

# Simplifying Root Cause Analysis in Kubernetes with StateGraph and LLM

Yong Xiang  
Tsinghua University  
Beijing, China

xiangyong13@mails.tsinghua.edu.cn

Charley Peter Chen  
Harmonic Inc  
Vancouver, Canada

charley.chen@harmonicinc.com

Liyi Zeng  
Peng Cheng Laboratory  
ShenZhen, China

zengly@pcl.ac.cn

Wei Yin  
Tsinghua University  
Beijing, China

yw@mail.tsinghua.edu.cn

Xin Liu  
Tsinghua University  
Beijing, China

liuxin19@mails.tsinghua.edu.cn

Hu Li  
Unaffiliated  
Beijing, China

lihu723@gmail.com

Wei Xu  
Tsinghua University  
Beijing, China

weixu@tsinghua.edu.cn

**Abstract**—Kubernetes, a notably complex and distributed system, utilizes an array of controllers to uphold cluster management logic through state reconciliation. Nevertheless, maintaining state consistency presents significant challenges due to unexpected failures, network disruptions, and asynchronous issues, especially within dynamic cloud environments. These challenges result in operational disruptions and economic losses, underscoring the necessity for robust root cause analysis (RCA) to enhance Kubernetes reliability. The development of large language models (LLMs) presents a promising direction for RCA. However, existing methodologies encounter several obstacles, including the diverse and evolving nature of Kubernetes incidents, the intricate context of incidents, and the polymorphic nature of these incidents. In this paper, we introduce SynergyRCA, an innovative tool that leverages LLMs with retrieval augmentation from graph databases and enhancement with expert prompts. SynergyRCA constructs a StateGraph to capture spatial and temporal relationships and utilizes a MetaGraph to outline entity connections. Upon the occurrence of an incident, an LLM predicts the most pertinent resource, and SynergyRCA queries the MetaGraph and StateGraph to deliver context-specific insights for RCA. We evaluate SynergyRCA using datasets from two production Kubernetes clusters, highlighting its capacity to identify numerous root causes, including novel ones, with high efficiency and precision. SynergyRCA demonstrates the ability to identify root causes in an average time of about two minutes and achieves an impressive precision of approximately 0.90.

**Index Terms**—Root Cause Analysis, Large Language Models, Incident Management, Retrieval-Augmented Generation, Graph, GPT-4o

## I. INTRODUCTION

Cluster orchestration platforms such as Kubernetes [1], Borg [2], ECS [3], and Twine [4] are widely utilized for managing an array of systems across diverse domains and applications [5]. Kubernetes, in particular, is a complex and distributed system made up of numerous components spread across multiple layers. It operates through a set of *controllers* [6] that implement cluster management logic based on the *state reconciliation principle* [6], [7]. These controllers continuously monitor the cluster state, striving to align the *current state* with the *desired state*. However, maintaining

consistency across these states is challenging due to unexpected failures [8], [9], network interruptions [10], and asynchrony issues [11], particularly when operating within complex, dynamic, and distributed cloud environments. Numerous documented failure incidents have resulted in significant operational disruptions and economic losses [12]–[17].

Root cause analysis (RCA) is crucial for the swift recovery of service availability and the enhancement of Kubernetes reliability. The rapid development and successful application [18]–[21] of LLMs present a promising avenue for RCA improvement. Ahamed et al. [22] conducted the first large-scale empirical study evaluating LLMs’ effectiveness in RCA, while Zhang et al. [23] demonstrated that using in-context examples can improve root cause recommendation quality. Additionally, Chen et al. [24] suggested employing in-context examples with a retrieval-augmented generation (RAG) framework for root cause categorization.

However, applying AI-based RCA in Kubernetes introduces several challenges. Firstly, the diverse array of Kubernetes incidents, varying in versions and configurations, continually evolves, complicating the RCA process. The fine-tuning method described by Ahamed et al. [22] is costly and challenging to update continuously, risking hallucinations [24]. Rule-based AI-enhancements like K8sGPT [25], which encodes Site Reliability Engineer (SRE) experience into analyzers, are difficult to maintain and require substantial expert labor for version updates.

Secondly, interpreting the wide range of incident contexts further complicates RCA. Chen et al. [24] utilized RAG to supply LLMs with incident-related data specific to a private service, and Zhang et al. [23] found that 20-shot in-context learning could achieve best performance. In our research, we bypass the Kubernetes incident handler and avoid directly importing examples into prompts, which restricts the LLM’s performance with limited examples. Instead, we propose structuring domain-specific runtime data in a graph database for dynamic querying to enhance LLM effectiveness.

Thirdly, the polymorphic nature of incident message se-

mantics poses an additional challenge for LLM analysis. For instance, in the error message “*Error: cannot find volume ‘foobar1’ to mount into container ‘foobar2’*,” the volume could be a ConfigMap, Secret or PVC. Drishti et al. [26] demonstrated that augmenting the LLM prompt with data from various stages of the software development lifecycle (SDLC) can improve RCA. Building upon this approach, we propose incorporating unstructured knowledge into the LLM through expert prompts.

In this paper, we introduce *SynergyRCA*, an innovative tool that utilizes off-the-shelf LLMs enhanced with retrieval augmentation from a graph database that captures the runtime states of Kubernetes clusters. This tool is further refined with expert prompts. SynergyRCA is widely applicable across various Kubernetes cluster instances and capable of analyzing a wide range of failure messages within incidents. We automatically construct a *StateGraph* to capture spatial and temporal relationships of entities within a Kubernetes instance and employ a *MetaGraph* to map the interconnections among various entity kinds. These graphs not only facilitate data management for LLM input but also act as a protective barrier, preventing direct LLM operations on the production cluster.

Upon the occurrence of a new incident, an LLM module, improved with expert prompts, predicts the most pertinent resource for examination. SynergyRCA then queries the Meta-Graph to uncover dependencies among resource entities and formulates a graph query to identify state dependencies on related entities within the StateGraph. This information is fed into the LLM, which subsequently generates a root cause report and suggests remediation commands.

SynergyRCA employs the essential principle of state reconciliation to identify root causes by comparing states through three key steps: (1) verifying the existence of the current state, such as checking for a directory’s presence on an NFS server; (2) ensuring the correctness of these states, such as confirming the resource quota is not exhausted; and (3) identifying discrepancies among states, such as verifying if a Pod and PVC exhibit a matching “Bound” status.

We implement SynergyRCA using GPT-4o and the Neo4j graph database [27]. We assess its performance with datasets from two production Kubernetes clusters of different versions, covering periods of one week and six months, respectively. SynergyRCA successfully identifies 18 and 20 types of root causes in these datasets, notably discovering five new types in the second dataset, underscoring its effectiveness. SynergyRCA operates efficiently, pinpointing root causes within an average of two minutes, and accurately identifies the root cause in 90% of cases.

This paper presents a comprehensive approach to LLM-based RCA with three key contributions:

- *Model RCA with Graph-based RAG*: We are pioneers in integrating graph databases with LLMs to generate RCA reports, advancing the accuracy and contextual understanding in RCA significantly.
- *End-to-End LLM-based RCA System*: We introduce SynergyRCA, an innovative end-to-end solution that auto-

matically builds graphs and leverages LLMs to precisely identify root causes and generate insightful RCA reports.

- *Real-World Evaluation on Kubernetes Clusters*: We demonstrate the practical utility and dependability of SynergyRCA in two real-world Kubernetes cluster scenarios.

The rest of the paper is organized as follows: Section II reviews related work, Section III details the design of SynergyRCA, Section IV discusses evaluation experiments and results, Section V addresses limitations and potential future work, and finally, Section VI concludes the paper.

## II. RELATED WORK

### A. Incident Management with AI

Traditional research in incident lifecycle automation has extensively employed machine learning (ML) techniques to enhance processes such as triaging [28]–[30], incident linking [31], [32], diagnosis [33]–[35], and mitigation [36]. Regarding RCA tasks, some studies have concentrated on identifying the problem’s root cause [37]–[40], while others have focused on determining the components involved [41]–[44]. Marco et al. [45] investigated real-world Kubernetes failures and developed a framework for executing fault injection campaigns to replicate these patterns. CausalIoT [46] introduced an anomaly detection system using device interaction graphs to identify abnormal states in IoT environments. RAPMiner [47] provided an anomaly localization mechanism for CDN systems, employing classification power-based redundant attribute deletion to eliminate non-root cause attributes. Despite their utility, these methods often involve intricate feature extraction processes and lack cross-platform applicability and task flexibility [48]. Some approaches, such as topology graph-based analyses [49]–[51], reconstruct graphs to model application topology and pinpoint potential anomaly root causes. Whereas these methods primarily focus on system components, our approach with StateGraph offers finer granularity by concentrating on entities and their interrelations.

### B. LLMs in Software System Reliability

Large Language Models (LLMs) are revolutionizing the landscape of software systems by offering innovative solutions for various tasks, including code generation [52], [53], code repair [20], [54], vulnerability repair [55], and bug fixing [21]. Our research focuses on enhancing software system reliability by using LLMs to automate and improve the accuracy of the RCA process.

Ahamed et al. [22] conducted the first large-scale study assessing the effectiveness of LLMs for RCA and mitigation of production incidents. Zhang et al. [23] introduced an in-context learning approach for automated RCA and performed an extensive analysis of over 100,000 incidents to tackle the high costs associated with fine-tuning LLMs. Drishti et al. [26] demonstrated that incorporating contextual data from various stages of the software development lifecycle (SDLC) enhances performance on critical tasks. Our work synthesizes these methodologies by employing a prompt-based approach with graph-based RAG and expert enhancements, thereby avoiding

the high costs of extensive fine-tuning and numerous in-context learning demonstrations.

RCAgent [56] presented a tool-augmented LLM autonomous agent designed for secure industrial use by deploying models internally to prioritize security over performance. K8sGPT [25] integrated SRE experience into its analyzers and utilized LLMs exclusively for generating natural language outputs, without providing a complete end-to-end solution. As a result, these approaches do not fully harness the capabilities of external models like GPT-4o, as achieved in our work.

LogConfigLocalizer [57] introduced a two-stage LLM-based strategy to identify root-cause configuration properties from Hadoop logs. RCACopilot [24] combined RAG and in-context learning to predict root cause categories, furnishing explanatory narratives by consolidating diagnostic information from incident handlers. Roy et al. [58] evaluated a ReAct agent for the dynamic collection of supplementary incident-related diagnostic data. While these methods emphasize failure localization, classification, and data collection, our approach utilizes RAG with graphs to conduct comprehensive, end-to-end RCA and offer remediation suggestions. Our work notably introduces the StateGraph, enabling LLMs to utilize structured runtime data for better contextual understanding in RCA. Conversely, GraphRAG [59] created knowledge graphs from raw text for community-oriented summaries but lacked the capability to capture rapidly changing system states, rendering it unsuitable for our StateGraph.

### III. SYNERGYRCA DESIGN

We introduce SynergyRCA, our LLM-based approach for RCA, which employs retrieval-augmented generation (RAG) from graph to enhance LLM performance. Section III-A provides an overview of our methodology. Section III-B details the integration of knowledge into LLMs for effective RCA. Finally, Section III-C and Section III-D explore the organization of knowledge within the StateGraph and MetaGraph, respectively.

#### A. Design Overview

*1) Core Concepts:* This section briefly introduces the fundamental concepts related to graphs utilized in SynergyRCA. Within a Kubernetes cluster and its related ecosystems, an *Entity* refers to a distinct and identifiable object in the system. A *Snapshot* captures the state of an entity at regular intervals, offering insights into system dynamics. If the entity is not an Event resource, the snapshot is referred to as the *State*, with uppercase convention applied to denote Snapshots. We construct a *StateGraph* to capture the dependencies among entities across both temporal and spatial dimensions. In the StateGraph, an *entity vertex* is created for each entity, and a *snapshot vertex* is defined for each snapshot, specifically a *STATE vertex* for State. A *statepath* represents a path through the StateGraph that includes only entity vertices, excluding any snapshot vertices. To provide a concise overview of the relationships among various *kinds* of entities, we derive a

*MetaGraph* from the StateGraph, depicting the overall structural connections. A *metapath* within the MetaGraph comprises paths typically consisting only of entity kinds, excluding snapshot kinds. These graphs are stored in Neo4j [27], a graph database well-suited for managing relationships. LLMs are employed to generate Cypher queries, Neo4j's intuitive query language [60], to facilitate data access and manipulation.

*2) SynergyRCA Workflow :* Fig. 1 illustrates the structure of our proposed methodology. Initially, we capture the spatial and temporal relationships among entities within the StateGraph and further delineate the dependencies between various entity kinds using the MetaGraph. These graphs form the foundational knowledge base for subsequent queries that augment LLM prompts.

The analysis procedure is managed by a standard *driver* program. Upon the occurrence of a new incident, the driver queries the StateGraph to match the incident's message, namespace, and timestamp, thereby identifying the source entity kind (the *srcKind*, e.g., Job or ReplicaSet) involved in the incident. This *srcKind*, along with the incident message, is fed into the *Triage* module, which utilizes an LLM to predict the root cause entity kind (the *destKind*, e.g., ResourceQuota) and any relevant resource kinds (the *interKinds*, e.g., Pod and Namespace).

Subsequently, the driver queries the MetaGraph to locate one or more metapaths connecting the *srcKind* to the *destKind*, potentially traversing through *interKinds*. Each metapath is transformed into a Cypher query by the *PathQueryGen* module, with the aid of an LLM for query generation. Executing these queries against the StateGraph yields a statepath (e.g., Job1-Namespace1-ResourceQuota1).

For each entity along the statepath, the driver re-queries the StateGraph to retrieve the current state of the entity (e.g., the details of RESOURCEQUOTA1). The state information for each entity is then input into the *StateChecker* module, which employs an LLM to extract a diagnostic summary relevant to the incident message.

Based on these diagnostic summaries, the *ReportGen* module, powered by an LLM, generates a root cause report and suggests remediation commands. Furthermore, the *ReportQualityChecker* module utilizes an LLM to assess how effectively the report's conclusion explains the error message and determines if further investigation is required.

If the current conclusion does not adequately explain the error message, the incident is re-routed back to the *Triage* module to identify a new *destKind* and *interKinds* for further analysis. This iterative process continues until a satisfactory explanation and remediation plan are achieved, or until the maximum number of trials is reached.

#### B. LLM-based RCA Design

In this section, we outline how to effectively integrate graph and expert knowledge into LLMs through prompts to conduct RCA. Following the workflow depicted in Fig. 1, we describe the design of each LLM module.

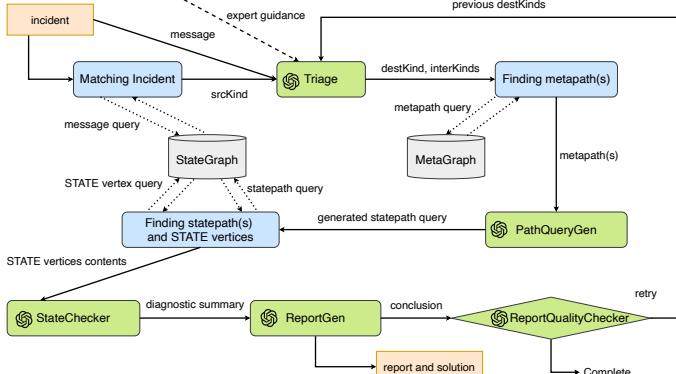


Fig. 1: Overview of SynergyRCA’s Retrieval Augmented LLM Workflow

1) **Triage:** Upon the occurrence of an incident, predicting the potential root cause location is essential. The *Triage* prompt, illustrated in Fig. 2, serves this purpose. Firstly, the LLM is provided with a specific starting point (*srcKind*) by querying the *StateGraph* to match the incident (line 5). Subsequently, the LLM is tasked with identifying all relevant resource kinds (*interKinds*) that might contribute to the incident (lines 6–7). Thirdly, the LLM determines a destination kind (*destKind*) most likely to be the root cause (lines 8–16). Finally, an in-context learning approach [23] instructs the LLM to strictly follow a JSON output format (lines 17–41).

To make *Triage* more effective, we incorporate entity taxonomy, an aspect of graph knowledge, along with expert guidance into its prompts, as shown in Fig. 3. Initially, a predefined list of existing entity kinds is given to the LLM to prevent the creation of non-existent kinds through hallucination (lines 2–5). Additionally, we provide expert guidance to prioritize the prediction focus on specific *destKinds* that may be the “usual suspects” in a specific cluster. For example, we direct attention to specific external kinds like `nfs`, `container`, and `image` (lines 7–11). Finally, we apply naming conventions to help the LLM infer *destKind*, and users can customize these conventions for their specific Kubernetes clusters (lines 12–16). For instance, in handling the error message “*Error: cannot find volume ‘gen-white-list-conf’ to mount into container ‘es-cronstab-job’.*” the presence of ‘-conf’ in the name indicates it is more likely a `ConfigMap` than a `Secret` or `PVC`.

*Triage* is a critical module that leverages LLM capabilities to minimize the need for user experience in Kubernetes. As a result, *Triage* can accurately identify `nfs` as the *destKind* and recommend `PVC` and `PV` as potential *interKinds* for `FailedMount-NoSuchFileDir` issues. *Triage* can be executed in a zero-shot manner using a sufficiently large LLM, such as GPT-4o, due to the widespread usage of Kubernetes and the availability of common error data from various Kubernetes versions. Additionally, deployment-specific errors are managed within the expert-prompt (lines 6–16).

2) **PathQueryGen:** We utilize the code generation capabilities of LLMs to develop *PathQueryGen*, which translates a

```

1 ### Task Description
2 Refer to the predefined resource kinds and naming convention. Analyze the
3 following Kubernetes error message that includes {involved_object}.
4 ### Analysis Steps
5 1. Recognize the {involved_object} as the starting point of the issue.
6 2. Identify the most critical Kubernetes native and external resources relevant to
7 the problem from the predefined resource kinds.
8 3. Determine the ‘destKind’, which is the resource kind most directly related to
9 the problem. Guidelines:
10 (1) Provide only one ‘destKind’; do not list multiple kinds.
11 (2) The ‘destKind’ should be the most crucial one for the error message. For
example, If a quota is exceeded, the ‘destKind’ should be ‘ResourceQuota’ .
13 (3) The ‘destKind’ must be in the relevant resources list and predefined resource
kinds.
15 (4) If ‘destKind’ is an external kind, it should always be in lowercase.
16 (5) If multiple possible kinds, infer the best match using naming conventions.
17 ### Output Formatting
18 Output the findings in JSON format encapsulated within triple backticks and the
19 ‘json’ specifier. The JSON output should not contain additional descriptions and
20 must follow the given structure:
21 ````json
22 {"SourceKind": "{involved_object}",
23 "DestinationKind": "destKind",
24 "RelevantResources": ["Resource1", "Resource2", ...,
25 "involved_object", "destKind"]
26 }
27 ``
28 ### Example
29 Sample Error Message:
30 ``
31 Error creating: pods "es-cronjob-1607245800-kmtgd" is
32 forbidden: exceeded quota: ...
33 ``
34 Sample Output:
35 ````json
36 {"SourceKind": "Job",
37 "DestinationKind": "ResourceQuota",
38 "RelevantResources": ["Job", "Pod", "Namespace",
39 "ResourceQuota"],
40 }
41 ``
42 ### Input
43 Analyze the following error message ensuring ‘destKind’ and ‘RelevantResources’
44 are strictly limited to the provided lists. {error_message}
  
```

Fig. 2: Prompt snippet in *Triage*

metapath into a Cypher query. *PathQueryGen* enhances the metapath by prepending the sequence (“*EVENT →Event →srcKind*”) to it. For example, the extended metapath for issues like `FailedMount-NoSuchFileDir` would be:

```

HasEvent, Event, EVENT, metadata_uid;
ReferInternal, Event, Pod, involvedObject_uid;
ReferInternal, Pod, PersistentVolumeClaim,
spec_volumes_persistentVolumeClaim_claimName;
ReferInternal, PersistentVolume,
PersistentVolumeClaim, spec_claimRef_uid;
UseExternal, PersistentVolume, nfs, spec_nfs_path;
  
```

*PathQueryGen* initiates by processing the error message and constructs a MATCH-WHERE clause for each edge, subsequently returning the nodes and relationships (akin to vertices and edges) along the path. For example, for the “`PV →PVC`” edge, it generates the following Cypher clause:

```

MATCH (pv:PersistentVolume)-[r4:ReferInternal]->(pvc
:PersistentVolumeClaim)
WHERE r4.key = 'spec_claimRef_uid'
  
```

```

1 ### Graph Knowledge of Entity Taxonomy
2 The predefined Kubernetes native resource kinds and external resource kinds are
3 listed below:
4 - k8s native resource kinds: {nativeKinds}
5 - External resource kinds: {externalKinds}
6 ### Expert Knowledge
7 Please note:
8 - For the k8s native kinds, focus on the commonly used kinds: ['ConfigMap',
9 'CronJob', 'DaemonSet', 'Deployment', 'Endpoints', 'Image', 'Job', ...].
10 - For the external kinds, prioritize focusing on: ['nfs', 'container', 'image',
11 'hostpath'].
12 - Naming Conventions:
13 (1) k8s external kinds are always lowercase, while most k8s native kinds are
14 capital case.
15 (2) '-conf' is more likely a ConfigMap, while '-cert' and '-token' are more likely
16 a Secret.

```

Fig. 3: Prompt snippet to import graph and expert knowledge

```

1 ### Analysis Steps
2 3. Chain MATCH Clauses Based on the Metapath:
3 (1) Continue the query by adding MATCH clauses for each part of the provided
4 metapath. For each segment of the metapath, use the node type (srcKind and
5 destKind) as the label for the source and destination node. Use the relationship type
6 (relType) as the label for the connecting relationship, and apply a WHERE clause
7 based on the 'key' property value (propertyName) specified for that relationship:
8 MATCH (startNode:srcKind)-[r1:relType]->(node1:destKind)
9 WHERE r1.key = 'propertyName'
10 (2) For consecutive relationships, increment the relationship alias sequentially to
11 use unique identifiers such as r1, r2, r3, etc. This ensures clarity when multiple
12 relationships are present in the MATCH pattern:
13 MATCH (node1:srcKind)-[r2:relType]->(node2:destKind)
14 WHERE r2.key = 'propertyName'
15 ... and so on for additional relationships.
16 (3) Ensure to use the same node alias for each node type, particularly if that node
17 type appears in multiple relationships to maintain consistency. For example:
18 MATCH (evt: EVENT),
19 MATCH (n1:Event)-[r1:HasEvent]->(evt: EVENT),
20 MATCH (n1:Event)-[r2:ReferInternal]->(n2: Pod)
21 ...
22 ### Example
23 Here's an example for clarity: ...

```

Fig. 4: Prompt snippet in *PathQueryGen*

Fig. 4 depicts the essential steps (lines 2–20) of *PathQueryGen*, which also employs an in-context learning approach (lines 22–23). We focus on generating Cypher queries specifically for statepaths due to their flexibility, whereas other Cypher queries tend to be simple and static.

*PathQueryGen* leverages the knowledge within the MetaGraph by utilizing a metapath. Even though LLMs like GPT-4o possess extensive knowledge of Kubernetes, they can still make errors when connecting specific kinds. For instance, GPT-4o might erroneously suggest a “PV → Node → nfs” path that does not exist. The metapath ensures that *PathQueryGen* generates accurate and executable Cypher queries in the StateGraph with high confidence.

3) **StateChecker:** We developed a powerful module, *StateChecker*, to extract diagnostic summaries from complex entity States, leveraging the semantic understanding capabilities of LLMs. Traditional methods often face challenges due to the

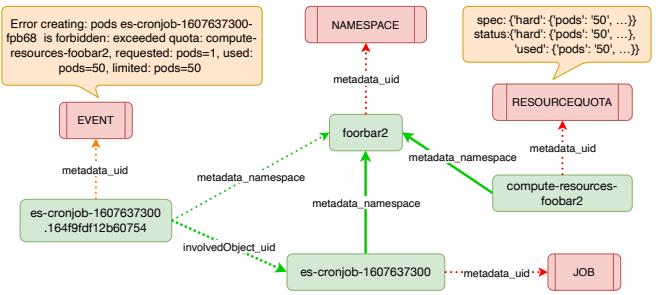


Fig. 5: An example statepath of FailedCreate-ExceedQuotaJob with Event and Snapshots extended.

intricate structure and data types present in States. This complexity is further complicated by varying keys across different entity kinds, and even within a single kind. *StateChecker* excels in extracting the most pertinent information, providing significant advantages in navigating these complexities.

We execute the generated Cypher query in StateGraph and get one or more instanced paths, known as *statepaths*. An example statepath of FailedCreate-ExceedQuotaJob (Job → Namespace → ResourceQuota, with Event and Snapshots extended) is illustrated in Fig. 5. We then query each entity in the statepath to access its STATE vertex through a HasState edge (e.g., Job → JOB), with the content stored as a JSON string termed StateJSON.

We then provide the *StateChecker* with portions of the StateJSON, typically the “spec” and “status” fields, along with the error message, and instruct it to summarize key observations using the prompt snippet shown in Fig. 6. The *StateChecker* assesses whether the error message correlates with the fields in the StateJSON. If they are related, it clarifies the connection and extracts the most pertinent JSON fragments; if they are not, it acknowledges their irrelevance (lines 2–16). It subsequently summarizes the key observations within 200 words (lines 18–33), as demonstrated in the diagnostic summary for RESOURCEQUOTA in Fig. 7.

To minimize hallucinations, we design the prompt to require the LLM to strictly adhere to the factual data present in the StateJSON, avoiding any fabrication or inference, even if discrepancies exist (lines 9–16). This approach prevents the LLM from inferring root causes based solely on the error message. If no STATE vertex is available for an entity, we directly include a key observation noting this absence in the *StateChecker*, bypassing the previous prompt.

4) **ReportGen:** Utilizing the diagnostic summary for each entity along the statepath, we employ three fundamental rules to ascertain the root cause of the error. These rules, rooted in the principle of state reconciliation, are applicable to systems beyond Kubernetes:

- 1) Each entity should have a current state represented by a STATE vertex.
- 2) The states of entities should be consistent (e.g., if a Job tries to create a Pod, the ResourceQuota should have

```

1 ### Analysis Steps
2 3. Conduct an evaluation to determine if there is any evidence (e.g., misconfigurations or errors) in the JSON fields, especially those which could align with the nature of the provided error message.
3 - If the error message seems to relate to the JSON data, clarify the connection and identify any anomalies or errors in the data.
4 - If the error message appears to be unrelated to the JSON data, clearly acknowledge this finding.
5 - Important Rule**:
6 (1) Your analysis must strictly adhere to the factual data provided in the JSON string. Do NOT create or fabricate any new JSON snippets.
7 (2) If the provided JSON cannot explain the root cause of the error message, clearly state that the JSON does not explain it. Do NOT infer the root cause from the error message, and do NOT attempt to justify or rationalize the error message.
8 (3) Report any discrepancies between the error message and the JSON data. Avoid making incorrect observation based on mismatched or incorrect data.
9 ...
10 ...
11 ## Output Formatting
12 Summarize key observations and any issues discovered in the JSON data in a concise manner (limit within 200 words).
13 Key points to include:
14 (1) Key Observations**: Highlight the most relevant findings ONLY based on the provided JSON data for the error message.
15 ...
16 (3) Relevant JSON Fragments**: Include only the most relevant parts or fragments from the JSON data that are crucial for understanding the issue.
17 ...
18 (b) Extract only the deepest fields directly related to the error message instead of including the full JSON structure.
19 ...
20 ...
21 ...
22 ...
23 ...
24 ...
25 ...
26 ...
27 ...
28 ...
29 ...
30 ...
31 ...
32 ...
33 ...

```

Fig. 6: Prompt snippet in *StateChecker*

```

1 RESOURCEQUOTA(25f8e934-6e18-5b9c-bee4-b4231a9d01a6):
2 Key Observations:
3 1. The error message indicates exceeding the pod quota in the 'foobar3' namespace.
4 2. The JSON object is a ResourceQuota for the 'foobar3' namespace.
5 3. The spec and status fields in the JSON confirm a limit of 50 pods and show that 50 pods are already in use.
6 Summary of Issues:
7 The error message is directly related to the JSON data. The ResourceQuota for the 'foobar3' namespace specifies a hard limit of 50 pods, and the status shows that 50 pods are currently in use. Therefore, any attempt to create additional pods will exceed the quota and be forbidden.
8 Relevant JSON Fragments:
9 ...
10 ...
11 ...
12 ...
13 ...
14 ...
15 ...
16 ...
17 ...
18 The JSON data clearly aligns with the error message by showing that the pod quota limit has been reached.

```

Fig. 7: An example diagnostic summary generated by *StateChecker*.

remaining quota).

- 3) Any inconsistency due to data discrepancy should be clearly reported.

We developed *ReportGen* to produce both a root cause report (lines 5–10) and recommend remediation commands (lines 11–15), as guided by the prompt in Fig. 8. A sample report is shown in Fig. 9, where the report accurately pinpoints the ResourceQuota as the root cause (lines 1–24) and

```

1 ### Task Description
2 Based on the previous analysis of {[kinds]}, summarize the root cause of the error message, and pinpoint out the most relevant parts.
3 ...
4 ### Analysis Steps
5 1. For each kind, faithfully summarize the findings based on evidences/facts only, and do not include any suspicions that are not verified. Then provide a score (0 ~10/10) to indicate how relevant it is to the error message.
6 2. Moreover, provide an overall score (0 ~10/10) to indicate how well the conclusion (and detailed summary if needed) can explain the root cause of the error message.
7 3. Furthermore, provide a resolution for the error with kubectl or bash command if applicable. Note: include crucial details such as resource names, IDs, and numbers that are pertinent to understanding the cause. The kubectl/bash command should incorporate the actual resource names, or namespaces, to achieve precision in execution.
8 ...
9 ...
10 ...
11 ...
12 ...
13 ...
14 ...
15 ...
16 ...

```

Fig. 8: Prompt snippet in *ReportGen*

```

1 {"summary": [
2   {"kind": "Job",
3     "explanation": "The Job object specifies creating a new pod in the 'foobar3' namespace, which is at its pod quota limit.",
4     "relevance_score": "9/10"
5   },
6   {"kind": "Namespace",
7     "explanation": "The Namespace 'foobar3' is active but lacks details on resource quotas, providing no direct evidence for the error.",
8     "relevance_score": "2/10"
9   },
10  {"kind": "ResourceQuota",
11    "explanation": "The ResourceQuota for the 'foobar3' namespace shows a hard limit of 50 pods, and the 'used' count is already at 50.",
12    "relevance_score": "10/10"
13  }],
14  "conclusion": "The error message is caused by the 'foobar3' namespace reaching its pod quota limit of 50, as confirmed by the ResourceQuota and Job objects.",
15  "overall_score": "9/10",
16  "resolution": "To resolve the error, increase the pod quota in the namespace 'foobar3': kubectl patch resourcequota compute-resources-foobar3 -n foobar3 -p '{\"spec\": {\"hard\": {\"pods\": \"60\"}}}'",
17  }

```

Fig. 9: An example RCA report and recommended solution for FailedCreate-ExceedQuotaJob.

suggests a specific kubectl command with the appropriate name and namespace (lines 25–28). This design builds on work by Ahmed et al. [22], which demonstrated that GPT-3.x models enhance their ability to propose mitigation plans when the root cause is clearly identified.

**5) ReportQualityChecker:** Finally, we assess the generated root cause report's effectiveness in explaining the error message and determine the need for further investigation. Using the prompt in Fig. 10, we direct the *ReportQualityChecker* to assign a score (lines 1–10) indicating whether additional investigation is necessary. This scoring approach, rather than a direct true/false decision, benefits from a longer chain-of-thought (CoT) [61] for reasoning. If the driver operates interactively, user validation can serve as an alternative to the *ReportQualityChecker*.

```

1  ### Analysis Steps
2 Finally, determine if further investigation is needed. Scoring and Investigation
3 Criteria:
4 1. If the conclusion alone can directly explain the root cause of the error message,
5 score it above 9/10, and set "further_investigation": False.
6 2. If the conclusion cannot solely explain the root cause, but in combination with
7 a detailed summary, they together can explain the root cause, score it above 7/10
8 and set "further_investigation": False.
9 3. If the conclusion and summary can only partially explain the root cause or are
10 merely relevant to it, score it below 5/10 and set "further_investigation": True.

```

Fig. 10: Prompt snippet in *ReportQualityChecker*

### C. Constructing StateGraph from Running Kubernetes Cluster

This section introduces *StateGraph*, a graph structure designed to capture the spatial and temporal relationships among entities within a Kubernetes environment. StateGraph enables the identification of connections among entities at specific timestamps and the tracking of entity state evolution over time. We begin by defining the key concepts fundamental to the design of StateGraph and then detail the construction process.

#### 1) Basic Concepts in StateGraph:

a) *Entities and Snapshots*: In the Kubernetes domain, an *entity* represents a unique and identifiable object or concept within the system. Examples of entities include Pods, PVCs, and NFS directories. Entities are classified into two categories: *k8s-native* entities, which are inherent to Kubernetes (e.g., Pod, ReplicaSet, Job, and CronJob) and stored in etcd, and *k8s-external* entities, which are not native to Kubernetes but are still relevant (e.g., containers, images, and NFS directories).

Entities may consist of one or more fields. An entity with a single field (e.g., IP address) is deemed *atomic*; if it has multiple fields, it is considered *composite*. Many entities, such as k8s-native entities like Pods or PVCs are composite and typically include fields like uid, name, namespace, and kind. Similarly, an NFS directory may have fields like path and server. Entities are uniquely identified by an identifier (e.g., uid) or a field combination (e.g., path and server).

We periodically collect data to capture the state of an entity at each timestamp, known as a *Snapshot*. For example, querying etcd for Pods every 5 minutes provides snapshots of the Pod's state at these intervals. However, rapid changes within an interval may result in inconsistent snapshots.

b) *Vertices and Edges*: The StateGraph consists of two vertex types: *entity vertices* and *snapshot vertices*. Lowercase letters represent k8s-external entities, capitalized letters denote k8s-native entities, and uppercase letters indicate snapshots. For example, POD represents a snapshot of a Pod. If the snapshot is not for an Event resource, it is also referred to as a *STATE vertex*.

StateGraph features two types of Entity-Entity edges. Edges labeled *ReferInternal* indicate native-to-native connections, while *UseExternal* denotes native-to-external connections. Entity-Snapshot edges also have two types: due to Event objects in etcd functioning like logs, we use the special *HasEvent* type; all other connections, whether native or external,

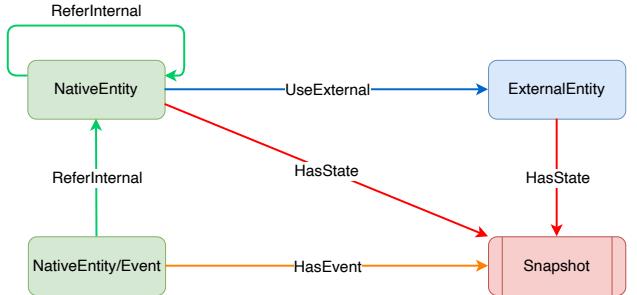


Fig. 11: Schematic diagram of StateGraph

are labeled *HasState*. These concepts and relationships form the foundation of the StateGraph, as illustrated in Fig. 11.

2) *Constructing the StateGraph*: The construction of the StateGraph involves a streamlined process consisting of collecting snapshots, removing duplicates, extracting and canonicalizing entities, and creating vertices and edges. An example snippet of a StateGraph is shown in Fig. 12.

- *Collecting snapshots*. We periodically collect snapshots from diverse components across multiple layers, including database (e.g., etcd in Kubernetes), compute runtime (e.g., containers in Nodes), storage systems (e.g., NFS servers and mount directories), network configurations (e.g., Calico, iptables), and images (e.g., images in Nodes and repositories).
- *Removing duplicates*. To ensure accurate tracking of state evolution, we eliminate consecutive duplicate snapshots while retaining the last one. This method allows us to trace changes, such as Deployment replica counts over time. Each snapshot maintains its associated time range.
- *Extracting entities*. Entities are extracted from each snapshot by identifying keys likely to represent or constitute an entity. Utilizing a rules and statistics-based method inspired by previous work [62], alongside human validation, we create a reliable reference. Keys that frequently appear with diverse values are identified as entity keys. We employ batch processing within Spark [63] to efficiently extract these key-value pairs.
- *Canonicalizing entities*. For consistent matching, entities are canonicalized. For example, transforming a Pod's reference to a PVC by its name into its canonical form (uid, name, namespace, kind) ensures uniformity across snapshots.
- *Creating vertices and edges*. For each snapshot, we designate a *primary* entity and establish connections to other referenced entities. A vertex is created for each entity, with *ReferInternal* or *UseExternal* edges connecting the primary entity to other referenced entities. Additionally, a vertex is created for the snapshot itself, linking it to the primary entity through *HasState* or *HasEvent* edges. These edges include properties such as the snapshot's time range and key. For a series of snapshots, we consolidate edges with identical source, destination, and key,

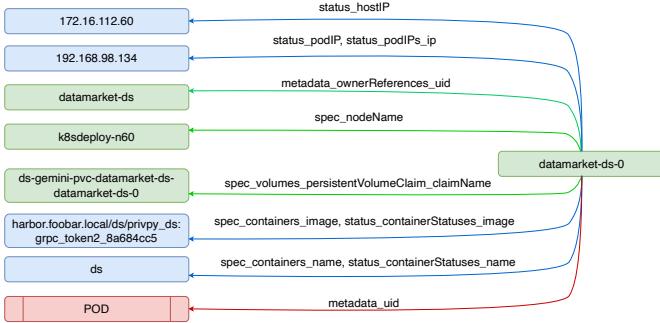


Fig. 12: An example StateGraph snippet for a single Pod snapshot.

using  $t_{\min}$  and  $t_{\max}$  to indicate the valid time range. Duplicated vertices are merged to maintain a clean and efficient graph structure.

#### D. Extracting MetaGraph from StateGraph

We construct a *MetaGraph*, similar to a class in object-oriented programming, to represent the overall structure of the StateGraph, which resembles an object. This MetaGraph is directly extracted from the StateGraph to ensure precision, avoiding extra or missing connections that may result from using alternative sources like Kubernetes documentation (e.g., Custom Resources). Such accuracy is essential for accurately identifying instantiated statepaths for a given metapath.

To form the MetaGraph, we create a distinct vertex for each entity kind in the StateGraph. For example, a *Pod* vertex represents all Pods. These vertices are grouped into three categories: *NativeEntity* for k8s-native entities (e.g., Pod, StatefulSet), *ExternalEntity* for external entities (e.g., container, image), and *Snapshot* for snapshots (e.g., POD for a Pod’s state). We retain the edge types from the StateGraph, categorizing them into four types: *ReferInternal*, *UseExternal*, *HasState*, and *HasEvent*. For instance, a ReferInternal edge between a *Pod* vertex and a *StatefulSet* vertex signifies a reference relationship.

The MetaGraph construction involves the following steps: First, from each triplet in the StateGraph (comprising a source vertex, edge, and a destination vertex), we extract a quadruplet in the format (srcKind, destKind, key, type). For example, a triplet involving a *Pod* and a *StatefulSet* may produce a quadruplet like (*Pod*, *StatefulSet*, *metadata\_ownerReferences\_name*, *ReferInternal*). Next, we remove duplicates from these quadruplets (or count their occurrences) to ensure that each unique relationship is represented once (or to collect frequency statistics). Then, we create vertices for each distinct srcKind and destKind. Subsequently, we establish edges between these vertices using the edge type (e.g., *ReferInternal*) and key (e.g., *metadata\_ownerReferences\_name*) to define the relationship nature. These steps result in a MetaGraph, as illustrated in Fig. 13, potentially containing additional edges due to further StateGraph snippets from other snapshots.

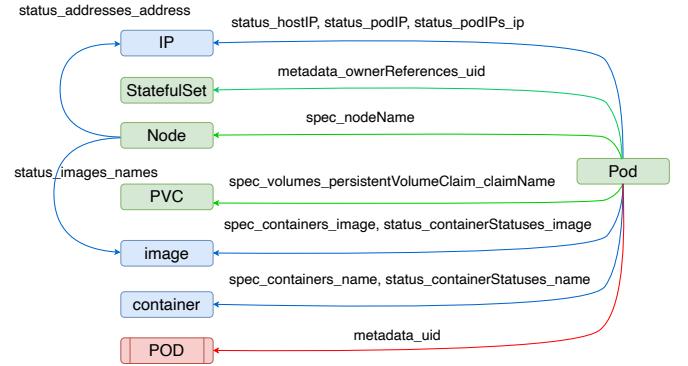


Fig. 13: An Example MetaGraph snippet encompassing the StateGraph snippet.

## IV. EVALUATION

In this section, we take a comprehensive approach to evaluate SynergyRCA, focusing on its effectiveness, module performance, and efficiency. Initially, we demonstrate SynergyRCA’s capability to accurately identify root causes within Kubernetes clusters, showcasing its real-world applicability. Next, we assess the performance of each LLM module in SynergyRCA, highlighting their precision and reliability. Finally, we analyze time and token costs to measure efficiency.

This comprehensive evaluation seeks to validate SynergyRCA’s practical applicability, the robustness of its modules, and its operational efficiency, reaffirming its value as a powerful RCA tool in dynamic environments such as Kubernetes clusters.

#### A. Implementation and Setup

We develop SynergyRCA for Kubernetes by leveraging GPT-4o as our LLM and Neo4j [27] for managing StateGraph and MetaGraph databases. Initially, we build the graphs using PySpark [63], [64] and the GraphFrames library [65], [66] for scalability and efficiency, then seamlessly integrate them into Neo4j for robust graph queries. We implement LLM modules in SynergyRCA using OpenAI’s Assistants API [67], [68] and executed GPT-4o on Microsoft Azure.

SynergyRCA comprises approximately 5900 lines of Python code, divided into 1620 lines for data collection, 2480 lines for constructing StateGraph and MetaGraph in PySpark, and 1760 lines for LLM-driven analysis using GPT-4o. Python code also generates bash commands for importing graphs into Neo4j, streamlining data integration.

For evaluation, we collect data from two production Kubernetes clusters of different sizes and versions. The first cluster, with 27 nodes running version 1.18, provides a week’s worth of data. The second cluster, with 88 nodes running version 1.21, offers six months of data. After removing consecutive duplicates, we retain datasets of 13.2GB and 118.8GB, respectively, ensuring high-quality data for evaluation. Incident owners and senior Kubernetes administrators [22], [23] are

invited to formulate the ground truth and evaluate the results to ensure assessment accuracy.

### B. Identifying Root Cause

We prepare test datasets comprising frequently occurring Kubernetes incidents with observable impacts, such as FailedCreate, FailedMount and Evicted, which arise from different reasons. Each reason can yield various error message types; for example, FailedCreate may include ExceedQuotaJob or ServiceAccount-NotFound scenarios. Table I provides a few examples. To ensure broad coverage, we randomly sampled these messages across different namespaces and timestamps.

The precision of SynergyRCA for two datasets is reported in Table II. For each error message, referred to as an *example*, SynergyRCA attempts to generate a report up to three trials. If SynergyRCA produces a report that reasonably explains the root cause based on the ground truth, it is counted as *correct*. Precision is defined as the percentage of correctly explained examples. The average precision achieved is 0.88 for dataset-1 and 0.92 for dataset-2.

Examples such as Evicted-LowOnResource, FailedCreate-ServiceAccountNotFound and FailedMount-NoSuchFileDir were well explained, while FailedCreate-ExceedQuotaReplicaSet, FailedMount-ObjectNotRegistered and FailedScheduling-UnBoundPVC exhibited lower precision. The primary reason is that there is a *discrepancy* in data due to inconsistent snapshot. For instance, while checking PERSISTENTVOLUMECLAIM, the status might show “bound,” whereas the error indicates “Unbound.” This is due to snapshots being captured every 5 minutes without ensuring consistency, potentially missing rapidly changing states. Similar issues occur in FailedCreate-ExceedQuota\* when a pod requests CPU or memory resources. If we tolerate inconsistent snapshots, precision increases above 0.95, highlighting SynergyRCA’s effectiveness using LLM.

### C. Effectiveness of Each LLM Module

We closely examine the performance of the LLM-based modules in SynergyRCA, with results shown in Table III.

For *Triage*, we assess its ability to propose the most relevant destKind, rating these proposals as high, moderate, low, or unrelated. Only high-related destKinds are considered correct, and we calculate their percentage as *Triage-Precision* (rows 1–2).

For *PathQueryGen*, we show the ability to generate runnable Cypher query, we take the percentage of successfully completed queries with expected output as *PathQueryGen-Precision* (row 3).

For *ReportGen*, we invite Kubernetes experts to evaluate the deduction quality from the diagnostic summary to conclusion. Deductions aligning with expert reasoning and leading to valid conclusions are deemed correct, calculated as *ReportGen-Conclusion-Precision* (row 4). Experts also assess the usefulness of recommended commands, with adherence to common

remediation practices deemed helpful, tallied as *ReportGen-Command-Precision* (row 5).

For *ReportQualityChecker*, we evaluate further investigation suggestions against expert labels. A false positive occurs when *ReportQualityChecker* labels a trial as True, indicating no further investigation is necessary, but the expert labels it as False, implying otherwise (row 6). Conversely, a false negative arises when *ReportQualityChecker* advises further investigation by labeling the trial as False, although the expert believes it True, requiring no additional analysis (row 7). False positives can hinder error message interpretation, while false negatives result in unnecessary use of computational resources. We consider both error rates to assess *ReportQualityChecker*’s effectiveness.

Due to the verbose nature of *StateChecker* summaries, which makes semantic quality difficult to quantify, we omit this module and consider the overall precision (e.g. 0.88) as a lower bound for its performance.

We calculate both the weighted mean and arithmetic mean for each metric. Essentially, each metric has a table similar to Table II. The weighted mean is computed as Total(#Correct)/Total(#Example) (e.g., 549/619), while the arithmetic mean is the average of the precision values (e.g., 0.88).

Table III shows the *mean Triage-Precision* ranges from 0.89 to 0.95 (row 1), indicating effective prediction of high-related destKinds. Without graph and expert guidance (i.e., Fig. 3), this precision drops to 0.79 to 0.91 (row 2), where determining the correct volume type (e.g., ConfigMap or Secret) or case (e.g., nfs or NFS) becomes challenging. The *mean PathQueryGen-Precision* of approximately 0.95 (row 3) indicates a low failure rate of 5%, demonstrating its robust ability to accurately convert metapaths into executable queries. The *mean ReportGen-Conclusion-Precision* ranges from 0.92 to 0.95 (row 4), demonstrating its ability to draw reliable conclusions, and *mean ReportGen-Command-Precision* spans from 0.94 to 0.97 (row 5), indicating its effectiveness in recommending useful remediation commands. The *mean ReportQualityChecker-FPR* ranges from 0.07 to 0.10 (row 6), suggesting possible oversight of cases needing retries due to occasional acceptance of a conclusion despite discrepancies. The *mean ReportQualityChecker-FNR*, from 0.04 to 0.07 (row 7), shows it rarely rejects correct reports, thus minimizing unnecessary computational resource usage.

### D. Time and Token Cost

Given an error message, SynergyRCA runs a max of three trials to explain the root cause, we call each trial as an *attempt*. We show the time and token costs of each attempt in Table IV and Table V for dataset-1 and dataset-2, respectively.

The average time cost across attempts is approximately 2 minutes. The average total token cost varies from 73K to 161K tokens, translating to roughly \$0.19 to \$0.41 based on GPT-4o pricing [69]. Importantly, about 99% of the total token cost comes from input prompt tokens. This high percentage is due to two main factors: (1) *StateChecker* processes the content of each STATE vertex, which constitutes the majority

TABLE I: Error message examples in two evaluation datasets

Reason	Type	Message
FailedCreate	NoVolumeToMount	Error: cannot find volume “gen-white-list-conf” to mount into container “es-crontab-job”
FailedCreate	ExceedQuotaJob	Error creating: pods “es-cronjob-1607637300-fpb68” is forbidden: exceeded quota: compute-resources-foobar3, requested: pods=1, used: pods=50, limited: pods=50
FailedCreate	ExceedQuotaReplicaSet	Error creating: pods “normal-es-s31-66bf4bb56d-zp2m4” is forbidden: exceeded quota: compute-resources-foobar5, requested: limits.memory=32Gi, used: limits.memory=5372Gi, limited: limits.memory=5400Gi
FailedCreate	ExceedQuotaStatefulSet	create Pod foobar4-c1-0 in StatefulSet foobar4-c1 failed error: pods “foobar4-c1-0” is forbidden: exceeded quota: compute-resources-foobar4, requested: pods=1, used: pods=50, limited: pods=50
FailedMount	NoSuchFileDir	(combined from similar events): MountVolume.SetUp failed for volume ‘pvc-ca00f7a6-fb99-49a1-9881-e1f98db9297d’: mount failed: exit status 32 Mounting command: systemd-run Mounting arguments: --description=Kubernetes transient mount for /var/lib/kubelet/pods/ae161f5fc87e-4c7a-aea5-51a46532ee7e/volumes/kubernetes.io~nfs/pvc-ca00f7a6-fb99-49a1-9881e1f98db9297d --scope -- mount -t nfs 172.16.112.63:/mnt/k8s_nfs_pv/foobar2-common-mysql-pvc-0-common-mysql-0-0-pvc-ca00f7a6fb99-49a1-9881-e1f98db9297d /var/lib/kubelet/pods/ae161f5fc87e-4c7a-aea5-51a46532ee7e/volumes/kubernetes.io~nfs/pvc-ca00f7a6-fb99-49a1-9881e1f98db9297d Output: Running scope as unit: run-r5389742587e44b6db650d4ea59fb94e5.scope mount.nfs: mounting 172.16.112.63:/mnt/k8s_nfs_pv/foobar2common-mysql-pvc-0-common-mysql-0-0-pvc-ca00f7a6-fb99-49a19881-e1f98db9297d failed, reason given by server: No such file or directory
FailedScheduling	UnboundPVC	pod has unbound immediate PersistentVolumeClaims (repeated 19 times)
FailedScheduling	NodeNotAvailable	0/87 nodes are available: 1 Too many pods, 3 node(s) had taint {node-role.kubernetes.io/master:}, that the pod didn’t tolerate, 38 node(s) were unschedulable, 45 node(s) didn’t match Pod’s node affinity/selector.

TABLE II: Precision of SynergyRCA in two datasets

Reason	Type	dataset-1			dataset-2		
		#Correct	#Example	Precision	#Correct	#Example	Precision
ClaimLost	PVLost	-	-	-	21	23	0.91
Evicted	LowOnResource	20	20	1.00	34	35	0.97
Evicted	NodeDiskPressure	24	24	1.00	24	24	1.00
Failed	AccessDenied	39	39	1.00	-	-	-
Failed	ArtifactNotFound	20	20	1.00	21	22	0.95
Failed	NetworkUnreachable	21	21	1.00	-	-	-
Failed	NoVolumeToMount	18	24	0.75	37	37	1.00
FailedCreate	ExceedQuotaJob	37	45	0.82	37	46	0.80
FailedCreate	ExceedQuotaReplicaSet	18	32	0.56	42	54	0.78
FailedCreate	ExceedQuotaStatefulSet	19	20	0.95	41	44	0.93
FailedCreate	ServiceAccountNotFound	32	32	1.00	40	40	1.00
FailedMount	ConfigMapNotFound	43	43	1.00	57	58	0.98
FailedMount	FailedSyncConfigMapCache	56	56	1.00	64	64	1.00
FailedMount	FailedSyncSecretCache	54	57	0.95	60	60	1.00
FailedMount	NoSuchFileDir	47	47	1.00	42	42	1.00
FailedMount	ObjectNotRegistered	14	24	0.58	34	34	1.00
FailedMount	PVCNotFound	-	-	-	39	56	0.70
FailedMount	SecretNotFound	56	56	1.00	55	55	1.00
FailedMount	ServiceAccountNotFound	-	-	-	25	25	1.00
FailedMount	StaleNFS	21	21	1.00	-	-	-
FailedScheduling	PVCNotFound	-	-	-	34	34	1.00
FailedScheduling	UnboundPVC	10	38	0.26	33	71	0.46
OutOfPods	NodeNotEnough	-	-	-	19	19	1.00
Total		549	619		759	843	
Average				0.88			0.92

TABLE III: Correctness metrics of each LLM module in two datasets

Metric (see text)	dataset-1		dataset-2	
	Weighted Mean	Arithmetic Mean	Weighted Mean	Arithmetic Mean
Triage-Precision	0.89	0.91	0.92	0.95
Triage-Precision (without knowledge)	0.84	0.79	0.91	0.90
PathQueryGen-Precision	0.95	0.95	0.95	0.94
ReportGen-Conlusion-Precision	0.92	0.94	0.93	0.95
ReportGen-Command-Precision	0.97	0.97	0.94	0.94
ReportQualityChecker-FPR	0.09	0.08	0.10	0.07
ReportQualityChecker-FNR	0.04	0.04	0.07	0.05

TABLE IV: Average time and token cost for each attempt of an error message in dataset-1

Reason	Type	TimeCost (sec)	PromptToken	CompletionToken	TotalToken
Evicted	LowOnResource	93.99	119179.05	1052.60	120231.65
Evicted	NodeDiskPressure	75.14	122300.46	919.25	123219.71
Failed	AccessDenied	110.46	136474.10	1486.08	137960.18
Failed	ArtifactNotFound	741.95	1335117.00	5178.65	1340295.65
Failed	NetworkUnreachable	152.99	114172.90	1930.10	116103.00
Failed	NoVolumeToMount	115.39	106318.08	1347.16	107665.24
FailedCreate	ExceedQuotaJob	101.72	112012.95	1048.98	113061.93
FailedCreate	ExceedQuotaReplicaSet	84.34	44067.00	945.89	45012.89
FailedCreate	ExceedQuotaStatefulSet	103.97	95695.17	1222.83	96918.00
FailedCreate	ServiceAccountNotFound	79.09	50227.56	903.00	51130.56
FailedMount	ConfigMapNotFound	46.86	27001.02	523.33	27524.35
FailedMount	FailedSyncConfigMapCache	42.59	19088.24	523.38	19611.62
FailedMount	FailedSyncSecretCache	85.98	49897.12	736.77	50633.89
FailedMount	NoSuchFileDir	147.71	189353.10	2506.44	191859.54
FailedMount	ObjectNotRegistered	57.87	45831.43	524.82	46356.25
FailedMount	SecretNotFound	42.49	35634.36	563.16	36197.52
FailedMount	StaleNFS	193.21	137606.00	2174.62	139780.62
FailedScheduling	UnboundPVC	82.26	128558.25	994.53	129552.78
<b>Average</b>		131.00	159362.99	1365.64	160728.63

TABLE V: Average time and token cost for each attempt of an error message in dataset-2

Reason	Type	TimeCost (sec)	PromptToken	CompletionToken	TotalToken
ClaimLost	PVLost	122.18	46828.45	1024.45	47852.90
Evicted	LowOnResource	107.53	61658.14	770.92	62429.05
Evicted	NodeDiskPressure	96.60	44501.46	668.75	45170.21
Failed	ArtifactNotFound	149.82	103740.09	1272.74	105012.83
Failed	NoVolumeToMount	147.08	60986.00	938.86	61924.86
FailedCreate	ExceedQuotaJob	99.52	54945.42	816.05	55761.47
FailedCreate	ExceedQuotaReplicaSet	131.67	165871.66	1104.28	166975.95
FailedCreate	ExceedQuotaStatefulSet	118.75	83234.30	1266.96	84501.26
FailedCreate	ServiceAccountNotFound	156.96	145011.48	975.20	145986.68
FailedMount	ConfigMapNotFound	90.06	27608.33	584.45	28192.78
FailedMount	FailedSyncConfigMapCache	108.28	44007.63	591.33	44598.96
FailedMount	FailedSyncSecretCache	129.67	54923.32	851.10	55774.41
FailedMount	NoSuchFileDir	130.71	78662.62	1597.67	80260.29
FailedMount	ObjectNotRegistered	105.05	33504.15	637.12	34141.26
FailedMount	PVCNotBound	119.72	99246.44	1135.55	100381.98
FailedMount	SecretNotFound	108.35	24171.76	591.29	24763.05
FailedMount	ServiceAccountNotFound	126.57	31199.46	882.65	32082.12
FailedScheduling	PVCNotFound	88.83	31818.56	554.21	32372.76
FailedScheduling	UnboundPVC	146.46	186477.81	1036.35	187514.15
OutOfPods	NodeNotEnough	89.69	65294.05	718.42	66012.47
<b>Average</b>		118.67	72184.56	900.92	73085.47

of the prompt tokens, as seen in scenarios like Failed-Create-ExceedQuotaReplicaSet where it examines REPLICASET, NAMESPACE and RESOURCEQUOTA, each containing spec and status fields. (2) The output of each module is significantly shorter than the input prompt. For example, *Triage* and *PathQueryGen* use a prompt of approximately 50 lines (604 tokens) but yield a 3-line JSON output (41 tokens) or a 15-line Cypher query (288 tokens).

We also observe that the Failed-ArtifactNotFound case in dataset-1 has a longer average running time due to the necessity of exhaustively verifying approximately 10.30 metapaths, whereas others typically only verify around 1.32 metapaths on average.

## V. DISCUSSION

We have demonstrated the effectiveness and efficiency of SynergyRCA. However, capturing consistent snapshots of fast-changing resources such as CPU and memory remains challenging, highlighting a limitation of the StateGraph. This challenge arises from inherent difficulties in data collection,

despite our efforts to estimate polling frequency. Implementing event-driven methods like triggers or logs could impose significant burdens on the target system (i.e., Kubernetes), potentially disrupting normal cluster operations.

While the state reconciliation principle is broadly applicable, some error messages necessitate more advanced checking strategies beyond merely verifying the STATE vertex's existence and content. For instance, addressing the Failed-Scheduling-NodeNotAvailable issue, as seen in Table I, requires aggregating data across all Nodes, even if *Triage* predicts the destKind as Node. Although foundational elements are in place, further enhancements are necessary to address more complex cases. We see the use of an agent [56] with tool-augmentation generation (TAG) as a promising approach to support diverse checking strategies and handle a wider range of scenarios, which we plan to explore in future work.

## VI. CONCLUSION

In this paper, we present SynergyRCA, an innovative tool utilizing cutting-edge LLMs such as GPT-4o for root cause analysis (RCA) in Kubernetes environments. SynergyRCA employs retrieval-augmented generation (RAG) from graph databases and expert prompts to boost LLM effectiveness, bypassing the need for expensive fine-tuning or extensive in-context learning demonstrations. Through a comprehensive evaluation on two real-world datasets from production Kubernetes clusters, we demonstrate that SynergyRCA can accurately identify root causes, achieving average precision of 0.88 and 0.92, and rapidly pinpointing issues in approximately 2 minutes on average. Our findings reveal that SynergyRCA identifies numerous root causes, including novel ones, thereby significantly advancing the state-of-the-art in RCA. This research highlights the potential of combining LLMs with graph databases to enhance reliability and effectiveness in dynamic cloud environments, as well as AIOps more broadly.

## REFERENCES

- [1] K8s-project, “Kubernetes:production-grade container orchestration,” <https://kubernetes.io/>, 2024, accessed: 2024-11-24.
- [2] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes, “Large-scale cluster management at google with borg,” in *Proceedings of the tenth european conference on computer systems*, 2015, pp. 1–17.
- [3] T. Melissaris, K. Nabar, R. Radut, S. Rehmtulla, A. Shi, S. Chandrashekhar, and I. Papapanagiotou, “Elastic cloud services: scaling snowflake’s control plane,” in *Proceedings of the 13th Symposium on Cloud Computing*, 2022, pp. 142–157.
- [4] C. Tang, K. Yu, K. Veeraraghavan, J. Kaldor, S. Michelson, T. Kooburat, A. Anburdurai, M. Clark, K. Gogia, L. Cheng *et al.*, “Twine: A unified cluster management system for shared infrastructure,” in *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, 2020, pp. 787–803.
- [5] CNCF, “Cloud native 2023: The undisputed infrastructure of global technology,” <https://www.cncf.io/reports/cncf-annual-survey-2023/>, 2023, accessed: 2024-11-24.
- [6] K8s-project, “Controllers and reconciliation,” [https://cluster-api.sigs.k8s.io/developer/providers/implementers-guide/controllers\\_and\\_reconciliation.html](https://cluster-api.sigs.k8s.io/developer/providers/implementers-guide/controllers_and_reconciliation.html), 2024, accessed: 2024-11-24.
- [7] B. Burns, B. Grant, D. Oppenheimer, E. Brewer, and J. Wilkes, “Borg, omega, and kubernetes,” *Communications of the ACM*, vol. 59, no. 5, pp. 50–57, 2016.
- [8] J. T. Gu, X. Sun, W. Zhang, Y. Jiang, C. Wang, M. Vaziri, O. Legunsen, and T. Xu, “Acto: Automatic end-to-end testing for operation correctness of cloud system management,” in *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023, pp. 96–112.
- [9] X. Sun, W. Ma, J. T. Gu, Z. Ma, T. Chajed, J. Howell, A. Lattuada, O. Padon, L. Suresh, A. Szekeres *et al.*, “Anvil: Verifying liveness of cluster management controllers,” in *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, 2024, pp. 649–666.
- [10] X. Sun, W. Luo, J. T. Gu, A. Ganesan, R. Alagappan, M. Gasch, L. Suresh, and T. Xu, “Automatic reliability testing for cluster management controllers,” in *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, 2022, pp. 143–159.
- [11] G. M. Diouf, H. Elbiaze, and W. Jaafar, “On byzantine fault tolerance in multi-master kubernetes clusters,” *Future Generation Computer Systems*, vol. 109, pp. 407–419, 2020.
- [12] H. Jacobs, “Kubernetes failure stories,” <https://k8s.af/>, 2024, accessed: 2024-11-24.
- [13] M. Cebula and B. Sherrod, “Weird ways to blow up your kubernetes,” *KubeCon North America (Nov. 2019)*, 10.
- [14] C. Madhu, “Preventing controller sprawl from taking down your cluster,” *KubeCon North America (Oct. 2022)*, 2022.
- [15] Github, “Zookeeper pod keeps crashing when scaling down and up,” <https://github.com/pravega/zookeeper-operator/pull/526>, 2024, accessed: 2024-11-24.
- [16] CASSKOP-370., “[bug] tidb operator unable to recover an unhealthy cluster even with manual revert,” <https://github.com/Orange-OpenSource/casskop/issues/370>, 2024, accessed: 2024-11-24.
- [17] Github, “Tidb operator unable to recover an unhealthy cluster even with manual revert,” <https://github.com/pingcap/tidb-operator/issues/4946>, 2024, accessed: 2024-11-24.
- [18] N. Jain, S. Vaidyanath, A. Iyer, N. Natarajan, S. Parthasarathy, S. Rajamani, and R. Sharma, “Jigsaw: Large language models meet program synthesis,” in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 1219–1231.
- [19] A. Mastropaoletti, L. Pascarella, and G. Bavota, “Using deep learning to generate complete log statements,” in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 2279–2290.
- [20] Z. Fan, X. Gao, A. Roychoudhury, and S. H. Tan, “Improving automatically generated code from codex via automated program repair,” *arXiv preprint arXiv:2205.10583*, 2022.
- [21] A. Mastropaoletti, S. Scalabrin, N. Cooper, D. N. Palacio, D. Poshyvanyk, R. Oliveto, and G. Bavota, “Studying the usage of text-to-text transfer transformer to support code-related tasks,” in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 2021, pp. 336–347.
- [22] T. Ahmed, S. Ghosh, C. Bansal, T. Zimmermann, X. Zhang, and S. Rajmohan, “Recommending root-cause and mitigation steps for cloud incidents using large language models,” in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 1737–1749.
- [23] X. Zhang, S. Ghosh, C. Bansal, R. Wang, M. Ma, Y. Kang, and S. Rajmohan, “Automated root causing of cloud incidents using in-context learning with gpt-4,” in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 2024, pp. 266–277.
- [24] Y. Chen, H. Xie, M. Ma, Y. Kang, X. Gao, L. Shi, Y. Cao, X. Gao, H. Fan, M. Wen *et al.*, “Automatic root cause analysis via large language models for cloud incidents,” in *Proceedings of the Nineteenth European Conference on Computer Systems*, 2024, pp. 674–688.
- [25] K8sGPT-project, “k8sgpt: Giving kubernetes superpowers to everyone,” <https://github.com/k8sgpt-ai/k8sgpt>, 2024, accessed: 2024-11-24.
- [26] D. Goel, F. Husain, A. Singh, S. Ghosh, A. Parayil, C. Bansal, X. Zhang, and S. Rajmohan, “X-lifecycle learning for cloud incident management using llms,” in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 2024, pp. 417–428.
- [27] Neo4j, “Neo4j graph database and analytics,” <https://neo4j.com/>, 2024, accessed: 2024-11-27.
- [28] A. P. Azad, S. Ghosh, A. Gupta, H. Kumar, P. Mohapatra, L. Eckstein, L. Posner, and R. Kern, “Picking pearl from seabed: Extracting artefacts from noisy issue triaging collaborative conversations for hybrid cloud services,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 440–12 446.
- [29] J. Chen, X. He, Q. Lin, Y. Xu, H. Zhang, D. Hao, F. Gao, Z. Xu, Y. Dang, and D. Zhang, “An empirical investigation of incident triage for online service systems,” in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 2019, pp. 111–120.
- [30] J. Chen, X. He, Q. Lin, H. Zhang, D. Hao, F. Gao, Z. Xu, Y. Dang, and D. Zhang, “Continuous incident triage for large-scale online service systems,” in *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2019, pp. 364–375.
- [31] Y. Chen, X. Yang, H. Dong, X. He, H. Zhang, Q. Lin, J. Chen, P. Zhao, Y. Kang, F. Gao *et al.*, “Identifying linked incidents in large-scale online service systems,” in *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, 2020, pp. 304–314.
- [32] S. Ghosh, K. Grover, J. Wong, C. Bansal, R. Namineni, M. Verma, and S. Rajmohan, “Dependency aware incident linking in large cloud systems,” in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 141–150.
- [33] C. Bansal, S. Renganathan, A. Asudani, O. Midy, and M. Janakiraman, “Decaf: Diagnosing and triaging performance issues in large-scale cloud services,” in *Proceedings of the ACM/IEEE 42nd International*

- Conference on Software Engineering: Software Engineering in Practice*, 2020, pp. 201–210.
- [34] C. Luo, J.-G. Lou, Q. Lin, Q. Fu, R. Ding, D. Zhang, and Z. Wang, “Correlating events with time series for incident diagnosis,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1583–1592.
- [35] V. Nair, A. Raul, S. Khanduja, V. Bahirwani, Q. Shao, S. Sellamanickam, S. Keerthi, S. Herbert, and S. Dhulipalla, “Learning a hierarchical monitoring system for detecting and diagnosing service issues,” in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 2029–2038.
- [36] J. Jiang, W. Lu, J. Chen, Q. Lin, P. Zhao, Y. Kang, H. Zhang, Y. Xiong, F. Gao, Z. Xu *et al.*, “How to mitigate the incident? an effective troubleshooting guide recommendation technique for online service systems,” in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 1410–1420.
- [37] Q. Lin, H. Zhang, J.-G. Lou, Y. Zhang, and X. Chen, “Log clustering based problem identification for online service systems,” in *Proceedings of the 38th International Conference on Software Engineering Companion*, 2016, pp. 102–111.
- [38] Y. Yuan, W. Shi, B. Liang, and B. Qin, “An approach to cloud execution failure diagnosis based on exception logs in openstack,” in *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*. IEEE, 2019, pp. 124–131.
- [39] C. Lee, T. Yang, Z. Chen, Y. Su, and M. R. Lyu, “Eadro: An end-to-end troubleshooting framework for microservices on multi-source data,” in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 1750–1762.
- [40] S. Zhang, P. Jin, Z. Lin, Y. Sun, B. Zhang, S. Xia, Z. Li, Z. Zhong, M. Ma, W. Jin *et al.*, “Robust failure diagnosis of microservice system through multimodal data,” *IEEE Transactions on Services Computing*, vol. 16, no. 6, pp. 3851–3864, 2023.
- [41] A. Ikrام, S. Chakraborty, S. Mitra, S. Saini, S. Bagchi, and M. Koçaoğlu, “Root cause analysis of failures in microservices through causal discovery,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 31 158–31 170, 2022.
- [42] L. Wang, C. Zhang, R. Ding, Y. Xu, Q. Chen, W. Zou, Q. Chen, M. Zhang, X. Gao, H. Fan *et al.*, “Root cause analysis for microservice systems via hierarchical reinforcement learning from human feedback,” in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 5116–5125.
- [43] M. Li, Z. Li, K. Yin, X. Nie, W. Zhang, K. Sui, and D. Pei, “Causal inference-based root cause analysis for online service systems with intervention recognition,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 3230–3240.
- [44] Z. Li, J. Chen, R. Jiao, N. Zhao, Z. Wang, S. Zhang, Y. Wu, L. Jiang, L. Yan, Z. Wang *et al.*, “Practical root cause localization for microservice systems via trace analysis,” in *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*. IEEE, 2021, pp. 1–10.
- [45] M. Barletta, M. Cinque, C. Di Martino, Z. T. Kalbarczyk, and R. K. Iyer, “Mutiny! how does kubernetes fail, and what can we do about it?” *arXiv preprint arXiv:2404.11169*, 2024.
- [46] J. Wang, Z. Li, M. Sun, B. Yuan, and J. C. Lui, “Iot anomaly detection via device interaction graph,” in *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2023, pp. 494–507.
- [47] C. Liu, Y. Liu, Z. Xu, and L. Dai, “Rapminer: A generic anomaly localization mechanism for cdn system with multi-dimensional kpis,” in *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2022, pp. 319–330.
- [48] L. Zhang, T. Jia, M. Jia, Y. Yang, Z. Wu, and Y. Li, “A survey of aiops for failure management in the era of large language models,” *arXiv preprint arXiv:2406.11213*, 2024.
- [49] Á. Brandón, M. Solé, A. Huélamo, D. Solans, M. S. Pérez, and V. Muntés-Mulero, “Graph-based root cause analysis for service-oriented and microservice architectures,” *Journal of Systems and Software*, vol. 159, p. 110432, 2020.
- [50] L. Wu, J. Tordsson, E. Elmroth, and O. Kao, “Microrca: Root cause localization of performance issues in microservices,” in *NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2020, pp. 1–9.
- [51] L. Wu, J. Bogatinovski, S. Nedelkoski, J. Tordsson, and O. Kao, “Performance diagnosis in cloud microservices using deep learning,” in *International Conference on Service-Oriented Computing*. Springer, 2020, pp. 85–96.
- [52] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, “Evaluating large language models trained on code,” *arXiv preprint arXiv:2107.03374*, 2021.
- [53] F. F. Xu, U. Alon, G. Neubig, and V. J. Hellendoorn, “A systematic evaluation of large language models of code,” in *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, 2022, pp. 1–10.
- [54] H. Joshi, J. C. Sanchez, S. Gulwani, V. Le, G. Verbruggen, and I. Radiček, “Repair is nearly generation: Multilingual program repair with llms,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 5131–5140.
- [55] M. Fu, C. Tantithamthavorn, T. Le, V. Nguyen, and D. Phung, “Vulrepair: a t5-based automated software vulnerability repair,” in *Proceedings of the 30th ACM joint european software engineering conference and symposium on the foundations of software engineering*, 2022, pp. 935–947.
- [56] Z. Wang, Z. Liu, Y. Zhang, A. Zhong, L. Fan, L. Wu, and Q. Wen, “Reagent: Cloud root cause analysis by autonomous agents with tool-augmented large language models,” *arXiv preprint arXiv:2310.16340*, 2023.
- [57] S. Shan, Y. Huo, Y. Su, Y. Li, D. Li, and Z. Zheng, “Face it yourselves: An llm-based two-stage strategy to localize configuration errors via logs,” *arXiv preprint arXiv:2404.00640*, 2024.
- [58] D. Roy, X. Zhang, R. Bhave, C. Bansal, P. Las-Casas, R. Fonseca, and S. Rajmohan, “Exploring llm-based agents for root cause analysis,” in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 2024, pp. 208–219.
- [59] Microsoft, “Welcome to graphrag,” <https://microsoft.github.io/graphrag/>, 2024, accessed: 2024-11-24.
- [60] Cypher, “Cypher manual for neo4j graph data platform,” <https://neo4j.com/docs/cypher-manual/current/introduction/>, 2024, accessed: 2024-11-27.
- [61] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [62] W. Xu, L. Huang, A. Fox, D. Patterson, and M. Jordan, “Largescale system problem detection by mining console logs,” in *Proceedings of SOSP*, vol. 9. Citeseer, 2009, pp. 1–17.
- [63] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, “Spark: Cluster computing with working sets,” in *2nd USENIX workshop on hot topics in cloud computing (HotCloud 10)*, 2010.
- [64] Spark, “Unified engine for large-scale data analytics,” <https://spark.apache.org/>, 2024, accessed: 2024-11-24.
- [65] A. Dave, A. Jindal, L. E. Li, R. Xin, J. Gonzalez, and M. Zaharia, “Graphframes: an integrated api for mixing graph and relational queries,” in *Proceedings of the fourth international workshop on graph data management experiences and systems*, 2016, pp. 1–8.
- [66] Spark, “Graphframes overview,” [https://graphframes.github.io/graphframes/docs/\\_site/index.html](https://graphframes.github.io/graphframes/docs/_site/index.html), 2024, accessed: 2024-11-24.
- [67] OpenAI, “Assistants api overview beta,” <https://platform.openai.com/docs/assistants/overview>, 2024, accessed: 2024-11-24.
- [68] Azure, “Quickstart: Get started using azure openai assistants (preview),” <https://learn.microsoft.com/en-us/azure/ai-services/openai/assistants-quickstart?tabs=command-line%2Ctypescript&pivots=programming-language-python>, 2024, accessed: 2024-11-24.
- [69] OpenAI, “Pricing: Simple and flexible. only pay for what you use.” <https://openai.com/api/pricing/>, 2024, accessed: 2024-11-24.