OpsEval: A Comprehensive Benchmark Suite for Evaluating Large Language Models' Capability in IT Operations Domain

Yuhe Liu
Tsinghua University & BNRist
Beijing, China

Changhua Pei
CNIC, CAS
Beijing, China

Longlong Xu
Bohan Chen
Tsinghua University & BNRist
Beijing, China

Mingze Sun
Tsinghua University & BNRist
Beijing, China

Zhirui Zhang
Beijing University of Posts and Telecommunications
Beijing, China

Yongqian Sun
Nankai University & TKL-SEHCI
Tianjin, China

Shenglin Zhang
Nankai University & HL-IT
Tianjin, China

Kun Wang
Tsinghua University & BNRist
Beijing, China

Haiming Zhang
Jianhui Li
Gaogang Xie
CNIC, CAS
Beijing, China

Xidao Wen
BizSeer
Beijing, China

Xiaohui Nie
CNIC, CAS
Beijing, China

Minghua Ma
Microsoft
Redmond, USA

Dan Pei*
Tsinghua University & BNRist
Beijing, China

Abstract

In recent decades, the field of software engineering has driven the rapid evolution of Information Technology (IT) systems, including advances in cloud computing, 5G networks, and financial information platforms. Ensuring the stability, reliability, and robustness of these complex IT systems has emerged as a critical challenge. Large language models (LLMs) that have exhibited remarkable capabilities in NLP-related tasks are showing great potential in AIOps, such as root cause analysis of failures, generation of operations and maintenance scripts, and summarizing of alert information. Unlike knowledge in general corpora, knowledge of Ops varies with the different IT systems, encompassing various private sub-domain knowledge, sensitive to prompt engineering due to various sub-domains, and containing numerous terminologies. Existing NLP-related benchmarks can not guide the selection of suitable LLMs for Ops (OpsLLM), and current metrics (e.g., BLEU, ROUGE) can not adequately reflect the question-answering (QA) effectiveness in the Ops domain. We propose a comprehensive benchmark suite, **OpsEval**, including an Ops-oriented evaluation dataset, an Ops evaluation benchmark, and a specially designed Ops QA evaluation method. Our dataset contains 7,334 multiple-choice questions and 1,736 QA questions. We have carefully selected and released 20% of the dataset written by domain experts in various sub-domains to assist current researchers in preliminary evaluations of OpsLLMs[1]. We test over 24 latest LLMs under various settings such as self-consistency, chain-of-thought, and in-context learning, revealing findings when applying LLMs to Ops. We also propose an evaluation method for QA in Ops, which has a coefficient of 0.9185 with human experts and is improved by 0.4471 and 1.366 compared to BLEU and ROUGE, respectively. Over the past one year, our dataset and leaderboard have been continuously updated.

CCS Concepts

• **Software and its engineering**; • **Computing methodologies** → **Artificial intelligence**;

Keywords

Large language models, Operations, Benchmark, Evaluation, Prompt engineering

ACM Reference Format:
Yuhe Liu, Changhua Pei, Longlong Xu, Bohan Chen, Mingze Sun, Zhirui Zhang, Yongqian Sun, Shenglin Zhang, Kun Wang, Haiming Zhang, Jianhui Li, Gaogang Xie, Xidao Wen, Xiaohui Nie, Minghua Ma, and Dan Pei. 2025. OpsEval: A Comprehensive Benchmark Suite for Evaluating Large Language Models' Capability in IT Operations Domain. In *33rd ACM International Conference on the Foundations of Software Engineering (FSE Companion '25), June 23–28, 2025, Trondheim, Norway.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3696630.3728572

1 Introduction

IT Operations (Ops) plays a pivotal role in ensuring efficient and reliable functioning of software systems, including cloud computing, 5G networks, and financial platforms. With the rapid expansion of the Internet, the scale and complexity of software systems have grown exponentially, making traditional operations increasingly challenging. To address these challenges, artificial intelligence-assisted operations have emerged as a transformative approach, often referred to as "AIOps" by Gartner [10]. AIOps uses artificial

*Dan Pei is the corresponding author.
[1]Data page is available at https://github.com/NetManAIOps/OpsEval-Datasets

FSE Companion '25, Trondheim, Norway

ACM ISBN 979-8-4007-1276-0/2025/06
https://doi.org/10.1145/3696630.3728572

intelligence to tackle critical software engineering tasks such as anomaly detection, fault diagnosis, and performance optimization.

In parallel, recent advances in large language models (LLMs) have further expanded the potential of intelligent solutions in software operations. The latest models, such as GPT-4o [16], GPT-4V [15], Meta-Llama-3 [2], and GLM-4 [30], have demonstrated exceptional generalization and task planning capabilities. As a result, these models have provided numerous opportunities to enhance downstream domain-specific applications. With its advanced text generation ability, LLM is well suited for Ops on tasks like question answering, information summarizing, and report analysis. Hereinafter, we refer to the LLM used for Ops as **OpsLLM**, regardless of whether they have been optimized specifically for Ops.

While there are benchmarks for assessing general-purpose NLP-related capabilities, no benchmark exists to evaluate the effectiveness of LLMs or OpsLLMs in Ops tasks. There is an urgent need for an Ops benchmark that informs us about the performance of current LLMs on Ops tasks. On the other hand, a good benchmark can significantly aid the optimization process of OpsLLMs tailored for the Ops domain. Nevertheless, due to the specialty of the Ops tasks, constructing an Ops benchmark presents the following challenges:

- **Sensitive data.** The Ops data is primarily sensitive and proprietary to companies, with very few publicly available data, making it difficult for any company to independently provide sufficient evaluation data to ensure confidence in the test results.
- **Sub-domains.** The Ops field spans many sub-domains, like 5G communications, cloud computing, and bank transactions, each requiring a mix of capabilities, or "tasks," such as network configuration or terminology explanation. The sheer number of sub-domains and tasks, combined with the absence of a systematic taxonomy, makes classifying questions challenging.
- **Prompt sensitivity.** Due to the relatively proprietary nature of the Ops, existing LLMs have not undergone specialized supervised fine-tuning (SFT) for instruct following within the Ops field, the evaluation results are more sensitive to prompt engineering. Designing appropriate prompts for robust and accurate evaluation is challenging.
- **QA metric.** Existing metrics like BLEU focus on linguistic similarity between model output and reference answers, which often fails to capture true performance in Ops tasks. In Ops, it's essential to assess whether the model's answers address key points in the reference and are supported by sufficient evidence, reflecting the precise meanings of domain-specific terms.

To address these issues, we propose **OpsEval**, a comprehensive benchmark suite for evaluating LLMs' capability in the IT Operations domain, focusing on tasks essential to maintaining and troubleshooting live systems in real-world production environments. First, to tackle the challenge of benchmark data mostly being private, we initiated an AIOps community, which has attracted dozens of companies to participate. We have selected 9 representative sub-domains from the community, allowing continuous data contributions from community members. We aggregate data under the same sub-domain to ensure robustness in evaluation. Additionally, we generated multi-choice (MC) and question-answering (QA) questions as supplements based on publicly available network management books. To address the challenge of classifying the numerous sub-domains and tasks in the Ops field, we employ model-based pre-clustering and manual review to annotate eight tasks and three abilities. Considering the prompt sensitivity of benchmark results, we systematically test model performance under self-consistency (SC), chain-of-thought (CoT), and few-shot in-context learning (ICL). Lastly, to address the inaccuracy of existing metrics in Ops QA evaluation, we design FAE-Score, which evaluates model responses based on fluency, accuracy, and evidence, with each criterion having its own dedicated assessment method.

The contributions of our paper are as follows:

(1) We introduce **OpsEval**, the first bilingual multi-task dataset in the Ops domain, covering 8 tasks and 3 abilities with 9,070 questions. To assist researchers in preliminary evaluating their OpsLLMs, we have carefully selected and released 20% of QAs from our benchmark licensed under CC-BY-NC-4.0, with the remaining 80% of undisclosed data preventing unfair evaluations due to data leakage [28].
(2) Based on the dataset, we introduce the OpsEval evaluation benchmark, conducting independent and robust evaluations with various prompting techniques and a specifically designed evaluation metric, FAE-Score. Compared to the commonly employed BLEU and ROUGE metrics, FAE-Score exhibits a more pronounced congruence with the evaluations of human experts. Specifically, FAE-Score attains a correlation coefficient 0.9175 with expert assessments, surpassing the coefficients of 0.6705 for BLEU and -0.3957 for ROUGE.
(3) Based on the results of OpsEval evaluation, we provide key observations and practical lessons to help domain practitioners make decisions such as whether existing models are sufficiently applicable within a specific sub-domain, the necessity for fine-tuning and whether model quantization compromises the effectiveness.

2 OpsEval Benchmark

Figure 1 shows the overall framework of OpsEval from construction to evaluation. We collected data from multiple sources and then preprocessed it to enhance its quality. Finally, we evaluated LLMs on the dataset using various prompt engineering techniques.

2.1 Data Collection

Our benchmark questions have been collected from various sources; we summarize them into four categories: company materials, certification exams and Ops textbooks. Each source is highly esteemed globally and reviewed by our Ops collaborators.

Company Materials. include production environment materials like Ops tickets and error logs , as well as internal documents and tests for Ops staff training. We have established cooperative relationships with 11 companies, covering various sectors like telecommunications, finance, and Ops service/tool providers, and received expert collaboration and Ops materials from them.

Table 2 shows the companies participating in the creation of OpsEval benchmark suite. Their industries include the Internet,

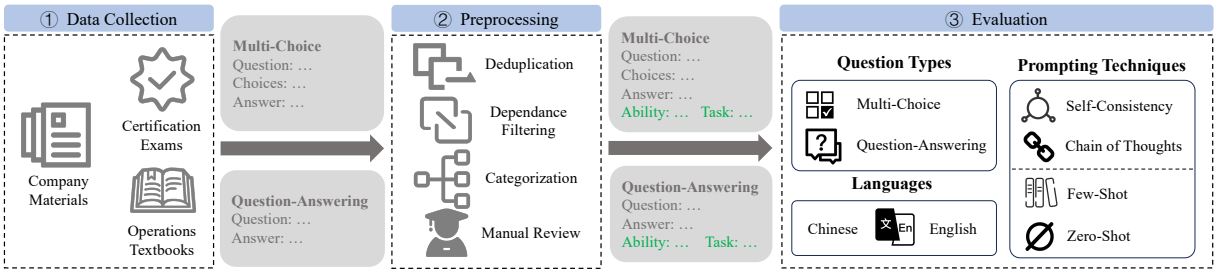

Figure 1: The framework of OpsEval.

Table 1: Overview of the question distribution in OpsEval by sub-domains, tasks and abilities.

(a) The number of questions in OpsEval, grouped by their sub-domains.

Sub-domain	Source	Type	Questions
Wired Network	Operation Textbooks	MC	3901
5G Communication	Certification Exams	MC	2615
		QA	1162
Oracle Database	Company Materials	MC	497
Log Analysis	Company Materials	QA	420
DevOps	Company Materials	QA	154
Private Cloud	Company Materials	QA	150
Securities Info.	Company Materials	MC	91
Hybrid Cloud	Company Materials	MC	40
Financial IT	Company Materials	MC	40
Total			9,070

(b) The distribution of different tasks and abilities of questions in OpsEval.

	Category	Percentage (%)
Task	Automation Scripts	3.3
	Monitoring and Alerting	5.2
	Performance Optimization	5.3
	Software Deployment	7.9
	Fault Analysis and Diagnostics	13.7
	Network Configuration	29.0
	General Ops Knowledge	20.2
	Miscellaneous	15.5
Ability	Knowledge Recall	49.8
	Analytical Thinking	39.9
	Practical Application	10.2

telecommunications, cloud computing, finance, and securities, and each company has dispatched at least two experts to participate in the OpsEval work.

Certification Exams. include knowledge assessments necessary for becoming an Ops staff and are naturally in the form of multiple-choice and question-answering questions. We obtained the relevant study guidebooks for these certification exams from public book websites and extracted sample questions from them as one of the sources for Ops questions.

Operations Textbooks. We first constructed a seeding keyword list for the Ops field and searched for related books. The textbooks contain relatively complete knowledge content, which can provide experts with materials for question creation, and some books themselves also include a certain number of exercises at the end of the chapters.

Table 2: Information of companies collaborating in OpsEval.

Organization	Domain	URL
BOSC	Financial IT	https://www.bosc.cn/zh/
Bizseer	Ops service provider	https://www.bizseer.com/
ChinaEtek	Internet	https://www.ce-service.com.cn/
Data Foundation	Internet	https://www.dfcdata.com.cn/
Guotai Junan	Securities	https://www.gtja.com/
Huawei	Communication	https://www.huawei.com/
Lenovo	Hybrid Cloud	https://www.lenovo.com/
Rizhiyi	Log Analysis	https://www.rizhiyi.com/
ZTE	Communication	https://www.zte.com.cn/china/
Zabbix	Ops service provider	https://www.zabbix.com/
Inspur	Ops service provider	https://www.inspur.com/
Total	11	

2.2 Preprocessing

We systematically carried out the preprocessing of our original data in the following stages:

Deduplication: Any repeated or highly similar questions are identified and removed to avoid redundancy in the test set. We calculate the cosine similarity of the question stems by bge-large-zh-v1.5 [29] to detect duplicate questions and identify pairs of questions with a similarity above a certain threshold (th=0.7).

Dependance Filtering: We have filtered out questions that rely on external images or document content to ensure the completeness of the question content itself. The filtering process was done by two parallel lists of empirical keywords in the question stems and the responses of GPT-3.5-turbo. The keyword list is listed below.

```
question_keywords = ['the figure', 'the scenario', 'the previous question']
fail_pred_keywords = ['unclear', 'scenario is not provided', 'cannot be determined', 'none of the options', 'none of the given options']
```

Question Categorization: We devise a categorization that captures many tasks that professionals confront in practical applications. The categorization process consists of two steps: automated screening and manual review. We first use GPT-4 for topic modeling to gain rough insights about the dataset and determine the relevance of each question to Ops, which resulted in more than 20 tasks but had an imbalanced distribution. We then involved dozens of experts during the manual review process to categorize the questions into eight tasks and three abilities.

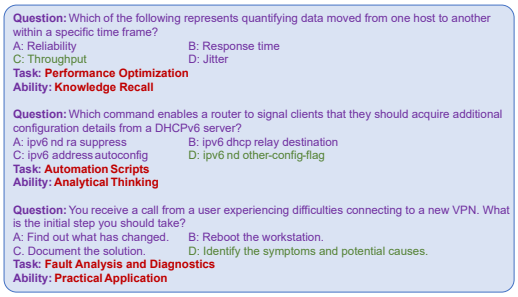

Figure 2: Three examples of the processed questions.

Tasks. The details of each task are as follows. *General Knowledge* pertains to foundational concepts and universal practices within the Ops domain. *Fault Analysis and Diagnostics* focuses on detecting and addressing discrepancies or faults within a network or system, and deducing the primary causes behind those disruptions. *Network Configuration* revolves around suggesting optimal configurations for network devices like routers, switches, and firewalls to ensure their efficient and secure operations. *Software Deployment* deals with the deployment and management of software applications throughout the network or system. *Monitoring and Alerts* harnesses monitoring tools to supervise network and system efficiency and implements alert mechanisms to notify administrators of emerging issues. *Performance Optimization* is centered on refining the network and system for peak performance and recognizing potential enhancement areas. *Automation Scripts* involves the formulation of automation scripts to facilitate processes and decrease manual intervention for administrators. *Miscellaneous* comprises tasks that do not strictly adhere to the aforementioned classifications or involve a combination of various tasks.

Abilities. Different questions require different levels of ability to answer. *Knowledge Recall* primarily test a model's capacity to recognize and recall core concepts and foundational knowledge, which are akin to situations where professionals identifies a standard procedure or recognize an issue based solely on previous knowledge. *Analytical thinking* necessitates a deeper level of thought, expecting the model to dissect a problem, correlate diverse pieces of information, and derive a coherent conclusion. It mirrors scenarios where professionals troubleshoot complex issues by leveraging their comprehensive understanding. *Practical Application* challenges a model to apply its knowledge or analytical conclusions to provide actionable recommendations for specific scenarios. It mirrors situations where professionals make decisions or suggest solutions based on in-depth analysis and expertise.

Figure 2 illustrates examples in our question set, demonstrating our classification methodology. The distribution of the questions across the tasks and ability levels is shown in Table 1b.

Manual Review: In the manual review step, we asked Ops experts from the industry to inspect the results of the previous three automated steps, including confirming duplicate and invalid questions and examining the classification results of GPT-4. In our work, an expert is defined as an individual with ten or more years of professional experience in their field, whether as an employee or a researcher. Experts were also asked to drop the questions unrelated to the Ops field. We split the dataset by n-folds and ensure each fold has at least two experts to review. The review process followed a standardized annotation guideline, which is available in our dataset repository. As listed in Table 1a, this quality enhancement process resulted in a refined test set of approximately 7,000 multi-choice and 2,000 question-answering questions.

Table 3: Models evaluated in this paper.

Model	#Parameters	Access	License
GPT-4/3.5-turbo	*undisclosed*	API	Proprietary
ERNIE-Bot-4.0	*undisclosed*	API	Proprietary
GLM4/GLM3-turbo	*undisclosed*	API	Proprietary
Meta-LLaMA-3	8B	Weights	Llama 3 Community
LLaMA-2	7/13/70B	Weights	Llama 2 Community
Qwen-Chat	7/14/72B	Weights	Qianwen LICENSE
Qwen1.5-Chat	14B	Weights	Qianwen LICENSE
InternLM2-Chat	7/20B	Weights	Apache-2.0
DevOps-Model	14B	Weights	Apache-2.0
Baichuan2-Chat	13B	Weights	Apache-2.0
ChatGLM3	6B	Weights	Apache-2.0
Mistral	7B	Weights	Apache-2.0
Gemma	2/7B	Weights	Gemma license
Claude-3-Opus	*undisclosed*	API	Proprietary
Qwen2-Instruct	7/72B	Weights	Qianwen LICENSE

2.3 Evaluation Settings

Multi-choice questions offer a structured approach with definitive answers. These questions are straightforward and provide a clear metric for assessment. We use **accuracy** as the metric. A choice-extracting function based on regular expressions is used to extract the predicted answer of LLMs. Then, we calculate the accuracy based on the extracted answer and the ground-truth labels.

Question-answering questions are evaluated using a metric designed specifically for OpsEval, called **FAE-Score**, which is explained in detail in the subsequent section. Additionally, we perform expert evaluations and calculate BLEU [17], ROUGE [12] and RAGAS [5] scores for comparison purposes, as reference to validate the accuracy of FAE-Score.

We use the same three criteria to evaluate the responses of various models for both FAE-Score and Expert Evaluation:

- **Fluency**. Assessment of the linguistic fluency in the model's output and compliance with the question-answering question's answering requirements.
- **Accuracy**. Evaluation of the precision and correctness of the model's output, including whether it adequately covers key points of the ground-truth answer.
- **Evidence**. Examine whether the model's output contains sufficient argumentation and evidential support to ensure the credibility and reliability of the answer.

In Expert Evaluation, we asked experts to score it between 0 and 3 for each criterion. During the scoring, the raw question, the detailed answer and its key points, and the output of an anonymous model are given at each iteration.

Prompting Techniques. We use various settings to evaluate LLMs on OpsEval to get a comprehensive overview of their performance. We evaluate LLMs in zero and few-shot (3-shot) settings. For each setting, we evaluate LLMs in four sub-settings of prompt engineering, that is, naive answers (Naive), self-consistency

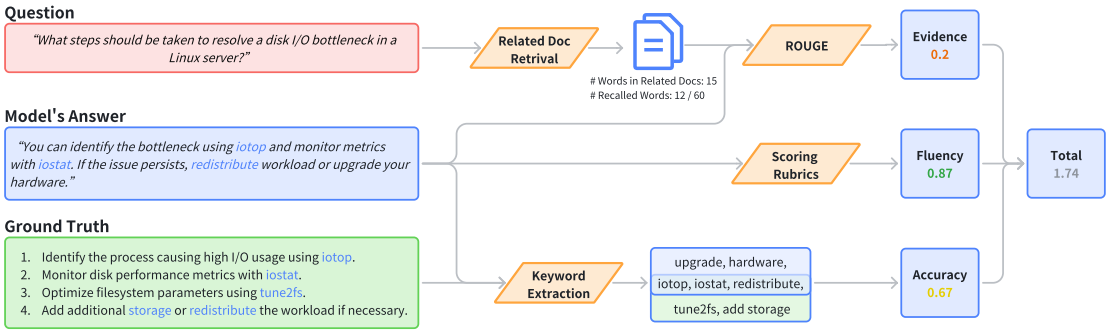

Figure 3: The FAE-Score pipeline.

(SC) [25], chain-of-thought (CoT) [27], self-consistency with chain-of-thought (CoT+SC). We set the number of queries in SC to **5**.

Models. We evaluate popular LLMs covering different weights from different organizations. The model selection was guided by specific criteria: We aimed to include the latest and most advanced large language models, with a particular focus on those capable of handling Chinese input. The detailed information of all 24 LLMs can be found in Table 3.

2.4 FAE-Score

Figure 3 shows the basic pipeline of our designed QA metric, FAE-Score. Here, we elaborate each evaluation methodology of each criterion.

Fluency. In Ops settings, the fluency of a model's output is crucial because the results are intended for human consumption by technical personnel. Unlike other generic benchmarks, the tasks in the Ops domain require clear and unambiguous communication, as the model's outputs may guide decision-making in production scenarios. To evaluate fluency in model outputs, we adapted the scoring rubrics methodology mentioned in Kim et al. [9]. We use Qwen2-72B-Instruct as the evaluation model, for its strong performance in general language generation [19] and its consistent multilingual capabilities. We assess the fluency of various model outputs based on grammar, coherence, clarity, appropriateness of style, and answer completeness, as shown in Figure 4.

Accuracy. Traditional metrics such as BLEU and ROUGE fall short in this vertical domain because they often fail to capture the key factual content within long-form responses. This results in inflated scores due to irrelevant word matches, making these metrics insufficient for accuracy evaluation in the highly specialized and knowledge-driven Ops context. To address these shortcomings, we take inspiration from Es et al. [5], using a keyword extraction method to evaluate the accuracy of model outputs. A judge model [14] is then employed to match the keywords from the model's response with the keywords from the standard answer. The final accuracy score is calculated by determining the F1-Score, which balances precision and recall for the matched keywords.

$$\textbf{Accuracy} = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{1}$$

1. **Grammatical Correctness (0-3 points)**:
• 0: Numerous grammatical errors that hinder comprehension.
• 1: Frequent errors that slightly disrupt the reading flow.
• 2: Minor grammatical errors, but the text remains easily readable.
• 3: Fluent and grammatically correct with no noticeable mistakes.
2. **Coherence and Consistency (0-3 points)**:
• 0: The output is disjointed, lacks logical flow, or contradicts itself.
• 1: Some inconsistencies or a lack of clear logical structure.
• 2: Mostly coherent, though minor clarity issues may be present.
• 3: The response is logically consistent and well-organized.
3. **Clarity of Expression (0-2 points)**:
• 0: The output is vague or ambiguous, making the response unclear.
• 1: Generally clear, though some areas may lack precision or clarity.
• 2: Clear, concise, and directly addresses the question or task.
4. **Style and Tone Appropriateness (0-2 points)**:
• 0: Inappropriate tone for the domain (e.g., overly casual or formal for the task).
• 1: Generally appropriate tone, but occasional mismatches with the task context.
• 2: Consistent tone that is well-suited to the operational context.
5. **Answer Completion (0-2 points)**:
• 0: The response is incomplete or significantly deviates from the expected format.
• 1: Response mostly follows the expected format but misses some details.
• 2: The response fully meets the structural and format requirements of the question.

Figure 4: Scoring rubrics for Fluency metric.

$$\mathrm{P} = \frac{\#\text{Matched Keywords}}{\#\text{Keywords in Model Output}} \tag{2}$$

$$\mathrm{R} = \frac{\#\text{Matched Keywords}}{\#\text{Keywords in Ground Truth}} \tag{3}$$

Evidence. Model responses must not only be accurate but also well-supported by relevant, authoritative information. To evaluate the evidence behind a model's response, we implement a ROUGE-based method to measure the overlap between the generated output and the content of related documents retrieved through similarity search. We used bge-large-zh [29] for document embedding and FAISS [4] for similarity search. By retrieving documents that closely match the question, we can assess whether the model's response appropriately references or aligns with this external information. We use ROUGE, as a recall-oriented metric, captures how much of the content in the relevant documents is reflected in the model's output. This ensures that the model does not simply generate plausible-sounding answers but grounds its responses in factual evidence from trusted sources.

$$\textbf{Evidence} = \mathrm{ROUGE}_{\mathrm{Recall}}(R, D) = \frac{\#\text{Overlapping Words}}{\#\text{Words in } D} \tag{4}$$

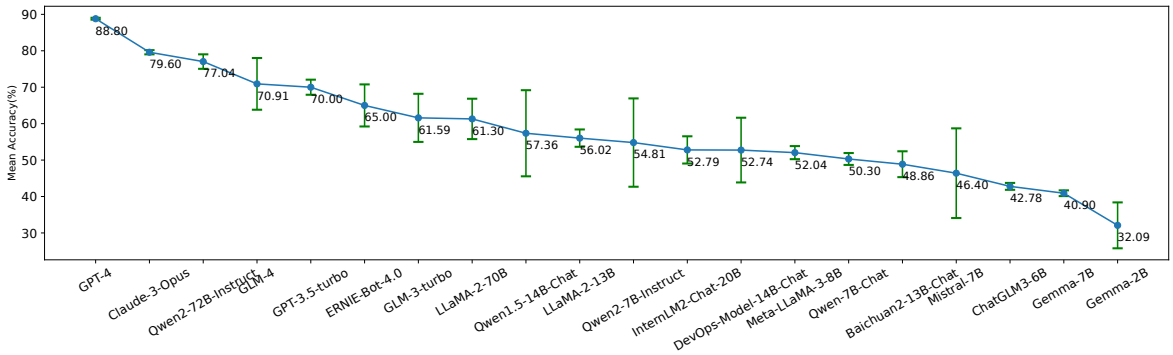

Figure 5: LLMs' overall performance on Wired Network Operations English test set (3-shot). Models are ranked based on their mean accuracy among different settings. The error bars represent the variance in the model's accuracy across different prompting techniques.

2.5 Open-source Policy

We released 20% of the OpsEval dataset to support research and community contributions. The subset was proportionally sampled across sources and sub-domains, with proprietary content reviewed to remove sensitive information. This sample dataset offers a clear view of question types and topics in the benchmark, helping researchers grasp the evaluation scope. It also supports local model evaluation for quicker iteration and can seed automatic QA generation [26], enriching Ops-specific data for future development. While this subset is available for users' self-evaluation, the complete dataset remains undisclosed. By ensuring that the test set answers are not leaked, we guarantee the reliability and non-leakage of the OpsEval benchmark.

3 Result Analysis

3.1 Overall Performance

The results of the few-shot evaluation with four settings on the Wired Network Operation test set are shown in Figure 5. [2] While closed source models like GPT-4 and Claude-3-Opus performs well on the OpsEval benchmark, open-sourced LLMs yield generally worse evaluation results than those in general domains like MMLU [7] and CEval [8]. This comparison highlights the necessity of explicitly fine-tuning OpsLLM for the Ops field. Recent open-sourced models like Qwen2-72B-Chat, exhibit competitive performance in multi-choice questions, thanks to their fine-tuning process and the quality of their training data. Furthermore, we observed significant variability in how different LLMs respond to various prompt engineering techniques. Given the critical importance of stability in the Ops domain, it is essential to consider a model's sensitivity to prompts when selecting foundation model. Further research into prompt engineering is needed to improve model performance and reliability in this domain.

Observations: 1) Few-shot and CoT can significantly increase performance if the model is tuned to adapt to these techniques, while SC may have little influence on highly consistent LLMs. 2) Smaller models with weaker abilities are less stable with advanced prompts. Simpler prompts work better for them.

Pratical Lesson: The choice of fundamental models should be a balance between their performance (average score) and robustness (variance) under different prompt settings.

3.2 Performance on Different Tasks and Abilities

To investigate how LLMs perform in each Ops sub-domain and each task, and to what extent they possess the general abilities, we summarize the result of different parameter-size groups of LLM in Figure 6. Regarding the eight tasks we tested, LLMs yield higher accuracy in General Knowledge tasks, while their performance drops and varies drastically in highly specialized tasks like Automation Scripts and Network Configuration, reflecting the impact of specialized corpus and domain knowledge on the performance of LLMs. By grouping LLMs by their parameter size, we find that while LLMs with 10B-30B parameters have higher accuracy in their best cases compared with LLMs with no more than 10B parameters, different 10B-20B LLMs' performance varies drastically. To provide systematic practical lessons for researchers in the operations domain on pre-training and fine-tuning OpsLLM, we have analyzed the error rates of LLMs across the 8 tasks and 3 abilities in Figure 7. By examining the focus areas across different categories, we have identified key research targets for capability training.

Observations: Among the 24 categories of results, models performed the worst in Analytical Thinking for Automation Scripts. This indicates that current models can only recall the learned scripts but struggle to infer their logical relationships. Similarly, Analytical Thinking showed the lowest performance across the three major tasks, indicating that current OpsLLM models still have some way to go before becoming foundational models for Ops Agents. Thus, researchers should focus on inference-related SFT (supervised fine-tuning) datasets.

Insights: 1) Among different sub-domains of Ops, 5G communication and database demand further pretraining and fine-tuning. 2) To be capable of an Ops agent, the foundation model must be able to make a connection between specialized domain knowledge.

[2] For results of the other sub-domains and settings, please check our official leaderboard website https://opseval.cstcloud.cn/content/leaderboard.

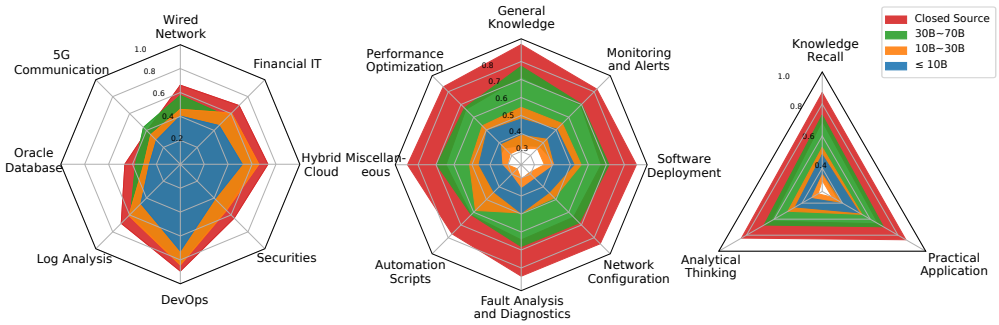

Figure 6: LLMs' performance on eight Ops sub-domains, eight tasks and three abilities. Each colored area presents the lower and upper bound of the corresponding parameter-size group.

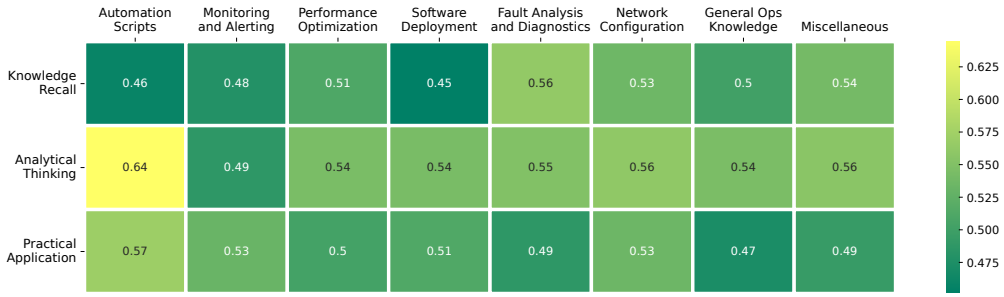

Figure 7: Heatmap of failure case distribution regarding tasks and abilities. The values represent the proportion of failure cases across all LLMs; yellower areas indicate higher failure rates.

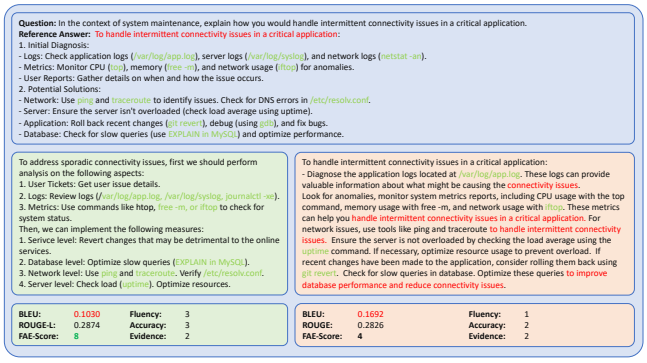

Figure 8: Case analysis on QA metrics.

3.3 Performance on Question-Answering

Table 4 presents the evaluation results of 200 question-answering English questions across four metrics: ROUGE, BLEU, RAGAS, FAE-Score, and Expert-Evaluation. To gain more insight into how different metrics perform in QA evaluation, we use Figure 8 as a case analysis. While BLEU and ROUGE are efficient in natural language comparison, they lack semantic information to determine which part of the context is more important than others. Knowing that a given benchmark evaluates QA based on BLEU/ROUGE, there is an obvious way to trick the metric: repeat patterns occurring in the question, gaining a higher possibility to match some patterns in the reference answer. Due to their lack of semantic information related to Ops and the potential hack, traditional metrics like BLEU are unsuitable for specialized benchmarks. Instead, with specialized prompting and seperately designed methodology for each criterion (Fluency, Accuracy and Evidence), FAE-Score can comprehensively evaluate models' QA performance, with the Accuracy metric picking up those important keywords and not be influenced by repeated words that contain no useful information, and the Evidence metric checking the recall of relevant supporting contents. In Section 4, we discuss the alignment between different metrics and expert evaluation, validating the effectiveness of FAE-Score in automated QA evaluation within the Ops domain.

Insight: In specialized domains, Ops specifically, traditional NLP metrics like BLEU and ROUGE cannot comprehend the key components in the reference answer, resulting their evaluation lacking practical significance. FAE-Score is suitable for large-scale qualitative evaluations in the Ops field.

3.4 Performance on Different Quantization parameters

We conducted experiments on different quantized versions of LLaMA-2-70B and obtained various results and conclusions. Figure 9 shows the accuracy of LLaMA-2-70B of different quantization parameters on English and Chinese questions. We do both zero-shot and few-shot evaluation with the naive setting.

Table 4: LLMs' performance on English network operations question-answering problems.

Model	ROUGE(%)	BLEU(%)	RAGAS(0-10)	Fluency		Accuracy		Evidence		FAE-Total	
				FAE	Expert	FAE	Expert	FAE	Expert	FAE	Expert
GPT-3.5-turbo	12.26	6.78	9.23	9.38	9.12	8.06	9.65	6.21	8.11	23.65	26.88
LLaMA-2-70B	7.74	4.2	6.04	8.69	8.25	7.71	8.79	9.08	8.98	25.48	26.02
LLaMA-2-13B	4.98	3.43	8.23	8.47	9.84	7.32	9.34	8.81	7.27	24.60	26.44
Chinese-Alpaca-2-13B	3.25	1.85	5.32	5.53	8.05	6.99	7.95	6.23	6.23	18.75	22.24
Baichuan-13B-Chat	4.76	0.35	7.93	7.16	7.98	8.71	7.84	6.66	7.31	22.53	23.13
Qwen-7B-Chat	11.82	4.33	4.92	7.63	5.82	6.42	7.27	6.57	5.37	20.62	18.47
ChatGLM2-6B	9.71	5.07	5.32	5.12	7.96	6.41	6.39	6.14	4.32	17.67	18.67
InternLM-7B	13.27	0.54	6.21	4.99	5.16	5.00	4.90	4.75	4.28	14.74	15.77
Chinese-LLaMA-2-13B	9.19	0.24	7.34	6.98	4.64	5.29	6.32	4.63	8.34	16.90	17.88

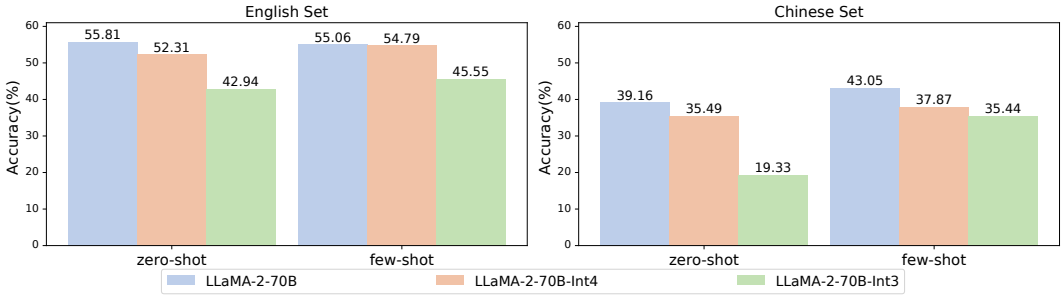

Figure 9: LLaMA-2-70B's performance of different quantization parameters. Both zero-shot and few-shot evaluations have been conducted on Wired Network Operations test set under the naive setting.

LLaMA2-70B-Int4 can achieve an accuracy close to LLaMA-2-70B without quantization. On English multi-choice questions, the accuracy of the GPTQ model with 4-bit quantization parameters is 3.50% lower in zero-shot evaluation and 0.27% in few-shot evaluation compared to LLaMA-2-70B. For Chinese questions, the accuracy of LLaMA2-70B-Int4 is 3.67% lower in zero-shot evaluation and 5.18% in few-shot evaluation compared to LLaMA-2-70B. However, LLaMA2-70B-Int3 has a performance degradation that cannot be ignored. On average, the accuracy of LLaMA2-70B-Int3 in English set has a 12.46% degradation compared to LLaMA-2-70B and a 9.30% degradation compared to LLaMA2-70B-Int4. Overall, although the performance of the INT4 version decreases in both English and Chinese, the decline does not exceed 10%. However, the performance drop in the INT3 version is more significant, requiring careful consideration in practical applications.

Practical Lesson: Quantization with more than 3 bits can effectively reduce computation and memory costs while preserving performance.

4 Validation

4.1 Benchmark Leakage Test

For the fairness of a benchmark suited for LLM, avoiding potential bias emerging from test set leakage is necessary. We adapted the methodology from [28] to perform a leakage test on OpsEval's dataset. We evaluate the LLM loss on samples from different datasets for several LLMs and calculate the average loss. For each dataset, we compare LLM loss on the test split (L_{test}) and a specially curated reference set (L_{ref}) generated by GPT-4, designed to mimic the testing dataset. While [28] only asked GPT-4 to generate similar questions to the GSM8K [3] dataset, we require GPT-4 to rewrite the question while preserving its original meaning. We define a key metric: $\Delta L = L_{test} - L_{ref}$, with a threshold of $\Delta L < 0$ indicating potential test data leakage. A negative ΔL suggests that the LLM's lower L_{test} comes from overfitting the test set rather than understanding the questions, indicating potential leakage. Figure 10 shows an example of how the metrics detect the data leakage. Table 5a shows the results of leakage measurement. In addition to the two standard evaluation benchmarks (CEval [8] and MMLU [7]), we conducted the same experiments on the alpaca dataset [23] and the

Table 5: Validation results.

(a) Measurement of potential test data leakage.

Dataset	L_{test}	L_{ref}	ΔL	≥ 0?
Alpaca	1.9940	2.3542	-0.3602	✗
Alpaca-GPT4	1.4988	1.7636	-0.3910	✗
CEval	2.5708	2.3099	0.2608	✓
MMLU	2.5475	2.1898	0.3577	✓
OpsEval	2.9854	2.6280	0.3050	✓

(b) Pearson correlation coefficients between Expert-Evaluation metrics and Automated metrics. Total is the sum of Fluency, Accuracy, and Evidence.

Metric	Total	Flu.	Acc.	Evi.
ROUGE	-0.44734	-0.49207	-0.40889	-0.31821
BLEU	0.47139	0.46369	0.55330	0.05977
RAGAS	0.57169	0.40029	0.51151	0.41928
FAE-Score	**0.91848**	**0.54757**	**0.81523**	**0.58160**

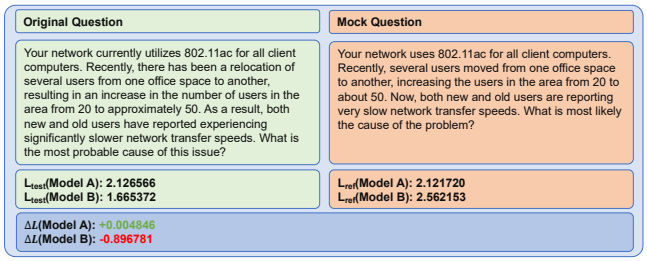

Figure 10: An example for leakage test.

Alpaca-GPT4 dataset [18], which is likely used in the pre-training of large models, using its ΔL as reference. This demonstrates the unbiased nature and non-leakage of the OpsEval test set.

4.2 Expert alignment of FAE-Score

Table 5b shows the correlation coefficients between various automated scoring metrics (ROUGE, BLEU, RAGAS, and FAE-Score) and Expert-Evaluation criteria. The results indicate that ROUGE and BLEU scores often misalign with Expert-Evaluation. This misalignment occurs because LLMs with poor performance may generate keywords that boost ROUGE and BLEU scores, while stronger LLMs might receive lower scores due to different wording from standard answers. While RAGAS [5] aligns better with experts than ROUGE and BLEU, there is still a gap between its scoring rankings for different models and expert judgement standards. In contrast, FAE-Score rankings closely match Expert-Evaluation, particularly with the Accuracy metric. This suggests that FAE-Score is more reliable in assessing the factual accuracy of LLMs' outputs. Notably, GPT-4's performance in factual accuracy is reflected in its strong alignment with the Accuracy metric.

5 Discussion

5.1 Automated QA generation

During the data collection process, we explored automating question-answer generation. Initially, we sampled QA pairs and manually evaluated their accuracy and domain relevance. Later, we utilized representative examples for few-shot learning, enabling GPT to generate and evaluate QA pairs automatically based on predefined criteria.

Recognizing that most existing benchmarks focus primarily on simple knowledge-based questions, we designed various task-specific templates to address this limitation. These templates require the model to complete specific fields within the template using the provided knowledge content, rather than generating entire questions and answers. This prompt engineering approach allows us to generate detailed and context-specific Ops tasks based on extensive operational knowledge while improving the model's instruction-following ability. By focusing on field-level completion, the overall structure of the QA remains consistent and accurate. Figure 11 shows the prompt template used for automatic QA generation, and Figure 12 illustrates some task cases. This approach ensures a more diverse and comprehensive evaluation of model capabilities while maintaining the relevance and quality of generated tasks.

```
You are an operations expert, and your task is to generate a question that adheres to the given template or follows a similar format, based on the provided knowledge content. The question template is as follows:
{question_template}
Here, {{}} represents parameters that need to be filled.

The knowledge points for generating the question are as follows:
{context}

When generating the question, you need to construct a business or operations scenario based on the knowledge content, and then use that scenario to populate the required fields. When creating the question, you need to adhere to the following constraints:
{constraints}

Return a JSON object containing the required fields from the template, as well as an answer field and an explanation field. The answer field should contain the answer to the question, and the explanation field should provide reasoning for the answer, explaining why it is correct. The terms used in your question and answer must match the given knowledge content, and you should not invent new terminology.

The reference format is as follows:
{json_example}

Your returned content should not start with ```json. Return the JSON object directly.
```

Figure 11: Prompt template for automated QA generation

Tool Selection

Task definition: The model should select and use the appropriate tool to solve user issues based on the provided operations scenario, tool descriptions, and user queries.

Template:
You can use various user-defined tools to solve the given user issues. Your task is to resolve the user's question based on the tool description.

Tool Description:
{tool_description}

User Question:
{user_question}

Business Information Reasoning

Task definition: The model needs to analyze the provided business information and solve user problems based on the given business knowledge and reasoning rules.

Template:
You are a core network operations engineer. Your task is to strictly follow the reasoning rules to provide an analysis result. Keep the reasoning process as concise as possible.

Input Information:
{input_info}
Business Knowledge:
{business_knowledge}

You need to answer in the specified format, filling in the content according to the reasoning rules. The response format should be in JSON and include two fields: reasoning process and analysis result.

Figure 12: Some automatically generated QAs, their task description, template and example question

5.2 Threats to Validity

The internal validity of this study is primarily influenced by the deployment parameters and prompts for the LLMs. Variations in these configurations may impact the evaluation results, and while we strive to follow best practices, some optimizations may not fully reflect real-world settings. The external validity is mainly limited by the datasets chosen. Our evaluation is based on specific datasets and Ops contexts, which may not generalize to other LLM deployment environments. In the future, we plan to expand OpsEval to include more datasets, scenarios, and deployment settings.

5.3 Future Work

Dataset Scale and Real-World Data. While privacy constraints limit real-world company data, our ongoing collaborations aim to expand the dataset with practical scenarios. Expanding the dataset with real-world scenarios remains a key focus, while the benchmark prioritizes robust evaluation over dataset scale. **Agent and RAG Introduction**: The inclusion of agents and Retrieval-Augmented Generation (RAG) techniques is constrained by the current large models' lack of foundational knowledge in operations. Our leaderboard will incorporate more complex tasks once open-source models possess sufficient operational capabilities. **More Balanced Distribution.** While the current sub-domain distribution in our work attempts to reflect the varying importance of different topics within the industry, we are actively cooperating with more collaborators to achieve a more balanced distribution.

6 Related Works

Table 6: A comparison of OpsEval with other popular datasets/benchmarks.

Dataset/Benchmark	Ops Domain	Open-sourced	Leaderboard
MMLU	✗	✓	✓
HELM	✗	✓	✓
BIG-bench	✗	✓	✓
SEAL	✓	✗	✓
CEval	✓	✓	✓
FLUE	✗	✓	✗
MultiMedQA	✗	✓	✗
CMB	✗	✓	✗
NetOps	✓	✗	✗
OWL	✓	✗	✗
OpsEval	✓	✓	✓

As LLMs evolve rapidly, their complex and varied capabilities are increasingly recognized. LLM specialized evaluation benchmarks can be divided into two categories: general ability benchmarks and domain-specific benchmarks.

General ability benchmarks assess the general abilities of LLMs across various tasks. These tasks evaluate LLMs' capacity for logical reasoning, general knowledge, common sense, and other similar abilities rather than being confined to a particular domain. MMLU [7] is a benchmark designed to measure knowledge acquired during pretraining by evaluating models exclusively in zero-shot and few-shot settings, covering 57 subjects across STEM. HELM [11] employs seven distinct metrics in 42 unique scenarios, offering a comprehensive evaluation of LLMs' capabilities across multiple dimensions. BIG-bench [22] comprises 204 tasks spanning a wide array of topics, with a particular focus on tasks deemed beyond the reach of current LLMs. SEAL [1] features private, expert evaluations of leading frontiers models. C-Eval [8] is a comprehensive Chinese evaluation suite designed to assess Chinese LLMs' advanced knowledge and reasoning abilities rigorously.

Domain-specific benchmarks evaluate the abilities of LLMs to handle tasks in specific fields. These benchmarks require LLMs to possess specialized knowledge in a specific domain and to respond in a manner consistent with the cognitive patterns of that field. Despite the rapid progression of LLMs in specialized domains, the evaluation metrics for these specific areas have received less attention. FLUE [20] is an open-source comprehensive suite of benchmarks, including new benchmarks across 5 NLP tasks in financial domain. MultiMedQA [21] is an extensive medical question-answering dataset, with questions derived from professional medical exams, research, and consultation records. CMB [24] includes multi-choice questions (CMB-Exam) and complex clinical questions based on real case studies (CMB-Clin). NetOps [13] focuses on evaluations in the network field, which is relevant to the field of Ops. NetOps includes multi-choice questions in both English and Chinese and a few question-answering questions. However, they only focus on wired network operations and while the dataset is released, they lack a benchmark that continuously updates the leaderboard. OWL [6] introduces Owl-Instruct and Owl-Bench datasets for IT operations, along with methods like HMCE for handling input length and a mixture-of-adapter for efficient tuning. However, it lacks a real-time updated leaderboard and does not provide a well-designed evaluation for IT operations QA tasks.

7 Conclusion

In this paper, we introduced **OpsEval**, the first comprehensive Ops benchmark suite designed for evaluating the performance of large language models (LLMs) in IT operations. We established a robust evaluation framework encompassing a wide range of sub-domains and tasks within Ops through rigorous data collection from multiple sources and meticulous preprocessing steps. Our benchmark includes a carefully selected set of 9,070 questions, which we have partially released to aid initial evaluations while protecting the integrity of the remaining dataset. It has undergone experiments in data leakage detection, ensuring its reliability. Our observations, supported by quantitative and qualitative results, highlight the need for a balanced approach to selecting fundamental models, considering both performance and robustness. During the QA evaluation, the FAE-Score emerges as a more reliable metric than traditional metrics, suggesting its potential as a replacement for manual labeling in large-scale quantitative evaluations. Our failure rate analysis across 8 tasks and 3 abilities provides researchers with crucial insights and prospects for future breakthroughs. The identified flexibility within the OpsEval framework presents opportunities for future exploration. This benchmark's adaptability facilitates the seamless integration of additional fine-grained tasks, providing a foundation for continued research and optimization of LLMs tailored for Ops.

Acknowledgments

This work was partially funded by the National Key Research and Development Program of China (No.2022YFB2901800), the National Natural Science Foundation of China (62202445, 62272249, 62302244), the Beijing National Research Center for Information Science and Technology (BNRist) key projects, the Fundamental Research Funds for the Central Universities (XXX-63253249), and the National Natural Science Foundation of China-Research Grants Council (RGC) Joint Research Scheme (62321166652).

The research was conducted and the paper was written when the author Mingze Sun was undergraduate student in Tsinghua University and author Xidao Wen was postdoc in Tsinghua University.

References

[1] Scale AI. 2024. SEAL Leaderboards. https://scale.com/leaderboard Accessed: 2024-06-03.

[2] AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[4] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[5] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Nikolaos Aletras and Orphee De Clercq (Eds.). Association for Computational Linguistics, St. Julians, Malta, 150–158. https://aclanthology.org/2024.eacl-demo.16

[6] Hongcheng Guo, Jian Yang, Jiaheng Liu, Liqun Yang, Linzheng Chai, Jiaqi Bai, Junran Peng, Xiaorong Hu, Chao Chen, Dongfeng Zhang, xu Shi, Tieqiao Zheng, liangfan zheng, Bo Zhang, Ke Xu, and Zhoujun Li. 2024. OWL: A Large Language Model for IT Operations. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=SZOQ9RKYJu

[7] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[8] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *arXiv e-prints* (2023), arXiv–2305.

[9] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. arXiv:2405.01535 [cs.CL]

[10] Andrew Lerner. 2017. AIOps Platforms—Gartner.

[11] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic Evaluation of Language Models. *arXiv e-prints* (2022), arXiv–2211.

[12] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[13] Yukai Miao, Yu Bai, Li Chen, Dan Li, Haifeng Sun, Xizheng Wang, Ziqiu Luo, Dapeng Sun, Xiuting Xu, Qi Zhang, Chao Xiang, and Xinchi Li. 2023. An Empirical Study of NetOps Capability of Pre-Trained Large Language Models. *CoRR* abs/2309.05557 (2023). https://doi.org/10.48550/arXiv.2309.05557

[14] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[15] OpenAI. 2023. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf

[16] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/ Accessed: 2024-06-01.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135

[18] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. *arXiv preprint arXiv:2304.03277* (2023).

[19] QwenLM. 2023. QwenLM/Qwen-7B. https://github.com/QwenLM/Qwen-7B

[20] Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2322–2335. doi:10.18653/v1/2022.emnlp-main.148

[21] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large Language Models Encode Clinical Knowledge. *arXiv preprint arXiv:2212.13138* (2022).

[22] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv e-prints* (2022), arXiv–2206.

[23] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

[24] Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023. CMB: A Comprehensive Medical Benchmark in Chinese. *arXiv e-prints* (2023), arXiv–2308.

[25] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL]

[26] Yizhong Wang, Yeganeh Kordi, and Swaroop et al. Mishra. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 13484–13508. doi:10.18653/v1/2023.acl-long.754

[27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL]

[28] Tianwen Wei and et.al. 2023. Skywork: A More Open Bilingual Foundation Model. arXiv:2310.19341 [cs.CL]

[29] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]

[30] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).

Received 2025-01-22; accepted 2025-03-25