

CoE-Ops: Collaboration of LLM-based Experts for AIOps Question-Answering

Jinkun Zhao¹ Yuanshuai Wang¹ Xingjian Zhang¹ Ruibo Chen¹ Xingchuang Liao¹ Junle Wang¹
Lei Huang^{1,2,✉}, Kui Zhang^{1,✉}, Wenjun Wu^{1,2,✉}

¹SKLCCSE, Institute of Artificial Intelligence, Beihang University

²Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Beihang University

{huangleiai, zhangkui, wuj09315}@buaa.edu.cn

Abstract—With the rapid evolution of artificial intelligence, AIOps has emerged as a prominent paradigm in DevOps. Lots of work has been proposed to improve the performance of different AIOps phases. However, constrained by domain-specific knowledge, a single model can only handle the operation requirement of a specific task, such as log parser, root cause analysis. Meanwhile, combining multiple models can achieve more efficient results, which have been proved in both previous ensemble learning and the recent LLM training domain. Inspired by these works, to address the similar challenges in AIOps, this paper first proposes a collaboration-of-expert framework(CoE-Ops) incorporating a general-purpose large language model task classifier. A retrieval-augmented generation mechanism is introduced to improve the framework's capability in handling both Question-Answering tasks with high-level(Code, build, Test, etc.) and low-level(fault analysis, anomaly detection, etc.). Finally, the proposed method is implemented in the AIOps domain, and extensive experiments are conducted on the DevOps-EVAL dataset. Experimental results demonstrate that CoE-Ops achieves a 72% improvement in routing accuracy for high-level AIOps tasks compared to existing CoE methods, delivers up to 8% accuracy enhancement over single AIOps models in DevOps problem resolution, and outperforms larger-scale Mixture-of-Experts (MoE) models by up to 14% in accuracy.

Index Terms—Collaboration of Experts, DevOps, AIOps, Ensemble Learning, Retrieval-augmented Generation.

I. INTRODUCTION

DevOps is a software engineering methodology designed to bridge the gap between software development (Dev) and IT operations (Ops) [2]. The comprehensive DevOps lifecycle comprises eight iterative phases: Plan, Code, Build, Test, Deploy, Release, Monitor, and Operation. Each phase operates cyclically and encompasses specific subtask categories, as illustrated in Fig. 1. With advancements in artificial intelligence (AI) and deep learning, emerging paradigms such as MLOps and AIOps have been proposed, representing two distinct approaches to integrating AI with DevOps. Specifically, AIOps employs machine learning models to optimize DevOps workflows [11] [12] [13].

In recent years, the rapid emergence of large language models has spurred research exploring the integration of LLMs with DevOps. These studies primarily focus on leveraging

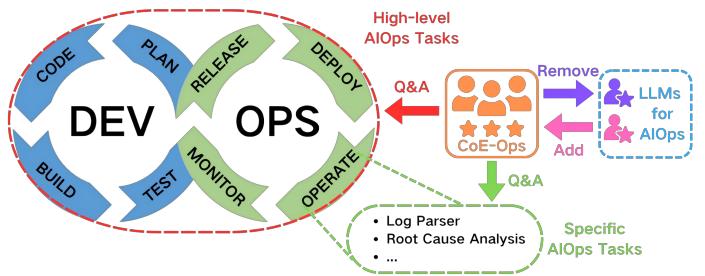


Fig. 1. Collaboration Scenarios of CoE-Ops Experts on Question-Answering Tasks at Various Levels within DevOps. CoE-Ops is capable of handling DevOps tasks across the entire life-cycle (high-level) and its sub-tasks (low-level), enabling flexible switching between AIOps experts.

LLMs to optimize DevOps workflows [17]. Within AIOps implementations that utilize LLMs for DevOps optimization, a critical challenge lies in selecting appropriate AIOps models for different AIOps tasks [1] [13] and enabling expert role-switching capabilities across heterogeneous workflows [3] [11] [19] which is not suitable for multi-agent system leveraging multiple AI agents with specialized and fixed roles [52]. Although domain-specific LLMs tailored for DevOps have emerged, current solutions face limitations due to their reliance on training data from specific domains [12] [15]. This results in inadequate coverage of all DevOps phases and their corresponding subtasks [17] [18], leading to deployment failures in unfamiliar scenarios [14] and representing a persistent bottleneck for AIOps advancements.

Based on the current limitations of AIOps, this paper formulates the following three research questions (RQs):

- **RQ1 (Effectiveness):** Can LLM ensembling mitigate the competency gaps between different LLMs?
- **RQ2 (Scalability):** How can LLM ensembling apply for multitask learning in the AIOps domain?
- **RQ3 (Efficiency):** Does the integration of smaller models via LLM ensembling enable performance that surpasses that of larger models?

To address the current challenges in AIOps regarding model selection [13], role-switching [19], and scalability [16], we propose the following solutions. First, we enhanced the existing Collaboration-of-Experts framework based on a two-stage expert routing mechanism [48] [49]. Subsequently, we inte-

✉ denotes corresponding author.

grated the ensemble learning concept and the refined framework into DevOps workflows, leveraging their inherent model-task scalability to enable dynamic selection, composition, and role-switching of AIOps experts (as illustrated in Fig. 1). Finally, to address high-level DevOps tasks, we incorporated the task classifier with retrieval-augmented generation. This combination enhances task classification by retrieving relevant contextual knowledge for target problems and integrating it into the prompt, thereby improving the framework's adaptability to complex operational scenarios.

Our key contributions are summarized as follows:

- A Collaboration-of-Expert framework CoE-Ops based on two-stage expert routing and a general-purpose large language model as task classifier, enabling dynamic switching across diverse AIOps task domains and LLM ensembles.
- An enhanced task classifier empowered by retrieval-augmented generation technology, specifically designed to address high-level task representations inherent in DevOps scenarios.
- Comprehensive empirical validation on DevOps-EVAL benchmarks with multiple task-expert configurations and over a dozen AIOps expert models, systematically validating CoE-Ops's dual scalability in task scalability and model scalability.

II. RELATED WORK

A. Development and Operations

DevOps is a collaborative, cross-domain software development methodology that emphasizes the automation of continuous delivery for software updates [3]. When integrated with artificial intelligence, its evolutionary trajectory bifurcates into two primary branches: MLOps and AIOps [16].

a) MLOps: MLOps focuses on applying DevOps practices to machine learning systems, aiming to establish seamless integration between diverse open source tools to enable fully automated execution of ML workflows, spanning dataset construction, model training, and deployment [4] [5]. With the recent emergence of large language models, LLMOps [7], an extension of MLOps tailored for LLM development and deployment, have gained momentum. LLMOps addresses the unique operational challenges of LLMs [8] and provides specialized tools for efficient data processing, model training, deployment, and maintenance [10]. However, both MLOps and LLMOps currently face limitations, including a lack of standardized practices, difficulties in maintaining model consistency and scalability [6] [9], and ambiguous evaluation criteria.

b) AIOps: In contrast, AIOps leverages AI and ML technologies to efficiently build and operate large-scale online services and applications in software engineering [11]. Most existing AIOps implementations rely on data from a limited number of domains [12] and predominantly employ supervised learning techniques [15]. Consequently, their proposed models are often confined to specific DevOps subdomains rather than being deployable across the entire ecosystem [17]. A critical

challenge for AIOps lies in selecting and integrating appropriate machine learning models [13] [19] to ensure adaptability to diverse use cases while fulfilling heterogeneous [18] and evolving requirements [16].

B. Ensemble Learning with Large Language Models

Ensemble learning with large language models involves the systematic utilization of multiple LLMs, each designed to handle user queries during downstream inference to capitalize on their individual strengths [20] [21]. Depending on the strategy for model integration, ensemble learning can be categorized into two paradigms: Mixture-of-Experts (MoE) and Collaboration-of-Experts (CoE).

a) Mixture-of-Experts: In recent years, MoE models have become a primary choice for foundation models [50] [51] due to their computational efficiency and strong generalization capabilities. In MoE systems, different expert modules possess distinct strengths, making efficient utilization a key challenge. FrugalGPT [22] and LLM-Blender [23] aggregate outputs from various experts to generate final results, while others adopt voting strategies to select the optimal output [24] [25] [26]. However, these expert modules cannot complete tasks independently, and the selection and generation processes lack interpretability. As a result, the Collaboration-of-Experts framework has increasingly drawn attention from researchers.

b) Collaboration-of-Experts: CoE primarily facilitates synergistic interactions among experts by selecting one or several optimal experts for a given input. Early efforts explored the use of sub-networks as expert models [27] [31]. With the proliferation of large-scale models, CoE has shifted focus toward incorporating diverse performance metrics, such as answer accuracy [29] [32], inference cost [28] [33] [34], and problem difficulty [30] [35]. A core research direction in CoE involves the design of routing algorithms for large models [29]. For instance, cascading networks [37] [40] have been proposed, or large models are represented as nodes [36] [39] or vector embeddings [41], with probabilistic methods [38] employed to predict routing outcomes. Recent studies further integrate reinforcement learning [42] [43] [44] [45] to refine expert routing strategies and introduce hardware-aware optimizations [46] [47] for efficient expert model loading. To address the lack of interpretability in routing decisions, a two-stage expert routing framework [48] [49] has been developed (as shown in Fig 2). This framework first categorizes input problems and then selects the most suitable expert for each category, thereby enhancing both the explainability of routing decisions and the scalability of the overall system.

III. PROBLEM FORMULATION

Before introducing the collaboration-of-experts paradigm into the AIOps domain, it is essential to delineate both the existing challenges within AIOps and the potential limitations this collaborative approach may encounter when addressing highly abstract AIOps tasks. The primary challenge in contemporary AIOps lies in effectively orchestrating diverse LLMs from distinct domains to address multifaceted operational

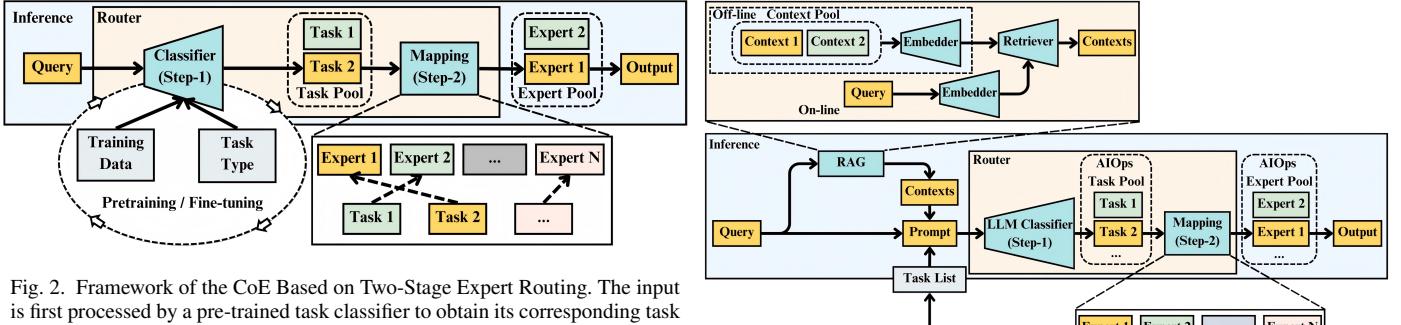


Fig. 2. Framework of the CoE Based on Two-Stage Expert Routing. The input is first processed by a pre-trained task classifier to obtain its corresponding task category label (Step-1). It is then routed to a designated expert model based on a pre-established "Task-Expert" mapping derived from existing benchmark (Step-2). Finally, it is the produce by the designated expert model.

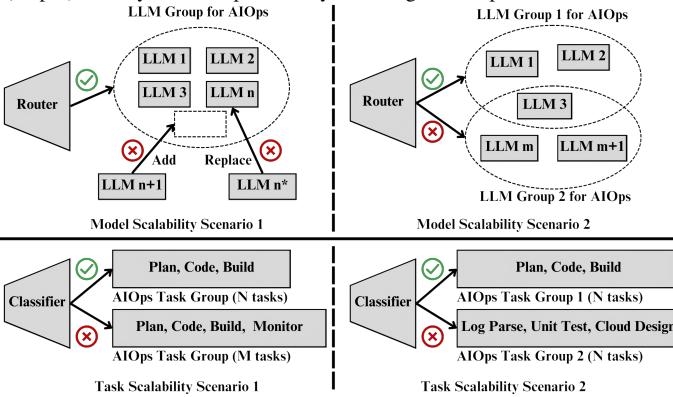


Fig. 3. AIOps Scenario Illustrating Model and Task Scalability. The end-to-end expert router is primarily constrained by model scalability. While the classifier in the two-stage expert routing approach improves model scalability, it is still constrained by task scalability.

requirements. For collaboration-of-experts frameworks, their scalability emerges as a critical concern within AIOps due to the field's broad spectrum of tasks and model heterogeneity. We categorize this scalability challenge along two dimensions: model scalability and task scalability. The corresponding operational scenarios for these dimensions are proposed in Fig. 3.

A. Model Scalability

For collaboration-of-experts with end-to-end expert routing, the scalability of model remains a critical challenge. As illustrated in Scenario 1, because the router directly employs LLMs as routing nodes, newly released AIOps LLMs cannot be dynamically incorporated into or replace old models in the router's LLM group. To address this issue, routers must undergo retraining whenever LLM group for AIOps are updated, incurring substantial computational overhead.

Furthermore, Scenario 2 demonstrates that when task contexts evolve, certain models in the existing group may become unsuitable as experts for emerging tasks. However, the router cannot transit to new expert groups tailored to the updated task requirements since it rigidly maps inputs to fixed AIOps experts. This necessitates costly retraining or fine-tuning of the router to adapt to new AIOps task scenarios.

With the emergence of collaboration-of-experts frameworks with two-stage expert routing such as Bench-CoE [49] and Composition of Experts [48], CoE can now dynamically

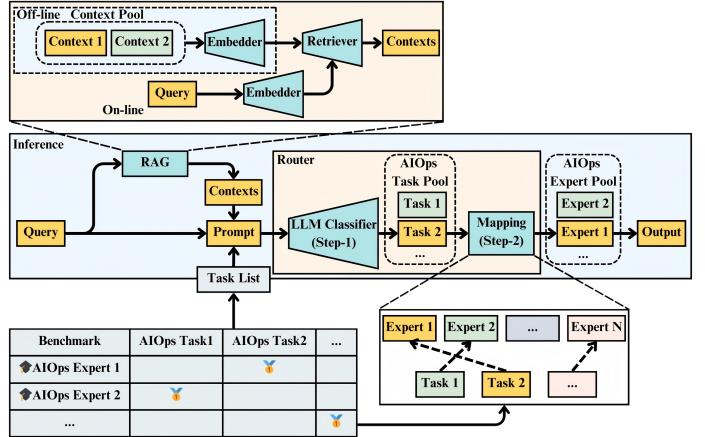


Fig. 4. Framework of CoE-Ops. CoE-Ops introduces improvements to Step-1 of the CoE based on two-stage expert routing. First, the discriminative model-based classifier is replaced with an LLM-based classifier. Subsequently, the prompt is enhanced by extracting a task list from benchmark datasets and employing Retrieval-Augmented Generation (RAG) technology to retrieve relevant context for the current input, thereby assisting the LLM-based classifier in classification.

reconfigure candidate AIOps LLMs through flexible mapping adjustments, thereby enhancing the scalability of model.

B. Task Scalability

While two-stage expert routers [48] [49] outperform their end-to-end counterparts in model scalability, they still exhibit significant limitations in task scalability. As demonstrated in Scenario 1, the output dimensionality of the two-stage expert routing remains fixed, since it employs discriminative models as classifiers. Consequently, when the number of classification tasks changes (from N to M), structural modifications and retraining are typically required.

Furthermore, Scenario 2 reveals that when task contexts evolve, the task classifier often fails to generalize to unseen tasks without retraining on in-domain data. This limitation stems from the classifier's reliance on parametric knowledge (memorized during training) rather than leveraging external knowledge sources, restricting its task scalability.

IV. METHODOLOGY

The framework of our proposed CoE-Ops is shown in Fig. 4. It consists of a two-stage expert routing mechanism which replaces discriminative models with general-purpose LLMs enhanced by retrieval-augmented generation capabilities.

A. Two-stage Expert Routing

CoE-Ops primarily improves upon the two-stage expert routing mechanism proposed in seminal works including Composition of Experts [48] and Bench-CoE [49]. During the original process of two-stage expert routing, the AIOps user's query is first classified by a pretrained or fine-tuned classifier to determine its task type. The query is then routed to the best-in-domain model for processing based on this label, as shown in Fig. 2.

The task classifier in the two-stage expert routing can be abstracted as (1) shows.

$$\hat{T} = \arg \max_{T \in \{T_1, T_2, \dots, T_n\}} P(T|X, \mathcal{C}), \quad (1)$$

where T represents the AIOps task, \mathcal{C} represents the classification model, and X denotes the current input from user.

In particular, within the two-stage expert routing architecture of the Collaboration of Experts, the cardinality of candidate AIOps experts should adhere to the bounds specified in (2), since each AIOps expert model demonstrates expertise in a minimum of one specialized AIOps domain.

$$2 \leq N_{\text{expert}} \leq N_{\text{task}}, \quad (2)$$

where N_{task} denotes the number of AIOps tasks.

Following AIOps task categorization by the classifier, input AIOps queries are dynamically routed to domain-specialized expert models through a "task-expert" allocation mechanism, as mathematically formalized in (3).

$$f : \mathcal{T} \rightarrow \mathcal{E}, \quad (3)$$

where $\mathcal{T} = \{T_1, T_2, \dots, T_M\}$ denotes the set of AIOps tasks, $\mathcal{E} = \{E_1, E_2, \dots, E_N\}$ denotes the set of AIOps experts, M indicates the count of AIOps tasks, and N indicates the count of AIOps experts.

When developing the "task-expert" allocation mechanism, it is necessary to establish a metric for evaluating the capability of each expert model across different task domains. For AIOps queries involving multiple-choice questions and question-and-answer formats, the answer accuracy of the expert model can serve as a suitable evaluation metric. This accuracy measurement, as shown in (4), provides a quantitative basis for assessing model performance.

$$\text{Accuracy}(M, T_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{I}(M(\mathcal{X}_{ij}) = \mathcal{A}_{ij}), \quad (4)$$

where N_i denotes the number of AIOps queries in the AIOps task T_i , M represents the expert model, \mathcal{X}_i stands for the AIOps queries in the AIOps task T_i , and \mathcal{A}_{ij} indicates the correct answer to the AIOps query \mathcal{X}_{ij} .

Upon construction of the capability assessment leaderboard, the expert model demonstrating superior accuracy within each task domain is designated as the optimal solution for the "task-expert" allocation, with formal validation provided in (5).

$$M_i^* = \arg \max_{M \in \mathcal{M}} \left(\frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{I}(M(\mathcal{X}_{ij}) = \mathcal{A}_{ij}) \right), \quad (5)$$

where M_i^* denotes the best AIOps model on task T_i .

B. Classifier with General-purpose LLM

Prompt 1 - Classifier with General-purpose LLM

You are a classifier that can categorize questions into specific tasks. Your job is to analyze the following given question and determine which task from the provided list it most likely belongs to.

The tasks are as follows: $\{\text{task list}\}$.

The question is:

- "{question}"
- A.{option_A}
- B.{option_B}
- C.{option_C}
- D.{option_D}".

Provide your answer in the format: "***Task: [selected task] ***".

To overcome the limitations inherent in conventional two-stage expert routing CoE frameworks, particularly their dependence on repeated classifier fine-tuning or retraining across distinct task scenarios, we implement a dual enhancement strategy. First, the classifier component is replaced by a general-purpose LLM operating in zero-shot mode, thereby eliminating fine-tuning requirements. Second, a structured task-list prompting mechanism (see Prompt 1) is integrated to ensure task scalability of the optimized architecture.

The enhanced framework enables dynamic adaptation to shifting task scenarios through prompt-based task list modification, eliminating the need for classifier pretraining or fine-tuning. This architectural innovation substantially reduces computational overhead while maintaining task scalability within the CoE paradigm.

The classification architecture of our framework, enhanced through the integration of prompt engineering and a general-purpose LLM, achieves formal abstraction as mathematically characterized in (6).

$$\hat{T} = \arg \max_{T \in \{T_1, T_2, \dots, T_n\}} P(T|X, P, \mathcal{L}_{\text{General}}), \quad (6)$$

where P denotes the prompt with the task list, $\mathcal{L}_{\text{General}}$ represents the general-purpose LLM.

Notably, unlike fine-tuned classifiers, using a general-purpose LLM as a classifier may yield an "unknown" class result. This reflects the LLM's effort to reduce hallucination by refusing to force-classify ambiguous inputs. Thus, after incorporating prompts and a general-purpose LLM, an additional "unknown" class is needed. Consequently, the number of output task classes is modified as shown in (7).

$$N_{\text{predict task}} = N_{\text{task}} + N_{\text{unk}}, \quad (7)$$

where N_{unk} denotes the number of tasks of unknown types (typically equals 1).

In this case, we need to select an extra expert model for the "unknown" class. Our selection strategy, as shown in (8), is to

choose the expert model with the highest average capability in all task domains to handle the "unknown" AIOps input.

$$M_{\text{unk}}^* = \arg \max_{M \in \mathcal{M}} \left(\frac{1}{N_{\text{total}}} \sum_{T_i \in \mathcal{T}} \sum_{j=1}^{N_i} \mathbb{I}(M(\mathcal{X}_{ij}) = \mathcal{A}_{ij}) \right), \quad (8)$$

where M_{unk}^* denotes the best AIOps model on "unknown" task.

Prompt 2 - AIOps Experts with Chain of thought

Please answer the following DEVOPS question.

The question is: $\{\text{question}\}$

The options are as follows:

- A. $\{\text{option}_A\}$
- B. $\{\text{option}_B\}$
- C. $\{\text{option}_C\}$
- D. $\{\text{option}_D\}$

Think step by step and then finish your answer with "the answer is (X)" where X is the correct letter choice.

For the expert models, we also avoid fine-tuning. Instead, we use prompts with chain of thought as the input. The prompt template is shown in Prompt 2. In the multiple-choice setting, to assess expert capabilities via answer accuracy, we ask the model to return answers in a fixed format.

C. LLM Classifier Enhanced with RAG

Simply replacing the classifier in the two-stage expert router with a general-purpose LLM carries risks. In AIOps domains with abstract or high-level task (like plan, build, code, etc.), the LLM may struggle to link inputs to tasks due to limited information. To address this, context needs to be introduced to help the LLM better understand the AIOps inputs, establish task-input connections, and improve AIOps task prediction.

In this condition, we integrated retrieval-augmented generation into the two-stage LLM routing. By retrieving similar questions and their categories to the input question, RAG aids the general-purpose LLM in determining the input's task category. This led to the improvement of the CoE-Ops framework in the scenarios with high-level AIOps tasks.

Similar to other RAG approaches, the RAG process in our CoE-Ops can be divided into two sub-phases: Off-line and On-line, as abstractly shown in (9).

$$P(o|q) = \sum_{c \in \mathcal{C}} P(a|q, c)P(c|q), \quad (9)$$

where q denotes the encoded vector of the query, c represents the encoded vector of the context, and o denotes the output of the LLM classifier.

During the Off-line stage, existing textual data is encoded, as shown in (10).

$$c = \text{Encoder}_{\text{RAG}}(C), \quad (10)$$

where C denotes the context data.

In the On-line stage, the input AIOps query is first encoded into a vector by the encoder, as shown in (11).

$$q = \text{Encoder}_{\text{RAG}}(Q), \quad (11)$$

where Q denotes the query data.

After obtaining the input AIOps query vector and knowledge base vectors, we perform retrieval to find the knowledge base vectors most similar to the input vector. The retrieval process is described by (12).

$$P(c|q) = \frac{\exp(\sin(q, c))}{\sum_{c \in \mathcal{C}} \exp(\sin(q, c))}. \quad (12)$$

The formula for the Retriever's similarity calculation is shown in (13).

$$\text{sim}(q, c) = q \cdot c. \quad (13)$$

After incorporating the RAG technique, we retrieve similar problems to the input question, using them as context in the prompt. The improved prompt is shown in Prompt 3.

Prompt 3 - Classifier with RAG

You are a classifier that can categorize questions into specific tasks. Your job is to analyze the following given question and determine which task from the provided list it most likely belongs to.

The tasks are as follows: $\{\text{task list}\}$.

The question is:

- $\{\text{question}\}$
- A. $\{\text{option}_A\}$
 - B. $\{\text{option}_B\}$
 - C. $\{\text{option}_C\}$
 - D. $\{\text{option}_D\}$

You can refer to the following examples of questions and their corresponding tasks to decide the current question's task: $\{\text{context}\}$

Provide your answer in the format: "***Task: [selected task] ***".

V. EXPERIMENT

A. Experimental Setup

To validate the effectiveness of our designed CoE-Ops in the complex domain of AIOps Question-Answering, we evaluated its performance using the DevOps-Eval¹ benchmark. DevOps-Eval is a comprehensive evaluation dataset specifically designed for large language models in the DevOps domain. This repository primarily contains a substantial collection of multiple-choice questions related to DevOps and AIOps, categorized into two subsets by language: DevOps-Eval English and DevOps-Eval Chinese. The DevOps-Eval English subset primarily covers low-level AIOps tasks, with its scope detailed in Table 1, while DevOps-Eval Chinese encompasses the comprehensive DevOps lifecycle, representing high-level AIOps tasks, as outlined in Tab. I.

¹<https://hf-mirror.com/datasets/codefuse-ai/CodeFuse-DevOps-Eval>

TABLE II
TASK-EXPERT MAPPING AND CLASSIFIER SETTINGS

Task Set A	Expert Set 1	Expert Set 2	Task Set B	Expert Set 3	Expert Set 4
Log Parser	Internlm-chat-7B ¹	Minstral-8b ²	Build	Internlm-chat-7b	Gemma-2-27b-it ²
Root Cause Analysis	CodeFuse-DevOps-Model-7B-Chat ¹	Minstral-8b	Code	Qwen2-7B-Instruct ¹	Doubaao-1.5-lite-32k ³
Time Series Anomaly Detection	CodeFuse-DevOps-Model-7B-Base ¹	Glm-4-flash ³	Deploy	Internlm-chat-7b	Doubaao-1.5-lite-32k
Time Series Classification	Internlm-7B ¹	Codegeex-4 ³	Monitor	Mathstral-7B-v0.1 ¹	Gemma-2-27b-it
Time Series Forecasting	Internlm-chat-7B	Minstral-8b	Operate	Qwen2-7B-Instruct	Gemma-2-27b-it
			Plan	Qwen2-7B-Instruct	Glm-4-flash ³
			Release	Mathstral-7B-v0.1	Gemma-2-27b-it
			Test	Qwen2-7B-Instruct	Doubaao-1.5-lite-32k

Classifier 1

DeepSeek-R1-Distill-Qwen-7B¹

Classifier 2

DeepSeek-V3²

[1]Deployed Locally

[2]Deployed through API, base url: <https://openrouter.ai/api/v1>

[3]Deployed through API, base url: <https://o3.fan/v1>

TABLE I
DATASET INFO OF DEVOPS-EVAL

DEVOPS-EVAL English ^a		DEVOPS-EVAL Chinese ^b	
Task	Sample	Task	Sample
LogParser	350	Build	218
RootCauseAnalysis	250	Code	1321
TimeSeriesAnomalyDetection	300	Deploy	255
TimeSeriesClassification	200	Monitor	216
TimeSeriesForecasting	320	Operate	2041
		Plan	66
		Release	212
		Test	228

^aCan be treated as "dataset with low-level tasks".

^bCan be treated as "dataset with high-level tasks".

To evaluate the performance of numerous expert models across diverse task domains, we established a comprehensive benchmark and constructed the "Task-Expert" mapping presented in Tab. II, where Task Set A represents a low-level AIOps task and Task Set B constitutes a high-level AIOps task. For Set 1 and Set 3 in Tab. II, we deploy the corresponding expert models locally for inference due to their moderate parameter size. Regarding Set 2 and Set 4 in Tab. II, the substantial parameter scale of these expert models precludes local deployment. We directly invoke these models via API interfaces provided by open-source platforms for inference, since our proposed CoE-Ops framework requires neither fine-tuning nor training of the models.

Notably, to verify that CoE-Ops framework possesses good task and expert extensibility, when switching among the four sets, we only modified the prompts and the "task-expert" mapping, without altering the model architecture or retraining and fine-tuning the models.

For experimental evaluation metrics, we employed numerical indicators including accuracy, precision, recall, and F1-score for classification and question-answering tasks to quantify the results. Additionally, we visualized the experimental outcomes using confusion matrix heatmaps and model capa-

bility radar charts.

B. RQ1: CoE-Ops Effectiveness Evaluation

To validate that our proposed CoE-Ops framework can balance capability disparities among models through ensemble learning across diverse model combinations, we applied CoE-Ops with different classifiers to expert collaborations (Expert Sets 1-4) on Task Set A and Task Set B from Tab. II. Specifically, we employed both the locally deployed DeepSeek-R1-Distill-Qwen-7B (Classifier 1) and the remotely accessed DeepSeek-V3 (Classifier 2) as task classifiers. For the RAG component, we employed the eval split from the DEVOPS-EVAL dataset as the context. The all-MiniLM-L6-v2 model was used as the encoder to encode both contexts and inputs into vector representations. Inputs were routed to corresponding AIOps experts within Expert Sets 1-4 based on the classification results.

We measured metrics such as answer accuracy for CoE-Ops and its utilized experts, and constructed capability radar charts for the models. The experimental results were subsequently organized and aggregated according to the Expert Sets. Specifically, results for Expert Set 1 are presented in Tab. III and Fig. 5, Expert Set 2 in Tab. IV and Fig. 6, Expert Set 3 in Tab. V and Fig. 7, and Expert Set 4 in Tab. VI and Fig. 8.

TABLE III
PERFORMANCE OF COE-OPS WITH EXPERT SET 1 ON DEVOPS-EVAL ENGLISH (TASK SET A)

Models	Acc(%)	Prec(%)	Rec(%)	F1(%)
Internlm-7B	35.07	37.05	35.07	34.36
Internlm-chat-7B	35.99	39.47	35.99	35.42
CodeFuse-7B-Base ^a	28.17	29.57	28.17	25.39
CodeFuse-7B-Chat ^b	30.56	31.71	30.56	30.36
CoE-Ops(Classifier 1)	40.07	42.40	40.07	39.6
CoE-Ops(Classifier 2)	44.08	46.82	44.08	43.58

^aModel's full name: CodeFuse-DevOps-Model-7B-Base.

^bModel's full name: CodeFuse-DevOps-Model-7B-Chat.

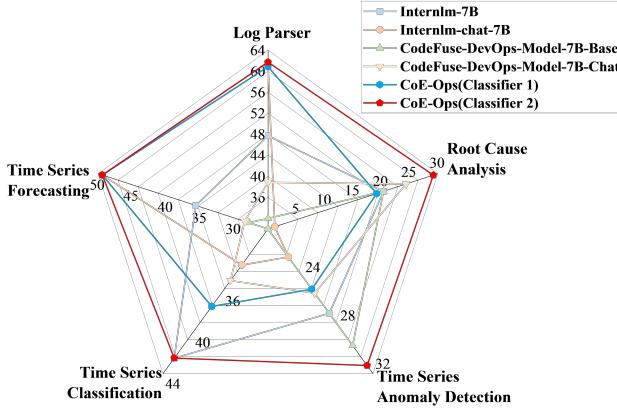


Fig. 5. Capability Radar Chart of CoE-Ops with Expert Set 1 on DevOps-EVAL English (TASK SET A)

As indicated in Tab. III, the CoE-Ops framework employing two classifiers demonstrates significant improvements over individual AIOps expert models across metrics including Accuracy, Precision, Recall, and F1-score. Specifically, Accuracy shows respective improvements of 4% and 8% compared to the best-performing standalone AIOps model. The effectiveness of the CoE-Ops framework is further validated in Fig. 5.

TABLE IV
PERFORMANCE OF COE-OPS WITH EXPERT SET 2 ON DEVOPS-EVAL ENGLISH (TASK SET A)

Models	Acc(%)	Prec(%)	Rec(%)	F1(%)
Glm-4-flash	62.54	64.50	62.54	63.16
Codegeex-4	54.44	63.84	54.44	58.65
Minstral-8b	68.38	69.07	68.38	68.70
CoE-Ops(Classifier 1)	69.15	71.13	69.15	70.10
CoE-Ops(Classifier 2)	70.49	72.29	70.49	71.31

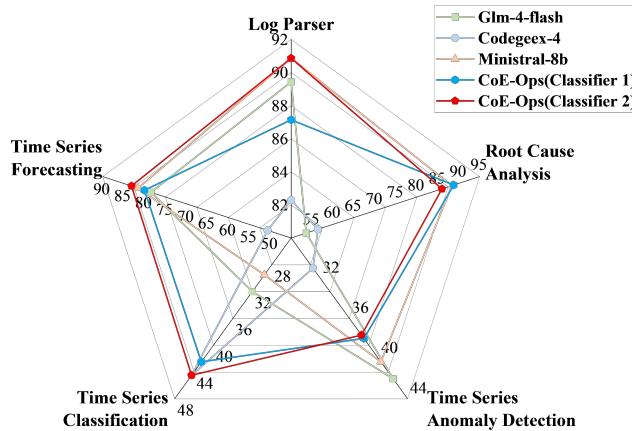


Fig. 6. Capability Radar Chart of CoE-Ops with Expert Set 2 on DevOps-EVAL English (TASK SET A)

As shown in Tab. IV, the CoE-Ops framework utilizing two classifiers achieves balanced capability enhancement across

varied expert configurations—both in quantity and type—on the same task. This demonstrates the scalability of our CoE-Ops framework with respect to model composition, as further evidenced in Fig. 5 and Fig. 6.

TABLE V
PERFORMANCE OF COE-OPS WITH EXPERT SET 3 ON DEVOPS-EVAL CHINESE (TASK SET B)

Models	Acc(%)	Prec(%)	Rec(%)	F1(%)
Internlm-chat-7b	54.2	53.63	54.20	53.56
Mathstral-7B-v0.1	62.74	62.77	62.74	62.47
Qwen2-7B-Instruct	63.57	64.44	63.57	63.32
CoE-Ops(Classifier 1)	64.52	64.93	64.52	64.24
CoE-Ops(Classifier 2)	64.14	64.44	64.14	63.86

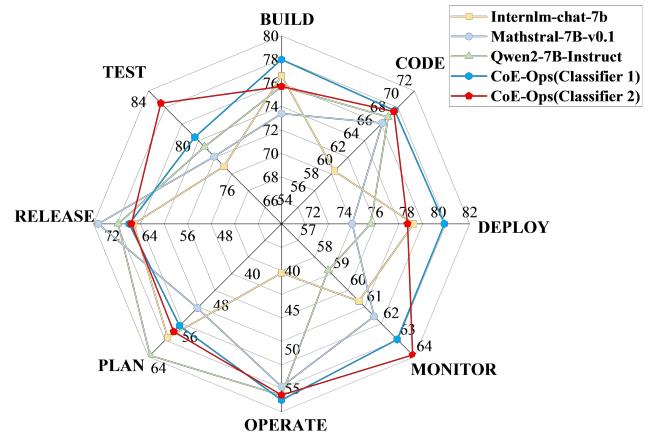


Fig. 7. Capability Radar Chart of CoE-Ops with Expert Set 3 on DevOps-EVAL Chinese (TASK SET B)

For high-level AIOps tasks such as Task Set B, despite their increased task classification difficulty, our CoE-Ops framework consistently outperforms individual AIOps expert models. This capability enhancement is evidenced by the analysis presented in Tab. V and Fig. 7.

TABLE VI
PERFORMANCE OF COE-OPS WITH EXPERT SET 4 ON DEVOPS-EVAL CHINESE (TASK SET B)

Models	Acc(%)	Prec(%)	Rec(%)	F1(%)
Doubao-1.5-lite-32k	73.21	73.73	73.21	73.47
Gemma-2-27b-it	74.22	74.13	74.22	74.14
Glm-4-flash	68.60	68.23	68.6	68.26
CoE-Ops(Classifier 1)	74.28	74.79	74.28	74.52
CoE-Ops(Classifier 2)	75.60	75.91	75.60	75.75

Similarly, by synthesizing results from Tab. V and Tab. VI, we observe that our CoE-Ops framework also exhibits model scalability on high-level AIOps tasks; it consistently enhances overall accuracy across model combinations involving both locally and remotely deployed models. This capability is further demonstrated in Fig. 7 and Fig. 8.

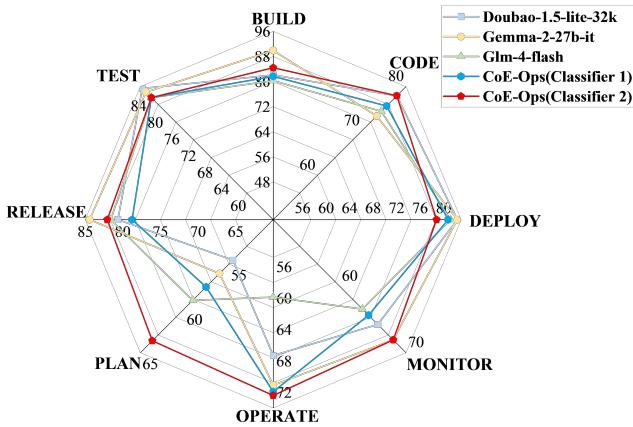


Fig. 8. Capability Radar Chart of CoE-Ops with Expert Set 4 on DevOps-EVAL Chinese (TASK SET B)

In summary, through comprehensive analysis of Accuracy, Recall, F1-Score, and model capability radar charts across diverse expert configurations on multiple AIOps tasks, we demonstrate the effectiveness of the CoE-Ops framework in balancing heterogeneous model capabilities while establishing its scalability across varying model compositions.

Answer to RQ1: Experimental results demonstrate that our proposed CoE-Ops framework effectively balances capability discrepancies among diverse models across various tasks and expert settings. This integration ultimately achieves an overall performance improvement of up to approximately 8%, confirming the effectiveness of our approach.

C. RQ2: Classifier Scalability Validation

Following the validation that our proposed CoE-Ops framework effectively balances capability disparities across different AIOps models, we conducted an ablation study on its core component, the Classifier, to assess its scalability for complex tasks in the AIOps domain. We evaluated two Classifiers employed by CoE-Ops (Classifier 1 and Classifier 2) on Task Set A and Task Set B, as detailed in Tab. II. Additionally, we tested the classification performance of a baseline Classifier without Retrieval-Augmented Generation enhancement. Task Set A and Task Set B differ in both the number of tasks and their hierarchical complexity. Testing on these two tasks thus allows coverage of the two Task Scalability Scenarios outlined in Section III.

We also evaluated the performance of the Bench-CoE framework, which utilizes a fine-tuned classifier, on both Task Set A and Task Set B in AIOps as a control. The general experimental results are presented in Tab. VII (for Task Set A) and Tab. VIII (for Task Set B).

Furthermore, to facilitate a more intuitive analysis of the classification performance of the two Classifiers employed by the CoE-Ops framework on individual tasks within Task Set A and Task Set B, we visualized their results using heatmaps.

TABLE VII
CLASSIFY PERFORMANCE ON DEVOPS-EVAL ENGLISH (TASK SET A)

Classifiers	Acc(%)	Prec(%)	Rec(%)	F1(%)
Random Select	20.00	-	-	-
Bench-CoE	62.46	52.69	62.46	55.35
Classifier 1 w/o RAG	77.11	87.66	77.11	81.52
Classifier 1	80.92	95.62	80.92	87.51
Classifier 2 w/o RAG	100	100	100	100
Classifier 2	100	100	100	100

TABLE VIII
CLASSIFY PERFORMANCE ON DEVOPS-EVAL CHINESE (TASK SET B)

Classifiers	Acc(%)	Prec(%)	Rec(%)	F1(%)
Random Select	12.5	-	-	-
Bench-CoE	4.94	11.86	4.94	0.83
Classifier 1 w/o RAG	13.91	32.65	13.91	14.66
Classifier 1	43.84	71.47	43.84	50.43
Classifier 2 w/o RAG	24.93	41.67	24.93	26.54
Classifier 2	77.22	79.79	77.22	77.22

The classification results for Task Set A are presented in Fig. 9 (Classifier 1) and Fig. 10 (Classifier 2), respectively. Similarly, the results for Task Set B are shown in Fig. 11 (Classifier 1) and Fig. 12 (Classifier 2).

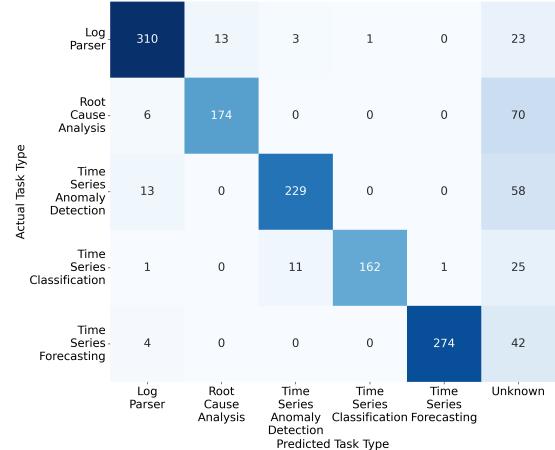


Fig. 9. Heatmap Visualization of Classifier 1's Confusion Matrix on DevOps-EVAL English (Task Set A)

Analysis of Tab. VII reveals that Classifier 1 and Classifier 2, implemented without fine-tuning or retraining, achieved strong classification performance in Task Set A. Their classification accuracy surpassed that of the Bench-CoE framework, which uses a fine-tuned classifier. In particular, Classifier 2 achieved the classification accuracy 100%, demonstrating its robust generalization capability. Furthermore, the classification accuracy of Classifier 1 showed a significant improvement after RAG integration. The heatmaps presented in Fig. 9 and

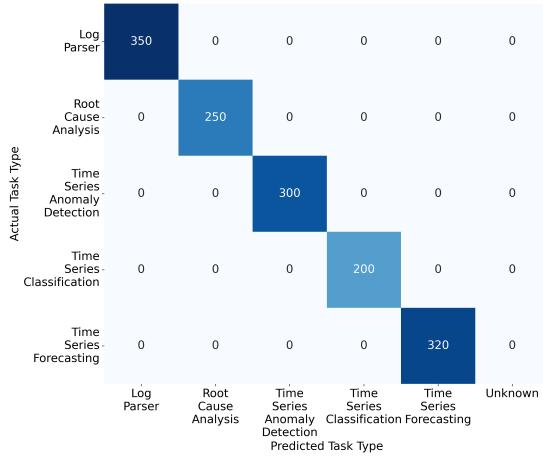


Fig. 10. Heatmap Visualization of Classifier 2’s Confusion Matrix on DevOps-EVAL English (Task Set A)

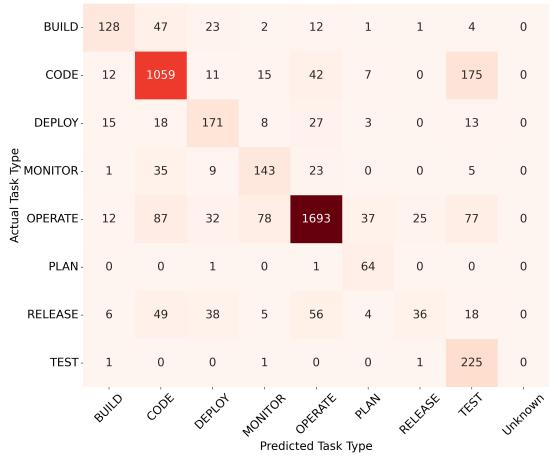


Fig. 12. Heatmap Visualization of Classifier 2’s Confusion Matrix on DevOps-EVAL Chinese (Task Set B)

Fig. 10 further validate the performance of both classifiers.

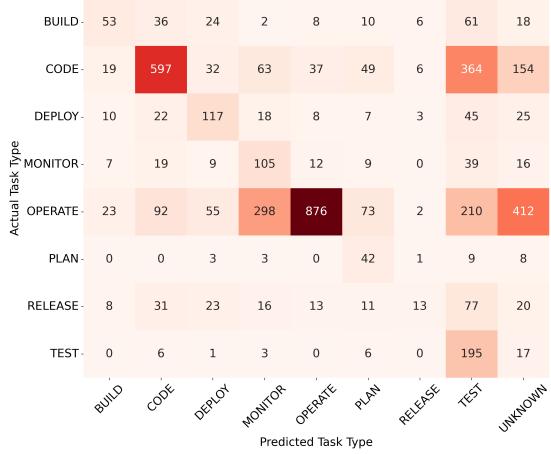


Fig. 11. Heatmap Visualization of Classifier 1’s Confusion Matrix on DevOps-EVAL Chinese (Task Set B)

A comparison of Tab. VII and Tab. VIII reveals that while the Bench-CoE framework, based on a fine-tuned classifier, demonstrates acceptable classification performance in the low-level AIOps task (Task Set A), its accuracy exhibits a marked degradation when the AIOps task scenario shifts to the high-level AIOps task (Task Set B). In contrast, although the performance of both Classifiers within our CoE-Ops framework also declined, their classification accuracy showed significant recovery, particularly for Classifier 2, upon augmentation with Retrieval-Augmented Generation technology. This robustly demonstrates the task scalability of our CoE-Ops framework within the complex AIOps task domain. This identical conclusion is further corroborated by the graphical evidence presented in Fig. 11 and Fig. 12.

Answer to RQ2: We design CoE-Ops, which employs a general-purpose large language model as the task classifier. This classifier is enhanced using prompting and Retrieval-Augmented Generation (RAG) techniques to adapt to the complex task scenarios in AIOps. We conduct classification experiments on both low-level and high-level tasks. The experimental results demonstrate that our CoE-Ops achieves significantly higher task classification accuracy compared to other ensemble learning methods in AIOps, showing improvements of 37.54% and 72.28%, respectively.

D. RQ3: Efficiency Validation

Following the validation of CoE-Ops’ effectiveness in balancing model capabilities and its classifier’s task scalability, we further compared CoE-Ops against other CoE and MoE models. Notably, the total parameter count of the mixtral-8x7b-instruct model reached approximately 56B, while the largest model deployed by our CoE-Ops utilized 27B parameters. We evaluated these models separately on Task Set A and Task Set B. Bench-CoE and Random-CoE (CoE with entirely random model routing) were tested on Task Set A as control groups, while Bench-CoE was not tested as a control group on Task Set B due to its poor classification performance. The experimental results are presented in Tab. IX and Tab. X, respectively, and are also visualized in the model capability radar charts shown in Fig. 13 and Fig. 14.

As indicated in Tab. IX, CoE-Ops demonstrates superior overall capability in the complex domain of AIOps compared to existing CoE and MoE models. Analysis combining Tab. IX and Tab. X reveals that CoE-Ops, leveraging an ensemble of smaller models, comprehensively surpasses large models such as mixtral-8x7b-instruct in terms of overall performance. This conclusion is further supported by the evidence presented in Fig. 13 and Fig. 14.

TABLE IX
PERFORMANCE OF COE AND MOE WITH EXPERT SET 2 ON DEVOPS-EVAL ENGLISH (TASK SET A)

Models	Acc(%)	Prec(%)	Rec(%)	F1(%)
Mixtral-8x7b-instruct	55.56	61.15	55.56	57.99
Random-CoE	59.15	62.63	59.15	60.84
Bench-CoE	68.94	70.30	68.94	69.58
CoE-Ops(Classifier 1)	69.15	71.13	69.15	70.10
CoE-Ops(Classifier 2)	70.49	72.29	70.49	71.31

TABLE X
PERFORMANCE OF COE AND MOE WITH EXPERT SET 4 ON DEVOPS-EVAL CHINESE (TASK SET B)

Models	Acc(%)	Prec(%)	Rec(%)	F1(%)
Mixtral-8x7b-instruct	65.26	66.89	65.26	65.94
CoE-Ops(Classifier 1)	74.28	74.79	74.28	74.52
CoE-Ops(Classifier 2)	75.60	75.91	75.60	75.75

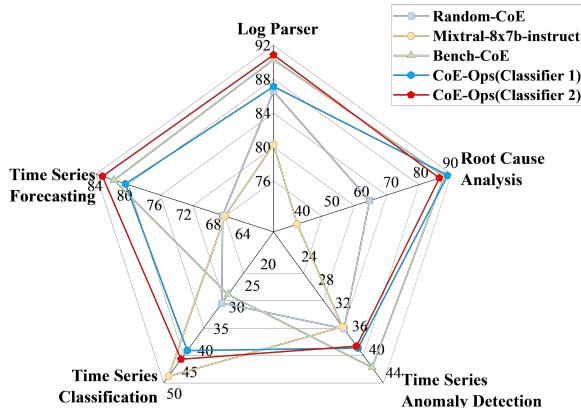


Fig. 13. Capability Radar Chart of Comparative Experiments on DevOps-EVAL English (Task Set A)

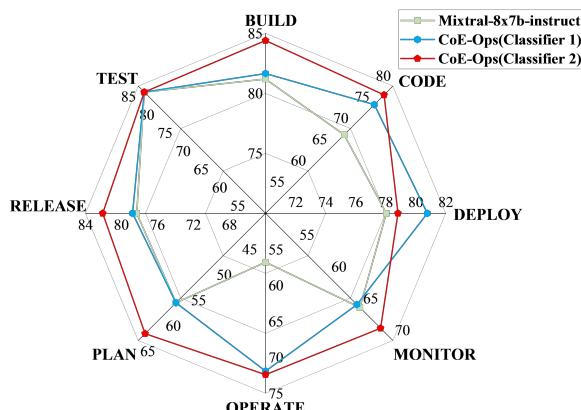


Fig. 14. Capability Radar Chart of Comparative Experiments on DevOps-EVAL Chinese (Task Set B)

Answer to RQ3: We conduct experiments comparing the performance of CoE-Ops integrated with small models against existing large models implemented via the MoE paradigm on the DEVOPS-EVAL dataset. The results experimentally demonstrate that, when appropriate small models are selected, CoE-Ops's integration of these small models achieves performance surpassing that of large models.

VI. THREATS TO VALIDITY

We acknowledge the following potential threats to the validity of our study and discuss our mitigation strategies:

a) Internal validity: Internal threats primarily center on the risks associated with large model API calls. To test as many large language models (LLMs) as possible, this study utilized both locally deployed models and API calls to access publicly available online models. However, this approach introduces risks such as invocation failure due to compromised API interfaces or credentials, or server crashes. To mitigate the internal threats arising from API call risks, we implemented additional program checkpoints during API invocation. When an API call fails—whether due to network connectivity issues, sensitivity of test data triggering content filters, or other causes—this mechanism allows us to resume the testing procedure from the checkpoint after troubleshooting the fault, thereby avoiding the need for complete retesting.

b) External validity: External threats primarily center on the specificity of task contexts. For the CoE framework, a significant risk lies in its limited extensibility across diverse task scenarios. Specifically, a CoE framework functioning effectively in one context may fail in others due to distributional shifts in training data. To address these external threats arising from task context specificity, our CoE-Ops framework leverages off-the-shelf general-purpose large models (without specialized training or fine-tuning) combined with advanced prompting techniques. This approach transcends the constraints of specific task contexts, enabling effective routing of expert models across both concrete and abstract domains.

c) Construct validity: Construct threats primarily center on hallucination issues introduced by the classification model. As our CoE-Ops framework employs a general-purpose large model—without specialized training or fine-tuning—as its classifier, it may exhibit hallucinations when processing high-level tasks. This presents a potential threat to the construct validity of our framework. To mitigate these construct threats, we employ Retrieval-Augmented Generation combined with prompt engineering to reduce hallucination in the classification model.

VII. CONCLUSION

To address the limitations of single AIOps expert models in mastering all DevOps domains and the challenges of ensemble learning in task switching within complex AIOps environments, this paper proposes CoE-Ops, a two-phase expert routing CoE framework based on a general large language model

classifier and Retrieval-Augmented Generation. By utilizing the general LLM classifier and prompts, CoE-Ops avoids the need for repeated training or fine-tuning during task scenario transitions, thereby enhancing its task scalability. Furthermore, the incorporation of RAG significantly strengthens its capability in handling tasks with highly abstract scenarios. In future work, we will explore the automated construction of AIOps expert capability rankings to achieve fully automated collaboration among AIOps experts. Additionally, we will integrate this framework with multi-agent systems to establish multi-tiered AIOps expert collaboration.

REFERENCES

- [1] Ebert, C., Gallardo, G., Hernantes, J. & Serrano, N. DevOps. *IEEE Software*. **33**, 94-100 (2016)
- [2] Jabbari, R., Ali, N., Petersen, K. & Tanveer, B. What is DevOps? A systematic mapping study on definitions and practices. *Proceedings Of The Scientific Workshop Proceedings Of XP2016*. pp. 1-11 (2016)
- [3] Leite, L., Rocha, C., Kon, F., Milojicic, D. & Meirelles, P. A survey of DevOps concepts and challenges. *ACM Computing Surveys (CSUR)*. **52**, 1-35 (2019)
- [4] Shah, P., Ahmad, N. & Beg, M. Towards MLOps: A DevOps Tools Recommender System for Machine Learning System. ArXiv Preprint ArXiv:2402.12867. (2024)
- [5] Kreuzberger, D., Kühl, N. & Hirschl, S. Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access*. **11** pp. 31866-31879 (2023)
- [6] Zarour, M., Alzabut, H. & Alsarayrah, K. MLOps best practices, challenges and maturity models: A systematic literature review. *Information And Software Technology*. pp. 107733 (2025)
- [7] Shan, R. & Shan, T. Enterprise LLMOps: Advancing Large Language Models Operations Practice. *2024 IEEE Cloud Summit*. pp. 143-148 (2024)
- [8] Diaz-De-Arcaya, J., López-De-Armentia, J., Miñón, R., Ojanguren, I. & Torre-Bastida, A. Large Language Model Operations (LLMops): Definition, Challenges, and Lifecycle Management. *2024 9th International Conference On Smart And Sustainable Technologies (SpliTech)*. pp. 1-4 (2024)
- [9] Tantithamthavorn, C., Palomba, F., Khomh, F. & Chua, J. MLOps, LLMops, FMOps, and Beyond. *IEEE Software*. **42**, 26-32 (2025)
- [10] Pahune, S. & Akhtar, Z. Transitioning from MLOps to LLMops: Navigating the Unique Challenges of Large Language Models. *Information*. **16**, 87 (2025)
- [11] Dang, Y., Lin, Q. & Huang, P. Aiops: real-world challenges and research innovations. *2019 IEEE/ACM 41st International Conference On Software Engineering: Companion Proceedings (ICSE-Companion)*. pp. 4-5 (2019)
- [12] Notaro, P., Cardoso, J. & Gerndt, M. A survey of aiops methods for failure management. *ACM Transactions On Intelligent Systems And Technology (TIST)*. **12**, 1-45 (2021)
- [13] Hua, Y. A systems approach to effective aiops implementation. (Massachusetts Institute of Technology, 2021)
- [14] Diaz-De-Arcaya, J., Torre-Bastida, A., Zárate, G., Miñón, R. & Almeida, A. A joint study of the challenges, opportunities, and roadmap of mlops and aiops: A systematic survey. *ACM Computing Surveys*. **56**, 1-30 (2023)
- [15] Mondru, A., Shreyas, R. & Anabathula, T. A Roadmap to Success: Strategies and Challenges in Adopting Aiops for it Operations. *International Journal Of Interpreting Enigma Engineers (IJIEE)*. **1** (2024)
- [16] Brahmamad, B. Beyond DevOps: The Evolution Toward Intelligent IT Operations with AIOps and MLOps. (2025)
- [17] Faraz Khan, A., Khan, A., Mohamed, A., Ali, H., Moolinti, S., Haroon, S., Tahir, U., Fazzini, M., Butt, A. & Anwar, A. LADs: Leveraging LLMs for AI-Driven DevOps. *ArXiv E-prints*. pp. arXiv-2502 (2025)
- [18] Krishnamurthy, D. & Neelanath, V. Establishing a Robust LLMops Framework for Intelligent Automation: Strategies and Best Practices. *2025 Emerging Technologies For Intelligent Systems (ETIS)*. pp. 1-5 (2025)
- [19] Mulongo, N. Key Performance Indicators of Artificial Intelligence For IT Operations (AIOPS). *2024 International Symposium On Networks, Computers And Communications (ISNCC)*. pp. 1-8 (2024)
- [20] Chen, Z., Li, J., Chen, P., Li, Z., Sun, K., Luo, Y., Mao, Q., Yang, D., Sun, H. & Yu, P. Harnessing Multiple Large Language Models: A Survey on LLM Ensemble. (2025), <https://arxiv.org/abs/2502.18036>
- [21] Varangot-Reille, C., Bouvard, C., Gourru, A., Ciancone, M., Schaeffer, M. & Jacquetin, F. Doing More with Less – Implementing Routing Strategies in Large Language Model-Based Systems: An Extended Survey. (2025), <https://arxiv.org/abs/2502.00409>
- [22] Chen, L., Zaharia, M. & Zou, J. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. (2023), <https://arxiv.org/abs/2305.05176>
- [23] Jiang, D., Ren, X. & Lin, B. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. (2023), <https://arxiv.org/abs/2306.02561>
- [24] Sukhbaatar, S., Golovneva, O., Sharma, V., Xu, H., Lin, X., Rozière, B., Kahn, J., Li, D., Wen-Yih, Weston, J. & Li, X. Branch-Train-Mix: Mixing Expert LLMs into a Mixture-of-Experts LLM. (2024), <https://arxiv.org/abs/2403.07816>
- [25] Si, C., Shi, W., Zhao, C., Zettlemoyer, L. & Boyd-Graber, J. Getting more out of mixture of language model reasoning experts. (2023), ArXiv Preprint ArXiv:2305.14628
- [26] Li, J., Zhang, Q., Yu, Y., Fu, Q. & Ye, D. More agents is all you need. (2024), ArXiv Preprint ArXiv:2402.05120
- [27] Zhang, Y., Chen, Z. & Zhong, Z. Collaboration of experts: Achieving 80% top-1 accuracy on imagenet with 100m flops. (2021), ArXiv Preprint ArXiv:2107.03815
- [28] Šakota, M., Peyrad, M. & West, R. Fly-swat or cannon? cost-effective language model choice via meta-modeling. *Proceedings Of The 17th ACM International Conference On Web Search And Data Mining*. pp. 606-615 (2024)
- [29] Shnitzer, T., Ou, A., Silva, M., Soule, K., Sun, Y., Solomon, J., Thompson, N. & Yurochkin, M. Large language model routing with benchmark datasets. (2023), ArXiv Preprint ArXiv:2309.15789
- [30] Ong, I., Almahairi, A., Wu, V., Chiang, W., Wu, T., Gonzalez, J., Kadous, M. & Stoica, I. Routellm: Learning to route llms with preference data, 2024. URL <Https://arxiv.Org/abs/2406.18665>.
- [31] Huang, S., Pan, J. & Zheng, H. CCoE: A Compact LLM with Collaboration of Experts. (2024), ArXiv Preprint ArXiv:2407.11686
- [32] Maurya, K., Srivatsa, K. & Kochmar, E. SelectLLM: Query-Aware Efficient Selection Algorithm for Large Language Models. (2024), ArXiv Preprint ArXiv:2408.08545
- [33] Stripelis, D., Hu, Z., Zhang, J., Xu, Z., Shah, A., Jin, H., Yao, Y., Avestimehr, S. & He, C. Polyrouter: A multi-llm querying system. ArXiv E-prints. pp. arXiv-2408(2024)
- [34] Stripelis, D., Hu, Z., Zhang, J., Xu, Z., Shah, A., Jin, H., Yao, Y., Avestimehr, S. & He, C. TensorOpera Router: A Multi-Model Router for Efficient LLM Inference. (2024), ArXiv Preprint ArXiv:2408.12320
- [35] Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Ruhle, V., Lakshmanan, L. & Awadallah, A. Hybrid llm: Cost-efficient and quality-aware query routing. (2024), ArXiv Preprint ArXiv:2404.14618
- [36] Guha, N., Chen, M., Chow, T., Khare, I. & Re, C. Smoothie: Label free language model routing. *Advances In Neural Information Processing Systems*. **37** pp. 127645-127672 (2024)
- [37] Hu, Q., Bieker, J., Li, X., Jiang, N., Keigwin, B., Ranganath, G., Keutzer, K. & Upadhyay, S. Routerbench: A benchmark for multi-llm routing system. (2024), ArXiv Preprint ArXiv:2403.12031
- [38] Zhang, T., Mehrdadfar, A., Dimitriadis, D. & Avestimehr, S. Leveraging uncertainty estimation for efficient llm routing. (2025), ArXiv Preprint ArXiv:2502.11021
- [39] Feng, T., Shen, Y. & You, J. Graphrouter: A graph-based router for llm selections. (2024), ArXiv Preprint ArXiv:2410.03834
- [40] Yue, Y., Zhang, G., Liu, B., Wan, G., Wang, K., Cheng, D. & Qi, Y. Masrouter: Learning to route llms for multi-agent systems. (2025), ArXiv Preprint ArXiv:2502.11133
- [41] Jitkrittum, W., Narasimhan, H., Rawat, A., Juneja, J., Wang, Z., Lee, C., Shenoy, P., Panigrahy, R., Menon, A. & Kumar, S. Universal Model Routing for Efficient LLM Inference. (2025), ArXiv Preprint ArXiv:2502.08773
- [42] Lu, K., Yuan, H., Lin, R., Lin, J., Yuan, Z., Zhou, C. & Zhou, J. Routing to the expert: Efficient reward-guided ensemble of large language models. (2023), ArXiv Preprint ArXiv:2311.08692.

- [43] Nguyen, Q., Hoang, D., Decugis, J., Manchanda, S., Chawla, N. & Doan, K. MetaLLM: A High-performant and Cost-efficient Dynamic Framework for Wrapping LLMs. (2024), ArXiv Preprint ArXiv:2407.10834.
- [44] Zhao, Z., Jin, S. & Mao, Z. Eagle: Efficient training-free router for multi-lm inference. (2024), ArXiv Preprint ArXiv:2409.15518.
- [45] Wang, X., Liu, Y., Cheng, W., Zhao, X., Chen, Z., Yu, W., Fu, Y. & Chen, H. Mixllm: Dynamic routing in mixed large language models. (2025), ArXiv Preprint ArXiv:2502.18482.
- [46] Prabhakar, R., Sivararamakrishnan, R., Gandhi, D., Du, Y., Wang, M., Song, X., Zhang, K., Gao, T., Wang, A., Li, X. & Others Sambanova sn40l: Scaling the ai memory wall with dataflow and composition of experts. *2024 57th IEEE/ACM International Symposium On Microarchitecture (MICRO)*. pp. 1353-1366 (2024)
- [47] Suo, J., Liao, X., Xiao, L., Ruan, L., Wang, J., Su, X. & Huo, Z. CoServe: Efficient Collaboration-of-Experts (CoE) Model Inference with Limited Memory. *Proceedings Of The 30th ACM International Conference On Architectural Support For Programming Languages And Operating Systems, Volume 2*. pp. 178-191 (2025)
- [48] Jain, S., Raju, R., Li, B., Csaki, Z., Li, J., Liang, K., Feng, G., Thakkar, U., Sampat, A., Prabhakar, R. & Others Composition of Experts: A Modular Compound AI System Leveraging Large Language Models. (2024), ArXiv Preprint ArXiv:2412.01868.
- [49] Wang, Y., Zhang, X., Zhao, J., Wen, S., Feng, P., Liao, S., Huang, L. & Wu, W. Bench-CoE: a Framework for Collaboration of Experts from Benchmark. ArXiv Preprint (2024), ArXiv:2412.04167.
- [50] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C. & Others Deepseek-v3 technical report. (2024), ArXiv Preprint ArXiv:2412.19437.
- [51] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C. & Others Qwen3 technical report. (2025), ArXiv Preprint ArXiv:2505.09388
- [52] Khan, U. & Kallinteris, N. Autonomous Multi-Agent LLMs in Agile Development: A Framework for AI-Driven Collaboration. (2025)