PFMR: ALLOCATING SUPPORT BY LOCAL EFFECT FOR TRACE-BASED ROOT CAUSE LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A single microservice failure can produce anomalies across a request path, making it difficult to decide which service to inspect first. Statistical support for a change and the magnitude of local deterioration carry different diagnostic information. We propose PFMR, a trace-based method that uses positive local effect to allocate two-window rank support across services. It updates a change reference with effect-weighted support, pooling the two signals within each service. Under additive execution, we derive a condition for separating local deterioration from propagated anomalies; a finite support ceiling ensures that weak local effects contribute vanishingly little at fixed sample counts. PFMR achieves 78.13% and 87.92% service-level Top-1 accuracy on FSE1422 and RCAEval240. At a fixed total support contribution, it improves Top-1 by 5.70 and 8.33 percentage points over a common coefficient and by 3.38 and 3.33 points over normalized additive fusion. Prediction-defined groups show the largest gains when the reference and support prefer different services and the support-preferred service has the larger local effect.

1 INTRODUCTION

A failure in one microservice can alter latency and errors throughout a request path. Distributed traces expose the affected services, but a pronounced anomaly may lie away from the fault source (Fang et al., 2026; Pham, 2026). SparseRCA combines relative and standardized deviations of exclusive latency, while TraceRCA combines suspicious trace sets with service-level propagation cues (Yao et al., 2024; Li et al., 2021). Both connect local behavior with broader diagnostic evidence. The remaining service-level decision is how to prioritize changes that differ in both strength and diagnostic relevance.

Statistical strength and local deterioration need not agree. A propagated anomaly can have strong support because it is repeatedly observed, even when its local effect is small; a locally deteriorating service may have a larger effect but fewer observations. Service ranking should therefore let the magnitude of local deterioration determine how strongly statistical support updates the initial priority.

The distinction is visible within a single request path. If a child call becomes slower, the caller's total duration rises even when its own execution is unchanged. Repetition makes that caller's anomaly statistically clear, but does not establish where the delay originated. Conversely, a change confined to a service's local work can be diagnostically meaningful before its operation has accumulated the strongest two-window contrast. We use local execution and error movement to describe the direction and size of deterioration, and rank evidence to describe how consistently it appears in the observed windows. Their roles are distinct at the service level.

PFMR allocates two-window rank support according to positive local effect. Invocation and state changes establish a reference priority; positive movement in local duration and error rate determines the subsequent support adjustment. Pooling effect and support separately within a service allows operations with different response magnitudes and observation counts to contribute to the same ranking decision. Under an additive execution model, removing observed child duration preserves local changes. The resulting bounds give a sufficient condition for ranking a locally deteriorating root above propagation-affected services. At fixed sample counts, a finite support ceiling limits the influence of arbitrarily small local effects across rank configurations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

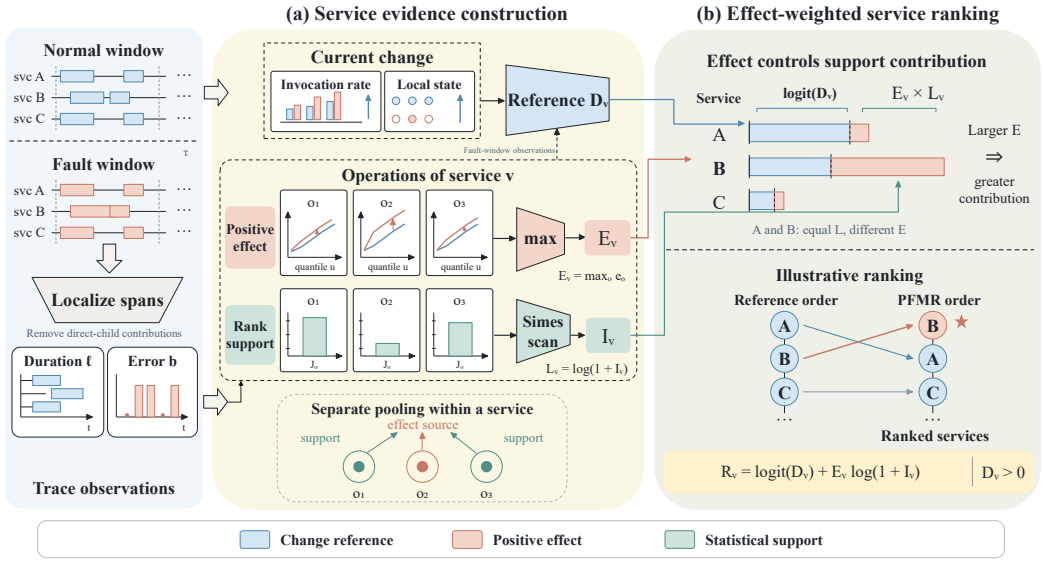

Figure 1: PFMR framework for effect-weighted service ranking. From normal and fault traces, PFMR constructs a change reference and separately pools positive local effects and statistical support within each service. The pooled effect controls how strongly support adjusts the reference ranking, as illustrated by services with equal log-support but different effects.

The evidence centers on how support is allocated. With the total contribution fixed, PFMR improves Top-1 over a common coefficient by 5.70/8.33 percentage points on FSE1422/RCAEval240, including when the control retains a positive-effect screen. A normalized additive rule with the same total contribution trails by 3.38/3.33 points. Prediction-defined groups show the largest gains when the reference and support prefer different services and the support-preferred service has greater effect. Together, these comparisons link service-level accuracy to the allocation of support across different degrees of local deterioration.

The paper makes three contributions:

- A service-level ranking rule that uses positive duration and error movement to determine how two-window rank support updates a change reference.
- A conditional separation result for local execution changes and a characterization of effect-dependent allocation under a fixed total support contribution.
- An evaluation on the complete FSE1422 and RCAEval240 populations, with 78.13% and 87.92% Top-1 accuracy and controlled evidence for service-specific allocation.

2 PFMR: EFFECT-WEIGHTED SERVICE RANKING

PFMR ranks services observed in normal and fault trace windows separated by a boundary τ. For each service v, a change reference D_v provides the initial priority, a positive local effect E_v measures deterioration, and statistical support I_v quantifies a two-window rank contrast. The ranking coordinate is

$$R_v = \text{logit}(D_v) + E_v \log(1 + I_v), \qquad D_v > 0. \tag{1}$$

Thus the effect determines how much log-support modifies the reference ranking. We write $L_v = \log(1 + I_v)$ for log-support and $T_v = E_v L_v$ for its contribution. The logarithm compresses the support scale while retaining differences between services.

2.1 MEASURING POSITIVE LOCAL EFFECT

The effect is built from local duration and error observations at each operation. Local duration subtracts the summed durations of a span's direct children from its duration and clips the result at zero. Local error retains a span error only when no direct child reports one. These coordinates focus the comparison on behavior attributed to the observed span; Appendix B specifies their construction.

For operation o, let $n_{0,o}, n_{1,o}$ be its normal and fault counts, $N_o = n_{0,o}+n_{1,o}$, and $\alpha_o = N_o/(N_o+1)$. For operations observed in both windows, transform local durations by $z = \log(1 + \ell/s_o)$, where s_o is a positive duration scale estimated from the observations. Let μ_0, μ_1 be the resulting empirical distributions. We measure positive duration movement by

$$C_o = \frac{W_1(\mu_0, \mu_1) + \bar{z}_1 - \bar{z}_0}{2}. \tag{2}$$

Here W_1 is the one-dimensional Wasserstein distance. Under quantile coupling, this expression averages only positive duration movement: upward shifts contribute, whereas downward shifts contribute zero (Ramdas et al., 2017). With local error rates ρ_0, ρ_1, the operation and service effects are

$$\delta_o = \frac{\alpha_o}{2}\left(1 - e^{-C_o}, [\rho_1 - \rho_0]_+\right), \qquad e_o = \|\delta_o\|_2, \qquad E_v = \max_{o \in \mathcal{A}_v} e_o, \tag{3}$$

where $\mathcal{A}_v$ is the set of operations belonging to service v. The bounded coordinates measure positive duration movement and error-rate increase. The maximum selects the strongest local effect within a service; an operation observed in only one window has $e_o = 0$.

The duration coordinate responds to both the size of a positive shift and the fraction of observations that move upward. A shift in the lower part of the duration distribution can therefore contribute alongside a change in the tail; a purely downward movement contributes nothing to the effect. Local error is a separate coordinate, so a service can register deterioration through errors even if its duration does not increase. These choices keep the effect tied to the behavior that the ranking seeks to prioritize, while allowing distinct operations to witness duration and error changes.

2.2 ESTABLISHING A REFERENCE FROM CURRENT OBSERVATIONS

The reference combines invocation-rate changes with positive local-state changes. Let $\lambda_{b,o} = n_{b,o}/T_b$ denote the invocation rate in phase b for window exposure T_b. Define the relative rate mismatch as $a_o = |\lambda_{1,o} - \lambda_{0,o}|/(\lambda_{1,o} + \lambda_{0,o})$ and the bounded state effect as $h_o = \min\{1, \sqrt{2}e_o\}$. Their combination is

$$d_o = \alpha_o\{a_o + (1 - a_o)h_o\}, \qquad D_v = \max_{o \in \mathcal{A}_v} \mathbf{1}[n_{1,o} > 0]d_o. \tag{4}$$

The rate mismatch contributes directly, while the matched fraction contributes through positive state change. The finite-count factor keeps D_v below one.

Only operations observed during the incident contribute to the reference. A newly appearing operation can contribute through its invocation rate; an operation seen only in the normal window contributes zero. For operations present in both windows, rate changes in either direction remain eligible.

2.3 POOLING STATISTICAL SUPPORT

For operations observed in both windows, support measures a covariance-adjusted contrast in the mean ranks of local duration and error, following the rank-transform multivariate two-sample perspective (Nath & Duran, 1983). Its nonnegative strength J_o has a sample-count ceiling $K_o = [\log\binom{N_o}{n_{0,o}} - \log 2]_+$ in both regular and singular covariance cases; Appendix B.2 specifies the statistic. The contrast detects rank-location changes in either direction, whereas the effect retains positive deterioration.

The covariance-adjusted mean-rank contrast summarizes the observed rank-location shift in local duration and error. Its count-dependent ceiling puts regular and singular covariance branches on a common finite range before service-level pooling. Positive transport measures the magnitude and direction of local duration movement, while the rank contrast supplies the strength of the observed

location shift. Their separate roles allow local deterioration to determine the contribution of rank evidence.

For a service with m_v two-window operations, order the strengths as $J_{(1)} \geq \cdots \geq J_{(m_v)}$. PFMR aggregates them using the Simes scan form (Simes, 1986):

$$I_v = \max_{k \in \mathcal{K}_v} \left[J_{(k)} - \log \frac{m_v}{k} \right]_+ . \tag{5}$$

The index set $\mathcal{K}_v$ contains the last position in each tie block, so every candidate represents a complete set of operations at or above a support level. The penalty falls as more operations support that level. If no operation appears in both windows, $I_v = 0$.

Separate pooling allows an operation with the largest effect and another with the strongest rank contrast to contribute to one service-level decision. Section 3 compares this rule with pooling that requires the effect to come from the selected support set.

Selection interpretation. On services with positive reference, Equation (1) orders the exponential tilt of reference odds $p_v \propto D_v/(1 - D_v)$ by $E_v L_v$. This is the standard KL-regularized selection solution (Wainwright & Jordan, 2008; Bissiri et al., 2016), stated in Appendix A.6. Its weights are diagnostic priorities. An equivalent stable score handles the ranking and zero-reference services as specified in Appendix B.

3 Local Effect and Support Allocation

Two properties connect the ranking rule to diagnosis. Local execution changes determine when effect weighting can distinguish a root from propagated anomalies; fixed-total allocation determines which services receive more support weight. The analysis assumes finite sets and nonnegative effects and supports. Derivations appear in Appendix A.

3.1 Local Changes and Diagnostic Separation

Consider an operation's coupled normal and fault executions. In an additive call model, write span duration as $Y_b = X_b + W_b + U_b$, where X_b is local execution, W_b is the summed duration of observed direct children, and U_b is additional local perturbation. If local execution is unchanged, the residual satisfies $|\ell_1 - \ell_0| \leq |U_1 - U_0|$ regardless of the change in W_b. Removing child duration therefore cancels downstream waiting while retaining local changes. At a fixed scale $s_o > 0$, the transformed difference is bounded by $|U_1 - U_0|/s_o$.

Suppose a root service r has an operation whose transformed duration quantiles rise by at least $\delta > 0$ on probability mass $\mu > 0$. At a competing service b, assume that each two-window operation admits a coupling with unchanged local execution, $|U_1 - U_0|/s_o \leq \varepsilon_b$, and positive local error-rate change at most η_b, where $\varepsilon_b, \eta_b \geq 0$. Equation (3) then gives

$$E_r \geq e_r^- := \frac{\alpha_o}{2}(1 - e^{-\mu\delta}), \qquad E_b \leq e_b^+ := \frac{1}{2}\sqrt{\varepsilon_b^2 + \eta_b^2}. \tag{6}$$

The lower bound follows from the root's positive quantile movement; the upper bound follows from the competing service's residual coupling and $1 - e^{-C} \leq C$.

Proposition 1 (Local-separation condition)**.** *Under these conditions, for $D_r, D_b > 0$, PFMR ranks r above b whenever*

$$\operatorname{logit}(D_b) - \operatorname{logit}(D_r) < e_r^- L_r - e_b^+ L_b. \tag{7}$$

The condition compares the initial reference gap with a lower bound on the root's supported deterioration and an upper bound on the competitor's contribution. Even strong support at b has no effect when it arises solely from downstream waiting with unchanged local errors: then $\varepsilon_b = \eta_b = 0$. For nonzero local perturbations, the bound explicitly scales their influence by L_b.

The comparison is between a root with supported local deterioration and a competitor whose observable change is carried by the call path. The reference may initially favor either service; the condition identifies when the root's effect-weighted support exceeds the remaining reference gap. It does not require support to be largest at the root. Instead, it requires the competing service's own execution and error changes to be small relative to the root's supported change. This is the setting in which a repeated propagation signal can be distinguished from the local source of deterioration.

3.2 LOCAL EFFECT CONTROLS RANK EVIDENCE

Rank support need not diminish with the magnitude of a positive shift. Constant normal observations and fault observations shifted upward by any $\delta > 0$, however small, have the same pooled ranks. Figure 2 shows how effect magnitude distinguishes these changes. A finite support ceiling extends the attenuation beyond this example: at fixed operation counts, let $K_v^{\max} = \max_o K_o$, with the empty maximum zero. The support scan satisfies

$$0 \le I_v \le K_v^{\max}, \qquad 0 \le T_v \le E_v \log(1 + K_v^{\max}). \tag{8}$$

As E_v approaches zero, its contribution vanishes uniformly over rank configurations at these counts, including different tie patterns and covariance ranks. In the separation condition, a competitor's contribution is consequently at most $e_b^+ \log(1 + K_b^{\max})$. This relates local perturbation to the largest support adjustment it can receive.

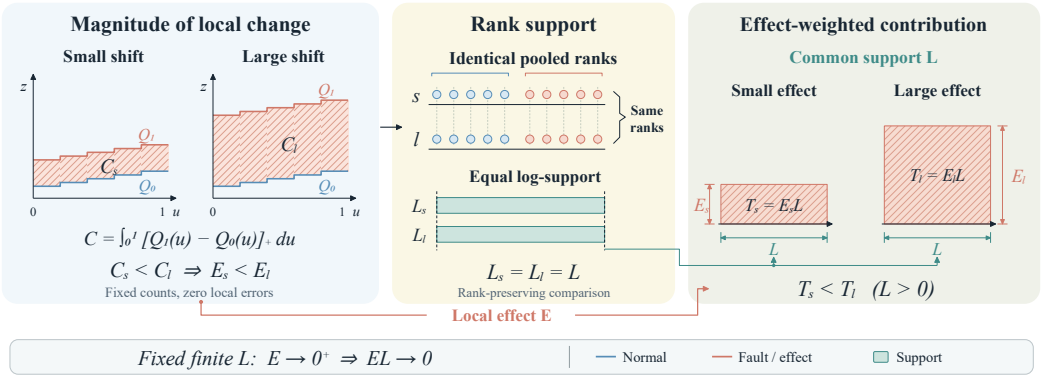

Figure 2: Local effect distinguishes rank-preserving changes. In this single-operation illustration with fixed sample counts and zero local errors, equal log-support accompanies different positive quantile movements. Effect weighting assigns different contributions and continuously attenuates them as local effect vanishes at fixed finite log-support.

Normalized additive fusion can assign positive support contribution to an arbitrarily small but positive effect. Under effect weighting, that contribution decreases continuously with effect magnitude. Appendix A.3 derives the comparison for the additive rule evaluated in Section 4.

3.3 ALLOCATION AT A FIXED TOTAL CONTRIBUTION

Define a common coefficient by

$$c = \frac{\sum_v E_v L_v}{\sum_v L_v}, \qquad \sum_v c L_v = \sum_v E_v L_v, \tag{9}$$

with $c = 0$ when the denominator is zero. Relative to this shared coefficient, PFMR redistributes the same total contribution toward services with $E_v > c$ and away from those with $E_v < c$. The exact pairwise correction appears in Appendix A.7; Section 4 tests the resulting ranking changes.

Conservation makes the comparison sensitive to *where* the contribution is assigned rather than to its aggregate size. For a service with $E_v > c$, PFMR raises the support adjustment above the common-coefficient rule; for $E_v < c$, it reduces that adjustment. The effect on a pair depends on both services' effect deviations and their respective log-support values, so allocation can reverse a reference preference without requiring either signal alone to rank the root first. The common-coefficient experiment holds the aggregate contribution fixed and tests whether this redistribution improves diagnostic ordering.

Pooling within a service. A shared perturbation can elicit responses of different magnitudes across a service's operations, which may also have different observation counts. Separate pooling retains the strongest effect and support even when they arise from different operations. Appendix A.8 constructs this case and characterizes when separate pooling equals a rule that requires effect and support to share a selected set. The 2×2 comparison evaluates both rules under two reference definitions.

Table 1: Localization accuracy (%) on the complete evaluation populations. Bold marks the best result per metric. Evaluation settings appear in Appendix C.2.

Method	Top-1	Top-3	Top-5	Avg@5
FSE1422				
BARO (Pham et al., 2024)	37.76	53.59	63.43	52.32
BARO-T (trace-series control)	21.24	33.12	47.47	33.88
MicroDig (trace) (Tao et al., 2024)	41.84	67.72	73.49	62.50
MicroRCA (Wu et al., 2020)	43.53	56.26	69.41	57.16
MicroRank (Yu et al., 2021)	5.56	21.59	30.10	20.04
SimpleRCA (Fang et al., 2026)	32.21	67.23	87.90	64.30
TraceRCA (Li et al., 2021)	42.55	65.75	74.82	62.52
SparseRCA (Yao et al., 2024)	20.60	43.04	55.91	40.59
PFMR	**78.13**	**91.49**	**94.37**	**89.13**
RCAEval240				
BARO (Pham et al., 2024)	35.42	86.25	91.25	76.75
RCD (Ikram et al., 2022)	35.00	43.33	50.42	42.67
ϵ-Diagnosis (Shan et al., 2019)	10.00	22.50	30.00	21.17
MicroRank (Yu et al., 2021)	11.25	36.67	52.50	33.58
PRISM (Pham, 2026)	74.58	93.75	97.08	90.08
Score Ordering (first point) (Orchard et al., 2025)	5.42	17.92	30.00	17.83
TraceRCA (Li et al., 2021)	29.17	57.92	73.33	54.08
SparseRCA (Yao et al., 2024)	38.75	65.00	77.50	61.50
PFMR	**87.92**	**97.50**	**99.17**	**95.75**

4 EXPERIMENTS

The experiments examine overall localization, the contribution of effect-dependent support allocation, and the prediction conditions under which it changes accuracy. A factorial comparison evaluates the reference and pooling choices.

4.1 EVALUATION SETUP

We evaluate on the complete FSE1422 population of the fault propagation-aware benchmark (Fang et al., 2026) and the complete RCAEval240 population (Pham et al., 2025). RCAEval240 comprises Online Boutique and Train Ticket under both RE2 and RE3. PFMR uses all available traces on each side of the normal/fault boundary. Comparisons include trace-based, metric-based, and multi-source methods; Appendix C.2 specifies their inputs and settings.

Top-k is the percentage of the evaluation population whose accepted root cause appears among the first k ranked services. We report Top-1, Top-3, Top-5, and Avg@5, the mean of Top-1 through Top-5. Errors and misses remain in the denominator. PFMR configurations share observations and service candidates; evaluation labels determine correctness after ranking. Confidence intervals use 20,000 paired case bootstrap resamples. Appendix C details mapping and scoring.

The internal comparisons hold the extracted operation statistics and candidate service set fixed. This makes each ranking rule face the same changes in invocation frequency, local duration, errors, and observation counts. The common-coefficient rule controls total support contribution, the positive-effect screen controls eligibility, and normalized additive fusion supplies a direct alternative to multiplicative allocation. Their paired differences ask whether service-specific effect magnitude changes which service is ranked first, rather than whether a different front end detects more anomalies.

Table 2: Five configurations on shared operation statistics (%). T1/T3/T5 denote Top-1/3/5 and A5 denotes Avg@5. Common E preserves the total additive support potential while replacing service-specific coefficients. Bold marks the best value.

	FSE1422				RCAEval240			
Configuration	T1	T3	T5	A5	T1	T3	T5	A5
PFMR	**78.13**	**91.49**	**94.37**	**89.13**	**87.92**	**97.50**	99.17	**95.75**
Change only	61.18	83.97	89.45	79.85	78.33	93.75	97.50	91.42
Support only	41.00	63.43	77.00	61.86	57.92	79.17	85.00	74.75
$E = 1$	76.30	90.44	93.32	87.78	78.75	94.58	**99.58**	92.17
Common E	72.43	88.40	92.62	85.60	79.58	95.83	98.33	92.83

4.2 Localization accuracy

Table 1 shows service-level Top-1 accuracy of 78.13% on FSE1422 and 87.92% on RCAEval240 for PFMR. These exceed the strongest listed comparisons, MicroRCA and PRISM, by 34.60 and 13.33 percentage points, respectively. PFMR reaches 91.49% and 97.50% Top-3 accuracy.

Two related trace-based methods examine complementary forms of evidence: local deviation in SparseRCA and suspicious trace sets in TraceRCA. Their Top-1 accuracies on FSE1422/RCAEval240 are 20.60%/38.75% and 42.55%/29.17%, respectively. PFMR ranks higher on both populations, with paired intervals in Table 6. We next compare ranking rules that share PFMR's service statistics.

4.3 The contribution of support allocation

Table 2 compares five rules using the same service statistics. Change only ranks by the reference; support only ranks by statistical strength; $E = 1$ adds unweighted log-support to the reference. The reference already includes local effect. Common E replaces service-specific effects with the shared coefficient in Equation (9), preserving the total additive support contribution.

Combining the signals. On FSE1422/RCAEval240, change only reaches 61.18%/78.33% Top-1 and support only reaches 41.00%/57.92%. Combining the reference with unweighted support raises accuracy to 76.30%/78.75%; effect weighting adds 1.83/9.17 percentage points. The paired 95% intervals are $[0.00, 3.66]$/$[5.42, 13.33]$.

Allocation beyond total weight and screening. PFMR exceeds common E by 5.70/8.33 percentage points, with paired intervals of $[3.80, 7.59]$/$[5.00, 12.08]$. A common coefficient restricted to positive-effect services has the same aggregate metrics as common E (Appendix C.3). Thus the observed gain persists when both the total contribution and eligibility for support are held fixed.

Common E matches PFMR's total support contribution for every evaluation instance. The paired gains therefore locate the benefit in the distribution of that contribution across services with different local effects.

Comparison with additive fusion. The normalized additive control divides the same total contribution equally between normalized effect and support, while retaining the reference and positive-effect eligibility. It reaches 74.75%/84.58% Top-1, compared with PFMR's 78.13%/87.92%. The paired differences are 3.38/3.33 percentage points, with intervals $[1.83, 4.92]$/$[0.83, 5.83]$; Avg@5 improves by 2.35/1.33 points. The result distinguishes effect-dependent support allocation from this direct additive fusion. Appendix C.4 gives the rule and complete metrics.

4.4 Where does allocation help?

The change and support rules choose different first services for 70.68% of FSE1422 and 50.00% of RCAEval240. Within these disagreement groups, PFMR gains 8.76/16.67 Top-1 percentage points over common E; the corresponding differences in agreement groups are -1.68/0.00 points.

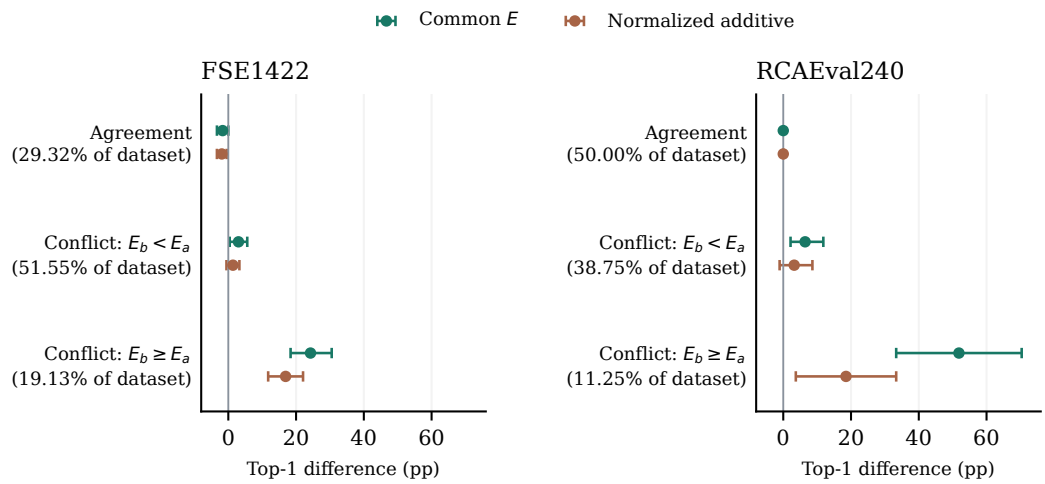

Figure 3: Allocation gains by prediction-defined effect relationship. In conflicts, a is the reference choice and b the support choice. Points show PFMR's within-group Top-1 difference from common E and normalized additive fusion; bars are paired 95% intervals. Percentages beside group labels give shares of the complete dataset. The effect relationship stratifies disagreements between the reference and support choices; correctness is evaluated within each group.

Table 3: Reference and pooling choices: Top-1 accuracy (%). The reference includes all operations or only those currently observed. Effect and support are pooled across the service or drawn from a common support set.

Reference	Pooling	FSE1422	RCAEval240
All contributions	Restricted	75.81	85.83
All contributions	Service	75.95	86.67
Current observations	Restricted	77.43	87.08
Current observations	Service	78.13	87.92

To relate these conflicts to local effect, let a and b denote the reference-only and support-only first services. We partition conflicts by $E_b < E_a$ and $E_b \geq E_a$ according to the two preferred services' effects, then evaluate correctness within each group. Figure 3 reports both conflict groups and agreement, each with its full group denominator.

When $E_b < E_a$, PFMR's gains over common E are 3.00/6.45 percentage points and its gains over additive fusion are 1.36/3.23 points. When $E_b \geq E_a$, these gains rise to 24.26/51.85 and 16.91/18.52 points. Direct comparisons between the disjoint groups confirm the larger gain in the latter group (Appendix C.6). The pattern links allocation gains to disagreement between the change reference and locally supported deterioration. The paired outcome comparison shows that the larger score corrections when $E_b \geq E_a$ also favor correct first choices, connecting the effect relationship to diagnostic value.

4.5 Reference and Pooling Choices

Table 3 crosses two choices: whether normal-only operations contribute to the reference and whether effect must arise from the selected support set. All four configurations share operation statistics. Restricting the reference to current observations yields higher Top-1 under either pooling rule.

Under the current-observation reference, separate pooling changes Top-1 by $+0.70/+0.83$ percentage points relative to restricted pooling, with paired intervals $[-0.14, 1.55]/[0.00, 2.08]$. Proposition 2 characterizes when separate pooling retains both maxima from different operations. Appendix C gives the complete metrics and fault-type strata.

5 Related work

Local deviations and trace evidence. SparseRCA models exclusive latency conditional on span patterns and child-call counts. It multiplies relative and standardized deviations for the same span, updates scores through the topology, and takes a service maximum (Yao et al., 2024). TraceRCA multiplies a suspicious service set's Jaccard score by a service's incoming–outgoing abnormal-trace count difference, then maximizes over sets (Li et al., 2021). Its set score measures abnormal-request association, while the within-set score examines whether abnormal invocations enter or leave a service. These methods already demonstrate the value of combining local and contextual signals; they differ in what the contextual term measures and at which level the combination is made. MicroRank integrates normal-operation calibration, trace abnormality, graph weights, and spectrum localization (Yu et al., 2021).

PFMR also combines local behavior with broader trace evidence, but its combination addresses support allocation across services. Positive duration movement and increased local errors measure effect, while a two-window mean-rank contrast measures support as a function of observation counts and covariance. The quantities may arise from different operations within a service; their pooled product adjusts a change reference. This service-level allocation is the object of the local-separation analysis and fixed-total comparisons.

The distinction matters when operation-level evidence is uneven within one service. An operation with a large local shift may be observed infrequently, while another operation of the same service supplies a strong rank contrast. PFMR pools those observations at the diagnostic unit that is ultimately ranked. Its positive-effect factor also distinguishes two rank contrasts of equal strength when their accompanying local changes differ in magnitude. The separation condition and the common-coefficient and additive comparisons examine these two consequences directly.

Combining evidence across telemetry. Other approaches draw on telemetry beyond individual trace operations. MicroRCA combines application and resource observations in a graph (Wu et al., 2020), and MicroDig uses a heterogeneous graph to distinguish calling from causal relationships (Tao et al., 2024). BARO combines robust anomaly scoring with change-point detection (Pham et al., 2024); ϵ-Diagnosis uses two-sample analysis (Shan et al., 2019); and RCD uses intervention indicators for causal discovery (Ikram et al., 2022).

Graph-free service ranking. PRISM separately pools internal and external component properties before combining their scores (Pham, 2026). Score Ordering ranks marginal anomalies and derives shortlist conditions for polytree models (Orchard et al., 2025). PFMR instead asks how a service's positive local effect should determine the contribution of statistical support to its ranking.

6 Conclusion

PFMR uses positive local deterioration to determine how two-window rank support changes service priority. The analysis gives a condition under which local execution changes separate a root from propagated anomalies and shows that weak effects have uniformly small support contributions at fixed sample counts. On FSE1422 and RCAEval240, service-specific allocation improves localization over fixed-total common-coefficient and normalized additive controls. The largest gains occur when the reference and support prefer different services and the support-preferred service has the larger local effect. Together, the theory and comparisons show why the diagnostic value of statistical support depends on the magnitude of local deterioration.

AI USE STATEMENT

AI tools assisted with writing, editing, and polishing the manuscript. The authors reviewed the text and are responsible for its content.

ETHICS STATEMENT

This study evaluates service diagnosis using existing microservice benchmark telemetry. Root-cause labels are used only to score rankings after inference. The method outputs an investigation priority order.

REPRODUCIBILITY STATEMENT

Section 2 defines the ranking rule, and Section 3 analyzes diagnostic separation and support allocation. Appendix A provides the derivations, Appendix B specifies the statistics and numerical conventions, and Appendix C records the evaluation settings and full results. Appendix C also defines the scoring and interval conventions used for the reported comparisons.

REFERENCES

Pier Giovanni Bissiri, Christopher C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130, 2016. doi: 10.1111/rssb.12158.

Aoyang Fang, Songhan Zhang, Yifan Yang, Haotong Wu, Junjielong Xu, Xuyang Wang, Rui Wang, Manyi Wang, Qisheng Lu, and Pinjia He. Rethinking the evaluation of microservice RCA with a fault propagation-aware benchmark. *Proceedings of the ACM on Software Engineering*, 3(FSE): 1605–1627, 2026. doi: 10.1145/3797100.

Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Kumar Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures in microservices through causal discovery. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31158–31170, 2022. doi: 10.52202/068431-2259.

Zeyan Li, Junjie Chen, Rui Jiao, Nengwen Zhao, Zhijun Wang, Shuwei Zhang, Yanjun Wu, Long Jiang, Leiqin Yan, Zikai Wang, Zhekang Chen, Wenchi Zhang, Xiaohui Nie, Kaixin Sui, and Dan Pei. Practical root cause localization for microservice systems via trace analysis. In *IEEE/ACM International Symposium on Quality of Service (IWQoS)*, pp. 1–10, 2021. doi: 10.1109/IWQOS52092.2021.9521340.

Ravinder Nath and Benjamin S. Duran. A robust test in the multivariate two-sample location problem. *American Journal of Mathematical and Management Sciences*, 3(3):225–249, 1983. doi: 10.1080/01966324.1983.10737125.

William Roy Orchard, Nastaran Okati, Sergio Hernan Garrido Mejia, Patrick Blöbaum, and Dominik Janzing. Root cause analysis of outliers with missing structural knowledge. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5399.

Luan Pham. Graph-free root cause analysis. *arXiv preprint arXiv:2601.21359*, 2026. URL https://arxiv.org/abs/2601.21359.

Luan Pham, Huong Ha, and Hongyu Zhang. BARO: Robust root cause analysis for microservices via multivariate bayesian online change point detection. *Proceedings of the ACM on Software Engineering*, 1(FSE):2214–2237, 2024. doi: 10.1145/3660805.

Luan Pham, Hongyu Zhang, Huong Ha, Flora Salim, and Xiuzhen Zhang. RCAEval: A benchmark for root cause analysis of microservice systems with telemetry data. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 777–780, 2025. doi: 10.1145/3701716.3715290.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017. doi: 10.3390/e19020047.

Huasong Shan, Yuan Chen, Haifeng Liu, Yunpeng Zhang, Xiao Xiao, Xiaofeng He, Min Li, and Wei Ding. ϵ-diagnosis: Unsupervised and real-time diagnosis of small-window long-tail latency in large-scale microservice platforms. In *The World Wide Web Conference*, pp. 3215–3222, 2019. doi: 10.1145/3308558.3313653.

R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3): 751–754, 1986. doi: 10.1093/biomet/73.3.751.

Lei Tao, Xianglin Lu, Shenglin Zhang, Jiaqi Luan, Yingke Li, Mingjie Li, Zeyan Li, Qingyang Yu, Hucheng Xie, Ruijie Xu, Chenyuan Hu, Canqun Yang, and Dan Pei. Diagnosing performance issues for large-scale microservice systems with heterogeneous graph. *IEEE Transactions on Services Computing*, 17(5):2223–2235, 2024. doi: 10.1109/TSC.2024.3402172.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Li Wu, Johan Tordsson, Erik Elmroth, and Odej Kao. MicroRCA: Root cause localization of performance issues in microservices. In *IEEE/IFIP Network Operations and Management Symposium (NOMS)*, pp. 1–9, 2020. doi: 10.1109/NOMS47738.2020.9110353.

Zhenhe Yao, Haowei Ye, Changhua Pei, Guang Cheng, Guangpei Wang, Zhiwei Liu, Hongwei Chen, Hang Cui, Zeyan Li, Jianhui Li, Gaogang Xie, and Dan Pei. SparseRCA: Unsupervised root cause analysis in sparse microservice testing traces. In *IEEE International Symposium on Software Reliability Engineering (ISSRE)*, pp. 391–402, 2024. doi: 10.1109/ISSRE62328.2024.00045.

Guangba Yu, Pengfei Chen, Hongyang Chen, Zijie Guan, Zicheng Huang, Linxiao Jing, Tianjun Weng, Xinmeng Sun, and Xiaoyun Li. MicroRank: End-to-end latency issue localization with extended spectrum analysis in microservice environments. In *Proceedings of the Web Conference 2021*, pp. 3087–3098, 2021. doi: 10.1145/3442381.3449905.

A SUPPORTING MATHEMATICAL ANALYSIS

Throughout, operation and service sets are finite, statistics are finite real values, and logarithms are natural. We use the convention $0 \log 0 = 0$ in KL divergence.

A.1 DIRECTIONAL DURATION MOVEMENT

Let p_i, q_j be empirical sample weights and γ_{ij} any coupling of the corresponding distributions. Decomposing the positive part gives

$$\sum_{ij} \gamma_{ij}[y_j - x_i]_+ = \frac{1}{2}\left(\sum_{ij} \gamma_{ij}|y_j - x_i| + \sum_j q_j y_j - \sum_i p_i x_i\right). \tag{10}$$

The signed term is fixed by the marginals, so a coupling that minimizes absolute transport cost also minimizes positive transport cost. This yields Equation (2) and its one-dimensional quantile form $C = \int_0^1 [Q_1(u) - Q_0(u)]_+ \, du$ (Ramdas et al., 2017). The quantity includes positive transformed-duration movement caused by a change in spread.

A.2 PROOF OF THE LOCAL-SEPARATION CONDITION

In the coupled executions of Section 3, removing observed child duration gives $\ell_b = [X + U_b]_+$. Because the positive-part map is 1-Lipschitz, $|\ell_1 - \ell_0| \le |U_1 - U_0|$. For nonnegative durations at a shared positive scale, the derivative of $\log(1 + \ell/s_o)$ is at most $1/s_o$; hence $|z_1 - z_0| \le |U_1 - U_0|/s_o$. Changes in summed child waiting can be arbitrary under this model, while U_b captures the remaining local perturbation.

For each competing operation, the assumed coupling has positive transport cost at most ε_b, so the minimum cost satisfies $C_o \le \varepsilon_b$. The inequalities $\alpha_o \le 1$ and $1 - e^{-C_o} \le C_o$ bound its duration coordinate by $\varepsilon_b/2$; its error coordinate is at most $\eta_b/2$. Taking their Euclidean norm and then the maximum over operations gives the upper bound in Equation (6). One-window operations have zero effect by definition.

For the root operation, integrating positive quantile movement over mass μ gives $C_o \ge \mu\delta$. The effect norm bounds its duration coordinate from above, and the service maximum bounds the operation effect from above, establishing the lower bound in Equation (6). Since $L_r, L_b \ge 0$, the two effect bounds imply $R_r - R_b \ge \operatorname{logit}(D_r) - \operatorname{logit}(D_b) + e_r^- L_r - e_b^+ L_b$. Equation (7) makes this margin positive. If it holds against every positive-reference competitor, r ranks first because all zero-reference services follow positive-reference services.

A.3 SMALL EFFECTS UNDER MULTIPLICATIVE AND ADDITIVE ALLOCATION

Consider n normal observations equal to one and n fault observations equal to $1 + \delta$, with $\delta > 0$ and no local errors. All ranks within each window coincide, but the window mean ranks differ. The covariance is zero and the contrast is nonzero; Equation (16) therefore gives $J = \log \binom{2n}{n} - \log 2$ for every positive δ. The effect is $\alpha\delta/[2(2+\delta)]$ and tends to zero with δ. Thus, at fixed n, effect-weighted support vanishes while rank support stays constant. A numerical check with $n = 20$ and amplitudes from 1 to 10^{-6} gives constant support 24.9563, while the effect falls from 0.16260 to 2.4390×10^{-7}.

For the additive control in Appendix C.4, $T^{\text{add}} = B(E/A + gL/C)/2$. Hold the other services fixed at positive total potential B_0 and consider $E > 0$ at fixed $L > 0$. Then $C = C_0 + L$ and $B = B_0 + EL$. By nonnegativity, $T^{\text{add}} \ge B_0 L/(2C)$, whereas PFMR assigns EL. Whenever $0 < E < B_0/(2C)$, the effect-weighted contribution is strictly smaller. The comparison isolates attenuation among services that remain eligible for support; the empirical evaluation tests its allocation consequences on the full populations.

Every operation strength is bounded by $K_v^{\max}$, and each scan penalty $\log(m_v/k)$ is nonnegative for $1 \le k \le m_v$. Therefore each threshold value, and hence I_v, is at most $K_v^{\max}$. Logarithm monotonicity gives Equation (8). For any $\epsilon > 0$, the choice $E_v < \epsilon/(1 + \log(1 + K_v^{\max}))$ ensures $T_v < \epsilon$ across all rank configurations with these counts.

A.4 Effect scale and reference domain

In pooled binary samples, ones and zeros differ in average rank by $N/2$. The positive normalized mean-rank shift is therefore $\alpha[\rho_1 - \rho_0]_+/2$, equal to the error coordinate in Equation (3). Both effect coordinates lie in $[0, \alpha/2]$, giving $0 \le E_v \le 1/\sqrt{2}$. Moreover, $a_o, h_o \in [0, 1]$ and $\alpha_o < 1$ imply $0 \le D_v < 1$ in Equation (4). Newly appearing operations contribute α_o through $a_o = 1$, normal-only operations contribute zero, and empty service aggregates are zero.

A.5 Complete support sets

For a nonempty operation subset H, define its support level as $[\min_{o \in H} J_o - \log(m/|H|)]_+$. Adding all operations whose strengths are at least this minimum leaves the minimum unchanged and reduces the size penalty. An upper-level set therefore attains the maximum over nonempty subsets; its endpoint is the final position of a tie block. The enlargement also cannot decrease the largest effect in the set, so complete threshold sets suffice in Equations (5) and (13).

A.6 The regularized selection solution

On the positive-reference set $\mathcal{V}_+$, normalize the reference odds as $p_v \propto D_v/(1 - D_v)$. For selection weights π, consider

$$\max_{\pi \in \Delta_{\mathcal{V}_+}} \left\{ \sum_v \pi_v E_v L_v - \mathrm{KL}(\pi \| p) \right\}. \tag{11}$$

Write $\mathcal{F}(\pi)$ for the objective in Equation (11) and $Z = \sum_v p_v e^{T_v}$. The standard exponential-tilt identity is $\log Z - \mathcal{F}(\pi) = \mathrm{KL}(\pi \| \pi^*)$, where $\pi^*_v = p_v e^{T_v}/Z$ (Wainwright & Jordan, 2008; Bissiri et al., 2016). Nonnegativity of KL, with equality only at $\pi = \pi^*$, gives the unique maximizer. Substituting the reference odds makes $\log \pi^*_v - R_v$ constant across positive-reference services. Equation (17) is strictly increasing in R_v, so the stable score has the same ordering.

A.7 Allocation and pooling

Let $R^c_v = \mathrm{logit}(D_v) + cL_v$. For positive-reference services, the exact correction is

$$\Delta_{r,b} = (R_r - R_b) - (R^c_r - R^c_b) = (E_r - c)L_r - (E_b - c)L_b. \tag{12}$$

Summing Equation (9) shows that the common coefficient conserves total contribution. If its denominator is zero, every nonnegative L_v is zero. At service v, the change from common to effect-dependent allocation is $(E_v - c)L_v$; taking the difference between two services gives Equation (12).

A.8 Effect and support across operations

For a complete support set $H_k = \{o : J_o \ge J_{(k)}\}$, let $L_k = \log(1 + [J_{(k)} - \log(m_v/k)]_+)$. Requiring the effect to come from that set gives

$$U_v = \max_{k \in \mathcal{K}_v} \left[\max_{o \in H_k} e_o \right] L_k. \tag{13}$$

Proposition 2 (Separate and restricted pooling)**.** *The restricted contribution satisfies* $0 \le U_v \le T_v$. *If* $E_v, L_v > 0$*, equality holds exactly when a support set attaining* L_v *also contains an operation attaining* E_v*. If either value is zero, both contributions are zero.*

To prove Proposition 2, set $q_k = \max_{o \in H_k} e_o$. Every restricted product satisfies $q_k L_k \le E_v L_v$. If both maxima are positive, equality requires one set to attain both $q_k = E_v$ and $L_k = L_v$; finiteness guarantees an attaining set. If either maximum is zero, all products vanish by nonnegativity.

Separate witnesses can arise under one service perturbation. In the model $Q_{1,o}(u) - Q_{0,o}(u) = a_o \theta_v$ with $a_o > 0$, each operation has $C_o = a_o \theta_v > 0$ when $\theta_v > 0$, but responses and observation counts may differ. For example, let two operations have constant normal local duration one, fault durations 1.8 and 1.05, and 5 and 100 observations per window, respectively. These correspond to the shared transformed shift $\theta_v = 1$ with $a_1 = \log(1.4)$ and $a_2 = \log(1.025)$. The first operation has the larger

effect (0.12987 versus 0.01213), whereas the second has stronger rank support (135.0601 versus 4.8363). A shared perturbation thus gives different witnesses for effect and support.

If all operation effects lie in $[e^-, e^+]$, then every nonempty support set has $q_k \geq e^-$. Choosing a set that attains L_v gives $U_v \geq e^- L_v$, while $T_v \leq e^+ L_v$. Subtracting yields $T_v - U_v \leq (e^+ - e^-)L_v$.

B Statistical specification and implementation

B.1 Local observations

For a span with duration Y_i, binary error B_i, and direct children $\mathrm{Ch}(i)$, the local coordinates are

$$\ell_i = \left[Y_i - \sum_{j \in \mathrm{Ch}(i)} Y_j \right]_+, \qquad b_i = \left[B_i - \max_{j \in \mathrm{Ch}(i)} B_j \right]_+, \tag{14}$$

Empty sums and maxima are zero. Under additive execution with nonoverlapping children, the duration residual equals exclusive duration; with overlapping children, Equation (14) defines the residual used by PFMR. Binary errors include recognized nonzero gRPC status codes. Subtracting the maximum child error retains an error not already reported by a direct child.

After unit normalization, we remove nonfinite observations and group the remaining samples by service and canonical operation. The normal and fault windows use $t < \tau$ and $t \geq \tau$, with exposures $T_0 = \max\{1, \tau - \min t\}$ and $T_1 = \max\{1, \max t - \tau\}$ from valid case timestamps. The duration scale s_o is the normal-window median of positive local durations, or the pooled positive median if that is unavailable, followed by one. Operations observed in only one window have zero effect and support.

Service attribution uses the observed service namespace. For a qualified RPC name, a leading slash is removed only when its method matches the method field. An interface changes attribution only when its normalized spelling uniquely identifies a service; otherwise the original attribution remains. The namespace contains services observed in traces and, for RCAEval, identifiers from metric column names.

B.2 The operation support statistic

To form the two-sample contrast, pool the local vectors (ℓ, b) across windows, assign average ranks to ties in each coordinate, and divide ranks by $N + 1$. For the resulting $u_{bi} \in \mathbb{R}^2$, compute

$$\Delta = \bar{u}_1 - \bar{u}_0, \qquad V = \frac{1}{n_0^2} \sum_i (u_{0i} - \bar{u}_0)(u_{0i} - \bar{u}_0)^\top + \frac{1}{n_1^2} \sum_i (u_{1i} - \bar{u}_1)(u_{1i} - \bar{u}_1)^\top. \tag{15}$$

Let $k = \mathrm{rank}(V)$, $k' = \mathrm{rank}([V, \Delta])$, and V^+ be the pseudoinverse. Let $K = [\log \binom{N}{n_0} - \log 2]_+$. The capped strength is

$$J = \begin{cases} K, & k' > k, \\ 0, & k' = k = 0, \\ \min\{K, -\log \Pr(\chi_k^2 \geq Q)\}, & Q = \Delta^\top V^+ \Delta, \quad k' = k > 0. \end{cases} \tag{16}$$

The combination count gives every branch the same upper range at fixed sample counts. A contrast outside the covariance image attains this ceiling. For $k = 1$ and $k = 2$, the uncapped tail strengths are $-\log 2 - \log \Phi(-\sqrt{Q})$ and $Q/2$, respectively. The ceiling is applied before sorting operations and scanning complete tie sets. Numerically, the combination count uses log-gamma functions, Q is clipped at zero, a sample anchor is removed before centering, and rank and pseudoinverse calculations use NumPy. Support measures mean-rank contrasts in either direction; Equation (3) retains only positive movement in the effect.

The rank contrast measures changes in pooled mean ranks, whereas positive transport measures upward duration movement. They enter the service ranking separately: positive movement contributes to the effect and reference, and the covariance-adjusted rank contrast determines the additional support.

B.3 NUMERICAL RANKING

The stable score below preserves the ordering of Equation (1) among positive-reference services:

$$S_v = \begin{cases} -\log(1 - D_v) + \log(1 - D_v + D_v e^{T_v}) = \log(1 + e^{R_v}), & D_v > 0, \\ 0, & D_v = 0. \end{cases} \tag{17}$$

Positive-reference scores are strictly positive and precede zero-reference services. If all references are zero, all scores are zero. Service names break exact ties in ascending order. For $T < 500$, the logarithmic increment is computed with `log1p` and `expm1`; larger values use log-add-exp. Reference values are capped at the representable float immediately below one. The same conventions cover zero effect, zero support, and empty aggregates.

C EVALUATION SETTINGS AND COMPLETE RESULT BREAKDOWNS

C.1 FIXED POPULATIONS AND SCORING

FSE1422 comprises the complete benchmark population. RCAEval240 combines RE2 Online Boutique, RE2 Train Ticket, RE3 Online Boutique, and RE3 Train Ticket, which contribute 37.50%, 37.50%, 12.50%, and 12.50% of its population. The configurations use the same evaluation windows and scoring conventions.

FSE scoring uses its accepted-answer lists and native ranking positions, including repeated services in baseline outputs. RCAEval maps metric identifiers to service prefixes, normalizes database names according to the benchmark, and retains each service's first occurrence. An accepted service absent from the ranking has infinite rank. Errors and misses remain in the denominator for all four metrics.

Trace observations retain their normal/fault assignments in consistent units. The normal window has $t < \tau$ and the fault window has $t \geq \tau$; both use all available observations. PFMR uses trace values and the observed service namespace. On RCAEval, service names from metric headers supplement the observed trace namespace. The configurations are scored after their service rankings are produced.

C.2 BASELINE CONFIGURATIONS

Each method contributes one fixed prediction set. Randomized FSE evaluations use seed zero; on RCAEval, RCD uses seed one and MicroRank, ϵ-Diagnosis, PRISM, and Score Ordering use seed zero. TraceRCA and SparseRCA are deterministic, with ties ordered by service name.

On FSE1422, MicroRank, TraceRCA, and SparseRCA use traces; MicroRCA uses traces and metrics; BARO uses metrics; and BARO-T uses series aggregated from traces. MicroDig uses trace observations, while SimpleRCA uses logs, metrics, and traces. Workload generators are excluded from the diagnostic candidate sets of TraceRCA and SparseRCA, although their observed calls remain available as evidence.

TraceRCA uses cross-service invocation latency and response status over the complete normal and fault windows. For a feature, a_0 and a_1 are its mean absolute standardized deviations in the two windows. A selected feature satisfies $a_1 - a_0 > 3a_0 + 1$; an invocation is abnormal when that feature's deviation exceeds one normal-period standard deviation. Candidate service sets cover at least 10% of abnormal traces, and the 100 highest distinct Jaccard-score levels are retained. A service receives the maximum, over sets, of Jaccard score times the absolute difference between its incoming and outgoing abnormal-trace counts.

SparseRCA uses complete normal and fault trace windows. Exclusive duration subtracts the union of direct-child intervals after clipping them to the parent. Gaussian latency models condition on normal-window service context and child-call counts; means are estimated by least squares and component variances by expectation maximization, with relative tolerance 10^{-8} and a 200-iteration limit. Unseen patterns receive reciprocal edit-distance weights after excluding the most dissimilar 5%, retaining all distance ties. Latencies are measured in microseconds and score smoothing is 10^{-8}. Three directed topology updates use coefficient 0.9; the service score is its maximum refined span score in the fault window.

Table 4: Five configurations on all four RCAEval subsets (%). Bold marks the best value within a subset and metric.

Subset	Configuration	Top-1	Top-3	Top-5	Avg@5
RE2-OB	PFMR	**87.78**	**98.89**	**100.00**	**95.78**
	Change only	82.22	97.78	**100.00**	94.22
	Support only	71.11	86.67	93.33	83.78
	$E = 1$	84.44	93.33	98.89	92.67
	Common E	84.44	**98.89**	**100.00**	95.33
RE2-TT	PFMR	**81.11**	**94.44**	97.78	**93.11**
	Change only	64.44	87.78	95.56	85.56
	Support only	37.78	68.89	77.78	62.67
	$E = 1$	63.33	92.22	**100.00**	87.33
	Common E	65.56	91.11	96.67	87.33
RE3-OB	PFMR	**100.00**	**100.00**	**100.00**	**100.00**
	Change only	90.00	93.33	93.33	92.67
	Support only	90.00	**100.00**	**100.00**	98.00
	$E = 1$	**100.00**	**100.00**	**100.00**	**100.00**
	Common E	90.00	96.67	96.67	95.33
RE3-TT	PFMR	**96.67**	**100.00**	**100.00**	**99.33**
	Change only	**96.67**	**100.00**	**100.00**	**99.33**
	Support only	46.67	66.67	66.67	60.67
	$E = 1$	86.67	**100.00**	**100.00**	97.33
	Common E	**96.67**	**100.00**	**100.00**	**99.33**

On RCAEval, BARO uses robust calibration from the normal period, RCD combines clustering-based discretization with conditional independence tests, and MicroRank combines service-call graph weights with trace spectra. The ϵ-Diagnosis comparison applies a squared correlation statistic to normal and fault metric observations, with thresholds bootstrapped from normal observations.

PRISM adds scores for internal and external properties after z-score calibration and maximum property pooling. Its metric inputs include latency percentiles, and exact score ties follow service-name order. Score Ordering combines IT anomaly scores with rescaled median CDF quantile scores; it calibrates on all normal observations and scores the first post-boundary observation.

C.3 Five configurations and four RCAEval subsets

The five main configurations share operation summaries and service candidates and differ in their ranking rules. Change only uses $-\log(1 - D_v)$, support only uses $\log(1 + I_v)$, $E = 1$ assigns unit coefficient to log-support, and common E uses $c = \sum_v E_v L_v / \sum_v L_v$, with value zero for a zero denominator. Common E preserves the total additive potential over all candidates. In both complete populations, it assigns zero potential to zero-reference services.

Table 4 reports all metrics for the four RCAEval subsets. PFMR's Top-1 accuracies are 87.78%, 81.11%, 100.00%, and 96.67%; the corresponding differences from common E are 3.33, 15.56, 10.00, and 0.00 percentage points. Table 5 gives paired Top-1 comparisons with the four controls, and Table 6 gives them for the complete baselines.

Paired intervals use 20,000 bootstrap samples of evaluation indices. PFMR and its comparator share each sampled index, and metric differences are recomputed for every sample. The reported interval spans the 2.5th to 97.5th percentiles. The resampling unit is an evaluation case in the indicated benchmark population.

Continuous magnitude and positive-effect eligibility. To distinguish continuous effect magnitude from positive-effect eligibility, this control assigns a common coefficient within the positive-effect

Table 5: Paired Top-1 outcomes against each core control. PFMR-only/control-only entries are percentages of the complete dataset for which only that side is correct. Differences and 95% paired intervals are in percentage points.

Dataset	Control	PFMR only	Control only	Difference	95% interval
FSE1422	Change only	20.68	3.73	+16.95	[14.56, 19.34]
FSE1422	Support only	40.93	3.80	+37.13	[34.25, 40.01]
FSE1422	$E = 1$	7.17	5.34	+1.83	[0.00, 3.66]
FSE1422	Common E	9.56	3.87	+5.70	[3.80, 7.59]
RCAEval240	Change only	10.00	0.42	+9.58	[5.83, 13.75]
RCAEval240	Support only	33.33	3.33	+30.00	[23.33, 36.67]
RCAEval240	$E = 1$	10.00	0.83	+9.17	[5.42, 13.33]
RCAEval240	Common E	8.33	0.00	+8.33	[5.00, 12.08]

Table 6: Paired Top-1 outcomes against the baselines in Table 1. Percentages use the complete-dataset denominator of Table 5; differences and 95% intervals are in percentage points.

Dataset	Baseline	PFMR only	Baseline only	Difference	95% interval
FSE1422	BARO	52.11	11.74	+40.37	[36.71, 43.95]
FSE1422	BARO-T	60.76	3.87	+56.89	[53.87, 59.85]
FSE1422	MicroDig	42.90	6.61	+36.29	[33.19, 39.45]
FSE1422	MicroRCA	41.84	7.24	+34.60	[31.36, 37.76]
FSE1422	MicroRank	73.63	1.05	+72.57	[70.11, 75.04]
FSE1422	SimpleRCA	53.45	7.52	+45.92	[42.69, 49.23]
FSE1422	TraceRCA	38.75	3.16	+35.58	[32.77, 38.40]
FSE1422	SparseRCA	60.69	3.16	+57.52	[54.57, 60.41]
RCAEval240	BARO	57.50	5.00	+52.50	[45.00, 60.00]
RCAEval240	RCD	57.08	4.17	+52.92	[45.42, 60.00]
RCAEval240	ϵ-Diagnosis	77.92	0.00	+77.92	[72.50, 82.92]
RCAEval240	MicroRank	77.08	0.42	+76.67	[71.25, 82.08]
RCAEval240	PRISM	22.08	8.75	+13.33	[6.25, 20.00]
RCAEval240	Score Ordering	83.33	0.83	+82.50	[77.08, 87.50]
RCAEval240	TraceRCA	62.50	3.75	+58.75	[51.25, 65.83]
RCAEval240	SparseRCA	54.58	5.42	+49.17	[41.67, 56.67]

set while preserving total additive potential. With $g_v = \mathbf{1}[E_v > 0]$, define

$$c_+ = \frac{\sum_v E_v L_v}{\sum_v g_v L_v}, \qquad T_v^{\text{gate}} = c_+ g_v L_v, \tag{18}$$

Set $c_+ = 0$ when the denominator vanishes. The reference, support, and stable scoring rule are unchanged. This control is evaluated alongside the other ranking configurations. Its Top-1/3/5 and Avg@5 match common E: 72.43/88.40/92.62/85.60% on FSE1422 and 79.58/95.83/98.33/92.83% on RCAEval240. PFMR-only/control-only outcomes account for 9.56/3.87% and 8.33/0.00% of the corresponding full populations. The paired Top-1 intervals are $[3.80, 7.59]$ and $[5.00, 12.08]$ percentage points. Aggregate metrics coincide, while the native rankings differ from common E on 1.20% and 1.67% of the two populations.

C.4 Direct additive fusion

To compare effect-dependent allocation with a direct fusion of continuous effect and support, let $B = \sum_v E_v L_v$, $A = \sum_v E_v$, $g_v = \mathbf{1}[E_v > 0]$, and $C = \sum_v g_v L_v$. The normalized additive rule is

$$T_v^{\text{add}} = \frac{B}{2}\left(\frac{E_v}{A} + \frac{g_v L_v}{C}\right). \tag{19}$$

Table 7: Multiplicative allocation and normalized additive fusion on identical service statistics (%). Both retain positive-effect eligibility and the same total contribution. Subset rows partition RCAEval240; bold marks the better value in each comparison.

Population	Allocation	Top-1	Top-3	Top-5	Avg@5
FSE1422	PFMR	**78.13**	**91.49**	**94.37**	**89.13**
	Normalized additive	74.75	89.73	93.11	86.78
RCAEval240	PFMR	**87.92**	**97.50**	**99.17**	**95.75**
	Normalized additive	84.58	96.67	98.75	94.42
RE2-OB	PFMR	**87.78**	**98.89**	**100.00**	**95.78**
	Normalized additive	86.67	**98.89**	**100.00**	95.56
RE2-TT	PFMR	**81.11**	**94.44**	**97.78**	**93.11**
	Normalized additive	73.33	92.22	96.67	89.78
RE3-OB	PFMR	**100.00**	**100.00**	**100.00**	**100.00**
	Normalized additive	**100.00**	**100.00**	**100.00**	**100.00**
RE3-TT	PFMR	**96.67**	**100.00**	**100.00**	**99.33**
	Normalized additive	**96.67**	**100.00**	**100.00**	**99.33**

Table 8: Top-1 accuracy (%) by prediction-defined disagreement between change-only and support-only first services. Share is the percentage of the complete dataset; each accuracy uses the full group denominator.

Dataset	Group	Share (%)	PFMR	Change	Support	$E = 1$	Common E
FSE1422	Conflict	70.68	72.64	47.96	19.40	69.35	63.88
FSE1422	Agreement	29.32	91.37	93.05	93.05	93.05	93.05
RCAEval240	Conflict	50.00	78.33	59.17	18.33	60.00	61.67
RCAEval240	Agreement	50.00	97.50	97.50	97.50	97.50	97.50

When $B = 0$, all contributions are zero. Otherwise $A, C > 0$, and the effect and support terms each distribute $B/2$. Consequently, $\sum_v T_v^{\text{add}} = B$ and zero-effect services receive no contribution. Normalizing each input by its sum gives the terms equal aggregate weight despite their different scales. The reference, observations, candidates, and stable scoring rule match PFMR's. A positive-effect service with zero support can receive an effect-only contribution under addition. The control is evaluated with the other configurations; labels are used later for scoring.

Table 7 reports all four metrics and the RCAEval subsets. PFMR-only/additive-only outcomes account for 5.98/2.60% of FSE1422 and 3.75/0.42% of RCAEval240. Top-1 differences are 3.38/3.33 percentage points, with intervals $[1.83, 4.92]$/$[0.83, 5.83]$; Avg@5 differences are 2.35/1.33 points, with intervals $[1.72, 3.02]$/$[0.42, 2.33]$. Among PFMR-only outcomes, its first service has the larger effect in 90.59%/88.89%; the corresponding proportions among additive-only outcomes are 97.30%/0.00%. The descriptors identify changed choices, and paired correctness identifies their diagnostic value.

C.5 Prediction groups and allocation descriptors

Table 8 reports all five configurations in the conflict and agreement groups. Membership depends only on the mapped first services under change-only and support-only ranking; correctness is evaluated afterward. Figure 3 further partitions conflicts by the effects of those two services.

Table 9 describes first-service changes through their effect and support relationships. Relative to $E = 1$, PFMR-only correct outcomes often favor larger effect despite lower support. Relative to common E, the descriptor also captures both directions of reallocation. The tables decompose the same evaluation results.

Table 9: Changed first-service predictions in which PFMR selects larger effect and lower support. Each entry is the percentage of matching pairs within its outcome group; a dash denotes an empty group. Same correctness includes both-correct and both-wrong cases.

Dataset	Control	PFMR only	Control only	Same correctness
FSE1422	$E = 1$	91.18	50.00	81.11
FSE1422	Common E	25.74	41.82	40.00
RCAEval240	$E = 1$	91.67	100.00	100.00
RCAEval240	Common E	20.00	–	–

Table 10: Effect relationships and allocation gains. All accuracies are within-group percentages; differences and paired 95% intervals are percentage points. The reference and support choices are a and b. Groups are fixed from predictions before outcome scoring.

Dataset	Group	Method	PFMR	Baseline	Gain	95% CI
FSE1422	Agreement	Common E	91.37	93.05	-1.68	[-3.36, 0.00]
		Additive	91.37	93.29	-1.92	[-3.36, -0.48]
FSE1422	Conflict: $E_b < E_a$	Common E	73.40	70.40	+3.00	[0.41, 5.59]
		Additive	73.40	72.03	+1.36	[-0.55, 3.27]
FSE1422	Conflict: $E_b \geq E_a$	Common E	70.59	46.32	+24.26	[18.38, 30.51]
		Additive	70.59	53.68	+16.91	[11.76, 22.06]
RCAEval240	Agreement	Common E	97.50	97.50	+0.00	[0.00, 0.00]
		Additive	97.50	97.50	+0.00	[0.00, 0.00]
RCAEval240	Conflict: $E_b < E_a$	Common E	81.72	75.27	+6.45	[2.15, 11.83]
		Additive	81.72	78.49	+3.23	[-1.08, 8.60]
RCAEval240	Conflict: $E_b \geq E_a$	Common E	66.67	14.81	+51.85	[33.33, 70.37]
		Additive	66.67	48.15	+18.52	[3.70, 33.33]

C.6 Effect relationships in prediction-defined conflicts

Let a and b be the first services under the native change-only and support-only rankings. After applying the evaluation service mapping, each prediction is assigned to agreement, conflict with $E_b < E_a$, or conflict with $E_b \geq E_a$. Exact effect ties enter the latter group; zero effects follow the same rule. Native ranking ties retain service-name order. Group membership is fixed from inference outputs before outcome labels are joined for scoring.

The resulting groups relate allocation gains to the effect ordering of the two preferred services. Their membership follows the reference and support first choices, and all three groups are evaluated on both populations against common E and normalized additive fusion. Table 10 reports group-level accuracy and paired differences. Each interval uses 20,000 within-group paired resamples, with independent random streams derived from seed 20260920 by dataset, group, and comparator.

We compare the gains in the two disjoint conflict groups by subtracting the $E_b \geq E_a$ gain from the $E_b < E_a$ gain. For common E, the contrasts are -21.26/-45.40 percentage points on FSE1422/RCAEval240, with intervals $[-27.95, -14.81]$/$[-64.99, -25.81]$. For additive fusion, they are -15.55/-15.29, with intervals $[-21.10, -10.13]$/$[-31.66, -0.48]$. Each contrast resamples the disjoint groups independently. Both comparisons place the larger advantage in conflicts where the support choice has at least as much local effect.

C.7 The complete mechanism comparison

Table 11 reports all four metrics for the 2×2 reference-by-pooling comparison. The all-contributions reference is $\max_o d_o$, and the current-observation reference is Equation (4). Pooling uses either T_v or the restricted U_v in Equation (13), both based on complete threshold sets. Operation statistics are

Table 11: All four metrics (%) for the same 2×2 configurations as Table 3. All/current specifies the reference; service/restricted specifies pooling.

	FSE1422				RCAEval240			
Configuration	T1	T3	T5	A5	T1	T3	T5	A5
All/restricted	75.81	90.79	94.44	88.19	85.83	97.08	98.75	95.00
All/service	75.95	90.72	94.87	88.41	86.67	97.08	98.75	95.17
Current/restricted	77.43	92.05	94.37	88.93	87.08	97.50	99.17	95.58
Current/service	78.13	91.49	94.37	89.13	87.92	97.50	99.17	95.75

Table 12: Complete FSE1422 fault-type strata: within-stratum Top-1 accuracy (%). Share is the percentage of the complete dataset. Errors and misses remain in the denominator.

Fault type	Share (%)	PFMR	Change	Support	$E = 1$	Common E
ContainerKill	6.26	61.80	60.67	29.21	66.29	62.92
DNSRandom	0.07	0.00	0.00	0.00	0.00	0.00
HTTPRequestAbort	4.22	73.33	70.00	58.33	83.33	76.67
HTTPRequestDelay	6.19	88.64	36.36	68.18	89.77	72.73
HTTPRequestReplaceMethod	13.36	86.32	60.00	43.16	82.11	75.79
HTTPRequestReplacePath	2.74	82.05	58.97	51.28	87.18	76.92
HTTPResponseAbort	3.09	81.82	79.55	43.18	88.64	90.91
HTTPResponseDelay	6.26	84.27	28.09	68.54	79.78	64.04
HTTPResponsePatchBody	0.28	50.00	0.00	50.00	50.00	50.00
HTTPResponseReplaceBody	3.59	92.16	88.24	47.06	94.12	96.08
HTTPResponseReplaceCode	16.24	93.51	81.39	48.48	92.64	89.61
JVMCPUStress	0.14	0.00	0.00	0.00	0.00	0.00
JVMException	3.02	95.35	88.37	44.19	83.72	88.37
JVMLatency	0.49	85.71	57.14	57.14	71.43	57.14
JVMMemoryStress	12.03	58.48	54.97	21.05	61.40	58.48
JVMMySQLLatency	0.14	0.00	0.00	0.00	0.00	0.00
JVMReturn	1.48	57.14	57.14	33.33	57.14	57.14
NetworkBandwidth	2.95	78.57	69.05	26.19	61.90	71.43
NetworkCorrupt	3.23	78.26	60.87	13.04	50.00	60.87
NetworkDelay	1.48	100.00	71.43	85.71	100.00	90.48
NetworkLoss	3.38	75.00	50.00	33.33	72.92	62.50
NetworkPartition	6.82	72.16	61.86	25.77	63.92	67.01
PodFailure	1.69	4.17	8.33	0.00	4.17	12.50
PodKill	0.70	50.00	50.00	0.00	60.00	50.00
TimeSkew	0.14	50.00	50.00	0.00	50.00	50.00

shared across all four configurations, allowing the pooling condition in Proposition 2 to be examined under each reference.

C.8 Fault-type strata

Tables 12 and 13 report every fault type in the two populations. Top-1 values use each stratum as the denominator; the share column uses the full dataset. Every row contains the five main configurations. The f1–f5 labels are code-level fault identifiers from the data release, used for reporting after inference.

Table 13: Complete RCAEval240 fault-type strata: within-stratum Top-1 accuracy (%). Share is the percentage of the complete dataset. Errors and misses remain in the denominator.

Fault type	Share (%)	PFMR	Change	Support	$E = 1$	Common E
cpu	12.50	93.33	90.00	56.67	83.33	90.00
delay	12.50	86.67	73.33	80.00	83.33	83.33
disk	12.50	76.67	66.67	43.33	63.33	70.00
f1	6.67	93.75	75.00	62.50	93.75	75.00
f2	4.17	100.00	100.00	60.00	100.00	100.00
f3	6.67	100.00	100.00	62.50	93.75	100.00
f4	5.00	100.00	100.00	75.00	83.33	100.00
f5	2.50	100.00	100.00	100.00	100.00	100.00
loss	12.50	86.67	60.00	46.67	66.67	60.00
mem	12.50	86.67	76.67	43.33	73.33	76.67
socket	12.50	76.67	73.33	56.67	73.33	70.00