Dear Editor and Reviewers,

We sincerely thank the Editor and all Reviewers for their thorough evaluation and constructive feedback on our manuscript. We have carefully addressed all concerns and suggestions, resulting in substantial improvements to both the technical content and presentation quality of the paper.

Below, we provide a point-by-point response to each Reviewer's comments. Changes in the manuscript are indicated by specific section and page references.

Summary of Major Revisions

- **Imbalance-Aware Learning:** Implemented quantile-based sample reweighting and stratified tail evaluation to address long-tailed effort distributions (Sections 3.6, 3.8; Abstract; Results)
- **Calibrated Baseline:** Replaced uncalibrated parametric models with scipy-fitted power-law baseline ensuring fair comparison (Section 2.1)
- **XGBoost Added:** Extended evaluation to include XGBoost alongside existing ensemble methods (Section 3.5; all results tables)
- **Statistical Significance Testing:** Added paired Wilcoxon tests with Holm-Bonferroni correction and Cliff's Delta effect sizes (Section 3.9, Table 2)
- **Feature Importance Analysis:** Conducted permutation importance analysis to address interpretability concerns (Section 4.10)
- **Ablation Study:** Systematic ablation quantifying preprocessing contributions (Section 4.6)

Response to Reviewer 1

Reviewer Comment	Our Response and Action Taken
R1.1: Novelty unclear beyond "unified pipeline" **Where revised:** Abstract (p.1); Introduction contributions (p.3, lines 128-130); Section 3.6 "Imbalance-Aware Training" (p.15); Results Section 4.X showing tail performance (p.28-30).	**Critically Revised.** We agree that a pipeline alone is insufficient novelty. We have repositioned the contribution to focus on a **benchmarking framework with imbalance-awareness**. Specifically, we addressed the long-tail distribution of effort data by integrating **quantile-based sample reweighting** (Section 3.6, Eq. 4). This moves beyond procedural harmonization to address the heteroscedastic nature of SEE data.

(Continued on next page)

(Continued from previous page)

Reviewer Comment	Our Response and Action Taken
R1.2: COCOMO II baseline fairness **Where revised:** Section 2.1 "Calibrated Size-Only Power-Law Baseline" (p.4-5, lines 142-175); Implementation Details paragraph explicitly mentions scipy (lines 171-175); Abstract (p.1, line 76).	**Implemented.** To ensure a fair comparison, we replaced default COCOMO II parameters with a **calibrated size-only power-law baseline** ($\log E = \alpha + \beta \log Size$) where (α, β) are optimized using `scipy.optimize.curve_fit` strictly on training folds of each seed. This ensures the baseline benefits from identical data availability as ML models.
R1.3: Modern datasets (DevOps/GitHub) **Where revised:** Section 6 paragraph on external validity (p.32); Scope & Limitations (Introduction, p.3); Future Work (Conclusions, p.38).	**Clarified & Added to Limitations.** We acknowledge the shift towards DevOps-based estimation. However, our scope focuses on early-stage estimation (LOC/FP/UCP) where runtime telemetry is unavailable. We have explicitly discussed this boundary in Section 6 "Threats to Validity" and added references to DevOps-based estimation studies.
R1.4: Additional metrics (MdMRE, MAPE, MdAE) **Where revised:** Section 2.3 "Evaluation Metrics" (p.6-7, Equations 4-9); All results tables now include MdMRE, MAPE, MdAE columns.	**Added.** We now report **Median MRE (MdMRE)**, **MAPE**, and **Median Absolute Error (MdAE)** alongside standard metrics. MdMRE/MdAE provide robust central tendency under heavy-tailed error distributions.
R1.5: Confidence intervals **Where revised:** Abstract (p.1); Section 3.7 Bootstrap CI methodology (p.16); Supplementary Tables S1-S2 (detailed CIs).	**Added.** All metrics reported as mean $\pm$ std across 10 random seeds. We additionally employ **bootstrap 95% confidence intervals** (1,000 iterations) for small-sample FP schema.
R1.6: Paper length reduction **Where revised:** Table 1 caption references Supplementary Materials (p.8); Detailed manifest in Table S1.	**Addressed.** We moved detailed provenance manifest (DOIs, URLs, MD5 hashes) to Supplementary Materials (Table S1), reducing main text length while maintaining full auditability.

(Continued on next page)

(Continued from previous page)

Reviewer Comment	Our Response and Action Taken
R1.7: Reproducibility (datasets + scripts) **Where revised:** Section 3.1 "Dataset Manifest" (p.7-8); Table 1 caption; Supplementary Table S1.	**Enhanced.** Table 1 now includes dataset summary with explicit deduplication percentages. Full provenance (source, year, DOI/URL, raw counts, deduplication rules, licenses, MD5 hashes) documented in Supplementary Table S1 with rebuild scripts for independent replication.

Response to Reviewer 2

Reviewer Comment	Our Response and Action Taken
R2.1: Aggregation definition unclear **Where revised:** Abstract (p.1, lines 78-79); Section 3.7 "Macro-Averaging Across Schemas" (p.16); All results tables clarify macro vs per-schema metrics.	**Defined.** We have formalized the aggregation metric. "Overall" results are calculated via **macro-averaging** across the three schemas (LOC, FP, UCP) to prevent the large LOC dataset ($n \approx 2765$) from overshadowing FP and UCP results: $m_{macro} = \frac{1}{3}\sum_{s\in\{LOC,FP,UCP\}} m^{(s)}$. Per-schema breakdowns provided in Tables 4-6.
R2.2: FP sample size ($n = 158$) unstable with 80/20 split **Where revised:** Section 3.3 Experimental Setup (p.12); FP results labeled as exploratory (Section 4.2); Limitations section (p.32).	**Methodology Updated.** We agree that 80/20 split on 158 samples is statistically fragile. We switched the validation protocol for FP schema to **Leave-One-Out Cross-Validation (LOOCV)**. We categorize FP findings as "exploratory" in Discussion.
R2.3: Dataset provenance and leakage control **Where revised:** Section 3.1 (p.7-8); Table 1 caption; Supplementary Table S1 with full audit trail.	**Added.** Dataset Manifest (Table S1) details source, DOI, original count, and cleaning logic for every dataset. Deduplication performed based on tuple matching {Project_Name, Size, Effort} to prevent leakage.
R2.4: Add robust metrics (MdAE/MdMRE)	**Added.** We now report Median Absolute Error (MdAE) and MdMRE alongside standard metrics, emphasizing MdAE as robust to extreme outliers typical in software engineering data.

(Continued on next page)

(Continued from previous page)

Reviewer Comment	Our Response and Action Taken
Where revised: Section 2.3 "Evaluation Metrics" (p.6-7); All results tables include MdMRE, MdAE columns.	
R2.5: Feature leakage (Developers variable) **Where revised:** Section 3.2 Preprocessing Pipeline (p.11); footnote clarifying feature exclusion criteria.	**Clarified.** We do **not** use Developers as an input feature precisely because it is often derived from Effort/Time (target leakage risk). We exclude any features derived from the target variable to prevent data leakage.

Response to Reviewer 3

Reviewer Comment	Our Response and Action Taken
R3.1: Introduction structure (What is known/missing/gap) **Where revised:** Introduction Section 1 (p.2-3, lines 88-105).	**Restructured.** Introduction now follows the recommended structure with explicit paragraphs: "What is known" (prior SEE research), "What is missing" (3 critical gaps), "Research gap" (our specific contribution), "Our approach" (4 concrete contributions).
R3.2: Related work comparison table **Where revised:** Section 5 "Related Work" (p.33-34); Table comparing methodological approaches.	**Enhanced.** We added a comprehensive comparison table (Table X) positioning our work relative to recent SEE studies, highlighting differences in provenance transparency, baseline fairness, and aggregation protocols.
R3.3: Assumptions & Limitations upfront **Where revised:** Introduction (p.3, lines 109-116); Section 6 "Threats to Validity" (p.32-33).	**Added.** Introduction now includes "Scope and limitations upfront" paragraph (5 constraints). Section 6 provides detailed threats to validity (internal, external, construct).
R3.4: Figure 1 description inadequate **Where revised:** Section 2.1 "Pipeline Overview" paragraph (p.5, lines 181-191); Figure 1 caption.	**Enhanced.** Pipeline Overview paragraph now provides detailed 4-stage description (Input, Preprocessing, Training, Evaluation). Figure caption expanded with methodology details.

(Continued from previous page)

Reviewer Comment	Our Response and Action Taken
R3.5: Paper organization **Where revised:** Introduction (p.3, lines 133-139).	**Added.** Introduction concludes with "Paper Organization" paragraph outlining all sections.

Response to Reviewer 4

Reviewer Comment	Our Response and Action Taken
R4.1: Introduction expanded with limitations **Where revised:** Introduction (p.2-3).	**Addressed.** See response to R3.1 and R3.3 above. Introduction now includes structured What is known/missing/gap framework plus explicit "Scope and limitations upfront" paragraph.
R4.2: Related work pros/cons + new citations **Where revised:** Section 5 "Related Work" (p.33-35); refs.bib updated.	**Enhanced.** Related Work section now includes pros/cons discussion of each approach category (parametric, ensemble, deep learning, transfer learning). Added recent citations including DOI 10.1109/TSMC.2025.3580086, 10.1109/TFUZZ.2025.3569741, 10.1109/TETCI.2025.3647653.
R4.3: XGBoost missing from evaluation **Where revised:** Section 3.5 "XGBoost (XGB)" (p.14); Table 1 Overall Performance (p.19); All per-schema tables (Tables 4-6); Statistical tests (Table 2, line 900 note).	**Experiments Expanded.** We added XGBoost to our benchmark suite. Results show that while XGBoost offers marginal gains in LOC schema, Random Forest remains most robust across small, heterogeneous datasets (FP/UCP).
R4.4: Post hoc statistical tests **Where revised:** Section 3.9 "Uncertainty & Significance Testing" (p.17-18); Table 2 "Post-hoc pairwise tests" (p.21); Section 4.4 discussion (p.23).	**Added.** We added **paired Wilcoxon signed-rank tests** with Holm-Bonferroni correction and **Cliff's Delta** (δ) effect sizes. RF outperforms calibrated baseline with large effects ($\delta = -0.52$, $p < 0.001$). Differences between RF and XGBoost are not statistically significant ($\delta = -0.08$, $p = 0.184$).

Response to Reviewer 8

Reviewer Comment	Our Response and Action Taken
R8.1: Limited novelty - framework is procedural **Where revised:** Abstract (p.1); Introduction contributions (p.3, lines 128-131); Sections 3.6 & 3.8 (p.15-17); Results tail analysis (p.28-30).	**Rebuttal & Enhancement.** We have addressed this by focusing on robustness under extreme imbalance. By demonstrating that **Imbalance-Aware Ensembles** (via quantile reweighting and stratified tail evaluation) significantly improve performance on the long-tail of large projects, we provide a **methodological contribution** beyond simple benchmarking. This shifts focus from claiming model superiority to establishing a reusable methodological artifact.
R8.2: Imbalance handling - suggests weighted/focal loss **Where revised:** Section 3.6 "Imbalance-Aware Training" (p.15); Section 3.8 "Stratified Evaluation" (p.17); Results Figure showing MAE by deciles (p.29); Discussion (p.31).	**Integrated.** We appreciate this vital suggestion. We incorporated **quantile-based sample reweighting** (Section 3.6, Eq. 4), assigning 4× weight to tail projects (90-100% effort). Stratified tail evaluation (Section 3.8) demonstrates that RF-weighted reduces tail MAE degradation from +296% (standard RF) to +232%. We acknowledge this as variance-stabilizing transformation aligning with principles in suggested reference (DOI: 10.1038/s41598-025-22853-y). Future work discusses focal loss adaptation for regression.
R8.3: Cross-schema transfer not explored **Where revised:** Introduction scope statement (p.3, line 112); Section 6 Limitations (p.32); Conclusions future work (p.38).	**Acknowledged as Limitation & Future Work.** Models trained independently per schema; no cross-schema generalization or transfer learning explored in this work. We explicitly state this constraint in Introduction "Scope and limitations" and Limitations section. Cross-schema transfer learning highlighted as future direction requiring feature alignment across LOC/FP/UCP paradigms.

We believe these comprehensive revisions substantially strengthen the manuscript's technical rigor, reproducibility, and contribution clarity. We are grateful for the Reviewers' insightful feedback, which has greatly improved the quality of our work.

Sincerely,
Nguyen Nhat Huy (corresponding author)
On behalf of all authors