING** (must fix for acceptance):

### 10.1 6 FATAL Issues - Must Fix

1. **E1: COCOMO II uncalibrated** - R1, R7, R8 (2-3 days to fix)
2. **E2: "Overall" aggregation undefined** - R6, R8 (0.5 day to fix)
3. **E3: Figures missing captions** - R5, R7 (1 day to fix)
4. **E4: FP n=24 protocol inappropriate** - R6, R7, R8 (1 day to fix)
5. **E5: Dataset manifest missing** - R7, R8 (1 day to fix)
6. **E6: Target leakage (Developers)** - R8 (0.5 day to fix)

Total FATAL fixes: 6-7 days

### 10.2 11 MAJOR Issues - Should Fix

7. **E7: Novelty unclear** - R1, R3, R4, R8 (1 day - rewrite)
8. **E8: Related Work lacks comparison** - R3, R4, R5 (1 day)
9. **E9: XGBoost missing** - R4, R7 (1-2 days)
10. **E10: Ablation study missing** - R5, R7 (1 day)
11. **E11: Interpretability unsupported** - R7 (1 day)
12. **E12: R<sup>2</sup> column "–"** - R6 (0.5 day)
13. **E13: Duplicate equation** - R6 (0.1 day - quick fix)
14. **E14: Assumptions section missing** - R3, R5 (1 day)
15. **E15: Generalization unclear** - R1, R5, R7 (1-2 days OR acknowledge)
16. **E16: Language quality** - R4, R7 (1 day - manual rewrite)
17. **E17: Data imbalance** - R8 (1 day - analysis)

Total MAJOR fixes: 9-11 days

### 10.3 3 MINOR Issues - Nice to Have

18. **E18: Additional metrics** - R1 (0.5 day)
19. **E19: Confidence intervals** - R1 (1 day)
20. **E20: Roadmap paragraph** - R5 (0.1 day)

Total MINOR fixes: 1.6 days

---

#### CRITICAL PATH RECOMMENDATION:

- **Days 1-3:** Fix FATAL issues (E1-E6)
- **Days 4-7:** Fix MAJOR issues (E7-E14)

- **Days 8-10:** Polish & integration (E15-E20)

**If you can only fix 10 issues, prioritize:** E1, E2, E3, E4, E5, E6, E7, E8, E10, E11  
**Likelihood of acceptance:**

- FATAL only (6 issues): 60-70%
- FATAL + key MAJOR (10 issues): 75-85%
- All issues (20): 85-90%