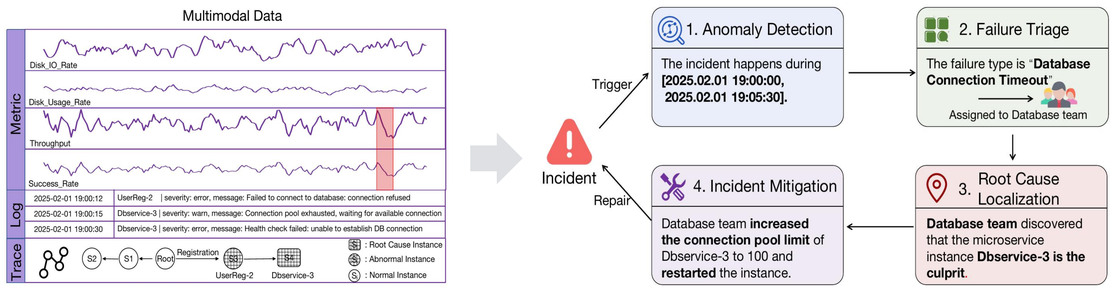
TrioXpert: An Automated Incident Management Framework for Microservice System

Yongqian Sun[1], Yu Luo[1], Xidao Wen[2], Yuan Yuan[3], Xiaohui Nie[4], Shenglin Zhang[1], Tong Liu[5], Xi Luo[5]

[1] Nankai University, [2] BizSeer, [3] National University of Defense Technology,
[4] Computer Network Information Center, Chinese Academy of Sciences, [5] Lenovo

INTRODUCTION



Objectives:

1. Enable end-to-end incident management for microservice systems across AD, FT, and RCL—focusing on timely, accurate diagnosis at production scale;
2. Unify multimodal observability (metrics, logs, traces) into a coherent workflow instead of siloed subtasks;
3. Deliver transparent, evidence-linked reasoning so engineers can verify conclusions and trust automation;

Motivating Study:

1. Metrics, logs, and traces provide complementary diagnostic signals, but logs/traces are highly redundant and require effective filtering;
2. A single LLM is unreliable and hard to interpret for complex incidents; collaborative / structured reasoning is needed;

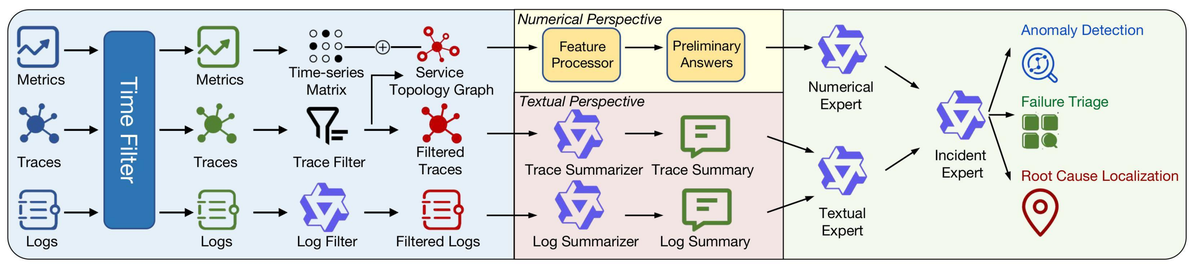
Challenges:

1. Semantic impoverishment in multimodal fusion;
2. Textual data overload in real-time incident management;
3. LLM limitations in complex and trust-critical incident management;

Contributions:

1. We propose ***TrioXpert***, the first end-to-end framework unifying AD, FT, and RCL via LLM-based collaborative reasoning, integrating metrics, logs, and traces through structured prompts for scalable, interpretable diagnostics.
2. We introduce modality-specific preprocessing and filtering to align heterogeneous observability and build a multi-view system state (numerical + textual), improving accuracy and interpretability.
3. We validate on two real microservice systems and a *Lenovo* production deployment, showing consistent gains, and release code, prompts, configurations, and data for reproducibility.

METHODOLOGY



(a). Multimodal Data Preprocessing *(b). Multi-Dimensional System Status Representation* *(c). LLMs Collaborative Reasoning*

Algorithm 1: Coordination Pipeline in *TrioXpert*

Data: $\mathcal{F}_{num}$: Numerical feature (from metrics and topology)
$\mathcal{F}_{txt}$: Textual feature (from logs and traces)
Result: $\mathcal{R}_{final}$: Final diagnostic decision (AD, FT, RCL)
$\mathcal{E}_{final}$: Final explanation reasoning chain
$\mathcal{R}_{num}, \mathcal{E}_{num} \leftarrow$ **NumericalExpert**$(\mathcal{F}_{num})$;
// Process metric-based features
$\mathcal{R}_{txt}, \mathcal{E}_{txt} \leftarrow$ **TextualExpert**$(\mathcal{F}_{txt})$; // Process logs and traces
$\mathcal{R}_{final}, \mathcal{E}_{final} \leftarrow$ **IncidentExpert**$(\mathcal{R}_{num}, \mathcal{E}_{num}, \mathcal{R}_{txt}, \mathcal{E}_{txt})$;
// Aggregate and reconcile
return $(\mathcal{R}_{final}, \mathcal{E}_{final})$

1. Multimodal Data Preprocessing: Metrics → time-series matrix M; traces → service-topology graph G with type-aware high-latency filtering to form filtered traces T; logs → two-stage LLM filtering to obtain incident-relevant logs L. All (M, G, T, L) are time-aligned for downstream use.

2. Multi-Dimensional System Status Representation: Numerical pipeline predicts next-step metrics and forms a deviation matrix, from which preliminary signals for AD/FT/RCL are derived. Textual pipeline summarizes T and L into compact incident cues. Together they provide a coherent multi-view system state.

3. ***LLMs Collaborative Reasoning:*** A Numerical Expert and a Textual Expert generate task-wise results and evidence; an Incident Expert reconciles them to finalize AD/FT/RCL with evidence-linked explanations. Conflicts are resolved by prioritizing numerical signals overall and logs over traces.

EXPERIMENTAL RESULTS

TABLE II
PERFORMANCE COMPARISON ON AD, FT, RCL, AND TIME. "-" MEANS THIS METHOD DOES NOT COVER THE TASK.

Methods	$\mathscr{D}1$										$\mathscr{D}2$									
	AD			FT			RCL			Efficiency	AD			FT			RCL			Efficiency
	Precision	Recall	F1	Precision	Recall	F1	Top@1	Top@3	Avg@5	Time (s)	Precision	Recall	F1	Precision	Recall	F1	Top@1	Top@3	Avg@5	Time (s)
TrioXpert	0.880	0.972	**0.924**	0.852	0.768	**0.807**	0.651	0.778	**0.773**	14.314	0.854	0.972	**0.909**	0.814	0.725	**0.767**	0.550	0.775	**0.750**	12.597
ART [1]	0.759	0.621	0.683	0.786	0.794	0.790	0.683	0.762	0.757	0.872	0.593	0.972	0.737	0.860	0.650	0.740	0.375	0.825	0.738	1.363
DiagFusion [2]	-	-	-	0.675	0.500	0.574	0.310	0.452	0.467	4.145	-	-	-	0.797	0.527	0.634	0.582	0.709	0.695	3.297
Eadro [4]	0.425	0.946	0.586	-	-	-	0.137	0.315	0.302	0.627	0.767	0.935	0.842	-	-	-	0.157	0.315	0.310	0.899
Hades [29]	0.866	0.863	0.865	-	-	-	-	-	-	0.104	0.867	0.868	0.868	-	-	-	-	-	-	0.415
MicroCBR [11]	-	-	-	0.667	0.796	0.726	-	-	-	0.278	-	-	-	0.629	0.678	0.653	-	-	-	0.306
PDiagnose [30]	-	-	-	-	-	-	0.615	0.692	0.685	4.342	-	-	-	-	-	-	0.037	0.296	0.285	9.919

TABLE III
THE EVALUATION RESULTS OF ABLATION STUDY.

Methods	$\mathscr{D}1$			$\mathscr{D}2$		
	AD: F1	FT: F1	RCL: Avg@5	AD: F1	FT: F1	RCL: Avg@5
TrioXpert	**0.924**	**0.807**	**0.773**	**0.909**	**0.767**	**0.750**
$\mathcal{A}1$	0.725	0.190	0.667	0.832	0.685	0.625
$\mathcal{A}2$	nan	0.261	0.238	nan	0.352	0.275
$\mathcal{A}3$	0.672	0.398	0.534	0.583	0.284	0.608
$\mathcal{A}4$	0.428	0.294	0.397	0.552	0.359	0.517
$\mathcal{A}5$	0.339	0.157	0.362	0.405	0.287	0.233

Overall performance. On both D1 and D2, TrioXpert surpasses six strong baselines across all three tasks—AD, FT, and RCL—with relative gains of 4.7%–57.7% (AD), 2.1%–40.6% (FT), and 1.6%–163.1% (RCL). In absolute terms, it achieves F1 > 0.9 for AD, F1 > 0.75 for FT, and Avg@5 > 0.75 for RCL. Despite LLM-based reasoning, the end-to-end processing time is < 15 s per case, satisfying typical production latency constraints. These results indicate that TrioXpert's modality-aware pipelines (metrics/logs/traces) and collaborative reasoning not only improve accuracy but also surface evidence-linked, interpretable diagnostics; notably, leveraging textual semantics in logs and traces provides complementary gains that prior methods underuse.

Ablation insights. Five controlled variants show that every major component is necessary. Removing textual pipelines (A1) or numerical pipeline (A2) degrades performance markedly; in A2, AD becomes undefined because textual data alone lack timestamp granularity—underscoring the role of metrics for temporal localization. Replacing multi-expert collaboration with a single LLM (A3) yields consistent drops under multimodal load. Disabling conflict resolution/aggregation (A4) or hallucination-mitigation prompts (A5) further harms robustness. Collectively, these studies validate the design: multimodal coverage + collaborative reasoning + structured prompts are all critical to TrioXpert's accuracy, stability, and interpretability.

Lenovo

Deployment at Lenovo: Deployed as a real-time diagnostic assistant on Lenovo's production microservice platform, the system markedly improves efficiency and precision: whereas the manual process typically involves **3** OCEs working **~2.5 hours** and iterating **~5** hypotheses, the system issues an initial diagnosis in **~26 seconds** and usually identifies the true root cause within **2** attempts. We present **3** representative incidents—disk-space exhaustion, goroutine leak, and proxy misconfiguration—and **2** Lenovo OCEs blind-reviewed the reasoning chains, confirming that the diagnoses were transparent, verifiable, and practically useful.

CONCLUSIONS

This work introduces ***TrioXpert***, an end-to-end framework for unified incident management in microservice systems through modality-specific preprocessing and collaborative LLM reasoning. By preserving semantic richness while addressing LLM limitations via structured prompts, it simultaneously achieves significant accuracy improvements across AD, FT, and RCL.The approach shows that modality-aware representations and structured LLM collaboration enable scalable, interpretable incident management.While designed for microservices, its principles can generalize to other complex, high-stakes systems requiring multisource analysis.

ACKNOWLEDGEMENT

This work is supported by the Advanced Research Project of China (31511010501), the National Natural Science Foundation of China (62272249, 62302244), the Fundamental Research Funds for the Central Universities (XXX-63253249), and the CCF-Lenovo Blue Ocean Research Fund.

Paper **Repo** **Profile**